BOSTON COLLEGE
Department of Economics
EC 327 Financial Econometrics
Prof. Baum, Mr. Zago, Spring 2014
Problem Set 2
Due Wednesday 12 February 2014
Total Points Possible: 210 points

Problem 13.5 (10 points)

(i) (10 pts) No, we cannot include age as an explanatory variable in the original model. Each person in the panel data set is exactly two years older on January 31, 1992 than on January 31, 1990. This means that $\Delta age_i = 2$ for all i. But the equation we would estimate is of the form $\Delta saving_i = \delta_0 + \beta_1 \Delta age_i + \ ...$, where δ_0 is the coefficient the year dummy for 1992 in the original model. As we know, when we have an intercept in the model we cannot include an explanatory variable that is constant across i; this violates Assumption MLR.3. Intuitively, since age changes by the same amount for everyone, we cannot distinguish the effect of age from the aggregate time effect.

Problem 13.6 (25 points)

(i) (10 points) Let FL be a binary variable equal to one if a person lives in Florida, and zero otherwise. Let y90 be a year dummy variable for 1990. Then, from equation (13.10), we have the linear probability model

$$arrest = \ \beta_0 \ + \ \delta_0 y90 \ + \ \beta_1 FL \ + \ \delta_1 y90 \times FL$$

The effect of the law is measured by δ_1, which is the change in the probability of drunk driving arrest due to the new law in Florida. Including $y90$ allows for aggregate trends in drunk driving arrests that would affect both states; including FL allows for systematic differences between Florida and Georgia in either drunk driving behavior or law enforcement.

(ii) (5 points) It could be that the populations of drivers in the two states change in different ways over time. For example, age, race, or gender distributions may have changed. The levels of education across the two states may have changed. As these factors might affect whether someone is arrested for drunk driving, it could be important to control for them. At a minimum, there is the possibility of obtaining a more precise estimator of δ_1 by reducing the error variance. Essentially, any explanatory variable that affects *arrest* can be used for this purpose. (See Section 6.3 for discussion.)

(iii) (10 points) Now the dependent variable is:

$arrest_{it}$ = #drivers arrested in county i and year t / # drivers licenced in county i and year t

so that the data structure now has one observation per county and year rather than one observation per driver. You could use the individual fixed effects estimator (or the first differences estimator) on this model, but if you did so, you could not include a FL dummy or interaction, as it would be collinear with the FL-county observations. Thus you might want to stick with pooled OLS and include a FL dummy, a 1990 dummy and the interaction that allows you to estimate the difference-in-differences model.

Problem C13.1 (25 points)

(i) (5 points) The F statistic (with 4 and 1,111 df) is about 1.16 and p-value $\approx$ .328, which shows that the living environment variables are jointly insignificant.

(ii) (5 points) The F statistic (with 3 and 1,111 df) is about 3.01 and p-value $\approx$ .029, and so the region dummy variables are jointly significant at the 5% level.

(iii) (10 points) After obtaining the OLS residuals, $\hat{u}$, from estimating the model in Table 13.1, we run the regression on $y74, y76, \ldots, y84$ using all 1,129 observations. The null hypothesis of homoskedasticity is $H_0 : \gamma_1 = 0, \gamma_2 = 0, \ldots, \gamma_6 = 0$. So we just use the usual F statistic for joint significance of the year dummies. The R-squared is about .0153 and F $\approx$ 2.90; with 6 and 1,122 df, the p-value is about .0082. So there is evidence of heteroskedasticity that is a function of time at the 1% significance level. This suggests that, at a minimum, we should compute heteroskedasticity-robust standard errors, t statistics, and F statistics. We could also use weighted least squares (although the form of heteroskedasticity used here may not be sufficient; it does not depend on *educ*, *age*, and so on). This test can also be performed using Stata's `robvar` command (see my Stata Tip 38, *Stata Journal* 6:4, 2006).

