JBS Advanced Quantitative Research Methods Module MPO-1A

Lent 2010
Thilo Klein
http://thiloklein.de

Computer Lab Session 4
Limited Dependent Variables and Panel Data

Contents

Qualitative and Limited Dependent Variables

The dependent variables have a restricted domain of possible outcomes.

We'll examine binary variables, censored and sample selection models. There are others but they're considered beyond the scope of this course.

Binary variables may have just two possible outcomes, which will be labelled as 1 ('success') or 0 ('failure'). The probability of success may differ among individuals and we are interested in modeling the possible causes of these differences.

Exercise 1. Logit model

Use eaef21.csv.

a) Observe the distribution of *ASVABC*. *ASVABC* accounts for the result of an ability test.[1]

b) Create the variable BACH for all those students with more than 12 years of education. So that, *BACH* = 1 if the respondent has a bachelor's degree (or higher degree) and 0 otherwise. Investigate if the probability of a respondent obtaining a bachelor's degree from a four-year college *(BACH*=1*)* is related to the respondent's score on *ASVABC*, by estimating a linear model and a logit model.

c) Plot the probabilities estimated.

d) In order to interpret the logit results, estimate the marginal effects at the mean value of ASVABC (default) and at values of 40, 55 and 70.

e) Give an interpretation of the OLS regression and explain why OLS is not a satisfactory estimation method for this kind of model.

f) For the logit model, generate the pseudo-R2. Hint: remember $pseudo - R^2 = 1 - \frac{\log L}{\log L_0}$.

Answers:

a)

- **hist(ASVABC)**
- **summary(ASVABC)**

b)

- **BACH <- ifelse(S>12, 1, 0)**
- **lm1b <- lm(BACH ~ ASVABC)**
- **glm1b <- glm(BACH ~ ASVABC, family=binomial(link=logit))**

c)

- **plot(lm1b$fitted ~ ASVABC)**

You see a constant marginal effect of an increase in ASVABC on the predicted probability of getting a degree.

- **plot(glm1b$fitted ~ ASVABC)**

With the logistic model the marginal effect changes with the value of the independent variable. From part a) we know most respondents has scores between 40 and 60. We can see that the marginal effect is greatest for scores between 45 and 60 (slope is

[1] There is a battery of ability exams called *ASVAB*. These are scaled with mean 50, standard deviation 10. The different assessments are as follow

ASVAB2	arithmetic reasoning
ASVAB3	word knowledge
ASVAB4	paragraph comprehension
ASVABC	composite of *ASVAB2* (with double weight), *ASVAB3* and *ASVAB4*

steepest), with an increase in score between 40 and 50 having increasing the probability of graduating from 20% to 50%, and the probability going up to 80% for a score of 60. The highest score in the sample was 66, corresponding to a probability of around 90%.

d)

- **glm1b$coef[2] * dlogis(c(1,40)%*%glm1b$coef)**
- **glm1b$coef[2] * dlogis(c(1,55)%*%glm1b$coef)**
- **glm1b$coef[2] * dlogis(c(1,70)%*%glm1b$coef)**

We can see that the marginal effect for a 1 point increase in the test score goes from 2.2% probability increase at ASVABC=40 to 3.5% at the mean value of ASVABC=50.9 and down to 0.8% at ASVABC=70.

e) The OLS regression suggests that a 1 point increase in test score results in a 2.8% increase in the probability of earning a degree over the whole range of values of ASVABC. This does not appear to be a sensible result for test scores at the low end of the distribution as very few of those at the low end of the spectrum earned bachelor's degrees so making variation in ASVABC in the low range of scores unlikely to affect the probability of graduation. Furthermore, the negative constant implies that all students with scores of 32 or less, of which there are a few in the sample, have *negative* probabilities of earning the degree – clearly nonsense (it is also possible to have LPM predicting probabilities > 1). Finally, the standard errors, t- and F-tests reported by OLS are invalid in this case as the disturbance term is not normally distributed.

f)

- **glm1f.null <- glm(BACH ~ 1, family=binomial(link=logit))**
- **1 - logLik(glm1b)[1]/logLik(glm1f.null)[1]**

