DEPARTMENT OF CIVIL ENGINEERING
UNIVERSITY OF ARKANSAS
COURSE SYLLABUS: **Data Analysis and Machine Learning (CVEG 563V-001 12383)**

Meeting Time: Monday & Wednesday 3:30-4:45PM, 01/11/2021 to 04/29/2021
Location: ENGR 0304

Instructor: Sarah Hernandez, PhD
(Office) 4159 Bell Engineering Center
(479) 575-4182
sarahvh@uark.edu
Office hours: By appointment, email, or Slack

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 and other broad disciplines are welcome. General topics include but are not limited to:

1. Sampling and experimental design- sampling methods and statistical properties of sample estimates
2. Statistical inference- hypothesis tests, nonparametric tests, goodness-of-fit measures
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)
6. Introduction to Matlab and KNIME software tools

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
Recommended 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. Sheskin, D., Handbook of Parametric and Nonparametric Statistical Procedures, Fifth Edition.
4. Bishop, C., Pattern Recognition and Machine Learning, Springer, 2006.
5. 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/
6. 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/)
Access to Blackboard Collaborate Ultra, a webcam and microphone

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

- *Homework (10%)*- typed; no late work accepted
- *Term Project (30%)*
- *Midterm (30%)*
- *Final Exam (30%)*

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 8-10 page journal style paper (Times New Roman, 11 point font, 1.15 spaced, double sided) that includes an analysis of a 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.* It is recommended to use a dataset collected through your thesis/dissertation work, however if this is not available to you, I will provide a dataset to you.

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.

Remote Delivery
This course is offered through remote delivery in real time (synchronous learning). Students are expected to attend live lectures at the scheduled time. Students will be encouraged to use webcams when asking questions. The instruction will share her screen for note taking. Notes will be written using a tablet and posted after class along with the video. The **Blackboard Collaborate software** will be used for live, online classes. Slack video conference will be used for Office Hours. Students can send a message on Slack to Dr. Hernandez during regular business hours for a response by chat or video. Students will be added to **Slack** using their UArk email address.

Blackboard
The online learning management system, Blackboard will be used for lecture delivery, assignments and grading.

Tentative Course Schedule

Week	***Topic***	***Reading***
Week 1 1/11	Review of statistical fundamentals	WKM Appendix A
	Descriptive statistics and properties of estimators	WKM 1.1-1.6
Week 2 1/18 (1/18 MLK Holiday)	Matlab data structures	FL Chapter 1 (pp. 11-28 and 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 1/25	Sampling: types of data, data needs, and sources of error	SM Chapter 2 (pp. 9-14) Chapter 3 (pp. 21-42)
	Experiments: sampling methods, standard errors, sample size	SM Chapter 4 (pp. 45-49; 54-57)
Week 4 2/1	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)
Week 5 2/8	Statistical inference (SI): Confidence intervals, hypothesis test for single sample	WKM Chapter 2 Intro and WKM 2.1-2.3
	SI: hypothesis testing for two populations	WKM 2.4
Week 6 2/15	SI: nonparametric methods	WKM 2.5
Week 7 2/22 (2/22-23 Holiday)	General Linear Model (GLM): Assumptions and Fundamentals, variables	WKM Chapter 3.1-3.3
Week 8 3/1	GLM: Estimating Coefficients and Goodness of fit	WKM 3.4, 3.5, 3.6, 3.9
	GLM: Model Building and Indicator Vars.	WKM 3.4-3.6; 3.11
Week 9 3/8	GLM: Data Transformations	WKM 3.13-3.14
	Midterm Exam	
Week 10 3/15	Logistic regression model and discrete outcome models	WKM 12.1-12.2; 13.1-13.4
	Estimation of multinomial logit (Fischer Iris Example)	WKM 13.5
Week 11 3/22 (3/25-26 Holiday)	Time Series Models and ARIMA	WKM 7.1-7.2 & 8.1-8.2
Week 12 3/29 (4/2 Holiday)	Machine Learning Part 1	
Week 13 4/5	Machine Learning Part 2	
Week 14 4/12	KNIME Examples (Fun with Flags!)	KNIME User Manual
	Machine Learning Ensemble Example (Inductive Signatures)	
Week 15 4/19	Class Presentations	**Project Due**
	Class Presentations	
Week 16 4/26	Class Presentations	
	Review	
Exams 5/3-5/7	**Final Exam**	