

ICPSR Regression (Pollins)

Problem Set 4

Doctor B.L.U.E., Data Detective

NOTE that you are given a different data set for each problem.

AGAIN, you may save time and effort by handing in **annotated** computer output. Simply **annotate** and **highlight your main results** on each page you give me so I can see **what** you did, **how** you did it, and what information you took from each page to use in your analysis.

Know Your Data

Data file: ANSCOMBE

1. [10 points] Use your chosen canned regression package and the data contained in the file ANSCOMBE to estimate the model $y_i = \beta_0 + \beta_1 x_i + e_i$ four times. Specifically, 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?**

Next : Use the matrix routines to calculate **hat values** for the cases in these data sets. Flag those you believe to have high leverage on your parameter estimates. **Hint**: These hat values are simply the diagonal elements in the **H** matrix. And **H** can be computed directly from **X**. Just identify influential cases. You will not be able to recommend treatment, because you have no substantive information about this data. Tell me which cases (if any) would bear closer scrutiny if you had substantive information about the data.

Outliers and Leverage

Data file: OUTLIERS

2. [30 points] 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 which 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 (i.e., the pros and cons). Numbers you may wish to employ include hat values, internally and externally studentized residuals, Cook's Distance

(D) and Belsley, et.al.'s DFFITs. There are others. You may also wish to produce a visual diagnostic with partial regression plots. You are **NOT** obligated to compute or discuss all these. Just give me enough information to come to a good conclusion about the leverage and influence of any outliers you identify, and tell me the kinds of things you would want to know about the substantive theory and/or these individual cases that would help you decide whether to leave them in or exclude them.

Hints: Again, you have *no substantive or theoretical* information on *this* data, so you cannot come to a final judgment regarding the advisability of excluding any observation(s). Just tell me what you would consider in determining which estimates to accept if you *were* familiar with the substantive question under study, and the data in your hands.

Collinearity

Data file: SINGULAR

3. [30 points] Use the data set named SINGULAR to estimate the model:

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \beta_7 X_{7i} + e_i$$

Use appropriate diagnostics to decide whether two or more of the independent variables in this model 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 under these difficult conditions. (You are **not** required to use all the techniques we considered.) **Hint:** Informed diagnostics can only take place when you have a good grasp of theory in your field, and knowledge of your cases. This is an exercise in detecting and dealing with collinear explanatory variables, *so read about the "Theory of Y" in the appendix to this assignment*. And remember, there is no "cure" for collinearity. Therefore, there is not a single correct model or set of estimates for you to find in this problem. Instead, I am looking for you to show me you know the correct **procedure** to **diagnose** the problem, locate the variables that are implicated, and **manage** the problem in a way that **still permits you to make a contribution to theoretical debates in your field**. This is what is meant by **Best Practice** in this case.

Autocorrelation and Heteroscedasticity

Data File: PRESPOLL

4a. [15 points] Autocorrelation. 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.

4b. [15 points] Using the 60 cross-sectional observations in the "GABB" vector, use it as your dependent variable, and correct for the heteroscedasticity you are likely to find.

I have 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

PRESPOLL

The variables are as follows

Political Science is sadly backward, so we have taken precautions in the measurement of Presidential Popularity by constructing alternative indicators. These measures are...

Y_1 : 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!"

Y_2 : The Gallup Method. Sixty annual observations of presidential popularity (believed to tilt toward Republicans). Measure: % respondents who agreed "This is a really great President!"

Y_3 : 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!"

Y_4 : 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:

X_2 : **ECONACT**, A general **index** of "Economic Activity", based on GNP or Capacity Utilization, or some such thing. 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%...

X_3 : **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%...

X_4 : **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 ...

COMPUTE NX₂ = X₂ - <rho> * LAG(X₂); I also suggest you not forget to "transform" the intercept by creating a variable NINT = 1 - <rho>. **Caution!!** SPSS mis-handles the "intercept through origin" option when you are supplying your own vector for the intercept term, as you do here. *IT IS SAFEST TO SIMPLY USE THE CANNED PACKAGE TO TRANSFORM THE VARIABLES, TAKE THOSE TRANSFORMED VECTORS & MATRICES INTO YOUR OLS PROGRAM, AND GET THE RIGHT ESTIMATES!*

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 (Yule-Walker in SAS, Prais-Winsten in Limdep.) 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 or Mr. Sweeney to learn what is available to us for this class.

5 points extra credit: Repeat steps 1-6 using the "Gallup" or "Caddell" data.

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

Heteroscedasticity The "GABB" Observations:

Use the PRESOLL data to estimate $Y_4 = b_1 + b_2X_2 + b_3X_3 + b_4X_4 + e$. As always, we begin with OLS and save our residuals.

Step 1) Begin with the Ordinary Least Squares package in SPSS for Windows, and save your residuals.

Step 2) Use any two of the many techniques we discussed to determine whether heteroscedasticity is present, and to identify the "offending X". These 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.... Again, you are NOT required to use *all* these may diagnostics. Just any two of your choice.