Exercise 2. Probit model

Use the data on LOANAPP.csv for this exercise, which provides information about loans applications drawn from data collected by the Federal Reserve Bank of Boston (The Cultural Affinity Hypothesis and Mortgage Lending Decisions, W.Hunter and M. Walker (1995), *The Journal of Real Estate Finance and Economics*).

a) Estimate a probit model of *approve* (a variable indicating if the loan was granted) on *white* (a variable indicating the race). Find the estimated probability of loan approval for both whites and nonwhites. How do these compare with the linear probability estimates?

b) Now, add the variables *hrat, obrat, loanprc, unem, male, married, dep, sch, cosign, chist, pubrec, mortlat1, mortlat2,* and *vr* to the probit model. It might be that white people present characteristics which lead to higher approval of loans. It might be instead that their higher rates are due to discrimination.

b1. Is there statistically significant evidence of discrimination against nonwhites?

b2. Family status is given by married and dep. (check with describe). Is there statistically significant evidence of discrimination against family status (married and dep)? [Hint: for the latter you have to apply a likelihood ratio test to test whether both variables are significant.]

b3. Estimate the same model, now reporting the marginal effects (use the **dnorm** command). How are these effects calculated for continuous and binary independent variables?

b4. Predict the events giving a success in the outcome as those cases with a predicted probability higher than 0.5. Compare these results with actual data.

c) Estimate the model from part b) by logit. Compare the coefficient on *white* to the probit estimate.

d) Compare the performance for all the models, consider actual and predicted values for different categories considering numbers of dependents (dep), and for 10 different categories made with proportions of non-housing obligations in household expenditure (obrat).

Answer:

a)

- **glm2a <- glm(approve ~ white, family=binomial(link=probit), data=loan)**

To get the probabilities for whites and non-whites from the probit estimates, we need to feed the estimates into the cumulative standard normal distribution. For whites we type:

- **pnorm(c(1,1)%*%glm2a$coef)**

For non-whites:

- **pnorm(c(1,0)%*%glm2a$coef)**

Now for OLS estimates:

- **lm2a <- lm(approve ~ white, data=loan)**

Get the predicted probabilities for white and non-white individuals:

- **c(1,1)%*%lm2a$coef; c(1,0)%*%lm2a$coef**

The above result is just the mean of the approve variable, conditional on the white variable. We could have got the same result just by summarizing approve:

- **by(loan$approve, loan$white, mean)**

Notice also that the probit probabilities are the same as the linear probability model (LPM) ones (only rounding errors make them slightly different). If the independent variables in a binary response model a mutually exclusive and exhaustive binary variables, then the predicted probabilities from LPM, logit and probit are simply the cell frequencies.

b) **b1.** The white variable is still highly significant so providing evidence that discrimination is a factor.

b2. H_0: beta_married=0 and beta_dep=0; H_1: at least one of their coefficients is non-zero. To test this run the model without the married and dep variables, making sure the sample is the same (so use **data=subset(loan, is.na(dep)==F)** in your probit command); generate the LR-statistic as the difference between the log-likelihoods of the two models. The null is rejected at 5%, so there is evidence of significant discrimination by family status.

b3. The effects are calculated at the mean values of continuous variables, unless you tell R to calculate them at other values. For binary independent variables, their marginal effects are calculated as the change in probability for an individual of changing from 0 in that variable to 1, with all other covariates at their mean values.

b4. Get the predicted values from the full model by typing **glm2b$fitted**. Simulate the outcomes by typing:

- **success <- ifelse(glm2b$fitted > 0.5, 1, 0)**

Compare predicted and actual outcomes using

- **table(success, glm2b$model$approve)**

The model fails in 11.36% of the cases, not bad.

c) Run the logit regression. To compare the estimated coefficients from logit and probit models, we need to multiply the logit estimates by 0.625. 0.938*0.625=0.586, compared to 0.520 from the probit estimation.

d) Follow the procedure we used to predict probit outcomes for the logit and LPM models. Create the 10 categories for obrat:

- **loan$obrat.cat <- cut(loan$obrat, breaks=quantile(loan$obrat, seq(0,1,.1)))**

To compare the predictive performance by number of dependents:

- **data <- data.frame(mean.approve=glm2b$model$approve, mean.app_LPM=lm2d$fitted, mean.app_probit=glm2b$fitted, mean.app_logit=glm2c.logit$fitted)**
- **INDICES <- data.frame(dependencies=glm2b$model$dep)**
- **FUN <- mean**
- **by(data, INDICES, FUN)**

To compare the predictive performance by non-housing obligation expenditure categories use the same command but with **obrat.cat** instead of **dep**. All models are guilty of over-predicting approval.

