ECON 6511: Advanced Applied Econometrics
Homework 3 Solutions

1. (Wooldridge, Chapter 13, Problem 3) Use the data in KIELMC.dta for this exercise.

   (a) The variable *dist* is the distance from each home to the incinerator site, in feet. Consider the model:

   $$\log(price) = \beta_0 + \delta_0 y81 + \beta_1 \log(dist) + \delta_1 y81 \cdot \log(dist) + u$$

   If building the incinerator reduces the value of homes closer to the site, what is the sign of δ_1? What does it mean if $\beta_1 > 0$?
   Answer: Other things equal, homes farther from the incinerator should be worth more, so $\delta_1 > 0$. If $\beta_1 > 0$, then the incinerator was located farther away from more expensive homes.

   (b) Estimate the model from part (a) and report the results in equation form. Interpret the coefficient on $y81 \cdot \log(dist)$. What do you conclude?
   Answer: The estimated equation is:

   $$\log(price) = 8.06 - 0.011 y81 + 0.317 \log(dist) + 0.048 y81 \cdot \log(dist)$$

   Before the incinerator is built, there should be no effect of being near the incinerator. The incinerator's effect is measured by the interaction $y81 \cdot \log(dist)$. While $\hat{\delta}_1 = 0.048$ is the expected sign, it is not statistically significant (t statistic ≈ 0.59). Coefficient implies that, after the incinerator, a 1% increase in distance from the incinerator is associated with an increase in home price of 0.05%.

   (c) Add age, age^2, $rooms$, $baths$, $\log(instst)$, $\log(land)$, and $\log(area)$ to the equation. Now, what do you conclude about the effect of the incinerator on housing values?
   Answer: When we add the list of housing characteristics to the regression, the coefficient on $y81 \cdot \log(dist)$ becomes .062 (se = .050). So the estimated effect is larger − the elasticity of price with respect to *dist* is .062 after the incinerator site was chosen − but its t statistic is only 1.24. The p-value for the one-sided alternative H1: $\delta_1 > 0$ is about .108, which is close to being significant at the 10% level.

   (d) Why is the coefficient on $\log(dist)$ positive and statistically significant in part (b) but not in part (c)? What does this say about the controls used in part (c)?
   Answer: The coefficient on $\log(dist)$ is 0.317 in (b) and 0.001 in (c). The coefficient is statistically significant in part (b) because it was biased upward. *dist* is positively correlated with the quality of housing (as measured by rooms, area etc.), i.e. the incinerator is located in areas with lower-quality housing, and these quality measures are positively

correlated with house prices. When these quality controls are included, the coefficient is less biased away from zero.

2. (Wooldridge, Chapter 13, Problem 5) Use the data in RENTAL.dta for this exercise. The data for the years 1980 and 1990 include rental prices and other variables for college towns. The idea is to see whether a stronger presence of students affects rental rates. The unobserved effects model is:

$$\log(rent_{it}) = \beta_0 + \delta_0 y90_t + \beta_1 \log(pop_{it}) + \beta_2 \log(avginc_{it}) + \beta_3 pctstu_{it} + a_i + u_{it}$$

where *pop* is city population, *avginc* is average income, and *pctstu* is student population as a percentage of city population (during the school year).

(a) Estimate the equation by pooled OLS and report the results in equation form. What do you make of the estimate on the 1990 dummy variable? What do you get for $\hat{\beta}_{pctstu}$?
Answer: Using pooled OLS we obtain:

$$\log(rent_{it}) = -0.569 + 0.262 y90_t + 0.041 \log(pop_{it}) + 0.571 \log(avginc_{it}) + 0.005 pctstu_{it}$$

The positive and very significant coefficient on $y90$ 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 4.95 shows that, at least based on the usual analysis, *pctstu* is very statistically significant.

(b) Are the standard errors you report in part (a) valid? Explain.
Answer: 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.

(c) Now, difference the equation and estimate by OLS. Compare your estimate of β_{pctstu} with that from part (b). Does the relative size of the student population appear to affect rental prices?
Answer: The equation estimated in differences is:

$$\Delta\log(rent_{it}) = 0.386 + 0.072 \Delta\log(pop_{it}) + 0.310 \Delta\log(avginc_{it}) + 0.0112 \Delta pctstu_{it}$$

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% (statistically significant at the 1 percent level). 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 $\Delta pctstu$.

(d) Obtain the heteroskedasticity-robust standard errors for the first-differenced equation in part (c). Does this change your conclusions?
Answer: The heteroskedasticity-robust standard error on $\Delta pctstu$ is about .0029, which is actually much smaller than the usual OLS standard error of 0.004. This only makes $pctstu$ even more significant (robust t statistic ≈ 3.82). Note that serial correlation is no longer an issue because we have no time component in the first-differenced equation, i.e. only one observation per city.

