R POS 417 Z Research Methods in Political Science II
Spring 2003
Tuesday & Thursday 8:15 - 9:35 in Lecture Center 12
Call # 7459

Instructor: Frederick Wood
E-mail: fw7855@albany.edu
Class Website: http://www.albany.edu/~fw7855/rpos417Z.htm

Office Hours: Monday 1:30pm - 2:30pm & Thursday 9:45am - 10:45am at the Political Science Contact Office in Ten Broeck Hall, Dutch Quad. I am also available by appointment.

Course Description: After a review of bivariate and multivariate OLS regression, we will examine how to diagnose and adjust for when your data violates the regression assumptions.

Prerequisite: You should have taken R POS 416 or a similar introductory statistics class that covered measurement, probability, sampling, and hypothesis testing. If you have not taken RPOS 416, you must meet with me to discuss whether you should take this class. This class meets the upper level writing intensive requirement and is required for the Public Policy major.

Texts and Materials:
- Kennedy, Peter. A Guide to Econometrics. MIT Press. 1998 Paperback ISBN: 0262611406. This book is not required for the class but it will serve as a useful reference in the future, especially if you are going to resell the Gujarati text or go on to graduate school.
- Articles from ERes [http://eres.ulib.albany.edu] are denoted by a “~”. (password: frink)
- A Zip Disk or several 1.44 MB floppy disks
- Calculator (one that has more than four functions)

Software: We will be using SPSS as the statistical package for this class. A student version is available for purchase; however, it is limited in the range of functions and number of cases that can be processed. For a number of classes we will convene in one of the computer labs on campus to allow you to become familiar with the program. I will also schedule and attend extra lab sessions to allow you to work on your research paper and class assignments.

Grading (A - E):
Participation 12%
6 Statistical Exercises at 7% each 48%
Term Paper Writing Process (hypothesis, literature review, research design, data collection, data processing/diagnosis, presentation, & final draft) 40%
Assignment # 1 Regression by hand review
Assignment # 2 Practicing Matrix Algebra
Assignment # 3 Heteroscedasticity
Assignment # 4 Autocorrelation
Assignment # 5 Collinearity
Assignment # 6 Outliers & Leverage

*** You can all thank Professor Brian Pollins from the Political Science department at The Ohio State University for introducing me to these exercises and datasets at the ICPSR Summer Program. ** Late papers or assignments can be allowed with the proper documentation from the Dean of Undergraduate Studies in LC 30. *** You must complete all of the assignments in order to receive a final grade for the course. If you are missing one or more assignments, you will be assigned an incomplete, regardless of whether or not you would pass the course with failing grades for the missing assignments. All work must be handed in by Friday, May 9th. I reserve the right to give a failing grade for excessive absences. **** Students with disabilities are responsible for making their needs known to the instructor, and seeking available assistance in a timely manner. ***** This syllabus should not be construed in any way as a contract. It is my intention to keep with the schedule laid out in this syllabus but I reserve the right to change or course topics, assignments, readings, and deadlines. Any changes will be announced in class.

Attendance Policy: While not technically included in the calculation of your grade, your attendance in the class and any computer lab sessions is essential for a number of reasons. First, not all of the material or techniques covered will come from assigned readings. Second, the required assignments will be periodically discussed in class and failure to complete the assignments on time will result in a grade of zero for that assignment. Finally, those students whose final grade resides on or near a border between grades may be rewarded for regular attendance, and it is impossible to participate in the class without physically being in the meeting room.

For those of you who will attend, I ask that you follow a few simple policies. First, respect the fact that this class has a definite start and end time. Be in your seats when class is due to begin and do not leave them until the class is over (usually denoted when I say “see you all next time”). Second, turn off your cell phones when you enter the classroom. You do not want me using your minutes while I have a conversation with the person who called you. Third, you should come to class prepared to pay attention and take notes. You will be asked to leave the room if you are listening to music, reading unrelated material, talking to other students around you, or sleeping. Finally, the reproduction, distribution, or publication (on-line or otherwise) of transcribed or taped versions of my lectures are not permitted without my consent. I thank you in advance for following these guidelines.

