**SC968 PANEL DATA METHODS FOR SOCIOLOGISTS**

**WORKSHEET FOR PRACTICAL SESSION 3, part 1**

**3.1.0 OBJECTIVES**

* To run fixed and random effects models with continuous dependent variables
* To use the Hausman test to test between fixed and random effects
* To consider issues of specification

**3.1.1 STARTING UP**

Open the file “SESSION2-2.dta” – should be familiar by now!

Tell STATA you have panel data, with individuals denoted by **pid** and years by **wave**

**3.1.2 CREATE A VARIABLE INDICATING HOURLY WAGES**

In this session, we will run panel data regressions with hourly wages as the dependent variable. First, we have to create a suitable variable.

Looking at the data set, you will see that it contains the following variables

**fimnl** Labour income last month

**jbhrs** No. of hours normally worked per week [nb, this excludes overtime]

**jbot** No. of overtime hours in normal week.

Create an hourly wage variable called wageh:

* Calculate normal hours per week (standard plus overtime), and make sure it’s “tidy”.
* Convert this into a monthly figure
* Tidy up monthly labour income variable to remove outlying and missing values (less than £1, or higher than £10000/month)
* Divide monthly labour income by monthly hours of work

Question: Constructed in this way, the variable has one conceptual difficulty (which we will ignore, but it’s worth thinking about). What is this difficulty?

When you’ve constructed the variable, summarize it and check that you have a mean of £7.62 and a SD of £5.04. If you don’t get this result, use my code from the do-file.

**3.1.3 CREATE VARIABLES FOR LONDON AND THE SOUTH EAST**

Now, create dummy variables indicating whether individuals live in London (**region** codes 1 and 2) or the South-East (code 3). Remember to code these variables as missing if the **region** code is missing.

**3.1.4 PANEL DATA REGRESSIONS**

We will regress hourly wages on age, age squared, whether the individual lives in London or the South East, their weekly hours of work (standard plus overtime), their health status [use the **badhealth** variable we created previously], whether they have a partner, whether they have a degree or secondary-level education; and wave.

Estimate OLS, fixed effects, between, and random effects, *for men aged between 21 and 60, who are employed or self-employed.* Women’s wages are more complicated to estimate, because the labour supply decision is more complicated for women.

Fill in the results in this table (two decimal places will do). Use stars to indicate significance (\*\*\* = 0.001, \*\* = 0.01 \* = 0.05)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | OLS | FE | BE | RE |
| Age | 0.64 \*\*\* |  |  |  |
| Age squared | -0.01 \*\*\* |  |  |  |
| London | 1.86 \*\*\* | -0.53 | 2.20 \*\*\* | 1.29 \*\*\* |
| South east | 0.80 \*\*\* |  |  |  |
| Weekly hours |  |  |  |  |
| Badhealth |  |  |  |  |
| Partner |  |  |  |  |
| Degree |  |  |  |  |
| Secondary |  |  |  |  |
| Wave |  |  |  |  |
| Constant |  |  |  |  |

Question: Why is the coefficient on **wave** positive and significant in each equation?

**3.1.5 INTERPRETING RESULTS**

Remember that in the lecture we discussed how within and between coefficients are measuring different things? Take a look at the coefficients on London and the South-East, and say why they are different between the four specifications.

How do all the other coefficients differ between specifications? What conclusions can you draw from these differences?

**3.1.6 TESTING BETEWEN FIXED AND RANDOM EFFECTS**

Run a Hausman test for random effects.

Question: What is the null hypothesis?

Question: Is the null hypothesis rejected?

**3.1.7 THINKING ABOUT SPECIFICATION**

There are many ways in which we might think about improving specification. Here’s one. Consider the coefficients on London and the south-east. At the moment, the coefficients in fixed and random effects equations are measuring different things – intuitively, random effects, which is a weighted effect of these, doesn’t really have a chance of producing anything sensible.

Create a set of four variables indicating whether an individual moves TO the south-east or London, or AWAY FROM these areas. Include them in regressions and see what happens.

**3.1.8 IF YOU HAVE TIME TO SPARE**

Repeat this exercise using log wages as dependent variable

**3.1.9 THE END OF THE SESSION**

Save your data set as SESSION3-1.dta.

**SC968 PANEL DATA METHODS FOR SOCIOLOGISTS**

**WORKSHEET FOR PRACTICAL SESSION 3, part 2**

**3.2.0 OBJECTIVES**

* To run fixed and random effects models with categorical dependent variables
* To use the Hausman test to test between fixed and random effects
* To consider issues of specification

**3.2.1 STARTING UP**

If you’re running on from Session 3-1, simply keep the data set open. Otherwise:

Open the file “SESSION3-1.dta”

Tell STATA you have panel data, with individuals denoted by **pid** and years by **wave**

**3.2.2 PRELIMINARY REGRESSIONS**

In this session, we will use the dichotomous variable which we have already created as an indicatory of risk of psychological morbidity (**PM**). Recall that PM was generated as **hlghq2 > 2**, and that about a quarter of our sample have **PM == 1**

First, run preliminary regressions using PM as the dependent variable, and using sex, age, age squared, partnership status, whether the person is unemployed/sick, and an indicator of bad health (all these are variables which we have derived previously and which are in the data set).

