Problem Set 2

Chris Conlon

Fall 2025

Econometrics I
NYU Stern

Professor Chris Conlon
Email: cconlon@stern.nyu.edu

I strongly encourage you read the documentation for `fixest` particularly around how to do interactions and nonlinear hypotheses. See: `https://lrberge.github.io/fixest/articles/fixest_walkthrough.html#interaction-terms`. This will make this homework assignment go much more smoothly.

1 Prediction errors

The Cornwell and Rupert data for this problem can be downloaded from `https://github.com/chrisconlon/applied_metrics/blob/master/Problem%20Sets-Fall/ps2/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. 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 (for men) and the sum of the two effects, ED and FEM × ED, for women. Re-estimate your model with this additional effect, and report your result. (Do this both for the continuous measure "years of schooling" and the "levels of schooling completed" dummies).

(c) Standard Errors. Can you compute the bias-corrected bootstrap confidence interval for the difference in log wages between men and women college graduates (controlling for all other variables)? Thinking about the right regression to run and the correct $g(\cdot)$ function is the tricky part.

2 Gasoline Demand

`https://github.com/chrisconlon/applied_metrics/blob/master/Problem%20Sets-Fall/ps2/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:

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

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

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

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

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

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 $\log \text{PD}, \log \text{PN}$ and $\log \text{PS}$ to your regression in part (b). Report the results. Now, test the hypothesis that the coefficients on the three variables are all zero.

3 Healthcare Data

Source: Riphahn, Wambach, Million (2003), “Incentive effects in the demand for healthcare: a bivariate panel count data estimation”

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

$$\log(\text{HHNINC}) = \beta_0 + \beta_1\,\text{AGE} + \beta_2\,\text{EDUC} + \beta_3\,\text{MARRIED} + \beta_4\,\text{HHKIDS} + \varepsilon$$

i.e., Test the hypothesis $\boldsymbol{\beta}_M = \boldsymbol{\beta}_F$, where $\boldsymbol{\beta}_M = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)'$ is the parameter vector for males, and $\boldsymbol{\beta}_F$ 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.

4 Spanish Dairy Data

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

$$YIT = \beta_1 + \beta_2 X1 + \beta_3 X2 + \beta_4 X3 + \beta_5 X4 + \epsilon$$

(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. (See if you can do this with a clever use of interactions in `fixest`).

(d) Compare the goodness of fit of: (1) constant returns to scale; (2) Cobb-Douglas; and (3) translog. How many degrees of freedom are in each model (number of observations minus number of parameters estimated plus number of linear restrictions imposed).

(e) 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 using the `i(·)` command or the | operator; (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).

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 $\boldsymbol{e}_1, \boldsymbol{e}_2$, and $\boldsymbol{e}_3$ (if any) could be from a regression of y on $\begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 \end{bmatrix}$? 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

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

where β_0 is the intercept, $\boldsymbol{\beta}$ is a column of the other parameter estimates, and the first column of $\boldsymbol{X}$ is a vector of 1 's as usual.

Hint: Show that $\boldsymbol{\beta} = \left(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\right)^{-1}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{y}}$, where $\widetilde{\boldsymbol{X}}$ drops the first column of $\boldsymbol{X}$ and for the other columns, $\widetilde{\boldsymbol{x}}_k = \boldsymbol{x}_k - n^{-1}\sum_{i=1}^{n} x_{ki}$. Similarly, $\widetilde{\boldsymbol{y}} = \boldsymbol{y} - n^{-1}\sum_{i=1}^{n} y_i$.