Graduate Examinations 2010-11 SC968-AS-MI

**SC968 Panel Data Methods for Sociologists**

**Time allowed: 2 hours**

**This paper consists of FIVE questions in total, split into THREE sections. Candidates must answer ONE question from each section.**

**All questions in sections A and B are worth 35 marks.**

**The question in section C is mandatory and worth 30 marks.**

**SECTION A**

1. Answer *all* parts of this question.

You are using a panel study which contains data on both children and their parents, to investigate the determinants of behavioural problems in children. The study focuses in particular on the effects on children of stress in their parents’ relationships.

In the data set, information on children’s behaviour comes from a questionnaire filled in by teachers. Scores range from 0 to 20, with zero indicating no behaviour problems whatsoever, and higher scores representing more problems. Frequencies for the outcome variable, and for a binary variable derived from this variable, are tabulated below.

|  |
| --- |
| Children’s behaviour – teacher assessment |
| Continuous variable |  | Dichotomous variable |
| Score | Percentage |  | Score | Percentage |
| 0 | 19.7 |  | <= 8 | 90.7 |
| 1 | 27.2 |  | >= 9 | 9.3 |
| 2 | 17.8 |  | Total | 100.0 |
| 3 | 10.0 |  |  |  |
| 4 | 7.6 |  |  |  |
| 5 | 4.0 |  |  |  |
| 6 | 2.3 |  |  |  |
| 7 | 2.1 |  |  |  |
| 8 | 2.0 |  |  |  |
| 9 | 1.9 |  |  |  |
| 10 | 1.8 |  |  |  |
| 11 | 1.3 |  |  |  |
| 12 | 1.1 |  |  |  |
| 13 | 0.7 |  |  |  |
| 14 | 0.4 |  |  |  |
| 15 | 0.1 |  |  |  |
| Total | 100.0 |  |  |  |

1. [8 marks] A researcher might choose to analyse the outcome variable (children’s behaviour) as a continuous variable, or to transform it to a dichotomous (binary) variable, such as the one tabulated above. Both approaches have their advantages: explain what these are.
2. [8 marks] If one is using cross-sectional data, Ordinary Least Squares is not suitable for analysing dichotomous outcomes. Similarly, using panel data, ordinary panel regression (xtreg in STATA) is not suitable for the analysis of dichotomous outcomes. Explain why this is.
3. [8 marks] Some preliminary estimates of the determinants of behavioural problems are shown in the table below. These are results from fixed effects models: the first treats the outcome of interest as continuous, while the second treats it as a dichotomous variable, and presents marginal effects from a fixed effects logit model. Interpret the results of these models, and explain why they are different.

|  |  |  |
| --- | --- | --- |
|  | Fixed-effects model treating outcome variable as continuous | Marginal effects from fixed-effects logit model |
| No stress in parental relationship (reference category) |  |  |
| Some stress in parental relationship  | 0.691\*\*\* | 0.037 |
| Severe stress in parental relationship | 1.862 \*\*\* | 0.239\*\*\* |
| Both parents educated below secondary level (reference category) |  |  |
| One or both parents educated to secondary level  | -0.032 | -0.070 \* |
| One or both parents have a degree or equivalent | -0.995 \*\*\* | -0.188 \*\*\* |
| Annual household income (£ thousands per year) | -0.008 | -0.005 |
| Number of siblings present in household | 1.05\*\*\* | 0.057 |
| Notes: based on a sample of children from two-parent families. Significance is denoted by asterisks: \* = 5%, \*\* = 1%, \*\*\* = 0.1%. |

1. [3 marks] What are the advantages of presenting marginal effects from a fixed effects logit model, rather than coefficients from the same model?
2. [8 marks] The results presented above are from fixed effects models. What insights might you gain from comparing them with estimates from between models?

1. Answer *all* parts of this question.
2. [9 marks] What is meant by “unobserved heterogeneity”? Explain, with the use of equations, why unobserved heterogeneity is an important issue in quantitative analysis, and why panel data analysis can help to control for its effects.
3. [3 marks] Explain the difference between a transition matrix and a transition probability matrix.
4. [6 marks] A sample of people are interviewed at two time points three years apart. At each interview respondents are asked questions about their job, if they have one, and their other activities; they are also asked about their earnings. The following transition matrix shows changes in people’s labour market status between the two interviews (retired people, students and a few “other” categories have been omitted). From the transition matrix, create a transition probability matrix, and fill in the values in the appropriate spaces in the second matrix.

|  |  |
| --- | --- |
| Transition matrix | Main activity 2nd interview |
|  |  | Employed or self-employed | Unemp-loyed | Caring for home/family | Total |
| Main activity 1st interview | Employed or self-employed | 350 | 50 | 100 | 500 |
| Unemployed | 10 | 30 | 10 | 50 |
| Caring for home/family | 80 | 20 | 150 | 250 |
| Total | 440 | 100 | 260 | 800 |

|  |  |
| --- | --- |
| Transition probability matrix | Main activity 2nd interview |
|  |  | Employed or self-employed | Unemp-loyed | Caring for home/family | Total |
| Main activity 1st interview | Employed or self-employed |  |  |  |  |
| Unemployed |  |  |  |  |
| Caring for home/family |  |  |  |  |
| Total |  |  |  |  |

1. [4 marks] Comment on the transition probabilities you have just calculated.
2. [4 marks] If you split the sample up between men and women, what differences might you expect to see in the transition matrices?
3. [9 marks] The procedure for formulating a transition matrix for income is less straightforward than the procedure used above, because income is a continuous variable. Explain what steps you would take to formulate a transition matrix in this case, and what options you would have.

