

Statistical Methods for Psychologists, Volume 2:

Introduction to Linear Statistical Models

Douglas G. Bonett
University of California, Santa Cruz

2021

© All Rights Reserved

Version 1.9

Contents

Chapter 1 Statistical Models with One Predictor Variable	1
1.1 Introduction	1
1.2 Population Scatterplot	1
1.3 Population Pearson Correlation	2
1.4 Quantitative Factors	5
1.5 Population Linear of Best Fit	7
1.6 Interpreting a Slope Coefficient	8
1.7 Comparison of β_1 and ρ_{yx}	9
1.8 Linear Statistical Model with One Predictor Variable	11
1.9 Parameter Estimation	12
1.10 Inferential Statistical Methods	14
1.11 Confidence Interval for β_1	15
1.12 Confidence Interval for a Conditional Mean	16
1.13 Prediction Interval for a Predicted y Score	17
1.14 Confidence Interval for ρ_{yx}	17
1.15 Confidence Interval for Coefficient of Determination	18
1.16 Confidence Interval for τ_e	19
1.17 Dummy Coded Predictor Variable	20
1.18 Point-biserial Correlation	20
1.19 Directional Two-sided Hypothesis Tests	21
1.20 Interval Hypothesis Test	23
1.21 Monotonicity Test	24
1.22 Dichotomizing a Quantitative Predictor Variable	25
1.23 Psychometric Validity	25
1.24 Sample Size Requirements for Desired Precision	26
1.25 Sample Size Requirements for Desired Power	29
1.26 Specifying Planning Values	30
1.27 Assumptions	33
1.28 Outliers	34
1.29 Data Transformations	35
1.30 Extrapolation Issues	37
1.31 Measurement Error	37
1.32 Distribution-free Methods	40
1.33 Bayesian Approach to Data Analysis	42
1.34 Using Information from a Prior Study	45
1.35 Data Visualization	46
Key Terms	49
Concept Questions	50
Data Analysis Problems	52
Suggested Readings	54

Chapter 2 Statistical Models with Multiple Predictor Variables	55
2.1 The General Linear Model	55
2.2 Interpreting Slope Coefficients in Nonexperimental Designs	56
2.3 Confounding Variable Bias	58
2.4 Analysis of Variance Table for a GLM	59
2.5 Confidence Interval for a Slope Coefficient	62
2.6 Confidence Interval for τ_e	63
2.7 Semipartial Correlation	63
2.8 Partial Correlation	65
2.9 Standardized Slope Coefficient	67
2.10 Directional Two-sided Hypothesis Test	68
2.11 Interval Hypothesis Test	69
2.12 Quadratic Model	69
2.13 Interaction Effects in a GLM	71
2.14 Confidence Interval for Conditional Slopes	74
2.15 Detecting Interactions in Nonexperimental Designs	75
2.16 Dummy Coding and Effect Coding	76
2.17 Comparing Two Correlations	80
2.18 Equivalence Test	82
2.19 Analysis of Covariance Model	82
2.20 Multiple Group Regression Model	86
2.21 Nested Model Comparisons	89
2.22 Simultaneous Confidence Intervals and Hypothesis Tests	91
2.23 Confidence Interval for a Conditional Mean	92
2.24 Prediction Interval for a Predicted y Score	92
2.25 Exploratory Model Selection	93
2.26 Influential Observations	95
2.27 Multicollinearity	96
2.28 Sample Size Requirements for Desired Precision	97
2.29 Sample Size Requirements for Desired Power	99
2.30 Assumptions	101
2.31 Distribution-free Methods	102
2.32 Bootstrap Standard Error	103
2.33 Bootstrap Confidence Interval	104
2.34 Missing Data	105
2.35 Data Visualization	105
2.36 Bayesian Approach to Data Analysis	110
2.37 Using Information from a Prior Study	112
2.38 GLM in Matrix Form	113
Key Terms	116
Concept Questions	117
Data Analysis Problems	119
Suggested Readings	122

Chapter 3 Statistical Models for Repeated Measurements	123
3.1 Between-subjects and Within-subjects Predictor Variables	123
3.2 Covariance Pattern Models	123
3.3 Prediction Error Covariance Matrices	125
3.4 Centering Predictor Variables	128
3.5 Modeling the Repeated Measurements	131
3.6 Data Transformations for Monotonic Time Trends	135
3.7 Random Coefficient Models	136
3.8 Composite Prediction Error Structure	140
3.9 Parameter Estimation	143
3.10 Sample Size Planning	146
3.11 Assumptions	146
3.12 Data Visualization	147
3.13 Bayesian Approach to Data Analysis	154
3.14 Using Information from a Prior Study	155
Key Terms	156
Concept Questions	156
Data Analysis Problems	158
Suggested Readings	160
Appendix A. Tables	161
Appendix B. Matrix Notation and Operations	167
Appendix C. Glossary	177
Appendix D. Answers to Concept Questions	183
Appendix E. Answers to Data Analysis Questions	195

Preface

This book (Volume 2) is designed for a one-semester or one-quarter course that introduces the basic concepts of linear statistical models with applications in psychological research. The reader is assumed to be familiar with the material in Volume 1. In Volume 1, the focus was on experimental designs. This book explains how linear models can be used to analyze data from an experimental design, but the focus is more on nonexperimental designs. Together, Volumes 1 and 2 provide a first-year course for psychology graduate students.

Like Volume 1, this book emphasizes confidence intervals rather than null hypothesis significance testing (NHST). Directional two-sided tests, interval tests, and equivalence tests are presented as useful alternatives to NHST. And like Volume 1, random sampling from finite populations is assumed. With a finite population assumption, the population parameters have a clear and explicit interpretation, which can then be used to obtain clear and explicit confidence interval interpretations.

Chapter 1 provides an introduction to the basic concepts and applications of a linear statistical model for the special case of one predictor variable. This chapter introduces the distinction between fixed and random predictor variables and explains how the model parameters are interpreted in each of these two cases. The concept of a dummy coded variable is introduced, and it is shown how the two-group designs in Chapter 2 of Volume 1 can be represented by a linear model with a dummy coded predictor variable. Confidence intervals are described for several different population parameters including the slope coefficient, Pearson correlation, point-biserial correlation, and conditional mean. Simple methods are presented to determine the sample size required to obtain a confidence interval with desired precision or a hypothesis test with desired power.

Chapter 2 introduces the general linear model (GLM) which can accommodate two or more predictor variables. All of the factors in Volume 1 are qualitative. The GLM is "general" in that it can accommodate qualitative factors as well as quantitative factors and covariates. All of the between-subjects designs in Chapter 3 of Volume 2 can be represented by a GLM. Using a GLM, it is possible to extend the between-subjects designs of Volume 1 to include quantitative factors and covariates. In studies with two or more quantitative predictor variables, other types of correlations (partial correlation, semi-partial correlation, and multiple correlation) are defined. Sample size planning is more complicated for a GLM than the confidence intervals and hypothesis tests in Volume 1, but simple sample size methods are presented for several important types of analyses.

Two general types of statistical models for repeated measures designs – the covariance pattern model and the random coefficient model – are described in Chapter 3 and can be used to analyze any of the within-subjects designs in Chapter 4 of Volume 1. The two statistical models in Chapter 3 are general and can accommodate quantitative within-subjects factors and within-subjects covariates as well as between-subjects factors and between-subjects covariates.

Like Volume 1, SPSS and R are referred to throughout Volume 2. The `statpsych` R package can be used to analyze most of the recommended statistical analyses and sample size methods that are not available in SPSS.

This book is much shorter than other graduate level linear model books that have been written for psychology students – most of the currently used linear model books are more than 600 pages long. The goal here is to present the material in a brief and concise manner with the expectation that the instructor will provide additional examples, demonstrations in the use of SPSS and R, or more statistical theory. To improve readability, citations are not included in the text. Instead, relevant references are given in the `statpsych` documentation, and suggested readings are given at the end of each chapter.

Each chapter ends with a set of key terms, concept questions, and data analysis problems. Appendix C contains a glossary of all key terms, Appendix D contains answers to all concept questions, and Appendix E contains answers to all data analysis problems. Appendix B provides a brief summary of matrix notation and matrix operations.

Douglas G. Bonett
Santa Cruz, California

Chapter 1

Statistical Models with One Predictor Variable

1.1 Introduction

Chapters 2 and 3 of Volume 1 focused on statistical methods for one quantitative response variable and one or more *qualitative* independent variables. This chapter will present statistical methods for assessing the relation between one quantitative response variable and one other variable that can be either quantitative or qualitative. Chapter 2 presents statistical methods for assessing the relation between one quantitative response variable and two or more quantitative or qualitative variables.

1.2 Population Scatterplot

Suppose two quantitative measurements, a y score and an x score, could be assigned to every person in a study population. The pairs of y scores and x scores can be displayed in a two-dimensional graph called a *scatterplot*. Each point in a population scatterplot represents a single person in the study population. A population scatterplot for Math SAT scores and Reading SAT scores is illustrated in Figure 1.1 for a study population of 2,000 students.

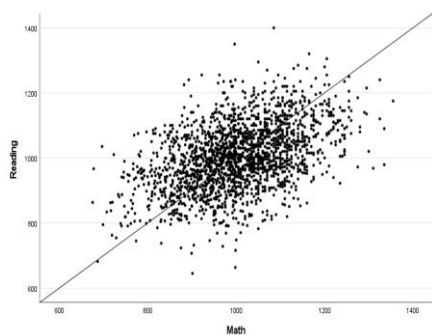


Figure 1.1 Population scatterplot of Math and Reading SAT scores

Figure 1.1 suggests that there is a positive relation between the Math SAT scores and the Reading SAT scores because students with above average Math SAT scores tend to also have above average Reading SAT scores, and students with below average Math SAT scores tend to also have below average Reading SAT scores.

The relation between Math and Reading SAT scores in Figure 1.1 appears to be approximately linear and could be summarized by a straight line. An example of a nonlinear relation is illustrated in Figure 1.2 for a study population of 1,500 high school students, where y is a self-esteem score and x is the number of hours per week spent using social media. A straight line does not provide a good description of the relation between self-esteem and hours of social media use.

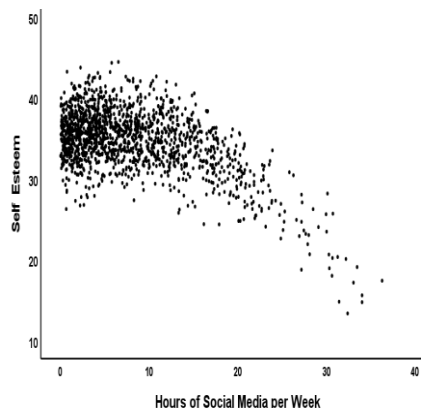


Figure 1.2 Population scatterplot of social media use and self-esteem

In some applications, one of the two quantitative variables could be viewed as a *predictor variable* and then the other quantitative variable would be the response variable. A predictor variable is any variable that is believed to be useful in predicting the value of a response variable. For example, hours of social media could be viewed as the predictor variable and self-esteem would be the response variable. Of course, some researchers could argue that self-esteem should be the predictor variable. It is traditional to call variable y the response variable and variable x the predictor variable. In a scatterplot, the vertical axis represents the response variable and the horizontal axis represents the predictor variable.

1.3 Population Pearson Correlation

A linear relation between two quantitative variables can be described numerically using a *Pearson correlation*. The Pearson correlation describes the degree to which two quantitative variables are linearly related. The Pearson correlation is not recommended if the relation is clearly nonlinear as in Figure 1.2. A method for linearizing certain types of nonlinear relations is described in section 1.28.

The population Pearson correlation is defined as

$$\rho_{yx} = \frac{\sigma_{yx}}{\sigma_y \sigma_x} \quad (1.1)$$

where σ_{yx} is the population *covariance* between the y scores and the x scores, σ_y is the population standard deviation of the y scores, and σ_x is the population standard deviation of the x scores. The population covariance is defined as

$$\sigma_{yx} = \sum_{i=1}^N \frac{(y_i - \mu_y)(x_i - \mu_x)}{N} \quad (1.2)$$

where μ_y is the population mean of the y scores and μ_x is the population mean of the x scores. Note that the value of $(y_i - \mu_y)(x_i - \mu_x)$ for person i will be positive if y_i and x_i are *both* above average or *both* below average. The value of $(y_i - \mu_y)(x_i - \mu_x)$ for person i will be negative if y_i is above average and x_i is below average or if y_i is below average and x_i is above average. The covariance is the mean of the $(y_i - \mu_y)(x_i - \mu_x)$ values. The covariance will have a positive value if the relation between x and y is positive (like Figure 1.1) and the covariance will have a negative value if the relation between x and y is negative.

The Pearson correlation has a possible range from -1 to 1. A ρ_{yx} value of -1 or 1 indicates a perfect linear association between x and y , and a ρ_{yx} value of 0 indicates no linear association between x and y . If the Pearson correlation equals 1 or -1, then all of the points in the scatterplot would fall on a straight line. Figure 1.3 gives several examples of population scatterplots of linear relations between x and y that have different Pearson correlation values. (*Note:* The APA Publication Manual states that correlations should be reported without a leading zero).

Note that the scatterplots for correlations in the .1 to .3 range are indistinguishable from a scatterplot for a correlation of 0. When the correlation is small, a participant could have a low, moderate, or high y score regardless of that person's x score. Alternatively, when the correlation is a large positive value, participants with high x scores tend to have high y scores and participants with low x scores tend to have low y scores. When the correlation is a large negative value, participants with high x scores tend to have low y scores and participants with low x scores tend to have high y scores.

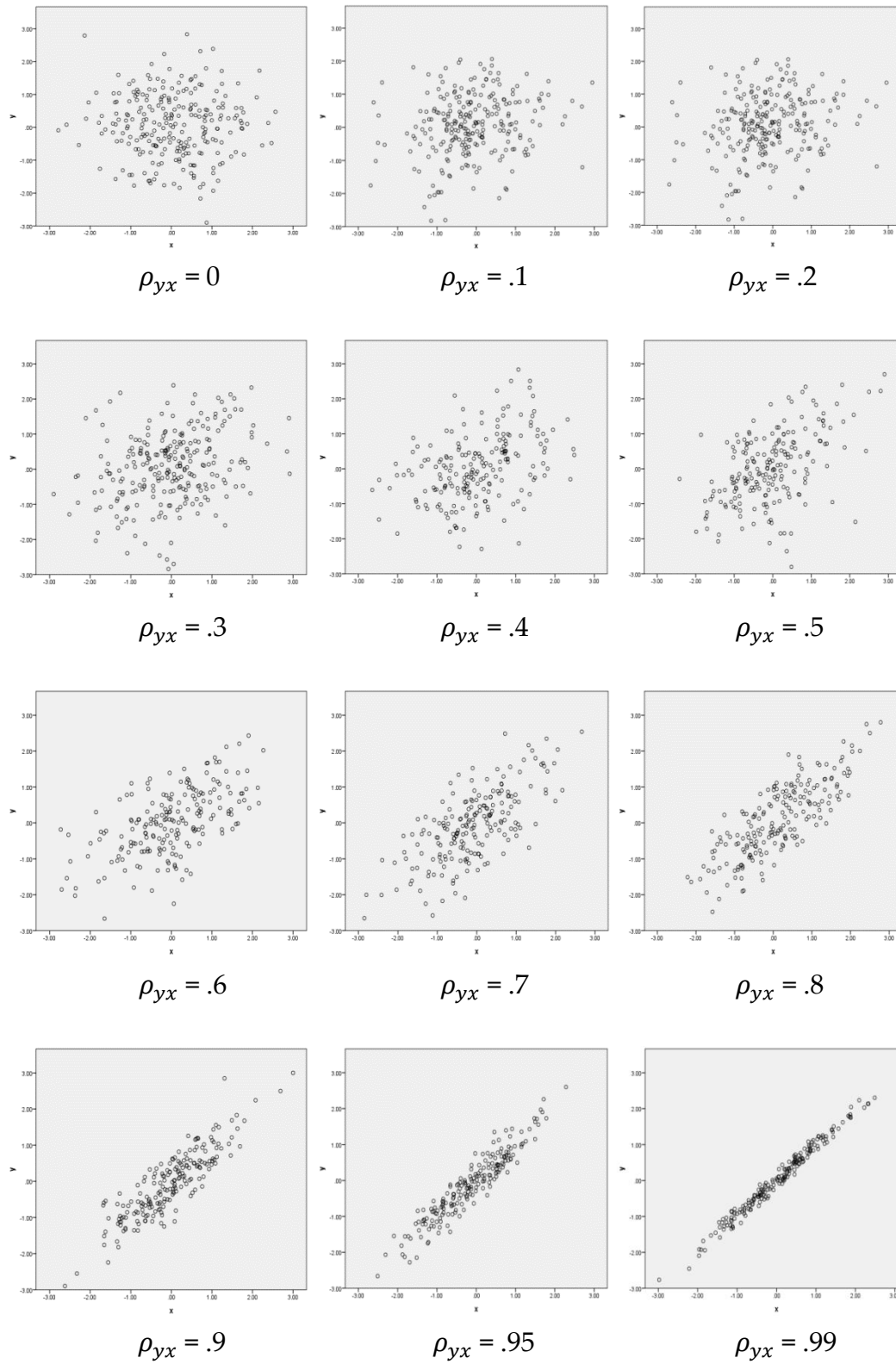


Figure 1.3. Scatterplots for nine Pearson correlation values

Pearson correlations of .1, .3, and .5 have traditionally been described as "small", "medium", and "large", respectively. The scatterplots in Figure 1.3 suggest that these descriptions are too generous. In Figure 1.3, we might describe a correlation of .3 as "small", a correlation of .5 as "medium", and a correlation of .8 as "large". Although it is tempting to describe a correlation of .3 as medium rather than small when reporting the results of a study, using exaggerated correlation interpretations is problematic in other situations. For example, when designing a study to detect a "medium" correlation, the required sample size is almost three times as large if a medium correlation is defined as .3 rather than .5. In other applications, the goal is to show that a correlation is "small" (see section 1.20), and the sample size will be almost ten times as large if a small correlation is defined as .1 rather than .3.

1.4 Quantitative Factors

A "factor" is a predictor variable with a specified (and usually small) number of levels. Recall from Chapter 3 of Volume 1 that a factor can be either a treatment factor or a classification factor. Only *qualitative* factors were considered in Volume 1. An example of a qualitative classification factor is "political affiliation" with levels: Democrat, Republican, and Independent. An example of a qualitative treatment factor is "type of psychotherapy" with levels: Individual, Group, and Teletherapy to which participants are randomly assigned.

The levels of a *quantitative* factor are interval scale or ratio scale values of the predictor variable. A quantitative factor can be a classification factor or a treatment factor. An example of a quantitative treatment factor is hours of training with levels 5, 10, and 15 hours, and an example of a quantitative classification factor is the number of living grandparents with levels 0, 1, 2, 3 and 4.

Each level of a qualitative or quantitative classification factor defines a subpopulation, and the mean of the response variable within each subpopulation is often an important parameter to examine. We will assume that a plot of the subpopulation means can be approximated by a straight line (this assumption will be relaxed later). An example of a population scatterplot with a quantitative classification factor is illustrated in Figure 1.4 where x is the elementary school grade level and y is a reading test score. In this example, each grade level represents a separate subpopulation of about 500 students. As in Figure 1.1, each point represents a single student.

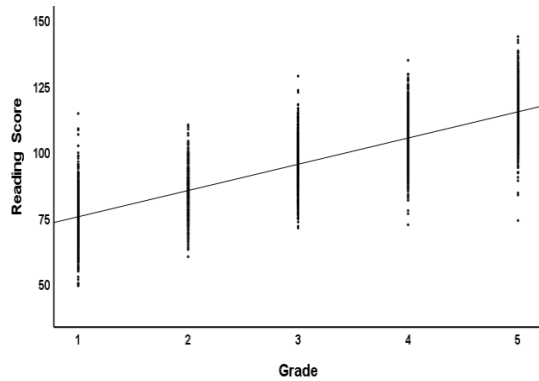


Figure 1.4. Scatterplot of reading test scores for five subpopulations of grades

In an experimental design with two or more treatment conditions, (see Chapters 2 and 3 of Volume 1) μ_j is the population mean of the response variable assuming everyone in the population had received treatment j . This same interpretation applies when the treatments are the levels of a quantitative factor. A population scatterplot like Figure 1.1 or Figure 1.4 does not exist when the factor is a treatment factor because the entire population is assumed to be assessed at a single level of the factor. But we can imagine a *hypothetical* population scatterplot for a quantitative treatment factor. Figure 1.5 illustrates a hypothetical population scatterplot for an experiment where 1,400 third grade students are assumed to read a story where each line of text is 3, 4, or 5 inches long and their reading speed is measured in words per minute. Within each factor level of this hypothetical scatterplot, each point represents the score of a single student but the scores of the *same* 1,400 students are plotted at each level of the treatment factor.

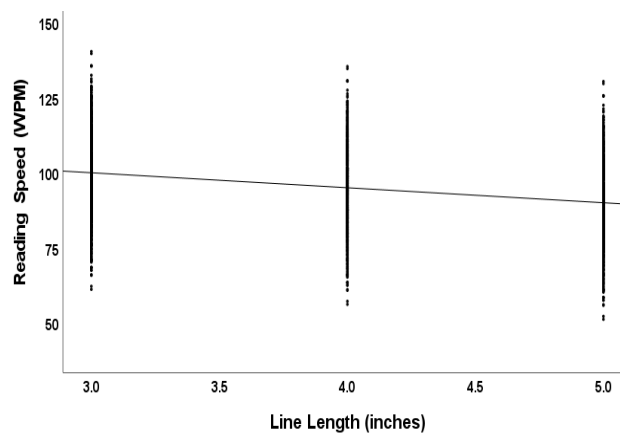


Figure 1.5. Hypothetical scatterplot of reading speed at three different text line lengths

1.5 Population Line of Best Fit

Each straight line in Figures 1.1, 1.2, and 1.3 is called the *line of best fit*. From elementary algebra, we know that a straight line can be defined by two parameters: the *y-intercept* and the *slope*. If y is the response variable and x is the predictor variable, then the slope of the line in the study population is defined as

$$\beta_1 = \sigma_{yx} / \sigma_x^2 \quad (1.3)$$

and the population y -intercept is defined as

$$\beta_0 = \mu_y - \beta_1 \mu_x. \quad (1.4)$$

The value $\beta_0 + \beta_1 x_i$ is called the *predicted y score* for person i . If x is a factor, then $\beta_0 + \beta_1 x_i$ is also the predicted mean of y . The predicted mean of y is also called a *conditional mean* of y at $\hat{x}$ and is denoted as $\mu_{y|\hat{x}}$ where $\hat{x}$ is the x value of interest. If x is a classification factor, then the conditional means are also referred to as subpopulation means. A *prediction error* (e_i) is the difference between the y score and the predicted y score for person i .

$$e_i = y_i - (\beta_0 + \beta_1 x_i) \quad (1.5)$$

Note that β_0 is the predicted y score for a person with an x score of 0. If an x score of 0 is impossible or uninteresting, then β_0 will not have a useful or interesting interpretation. If the x scores are *mean centered* by subtracting μ_x from each x score, then β_0 becomes the predicted y score for a person with an average x score.

The study population slope (β_1) and y -intercept (β_0) have the property of producing prediction errors that have a mean value of 0 in the study population. Given this property of the prediction errors, the population standard deviation of the prediction errors is

$$\sigma_e = \sqrt{\sum_{i=1}^N e_i^2 / N} \quad (1.6)$$

and σ_e^2 is the *prediction error variance*. The values of β_1 and β_0 have the unique property of minimizing σ_e (or σ_e^2). No other value of β_1 and β_0 , as defined in Equations 1.3 and 1.4, will produce a smaller value of σ_e . The prediction error standard deviation describes how accurately x predicts y .

An alternative measure of predictive accuracy is the *mean absolute prediction error* (MAPE). The population MAPE is defined as

$$\tau_e = \sum_{i=1}^N |e_i|/N \quad (1.7)$$

which is arguably easier to interpret than σ_e because τ_e is simply the average absolute prediction error.

The *coefficient of determination* is defined as $1 - \sigma_e^2/\sigma_y^2$ and describes the proportion of y score variance that is predictable from the x scores. The coefficient of determination has a possible range of 0 to 1 where 0 indicates that none of the y score variance is predictable from the x scores and 1 indicates that the y scores are perfectly predicted by the x scores. It can be shown that the squared Pearson correlation is a coefficient of determination ($\rho_{yx}^2 = 1 - \sigma_e^2/\sigma_y^2$).

1.6 Interpreting a Slope Coefficient

The slopes of the straight lines in Figures 1.1, 1.4, and 1.5 each have a different interpretation. For scatterplots like Figure 1.1 where x is a quantitative variable and not a factor, β_1 describes the change in the predicted y score that is *associated* with a 1-unit increase in x . In some applications it is more informative to describe the change in predicted y scores for a k -unit increase in x . The change in the predicted y score associated with a k -unit increase in x is equal to $k\beta_1$. Suppose the slope of the line in Figure 1.1 is 0.4. Then every 1-point increase in a Math SAT score is associated with a 0.4-point increase in the predicted Reading SAT score. Alternatively, every 10-point increase in a Math SAT score is associated with a 4-point increase in the predicted Reading SAT score.

In scatterplots like Figure 1.4 where x is a quantitative classification factor, β_1 describes the change in the population conditional mean (subpopulation mean) of the y scores that is *associated* with a 1-unit increase in x , within the range of x values used in the study. Suppose the slope of the line in Figure 1.3 is 9.95. Then each one-grade increase is associated with a 9.95 increase in the predicted subpopulation mean reading score. A k -grade increase is associated with a $k \times 9.95$ increase in the predicted subpopulation mean reading score. For example, we would expect a $2 \times 9.95 = 19.90$ increase in the predicted subpopulation mean reading score going from grade 1 to grade 3, or grade 2 to 4, or grade 3 to 5.

In an experimental design where x is a quantitative treatment factor, β_1 describes the change in the population conditional mean of y that would be *caused* by any 1-unit increase in the value of x . In Figure 1.5 where the slope is approximately -4.8, a 1-inch increase in the line length (up to 5 inches) would cause the population mean reading speed of the 1,400 third-grade students to decrease about 4.8 wpm. In some applications it is more informative to describe the slope in terms of a k -unit change in x rather than a 1-unit change in x . For example, any $\frac{1}{4}$ -inch increase in the line length would cause the population mean reading speed of the 1,400 third-grade students to decrease about $4.8/4 = 1.2$ wpm.

When interpreting the slope of a straight line for scatterplots like Figure 1.1, the interpretation applies to the range of x scores in the study population. When interpreting the slope of a line for scatterplots like Figure 1.4 where x is a classification factor, the interpretation applies to the range of x values that were used in the study and exist in the study population. For example, if grades 1, 3, and 6 were used in a study, the slope would describe the change in the conditional mean (subpopulation mean) for grades 1, 2, 3, 4, 5, and 6. When interpreting the slope of a line for scatterplots like Figure 1.5 where x is a treatment factor, the interpretation applies to any value of x that is within the range of x values used in the experiment.

1.7 A Comparison of β_1 and ρ_{yx}

Both β_1 and ρ_{yx} describe the strength of a linear relation between y and x but they have different interpretations. To interpret β_1 , it is necessary to understand the scale of both y and x . Suppose y is a fifth-grade mathematics achievement test score, x is the amount of supplemental math instruction funding a school receives, and $\beta_1 = 2.5$. Without knowing the scales of y and x it is impossible to assess the effect of funding on mathematics performance. If x was measured in millions of dollars and y was measured on a 0 to 100 scale, a slope of 0.25 would be considered small and unimportant (i.e., every one million dollars invested is associated with a mere 0.25-point increase in reading performance). But if x was measured in thousands of dollars, then a slope of 0.25 might suggest a very strong and important association between funding and mathematics performance.

The Pearson correlation is a unitless measure of association that does not require knowledge of the scales of y and x . For example, if the correlation between funding and

mathematics performance is .60, we could conclude that the relation is moderately strong without knowing how mathematics performance or funding was measured.

The Pearson correlation can also be interpreted as a *standardized slope*. From Equation 1.1, the covariance can be expressed as $\sigma_{yx} = \rho_{yx}\sigma_y\sigma_x$ and substituting $\rho_{yx}\sigma_y\sigma_x$ for the covariance in Equation 1.3 gives $\beta_1 = \rho_{yx}\sigma_y/\sigma_x$ which shows that the slope is related to the Pearson correlation. Solving for ρ_{yx} gives $\rho_{yx} = \beta_1\sigma_x/\sigma_y$. If both y and x are standardized so that $\sigma_x = \sigma_y = 1$, then $\rho_{yx} = \beta_1$. In a nonexperimental design ρ_{yx} describes the change in the predicted *standardized* y score associated with a one *standard deviation* increase in x . For example, suppose y is a measurement of physics knowledge, x is a measurement of interest in physics, and $\rho_{yx} = .25$. If x is assumed to be the predictor variable, we could say that a one standard deviation increase in physics knowledge is associated with a .25 standard deviation increase in the physics interest score.

The value of ρ_{yx} is unchanged if all the y scores or x scores are multiplied or divided by any positive number or if any number is added to or subtracted from all the y scores or all the x scores. This can be proven using the covariance algebra results in Appendix B. Adding or subtracting a number from all the y scores or all the x scores will not change the value of the β_1 . However, multiplying the x scores by a number greater than 1 will make β_1 smaller, and multiplying the y scores by a number greater than 1 will make β_1 larger. The exact effect of rescaling the y scores and x scores on β_1 can be determined using the covariance algebra results in Appendix B.

Reducing the range of x scores in the study population will reduce the Pearson correlation but will not affect the slope. In Figure 1.1, the Pearson correlation is about .45 and the slope is about .40. If only students with Math SAT scores above 1100 are included in the study population (see Figure 1.6), the Pearson correlation is reduced to .23 while the slope is virtually unchanged. When x is a factor, as in Figure 1.4 and Figure 1.5, then it is easy to see that the slope would be unchanged if the range of x values is reduced.

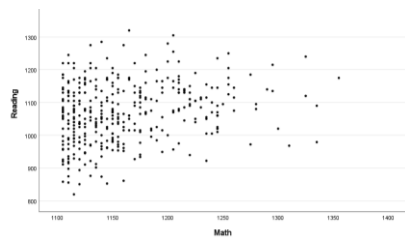


Figure 1.6 Population scatterplot of Math SAT scores (> 1100) and Reading SAT scores

1.8 Linear Statistical Models with One Predictor Variable

In most applications, it will not be practical or possible to examine the relation between the y scores and the x scores in the entire study population. Instead, the researcher will obtain a random sample of size n from a study population of size N . Simple random sampling is used if x is not a factor. With simple random sampling, a random sample of participants is obtained from the study population and each participant is assigned a y score and an x score.

If x is a quantitative classification factor, either simple random sampling or stratified random sampling can be used. With simple random sampling, a random sample of participants is obtained and each participant is then classified into the levels of x . With stratified random sampling, the study population is stratified into subpopulations and a random sample of participants is obtained from each level of x . Stratified random sampling is usually more difficult to implement, but it allows the researcher to specify the number of participants at each level of x . If x is a quantitative treatment factor, simple random sampling is used and the participants are then randomly assigned to the levels of the treatment factor.

Unlike a mathematical model, such as $y = a + bx$, that describes an exact relation between x and y , a statistical model is used to describe an imperfect relation between x and y . A *simple linear statistical model* can be used to describe an assumed linear relation between a quantitative response variable (y) and one predictor variable (x) in a random sample of n participants. The adjective "simple" is used to indicate that there is only one predictor variable. Linear statistical models with two or more predictor variables are introduced in Chapter 2.

There are two versions of the simple linear statistical model: a *random- x* model and a *fixed- x* model. In a fixed- x model, x is a classification or treatment factor where the values of x are predetermined by the researcher. In the random- x model, x is not a factor and each participant in a random sample is assigned a pair of x and y scores. In a random- x model, the values of the x scores observed in the sample will not be known in advance.

The following simple linear statistical model describes an assumed linear relation between x and y for participant i ($i = 1$ to n)

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad (1.8)$$

where β_0 (the population y -intercept) and β_1 (the population slope) are the parameters of the statistical model. As explained in section 1.5, $\beta_0 + \beta_1 x_i$ is the predicted y score for participant i , and $e_i = y_i - (\beta_0 + \beta_1 x_i)$ is the prediction error for participant i .

At this point, it might seem unnecessary to define a statistical model that merely specifies a linear relation between x and y for a random sample of participants. Likewise, statistical models for all the designs in Volume 1 could have been defined, but the parameters of these models would contain only population means and it is more straightforward to simply describe hypothesis tests and confidence intervals for the population means. The usefulness of a statistical model will become clear in Chapter 2 where Equation 1.8 is extended to include two or more predictor variables, predictor variables that are qualitative or quantitative, nonlinear effects of the predictor variables, and interaction effects of the predictor variables.

1.9 Parameter Estimation

The formulas for estimating β_0 and β_1 from a random sample of size n are the same for random- x and fixed- x models. The estimate of β_1 is

$$\hat{\beta}_1 = \hat{\sigma}_{yx} / \hat{\sigma}_x^2 \quad (1.9)$$

and the estimate of β_0 is

$$\hat{\beta}_0 = \hat{\mu}_y - \hat{\beta}_1 \hat{\mu}_x \quad (1.10)$$

where the estimated means of the y scores and x scores are

$$\hat{\mu}_y = \sum_{i=1}^n y_i / n \quad (1.11a)$$

$$\hat{\mu}_x = \sum_{i=1}^n x_i / n, \quad (1.11b)$$

the estimated variances of the y scores and x scores are

$$\hat{\sigma}_y^2 = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_y)^2}{n - 1}, \quad (1.12a)$$

$$\hat{\sigma}_x^2 = \sum_{i=1}^n \frac{(x_i - \hat{\mu}_x)^2}{n - 1}, \quad (1.12b)$$

and the estimated covariance between the y scores and x scores is

$$\hat{\sigma}_{yx} = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_y)(x_i - \hat{\mu}_x)}{n - 1}. \quad (1.13)$$

The *estimated y score* for participant i is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. The estimated prediction error (also called the *residual*) for participant i is $\hat{e}_i = y_i - \hat{y}_i$. Equations 1.9 and 1.10 are called *ordinary least squares* (OLS) estimates, and they are the unique values that minimize $\sum_{i=1}^n \hat{e}_i^2$. *Generalized least squares* (GLS) estimation is introduced in Chapter 3.

Some researchers prefer to center the predictor variable by subtracting some value of x , denoted as $\dot{x}$, from each x score. When the predictor variable has been mean centered ($\dot{x} = \hat{\mu}_x$), then $\hat{\beta}_0$ is the estimated predicted y score at $x = \hat{\mu}_x$ which could be more useful than the estimated predicted y score at $x = 0$. Centering does not change the value of $\hat{\beta}_1$.

An estimate of the population Pearson correlation is

$$\hat{\rho}_{yx} = \frac{\hat{\sigma}_{yx}}{\hat{\sigma}_y \hat{\sigma}_x} \quad (1.14)$$

where $\hat{\sigma}_{yx}$ is the estimated covariance, $\hat{\sigma}_y$ is the estimated standard deviation of the y scores, and $\hat{\sigma}_x$ is the estimated standard deviation of the x scores.

The mean of the estimated prediction errors is always 0, which is a property of OLS estimates. An estimate of the population prediction error variance is

$$\hat{\sigma}_e^2 = \frac{\sum_{i=1}^n \hat{e}_i^2}{n - 2} \quad (1.15)$$

where $\hat{e}_i$ is the residual for the i^{th} participant. The estimate of σ_e^2 is also denoted as MS_E in some statistical formulas. The square root of Equation 1.15 ($\hat{\sigma}_e$) estimates the standard deviation of the population prediction errors and describes how accurately the y scores are predicted by the x scores in the sample.

An estimate of the population MAPE is

$$\hat{\tau}_e = \sum_{i=1}^n |\hat{e}_i| / n \quad (1.16)$$

where $\hat{e}_i$ is the residual for the i^{th} participant. Some researchers prefer reporting $\hat{\tau}_e$ rather than $\hat{\sigma}_e$ because $\hat{\tau}_e$ has a simple interpretation. Both $\hat{\tau}_e$ and $\hat{\sigma}_e$ provide useful information regarding the relation between x and y that is distinct from the information provided by $\hat{\rho}_{yx}$ and $\hat{\beta}_1$.

Example 1.1. A random sample of 150 UCSC undergraduate students was obtained from a university database of about 18,000 undergraduate students. The 150 students were asked to rate their overall satisfaction with their UCSC education and to also rate their satisfaction with the quality of the academic advising they received. Advising satisfaction (scored on a 1 to 5 scale) is the predictor variable (x) and education satisfaction (scored on a 1 to 10 scale) is the response variable (y). The estimated slope is $\hat{\beta}_1 = 1.7$ and the estimated y -intercept is $\hat{\beta}_0 = 2.1$. In the sample of 150 students, every 1-point increase in advising satisfaction is associated with a 1.7-point increase in education satisfaction. The predicted education satisfaction score for a student with an advising satisfaction rating of 4 is $2.1 + 1.7(4) = 8.9$. The estimated correlation between advisor satisfaction and education satisfaction is $\hat{\rho}_{yx} = .60$. These results apply only to the sample of 150 students.

1.10 Inferential Statistical Methods

Researchers are interested in the values of population parameters (e.g., β_1 , ρ_{yx} , τ_e), but they are usually able to obtain only estimates of these parameters because the study population size is either too large or the measurement process is too costly. A parameter estimate can be misleading because it can be greater than or less than the population value as a result of random sampling error, and the direction of the sampling error will be unknown. Furthermore, the magnitude of the sampling error will be unknown. Thus, the estimate might be too small or too large, and it might be close to or very different from the population parameter value. In Example 1.1, the estimated slope was 1.7, and the estimated correlation was .60. However, these estimates were computed from a random sample and they contain sampling error of unknown direction and unknown magnitude. The researcher would like to say something about the values of β_1 and ρ_{yx} in the study population of 18,000 UCSC students.

A parameter estimate combined with its standard error, both of which are computed from a random sample, can be used to make a statement regarding the value of an unknown population parameter. Recall from Chapter 1 of Volume 1 that inferential statistical methods can be used to make different types of statements about a population parameter based on an estimate of the parameter and its standard error. One type of statement is in the form of a confidence interval, and another type of statement is in the form of a hypothesis test. Confidence intervals for β_1 , $\mu_{y|x}$, ρ_{yx} , and τ_e are presented in sections 1.11, 1.12, 1.14, and 1.16. Hypothesis testing methods are presented in sections 1.19 and 1.21.

1.11 Confidence Interval for β_1

A $100(1 - \alpha)\%$ confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{\alpha/2;df} SE_{\hat{\beta}_1} \quad (1.17)$$

where $SE_{\hat{\beta}_1} = \sqrt{\frac{\hat{\sigma}_e^2}{\hat{\sigma}_x^2(n-1)}}$ is the estimated standard error of $\hat{\beta}_1$, $t_{\alpha/2;df}$ is a two-sided critical t -value, and $df = n - 2$.

The value of $SE_{\hat{\beta}_1}$ depends on the variance of the x scores with greater variability in the x scores resulting in a smaller value of $SE_{\hat{\beta}_1}$. This fact is useful when planning a study. Specifically, obtaining a wider range of x values will provide a more accurate estimate of β_1 for a given total sample size. The value of $SE_{\hat{\beta}_1}$ also depends on the sample size with larger sample sizes tending to produce smaller standard errors. It can be shown that $\hat{\sigma}_e^2 = \frac{(n-1)\hat{\sigma}_y^2(1-\hat{\rho}_{yx}^2)}{n-2}$. This expression for $\hat{\sigma}_e^2$ shows that $SE_{\hat{\beta}_1}$ depends on the value of $\hat{\rho}_{yx}^2$ with a larger value of $\hat{\rho}_{yx}^2$ resulting in a smaller standard error. Thus, a $100(1 - \alpha)\%$ confidence interval for β_1 tends to be narrower for a larger sample size, a wider range of x scores, and a stronger linear association between x and y . SPSS and R will compute Equation 1.17.

Example 1.2. In Example 1.1, the standard error of the slope estimate ($SE_{\hat{\beta}_1}$) is 0.54. The critical t -value for $\alpha = .05$ and $df = 150 - 2 = 148$ is 1.98. The 95% confidence interval for the population slope is $1.7 \pm 1.98(0.54) = [0.63, 2.77]$. The researcher is 95% confident that for this population of about 18,000 UCSC students, every 1-point increase in advising satisfaction is associated with a 0.63 to 2.77 increase in education satisfaction.

Example 1.3. Sixty students were randomly selected from a research participant pool consisting of about 4,000 students. Each student was given a fitness tracker to wear for one week. The response variable was the total number of steps taken over a 5-day period. The students were randomly assigned to three groups and were told, in groups 1 to 3 respectively, that they would receive an award of \$5, \$15, or \$25 if they could reach 50,000 steps in five days. The estimated y -intercept is $\hat{\beta}_0 = 25,243$, which is the predicted number of steps with no reward. The estimated slope and its standard error are $\hat{\beta}_1 = 1,326$ and $SE_{\hat{\beta}_1} = 147.5$. The critical t -value for $\alpha = .05$ and $df = 60 - 2 = 58$ is 2.00. The 95% confidence interval for the population slope is $1,326 \pm 2.00(147.5) = [1,031, 1,621]$. The researcher is 95% confident that for this study population of 4,000 college students, every additional dollar of reward, in the range of \$5 to \$25, will cause a 1,031 to 1,621 increase in the total number of steps taken during a 5-day period.

In an experimental design with an α -level quantitative factor, an alternative confidence interval for β_1 uses the confidence interval for a linear contrast of means described in section 3.3 of Volume 1. It can be shown that β_1 in an experimental design with a single quantitative factor is equal to $\sum_{j=1}^a v_j \mu_j$ where $v_j = (x_j - \mu_x) / \sum_{j=1}^a (x_j - \mu_x)^2$ and $\mu_x = \sum_{j=1}^a x_j / a$. The `ci.slope.mean.bs` function will compute a confidence interval for β_1 for a quantitative factor that does not assume equal variances.

1.12 Confidence Interval for a Conditional Mean

If x is a quantitative factor, we might want to estimate the population conditional mean of y at $\dot{x}$, where $\dot{x}$ is some value of x within the range of x values that were used in the study. The conditional mean of y at $\dot{x}$ is equal to $\mu_{y|\dot{x}} = \beta_0 + \beta_1 \dot{x}$. A $100(1 - \alpha)\%$ confidence interval for $\mu_{y|\dot{x}}$ is

$$\hat{\mu}_{y|\dot{x}} \pm t_{\alpha/2; df} SE_{\hat{\mu}_{y|\dot{x}}} \quad (1.18)$$

where $\hat{\mu}_{y|\dot{x}} = \hat{\beta}_0 + \hat{\beta}_1 \dot{x}$, $SE_{\hat{\mu}_{y|\dot{x}}} = \hat{\sigma}_e \sqrt{\frac{1}{n} + \frac{(\dot{x} - \hat{\mu}_x)^2}{\hat{\sigma}_x^2(n-1)}}$, and $df = n - 2$. Note that the standard error of $\hat{\mu}_{y|\dot{x}}$ is smallest when $\dot{x} = \hat{\mu}_x$. Note also that if $\dot{x} = 0$, then Equation 1.18 is a confidence interval for β_0 . SPSS and R will compute Equation 1.18.

Suppose participants are randomly assigned to receive 1, 5, or 9 hours of training. The confidence interval for $\mu_{y|\dot{x}}$ can be computed for any value of $\dot{x}$ in the range of 1 to 9. If x is a quantitative classification factor, it usually makes sense to compute a confidence interval for $\mu_{y|\dot{x}}$ only for values of $\dot{x}$ that are in the range of x values used in the study assuming these values also exist in the study population.

Another way to compute Equation 1.18 is to center the x scores around $\dot{x}$ (i.e., subtract $\dot{x}$ from every x score) because then β_0 will equal the mean of y at $\dot{x}$ and the confidence interval for β_0 (from SPSS or R) will be the same as Equation 1.18.

Example 1.4. In Example 1.3, the researcher wants to estimate the population mean number of steps if all 4,000 students had been told they would receive a \$10 reward. The estimated mean number of total steps at $\dot{x} = 10$ is $25243 + 1326(10) = 38,503$ with an estimated standard error of 1406.2. The 95% confidence interval for the population mean number of total steps for a \$10 reward is $38503 \pm 2.00(1406.2) = [35691, 41315]$.

1.13 Prediction Interval for a Predicted y Score

The simple linear statistical model (Equation 1.8) is also used to select applicants for jobs and training programs. In these applications, y is usually a measure of performance on the job or in the training program, and x is variable that should be at least moderately correlated with y . In some applications, the measurement of y is costly or difficult to obtain, while x is relatively easy to obtain. In these types of applications, β_0 and β_1 could be estimated using a random sample of applicants in which it is possible to obtain measurements of both y and x . The estimates of β_0 and β_1 can then be used to estimate y scores for new applicants. The predicted y score for a new applicant is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \hat{x}$ where $\hat{x}$ is the x score of the new applicant. A range of plausible y scores for the new applicant is called a *prediction interval* and is computed using the following formula

$$\hat{y} \pm t_{\alpha/2; df} SE_{\hat{y}} \quad (1.19)$$

where $SE_{\hat{y}} = \hat{\sigma}_e \sqrt{1 + \frac{1}{n} + \frac{(\hat{x} - \hat{\mu}_x)^2}{\hat{\sigma}_x^2(n-1)}}$ and $df = n - 2$. Note that the prediction interval width is narrowest for $\hat{x} = \hat{\mu}_x$. The prediction interval can never be narrower than the confidence interval for $\mu_{y|\hat{x}}$. SPSS and R will compute Equation 1.19.

Example 1.5. Fifty randomly selected job applicants who had passed an initial screening review were then given a newly developed job screening test. The screening test (x) was scored on a 1 to 100 scale. All fifty applicants were hired and their job performance (y) was assessed (on a scale of 0 to 10) three months later using a team of expert evaluators. The x scores were mean centered around $\hat{\mu}_x = 54.1$ and the estimates of β_0 and β_1 are 4.85 and .071, respectively. A new applicant was given the screening test and obtained a score of 78. The estimated y score for this applicant is $4.85 + .071(78 - 54.1) = 6.55$. The 95% prediction interval for the job performance score of this one applicant is [5.75, 7.35].

1.14 Confidence Interval for ρ_{yx}

An approximate confidence interval for ρ_{yx} , which assumes a random- x model, is obtained in two steps. First, a $100(1 - \alpha)\%$ confidence interval for a transformed correlation estimate is computed

$$\left(\hat{\rho}_{yx}^* - \frac{\hat{\rho}_{yx}}{2(n-1)} \right) \pm z_{\alpha/2} \sqrt{\frac{1}{n-3}} \quad (1.20)$$

where $\hat{\rho}_{yx}^* = \ln\left(\frac{1 + \hat{\rho}_{yx}}{1 - \hat{\rho}_{yx}}\right)/2$ is called the *Fisher transformation* of $\hat{\rho}_{yx}$, $z_{\alpha/2}$ is a critical two-sided z-value, $\sqrt{\frac{1}{n-3}}$ is the standard error of $\hat{\rho}_{yx}^*$, and $\frac{\hat{\rho}_{yx}}{2(n-1)}$ is a bias correction. Let ρ_L^* and ρ_U^* denote the endpoints of Equation 1.20. Reverse transforming the endpoints of Equation 1.20 gives the following lower confidence limit for ρ_{yx}

$$LL = \frac{\exp(2\rho_L^*) - 1}{\exp(2\rho_L^*) + 1} \quad (1.21a)$$

and the following upper confidence limit for ρ_{yx}

$$UL = \frac{\exp(2\rho_U^*) - 1}{\exp(2\rho_U^*) + 1}. \quad (1.21b)$$

The sampling distribution of $\hat{\rho}_{yx}$ is not well approximated by a normal distribution unless the sample size is very large or if ρ_{yx} is close to zero. Remarkably, the sampling distribution of $\hat{\rho}_{yx}^*$ is closely approximated by a normal distribution in small samples and virtually any value of ρ_{yx} . SPSS and R will compute Equations 1.21a and 1.21b from raw data. The `ci.cor` function in the `statpsych` package will compute Equations 1.21a and 1.21b using an estimate of the Pearson correlation.

Example 1.6. The estimated Pearson correlation was .60 in Example 1.1 where the relation between advising satisfaction and education satisfaction was examined in a random sample of 150 students taken from a study population of 18,000 UCSC undergraduate students. The Fisher transformed correlation is $\hat{\rho}_{yx}^* = 0.693$ and Equation 1.20 gives the following 95% confidence interval.

$$(0.693 - .002) \pm 1.96\sqrt{1/147} = [0.529, 0.853]$$

Reverse transforming 0.529 gives a Pearson correlation of .485 and reverse transforming .853 gives a Pearson correlation of about .692. The 95% confidence interval for ρ_{yx} is [.485, .692]. The researcher is 95% confident that the Pearson correlation between education satisfaction and advising satisfaction in the study population of 18,000 UCSC students is between .485 and .692.

1.15 Confidence Interval for a Coefficient of Determination

The coefficient of determination in a random- x model is equal to the squared Pearson correlation (ρ_{yx}^2) and describes the proportion of variance of the y scores that can be predicted from knowing a person's x score. In fixed- x models with a quantitative factor,

the population coefficient of determination is also equal to ρ_{yx}^2 but will be denoted as η^2 . Although $\eta^2 = \rho_{yx}^2$, different symbols are used because the confidence intervals for η^2 and ρ_{yx}^2 are computed differently. In Chapter 3 of Volume 1, η^2 was used to describe the proportion of variance in the y scores that is predictable from a qualitative factor rather than a quantitative factor.

In a random- x model, if the confidence interval for ρ_{yx} does not include 0, then a confidence interval for ρ_{yx}^2 can be obtained by simply squaring the endpoints of Equations 1.21a and 1.21b. In Example 1.6, the 95% confidence interval for ρ_{yx} was [.486, .694]. Squaring the lower and upper limits gives a 95% confidence interval for ρ_{yx}^2 of [.236, .482] indicating that 23.6% to 48.2% of the variance in the education satisfaction scores in the study population can be predicted from the advising satisfaction scores.

If the confidence interval for ρ_{yx} includes 0, then the lower limit for ρ_{yx}^2 is set to 0 and the upper limit for ρ_{yx}^2 is equal to the largest squared endpoint. For example, if the confidence interval for ρ_{yx} is [-.40, .07], the confidence interval for ρ_{yx}^2 is [0, .16].

The confidence interval for η^2 in a fixed- x model is not the same as a confidence interval for ρ_{yx}^2 in a random- x model. The confidence interval for η^2 does not have a simple formula but can be obtained using the `ci.etasqr` function in the `statpsych` package. The `ci.etasqr` function requires the values of df_1 and df_2 where df_2 is the Error (or Residual) degrees of freedom. In a design with a single quantitative factor, $df_1 = 1$.

1.16 Confidence Interval for τ_e

An approximate $100(1 - \alpha)\%$ confidence interval for the population MAPE (τ_e) is

$$\exp[\ln(c\hat{\tau}_e) \pm z_{\alpha/2}SE_{\ln(\hat{\tau}_e)}] \quad (1.22)$$

where $c = n/(n - 3/2)$, $df = n - 2$, and $SE_{\ln(\hat{\tau}_e)} = \sqrt{(\frac{\sum_{i=1}^n \hat{e}_i^2}{(n-1)\hat{\tau}_e^2} - 1)/df}$. A less biased estimate of τ_e is $c\hat{\tau}_e$. Recall that τ_e is easier to interpret than σ_e . Furthermore, the traditional confidence interval for σ_e is extremely sensitive to nonnormality of the prediction errors. The `ci.mape` function in the `statpsych` package will compute Equation 1.22.

1.17 Dummy Coded Predictor Variable

Consider a study with a two-level *qualitative* predictor variable, such as male and female or Treatment A and Treatment B. This type of study was described in Chapter 2 of Volume 1. A two-level qualitative predictor variable can be used in a simple linear statistical model if the two levels are converted into numerical values. A *dummy coded* predictor variable sets $x_i = 1$ for every participant in level 1 of the qualitative factor (e.g., male, Treatment A) and sets $x_i = 0$ for every participant in level 2 of the qualitative factor (e.g., female, Treatment B).

The conditional mean of y at $\dot{x} = 0$ is $\mu_{y|0} = \beta_0 + \beta_1(0) = \beta_0$ and the conditional mean of y at $\dot{x} = 1$ is $\mu_{y|1} = \beta_0 + \beta_1(1) = \beta_0 + \beta_1$. It follows that $\mu_{y|1} - \mu_{y|0} = \beta_1$. Using the notation in Chapter 2 of Volume 1, $\mu_{y|0} = \mu_2$, $\mu_{y|1} = \mu_1$, and then $\beta_1 = \mu_1 - \mu_2$. The confidence interval for β_1 using Equation 1.17 when x is a dummy coded variable is identical to the equal-variance confidence interval for $\mu_1 - \mu_2$ described in Chapter 2 of Volume 1. A hypothesis test of $\beta_1 = 0$ when x is a dummy coded variable will give the same results (same test statistic value and same p -value) as the equal-variance independent-samples t -test described in Chapter 2 of Volume 1.