(iv) (5 points) Adding $y74 \cdot educ, y84 \cdot educ$ allows the relationship between fertility and education to be different in each year; remember, the coefficient on the interaction gets added to the coefficient on *educ* to get the slope for the appropriate year. When these interaction terms are added to the equation, $R^2 \approx .137$. The F statistic for joint significance (with 6 and 1,105 *df*) is about 1.48 with p-value$\approx$.18. Thus, the interactions are not jointly significant at even the 0.1 level. This is a bit misleading, however. An abbreviated equation (which just shows the coefficients on the terms involving *educ*) is

$$\widehat{kids} = -8.48 - .023educ + ... - .056y74 \cdot educ - .092y76 \cdot educ$$

$$-.152y78 \cdot educ - .098y80 \cdot educ - .139y82 \cdot educ - .176y84 \cdot educ.$$

Three of the interaction terms, $y78 \cdot educ, y82 \cdot educ$, and $y84 \cdot educ$ are statistically significant at the 0.05 level against a two-sided alternative, with the p-value on the latter being about .012. The coefficients are large in magnitude as well. The coefficient on *educ* which is for the base year, 1972 is small and insignificant, suggesting little if any relationship between fertility and education in the early seventies. The estimates above are consistent with fertility becoming more linked to education as the years pass. The F statistic is insignificant because we are testing some insignificant coefficients along with some significant ones.

Problem C13.2 (40 points)

(i) (5 points) The coefficient on y85 is roughly the proportionate change in wage for a male (female = 0) with zero years of education (educ = 0). This is not especially useful because the U.S. working population without any education is a small group; such people are in no way typical.

(ii) (10 points) The basic equation can be estimated, using factor variable notation, as

```
reg lwage  c.educ##i.y85 c.exper##c.exper union i.female##i.y85
```

What we want to estimate is $\theta_0 = \delta_0 + 12\delta_1$; this is the change in the intercept for a male with 12 years of education, where we also hold other factors fixed. Write $\delta_0 = \theta_0 - 12\delta_1$ to get

$$log(wage) = \beta_0 + \theta_0 y85 + \beta_1 educ + \delta_1 y85 \cdot (educ - 12) +$$

$$+\beta_2 expr + \beta_3 exper^2 + \beta_4 union + \beta_5 female + \delta_5 y85 \cdot female + u.$$

Therefore, we simply replace $y85 \cdot educ$ with $y85 \cdot (educ - 12)$, and then the coefficient and standard error we want is on $y85$. These turn out to be $\hat{\theta}_0 = .339$ and $se(\hat{\theta}_0) = .034$. Roughly, the nominal increase in wage is 33.9 %, and the 95% confidence interval is 33.9 ± 1.96(3.4), or about 27.2 % to 40.6 %.
This can also be easily done in Stata using the `lincom` command:

```
lincom 1.y85 + 12*1.y85#c.educ
```

The coefficient displayed by `lincom` is identical to that constructed by the method described above.

(iii) (5 points) Recalling how logarithms work,

```
g lrwage = cond(y85,lwage - log(1.65), lwage)
```

Only the coefficient on y85 differs from equation (13.2). The new coefficient is about -0.383 (se $\approx$ .124). This shows that real wages have fallen over the seven year period, although less so for the more educated. For example, the proportionate change for a male with 12 years of education is $-.383 + .0185(12) = -.161$, or a fall of about 16.1%. For a male with 20 years of education there has been almost no change [-.383 + .0185(20) = 0.013].

(iv) (5 points) The R-squared when log(rwage) is the dependent variable is .356, as compared with .426 when log(wage) is the dependent variable. If the SSRs from the regressions are the same, but the R-squares are not, then the total sum of squares must be different. This is the case, as the dependent variables in the two equations are different.

(v) (5 points) From `tabstat union, by(y85)`, in 1978, about 30.6% of workers in the sample belonged to a union. In 1985, only about 18% belonged to a union. Therefore, over the seven-year period, there was a notable fall in union membership.

(vi) (5 points) When `i.y85#i.union` is added to the equation, its coefficient and standard error are about -0.0004 (se $\approx$ .06104). This is practically very small and the t statistic is almost zero. There has been no change in the union wage premium over time.

