DEPARTMENT OF CIVIL ENGINEERING
UNIVERSITY OF ARKANSAS
COURSE SYLLABUS: **Data Analysis and Machine Learning (CVEG 563V-001)**

Meeting Time: Monday, Wednesday, and Friday, 12:55 – 1:45PM
Final Exam: TBD
Location: Bell 2268

Instructor: Sarah Hernandez, PhD
(Office) 4159 Bell Engineering Center
(479) 575-4182
sarahvh@uark.edu
Office hours: Monday and Wednesday 3-5PM or whenever door is open

Course Description
The purpose of this course is to provide students with a solid background in the application of common statistical/econometric analysis techniques and related statistical modeling. This course emphasizes the empirical application of statistical techniques, but underlying theories and their limitations will also be discussed and simple derivations will be performed in class. The class will focus on applications of modeling techniques through the use of technical computing software including Matlab and KNIME. Students from all areas of engineering, supply chain management, and other broad disciplines are welcome. General topics include but are not limited to:

1. Survey sampling- sampling methods and statistical properties of survey sample estimates
2. Statistical inference- hypothesis tests, nonparametric tests, goodness-of-fit
3. Regression and time series modelling- estimation methods and model assumptions
4. Machine Learning I- supervised learning (classification and regression trees, neural networks, support vector machines)
5. Machine Learning II- unsupervised learning (clustering, mixture models)

Course Objectives
By the end of this course students should be able to…
1. Select and apply appropriate statistical and econometric models and analytical tools
2. Interpret the results of statistical and econometric models used in civil engineering analyses
3. Critique statistical and econometric models used in research

Materials
Textbooks:
1. Washington, S., Karlaftis, M., and Mannering, F. (2011). Statistical and Econometric Methods for Transportation Data Analysis, 2nd Edition, Chapman and Hall/CRC. *(denoted WKM in reading schedule)*
2. Stopher, P. R., & Meyburg, A. H. (1979). Survey sampling and multivariate analysis for social scientists and engineers. Lexington, Mass: Lexington Books. *(denoted SM in reading schedule)*
3. Bishop, C., Pattern Recognition and Machine Learning, Springer, 2006.
4. Barber, D., Bayesian Reasoning and Machine Learning, Cambridge University Press, 2012. Available online for free at http://www.cs.ucl.ac.uk/staff/d.barber/brml/
5. Fitzpatrick and Ledeczi, Computer Programming with MATLAB, online at http://cs103.net *(denoted FL in reading schedule)*

Computing:
MatLab (students can obtain for free through IT services)
KNIME (free, open source available at https://www.knime.org/)

Student Evaluation
The following *tentative* weighting scheme and assignments will be applied:

- *Homework (50%)*- typed; no late work accepted
- *Term Project (20%)*
- *Midterm (15%)*
- *Final Exam (15%)*

Tentative Grading Scale: A: 90-100%, B: 80-89%, C: 70-79%, D: 60-69%, F: less than 59%

Term Project
The term project should be a 10-12 page paper (Times New Roman, 11 point font, 1.15 spaced, double sided) that includes an analysis of a large dataset using one of the techniques covered in class (or other methods approved by the instructor). It should include: 1) a brief motivation for the problem studied, 2) a brief literature review (with at least five recent papers published in peer reviewed journals), 3) a salient statistical overview of the data, 4) a brief discussion of the methodology, 5) a summary of the results (numerical and written), and 6) a short conclusion. For your analysis, you may use Matlab, KNIME, SAS, Stata, or SPSS. A one page proposal is due in week 9 and the final report is due week 16.

Academic Integrity and Emergency Procedures:
Each University of Arkansas student is required to be familiar with and abide by the University's 'Academic Integrity Policy' which may be found at http://provost.uark.edu/. Students with questions about how these policies apply to a particular course or assignment should immediately contact me.

In addition, many types of emergencies can occur on campus. Instructions for specific emergencies such as severe weather, active shooter, or fire can be found at http://emergency.uark.edu/. If the University is closed, class is cancelled.

Tentative Course Schedule

Week	***Topic***	***Reading/Assignments***
Week 1	Introduction and Statistical Fundamentals	WKM Appendix A
	Descriptive statistics and properties of estimators	WKM 1.1-1.6
Week 2	Matlab introduction	FL Chapter 1 (pp. 11-28)
	Matlab data structures (matrix, array, vector, struct, cell, etc)	FL Chapter 1 (pp. 33-60) and Chapter 2 (pp. 196-227)
	Matlab plotting	WKM 1.6; FL Chapter 1 (pp. 29-30) and Chapter 2 (pp. 97-101)
Week 3	Surveys: types of data, data needs, and sources of error	SM Chapter 2 (pp. 9-14)
	Surveys: sampling methods	SM Chapter 3 (pp. 21-42)
	Surveys: standard errors, sample size	SM Chapter 4 (pp. 45-49; 54-57)
Week 4	Matlab procedural programming and scripts	FL Chapter 2 (pp. 62 -79)
	Matlab programming- if, switch, and loops	FL Chapter 2 (pp. 113-133; 139-195)
	Statistical inference (SI): Introduction & confidence intervals	WKM Chapter 2 Intro and WKM 2.1
Week 5	SI: hypothesis testing for single population	WKM 2.2-2.3
	SI: two populations	WKM 2.4
	SI: nonparametric methods	WKM 2.5
Week 6	General Linear Model (GLM): Assumptions and Fundamentals	WKM Chapter 3.1-3.2
	GLM: Variables	WKM 3.3
	GLM: Estimating Beta for variables	WKM 3.4, 3.5, 3.6
Week 7	GLM: Goodness-of-fit measures	WKM 3.9
	GLM: Model building strategies	WKM 3.11
	GLM extensions: Tobit and Box-Cox	WKM 3.13 and 3.14
Week 8	Matlab exercise: Loops for plotting	(inductive signature example)
	Matlab exercise: Statistical hypothesis testing	(GVW distribution comparisons)
	Matlab exercise: Box-Cox Regression	(truck count and weather data)
Week 9	Midterm Exam Review	
	Midterm Exam	WKM Ch. 2-3, Appendix A
	Project Discussions	
Week 10	Spring Break! No Class	
Week 11	Logistic regression model	WKM 12.1-12.2
	Overview of discrete outcome models	WKM 13.1-13.4
	Estimation of multinomial logit	WKM 13.5
Week 12	Simultaneous Equation Models	WKM 5
	Time Series Models and ARIMA	WKM 7.1-7.2 & 8.1-8.2
	Matlab exercise: Time series model	
Week 13	Machine Learning overview	Readings distributed in class
	Machine learning I: classification and regression trees (CART)	
	Machine learning I: neural networks & support vector machines	
Week 14	KNIME: Introduction	
	KNIME: Data sorting exercise	(prepare data for training and testing)
	KNIME: Supervised learning methods	(CART for axle based vehicle class)
Week 15	Machine learning II: Clustering	
	Machine learning II: Gaussian Mixture Models	
	KNIME: Unsupervised learning	(clustering)
Week 16	Class presentations	
	Class presentations	
	Dead Day	
Exams	Final Exam	