1.18 Point-biserial Correlation

A point-biserial correlation is a correlation between a quantitative variable and a dummy coded variable. A point-biserial correlation is a standardized measure of effect size for a two-group experimental or nonexperimental design. A point-biserial correlation has a possible range of -1 to 1 and might be easier to interpret than $\mu_1 - \mu_2$ in some applications. A point-biserial correlation can be defined in terms of a standardized mean difference. Recall from Chapter 2 of Volume 1 that the following estimate of a standardized mean difference uses an unweighted average of the variances and is appropriate for experimental designs

$$\hat{\delta} = (\hat{\mu}_1 - \hat{\mu}_2) / \sqrt{(\hat{\sigma}_1^2 + \hat{\sigma}_2^2) / 2}. \quad (1.23)$$

The following estimate of a population standardized mean difference (also known as Cohen's d) uses a weighted average of the variances and is appropriate for nonexperimental designs with simple random sampling

$$\hat{\delta}' = (\hat{\mu}_1 - \hat{\mu}_2) / \sqrt{[(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2] / (n_1 + n_2 - 2)} \quad (1.24)$$

For experimental designs, an estimate of the population point-biserial correlation can be expressed as the following function of $\hat{\delta}$

$$\hat{\rho}_{pb} = \hat{\delta} / \sqrt{\hat{\delta}^2 + 4} \quad (1.25)$$

and for nonexperimental designs, an estimate of the population point-biserial correlation can be expressed as the following function of $\hat{\delta}'$

$$\hat{\rho}'_{pb} = \hat{\delta}' / \sqrt{\hat{\delta}'^2 + \frac{n-2}{n\hat{\pi}(1-\hat{\pi})}} \quad (1.26)$$

where $n = n_1 + n_2$ and $\hat{\pi} = n_1/n$.

The confidence intervals for δ or δ' given in Chapter 2 of Volume 1 are used to construct a confidence interval for a population point-biserial correlation by substituting the estimated standardized mean difference in Equations 1.25 or 1.26 with a lower confidence limit and an upper confidence limit for a standardized mean difference. The `ci.pbcor` function in the `statpsych` package will compute a confidence interval for both types of point-biserial correlations. The confidence intervals for δ and δ' in Chapter 2 of Volume 1 do not assume equal population variances. However, the confidence interval for δ' does assume equal population variances if stratified random sampling is used.

Example 1.7. A researcher obtained a random sample of 200 employees from a corporation with about 8,500 employees. A happiness scale was given to the 200 employees. The 200 employees were classified into two groups: those who exercised regularly ($n_1 = 47$) and those who did not ($n_2 = 153$). The estimate of δ' is 0.35 with a 95% confidence interval of [0.02, 0.68]. The estimate of ρ_{pb} is $0.35 / \sqrt{.35^2 + 5.51} = 0.147$. The 95% lower and upper limits for ρ_{pb} are $LL = 0.02 / \sqrt{0.02^2 + 5.51} = .008$ and $UL = 0.68 / \sqrt{0.68^2 + 5.51} = .278$. The researcher is 95% confident that the point-biserial correlation between happiness and the dichotomous measure of exercise in the population of 8,500 employees is between .008 and .278. This study is not an experiment and does not provide evidence of a causal relation between exercise and happiness, and the researcher should not claim that people will become happier if they begin exercising.

1.19 Directional Two-sided Hypothesis Tests

In Volume 1, directional two-sided hypothesis tests are used to decide if a population parameter (e.g., population mean, difference in population means, linear contrast of population means) is greater than some value or less than some value. If the population parameter is greater than some value, this could provide support for a particular theory

or course of action. If the population parameter is less than some value, that would provide support for another theory or another course of action. A directional two-sided test can be applied to any of the population parameters that have been defined in this chapter (e.g., ρ_{yx} , β_1 , $\mu_{y|x^*}$, ρ_{pb} , ρ_{yx}^2 , η^2 , τ_e). For example, the following hypotheses can be specified regarding β_1

$$H_0: \beta_1 = h$$

$$H_1: \beta_1 > h$$

$$H_2: \beta_1 < h$$

where h is some number specified by the researcher. In virtually all applications, H_0 is known to be false (because it is extremely unlikely that β_1 will exactly equal h) and the goal is to decide if H_1 is true or if H_2 is true.

A $100(1 - \alpha)\%$ confidence interval for β_1 can be used to choose $H_1: \beta_1 > h$ or $H_2: \beta_1 < h$. If the upper limit of the confidence interval is less than h , then H_0 is rejected and $H_2: \beta_1 < h$ is accepted; if the lower limit of the confidence interval is greater than h , then H_0 is rejected and $H_1: \beta_1 > h$ is accepted. If the confidence interval includes h , then H_0 cannot be rejected and the results are inconclusive. Confidence intervals for ρ_{yx} , $\mu_{y|x^*}$, ρ_{pb} , ρ_{yx}^2 , η^2 or τ_e can be used to perform directional two-sided hypothesis tests for those population parameters. Recall from Volume 1 that the probability of accepting H_1 when H_2 is true or accepting H_2 when H_1 is true is called a directional error, and the probability of making a directional error is at most $\alpha/2$.

A directional two-sided hypothesis test for $H_0: \beta_1 = h$ is often performed using a test statistic rather than a confidence interval. The test statistic is

$$t = (\hat{\beta}_1 - h)/SE_{\hat{\beta}_1} \quad (1.27)$$

and the following decision rules are used.

reject H_0 and accept $H_1: \beta_1 > h$ if $t > t_{\alpha/2; df}$

reject H_0 and accept $H_2: \beta_1 < h$ if $t < -t_{\alpha/2; df}$

fail to reject H_0 if $|t| < t_{\alpha/2; df}$

A directional two-sided test for $H_0: \rho_{yx} = h$ uses the following test statistic

$$z = (\hat{\rho}_{yx}^* - h^*)/\sqrt{1/(n-3)} \quad (1.28)$$

and the following decision rules

reject H_0 and accept $H_1: \rho_{yx} > h$ if $z > z_{\alpha/2}$

reject H_0 and accept $H_2: \rho_{yx} < h$ if $z < -z_{\alpha/2}$

fail to reject H_0 if $|z| < z_{\alpha/2}$.

where h^* is a Fisher transformation of h .

SPSS and R compute Equation 1.27 for $h = 0$. Note that $\beta_1 = 0$ implies $\rho_{yx} = 0$, $\beta_1 > 0$ implies $\rho_{yx} > 0$, and $\beta_1 < 0$ implies $\rho_{yx} < 0$ so that Equation 1.27 can also be used to test $H_0: \rho_{yx} = 0$. If x is dummy coded, then Equation 1.27 can be used to test $H_0: \rho_{pb} = 0$.

SPSS and R will compute a p -value for the test of $H_0: \beta_1 = 0$, which also applies to the test of $H_0: \rho_{yx} = 0$ and the test of $H_0: \rho_{pb} = 0$. Recall from Volume 1 that it is a common practice to report the results of a statistical test to be “significant” if the p -value is less than .05 and “nonsignificant” if the p -value is greater than .05. Remember that a significant result does not necessarily indicate that the population parameter is meaningfully different from the hypothesized value, and that a nonsignificant result does not imply that the null hypothesis is true. It is a common mistake to test $H_0: \beta_1 = 0$ or $H_0: \rho_{yx} = 0$ and claim that there is “no relation” between x and y if the hypothesis testing results are nonsignificant. It is not possible to show that a population correlation or slope is equal to zero, but the interval hypothesis test described in section 1.20 can be used to show that the population correlation is less than some value that would be considered small or unimportant.

1.20 Interval Hypothesis Test

In some studies the goal is to show that the relation between x and y is small or unimportant. In these applications, the following hypotheses regarding a population Pearson correlation can be specified

$$H_1: |\rho_{yx}| < h \quad H_2: |\rho_{yx}| \geq h$$

where $|\rho_{yx}|$ denotes the absolute value of ρ_{yx} and h is a researcher-specified number that represents a small or unimportant correlation value. These types of hypotheses are evaluated using an *interval hypothesis test*. Note that H_1 implies ρ_{yx} is between $-h$ and h and H_2 implies that ρ_{yx} is less than $-h$ or greater than h . A confidence interval for ρ_{yx} can be used to perform an interval hypothesis test. H_1 is accepted if the confidence interval is

completely within the $-h$ to h range, and H_2 is accepted if the confidence interval is completely outside the $-h$ to h range. The results are inconclusive if the confidence interval includes the values h or $-h$. If x is a dummy coded variable, a confidence interval for ρ_{pb} could be used to perform an interval hypothesis test.

1.21 Monotonicity Test

In a fixed- x model (experimental or nonexperimental design), the slope (β_1) provides a useful description of how the conditional mean ($\mu_{y|x}$) changes linearly as a function of x . Specifically, β_1 describes the change in $\mu_{y|x}$ for any 1-unit increase in x across the range of x values that were used in the study. If the relation between x and $\mu_{y|x}$ is *monotonic* rather than linear, then the change in $\mu_{y|x}$ will not necessarily be the same for every 1-unit increase in x . For example, if the levels of x are first grade, second grade, and third grade students, the change in $\mu_{y|x}$ could be larger from first grade to second grade than from second grade to third grade. Also, if x is measured on an ordinal scale rather than an interval scale, then β_1 can be misleading and it is more appropriate to test for monotonicity.

The relation between x and $\mu_{y|x}$ is *monotonic increasing* if $\mu_{y|x}$ increases or is unchanged for each 1-unit increase in x . The relation between x and $\mu_{y|x}$ is *monotonic decreasing* if $\mu_{y|x}$ decreases or is unchanged for each 1-unit increase in x .

A simple way to test for monotonicity is to compute confidence intervals for all adjacent pairs of x values. For example, to determine if $\mu_{y|x}$ is a monotonically related to x where x has four levels, compute confidence intervals for $\mu_{y|1} - \mu_{y|2}$, $\mu_{y|2} - \mu_{y|3}$, and $\mu_{y|3} - \mu_{y|4}$, where the subscripts 1, 2, 3, and 4 refer to the levels of x . If at least one upper confidence limit is less than 0 and no lower confidence limit is greater than 0, then conclude that if $\mu_{y|x}$ is a monotonically increasing function of x . If at least one lower confidence limit is greater than 0 and no upper confidence limit is less than 0, then conclude that if $\mu_{y|x}$ is a monotonically decreasing function of x . If any lower confidence limit is greater than 0 and any upper confidence limit is less than 0, then conclude that $\mu_{y|x}$ is not monotonically related to x . A Bonferroni adjustment to the α value in the confidence intervals is recommended. The `test.mono.mean.bs` function in the `statpsych` package will compute Bonferroni simultaneous confidence intervals for population mean differences

at adjacent pairs of x values. The `test.mono.median.bs` function is recommended when the response variable is highly skewed within groups.

Example 1.8. A stratified random sample of 40 participants from each of four subpopulations (high school degree, AA degree, BA degree, and graduate degree) was obtained and each participant was asked to rate a plan to use a one-quarter cent sale tax increase to expand the bicycle and walking paths in the county. Participants rated the plan on a 1 to 10 scale. The 95% Bonferroni confidence intervals for $\mu_{y|1} - \mu_{y|2}$, $\mu_{y|2} - \mu_{y|3}$, and $\mu_{y|3} - \mu_{y|4}$ are $[-1.24, -0.98]$, $[-0.87, -0.56]$, and $[-0.45, .21]$, respectively. These results indicate that the subpopulation approval means are a monotonically increasing function of education.

1.22 Dichotomizing a Quantitative Predictor Variable

Some researchers will dichotomize a quantitative predictor variable in an effort to simplify the analysis. With a dichotomous predictor variable and a quantitative response variable, the data could be analyzed using an independent samples t -test (see Chapter 2 of Volume 1). However, a test of $H_0: \beta_1 = 0$ using the original quantitative predictor variable can be substantially more powerful than the independent-samples t -test.

In some cases, the loss of power can be justified if the dichotomized predictor variable is more meaningful than the quantitative predictor variable. For example, instead of using the number of cigarettes smoked per day as a predictor variable, it might be more meaningful to classify participants as smokers or nonsmokers. Or instead of using a quantitative measure of depression, a clinical threshold could be used to classify participants as depressed or non-depressed.

1.23 Psychometric Validity

In the physical sciences, there is no question that a ruler measures the attribute of “length” or a clock measures the attribute of “time”. In contrast, there could be serious doubts that a psychological scale of an attribute such as intellectual ability, personality, interests, or attitudes, is actually measuring what it claims to measure. *Psychometric validity* is an assumption that a psychological scale is related to the psychological attribute it claims to measure.

The psychometric validity of a psychological scale can never be determined exactly, and it is misleading to claim that some measurement is “valid”. It is only possible to

accumulate various kinds of validity evidence, and only after sufficiently compelling evidence has been obtained can the researcher cautiously claim that the psychological measurement has its intended interpretation. One type of psychometric validity evidence is referred to as *construct validity*.

Construct validity evidence can be obtained from a confidence interval for a correlation between the scale in question (x) and the scale of some other psychological attribute (y). The magnitude of the correlation between y and x , assessed using a confidence interval, should be consistent with theoretical expectations. For example, a measure of “self-esteem” is expected to have a moderately large positive correlation with a measure of “confidence” and a weaker positive correlation with a measure of “optimism”. Substantial subject-matter expertise is required to specify an attribute that should have a small, moderate, or large correlation with the scale in question.

Some psychologists use the term “discriminant validity” evidence to describe a correlation between y and x that is correctly predicted to be small. The term “convergent validity” evidence is used to describe a correlation with between y and x that is correctly predicted to be moderate or large. Note that in a test of $H_0: \rho_{yx} = 0$, a significant result is not evidence of convergent validity and a nonsignificant result is not evidence of discriminant validity.

Criterion validity evidence can be obtained from a confidence interval for a correlation between x and y where x is the scale in question and y is some objective measure of the attribute that x claims to measure. For example, a researcher could examine the Pearson correlation between a questionnaire (x) designed to measure “job aptitude” and subsequent job performance (y). Larger correlations between the scale in question and the criterion provide stronger evidence of criterion validity. In a simple linear regression model, confidence intervals for β_1 and τ_e can also provide useful criterion validity evidence with larger values of $|\beta_1|$ and smaller values of τ_e providing greater evidence of criterion validity.

1.24 Sample Size Requirements for Desired Precision

Larger sample sizes give narrower confidence intervals, and it is possible to approximate the sample size that will give the desired confidence interval width (upper limit minus

lower limit) with a desired level of confidence. In the fixed- x model, the required total sample size to estimate β_1 with desired precision and confidence is approximately

$$n = 4(\tilde{\sigma}_e^2/\sigma_x^2)(z_{\alpha/2}/w)^2 + 1 + z_{\alpha/2}^2/2 \quad (1.29)$$

where $\tilde{\sigma}_e^2$ is a planning value of the within-group error variance and σ_x^2 is the known variance of the researcher-selected x values. Planning values are obtained from published reports, pilot studies, or expert opinion (see section 1.20 of Volume 1 for more details).

Equation 1.29 shows that larger sample sizes are required with higher levels of confidence, narrower desired confidence interval widths, a larger within-group error variance, and a smaller range of x values. Using a wider range of x values (which increases σ_x^2) will result in a smaller sample size requirement. Sample size formulas do not generally give an integer value for n and it is customary to always round the result up to the nearest integer. The `size.ci.slope` function in the `statpsych` package will compute Equation 1.29.

Example 1.9. A researcher wants to assess the relation between grade level and the number of texts a student sends per day. Students will be randomly selected from grades, 5, 7, and 9 ($\sigma_x^2 = 2.66$). The results of a pilot study were used to set $\tilde{\sigma}_e^2 = 300$ texts. The researcher would like to obtain a 95% confidence interval for the population slope that has a width of 5 texts. The required total sample size is about $n = 4(300/2.66)(1.96/5)^2 + 1 + 1.96^2/2 = 72.1 \approx 73$ and the researcher will obtain a random sample of 24 students from each of the three grade levels.

In the fixed- x model, the required total sample size to estimate $\mu_{y|\dot{x}}$ with desired precision and confidence is approximately

$$n = 4[\tilde{\sigma}_e^2\{1 + (\dot{x} - \mu_x)^2/\sigma_x^2\}](z_{\alpha/2}/w)^2 + 1 + z_{\alpha/2}^2/2 \quad (1.30)$$

where $\dot{x}$ is the value of x at which the mean of y will be estimated, and μ_x is the known mean of the fixed x values. Larger sample sizes are needed to estimate $\mu_{y|\dot{x}}$ for $\dot{x}$ values that are further from μ_x . Like Equation 1.29, using a wider range of x values will result in a smaller sample size requirement. The `size.ci.condmean` function in the `statpsych` package will compute Equation 1.30.

Example 1.10. The EPA estimates that lead in drinking water is responsible for more than 500,000 new cases of learning disabilities in children each year. A researcher wants to estimate the mean drinking water lead concentration in San Francisco apartment buildings that were built in the 1920s, 1930, 1940s, and 1950s using a random sample of buildings from each decade. Based on results from a previous study, $\tilde{\sigma}_e^2$ was set at 120. The researcher believes that lead concentration is linearly related to the year the apartment was built. The researcher will use $x_1 = 1920$, $x_2 = 1930$, $x_3 = 1940$, and $x_4 = 1950$ where $\mu_x = 1935$ and $\sigma_x^2 = 125$. The researcher would like to obtain a 95% confidence interval for the population mean lead concentration (in ppb) at each decade with a width of 5 ppb. For 1920 and 1950 (where $(x^* - \mu_x)^2 = 15^2$), the approximate total sample size requirement is $n = 4[120(1 + 15^2/125)](1.96/5)^2 + 1 + 1.96^2/2 = 209.4 \approx 210$, or about 53 apartment buildings in each age group. This sample size provides desired precision in 1920 and 1950 and better than desired precision for any year between 1920 and 1950.

In the random- x model, a sample size formula for estimating β_1 with desired precision is not very useful because it requires the researcher to specify a planning value for the correlation between x and y as well as planning values for the variances for both x and y , which could be difficult to specify. In a random- x model, ρ_{yx} is often the parameter of primary interest. The required sample size to estimate ρ_{yx} with desired precision and confidence is approximately

$$n = 4(1 - \tilde{\rho}_{yx}^2)^2(z_{\alpha/2}/w)^2 + 3 \quad (1.31)$$

where $\tilde{\rho}_{yx}$ is a planning value of the Pearson correlation between variables y and x . A smaller value of $\tilde{\rho}_{yx}^2$ will give a larger sample size requirement. The researcher typically obtains a range of possible planning values for the Pearson correlation from previous research, and using the planning value closest to zero gives a conservatively large sample size requirement. The `size.ci.cor` function in the `statpsych` package computes Equation 1.31 as an initial approximation and then applies an adjustment for greater accuracy.

Example 1.11. A researcher wants to assess the Pearson correlation between social skills and job performance in a study population of 2,850 service technicians who work for a large international company. Job performance ratings are available for all 2,850 employees, but the social skills assessment is costly to administer and will be given only to a random sample of employees. Previous research in related areas suggests that the Pearson correlation between social skills and job performance could be as low as .3 and the researcher decided to use .3 as the correlation planning value. The researcher would like to estimate the population Pearson correlation with 95% confidence and would like the 95% confidence interval to have a width of about 0.2. The approximate sample size requirement is $n = 4(1 - .3^2)^2(1.96/0.2)^2 + 3 = 321.1 \approx 322$.

The `size.ci.pbcor` function in the `statpsych` package computes the sample size required to obtain a confidence interval for a population point-biserial correlation with desired precision. This function requires a planning value for the proportion of population members who belong to one of the two subpopulations.

If sampling can be performed in two stages, the confidence interval width obtained in the first-stage sample can be used to determine how many additional participants are needed in the second stage to obtain a confidence interval with the desired width (see Equation 1.18 in Volume 1). The `size.ci.second` function in the `statpsych` package will compute a second-stage sample size requirement. Note that this technique does not guarantee a desired confidence interval width in the combined sample.

Example 1.12. A 95% confidence interval for η^2 had a width of .46 in a study with 25 participants. The results of this study are not very informative because the confidence interval is too wide. The researcher would like to obtain a 95% confidence interval for η^2 that has a width of .2. To achieve this goal, the number of additional participants required is $[(.46/.2)^2 - 1]25 = 107.3 \approx 108$.

1.25 Sample Size Requirements for Desired Power

Increasing the sample size will increase the power of the test, and increasing the power of the hypothesis test will decrease the probability of obtaining an inconclusive result. In the fixed- x model, the required sample size to test $H_0: \beta_1 = h$ for a given α value and with desired power is approximately

$$n = (\tilde{\sigma}_e^2 / \sigma_x^2) (z_{\alpha/2} + z_\beta)^2 / (\tilde{\beta}_1 - h)^2 + 1 + z_{\alpha/2}^2 / 2 \quad (1.33)$$

where $\tilde{\beta}_1$ is the anticipated value of β_1 , and z_β is a critical one-sided z -value for $\beta = 1 - \text{power}$ (β is the symbol, not to be confused with a slope coefficient, that is typically used to denote the probability of obtaining an inconclusive result). Equation 1.33 shows that greater desired power (smaller values of β), smaller α levels, and smaller values of $(\tilde{\beta}_1 - h)^2$ require larger sample sizes. The `size.test.slope` in the `statpsych` package will compute Equation 1.33.

The required sample size to test $H_0: \rho_{yx} = h$ in a directional two-sided test for a given α value and with desired power is approximately

$$n = (z_{\alpha/2} + z_\beta)^2 / (\tilde{\rho}_{yx}^* - h^*)^2 + 3 \quad (1.34)$$

where $\tilde{\rho}_{yx}^*$ is a Fisher transformation of a planning value for ρ_{yx} , and h^* is a Fisher transformation of h . Recall that a test of $H_0: \beta_1 = 0$ is equivalent to a test of $H_0: \rho_{yx} = 0$ and Equation 1.34 is easier to implement than Equation 1.33 in this special case. The `size.test.cor` function in the `statpsych` package will compute Equation 1.34. In applications where the sample size is predetermined, the `power.cor` function in the `statpsych` package will estimate the power of the test for the given sample size, α value, and correlation planning value.

Example 1.13. A researcher wants to test the null hypothesis that reaction time is unrelated to recall accuracy ($H_0: \rho_{yx} = 0$) in a college student population with $\alpha = .05$ and power of .95. The researcher believes that the population Pearson correlation is about -.5. The approximate sample size requirement is $n = (1.96 + 1.65)^2 / (-.549 - 0)^2 + 3 = 46.2 \approx 47$.

The required sample size to perform an interval test of $H_1: |\rho_{yx}| < h$ for a given α value and with desired power is approximately

$$n = (z_\alpha + z_{\beta/2})^2 / (h^* - |\tilde{\rho}_{yx}^*|)^2 + 3 \quad (1.35)$$

where $\tilde{\rho}_{yx}^*$ is a Fisher transformation of a planning value for ρ_{yx} , h^* is a Fisher transformation of h , and $|\rho_{yx}|$ must be less than h . Note that z_α is a critical one-sided z-value and $z_{\beta/2}$ is a critical two-sided z-value. In this type of interval hypothesis test, power is the probability of correctly accepting $H_1: |\rho_{yx}| < h$. The sample size requirement for an interval test can be prohibitively large if h is set to a small value. The `size.interval.cor` function in the `statpsych` package will compute Equation 1.35.

Example 1.14. A researcher wants to show that a new measure of academic self-efficacy is at most weakly related to a measure of optimism, and the researcher believes the population Pearson correlation is about .1. The researcher wants to test $H_1: |\rho_{yx}| < .3$ against $H_2: |\rho_{yx}| \geq .3$ with $\alpha = .05$ and power of .9. With a correlation planning value of $\tilde{\rho}_{yx}^* = .1$, the approximate sample size requirement is $n = (1.65 + 1.65)^2 / (.31 - .10)^2 + 3 = 249.9 \approx 250$.

1.26 Specifying Planning Values

The correlation planning value that is used to approximate the sample size required to estimate ρ_{yx} with a desired confidence interval width (Equation 1.31) is a subjective estimate of the *sample* correlation that is likely to be observed in the planned study. The correlation planning value that is used to approximate the sample size required to test a

hypothesis regarding the value of ρ_{yx} with desired power (Equation 1.34) is a subjective estimate of the *population* correlation. In practice, the researcher will not know the value of the population correlation (needed for Equation 1.34) or what the sample correlation will happen to be in a planned study (needed for Equation 1.31).

Subjective correlation planning values can be obtained from expert opinion, pilot studies, or a review of published studies that have used the same y and x variables that will be used in the planned study. If a prior study has used the same y and x variables that will be used in the planned study, and the prior study has sampled from a study population that is similar to the study population in the planned study, the sample correlation from the prior study could be used to specify the correlation planning value in Equations 1.31 and 1.34. However, the sample correlation from the prior study will contain sampling error and could overstate the value of the population correlation required in Equation 1.34. When using Equation 1.34, we could compute a confidence interval for the population correlation in the prior study and set the correlation planning value to the lower limit or upper limit that is closest to 0. Unless the sample size in the prior study is large, an 80% or 90% confidence interval could be used to determine the correlation planning value.

Example 1.15. A job satisfaction questionnaire and a leadership toxicity questionnaire will be given to a random sample of employees at a large public university. The estimated correlation between job satisfaction and leadership toxicity was $-.56$ at a similar public university using sample of 120 employees. The researcher used this correlation to compute a 90% confidence interval and obtained $[-.65, -.44]$. The researcher set the correlation planning value at $-.44$ (the Fisher transformed value is $-.472$). To test $H_0: \rho_{yx} = 0$ with $\alpha = .05$ and a desired power of $.95$, the required sample size is approximately $n = (1.96 + 1.65)^2 / (-.472 - 0)^2 + 3 = 61.5 \approx 62$. Note that a large sample size is not needed to reject $H_0: \rho_{yx} = 0$ if the population correlation is actually $-.44$. A larger sample size is needed to accurately estimate the population correlation with desired precision as illustrated in Example 1.16.

The correlation planning value in Equation 1.31 should be a prediction of what the *sample* correlation will be in the planned study. If the sample correlation from a prior study is used as a correlation planning value in Equation 1.31, the sample correlation in the planned study could be smaller than the correlation planning value and then the confidence interval in the planned study will be wider than desired. The sample correlation from a prior study can be used to predict the value of the sample correlation

in a planned study using the following prediction interval, which would be reverse transformed using Equations 1.21a and 1.21b

$$(\hat{\rho}_0^* - \frac{\hat{\rho}_0}{2(n_0 - 1)}) \pm z_{\alpha/2} \sqrt{\frac{1}{n_0 - 3} + \frac{1}{n - 3}} \quad (1.36)$$

where n_0 is the sample size in the prior study, $\hat{\rho}_0$ is the sample correlation in the prior study, and n is an initial approximation to the required sample size in the planned study. The `pi.cor` function in the `statpsych` package can compute Equation 1.36 and the reverse transformations. Unless the sample size in the prior study is large, the confidence level could be set to .10 or .20 in Equation 1.36 rather than the traditional .05. The initial sample size approximation in Equation 1.36 is obtained using Equation 1.31 with a correlation planning value set to the confidence interval limit that is closest to 0. Equation 1.31 is recomputed using the prediction limit from Equation 1.36 that is closest to 0 to obtain a revised sample size approximation. These steps can be repeated a few times to obtain the final sample size requirement. The `size.ci.cor.prior` function in the `statpsych` package will compute these multi-step computations.

Example 1.16. The researcher in *Example 1.15* also wants to obtain a 95% confidence interval for the population correlation between job satisfaction and leadership toxicity that has a width of about .2. The 90% confidence interval for the population correlation in the prior study is [-.654, -.445]. Using -.445 as an initial correlation planning value gives an initial sample size requirement of 250 for a 95% confidence and a width of .2. The prediction limit was computed using the prior sample correlation of -.56 and an initial sample size of 250. The 90% prediction limit is [-.672, -.419]. Using -.419 as the revised correlation planning value gives a revised sample size requirement of 264 for 95% confidence and a width of .2. The prediction interval was recomputed for a planned sample size of 264, which gave a revised correlation planning value of -.420. The final sample size requirement is 263 for 95% confidence and a width of .2.

The sample size required to estimate a population slope or conditional mean with desired precision (Equations 1.29 and 1.30), and the sample size required to test a hypothesis about a population slope with desired power (Equation 1.33) require a variance planning value ($\tilde{\sigma}_e^2$). The strategies for specifying a variance planning value described in section 1.31 of Volume 1 can be used to specify the value of $\tilde{\sigma}_e^2$. In addition to specifying a variance planning value, Equation 1.33 also requires an effect size planning value ($\tilde{\beta} - h$). Two different approaches can be used to specify the effect size planning value in Equation 1.33. One approach sets the planning value of the slope to its most likely value

based on expert opinion or prior research. Another approach sets the value of $|\tilde{\beta} - h|$ to the smallest value that would represent a useful or interesting effect size (referred to as the "minimally interesting effect size").

1.27 Assumptions

A $100(1 - \alpha)\%$ confidence interval will include the unknown value of a population parameter in about $100(1 - \alpha)\%$ of all possible samples of size n if certain assumptions have been satisfied. However, if certain assumptions are not satisfied, a $100(1 - \alpha)\%$ confidence interval might include the population parameter value in much less than $100(1 - \alpha)\%$ of all possible samples, and thus a $100(1 - \alpha)\%$ confidence interval computed from any one sample would be interpreted with excessive confidence.

The confidence interval and hypothesis testing methods for β_1 , $\mu_{y|x}$, η^2 , and τ_e are based on the following five assumptions.

- 1) The sample is a random sample from the study population (*random sampling assumption*).
- 2) The n participants respond independently of one another (*independence assumption*).
- 3) The relation between y and x is linear (*linearity assumption*).
- 4) The variability of the prediction errors is constant across the values of x (*prediction error homoscedasticity assumption*).
- 5) The prediction errors have an approximate normal distribution in the study population (*prediction error normality assumption*).

Violating the random sampling assumption can render hypothesis testing and confidence interval results uninterpretable. Violating the independence assumption can result in p -values that are too small and confidence intervals that are too narrow. Violating the linearity assumption can affect the interpretability of the parameters. For example, β_1 describes the change in y for a 1-unit increase in x only if the relation between y and x is linear. Unless the sample size is small, the test and confidence interval for β_1 should perform properly with moderate violations of the prediction error normality assumption. Any negative effects of prediction error nonnormality are reduced with larger sample sizes. Violating the homoscedasticity assumption can result in p -values that are too small and confidence intervals that are too narrow regardless of the sample size. Data

transformations (section 1.28) can sometimes reduce heteroscedasticity and prediction error nonnormality. A general method for dealing with heteroscedasticity is discussed in Chapter 2.

The confidence intervals for ρ_{yx} and ρ_{yx}^2 assume random sampling, independence among participants, and *bivariate normality*. Bivariate normality implies linearity, prediction error homoscedasticity, and prediction error normality. Bivariate normality also implies normality of the y scores and normality of the x scores. If any of the implications of bivariate normality cannot be satisfied, then the bivariate normality assumption cannot be satisfied. The linearity and equal prediction error variance assumptions can be visually assessed by examining a scatterplot of x and y . Skewness and kurtosis estimates of the residuals are helpful in assessing the prediction error normality assumption. To assess the bivariate normality assumption, also examine the skewness and kurtosis estimates of the y scores and x scores.

The performance of the confidence interval for ρ_{yx} can be unsatisfactory when the bivariate normality assumption is violated. The effect of violating the bivariate normality assumption is more pronounced for larger values of $|\rho_{yx}|$ regardless of the sample size. A violation of the bivariate normality assumption has almost no negative effects on a directional two-sided test when the null hypothesis is $H_0: \rho_{yx} = 0$. The `sim.ci.cor` function in the `statpsych` package can be used to assess the performance of the confidence interval for ρ_{yx} for different distributions of y and x , different sample sizes, and different values of ρ_{yx} .

1.28 Outliers

Unequal variances and nonnormality are sometimes due to one or more participants who have unusually small or large y scores or x scores. In some studies, it is an unusual combination of y scores and x scores that can lead to problems. Participants with unusual y scores or unusual x scores are referred to as *univariate outliers*. Participants with unusual combinations of y and x scores are referred to as *bivariate outliers*.

Sometimes the outlier is the result of a data entry error (e.g., a .73 was incorrectly entered as 73). Of course, all data entry errors must be corrected. An unusual point on the scatterplot might also be the result of a measurement device malfunction, a participant who did not understand the instructions, or a participant who does not belong to the

study population of interest. In these situations, it is appropriate to remove the outlying participant from the analysis. If a participant is removed from the analysis, the researcher must disclose this information along with the justification.

Bivariate outliers can increase or decrease the estimated slope and correlation. In the examples below, the bivariate outlier in the upper right corner of the first scatterplot has increased the correlation. The bivariate outlier in the lower right corner of the second scatterplot has decreased the correlation.

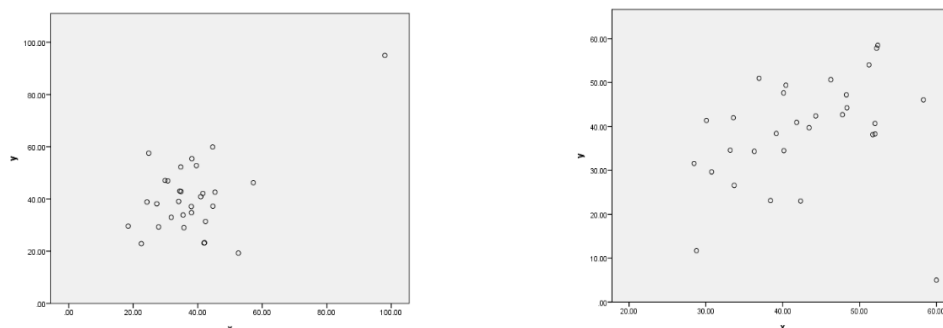


Figure 1.7 Scatterplots with outliers. Left: $\hat{\rho}_{yx} = .1$ without outlier and $\hat{\rho}_{yx} = .6$ with outlier. Right: $\hat{\rho}_{yx} = .2$ with outlier and $\hat{\rho}_{yx} = .6$ without outlier

1.29 Data Transformations

If the scatterplot suggests a nonlinear relation between y and x , transforming x could help linearize the relation. Logarithmic, square-root, or reciprocal transformations of the x scores are commonly used for this purpose. For example, in the scatterplot below (left), the relation between x and y appears to be nonlinear. The relation between $\sqrt{x}$ and y in the scatterplot below (right) appears to be approximately linear. In some studies, it will be necessary to transform both y and x to obtain approximate linearity.

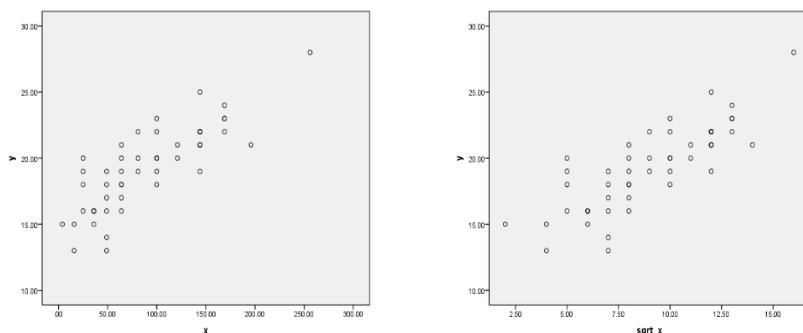


Figure 1.8 Scatterplot of x and y (left). Scatterplot of $\sqrt{x}$ and y (right)

If the residuals show nonnormality or if the variance of y appears to increase with greater values of x (which would be a symptom of unequal prediction error variance), then a logarithmic, square root, or reciprocal transformation of the y scores might be helpful. A square root transformation could be helpful if y_i is a frequency count for participant i (number of cigarettes smoked per day, number of facts correctly recalled, etc.), and a logarithmic or reciprocal transformation could be helpful if y_i is a time measurement (e.g., reaction time, time to complete a task, etc.) for participant i . In some studies, it will be necessary to transform both y and x to reduce residual nonnormality or unequal residual variability.

The scatterplot below (left) shows increasing variability of y with greater values of x . A log transformation of y appears to have corrected the unequal variance problem as shown in the scatterplot below (right).

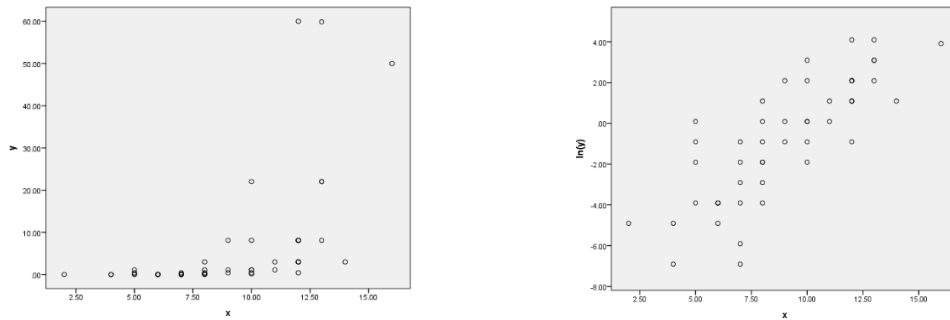


Figure 1.8 Scatterplot of x and y (left). Scatterplot of x and $\ln(y)$ (right)

If y or x is log-transformed, the slope coefficient (β_1) takes on a special meaning. If $\ln(x)$ is used to predict y , a 1% increase in x is associated with a $\ln(1.01)\beta_1$ change in y and a $k\%$ increase in x is associated with a $\ln(1 + k/100)\beta_1$ change in y . If x is used to predict $\ln(y)$, a 1-unit increase in x is associated with a $100(e^{\beta_1} - 1)\%$ change in y , and a k -point increase in x is associated with a $100(e^{k\beta_1} - 1)\%$ change in y . If $\ln(x)$ is used to predict $\ln(y)$, a 1% increase in x is associated with a $100(1.01^{\beta_1} - 1)\%$ change in y , and a $k\%$ increase in x is associated with a $100[(1 + \frac{k}{100})^{\beta_1} - 1]\%$ change in y .

In some cases, the researcher will hypothesize a specific type of nonlinear relation between y and x . For example, suppose the strength of a physical stimulus (x) is believed to predict the perceived magnitude of the sensation (y) by the nonlinear relation $y = ax^b$ where a and b are unknown values that the researcher wants to estimate for a particular

type of stimulus. This nonlinear relation can be expressed in terms of the following linear statistical model with transformed variables

$$\ln(y_i) = \beta_0 + \beta_1 \ln(x_i) + e_i \quad (1.36)$$

where $\beta_0 = \ln(a)$ and $\beta_1 = b$. Or suppose that performance on a complex task (y) is believed to be related to hours of practice (x) by the following nonlinear relation $y = \frac{x}{b + ax}$ where a and b are unknown values that the researcher wants to estimate for a particular type of task. This nonlinear relation can be expressed in terms of the following linear regression model with transformed variables

$$1/y_i = \beta_0 + \beta_1(1/x_i) + e_i \quad (1.37)$$

where $\beta_0 = a$ and $\beta_1 = b$.

1.30 Extrapolation Issues

In fixed- x models, the interpretation of confidence intervals and hypothesis tests should be restricted to the range of x values that were used in the sample. Generalizing the results to other values of x that are outside the range of the sample x values is called *extrapolation* and should be done very cautiously and with arguments to justify a claim that all assumptions will extend to these other x values. Extrapolating to values of x that are only slightly less than the smallest sample x value or slightly greater than the largest sample x value, could be easily justified in some applications. Extrapolating to x values that are far outside the range of sample x values is difficult to justify in most applications. In random- x models, confidence interval and hypothesis testing results hold for all values of x and y in the study population if the bivariate normality assumption is plausible because the bivariate normality assumption implies a linear relation between y and x for all values of y and x in the study population.

1.31 Measurement Error

In Chapter 4 of Volume 1, measurement error was defined as the difference between a true score and an observed score for some attribute. A reliability coefficient was defined as one way to quantify the amount of measurement error. A reliability coefficient has a range of 0 to 1 where a value of 1 indicates no measurement error.

Measurement error in y or x will attenuate the Pearson correlation. The amount of attenuation in ρ_{yx} is

$$\rho_{yx} = \rho_{yx}^{\#} \sqrt{\rho_y \rho_x} \quad (1.38)$$

where ρ_y is the reliability of the y scores, ρ_x is the reliability of the x scores, and $\rho_{yx}^{\#}$ is the Pearson correlation between the true x scores and the true y scores. For example, if the reliability of the y measurements is .6 and the reliability of the x measurements is .7, the correlation between the observed x scores and observed y scores would be about $\sqrt{.6 \times .7} = .648$ times as large as the correlation between the true x scores and the true y scores.

In a random- x model, measurement error in x will attenuate the slope coefficient. The amount of attenuation in β_1 is

$$\beta_1 = \beta_1^{\#} \rho_x \quad (1.39)$$

where $\beta_1^{\#}$ is the slope coefficient for the line relating the true x scores to the observed y scores. Measurement error in the x scores in a fixed- x model is also possible. For example, in an experiment where participants are given a cognitive performance task in one of four room temperature conditions (60° F, 70° F, 80° F, and 90° F), the actual room temperature could vary randomly across participants within each of the four treatment conditions. However, this specific type of measurement error will not attenuate the slope coefficient.

Measurement error in y will attenuate η^2 , δ , and ρ_{pb} as shown below

$$\eta^2 = \eta^{\#2} \rho_y \quad (1.40)$$

$$\delta = \delta^{\#} \sqrt{\rho_y} \quad (1.41)$$

$$\rho_{pb} = \rho_{pb}^{\#} \sqrt{\rho_y} \quad (1.42)$$

where $\eta^{\#2}$ is the proportion of true y score variance that is predicable from x , $\delta^{\#}$ is a standardized mean difference in true y scores, and $\rho_{pb}^{\#}$ is the point-biserial correlation between a dummy coded variable and the true y scores.

In fixed- x or random- x models, measurement error in y inflates the prediction error variance which increases the confidence interval widths for β_1 and $\mu_{y|\dot{x}}$. An increase in the prediction error variance decreases the power of tests for β_1 and $\mu_{y|\dot{x}}$. Measurement error in y also increases the prediction interval width for $\hat{y}$.

Small correlations and slopes are common in psychological research, and measurement error is partly to blame. If a large-sample estimate of ρ_y or ρ_x is available, correlation and slope estimates can be disattenuated using the above attenuation formulas. For example, an estimate of a disattenuated Pearson correlation is $\hat{\rho}_{yx}^{\#} = \hat{\rho}_{yx} / \sqrt{\hat{\rho}_y \hat{\rho}_x}$, and an estimate of a disattenuated slope coefficient is $\hat{\beta}_1^{\#} = \hat{\beta}_1 / \hat{\rho}_x$. Confidence intervals for population disattenuated correlations and slopes are discussed in Volume 4.

As illustrated in Equations 1.38 to 1.42, the detrimental effects of measurement error are most serious when the reliability of a variable is low. When analyzing any variable that contains measurement error, it is important to know its reliability so that the results of a statistical analysis can be interpreted with the appropriate degree of caution. The within-subjects ANOVA (see section 4.19 of Volume 1) can be used to estimate the reliability of a single measurement or the reliability of a sum (or average) of two or more measurements. If two equally reliable measurements (y_1 and y_2) of the same attribute are obtained from a random sample, the estimated Pearson correlation between y_1 and y_2 is also an estimate of the reliability of a single measurement. The two measurements could be scores for two forms of a test (to assess alternative-form reliability), scores on the same test or questionnaire on two occasions (to assess test-retest reliability), or the ratings from two raters (to assess interrater reliability). A confidence interval for the population Pearson correlation (section 1.14) can be used to obtain a confidence interval for the population reliability of a single measurement.

If the two measurements are ratings from two different raters, some studies will use the average rating in a statistical analysis because the average rating will be more reliable than a single rating. The Spearman-Brown formula (see section 4.19 of Volume 1) can be used to estimate the reliability of an average of two equally reliable ratings using the following formula

$$\rho_2 = 2\rho_1 / (1 + \rho_1) \quad (1.43)$$

where ρ_1 is the reliability of a single rating (the Pearson correlation) and ρ_2 is the reliability of an average of two ratings. If the variances of y_1 and y_2 are equal, then ρ_2 is equal to Cronbach's alpha reliability that was defined in section 4.19 of Volume 1; otherwise, ρ_2 is equal to Cronbach's alpha for standardized variables. A confidence interval for ρ_2 can be obtained by first computing a confidence interval for ρ_1 and then plugging the lower and upper confidence interval limits into Equation 1.43.

Example 1.17. A researcher obtained a random sample of 200 episodes of several different children's television shows. Two raters independently rated each episode on a 1 to 10 scale in terms of the amount and degree of gender stereotyping. The average rating will be used as a predictor variable in a regression analysis. The estimated Pearson correlation between the two ratings is .820 with a 95% confidence interval of [.768, .860]. The researcher is 95% confident that the reliability of a single rater is between .768 and .860. The reliability of the average rating is $2(.82)/(1 + .82) = .901$ with a 95% confidence interval of [.869, .925]. The researcher is 95% confident that the reliability of the average rating is between .869 and .925.

Although measurement error can be a serious problem, measurement error is not a problem in regression models where an observed x score is used to predict an observed y score. For example, a student's ACT score might be used to predict that student's college GPA. In these applications, β_1 (and not $\beta_1^\#$) is the appropriate slope coefficient.

1.32 Distribution-free Methods

Correlation. The Pearson correlation describes the magnitude of a *linear* relation between x and y . A *Spearman* correlation can be used to describe the magnitude of a *monotonic* relation between two quantitative variables. In a perfect positive monotonic relation, the value of y will increase but never decrease as the value of x is increased. In a perfect negative monotonic relation, the value of y will decrease but never increase as x is increased. A scatterplot of a less than perfect positive monotonic relation is given below.



Figure 1.9 Scatterplot of a positive monotonic relation between x and y

The Pearson correlation can understate the relation between x and y if the relation is monotonic rather than linear. For example, in the above scatterplot where the linearity assumption appears suspect, the Pearson correlation is .48 and the Spearman correlation is .51. The Spearman correlation is also less affected by outliers than a Pearson correlation.

A Spearman correlation is computed by replacing the x scores with their rank orders, replacing the y scores with their rank orders, and then computing a Pearson correlation on the rank transformed scores. For example, in a sample of $n = 5$, the x scores $x_1 = 3.1$, $x_2 = 2.9$, $x_3 = 5.7$, $x_4 = 1.5$, $x_5 = 10.5$ and y scores $y_1 = 9.2$, $y_2 = 9.8$, $y_3 = 12.2$, $y_4 = 4.5$, $y_5 = 25.3$ would be replaced with the rank-transformed scores $x_1 = 3$, $x_2 = 2$, $x_3 = 4$, $x_4 = 1$, $x_5 = 5$ and $y_1 = 2$, $y_2 = 3$, $y_3 = 4$, $y_4 = 1$, $y_5 = 5$.

The estimated Spearman correlation is denoted as $\hat{\rho}_s$, and an approximate standard error for $\hat{\rho}_s$ is

$$SE_{\hat{\rho}_s} = \sqrt{(1 - \hat{\rho}_s^2)^2(1 + \hat{\rho}_s^2/2)/(n - 3)}. \quad (1.44)$$

A confidence interval for the population Spearman correlation (ρ_s), which does not assume bivariate normality of the x and y scores, is obtained by first computing

$$\hat{\rho}_s^* \pm z_{\alpha/2} SE_{\hat{\rho}_s} / (1 - \hat{\rho}_s^2) \quad (1.45)$$

where $\hat{\rho}_s^*$ is a Fisher-transformed Spearman correlation. The upper and lower limits in Equation 1.45 are then reverse transformed using Equations 1.21a and 1.21b to give a confidence interval for a population Spearman correlation. The `ci.spear` function in the `statpsych` package will compute a confidence interval for a population Spearman correlation. The `size.ci.spear` function in the `statpsych` package will compute the sample size required to obtain a confidence interval for a population Spearman correlation with desired precision. The `sim.ci.spear` function in the `statpsych` package can be used to assess the performance of the confidence interval for ρ_s for different distributions of y and x , different sample sizes, and different values of ρ_s .

The Spearman correlation is also used to describe the relation between two variables that have ordered categories with a small number of categories (e.g., strongly disagree, disagree, agree, strongly agree), but Equation 1.44 is not appropriate in these applications and a more complicated standard error formula is required.

Slope. The traditional confidence interval for β_1 (Equation 1.17) assumes approximate normality of the prediction errors and equal prediction error variance across the values of x . If either of these two assumptions are violated, Equation 1.17 can have a coverage probability that is less than $1 - \alpha$. The *Theil-Sen* confidence interval for β_1 does not assume normality of the prediction errors or equality of prediction error variance. The Theil-Sen

confidence interval is attractive because it will have a coverage probability close to $1 - \alpha$ with nonnormality and unequal variance of the prediction errors and will be only slightly wider than Equation 1.17 when all of the assumptions of Equation 1.17 have been satisfied.

The Theil-Sen confidence interval for β_1 is obtained from "slope scores" for all $n(n - 1)/2$ pairs of participants, where the slope score for participants i and j is $(y_i - y_j)/(x_i - x_j)$. The Theil-Sen estimate of β_1 is the median of the $n(n - 1)/2$ slope scores ignoring any pair of participants for which $x_i = x_j$. The Theil-Sen confidence interval for β_1 is computed by rank ordering the slope scores from smallest to largest where $b_{(1)}$ is the smallest slope score, $b_{(2)}$ is the next smallest slope score, etc. (ignoring any pair of participants for which $x_i = x_j$). A $100(1 - \alpha)\%$ Theil-Sen confidence interval for β_1 is

$$[b_{(L)}, b_{(U)}] \quad (1.46)$$

where $L = (n(n - 1)/2 - o)/2$, $U = L + o + 1$, $o = z_{\alpha/2} \sqrt{n(n - 1)(2n + 5)/18}$, and o is rounded to the nearest integer.

An approximate standard error for the Theil-Sen estimate of β_1 is

$$(b_{(U)} - b_{(L)})/2z_{\alpha/2}. \quad (1.47)$$

Equation 1.46 is usually computed using $\alpha = .05$. An estimate of the y -intercept is $\hat{\theta}_y - \hat{\beta}_1 \hat{\theta}_x$ where $\hat{\beta}_1$ is the Theil-Sen estimate of β_1 , $\hat{\theta}_y$ is the estimated median of the y scores, and $\hat{\theta}_x$ is the estimated median of the x scores. The `ci.theil` function in the `statpsych` package will compute Equations 1.45 and 1.46.

1.33 Bayesian Approach to Data Analysis

As explained in section 1.32 of Volume 1, Bayesian data analysis is an alternative to classical data analysis. Bayesian methods can combine a subjective prior distribution of plausible population slope values with sample information to obtain a subjective posterior distribution of plausible population slope values. When estimating β_1 , it is common to assume that the prior distribution is a normal distribution and the population prediction errors have a normal distribution. With these two assumptions, it can be shown that the posterior distribution will also have a normal distribution. The Bayesian estimate of β_1 is the mean of the posterior distribution.

Let $\tilde{\beta}_1$ denote the subjective mean of the prior normal distribution and let $\sigma_{\tilde{\beta}_1}$ denote the subjective standard deviation of the prior normal distribution. The sample slope ($\hat{\beta}_1$) is defined in Equation 1.9, and its standard error ($SE_{\hat{\beta}_1}$) is defined in Equation 1.17. The mean of the posterior normal distribution is denoted as $\tilde{\beta}_1^*$ and is equal to

$$\tilde{\beta}_1^* = [(1/\sigma_{\tilde{\beta}_1}^2)\tilde{\beta}_1 + (1/SE_{\hat{\beta}_1}^2)\hat{\beta}_1]/(1/\sigma_{\tilde{\beta}_1}^2 + 1/SE_{\hat{\beta}_1}^2) \quad (1.48)$$

which is the Bayesian estimate of β_1 . The standard deviation of the posterior normal distribution is denoted as $\sigma_{\tilde{\beta}_1^*}$ and is equal to

$$\sigma_{\tilde{\beta}_1^*} = \sqrt{\frac{1}{1/\sigma_{\tilde{\beta}_1}^2 + 1/SE_{\hat{\beta}_1}^2}}. \quad (1.49)$$

An approximate $100(1 - \alpha)\%$ credible interval for β_1 is

$$\tilde{\beta}_1^* \pm z_{\alpha/2}\sigma_{\tilde{\beta}_1^*} \quad (1.50)$$

and is the Bayesian alternative to the confidence interval for β_1 given in Equation 1.17. The `ci.bayes.normal` function in the `statpsych` package will compute Equations 1.48, 1.49, and 1.50.

The Bayesian estimate of β_1 is biased in small samples and the bias can be large if $\tilde{\beta}_1$ is not similar to β_1 . Furthermore, a 95% credible interval can have a true coverage probability that is far less than .95 unless the sample size is large and $\tilde{\beta}_1$ is similar to β_1 . As can be seen in Equations 1.48 and 1.49, if the sample size is large then $\tilde{\beta}_1^*$ will be similar to $\hat{\beta}_1$, $\sigma_{\tilde{\beta}_1^*}$ will be similar to $SE_{\hat{\beta}_1}$, and then a $100(1 - \alpha)\%$ credible interval will be similar to a $100(1 - \alpha)\%$ confidence interval. Or if a non-informative prior distribution is specified by setting $\sigma_{\tilde{\beta}_1}$ to a large value, then a $100(1 - \alpha)\%$ credible interval will be similar to a $100(1 - \alpha)\%$ confidence interval.

Example 1.18. In *Example 1.1*, 150 students rated their satisfaction with their UCSC education (y) and their satisfaction with the quality of the academic advising (x). The purpose of the study is to estimate the population slope (β_1). Prior to data collection, suppose the researcher has an opinion about the plausible values of β_1 , and the researcher's subjective probabilities for these plausible values can be characterized by a normal distribution with a mean of 2 and a standard deviation of 0.25. After the data are collected, the researcher obtains $\hat{\beta}_1 = 1.70$ and $SE_{\hat{\beta}_1} = 0.54$. This sample information alters the researcher's opinion about the plausible values of β_1 .

