SPEA Connect P507 (17024) – Spring 2017
Data Analysis and Modeling for Public Affairs

Tutorial: Tue, 7:00-8:30 pm via Zoom
Class Number: 17024
Instructor: Professor Barry Rubin
Office: SPEA 329
Phone: (812) 855-4556/327-3852
E-Mail: rubin@indiana.edu
Office Hrs: Tue, 8:30-9:30 pm, or by appointment

Lab: Thu, 7:00-8:30 pm via Zoom
TAs: Mackenzie Higgins & Venkat Nadella
Mack: (317) 217-9000, mackhigg@iu.edu
Venkat: (812) 369-1725, vnadella@indiana.edu
TA Office Hours: Thu, 8:30-10:00 pm & Sunday, 4:00-6:00 pm, or by appointment

Course Description and Objectives

SPEA Connect P507 provides students of public and environmental affairs and related disciplines with a detailed, intermediate-level perspective on statistical concepts and techniques for analyzing and modeling complex systems. The course content includes estimating the parameters of such models based on existing data, testing hypotheses about these systems, and forecasting. The context of the course is the application of these techniques to problems and policies in public and environmental affairs. Multivariate regression analysis is one of the primary tools for statistical modeling for purposes of policy analysis, program evaluation, simulation of systems, and general forecasting. Thus, most of the course is devoted to single equation regression models and the extension of these models to a variety of situations. A prerequisite for the class is a graduate-level, introductory statistics course that includes coverage of the simple (two-variable) regression model and an introduction to multivariate regression.

The primary learning outcomes for the course are to enable students to:

- explain how regression techniques can be used for data analysis and applied problem solving;
- apply regression techniques to such problems using the appropriate technology;
- explain the assumptions required to use regression analysis, the impact of violating these assumptions, and how to correct for such violations;
- model and forecast the behavior of systems using regression tools;
- comprehend and critically analyze the results of research employing these tools as presented in academic/professional journals, public/non-profit sector reports, and policy analyses; and
- apply several extensions to the basic multivariate regression analysis framework, including logit analysis,
- analyze panel data with fixed and random effects models, and difference-in-differences techniques.

The major topics that will be covered during the semester are:

- the general (multivariate) linear model;
- summary statistics, hypothesis testing, and implicitly linear models;
- multicollinearity, the use of dummy variables, and panel data regression models;
- violations of the general linear model assumptions and errors of specification; and
- logit models

Most of the class content is provided via a modular design that has been implemented in Indiana University’s Canvas teaching and learning environment. Each module is based on a set of class notes, hyperlinked instructional videos that explain specific topics in greater depth, screen capture videos that illustrate several of the major regression techniques and problems, and the accompanying readings in the required textbook. Students are expected to do the readings prior to accessing each module, and may find that reviewing the readings again after completing each module is necessary.
There is also a Tutorial Session with the instructor, scheduled for 7:00 to 8:30 pm every Tuesday evening, during which additional explanations of critical course concepts will be provided along with demonstrations of applying the SAS software to address regression issues. In addition, there is also a Lab Session on Thursday evenings from 7:00 to 8:30 pm, during which the TA will provide assistance with learning SAS as applied to the course content, additional illustrations of concepts, homework assignments, review sessions for exams, and going over the answers for the homework exercises and midterm. Students must either attend the respective Zoom Meetings in person or view the recordings of these meetings. Both of these Tutorial/Lab sessions will allow for student, instructor, and TA interaction via the Zoom software that is utilized for the SPEA Connect program. The Tutorials and Lab Sessions will be digitally recorded and made available for any students who cannot be virtually present for the actual Zoom Meetings. We will also make use of Canvas and email for student-to-student, student-to-instructor, and student-to-TA interaction. Students can use the forum to post questions, address issues, and assist in establishing groups for homework exercises and the project.

All scheduled office hours for both the instructor and the TA will utilize the Zoom software and will be publicly accessible so that multiple students may participate simultaneously. These can be joined at the start of the office hour session or at any time during that period. Office hours by appointment will be private, one-on-one sessions unless a group meeting is specifically requested.