Exercise 3. Tobit model

Use the data in FRINGE.csv to estimate the pension earned considering individual information, based on US data in 1977. (Source: F.Vella (1993) "A Simple Estimator for Simultaneous Models with Censored Endogenous Regressors". *International Economic Review* 34, 441-457. The paper presents information needed to estimate the trade-off between wages and fringe benefits.).

a) For what percentage of the workers in the sample is *pension* equal to zero? What is the range of *pension* for workers with nonzero pension benefits? Why is a Tobit model appropriate for modeling *pension?*

b) Estimate a tobit model explaining *pension* in terms of *exper, age, tenure, educ, depends, married, white,* and *male.* Do whites and males have statistically significant higher expected pension benefits?

c) Use the results from part b) to estimate the difference in expected pension benefits for a white male and a nonwhite female, both of whom are 35 years old, single with no dependents, have 16 years of education, and 10 years of experience.

d) Add *union* to the Tobit model and comment on its significance.

e) Apply the Tobit model from part d) but with *peratio*, the pension-earnings ratio, as the dependent variable. (Notice that this is a fraction between zero and one, but, while it often takes on the value zero, it never gets close to being unity. Thus, a Tobit model bound by 0 is fine as an approximation). Does gender or race have an effect on the pension-earnings ratio?

Answer:

a)

- **nopension <- ifelse(fringe$pension==0, 1, 0)**
- **table(nopension)**
- **summary(fringe$pension[nopension==0])**

172 workers in the sample of 616 receive no pension benefits. For the 444 workers that do receive pension benefits, these benefits range from $7.28 to $2,280.27. The

Tobit model is suitable here as we have a significant proportion of the sample with no pension benefits, and a fairly wide spread among those that do receive pension benefits.

b)

- **library(VGAM)**
- **vglm3b <- vglm(pension ~ exper + age + tenure + educ + depends + married + white + male, data=fringe, tobit(Lower=0), trace=TRUE)**

The coefficients on both male and white are positive, although only the one on male is significant. They are both jointly significant however as we can see if we conduct a LR-test:

- **D <- 2*(logLik(vglm3b)[1] - logLik(vglm3b2)[1])**
- **pchisq(q=D, df=2, lower.tail=F)**

c)

$$(1) \quad E[y \mid x] = P(y>0 \mid x) \cdot E[y \mid y>0, x] = \Phi\left(x\beta / \sigma\right) \cdot E[y \mid y>0, x] =$$

$$= \Phi\left(\sum x_j \beta_j / \sigma\right) \cdot \sum x_j \beta_j + \sigma \cdot \phi\left(\sum x_j \beta_j / \sigma\right)$$

We use equation (1). First we need to calculate $x\beta$ to estimate the expected benefit for a white male with the given characteristics. We start with the **white male**: *exper* = *tenure* = 10, *age* = 35, *educ* = 16, *depends* = 0, *married* = 0, *white* = 1, and *male* = 1. Using our shorthand, we have

$x\hat{\beta}$ = -1,252.5 + 5.20(10) – 4.64(35) + 36.02(10) + 93.21(16) + 144.09 + 308.15 = 940.90.

Therefore, with $\sigma^2 = 677.74$ we apply equation (1):

$$E(pension \mid x) = \Phi\left(\frac{940.9}{677.74}\right) \cdot (940.9) + (677.74) \cdot \phi\left(\frac{940.9}{677.74}\right) = 966.40$$

For a **non-white female** with the same characteristics,

$x\hat{\beta}$ = - 1,252.5 + 5.20(10) – 4.64(35) + 36.02(10) + 93.21(16) = 489.07.

Therefore, her predicted pension benefit is

$$E(pension \mid x) = \Phi\left(\frac{488.66}{677.74}\right) \cdot (488.66) + (677.74) \cdot \phi\left(\frac{488.66}{677.74}\right) = 582.10$$

The difference between the white male and nonwhite female is 966.40 – 582.10 = $384.30.