(continued)

The researcher's subjective probabilities about the value of β_1 can now be characterized by a normal distribution with a mean of $\tilde{\beta}_1^* = [(1/0.25^2)2 + (1/0.54^2)1.70]/(1/0.25^2 + 1/0.54^2) = 1.95$ and a standard deviation of $\sigma_{\tilde{\beta}_1^*} = \sqrt{\frac{1}{1/0.25^2 + 1/0.54^2}} = 0.052$. The researcher's 95% credible interval for β_1 has a lower limit of $LL = 1.95 - 1.96(0.227) = 1.51$ and an upper limit of $UL = 1.95 + 1.96(0.227) = 2.39$.

The 95% credible interval in Example 1.18 has been influenced by the researcher's subjective prior distribution. A different researcher with a different subjective prior distribution will have a different 95% credible interval when analyzing the same data. The sensitivity of a credible interval to the specification of a subjective prior distribution is one of the main criticisms of the Bayesian approach. This criticism is more compelling in the context of scientific research than applied research. However, the use of a skeptical prior distribution (see section 3.23 of Volume 1) has important applications in scientific research. Suppose, for a specific study, that a value of β_1 that is between -4.0 and 4.0 would be considered by most experts to represent a small or unimportant relation between x and y , and suppose a skeptic is 95% confident that β_1 is between -4.0 to 4.0. For this study, a skeptical prior distribution could be characterized by a normal distribution with a mean of 0 and a standard deviation of $4/1.96 = 2.04$. Suppose the sample value of β_1 is 8.6 with a standard error of 3.8 (95% CI [1.0, 16.2]). Using the `ci.bayes.normal` function gives a Bayesian estimate of β_1 equal to 1.87 with a 95% credible interval of [-1.60, 5.33]. Unlike the 95% confidence interval, the 95% credible interval includes 0 and we cannot conclude that there is a linear relation between x and y . In this example, the strength of the relation between y and x in the sample was not sufficient to overcome our initial skepticism.

A Bayesian credible interval with a skeptical prior distribution is easier to apply to a Pearson correlation than a slope because it is usually easier to specify a correlation value that would be considered to be small or unimportant. In psychological research, a correlation between -.2 and .2 is usually considered to be "small". If a skeptic is 95% confident that the population correlation is between -.2 and .2, a skeptical prior distribution could be characterized by a normal distribution with a mean of 0 and a standard deviation of $.2/1.96 = 1.02$. A normal prior distribution can be used because the sampling distribution of a correlation estimate is approximately normal if the population correlation is small. The `ci.bayes.cor` function will compute a Bayesian credible interval for a Pearson correlation with a skeptical prior distribution.

1.34 Using Information from a Prior Study

Classical methods can be used to combine a parameter estimate from a prior study with a parameter estimate in a new study to obtain a more informative and generalizable confidence interval for the population parameter. The classical approach also provides important information regarding the replicability of the prior results. The sample size in the new study should be at least as large as the sample size in the prior study, and the study population in the new study should be similar to the study population in the prior study.

If a prior study reports the results for a Pearson correlation, that study could be replicated in a follow-up study in an effort to assess replicability and obtain a narrower confidence interval for the Pearson correlation that describes both the prior study population and the follow-up study population. If an estimate of a Pearson correlation has been reported in a prior study, the `replicate.cor` function in the `vcmeta` package will compute confidence intervals for $\rho_{yx[1]}$, $\rho_{yx[2]}$, $\rho_{yx[1]} - \rho_{yx[2]}$, and $(\rho_{yx[1]} + \rho_{yx[2]})/2$ where $\rho_{yx[j]}$ is the population Pearson correlation in study j . The `replicate.spear` function in the `vcmeta` package can be used to combine and compare Spearman correlations.

If an estimate of a slope coefficient and its standard error have been reported in a prior study, that study could be replicated in a different study population using the same response variable and the same predictor variable. The `replicate.slope` function in the `vcmeta` package will compute confidence intervals for $\beta_{1[1]}$, $\beta_{1[2]}$, $\beta_{1[1]} - \beta_{1[2]}$, and $(\beta_{1[1]} + \beta_{1[2]})/2$ where $\beta_{1[j]}$ is the population slope coefficient in study j .

The confidence intervals for $\rho_{yx[1]} - \rho_{yx[2]}$ and $\beta_{1[1]} - \beta_{1[2]}$, can be used to assess equivalence (see section 2.16) and replicability (see section 4.2 of Volume 5).

Example 1.19. A study reported a Pearson correlation of .424 between hours of social media use and anxiety in a sample of 50 public high school students with a 95% confidence interval of [.165, .628]. The study was replicated at another public high school using a sample of 100 students. The estimated Pearson correlation in the follow-up study was .497. The 95% confidence interval for $\rho_{yx[1]} - \rho_{yx[2]}$ is [-.316, .149], which includes 0 and justifies combining the correlations from the prior and follow-up studies. The 95% confidence interval for $(\rho_{yx[1]} + \rho_{yx[2]})/2$ is [.328, .575]. We can be 95% confident that the average Pearson correlation between hours of social media use and anxiety in the two high school study populations is between .328 and .575.

1.35 Data Visualization

A scatterplot of the paired y scores and x scores can be used to visually assess the linearity and homoscedasticity assumptions. The values of the response variable scores are represented on the vertical axis, and the values of the predictor variable scores are represented on the horizontal axis. The scatterplot can also reveal bivariate outliers that might go unnoticed when examining the y scores or x scores separately (see section 1.27). A scatterplot is not effective in revealing nonnormality of the y scores or x scores. Box plots of the y scores and x scores can reveal skewness and univariate outliers. If the sample size is sufficiently large, histograms of the y scores and x scores can illustrate the shape of each distribution. Dot plots of the y scores and x scores are also useful in the data screening process.

As illustrated in the scatterplots of Figure 1.3, it is difficult to distinguish between correlations that are in the 0 to .3 range. Adding a line of best fit can help reveal the magnitudes of small associations. Confidence interval and prediction interval bands can also be added to a scatterplot. In fixed- x models, both confidence interval and prediction interval bands can be informative. In random- x models, prediction interval bands are most useful. A scatterplot with a 95% confidence interval band (left) and a scatterplot with a 95% prediction interval band (right) for the 4-group experiment in Data Analysis Problem 1-1 are shown below.

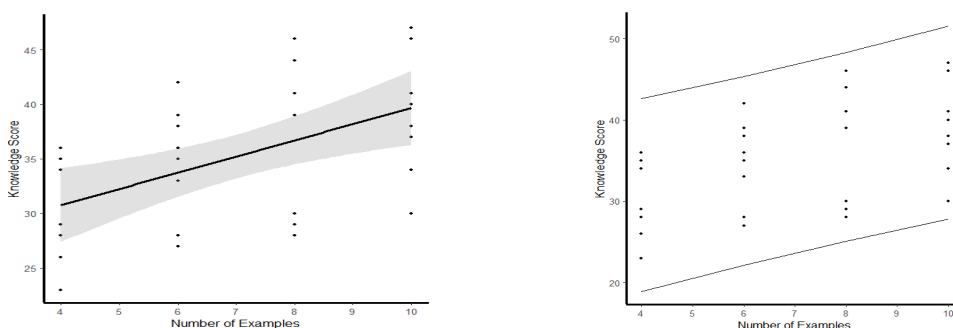


Figure 1.10 Scatterplots with 95% confidence interval band (left) and 95% prediction interval band (right)

In an experimental design, the 95% confidence interval at a particular value of the independent variable is the range of plausible values for the *population mean* of the dependent variable if everyone in the study population had received that particular value of the independent variable. For example, if everyone in the study population had been shown 7 examples, we can be 95% confident that the population mean knowledge score

would be between about 32 and 36. The prediction interval band is much wider than the confidence interval band and represents a range of plausible values for *one* randomly selected participant from the study population. For example, if a one person is selected from the study population and is shown 7 examples, we can be 95% confidence that that this person's knowledge score would be between about 24 to 45.

Although a scatterplot of the y scores and x scores can reveal a violation of the homoscedasticity assumption, a visual assessment of this assumption is easier using a *residual plot*. In a residual plot, the values of the residuals are represented on the vertical axis, and the values of the x scores are represented on the horizontal axis. Ideally, the variability of the residuals should not be noticeably related to the x scores. Other types of residual plots are described in the next chapter that are useful in models with more than one predictor variable. An example of a residual plot is shown below. In this example, there are no clear indications of heteroscedasticity.

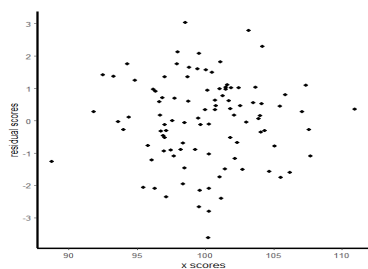


Figure 1.11 Example of a residual scatterplot with no indication of heteroscedasticity

Scatterplots can misrepresent the magnitude of the relation between y and x if there are many tied y scores or many tied x scores. For example, the following scatterplot has been computed for a sample of $n = 200$ participants where each variable has been measured on a 1-9 rating scale. The estimated Pearson correlation between x and y is about .70, but the scatterplot does not accurately illustrate the strength of the relation.

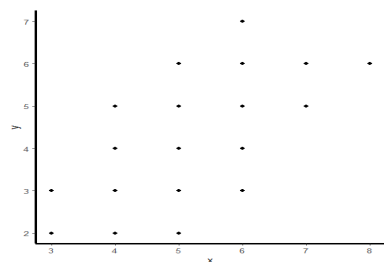


Figure 1.12 Example of a scatterplot with many tied scores

With many tied scores, the relation between the two variables can be more clearly seen by adding *jitter* (a small random value) to each pair of observations. SPSS and R have options to add jitter to a scatterplot. A scatterplot of the above data with jitter illustrates the strength of the relation between y and x more clearly.

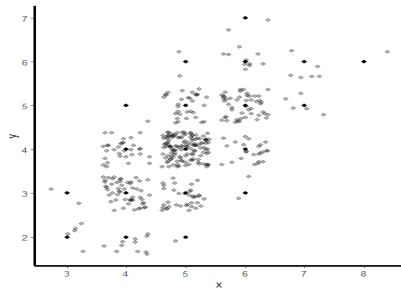


Figure 1.13 Scatterplot of data from Figure 1.12 with added jitter

If the predictor variable is a quantitative classification or treatment factor, the confidence interval chart described in Volume 1 can be used to visualize a trend in means or medians. A confidence interval chart for a quantitative factor can include lines that connect the points to better illustrate the trend. An example of a confidence interval chart with connecting lines is shown below for a 4-group experiment where participants received 10, 15, 20, or 25 minutes of study time and were later asked to recall the studied material.

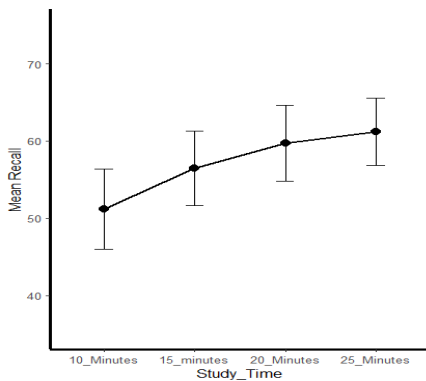


Figure 1.14 Example of a confidence interval chart for a quantitative factor

Key Terms

scatterplot
predictor variable
Pearson correlation
covariance
quantitative factor
qualitative factor
slope
 y -intercept
predicted y score
prediction error
conditional mean
mean centering
prediction error variance
MAPE
coefficient of determination
simple linear statistical model
random- x model
fixed- x model
residuals
ordinary least squares (OLS) estimates
Fisher transformation
dummy coded predictor variable
point-biserial correlation
interval hypothesis test
monotonicity test
criterion validity
construct validity
linearity assumption
equal prediction error variance assumption
prediction error normality assumption
bivariate normality assumption
outlier
data transformation
extrapolation
Spearman correlation
Theil-Sen confidence interval

Concept Questions

1. What information does a sample scatterplot provide?
2. What are the confidence interval assumptions for β_1 ?
3. What are the confidence interval assumptions for ρ_{yx} ?
4. What effect does the planning value of ρ_{yx} have on the sample size requirement for ρ_{yx} ?
5. How is β_1 interpreted in a nonexperimental random- x model?
6. How is β_1 interpreted in an experimental fixed- x model?
7. How is $\mu_{y|\hat{x}}$ interpreted in an experimental fixed- x model?
8. If the estimated y -intercept is 10 and the estimated slope is 5, what is the predicted y_i score at $x_i = 20$?
9. How is eta-squared interpreted?
10. How does the prediction error variance affect the width of the confidence interval for β_1 ?
11. If the estimate of β_1 is 5.1, the 95% critical t -value is 3.0, and the standard error is 0.8, what is the 95% confidence interval for β_1 ?
12. Explain the difference between a confidence interval for $\mu_{y|\hat{x}}$ and a prediction interval for $\hat{y}$.
13. Draw a scatter plot of six observations where the correlation is about 0.9.
14. Draw a scatter plot of ten observations where the correlation is about -0.5.
15. How does the sample size affect the width of a confidence interval for β_1 ?
16. How does the sample size affect the width of a confidence interval for ρ_{yx} ?
17. How is β_1 interpreted in a simple linear regression model if x is dummy coded?
18. What kind of data are used to compute a point-biserial correlation?
19. How would the value of $\hat{\rho}_{yx}$ change if the x scores were all multiplied by 10?

20. How would the value of $\hat{\beta}_1$ change if the x scores were all multiplied by 10?
21. How would the value of $\hat{\beta}_1$ change if the y scores were all multiplied by 10?
22. What transformations are sometimes helpful in linearizing a nonlinear relation?
23. What transformations are sometimes helpful in normalizing residuals or stabilizing the error variance?
24. What are two general types of psychometric validity evidence?
25. How does the desired power of a statistical test affect the sample size requirement?
26. How does the α level affect the sample size requirement?
27. How does the desired confidence interval width affect the sample size requirement?
28. How does measurement error in the y scores and x scores affect the Pearson correlation?
29. How does measurement error in the y scores and x scores affect the slope?
30. What are some advantages of using the Theil-Sen confidence interval for a population slope?
31. What are some advantages of using a Spearman correlation rather than a Pearson correlation?

32.

Participant	y score	x score
1	5	20
2	7	25
3	12	30

Use the above sample data ($n = 3$) to compute the following:

- | | |
|------------------------|--------------------------------------|
| a) $\hat{\mu}_y$ | f) $\hat{\rho}_{yx}$ |
| b) $\hat{\mu}_x$ | g) $\hat{\beta}_1$ |
| c) $\hat{\sigma}_y^2$ | h) $\hat{\beta}_0$ |
| d) $\hat{\sigma}_x^2$ | i) $\hat{y}_1, \hat{y}_2, \hat{y}_3$ |
| e) $\hat{\sigma}_{yx}$ | j) $\hat{e}_1, \hat{e}_2, \hat{e}_3$ |

Data Analysis Problems

1-1. A large organization has been using a sexual harassment training program that was originally designed for use by both men and women. A new training program designed specifically for men is being evaluated. The new training program uses numerous short but detailed and realistic examples of men sexually harassing women, with each example followed by a summary of the important aspects of the harassment issues. A random sample of 32 male employees was obtained from an organization's employee database which contains the names of about 4,000 male employees. The 32 participants were randomly assigned into four groups of equal size with each group receiving 4, 6, 8, or 10 examples. After reading the examples, all 32 participants were given a test to assess their understanding of sexual harassment concepts and facts (scored on a 0 to 50 scale with larger values representing higher levels knowledge and understanding). A mandatory training program that uses more examples might be more effective but it would also be more time consuming. The organization wants to determine the relation between the number of examples used and the test score. The sample data are shown below.

Group 1 (4):	36	26	35	23	28	29	34	29
Group 2 (6):	38	27	39	42	33	35	36	28
Group 3 (8):	39	44	41	30	29	46	39	28
Group 4 (10):	47	34	38	40	41	37	30	46

- Describe the study population.
- Is this an experimental or a nonexperimental design? Why?
- What is the response variable and what is the predictor variable in this study?
- Is x fixed or random? Why?
- Report the hypothesis testing results in APA style for a test of $H_0: \beta_1 = 0$.
- Compute a 95% confidence interval for β_1 and interpret the results.
- Compute a 95% prediction interval for the test score of a single employee who is given $x = 5$ examples, and also compute a 95% confidence interval for the population mean test score if all 4,000 employees had all been given five examples.
- The organization plans to replicate this study in a population of employees in another city with training that consists of 2, 5, and 8 examples ($\sigma_x^2 = 6.0$). How many employees should be sampled to obtain a 95% confidence interval for β_1 that has a width of about 1.0? Use the MS_E from this study as the planning value of the within-group error variance.

1-2. One hundred first-grade and second-grade students were randomly selected from the approximately 1,200 first-grade and second-grade students who attend Los Angeles elementary schools. All students participate in a parent-student conference each year. After the required conference, the parents of the 100 randomly selected students were told that they had been randomly selected to participate in an important research study. The parent (or parents) of each child recorded for one week how many hours they used any type of screen device and how many hours their child used any type of screen device. In two-parent families, the screen times for the two parents were averaged. The parent's screen time is believed to predict the child's screen time. The HW1-2.sav file contains the sample data.

- a) Describe the study population.
- b) Is this an experimental or a nonexperimental design?
- c) What is the response variable and what is the predictor variable in this study?
- d) Is x fixed or random?
- e) Examine the scatterplot of the sample data. Are there any obvious outliers or indications of nonlinearity?
- f) Compute a 95% confidence interval for β_1 (average parent score as the predictor variable) and interpret the result.
- g) Compute a 95% confidence interval for ρ_{yx} . Compute a 95% confidence interval for ρ_{yx}^2 . Interpret the results for both confidence intervals.
- h) Use the 95% confidence interval for ρ_{yx} to test $H_0: \rho_{yx} = .3$ for $\alpha = .05$. What is your decision?
- i) The researcher plans to replicate this study in New York. How many families are needed to obtain a 95% confidence interval for ρ_{yx} that has a width of about 0.25? Use the lower 95% limit (the endpoint closest to zero in this case) of ρ_{yx} from this study as your planning value for ρ_{yx} .
- j) Compute the skewness and excess kurtosis for the parent and child screen times and report the values.

1-3. A random sample of 250 undergraduate students was obtained from a public university student directory of about 15,000 undergraduate students. The 250 students were classified as first-generator or continuing-generation students. All 250 students were given an academic self-efficacy questionnaire. The sample means, sample standard deviations, and sample sizes are given below.

First-generation	Continuing-generation
$\hat{\mu}_1 = 47.2$	$\hat{\mu}_2 = 56.3$
$\hat{\sigma}_1 = 8.34$	$\hat{\sigma}_2 = 6.92$
$n_1 = 52$	$n_2 = 198$

a) Describe the study population.

b) Use the `ci.pbcor` function in the `statpsych` package to compute a 95% confidence interval for a weighted variance point-biserial correlation. Interpret the results.

Suggested Readings

Berk, R. A. (2004). *Regression analysis: A constructive critique*. Sage. **Chapters 3-5**

Cohen, J. , Cohen, P. West, S. G., & Aiken (2003). *Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.)*. Erlbaum. **Chapter 2**

Darlington, R.B., & Hayes, A. F. (2017). *Regression analysis and linear models: Concepts, applications, and implementation*. Guilford. **Chapter 2**

Fox, J. (2016). *Applied regression analysis and generalized linear models (3rd ed)*. Sage. **Sections 3.2, 5.1, 6.1, and 7.1**

Graybill, F. A., & Iyer, H. K. (1994). *Regression analysis: Concepts and applications*. Duxbury. http://www.stat.colostate.edu/~hari/regression_book/ **Chapter 3**

Chapter 2

Statistical Models with Multiple Predictor Variables

2.1 The General Linear Model

All of the statistical methods in Chapter 1 were limited to the special case of one predictor variable where the predictor variable could be either quantitative or dichotomous. In this chapter, we consider linear statistical models with more than one predictor variable. Some or all of the predictor variables can be quantitative, and some or all of the predictor variables can be qualitative. Furthermore, the qualitative predictor variables can have more than two levels.

The relation between one quantitative response variable (y) and $s \geq 1$ predictor variables ($x_1, x_2, \dots, x_s$) for participant i ($i = 1$ to n) can be represented by the following *general linear model* (GLM)

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_s x_{si} + e_i \quad (\text{Model 2.1})$$

where y_i is the quantitative response variable score for participant i , and x_{ki} is the score for participant i on predictor variable k .

The value $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_s x_{si}$ is the predicted y score for participant i , and $e_i = y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_s x_{si})$ is the prediction error for participant i . The coefficients $\beta_0, \beta_1, \dots, \beta_s$ in Model 2.1 are the unknown population parameters to be estimated from sample data. The coefficient β_0 is the y -intercept and $\beta_1, \dots, \beta_s$ are the slope coefficients.

Each predictor variable in a GLM is either fixed or random. Random predictor variables are always quantitative, but fixed predictor variables can be quantitative (e.g., 10 hours, 20 hours, or 30 hours of training; 0, 1, 2, or 3 siblings) or qualitative (e.g., Treatment A, Treatment B, or Treatment C; Democrat, Republican, or Independent). The fixed predictor variables can be treatment factors with levels to which participants are randomly assigned (e.g., Treatment A, Treatment B, or Treatment C; 10 hours, 20 hours, or 30 hours of training) or classification factors with levels that represent existing characteristics of the study population (e.g., 0, 1, 2, or 3 siblings; Protestant, Catholic, or Other).

In Chapter 1, we saw how a 2-level qualitative predictor variable (e.g., male/female) could be included in the model by dummy coding the predictor variable. An a -level predictor variable can be converted into $a - 1$ dummy coded or effect coded variables as will be explained in section 2.15. A GLM where all predictor variables are dummy coded or effect coded variables is called an *analysis of variance (ANOVA) model*; a GLM that has only quantitative predictor variables is called a *multiple regression model*; and a GLM that has one or more dummy coded or effect coded predictor variables and one or more quantitative predictor variables is called an *analysis of covariance (ANCOVA) model*.

2.2 Interpreting Slope Coefficients in Nonexperimental Designs

In an experiment with a quantitative treatment factor (x_k), β_k describes the change in the population mean of y that will be *caused* by any 1-unit increase in x_k (within the range of x_k values used in the experiment). In nonexperimental designs with fixed quantitative predictor variables, the slope coefficient for x_k describes the change in the population mean of y that is *associated* with a 1-unit increase in x_k , controlling for the other $s - 1$ predictor variables. In nonexperimental designs with random quantitative predictor variables, the slope coefficient for x_k describes the change in the predicted y score *associated* with a 1-unit increase in x_k , controlling for the other $s - 1$ predictor variables.

The meaning of the phrase “controlling for” is not obvious and requires explanation. Consider the following model with two predictor variables

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \quad (\text{Model 2.2})$$

Now consider a model where x_1 is treated as a response variable, x_2 is treated as a predictor variable, and e_{x_1} is the resulting prediction error. The e_{x_1} prediction error reflects the part of x_1 that is not linearly associated with x_2 . It can be shown that β_1 in Model 2.2 is equal to β_1 in the following model

$$y_i = \beta_0 + \beta_1 e_{x_{1i}} + \beta_2 x_{2i} + e_i. \quad (\text{Model 2.3})$$

Model 2.3 shows that β_1 does not describe the relation between x_1 and y , as is commonly assumed, but actually describes the relation between e_{x_1} and y . The meaning of the e_{x_1} variable, which may not be obvious in many applications, is needed to interpret the estimate and confidence interval for β_1 . In general, each β_k in a GLM describes the relation between y and e_{x_k} where e_{x_k} represents the part of x_k that is not linearly related

to any of the other predictor variables, and it is in this sense that β_k describes the association between x_k and y “controlling for” all other predictor variables in the model.

The interpretation of β_k can be especially difficult in a nonexperimental design where two or more predictor variables measure similar attributes. For example, if x_1 is “depression” and x_2 is “psychological well-being”, e_{x_1} represents some component of depression that is unrelated to psychological well-being and that component could be very difficult to explain. However, in some applications, β_k will become more meaningful when certain predictor variables are added to the model. For example, if x_1 is a measure of spatial ability that involves complicated verbal instructions and x_2 is a measure of reading comprehension, β_1 describes the relation between y and the component of the spatial ability measure that is unrelated to reading comprehension (e_{x_1}). In this application, e_{x_1} could represent a purer measure of spatial ability than x_1 .

If x_1 is the predictor variable of primary interest, controlling for a categorical demographic variable often provides a more meaningful or interesting interpretation of β_1 . For example, if a dummy coded predictor variable (x_2) for gender is added to the model, then β_1 describes the slope of the line relating x_1 to y *within* each gender category. If there are mean differences in the y scores or x scores across groups, the within-group slope could be more meaningful than the slope of the line relating x_1 to y ignoring group differences (see Figure 2.1).

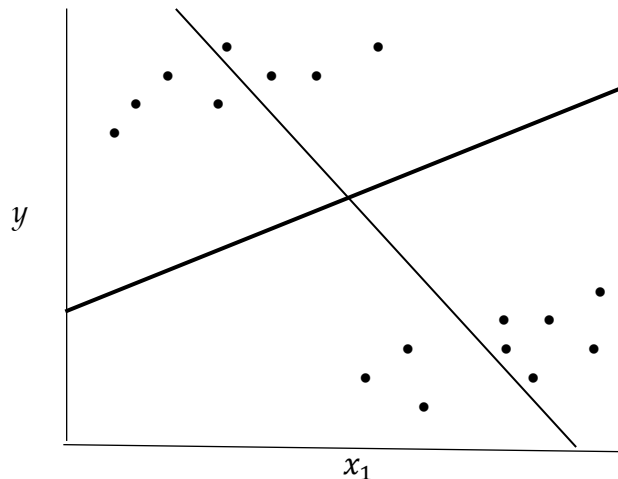


Figure 2.1 Slope for x_1 in a model without x_2 (thin line) and slope for x_1 in a model with x_2 (thick line)

2.3 Confounding Variable Bias

The value of the slope coefficient for x_1 can change substantially if another variable is added to the model. If x_2 has a causal effect on both x_1 and y , then x_2 is a *confounding* variable for x_1 . However, if x_2 is related to x_1 because x_1 has a causal effect only on x_2 , then x_2 is called a *mediator variable* rather than a confounding variable. The analysis of mediator variables is discussed in Volume 4.

Consider a statistical model with one predictor variable

$$y_i = \ddot{\beta}_0 + \ddot{\beta}_1 x_{1i} + \ddot{e}_i \quad (\text{Model 2.4})$$

and the following model that includes x_2 as an additional predictor variable

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \quad (\text{Model 2.5})$$

where the double dots in Model 2.4 indicate that the parameters and prediction errors are not necessarily identical to those in Model 2.5. If x_1 in Model 2.4 is confounded with some other variable, then $\ddot{\beta}_1$ will contain *confounding variable bias*. The confounding variable bias in $\ddot{\beta}_1$ can be described in terms of how much its value changes if a confounding variable is added to the model. Suppose x_2 is a confounding variable for x_1 and is added to Model 2.4 to produce Model 2.5. It can be shown that

$$\ddot{\beta}_1 = \beta_1 + (\beta_2 \rho_{x_1 x_2} \sigma_{x_2} / \sigma_{x_1}) \quad (2.1)$$

where the term in parentheses is the amount of confounding variable bias in $\ddot{\beta}_1$ due to the omission of x_2 . Note that the amount of confounding variable bias in $\ddot{\beta}_1$ due to x_2 depends on the magnitude of the correlation between x_2 and x_1 as well as the magnitude of β_2 . The magnitude of the confounding variable bias due to x_2 will be small if either β_2 or $\rho_{x_1 x_2}$ is small.

Adding a predictor variable will usually shrink the slope coefficients for the other predictor variables in the model towards 0. Occasionally, adding a predictor variable will have the opposite effect. In these rare cases, the added predictor variable is called a *suppressor variable* because failure to include it in the model will suppress the value of a slope coefficient for at least one other predictor variable.

Confounding variable bias for a particular slope coefficient due to one or more confounding variables can be removed by including those confounding variables, usually called *control variables*, as additional predictor variables in the model. In a nonexperimental design, if the researcher can present a convincing argument that the most important control variables for x_1 have been included in the model, β_1 might tentatively and cautiously be interpreted as a description of a causal relation between x_1 and y . However, providing such an argument will require substantial subject-matter expertise, and the premises of the argument will almost certainly be challenged by other experts. Furthermore, if any of the control variables for x_1 have measurement error, which is the rule rather than the exception, then only some of the confounding variable bias in β_1 will be removed. Statistical methods that correct for measurement error in the control variables are described in Volume 4.

Confounding variable bias is not a problem in an experimental design. In an experimental design, participants are randomly assigned to the levels of some treatment factor. The randomization process guarantees that *all* excluded predictor variables will be unrelated to the treatment predictor variables, and therefore these excluded variables cannot be confounding variables. In an experimental design with no differential nonrandom attrition (see section 2.17 of Volume 1), the slope coefficient for a particular treatment predictor variable will describe a causal relation between the treatment predictor variable and the response variable.

2.4 Analysis of Variance Table for a GLM

The variance of the y scores in a GLM can be decomposed into two sources of variability: the variance of the prediction errors (also called *error variance* or *residual variance*) and the variance of y that is predictable from the s predictor variables (also called *model variance*). The decomposition of the variance of the y scores can be summarized in an analysis of variance (ANOVA) table as shown below.

Source	SS	df	MS	F
MODEL	SS_M	$df_M = s$	$MS_M = SS_M/df_M$	MS_M/MS_E
ERROR	SS_E	$df_E = n - s - 1$	$MS_E = SS_E/df_E$	
TOTAL	SS_T	$df_T = n - 1$	$MS_T = SS_T/df_T$	

The estimated predicted y score for person i is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \dots + \hat{\beta}_s x_{si}$ and the estimated prediction error (residual) for person i is $\hat{e}_i = y_i - \hat{y}_i$ where $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_s$ are OLS estimates. The sum of squares (SS) values in the ANOVA table are given below.

$$SS_T = \sum_{i=1}^n (y_i - \hat{\mu}_y)^2 \quad (2.2)$$

$$SS_E = \sum_{i=1}^n \hat{e}_i^2 \quad (2.3)$$

$$SS_M = SS_T - SS_E \quad (2.4)$$

MS_E is an estimate of the variance of the prediction errors in the study population (σ_e^2), and the MS_T is an estimate of the variance of the y scores in the study population (σ_y^2).

If all predictor variables in the GLM are random, a *multiple correlation* between y and the s predictor variables could be an interesting population parameter to estimate. A multiple correlation is equal to a Pearson correlation between the y scores and the predicted y scores. Recall from Chapter 1 that the OLS estimate of β_1 minimizes the sum of the squared residuals. The OLS estimates of $\beta_1, \beta_2, \dots, \beta_s$ minimize the sum of the squared residuals and also define a unique linear function of the s predictor variables that maximizes the value of the Pearson correlation between y_i and $\hat{y}_i$ where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \dots + \hat{\beta}_s x_{si}$.

In the random- $\mathbf{x}$ case, the population multiple correlation is denoted as $\rho_{y,\mathbf{x}}$ where $\mathbf{x}$ denotes the set of s predictor variables. The population squared multiple correlation $\rho_{y,\mathbf{x}}^2$, referred to as the *coefficient of multiple determination*, is equal to $1 - \sigma_e^2 / \sigma_y^2$. The coefficient of multiple determination describes the proportion of the response variable variance that can be predicted from the s predictor variables.

An estimate of $\rho_{y,\mathbf{x}}^2$ (reported as “R-squared” in most statistical packages) is obtained from the SS values in the ANOVA table

$$\hat{\rho}_{y,\mathbf{x}}^2 = 1 - SS_E / SS_T. \quad (2.5)$$

The confidence interval for $\rho_{y,\mathbf{x}}^2$ does not have a simple formula but it can be obtained using the `ci.rsqr` function in the `statpsych` package.

The estimated squared multiple correlation has a positive bias (i.e., it tends to be too large), and the bias can be substantial when df_E is small. The following *adjusted squared*

multiple correlation (reported as "adjusted R-squared" in most statistical packages) has less positive bias and is equal to

$$adj \hat{\rho}_{y,x}^2 = 1 - df_T(1 - \hat{\rho}_{y,x}^2)/df_E = 1 - MS_E/MS_T. \quad (2.6)$$

In the fixed- $\mathbf{x}$ case, the following estimate of the coefficient of multiple determination

$$\hat{\eta}^2 = 1 - SS_E/SS_T \quad (2.7)$$

is an estimate of η^2 and

$$adj \hat{\eta}^2 = 1 - df_T(1 - \hat{\eta}^2)/df_E = 1 - MS_E/MS_T \quad (2.8)$$

is a less positively biased estimate of η^2 . Although $\eta^2 = \rho_{y,x}^2$, different symbols are used because a confidence interval for η^2 in the fixed- $\mathbf{x}$ model will be different than a confidence interval for $\rho_{y,x}^2$ in the random- $\mathbf{x}$ model. The confidence interval for η^2 in the fixed- $\mathbf{x}$ case does not have a simple formula but it can be obtained using the `ci.etasqr` function in the `statpsych` package.

Both $adj \hat{\rho}_{y,x}^2$ and $adj \hat{\eta}^2$ are functions of $MS_E = SS_E/df_E$. Although SS_E can never increase when a predictor variable is added to the model, MS_E can increase because the decrease in df_E could be relatively greater than the decrease in SS_E . Unlike $\hat{\rho}_{y,x}^2$, which can never decrease when a predictor variable is added to the model, $adj \hat{\rho}_{y,x}^2$ and $adj \hat{\eta}^2$ can decrease when a predictor variable is added to the model because they are functions of MS_E .

The F statistic from the ANOVA table is used to test the omnibus null hypothesis $H_0: \beta_1 = \beta_2 = \dots = \beta_s = 0$ against an alternative hypothesis that at least one population slope coefficient is nonzero. H_0 is rejected if the p -value for the F statistic is less than α . These null and alternative hypotheses are equivalent to testing $H_0: \rho_{y,x}^2 = 0$ against $H_1: \rho_{y,x}^2 > 0$ (or testing $H_0: \eta^2 = 0$ against $H_1: \eta^2 > 0$ since $\eta^2 = \rho_{y,x}^2$). A statistical test that allows the researcher to simply decide if $H_0: \rho_{y,x}^2 = 0$ can or cannot be rejected does not provide useful scientific information because the researcher knows, before any data have been collected, that H_0 is almost certainly false and hence H_1 is almost certainly true. Although the researcher knows that $\rho_{y,x}^2$ will almost never equal 0, the value of $\rho_{y,x}^2$ will not be known and therefore a confidence interval for $\rho_{y,x}^2$ (or η^2) provides useful scientific or practical information.

A confidence interval for $\rho_{y,x}^2$ can be used to test $H_1: \rho_{y,x}^2 \leq h$ against $H_2: \rho_{y,x}^2 > h$ where h is a value, such as .05, that would be considered a "small" effect size. If the lower limit is greater than h , then accept H_2 ; if the upper limit is less than h then accept H_1 ; otherwise, the results are inconclusive. Accepting H_1 is evidence that all of the population slope values represent weak relations. Accepting H_2 is evidence that at least one slope is meaningfully large. This test can also be applied to η^2 .

Example 2.1. A random sample of 150 college students completed a survey that measured resilience (y), compassion (x_1), problem solving (x_2), optimism (x_3), and physical fitness (x_4). The estimated squared multiple correlation between resilience and the four predictor variables is .247, and the adjusted estimate of the squared multiple correlation is .226. The 95% confidence interval for the population squared multiple correlation is [.115, .351] indicating that 11.5% to 35.1% of the variance in resilience scores in the population of college students can be predicted from their scores on the compassion, problem solving, optimism, and physical fitness measures.

2.5 Confidence Interval for Slope Coefficient

A $100(1 - \alpha)\%$ confidence interval for β_k (the population slope coefficient corresponding to x_k) is

$$\hat{\beta}_k \pm t_{\alpha/2, df_E} SE_{\hat{\beta}_k} \quad (2.9)$$

where $\hat{\beta}_k$ is the OLS estimate of β_k (see Equation 2.32) and $df_E = n - s - 1$. The standard error of $\hat{\beta}_k$ can be expressed as

$$SE_{\hat{\beta}_k} = \sqrt{MS_E / [(1 - \hat{\rho}_{x_k, x_0}^2) \hat{\sigma}_{x_k}^2 (n - 1)]} \quad (2.10)$$

where $\hat{\rho}_{x_k, x_0}^2$ is the estimated squared multiple correlation between x_k and the other $s - 1$ predictor variables, and $\hat{\sigma}_{x_k}^2$ is the estimated variance of x_k .

The MS_E value in Equation 2.10 can be expressed as $df_T \hat{\sigma}_y^2 (1 - \hat{\rho}_{y,x}^2) / df_E$. This expression is informative because it shows that a larger value of the multiple correlation between the response variable and all predictor variables ($\hat{\rho}_{y,x}^2$) will give a smaller value of MS_E , which reduces the value of $SE_{\hat{\beta}_k}$. Furthermore, the value of $SE_{\hat{\beta}_k}$ depends on the multiple correlation between predictor variable k and the other predictor variables ($\hat{\rho}_{x_k, x_0}^2$) where a larger value of $\hat{\rho}_{x_k, x_0}^2$ increases the value of $SE_{\hat{\beta}_k}$. Thus, correlated predictor variables will inflate the value of $SE_{\hat{\beta}_k}$. A larger value of $SE_{\hat{\beta}_k}$ will increase the width of the confidence

interval for β_k and reduce the power of the test of $H_0: \beta_k = 0$. If the predictor variables are moderately correlated, it is not uncommon to reject $H_0: \beta_1 = \beta_2 = \dots = \beta_s = 0$ but then fail to reject $H_0: \beta_k = 0$ for every predictor variable in the model.

Example 2.2. A researcher obtained a random sample of $n = 236$ participants from a business directory containing contact information for about 50,000 working adults. All participants in the sample were given a life satisfaction questionnaire (y), a neuroticism questionnaire (x_1), a conscientiousness questionnaire (x_2), and an industriousness questionnaire (x_3). All three predictor variables were moderately correlated with each other. In this study, the researcher is primarily interested in the effect of industriousness on life satisfaction controlling for neuroticism and conscientiousness. The researcher can be 95% confident that in the study population of 50,000 working adults, a 1-point increase in industriousness is associated with a 0.148 to 0.456 increase in life satisfaction, controlling for neuroticism and conscientiousness.

Predictor	$\hat{\beta}_k$	$SE_{\hat{\beta}_k}$	95% CI for β_k
neurotic	-0.553	0.088	[-0.729, -0.377]
conscientious	-0.301	0.117	[-0.536, -0.065]
industrious	0.302	0.077	[0.148, 0.456]

2.6 Confidence Interval for τ_e

The mean absolute prediction error (τ_e) in a GLM describes the predictive accuracy of the s predictor variables and is a useful supplement to the squared multiple correlation. An approximate $100(1 - \alpha)\%$ confidence interval for τ_e in a GLM is given in Equation 1.22 with $df = n - s - 1$ and a bias adjustment of $c = n/(n - (s + 2)/2)$. The `ci.mape` function in the `statpsych` package will compute Equation 1.22 for a GLM.

Example 2.3. The life satisfaction questionnaire in *Example 2.2* has a range of 5 to 35. The estimate of τ_e is 3.81 and the 95% confidence interval for τ_e is [3.60, 4.09]. We can be 95% confident that the mean absolute prediction error using neuroticism, conscientiousness, and industriousness to predict life satisfaction is between 3.60 and 4.09 in the study population of 50,000 working adults.

2.7 Semipartial Correlation

A *semipartial correlation* (or *part correlation*) is an especially useful measure of effect size in a random-x model. The semipartial correlation between x_1 and y controlling for $x_2 \dots x_s$ is denoted as $\rho_{y(x_1, \mathbf{x}_0)}$ where $\mathbf{x}_0$ represents a set of control variables $x_2 \dots x_s$. A semipartial correlation is a Pearson correlation between y and e_{x_1} where e_{x_1} is the prediction error in

a model that predicts x_1 from $x_2 \dots x_s$. A semipartial correlation between x_1 and y controlling for $x_2 \dots x_s$ describes the change in a standardized y score that is associated with a one standard deviation increase in e_{x_1} . The control variables are assumed to be confounding variables and not mediator variables.

An estimate of a semipartial correlation between x_1 and y controlling for $x_2 \dots x_s$ can be obtained by first computing the residuals in a GLM where x_1 is the response variable and $x_2 \dots x_s$ are the predictor variables. The Pearson correlation between the y scores and the x_1 residuals ($\hat{e}_{x_1}$) is an estimate of the semipartial correlation. In the special case of one control variable (x_2), the estimated semipartial correlation between y and x_1 can be expressed in terms of Pearson correlation estimates as shown below.

$$\hat{\rho}_{y(x_1.x_2)} = (\hat{\rho}_{yx_1} - \hat{\rho}_{yx_2}\hat{\rho}_{x_1x_2})/\sqrt{1 - \hat{\rho}_{x_1x_2}^2} \quad (2.11)$$

It can be shown that the squared semipartial correlation between y and x_k is equal to

$$\rho_{y(x_k.x_0)}^2 = \rho_{y.x}^2 - \rho_{y.x_0}^2 \quad (2.12)$$

where $\mathbf{x}_0$ is the set of all predictor variables in the model excluding x_k , and $\mathbf{x}$ is the set of all predictor variables in the model. This difference is referred to as ΔR^2 in APA journals. The squared semipartial correlation between y and x_k describes the proportion of y score variance that is predictable from the e_{x_k} scores.

In a random- $\mathbf{x}$ model, a semipartial correlation could be computed for each predictor variable controlling for all other predictor variables in the model. These semipartial correlations are conceptually similar to slope coefficients because they describe the relation between x_j and y after the linear effects of all other predictor variables have been removed from x_k . A semipartial correlation is a standardized measure of effect size that has a possible range of -1 to 1 and will be easier to interpret than a slope coefficient if the scales of x_k and y are unfamiliar to the intended audience.

A confidence interval for $\rho_{y(x_1.x_0)}$ is obtained in two steps. First, an approximate $100(1 - \alpha)\%$ confidence interval for a transformed semipartial correlation estimate is computed

$$\hat{\rho}_{y(x_1.x_0)}^* \pm z_{\alpha/2}\sqrt{var} \quad (2.13)$$

where $var = (\hat{\rho}_{y,x}^4 - 2\hat{\rho}_{y,x}^2 + \hat{\rho}_{y,x_0}^2 - \hat{\rho}_{y,x_0}^4 + 1) / (1 - \hat{\rho}_{y(x_1,x_0)}^2)^2 / (n - 3)$ and $\hat{\rho}_{y(x_1,x_0)}^* = \ln\left(\frac{1 + \hat{\rho}_{y(x_1,x_0)}}{1 - \hat{\rho}_{y(x_1,x_0)}}\right) / 2$. Let ρ_L^* and ρ_U^* denote the endpoints of Equation 2.13. Next, reverse transforming the endpoints of Equation 2.13 gives the following lower confidence limit for $\rho_{y(x_1,x_0)}$

$$LL = [\exp(2\rho_L^*) - 1] / [\exp(2\rho_L^*) + 1] \quad (2.14a)$$

and the following upper confidence limit for $\rho_{y(x_1,x_0)}$

$$UL = [\exp(2\rho_U^*) - 1] / [\exp(2\rho_U^*) + 1]. \quad (2.14b)$$

If the most important confounding variables are included as control variables in a semipartial correlation, the semipartial correlation could potentially describe a causal relation between y and x_1 . The `ci.spcor` function in the `statpsych` package will compute a confidence interval for a population semipartial correlation.

2.8 Partial Correlation

A *partial correlation* is another measure of effect size for a random- x model. A partial correlation between x_k and y removes the linear effects of one or more control variables from *both* x_k and y . In comparison, a semipartial correlation between x_k and y removes the linear effects of one or more control variables from only x_k . As with semipartial correlations, the control variables are assumed to be confounding variables and not mediator variables.

Let x_1 be the predictor variable of interest and let $\mathbf{x}_0$ denote a set of control variables $x_2 \dots x_s$. A partial correlation between x_1 and y controlling for $\mathbf{x}_0$ is denoted as $\rho_{yx_1.\mathbf{x}_0}$. A partial correlation is a Pearson correlation between e_{x_1} and e_y where e_{x_1} represents the prediction errors in a model that predicts x_1 from $\mathbf{x}_0$ and e_y represents the prediction errors in a model that predicts y from $\mathbf{x}_0$. A partial correlation between x_1 and y describes the change in a standardized e_y score that is associated with a one standard deviation increase in e_{x_1} . The squared partial correlation between y and x_1 describes the proportion of variance in the e_y scores that is predictable from the e_{x_1} scores.

In the special case of one control variable (x_2), the estimated partial correlation between y and x_1 can be expressed in terms of Pearson correlation estimates as shown below

$$\hat{\rho}_{yx_1.x_2} = (\hat{\rho}_{yx_1} - \hat{\rho}_{yx_2}\hat{\rho}_{x_1x_2}) / \sqrt{(1 - \hat{\rho}_{yx_2}^2)(1 - \hat{\rho}_{x_1x_2}^2)} \quad (2.15)$$

and is larger than the semipartial correlation unless $\hat{\rho}_{yx_2} = 0$. The squared partial correlation between y and x_k for any number of control variables is

$$\rho_{yx_1.x_0}^2 = (\rho_{y.x}^2 - \rho_{y.x_0}^2) / (1 - \rho_{y.x_0}^2) \quad (2.16)$$

where $\rho_{y.x}^2$ and $\rho_{y.x_0}^2$ are defined in Equation 2.12.

A partial correlation may be more interesting than a semipartial correlation if it is important to remove the effects of the control variables from both y and x_1 . For example, if social skills and problem-solving skills are measured in a sample of children, age six to nine, the correlation between these two variables could be misleadingly high because both variables are confounded with age, and it would be desirable to remove the effect of age from both social skills and problem-solving skills.

A confidence interval for $\rho_{yx_1.x_0}$ is obtained in two steps. First, a $100(1 - \alpha)\%$ confidence interval for a transformed partial correlation is computed

$$(\hat{\rho}_{yx_1.x_0}^* - \frac{\hat{\rho}_{yx_1.x_0}}{2(n-1)}) \pm z_{\alpha/2} \sqrt{1/(n - s_2 - 3)} \quad (2.17)$$

where $\hat{\rho}_{yx_1.x_0}^* = \ln\left(\frac{1 + \hat{\rho}_{yx_1.x_0}}{1 - \hat{\rho}_{yx_1.x_0}}\right)/2$ and s_2 is the number of control variables. Let ρ_L^* and ρ_U^* denote the endpoints of Equation 2.17. Next, reverse transforming the endpoints of Equation 2.17 gives the following lower confidence limit for $\rho_{yx_1.x_0}$

$$LL = [\exp(2\rho_L^*) - 1] / [\exp(2\rho_L^*) + 1] \quad (2.18a)$$

and the following upper confidence limit for $\rho_{yx_1.x_0}$

$$UL = [\exp(2\rho_U^*) - 1] / [\exp(2\rho_U^*) + 1]. \quad (2.18b)$$

If the most important confounding variables are included as control variables in a partial correlation, the partial correlation could potentially describe a causal relation between y and x_1 . The `ci.cor` function in the `statpsych` package will compute a confidence interval for a population partial correlation.

Example 2.4. A psychometric validation study examined the correlation between a new 3-D spatial ability scale and a 2-D spatial ability scale. Both scales have detailed written instructions, and it is likely that both scales are affected by reading comprehension. The 2-D and 3-D spatial ability scales and a measure of reading comprehension were given to a random sample of 135 community college students taken from a study population of about 1,900 students. The sample partial correlation between the 3-D and 2-D scales, controlling for reading comprehension, is .70. The 95% confidence interval for the population partial correlation is [.60, .78]. We can be 95% confident that in the study population of 1,900 community college students, the correlation between the 2-D and 3-D spatial ability scales is between .60 and .78 after removing the effect of reading comprehension from both scales.

2.9 Standardized Slope Coefficient

An estimated *standardized slope*, denoted as $\underline{\hat{\beta}}_k$, is computed in a multiple regression model where all of the variables have been standardized. The standardized y scores are $y_i^* = (y_i - \hat{\mu}_y)/\hat{\sigma}_y$ and the standardized x_k scores are $x_{ki}^* = (x_{ki} - \hat{\mu}_{x_k})/\hat{\sigma}_{x_k}$. Recall that β_k describes the relation between y and e_{x_k} where e_{x_k} is the part of x_k that is not linearly related to the other predictor variables in the model. If the predictor variables are standardized and if x_k^* is predicted from all other predictor variables in the model, $e_{x_k^*}$ represents the part of x_k^* that is not linearly related to any of the other predictor variables. Then $\underline{\hat{\beta}}_k$ describes the change in a standardized y score that is associated with a 1-point increase in $e_{x_k^*}$. SPSS will compute standardized slopes for every predictor variable in the model, but it is not necessary or appropriate to compute a standardized slope for a dummy or effect coded variable (see section 2.15), a squared variable (see section 2.10), or a product variable (see section 2.11).

A standardized slope estimate for x_k is often interpreted as the change in a standardized y score that is associated with a one standard deviation increase in x_k , controlling for the other predictor variables. However, the phrase "controlling for" is often misunderstood in the context of standardized variables. It can be shown that the standard deviation of the $e_{x_k^*}$ scores is equal to $\sqrt{1 - \hat{\rho}_{x_k, x_0}^2}$ which will be less than 1 unless $\hat{\rho}_{x_k, x_0}^2 = 0$. For example, if $\hat{\rho}_{x_k, x_0}^2 = .75$ then the standard deviation of the $e_{x_k^*}$ scores is equal to $\sqrt{1 - \hat{\rho}_{x_k, x_0}^2} = 0.5$ so that a 1-point increase in $e_{x_k^*}$ corresponds to a 2 standard deviation increase in $e_{x_k^*}$. The fact that a 1-point increase in $e_{x_k^*}$ usually corresponds to more than a one standard deviation increase in $e_{x_k^*}$ makes the standardized slope difficult to interpret. If the model

has only one predictor variable or if the predictor variables are uncorrelated, the standardized slope for x_k is equal to the Pearson correlation between y and x_k .

An estimated standardized slope for x_k can be expressed as $\underline{\hat{\beta}}_k = \hat{\rho}_{y(x_1, x_0)} / \sqrt{1 - \hat{\rho}_{x_k, x_0}^2}$.

Unless $\hat{\rho}_{x_k, x_0}^2 = 0$, the standardized slope will be greater than the corresponding semipartial correlation. Unlike a semipartial correlation, which has a possible range of -1 to 1, a standardized slope can have a value that is far outside this range when $\hat{\rho}_{x_k, x_0}^2$ is large. Although standardized slope estimates are popular, it is usually better to report semipartial correlations along with their confidence intervals. A confidence interval for a population standardized slope is discussed in Volume 4.

An unstandardized slope can be transformed into a standardized slope. The standardized slope for x_k can be expressed as $\underline{\hat{\beta}}_k = \hat{\beta}_k \hat{\sigma}_{x_k} / \hat{\sigma}_y$.

Example 2.5. In Example 2.2, life satisfaction (y), neuroticism (x_1), conscientiousness (x_2), and industriousness (x_3) were measured in a sample of $n = 236$ employees taken from a study population of about 50,000 employees. The estimated standardized slopes and semipartial correlations are given below. Note that the standardized slopes are larger in absolute value than the semipartial correlations. The estimated squared multiple correlation is .230 (which is needed to compute the confidence interval for the population semipartial correlation). The confidence interval for the population standardized slope was computed using the `lavaan` R package.

Predictor	Semipartial Correlation	Standardized Slope	95% CI for Semipartial Correlation	95% CI for Standardized Slope
neuroticism	-.355	-0.424	[-.456, -.244]	[-0.554, -0.294]
conscientious	-.144	-0.220	[-.253, -.031]	[-0.479, -0.217]
industrious	.221	0.310	[.108, .328]	[0.170, 0.450]

2.10 Directional Two-sided Hypothesis Tests

Directional two-sided hypothesis tests can be specified in terms of β_k , $\underline{\beta}_k$, $\rho_{y(x_k, x_0)}$ or ρ_{yx_k, x_0} . As explained in section 1.19, a confidence interval can be used to perform a directional two-sided test. In the most typical case where the null hypothesis value is 0, a confidence interval or test statistic for β_k can be used to perform a directional two-sided hypothesis test for β_k , $\underline{\beta}_k$, $\rho_{y, x'}$, or $\rho_{y, x}$ because $\beta_k > 0$ implies $\underline{\beta}_k > 0$, $\rho_{y, x} > 0$ and $\rho_{y, x} > 0$. Likewise, $\beta_k < 0$ implies $\underline{\beta}_k < 0$, $\rho_{y, x} < 0$ and $\rho_{y, x} < 0$.

2.11 Interval Hypothesis Test

In some studies, the goal is to show that the relation between y and x_k , controlling for the other predictor variables in the model, is small or unimportant. An interval hypothesis test (see section 1.20) is useful in these applications. Although an interval hypothesis test can be specified in terms of β_k , $\underline{\beta}_k$, $\rho_{y(x_k.x_0)}$ or $\rho_{yx_k.x_0}$, it is usually easiest to specify the value of h for a semipartial or partial correlation. For example, the following hypotheses regarding a population semipartial correlation between y and x_k can be specified

$$H_1: |\rho_{y(x_k.x_0)}| < h \quad H_2: |\rho_{y(x_k.x_0)}| \geq h$$

where h is a number that represents a small or unimportant semipartial correlation value (e.g., $h = .20$). A confidence interval for $\rho_{y(x_k.x_0)}$ can be used to perform the test. H_1 is accepted if the confidence interval is completely within the $-h$ to h range, and H_2 is accepted if the confidence interval is completely outside the $-h$ to h range. The results are inconclusive if the confidence interval includes the values h or $-h$. Interval hypothesis tests are most informative when h is a small value, but then a large sample size will be needed to obtain a confidence interval that is completely within the narrow $-h$ to h range.

