Graduate Examinations 2009-10 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.

Below are two tables. One shows changes in marital status from 1991 to 1996 and the other shows changes from 1996 to 2001.

|  |  |
| --- | --- |
|  | 1996 |
| 1991 |  | Cohabiting | Single | Ex-partnered |
| Cohabiting | 777 | 10 | 64 |
| Single | 72 | 162 | 5 |
| Ex-partnered | 28 | 0 | 110 |

|  |  |
| --- | --- |
|  | 2001 |
| 1996 |  | Cohabiting | Single | Ex-partnered |
| Cohabiting | 803 | 9 | 65 |
| Single | 62 | 105 | 5 |
| Ex-partnered | 32 | 1 | 146 |

1. [6 marks] Calculate the transition probability matrices for these data.
2. [4 marks] Draw a decision tree showing all possible transitions
3. [8 marks] What is the transition probability of moving from partnered in 1991 to divorced, separated or widowed in 2001? And the transition probability of moving from partnered in 1991 to single in 2001?
4. [5 marks] If 69% are partnered, 20% single and 11% divorced, separated or widowed, then what is the probability of being partnered in 2001 if single in 1991?
5. [6 marks] What problems are there with these data? Are there any other issues to consider when interpreting the matrices?
6. [6 marks] If a researcher wanted to examine changes in household income following the end of a cohabiting union, what type of transition matrix would you recommend they use, and why?
7. Answer *all* parts of this question.
8. [6 marks] Explain the concepts of within- and between-group variation, in the context of panel data.
9. [5 marks] In a panel data set, give an example of a variable for which we would expect:
	1. All the variation to be between groups
	2. Most of the variation to be between groups
	3. Substantial variation both within and between groups
	4. Most or all of the variation to be within groups
10. [7 marks] Explain the properties, and the relative advantages and disadvantages, of the fixed effects, the between group, and the random effects estimators.
11. [10 marks] A colleague consults you. He has attempted to estimate the determinants of monthly earnings, for a sample of men aged 25-60, who work for at least one hour per week. He cannot understand why the coefficients should be so different between the fixed effects and between group estimators. What would you tell him?

|  |  |  |  |
| --- | --- | --- | --- |
|  | Fixed effectsestimates |  | Between group estimates |
|  | Coeff. | t-stat | P-value |  | Coeff. | t-stat | P-value |
| Age | 217.02 | 20.94 | 0.00 |  | 111.31 | 4.84 | 0.00 |
| Age squared | -1.56 | -12.41 | 0.00 |  | -1.29 | -4.61 | 0.00 |
| Weekly hours of work | 13.59 | 13.25 | 0.00 |  | 25.11 | 10.26 | 0.00 |
| Has a partner | -28.22 | -0.88 | 0.38 |  | 259.18 | 3.79 | 0.00 |
| Education: degree | 54.88 | 0.58 | 0.56 |  | 1079.28 | 14.98 | 0.00 |
| Education: secondary | -48.45 | -1.12 | 0.26 |  | 377.71 | 7.24 | 0.00 |
| Number of children | 63.70 | 5.06 | 0.00 |  | 0.08 | 0.00 | 1.00 |
| Living in Scotland | 442.73 | 1.99 | 0.05 |  | -6.72 | -0.08 | 0.94 |
| Constant | -4889.13 | -23.26 | 0.00 |  | -2264.60 | -5.04 | 0.00 |
| No. of observations | 4513 |  |  |  | 5413 |  |  |

1. [7 marks] Looking at the regression in part (d), how might you suggest that your colleague improves this specification?

**SECTION B**

3. Answer *all* parts of this question.

1. [12 Marks] Why are random coefficients models useful for modelling development over time? Use equations and figures where necessary to illustrate your answer.
2. [12 marks] Describe two non-linear growth functions and give a hypothetical research question where such a growth function is appropriate. Use equations and figures where appropriate. Extra marks will be given for describing functions that were not outlined in the lecture.
3. [11 marks] A researcher intends using data from a 10-year nationally representative panel study to investigate obesity over the life course. What type of model would you suggest to them? Explain what assumptions underlie the model and how you would test for any violations of your assumptions.

4. Answer *all* parts of this question.

In a study to compare two interventions for work stress, workers with high GHQ scores were randomised to receive counselling or an information leaflet. They were followed up until either their GHQ scores fell within the normal range or 10 weeks had passed from the start of the intervention trial. Times to improvement in GHQ scores for a random sample assigned counselling were as follows

2 2 3 5 8 9 10 10\* 10\* 10\*

(\* A star indicates a right-censored observation)

1. [6 marks] Explain what is meant by a right-censored observation. Also explain the meaning of the term hazard function.
2. [8 marks] Construct a Kaplan-Meier survival curve for the counselling group and show it on a suitable graph.
3. [8 marks] The full trial involved 60 workers, with 35 randomised to the counselling group and 25 to the leaflet group. The figure below shows Kaplan-Meier estimates of survival functions for times to improvement in GHQ scores for the two groups, and the results of a log-rank test; the test statistic was 2.90 on 1 d.f. (p = 0.072).

Use the diagram to estimate the median time to improvement for the two groups. Is there a difference between the survival patterns of the two groups? Comment on the results of the log-rank test.



1. [10 marks] Previous studies have suggested that age, marital status and social support are important factors in predicting recovery. Cox proportional hazards regression analysis was used to adjust times to improvement in GHQ scores for these variables.

The following tables show abbreviated computer output from two Cox regression models. In Analysis 1, a simple regression of times to improvement in GHQ scores on group (counselling vs. leaflet) alone was performed. Analysis 2 involved a multiple regression of times to improvement in GHQ scores on age (years), marital status (coded 0 = cohabiting, 1 = single, 2 = divorced, separated or widowed), social support (score 0-30) and group (coded as 0 = leaflet and 1 = counselling).

How well can an individual's time to improvement in GHQ be predicted from a multiple Cox regression model including age, marital status, social support and group? Comment on the regression coefficients from this model. Compare the results obtained from the multiple regression model with those derived from the regression on group alone.

**Analysis 1**

**Analysis 2**

**SECTION C.**

**NOTE that this section is compulsory.**

[30 marks]

1. The government’s Department for Children, Schools and Families has invited you to prepare a proposal for a programme of research into the determinants of divorce and separation. The DCSF is particularly interested in the role that poverty, and other factors linked with poverty, play as determinants of divorce and separation.

Draft a proposal which

1. begins by making the case for using panel data for the research, and explaining how panel data can inform this question in ways which cross-sectional data could not.

The proposal should also contain:

1. Details of the research questions which you propose to address, together with associated hypotheses.
2. A description of the methods which you propose to use for the analysis (which would include the methods taught in SC968, but which need not be restricted to these methods), and the particular aspects of the research question which will be addressed by each of these methods. The potential strengths and limitations of each of the methods should be pointed out.
3. Details of how you will define the variable(s) describing divorce and/or separation. Where you are proposing multivariate analysis, you should also describe the explanatory variables which you will be seeking to use. This set of variables should not be restricted to the variables present in the SC968 teaching data set, but should be limited to variables which one might reasonably expect to be present in a household data set such as the BHPS.

Rather than being answered in sections, this question should be answered as a cohesive whole. The question carries 30 marks in total. Six marks will be awarded for each of criteria (a) to (d), with an additional 6 marks for clarity and originality.