[If we had just done a linear regression, we would have added the coefficients on *white* and *male* to obtain the estimated difference. Following this procedure we would calculate a difference of about 114.94 + 272.95 = 387.89, which is similar to the Tobit estimate. In fact, provided that we focus on partial effects, Tobit and a linear model often give similar answers for explanatory variables near the mean values.]

d) The coefficient on union is large and highly significant. To see what difference it makes in terms of pension benefits, you can go through the procedure described in part c).

e) Neither male nor white are significant, individually or jointly, thus suggesting that these variables aren't useful predictors of pension benefits as proportion of earnings. It would thus seem that white males have larger pension benefits than non-white females simply because they earn more on average.

Exercise 4. Heckman model (optional)

Use the MROZ.csv dataset. Source: The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions, Econometrica (1987), taken from the University of Michigan Panel Study of Income Dynamics for the year 1975. It consists of data about women in the labour market, and the purpose of this exercise is to estimate the rate of return to education. This can be done by estimating an equation with log(wage) in the left hand side and education in the right hand side. Other variables may be included as well. Typically, experience and experience squared are included. The rate of return to schooling is estimated by the coefficient which multiplies the variable *education*.

a) Using the 428 women who were in the workforce, estimate the return to education by OLS including *exper, exper2, nwifeinc, age, kidslt6,* and *kidsge6* as explanatory variables. Report your estimate on *educ* and its standard error.

b) Now, estimate the return to education by **heckit** from library **sampleSelection**, where all exogenous variables show up in the second-stage regression. In other words, the regression is *log(wage)*, on *educ, exper, exper2, nwifeinc, age, kidslt6,* and *kidsge6* and $\hat{\lambda}$. Calculate $\hat{\lambda}$ considering *exper, exper2, nwifeinc, age, kidslt6,* and *kidsge6* as explanatories. Compare the estimated return to education and its standard error to that from part a).

c) Using only the 428 observations for working women, regress $\hat{\lambda}$ on educ, *exper, exper2, nwifeinc, age, kidslt6,* and *kidsge6*. How big is the R-squared? How does this help explain your findings from part b? [Hint: think multicollinearity].

d) Finally, estimate the return to education by heckit, where restrictions are applied in the second stage. In other words, the regression is *log(wage)*, on *educ, exper, exper2* and $\hat{\lambda}$. Calculate $\hat{\lambda}$ considering *exper, exper2, nwifeinc, age, kidslt6,* and *kidsge6* as explanatories. Compare the estimated return to education and its standard error to that from part a).

Note: All datasets except for exercise 1 were collected by J. Wooldridge, *Introductory Econometrics, 3e.*

Panel Data

Panel data consists of information obtained for a group of individuals randomly selected at a sequence of points in time.

For this reason, we cannot assume that the observations are independently distributed along time. There will be unobserved effects having an impact on the dependent variable in time t which will probably have a similar effect in time t+1.

Typically, panel data are datasets containing a large amount of individuals (big N) and a small number of data points (small T≥2).

Notation:

$$y_{it} = \beta_0 + \beta_1 \cdot x_{1it} + \beta_2 \cdot x_{2it} + ... + \beta_k \cdot x_{k,it} + \nu_{it}$$
$$\nu_{it} = a_i + d_t + u_{it}$$
$$i = 1,...,N; \quad t = 1,...,T.$$

Note that ν_{it} represents the perturbations and is a composed effect with a time-invariant, individual, component (a_i, also called individual heterogeneity), a component which is common for all individuals and changes each year (d_t), and classic perturbations as we've seen so far (u_{it}, also called time-varying errors). d_t is frequently not considered, assuming implicitly that there's no defined pattern in changes along time. If a_i and d_t are included in the model the estimator is known as the two-way estimator. β_0 collects an average of individual effects, and so a_i and d_t are defined in terms of differences with this average.

There are various ways to treat econometrically this model, as you could see in the lecture. Each of them will have its flaws and advantages and will produce a different set of estimators for the coefficients and their standard deviations. Some, however, are similar under special conditions, and we find some models producing identical estimators in all classes. In this set of exercises we will cover the following ones: a) *pooling estimator*; b) *first-differencing* estimators, DE; c) *fixed-effects* estimator (or within estimator, FE); d) *between estimator*, BE; e) *random-effects* estimator, RE; f) the *least squares dummy-variable* estimation (LSDV) and g) the *two-way* estimator. Of course, all of them intend to estimate the same betas, but, according to the circumstances, some of them will be biased. In order to select the most appropriate one, it is necessary to observe how the dataset that we're using behaves according to classic assumptions.