Writing Intensive Properties: These courses use writing as an important tool in the discipline studied and are not designed primarily to teach the technical aspects of writing. The emphasis is on using writing as a means of sharpening critical thinking in and understanding of the subject. Approved courses must meet each of the following four criteria:

1. A Substantial Body of Finished Work: a total of 20+ double-spaced pages (paper + assignments)
2. Opportunity for Students to Receive Assistance in Progress: visits to the Writing Center (HU-140) or conferences with the instructor.
3. Opportunity to Revise Some Pieces: this will be accomplished during the research paper process.
4. Response to Student Writing: extended comments from the instructor.
Paper Format: The term paper will be between 13 and 16 full pages in length, double spaced, 12 point Times New Roman font, with 1 inch margins all around. On a title page, you will include your name, date, my name, the name of the class, day and time the class meets, and a title for your essay if you wish. There is no need to place your paper in a cover or folder, just use a staple. Spelling, grammar, and the format proscribed above will be considered in the grading of your paper. Papers that do not conform to the above format may be returned to the student. All students should photocopy or have a disk copy of all work handed in during class. Late papers and assignments will be penalized in proportion to their lateness. All students are responsible for learning what constitutes plagiarism and the proper procedures for the citation of other’s works.

Students will make a brief Power Point presentation of their term paper’s findings in class. Presentations will last approximately 5 minutes and will be followed by 5 minutes of class discussion and questioning. You must present your findings to the class in order to hand in your paper for grading.

Course Outline- due to the nature of the course we will not follow a rigid reading schedule. You should stay ahead of the material as it is discussed in class. You should be working continuously on your term paper.

1. Review Bivariate Regression and Correlation
Read: Gujarati Chapters 1 – 6.
Applications:

2. Multivariate Regression
Read: Gujarati Chapters 7 – 8 & Appendix C.
Applications:

3. Heteroscedasticity & Autocorrelation
Read: Gujarati Chapters 11 – 12.
   ~ Downs, George & David Rocke (1979)“Interpreting Heteroscedasticity” APSR 23: 816 – 823
Applications:

Read: Gujarati Chapters 10 & 13.
Fox Sections 3.1, 3.2

5. The Analysis of Residuals, Outliers, Leverage, Non-Linearity, Non-Normality
Read: Fox Sections 4, 5, 7, and Appendices 4.1 - 4.4 & 6.1

Assignment # 1
Wood: Spring 2003 RPOS 417 Z

Regression review by pencil & calculator.

The following exercises should be calculated without the use of statistical programs such as Excel or SPSS. Pencil and paper should be used to show your work. Please highlight your final answers and use paragraphs where an explanation is necessary.

1. In studying the effect of military spending on Gross National Product within the European Community. You have data for all of the EU members. Does this mean that you have data for a sample or for the population and explain why? (10 pts)

2. Use the following data to estimate the relationship between Birth Rates, Per Capita Income, and Race for nine U.S. regions. Estimate BirthRate = b1 + b_income + b_race + e as a multivariate regression.

<table>
<thead>
<tr>
<th>Region</th>
<th>Live Births/1000 Pop.</th>
<th>Income/Cap ($)</th>
<th>% African-American</th>
</tr>
</thead>
<tbody>
<tr>
<td>New England</td>
<td>22.5</td>
<td>3998</td>
<td>3.3</td>
</tr>
<tr>
<td>Middle Atlantic</td>
<td>21.5</td>
<td>4193</td>
<td>10.6</td>
</tr>
<tr>
<td>East North Central</td>
<td>24.2</td>
<td>3928</td>
<td>9.6</td>
</tr>
<tr>
<td>West North Central</td>
<td>24.0</td>
<td>3492</td>
<td>4.3</td>
</tr>
<tr>
<td>South Atlantic</td>
<td>24.2</td>
<td>3304</td>
<td>20.9</td>
</tr>
<tr>
<td>East South Central</td>
<td>24.4</td>
<td>2724</td>
<td>20.3</td>
</tr>
</tbody>
</table>
West South Central | 25.4 | 3122 | 15.8
Mountain | 27.3 | 3267 | 2.2
Pacific | 23.6 | 4134 | 5.7

a. Present your calculations for each of the three model parameters. (15 pts)

