The University of Texas at Austin
ECO 348K (Advanced Econometrics)
Prof. Jason Abrevaya
Fall 2016

PROBLEM SET #5 (due Tuesday, November 8th, 8am)

1. (Adapted from Wooldridge 13.C3) Use the data in **kielmc.dta** for this question. (Recall the incinerator example discussed in class.)

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

   $$\ln(price) = \beta_1 + \beta_2 \ln(dist) + \delta_1 y81 + \delta_2 y81 \cdot \ln(dist) + u.$$

   If building the incinerator reduces the value of homes closer to the site, what is the sign of δ_2? What does it mean if $\beta_2 > 0$?

   (b) Estimate the model in part (a) and interpret your estimate of δ_2.

   (c) Re-run the regression with *age*, age^2, *rooms*, *baths*, $\ln(intst)$, $\ln(land)$, and $\ln(area)$ added to the model. What do you conclude about the effect of the incinerator on housing values? Explain the difference or similarity to your result from part (b).

2. Use the **children_sample.dta** dataset that's available on the course website. These data are from the 2003 Community Tracking Study Household Survey. This particular subsample focuses on grown children, ages 18 through 30, for whom we have information about both parents (mother and father). The outcome of interest is **bmi** (body mass index, defined as weight (in kilograms) divided by height (in meters) squared). The **x** variables that we will consider are **educ** (years of education), **age** (in years), **mombmi** (mother's BMI), and **dadbmi** (father's BMI). Focus only on the subsample of white male children (command **keep if white & male**, leaving you with 770 observations).

   (a) Using the **tabstat** command, report the sample average and the 10%, 25%, 50% (median), 75%, and 90% quantiles for the **bmi** variable.

   (b) Looking at the histogram of **bmi** (command **hist bmi**), do the relative values of the sample average and the sample median make sense?

   (c) Run a linear regression of **bmi** on the **x** variables. Briefly describe what the slope estimates say about the relationship between **bmi** and the **x** variables.

   (d) Run a median regression (LAD) of **bmi** on the **x** variables using the **sqreg** command in Stata. Use 500 bootstrap replications (e.g. with the **reps(500)** option).

      i. Interpret the slope estimate on **educ**.

ii. What is the effect of a one unit increase in both mother's BMI and father's BMI? Provide a 90% confidence interval for this effect.

iii. Run the exact same LAD regression again (again with 500 replications). Explain why the standard errors and z-statistics are different from before.

(e) Now simultaneously run the quantile regressions, again with **sqreg** and with 500 bootstrap replications, for the 10%, 25%, 50%, 75%, and 90% quantiles.

i. Looking at your slope estimates for the different quantiles, are there any interesting differences that appear?

ii. Based on your results, do you think that the original linear regression (from part (c)) had heteroskedastic errors? Explain why or why not.

iii. Provide a p-value for the null hypothesis that the **age** slope is the same for the five quantile regressions.

iv. Provide a p-value for the null hypothesis that the **mombmi** and **dadbmi** slopes are the same in the 50% and 90% quantile regressions. (To be clear, you're jointly testing if **mombmi** is the same for the 50% and 90% quantiles and **dadbmi** is the same for the 50% and 90% quantiles.)

v. Using the 10% and 90% quantile estimates, provide an 80% predictive interval for **bmi** of an individual with the average values for each of the **x** variables. How does this interval compare to the 80% (unconditional) predictive interval that would be formed from the sample quantiles in part (a)?

3. Consider a probit model with an interaction term, specifically

$$\Pr(y = 1|x, z) = \Phi(\beta_1 + \beta_2 x + \beta_3 z + \beta_4 xz).$$

(a) What is the partial effect of x evaluated at some given values $x = x^*$ and $z = z^*$?

(b) Figure out the "interaction effect," which is defined by $\frac{\partial \Pr(y=1|x,z)}{\partial x \partial z}$. Is the sign of the interaction effect the same as the sign of β_4 (i.e. the slope on the interaction variable)? Is this different from what we would have in a linear regression model?

4. (Adapted from Wooldridge 17.C2) Use the **loanapp.dta** dataset for this question. These data were used in a 1996 *American Economic Review* paper to analyze mortgage loan approvals in Boston. The dependent variable of interest is the binary variable *approve*, equal to 1 if the loan applicant was approved for the loan.

(a) Estimate a linear probability model (LPM) of *approve* on *white*. Then, estimate a probit model of *approve* on *white*. For the probit model, determine the estimated approval probability for whites and the estimated approval probability for nonwhites. How does this compare to the LPM estimates? Explain. Do you think it would matter if we picked some other distributional assumption on u (i.e. other than the normal distribution)?

(b) Add the variables *hrat*, *obrat*, *loanprc*, *unem*, *male*, *married*, *dep*, *sch*, *cosign*, *chist*, *pubrec*, *mortlat*1, *mortlat*2, and *vr* and re-run both the LPM and the probit model.

i. Do you find statistically significant discrimination against non-whites?

ii. For the LPM model, interpret the slope estimate on *white*.

iii. For the probit model, provide both the partial effect of *white* at the average covariate values (PEA) and the average partial effect of *white* (APE). Explain these in words. How do the estimates compare to the LPM estimate?

iv. Repeat parts ii. and iii. for the *obrat* variable. Note that *obrat* is a value between 0 and 100, indicating the percentage of total income that is dedicated to other loan obligations (e.g. credit cards, cars, etc).

v. Use the **margins** command (with the **at** option) in order to figure out the average predicted approval probabilities at *obrat* values of $10, 20, 30, 40, 50$. (For fun, you can also do the **marginsplot** command afterwards to see things graphically—you do not need to print out the graph for the assignment.)

vi. Use the **margins** command (with the **at** option and the **dydx** option) in order to figure out the average partial effect of *obrat* at *obrat* values of $10, 20, 30, 40, 50$. How do the partial effects change as *obrat* increases? Does this make sense?

(c) Suppose that you are considering dropping all of the variables from the part (b) probit model that have z-statistics below 2 in magnitude.

i. Write down the associated null hypothesis that you'd want to test.

ii. What is the p-value for the Wald test associated with this null hypothesis?

iii. What is the p-value for the LR test associated with this null hypothesis?

iv. What would you conclude based upon these tests? Do they differ in their conclusion?