(vii) (5 points) Parts (v) and (vi) are not at odds. They imply that while the economic return to union membership has not changed (assuming we think we have estimated a causal effect), the fraction of people reaping those benefits has fallen.

Problem C13.4 (20 points)

(i) (5 points) In addition to male and married, we add the variables *head*, *neck*, *upextr*, *trunk*, *lowback*, *lowextr*, and *occdis* for injury type, and *manuf* and *construc* for industry. The coefficient on $afchnge * highearn$ becomes .231 (se $\approx$ .070), and so the estimated effect and t statistic are now larger than when we omitted the control variables. The estimate 0.231 implies a substantial response of *durat* to the change in the cap for high-earnings workers.

(ii) (5 points) The R-squared is about 0.041, which means we are explaining only 4.1% of the variation in log(*durat*). This means that there are some very important factors that affect log(*durat*) that we are not controlling for. While this means that predicting log(*durat*) would be very difficult for a particular individual, it does not mean that there is anything biased about $\hat{\delta}_1$: it could still be an unbiased estimator of the causal effect of changing the earnings cap for workers' compensation.

(iii) (10 points) The estimated equation using the Michigan data is

$$\widehat{log(durat)} = \underset{(.057)}{1.413} + \underset{(.085)}{.097}afchnge + \underset{(.106)}{.169}highearn + \underset{(.154)}{.192}afchnge \times highearn$$

$$n = 1524, R^2 = .012$$

The estimate of δ_1, .192, is remarkably close to the estimate obtained for Kentucky (.191). However, the standard error for the Michigan estimate is much higher (.154 compared with .069). The estimate for Michigan is not statistically significant at even the 10% level against $\delta_1 > 0$. Even though we have over 1,500 observations, we cannot get a very precise estimate. (For Kentucky, we have over 5,600 observations.)

Problem C13.5 (30 points)

It is useful to declare these as panel data using

```
xtset city year, delta(10)
```

to make it clear that the data are observed every ten years, not every year.

(i) (10 points) Using pooled OLS we obtain

$$\widehat{log(rent)} = -.569 + .262d90 + .041log(pop) + .571log(avginc) + .0050pctstu$$

$$n = 128, R^2 = .861.$$

The positive and very significant coefficient on $d90$ simply means that, other things in the equation fixed, nominal rents grew by over 26% over the 10 year period. The coefficient on *pctstu* means that a one percentage point increase in *pctstu* increases rent by half a percent (.5%).
The t statistic of five shows that, at least based on the usual analysis, *pctstu* is very statistically significant.

(ii) (5 points) The standard errors from part (i) are not valid, unless we think a_i does not really appear in the equation. If a_i is in the error term, the errors across the two time periods for each city are positively correlated, and this invalidates the usual OLS standard errors and t statistics.

(iii) (10 points) The equation estimated in differences is:

$$\Delta log(rent) = .386 + .072\Delta log(pop) + .310log(avginc) + .0112\Delta pctstu$$

$$n = 64, R^2 = .322.$$

Interestingly, the effect of *pctstu* is over twice as large as we estimated in the pooled OLS equation. Now, a one percentage point increase in *pctstu* is estimated to increase rental rates by about 1.1%. Not surprisingly, we obtain a much less precise estimate when we difference (although the OLS standard errors from part (i) are likely to be much too small because of the positive serial correlation in the errors within each city). While we have differenced away a_i, there may be other unobservables that change over time and are correlated with $\Delta pctstu$.

(iv) (5 points) The heteroskedasticity-robust standard error on $\Delta pctstu$ is about .0028, which is actually much smaller than the usual OLS standard error. This only makes *pctstu* even more significant (robust t statistic ≈ 4). Note that serial correlation is no longer an issue because we have no time component in the first-differenced equation.