Hint: If your estimate for $b_1$ is less than 35.0 or greater than 38, or if you estimate for $b_2$ is less than -.01 or greater than 0.0 or if your estimate for $b_3$ is less than –0.20 or greater than 0.0 then you have miscalculated.

b. Once you have calculated the parameters, describe verbally what $b_1$, $b_{\text{Income}}$, & $b_{\text{Race}}$ say about the relationship between income, race and births. (15 pts)

c. If income were measured in thousands of dollars instead of dollars would your coefficient estimates from part a change? If so, how? (10 pts)

d. Given your estimates from part a, calculate the residuals for the nine regions. Do you see any patterns in these residual that suggest another explanatory factor might be added to this model? Hint: concentrate on birthrates. (10 pts)

3. Imagine you have calculated the following estimates:

\[ Y = 21.7 + 0.83X_2 + 1.07X_3 + e \]
\[ R^2 = 0.34 \]
\[ (0.38) \quad (0.78) \quad N = 25 \]

Values in parentheses are the standard errors of the partial slope coefficients.

a. Test the following hypothesis: $H_0 : 2 = 0$, $H_0 : 3 = 0$. Did you choose a one-tailed or two-tailed test? Why? On what basis did you choose your significance level? (20 pts)

b. What would be the change in the predicted value of $Y$ if $X_2$ increased by 14 units? (10 pts)

c. Is your answer to part b affected by the value of $X_3$, or is your answer independent of values of $X_3$? (10 pts)
Practicing Matrix Algebra!!!
The following exercises should be calculated without the use of statistical programs such as Excel or SPSS. Pencil and paper should be used to show your work. Please highlight your final answers and use paragraphs where an explanation is necessary.

1. (15 pts.) Let:

\[
\begin{align*}
A &= \begin{pmatrix}
5 & 4 & 7 \\
-3 & 1 & 2 \\
8 & 0 & 4
\end{pmatrix}, \\
B &= \begin{pmatrix}
1 & 1 & 2 \\
2 & 0 & 9 \\
6 & 8 & 7
\end{pmatrix}, \\
C &= \begin{pmatrix}
1 & 2 & 6 \\
4 & 3 & 9 \\
8 & 0 & 4
\end{pmatrix}
\]

a. What is the order of each Matrix
b. Find \( A + B \)
c. Find \( AB \)
d. Find \( CA \)
e. Find \( A', B', \) and \( C' \) (the ‘ means transpose)

2. (15 pts.) Let:

\[
\begin{align*}
A' &= \begin{pmatrix}
6 & 7 & 8 & 1 \\
4 & 5 & 2 & 3
\end{pmatrix}, \\
B &= \begin{pmatrix}
7 & 2 & 9 \\
9 & 3 & 0
\end{pmatrix}, \\
C &= \begin{pmatrix}
3 & 4 & 8 \\
1 & 9 & 4 \\
0 & 3 & 0
\end{pmatrix}
\]

a. Find \( AB \) Hint: You have been given \( A' \) not \( A \)
b. Find \( BC \)
c. Find \( C'C \)
d. Find \( CC' \)

3. (10 pts.) Let:

\[
\begin{align*}
X &= \begin{pmatrix}
1 & 4 & 2 \\
1 & 3 & 6 \\
1 & 9 & 0
\end{pmatrix}, \\
y &= \begin{pmatrix}
7 \\
4 \\
3
\end{pmatrix}
\]

a. Find \( X'X \)
b. Find \( X'y \)
4. (10 pts.) Show that if

\[ A = \begin{pmatrix} 2 & 3 \\ 1 & 4 \end{pmatrix} \quad \text{then} \quad A^{-1} = \begin{pmatrix} 4/5 & -3/5 \\ -1/5 & 2/5 \end{pmatrix} \]