2.12 Quadratic Model

If a nonlinear relation between y and x cannot be linearized using transformations, the following *quadratic model* will be appropriate when the relation between y and x can be characterized by a curve with a single bend.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i \quad (\text{Model 2.6})$$

In a quadratic model, the slope of the line relating x to y varies across the levels of x . Specifically, the slope of the line at $x = \dot{x}$ is equal to $\beta_1 + 2\beta_2 \dot{x}$, which may be estimated by replacing β_1 and β_2 with their estimates. It is standard practice to mean center x in a quadratic model. If x is mean centered, then β_1 describes the slope of the line at the mean of x . Mean centering x will not change the estimate of β_2 or its standard error.

A quadratic model implies that the direction of the relation between x and y changes sign at some value of x . In some applications it could be informative to estimate the value of x where the relation between x and y changes direction. If x has been mean centered, the estimated change point for the non-centered x is equal to $-\hat{\beta}_1/2\hat{\beta}_2 + \hat{\mu}_x$ (if x has not been

mean centered, the estimated change point is $-\hat{\beta}_1/2\hat{\beta}_2$). If the change point is outside the range of typical x scores, this implies that the direction of the relation is constant across typical x scores and this could simplify the interpretation of results. A confidence interval for the population value of the change point is discussed in Volume 4.

To estimate the amount of curvature in the nonlinear relation between x and y , the slope of the line can be compared at low (x_L) and high (x_H) values of the mean centered x . The difference in slopes at low and high values of the mean centered x is equal to $(\beta_1 + \beta_2 x_H) - (\beta_1 + \beta_2 x_L) = 2(x_H - x_L)\beta_2$ where x_L and x_H are values specified by the researcher. A confidence interval for $2(x_H - x_L)\beta_2$ is obtained by multiplying the endpoints of a confidence interval for β_2 by $2(x_H - x_L)$.

In some applications the researcher might be content with simply testing the following directional hypotheses about the values of β_1 and β_2 .

$$\begin{array}{lll} H_0: \beta_1 = 0 & H_1: \beta_1 > 0 & H_2: \beta_1 < 0 \\ H_0: \beta_2 = 0 & H_1: \beta_2 > 0 & H_2: \beta_2 < 0 \end{array}$$

Confidence intervals for β_1 and β_2 may be used to select H_1 or H_2 . Deciding $\beta_1 > 0$ implies that there is a positive relation between y and x , and deciding $\beta_1 < 0$ implies that there is a negative relation between y and x . Deciding $\beta_2 > 0$ implies that the predicted y scores follow a curve that bends up, and deciding $\beta_2 < 0$ implies that the predicted y scores follow a curve that bends down. It is common practice to mean center x in a quadratic model because doing so reduces the standard error for β_1 and increases the power of the test of $H_0: \beta_1 = 0$ without affecting the power of the test of $H_0: \beta_2 = 0$.

Example 2.6. A random sample of psychology students was obtained from a volunteer pool. The sample was randomized into four team-size conditions with ten study teams per condition. The four conditions used teams of 2, 4, 6, or 8 students. The dependent variable is performance on a research project (scored 0 to 100). The estimates of β_1 and β_2 in a quadratic model are 7.23 and -0.65, respectively. The 95% confidence intervals for β_1 and β_2 are [4.82, 9.64] and [-0.79, -0.34], respectively. The maximum project score occurs at $x = 7.23/[2(.65)] = 6.4$ which suggests that teams of 6 or 7 are optimal for the type of project used in this study.

2.13 Interaction Effects in a GLM

Two-way Interactions

The strength of the relation between x_1 and y might depend on the value of a second predictor variable x_2 . When this occurs, it is also the case that the relation between x_2 and y will depend on the value of x_1 . In these situations, we say that x_1 and x_2 “interact”. In applications where x_1 and x_2 interact and the relation between x_1 and y is of primary interest, it could also be said that x_2 “moderates” the relation between x_1 and y .

The interaction effect for x_1 and x_2 can be included in a GLM by adding a new predictor variable that is the product of x_1 and x_2 . An example of a GLM with two predictor variables and their interaction is shown below.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + e_i \quad (\text{Model 2.7})$$

where β_3 describes the two-way interaction for x_1 and x_2 .

If the confidence interval for β_3 does not include 0, the values of β_1 and β_2 may be uninteresting. This is analogous to the problem of interpreting main effects in a two-factor design (section 3.13 in Volume 1) when an interaction effect has been detected. The nature of the interaction effect in a GLM can be understood by examining a *conditional slope* (also called a *simple slope*) for x_1 at low and high values of x_2 or a conditional slope for x_2 at low and high values of x_1 . The conditional slope for x_1 can be defined by factoring x_{1i} out of the $\beta_1 x_{1i}$ and $\beta_3 x_{1i} x_{2i}$ terms in Model 2.7 as shown below

$$y_i = \beta_0 + \beta_2 x_{2i} + (\beta_1 + \beta_3 x_{2i}) x_{1i} + e_i$$

where the term in parentheses is the conditional slope for x_1 . Factoring x_{2i} out of the $\beta_2 x_{2i}$ and $\beta_3 x_{1i} x_{2i}$ terms shows that conditional slope for x_2 is $\beta_2 + \beta_3 x_{1i}$.

When estimating a conditional slope for x_1 at x_2 where x_2 is a dummy coded variable, the conditional slope for x_1 could be computed at each level of x_2 . This can be accomplished two different ways. One way to estimate the conditional slope for x_1 is to compute $\hat{\beta}_1 + \hat{\beta}_3 x_2$ for each value of x_2 . Alternatively, a multiple group regression model (described in section 2.20) can be used. In a multiple group regression model where x_2 is dummy coded, x_2 and $x_{1i} x_{2i}$ are dropped from the model and β_1 is estimated in the $x_2 = 0$ group and then in the $x_2 = 1$ group.

In a multiple group analysis, the confidence interval for each conditional slope is simply the confidence interval for β_1 computed within each group. However, if the conditional slope is estimated as $\hat{\beta}_1 + \hat{\beta}_3 x_2$, then the computation of a confidence interval for $\beta_1 + \beta_3 x_2$ is more complicated and is described in section 2.12. In addition to computational simplicity, another advantage of the multiple group analysis is that tests and confidence intervals for the conditional slopes do not assume equal prediction error variances across the levels of the fixed factor.

Suppose x_1 and x_2 are both dummy coded variables and there are no other predictor variables in the model. The conditional slope for x_1 at $x_2 = 0$ is β_1 and the conditional slope for x_1 at $x_2 = 1$ is $\beta_1 + \beta_3$. The conditional slope for x_2 at $x_1 = 0$ is β_2 and the conditional slope for x_2 at $x_1 = 1$ is $\beta_2 + \beta_3$. These conditional slopes are identical to the simple main effects in a 2×2 factorial design that was described in Chapter 3 of Volume 1. If x_1 and x_2 are both dichotomous predictor variables, then the confidence interval for a linear contrast of population means (Equation 3.1 of Volume 1) is recommended because it is easier to apply and does not assume equality of variances.

When estimating a conditional slope for x_1 at x_2 where x_2 is a random predictor variable, it is common practice to report the conditional slopes at one standard deviation above the sample mean of x_2 and at one standard deviation below the sample mean of x_2 (assuming these low and high values are not implausible). If x_2 is highly skewed, setting x_2 at its 25th and 75th percentiles would be better choices.

Mean centering has no effect on the estimate of β_3 or the confidence interval for β_3 . Also, the conditional slope for x_1 at a mean centered value of x_2 is identical to the conditional slope for x_1 at the corresponding non-centered value of x_2 . Likewise, the conditional slope for x_2 at a mean centered value of x_1 is identical to the conditional slope for x_2 at the corresponding non-centered value of x_1 .

In applications where the interaction effect in Model 2.7 has important practical or theoretical implications, be sure to verify that both x_1 and x_2 are linearly related to y before analyzing Model 2.7. If x_1 and x_2 are at least moderately correlated, then a quadratic relation between x_1 and y or x_2 and y could be mistaken for an interaction effect.

The interaction between x_1 and x_2 can be "removable" or "non-removable" (see section 3.24 of Volume 1). A removable interaction effect can potentially be reduced to zero by

some monotonic transformation of the y scores. To determine if the interaction is removable or non-removable, it is necessary to first determine the value of x_2 where the estimated conditional slope for x_1 ($\hat{\beta}_1 + \hat{\beta}_3 x_2$) changes sign and the value of x_1 where the estimated conditional slope for x_2 ($\hat{\beta}_2 + \hat{\beta}_3 x_1$) changes sign. The x_2 change point for the conditional slope for x_1 is $-\hat{\beta}_1/\hat{\beta}_3$, and the x_1 change point for the conditional slope for x_2 is $-\hat{\beta}_2/\hat{\beta}_3$. If the x_1 and x_2 change points are both outside their ranges of observed values, then the interaction effect is potentially removable; otherwise, the interaction is non-removable. For example, if $\hat{\beta}_1 = 4.52$, $\hat{\beta}_2 = 1.87$, and $\hat{\beta}_3 = -0.56$, the x_1 change point is $-\hat{\beta}_2/\hat{\beta}_3 = 1.87/.56 = 3.34$ and the x_2 change point is $-\hat{\beta}_1/\hat{\beta}_3 = 4.52/.56 = 8.07$. If 3.34 is outside the observed range of x_1 scores *and* 8.07 is outside the observed range of x_2 scores, then the interaction is potentially removable with some transformation of the y scores. But if 3.34 is within the observed range of x_1 scores *or* if 8.07 is within the range of observed x_2 scores, then the interaction between x_1 and x_2 is non-removable.

Three-way Interaction

A three-way interaction for variables x_1 , x_2 and x_3 is specified by creating a new predictor variable that is the product of x_1 , x_2 and x_3 . An example of a GLM with three predictor variables, all two-way interactions, and the three-way interaction is shown below.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} \quad (\text{Model 2.8})$$

$$+ \beta_4 x_{1i} x_{2i} + \beta_5 x_{1i} x_{3i} + \beta_6 x_{2i} x_{3i} + \beta_7 x_{1i} x_{2i} x_{3i} + e_i$$

where β_4 describes the two-way interaction for x_1 and x_2 , β_5 describes the two-way interaction for x_1 and x_3 , β_6 describes the two-way interaction for x_2 and x_3 , and β_7 describes the three-way interaction for x_1 , x_2 and x_3 .

The analysis strategy for a GLM with a three-way interaction follows the same basic principles for analyzing a three-factor design (see section 3.16 of Volume 1). If a three-way interaction is detected, that implies that each two-way interaction differs across the levels of a third variable. The conditional two-way interaction effects are analogous to simple two-way interaction effects in a three-factor design. The conditional interaction effect for x_1 and x_2 is $\beta_4 + \beta_7 x_3$, the conditional interaction effect for x_1 and x_3 is $\beta_5 + \beta_7 x_2$, and the conditional interaction effect for x_2 and x_3 is $\beta_6 + \beta_7 x_1$. Estimates of these conditional interaction effects can be obtained by replacing the population values with their estimates.

The conditional interaction at the levels of a dummy coded variable can be obtained two ways. For example, if x_3 is a dummy coded variable, then the conditional interaction for x_1 and x_2 is β_4 at $x_3 = 0$ and $\beta_4 + \beta_7$ at $x_3 = 1$. Estimates of these conditional interaction effects can be obtained by replacing the population values with their estimates. Alternatively, a multiple group regression analysis (section 2.17) could be used to assess the conditional interaction for x_1 and x_2 in the $x_3 = 0$ group and the $x_3 = 1$ group.

A three-way interaction also implies that the slope for, say x_1 , changes across combinations of values of x_2 and x_3 . Instead of analyzing conditional two-way interactions, it might be more interesting to examine the conditional slope for x_1 at four combinations of low and high values of x_2 and x_3 . The conditional slopes for one variable at combinations of low and high values of the other two variables are analogous to simple-simple main effects in a three-factor design (see section 3.14 of Volume 1).

The conditional slope for x_1 at specified values of x_2 and x_3 is $\beta_1 + \beta_4x_2 + \beta_5x_3 + \beta_7x_2x_3$, the conditional slope for x_2 at specified values of x_1 and x_3 is $\beta_2 + \beta_4x_1 + \beta_5x_3 + \beta_7x_1x_3$, and the conditional slope for x_3 at specified values of x_1 and x_2 is $\beta_3 + \beta_5x_1 + \beta_6x_2 + \beta_7x_1x_2$. Estimates of these conditional slopes can be obtained by replacing the population values with their estimates. Alternatively, if both of the predictor variables being conditioned on are dummy coded variables, then a multiple group regression analysis can be used to estimate the conditional slope within each of the four groups ($x_1 = 0$ and $x_2 = 0$, $x_1 = 1$ and $x_2 = 0$, $x_1 = 0$ and $x_2 = 1$, $x_1 = 1$ and $x_2 = 1$).

2.14 Confidence Interval for Conditional Slopes

In a model that contains a two-way interaction for x_1 and x_2 , a $100(1 - \alpha)\%$ confidence interval for the conditional slope of x_1 at $\dot{x}_2$ (where $\dot{x}_2$ is a particular value of x_2) is

$$\hat{\beta}_1 + \hat{\beta}_3\dot{x}_2 \pm t_{\alpha/2;df_E}SE_{\hat{\beta}_1 + \hat{\beta}_3\dot{x}_2} \quad (2.19)$$

where $df_E = n - s - 1$, $SE_{\hat{\beta}_1 + \hat{\beta}_3\dot{x}_2} = \sqrt{SE_{\hat{\beta}_1}^2 + \dot{x}_2^2 SE_{\hat{\beta}_3}^2 + 2\dot{x}_2 cov(\hat{\beta}_1, \hat{\beta}_3)}$, and $cov(\hat{\beta}_1, \hat{\beta}_3)$ is the estimated covariance between $\hat{\beta}_1$ and $\hat{\beta}_3$. SPSS and R have options to compute the estimated covariances among slope estimates. A $100(1 - \alpha)\%$ confidence interval for the conditional slope of x_2 at $\dot{x}_1$ is

$$\hat{\beta}_2 + \hat{\beta}_3\dot{x}_1 \pm t_{\alpha/2;df}SE_{\hat{\beta}_2 + \hat{\beta}_3\dot{x}_1} \quad (2.20)$$

where $SE_{\hat{\beta}_2 + \hat{\beta}_3 \dot{x}_1} = \sqrt{SE_{\hat{\beta}_2}^2 + \dot{x}_1^2 SE_{\hat{\beta}_3}^2 + 2\dot{x}_1 cov(\hat{\beta}_2, \hat{\beta}_3)}$. The `ci.condslope` function in the `statpsych` package will compute Equations 2.19 and 2.20.

An easier way to compute Equation 2.19 or 2.20 is to center x_1 around $\dot{x}_1$ and center x_2 around $\dot{x}_2$. If x_1 is centered around $\dot{x}_1$ then β_2 in Model 2.7 becomes $\beta_2 + \beta_3 \dot{x}_1$, and if x_2 is centered around $\dot{x}_2$ then β_1 in Model 2.7 becomes $\beta_1 + \beta_3 \dot{x}_2$. Converting β_1 and β_2 into conditional slopes is a simple way to compute Equations 2.19 and 2.20 because statistical packages will compute an estimate, standard error, test statistic, p -value, and confidence interval for β_1 and β_2 .

Equation 2.20 can also be used to obtain confidence intervals for conditional two-way interactions. For example, to compute the conditional interaction effect for x_1 and x_2 at $\dot{x}_3$, replace $\hat{\beta}_1$ with $\hat{\beta}_4$, replace $\hat{\beta}_3$ with $\hat{\beta}_7$, and replace $\dot{x}_1$ with $\dot{x}_3$. Alternatively, the conditional interaction effect for x_1 and x_2 at $\dot{x}_3$ could be obtained by centering x_3 around $\dot{x}_3$ and then β_4 becomes the conditional interaction effect for x_1 and x_2 at $\dot{x}_3$.

A general method that can be used to compute a confidence interval for a conditional slope at a combination of low and high values of two other variables can be obtained using Equation 2.38 in section 2.31. Alternatively, the method of re-centering can be used. For example, the conditional slope for x_1 at a low value of x_2 and a low value of x_3 can be obtained by centering x_2 around the low value of x_2 and centering x_3 around the low value of x_3 . After x_2 and x_3 have been centered this way, β_1 becomes the conditional slope for x_1 at the low value of x_2 and the low value of x_3 .

2.15 Detecting Interactions in Nonexperimental Designs

In two-factor and three-factor experimental designs, the interaction effects are routinely examined and reported. In nonexperimental designs, however, an interaction effect can be difficult to detect. As explained in section 3.24 of Volume 1, the sample size required to detect an interaction effect is larger than the sample size required to detect a main effect. Furthermore, the classification factors in a nonexperimental factorial design can have highly dissimilar sample sizes across the factor levels, and as explained in section 2.15 of Volume 1, the power of a hypothesis test is reduced when the sample sizes are unequal. Consequently, a large sample size may be required to detect a two-way interaction between two classification factors in a nonexperimental design. A three-way

interaction for three classification factors is even more difficult to detect. Likewise, a large sample size may be required to detect interaction effects for random predictor variables in a multiple regression model.

2.16 Dummy Coding and Effect Coding

An a -level qualitative predictor variable (e.g., Democrat/Republican/Independent, Treatment/Control, Freshman/Sophomore/Junior/Senior, etc.) can be used as a predictor variable in a GLM if it is first converted into $a - 1$ dummy coded variables. In Chapter 1 a single dummy coded variable was used to represent a dichotomous ($a = 2$) predictor variable. For an a -level factor, the dummy coded variable k ($k = 1$ to $a - 1$) is equal to 1 in level k and 0 otherwise. Two dummy coded variables x_1 and x_2 are needed for a 3-level factor as shown below.

level	x_1	x_2
1	1	0
2	0	1
3	0	0

In this example, all participants with level 1 of the predictor variable are assigned an x_1 score of 1 and an x_2 score of 0, all participants with level 2 of the predictor variable are assigned an x_1 score of 0 and an x_2 score of 1, and all participants with level 3 of the predictor variable are assigned an x_1 score of 0 and an x_2 score of 0. The level for which all the dummy codes are equal to 0 is called the *reference level*. In the above example, level 3 is the reference level.

The GLM for a 3-level qualitative factor is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \quad (\text{Model 2.9})$$

where $x_{1i} = 1$ if participant i belongs to level 1 (0 otherwise) and $x_{2i} = 1$ if participant i belongs to level 2 (0 otherwise). It can be shown that $\beta_0 = \mu_3$, $\beta_1 = \mu_1 - \mu_3$ and $\beta_2 = \mu_2 - \mu_3$. Equation 2.10 can be used to obtain confidence intervals for β_1 and β_2 , which are confidence intervals for $\mu_1 - \mu_3$ and $\mu_2 - \mu_3$, respectively. Note that $\beta_1 - \beta_2 = (\mu_1 - \mu_3) - (\mu_2 - \mu_3) = \mu_1 - \mu_2$ and so a confidence interval for $\mu_1 - \mu_2$ is obtained from a confidence interval for $\beta_1 - \beta_2$. Equation 2.33 in section 2.31 can be used to obtain a confidence interval for $\beta_1 - \beta_2$. In general, a GLM for a single a -level qualitative factor that has been dummy coded, the parameters are $\beta_0 = \mu_a$ and $\beta_k = \mu_k - \mu_a$ ($k = 1$ to $a - 1$).

Instead of using Equation 2.38 to obtain additional pairwise comparisons, it is usually easier to use a new set of dummy variables that will have a different reference level. In the above example with $a = 3$ levels, the following two alternative dummy coded variables could be used.

level	x_1	x_2
1	0	0
2	1	0
3	0	1

The reference level is now level 1 and the slope coefficient for x_1 is equal to $\mu_1 - \mu_2$.

A GLM can have more than one qualitative factor. Consider the simplest case of two qualitative factors (Factor A and Factor B) that each have two levels (i.e., a 2×2 factorial design). The two levels of Factor A are denoted as a_1 and a_2 , and the two levels of Factor B are denoted as b_1 and b_2 . The interaction between the two factors is coded by taking the product of the dummy codes for each factor. The dummy codes are shown below for a 2×2 factorial design.

A	B	x_1	x_2	x_1x_2
1	1	1	1	1
2	1	0	1	0
1	2	1	0	0
2	2	0	0	0

The GLM for a 2×2 factorial design is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + e_i \quad (\text{Model 2.10})$$

and the model parameters have the following definitions.

$\beta_0 = \mu_{22}$	(mean at a_2 and b_2)
$\beta_1 = \mu_{12} - \mu_{22}$	(simple main effect of A at b_2)
$\beta_2 = \mu_{21} - \mu_{22}$	(simple main effect of B at a_2)
$\beta_3 = \mu_{11} - \mu_{12} - \mu_{21} + \mu_{22}$	(AB interaction effect)

The other simple main effects and the two main effects are defined from the following linear functions of the model parameters.

$$\begin{aligned}
\beta_1 + \beta_3 &= \mu_{11} - \mu_{21} && \text{(simple main effect of A at } b_1) \\
\beta_2 + \beta_3 &= \mu_{11} - \mu_{12} && \text{(simple main effect of B at } a_1) \\
\beta_1 + \beta_3/2 &= (\mu_{11} + \mu_{12})/2 - (\mu_{21} + \mu_{22})/2 && \text{(main effect of A)} \\
\beta_2 + \beta_3/2 &= (\mu_{11} + \mu_{21})/2 - (\mu_{12} + \mu_{22})/2 && \text{(main effect of B)}
\end{aligned}$$

Equation 2.38 can be used to obtain a confidence interval for any linear function of model parameters. However, if all of the predictor variables in a GLM are qualitative factors, it would be much easier to specify contrast coefficients for a particular effect using Equation 3.1 of Volume 1. The above explanations of the slope parameters are helpful when interpreting the slope coefficients for the qualitative predictor variables in a GLM that contains both qualitative and quantitative predictor variables because in these applications Equation 3.1 of Volume 1 is not applicable.

If the product variable is removed from Model 2.10 because the interaction effect is assumed to be small or unimportant, then β_1 becomes the main effect of Factor A and β_2 becomes the main effect of Factor B.

With *effect coding*, the predictor variables have values of 1, 0, and -1 (or just 1 and -1 if there are only two categories). Effect coding is also referred to as *deviation coding*. As with dummy coding, $a - 1$ effect coded variables are needed to code a qualitative factor with $a \geq 2$ levels. The two effect coded variables for a qualitative factor with three levels are shown below.

level	x_1	x_2
1	1	0
2	0	1
3	-1	-1

For a qualitative factor with a levels, the effect coded variable k ($k = 1$ to $a - 1$) is equal to 1 in level k , -1 in the last level, and 0 otherwise. In this general case, $\beta_0 = (\mu_1 + \mu_2 + \dots + \mu_a)/a$ and $\beta_k = \mu_k - \beta_0$. If β_k is multiplied by $a/(a - 1)$, then β_k describes the difference between μ_k and the average of the other means. For example, with $a = 3$, $(3/2)\beta_1 = \mu_1 - (\mu_2 + \mu_3)/2$ which might be more meaningful than $\mu_1 - \beta_0$.

To obtain a confidence interval for a pairwise mean difference, say $\mu_1 - \mu_2$, it would be necessary to compute a confidence interval for $\beta_1 - \beta_2$, which equals $(\mu_1 - \beta_0) - (\mu_2 - \beta_0)$

$= \mu_1 - \mu_2$. However, pairwise comparisons involving the last category are not obvious. For $a = 3$, $\mu_1 - \mu_3 = 2\beta_1 + \beta_2$ and $\mu_2 - \mu_3 = \beta_1 + 2\beta_2$.

If the factor has $a = 2$ levels, only one effect coded variable is required and the model is $y_i = \beta_0 + \beta_1 x_i + e_i$ with participants at level 1 assigned an x_i score of 1 and participants at level 2 assigned an x_i score of -1. In this model $\beta_0 = (\mu_1 + \mu_2)/2$ and $\beta_1 = (\mu_1 - \mu_2)/2$. Note that the estimate and confidence interval for β_1 must be multiplied by 2 to obtain an estimate and confidence interval for $\mu_1 - \mu_2$.

The effect coded variables are shown below for a 2×2 factorial design.

A	B	x_1	x_2	$x_1 x_2$
1	1	1	1	1
2	1	-1	1	-1
1	2	1	-1	-1
2	2	-1	-1	1

With effect coding, the parameters of Model 2.10 have the following definitions.

$$\begin{aligned}
 \beta_0 &= (\mu_{11} + \mu_{12} + \mu_{21} + \mu_{22})/4 && \text{(grand mean)} \\
 \beta_1 &= (\mu_{11} + \mu_{12})/4 - (\mu_{21} + \mu_{22})/4 && \text{(main effect of A divided by 2)} \\
 \beta_2 &= (\mu_{11} + \mu_{21})/4 - (\mu_{12} + \mu_{22})/4 && \text{(main effect of B divided by 2)} \\
 \beta_3 &= (\mu_{11} - \mu_{12} - \mu_{21} + \mu_{22})/4 && \text{(AB interaction effect divided by 4)}
 \end{aligned}$$

Note that the estimates and confidence intervals for β_1 and β_2 must be multiplied by 2 to obtain the estimates and confidence intervals for the two main effects. The estimate and confidence interval for β_3 must be multiplied by 4 to obtain the estimate and confidence interval for the interaction effect.

As noted above, if dummy coding is used and the product variable is removed from Model 2.10, β_1 becomes the main effect of A and β_2 becomes the main effect of B. But if effect coding is used and the product variable is removed from the model, the interpretations of β_1 and β_2 are unchanged.

In a 2×2 factorial design where it is important to assess the interaction effect and the main effects, effect coding can be used to report the t -value and p -value for each effect. If the interaction is nonsignificant or small, the analysis would typically focus on the main effects; otherwise, simple main effects would be analyzed. With effect coding, the lower

and upper confidence limits must be multiplied by 2 for the main effects and multiplied by 4 for the interaction effect. Dummy coding is preferred to effect coding if the model contains just one qualitative factor or in factorial designs where the model does not include any interaction effects. If simple main effects need to be examined, dummy coding will produce some of the simple main effects, and then the other simple main effects can be obtained by changing the reference level. It is easy to change the reference level in SPSS and R.

When analyzing factorial designs using statistical software, it is important to know how the factors have been coded. In the SPSS-GLM procedure, any variable that is specified as a factor will be dummy coded. If parameter estimates are requested in a two-factor design where the model includes an interaction effect, it is important to remember the results for Factor A will be the simple main effects of Factor A at the last level of Factor B, and the results for Factor B will be simple main effects of Factor B at the last level of Factor A. However, the ANOVA summary table in the SPSS-GLM procedure gives the hypothesis testing results for main effects and not simple main effects. The Marginal Means option in the SPSS-GLM procedure will provide estimates and confidence intervals for main effects in factorial designs. R has options to define a qualitative factor using dummy coding or effect coding.

2.17 Comparing Two Correlations

Let $\rho_{yx[1]}$ represent the population correlation between y and x in study population 1, and let $\rho_{yx[2]}$ represent the population correlation between y and x in study population 2. A random sample from study population 1 will be used to estimate $\rho_{yx[1]}$, and a random sample from study population 2 will be used to estimate $\rho_{yx[2]}$. Let $\hat{\rho}_{yx[1]}$ represent the estimated correlation from sample 1, and let $\hat{\rho}_{yx[2]}$ represent the estimated correlation from sample 2. Let L_1 and U_1 denote the lower and upper $100(1 - \alpha)\%$ confidence interval endpoint computed from sample 1, and let L_2 and U_2 denote the lower and upper $100(1 - \alpha)\%$ confidence interval endpoint computed from sample 2. The lower and upper $100(1 - \alpha)\%$ confidence interval endpoints for $\rho_{yx[1]} - \rho_{yx[2]}$ are given below.

$$LL = \hat{\rho}_{yx[1]} - \hat{\rho}_{yx[2]} - \sqrt{(\hat{\rho}_{yx[1]} - L_1)^2 + (\hat{\rho}_{yx[2]} - U_2)^2} \quad (2.21a)$$

$$UL = \hat{\rho}_{yx[1]} - \hat{\rho}_{yx[2]} + \sqrt{(\hat{\rho}_{yx[1]} - U_1)^2 + (\hat{\rho}_{yx[2]} - L_2)^2} \quad (2.21b)$$

Equations 2.21a and 2.21b are appropriate for Pearson, Spearman, point-biserial correlations, semipartial, partial, and squared multiple correlations. The `ci.cor2` function in the `statpsych` package will compute Equations 2.21a and 2.21b for Pearson correlations. The `ci.cor2.gen` function in the `statpsych` package will compute Equations 2.21a and 2.21b for any type of correlation if the confidence interval for each correlation has already been computed.

Example 2.7. The estimated correlation between problem solving and social skills is .64 in a random sample of 200 first-grade students (95% CI [.55, .71]) and .31 in a random sample of 200 sixth-grade students (95% CI [.18, .43]). A 95% confidence interval for the difference in population correlations between first graders and sixth graders is [.18, .48]. This result indicates that the population correlation between problem solving and social skills is .18 to .48 greater in the study population of first graders than in the study population of sixth graders.

Now let ρ_{yx_1} represent the population Pearson correlation between y and x_1 , and let ρ_{yx_2} represent the population Pearson correlation between y and x_2 where x_1 and x_2 are two different predictor variables and all three variables (y , x_1 , x_2) are measured in the *same* sample. The lower and upper $100(1 - \alpha)\%$ confidence interval endpoints for $\rho_{yx_1} - \rho_{yx_2}$ are

$$LL = \hat{\rho}_{yx_1} - \hat{\rho}_{yx_2} - \sqrt{(\hat{\rho}_{yx_1} - L_1)^2 + (\hat{\rho}_{yx_2} - U_2)^2 - 2c(\hat{\rho}_{yx_1} - L_1)(\hat{\rho}_{yx_2} - U_2)} \quad (2.22a)$$

$$UL = \hat{\rho}_{yx_1} - \hat{\rho}_{yx_2} + \sqrt{(\hat{\rho}_{yx_1} - U_1)^2 + (\hat{\rho}_{yx_2} - L_2)^2 - 2c(\hat{\rho}_{yx_1} - U_1)(\hat{\rho}_{yx_2} - L_2)} \quad (2.22b)$$

where $c = [(\hat{\rho}_{x_1x_2} - \hat{\rho}_{yx_1}\hat{\rho}_{yx_2}/2)(1 - \hat{\rho}_{yx_1}^2 - \hat{\rho}_{yx_2}^2 - \hat{\rho}_{x_1x_2}^2) + \hat{\rho}_{x_1x_2}^3]/[(1 - \hat{\rho}_{yx_1}^2)(1 - \hat{\rho}_{yx_2}^2)]$. Equations 2.22a and 2.22b apply only to Pearson correlations. The `ci.cor.dep` function in the `statpsych` package will compute Equations 2.22a and 2.22b.

Example 2.8. In a random sample of $n = 166$ social workers, the estimated correlation between job performance and creativity is .396 (95% CI [.259, .517]). The estimated correlation between job performance and college GPA is .179 (95% CI [.027, .323]). The estimated correlation between creativity and GPA is .088. A 95% confidence interval for the difference in the correlation between job performance and creativity and the correlation between job performance and GPA is [.013, .416].

2.18 Equivalence Test

A confidence interval for $\rho_{yx[1]} - \rho_{yx[2]}$ can be used to perform an equivalence test for two population correlations. The null and alternative hypotheses are given below

$$H_1: |\rho_{yx[1]} - \rho_{yx[2]}| < h \quad H_2: |\rho_{yx[1]} - \rho_{yx[2]}| \geq h$$

where h is a number specified by the researcher that represents a small or unimportant difference between the two population correlations. Note that $H_1: |\rho_{yx[1]} - \rho_{yx[2]}| < h$ can be expressed as $H_1: -h < \rho_{yx[1]} - \rho_{yx[2]} < h$ where the interval $-h$ to h is the region of practical equivalence. H_1 is accepted if the confidence interval for $\rho_{yx[1]} - \rho_{yx[2]}$ is completely within the $-h$ to h range, and H_2 is accepted if the confidence interval is completely outside the $-h$ to h range. The results are inconclusive if the confidence interval includes the values h or $-h$. Equivalence tests are usually performed using a 90% confidence interval rather than a 95% confidence interval. Likewise, a confidence interval for $\rho_{yx_1} - \rho_{yx_2}$ (Equations 2.22ab) can be used to show the equivalence of ρ_{yx_1} and ρ_{yx_2} . An equivalence test can also be applied to slopes but then it usually more difficult to specify the value of h .

Example 2.9. A new reading readiness test for 5-year-old children was given to a random sample of boys and a random sample of girls. A standardized reading performance was given to these students one year later. In this study, the researchers believe that the new reading readiness test could be gender biased and will require modifications if the correlation between the reading readiness test and the reading performance test differs by more than .15 in the populations of male and female children. A 90% confidence interval for $\rho_{yx[1]} - \rho_{yx[2]}$ is [-0.26, -0.19] and suggests that the new reading readiness test could be gender biased because there is at least a .19 difference in the two population correlations.

2.19 Analysis of Covariance Model

An ANCOVA model is a GLM with one or more qualitative factors and one or more quantitative predictor variables. The quantitative predictor variables in the ANCOVA

model are often referred to as *covariates*. In experimental designs, covariates are used primarily to reduce prediction error variance and secondarily because the researcher is interested in how the covariates relate to the dependent variable or interact with the treatment factor(s). Reducing prediction error variance has the beneficial effects of narrowing the confidence interval widths and increasing the power of the hypothesis tests. Including one or more covariates in an experiment can substantially reduce the sample size requirement.

Consider an ANCOVA model for a 2-level factor and one covariate. The following model includes one dummy coded variable (x_1), one covariate (x_2), and their product ($x_3 = x_1x_2$)

$$y_i = \beta_0 + \beta_1x_{1i} + \beta_2x_{2i} + \beta_3x_{3i} + e_i. \quad (\text{Model 2.11})$$

The covariate can be mean centered by subtracting the overall sample covariate mean from every covariate score. If a confidence interval for β_3 includes 0 or if the confidence interval suggests that the interaction effect could be small, a confidence interval or test for β_2 would be examined. Otherwise, the simple main effects of the 2-level factor would be examined at two or more values of the covariate. A simple main effect of the 2-level factor at a covariate value of $\bar{x}_2$ is $\beta_1 + \beta_3\bar{x}_2$. The conditional slope for the covariate (x_2) within each level of x_1 could also be of interest. The conditional slope for x_2 at $x_1 = 1$ is $\beta_2 + \beta_3$ and the conditional slope for x_2 at $x_1 = 0$ is β_2 .

Example 2.10. A researcher suspects that multiple attempts to recall learned information, without performance feedback, will improve recall performance. A two-group experiment was conducted where all participants viewed a 50-minute video lecture. The first group was given a short test over the lecture material without performance feedback every day for five days. The second group was not tested during this 5-day period. Ten days later, all participants were given a test of the lecture material (scored 0 to 100). The ANOVA model is $y_i = \beta_0 + \beta_1x_{1i} + e_i$ where x_1 is a dummy coded variable. The 95% confidence interval for β_1 is [1.1, 18.8], suggesting that attempts to recall information without feedback will improve retention of learned material. The researcher also obtained the GPA for each participant, which is believed to correlate with the final test score. The ANCOVA model is $y_i = \beta_0 + \beta_1x_{1i} + \beta_2x_{2i} + \beta_3x_{3i} + e_i$ where x_1 is the dummy coded variable and x_2 is GPA. The interaction effect appeared to be small, and a 95% confidence interval for β_1 is [6.7, 13.2]. This confidence interval is considerably narrower than the confidence interval for β_1 using the ANOVA model.

The ANCOVA model is also used in nonexperimental designs where participants have self-selected into two or more treatment conditions. When participants self-select into treatment conditions, the slope coefficients for the treatment indicator variables will contain confounding variable bias because the characteristics of the participants in different treatment conditions could differ due to the causal effects of one or more unmeasured attributes that also have a causal effect on the response variable. If the most important confounding variables can be included in the model as covariates, the confounding variable bias could be substantially reduced and then the slope coefficients for the treatment effects will more closely approximate a possible causal relation between the treatment variable and the response variable.

An ANCOVA model is also useful in an experimental design where the random assignment process was not carried out properly. With imperfect random assignment, the population means of one or more covariates could differ among the treatment conditions. Including these covariates in the analysis could improve the interval validity of the study.

The ANCOVA model is most effective in a nonexperimental design or a flawed experimental design if the means of the covariates do not differ substantially across groups. We typically assume that each covariate is linearly related to the response variable. If the groups differ substantially in their mean covariate values, then the linearity assumption would need to hold over a wide range of covariate values. If the linearity assumption cannot be justified over the observed range of covariate values, then the ANCOVA model will not properly control for the confounding variables.

Example 2.11. A researcher obtained a sample of retirees who began working crossword puzzles after retirement and another sample of retirees who did not take up crossword puzzles. The ANOVA model is $y_i = \beta_0 + \beta_1 x_{1i} + e_i$ where y is a measure of cognitive ability and x_1 is a dummy coded variable. The researcher obtained a 95% confidence interval for β_1 of [8.2, 12.5]. This result suggests that the average cognitive ability is higher for retirees who began working crossword puzzles after retirement. However, it is possible that retirees who take up crossword puzzles after retirement had greater cognitive ability prior to retirement than retirees who do not take up crossword puzzles. Using years of education as a covariate, the ANCOVA model is $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}x_{2i} + e_i$ where x_1 is the dummy coded variable and x_2 is years of college education. The confidence interval for β_3 included zero, and the 95% confidence interval for β_1 is [-0.4, 2.1]. The result for β_1 is inconclusive but the upper limit of 2.1 points suggests that taking up crossword puzzles after retirement could, at best, have a very small effect on cognitive ability.

The ANCOVA has another important application in studies that compare two different treatments (e.g., Treatment 1 and Treatment 2) and certain participants are expected to benefit more from Treatment 1 than Treatment 2. In these situations, it may be difficult to ethically justify an experiment that randomly assigns participants to treatment conditions. However, if a “priority score” can be assigned to each participant that assesses the degree to which a participant might benefit from Treatment 1, all participants with priority scores above some cut point would be assigned to Treatment 1 and all participants with priority scores below the cut point would be assigned to Treatment 2. It is a remarkable fact that if the priority score is used as a covariate in a two-group ANCOVA, the slope coefficient for the treatment dummy variable will describe the *causal* effect of treatment on the response variable even though participants were not randomly assigned to treatments. The causal interpretation assumes a linear relation between the priority scores and the response variable and no priority $\times$ treatment interaction. This type of design is called a *regression discontinuity design*.

Example 2.12. A new program to provide low-income college-bound students with financial aid application training was evaluated using a random sample of 100 low-income high school students in the Los Angeles school district. All students from families making less than \$40,000 per year received the financial aid application training, and all other students received the usual assistance from their high school counselor. The response variable was the amount of financial aid obtained. The 95% confidence interval for the dummy variable slope coefficient is [\$602, \$918] indicating that if all low-income college-bound students in the Los Angeles school district had received the financial aid application training, their mean financial aid would be \$602 to \$918 higher than if they had received the typical assistance provided by a high school counselor. Note that the researcher was able to make a causal claim about the effectiveness of the new program even though students were not randomly assigned to groups.

Although the regression discontinuity design can reduce the “ethical costs” of a study, the regression discontinuity design requires more than twice as many participants as the corresponding two-group experimental design to achieve the same hypothesis testing power and confidence interval precision. Thus, the regression discontinuity design will have lower ethical costs but greater sampling costs (e.g., measurement costs, cost of administering treatment, payment for participation) than a two-group experimental design.

The sampling costs of the regression discontinuity design can be substantially reduced by specifying an “indifference range” of priority scores for which either treatment could

be beneficial and then randomly assigning participants with priority scores within this range to one of the two treatment conditions. All participants with priority scores below the lower limit of the indifference range are assigned to Treatment 2, and all participants with priority scores above the upper limit of the indifference range are assigned to Treatment 1. Increasing the width of the indifference range decreases the sample size requirements of the regression discontinuity design. For example, with a wide indifference range that includes about 50% of the priority scores, the sample size requirement for the regression discontinuity design only requires about 30% more participants than the corresponding two-group experimental design. If an indifference range is used, the detrimental effects of violating the linearity assumption and the assumption of no priority score by treatment interaction are less serious.

In experimental and nonexperimental designs, measurement error in the covariates will attenuate the correlation between the response variable and each covariate. In an experimental design, this attenuation reduces the benefits of using a covariate in terms of hypothesis testing power or confidence interval precision for the treatment effects. In nonexperimental designs, where confounding variables are included in the model as covariates, this attenuation is much more serious because the effects of the confounding variables will not be completely removed. Methods for reducing the negative effects of covariate measurement error in nonexperimental designs are discussed in Volume 4. Measurement error in the priority scores does not introduce any problems in a regression discontinuity design.

2.20 Multiple Group Regression Model

In a between-subjects design with m groups and s quantitative predictor variables, the following multiple regression model can be analyzed separately within each group ($j = 1$ to m)

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \beta_{2j}x_{2ij} + \cdots + \beta_{sj}x_{sij} + e_{ij} \quad (\text{Model 2.12})$$

and then the slope coefficients and y -intercepts can be compared across groups. The m groups can represent the levels of a single between-subjects factor or the combination of levels of two or more between-subjects factors. Each factor can be a between-subject treatment factor or a between-subjects classification factor.

The multiple group regression model is an alternative to the ANCOVA model that can be used with any of the between-subjects designs described in Chapter 3 of Volume 1. Each slope coefficient in the multiple group regression model is a conditional effect of a quantitative predictor variable at group j .

Each predictor variable in Model 2.12 can be grand-mean centered using the mean computed from all m groups. Grand-mean centering the predictor variables will change the interpretation of the y -intercepts but will not affect the values of the slope coefficients.

A $100(1 - \alpha)\%$ confidence interval for a linear contrast of the k^{th} parameter values ($k = 0$ to s) in Model 2.12 across the m groups is given below

$$\sum_j^m v_j \hat{\beta}_{kj} \pm t_{\alpha/2; df} \sqrt{\sum_j^m v_j^2 SE_{\hat{\beta}_{kj}}^2} \quad (2.23)$$

where $df = (\sum_j^m v_j^2 SE_{\hat{\beta}_{kj}}^2)^2 / [\sum_j^m v_j^4 SE_{\hat{\beta}_{kj}}^4 / (n_j - s - 1)]$ and the v_j values are specified by the researcher.

The `ci.lc.reg` function in the `statpsych` package will compute Equation 2.23 using the parameter estimates and standard errors that are reported in a linear regression program within each group. Equation 2.23 does not assume equal prediction error variances across groups. To apply Equation 2.23, first estimate the y -intercept and slope(s) of a linear regression model within each of the m groups. The `ci.lc.reg` function requires the estimate and its standard error for a specific parameter within each group, the group sample sizes, and the v_j coefficients.

In an experimental design, the random assignment of participants to treatment conditions implies that the population means of all covariates will be equal across the treatment conditions. If the population slopes in Model 2.12 are assumed to be approximately equal across the treatment conditions, then a linear contrast of the population y -intercepts ($\sum_j^m v_j \beta_{0j}$) can be interpreted in the same way as the linear contrast of population means ($\sum_j^m v_j \mu_j$) described in Chapter 3 of Volume 1. If any of the predictor variables are at least moderately correlated with the response variable, a confidence interval for $\sum_j^m v_j \beta_{0j}$ can be considerably narrower than the confidence interval for $\sum_j^m v_j \mu_j$, and hypothesis tests for $\sum_j^m v_j \beta_{0j}$ can have considerably more power than hypothesis tests for $\sum_j^m v_j \mu_j$.

The parameters of an ANCOVA model can be defined in terms of the parameters of a multiple group regression model. For example, in the following two-group linear regression model with one covariate

$$y_{i1} = \beta_{01} + \beta_{11}x_{12i} + e_{i1} \quad (\text{Model 2.13a})$$

$$y_{i2} = \beta_{02} + \beta_{12}x_{12i} + e_{i2} \quad (\text{Model 2.13b})$$

where the second subscript indicates group membership (1 or 2), it can be shown that β_1 , β_2 , and β_3 in Model 2.11 can be defined in terms of the parameters in Models 2.13a and 2.13b. Specifically, $\beta_1 = \beta_{12}$, $\beta_2 = \beta_{01} - \beta_{02}$, and $\beta_3 = \beta_{11} - \beta_{12}$. In Models 2.13a and 2.13b, β_{11} and β_{21} are the conditional slopes for x_1 at each level of the dummy variable in Model 2.10.

A confidence interval for $\beta_{11} - \beta_{12}$ can be obtained from Equation 2.23. Suppose the confidence interval for $\beta_{11} - \beta_{12}$ does not include 0. This implies a covariate by group interaction, and a confidence interval for $\beta_{01} - \beta_{02}$ should be computed for both low and high values of the covariate. To assess $\beta_{01} - \beta_{02}$ at a low value of the covariate, center x_1 around the low value and apply Equation 2.23. To assess $\beta_{01} - \beta_{02}$ at a high value of the covariate, re-center x_1 around the high value and apply Equation 2.23 again.

Now suppose the confidence interval for $\beta_{11} - \beta_{12}$ includes 0. It is customary to make a simplifying assumption that these population slopes are similar, and then it would be appropriate to compute a confidence interval for $\beta_{01} - \beta_{02}$ using Equation 2.23 after x_1 has been grand-mean centered.

It can be challenging to specify the main effects and simple main effect in an ANCOVA model with more than one factor. These effects are more easily analyzed using a multiple group regression model and Equation 2.23. The multiple group regression approach has the added benefit of not requiring equal prediction error variances across groups.

In a multiple-group regression analysis, it might be informative to estimate a squared multiple correlation within each group. Pairwise comparisons of squared multiple correlations might also be informative. After computing an estimate and confidence interval for a squared multiple correlation within each group (using the `ci.rsqr` function in the `statpsych` package), Equations 2.21a and 2.21b can be used to obtain a confidence

interval for any pairwise comparison of population squared multiple correlations (which can be computed using the `ci.cor2.gen` function in the `statpsych` package).

In addition to an examination of the squared multiple correlation within each group, an examination of the MAPE within each group could also be informative. After computing an estimate and confidence interval for a population MAPE within each group (using the `ci.mape` function in the `statpsych` package), the `ci.mape2` function in the `statpsych` package can be used to compute a confidence interval for any pairwise comparison of population MAPEs.

2.21 Nested Model Comparisons

Consider the following two GLMs

$$y_i = \ddot{\beta}_0 + \ddot{\beta}_1 x_{1i} + \ddot{\beta}_2 x_{2i} + \ddot{e}_i \quad (\text{Model 2.14})$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + e_i \quad (\text{Model 2.15})$$

where Model 2.14 is said to be *nested* in Model 2.15 because every predictor variable in Model 2.14 is also in Model 2.15. The double-dot notation in Model 2.14 is a reminder that β_0 , β_1 , β_2 , and e_i do not necessarily have the same values in Models 2.14 and 2.15. Model 2.14 is called the *reduced model* and Model 2.15 is called the *full model*. In general, the reduced model can have s_1 predictor variables and the full model will have an additional s_2 predictor variables so that the total number of predictor variables in the full model is $s = s_1 + s_2$.

The variance of the y scores can be decomposed into MODEL and ERROR sources as explained in section 2.4 for both the reduced model and the full model. Let $SS_{E:R}$ represent the ERROR sum of squares for the reduced model, and let $SS_{E:F}$ represent the ERROR sum of squares for the full model. The total sum of squares (SS_T) is the same in both models. The ERROR degrees of freedom is $n - s_1 - 1$ for the reduced model and $n - s - 1$ for the full model.

A *model comparison test* is a test of the omnibus null hypothesis

$$H_0: \beta_{s_1+1} = \beta_{s_1+2} = \dots = \beta_s = 0$$

which states that the population slopes for all of the additional s_2 predictor variables in the full model are equal to zero. The following F test statistic is used to test this omnibus null hypothesis

$$F = \frac{(SS_{E:R} - SS_{E:F})/s_2}{SS_{E:F}/(n - s - 1)} \quad (2.24)$$

and the degrees of freedom for this F test are s_2 and $n - s - 1$. This F statistic and its corresponding p -value can be obtained in SPSS and R. The test of the model comparison omnibus hypothesis has the same limitations as the omnibus test described in section 2.4. Specifically, a failure to reject H_0 does not imply that H_0 is true and a rejection of H_0 does not provide useful scientific information about the signs or magnitudes of the additional s_2 population slope coefficients that are in the full model.

Despite the limitations of the model comparison F test, model comparisons are popular because the difference between the coefficients of determination for the full and reduced models provide a description of the proportion of variance in the y scores that is predictable from the s_2 additional predictor variables in the full model, controlling for the s_1 predictor variables in the reduced model. A model comparison is often used when the s_2 additional predictor variables in the full model are the predictor variables of primary theoretical interest, and the s_1 predictor variables in the reduced model are believed to be confounding variables.

The difference in the full and reduced coefficients of determination is reported in APA journals as ΔR^2 and can be computed using either of the following formulas

$$\Delta R^2 = (SS_{E:R} - SS_{E:F})/SS_T = \hat{\rho}_{y,x}^2 - \hat{\rho}_{y,x_0}^2 \quad (2.25)$$

where $\hat{\rho}_{y,x}^2$ is the estimated squared multiple correlation for the full model and $\hat{\rho}_{y,x_0}^2$ is the estimated squared multiple correlation for the reduced model. If all of the s_2 additional predictor variables in the full model are random rather than fixed, then ΔR^2 is also referred to as a *squared multiple semipartial correlation*. Note that if the full model contains only one additional predictor variable ($s_2 = 1$), then ΔR^2 is equal to the squared semipartial correlation described in section 2.10. It is important to remember that ΔR^2 is a sample estimate that contains sampling error of unknown magnitude. Ideally, a confidence interval for the population ΔR^2 value would be reported when comparing full and reduced models, but no simple confidence interval method is currently available if

$s_2 > 1$. If the ERROR degrees of freedom is large (at least several hundred), then the sampling error in ΔR^2 should be small and the ΔR^2 value could cautiously be interpreted as an approximation to the population ΔR^2 value.

In applications where the slope coefficients for each of the s_2 predictor variables in the full model have theoretical importance, the recommended alternative to the model comparison F test is to simply analyze the full model and report a confidence interval and p -value for each of the slope coefficients in the full model.

2.22 Simultaneous Confidence Intervals and Hypothesis Tests

In applications where the values of the slopes, semipartial correlations, or partial correlations for $g \leq s$ of the predictor variables have important theoretical or practical implications, it is important to report Bonferroni simultaneous confidence intervals for these g population parameters. Suppose a theory would be supported if the population semipartial correlations for x_1 and x_2 are at least moderately large and positive and the population semipartial correlation for x_3 is at least moderately large and negative, then it would be appropriate to compute Bonferroni simultaneous confidence intervals for these three population semipartial correlations. It is usually not necessary to compute simultaneous confidence intervals for two or more control variables or covariates.

In some applications, a particular theory would be supported if the signs of the slope coefficients agreed with theoretical expectations. The sign of a semipartial correlation and a partial correlation for x_k is always the same as the sign of the slope for x_k . For example, a theory could be supported if the slope for x_1 is positive and the slope for x_2 is negative. Directional two-sided hypothesis tests can be used to make decisions about the signs of the slope coefficients. Although Bonferroni simultaneous confidence intervals can also be used to test multiple directional two-sided hypotheses to keep the familywise directional error rate (FWDER) for all g tests at or below $\alpha/2$, the Holm procedure described in Chapter 3 of Volume 1 provides greater power and also keeps the FWDER at or below $\alpha/2$. In the special case where $g = 2$, if $H_0: \beta_1 = \beta_2 = 0$ can be rejected using the F test described in section 2.4 or the model comparison test described in section 2.19, then follow-up directional two-sided tests for β_1 and β_2 can be performed without any adjustment to α . This approach for $g = 2$ is an adaptation of the Fisher protected test described in Chapter 3 of Volume 1. It is usually not necessary to apply the Holm procedure for tests of two or more control variables or covariates.

2.23 Confidence Interval for a Conditional Mean

In section 1.12, an estimate and confidence interval for a conditional mean was given for a linear statistical model with a single quantitative factor. In a GLM, the estimated conditional mean of y at $\hat{x}_1, \hat{x}_2, \dots, \hat{x}_s$ is

$$\hat{\mu}_{y|\hat{\mathbf{x}}} = \hat{\beta}_0 + \hat{\beta}_1\hat{x}_1 + \hat{\beta}_2\hat{x}_2 + \dots + \hat{\beta}_s\hat{x}_s \quad (2.26)$$

and a confidence interval for $\mu_{y|\hat{\mathbf{x}}}$ is given in Equation 2.40 of section 2.37. SPSS and R have options to compute Equations 2.26 and 2.40.

In a GLM, conditional means are especially useful in ANCOVA models that have quantitative or qualitative factors with one or more covariates. The estimated conditional mean at each level of a factor can also be used in to produce a bar chart (see section 3.26 in Volume 1) of conditional means with confidence interval lines. Typically, Equation 2.26 is computed for each level of a factor with the covariates set at their average values. If the model contains an interaction between a factor and a covariate, then Equation 2.26 could be computed for each factor level at one standard deviation below the covariate mean and one standard deviation above the covariate mean.

