

# Problem Set 2

Paul T. Scott

Econometrics I  
NYU Stern

Professor Paul T. Scott  
Email: [ptscott@stern.nyu.edu](mailto:ptscott@stern.nyu.edu)

## Problem 1 Replicating Results from Class

The Cornwell and Rupert data for this problem can be downloaded from

<http://ptscott.com/teaching/data/cornwell-rupert.csv>

Source: Cornwell and Rupert, (1988) “Efficient estimation with panel data: an empirical comparison of instrumental variables estimators”

(a) For this part of the assignment, you are to replicate the regression “Mincerian Regression, Cornwell and Rupert Data” from the Linear Regression slides by obtaining the same coefficients and standard errors. If that is possible (which is often not the case), other differences in reported results can usually be explained. As part of your submission for this assignment, include the specific estimation results that you obtained for this regression.

Now that you have replicated the regression, we’ll consider a couple of minor extensions.

(b) Functional Form. The example thus far computes a single, generic effect of education on LWAGE. We’re interested in determining if there is a different effect for men ( $FEM=0$ ) and women ( $FEM=1$ ). One compact way to do this is to add an interaction term,  $FEM*ED$  to the model. The different effects are the coefficient on  $ED$  which is for men and the sum of the two effects,  $ED$  and  $FEM*Ed$ , for women. Reestimate your model with this additional effect, and report your result.

(c) Standard Errors. Just as an experiment, I’d be interested to compute the standard errors for the model in 1 using bootstrapping. Do this, using 100 replications, and report/discuss the difference you observe.

## Problem 2 Gasoline Demand

This part of the exercise is based on the gasoline data, which you can download as

<http://ptscott.com/teaching/data/gasoline.csv>

Source: Compiled by Professor Chris Bell, Department of Economics, University of North Carolina, Asheville from bea.gov and bls.gov.

Note, there is some ambiguity as to how to obtain the dependent variable in this data set. Use the following as a guide:

$$\begin{aligned}G &= ((\text{GASEXP}/\text{GASP})/\text{POP}) \times 1000000 \\ \text{LOGG} &= \log(G) \\ \text{LOGPG} &= \log(\text{GASP}) \\ \text{LOGY} &= \log(\text{INCOME}/\text{POP})\end{aligned}$$

(a) Estimate by least squares a version of the regression model of which looks as follows:

$$\begin{aligned}\text{LOGG} &= \beta_1 + \beta_2(\text{LOGPG upto 1973, 0 else}) + \beta_3(\text{LOGPG after 1973, 0 else}) \\ &+ \beta_4\text{LOGY} + \beta_5(\text{YEAR} - 1952) + \epsilon.\end{aligned}$$

(b) Now, use least squares to fit the coefficients of the model:

$$\begin{aligned}\text{LOGG} &= \beta_1 + \beta_2\text{LOGPG} + \beta_3(\text{LOGPG after 1973, 0 else}) \\ &+ \beta_4\text{LOGY} + \beta_5(\text{YEAR} - 1952) + \epsilon.\end{aligned}$$

Report the least squares coefficients for both cases.

(c) Could you have computed the least squares regression in (b) from (a)? If so, show how, algebraically. If not, why not? Describe this model in terms of the relationship between price and quantity that it implies.

(d) We now examine whether the three aggregate price indexes, PD = durables price, PN = nondurables price, PS = services price, are significant explanatory variables in the equation. Add logPD, logPN and logPS to your regression in part (b). Report the results. Now, test the hypothesis that the coefficients on the three variables are all zero.

### Problem 3 Health Care

Use the health care data:

<http://ptscott.com/teaching/data/healthcare.csv>

Source: Riphahn, Wambach, Million (2003), "Incentive effects in the demand for health care: a bivariate panel count data estimation"

(a) Test whether different regressions apply for men and women , using this model for log of income,

$$\log(\text{HHNINC}) = \beta_1 + \beta_2\text{AGE} + \beta_3\text{EDUC} + \beta_4\text{MARRIED} + \beta_5\text{HHKIDS} + \epsilon$$

i.e., Test the hypothesis  $\beta_M = \beta_F$ , where  $\beta_M = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)'$  is the parameter vector for males, and  $\beta_F$  is the corresponding parameter vector for females.

(b) Now, add at least one variable to the model and carry out the test again with your expanded model. Report all relevant results.

### Problem 4 Milk Production

This exercise is based on the Spanish dairy data,

<http://ptscott.com/teaching/data/dairy.csv>

Source: Alvarez, Arias, Orea (2006) “Explaining differences in milk quota values: the role of economic efficiency”

The data in the dairy data file are already in logs, so the regression model is

$$YIT = \beta_1 + \beta_2X1 + \beta_3X2 + \beta_4X3 + \beta_5X4 + \epsilon.$$

- (a) Compute the coefficients of this regression and report your results.
- (b) The constant returns to scale hypothesis is  $\beta_2 + \beta_3 + \beta_4 + \beta_5 = 1$ . Carry out a test of the hypothesis and report all results.
- (c) These data have been used in many studies to study functional form in production. In part (a), you fit a Cobb-Douglas model. A translog model would include all unique squares and cross products,  $x_1 * x_1, x_1 * x_2$ , etc. Fit a translog model, and test the hypothesis of the Cobb-Douglas model as a restriction on the translog model.
- (d) These data are a panel spanning 6 years. There might have been technological change in those 6 years. There are time dummy variables in the data set, YEAR93, ..., YEAR98. Add the time effects (dropping one of the dummy variables, of course) to your regression, and examine the results to see if there is evidence that the production shifted over time. Test the joint hypothesis that the time effects are all zero in the context of your translog model of part (c).

### Problem 5 - OLS Residuals

- (a) Consider the following table of data and potential residuals

$y$	$x_1$	$x_2$	$e_1$	$e_2$	$e_3$
?	1	0	1	2	3
?	1	-1	-3	-1	-2
?	1	1	2	-1	1

Which of the potential vectors of residuals  $e_1$ ,  $e_2$ , and  $e_3$  (if any) could be from a regression of  $y$  on  $[x_1 \ x_2]$ ? Explain.

(b) Show that the estimated OLS parameters are unchanged if the dependent variable and all dependent variables are transformed by subtracting their means.

That is, let

$$\begin{bmatrix} \beta_0 \\ \boldsymbol{\beta} \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y},$$

where  $\beta_0$  is the intercept,  $\boldsymbol{\beta}$  is a column of the other parameter estimates, and the first column of  $\mathbf{X}$  is a vector of 1's as usual. Show that  $\boldsymbol{\beta} = (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}'\widetilde{\mathbf{y}}$ , where  $\widetilde{\mathbf{X}}$  drops the first column of  $\mathbf{X}$  and for the other columns,  $\widetilde{\mathbf{x}}_k = \mathbf{x}_k - n^{-1} \sum_{i=1}^n x_{ki}$ . Similarly,  $\widetilde{\mathbf{y}} = \mathbf{y} - n^{-1} \sum_{i=1}^n y_i$ .