Exercise 1. Least Square Dummy Variable estimation, incl. dummies per time-period

Use FERTIL1.csv which contains information for women for even years from 1972 to 1984. Observe the content of the variables. The aim is to explain the total number of kids born to a woman (*kids*).

- **a)** After controlling for all other observable factors, what has happened to fertility rates over time?
- **b)** South is the base group. Test whether region of the country at age 16 has an effect on fertility.
- **c)** Test whether other living environment characteristics at 16 have an effect on fertility.
- **d)** Heteroskedasticity. Has the variance of error u changed over time? [Hint: run $\hat{u}^2 = \gamma_0 + \gamma_1 \cdot y74 + \gamma_2 \cdot y76 + ... + \gamma_6 \cdot y84 + \varepsilon$].

e) Add interaction terms y74*educ, … y84*educ to the model estimated. Explain what these represent. Are they jointly significant?

Answer:

Note that a number of variables give the characteristics of the women when they were sixteen.

a)

- **fert$year <- as.factor(fert$year)**
- **lm1a <- lm(kids ~ educ + age + agesq + black + east + northcen + west + farm + othrural + town + smcity + year, data=fert); shccm(lm1a)**

We can see that the dummy variables for 1982 and 1984 are significant and negative, suggesting a significant fall in fertility in those years compared to the base year of 1972 not explained by the other covariates. Keep in mind that the model assumes that the effects of the explanatory variables have remained constant over time.

b)

- **linearHypothesis(lm1a, c("east=0","northcen=0","west=0"), vcov=hc0)**

The region dummies are jointly significant at 5%. Northcen is the only one that's significant by itself.

c)

- **linearHypothesis(lm1a, c("farm","othrural","town","smcity"), vcov=hc0)** # large city is omitted category

The variables are jointly insignificant. This suggests that living environment at 16 didn't play a role in fertility.

d)

- **lm1d <- lm(lm1a$res^2 ~ year, data=fert)**
- **lht(lm1d, c("year74", "year76", "year78", "year80", "year82", "year84"), vcov=hc0)**

The year dummies are jointly significant, suggesting that the variance of the error has changed over time, so heteroscedasticity is an issue and robust standard errors should be used.

e)

- **lm1e <- lm(kids ~ educ + age + agesq + black + east + northcen + west + farm + othrural + town + smcity + year + educ:year, data=fert)**
- **lht(lm1e, c("educ:year74","educ:year76","educ:year78","educ:year80", "educ:year82","educ:year84"),vcov=hc0)**

Though the interactions are jointly insignificant, the ones for 82 and 84 are significant at 10% and 5%, suggesting that perhaps there is a stronger effect from education on fertility in 1984 than in 1972.

Exercise 2. Pooled OLS (POLS), difference-equations and fixed effects

Use MURDER.csv, a state level dataset on murder rates (mrdrte) in the US, unem captures unemployment rates and executions (exec).

a) Consider the model: $mrdrte_{it} = d_t + \beta_1 \cdot exec_{it} + \beta_2 \cdot unem_{it} + a_i + u_{it}$. If past executions of convicted murderers have a deterrent effect, what should be the sign of β_1? What sign should β_2 have? Explain.

b) Using just years 1990 and 1993, estimate the equation from part a) considering time effects. Ignore the serial correlation problem in the composite errors. Do you find any evidence for a deterrent effect?

c) 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. Now, is there evidence of a deterrent effect? How strong?

d) Use the heteroscedasticity-robust standard error for the estimations in part c).

e) Find the state that has the largest number for the execution variable (exec) in 1993. How much higher is this value from the next highest value?

f) Estimate the equation, dropping Texas from the analysis. Compute the usual and heteroskedasticity-robust standard errors. Now, what do you find?

g) Finally, use all data. Estimate the two-way fixed effects with robust standard errors and conclude.

