Problem Set 4 -- BUAN/MIS 6356

Turn in your code as ps4.R in eLearning. Add your answers to the interpretations as comments within your code. Your code should be able to be run. All the data files can be found in wooldridge.db

This problem set is due Tuesday. Oct 23rd at 11:59pm

4.1 Using the HPRICE1 data, find the best model for the housing price that you can using the AIC and BIC.

4.2 Using the GPA2 data, find the best model for the college gpa that you can using the AIC and BIC.

4.3 Using the MLB1 data, find the best model for the log of salary that you can using the AIC and BIC.

4.4 Use the data in RENTAL.RAW for this exercise. The data on rental prices and other variables for college towns are for the years 1980 and 1990. 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).

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

(ii) Are the standard errors you report in part (i) valid? Explain.

(iii) Now, difference the equation and estimate by OLS. Compare your estimate of β_{pctstu} with that from part (i). Does the relative size of the student population appear to affect rental prices?

(iv) Estimate the model by fixed effects to verify that you get identical estimates and standard errors to those in part (iii).

4.5 Use the state-level data on murder rates and executions in MURDER.RAW for the following exercise.

(i) Consider the unobserved effects model

$$mrdrte_{it} = \eta_t + \beta_1 exec_{it} + \beta_2 unem_{it} + a_i + u_{it},$$

where η_t simply denotes different year intercepts and a_i is the unobserved state effect. If past executions of convicted murderers have a deterrent effect, what should be the sign of β_1? What sign do you think β_2 should have? Explain.

(ii) Using just the years 1990 and 1993, estimate the equation from part (i) by pooled OLS. Ignore the serial correlation problem in the composite errors. Do you find any evidence for a deterrent effect?

(iii) Now, using 1990 and 1993, estimate the equation by fixed effects. You may use first differencing since you are only using two years of data. Is there evidence of a deterrent effect? How strong?

(iv) Compute the heteroskedasticity-robust standard error for the estimation in part (ii).

(v) Find the state that has the largest number for the execution variable in 1993. (The variable *exec* is total executions in 1991, 1992, and 1993.) How much bigger is this value than the next highest value?

(vi) Estimate the equation using first differencing, dropping Texas from the analysis. Compute the usual and heteroskedasticity-robust standard errors. Now, what do you find? What is going on?

(vii) Use all three years of data and estimate the model by fixed effects. Include Texas in the analysis. Discuss the size and statistical significance of the deterrent effect compared with only using 1990 and 1993.

4.6 Use the data in AIRFARE.RAW for this exercise. We are interested in estimating the model

$$\log(fare_{it}) = \eta_t + \beta_1 concen_{it} + \beta_2 \log(dist_i) + \beta_3 [\log(dist_i)]^2 + a_i + u_{it}, t = 1, \ldots, 4,$$

where η_t means that we allow for different year intercepts.

(i) Estimate the above equation by pooled OLS, being sure to include year dummies. If $\Delta concen = .10$, what is the estimated percentage increase in *fare*?

(ii) What is the usual OLS 95% confidence interval for β_1? Why is it probably not reliable? If you have access to a statistical package that computes fully robust standard errors, find the fully robust 95% CI for β_1. Compare it to the usual CI and comment.

(iii) Describe what is happening with the quadratic in log(*dist*). In particular, for what value of *dist* does the relationship between log(*fare*) and *dist* become positive? [*Hint:* First figure out the turning point value for log(*dist*), and then exponentiate.] Is the turning point outside the range of the data?

(iv) Now estimate the equation using fixed effects. What is the FE estimate of β_1?

(v) Name two characteristics of a route (other than distance between stops) that are captured by a_i. Might these be correlated with $concen_{it}$?

(vi) Are you convinced that higher concentration on a route increases airfares? What is your best estimate?

4.7 Use the data in LOANAPP.RAW for this exercise; see also 2.16 in Problem Set 2

(i) Estimate a logit model of *approve* on *white*. Find the estimated probability of loan approval for both whites and nonwhites. How do these compare with the linear probability estimates?

(ii) Now, add the variables *hrat*, *obrat*, *loanprc*, *unem*, *male*, *married*, *dep*, *sch*, *cosign*, *chist*, *pubrec*, *mortlat1*, *mortlat2*, and *vr* to the probit model. Is there statistically significant evidence of discrimination against nonwhites?

4.8 Use the data set in ALCOHOL.RAW, obtained from Terza (2002), to answer this question. The data, on 9,822 men, includes labor market information, whether the man abuses alcohol, and demographic and background variables. In this question you will study the effects of alcohol abuse on *employ*, which is a binary variable equal to one if the man has a job. If $employ = 0$ the man is either unemployed or not in the workforce.

(i) What fraction of the sample is employed at the time of the interview? What fraction of the sample has abused alcohol?

(ii) Run the simple regression of *employ* on *abuse* and report the results in the usual form, obtaining the heteroskedasticity-robust standard errors. Interpret the estimated equation. Is the relationship as you expected? Is it statistically significant?

(iii) Run a glm-logit of *employ* on *abuse*. Do you get the same sign and statistical significance as in part (ii)? How does the average marginal effect for the logit compare with that for the linear probability model?

(iv) Obtain the fitted values for the LPM estimated in part (ii) and report what they are when $abuse = 0$ and when $abuse = 1$. How do these compare to the probit fitted values, and why?

(v) To the LPM in part (ii) add the variables *age*, *agesq*, *educ*, *educsq*, *married*, *famsize*, *white*, *northeast*, *midwest*, *south*, *centcity*, *outercity*, *qrt*1, *qrt*2, and *qrt*3. What happens to the coefficient on *abuse* and its statistical significance?

(vi) Estimate a glm-logit model using the variables in part (v). Find the marginal effect of *abuse* and its *t* statistic. Is the estimated effect now identical to that for the linear model? Is it "close"?

(vii) Variables indicating the overall health of each man are also included in the data set. Is it obvious that such variables should be included as controls? Explain.

(viii) Why might *abuse* be properly thought of as endogenous in the *employ* equation? Do you think the variables *mothalc* and *fathalc*, indicating whether a man's mother or father were alcoholics, are sensible instrumental variables for *abuse*?

4.9 Refer to the data in FERTIL1.RAW to estimate a linear model for *kids*, the number of children ever born to a woman.

(i) Estimate a Poisson regression model for *kids*, using *educ, age, age-squared, black, east, northcen, west, farm, othrural, town, smcity, y74, y76, y78, y80, y82,* and *y84*. Interpret the coefficient on *y82*.

(ii) What is the estimated percentage difference in fertility between a black woman and a nonblack woman, holding other factors fixed?

(iii) Compute the fitted values from the Poisson regression and obtain the *R*-squared as the squared correlation between $kids_i$ and $\widehat{kids}_i$ .. Compare this with the *R*-squared for the linear regression model.