Example 2.13. In *Example 1.3*, 60 randomly selected students were given a \$25 fitness tracker to keep. The participants were given the opportunity to sell their tracker to the researcher. The response variable was the dollar amount they wanted for their tracker. The students were randomly assigned to three groups and were told, in groups 1, 2 and 3 respectively, that they would be paid immediately, in 7 days, or in 14 days. Now suppose each participant was given a food insecurity questionnaire at the beginning of the study. The food insecurity scores had a mean of 12.5 and a standard deviation of 4.2. The estimated parameters for payment delay (x_1), mean centered food insecurity (x_2) and the interaction (x_1x_2) are $\hat{\beta}_0 = 9.1$, $\hat{\beta}_1 = 0.72$, $\hat{\beta}_2 = -1.43$, and $\hat{\beta}_3 = 0.07$. The estimated conditional mean asking price for 0, 7, and 14 days of payment delay at a low food insecurity score of $12.5 - 4.2 = 8.3$ are \$15.10 (95% CI [14.7, 15.5]), \$20.10 (95% CI [19.8, 20.4]), and \$25.20 (95% CI [24.8, 25.6]), respectively. The estimated conditional mean asking price for 0, 7, and 14 days of payment delay at a high food insecurity score of $12.5 + 4.2 = 16.7$ are \$3.10 (95% CI [2.7, 3.5]), \$8.10 (95% CI [7.8, 8.4]), and \$13.20 (95% CI [12.8, 13.6]), respectively.

2.24 Prediction Interval for a Predicted y Score

In section 1.13, a linear statistical model with one predictor variable was used to predict the y score for a new person in the study population, and a prediction interval (Equation 1.19) for the predicted y score was given. The GLM can be used to predict the y score for

a new person of the study population using two or more predictor variables. Compared to the prediction of a y score using a single predictor variable, the y score can often be predicted more accurately using two or more predictor variables. As explained in section 1.13, in applications where the response variable is costly or difficult to measure, but the measurements of the predictor variables are relatively easy to obtain, a random sample of participants could be obtained where it is feasible to obtain a measurement of the costly response variable along with measurements of the predictor variables. The parameters of a GLM are then estimated using the sample data, and the predicted y score for a new person in the study population is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1\hat{x}_1 + \hat{\beta}_2\hat{x}_2 + \dots + \hat{\beta}_s\hat{x}_s$ where $\hat{x}_1, \hat{x}_2, \dots, \hat{x}_s$ are the predictor variables scores for the new person. A prediction interval for the predicted y score can be computed using Equation 2.41 in section 2.37. SPSS and R have options to compute the predicted y score and Equation 2.41 for a new person.

Example 2.14. In Example 1.5, fifty randomly selected job applicants who had passed an initial screening review were given a newly developed job screening test. The screening test (x_1) was scored on a 1 to 100 scale. Now suppose college GPA (x_2) and relevant job experience (1 = yes, 0 = no) were also recorded for each participant. All fifty applicants were hired and their job performance (y) was assessed on a scale of 0 to 10 three months later using a team of expert evaluators. The x_1 scores were mean centered and the estimates of β_0 , β_1 , β_2 , and β_3 are -1.84, .058, 1.26, and 1.82, respectively. A new applicant was given the screening test and obtained a score of 78. The new applicant had a college GPA of 3.54 and relevant job experience. The estimated y score for this applicant is $-1.84 + .058(78) + 1.26(3.4) + 1.82(1) = 8.78$. The 95% prediction interval for the job performance score of this one applicant is [8.4, 9.1], which is narrower than the prediction interval in Example 1.5 where only x_1 was used to predict y .

2.25 Exploratory Model Selection

In applications where the researcher has measured many potential predictor variables and wants to determine which of these variables are most useful in predicting y , SPSS and R can perform an exploratory model selection using methods (e.g., forward inclusion, backwards elimination, stepwise) that will sift through a set of candidate predictor variables and attempt to select the best predictor variables. Although these exploratory procedures are appealing, they can produce very misleading results. In general, the p -values for the selected predictor variables will be too small and the confidence intervals for the slope coefficients of the selected predictor variables will be too narrow. Consequently, the probability of making a directional error in a hypothesis

test can be much greater than $\alpha/2$ and the coverage probability of a 95% confidence interval can be much less than .95.

If a large sample is available, the sample can be randomly divided into two samples. The exploratory model selection is then performed in one sample (the *training sample*) and the predictor variables selected in the training sample are used in the second sample (the *test sample*). The results in the test sample will not be misleading and only the results in the test sample are reported. Note however that the parameter estimates (e.g., $\hat{\rho}_{y.x}^2$, $\hat{\rho}_{y(x_1.x_0)}$, $\hat{\beta}_k$, $\underline{\hat{\beta}}_k$) obtained in the training sample tend to shrink towards 0 in the test sample, and the shrinkage can be substantial if the number of candidate predictor variables is large.

Exploratory model selection is diametrically opposed to the confirmatory methods described in all of Volume 1 and sections 2.1 - 2.21 of this chapter. In a confirmatory analysis, only theoretically motivated predictor variables are measured and all of these predictor variables are included in the statistical model. Of course, the researcher hopes that the confidence intervals for all of the predictor variable slopes will exclude 0 and will have meaningfully large values, but some of the confidence intervals may include 0. In a confirmatory analysis, the results for *all* predictor variables are reported.

In a confirmatory analysis, it is tempting to remove any predictor variable with a large p -value and report the results only for a reduced model that contains only those predictor variables with p -values less than .05. Predictor variables with large p -values could be removed one at a time. With correlated predictor variables, removing one predictor variable with a large p -value can reduce the p -value for one or more other predictor variables. Although this "semi-exploratory" approach is not ideal, the p -values and confidence intervals for the retained predictor variables may not be too misleading if only a few predictor variables have been removed from the model.

A semi-exploratory analysis could also be used in nonexperimental designs where interaction effects are added one at a time, and only those interactions with p -values less than .05 are retained. The p -values and confidence intervals in the final model may not be too misleading if only a few theoretically justified interaction effects are examined. If one or more variables are removed from or added to the planned model, it is important to disclose this information when reporting the results. Many (most?) published studies that are described as confirmatory are actually semi-exploratory.

The p -values and confidence intervals in a semi-exploratory analysis can be improved by adjusting the standard error for each slope estimate. The adjusted standard error uses the MS_E from the full model that includes all of the planned predictor variables. The adjusted standard error for each slope coefficient in the reduced model is obtained by multiplying each standard error by $\sqrt{MS_{E:F}/MS_{E:R}}$ where $MS_{E:F}$ is the MS_E for the full model and $MS_{E:R}$ is the MS_E for the reduced model. The t -value, p -value, and confidence interval for each slope can then be recomputed using the adjusted standard error with $df = n - s - 1$ where s is the number of planned predictor variables. The `adj.se` function in the `statpsych` package will compute the adjusted standard error, t -value, p -value, and confidence interval for each slope coefficient.

2.26 Influential Observations

A participant with an unusually large residual ($\hat{e}_i = y_i - \hat{y}_i$) may excessively influence the OLS estimates of one or more β_k values. However, the residuals can be misleading because the OLS estimates of β_k have the property of minimizing the sum of the squared residuals, and this tends to reduce the size of the residual for an outlier participant. A better approach is to compute $\hat{y}_i$ for participant i using only the data from the other $n - 1$ participants and then subtract this predicted y score from y_i . These are called *deleted residuals*. The deleted residuals can be made easier to interpret by dividing them by their standard errors. Deleted residuals divided by their standard errors are called *studentized deleted residuals* and are options in SPSS and R.

A statistical test to detect participants with unusually large residuals is based on the fact that studentized deleted residuals follow a t -distribution if the assumptions of the GLM are satisfied. If the absolute value of a studentized deleted residual is greater than the critical t -value for $n - s - 2$ degrees of freedom and a Bonferroni adjusted alpha level of α/n , then that participant is declared to be an outlier, and the scores for that participant should be examined for possible data entry errors or other problems.

A participant can have excessive influence on a particular $\hat{\beta}_k$ value even though that participant does not have a large studentized deleted residual. The influence of participant i on $\hat{\beta}_k$ can be assessed by comparing the value of $\hat{\beta}_k$ when participant i has been included in the analysis with the value of $\hat{\beta}_k$ when participant i has been omitted from the analysis and then dividing the difference of these two estimates by the standard error (computed with participant i omitted) of $\hat{\beta}_k$. This measure of influence for each β_k

is called DFBETAS. Participants with an unusually large DFBETAS value on any predictor variable should be examined for possible data entry errors or other problems. If a model contains a product of two quantitative predictor variables to assess an interaction effect, it is important to closely examine the DFBETAS value for the slope of the product variable because an observed interaction might be caused by just one or two participants who have an unusually low or high score on the product variable.

Cook's D (also referred to as *Cook's distance*) for participant i describes the influence of participant i on all of the parameter estimates. Participants with unusually large Cook's D values should be examined for possible data entry errors or other problems. A useful strategy to identify influential observations is to first examine the studentized deleted residuals and Cook's D for each participant. If a participant has an unusually large Cook's D , then follow up with an examination of the DFBETAS for each predictor variable.

2.27 Multicollinearity

Multicollinearity is the extent to which the s predictor variables in a GLM are correlated. As can be seen in Equation 2.11, high multicollinearity (i.e., a large value of $\hat{\rho}_{x_k, \mathbf{x}}^2$) increases the standard error of $\hat{\beta}_k$ which reduces the power of the test and increases the width of confidence interval for each β_k . Multicollinearity is often described in terms of a *variance inflation factor* for x_k which is defined as $VIF_k = 1/(1 - \hat{\rho}_{x_k, \mathbf{x}}^2)$. The value of VIF_k describes how much the squared standard error for $\hat{\beta}_k$ is inflated due to multicollinearity. Note that $\hat{\rho}_{x_k, \mathbf{x}}^2 = 1 - \frac{1}{VIF_k}$. SPSS also reports a *tolerance* statistic which is equal to $1 - \hat{\rho}_{x_k, \mathbf{x}}^2$.

High multicollinearity also reduces the generalizability of confidence interval and hypothesis testing results for β_k . Consider a GLM where x_1 and x_2 are highly correlated. In the scatterplot of x_1 with x_2 (see Figure 1.3), the sample points would be concentrated in a very small region of the graph. In this situation, the confidence interval and hypothesis testing results for β_1 and β_2 would apply only to this small region of x_1 and x_2 combination values. Alternatively, suppose the correlation between x_1 and x_2 is small (low multicollinearity). The scatterplot of x_1 with x_2 would have sample points that are spread out over a large region of the graph, and the conclusions regarding β_1 and β_2 would apply to this large region of x_1 and x_2 combination values.

Multicollinearity also affects the interpretability of the β_k values. Recall that β_k describes the relation between y and e_{x_k} where e_{x_k} represents the part of x_k that is not linearly

related to all other predictor variables. An interpretation of β_k requires an understanding of the psychological meaning of e_{x_k} . With large $\hat{\rho}_{x_k, \mathbf{x}}^2$ values, e_{x_k} could be difficult to interpret, and hence β_k would be difficult to interpret. However, if VIF_k is close to 1 (or if $\hat{\rho}_{x_k, \mathbf{x}}^2$ is close to 0), e_{x_k} and x_k are then essentially the same and β_k will be easy to interpret. Multicollinearity does not affect the interpretation of $\rho_{y, \mathbf{x}}^2$, $\mu_{y|x^*}$, or predicted y scores.

2.28 Sample Size Requirements for Desired Precision

Coefficient of Determination. The required sample size to estimate a population squared multiple correlation with desired confidence and precision is approximately

$$n = 16\tilde{\rho}_{y, \mathbf{x}}^2(1 - \tilde{\rho}_{y, \mathbf{x}}^2)^2(z_{\alpha/2}/w)^2 + s + 2 \quad (2.27)$$

where $\tilde{\rho}_{y, \mathbf{x}}^2$ is a planning value for the population squared multiple correlation, w is the desired width of the confidence interval, and s is the number of predictor variables. Equation 2.27 is maximized with $\tilde{\rho}_{y, \mathbf{x}}^2 = 1/3$, and this planning value could be used to give a conservatively large sample size requirement. The `size.ci.rsqr` function in the `statpsych` package will compute Equation 2.27 as a first-step approximation and then apply an adjustment for greater accuracy. Equation 2.27 can also be used as a first-step approximation for η^2 although it will overstate the sample size requirement. The `size.ci.etasqr` function in the `statpsych` package will compute Equation 2.27 as a first-step approximation and then applies an adjustment for greater accuracy.

Example 2.16. A researcher wants to estimate the squared multiple correlation between a measure of gender ideology and five predictor variables in a study population of 3,500 college freshmen. The researcher believes the squared multiple correlation will be about .3 and would like the 95% confidence interval to have a width of about .2. The approximate sample size requirement is $n = 16(.3)(.7^2)(1.96/.2)^2 + 5 + 2 = 232.9 \approx 233$. The `size.ci.rsqr` function gives a more accurate sample size requirement of 227.

Partial correlation. The required sample size to estimate a partial correlation with desired confidence and precision is approximately

$$n = 4(1 - \tilde{\rho}_{yx_1, \mathbf{x}_0}^2)^2(z_{\alpha/2}/w)^2 + s_2 + 3 \quad (2.28)$$

where $\tilde{\rho}_{yx_1, \mathbf{x}_0}^2$ is a planning value for the population squared partial correlation between y and x_1 controlling for s_2 control variables, and w is the desired width of the confidence

interval. The `size.ci.cor` function in the `statpsych` package will compute Equation 2.28 as a first-step approximation and then applies an adjustment for greater accuracy.

ANCOVA. Recall from Chapter 3 of Volume 1 that the main effects, interaction effects, and simple effects in an ANOVA model with m groups can all be expressed as a linear contrast of means $\sum_{j=1}^m v_j \mu_j$. In an ANCOVA model, these effects are defined in terms of the y -intercepts. The required sample size per group to estimate a linear contrast of m population y -intercepts with desired confidence and precision is approximately

$$n_j = 4\tilde{\sigma}^2(1 - \tilde{\rho}_{y,x}^2)(1 + \frac{\tilde{\delta}_{cov}^2}{4})(\sum_{j=1}^m v_j^2)(z_{\alpha/2}/w)^2 + \frac{z_{\alpha/2}^2}{2m^*} + s_2 \quad (2.29)$$

where s_2 is the number of covariates, $\tilde{\sigma}^2$ is a planning value of the average within-group variance, $\tilde{\rho}_{y,x}^2$ is a planning value of the within-group squared multiple correlation between the response variable and the covariates, $\tilde{\delta}_{cov}$ is a planning value of the largest pairwise standardized covariate mean difference among all covariates, and m^* is the number of nonzero v_j values. In an experiment where the population covariate means are assumed to be equal across treatment conditions, $\tilde{\delta}_{cov}^2$ would be set to 0. Note that a larger value of $\tilde{\delta}_{cov}^2$ requires a larger sample size. The `size.ci.lc.ancova` function in the `statpsych` package will compute Equation 2.29. The `size.ci.ancova2` function in the `statpsych` package will compute Equation 2.29 for the special case of two groups and can solve for unequal sample sizes.

Equation 2.29 is general and may be used for single-factor designs with m levels or multi-factor designs with a total of m treatment combinations. The MS_E in an ANCOVA is an estimate of $\sigma^2(1 - \rho_{y,x}^2)$. The MS_E from a prior ANCOVA that used the same covariates and response variable as the planned study can be used in place of $\tilde{\sigma}^2(1 - \tilde{\rho}_{y,x}^2)$ in Equation 2.29.

Example 2.17. A researcher wants to estimate $(\mu_{11} + \mu_{12})/2 - (\mu_{21} + \mu_{22})/2$ in a 2×2 factorial experiment with 95% confidence and a desired confidence interval width of 3.0. One covariate will be used. After reviewing the literature, $\tilde{\sigma}^2$ was set at 8.0 and $\tilde{\rho}_{y,x}^2$ was set at .2 ($MS_E = 6.4$). The contrast coefficients are 1/2, 1/2, -1/2, and -1/2. The approximate sample size requirement per group is $n_j = 4(8.0)(1 - .2)(1)(1/4 + 1/4 + 1/4 + 1/4)(1.96/3.0)^2 + 1.96^2/8 + 1 = 12.4 \approx 13$.

Semipartial correlation. A sample size formula for a semipartial correlation is not practical because it requires a planning value for both $\rho_{y,x}^2$ and $\rho_{y(x_1, x_0)}$. If sampling can

be obtained in two stages, the confidence interval width for $\rho_{y(x_1, x_0)}$ in a first-stage sample can be used in Equation 1.32 to determine the number of additional participants to sample in the second stage to achieve the desired confidence interval width.

Slope coefficient. A sample size formula for estimating β_k in a general linear model is not useful because it requires a planning value of $\sigma_{x_k}^2$ and the squared multiple correlation between the predictor variable of interest and all other predictor variables. These planning values are usually difficult to specify. However, if β_k has been previously estimated in a model with the same response variable and predictor variables that will be used in the planned study, the standard error of $\hat{\beta}_k$ in the prior study can be used to approximate the sample size required for desired confidence and precision in the planned study as shown below

$$n = 4n_0 SE_{\hat{\beta}_k}^2 (z_{\alpha/2}/w)^2 \quad (2.30)$$

where n_0 is the sample size in the prior study. The `size.ci.gen` function in the `statpsych` package will compute Equation 2.30. Alternatively, if sampling can be obtained in two stages, the confidence interval width for β_k in a first-stage sample can be used in Equation 1.32 to determine the number of additional participants to sample in the second stage to achieve the desired confidence interval width.

Mean Absolute Prediction Error. The `size.ci.mape` function in the `statpsych` package computes the sample size required to obtain a confidence interval for a population mean absolute prediction error with desired precision. This function is especially useful in studies where the primary goal is to assess the predictive accuracy of a statistical model. This function requires a planning value for the population MAPE. The MS_E from a previous study can be used to specify a MAPE planning value using the approximation $MAPE \approx .8\sqrt{MS_E}$. This approximation is very accurate if the prediction errors have an approximate normal distribution.

2.29 Sample Size Requirements for Desired Power

Partial correlation. The required sample size to test $H_0: \rho_{yx_1, x_0} = h$ for a given α value and with desired power is approximately

$$n = (z_{\alpha/2} + z_\beta)^2 / (\tilde{\rho}_{yx_1, x_0}^* - h^*)^2 + s_2 + 3 \quad (2.31)$$

where $\tilde{\rho}_{yx_1.x_0}^*$ is a Fisher transformation of a planning value for $\rho_{yx_1.x_0}$, h^* is a Fisher transformation of h , and s_2 is the number of control variables. The `size.test.cor` function in the `statpsych` package will compute Equation 2.31 as a first-step approximation and then applies an adjustment for improved accuracy.

Example 2.18. A researcher wants to assess the correlation between the amount of social media use and self confidence in a study population of about 5,000 high school male students. GPA and the number of close friends will be used as two control variables. After reviewing the literature, the researcher decided to use .4 as the planning value of the partial correlation. The researcher wants to test the null hypothesis of a zero partial correlation and wants the test to have power of .9. The approximate sample size requirement is $n = (1.96 + 1.28)^2 / (.424 - 0)^2 + 2 + 3 = 63.4 \approx 64$.

ANCOVA. The required sample size per group to test a linear contrast of population y -intercepts in an ANCOVA model with m groups and s_2 covariates for a specified value of α and with desired power is approximately

$$n_j = \tilde{\sigma}^2(1 - \tilde{\rho}_{y.x}^2)(1 + \frac{\tilde{\delta}_{cov}^2}{4})(\sum_{j=1}^m v_j^2)(z_{\alpha/2} + z_\beta)^2 / (\sum_{j=1}^m v_j \tilde{\beta}_{0j})^2 + \frac{z_{\alpha/2}^2}{2m^*} + s_2 \quad (2.32)$$

where $\sum_{j=1}^m v_j \tilde{\beta}_{0j}$ is the planning value of the effect size and the other terms are defined in Equation 2.29. In an experimental design where the population covariate means are assumed to be equal across the treatment conditions, $\tilde{\delta}_{cov}$ is set to 0 and the effect size will be the same for an ANCOVA model and an ANOVA model. A larger value of $\tilde{\delta}_{cov}^2$ requires a larger sample size. In nonexperimental designs, $\sum_{j=1}^m v_j \tilde{\beta}_{0j}$ may be difficult to specify because the effect size depends on how the covariates correlate with the indicator variables. In nonexperimental design applications of the ANCOVA, the value of $\sum_{j=1}^m v_j \tilde{\beta}_{0j}$ could be set to smallest value of theoretical or practical importance. The `size.test.lc.ancova` function in the `statpsych` package will compute Equation 2.31. The `size.test.ancova2` function in the `statpsych` package will compute Equation 2.31 for the special case of two groups and can solve for unequal sample sizes.

Example 2.19. A researcher wants to test $H_0: \frac{\mu_1 + \mu_2 + \mu_3 + \mu_4}{4} - \mu_5$ in a one-factor experiment with power of .90, $\alpha = .05$, an anticipated value for $\frac{\mu_1 + \mu_2 + \mu_3 + \mu_4}{4} - \mu_5$ of 4.0, two covariates, a planning value of 25 for the average within-group variance, and a planning value of the squared multiple correlation between the covariates and the response variable of .3 ($MS_E = 17.5$). The contrast coefficients are 1/4, 1/4, 1/4, 1/4, and -1. The approximate sample size requirement per group is $n_j = 25(1 - .3)(1)(1/16 + 1/16 + 1/16 + 1/16 + 1)(1.96 + 1.28)^2 / 4.0^2 + 1.96^2 / 10 + 2 = 16.7 \approx 17$.

Slope coefficient. A sample size formula for testing $H_0: \beta_k = 0$ in a general linear model is not useful because it requires a planning value of $\sigma_{x_k}^2$ and a planning value for the squared multiple correlation between the predictor variable of interest and all other predictor variables. These planning values are usually difficult to specify. However, if β_k has been previously estimated in a model with the same response variable and predictor variables that will be used in the planned study, the standard error of $\hat{\beta}_k$ in the prior study can be used to approximate the sample size required for desired power in the planned study as shown below

$$n = n_0 SE_{\hat{\beta}_k}^2 (z_{\alpha/2} + z_\beta)^2 / \tilde{\beta}_k^2 \quad (2.33)$$

where n_0 is the sample size in the prior study and $\tilde{\beta}_k$ is a planning value of β_k . The `size.test.gen` function in the `statpsych` package will compute Equation 2.33. The planning value of β_k could be specified based on the value of $\hat{\beta}_k$ in the prior study. But if n_0 is small, it is possible that $\hat{\beta}_k$ has substantially overestimated the value of β_k . An alternative approach is to set the planning value of β_k to the smallest value that would have practical or scientific importance.

2.30 Assumptions

In fixed-x models, tests and confidence intervals for slopes and conditional slopes, and confidence intervals for η^2 and τ_e assume: 1) random sampling, 2) independence among participants, 3) linearity between y and every predictor variable (linearity assumption) in models without quadratic effects, 4) constant variability of the prediction errors across the values of every predictor variable (prediction error homoscedasticity assumption), and 5) approximate normality of the prediction errors in the study population (prediction error normality assumption). The effects of violating these assumptions are summarized in section 1.26. The linearity assumption does not apply in quadratic models or models with interaction effects.

Scatterplots of y with each predictor variable are useful in assessing the linearity assumption. Scatterplots of the studentized deleted residuals with each predictor variable (called *residual plots*) are helpful in assessing the prediction error homoscedasticity assumption. A single scatterplot of the studentized deleted residuals with the predicted y scores ($\hat{y}$) can reveal violations of the linearity and prediction error homoscedasticity

assumptions. Skewness and kurtosis estimates of the residuals are useful in assessing the prediction error normality assumption. Transforming the response variable or predictor variables may reduce prediction error nonnormality or nonlinearity.

In random- $\mathbf{x}$ models, tests and confidence intervals for slopes, conditional slopes, and τ_e requires the same assumptions given above for the fixed- $\mathbf{x}$ models. Confidence intervals for a squared multiple correlation, partial correlation, or semipartial correlation in random- $\mathbf{x}$ models require an additional assumption of *multivariate normality*. However, these confidence intervals should perform properly under a less restrictive assumption of bivariate normality. Specifically, the confidence interval for the squared multiple correlation requires y and $\hat{y}$ to have an approximate bivariate normal distribution, the confidence interval for a semipartial correlation requires y and e_{x_k} to have an approximate bivariate normal distribution, and the confidence interval for a partial correlation requires e_y and e_{x_k} to have an approximate bivariate normal distribution. To assess these bivariate normality assumptions, first compute the predicted y scores ($\hat{y}$) and estimated prediction errors ($\hat{e}_y$ and $\hat{e}_{x_k}$). Next, check for a linear relation between the appropriate pair of variables (i.e., y and $\hat{y}$, y and $\hat{e}_{x_j}$, $\hat{e}_y$ and $\hat{e}_{x_k}$) and check for skewness and kurtosis of the appropriate pair of variables. Transforming the response variable or predictor variables may reduce nonnormality.

2.31 Distribution-free Methods

OLS estimates of $\beta_0, \beta_1, \dots, \beta_s$ in a GLM minimize the sum of the squared residuals ($\sum_{i=1}^n \hat{e}_i^2$). An alternative to OLS estimation is *least absolute deviation* (LAD) estimation. LAD estimates of $\beta_0, \beta_1, \dots, \beta_s$ minimize the sum of absolute residuals ($\sum_{i=1}^n |\hat{e}_i|$). Unlike OLS estimation, LAD estimation is computationally intensive and the LAD estimates of $\beta_0, \beta_1, \dots, \beta_s$ cannot be expressed as simple formulas.

In a fixed- $\mathbf{x}$ model, the predicted y score for a particular set of $x_1, x_2, \dots, x_s$ values using OLS estimation is an estimate of the mean of y at those values of $x_1, x_2, \dots, x_s$. With LAD estimation in a fixed- $\mathbf{x}$ model, the predicted y score is an estimate of the median of y at those values of $x_1, x_2, \dots, x_s$. Large residuals are much less influential on LAD estimates than OLS estimates, and LAD estimation is attractive in applications where there is no justification for removing participants with unusually large residuals.

LAD estimates can be obtained using *quantile regression* where the "tau" parameter is set to .5. SPSS and the `rq` function in the `quantreg` package can be used to perform quantile regression. Quantile regression does not assume prediction error normality, and quantile regression programs also have an option to relax the homoscedasticity assumption. Thus, the traditional diagnostic methods for OLS estimation that check for nonnormality, heteroscedasticity, and influential observations are not necessary with LAD estimation. Like the GLM, quantile regression is general and can be used for any of the models described in this chapter. However, the currently available LAD-based hypothesis tests and confidence intervals for $\beta_0, \beta_1, \dots, \beta_s$ require larger sample sizes than the OLS methods. Although there is currently no consensus regarding the minimum sample size requirement for LAD-based hypothesis tests and confidence intervals, these methods should perform properly if $n - s > 100$.

2.32 Bootstrap Standard Error

The standard error of $\hat{\rho}_{yx}^*$ assumes that y and x have a bivariate normal distribution. The standard errors of $\hat{\rho}_{y(x_1, x_0)}^*$, $\hat{\rho}_{yx_1, x_0}$, and $\underline{\beta}_k$ assume that y and all s predictor variables have a multivariate normal distribution. These standard errors can be much too small in non-normal distributions regardless of the sample size. In applications where the normality assumption is questionable, the standard errors in Equations 1.14, 2.13 and 2.17 can be replaced with *bootstrap standard errors* to obtain better confidence intervals. A confidence interval for standardized slopes with bootstrap standard errors can be obtained using the `sem` function in the `lavaan` package (see section 1.14 of Volume 4). If the normality assumption is violated because of leptokurtosis, a large sample size (100 or more) is needed to obtain a useful bootstrap standard error.

To compute a bootstrap standard error, the first step is to take multiple *bootstrap samples* from the original sample of size n . A bootstrap sample is a sample of size n taken from the original sample *with replacement*. When sampling with replacement, the same participant can appear more than once in any given bootstrap sample. The number of possible bootstrap samples is enormous for typical sample sizes, and it is customary to randomly sample 500 (or more) of all possible bootstrap samples when computing a bootstrap standard error. For example, suppose the sample size is $n = 9$. The participant IDs (1 to 9) are shown below for the first five randomly selected bootstrap samples. Note

that a participant can appear more than once within each bootstrap sample (e.g., participants 4 and 8 occur twice in the first bootstrap sample).

Bootstrap Sample	Participant IDs
1	1 2 3 4 4 5 6 7 8 8
2	2 3 4 5 6 6 7 8 8 9
3	1 3 4 5 5 5 6 7 8 9
4	1 3 4 5 6 6 7 8 9 9
5	1 2 2 2 3 4 5 6 8 8
⋮	⋮

After obtaining a random sample of 500 or more bootstrap samples, compute $\hat{\rho}_{yx}^*$, $\hat{\rho}_{y(x_1, \mathbf{x}_0)}^*$ or $\hat{\rho}_{yx_1, \mathbf{x}_0}^*$ in each bootstrap sample. The bootstrap standard error is simply the standard deviation of the parameter estimates from the bootstrap samples. The bootstrap standard error then could be used in place of the standard errors in Equations 1.14, 2.13, or 2.17.

2.33 Bootstrap Confidence Interval

The confidence intervals for some parameters (e.g., η^2 and $\rho_{y, x}^2$) do not use standard errors and may not perform properly when their assumptions have been violated. For some parameters (e.g., η^2 for within-subjects designs, squared multiple semipartial correlation, and indirect effects in the path models described in Volume 4), acceptable confidence intervals are not currently available. For some parameters, the *percentile bootstrap confidence interval* could be useful.

Although 500 or more bootstrap samples is usually adequate for a bootstrap standard error, a percentile bootstrap confidence interval requires a much larger number of bootstrap samples. For $\alpha = .05$, at least 10,000 bootstrap samples should be obtained, and smaller values of α will require more bootstrap samples. Furthermore, the percentile bootstrap confidence interval assumes the distribution of parameter estimates from the bootstrap samples is at most moderately skewed, and a large sample size (100 or more) is usually needed to satisfy this requirement.

To compute a percentile bootstrap confidence interval, compute the parameter estimate in each bootstrap sample as explained above for the bootstrap standard error. The lower limit of a $100(1 - \alpha)\%$ percentile bootstrap confidence interval is the $\alpha/2$ percentile of the bootstrap distribution of parameter estimates, and upper limit of a $100(1 - \alpha)\%$ percentile bootstrap confidence interval is the $1 - \alpha/2$ percentile of the bootstrap distribution of parameter estimates.

2.34 Missing Data

All of the hypothesis tests and confidence intervals described in this chapter assume that every participant produces a score for the response variable and all s predictor variables. If a participant is missing a response variable score or a score on any of the s predictor variables, most computer programs will eliminate that participant from the analysis. This approach to missing data is called *listwise deletion*.

Let w_k represent a dummy variable that codes missing vs. non-missing for variable k . If w_k is unrelated to all other variables in the model, and if w_k is also *assumed* to be unrelated to variable k and all other variables that have not been included in the model, then the missing scores for variable k are said to be *missing completely at random* (MCAR). If w_k is related to one or more variables that have been included in the model and is *assumed* to be unrelated to variable k and all variables that have not been included in the model, then the missing scores for variable k are said to be *missing at random* (MAR).

Although it is possible to check for relations between w_k and all other variables in the model, it is not possible to determine if w_k is related to variable k or variables not in the model. If w_k is found to be unrelated to all variables in the model, then the missing data for variable k *might be* MCAR. If w_k is found to be related to one or more other variables in the model, then the missing data for variable k are not MCAR but *might be* MAR.

If the missing responses are MCAR, then the reduced sample after listwise deletion remains a random sample from the original study population, and inferential methods computed from the reduced sample will provide a description of the original study population. One drawback of listwise deletion is that the sample size is reduced and this leads to wider confidence intervals and hypothesis tests with less power. If the missing data are assumed to be MCAR or MAR, then listwise deletion is not required and all available data from all participants can be analyzed using a *full information maximum*

likelihood (FIML) estimation procedure which can be obtained using the `sem` function in the `lavaan` package (see section 1.13 of Volume 4).

2.35 Data Visualization

Dot plots, box plots, and histograms of the response variable and each predictor variable can be used in the data screening process to detect univariate outliers and visually assess the distribution of each variable. Scatterplots of all pairs of variables can be used to detect bivariate outliers and visually assess the linearity assumption. In models with two or more predictor variables, a *matrix scatterplot* is a compact graphical display of scatterplots for all pairs of variables. An example of a matrix scatterplot for the response variable and two predictor variables in Data Analysis Problem 2-2 is shown below.

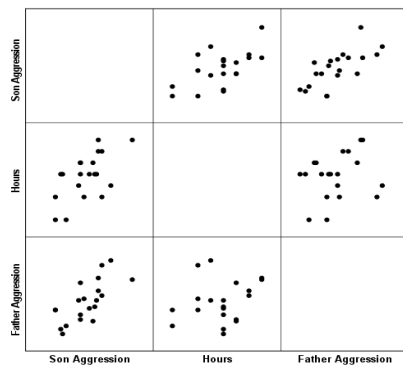


Figure 2.2 Matrix scatterplot

The multiple correlation in section 2.4 is defined as the Pearson correlation between y and $\hat{y}$ where $\hat{y}$ is the y score that is predicted from the values of the $x_1, x_2, \dots, x_s$ predictor variables. A scatterplot for y and $\hat{y}$ visually describes the magnitude of the multiple correlation and can reveal possible problems with nonlinearity and bivariate outliers. An example of a scatter plot for y and $\hat{y}$ is given below for Data Analysis Problem 2.2.

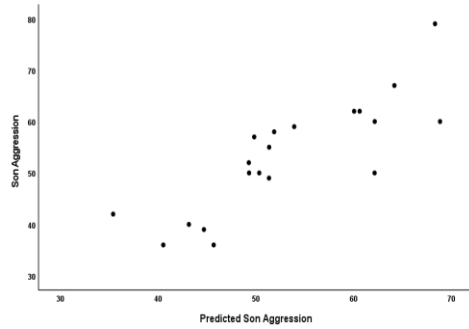


Figure 2.3 Scatterplot of y and $\hat{y}$

As explained in section 2.2, the slope coefficient for predictor variable x_j describes the relation between y and x_j "controlling for" all other predictor variables in the model. Specifically, the estimated slope coefficient for x_j describes the relation between y and $\hat{e}_{x_j}$ where $\hat{e}_{x_j}$ is the component of x_j that is not linearly predictable from the other predictor variables in the model. Also, the estimated semipartial correlation between y and x_j (see section 2.7) is equal to the Pearson correlation between y and $\hat{e}_{x_j}$. To obtain $\hat{e}_{x_j}$, use a linear model program to predict x_j from all other predictor variables in the model and then compute $\hat{e}_{x_j} = x_j - \hat{x}_j$. A scatterplot of y and $\hat{e}_{x_j}$, referred to here as a *semipartial correlation plot*, is useful in visually assessing the magnitude of a semipartial correlation or slope coefficient and can also reveal outliers or nonlinearity. An example of a semipartial correlation is shown below for Data Analysis Problem 2-2 where the vertical axis is son's aggression and the horizontal axis is hours of violent video game play, controlling for father's aggression. In this study, we would also want to examine a second semipartial correlation plot where the horizontal axis is father's aggression, controlling for hours of video game play.

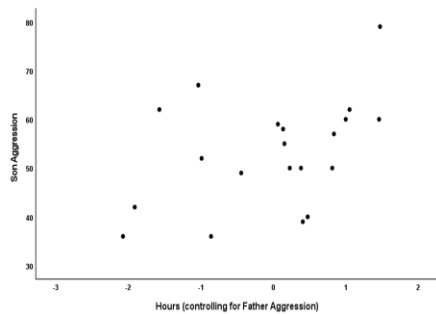


Figure 2.4 Example of semipartial scatterplot

The estimated partial correlation between y and x_j , controlling for one or more other variables is equal to the Pearson correlation between $\hat{e}_y$ and $\hat{e}_{x_j}$ (see section 2.8). A scatterplot of $\hat{e}_y$ and $\hat{e}_{x_j}$, referred to as a *partial correlation plot*, is useful in visually assessing the magnitude

of a partial correlation and can also reveal outliers or nonlinearity. An example of a partial correlation plot is shown below for Data Analysis Problem 2-2 where the vertical axis is son's aggression, controlling for father's aggression, and the horizontal axis is hours of violent video game play, controlling for father's aggression. We would also want to examine a second partial correlation plot where the horizontal axis is father's aggression, controlling for hours of video game play.

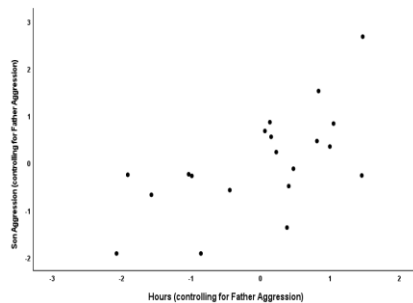


Figure 2.5 Example of partial scatterplot

A scatterplot of studentized deleted residuals with each predictor variable can be used to visually assess the homoscedasticity assumption. These scatterplots can also reveal outliers and nonlinearity. Ideally, there will be no obvious increase or decrease in the spread of the residuals across the values of a predictor variable. Residual plots for the two predictor variables in Data Analysis Problem 2-2 are shown below and do not show any obvious assumption violations.

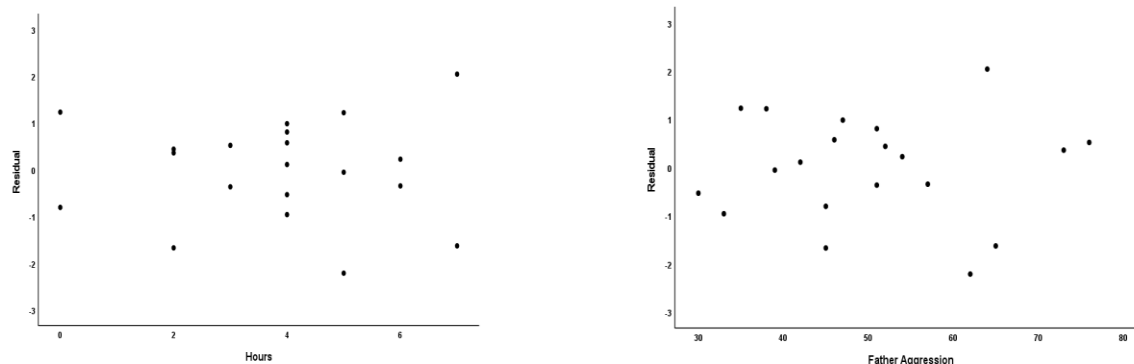


Figure 2.6 Residual scatterplots for two predictor variables

A scatterplot of studentized deleted residuals with the predicted y scores is another way to assess the homoscedasticity assumption and check for outliers. This type of scatterplot is shown below for Data Analysis Problem 2-2.

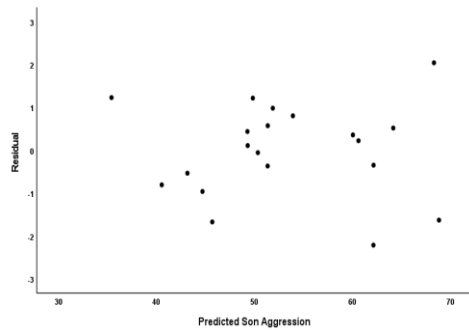


Figure 2.6 Scatterplot of studentized residuals with the predicted response variable

Unless the sample size is small, a histogram of the studentized deleted residuals can provide information about the shape of the prediction error distribution. A dot plot of the residuals can help identify outliers. A dot plot of the studentized deleted residuals for Data Analysis Problem 2-2 is shown below. This plot suggests that there is one participant with a residual less than -2 and one participant with a residual greater than 2. Although studentized deleted residual values greater than 2 in absolute value are expected in about 5% of the participants, it would be prudent to examine these two participants for possible data entry errors or other problems.

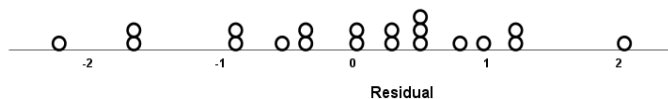


Figure 2.7 Dot plot of studentized residuals

A dot plot of Cook's D values can help identify participants who are having the greatest influence on the estimated y -intercept and slopes. A dot plot of Cook's D values for Data Analysis Problem 2-2 is shown below. Cook's D values less than $4/n$ are usually not a concern. In this example where $4/20 = .2$, it would be prudent to examine the data for the participants with the two largest Cook's D values and check for possible data entry errors or other problems.

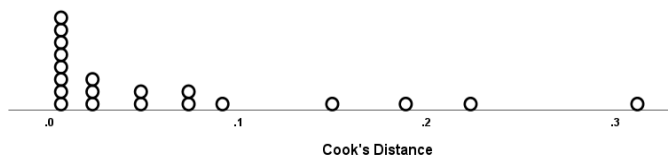


Figure 2.8 Dot plot of Cook's D values

In an ANCOVA model, a multiple-group scatterplot for each covariate can be used to visually assess the strength and form of the relation between each covariate and the response variable. Different colors or symbols can be used to distinguish the paired observations from each group. It is helpful to include a line of best fit within each group to visually assess differences in slopes. A two-group scatterplot is shown below for Data Analysis Problem 2-4 where GPA is the only covariate.

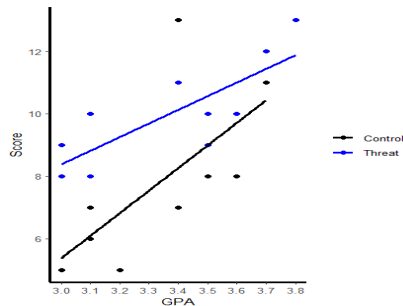


Figure 2.9 Example of a two-group scatterplot with regression lines

A note of caution is warranted here. Although data visualization methods are useful for gaining important insights about relations among the model variables, there is a danger that these visualization methods can lead to a variety of unintended exploratory model modifications that can produce misleading results (e.g., p -values that are too small and confidence intervals that are too narrow). The statistical models in this volume are especially susceptible to the negative effects of exploratory model modifications, which are the main contributors to the replication crisis in psychology. Researchers who use any of the methods in this volume are urged to preregister their planned study (see section 1.19 of Volume 1). The *Open Science Framework* provides a platform for researchers to publicly register a detailed description of their study and all of the planned analyses prior to data collection.

2.36 Bayesian Approach to Data Analysis

Specifying an informative prior distribution for β_k in a GLM is extremely difficult because the value of β_k depends on the (usually unknown) correlations among the predictor variables. Except in very special cases, an informative prior distribution for β_k would not be taken seriously by other researchers. Most Bayesian analyses of a GLM use a non-informative prior for each β_k . A non-informative prior for β_k would typically be a normal distribution with a mean of 0 and a very large standard deviation so that the estimate of β_k will be determined only by the sample data. With a non-informative prior, a Bayesian

credible interval for β_k is the same as a classical confidence interval for β_k (Equation 2.9). Most Bayesian statisticians will simply use classical multiple regression software (e.g., the `lm` function in R or the Linear Regression procedure in SPSS) to compute confidence intervals for each β_k and then report the confidence interval results as credible intervals.

A skeptical prior is one type of informative prior that is not difficult to implement in a GLM. If the researcher can specify a value for β_k that represents a weak association, the `ci.bayes.normal` function can be used to compute a credible interval for β_k . Suppose a slope coefficient for x_k that is less than 10 in absolute value is believed to represent a weak association. If a skeptic is 95% confident that β_k is between -10 and 10, then the prior mean would be set to 0, and the prior standard deviation would be set to 10/1.96. The `ci.bayes.normal` requires an estimate of β_k and its standard error, which can be obtained from any classical GLM software package. This function also requires the prior mean (e.g., 0) and prior standard deviation (e.g., 10/1.96) as input. A skeptical prior might more easily be applied to a standardized slope because it is easier to specify a "small" value of the standardized slope. For example, a standardized slope coefficient for x_k that is less than .25 in absolute value might be viewed as a weak association. Then the prior mean would be set to 0, and the prior standard deviation would be set to .25/1.96. The `ci.bayes.normal` requires an estimate of the standardized slope and its standard error. The standard error of a standardized slope can be obtained from the `lavaan` R package.

The `ci.bayes.normal` function is very general and can accept different types of β_k estimates, such as OLS (Equation 2.32) or LAD (see section 2.32) estimates. Alternative standard errors can also be used, such as the heteroscedastic-robust standard error for an OLS estimate (Equation 2.42) or the nonnormality-robust standard error for a standardized slope (an option in the `lavaan` package).

A skeptical prior can also be applied to a semipartial correlation. It is usually easy to specify a "small" value of a semipartial correlation for the k^{th} predictor variable. For example, if a skeptic is 95% confident that a semipartial correlation is between -.2 and .2, the mean of the prior distribution would be set at 0 and the standard deviation would be set at .2/1.96. The `ci.bayes.spcor` function can be used to estimate a semipartial correlation with a skeptical prior. This function assumes the mean of the prior distribution is 0, and the user is required to specify the prior standard deviation (e.g., .2/1.96 = .102). This function also requires an estimate of the semipartial correlation and

its standard error, which can be obtained from the `ci.spcor` function in the `statpsych` package.

Example 2.16. In *Example 2.2*, neuroticism, conscientiousness, and industriousness are used to predict life satisfaction. The estimated semipartial correlation for neuroticism is $-.355$ with a standard error of $.054$. The researcher's skeptical prior opinion for the population semipartial correlation is characterized by a normal distribution with a mean of 0 and a standard deviation of $.102$ (which implies a $.95$ probability that the population semipartial correlation is between $-.2$ and $.2$). The mean of the posterior distribution is $-.282$, and the 95% credible interval is $[-.366, -.194]$. The estimated standardized slope is $-.424$ with a standard error of $.066$. The researcher's skeptical prior opinion for the population standardized slope is characterized by a normal distribution with a mean of 0 and a standard deviation of $.127$ (which implies a $.95$ probability that the population standardized slope is between $-.25$ and $.25$). Even with skeptical priors, the 95% credible intervals suggests that there is a small to moderate association between life satisfaction and the part of neuroticism that is not linearly related to conscientiousness and industriousness. This analysis would be repeated for the other two predictor variables.

Skeptical priors should not be used in applications where the primary goal of the study is to show that the k^{th} predictor variable, controlling for the other predictor variable, has a small or unimportant effect on the response variable. For example, a skeptical prior should not be used if the purpose of a study is to show similarity in mean resilience for men and women after controlling for specific demographic variables.

2.37 Using Information from a Prior Study

Classical statistical methods can be used to combine and compare information from a prior study with information in a new study. The prior information is in the form of a parameter estimate and its standard error. Combining the parameter estimate from a prior study with a new study can produce a substantially narrower (and hence more informative) confidence interval along with greater generalizability.

In studies that use multiple predictor variables, the parameter of primary interest might be a partial correlation or a slope coefficient. A confidence interval for the difference in parameters for the prior and new study populations can be used to assess equivalence (see section 2.16) and replicability (see section 4.2 of Volume 5). Tests of equivalence and replicability typically use a $1 - 2\alpha$ rather than a $1 - \alpha$ level of confidence.

If an estimate of a partial correlation has been reported in a prior study, the `replicate.cor` function in the `vcmeta` package will compute confidence intervals for $\rho_{yx_k.x_0[1]}$, $\rho_{yx_k.x_0[2]}$, $\rho_{yx_k.x_0[1]} - \rho_{yx_k.x_0[2]}$, and $(\rho_{yx_k.x_0[1]} + \rho_{yx_k.x_0[2]})/2$ where $\rho_{yx_k.x_0[j]}$ is the population partial correlation between y and x_k in study j . The new study must use the exact same response variable and exact same set of predictor variables that were used in the prior study; otherwise, the partial correlations between y and x_k in the two study populations will not be comparable.

Example 2.15. A study reported a partial correlation of .385 between the grade in a research methods course (y) and research self-efficacy (x_1), controlling for GPA and general self-confidence, in a sample of 80 college students with a 95% confidence interval of [.175, .558]. The study was replicated at another university using a sample of 150 students. The estimated partial correlation in the follow-up study was .369. The 90% confidence interval for $\rho_{yx_k.x_0[1]} - \rho_{yx_k.x_0[2]}$ is [-.189, .210], which includes 0 and justifies combining the partial correlations from the prior and follow-up studies. The estimate of $(\rho_{yx_k.x_0[1]} + \rho_{yx_k.x_0[2]})/2$ is .377 with a 95% confidence interval of [.262, .482]. We can be 95% confident that the average partial correlation between grade in a research methods course and research self-efficacy, controlling for GPA and general self-confidence, in the two university populations is between .262 and .482.

If a slope coefficient has been reported in a prior study, the estimate can be compared and combined with an estimate of the same slope coefficient in a new study. The new study must use the exact same response variable and exact same set of predictor variables that were used in the prior study; otherwise, the slope coefficients for a given predictor variable in the two study populations will not be comparable. The `replicate.slope` function in the `vcmeta` package will compute confidence intervals for $\beta_{k[1]}$, $\beta_{k[2]}$, $\beta_{k[1]} - \beta_{k[2]}$, and $(\beta_{k[1]} + \beta_{k[2]})/2$ where $\beta_{k[j]}$ is the population slope coefficient for predictor variable k in study j .

Example 2.16. In *Example 2.2*, neuroticism, conscientiousness, and industriousness were used to predict life satisfaction. The estimated slope for neuroticism was -.553 with a standard error of .088. The study was replicated at another organization using a sample of 400 employees. The estimated slope for neuroticism was -.451 with a standard error of .067. The 90% confidence interval for $\beta_{1[1]} - \beta_{1[2]}$ is [-.284, .080], which includes 0 and justifies combining the slopes from the prior and follow-up studies. The estimate of $(\beta_{1[1]} + \beta_{1[2]})/2$ is -.502 with a 95% confidence interval of [-.611, -.393]. This confidence interval is narrower than the confidence interval reported in the prior study, and this confidence interval describes both study populations. This analysis would be repeated for the conscientiousness and industriousness slopes.

The command `replicate.cor(.05, .385, 80, .369, 150, 2)` will produce the results in Example 2.15, and the command `replicate.slope(.05, -.553, .088, 236, -.451, .067, 400, 3)` will produce the results in Example 2.16. The 2 in `replicate.cor` is the number of control variables, and the 3 in `replicate.slope` is the number of predictor variables.

2.38 GLM in Matrix Form

Appendix B provides a review of matrix notation and matrix operations. A GLM model for a random sample of n participants can be expressed using matrix notation as follows

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \quad (\text{Model 2.16})$$

where $\mathbf{y}$ is an $n \times 1$ vector of response variable scores, $\mathbf{e}$ is an $n \times 1$ vector of prediction errors, $\boldsymbol{\beta}$ is a $(s + 1) \times 1$ vector of unknown population parameters, and $\mathbf{X}$ is an $n \times (s + 1)$ *design matrix*. The first column of $\mathbf{X}$ is usually an $n \times 1$ vector of ones to code the y -intercept and the other s columns contain the values of s predictor variables.

To illustrate the matrix structure of Model 2.26, consider a model with $s = 1$ predictor variable and a random sample of $n = 7$. The elements of $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ for this example are shown below.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{bmatrix} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{12} \\ 1 & x_{13} \\ 1 & x_{14} \\ 1 & x_{15} \\ 1 & x_{16} \\ 1 & x_{17} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \end{bmatrix}$$

The OLS estimate of $\boldsymbol{\beta}$ for Model 2.16 is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \quad (2.32)$$

the estimated prediction errors (residuals) are

$$\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}, \quad (2.33)$$

and an estimate of the residual variance (σ_e^2) is

$$MS_E = \hat{\mathbf{e}}'\hat{\mathbf{e}}/(n - s - 1). \quad (2.34)$$

The k^{th} element of $\hat{\boldsymbol{\beta}}$ is $\hat{\beta}_k$. The standard error of $\hat{\beta}_k$ is equal to the square root of the k^{th} diagonal element of the following covariance matrix of $\hat{\boldsymbol{\beta}}$

$$\text{cov}(\hat{\boldsymbol{\beta}}) = MS_E(\mathbf{X}'\mathbf{X})^{-1}. \quad (2.35)$$

A $100(1 - \alpha)\%$ confidence interval for β_k is

$$\hat{\beta}_k \pm t_{\alpha/2; df_E} SE_{\hat{\beta}_k} \quad (2.36)$$

where $SE_{\hat{\beta}_k}$ is the square root of the k^{th} diagonal element of $\text{cov}(\hat{\boldsymbol{\beta}})$ and $df_E = n - s - 1$. SPSS and R will compute Equations 2.32, 2.35 and 2.36.

A linear function of the elements of $\boldsymbol{\beta}$ will be of interest in some applications. Conditional slopes (e.g., $\beta_1 + \beta_3 \dot{x}_2$) or the slope of the line at $x = \dot{x}$ in a quadratic model (i.e., $\beta_1 + 2\beta_2 \dot{x}$) are two examples of linear functions of $\boldsymbol{\beta}$. A linear function of $\boldsymbol{\beta}$ can be expressed as $\mathbf{q}'\boldsymbol{\beta}$ where $\mathbf{q}$ is an $(s + 1) \times 1$ vector of coefficients. For example, the linear function $\beta_1 + 5\beta_3$ can be expressed as $\mathbf{q}'\boldsymbol{\beta}$ where $\mathbf{q}' = [0 \ 1 \ 0 \ 5]$ and $\boldsymbol{\beta}' = [\beta_0 \ \beta_1 \ \beta_2 \ \beta_3]$.

The estimated standard error of $\mathbf{q}'\hat{\boldsymbol{\beta}}$ is

$$SE_{\mathbf{q}'\hat{\boldsymbol{\beta}}} = \sqrt{\mathbf{q}' \text{cov}(\hat{\boldsymbol{\beta}}) \mathbf{q}} \quad (2.37)$$

and a $100(1 - \alpha)\%$ confidence interval for $\mathbf{q}'\boldsymbol{\beta}$ is

$$\mathbf{q}'\hat{\boldsymbol{\beta}} \pm t_{\alpha/2; df_E} SE_{\mathbf{q}'\hat{\boldsymbol{\beta}}} \quad (2.38)$$

where $df_E = n - s - 1$. The `ci.lc.glm` function in the `statpsych` package will compute Equations 2.37 and 2.38.