Answer:

a) You would expect β_1 to be negative and β_2 to be positive if we think that a stronger economy should result in less crime.

b)

- **murder.y <- subset(murder, d90==1 | d93==1)**
- **lm2b <- lm(mrdrte ~ d93 + exec + unem, data=murder.y); shccm(lm2b)**

The coefficient on exec is positive (!) and insignificant, suggesting that there is no deterrent effect. The coefficient on unem is significant and positive as expected.

c) Either: LSDV (Least Squares Dummy Variable estimator)

- **lm2c.LSDV <- lm(mrdrte ~ d93 + exec + unem + id, data=murder.y)**

Or: Time-demeaning ("within" estimator)

- **library(plm)**
- **plm2c.within <- plm(mrdrte ~ d93 + exec + unem, data=murder.y , model="within", effect="individual", index=c("id","year"))**

Or: First differencing ("first-differencing" estimator)

- **plm2c.fd <- plm(mrdrte ~ d93 + exec + unem, data=murder.y, model="fd", effect="individual", index=c("id","year"))**

Now the coefficient on exec is negative and significant but small (1 more execution reducing number of murders per 100,000 by 0.1). The coefficient on unem is no longer significant.

d) Use the **vcov=hc0** option. The coefficient on exec is now significant at 1%.

e)

- **tail(murder[order(murder$year, murder$exec),])**

Texas has the highest number of executions, followed by Virginia.

f)

- **plm2f.within <- plm(mrdrte ~ d93 + exec + unem, data=subset(murder.y, state!="TX"), model="within", effect="individual", index=c("id","year")); coeftest(plm2f.within, vcov=hc0)**

With Texas out of the sample there is no evidence for a deterrent effect. The coefficient is lower and the standard errors are much larger as we've reduced the variation in the explanatory variables by a lot by dropping Texas.

g)

- **plm2g.within <- plm(mrdrte ~ exec + unem, data=murder , model="within", effect="twoways", index=c("id","year"))**

Neither exec nor unem are significant at 5% when using all data for all years. There is no strong evidence of a deterrent effect, although exec is significant at 10%. There is also no support to the idea that a better economy leads to less crime from this sample.

Exercise 3. Panel Data, Random Effects, Fixed Effects and First Differences

Use the data in wagepan.csv. This data set contains annual information on wages, education, experience and other demographic and socio-economic variables for 545 men that worked in every year from 1980 through 1987. (From Vella, F. and M. Verbeek (1998), "Whose Wages do Unions Raise? A Dynamic Model of Unionism and Wage Rate Determination for Young Men", *Journal of Applied Econometrics* 13, 163-183).

a) Use the command **pdata.frame** to define the person and time identifiers. Obtain summary statistics for *lwage, educ, black, hisp, exper, married* and *union*. What do you observe? Which variables do not change over time?

b) Estimate a wage equation with *lwage* as the dependent variable and *educ*, *black*, *hisp*, *exper*, *expersq*, *married, union* and dummies to collect the effect of the years as the explanatory variables. Use simple OLS. In this model the coefficient that multiplies *educ* is interpreted as the rate of return to schooling. Comment on the results. In particular, is this a panel data estimator?

c) The estimated OLS standard errors are wrong if the individual errors are correlated over time, for example due to unobserved individual heterogeneity that is constant over time. Can you explain why does this occur? In order to get an indication whether the errors are correlated over time, we will look at the correlations of the residuals over time. In order to do this generate the OLS residuals, create lagged values of these residuals per individual and use the **cor** or **rcorr** commands. With the latter you can get p-values for the test whether the correlations are equal to zero. What do you conclude about the correlations?

d) Adjust the standard errors for the correlation of the residuals over time per individual. Are these standard errors different from the simple OLS ones?

e) In R, the *plm* command is for random and fixed effects panel data regressions. Estimate the model as in b), allowing for random and fixed unobserved individual effects.

Are there any differences between random effects and OLS? What are the differences between the fixed effects regression results and the other two sets of results? How would you explain the return to being married? What is the return to being unionised?

f) The Hausman test for fixed versus random effects can be easily performed in R using the **phtest** command. What do you conclude from the test result?

g) Estimate the model in first differences. Check the autocorrelation structure of the residuals in the first-differenced model. Do this in the same way as in c). (You will first have to estimate the model without the robust standard errors). What do you conclude?

h) Adjust the standard errors in the first-difference model using robust standard errors taking into account the clustering of the data (the errors are correlated over time

for every individual). Comment on the results and discuss the differences with the fixed effects results.

i) Estimate the between estimator. What does this estimator collect?