Course Requirements and Grading Criteria

There will be two examinations, each equally weighted. These exams are scheduled for the weeks of Feb 27 and May 2. Each exam will combine multiple choice questions, short answer essays, and problems. The exams will be comprehensive to the extent that concepts covered in previous sections of the course are required to understand those covered later. The exams will be open book and open note, and delivered in an online format. Once a student begins the exam he/she will have two and one-half hours to complete it. Students are on their honor to not collaborate with anyone else on the exams.

Six homework exercises will be assigned during the course. These will utilize computer software (primarily SAS Version 9.4 via IUanyWare) to apply the regression techniques covered in the lecture and readings. Students are required to work collaboratively in teams of two on these homework assignments, using Zoom, phone conferencing, Skype, or other technologies to overcome the impediments of distance. For those students who may be in the same geographic location, face-to-face interaction to complete the exercise is also acceptable. The answers to the assignment will be submitted by one student in the team on behalf of both students. However, each student should be able to successfully utilize the problem-solving approaches used on the homework assignments. Late homework assignments will be accepted with a 10 percent penalty per day late, up to three days (including weekends). Assignments will be due in the Canvas Assignments Area by 5:00 p.m. on the date specified. Homework answers must be word-processed and professional quality.

A major project will be required which applies multivariate regression analysis to a data set that student will be responsible for identifying and assembling. Detailed information on the project will be provided at the beginning of March in a separate document in Canvas. Students are required to work in groups of two or three on this project, with Zoom and other technologies again utilized to allow for appropriate communications and interaction.

Students must be familiar with desktop computing software (including Microsoft Word and Excel), IUanyWare, IU email, and Canvas. For those students who have not used Zoom, instructions for doing so will be provided prior to the first Tutorial on January 10th. Email will serve as an important means of communication for the class. Each student must also have access to appropriate computer technology and the Internet, as specified in the SPEA Connect MPA requirements.
The following grading criteria will be used:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Midterm Exam</td>
<td>20%</td>
</tr>
<tr>
<td>Final Exam</td>
<td>20%</td>
</tr>
<tr>
<td>Data Set Analysis Project</td>
<td>25%</td>
</tr>
<tr>
<td>Homework Exercises</td>
<td>30%</td>
</tr>
<tr>
<td>Class Participation</td>
<td>5%</td>
</tr>
</tbody>
</table>

Text and Reading Assignments


Additional readings identified in the syllabus will be available via Canvas. Note that the course syllabus is somewhat flexible and additional topics or content may be added depending on the interests of the class, or as additional helpful material becomes available.

---

**LEARNING OUTCOMES BY MODULE**

**SECTION 1: THE GENERAL (MULTIVARIATE) LINEAR MODEL**

**Module 1-1: The General Linear Model (GLM)**

1. Describe the structure of the two-variable and general linear regression functions.
2. Explain the difference between the Population Regression Function (PRF) and the Sample Regression Function (SRF), and the difference between the random error term in the PRF and the residual term in the SRF.
3. Explain the formulas for the two-variable and GLM regression functions.
4. Interpret the parameters of the PRF and the parameter estimates of the SRF.

**Module 1-2: Assumptions and Estimation of the GLM**

1. Explain the six assumptions of the GLM and state these in non-mathematical form.
2. Explain the Ordinary Least Squares (OLS) estimation procedure and describe the major elements of this procedure.
3. Explain the concept of a Best Linear Unbiased Estimator (BLUE) as it relates to the GLM.
4. Explain the concepts and importance of the variance, estimated variance, and the standard errors of the OLS parameter estimates.
5. Describe how the true standard error of the parameter estimates differs from the estimated standard error, and apply the formula for the estimated standard error.
6. Estimate the GLM sample regression function and interpret all parameter estimates and standard errors.

**Module 1-3: Strength of Relationship and Hypothesis Testing**

1. Describe the derivation of the Coefficient of Multiple Determination (R-squared) and its importance, and interpret its value and practical significance.
2. Explain the limitations of the raw R-squared and when to use the adjusted R-squared, and interpret the value of the adjusted R-squared.
3. Test the overall multivariate regression for statistical significance, interpret the null and alternative hypotheses, reject or not reject the null hypothesis, apply and interpret the F-statistic for this purpose.
4. Test each parameter estimate for statistical significance, explain what rejecting the null hypothesis for each partial regression coefficient means, and apply and interpret the t-statistic for this purpose.
5. Utilize the p-value for hypothesis testing of regression parameter estimates with the t-statistic, and for hypothesis testing of the overall regression model using the F-statistic.
6. Test for statistical significance across two multivariate regression equations with the same dependent variable.
7. Apply SAS PROC REG to carry out a multivariate regression.