Step 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. **Caution!!** SPSS mis-handles the "intercept through origin" option when you are supplying your own vector for the intercept term, as you do here. *IT IS SAFEST TO SIMPLY USE THE CANNED PACKAGE TO TRANSFORM THE VARIABLES, TAKE THOSE TRANSFORMED VECTORS & MATRICES INTO YOUR OLS PROGRAM, AND GET THE RIGHT ESTIMATES!*

Remember: Always re-check your residuals to make sure you have corrected the problem. You're not happy until those residuals are white noise. And always interpret your results statistically *and* substantively!

You now know **Best Practice** techniques to diagnose autocorrelation and heteroscedasticity. You also now know how to obtain BLUE estimates when these problems are indicated. In your own work, however, **REMEMBER** that the presence of these problems may well indicate model mis-specification, and you should always think substantively about the possible cause of these pathologies. *Do not blindly correct for the problem by pushing canned-package buttons. Remember, in your own work...*

AN OUNCE OF GOOD THEORY IS WORTH SEVERAL POUNDS OF ESTIMATION.

THINK FIRST AND SAVE A TREE

Political Science 686
Pollins Regression Class
"Theory of Y" Appendix
Assignment 4, Problem 3 on Multicollinearity

Y is a phenomenon of great interest in your field, and several academic poohbahs and mooseheads in your field gained tenure at places like Harvard and Stanford (not to mention megabuck NSF grants) by developing a grand Theory of Y. The implications of this theory (which the mooseheads never condescended to test empirically) are as follows:

- 1) Y is a linear, additive function of several explanatory factors, *perhaps* as many as seven ($X_1 \dots X_7$).
- 2) *If* factors X_1 and X_3 have any influence, it is negative. (This is a point of heated controversy that had Professor Imso Selfimpressed writing a scathing, almost *ad hominem* critique of Professor Brainz Bean-Calcified in a recent issue of *The Journal of Irreproducible Results*.)
- 3) The Conventional Wisdom in your field accepts that X_4 matters, and that its associated coefficient should be around 55.0.
- 4) Various small-fry struggling to attain at least tenure, if not a major reputation, have published pieces variously 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 capable of nothing but the pseudo-concretization of reality for no purpose other than the reproduction of a rapaciously exploitive power structure?

Wow! What a mess! (Welcome to academia.) Your mission, should you decide to accept it (as though you had a choice!) is to untangle this disciplinary imbroglio, making and breaking the careers of many others while you claw your way to the top of your field. You *know* you can make a contribution to your field's knowledge about The Theory of Y! Your data awaits....

1. Assume you've been in a matrix subroutine, i.e., in between
MATRIX.
and
ENDMATRIX.

Notes on moving data
and stat results from
SPSS matrix into
SPSS data window.

2. Also assume that in that subroutine you've created matrices named Y, X, PREDICTY, and E.

3. You may save them all to the same file, provided that they all have the same number of rows.

4. The SAVE subroutine must have at least these three elements:

- A. The SAVE command followed by, in curly brackets, the names of the matrices you want to save.
- B. The OUTFILE subcommand, i.e., location to save them. I always use a new file (I think you have to).
- C. The VARIABLE subcommand, i.e., the name those matrices will take on in the new file; each column in a multi-column vector must have a name.

Example:

```
SAVE {Y, X, PREDICTY, E}
  /OUTFILE='C:\Program Files\Spss\ICPSRSession#2\ProblemSet#4\PrespollFixSave.sav'
  /VARIABLES=Y4, INTERCPT, X2, X3, X4, PREDICTY, RESIDS.
```

5. Note: SPSS will automatically create the new file and store the matrices there as column vectors. Each column from the matrix becomes a column (with its own variable name) in the new data file.

6. Note that in the example matrix "Y" has been named variable "Y4." And note that the four columns, of matrix "X" were, in the OUTFILE, each given their own name, i.e, "INTERCPT," "X2," "X3," and "X4." Similarly, note that matrix "PREDICTY" was given the same name in the new OUTFILE, and note that matrix "E" was named "RESIDS."

7. Now, when you want to use it, use the GET FILE command (before the next matrix subroutine) and use the GET command (in the new matrix subroutine) to restructure the matrix for the new subroutine.

Example:

```
GET FILE='C:\Program Files\Spss\ICPSRSession#2\ProblemSet#4\PrespollFixSave.sav'.
```

```
MATRIX.
GET RESIDS
  /VARIABLES=RESIDS
  /NAMES=VARNAMES
  /MISSING=ACCEPT
  /SYSMIS=0.
GET Y
  /VARIABLES=Y4
  /NAMES=VARNAMES
  /MISSING=ACCEPT
  /SYSMIS=0.
GET X
  /VARIABLES=INTERCPT X2, X3, X4
  /NAMES=VARNAMES
  /MISSING=ACCEPT
  /SYSMIS=0.
END MATRIX.
```