

ICPSR Regression (Pollins)
Summer, 2003
Problem Set 3

Mastering Regression: Write Your Own Rosetta Stone
(due in class on Thursday, August 8)

At Last! On to Computers!

You are going to write your own estimation program using matrix algebra. You will use your program to solve the estimation problems you worked by hand in assignment #1.

The matrix algebra routines in SPSS for Windows are adequate to complete this assignment. I will provide you with a menu of useful commands for SPSS.

1. [40 points] Return to problem #2 from assignment 1 (where you looked at the relationship between birthrates, income and race). Write a program in matrix algebra to compute model parameters, their standard errors, and t-statistics. **Note:** To receive the Full 40 Points you must make your code *general*; i.e., not specific to the problem. No big deal. This involves no more than figuring out how to read in an existing data file and getting your program to recognize the number of observations and the number of parameters being estimated.

Hints:

i) The computational steps you followed to solve problem #7a-7f in assignment 2 can be programmed in the matrix routines of SPSS for Windows (or Stata, SAS, LIMDEP, Shazam, Fortran among others. But we only assure support for SPSS).

ii) Concerning the management of your data: It could be easiest to create a data file containing the nine (9) observations given in the appendix to assignment 1 for that problem, then read that file into your program. Alternatively, you could enter the observations for the variables as vectors and matrices within the matrix routine of your chosen package. **Remember:** you need a vector of 1's in your data matrix [**X**] in order to compute the intercept! This hint is central to making your code general for all OLS estimation problems, not just specific to this assignment. Look at your **X** matrix and **THINK**: How does this **X** matrix tell me the number of observations I have for this problem? How does this **X** matrix tell me the number of parameters I am estimating for this problem? Now you know what to instruct your program to look for in order to make this general!

iii) This is much less complicated than you might think. 20-30 lines of active code could be plenty. I am not looking for computational elegance. I just want to see that you can do it. Be liberal with the use of comments in your code, and try hard to make your code general to any problem, not just specific to this one (there are plenty of hints toward this end in the handouts.)

***** To save your time and energy, you MAY hand in annotated computer output. In any case (and every case) just be sure to **label all** your matrices clearly. Also, highlight your main results and key intermediate steps. There are many ways to write this program, and I can only decipher your work if you tell me at EACH step exactly what you believe you are doing. *****

2. [40 points] Use the same data and model that you just worked with in problem 1 of this assignment. In assignment #1, you "stargazed" at the residuals to look for substantively interesting patterns that might help you refine your hypothesis about the relationship between income, race, and birthrates for nine (9) U.S. regions. CODE a qualitative independent variable (i.e. a "dummy" variable) that might capture something new about this relationship (urban vs. rural regions, sun-belt vs rust-belt, catholic vs. non-catholic, whatever...I am flexible.) In other words, try to improve the specification of your model, based on your analysis of the residuals. ADD this new variable to your data set. RE-ESTIMATE your model and compare your results to those obtained for problem 1 in this assignment. Now interpret your results both substantively and statistically. Does your new model do any better? Do your coefficient estimates change very much after you add your new variable? What does this imply about the "independence" between your new variable and those that were in the original model? You should use your matrix routines to **compute and compare the adjusted R^2** across the two equations. Hey, why do you want to compare the adjusted R^2 instead of the regular, good, old R^2 ?

Hints:

i) Some changes in your computer code will probably be necessary since you are now dealing with 3 independent variables, i.e., 4 columns in your matrix [**X**] rather than 3.

ii) I **strongly** advise that you try the following dummy variable in your model respecification. Hypothesis derived from theory: "People in economically depressed regions (the Rust Belt) work longer hours for their meager incomes, and are stressed-out about their low incomes. This worry and stress leads them to have fewer babies." Now, see whether this hypothesis holds water.

Rust Belt = 1

Sun Belt = 0

New England

S. Atlantic

Mid Atlantic

W.S. Central

E.N. Central

Mountain

W.N. Central

Pacific

E.S. Central

iii) Don't forget to **INTERPRET** your coefficients substantively *and* statistically!

3. [10 points] Use your matrix algebra program to solve problem #6 from Assignment 1. Again, this is a problem you solved by hand in assignment 1, using scalar algebra. **So go back and look at that problem.** Recall that you were asked to find b_k , σ_{bk} , R^2 , and σ_e . Show on a page of canned output that your results for b_k , σ_{bk} , R^2 , and σ_e match the canned package's output for b_k , σ_{bk} , R^2 , and σ_e .

Hints:

i) It may be easiest to create a new data file containing the observations given for this problem, then read them into your matrix routine. Or you might find it to be more convenient to enter the observations in matrix form at the beginning of your program. That choice is yours.

ii) Remember that the number of rows and columns in the matrix will correspond to the number of observations and variables. These will be different from the previous problem.

4. [10 points, or more for the ambitious] Use the canned regression program in SPSS for Windows (or PROC GLM or PROC REG in SAS, OLS in Shazam, or CRMDEL in Limdep.) to solve **problem 2** in this assignment. Compare your results here with those obtained through your matrix algebra program. Of course, the numbers produced by any canned package and your program should be **identical** to at least a few decimal places (rounding error may create differences beyond that.) Points will be awarded here by highlighting each correspondence between a *label used by a canned package* (e.g. "sum of squared residuals," "number of observations," "adjusted r-squared," "coefficient," "standard error," ...anything is fair game) and whatever label or matrix name you gave to a number in your program. **In other words, highlight the variable name and the result yielded by your package and show the corresponding output and result in the canned package so we can see they are the same. Try a Table that looks like this:**

<u>Item Match #</u>	<u>My Exact Var Name</u>	<u>My Computed Value</u>	<u>Package Computed Val</u>	<u>SPSS. Exact Var Name</u>
#1	RSQD	.628	.628	R Square

And mark that "Item Match #" on the SPSS output and on the output for your own program.

One point will be awarded here for each valid correspondence. Yes, you could win more than ten (10) points on this problem. What is the point of exercise #4? Well, this is your Rosetta Stone. If you complete this task, you will understand most numbers produced by any regression package because you will have written a program of your own capable of calculating those numbers.

(END ASSIGNMENT 3)

CONGRATULATIONS! You have programmed your own estimator. This is something very few practitioners know how to do. **Students:** Be sure to keep a copy of your computer program (your very own matrix routine to do OLS). You will be using it again for Assignment 4.