Hint: show how to compute \( A^{-1} \) and show that \( AA^{-1} = I \)

5. The following data was collected to examine the relationship between international cooperation (C) and trade (T).

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Case 2</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>Case 3</td>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>

a. Estimate the parameters of the model \( T_1=b_1+b_2C_i=e_i \) using the matrix formula \( b = (X'X)^{-1}X'y \) (10 pts)

Hint: Do not forget that your model includes and intercept term. This means that a vector of 1’s will be inserted into your data matrix. \( b_1 : 2.0 < b_1 < 3.0 \) & \( 1.0 < b_2 < 1.5 \)

b. Calculate the model’s predicted values for Trade (T) (5 pts)

Hint: \( \hat{y} = Xb \) & \( 8.0 < \hat{y} < 8.5 \)

c. Calculate the vector of residuals (5 pts.)

Hint: \( e = y - \hat{y} \) & \( 0.0 < e < 0.2 \)

d. Calculate the standard error of the estimate. (5 pts.)

Hint: \( s_e = \sqrt{s_e^2} \); where \( s_e^2 = (e'e) / df \) and where \( e'e \). Finally, \( 0.30 < s_e < 0.50 \).

e. Calculate the standard error of the model parameters. (5 pts)

Hint: the variance of parameter estimates is found on the main diagonal of the matrix \( s_e^2(X'X)^{-1} \) and \( 0.3 < s_{b1} < 0.5 \) & \( 0.01 < s_{b2} < 0.10 \)

f. Calculate the t-statistics for model parameters (5 pts.)
g. Given that $\Sigma(y - \bar{y})^2 = 32.6666$, calculate $R^2$ (5pts.)

Hint $R^2 > 0.98$

h. What have you learned about the relationship between International Cooperation and Trade? (10 pts.)

Hint: Discuss overall model results and individual parameters. Include the strength, direction and significance of the relationships, and interpret your coefficients. Ignore that fact that your N is very low to make inferences.

Assignment # 3

Wood: Spring 2003  RPOS 417 Z  

**Heteroscedasticity**

1. Use the PRESPOOL data to estimate $Y_4 = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$. Begin with OLS and save your residuals.

2. Use any two of the techniques discussed to determine whether heteroscedasticity is present, and to identify the “offending $X$”. The techniques include residual plots, condition number of the matrix, the Cook-Weisberg test, the Glejser test, the White test, the Goldfeld-Quandt, and the Breusch-Pagan-Godfrey test. You are not required to use all of these tests, just two.

3. Once you are satisfied you have identified the proper $X$ variable and its functional relationship to the error variance, transform your data accordingly, re-estimate the model, check the new residuals to see they are white noise, and compare your GLS results to your original OLS results.
You have three possible sets of observations on your dependent variable, given that different polls yield different results (and, I will tell you, different time-dependent problems in the error terms). These three are labeled "Harris", "Gallup", or "Caddell". You are required to analyze the model using the "Harris" data.

Model for the "Harris" Autocorrelation Problem: $Y_1 = b_1 + b_2X_2 + b_3X_3 + b_4X_4 + e$

Professor Pollins has painstakingly collected data concerning macroeconomic performance in the U.S. and Presidential Popularity; there are 60 observations in the set. The data set name is PRESPOULL. The variables are as follows:

Y1: The Harris Method. Sixty annual observations of presidential popularity (suspected to make Dems look good). Measure: % respondents who agreed, "This is a really great President!"

Y2: The Gallup Method. Sixty annual observations of presidential popularity (believed to tilt toward Republicans). Measure: % respondents who agreed, "This is a really great President!"

