

## EC968 Course title

### Term Paper Titles

Students are encouraged to submit a term paper in each course for which term paper assessment is available. Four questions are provided below; choose *one*.

Details of assessment procedures are contained in the Postgraduate Economics Handbook, available from room 5B.206, and on the web at <http://www.essex.ac.uk/economics/documents/handbooks/ghb1112.pdf>. Be sure to read the sections on **A Guide to Good Practice in Assessed Work** and **Making the Most of your Essays, Term papers and Projects**, and the sections on plagiarism in particular.

Return the assessment to Claire Cox, in the Economics Department by 12:00 (midday) Friday 4 May 2012.

#### Question 1

Consider the following dynamic panel data model:

$$y_{it} = \alpha_0 + \mathbf{z}_i \alpha + \mathbf{x}_{it} \beta + \gamma y_{it-1} + u_i + \varepsilon_{it}, \quad i = 1, \dots, n; t = 1, \dots, T$$

where  $u_i$  and  $\varepsilon_{it}$  are mutually independent unobservables, independent of the observed covariate vectors  $\mathbf{z}_i$  and  $\mathbf{x}_{i1} \dots \mathbf{x}_{iT}$ ;  $\varepsilon_{it}$  is serially independent;  $\text{var}(u_i) = \sigma_u^2$ ; and  $\text{var}(\varepsilon_{it}) = \sigma_\varepsilon^2$ .

- Describe in detail the Arellano-Bond GMM estimator for this model.
- Explain how the parameters  $\alpha_0$  and  $\alpha$  can be estimated after GMM is used to estimate  $\beta$  and  $\gamma$ .
- How would you expect the Arellano-Bond estimator to perform if the true value of  $\gamma$  is close to 1? Explain.
- Use Stata to carry out a small-scale Monte Carlo simulation of the Arellano-Bond estimator for the following model:

$$y_{it} = x_{it} \beta + \gamma y_{it-1} + u_i + \varepsilon_{it}, \quad i = 1, \dots, 500; t = 1, \dots, 8$$

where:  $x_{it}$  is distributed as  $N(0,1)$ , independently across individuals and time periods;  $y_{i0}$  is normally distributed across individuals with mean zero and variance  $1.5 / (1-\gamma^2)$ ;  $\beta = 1$ ;  $\sigma_\varepsilon^2 = 0.2$ . Carry out the simulation for  $\gamma = 0.1, 0.5, 0.9$ . Comment on the results. Include a listing of your .do file in your answer.

*Reading:* The Monte Carlo simulation requires some initiative. You will need to use the Stata manuals or help screens relating to the generation of random numbers. You will also need to know about loops; alternatively, if you decide to use the `simulate` command, you will need to read about the Stata `program` command.

Arellano & Bond (1991), 'Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations', *Review of Economic Studies* **58**, 277–297.

## Question 2

Define the following variables:

$y_{it} = 1$  if, at year  $t$ , person  $i$  is unemployed, and 0 otherwise.

$\mathbf{x}_{it}$  = vector of variables representing gender, year of birth, state of health, educational attainment, age and region.

You should construct your own BHPS panel dataset containing variables of this type.

- Using the linear probability model, estimate the regression of  $y$  on  $\mathbf{x}$  using: within-group regression and random effects regression. Carry out a test to determine which is the more appropriate. Calculate the within-sample predicted probability of unemployment for each individual and summarise the distribution of predicted probabilities.
- Write a critique of the linear probability model in this context. Give an appraisal of your results at part (a) in the light of this critique.
- Explain in detail the conditional logit model. What are its advantages and disadvantages in this case?
- Estimate a random effects logit version of your model using `xtlogit` in Stata. Compare the results with your conditional logit estimates. Explain how you can discriminate between them statistically by means of a formal test. Interpret the test result.

*Reading:* Course lecture notes and Hsiao, chapter 7; Stata documentation. BHPS data (produced by ISER) are available from the UK Data Archive (<http://www.data-archive.ac.uk>). See also the BHPS pages at <http://www.iser.essex.ac.uk/survey/bhps> for documentation and other information.

