**Problem Set 1**
**ECON 60330**
**(Due: 5pm, Thursday, January 30, 2014)**

Bill Evans
Spring 2014

The first four problems are questions about the OLS model. Some are more difficult than others, but these are a good review of topics from Econometrics I.

1. Consider a multivariate model of the form $Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$ where $X_1$ is (n x k1) and $X_2$ is (n x k2). Using the definition of partition inverses, show that if $X_1$ and $X_2$ are uncorrelated, one can obtain an estimate for $\beta_1$ by simply regressing $Y$ on $X_1$.

2. In a seminar one day, someone from the back of the room asks why authors do not directly test for omitted variables bias. Specifically, the questioner asks why for the linear model Y=Xβ+ε estimated by OLS, does someone not take the estimated residuals $\hat{\varepsilon}$ and estimate a second regression, $\hat{\varepsilon} = X\pi + v$ and test the null Ho: π=0 to see whether X and ε are uncorrelated. Is this a good idea? Why or why not? Explain your work.

3. Suppose one is interested in estimating a linear model of the form y=xβ+ε but the outcome of interest is measured with error. Suppose the i[th] observation for y is $y_i$ and the measured value of $y_i$ is $y_i^* = y_i + \nu_i$ where $\nu_i$ is an i.i.d. error and $E[\nu_i \mid x_i, y_i] = 0$ and $\text{var}[\nu_i] = \sigma_\nu^2$ for all i and $\text{cov}(\nu_i, \nu_j)$ for i≠j. Let $\hat{\beta}^* = (X'X)^{-1}X'Y^*$ be the OLS estimate for β using the mismeasured value of y and $\hat{\beta} = (X'X)^{-1}X'Y$ be the estimate if y was available. Show that in the presence of measurement error, $\hat{\beta}^*$ is still and unbiased estimate but $Var(\hat{\beta}^*) > Var(\hat{\beta})$.

4. Given a linear model of the form y=xβ+ε, and a sample with n observations, we can easily show that the $R^2 = \hat{y}'M\hat{y} / y'My$ where $M = (I_n - \frac{1}{n} i_n i_n')$. Show that the $R^2$ is the squared correlation coefficient between $y$ and $\hat{y}$. HINT: Start from the definition of the squared correlation coefficient and re-write as $R^2$.

5. Consider the one-way fixed-effects model $Y_{it} = \alpha + X_{it}\beta + u_i + \varepsilon_{it}$ where X is a scalar and the data set has balanced panels (there are T observations for all blocks i). Using summation notation, show that if there is no within-panel variation in X then that 'block' of data is not used in the estimation of β.

6. Consider the balanced panel fixed-effect model we have outlined in class. There are n panels with T observations per panel and NT observations in total. In the least-square dummy variable version of the model, we write the equation as

   $$Y = X\beta + D\alpha + \varepsilon$$

   Where D is the NT x N matrix of dummy variables for each group. In class, we assumed the data

was sorted by group then year, which meant that $D = I_n \otimes i_t$. Suppose instead that the data is sorted by year then group. Write an equation for D in this context. Using partitioned inverses, we know the estimate for $\beta$ is $\hat{\beta} = (X'MX)^{-1}(X'MY)$ where $M = I_{nt} - D(D'D)^{-1}D'$. What does M look like when the data is sorted by year then group?

7. Computer problem. For this problem, you will use the same data from the class example about the rate of return to tenure. That data is found on the class web page and it is named psid1.dta.

a. Using xtreg, Replicate the fixed-effect estimates of the returns to tenure from the class handout. The model should be a regression of ln(hourly wage) on tenure, tenure squared, age, age squared, and union status controlling for individual fixed effects.

b. Sort the data by id, and using the egen command, generate within-panel means of all the Y's and X's used in the previous problem. Next, generate the $\tilde{y}_{it} = y_{it} - \overline{y}_i$ and $\tilde{x}_{it} = x_{it} - \overline{x}_i$ for all the X's. [Egen produces descriptive statistics and replaces them for all observations in the data set. So for example, egen unionm=mean(union) would add the mean of union for all observations. Likewise, by id: egen unionm=mean(union) adds the within-id mean of union back to the data set.] Run a fixed-effects regression by regressing $\tilde{y}_{it}$ on all the $\tilde{x}_{it}$.

   b1. What are the means of $\tilde{y}_{it}$ on all the $\tilde{x}_{it}$ out to 6 decimal places?

   b2. Should you include a constant in this model?

   b3. The parameter estimates in this case should be the same as in part a, but why are the standard errors in this model much smaller than in the case for xtreg?

c. Next, run a fixed effect similar to that in part a) but include "union" as the sole covariate. What is the rate of return to union status and the estimated standard error on this parameter?

d. Using the within-id mean of union status (unionm) generated in part b), define a dummy variable that equals 1 if this mean is either zero or one. Call this variable nochange_union

```
gen nonchange_union=(unionm==1 | unionm==0)
```

Next, re-run the model from part c) but only include people with within panel variation in union status (nochange=0). Compare the coefficient on union in parts c) and d).

e. Next, re-run the model from part a) but only include people with within panel variation in union status (nochange=0). Compare the coefficient on union in parts e) and a). Why is it that there was no change in the coefficient on union in part d) but there was some change in the coefficient in part e?