Y3: The whatever-happened-to-Pat Caddell method. Sixty annual observations of presidential popularity designed to make Harvard Johns look like brilliant campaign consultants. Measure: % respondents who agreed, "This is a really great President!"

Y4: The 1-976-GABB Method. Sixty cross-sectional observations. One caller from each of the 50 states and 10 Canadian provinces was asked their assessment of the President. Measure: Response given to question: "On a scale of 0 - 100, where 100 is your favorite President ever, how would you rate the individual now in office?"

For the time-series problems, each of the three independent variables was measured on the same years that the polls were taken. For the heteroscedasticity problem, they were
measured cross-sectionally on 12/5/97 in each of the fifty states and ten provinces. Just use the same independent variables for the autocorrelation and the heteroscedasticity exercises. The independent variables are:

X2: ECONACT, A general index of "Economic Activity", based on GNP or Capacity Utilization, or some such measure. Larger Numbers mean more economic activity. An increase of one unit corresponds to a 0.25 point increase in the growth rate of GDP. To gauge the substantive effect, imagine that GDP growth were to improve by one point, say from 3% to 4%.

X3: INTOPT, An index of optimism/pessimism in capital markets. This is a much more sophisticated measure of the money side of the economy than simple interest rate indicators often cited by trenchant analysts like the Mclaughlin Group or G. Gordon Liddy. Higher numbers mean greater optimism. An increase of one unit corresponds to a 0.25 point decrease in the inflation rate. To gauge the substantive effect, imagine that inflation were to decline by one point, say from 2% to 1%.

X4: UNEMP, An employment indicator that the House Labor Committee refuses to even acknowledge because it yields results that contradict Reaganomics. This data is particularly despised by Newt Gingrich and other knights of the new War on Poverty because it assumes that poor people are actually willing to work, and therefore should be classified as "unemployed" rather than as "shiftless welfare cheats sponging off deservedly super-rich American patriots". Higher numbers mean the general employment situation is worse than when low values of the measure are observed. An increase of one unit corresponds to a 0.25 point increase in the unemployment rate. To gauge the substantive effect, imagine that unemployment were to rise by one point, say from 5% to 6%.

Prevailing theory in the field of American Politics tells us that presidential popularity should rise and fall in step with macroeconomic performance. I expect you to surmise what this implies about the expected direction of each of the partial slope coefficients in the model.

**Autocorrelated Residuals in a Time Series**

Using the "Harris" data, take the following six steps:

**Step 1)** Begin with OLS, and save your residuals.

You could estimate Rho at this point if you were so inclined. Recall that an unbiased estimate of rho can be obtained from the very simple, bivariate model: \( e_t = \rho e_{t-1} + \mu \).
Step 2) Compute autocorrelations between \( e_t \) and \( e_{t-s} \), and from these results construct a correlogram. OR, you can find a menu item in SPSS for windows that will compute and display the ACF and PACF for you.

Step 3) Study these results and formulate a model of the error... AR(1), MA(2), etc. Then, using SPSS for Windows, use the appropriate time-series diagnostics to test this hypothesis.

Step 4) When you are satisfied that you have correctly identified the error process, transform your data accordingly and re-estimate the original model using OLS. **Hint:** A useful transformation statement could be ...

\[
\text{COMPUTE NX2 = X2 - <rho> * LAG(X2);}
\]

I also suggest you not forget to "transform" the intercept by creating a variable \( \text{NINT} = 1 - <\text{rho}> \). **Caution!!** SPSS mishandles the "intercept through origin" option when you are supplying your own vector for the intercept term, as you do here.

Step 5) Compare these results to your original OLS results. Are the new residuals really white noise as they should be? Once you are satisfied that your estimates are BLUE, tell me what happened to the parameters and their standard errors. Why do you suppose it worked out this way?