Answer:

a) (See script-file). *lwage* and *exper* vary quite widely over time and between individuals. Education and ethnicity do not vary over time. Marital and union status varies considerably over time, suggesting that quite a few of the individuals change marital and union status over the time period.

b)

- **lm3b <- lm(lwage ~ year + educ + black + hisp + exper + married + union, data=wage); shccm(lm3b)**

The return to an extra year of education is 9% on average. Blacks earn 14% less on average than others. Married people earn around 11% more on average, and union members 18% more.
This is a panel data estimator as there are time effects, although there are no individual effects.

c) Earnings are likely to be affected by individual ability, which can be argued to stay constant over time, suggesting that the errors will be correlated for the same individuals over time.

```
e <- lm3b$res
e_1 <- unlist( by(e, wage$nr, function(x) c(NA, x[-length(x)])) )
e_2 <- unlist( by(e, wage$nr, function(x) c(NA, NA, x[-c(7:8)])) )
e_3 <- unlist( by(e, wage$nr, function(x) c(NA, NA, NA, x[-c(6:8)])) )
e_4 <- unlist( by(e, wage$nr, function(x) c(rep(NA,4), x[-c(5:8)])) )
C <- cbind(e, e_1, e_2, e_3, e_4)
```

- **library(Hmisc)**

```
rcorr(C)
```

There are significant persistent correlations in the errors for the individuals over time (the third matrix in the **rcorr** output gives the p-values for the null hypothesis that the correlation is equal to zero). This supports the idea of a time-invariant individual ability component of the errors, meaning that the standard errors in part b) are invalid.

d)

- **source("http://thiloklein.de/clmclx.R"); clx**
- **clx(lm3b, 1, wage$nr)**

The clustered standard errors take the fact that the errors may not be independent over time for individuals into account and are robust to heteroscedasticity and serial correlation. As you can see they are a fair bit larger than the standard OLS errors, but this doesn't make a substantive difference to the results (apart from the year dummies).

e)

- **lm3e.re <- plm(lwage ~ year+educ+black+hisp+exper+married+union, data=wage, model="random", effect="individual",index=c("nr","year"))**

The coefficients on educ black and hisp are similar to those from pooled OLS, but the ones on exper, expersq, married and union are quite a bit different.

- **lm3e.fe <- plm(lwage ~ year+educ+black+hisp+exper+married+union, data=wage, model="within", effect="individual", index=c("nr","year"))**

Variables that don't vary over time are dropped, as is the dummy for 1987 as it is perfectly collinear with the other year dummies and exper (as experience increases by one year every year). The coefficients on married and union are smaller than in

the OLS and RE estimates, suggesting that unobserved individual ability is positively correlated with these variables.

f)

- **phtest(lm3e.fe, lm3e.re)**

The null hypothesis of there being no systematic difference in the coefficients (that unobserved heterogeneity is uncorrelated with the regressors) is not rejected at 5%. The random effects specification is then the preferred one.

g)

- **lm3e.fd <- plm(lwage ~ year+educ+black+hisp+exper+married+union, data=wage, model="fd", effect="individual", index=c("nr","year"))**
- **e <- lm3e.fd$res**
- **acf(e)**

There is significant correlation between the first lags of the error term. We need to use clustered standard errors.

h) The coefficients on *married* and *union* are smaller than those in the FE estimate, with only *union* significant at 10%.

i)

- **lm3e.fd <- plm(lwage ~ year+educ+black+hisp+exper+married+union, data=wage, model="between", effect="individual",index=c("nr","year"))**

This estimator is usually thought to estimate the long-run effect as what it does is basically calculate averages over all variables per individual (along time), and then run a cross-section with these averages.