3. (Wooldridge, Chapter 13, Problem 14) Use the data in JTRAIN3.dta for this question.

(a) Estimate the simple regression model $re78 = \beta_0 + \beta_1 train + u$, and report the results in equation form. Based on this regression, does it appear that job training, which took place in 1976 and 1977, had a positive effect on real labor earnings in 1978?
Answer: The simple regression estimates are:

$$re78 = 21.55 - 15.20 train$$

The estimated training effect is actually negative (!), large in absolute value − some \$15,200 less in real earnings in 1978 (just about one standard deviation, which is \$15,633). Further, using either t statistic, the coefficient is very statistically significant. The negative coefficient is likely due to the fact that lower skilled workers were more likely to receive a training grant (positive correlation between $train$ and omitted variables related to ability) and real labor earnings

(b) Now use the change in real labor earnings, $cre = re78 - re75$, as the dependent variable. (We need not difference $train$ because we assume there was no job training prior to 1975. That is, if we define $ctrain = train78 - train75$ then $ctrain = train78$ because $train75 = 0$.) Now what is the estimated effect of training? Discuss how it compares with the estimate in part (a).
Answer: With $cre = re78 - re75$, the simple regression results are:

$$cre = 2.49 + 2.33 train$$

Now the estimated training effect is positive (about \$2,330) and statistically significant. This is a huge difference compared with the original OLS estimate. It seems pretty clear that lower skilled workers were more likely to be in a job training program. The

differencing, hopefully, eliminates most of the self-selection bias; it certainly seems to eliminate much of it.

(c) Find the 95% confidence interval for the training effect using the usual OLS standard error and the heteroskedasticity-robust standard error, and describe your findings.
Answer: The 95% CI using the usual OLS standard error is about 0.74 to 3.92. Using the heteroskedasticity-robust standard error, the 95% CI is about 1.06 to 3.59. The robust CI is narrower. To be cautious, we might want to use the wider confidence interval. (In addition, we should probably control for other factors in the FD regression.)

4. (Wooldridge, Chapter 13, Problem 8) VOTE2.data includes panel data on House of Representatives elections in 1988 an 1990. Only winners from 1988 who are also running in 1990 appear in the sample; these are the incumbents. An unobserved effects model explaining the share of the incumbent's vote in terms of expenditures by both candidates is:

$$vote_{it} = \beta_0 + \delta_0 d90_t + \beta_1 \log(inexp_{it}) + \beta_2 \log(chexp_{it}) + \beta_3 incshr_{it} + a_i + u_{it},$$

where $incshr_{it}$ is the incumbent's share of total campaign spending (in percentage form). The unobserved effect a_i contains characteristics of the incumbent–such as "quality"–as well as things about the district that are constant. The incumbent's gender and party are constant over time, so these are subsumed in a_i. We are interested in the effect of campaign expenditures on election outcomes.

(a) Difference the given equation across the two years and estimate the differenced equation by OLS. Which variables are individually significant at the 5% level against a two-sided alternative?
Answer: The estimated equation using differences is

$$\Delta vote_{it} = -2.56 - 1.29\Delta\log(inexp_{it}) - 0.599\Delta\log(chexp_{it}) + 0.156\Delta incshr_{it}$$

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

(b) In the equation from part (a), test for joint significance of $\Delta\log(inexp)$ and $\Delta\log(chexp)$. Compute by hand and report the F-statistic.
Answer: The F statistic (with 2 and 153 df) is $F = \frac{(9282.27312 - 9102.29244)/2}{9102.29244/153} = 1.51$ with p-value $\approx .224$. Therefore, $\Delta\log(inexp)$ and $\Delta\log(chexp)$ are jointly insignificant at even the 20% level.

(c) Reestimate the equation from part (a) using $\Delta incshr$ as the only independent variable. Interpret the coefficient on $\Delta incshr$. For example, if the incumbent's share of spending

increases by 10 percentage points, how is this predicted to affect the incumbent's share of the vote?

Answer: The simple regression equation is:

$$\Delta vote = -2.68 + 0.218\Delta incshr$$

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

(d) Redo part (c), but now use only pairs that repeat challengers. [This allows us to control for characteristics of challengers as well, which would be in a_i.]

Answer: Using the 33 elections with repeat challengers we obtain

$$\Delta vote = -2.25 + 0.092\Delta incshr$$

The estimated effect is notably smaller and, not surprisingly, the standard error is much larger than in part (c). This suggests omitted variables were biasing the initial coefficient upwards. While the direction of the effect is the same, it is not statistically significant (p-value $\approx .14$ against a one-sided alternative). This estimate implies that a 10 percentage point increase in the incumbent's share of total spending increases the percent of the incumbent's vote by 0.9 percentage points.