Step 6) IFF the local, available SPSS package includes a canned GLS routine(s), you should compare your GLS results to those given by a canned GLS routine. You may use the default, canned solution. Feel free to try other options like MLE for sake of comparison. If the local, available SPSS package does not include a canned GLS routine, you are hereby absolved of step 6. Check with me to learn what is available to us for this class.
Collinearity

1. Use the file SINGULAR to estimate the model:

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + e_i \]

Use appropriate diagnostics to decide whether two or more of the independent variables are collinear. Use techniques like variable selection, joint hypothesis testing, factor analysis, and good judgment to decide how to get the best estimates you can.

Informed diagnostics can only take place when you have a good grasp of the theory in your field and knowledge of your cases. This is an exercise in detecting and dealing with collinear explanatory variables, so read the “Theory of Y” below. Remember there is no cure to collinearity. There is not a single correct model or set of estimates for you to find in this problem. Instead, you should explain and describe the procedure that you employed to diagnose the problem, locate the variables that are implicated, and manage the problem that still permits you to contribute to the theoretical debates in your field.

The Theory of Y

Y is a phenomenon of great interest in your field, and several academic names have gained tenure at places we all think we want to teach at by developing a grand Theory of Y. The implications of this theory (which no one has tested empirically) are as follows:

1. Y is a linear, additive function of several explanatory factors, perhaps as many as seven.
2. If factors \( X_1 \) and \( X_3 \) have any significance, it is negative.
3. The Conventional Wisdom in your field accepts that \( X_4 \) matters, and that its associated coefficient should be around 55.0.
4. Various tenure-seekers have published numerous pieces contending that \( X_2, X_5, X_6, \) and \( X_7 \) influence Y positively. \( X_6 \) and \( X_7 \) are particularly controversial, since their purported effect on Y was inspired by the writings of Michel Foucault. The Foucauldians in your field are deeply conflicted: What if the Late Great Deconstructionist himself were vindicated using regression analysis – a “method” we know to be a self-deluding tool of the ruling class (namely Big Ten programs) capable of nothing but the pseudo-concretization of reality for no purpose other than the reproduction of a rapaciously exploitive power structure.

Welcome to academia and the “battle” between theory and empiricism!
Know your data: Outliers and Leverage.

1. Using SPSS and the file ANSCOMBE to estimate the model $y_i = \beta_0 + \beta_1 x_i + e_i$ four times. Regress $y_1$ on $x_1$, $y_2$ on $x_1$, $y_3$ on $x_1$, and $y_4$ on $x_4$. Note the parameter estimates, the error variance, and $R^2$. Next, construct scatterplots for each of these bivariate relationships. Comment. Which of these estimates would you believe? Which fail to convince you and why? (40 pts.)

2. Estimate the model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$ using the variables contained in the data set OUTLIERS. Employ at least two statistics that indicate potential leverage and influence to identify these outliers, if any. Discuss the confidence you have in these parameter estimates, given what you have found out about the influence of particular observations. You may re-estimate the model excluding one or more observations that may have “undue” influence, in your judgment. Discuss the consequences of such exclusion (good and bad). Statistics you may want to use include hat values, internally and externally studentized residuals, Cook’s Distance ($D$) and Belsey, et al.’s DFFITS. This list is not exhaustive. You may also wish to produce a visual diagnostic with partial regression plots. You are not obligated to compute or discuss all these. Just include sufficient information to come to a good conclusion about the leverage and influence of any outliers that you identify. What would you want to know about the substantive theory and/or the individual cases that would help you decide whether to leave them in or exclude them? (60 pts.)

Data file: ANSCOMBE
Number of Observations: 11
Number of Variables: 7
Sequence of Read: CASE, X1, Y1, Y2, Y3, X4, Y4

Data file: OUTLIERS
Number of Observations: 50
Number of Variables: 4
Sequence of Read: Y, X1, X2, X3

Data file: SINGULAR
Number of Observations: 50
Number of Variables: 8
Sequence of Read: Y, X1, X2, X3, X4, X5, X6, X7

Data file: PRESPOLL
Number of Observations: 60
Number of Variables: 8
Sequence of Read: CASE, Y1, Y2, Y3, Y4, X2, X3, X4