Estimate the regressions over a sample of people aged 18-65 (inclusive) using the **xtlogit** command, for both fixed and random effects. Note that random effects takes much longer to run.

**3.2.3 HAUSMAN TEST**

After running the preliminary regressions (or if you’ve read ahead, while you’re doing them) do the Hausman test (by now, hopefully, familiar). What do the results tell you?

**3.2.4 CHANGES IN PARTNERSHIP STATUS**

Remember from the lectures that it can be more instructive to look at *changes* in status, rather than the status itself.

Create variables **get\_pnr** and **lose\_pnr** indicating whether an individual gets or loses a partner.

Tabulate these new variables to check you have 534 people who get a partner in the sample, and 454 people who lose one. If your answers don’t concur, use the code in the do-file.

Now, repeat the regressions using **get\_pnr** and **lose\_pnr** as well as the **partner** variable. Look at the coeffiecients on these variables, and compare them with the coefficients from the previous regressions which used only **partner**. What do the results tell you?

**3.2.5 MORE DETAILS ON PARTNERSHIP STATUS**

The above section raises the question: are changes in PM associated with changes in partnership status associated with partnership splits or with widowhood?

Test this by replacing the **lose\_pnr** variable with two new variables: **lose\_pnr\_wdw** indicating widowhood and **lose\_pnr\_spl** indicating marital or partnership splits.

There are several different ways you might do this – have a go yourself, then look at the do-file to see how I did it. My way isn’t the only way, but it does ensure that everyone with **lose\_pnr == 1** in the previous regression has either **lose\_pnr\_wdw** or **lose\_pnr\_spl** equal to one – which may not be the case with other ways of operationalising the variable.

Run **xtlogit** FE and RE regressions, and look at the results. What do they tell you?

**3.2.6 COMPARING XTLOGIT AND XTPROBIT**

Remember that there is no FE command for **xtprobit**. However, you can compare random effects models for **xtlogit** and **xtprobit.** For the model you have run most recently, write down coefficients from xtlogit and xtprobit, and check that they are in the same ball park using Amemiya’s scaling factor.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **xtprobit** | Probit x 1.6 | **xtlogit** |
| Age | 0.036 \*\*\* | 0.058 | 0.063 \*\*\* |
| Age squared |  |  |  |
| Female |  |  |  |
| Partner |  |  |  |
| Get partner |  |  |  |
| Lose partner (widow) |  |  |  |
| Lose partner (split) |  |  |  |
| Unemployed/sick |  |  |  |
| Bad health |  |  |  |
| Constant |  |  |  |

**3.2.7 THINKING ABOUT CHILDREN AND GENDER**

In the lecture, we talked about the relationship between children and mental health, showing that the variable **nch02** (number of children between 0 and 2) was associated with an increased risk of PM – especially among women.

Our data set also contains a variable **agechy** indicating the age of the youngest child in the household. This will enable us to explore the effect of young children in more detail.

Create three dichotomous variables **agey0, agey1** and **agey2** which indicate whether the youngest child in the household is aged 0, 1 or 2 respectively. Include these in fixed and random effects logit regressions. What do you find?

Remember that if a particular coefficient is not significant, this may be because we are using a pooled sample of men and women, and a significant effect exists only for women.

There are two ways to examine the differences between these effects. One is to estimate the regressions for men and women separately. The other is to include gender interaction effects. We’ll do both, and observe that they are completely equivalent.

Create a FULL SET of interaction coefficients.

You need to create:

**f\_age**

**f\_age2**

 **f\_partner**

**f\_get\_pnr
f\_lose\_pnr\_wdw
f\_lose\_pnr\_spl
f\_ue\_sick
f\_badhealth
f\_agey0
f\_agey1
f\_agey2**

Run FE regression including the full set of interaction terms [regression 1]

Now, run FE for men and women separately [regressions 2a and 2b]

Check that the estimated coefficients are identical in regressions [1] and [2a] (ie, the “basic” coefficients in the interacted equation are identical to the “men’s” coefficients in regression [2a].

Check also that the sum of each “basic” coefficient plus its “interacted” counterpart in regression [1] is equal to the “women’s” coefficient in regression [2b].

Now, do the same for RE.

There are two advantages to the “fully interacted” approach. First, we can see at a glance the coefficients which are significantly different between men and women – they are the ones where the interaction term is significantly different from zero.

Second, we can constrain the coefficients for men and women to be identical in the cases where they are not significantly different from each other, making for a more parsimonious specification.

Do this now, write down your results, and you’re finished.

|  |  |  |
| --- | --- | --- |
|  | FE | RE |
|  | Cofficient | Interaction (if included) | Cofficient | Interaction (if included) |
| Age |  |  |  |  |
| Age squared |  |  |  |  |
| Female |  |  |  |  |
| Partner |  |  |  |  |
| Get partner |  |  |  |  |
| Lose partner (widow) |  |  |  |  |
| Lose partner (split) |  |  |  |  |
| Unemployed/sick |  |  |  |  |
| Bad health |  |  |  |  |
| Age youngest = 0 |  |  |  |  |
| Age youngest = 1 |  |  |  |  |
| Age youngest = 2 |  |  |  |  |
| Constant |  |  |  |  |

**3.2.8 AT THE END OF THE SESSION**

Save your data set as SESSION3-2.dta.