Assuming the first column of the design matrix is a vector of ones, the predicted y score at $\dot{\mathbf{x}}' = [1 \ \dot{x}_1 \ \dot{x}_2 \ \dots \dot{x}_s]$ is

$$\hat{y} = \dot{\mathbf{x}}'\hat{\boldsymbol{\beta}} \quad (2.39)$$

If one or more of the predictor variables are factors, $\hat{y}$ also may be interpreted as the estimated conditional population mean of y at $\dot{\mathbf{x}}$, which will denoted as $\hat{\mu}_{y|\dot{\mathbf{x}}}$. A $100(1 - \alpha)\%$ confidence interval for $\mu_{y|\dot{\mathbf{x}}}$ is

$$\hat{\mathbf{x}}'\hat{\boldsymbol{\beta}} \pm t_{\alpha/2;df_E} \sqrt{\hat{\mathbf{x}}' cov(\hat{\boldsymbol{\beta}}) \hat{\mathbf{x}}} \quad (2.40)$$

which is a generalization of Equation 1.18. A $100(1 - \alpha)\%$ prediction interval for a predicted y score at $\hat{\mathbf{x}}$ is

$$\hat{\mathbf{x}}'\hat{\boldsymbol{\beta}} \pm t_{\alpha/2;df_E} \sqrt{MS_E + \hat{\mathbf{x}}' cov(\hat{\boldsymbol{\beta}}) \hat{\mathbf{x}}} \quad (2.41)$$

which is a generalization of Equation 1.19. In Equations 2.40 and 2.41, $df_E = n - s - 1$. SPSS and R have options to compute Equations 2.40 and 2.41.

The estimated covariance matrix of $\hat{\boldsymbol{\beta}}$ given in Equation 2.35 assumes the prediction errors have equal variances. In ANOVA and ANCOVA models where the group sample sizes are not equal, the use of Equations 2.36, 2.38, 2.40, and 2.41 can give misleading results if the prediction error variances differ across groups. An alternative estimate of the covariance matrix of $\hat{\boldsymbol{\beta}}$ for ANOVA and ANCOVA models is given below

$$cov(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\hat{\mathbf{V}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad (2.42)$$

where $\hat{\mathbf{V}}$ is a diagonal matrix with estimates of the group prediction error variances along the diagonal. An alternative standard error of $\hat{\beta}_k$ that does not assume equal prediction error variances across groups is the square root of the k^{th} diagonal element in Equation 2.42. This alternative standard error can be used in Equations 2.36 and 2.37. The accuracy of Equations 2.36 and 2.37 can be further improved by replacing the degrees of freedom $(n - s - 1)$ with Satterthwaite degrees of freedom.

In random- $\mathbf{x}$ models where the prediction error variances are not assumed to be constant across the levels of every predictor variable, MacKinnon-White HC3 variance estimates are recommended for the diagonal elements of $\hat{\mathbf{V}}$ of Equation 2.42. SPSS and R have options to compute a confidence interval for β_k using Equation 2.42 with MacKinnon-White variance estimates.

Key Terms

general linear model (GLM)

ANOVA model

multiple linear regression model

ANCOVA model
confounding variable
confounding variable bias
multiple correlation
coefficient of multiple determination
adjusted squared multiple correlation
conditional slope
quadratic model
semipartial correlation
partial correlation
standardized slope
regression discontinuity design
residual plot
multivariate normal distribution
studentized deleted residual
DFBETAS
Cooks' D
multicollinearity
variance inflation factor
LAD estimation
bootstrap standard error
percentile bootstrap confidence interval
matrix scatterplot
semipartial correlation plot
partial correlation plot

Concept Questions

1. Explain why slope coefficients for a treatment factor in an experimental design do not have confounding variable bias.
2. Why is it important in a nonexperimental design to include confounding variables in the GLM?
3. Why is the adjusted squared multiple correlation estimate preferred to the unadjusted estimate?
4. How does the value of the squared multiple correlation between y and the predictor variables affect the width of a confidence interval for β_j ? Why?

5. When would a confidence interval for η^2 , rather than $\rho_{y.x}^2$, be reported?
6. How does the value of the squared multiple correlation between x_j and the other predictor variables affect the width of a confidence interval for β_j ? Why?
7. Why is it informative to estimate the conditional slope for x_1 at low and high values of x_2 when x_1 and x_2 interact?
8. If a categorical predictor variable has four levels, how many dummy coded predictor variables are needed and what values of the dummy codes define each of the four levels?
9. What are two ways to compute the conditional slopes for x_1 at the levels of x_2 if x_2 is a factor?
10. How is an interaction effect incorporated into a GLM?
11. How is a quadratic effect incorporated into a GLM?
12. Why will an ANCOVA potentially require a smaller sample size than an ANOVA in an experimental design?
13. When would a partial or semipartial correlation be preferred to a squared multiple correlation?
14. How are semipartial correlations useful in a regression analysis?
15. How is β_1 interpreted in a two-group experiment where group membership is effect coded?
16. How are interaction effects in a 2×3 design coded if both factors are qualitative?
17. How is the interaction effect in a 2×3 design coded if both factors are quantitative?
18. Under what conditions would conditional slopes be interesting?
19. What is the main difference between a partial correlation and a semipartial correlation?
20. What is the primary reason for using an ANCOVA rather than an ANOVA in experimental designs?
21. What is the primary reason for using an ANCOVA rather than an ANOVA in nonexperimental designs?

22. What should be done if a particular participant has an unusually large residual, DFBETAS value for any slope coefficient, or Cook's D value?
23. What are some simple ways to assess the bivariate normality assumption?
24. What is the main advantage of a regression discontinuity design?
25. How does the planning value of a partial correlation affect the sample size requirement for estimating the partial correlation with desired precision?
26. Why are slope coefficients sometimes difficult to interpret when the predictor variables are highly correlated?
27. Suppose the sample size requirement for a linear contrast of means with desired power in an experimental design is 40 per group. Suppose that high school GPA is correlated .5 with the response variable. What would the sample size requirement be if GPA was used as the covariate?
28. Explain how the multiple correlation, the semipartial correlation, and the partial correlation can be expressed in terms of a Pearson correlation between two variables.
29. Fill in the blanks in the following ANOVA table where the total sample size is $n = 100$ and the GLM has $s = 5$ predictor variables.

Source	SS	df	MS	F
MODEL	210	—	—	—
ERROR	564	—	—	
TOTAL	—	—		

30. Can adding a predictor variable decrease the value of $\hat{\rho}_{y,x}^2$ or $\hat{\eta}^2$. Why? Can adding a predictor variable decrease the value of $adj \hat{\rho}_{y,x}^2$? Why?
31. Why is the interpretation of β_j simplified if VIF_j for x_j is close to 1.0?

Data Analysis Problems

2-1. A random sample of 12 students was taken from a list of 652 fifth-grade students. Each of the 12 students was given a mathematics self-efficacy questionnaire followed by a comprehensive fifth-grade mathematics exam that had no time limit. In previous studies, the researcher noticed that students with higher levels of math self-efficacy tended to work longer on untimed tests and which could lead to higher test scores. The

time (in minutes) each student worked on the mathematics exam was also recorded. The data are given below.

Student	1	2	3	4	5	6	7	8	9	10	11	12
Math test score	20	36	72	40	95	71	65	48	55	85	92	45
Math SE score	4	5	12	8	19	10	8	8	10	12	15	7
Time	32	36	112	60	130	121	65	48	105	120	128	45

- Describe the study population.
- What is the control variable?
- Compute a 95% confidence interval for the population partial correlation between mathematics self efficacy and mathematics examination performance controlling for time spent on the mathematic examination. Interpret the result.
- The results of this study are potentially interesting but the confidence interval was much too wide to report in a scientific publication. The researcher wants to replicate this study and compute a 95% confidence interval that has a width of about 0.25. How many students should the researcher sample? Use the 95% lower limit (i.e., the endpoint closest to zero in this case) of the partial correlation from this study as a planning value.

2-2. Twenty male first-year college students and their fathers were randomly selected from a university orientation for all incoming male students. A trait aggression questionnaire measured on a 0 to 100 scale was given to the sample of 20 male students and their fathers. The sons were also asked to estimate the average number of hours per day, during their summer break, that they had played any type of violent video game. The sample results are given below where y = son's aggression score, x_1 = average hours per day of violent video game playing, and x_2 = father's aggression score.

Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
y	50	62	62	50	49	60	40	36	79	39	36	55	52	42	59	50	57	60	58	67
x_1	5	6	2	5	3	6	4	2	7	4	0	4	2	0	4	4	5	7	4	3
x_2	62	54	73	39	51	57	30	45	64	33	45	46	52	35	51	42	38	65	47	76

- Describe the study population.
- What is the response variable and what are the two predictor variables in this study?
- Compute 95% confidence intervals for the two population slope coefficients and interpret the results.

- d) Compute a 95% confidence interval for the population squared multiple correlation and interpret this result.
- e) Compute 95% confidence intervals for semipartial correlation between y and x_1 and y and x_2 . Interpret these results.
- f) Another researcher wants to conduct a similar study using a sample of male community college students and their fathers. The researcher suspects that the population squared correlation is about 0.7 and would like the 95% confidence interval for the squared multiple correlation to have a width of .2. How many male first-year students should be sampled?
- g) Compute a scatterplot for each pair of variables.

2-3. Thirty-six participants were randomly selected from a volunteer pool consisting of about 4,000 undergraduate students. The participants were randomized into one of nine treatment conditions in a 3×3 factorial experiment. The participants each played a computer checker game that was set to a difficulty level of 1 (easiest), 5, or 9 (hardest) and played while 0, 1, or 3 other students watched. Heart rate (beats per minute) 2 minutes into the game was determined for each participant. The sample data are shown below.

		Difficulty Level											
		1				5				9			
		=====											
Spectators	0	67	64	68	70	70	66	71	74	74	78	75	65
	1	73	71	69	79	78	74	75	81	83	80	75	80
	3	75	74	72	82	80	83	75	90	87	84	91	85

- a) Describe the study population.
- b) What is the response variable and what are the predictor variables in this experiment?
- c) Compute a 95% confidence interval for the two main effects and the interaction effect. Interpret the results. (Both predictor variables have been centered in the SPSS data file.)
- d) Use Equation 1.32 to determine the number of participants that should be added to this study to cut each 95% confidence interval width in half (i.e., set $w_0/w = 2$).

2-4. "Stereotype threat" is believed to affect members of a stereotyped group who feel extra pressure in situations where their behavior might confirm the negative reputation of their group. A random sample of 20 female participants was obtained from a university

database of about 1,200 undergraduate students. The 20 women were randomized into two groups of equal size and asked to complete a mathematics test. Participants in group 1 were simply told to complete the test. Participants in group 2 were (falsely) told, “Men typically perform better on this particular test than women and the purpose of this study is to help us understand why.” The college GPA for each student was used as a covariate. The test scores and GPAs are shown below.

Group	Variable	Scores									
1	Test Score	10	8	9	12	10	8	9	10	13	11
	GPA	3.6	3.1	3.5	3.7	3.5	3.0	3.0	3.1	3.8	3.4
2	Test Score	7	8	11	6	5	8	9	13	7	5
	GPA	3.4	3.6	3.7	3.1	3.2	3.5	3.5	3.4	3.1	3.0

- Describe the study population.
- Use a GLM with centered GPA and dummy coding for group (Instruction). Perform a test of the GPA $\times$ Group interaction. Can an analysis of the main effect of Group be justified? Why?
- Compute a 95% confidence interval for the main effect of Instruction. Interpret the result.
- The researcher wants to replicate this experiment using a larger sample with the goal of obtaining a narrower confidence interval for the population mean difference. How many participants per group are needed to obtain a 95% confidence interval for the population mean difference that has a width of 1.5 using GPA as a covariate. Use the MS_E from this study as a planning value of the average within-group residual variance.

Suggested Readings

- Cohen, J., Cohen, P. West, S. G., & Aiken (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Erlbaum. **Chapters 3 - 10**
- Darlington, R.B., & Hayes, A. F. (2017). *Regression analysis and linear models: Concepts, applications, and implementation*. Guilford. **Chapters 3 and 4**
- Fox, J. (2016). *Applied regression analysis and generalized linear models* (3rd ed.). Sage. **Sections 3.3, 5.2, 6.2, 7.2, and 7.3**
- Graybill, F. A., & Iyer, H. K. (1994). *Regression analysis: Concepts and applications*. Duxbury. http://www.stat.colostate.edu/~hari/regression_book/ **Chapter 4**

Chapter 3

Statistical Models for Repeated Measurements

3.1 Between-subjects and Within-subjects Predictor Variables

In a repeated measures design, each participant is measured on $r \geq 2$ occasions. The pretest-posttest design and within-subjects experimental design described in Chapter 4 of Volume 1 are special cases of the repeated measured design. In a pretest-posttest design, the response variable is measured on one or more occasions prior to treatment and then on one or more occasions following treatment. In a within-subjects experimental design, the response variable for each participant is measured under all treatment conditions (usually in counterbalanced order). The longitudinal design is also a special case of the repeated measures design. In a longitudinal design, the response variable for each participant is measured on two or more occasions with no treatment conditions.

Two types of predictor variables can be used in a repeated measures design. One type of predictor variable is a between-subjects predictor variable (also called a time-invariant predictor variable) that describes differences in y scores among participants. The values of a between-subjects predictor variable do not vary over the repeated measures. Gender, political affiliation, and a between-subjects treatment are a few examples of between-subjects predictor variables. All of the predictor variables in Chapters 1 and 2 are

between-subjects predictor variables because they are used to describe differences in y scores among the participants at a single point in time. Another type of predictor variable is the within-subjects predictor variable (also called a time-varying predictor variable) that describes how the y scores change across the repeated measurements. The statistical methods described in this chapter can accommodate both between-subjects and within-subjects predictor variables.

3.2 Covariance Pattern Models

In a GLM, each prediction error (e_i) corresponds to a different participant and the independence assumption implies that the prediction errors are uncorrelated. In a repeated measures design, each participant will have multiple y scores and there is a corresponding prediction error for each of these y scores. It is not reasonable to assume that the multiple prediction errors for each participant are uncorrelated. A *covariance pattern model* is useful in repeated measures designs because it can accommodate correlated prediction errors.

In section 4.12 of Volume 1, the data file for a repeated measures study was arranged in either a "wide format" or a "long format". In a wide format, each row contains the r measurements of a single participant. The covariance pattern model requires the data file to be in long format where the r repeated measurements for a participant appear in r different rows. Some traditional repeated measures statistical methods require the data to be in wide format and will use listwise deletion to eliminate any participant from the analysis that has missing data on one or more of the repeated measures. In comparison, a covariance pattern model will use all of the available data for each participant.

Consider a very basic covariance pattern model that has only one between-subjects predictor variable and one within-subjects predictor variable. Suppose a sample of 30 first grade boys and girls are given a reading ability test in September, October, November, and December. The covariance pattern model for this two-group repeated measures design is

$$y_{ij} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2ij} + e_{ij} \quad (\text{Model 3.1})$$

where $i = 1$ to 30 , $j = 1$ to 4 , x_{1i} is a dummy coded variable for gender, and x_{2ij} has values 1, 2, 3, 4 corresponding to the September, October, November, and December time

periods. Note that the i subscript specifies a particular participant and the j subscript specifies a particular occasion. Note also that the dummy coded variable does not have a j subscript because it is a between-subjects predictor variable that describes gender differences among the participants. The within-subjects predictor variable (x_{2ij}) has both an i subscript and a j subscript because the values of this predictor variable vary across the repeated measures.

In Model 3.1, β_1 describes the gender difference in population mean reading scores, averaged over the four months, and β_2 describes the linear within-gender change in population mean reading scores over the four time periods. Model 3.1 assumes that β_2 is equal for boys and girls. This equal slope assumption could be relaxed by adding a group by time interaction effect to the model.

A general covariance pattern model can have all between-subjects predictors, all within-subjects predictors, or a combination of between-subjects and within-subjects predictors. The between-subjects and within-subjects predictors can be dummy coded or effect coded variables, fixed quantitative variables, or random quantitative variables. The predictor variables can be squared variables to describe quadratic effects or product variables to describe interaction effects. A product variable can be a product of two between-subjects predictor variables or a product of two within-subjects predictor variables. The model can also include a "cross-level interaction" which is the product of a between-subjects predictor variable and a within-subjects predictor variable.

The covariance pattern model has $n \times r$ prediction errors ($e_{11}, \dots, e_{1r}, e_{21}, \dots, e_{2r}, \dots, e_{n1}, \dots, e_{nr}$). The r prediction errors for each participant in the covariance pattern model are assumed to be correlated and possibly also have unequal variances. The variances and covariances of the prediction errors in a covariance pattern model can be represented by a *prediction error covariance matrix* as described in section 3.3.

Repeated measurements can be *balanced* or *unbalanced*. The repeated measurements are balanced if each participant is measured on the same set of time points. The repeated measurements are unbalanced if participants are measured on different sets of time points (e.g., one participant is measured at time points 1, 2, and 4 while another participant is measured at time points 1, 3, 8, and 15). Random missing data is allowed with both balanced and unbalanced repeated measurements.

3.3 Prediction Error Covariance Matrices

A prediction error covariance matrix is a symmetric matrix with variances of the prediction errors in the diagonal elements and covariances among pairs of prediction errors in the off-diagonal elements (see section B.3 in Appendix B). In the GLM where the n prediction errors are assumed to be uncorrelated and have the same variance, the prediction error covariance matrix for the $n \times 1$ vector of prediction errors ($\mathbf{e}$) has the following *scaled identity* structure

$$\text{cov}(\mathbf{e}) = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

which can be expressed more compactly as $\text{cov}(\mathbf{e}) = \sigma^2 \mathbf{I}_n$ where $\mathbf{I}_n$ is an $n \times n$ identity matrix (see section B.1 in Appendix B).

The prediction error covariance matrix in a covariance pattern model has the following block matrix structure

$$\text{cov}(\mathbf{e}) = \begin{bmatrix} \mathbf{\Sigma} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{\Sigma} \end{bmatrix}$$

which can be expressed more compactly as $\text{cov}(\mathbf{e}) = \mathbf{I}_n \otimes \mathbf{\Sigma}$ where $\mathbf{\Sigma}$ is an $r \times r$ covariance matrix for the r prediction errors for a participant, each $\mathbf{0}$ is an $r \times r$ matrix of zeros, and $\otimes$ denotes the Kronecker product (see section B.2 in Appendix B). For example, with $r = 3$

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}$$

where σ_1^2 is the prediction error variance for occasion 1, σ_2^2 is the prediction error variance for occasion 2, σ_3^2 is the prediction error variance for occasion 3, σ_{12} is the covariance of prediction errors for occasions 1 and 2, σ_{13} is the covariance of prediction errors for occasions 1 and 3, and σ_{23} is the covariance of prediction errors for occasions 2 and 3.

The above covariance matrix is referred to as an *unstructured* covariance matrix because there are no constraints imposed on the values of the variances or covariances. An unstructured covariance matrix requires the estimation of r variances and $r(r - 1)/2$ covariances, or a total of $r(r + 1)/2$ parameters.

A covariance matrix where all variances are constrained to be equal and all correlations are constrained to be equal is called a *compound-symmetric covariance matrix* (also called an *exchangeable* covariance matrix) and is illustrated below for $r = 4$. A compound-symmetric covariance matrix requires the estimation of only two parameters (σ^2 and ρ). A compound-symmetric covariance structure might be justified in a within-subjects experiment where participants are measured under r treatment conditions in random or counterbalanced order.

$$\Sigma = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \sigma^2 \end{bmatrix}$$

If the response variable is measured over r equally or nearly equally spaced time intervals, as is often the case in a longitudinal or pretest-posttest design, a *first-order autoregressive* (AR-1) covariance structure could be justified. The AR-1 covariance structure constrains all the variances to be equal and constrains the correlations so that they decrease exponentially in magnitude as the separation in time increases. An example of an AR-1 structure for $r = 4$ is shown below.

$$\Sigma = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \rho^2\sigma^2 & \rho^3\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \rho^2\sigma^2 \\ \rho^2\sigma^2 & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ \rho^3\sigma^2 & \rho^2\sigma^2 & \rho\sigma^2 & \sigma^2 \end{bmatrix}$$

Like a compound-symmetric covariance matrix, an AR-1 covariance matrix requires the estimation of only two parameters (σ^2 and ρ).

If the r prediction errors are assumed to be uncorrelated but have unequal variances, these assumptions imply a *diagonal* prediction error covariance matrix illustrated below for $r = 3$.

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix}$$

The unstructured prediction error matrix will be used in section 3.9 to estimate robust standard errors for the parameter estimates of a covariance structure model. The compound symmetric matrix has important implications in the random coefficient model that is described in section 3.7. The AR-1 and diagonal error covariance structures can also be applied in random coefficient model to relax certain assumptions.

3.4 Centering the Predictor Variables

Consider a covariance pattern model that has only time (x_{1ij}) as a predictor variable

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + e_{ij} \quad (\text{Model 3.2})$$

where β_1 is the slope of the line relating x_1 to y and β_0 is the y -intercept. Suppose x_{1ij} was coded 1, 2, 3, 4, 5 to represent five possible weeks when a participant could be measured. With x_{1ij} coded this way, β_0 describes the predicted y score for $x_{1ij} = 0$ which would correspond to one week *prior* to the start of the study. If x_{1ij} had instead been *baseline centered* so that x_{1ij} was coded 0, 1, 2, 3, 4, then β_0 would describe the predicted y score for the first week of the study. Alternatively, the time variable could be mean centered. After mean centering, the values of x_{1ij} are -2, -1, 0, 1, 2 and then β_0 would describe the predicted y score for the third week of the study.

Consider the following covariance pattern model that has one between-subjects predictor variable and one within-subjects predictor variable

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2i} + e_{ij} \quad (\text{Model 3.3})$$

where x_{1ij} is the time value for participant i on occasion j and x_{2i} is a between-subjects predictor variable. As an example, if x_{1ij} is a baseline centered week variable, x_{2i} is an ACT score for participant i , and y_{ij} is a chemistry quiz score in week j for participant i ,

then β_0 describes the predicted chemistry quiz score at week 1 for participants with an ACT score of 0. This y -intercept is not meaningful because an ACT score of 0 is impossible. However, if the ACT scores are mean centered, then β_0 describes the predicted chemistry quiz score at week 1 for participants with an average ACT score.

If a product of two quantitative between-subjects predictor variables is included in the model to assess an interaction effect, these predictor variables should be mean centered before computing the product variable. Also, if a squared between-subjects predictor variable is included in the model to assess a quadratic effect, the predictor variable should be mean centered before computing the squared variable. It is usually a good idea to mean center all quantitative between-subjects predictor variables.

Quantitative within-subjects predictor variables (other than time) that have average values that differ across participants usually should be *person centered* rather than mean centered. For example, if happiness is a within-subjects predictor variable across $r = 4$ repeated measurements, the average of the four happiness scores for each participant could vary substantially across participants. The happiness scores can be person centered by computing the average happiness score for each participant. Then, for each participant, subtract the average happiness score from each of the repeated happiness scores.

Consider the following covariance pattern model that adds a second within-subjects predictor variable to Model 3.2

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + e_{ij} \quad (\text{Model 3.4})$$

where x_{1ij} is a baseline centered time value for participant i on occasion j , and x_{2ij} is a within-subjects predictor variable for participant i on occasion j . In this model, β_2 describes the slope of the line relating y to x_2 across all time periods and all participants.

If participants have substantially different x_2 scores, β_2 will be a misleading description of the relation between y and x_2 within each person. Consider the scatterplot in Figure 3.1 for two participants who have been measured on seven occasions and have substantially different x_2 scores. The line of best fit for x_2 has a negative slope while the line of best fit is positive after person centering the x_2 scores. When a within-subjects predictor variable (x_2) has been person-centered, the slope coefficient for x_2 describes the within-person relation between y and x_2 .

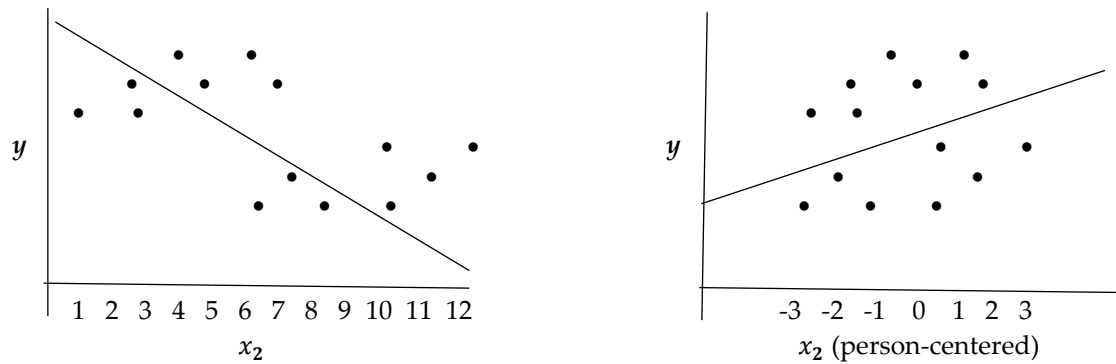


Figure 3.1 Scatterplots without (left) and with (right) person-centered x_2 scores

Some of the variability in y can usually be predicted by between-person differences in x_2 , but these between-person differences are lost when x_2 is person centered. This lost information can be recovered by simply adding another predictor variable to the model that represents the average within-subjects predictor score for each participant as shown below

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3i} + e_{ij} \quad (\text{Model 3.5})$$

where x_{2ij} is the person-centered within-subjects predictor variable score for participant i on occasion j , and x_{3i} is the average within-subjects predictor score for participant i . Note that x_3 is a between-subjects predictor variable. In Model 3.5, β_2 describes the slope of the line relating y to x_2 *within* persons and β_3 describes the slope of the line relating y to x_2 *across* persons.

To illustrate the computation of x_2 and x_3 in Model 3.5, consider the following hypothetical data for the first two participants where x_1 is the time variable (baseline centered) and x_2 is an uncentered within-subjects predictor variable. Participant 1 was measured on three occasions and participant 2 was measured on four occasions.

Participant	y	x_1	x_2
1	15	0	7
1	19	1	9
1	22	2	11
2	23	0	16
2	27	1	19
2	34	2	25
2	35	3	28

The mean of the x_2 scores for participant 1 is $(7 + 9 + 11)/3 = 9$ and the mean of the x_2 scores for participant 2 is $(16 + 19 + 25 + 28)/4 = 22$. To person center x_2 , subtract 9 from the x_2 scores for participant 1 and subtract 22 from the x_2 scores for participant 2. The person centered x_2 scores are given below along with a new between-subjects variable (x_3) that represents the person averages of x_2 .

Participant	y	x_1	x_2	x_3
1	15	0	-2	9
1	19	1	0	9
1	22	2	2	9
2	23	0	-6	22
2	27	1	-3	22
2	34	2	3	22
2	35	3	6	22

3.5 Modeling the Repeated Measurements

As noted above, all of the predictor variable types (e.g., quantitative, dummy or effect coded, squared quantitative, and product) that were used in the GLM in Chapter 2 to represent differences among the n participants can also be included as between-subjects predictor variables in a covariance pattern model. Within-subjects predictor variables can be included in a covariance pattern model to represent differences among the r repeated measures within participants. Several basic types of within-subjects predictor variables are described below for longitudinal designs, pretest-posttest designs, and within-subjects experimental designs.

In a longitudinal design, the covariance pattern model will usually include a within-subjects predictor variable that represents points in time. For example, suppose a social skills score is obtained for a sample of kindergarten students in four consecutive months. Hypothetical data for the first two students are shown below.

Participant	Month	SocSkill
1	1	24
1	2	35
1	3	28
1	4	19
2	1	30
2	2	39
2	3	32
2	4	29
⋮	⋮	⋮

If the social skills response variable is assumed to have a quadratic trend over time, month-squared could be included as an additional within-subjects predictor variable. The mean centered month variable and its squared value for the two students are shown below.

Participant	Month	Month ²	SocSkill
1	-1.5	2.25	24
1	-0.5	0.25	35
1	0.5	0.25	28
1	1.5	2.25	19
2	-1.5	2.25	30
2	-0.5	0.25	39
2	0.5	0.25	32
2	1.5	2.25	29
⋮	⋮	⋮	⋮

A covariance pattern model with only month as a predictor variable implies a linear relation between month and social skills as illustrated in Figure 3.2a. If month-squared is added to the model, then the model implies a quadratic relation between month and social skills. Recall from Chapter 2 that a quadratic relation can describe a curve that with a single bend. An example of a quadratic relation is illustrated in Figure 3.2b.

In pretest-posttest designs, a dummy variable could be added to represent the pre-treatment and post-treatment time periods. For example, suppose the social skills of each kindergarten student in the sample is measured every month for three months before exposure to a social skills training program and then every month for two months following training. Hypothetical data for the first two students are shown below. A dummy coded variable for treatment is equal to 0 in months 1, 2, and 3 and equal to 1 in months 3 and 4.

Participant	Month	Treatment	SocSkill
1	1	0	24
1	2	0	35
1	3	0	28
1	4	1	24
1	5	1	35
2	1	0	30
2	2	0	39
2	3	0	32
2	4	1	30
2	5	1	39
⋮	⋮	⋮	⋮

If a covariance pattern model for a longitudinal or pretest-posttest design only includes the treatment dummy variable as a predictor variable, the model implies a horizontal trend prior to treatment and a jump after treatment that remains horizontal as illustrated

in Figure 3.2c. If both month and the treatment dummy variable are included as predictor variables, the model implies a linear trend prior to treatment with a jump following treatment with a linear trend following treatment that has the same slope as during pretreatment time periods (see Figure 3.2d). If the pretreatment and posttreatment slopes are expected to differ (see Figure 3.2e), then the product of month and treatment can be added to the model as shown below.

Participant	Month	Treatment	Month x Treatment	SocSkill
1	1	0	0	24
1	2	0	0	35
1	3	0	0	28
1	4	1	4	24
1	5	1	5	35
2	1	0	0	30
2	2	0	0	39
2	3	0	0	32
2	4	1	4	30
2	5	1	5	39
⋮	⋮	⋮	⋮	⋮

In Figure 3.2e, there is a jump and a change in the slope following treatment. If the pretreatment slope is assumed to differ from the posttreatment slope but the two lines are assumed to connect (see Figure 3.2f), this pattern can be modeled by removing the dummy variable for treatment and including only the month and month x treatment variables in the model. A model that will produce predicted values like Figure 3.2f is called a "spline model".

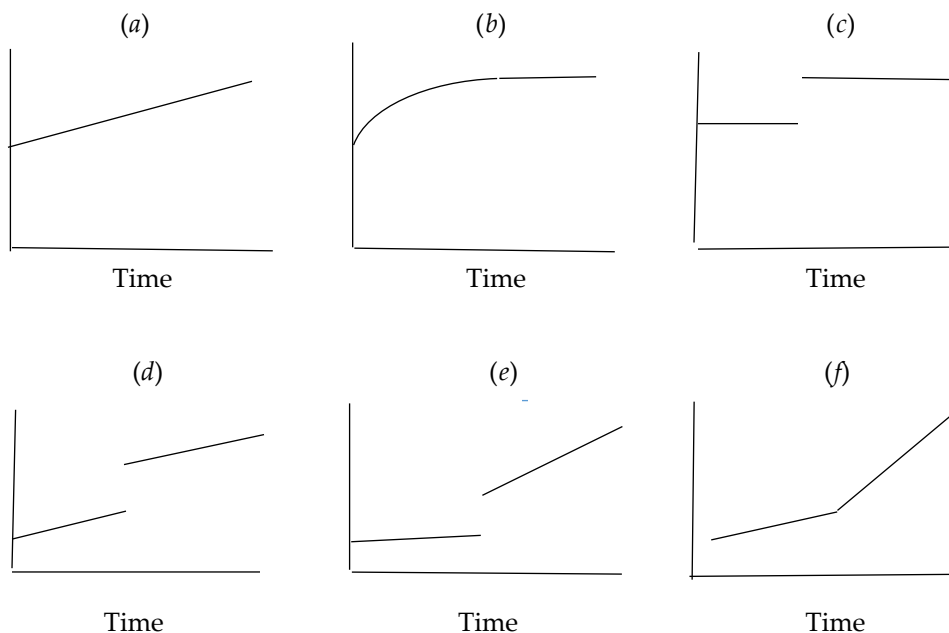


Figure 3.2. Examples of time trends

In within-subject experiments, where participants are measured under all treatment conditions of a qualitative factor, $a - 1$ dummy coded or effect coded variables are needed to represent the a -level treatment factor. For example, with $a = 3$ treatments, the data file could include two dummy coded variables as shown below with hypothetical response variable scores for the first two participants.

Participant	dummy1	dummy2	Response
1	1	0	61
1	0	1	57
1	0	0	78
2	1	0	54
2	0	1	62
2	0	0	48
:	:	:	:

Within-subjects covariates can be included in longitudinal, pretest-posttest or within-subjects experimental designs. For example, in the above within-subjects experiment, suppose the response variable is the number of questions answered correctly after reading three short reports and suppose that participants differ in the length of time they read each report. If reading comprehension is related to reading time, reading time could be included as a within-subjects covariate as illustrated below for the first two participants.

Participant	dummy1	dummy2	Minutes	Score
1	1	0	3.5	61
1	0	1	4.8	57
1	0	0	6.1	78
2	1	0	4.6	54
2	0	1	5.9	62
2	0	0	4.1	48
:	:	:	:	:

In longitudinal and pretest-posttest designs, a *lagged* within-subjects covariate is useful if the covariate has a delayed effect on the response variable. With a *one-period* lagged covariate, the value of the response variable at each point in time is predicted from the value of the covariate at the prior point in time. For example, suppose a sample of first-year college students agree to report their number of close friends and their loneliness each month for six months. The researcher believes that the number of close friends reported in the prior month is a predictor of loneliness in the current month. Hypothetical data are given below where the friend variable has been lagged one month. With a one-period lagged predictor variable, the first time period (month = 1) is excluded from the analysis because the lagged predictor variable value is unavailable at time 1.

Participant	Month	Friends	FriendsLag	Loneliness
1	1	0	--	48
1	2	1	0	46
1	3	2	1	44
1	4	3	2	44
1	5	6	3	38
1	6	3	6	30
2	1	2	--	39
2	2	3	2	37
2	3	4	3	35
2	4	4	4	30
2	5	5	4	28
2	6	2	5	20
:	:	:	:	:

In all of the above examples, between-subjects predictor variables can be added to the model to describe quantitative or qualitative differences among the participants. For example, participant gender could be included as a between-subjects predictor variable in the above loneliness study. Suppose the first participant is male and the second participant is female. The data file could include a dummy variable coding gender as shown below. Note that the value of the between-subjects dummy variable coding gender is unchanged across all five time periods (because gender does not change over time) and differs only across participants.

Participant	Month	Gender	Friends	FriendsLag	Loneliness
1	1	1	0	--	48
1	2	1	1	0	46
1	3	1	2	1	44
1	4	1	3	2	44
1	5	1	6	3	38
1	6	1	3	6	30
2	1	0	2	--	39
2	2	0	3	2	37
2	3	0	4	3	35
2	4	0	4	4	30
2	5	0	5	4	28
2	6	0	2	5	20
:	:	:	:	:	:

Example 3.1. One hundred college students were randomly selected from a study population of about 32,000 students. The students were given a cultural sensitivity questionnaire twice a year (October and March) for two years. During the summer after the first year, the students were given course credit to complete an online cultural sensitivity training program. A covariance pattern model was used to analyze the data with time (coded 0, 1, 2, 3) as a within-subjects predictor variable and a dummy variable coding the summer intervention as another within-subjects predictor variable. An unstructured error covariance matrix was used. The 95% confidence interval for the dummy variable slope coefficient is [3.62, 5.34]. We can be 95% confident that if all 32,000 students were measured on those four time periods and received training during the summer, their mean cultural sensitivity score would increase 3.62 to 5.34 points after training.

3.6 Data Transformations for Monotonic Time Trends

Recall from section 2.10 that a quadratic relation can describe a curve that has a single bend. An example of a quadratic relation is illustrated in Figure 3b. If the relation between the time variable (e.g., month) and the response variable is nonlinear but monotonic, as in Figure 3.2b, then using $\sqrt{\text{month}}$ as a single predictor variable instead of the two predictor variables, month and month-squared, could simplify the analyses.

Four monotonic time trends are illustrated in Figure 3.3. In Figures 3.3a, 3.3b, and 3.3c, a quadratic model with time and time-squared as predictor variables could be used, but these three types of monotonic relations can often be linearized using a data transformation. As in Figure 3.2b, the relation in Figure 3.3a could be linearized using $\sqrt{\text{time}}$ as a predictor variable. The monotonic relation in Figure 3.3b could be linearized using $1/\text{time}$, $1/\sqrt{\text{time}}$, or $1/\sqrt[4]{\text{time}}$ as a predictor variable. The monotonic relation in Figure 3.3c could be linearized using time as the predictor variable and using $\sqrt{y}$ or $\ln(y)$ as the response variable.

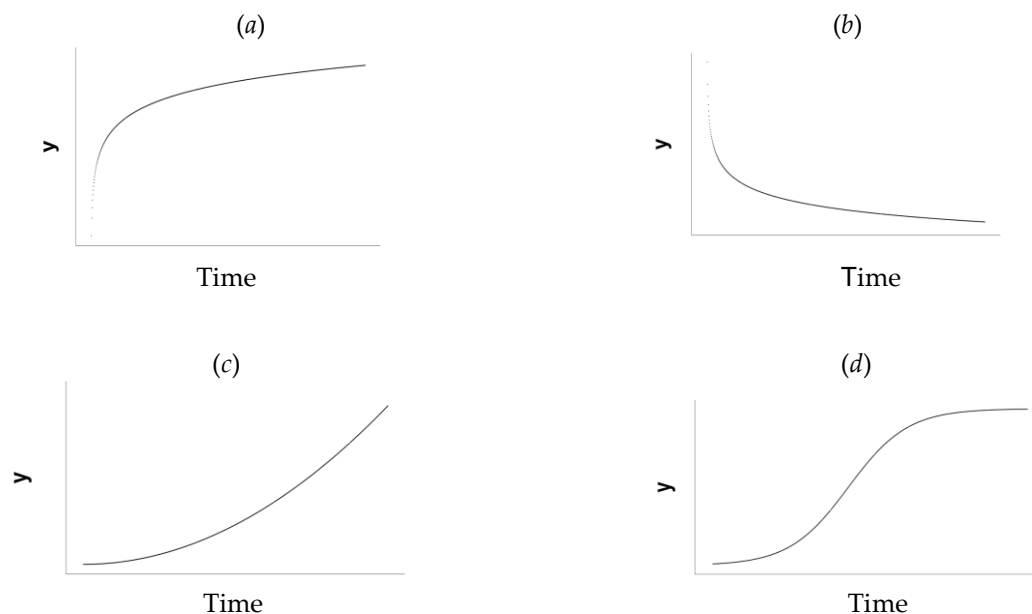


Figure 3.3 Examples of monotonic time trends

The relation in Figure 3.3d has two bends and can be described by a *cubic* relation using time, time-squared, and time-cubed as three predictor variables. However, the relation in

Figure 3.3d is monotonic and could be linearized using time as a single predictor variable and applying the response variable transformation $\ln[y/(y_{max} - y + c)]$ where y_{max} is the largest y score and c is a small positive number to avoid division by 0.

3.7 Random Coefficient Models

A *random coefficient model*, which is also referred to as a *multi-level model*, a *hierarchical linear model*, or a *mixed linear model*, is an alternative to the covariance pattern model for repeated measures designs. Like the covariance pattern model, a random coefficient model can include between-subjects and within-subjects predictor variables. All of the methods for modeling the repeated measurements in a covariance pattern model can be applied to a random coefficient model. All of the issues regarding the centering of predictor variables in a covariance pattern model also apply to a random coefficient model. Like the covariance pattern model, the random coefficient model will use all of the available data for each participant.

In repeated measures designs, a random coefficient model can be expressed in terms of a *level-1 model* that includes only within-subjects predictor variables and one or more *level-2 models* that include only between-subjects predictor variables. Consider a longitudinal study where the researcher believes that the response variable (y) changes linearly over time. A level-1 model that includes only the time predictor variable (x_1) is given below

$$y_{ij} = b_{0i} + b_{1i}x_{1ij} + e_{ij} \quad (\text{Model 3.6})$$

where the values of x_{1ij} are the time points at which participant i was measured. Note that the parameters of the level-1 model contain an i subscript to indicate that participants have their own y -intercept (b_{0i}) and slope (b_{1i}) values. The prediction errors (e_{ij}) in the level-1 model are assumed to be uncorrelated across participants and occasions. The prediction errors in the level-1 model are also assumed to have equal variances across participants and occasions (but there are options to relax these assumptions).

The n participants are assumed to be a random sample from some specified study population of N people. The level-1 model indicates that each of the N participants have their own y -intercept and slope values. The level-1 model describes a random sample of n participants and thus the b_{0i} and b_{1i} values ($i = 1$ to n) are a random sample of a

population of y -intercept and slope values. In the same way that a statistical model describes a random sample of y scores, statistical models can be used to describe a random sample of b_{0i} and b_{1i} values. The statistical models for b_{0i} and b_{1i} are the level-2 models.

The following level-2 models for b_{0i} and b_{1i} are "intercept-only" models because they have no predictor variables

$$b_{0i} = \beta_0 + u_{0i} \quad (\text{Model 3.7a})$$

$$b_{1i} = \beta_1 + u_{1i} \quad (\text{Model 3.7b})$$

where u_{0i} and u_{1i} are *parameter prediction errors* for the random values of b_{0i} and b_{1i} , respectively. These two parameter prediction errors can be correlated with each other but they are assumed to be uncorrelated with the level-1 prediction errors (e_{ij}). The n parameter prediction errors for b_{0i} are assumed to be uncorrelated with each other and have a variance equal to $\sigma_{u_0}^2$. Likewise, the n parameter prediction errors for b_{1i} are assumed to be uncorrelated with each other and have a variance equal to $\sigma_{u_1}^2$. In Models 3.7a and 3.7b, the variance of the b_{1i} values in the population is described by $\sigma_{u_1}^2$, and the variance of the b_{0i} values is described by $\sigma_{u_0}^2$. The variability of the y -intercept values (b_{0i}) is usually interesting only if the variability of the slope values (b_{1i}) is small. The graphs in Figure 3.4 illustrate a sample of $n = 5$ participants where the slope variance is large (left) and the slope variance is small (right).

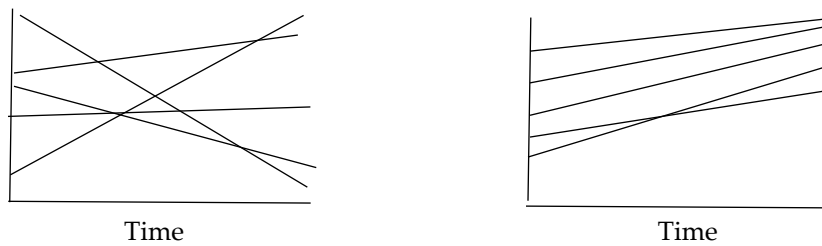


Figure 3.4 Examples of large slope variance (left) and small slope variance (right)

In Model 3.7a, β_0 is the population mean of the y -intercepts, and in Model 3.7b, β_1 is the population mean of the slope coefficients. Mixed linear model programs will compute estimates of β_0 , β_1 , $\sigma_{u_0}^2$, and $\sigma_{u_1}^2$ along with confidence intervals for β_0 , β_1 , $\sigma_{u_0}^2$, and $\sigma_{u_1}^2$. SPSS computes confidence intervals for the variances of the random coefficients while

the `lme` function in R computes confidence intervals for the standard deviations of the random coefficients. Standard deviations are easier to interpret than variances. When using SPSS, taking the square roots of the confidence interval endpoints for variances gives confidence intervals for standard deviations.

The level-2 models (Models 3.7a and 3.7b) can be substituted into the level-1 model to give the following *composite model*

$$\begin{aligned} y_{ij} &= (\beta_0 + u_{0i}) + (\beta_1 + u_{1i})x_{1ij} + e_{ij} \\ &= \beta_0 + \beta_1 x_{1ij} + e_{ij}^* \end{aligned} \quad (\text{Model 3.8})$$

where $e_{ij}^* = u_{0i} + u_{1i}x_{1ij} + e_{ij}$ is the *composite prediction error* for participant i at time j . It is important to note that the y -intercept (β_0) and the slope (β_1) in Model 3.8 are identical to the y -intercept (β_0) and slope (β_1) in a covariance pattern model with x_{1ij} as a within-subjects predictor variable.

In a random coefficient model, the y -intercepts (b_{0i}) are almost always assumed to be random. If the estimate of $\sigma_{u_0}^2$ is not small, this indicates that there could be interesting variability in the y -intercepts among people in the study population. One or more predictor variables can be included in Model 3.7a in an effort to explain some of the variability in the y -intercepts.

If the estimate of the slope variability ($\sigma_{u_1}^2$) is small, this variance could then be assumed to equal 0 and Model 3.7b reduces to $b_{1i} = \beta_1$. This level-2 model implies that the slope coefficient for x_1 is the same for everyone in the population and is equal to β_1 . In this case, the slope coefficient is said to be "fixed" rather than random.

If the estimate of $\sigma_{u_1}^2$ is not small, this indicates that there is potentially interesting variability in the slope coefficients among people in the study population. One or more predictor variables can be included in Model 3.7b in an effort to explain some of the variability in the slope coefficients. Any predictor variable that is used in the level-2 slope model is almost always used in the level-2 y -intercept model because any variable that is related to the slope is almost always also related to the y -intercept. As can be seen in Figure 3.3, participants with large positive slopes tend to have small y -intercepts and participants with large negative slopes tend to have large y -intercepts.

Consider a study where vocabulary size is measured in a sample of preschool children each month for five consecutive months. In Model 3.6, y_{ij} is vocabulary size for child i in month j and x_{ij} is the month ($j = 1, 2, 3, 4$, or 5) in which the vocabulary size was measured for child i . Suppose the standard deviation of the random slopes is large. This suggests that the slope coefficient for each child (b_{1i}) could vary considerable across children with some children having large gains and others having very little gain. The researcher suspects that younger children are more likely to have larger gains than older children. The following level-2 models for b_{0i} and b_{1i} in Model 3.6 could then be specified.

$$b_{0i} = \beta_0 + \beta_{02}x_{2i} + u_{0i} \quad (\text{Model 3.9a})$$

$$b_{1i} = \beta_1 + \beta_{12}x_{2i} + u_{1i} \quad (\text{Model 3.9b})$$

where x_{2i} is a between-subjects predictor variable that is equal to the age of child i at the beginning of the study.

Substituting Models 3.9a and 3.9b into the level-1 model (Model 3.6) gives the following composite model

$$y_{ij} = \beta_0 + \beta_1x_{1ij} + \beta_{02}x_{2i} + \beta_{12}x_{1ij}x_{2i} + e_{ij}^* \quad (\text{Model 3.10})$$

where $e_{ij}^* = u_{0i} + u_{1i}x_{1ij} + e_{ij}$. Note that a level-2 predictor variable of a random slope produces an interaction effect in the composite model. In this example, the composite model includes an age x time interaction effect which describes a relation between vocabulary size and time that depends on the child's age. The parameters in Model 3.10 are identical to the parameters in a covariance pattern model with time, age, and an age x time interaction as predictor variables (in a covariance pattern model, β_{02} in Model 3.10 would be labeled β_2 , and β_{12} in Model 3.10 would be labeled β_3). The parameters in a composite model are referred to as "fixed effects" in the output of a mixed linear model program.

3.8 Composite Prediction Error Covariance Structure

In a covariance pattern model, the researcher can specify the structure of the predictor error covariance matrix (with an unstructured error covariance matrix as an option). The random coefficient model does not have this flexibility. In a random coefficient model, the composite predictor error covariance structure depends on the variances and

covariances of the random coefficients. This is a limitation of the random coefficient model because the covariance structure of the composite prediction errors might be very unrealistic. If the covariance structure of the composite prediction errors does not accurately describe the true covariance structure, then the hypothesis testing results and confidence interval results in a random coefficient model can be very misleading.

To illustrate how the variances and covariances of the random coefficient model define the covariance structure of the composite prediction errors, consider a simple level-1 model like Model 3.6 where the slope is fixed and only the y -intercepts are random. The level-2 models are given below.

$$\begin{aligned} b_{0i} &= \beta_0 + u_{0i} \\ b_{1i} &= \beta_1 \end{aligned}$$

Substituting these level-2 models into the level-1 model gives the following composite model

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + e_{ij}^*$$

where $e_{ij}^* = u_{0i} + e_{ij}$. Using covariance algebra (see section B.5 in Appendix B) and assuming that u_{0i} and e_{ij} are uncorrelated, the variance of e_{ij}^* at time j is $\sigma_{u_0}^2 + \sigma_e^2$ and the covariance between two composite prediction errors at time j and time j' is $\sigma_{u_0}^2$. Thus, the composite prediction errors in this model are assumed to have a compound-symmetric covariance structure where all the variances are equal and all covariances are equal. The assumption of equal covariances is usually not realistic in longitudinal and pretest-posttest designs. The assumption of equal variances at each time point could also be unrealistic.

Now consider Model 3.6 where both the y -intercept and slope are random rather than fixed. The level-2 models are given in Models 3.7a and 3.7b. Substituting these level-2 models into the level -1 model gives Model 3.8 where the composite prediction errors are $e_{ij}^* = u_{0i} + u_{1i}x_{1ij} + e_{ij}$. Assuming u_{0i} and e_{ij} are uncorrelated, u_{1i} and e_{ij} are uncorrelated, but allowing u_{0i} and u_{1i} to be correlated, the variance of e_{ij}^* at time j is

$$\text{var}(e_{ij}^*) = \sigma_{u_0}^2 + 2x_{1j}\sigma_{u_0,u_1}^2 + x_{1j}^2\sigma_{u_1}^2 + \sigma_e^2 \quad (3.1)$$

and the covariance between two composite prediction errors at time j and time j' is

$$\text{cov}(e_{ij}^*, e_{ij'}^*) = \sigma_{u_0}^2 + (x_{1j} + x_{1j'})\sigma_{u_0, u_1}^2. \quad (3.2)$$

Note that $\text{var}(e_{ij}^*)$ is now assumed to be a quadratic function of x_1 which is usually an unrealistic assumption.

To illustrate the covariance structure implied by these composite prediction errors, suppose the estimates of $\sigma_{u_0}^2$, $\sigma_{u_1}^2$, σ_e^2 , and σ_{u_0, u_1}^2 are 5.3, 2.4, 3.8, and -1.3, respectively. Next, assume that the participants were measured on four equally spaced time points with baseline centering so that $x_{1i1} = 0$, $x_{1i2} = 1$, $x_{1i3} = 2$, and $x_{1i4} = 3$. Plugging these values into Equations 3.1 and 3.2 gives the following 4 x 4 composite prediction error covariance matrix.

$$\hat{\Sigma} = \begin{bmatrix} 9.1 & 4.0 & 2.7 & 1.4 \\ 4.0 & 8.9 & 1.4 & 0.1 \\ 2.7 & 1.4 & 13.5 & -1.2 \\ 1.4 & 0.1 & -1.2 & 22.9 \end{bmatrix}$$

With the assumption of a random y -intercept and a random slope, the resulting composite prediction error covariance matrix assumes that the variances decrease from 9.1 to 8.9 in periods 1 and 2 and then increase from 13.5 to 22.9 in periods 3 and 4. Note also that the covariance for periods 3 and 4 is assumed to be negative. This pattern of variances and covariances could be very poor approximation to the true prediction error covariance structure.

If the prediction errors in the level-1 model are assumed to be uncorrelated and have equal variances, the composite prediction error covariance matrix in a random coefficient model will often be a poor approximation to the true composite prediction error covariance matrix. To address this problem, mixed linear model programs have options to specify more realistic level-1 prediction error covariance structures. In longitudinal and pretest-posttest designs with equally or nearly equally spaced time points, a first-order autoregressive prediction error structure for the level-1 prediction errors usually gives a more realistic composite prediction error covariance matrix. Alternatively, the level-1 prediction errors could be assumed to have unequal variances by specifying a diagonal covariance structure for the level-1 prediction errors. Although mixed linear model programs have options to specify more general level-1 prediction error covariance

structures, the parameter estimates of these covariance structures can be highly correlated, or sometimes perfectly correlated, with the estimates of the variances the random model coefficients and then the mixed linear model program will not be able to provide unique estimates of the covariance structure parameters and variances of the random coefficients.

If research questions can be answered from tests or confidence intervals for the slope coefficients in the composite model, then a covariance pattern model would be preferred because the covariance pattern model does not make unrealistic assumptions about the structure of the prediction error covariance matrix. Alternatively, if certain research questions require an analysis of the variability of any random coefficient, then a random coefficient model is needed.

Example 3.2. The data in Example 3.1 were reanalyzed using a random coefficient model with a random y -intercept and a random slope for time (baseline centered). The random y -intercepts and slopes were assumed to be correlated. A first-order autoregressive prediction error structure was used. The 95% confidence interval for the training intervention effect is [3.78, 5.93] and is similar to the covariance pattern model result. The 95% confidence interval for the standard deviation of the y -intercepts is [6.95, 8.42], and the 95% confidence interval for the standard deviation of the slope coefficients for the time predictor variable is [0.45, 1.12]. We can be 95% confident that if all 32,000 students were measured on those four time periods and received training during the summer, their predicted time 1 cultural sensitivity scores (i.e., the y -intercepts) would have a standard deviation between 6.95 to 8.42 and their slopes would have a standard deviation between 0.45 to 1.12.

3.9 Parameter Estimation

A covariance pattern model and the composite random coefficient model can be expressed in matrix form for a random sample of n participants as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \quad (\text{Model 3.11})$$

where $\mathbf{y}$ is an $k \times 1$ vector of observations, $\mathbf{X}$ is a $k \times (s + 1)$ design matrix, $\boldsymbol{\beta}$ is a $(s + 1) \times 1$ vector of fixed parameters (containing one y -intercept and s slope coefficients), and $\mathbf{e}$ is $k \times 1$ vector of prediction errors. Model 3.11 assumes the data are in long format with k rows. The number of rows will equal $\sum_i^n r_i$ where r_i is the number of repeated measurements for participant i .