Exercise 4. Optional

The file MATHPNL.csv contains panel data information about how different schools districts perform in an exam on Maths in Michigan called the fourth grade math test. Individuals will be, then, school districts in Michigan, and information is available for the years 1992 through 1998. The response variable of interest is *math4*, the percentage of fourth graders in an US district receiving a passing score on a standardized math test. The key explanatory variable is rexpp (real expenditures per pupil in the district). Amounts are in 1997 dollars. The spending variable will appear in logarithmic form. Other variables considered are *lunch* (% eligible for free lunch), *enrol* (school enrollment).

a) Consider the model:

$$math4_{it} = \delta_1 \cdot y93_t + ... + \delta_6 \cdot y98_t + \beta_1 \cdot \log(rexpp_{it}) + \beta_2 \cdot \log(enrol_{it}) + \beta_3 \cdot lunch_{it} + a_i + u_{it}$$

where enroll is total district enrollment and lunch is the percentage of students in the district eligible for the school lunch program (so lunch is a pretty good measure of the district-wide poverty rate.) " $\beta_1/10$ is the percentage point change in math4 when real per-student spending increases by roughly 10%". Explain. (This part is answered at the end of this exercise).

b) Use first differencing to estimate the model in part a). Interpret the coefficient pertaining to the spending variable.

c) Now, add one lag of the spending variable to the model. Now the model is

$$math4_{it} = \delta_1 \cdot y94_t + ... + \delta_6 \cdot y98_t + \beta_1 \cdot \log(rexpp_{it}) + \beta_2 \cdot \log(rexpp_{i,t-1}) + \beta_3 \cdot \log(enrol_{it}) + \beta_4 \cdot lunch_{it} + a_i + u_{it}$$

where the first available year (the base year) is 1993, as the lagged spending

variable makes us to lose one observation point (the one corresponding to year 1992). Estimate the model and report the usual standard errors.

d) Is the sign of the lunch coefficient what you expected? Interpret the magnitude of the coefficient. Would you say that the district poverty rate has a big effect on test pass rates?

e) Obtain the OLS residuals, v_{it}, and its lagged values. Compute a test for AR(1) serial correlation using the regression v_{it} on $v_{i,t-1}$. For this purpose, only use the years 1994 through to 1998 in the regression. Verify that there is strong positive serial correlation and discuss why. Note: AR(1) is a model such as the following: $v_{it} = \phi \cdot v_{i,t-1} + \varepsilon_{it}$

f) Re-estimate using first differencing. Note that you lose another year data, so you are only using changes starting in 1994. Discuss the coefficients and significance on the current and lagged spending variables.

g) Obtain heteroskedasticity-robust standard errors for the first-differenced regression in part iii. How do these standard errors compare with those from part iii for the spending variables?

h) Estimate correcting for both heteroskedasticity and serial autocorrelation: the fully robust estimator.

i) Estimate the equation by fixed effects. Is the lagged spending variable still significant?

j) Why do you think, in the fixed effects estimation, the enrollment and lunch program variables are jointly insignificant?

k) Define the total, or long-run effect of spending (i.e.: that one which occurs when spending increases in 1 in all periods) as $\theta = \beta_1 + \beta_2$. Use the substitution $\beta_1 = \theta - \beta_2$ to obtain a standard error for $\hat{\theta}$. The solution to this point follows below and in the do-file.

l) **[OPTIONAL]** Alternative ways to work with heteroskedasticity and autocorrelation with panel data:

Answer:

a) Ceteris paribus (holding other variables fixed):

$\Delta math4_{it} = \beta_1 \cdot \Delta \log(rexpp_{it}) = \frac{\beta_1}{100} \cdot [100 \cdot \Delta \log(rexpp_{it})] \approx \frac{\beta_1}{100} \cdot [\%\Delta rexpp_{it}]$. So if $\%\Delta rexpp_{it}$ =10, then $\Delta math4_{it} = \frac{\beta_1}{100} \cdot [10] = \frac{\beta_1}{10}$.

xi.

$$math4_{it} = \delta_1 \cdot y94_t + ... + \delta_6 \cdot y98_t + (\theta - \beta_2) \cdot \log(rexpp_{it}) + \beta_2 \cdot \log(rexpp_{i,t-1}) + \beta_3 \cdot \log(enrol_{it}) + \beta_4 \cdot lunch_{it} + a_i + u_{it}$$

$$math4_{it} = \delta_1 \cdot y94_t + ... + \delta_6 \cdot y98_t + \theta \cdot \log(rexpp_{it}) + \beta_2 \cdot \left(\log(rexpp_{i,t-1}) - \log(rexpp_{i,t})\right) + \beta_3 \cdot \log(enrol_{it}) + \beta_4 \cdot lunch_{it} + a_i + u_{it}$$