Module 1-4: Implicitly Linear Models and Variable Transformations

1. Describe what makes a regression equation “implicitly linear” rather than truly nonlinear.
2. Apply standard algebraic transformations to regression equations, particularly those related to logarithmic and semi-logarithmic functions.
3. Interpret the parameter estimates of regression equations that utilize logarithmically transformed variables in terms of percentage change and elasticities.

Module 1-5: Multicollinearity

1. Explain the concept of perfect multicollinearity and how it relates to linear dependence, and why the presence of perfect multicollinearity results in the failure of the OLS technique to generate regression parameter estimates.
2. Describe the concept of near multicollinearity, why it is more problematic than perfect multicollinearity, and the implication of near multicollinearity for the standard errors of the parameter estimates and their t-statistics.
3. Explain why dropping relevant variables from a regression equation is not an acceptable solution, even if it makes the near multicollinearity disappear.
4. Apply appropriate tests for near multicollinearity.
5. Solve tractable near multicollinearity problems.
6. Explain the consequences of, and subsequent appropriate actions for, intractable near multicollinearity problems.
7. Apply SAS PROC REG to test for near multicollinearity issues.

Module 1-6: Dummy Variables

1. Explain the advantages and issues involved in using dummy (dichotomous or binary) variables as explanatory elements in a regression analysis.
2. Interpret the parameter estimates for both dummy variables and interaction terms incorporating both dummy and quantitative variables.
3. Specify a regression equation with the proper number and coding of dummy variables to represent categorical variables, to control for seasonality, and to allow for piecewise linear regression.

Module 1-7: Panel Data: Pooled OLS, Fixed Effects, and Random Effects Models

1. Explain the difference between pooled cross-sectional/time-series data and panel data.
2. Describe why fixed effects and random effects models are important in analyzing panel data.
3. Apply and interpret both a fixed effects model and a random effects model with SAS PROC PANEL to analyze panel data, and identify the model that best fits the circumstances utilizing the Hausman Test.
4. Carry out and interpret a Difference-in-Differences analysis as a quasi-experimental research design using regression.

SECTION 2: VIOLATIONS OF ASSUMPTIONS AND OTHER PROBLEMS IN THE GLM

Module 2-1: Heteroscedasticity

1. Describe the concept and causes of heteroscedasticity.
2. Describe the effects of heteroscedasticity on the General Linear Model, particularly those related to the minimum variance property, biased estimates of the standard errors of the parameter estimates, and the t and F statistics.
3. Explain the theory behind Weighted Least Squares (WLS) estimation and how this may correct for heteroscedasticity, and apply the WLS estimation technique for transforming the data to remove the heteroscedasticity and then estimating the transformed model with OLS.
4. Specify appropriate ad hoc assumptions that can be implemented via WLS to potentially eliminate heteroscedasticity.
5. Apply appropriate tests for heteroscedasticity via SAS and Stata, including White’s test and the Goldfeld-Quandt test.
6. Apply the WLS technique to eliminate heteroscedasticity from actual data sets using the SAS PROC REG procedure and supplemental SAS code for carrying out such corrections, and the analogous procedures in Stata.
7. Apply the SAS PROC MODEL procedure and the analogous procedure in Stata to eliminate heteroscedasticity when possible, or to generate White’s Heteroscedasticity Consistent Errors to correct the t and F statistics when the heteroscedasticity cannot be eliminated.

Module 2-2: Autocorrelation

1. Explain the concept and causes of autocorrelation, and describe its effects on the General Linear Model, particularly those related to the minimum variance property, biased estimates of the standard errors of the parameter estimates, and the t and F statistics.
2. Apply the WLS estimation technique by transforming the data to remove the autocorrelation and then estimating the transformed model with OLS.
3. Apply the Durbin-Watson test for autocorrelation via SAS and Stata, and understand how to interpret its values.
4. Execute the Cochrane-Orcutt iterative procedure for estimating the autocorrelation coefficient, transforming the data via this coefficient to remove the autocorrelation, refining the estimate of this coefficient, and then estimating the transformed model to implement the WLS procedure.
5. Apply the WLS technique to eliminate autocorrelation from actual data sets using the SAS PROC AUTOREG procedure and the analogous procedure in Stata.