The prediction error covariance matrix in Model 3.11 is $cov(\mathbf{e}) = \mathbf{I}_n \otimes \mathbf{\Sigma}$. In a covariance pattern model, the researcher can specify the structure of $\mathbf{\Sigma}$. In a random coefficient model, $\mathbf{\Sigma}$ is determined by the variances and covariances of the random coefficients. Let $\hat{\mathbf{V}} = \mathbf{I}_n \otimes \hat{\mathbf{\Sigma}}$ where $\hat{\mathbf{\Sigma}}$ is a sample estimate of $\mathbf{\Sigma}$. An estimate of $\mathbf{\Sigma}$ is obtained by computing an OLS estimate of $\boldsymbol{\beta}$ (Equation 2.32 applied to data in long format) and the vector of estimated prediction errors using Equation 2.33. The sample variances and covariances are then computed from these estimated prediction errors. The method of *generalized least squares* (GLS) can be used to estimate $\boldsymbol{\beta}$ in Model 3.11. The GLS estimate of $\boldsymbol{\beta}$ is given below.

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y} \quad (3.3)$$

The GLS estimate of $\boldsymbol{\beta}$ can be used to obtain a revised estimate of the prediction errors and a revised estimate of $\mathbf{\Sigma}$ is then computed from these revised prediction error estimates. Equation 3.3 can be recomputed using the revised $\hat{\mathbf{V}} = \mathbf{I}_n \otimes \hat{\mathbf{\Sigma}}$. This process can be continued until the GLS estimate of $\boldsymbol{\beta}$ stabilizes. The stabilized GLS estimates are called *restricted maximum likelihood* (REML) or *maximum likelihood* (ML) estimates depending on how $\hat{\mathbf{\Sigma}}$ is computed. With REML, $\hat{\mathbf{\Sigma}}$ is computed using variances and covariances that are divided by $n - s - 1$, where s is the total number of predictor variables. With ML, $\hat{\mathbf{\Sigma}}$ is computed using variances and covariances that are divided by n . REML estimates are the default in most mixed linear model programs and are preferred to ML estimates. However, ML estimates are needed to perform nested model comparison hypothesis tests (see section 2.19) for two random coefficient models that have a different number of fixed effects but are assumed to have the same prediction error covariance structure.

The covariance matrix of $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ is

$$cov(\hat{\boldsymbol{\beta}}_{\text{GLS}}) = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1} \quad (3.4)$$

and the standard error of a particular slope estimate (β_k) is equal to the square root of the k^{th} diagonal element in Equation 3.4. Equation 3.4 applies to REML and ML estimates.

The REML/ML estimates are biased, and the bias can be substantial in small samples. The small-sample bias of the REML/ML estimates are usually larger when $\hat{\mathbf{V}}$ contains many estimated parameters (like an unstructured covariance matrix) rather than just a few estimated parameters (like a compound symmetric or AR-1 covariance matrix).

An approximate $100(1 - \alpha)\%$ confidence interval for β_k is given below.

$$\hat{\beta}_k \pm t_{\alpha/2;df} SE_{\hat{\beta}_k} \quad (3.5)$$

Equation 3.5 can be used to test a directional two-sided test as described in section 1.19. Equation 3.5 will perform properly if the bias in $\hat{\beta}_k$ is small and $\hat{\mathbf{V}}$ in Equation 3.4 closely approximates the true prediction error covariance matrix. In a covariance pattern model, an unstructured error matrix will closely approximate the true prediction error covariance matrix, but then $\hat{\beta}_k$ might have large bias. On the other hand, using a structured prediction error covariance matrix will reduce the bias in $\hat{\beta}_k$, but then $SE_{\hat{\beta}_k}$ might be too big or too small.

In a random coefficient model, the number of estimated parameters in $\hat{\mathbf{V}}$ is typically small so that $\hat{\beta}_k$ should have small bias, but $SE_{\hat{\beta}_k}$ is typically too big or too small because the prediction error covariance matrix in the composite model is seldom a good approximation to the true prediction error covariance matrix.

The confidence interval and hypothesis testing results for a covariance pattern or random coefficient model using REML/ML estimation can be misleading in small samples. An alternative method of estimation is the following OLS estimate of β

$$\hat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \quad (3.6)$$

with covariance matrix

$$cov(\hat{\beta}_{OLS}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad (3.7)$$

where $\hat{\Sigma}$ in $\hat{\mathbf{V}} = \mathbf{I}_n \otimes \hat{\Sigma}$ is an unstructured prediction error covariance matrix. Unlike the REML/ML estimates, the OLS estimate (Equation 3.6) is unbiased even in small samples. Equation 3.7 is referred to as a "cluster-robust" covariance matrix that accommodates unequal variances and correlations among the prediction errors. Unlike the standard errors from Equation 3.4, the standard errors from Equation 3.7 do not tend to be too big or too small.

Equation 3.6 is obtained by applying the linear model in Chapter 2 to repeated measures data that are in long format. Using Equation 3.6 and 3.7 can then be used in Equation 3.5 to obtain a confidence interval for each β_k . When an OLS estimate with a cluster-robust

standard error is used in Equation 3.5, the recommended value of df uses a Satterthwaite approximation (this approach can be implemented in R using the `lm` function along with the `conf_int` function in the `clubSandwich` package). A confidence interval for β_k using an OLS estimate, a cluster-robust standard error, and a Satterthwaite adjustment to the degrees of freedom should perform well if the sample size is several times larger than the number of repeated measurements.

3.10 Sample Size Planning

Practical sample size formulas for slope coefficients in covariance pattern models and random coefficient models are not available. Although sample size formulas for desired confidence interval precision or desired hypothesis testing power could be derived, they would require planning values that most researchers would have great difficulty specifying. For example, a planning value of the squared multiple correlation between the predictor variable of interest and all other predictor variables would be difficult to specify. Furthermore, small changes in this planning value will give large changes in the sample size requirement.

If sampling can be performed in two stages, then Equation 1.32 could be used to approximate the required number of participants in the second stage sample to obtain a desired confidence interval width for β_k , where β_k is a specific slope coefficient in a covariance pattern model or a random coefficient model. Equation 1.32 is typically applied to the most important slope coefficient.

3.11 Assumptions

Covariance Pattern Model

Hypothesis tests and confidence intervals in covariance pattern models assume the n participants have been randomly sampled from some population (the random sampling assumption) and the responses from one participant are uncorrelated with the responses of any other participant (the independence assumption). The multiple responses from each participant are allowed to be correlated. The prediction errors are assumed to have an approximate normal distribution in the population (the prediction error normality assumption).

When REML/ML is used with any type of structured prediction error covariance matrix (e.g., compound symmetric, autoregressive), the structured covariance matrix is assumed to closely approximate the true prediction error covariance matrix. This assumption is not relevant when OLS estimation with cluster-robust standard errors is used.

Like the GLM of Chapter 2, the covariance pattern model assumes constant variability of the prediction errors across the levels of every between-subjects predictor variable. In models that do not include quadratic effects, every predictor variable is assumed to be linearly related to the response variable. All of the issues regarding simultaneous confidence intervals and tests, exploratory model selection, influential observations, and multicollinearity in the GLM also apply to the covariance pattern model.

Random Coefficient Model

Like covariance pattern models, random coefficient models assume random sampling, independence among participants, and approximate prediction error normality. In a random coefficient model when REML/ML is used to estimate the fixed effects, the prediction error covariance matrix for the composite model is assumed to closely approximate the true composite prediction error covariance matrix.

REML is typically used to obtain estimates, tests, and confidence intervals for the variances of the random coefficients. Hypothesis tests and confidence intervals for the random coefficient variances assume that each random coefficient has an approximate normal distribution in the study population (the *random coefficient normality assumption*). Hypothesis tests and confidence intervals for the random coefficient variances are very sensitive to minor violations of the random coefficient normality assumption, and increasing the sample size will not mitigate the problem. A confidence interval for a random coefficient variance can have a true coverage probability that is far less than $1 - \alpha$ if the distribution of person-level coefficient values in the population is leptokurtic, regardless of sample size. Furthermore, a large sample size is often needed to obtain a usefully narrow confidence interval for the population variance of a random coefficient.

Like the GLM, the random coefficient model assumes constant variability of the prediction errors across the levels of every between-subjects predictor variable. In models that do not include quadratic effects, every predictor variable is assumed to be linearly related to the response variable. All of the issues regarding simultaneous confidence

intervals and tests, exploratory model selection, influential observations, and multicollinearity in the GLM also apply to the random coefficient model.

3.12 Data Visualization

Compared to the statistical models in Chapters 1 and 2, data visualization for the statistical models in this chapter is more complicated because the models can have both between-subjects and within-subjects predictor variables. Data visualization methods for between-subjects predictor variables require the data file to be converted from long format to wide format (SPSS and R have procedures that will perform this operation). With the data file in wide format, the response variable and every within-subjects predictor variable will appear as separate columns (variables) in the data file with each column corresponding to a specific time period. The next step is to compute the average of the response variable and the average of each within-subjects predictor variable across the repeated measurements. Now the data visualization methods described in Chapter 2 for the GLM can be applied to the average response variable, the average of each within-subjects predictor variable, and the between-subjects predictor variables. Some of these methods are illustrated below for 60 fifth-grade girls in Data Analysis Problem 3-3 where self-esteem is the response variable. Physical activity and program participation are within-subjects predictor variable.

A matrix scatterplot provides a compact description of the direction and magnitude of the bivariate relations among all of the variables. Additional information can be obtained from a semipartial correlation plot. With the data in wide format, the vertical axis of the semipartial plot is the response variable averaged over the repeated measurements. After controlling for all other predictor variables (see section 2.35), the horizontal axis can be a within-subjects predictor variable averaged over the repeated measurements or a between-subjects predictor variable. A semipartial correlation plot of average self-esteem with average physical performance (after removing the effect of program participation from average physical performance) is shown below.

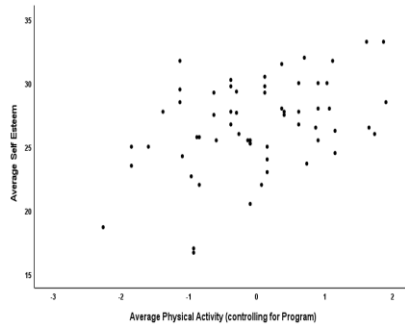


Figure 3.5 Example of a semipartial correlation plot

When one of the control variables in a semipartial correlation plot is a qualitative factor (e.g., program participation), a multiple-group scatterplot with lines of best fit provides useful information about the relation within each group. If the slopes of the lines are substantially different, this suggests a need to include an interaction effect in the model.

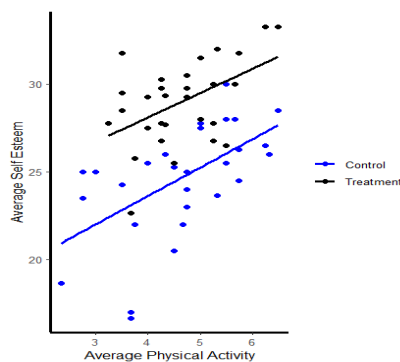


Figure 3.6 Example of a two-group scatterplot

Homoscedasticity and outliers can be visually assessed by a scatterplot with studentized deleted residuals on the vertical axis and a between-subjects predictor variable, an averaged within-subjects predictor variable, or the predicted average response variable on the horizontal axis as shown below.

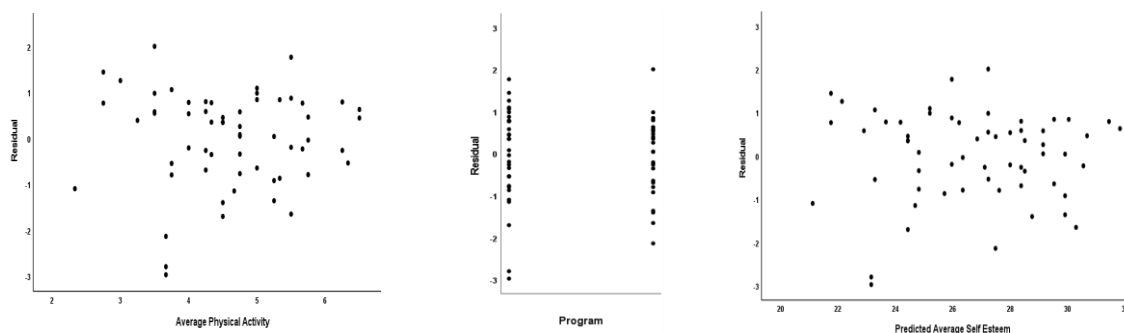
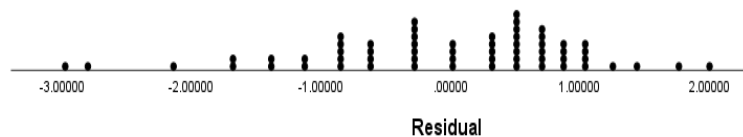
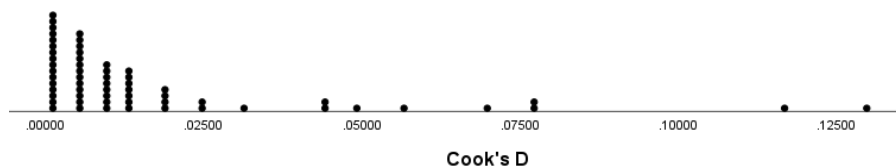


Figure 3.7 Examples of residual scatterplots

A histogram, box plot, or dot plot of studentized deleted residuals is also useful as shown below. In this example, three participants have residuals that are less than -2, and the data for these participants should be examined for possible data entry error or other problems.

**Figure 3.8** Dot plot of studentized deleted residuals

A box plot or dot plot of Cook's D values can help identify participants who have an unusually large influence on the parameter estimates. Participants with Cook's D values greater than $4/n$ ($4/60 = .067$ in this example) should be examined for possible data entry errors or other problems. In this example, the data for the five participants with the largest D values should be examined for possible data entry error or other problems.

**Figure 3.9** Dot plot of Cook's D values

To assess relations between the response variable and the within-subjects predictor variables, the data file must be in long format. Then these relations can be visually assessed using scatterplots. Several graphs are described below for Data Analysis Problem 3-2, where the social skills of six children in a group setting are measured each week for 18 weeks. Gender and vocabulary in week 1 are between-subjects predictor variables. Group size and week are within-subjects predictor variables. Scatterplots of social skills with the two within-subjects predictor variables are shown below.

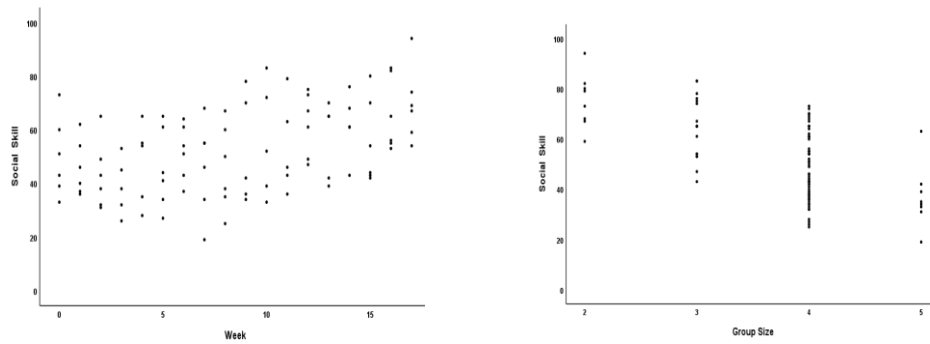


Figure 3.10 Scatterplots for two within-subjects predictor variables

In a covariance pattern or random coefficient model, the slope coefficient for group size will describe the relation between social skill and the component of group size that is not linearly related to the other predictor variables (i.e., gender, vocabulary, and week). A semipartial correlation plot for group size shown below and illustrates the direction and strength of the relation between social skill and group size controlling for gender, vocabulary, and week.

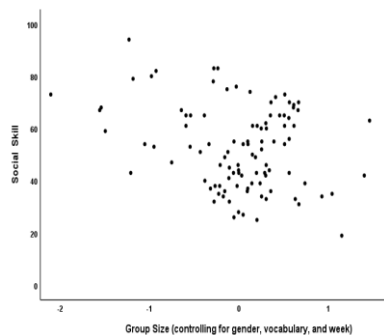


Figure 3.11 Semipartial correlation plot

If there are any between-subject qualitative factors (e.g., gender, treatment), the relation between the repeated y scores and a within-subjects predictor variable can be assessed using a multiple-group scatter plot. An alternative to the above semipartial correlation plot is the following semipartial correlation plot which shows the relation between social skill and group size for boys and girls separately after controlling for vocabulary size and week.

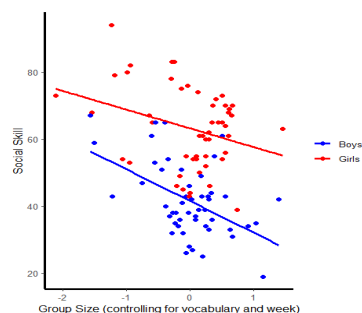


Figure 3.12 Two-group semipartial correlation plot

To assess the homoscedasticity assumption, studentized deleted residuals can be plotted with each predictor variable and the predicted response variable as shown below.

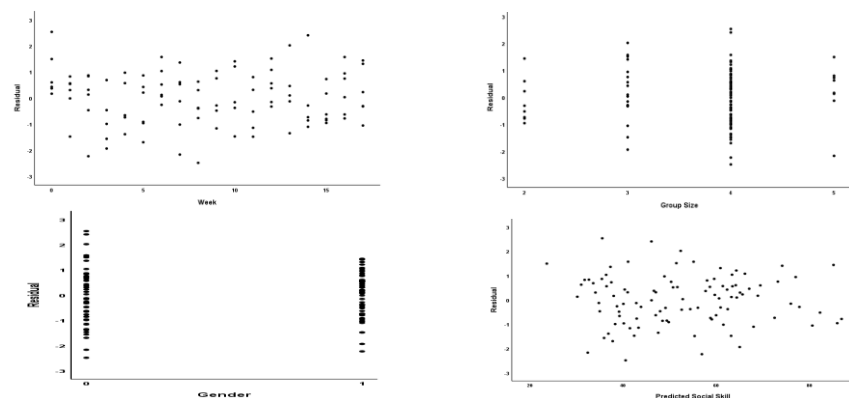


Figure 3.13 Residual scatterplots

A box plot or dot plot of studentized deleted residuals can identify participants who have unusually large residuals. In the dot plot below, six participants have studentized deleted residuals that are less than -2 or greater than 2, and the data for these participants should be examined for possible data entry errors or other problems.

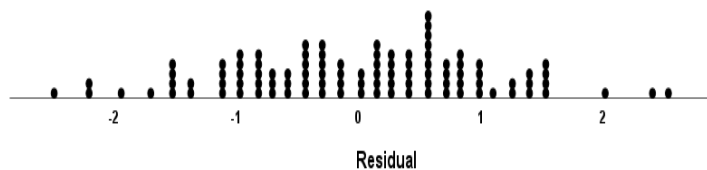


Figure 3.14 Dot plot of studentized deleted residuals

When Cook's D is computed from data in long format, any response that has a D -value greater than $4/(n \times r)$, where r is the number of repeated measurements, should be examined. In the following example, Cook's D was computed for the $r = 18$ repeated measurements of the $n = 6$ children, and we should examine a response at any time point

that has a D -value greater than $4/(6 \times 18) = .037$ for possible data entry errors or other problems. In this example, the data for the participants with the five or six largest D -values should be examined.

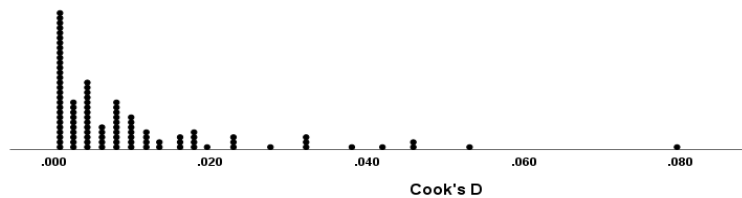


Figure 3.15 Dot plot of Cook's D values

The studentized deleted residuals and Cook's D values are computed from OLS estimates using multiple regression software such as the `lm` function in R or the Linear Regression option in SPSS. As explained in section 3.9, the OLS estimates are unbiased and have good small-sample properties.

In addition to graphs described above, the relation between the response variable and each within-subjects predictor variable can be examined using a scatterplot for each subject. These subject-level scatterplots can reveal outliers and other problems that could go unnoticed in the above graphs. Scatterplots showing the relation between social skill and time (in weeks) are given below for all six children in the study.

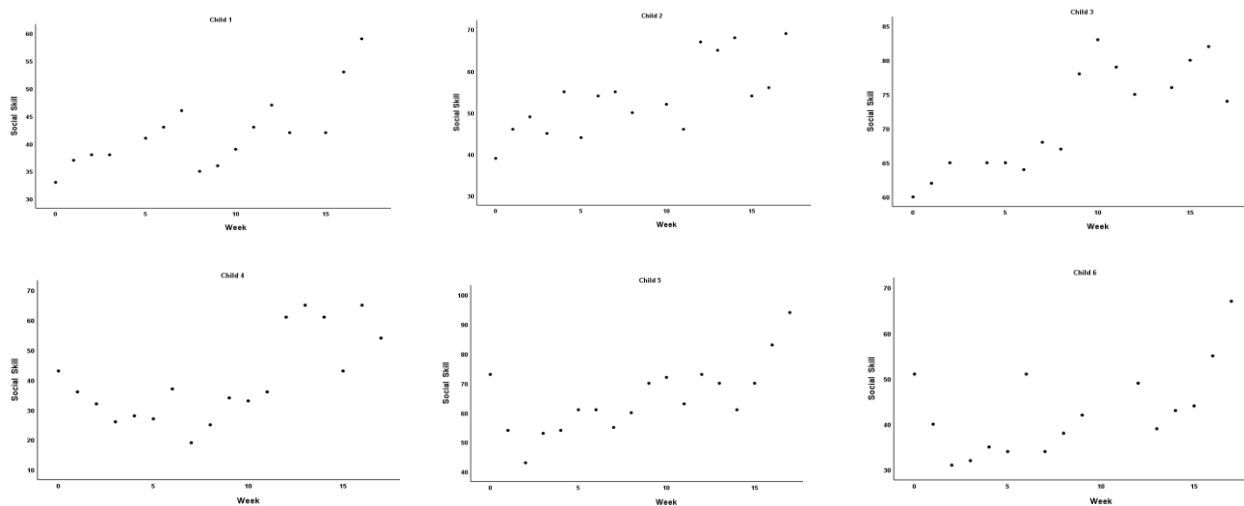


Figure 3.16 Scatterplots for six participants

The data screening methods described in section 1.34 of Volume 1 can be tedious and time-consuming with repeated measurements of multiple variables. If any variable is a

score on a multi-item questionnaire, then individual item responses should also be screened. The data visualization methods proposed here can reduce data screening time and effort by highlighting a small subset of participants who appear to have anomalous data.

Data visualization should not be used to justify the automatic discarding of participants who have large studentized deleted residuals or large Cook's D values. Likewise, participants who appear to be bivariate outliers should not automatically be discarded. If the data screening methods in section 1.34 of Volume 1 do not reveal any data entry errors or other problems for the participants in question, discarding these outlier participants is not recommended. A common strategy involves analyzing the data with and then without the outlier participants. Ideally the main conclusions will be the same in the two analyses; otherwise, it will be necessary to report the results from both analyses.

A common mistake in data visualization is to over-interpret visual indications of outliers, nonlinearity, heteroscedasticity, and nonnormality. Graphs contain random sampling error, and this sampling error can produce patterns in the sample data that appear to deviate from the ideal patterns. Even if all assumptions have been perfectly satisfied, the graphs will show false signs of outliers, nonlinearity, and heteroscedasticity. These false signs are common when the sample size is small. Unless the sample size is large, only the most obvious problems need to be investigated.

Although the data visualization methods presented here are recommended as a way to detect and correct problems with the data, these methods can also result in unintended semi-exploratory model modifications that can lead to misleading hypothesis testing and confidence interval results (see section 2.25). For example, a preliminary examination of the scatterplots might suggest that one or more predictor variables are not useful, and the researcher decides to exclude these variables from the model without disclosing this decision when presenting the results. Or the researcher might decide to include or exclude interaction effects based on a preliminary screening of the data. One way to avoid undisclosed semi-exploratory model modifications is to preregister the data analysis plans prior to data collection. The preregistration report will include a description of the response variable, all predictor variables, and all interaction effects that will be included in the analysis.

3.13 Bayesian Approach to Data Analysis

As in a GLM, specifying an informative prior distribution for β_k in a covariance pattern model or random coefficient model is extremely difficult because the value of β_k is very sensitive to the correlations among the predictor variables. Researchers are often surprised that an estimate of β_k is very different and sometimes has the opposite sign from what they had expected. Bayesian methods have focused more on the random coefficient model than the covariance pattern model. As in a GLM, a typical Bayesian analysis of a random coefficient model will use a non-informative prior for each β_k . A non-informative prior for β_k would typically be a normal distribution with a mean of 0 and a very large standard deviation so that the estimate of β_k will be determined only by the sample data. However, a Bayesian analysis can also impose informative prior information on the variances of the random components, which can result in credible intervals for the variances that are substantially narrower than the classical confidence intervals.

Most applications of a covariance pattern model or a random coefficient model involves semi-exploratory model modifications as a result of a preliminary screening of the data, the exploratory selection of a prediction error covariance structure in the covariance pattern model, an exploratory selection of random slopes in a random coefficient model, and the removal of predictor variables that do not have p -values less than .05. The final hypothesis testing and confidence interval results from these semi-exploratory analyses can be misleading: the estimate of each β_k is inflated, the p -values are too small, and the confidence intervals are too narrow. The use of skeptical prior for β_k will shrink β_k towards 0 and can give more realistic results in a semi-exploratory analysis. A skeptical prior for each β_k is easy to implement in a covariance pattern model or a random coefficient model if the researcher can specify a value for β_k that represents a weak association. The `ci.bayes.normal` function can be used to compute a credible interval for β_k with a skeptical prior as explained in section 2.36.

Example 3.3. In *Example 3.2*, a random coefficient model was used to estimate the effect of a cultural sensitivity training program. The estimated program effect was 5.48 with a standard error of 0.548. The 95% confidence interval for the program effect was [3.78, 5.93]. These results might be optimistic due to some exploratory model modifications. The researcher believes that a 2-point difference in means (training vs. control) is a small effect. If a skeptic is 95% confident that the population mean difference is between -2 and 2, the mean of the prior normal distribution can

be set to 0 with a standard deviation of $2/1.96 = 1.02$. The resulting posterior mean for the training effect is 3.77 with a 95% credible interval of [2.83, 4.72].

3.14 Using Information from a Prior Study

Published results from covariance pattern or random coefficient models will likely have used semi-exploratory model modifications that tend to exaggerate the importance of the predictor variables. Also, the review process for most psychology journals still focuses on p -values. Tests for slope coefficients that have p -values slightly less than .05 will have confidence intervals (if reported) that barely exclude 0 and could be uselessly wide. For these reasons, it is important to replicate covariance pattern or random coefficient model results in a new sample. In the absence of non-replication evidence (see Chapter 4 of Volume 5 for more details), it is appropriate to combine the results of a prior study with a follow-up study. The confidence intervals from the combined studies can be substantially narrower than those reported in the original study, and the combined results describe the study population from both the original study and the follow-up study. The `replicate.slope` function in the `vcmeta` package can be used to compare and combine a slope coefficient from a covariance pattern or random coefficient model in an original study with the corresponding slope coefficient in a follow-up study.

Example 3.4. A researcher has replicated the study in *Example 3.1*. In the original study of 100 college students, which used a covariance pattern model, the estimated program effect was 4.48 with a standard error of 0.439. The 95% confidence interval was [3.62, 5.34]. In the follow-up study of 250 college students at another university, the estimated program effect was 3.86 with a standard error of 0.289. The 90% confidence interval for $\beta_{1[1]} - \beta_{1[2]}$ is [-0.34, 1.39], which includes 0 and justifies combining the slopes from the original and follow-up studies. The estimate of $(\beta_{1[1]} + \beta_{1[2]})/2$ is .422 with a 95% confidence interval of [3.70, 4.74]. This confidence interval is narrower than the confidence interval reported in the original study, and this confidence interval describes the study populations at both universities.

Key Terms

between-subjects (time-invariant) predictor variables
 within-subjects (time-varying) predictor variables
 covariance pattern model
 prediction error covariance matrix
 unstructured covariance matrix
 compound-symmetric covariance matrix

first-order autoregressive covariance matrix (AR-1)
 Toeplitz covariance matrix
 baseline centering
 person centering
 random coefficient model (multilevel model)
 level-1 model
 level-2 model
 composite model
 generalized least squares estimate
 random coefficient normality assumption
 AIC
 cluster-robust standard errors

Concept Questions

1. What are the main differences between a covariance pattern model and a random coefficient model?
2. When would a random coefficient model be preferred to a covariance pattern model?
3. When would a covariance pattern model be preferred to a random coefficient model?
4. What are the advantages and disadvantages of using an unstructured prediction error covariance matrix in a covariance pattern model?
5. What are the advantages and disadvantages of using a structured prediction error covariance matrix in a covariance pattern model?
6. What is the structure of each block of the composite prediction error covariance matrix if a random coefficient model has a random y -intercept but no random slopes?
7. If a predictor variable is used in the level-2 model for a random slope, how will that predictor variable appear in the composite model?
8. What is a possible negative consequence of treating a slope coefficient as random rather than fixed on the composite prediction error covariance matrix?
9. How can a quadratic time trend be modeled?

10. How can a 3-level within-subjects treatment factor be modeled in a covariance pattern model or a random coefficient model?
11. When would a lagged predictor variable be used?
12. What covariance structures should be considered in a within-subjects experiment?
13. What covariance structures should be considered in a longitudinal design?
14. How would an interaction between two within-subjects predictor variables be modeled in a random coefficient model?
15. How would an interaction between two between-subjects predictor variables be modeled in a random coefficient model?
16. What are the advantages of using a covariance pattern model to analyze a within-subjects experiment over the traditional methods?
17. What are the recommended types of centering for a "time" variable?
18. What is the recommended type of centering for a within-subjects predictor variable?
19. If the variance of a random slope for predictor variable x_1 is large, what does this suggest?
20. What is the effect of violating the random coefficient normality assumption on confidence intervals for the parameter variances?
21. Explain the difference between a wide data format and a long data format.

Data Analysis Problems

3-1. Twenty employees were randomly selected from a Silicon Valley employee database of about 14,000 employees. The twenty employees agreed to fill out a mood questionnaire on Monday, Wednesday, and Friday of a particular week. Day of week was baseline centered (Monday = 0, Wednesday = 2, Friday = 4). The `HW3-1.sav` file contains the sample data with variable names `employee`, `mood`, and `DOW`.

- a) Describe the study population.

b) Use R to compute a 95% confidence interval for the population slope using OLS estimation with cluster-robust standard errors and Satterthwaite degrees of freedom. Interpret the results. Also, report the t -value, degrees of freedom, and p -value.

c) Use a random coefficient model with scaled identity level-1 prediction error covariance structure, a random y -intercept, and a random slope. Allow the random y -intercept and slope parameters to be correlated. Compute 95% confidence intervals for the standard deviations of the random y -intercept and slope coefficients. Interpret the results.

3-2. A random sample of six preschool children were obtained from a listing of about 500 preschool children in San Jose whose parents agreed to let their children participate in certain types of educational studies. A social skill assessment was given to each of the six children while interacting with other children in a group each week for 18 consecutive weeks. The size of the group was used as a within-subjects covariate. For some children, social skill was not measured for all 18 weeks due to illness. Vocabulary scores in week 1 and the child's gender were used as between-subjects predictor variables. Week was baseline centered. The HW3-2.sav file contains the sample data with variable names `child`, `vocab`, `sex`, `grpsize`, `week`, and `socskill` (`sex` is dummy coded girl = 1 and boy = 0). The person-centered group size (`cengrpsize`) and mean group size per student (`mgrpssize`) are also included in the data file.

a) Describe the study population.

b) Use R to compute 95% confidence intervals for the population slopes using OLS estimation with cluster-robust standard errors and Satterthwaite degrees of freedom. Interpret the results. Also report the t -values, degrees of freedom, and p -values.

c) Use a random coefficient model with scaled identity level-1 prediction error covariance structure, a random y -intercept, and a random slope for week. Allow the random y -intercept and slope parameters to be correlated. Compute 95% confidence intervals for the standard deviations of the random y -intercept and slope coefficients. Interpret the results.

3-3. A random sample of 60 fifth grade girls was obtained from a Los Angeles school district and then randomly divided into two groups of equal size. One group served as the control and the other group was given a new mentoring program in the seventh and eighth grades designed to build self-esteem in adolescent girls. Self-esteem was measured at the end of grades 5, 6, 7, and 8. Not all girls were measured in all four years. A measure of physical activity was also measured each year as a within-subjects predictor variable. The HW3-3.sav file contains the sample data with variable names `girl`, `Program`,

PhysAct, Grade, and SelfEst (Program is dummy coded with new program = 1 and control = 0).

- a) Describe the study population.
- b) Use R to compute 95% confidence intervals for the population slopes using OLS estimation with cluster-robust standard errors and Satterthwaite degrees of freedom. Interpret the results. Also report the t -values, degrees of freedom, and p -values.
- c) Use a random coefficient model with a scaled identity level-1 prediction error covariance structure, a random y -intercept, and a random slope for grade. Allow the random y -intercept and slope parameters to be correlated. Report the p -values for the test of the y -intercept variance equal to zero and the test of the random slope variance equal to zero. If the p -value for a hypothesis test is greater than .05, then do not report the 95% confidence interval for that variance.

3-4. Previous research has shown that scores on a driving simulator are lower when the driver uses a cell phone. A new study examined the effects of traffic conditions and the effects of sending or receiving a call on driving performance. Twelve participants were randomly sampled from a volunteer pool of 2,560 cell phone using undergraduate students and were tested in a driving simulator under two conditions. Each the participant was required to send a call and also answer a call on their own cell phone. Driving performance scores (higher scores represent better performance) were obtained for 2-minute periods following the sending and receiving of calls. The total amount of time spent sending and receiving a call was used as a between-subjects covariate. The HW3-4 .sav file contains the sample data with variable names `student`, `SvsR`, `time`, and `score`. The `SvsR` variable is dummy coded with `send` = 1 and `receive` = 0.

- a) Describe the study population.
- b) Use R to compute 95% confidence intervals for the population slopes using OLS estimation with cluster-robust standard errors and Satterthwaite degrees of freedom. Interpret the results. Also report the t -values, degrees of freedom, and p -values.

Suggested Readings

Cohen, J., Cohen, P. West, S. G., & Aiken (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Erlbaum. **Section 15.4**

Fox, J. (2016). *Applied regression analysis and generalized linear models* (3rd ed). Sage.

Chapter 13

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. Wiley.

Chapters 7 and 8

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Wiley. **Chapter 6**

Singer, J. D. & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford. **Chapters 1 - 7**

Appendix A. Tables

Table 1a Two-sided critical z-values ($z_{\alpha/2}$)

$1 - \alpha$				
.80	.90	.95	.99	.999
1.28	1.65	1.96	2.58	3.29

Table 1b One-sided critical z-values (z_{α})

$1 - \alpha$

.80	.90	.95	.99	.999
0.84	1.28	1.65	2.33	3.09

R Functions

Use `qnorm(1 -  $\alpha/2$ )` for 2-sided critical z-value.

Example: $\alpha = .005$

```
qnorm(1 - .005/2)
[1] 2.807034
```

Use `qnorm(1 -  $\alpha$ )` for 1-sided critical z-value.

Example: $\beta = .25$

```
qnorm(1 - .25)
[1] 0.6744898
```

Use `2*(1 - pnorm(abs(z)))` to compute 2-sided p -value for z statistic.

Example: $z = 2.32$

```
2*(1 - pnorm(2.32))
[1] 0.02034088
```

Table 2 Two-sided critical t -values ($t_{\alpha/2;df}$)

	1 - α			
df	.90	.95	.99	.999
1	6.31	12.71	63.66	636.62
2	2.92	4.30	9.93	31.60
3	2.35	3.18	5.84	12.92
4	2.13	2.78	4.60	8.61
5	2.02	2.57	4.03	6.86
6	1.94	2.45	3.71	5.96
7	1.90	2.37	3.50	5.41
8	1.86	2.31	3.36	5.04
9	1.83	2.26	3.25	4.78
10	1.81	2.23	3.17	4.59
11	1.80	2.20	3.11	4.48
12	1.78	2.18	3.06	4.32
13	1.77	2.16	3.01	4.22
14	1.76	2.15	2.98	4.14

15	1.75	2.13	2.95	4.07
16	1.75	2.12	2.92	4.02
17	1.74	2.11	2.90	3.97
18	1.73	2.10	2.88	3.92
19	1.73	2.09	2.86	3.88
20	1.73	2.09	2.85	3.85
25	1.71	2.06	2.79	3.73
30	1.70	2.04	2.75	3.65
35	1.69	2.03	2.72	3.59
40	1.68	2.02	2.70	3.55
60	1.67	2.00	2.66	3.46
∞	1.65	1.96	2.58	3.29

R Functions

Use `qt(1 -  $\alpha$ /2, df)` for 2-sided critical t -value.

Example: $\alpha = .05$ and $df = 22$.

```
qt(1 - .05/2, 22)
[1] 2.073873
```

Use `2*(1 - pt(abs(t), df))` to compute 2-sided p -value for t statistic.

Example: $t = -2.43$ and $df = 9$.

```
2*(1 - pt(2.43, 9))
[1] 0.03798256
```

Table 3 Two-sided Bonferroni critical z -values ($z_{\alpha^*/2}$)

Simultaneous Confidence Level	Number of Confidence Intervals								
	2	3	4	5	6	7	8	9	10
.90	1.96	2.13	2.24	2.31	2.40	2.45	2.50	2.54	2.58
.95	2.24	2.39	2.50	2.58	2.64	2.69	2.74	2.77	2.81
.99	2.81	2.94	3.02	3.09	3.15	3.19	3.23	3.26	3.39

R Function

Use `qnorm(1 -  $\alpha/(2 * g)$ )` for 2-sided critical z-value and g simultaneous CIs.

Example: $\alpha = .05$ and $g = 4$.

```
qnorm(1 - .05/(2*4))
[1] 2.497705
```

Two-sided Bonferroni critical t -values ($t_{\alpha/2;df}$)

R Function

Use `qt(1 -  $\alpha/(2 * g)$ )` for 2-sided critical t -value and g simultaneous CIs.

Example: $\alpha = .05$, $g = 4$, and $df = 15$

```
qt(1 - .05/(2*4), 15)
2.836627
```

Table 4 Critical F values ($\alpha = .05$, upper tail)

df_{error}	df_{effect}			
	1	2	3	4
2	18.51	19.00	19.16	19.25
3	10.13	9.55	9.28	9.12
4	7.71	6.94	6.59	6.39
5	6.61	5.79	5.41	5.19
6	5.99	5.14	4.76	4.53
7	5.59	4.74	4.35	4.12
8	5.32	4.46	4.07	3.84
9	5.12	4.26	3.86	3.63
10	4.96	4.10	3.71	3.48
11	4.84	3.98	3.59	3.36
12	4.75	3.89	3.49	3.26
13	4.67	3.81	3.41	3.18

14	4.60	3.74	3.34	3.11
15	4.54	3.68	3.29	3.06
16	4.49	3.63	3.24	3.01
17	4.45	3.59	3.20	2.96
18	4.41	3.55	3.16	2.93
19	4.38	3.52	3.13	2.90
20	4.35	3.49	3.10	2.87
∞	3.84	3.00	2.61	2.37

R Functions

Use `qf(1 -  $\alpha$ , df1, df2)` for critical F value.

Example: $\alpha = .01$, $df1 = 2$, and $df2 = 27$.

```
qf(1 - .01, 2, 27)
[1] 5.488118
```

Use `1 - pf(f, df1, df2)` to compute p -value for F statistic.

Example: $f = 4.74$, $df1 = 2$, and $df2 = 10$.

```
1 - pf(4.74, 2, 10)
[1] 0.03564965
```

Table 5 Fisher correlation transformation

ρ_{yx}	ρ_{yx}^*	ρ_{yx}	ρ_{yx}^*	ρ_{yx}	ρ_{yx}^*	ρ_{yx}	ρ_{yx}^*	ρ_{yx}	ρ_{yx}^*
0.000	0.000	0.200	0.203	0.400	0.424	0.600	0.693	0.800	1.099
0.005	0.005	0.205	0.208	0.405	0.430	0.605	0.701	0.805	1.113
0.010	0.010	0.210	0.213	0.410	0.436	0.610	0.709	0.810	1.127
0.015	0.015	0.215	0.218	0.415	0.442	0.615	0.717	0.815	1.142
0.020	0.020	0.220	0.224	0.420	0.448	0.620	0.725	0.820	1.157
0.025	0.025	0.225	0.229	0.425	0.454	0.625	0.733	0.825	1.172
0.030	0.030	0.230	0.234	0.430	0.460	0.630	0.741	0.830	1.188
0.035	0.035	0.235	0.239	0.435	0.466	0.635	0.750	0.835	1.204
0.040	0.040	0.240	0.245	0.440	0.472	0.640	0.758	0.840	1.221
0.045	0.045	0.245	0.250	0.445	0.478	0.645	0.767	0.845	1.238
0.050	0.050	0.250	0.255	0.450	0.485	0.650	0.775	0.850	1.256
0.055	0.055	0.255	0.261	0.455	0.491	0.655	0.784	0.855	1.274
0.060	0.060	0.260	0.266	0.460	0.497	0.660	0.793	0.860	1.293
0.065	0.065	0.265	0.271	0.465	0.504	0.665	0.802	0.865	1.313
0.070	0.070	0.270	0.277	0.470	0.510	0.670	0.811	0.870	1.333
0.075	0.075	0.275	0.282	0.475	0.517	0.675	0.820	0.875	1.354
0.080	0.080	0.280	0.288	0.480	0.523	0.680	0.829	0.880	1.376
0.085	0.085	0.285	0.293	0.485	0.530	0.685	0.838	0.885	1.398

0.090	0.090	0.290	0.299	0.490	0.536	0.690	0.848	0.890	1.422
0.095	0.095	0.295	0.304	0.495	0.543	0.695	0.858	0.895	1.447
0.100	0.100	0.300	0.310	0.500	0.549	0.700	0.867	0.900	1.472
0.105	0.105	0.305	0.315	0.505	0.556	0.705	0.877	0.905	1.499
0.110	0.110	0.310	0.321	0.510	0.563	0.710	0.887	0.910	1.528
0.115	0.116	0.315	0.326	0.515	0.570	0.715	0.897	0.915	1.557
0.120	0.121	0.320	0.332	0.520	0.576	0.720	0.908	0.920	1.589
0.125	0.126	0.325	0.337	0.525	0.583	0.725	0.918	0.925	1.623
0.130	0.131	0.330	0.343	0.530	0.590	0.730	0.929	0.930	1.658
0.135	0.136	0.335	0.348	0.535	0.597	0.735	0.940	0.935	1.697
0.140	0.141	0.340	0.354	0.540	0.604	0.740	0.950	0.940	1.738
0.145	0.146	0.345	0.360	0.545	0.611	0.745	0.962	0.945	1.783
0.150	0.151	0.350	0.365	0.550	0.618	0.750	0.973	0.950	1.832
0.155	0.156	0.355	0.371	0.555	0.626	0.755	0.984	0.955	1.886
0.160	0.161	0.360	0.377	0.560	0.633	0.760	0.996	0.960	1.946
0.165	0.167	0.365	0.383	0.565	0.640	0.765	1.008	0.965	2.014
0.170	0.172	0.370	0.388	0.570	0.648	0.770	1.020	0.970	2.092
0.175	0.177	0.375	0.394	0.575	0.655	0.775	1.033	0.975	2.185
0.180	0.182	0.380	0.400	0.580	0.662	0.780	1.045	0.980	2.298
0.185	0.187	0.385	0.406	0.585	0.670	0.785	1.058	0.985	2.443
0.190	0.192	0.390	0.412	0.590	0.678	0.790	1.071	0.990	2.647
0.195	0.198	0.395	0.418	0.595	0.685	0.795	1.085	0.995	2.994

$$\rho_{yx}^* = \ln[(1 + \rho_{yx}) / (1 - \rho_{yx})] / 2$$

$$\rho_{yx} = [\exp(2\rho_{yx}^*) - 1] / [\exp(2\rho_{yx}^*) + 1]$$

R Computations

```
r = .6
log((1 + r)/(1 - r))/2
[1] 0.6931472
```

```
f = .6931472
(exp(2*f) - 1)/(exp(2*f) + 1)
[1] 0.6
```

Appendix B. Matrix Notation and Operations

B.1. Matrix Notation

An $r \times c$ *matrix* is a rectangular array of elements with r rows and c columns. An $r \times c$ matrix is said to be of *order* $r \times c$. A matrix is usually denoted by a capital letter printed in a boldface font (e.g., **A**, **B**, **X**). The elements of the matrix are represented by lower case letters with a double subscript (e.g., a_{jk} , b_{jk} , x_{jk}). For example, in the matrix **X**, x_{13} is the element in the first row and the third column. A 4×3 matrix **X** is shown below.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \end{bmatrix}$$

A matrix with a single row is called a *row vector* and a matrix with a single column is called a *column vector*. Vectors are usually represented by lower case letters printed in a boldface font (e.g., **a**, **b**, **x**). The elements of the vector are represented by lower case letters with a single subscript (e.g., a_j , b_j , x_j). A 3×1 column vector **y** and a 1×4 row vector **h** are shown below.

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

$$\mathbf{h} = [h_1 \ h_2 \ h_3 \ h_4].$$

It is sometimes necessary to refer to a particular row or column of a matrix. These row or column vectors are represented by a subscripted lower case letter in a boldface font. For instance, the j^{th} row vector in the above 4×3 matrix **X** would be noted as **x_j** .

A *square matrix* has the same number of rows as columns. A square matrix where the *diagonal elements* (the elements where the two subscripts are equal) are nonzero and the *off-diagonal elements* (the elements where the two subscripts are not equal) are zero is called a *diagonal matrix*. A 3×3 diagonal matrix **D** is shown below.

$$\mathbf{D} = \begin{bmatrix} d_{11} & 0 & 0 \\ 0 & d_{22} & 0 \\ 0 & 0 & d_{33} \end{bmatrix}$$

The *identity matrix* is a special type of diagonal matrix where all diagonal elements are equal to 1. The identity matrix is usually represented as **I** or **I_n** where n is the order of the identity matrix.

A square matrix where the jk^{th} element is equal to the kj^{th} element is called a *symmetric matrix*. A symmetric 3×3 matrix is shown below.

$$\mathbf{S} = \begin{bmatrix} 14 & 5 & 2 \\ 5 & 20 & 8 \\ 2 & 8 & 11 \end{bmatrix}$$

A *one vector* is a row or column vector in which every element is equal to 1 and is represented as the number one printed in a boldface font. A 1×3 one vector is shown below.

$$\mathbf{1} = [1 \quad 1 \quad 1]$$

B.2. Matrix Operations

The *transpose* of a matrix $\mathbf{X}$ is represented as $\mathbf{X}'$ (or $\mathbf{X}^T$). The transpose of a matrix is obtained by interchanging the rows and columns of the matrix. For example, if

$$\mathbf{X} = \begin{bmatrix} 4 & 6 \\ 7 & 1 \\ 3 & 9 \end{bmatrix}$$

then

$$\mathbf{X}' = \begin{bmatrix} 4 & 7 & 3 \\ 6 & 1 & 9 \end{bmatrix}.$$

Note that the jk^{th} element in $\mathbf{X}$ is equal to the kj^{th} element in $\mathbf{X}'$. Most vectors in statistical formulas are assumed to be column vectors. Row vectors, when needed, are obtained by taking the transpose of a column vector.

If two matrices $\mathbf{A}$ and $\mathbf{B}$ are of the same order, the two matrices are then *conformable for addition or subtraction*, and $\mathbf{A} + \mathbf{B}$ is a matrix with element $a_{jk} + b_{jk}$ in the j^{th} row and the k^{th} column, as illustrated below for the sum of two 3×2 matrices.

$$\begin{aligned} \mathbf{A} + \mathbf{B} &= \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & a_{13} + b_{13} \\ a_{21} + b_{21} & a_{22} + b_{22} & a_{23} + b_{23} \end{bmatrix} \end{aligned}$$

Likewise, $\mathbf{A} - \mathbf{B}$ is a matrix with element $a_{jk} - b_{jk}$ in the j^{th} row and the k^{th} column, as illustrated below for the difference of two 3×2 matrices.

$$\begin{aligned} \mathbf{A} - \mathbf{B} &= \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} - \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} - b_{11} & a_{12} - b_{12} & a_{13} - b_{13} \\ a_{21} - b_{21} & a_{22} - b_{22} & a_{23} - b_{23} \end{bmatrix} \end{aligned}$$

To multiply a matrix by a *scalar* (i.e., a single number), simply multiply each element in the matrix by the scalar. Scalars are represented by italicized lower case letters in a non-boldface font. To illustrate, if $b = 2$ and

$$\mathbf{A} = \begin{bmatrix} 4 & 7 & 3 \\ 6 & 1 & 9 \end{bmatrix}$$

then

$$b\mathbf{A} = \begin{bmatrix} 8 & 14 & 6 \\ 12 & 2 & 18 \end{bmatrix}.$$

Some statistical formulas involve the subtraction of a scalar from an $n \times 1$ vector. The result is obtained by first multiplying the scalar by an $n \times 1$ one vector and then taking the difference of the two vectors. For example, if $\mathbf{y}$ is a 3×1 vector and m is a scalar, then $\mathbf{y} - m$ is

$$\mathbf{y} - m\mathbf{1} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} - m \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} y_1 - m \\ y_2 - m \\ y_3 - m \end{bmatrix}$$

The *dot product* of a $n \times 1$ vector $\mathbf{a}$ with a $n \times 1$ vector $\mathbf{b}$ is

$$\mathbf{a}'\mathbf{b} = a_1b_1 + a_2b_2 + \dots + a_nb_n$$

Note that $\mathbf{a}'\mathbf{b} = \mathbf{b}'\mathbf{a}$. For instance, if $\mathbf{a}' = [4 \ 3 \ 2]$ and $\mathbf{b}' = [6 \ 1 \ 4]$, then $\mathbf{a}'\mathbf{b} = 4(6) + 3(1) + 2(4) = 35$.

Two $n \times 1$ vectors, $\mathbf{a}$ and $\mathbf{b}$, are said to be *orthogonal* if $\mathbf{a}'\mathbf{b} = 0$. For instance, if $\mathbf{a}' = [.5 \ .5 \ -1]$ and $\mathbf{b}' = [1 \ -1 \ 0]$, then $\mathbf{a}$ and $\mathbf{b}$ are orthogonal because $\mathbf{a}'\mathbf{b} = (.5)(1) + (.5)(-1) + (-1)(0) = 0$. A vector $\mathbf{b}$ is *normalized* if $\mathbf{b}'\mathbf{b} = 1$. Two vectors are *orthonormal* if they are normalized and orthogonal.

Two matrices $\mathbf{A}$ and $\mathbf{B}$ can be multiplied if they are *conformable for multiplication*. To compute the matrix product $\mathbf{AB}$, the number of columns of $\mathbf{A}$ must equal the number of rows of $\mathbf{B}$. In general, if $\mathbf{A}$ is $r \times n$ and $\mathbf{B}$ is $n \times c$, then the matrix product $\mathbf{AB}$ is an $r \times c$ matrix. The jk^{th} element in the $r \times c$ product matrix is equal to the dot product $\mathbf{a}_j\mathbf{b}_k$ where $\mathbf{a}_j$ is the j^{th} row vector of matrix $\mathbf{A}$ and $\mathbf{b}_k$ is the k^{th} column vector of matrix $\mathbf{B}$. For instance,

the matrices **A** and **B** shown below are conformable for computing the product **AB** because **A** is 2×3 and **B** is 3×4 so that the product will be a 2×4 matrix.

$$\mathbf{A} = \begin{bmatrix} 4 & 7 & 3 \\ 6 & 1 & 9 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 1 & 2 & 1 & 4 \\ 5 & 4 & 3 & 1 \\ 4 & 2 & 3 & 2 \end{bmatrix}$$

Each of the $2 \times 4 = 8$ elements of the **AB** matrix is a dot product. For example, the element in row 1 and column 1 of the product **AB** is

$$[4 \ 7 \ 3] \begin{bmatrix} 1 \\ 5 \\ 4 \end{bmatrix} = 4(1) + 7(5) + 3(4) = 51$$

and the element in row 2 and column 3 of **AB** is

$$[6 \ 1 \ 9] \begin{bmatrix} 1 \\ 3 \\ 3 \end{bmatrix} = 6(1) + 1(3) + 9(3) = 36.$$

After computing all 8 dot products, the following result is obtained.

$$\mathbf{AB} = \begin{bmatrix} 51 & 42 & 34 & 29 \\ 47 & 34 & 36 & 43 \end{bmatrix}$$

Unlike scalar multiplication where $ab = ba$, the matrix product **AB** does not in general equal **BA**. Regarding the matrix product **AB**, we can say that **B** is *pre-multiplied* by **A** or that **A** is *post-multiplied* by **B**. The product of matrix **A** with itself is denoted as **A**².

The transpose of a matrix product is equal to the product of the transposed matrices in reverse order. Specifically, $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$.

The product of three matrices **ABC** requires **A** and **B** to be conformable for multiplication and also requires **B** and **C** to be conformable for multiplication. The product **ABC** can be obtained by first computing **AB** and then post-multiplying the result by **C**, or by first computing **BC** and then pre-multiplying the result by **A**.