Problem C13.7 (35 points)

(i) (10 points) Pooling across semesters and using OLS gives

$$\widehat{trmgpa} = -1.75 - .058 spring + .00170 sat - .0087 hsperc + .350 female - .254 black -$$

$$- .023 white - .035 frstsem - .00034 tothrs + 1.048 crsgpa - .027 season$$

$$n = 732, R^2 = .478, \overline{R}^2 = .470.$$

The coefficient on *season* implies that, other things fixed, an athlete's term GPA is about .027 points lower when his/her sport is in season. On a four point scale, this a modest effect (although it accumulates over four years of athletic eligibility). However, the estimate is not statistically significant (t statistic≈.55).

(ii) (10 points) The quick answer is that if omitted ability is correlated with *season* then, as we know from Chapters 3 and 5, OLS is biased and inconsistent. The fact that we are pooling across two semesters does not change that basic point.
If we think harder, the direction of the bias is not clear, and this is where pooling across semesters plays a role. First, suppose we used only the fall term, when football is in season. Then the error term and season would be negatively correlated, which produces a downward bias in the OLS estimator of β_{season}. Because β_{season} is hypothesized to be negative, an OLS regression using only the fall data produces a downward biased estimator. [When just the fall data are used,

$\widehat{\beta_{season}} = -.116(se = .084)$, which is in the direction of more bias.] However, if we use just the spring semester, the bias is in the opposite direction because ability and season would be positive correlated (more academically able athletes are in season in the spring). In fact, using just the spring semester gives $\widehat{\beta_{season}} = .00089(se = .06480)$, which is practically and statistically equal to zero. When we pool the two semesters we cannot, with a much more detailed analysis, determine which bias will dominate.

(iii) (10 points)The variables sat, $hsperc$, $female$, $black$, and $white$ all drop out because they do not vary by semester. The intercept in the first-differenced equation is the intercept for the spring. We have:

$$\Delta\widehat{trmgpa} = -.237 + .019\Delta frstsem + .012\Delta tothrs + 1.136\Delta crsgpa - .065 season.$$

$$n = 366, R^2 = .208, \overline{R}^2 = .199.$$

Interestingly, the in-season effect is larger now: term GPA is estimated to be about .065 points lower in a semester that the sport is in-season. The t statistic is about 1.51, which gives a onesided p-value of about .065.

(iv) (5 points) One possibility is a measure of course load. If some fraction of student-athletes take a lighter load during the season (for those sports that have a true season), then term GPAs may tend to be higher, other things equal. This would bias the results away from finding an effect of season on term GPA.

Problem C13.8 (25 points)

(i) (10 points) The estimated equation using differences is:

$$\widehat{\Delta vote} = -2.56 - 1.29\Delta log(inexp) - .599\Delta log(chexp) + .156\Delta incshr$$

$$n = 157, R^2 = .244, \overline{R}^2 = .229.$$

Only $\Delta incshr$ is statistically significant at the 5% level (t statistic $\approx$ 2.44, p-value $\approx$ .016). The other two independent variables have t statistics less than one in absolute value.

(ii) (5 points) The F statistic (with 2 and 153 df) is about 1.51 with p-value $\approx .224$. Therefore, $\Delta log(inexp)$ and $\Delta log(chexp)$ are jointly insignificant at even the 20% level.

(iii) (5 points) The simple regression equation is

$$\widehat{\Delta vote} = \underset{(1.00)}{-2.68} + \underset{(.085)}{.092} \quad \Delta incshr$$

$$n = 33, R^2 = .037$$

This equation implies that a 10 percentage point increase in the incumbent's share of total spending increases the percent of the incumbents vote by about 2.2 percentage points.

(iv) (5 points) Using the 33 elections with repeat challengers we obtain

$$\widehat{\Delta vote} = \underset{(.63)}{-2.25} + \underset{(.032)}{2.18} \quad \Delta incshr$$

$$n = 157, R^2 = .229$$

The estimated effect is notably smaller and, not surprisingly, the standard error is much larger than in part (iii). While the direction of the effect is the same, it is not statistically significant (p-value $\approx .14$ against a one-sided alternative).