Module 2-3: Errors of Specification and Proxy Variables

1. Describe the impacts of omitting a relevant variable or including an irrelevant variable in a regression with respect to the properties of the estimators, estimates of the standard errors, and t-statistics.
2. Carry out a standardized procedure for appropriately specifying a regression equation, and working from the initial model specification to the final model specification.
3. Explain the rationale for using a proxy variable in a regression, and the effects of using a proxy variable on the properties of the estimators, estimates of the standard errors, and t-statistics.
4. Identify and deploy appropriate proxy variables in a multivariate regression.

SECTION 3: ALTERNATIVE FORMS AND EXTENSIONS OF REGRESSION ANALYSIS

Module 3-1: The Linear Probability and Logit Models

1. Describe the linear probability model and estimate such a model with SAS and Stata.
2. Describe the limitations of the linear probability model with respect to violations of GLM assumptions and other issues, and explain why OLS is inappropriate for use with a dichotomous dependent variable.
3. Interpret the logit model and explain why it is superior to the linear probability model for dichotomous dependent variables.
4. Estimate a logit model using SAS PROC LOGISTIC and the equivalent Stata procedure; interpret the parameter estimates, the Log-Likelihood statistic, Wald Chi-Square statistics, the pseudo-R-squared, odds ratios, Akaike Information Criterion, and concordance/discordance outcomes.
5. Explain why it is necessary to employ illustrative values of the independent variables to obtain predicted values of dependent variable probabilities, and use these predicted values to assist with the interpretation of logit model estimation results.

Module 3-2: Autoregressive and Distributed Lag Models

1. Explain the concept of lagged variables in a regression, and why lagged variables are useful or necessary in the context of correctly specifying the regression equation.
2. Explain the concepts of autoregressive and distributed lag models, the differences between these, and the impacts of using such models on the properties of the estimators, estimates of the standard errors, t-statistics, and Durbin-Watson statistic.
**WEEK MODULE, TOPICS, READINGS, AND EXERCISES**

**SECTION 1: THE GENERAL (MULTIVARIATE) LINEAR MODEL**

**Jan 9-15 Module 1-1: The General Linear Model (GLM)**
Readings: G&P, scan\view pp. 34-48, 55-85, 97-102, and 107-134.

Jan 10 - Tutorial: Introduction to the Class, The General Linear Model
Jan 12 - Lab: Review of SAS

**Jan 16-22 Module 1-2: Assumptions and Estimation of the GLM**
Readings: G&P, pp. 188-196.

**Module 1-3: Strength of Relationship and Hypothesis Testing**

Jan 17 - Tutorial: Assumptions, Estimation and Strength of Relationship
Jan 19 - Lab: SAS Examples and Interpretation of Regression Output

Exercise 1 Assigned (Due Monday, Jan 30, by 5:00 p.m.)

**Jan 23-29 Module 1-4: Implicitly Linear Models and Variable Transformations**
Readings: G&P, pp. 159-173 and 207-213.

Jan 24 - Tutorial: Hypothesis Testing in the GLM
Introduction to Implicitly Linear Models

Jan 26 - Lab: Hypothesis Testing in SAS; Exercise 1 Help Session
<table>
<thead>
<tr>
<th>WEEK</th>
<th>MODULE, TOPICS, READINGS, AND EXERCISES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 30 - Feb 5</td>
<td><strong>Module 1-5: Multicollinearity</strong></td>
</tr>
<tr>
<td></td>
<td>Readings: G&amp;P, pp. 320-337.</td>
</tr>
<tr>
<td></td>
<td>Jan 31- Tutorial: Implicitly Linear Models and Variable Transformations Intro to Multicollinearity</td>
</tr>
<tr>
<td>Feb 2</td>
<td>Lab: Implicitly Linear Models in SAS; Exercise 1 Reviewed</td>
</tr>
<tr>
<td>Exercise 2 Assigned (Due Monday, Feb 13, by 5:00 p.m.)</td>
<td></td>
</tr>
</tbody>
</table>

| Feb 6-12 | **Module 1-5: Multicollinearity (continued)** |
|          | Readings: G&P, pp. 337-351. |
|          | Application Article Assigned via Canvas. |
| Feb 7    | Tutorial: Multicollinearity |
| Feb 9    | Lab: Testing for Multicollinearity; Exercise 2 Help Session |