**SECTION B**

1. Answer *all* parts of this question.
2. [18 marks] Describe the similarities and differences between fixed effects, random effects and random coefficients models. Use equations, diagrams and examples to illustrate your answer.
3. [12 marks] A researcher asks a statistician to run a model on panel data to investigate the effect of age, gender and having children on wages. The statistician runs a random coefficients model of monthly income (income) on age in years, a dummy variable for female gender, a dummy variable for children (kids) and an interaction term between gender and children (femaleXkids). Below is the Stata output that the statistician produced. Explain to the researcher the results of the model, describing what each of the coefficient estimates represents in non technical language.
4. [5 marks] At the mean value of income, the variance partition coefficient for the null model with no independent variables is 0.47. Estimate the variance partition coefficient for the model below and comment on the meaning of the difference between the two values.

Mixed-effects ML regression Number of obs = 26968

Group variable: pid Number of groups = 3377

 Obs per group: min = 1

 avg = 8.0

 max = 15

 Wald chi2(4) = 506.78

Log likelihood = -221282.89 Prob > chi2 = 0.0000

------------------------------------------------------------------------------

 income | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

 age | 14.02087 .8516294 16.46 0.000 12.35171 15.69003

 female | -245.7615 29.53477 -8.32 0.000 -303.6486 -187.8744

 kids | 114.432 27.77625 4.12 0.000 59.99152 168.8724

 femaleXkids | -318.6507 38.71922 -8.23 0.000 -394.539 -242.7625

 \_cons | 439.7345 32.66486 13.46 0.000 375.7126 503.7565

------------------------------------------------------------------------------

------------------------------------------------------------------------------

 Random-effects Parameters | Estimate Std. Err. [95% Conf. Interval]

-----------------------------+------------------------------------------------

pid: Independent |

 var(age) | 395.6174 13.84975 369.3828 423.7154

 var(kids) | 141341.8 15455.46 114075.7 175124.9

 var(\_cons) | .0000456 .0000411 7.80e-06 .0002663

-----------------------------+------------------------------------------------

 var(Residual) | 603945.8 5631.745 593008 615085.3

------------------------------------------------------------------------------

LR test vs. linear regression: chi2(3) = 11847.79 Prob > chi2 = 0.0000

4. Answer *all* parts of this question.

 (a) [12 marks] Describe what a Cox regression model is. Include in your answer an explanation of why it is a semi-parametric model and why it is also called the Cox proportional hazards model. You may include equations and figures to illustrate your answer.

1. [5 marks] A researcher is interested in the factors that predict job retention in a call centre. She carries out a Cox regression model of time in post until quitting job. The independent variables are a dummy variable for female gender and a continuous variable for starting salary in units of a thousand pounds. Write a summary of the evidence based on the Stata output below.

Cox regression -- Breslow method for ties

No. of subjects = 50 Number of obs = 50

No. of failures = 29

Time at risk = 3070

 LR chi2(2) = 11.40

Log likelihood = -86.425061 Prob > chi2 = 0.0033

------------------------------------------------------------------------------

 \_t | Haz. Ratio Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

 salary | .5561377 .1362697 -2.39 0.017 .3440437 .8989822

 female | .4226076 .1818522 -2.00 0.045 .1818264 .9822399

------------------------------------------------------------------------------

1. [5 marks] The survival curve for these data is shown below. What does it show about the length of time people stay in their Call Centre jobs?



1. The researcher tests the proportional hazards assumption of the Cox model by adding a term called femXtime, which is the interaction between female and log(time).
2. [2 marks] What is the proportional hazards assumption?
3. [2 marks] What is the effect of ignoring a violation of the proportional hazards assumption?
4. [2 marks] What do you conclude from the results below?
5. [3 marks] What would you recommend that the researcher does next, and why?
6. [4 marks] Suppose starting salary violated the proportional hazards assumption. What would you recommend in that case?

Cox regression -- Breslow method for ties

No. of subjects = 50 Number of obs = 50

No. of failures = 29

Time at risk = 3070

 LR chi2(3) = 34.89

Log likelihood = -74.681666 Prob > chi2 = 0.0000

------------------------------------------------------------------------------

 \_t | Haz. Ratio Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

 salary | .5082946 .1391654 -2.47 0.013 .2972117 .8692907

 female | 4773.799 8928.174 4.53 0.000 122.1542 186560.5

 femXtime | .0990135 .0484784 -4.72 0.000 .0379257 .2584966

------------------------------------------------------------------------------

**SECTION C.**

**NOTE that this section is compulsory.**

[30 marks]

1. The Department of Health has announced that it intends to fund a research project investigating the effects of shift-working on mental health; in particular, on whether people’s mental health is adversely affected by (a) working at night, and by (b) irregular working hours.

Your institution has decided to apply for this research funding. Your colleague has made a first draft of a proposal, but was obviously in a great hurry when he wrote it. His proposal as it stands reads:

“This is an extremely important and timely research question. For this analysis we propose to use the British Household Panel Survey (BHPS). The BHPS contains two variables which measure mental health: the Likert scale, which ranges from 0 to 36, and the Caseness scale, which ranges from 0 to 12. It also contains information on whether people work at night, and a question asking whether people work irregular shifts. This means we will be able to investigate the relationship between shift working and mental health.”

This proposal clearly needs a great deal more work before it has any chance whatsoever of attracting funding. It is terrible in so many ways: it hasn’t mentioned any hypotheses for why a relationship might exist between shift working and mental health; it hasn’t mentioned what other variables you might control for in regressions; and worst of all, it hasn’t said anything about the analytical techniques you might employ, or what insights panel data analysis might offer above cross-sectional analysis.

Your colleague has asked for your suggestions. Write him a memo suggesting how he might improve the proposal.

There are no “right” or “wrong” answers to this question: we are looking for a sensible approach, clear explanations and originality.