If **A** is a square matrix, then the *matrix inverse* of **A** is represented as **A**⁻¹. If the inverse of **A** exists, then **AA**⁻¹ = **I**. This result is a generalization of scalar arithmetic where $x(1/x) = 1$, assuming $x \neq 0$ so that the inverse of x exists. Computing a matrix inverse is tedious, and

the amount of computational effort increases as the size of the matrix increases, but inverting a 2×2 matrix is not difficult. The inverse of a 2×2 matrix $\mathbf{A}$ is

$$\mathbf{A}^{-1} = (1/d) \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$$

where $d = a_{11}a_{22} - a_{12}a_{21}$ is called the *determinant* of $\mathbf{A}$. If the determinant of a matrix is zero, then the matrix is *singular*; otherwise, the determinant is *non-singular*. The matrix inverse is defined only for nonsingular matrices.

Inverting a diagonal matrix $\mathbf{D}$ of any order is simple. The inverse of $\mathbf{D}$ is equal to a diagonal matrix where the j^{th} diagonal element is equal to $1/d_j$.

The *trace* of a square $n \times n$ matrix $\mathbf{A}$, denoted as $tr(\mathbf{A})$, is defined as the sum of its n diagonal elements.

$$tr(\mathbf{A}) = a_{11} + a_{22} + \dots + a_{nn}$$

For instance if

$$\mathbf{V} = \begin{bmatrix} 14 & 5 & 2 \\ 9 & 20 & 8 \\ 7 & 8 & 11 \end{bmatrix}$$

then $tr(\mathbf{V}) = 14 + 20 + 11 = 45$. For two square matrices $\mathbf{A}$ and $\mathbf{B}$, $tr(\mathbf{AB}) = tr(\mathbf{BA})$.

For two matrices $\mathbf{A}$ and $\mathbf{B}$ of the same order, *matrix equality* $\mathbf{A} = \mathbf{B}$ indicates that $a_{jk} = b_{jk}$ for every value of j and k . *Matrix inequality* $\mathbf{A} \neq \mathbf{B}$ indicates that there is at least one element in $\mathbf{A}$ that is not equal to its corresponding element in $\mathbf{B}$.

The *Kronecker product* of two matrices, an $m \times n$ matrix $\mathbf{A}$ and a $p \times q$ matrix $\mathbf{B}$, is defined to be the $mp \times nq$ matrix

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix}$$

which is obtained by replacing each element a_{jk} with the $p \times q$ matrix $a_{jk}\mathbf{B}$. For example, if

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & -1 \end{bmatrix} \text{ and } \mathbf{b} = [1 \ 2 \ 3]$$

then

$$\mathbf{A} \otimes \mathbf{b} = \begin{bmatrix} 1 & 2 & 3 & 2 & 4 & 6 \\ 3 & 6 & 9 & -1 & -2 & -3 \end{bmatrix}.$$

The Kronecker product of an identity matrix and another matrix has the following simple form

$$\mathbf{I} \otimes \mathbf{B} = \begin{bmatrix} \mathbf{B} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{B} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{B} \end{bmatrix}$$

where $\mathbf{0}$ is a matrix of zeros that has the same order as $\mathbf{B}$.

The transpose of a Kronecker product of two matrices is

$$(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'$$

and the inverse of a Kronecker product of two matrices is

$$(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}.$$

In some statistical formulas, it is convenient to rearrange the elements of an $r \times c$ matrix $\mathbf{A}$ into an $rc \times 1$ column vector $\mathbf{a}$. This is done by stacking the c column vectors (which are each $r \times 1$) of $\mathbf{A}$ ($\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_c$) one under the other as shown below

$$\mathbf{a} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_c \end{bmatrix}$$

The conversion of a matrix $\mathbf{A}$ into a vector is denoted as $vec(\mathbf{A})$. To illustrate, consider the following 4×2 matrix $\mathbf{Y}$ and $vec(\mathbf{Y})$. The $mat(\mathbf{y})$ function returns the vector into its original matrix form.

$$\mathbf{Y} = \begin{bmatrix} 14 & 2 \\ 9 & 8 \\ 7 & 11 \\ 16 & 34 \end{bmatrix} \quad vec(\mathbf{Y}) = \begin{bmatrix} 14 \\ 9 \\ 7 \\ 16 \\ 2 \\ 8 \\ 11 \\ 34 \end{bmatrix} = \mathbf{y}$$

The vectorization of a matrix triple product can be expressed as

$$\text{vec}(\mathbf{ABC}) = (\mathbf{C}' \otimes \mathbf{A})\text{vec}(\mathbf{B})$$

and the vectorization of a matrix product $\mathbf{AB}$ follows from this result by setting $\mathbf{C} = \mathbf{I}_p$

$$\text{vec}(\mathbf{AB}) = (\mathbf{I}_p \otimes \mathbf{A})\text{vec}(\mathbf{B})$$

where p is the number of columns of $\mathbf{B}$.

B.3. Covariance and Correlation Matrices

A *covariance matrix* is a symmetric matrix with variances in the diagonal elements and covariances in the off-diagonal elements. If r response variables $\mathbf{y}' = [y_1 \ y_2 \ \dots \ y_r]$ are measured for each person in a random sample of n people, the estimated variance for the j^{th} response variable is

$$\hat{\sigma}_j^2 = \sum_{i=1}^n (y_{ij} - \hat{\mu}_j)^2 / (n - 1)$$

and the estimated covariance between the j^{th} and k^{th} response variables is

$$\hat{\sigma}_{jk} = \sum_{i=1}^n (y_{ij} - \hat{\mu}_j)(y_{ik} - \hat{\mu}_k) / (n - 1)$$

where $\hat{\sigma}_{jk} = \hat{\sigma}_{kj}$.

The r variances and the $r(r - 1)/2$ covariances of $\mathbf{y}' = [y_1 \ y_2 \ \dots \ y_r]$ can be summarized in an $r \times r$ covariance matrix denoted as $\text{cov}(\mathbf{y})$. For example, with $r = 3$ the covariance matrix is

$$\text{cov}(\mathbf{y}) = \begin{bmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \hat{\sigma}_{13} \\ \hat{\sigma}_{12} & \hat{\sigma}_2^2 & \hat{\sigma}_{23} \\ \hat{\sigma}_{13} & \hat{\sigma}_{23} & \hat{\sigma}_3^2 \end{bmatrix}.$$

The *generalized variance* of a vector $\mathbf{y}$ is defined as the determinant of $\text{cov}(\mathbf{y})$.

The estimated covariance between the j^{th} and k^{th} measurements is also equal to $\hat{\rho}_{jk}\hat{\sigma}_j\hat{\sigma}_k$ where $\hat{\rho}_{jk}$ is the estimated Pearson correlation between the two response variables. The correlations can be summarized in an $r \times r$ *correlation matrix* denoted as $\text{cor}(\mathbf{y})$. The

diagonal elements in the correlation matrix are all equal to 1. For example, with $r = 3$ the correlation matrix is

$$\text{cor}(\mathbf{y}) = \begin{bmatrix} 1 & \hat{\rho}_{12} & \hat{\rho}_{13} \\ \hat{\rho}_{12} & 1 & \hat{\rho}_{23} \\ \hat{\rho}_{13} & \hat{\rho}_{23} & 1 \end{bmatrix}.$$

If there are n sets of response variables where the i^{th} set has a covariance matrix $\mathbf{S}_i$, and response variables from different sets are assumed to be uncorrelated, the covariance matrix for all response variables has the following block-diagonal form

$$\begin{bmatrix} \mathbf{S}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{S}_n \end{bmatrix}$$

where each $\mathbf{0}$ represents a matrix of zeros. If all of the n covariance matrices are equal to $\mathbf{S}$, the above covariance matrix can then be expressed as $\mathbf{I} \otimes \mathbf{S}$.

Let $\mathbf{S}$ represent an $r \times r$ covariance or correlation matrix (i.e., $\mathbf{S} = \text{cov}(\mathbf{y})$ or $\text{cor}(\mathbf{y})$). There exists an $r \times r$ matrix $\mathbf{L}$ such that

$$\mathbf{S} = \mathbf{L}'\mathbf{V}\mathbf{L}$$

where $\mathbf{V}$ is an $r \times r$ diagonal matrix and the columns of $\mathbf{L}$ are orthonormal. The columns of $\mathbf{L}$ are the *eigenvectors* of $\mathbf{S}$, and each diagonal element of $\mathbf{V}$ is the *eigenvalue* of its corresponding eigenvector. Each eigenvector is a set of coefficients that define a linear function of the r variables, and its corresponding eigenvalue is equal to the variance of that linear function. The sum of the r eigenvalues is equal to $\text{tr}(\mathbf{S})$. If any eigenvalue is equal to 0, then the covariance or correlation matrix is singular and cannot be inverted.

B.4. Variance of a Linear Function of Variables

A linear function of r variables can be expressed as $\sum_{j=1}^r q_j y_j$ where the q_j coefficients are known numbers. This linear function can be expressed in matrix notation as $\mathbf{q}'\mathbf{y}$ where $\mathbf{q}' = [q_1 \ q_2 \ \dots \ q_r]$ and $\mathbf{y}' = [y_1 \ y_2 \ \dots \ y_r]$. The variance of $\mathbf{q}'\mathbf{y}$ can be expressed in matrix notation as $\mathbf{q}'\text{cov}(\mathbf{y})\mathbf{q}$. For example, the variance of $y_1 + y_2 + y_3$ is

$$\begin{aligned}
& [1 \ 1 \ 1] \begin{bmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \hat{\sigma}_{13} \\ \hat{\sigma}_{12} & \hat{\sigma}_2^2 & \hat{\sigma}_{23} \\ \hat{\sigma}_{13} & \hat{\sigma}_{23} & \hat{\sigma}_3^2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \\
&= [(\hat{\sigma}_1^2 + \hat{\sigma}_{12} + \hat{\sigma}_{13}) \quad (\hat{\sigma}_{12} + \hat{\sigma}_2^2 + \hat{\sigma}_{23}) \quad (\hat{\sigma}_{13} + \hat{\sigma}_{23} + \hat{\sigma}_3^2)] \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \\
&= \hat{\sigma}_1^2 + \hat{\sigma}_2^2 + \hat{\sigma}_3^2 + 2\hat{\sigma}_{12} + 2\hat{\sigma}_{13} + 2\hat{\sigma}_{23}.
\end{aligned}$$

The variance of a linear function of parameter estimates (i.e., $\mathbf{q}'\hat{\boldsymbol{\beta}}$) is similarly defined as $\mathbf{q}'cov(\hat{\boldsymbol{\beta}})\mathbf{q}$. For instance, the variance of $\hat{\beta}_1 - \hat{\beta}_2$ is

$$\begin{aligned}
& [1 \ -1] \begin{bmatrix} \hat{\sigma}_{\hat{\beta}_1}^2 & \hat{\sigma}_{\hat{\beta}_1\hat{\beta}_2} \\ \hat{\sigma}_{\hat{\beta}_1\hat{\beta}_2} & \hat{\sigma}_{\hat{\beta}_2}^2 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \\
&= [\hat{\sigma}_{\hat{\beta}_1}^2 - \hat{\sigma}_{\hat{\beta}_1\hat{\beta}_2} \quad \hat{\sigma}_{\hat{\beta}_1\hat{\beta}_2} - \hat{\sigma}_{\hat{\beta}_2}^2] \begin{bmatrix} 1 \\ -1 \end{bmatrix} \\
&= \hat{\sigma}_{\hat{\beta}_1}^2 - \hat{\sigma}_{\hat{\beta}_1\hat{\beta}_2} - \hat{\sigma}_{\hat{\beta}_1\hat{\beta}_2} + \hat{\sigma}_{\hat{\beta}_2}^2 \\
&= \hat{\sigma}_{\hat{\beta}_1}^2 + \hat{\sigma}_{\hat{\beta}_2}^2 - 2\hat{\sigma}_{\hat{\beta}_1\hat{\beta}_2}.
\end{aligned}$$

B.5. Covariance Algebra

Covariance algebra is used to define variances and covariances of linear functions of variables. Let y_1, y_2, y_3 , and y_4 be scalar variables and let a, b, c , and d be scalar constants (positive or negative). It will be convenient to express the variances of y_1, y_2, y_3 , and y_4 as $var(y_1), var(y_2), var(y_3)$, and $var(y_4)$ rather than $\sigma_{y_1}^2, \sigma_{y_2}^2, \sigma_{y_3}^2$, and $\sigma_{y_4}^2$, and to express the covariance of y_1 and y_2 as $cov(y_1, y_2)$ rather than $\sigma_{y_1y_2}$ to avoid double subscripts. Some basic covariance algebra results are summarized below.

$$\begin{aligned}
cov(y_1, y_2) &= cov(y_2, y_1) \\
cov(a, y_1) &= 0 \\
cov(y_1, y_1) &= var(y_1) \\
var(a + by_1) &= b^2 var(y_1) \\
cov(a + by_1, c + dy_2) &= bd[cov(y_1, y_2)] \\
cov(y_1 + y_2, y_3 + y_4) &= cov(y_1, y_3) + cov(y_1, y_4) + cov(y_2, y_3) + cov(y_2, y_4) \\
var(y_1 + y_2) &= var(y_1) + var(y_2) + 2cov(y_1, y_2) \\
var(y_1 - y_2) &= var(y_1) + var(y_2) - 2cov(y_1, y_2)
\end{aligned}$$

B.5. Some Matrix Commands in R

Dimension of matrix A :	<code>dim(A)</code>
Transpose matrix A :	<code>t(A)</code>
Add matrices A and B :	<code>A + B</code>
Subtract matrix B from A :	<code>A - B</code>
Multiply matrices A and B :	<code>A %*% B</code>
Multiply scalar b and matrix A :	<code>b*A</code>
Inverse of matrix A :	<code>solve(A)</code>
Kronecker product of matrices A and B :	<code>A %x% B</code>
Determinant of matrix A :	<code>det(A)</code>
Diagonal of matrix A :	<code>diag(A)</code>
Trace of matrix A :	<code>sum(diag(A))</code>
Eigenvalues and eigenvectors of A :	<code>eigen(A)</code>
Horizontal concatenation of A and B :	<code>cbind(A, B)</code>
Vertical concatenation of A and B :	<code>rbind(A, B)</code>

Appendix C. Glossary

adjusted squared multiple correlation – an estimate of the population squared multiple correlation that is less biased than the sample squared multiple correlation

AIC – Akaike information criterion; a measure of model fit that is used in covariance pattern models and multi-level models to choose between competing models and covariance structures

ANCOVA model – a special case of the general linear model where some predictor variables are indicator variables and some predictor variables are quantitative variables

ANOVA model – a special case of the general linear model where all predictor variables are indicator variables

baseline centering – a method of centering a time variable such that the first time period is equal to 0

between-subjects predictor variables – predictor variables in a repeated measures design that do not vary over time (e.g., gender, ethnicity); also referred to as time-invariant predictor variable

bivariate normality assumption – the assumption that two quantitative variables have an approximate bivariate normal distribution; this assumption implies that both variables have an approximate normal distribution, the two variables are linearly related, and the variability of one variable is approximately equal across the levels of the other variable

bootstrap standard error – a computationally intensive method to obtain approximate distribution-free standard errors

cluster-robust standard error – a standard error that assumes correlated prediction errors for the repeated measurements of each participant.

coefficient of determination – a squared Pearson correlation or eta-squared; describes the proportion of variance of the response variable that is predictable from the predictor variable

coefficient of multiple determination – a measure (e.g., R-squared and eta-squared) of the proportion of response variable variance that is predictable from all of the predictor variables

confounding variable – a variable that has a causal effect on the response variable and a predictor variable

composite model – a multi-level model where the level-2 models have been substituted into the level-1 model

compound-symmetric covariance matrix – a covariance matrix that is assumed to have equal variances and equal covariances

conditional slope – the slope coefficient for one predictor variable at a specified level of a second predictor variable; used when there is an interaction between the two predictor variables (also called a *simple slope*)

confounding variable bias – the bias in a slope coefficient that results when a confounding variable has been omitted in a statistical model

construct validity – the extent to which the magnitude of the correlation between the measurements of two attributes agrees with theoretical expectations

Cooks' D – a measure of the influence of each participant on all of the parameter estimates in a general linear model

covariance – a measure of association between two quantitative variables which is used in the definition of a regression slope and Pearson correlation

covariance pattern model – a linear statistical model for repeated measures data where the covariance structure of the prediction errors can be specified by the researcher

criterion validity – the extent to which a measurement of some attribute correlates with an objective criterion that should be related to the attribute

data transformation – used to reduce nonnormality, heteroscedasticity, or nonlinearity; commonly used data transformations are the log, square-root, and reciprocal transformations

design matrix – an n by $s + 1$ matrix with ones in the first column and s predictor variables in the other s columns

DFBETAS – a measure of the influence of each participant on each of the parameter estimates in a general linear model

dummy coded predictor variable – a predictor variable that has values of 0 and 1

eta-squared coefficient – a coefficient of determination for a fixed- x statistical model

equal prediction error variance assumption – the assumption in a linear statistical model that the prediction errors have approximate equal variability across the levels of each predictor variable

extrapolation – predicting a response variable for values of the predictor variable(s) that were not observed in the sample

FIML – full information maximum likelihood; a method of estimating parameters that allows for missing data

first-order autoregressive covariance structure – a prediction error covariance structure that is useful in longitudinal and pretest-posttest designs

Fisher transformation – a transformation of the Pearson correlation estimate is used to compute a confidence interval for the population Pearson correlation

fixed-x model – a statistical model where the values of predictor variables in the sample were predetermined

general linear model – a linear statistical model for a quantitative response variable where the predictor variables can be indicator variables, quantitative variables, or any combination of indicator and quantitative predictors

generalize least squares estimate – a method of estimating parameters of a statistical model that does not assume uncorrelated and equal variance prediction errors

interval hypothesis test – a test that a population parameter is within a specified (usually small) range of values

LAD estimation – a method of parameter estimation that minimized the absolute deviations of the prediction errors

level-1 model – the component of a multi-level model for repeated measures that describes the relation between the response variable and the within-subjects predictor variables

level-2 model – the component of a multi-level model for repeated measures that describes the relation between the random coefficients and between-subjects predictor variables

linearity assumption – the assumption that the response variable in a linear regression model is linearly related to each predictor variable

MAPE – a measure of the predictive accuracy of a general linear model

MAR – missing at random; missingness for a given variable that is unrelated to that variable and all variables not included in the model

matrix scatterplot – a scatterplot for all pairs of three or more variables

MCAR – missing completely at random; missingness for a given variable that is unrelated to that variable, all variables included in the model, and all variables not included in the model

model variance – the variance of the response variable that is predictable from all of the predictor variables

monotonicity test – a test that some parameter (e.g., a population mean) has a monotonic relation with an ordinal or quantitative factor

multicollinearity – the extent to which the predictor variables are correlated

multiple correlation – a measure of association, with a possible range of 0 to 1, between one response variable and two or more predictor variables

multiple linear regression model – a special case of the general linear model where all predictor variables are quantitative variables

multivariate normal distribution – a distribution assumption for two or more quantitative variables; multivariate normality implies normality of each individual variables, linear relations among every pair of variables, and equal residual variability for every pair of variables

ordinary least squares (OLS) estimates – a method of estimating the y -intercept and slope parameters such that the sum of the squared residuals are minimized

outlier – an unusually large or small y score or x score

partial correlation – the correlation between the part of a predictor variable that is not linearly related to the control variables (a residual predictor variable) with the part of the response variable that is not linearly related to the control variables (a residual response variable)

partial correlation plot – a scatterplot of a residual response variable with a residual predictor variable residual

person centering – a type of centering where each participant's mean score for a within-subjects predictor variable is subtracted from that within-subjects predictor variable

Pearson correlation – a measure of association between two quantitative variables that has a possible range of -1 to 1 where 0 indicates no linear relation and -1 or 1 indicate perfect linear association

point-biserial correlation – a measure of association between a dichotomous variable and a quantitative variable that has a possible range of -1 to 1

prediction error covariance matrix – the covariance matrix among the prediction errors of two or more response variables

prediction error normality assumption – the assumption in a linear statistical model that the prediction errors have an approximate normal distribution

prediction error variance – the variance of the prediction errors; the variance of the response variable that is not predictable from all of the predictor variables

predictor variable – a quantitative or dummy variable in a statistical model that is assumed to predict some of the variation in the response variable

quadratic model – a linear statistical model that has a quantitative variable and its squared value as predictor variables

qualitative factor – a factor with qualitatively different levels

quantitative factor – a factor with levels that are quantities

random coefficient model (multi-level model) – a statistical model where some of the model parameters are assumed to be random variables

random coefficient normality assumption – the assumption in statistical models with one or more random coefficients that the distribution of each random coefficient has an approximate normal distribution

random-x model – a statistical model where the values of predictor variables in the sample were not predetermined

regression discontinuity design – a nonexperimental design where participants with a high priority score receive a treatment and participants with low priority scores do not receive treatment; the causal effect of treatment can be assessed by using the priority score as a covariate

residual plot – a plot of the residuals (vertical axis) with a predictor variable (horizontal axis); used to visually assess a violation of the equal prediction variance assumption

scatterplot – a graphical method that illustrates the strength and form of the relation between two quantitative variables; bivariate outliers can also be detected in a scatterplot

semipartial correlation – the correlation between the component of a predictor variable that is not linearly related to the other predictor variables with the response variable

semipartial correlation plot – (also called a regression plot) a scatterplot of the response variable with a residual predictor variable

simple linear regression model – a statistical model that describes an assumed linear relation between one quantitative response variable and one predictor variable

slope – describes the change in the response variable for a 1-unit increase in a predictor variable

Spearman correlation – a measure of association that describes the strength of a monotonic relation between two quantitative variables (also used to describe the relation between two ordinal variables)

standardized slope – a slope coefficient that has been estimated from standardized predictor and response variables

studentized deleted residual – a residual for participant i that is computed from a model where the parameters are estimated excluding participant i and then divided by the standard error of the residual; used to detect outlier participants

Theil-Sen confidence interval – a distribution-free confidence interval for a population slope in a simple linear regression model

Toeplitz covariance structure – a covariance matrix where the covariances are assumed to be equal within each diagonal band

unstructured covariance matrix – a covariance matrix with no restrictions imposed on the variances or covariances

variance inflation factor – a measure of multicollinearity

within-subjects predictor variables – predictor variables in a repeated measures design that vary over time; also referred to as time-varying predictor variables

Appendix D. Answers to Concept Questions

Chapter 1

1. What information does a sample scatterplot provide?

The linearity of the relation, the strength of relation, the direction of the relation, and the possible presence of bivariate outliers

2. What are the confidence interval assumptions for β_1 ?

Random sample, independence among participants, linear relation between y and x , approximate normality of prediction errors, and prediction error homoscedasticity

3. What are the confidence interval assumptions for ρ_{yx} ?

Random sample, independence among participants, approximate bivariate normality of y and x

4. What effect does the planning value of ρ_{yx} have on the sample size requirement for ρ_{yx} ?

A smaller absolute planning value for ρ_{yx} gives a larger sample size requirement

5. How is β_1 interpreted in a nonexperimental random- x model?

β_1 describes the change in y associated with a 1-unit increase in x

6. How is β_1 interpreted in an experimental fixed- x model?

β_1 describes the change in the population mean of y caused by a 1-unit increase in x

7. How is $\mu_{y|x^*}$ interpreted in an experimental fixed- x model?

$\mu_{y|x^*}$ is the population mean of y if everyone in the population had an x score of x^*

8. If the estimated y -intercept is 10 and the estimated slope is 5, what is the predicted y_i score at $x_i = 20$?

$$10 + 5(20) = 110$$

9. How is eta-squared interpreted?

The proportion of the variance in the y scores that is predictable from all of the predictor variables in the model (the predictor variables are assumed to be fixed rather than random)

10. How does the prediction error variance affect the width of the confidence interval for β_1 ?

A smaller prediction error variance gives a narrower confidence interval width for the population slope

11. If the sample estimate of β_1 is 5.1, the 95% critical t -value is 3.0, and the standard error is 0.8, what is the 95% confidence interval for β_1 ?

$$5.1 \pm 3.0(0.8) = [2.7, 7.5]$$

12. Explain the difference between a confidence interval for $\mu_{y|x}$ and a prediction interval for $\hat{y}$.

A confidence interval for $\mu_{y|x}$ describes a plausible range for the population mean of y at the specified value of x . A prediction interval for $\hat{y}$ describes a plausible range of y scores for a single person who has the specified value of x .

13. Draw a scatter plot of 6 observations where the correlation is about 0.9.



14. Draw a scatter plot of 10 observations where the correlation is about -0.5.



15. How does the sample size affect the width of a confidence interval for β_1 ?

A larger sample size tends to give a narrower confidence interval width

16. How does the sample size affect the width of a confidence interval for ρ_{yx} ?

A larger sample size tends to give a narrower confidence interval width

17. How is β_1 interpreted in a simple linear regression model if x is dummy coded?

β_1 is equal to the population mean for the factor level that is coded 1 minus the population mean for the factor level this is coded 0.

18. What kind of data are used to compute a point-biserial correlation?

One variable is a dichotomous variable and one variable is a quantitative variable

19. How would the value of $\hat{\rho}_{yx}$ change if the x scores were all multiplied by 10?

No change because the correlation is a unitless measure of association (the standard deviations of x and y , which are multiplied in the denominator of $\hat{\rho}_{yx}$, each become 10 times larger and the covariance, which is the numerator of $\hat{\rho}_{yx}$, becomes 100 times larger).

20. How would the value of $\hat{\beta}_1$ change if the x scores were all multiplied by 10?

$\hat{\beta}_1$ would become 10 times smaller because a 1-point increase in the new x would correspond to a smaller change in the old x (the variance of x , which is the denominator of $\hat{\beta}_1$ becomes 100 times larger and the covariance, which is the numerator of $\hat{\beta}_1$ becomes 10 times larger).

21. How would the value of $\hat{\beta}_1$ change if the y scores were all multiplied by 10?

$\hat{\beta}_1$ would become 10 times larger because a 1-point increase in x would be associated with a larger change in y (the covariance, which is the numerator of $\hat{\beta}_1$, becomes 10 times larger).

22. What transformations are sometimes helpful in linearizing a nonlinear relation?

A square-root, log, or a reciprocal transformation of the x scores (but sometimes both x and y need to be transformed)

23. What transformations are sometimes helpful in normalizing residuals or stabilizing the error variance?

A square-root, log, or a reciprocal transformation of the y scores (but sometimes both x and y need to be transformed)

24. What are two general types of psychometric validity evidence?

Construct validity evidence and criterion validity evidence

25. How does the desired power of a statistical test affect the sample size requirement?

Larger desired power requires a larger sample size

26. How does the α level affect the sample size requirement?

A smaller α level requires a larger sample size

27. How does the desired confidence interval width affect the sample size requirement?

A narrower desired confidence interval width requires a larger sample size

28. How does measurement error in the y scores and x scores affect the Pearson correlation?

Measurement error in the y scores or the x scores will attenuate the Pearson correlation

29. How does measurement error in the y scores and x scores affect the slope?

Measurement error in the y scores has no effect on the slope but measurement error in the x scores will attenuate the slope

30. What are some advantages in using the Theil-Sen confidence interval for a population slope?

The Theil-Sen confidence interval does not require prediction error normality or prediction error homoscedasticity.

31. What are some advantages of using a Spearman correlation rather than a Pearson correlation?

The Spearman correlation assume a monotonic relation while the Pearson correlation requires a linear relation. The Spearman correlation is less sensitive to outliers than the Pearson correlation.

32. Participant y score x score

1	5	20
2	7	25
3	12	30

- | | | | |
|-------------------------------|-----------------------------|--------------------|--------------------|
| a) $\hat{\mu}_y = 8$ | f) $\hat{\rho}_{yx} = .971$ | | |
| b) $\hat{\mu}_x = 25$ | g) $\hat{\beta}_1 = 0.7$ | | |
| c) $\hat{\sigma}_y^2 = 13$ | h) $\hat{\beta}_0 = -9.5$ | | |
| d) $\hat{\sigma}_x^2 = 25$ | i) $\hat{y}_1 = 4.5$ | $\hat{y}_2 = 8.0$ | $\hat{y}_3 = 11.5$ |
| e) $\hat{\sigma}_{yx} = 17.5$ | j) $\hat{e}_1 = 0.5$ | $\hat{e}_2 = -1.0$ | $\hat{e}_3 = 0.5$ |

Chapter 2

1. Explain why slope coefficients for a treatment factor in an experimental design do not have confounding variable bias.

When participants are randomly assigned to treatments, the treatment indicator variables are uncorrelated with every possible confounding variable and therefore there can be no confounding variables and no confounding variable bias

2. Why is it important in a nonexperimental design to include confounding variables in the GLM?

Because confounding variables are controlled for when they are included in the model.

3. Why is the adjusted squared multiple correlation estimate preferred to the unadjusted estimate?

The adjusted squared multiple correlation has less positive bias.

4. How does the value of the squared multiple correlation between y and the predictor variables affect the width of a confidence interval for β_j ? Why?

A larger correlation between y and the predictor variables gives a smaller error for β_j , and a smaller standard error gives a narrower confidence interval for β_j .

5. When would a confidence interval for η^2 , rather than $\rho_{y,x}^2$, be reported?

η^2 would be reported in a fixed-x model and $\rho_{y,x}^2$ would be reported in a random-x model

6. How does the value of the squared multiple correlation between x_j and the other predictor variables affect the width of a confidence interval for β_j ? Why?

A larger multiple correlation between x_j and the other predictor variables increases the standard error for β_j , and a larger standard error gives a wider confidence interval for β_j .

7. Why is it informative to estimate the conditional slope for x_1 at low and high values of x_2 when x_1 and x_2 interact?

If x_1 and x_2 interact, then the strength of the relation between y and x_1 differs across the levels of x_2 and so it is informative to assess the relation between y and x_1 at low and high values of x_2 .

8. If a categorical predictor variable has four levels, how many dummy coded predictor variables are needed and what values of the dummy codes define each of the four levels?

Three dummy coded variables are needed with the following values:

Group	x_1	x_2	x_3
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	0

9. What are two ways to compute the conditional slopes for x_1 at the levels of x_2 if x_2 is a factor?

Compute the slope for x_1 within each level of x_2 or use the conditional slope formula $\hat{\beta}_1 + \hat{\beta}_3 x_2^*$

10. How is an interaction effect incorporated into a GLM?

By creating a product of two predictor variables and adding the product as an additional predictor variable in the model

11. How is a quadratic effect incorporated into a GLM?

By squaring the predictor variable and adding the squared variable as an additional predictor variable in the model

12. Why will an ANCOVA potentially require a smaller sample size than an ANOVA in an experimental design?

When the covariate is more than weakly related to the response variable, the error variance is reduced and the required sample size is a function of the error variance

13. When would a partial or semipartial correlation be preferred to a squared multiple correlation?

When the researcher wants to assess the relation between a single predictor variable and the response variable controlling for other predictor variables in the model

14. How are semipartial correlations useful in a regression analysis?

The semipartial correlation for each predictor variable describes the standard deviation change in the response variable associated with a one standard deviation increase in the component of the predictor variable that is unrelated to the other predictor variables. The semipartial correlation is a unitless measure of effect size

15. What is the main advantage of a regression discontinuity design?

Can be used to assess the causal effect of treatment without randomly assigning participants to treatment conditions.

16. How are interaction effects in a 2×3 design coded if both factors are qualitative?

The one indicator variable for the 2-level factor is multiplied by each of the two indicator variables for the 3-level factor and these two product variables are additional predictor variables in the model

17. How is the interaction effect in a 3×3 design coded if both factors are quantitative?

The quantitative 3-level variable is multiplied by the quantitative 4-level variable and this one product variable is an additional predictor variables in the model

18. Under what conditions would conditional slopes be interesting?

When there is an interaction effect between two predictor variables

19. What is the main difference between a partial correlation and a semipartial correlation?

The partial correlation removes the effects of the control variables from both the response variable and the predictor variable of interest while the semipartial correlation removes the effects of the control variables only from the predictor variable of interest.

20. What is the primary reason for using an ANCOVA rather than an ANOVA in experimental designs?

To reduce error variance which will increase the power of the test and decrease the confidence interval widths

21. What is the primary reason for using an ANCOVA rather than an ANOVA in nonexperimental designs?

To control for confounding variables

22. What should be done if a particular participant has an unusually large residual, DFBETAS value for any slope coefficient, or Cook's D value?

Check the data for that participant for possible data entry errors

23. What are some simple ways to assess the bivariate normality assumption?

Check for skewness and kurtosis of each variable separately, use a scatterplot to check for a linear relation and equal prediction error variability

24. Suppose a 95% confidence interval for $\rho_1 - \rho_2$ is [.132, .264] where ρ_1 is the Pearson correlation between academic self efficacy and grades for female high school students and ρ_2 is the Pearson correlation between academic self efficacy and grades for male high school students. How would you interpret this result?

The researcher can be 95% confident that the Pearson correlation between academic self efficacy and grades is .132 to .254 greater in the study population of female high school student than in the study population of male high school students.

25. How does the planning value of a partial correlation affect the sample size requirement for estimating the partial correlation with desired precision?

A smaller correlation planning value gives a larger sample size requirement

26. Why are slope coefficients sometimes difficult to interpret when the predictor variables are highly correlated?

When the predictor variables are highly correlated, the component of a predictor variable that is unrelated to the other predictor variables may not have an obvious interpretation which makes the slope coefficients difficult to interpret.

27. Suppose the sample size requirement for a linear contrast of means with desired power in an experimental design is 40 per group. Suppose that high school GPA is correlated .5 with the response variable. What would the sample size requirement be if GPA was used as the covariate?

When comparing the sample size formulas for a linear contrast of means with a covariate and without a covariate, we observe that the required sample size with a covariate is equal to $(1 - \tilde{\rho}_{y,x}^2)$ times the sample size without a covariate plus the number of covariates. Knowing that the sample is 40 per group without any covariates, the required sample size per group with one covariate would be $40(1 - .25) + 1 = 31$.

28. Explain how the multiple correlation, the semipartial correlation, and the partial correlation can be expressed in terms of a Pearson correlation between two variables.

The multiple correlation is a Pearson correlation between y and the predicted y scores, the partial correlation is a Pearson correlation between the residual y scores and the residual x scores, and the semipartial correlation is a Pearson correlation between the y scores and the residual x scores.

29. Fill in the blanks in the following ANOVA table where the total sample size is $n = 100$ and the GLM has $s = 5$ predictor variables.

Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>
MODEL	210	5	42	7
ERROR	564	94	6	
TOTAL	774	99		

30. Can adding a predictor variable decrease the value of $\hat{\rho}_{y,x}^2$. Why? Can adding a predictor variable decrease the value of $adj \hat{\rho}_{y,x}^2$? Why?

No, because $\hat{\rho}_{y,x}^2$ are computed from the Error SS and adding a predictor variable can decrease but never increase the Error SS and hence $\hat{\rho}_{y,x}^2$ can never decrease when a predictor variable is added to the model. Yes, because $adj \hat{\rho}_{y,x}^2$ is computed using the Error MS which can increase when a predictor variable is added to the model.

31. Why is the interpretation of β_j simplified if VIF_j for x_j is close to 1.0?

If VIF_j for x_j is close to 1.0, then x_j is at most very weakly correlated with the other predictor variables and then the component of x_j that is unrelated to the other predictor variables is essential equal to x_j . Hence β_j will then describe the relation between x_j and y .

Chapter 3

1. What are the main differences between a covariance pattern model and a random coefficient model?

A random coefficient model allows the y -intercept and one or more slope coefficients to vary randomly across persons and the variance of these random coefficients can provide useful information. A random coefficient model can also be used when participants are measured on different sets of time periods. The covariance pattern model requires all participants to be measured on the same set of time periods with missing data on some time periods allowed.

2. When would a random coefficient model be preferred to a covariance pattern model?

When participants are measured on different sets of time points or when the variance of the random coefficients provides useful information

3. When would a covariance pattern model be preferred to a random coefficient model?

If all participants are measures on the same set of time points and variability of person-level y -intercepts and slopes is not interesting.

4. What are the advantages and disadvantages of using an unstructured prediction error covariance matrix in a covariance pattern model?

The advantage is that the standard errors will be estimated more accurately and the disadvantage is that the sampling distribution of the slope estimates will be less accurately approximated by a normal distribution in small samples

5. What are the advantages and disadvantages of using a structured prediction error covariance matrix in a covariance pattern model?

The disadvantage is that the standard errors will be estimated less accurately and the advantage is that the sampling distribution of the slope estimates will be more accurately approximated by a normal distribution in small samples

6. What is the structure of each block of the composite prediction error covariance matrix if a random coefficient model has a random y -intercept but no random slopes?

compound symmetric (which could be an unrealistic representation of the true prediction error covariance matrix)

7. If a predictor variable is used in the level-2 model for a random slope, how will that predictor variable appear in the composite model?

As an interaction between the level-2 predictor variable and the level-1 predictor variable

8. What is a possible negative consequence of treating a slope coefficient as random rather than fixed on the composite prediction error covariance matrix?

The implied variances and covariances of the composite prediction error covariance matrix could be an unrealistic representation of the true prediction error covariance matrix

9. How can a quadratic time trend be modeled?

Include a squared time variable in the model

10. How can a 3-level within-subjects treatment factor be modeled in a covariance pattern model or a random coefficient model?

Include two dummy coded within-subjects predictor variables in the level-1 model

11. When would a lagged predictor variable be used?

In a longitudinal design where the predictor variable has a delayed effect on the response variable

12. What covariance structures should be considered in a within-subjects experiment?

Compound-symmetric, heteroscedastic compound-symmetric, unstructured

13. What covariance structures should be considered in a longitudinal design?

First-order autoregressive, heteroscedastic first-order autoregressive, Toeplitz, heteroscedastic Toeplitz, unstructured

14. How would an interaction between two within-subjects predictor variables be modeled in a random coefficient model?

By including the product of the two within-subjects predictor variables in the level-1 model

15. How would an interaction between two between-subjects predictor variables be modeled in a random coefficient model?

By including the product of the two between-subjects predictor variables in a level-2 model

16. What are the advantages of using a covariance pattern model to analyze a within-subjects experiment over the traditional methods?

If there is missing data, all of the data are used in a covariance pattern model while participants with any missing data are eliminated in a traditional analysis. Also, the traditional methods assume either compound symmetric or an unstructured prediction error covariance matrix while the covariance pattern model can use other covariance structures.

17. What are recommended types of centering for a "time" variable?

baseline centering or mean centering

18. What is the recommended type of centering for a within-subjects predictor variable?

person centering

19. If the variance of a random slope for predictor variable x_1 is large, what does this suggest?

There are other predictor variables that might be interacting with x_1

20. What is the effect of violating the random coefficient normality assumption on confidence intervals for the parameter variances?

A violation of this assumption will result in misleading confidence intervals for the parameter variances regardless of sample size.

21. Explain the difference between a wide data format and a long data format.

With a wide data format, each row corresponds to a single participant and the repeated measurements for each participant given in different columns. With a long data format, each row corresponds to a single occasion for a given participant.

Appendix E. Answers to Data Analysis Problems

Problem 1-1

a) Describe the study population.

4,000 male employees from the employee database

b) Is this an experimental or a nonexperimental design? Why?

Experimental design because the participants were randomly assigned to the levels of the independent variable

c) What is the response variable and what is the predictor variable in this study?

The response variable is the score on the sexual harassment test. The predictor variable (independent variable) is the number of examples.

d) Is x fixed or random? Why?

Fixed because the values of the independent variables (the number of examples) were predetermined.

e) Report the hypothesis testing results in APA style for a test of $H_0: \beta_1 = 0$.

$t(30) = 3.36, p = .002$. Reject the null hypothesis and accept the alternative hypothesis that the population slope is greater than 0.

f) Compute a 95% confidence interval for β_1 and interpret the results.

95% CI: [0.58, 2.38]. The researcher can be 95% confident that every one additional example will cause the mean test score to increase 0.58 to 2.38 points in the population of 4,000 male employees.

g) The organization plans to replicate this study in a population of employees in another city with training that consists of 2, 5, and 8 examples ($\sigma_x^2 = 6.0$). How many employees should be sampled to obtain a 95% confidence interval for β_1 that has a width of about 1.0? Use the MS_E from this study as the planning value of the within-group error variance.

```
size.ci.slope(.05, 31.1, 6, 1)
Sample size
      83
```

83 employees should be randomly selected and then randomly assigned to the three groups (if 84 employees are sampled then 28 could be assigned to each group)

Problem 1-2

a) Describe the study population.

The 1,200 1st and 2nd grade students who attend Los Angeles elementary schools

b) Is this an experimental or a nonexperimental design?

Nonexperimental

c) What is the response variable and what is the predictor variable in this study?

The response variable is the child screen time and the predictor variable is the parent screen time.

d) Is x fixed or random?

Random because the parent screen time values were observed in the sample and not predetermined

e) Examine the scatterplot of the sample data. Are there any obvious outliers or indications of nonlinearity?

No obvious outliers or indications of nonlinearity

f) Compute a 95% confidence interval for β_1 (average parent score as the predictor variable) and interpret the result.

95% CI: [0.26, 0.51]. The researcher can be 95% confident that every 1-hour increase in parent screen time is associated with a 0.26 to 0.55 hour increase in child screen time in the population of 1,200 1st and 2nd grade Los Angeles children.

g) Compute a 95% confidence interval for ρ_{yx} . Compute a 95% confidence interval for ρ_{yx}^2 . Interpret the results for both confidence intervals.

95% CI: [.362, .651]. The researcher can be 95% confident that the Pearson correlation between parent screen time and child screen time is between .362 and .651 in the population of 1,200 1st and 2nd grade Los Angeles children.

h) Use the 95% confidence interval for ρ_{yx} to test $H_0: \rho_{yx} = .3$ for $\alpha = .05$. What is your decision?

The 95% lower limit is greater than .3 and so we can reject $H_0: \rho_{yx} = .3$ and accept $\rho_{yx} > .3$

i) The researcher plans to replicate this study in New York. How many families are needed to obtain a 95% confidence interval for ρ_{yx} that has a width of about 0.25? Use the lower 95% limit (the endpoint closest to zero in this case) of ρ_{yx} from this study as your planning value for ρ_{yx} .

```
size.ci.cor(.05, .362, 0, .25)
Sample size
```

A random sample of 188 families should be obtained.

j) Compute the skewness and kurtosis (or excess kurtosis if using SPSS) for the parent and child screen times and report the values.

child skewness = -0.31

parent skewness = 0.28

child kurtosis = 2.97 (excess kurtosis = $2.97 - 3 = -0.03$)

parent kurtosis = 2.97 (excess kurtosis = $2.97 - 3 = -0.03$)

Problem 1-3

a) Describe the study population.

The 15,000 undergraduate students.

b) Compute a 95% confidence interval for a weighted variance point-biserial correlation. Interpret the result.

95% CI: [-.548, -.347]. The researcher can be 95% confident that the point-biserial correlation between education generation status and academic self-efficacy is between .347 and .548 in the study population of 15,000 undergraduate students. (*Note: Unless the group variable has a natural ordering, the sign of a point-biserial correlation is arbitrary and depends only which group is called group 1*)

Problem 2-1

a) Describe the study population.

652 5th grade students

b) What is the control variable?

Time

c) Compute a 95% confidence interval for the population partial correlation between mathematics self efficacy and mathematics examination performance controlling for time spent on the mathematic examination. Interpret the result.

95% CI: [.024, .887]. The researcher can be 95% confident that the correlation between mathematics self efficacy and math performance controlling for the amount of time required to complete the math exam, is between .024 and .887. This confidence interval is very wide and the study should be replicated using a much larger sample.

d) The results of this study are potentially interesting but the confidence interval was much too wide to report in a scientific publication. The researcher wants to replicate this study and compute a 95% confidence interval that has a width of about 0.25. How many students should the researcher sample? Use the 95% lower limit (i.e., the endpoint closest to zero in this case) of the partial correlation from this study as a planning value.

```
size.ci.cor(.05, .0238, 1, .25)
Sample size
      248
```

A random sample of 248 5th grade students should be obtained.

Problem 2-2

a) Describe the study population.

All incoming male students at the university where the study was conducted

b) What is the response variable and what are the two predictor variables in this study?

The response variable is the son's aggression score; the father's aggression score and the son's average video game play per day each are the two predictor variables.

c) Compute 95% confidence intervals for the two population slope coefficients and interpret the results.

father aggression - 95% CI: [0.265, 0.760]. The researcher can be 95% confident that in the population of incoming male students, every 1-point increase in the father's aggression score is associated with a 0.265 to 0.760 increase in the son's aggression score, controlling for son's hours of violent video game play.

hours game play - 95% CI: [0.967, 4.188]. The researcher can be 95% confident that in the population of incoming male students, every 1-hour increase in son's average daily violent video game playing is associated with a 0.967 to 4.188 increase in the son's aggression score, controlling for father's aggression.

d) Compute a 95% confidence interval for the population squared multiple correlation and interpret this result.

95% CI: [.337, .855]. The researcher can be 95% confident that in the population of incoming male students, 33.7% to 85.5% of the variance in son's aggression scores is predictable from father's aggression scores and son's hours of violent video game play.

e) Compute 95% confidence intervals for semipartial correlation between y and x_1 and y and x_2 . Interpret these results.

father aggression - 95% CI: [.252, .790]. The researcher can be 95% confident that in the population of incoming male students, the correlation between the son's aggression scores and the father's aggression scores is between .242 and .790, after removing the effect of son's hours of violent video game playing from father's aggression scores.

hours game play - 95% CI: [.136, .682]. The researcher can be 95% confident that in the population of incoming male students, the correlation between the son's aggression scores and son's average daily violent video game use is between .136 and .682, after removing the effect of father's aggression from son's hours of violent video game playing.

f) Another researcher wants to conduct a similar study using a sample of male community college students and their fathers. The researcher suspects that the population squared correlation is about 0.7 and would like the 95% confidence interval for the squared multiple correlation to have a width of .2. How many male first-year students should be sampled?

```
size.ci.rsqr(.05, .7, 2, .2)
Sample size
      104
```

A random sample of 104 community college male students should be obtained.

Problem 2-3

a) Describe the study population.

The 4,000 undergraduate students

b) What is the response variable and what are the predictor variables in this experiment?

The response variable is heart rate during the checkers game. The two predictor variables are game difficulty level and number of spectators.

c) Compute a 95% confidence interval for the two main effects and the interaction effect. Interpret the results. (Both predictor variables have been centered in the SPSS data file.)

The 95% confidence interval for the interaction effect includes 0 and so the main effects will be examined.

number spectators - 95% CI: [2.50, 4.70]. The researcher can be 95% confident that in the population of 4,000 college students, every one additional spectator will cause the mean heart rate during the checkers game to increase 2.5 to 4.7 beats per minute.

game difficulty - 95% CI: [0.55, 1.39]. The researcher can be 95% confident that in the population of 4,000 college students, every 1-point increase in the game difficulty level will cause the mean heart rate during the checkers game to increase 0.55 to 1.33 beats per minute.

(Note that it was not necessary to include the phrase "controlling for" because participants were randomly assigned to the treatment conditions. Random assignment controls for all possible confounding variables and permits a causal interpretation.)

d) Use Equation 1.35 to determine the number of participants that should be added to this study to cut each 95% confidence interval width in half (i.e., set $w_0/w = 2$).

$$(2^2 - 1)36 = 108$$

A random sample of 108 undergraduates from the same university should be sampled and added to the original sample of 36 students.

Problem 2-4

a) Describe the study population.

1,200 female undergraduate students

b) Use a GLM with centered GPA and dummy coding for group (Instruction). Perform a test of the GPA $\times$ Group interaction. Can an analysis of the main effect of Group be justified? Why?

The 95% confidence interval for the interaction effect includes 0 and so the main effect of Instruction can be examined.

c) Compute a 95% confidence interval for the effect of Instruction. Interpret the result.

95% CI: [0.46, 3.51] The researcher can be 95% confident that in the population of 1,200 female college students, the mean test score would be 0.46 to 3.51 larger if they had all had been given the neutral instruction than if they had all been given the stereotype threat instruction.

d) The researcher wants to replicate this experiment using a larger sample with the goal of obtaining a narrower confidence interval for the population mean difference. How many participants per group are needed to obtain a 95% confidence interval for the population mean difference that has a width of 1.5 using GPA as a covariate. Use the MS_E from this study as a planning value of the average within-group residual variance.

```
v = c(1, -1)
size.ci.lc.ancova(.05, 2.58, 1, 0, 1.5, v)
Sample size per group
38
```

A random sample of 76 female participants should be obtained and then randomly divided into two groups with 38 per group.

Problem 3-1

a) Describe the study population.

The 14,000 Silicon Valley employees in the database

b) Use R to compute a 95% confidence interval for the population slope using OLS estimation with cluster-robust standard errors and Satterthwaite degrees of freedom. Interpret the results. Also, report the t -value, degrees of freedom, and p -value.

95% CI: [0.159, 1.02] The researcher can be 95% confident that the mean mood score of the 14,000 employees increases 0.159 to 1.02 each day from Monday to Friday. $t(19) = 2.87, p = .010$.

c) Use a random coefficient model with uncorrelated and homoscedastic level-1 prediction errors, a random y -intercept, and a random slope. Allow the random y -intercept and slope parameters to be correlated. Report the 95% confidence intervals for the random y -intercept standard deviation and random slope standard deviation.

```
y-intercept 95% CI: [1.28, 3.05]
slope       95% CI: [0.53, 1.22]
```

The researcher can be 95% confident that the standard deviation of the employee-specific y -intercepts is between 1.28 and 3.05 and the standard deviation of the employee-specific slopes is between 0.53 and 1.22 for the 14,000 employees.

Problem 3-2**a) Describe the study population.**

500 preschool children in San Jose

b) Use R to compute 95% confidence intervals for the population slopes using OLS estimation with cluster-robust standard errors and Satterthwaite degrees of freedom. Interpret the results. Also report the t -values, degrees of freedom, and p -values.

gender – 95% CI: [-5.38, 26.85]. The results are inconclusive. $t(1.37) = 4.61, p = .087$.

week – 95% CI: [0.32, 1.53]. The researcher can be 95% confident that the mean social skills score for all 500 preschool children in the population increased 0.52 to 1.53 each week. $t(4.45) = 4.09, p = .012$.

vocabulary – 95% CI: [-0.41, 1.78]. The results are inconclusive. $t(1.40) = 4.17, p = .094$.

group size – 95% CI: [-11.12, -3.28]. The researcher can be 95% confident that every one additional child in group is associated with a 3.28 to 11.12 point decrease in social skills score for all preschool children in the population. $t(4.22) = -4.99, p = .006$.

average group size – 95% CI: [-49.69, 31.99]. The results are inconclusive. $t(1.49) = 1.32, p = .354$.

c) Use a random coefficient model with uncorrelated and homoscedastic level-1 prediction errors, a random y -intercept, and a random slope for week. Allow the random y -intercept and slope parameters to be correlated. Report the 95% confidence intervals for the random y -intercept standard deviation and random slope standard deviation.

y -intercept 95% CI: [3.06, 13.46]

slope 95% CI: [0.27, 1.25]

These confidence intervals are very wide and a larger sample size is needed to obtain more accurate estimates of the random coefficient standard deviations.

Problem 3-3**a) Describe the study population.**

All girls in a Los Angeles school district

b) Use R to compute 95% confidence intervals for the population slopes using OLS estimation with cluster-robust standard errors and Satterthwaite degrees of freedom. Interpret the results. Also report the t -values, degrees of freedom, and p -values.

program – 95% CI: [4.10, 7.20]. The researcher can be 95% confident that the mean self esteem score for all Los Angeles girls in the population would be 4.10 to 7.20 greater if they all received the mentoring program in than if they had all had not received the mentoring program. $t(53.8) = 7.33, p < .001$.

average physical activity – 95% CI: [0.36, 1.72]. For the population of Los Angeles girls, the researcher can be 95% confident that every 1-point increase in a girl's average physical activity across grades 5 through 8 is associated with a 0.36 to 1.72 point increase in self esteem. $t(22.7) = 3.15, p = .004$.

physical activity – 95% CI: [0.86, 1.75]. For the population of Los Angeles girls, the researcher can be 95% confident that every 1-point increase in a girl's physical activity in a given grade is associated with a 0.86 to 1.75 point increase in self esteem in that grade. $t(26.0) = 6.05, p < .001$.

grade – 95% CI: [-3.51, -2.15]. For the population of Los Angeles girls, the researcher can be 95% confident that the mean self esteem score decreases 2.15 to 3.51 each year from 5th grade to 8th grade. $t(41.8) = -8.40, p < .001$.

c) Use a random coefficient model with uncorrelated and homoscedastic level-1 prediction errors, a random y -intercept, and a random slope for grade. Allow the random y -intercept and slope parameters to be correlated. Report the 95% confidence intervals for the random y -intercept standard deviation and random slope standard deviation.

y -intercept 95% CI: [0.00, 61.36]
slope 95% CI: [0.48, 2.13]

These confidence intervals are very wide and a larger sample size is needed to obtain more accurate estimates of the random coefficient standard deviations.

Problem 3-4

a) Describe the study population.

The 2,560 undergraduate students.

b) Use R to compute 95% confidence intervals for the population slopes using OLS estimation with cluster-robust standard errors and Satterthwaite degrees of freedom. Interpret the results. Also report the t -values, degrees of freedom, and p -values.

send vs. receive – 95% CI: [-9.76, -2.57]. In the population of 2,560 undergraduate students, the mean driving performance score would be 2.57 to 9.76 higher under a receive text condition than under a send text condition. $t(11.0) = -3.78, p = .003$

time – 95% CI: [-0.404, -0.015]. In the population of 2,560 undergraduate students, every one additional second total sending and receiving time is associated with a .015 to .404 decrease in the driving performance score. $t(3.7) = -3.08, p = .041$