| Feb 13-19 | **Module 1-6: Dummy Variables** |
| Feb 14    | Tutorial: Dummy Variables |
| Feb 16    | Lab: SAS Intermediate Lab; Exercise 2 Reviewed |
| Exercise 3 Assigned (Due Monday, Feb 27, by 5:00 p.m.) |

| Feb 20-26 | **Module 1-7: Panel Data - Pooled OLS, Fixed Effects, and Random Effects Models; Difference-In-Differences Models** |
|           | Application Article Assigned via Canvas. |
| Feb 21    | Tutorial: Panel Data |
| Feb 23    | Lab: Panel Data in SAS; Exercise 3 Help Session |
WEEK MODULE, TOPICS, READINGS, AND EXERCISES

Feb 27 - Mar 5 Review and Midterm Exam

Feb 28 - (Tues) Lab: Midterm Review; Exercise 3 Reviewed
(note there is no Tutorial this week)

Mar 2-4 - (Thursday-Saturday) Midterm Examination via Canvas, 2.5 hour
time period selected by student

SECTION 2: VIOLATIONS OF ASSUMPTIONS AND OTHER PROBLEMS IN THE GLM

Mar 6-12 Module 2-1: Heteroscedasticity


Mar 7 - Tutorial: Introduction to Heteroscedasticity
Data Set Analysis Project Assigned
(Due Friday, April 28, by midnight)

Mar 9 - Lab: Midterm Exam Returned, Data Set Analysis Project Example

Mar 13-19 Spring Break

Mar 20-26 Module 2-1: Heteroscedasticity (continued)


Mar 21 - Tutorial: Heteroscedasticity

Mar 23 - Lab: Testing and Correcting for Heteroscedasticity

Exercise 4 Assigned (Due Monday, April 3, by 5:00 p.m.)
<table>
<thead>
<tr>
<th>WEEK</th>
<th>MODULE, TOPICS, READINGS, AND EXERCISES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mar 27-Apr 2</td>
<td><strong>Module 2-2: Autocorrelation</strong></td>
</tr>
<tr>
<td>Mar 28</td>
<td>Tutorial: Introduction to Autocorrelation</td>
</tr>
<tr>
<td>Mar 30</td>
<td>Lab: Testing for Autocorrelation; Exercise 4 Help Session</td>
</tr>
<tr>
<td>Apr 3-9</td>
<td><strong>Module 2-2: Autocorrelation (continued)</strong></td>
</tr>
<tr>
<td>Apr 4</td>
<td>Tutorial: Autocorrelation</td>
</tr>
<tr>
<td>Apr 6</td>
<td>Lab: Correcting for Autocorrelation; Exercise 4 Reviewed</td>
</tr>
<tr>
<td></td>
<td>Exercise 5 Assigned (Due Monday, April 17, by 5:00 p.m.)</td>
</tr>
<tr>
<td>Apr 10-16</td>
<td><strong>Module 2-3: Errors of Specification and Proxy Variables</strong></td>
</tr>
<tr>
<td>Apr 11</td>
<td>Tutorial: Specification Errors</td>
</tr>
<tr>
<td>Apr 13</td>
<td>Lab: Data Analysis Project Help Session; Exercise 5 Help Session</td>
</tr>
</tbody>
</table>
SECTION 3: ALTERNATIVE FORMS AND EXTENSIONS OF REGRESSION ANALYSIS

Apr 17-23  Module 3-1: The Linear Probability and Logit Models

Readings: G&P, pp. 552-566.

Apr 18 - Tutorial: Logit Models

Apr 20 - Lab: SAS Examples of Logit Models; Exercise 5 Reviewed; Exercise 6 Help Session

Exercise 6 Assigned (Due Monday, April 24, by 5:00 p.m.)

Apr 24-30  Module 3-2: Autoregressive and Distributed Lag Models

Data Analysis Project PowerPoint Presentations and Reports


Apr 25 - Tutorial: Autoregressive and Distributed Lag Models

Research Presentation

Apr 27 - Lab: Final Exam Review; Exercise 6 Reviewed

Apr 28 – (Fri) Data Analysis Project Report and Executive Summary

Due by midnight

May 3-6  (Wed-Sat) Final Examination via Canvas, 2.5 hour time period selected by student