## Question 3

- Provide a clear explanation of what is meant by “left censored” and “right censored” survival times, and illustrate your answer with some examples of how each may arise in a social science context.
- Suppose that you have continuous time unemployment spell data. The data were derived using a stock sample with “follow-up” (i.e. interviews some time after the stock sampling date). You also know the date of the interview, at which time information about characteristics were collected, and whether or not the spell in progress at the stock sampling date was still in progress and, if not, the date the spell ended. By deduction, you can calculate the length of time between the stock sample date and the date at which each person was last observed to be unemployed (the interview date for those still unemployed; or some date between the stock sample date and interview date for those who got a job). However, you don’t know the date at which each person’s spell began, and nor therefore the length of each person’s unemployment spell in total from start until last observed. With reference to expressions for the sample log-likelihood function, show that it is possible to estimate the parameters of an Exponential hazard regression model in this case. Also discuss, giving reasons, whether you could estimate a Weibull model with the same data.

- (c) [adapted from Wooldridge (2002, Ex. 20.3)] Assume that you have a random sample from the inflow to the state, and *all* survival times are right-censored.
- (i) Write down the sample log-likelihood function for this situation.
  - (ii) Derive the special case of likelihood function given in (i) when survival times follow the Gompertz distribution. [Recall that the Gompertz model has hazard function  $\theta(t, X) = \lambda \exp(\gamma t)$ , where  $\lambda = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)$  and shape parameter  $\gamma > 0$ .]
  - (iii) Consider the Gompertz model in which the covariate vector  $X$  only contains a constant. Show that the Gompertz log likelihood cannot be maximized for real numbers  $\beta_0$  and  $\gamma$ .
  - (iv) From (iii), what do you conclude about estimating duration models from inflow sample data when all survival times are right censored?

### Reading

As for Question 4 below.

### Question 4

The standard accelerated failure time (AFT) model for single-spell duration data has the form

$$\ln(T) = \beta'X + z$$

where  $T$  is an individual's survival time,  $\beta$  is a vector of coefficients,  $X$  is a vector of time-invariant explanatory variables (including a constant), and  $z$  is an error term.

- (a) Given an estimate of a particular regression coefficient, call it  $\beta_k$ , how would you interpret its sign and magnitude? Why might you instead interpret the impact of a given explanatory variable using estimates of the corresponding 'time ratio', i.e.  $\exp(\beta_k)$ ?

Consider the following table of estimates of several AFT models of the time between being released from prison and being arrested again ('criminal recidivism'). The data come from Chung, Schmidt and Witte ('Survival analysis: a survey', *Journal of Quantitative Criminology*, vol. 7, pp. 57–98): see 'recid.dta', a Stata data set downloadable from <http://www.stata.com/texts/eacsap/>. The data refer to 1465 convicts from North Carolina (NC) released from prison during the period 1 July 1977 and 30 June 1978. You should use these data to illustrate your answers, especially to parts (d) and (e).

- (b) Focusing on the *lognormal* model, provide an extended commentary on the model estimates, discussing the estimated impacts of each of the explanatory variables, and how the recidivism hazard rate varies with time since release from prison.
- (c) Briefly discuss whether the Weibull model leads to different conclusions from those derived in (b).
- (d) Discuss, with supporting evidence, other models and specifications that one might use to investigate the time to recidivism further using these data.
- (e) Of the various models, which is your preferred specification? Give reasons.

*Reading*

Jenkins, S.P. (2005a). *Survival Analysis*, unpublished manuscript, Institute for Social and Economic Research, University of Essex. Downloadable from

<http://www.iser.essex.ac.uk/files/teaching/stephenj/ec968/pdfs/ec968lnotesv6.pdf>

Jenkins, S.P. (2005b). *Lessons* (to accompany *Survival Analysis* op. cit.), 9 pdf files, Institute for Social and Economic Research, University of Essex. Downloadable from <http://www.iser.essex.ac.uk/resources/survival-analysis-with-stata-module-ec968>

Wooldridge, J.M. (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge MA, chapter 20. HB 71.6.P2

And other course reading, as appropriate.

**Accelerated Failure Time Models of Criminal Recidivism**

	Lognormal	Weibull
In NC prisoner work programme *	-0.063 (0.120)	-0.113 (0.113)
Number of prior convictions	-0.137 (0.021)	-0.110 (0.017)
Time served (months)	-0.019 (0.003)	-0.017 (0.002)
Had felony sentence *	0.444 (0.145)	0.372 (0.132)
Had alcohol problems *	-0.635 (0.144)	-0.555 (0.132)
Had drug history *	-0.298 (0.133)	-0.349 (0.122)
Black *	-0.543 (0.117)	-0.563 (0.111)
Married when incarcerated *	0.341 (0.140)	0.188 (0.136)
Years of schooling	0.023 (0.025)	0.029 (0.024)
Age (months)	0.004 (0.001)	0.005 (0.001)
Intercept	4.099 (0.348)	4.222 (0.341)
$\alpha$		0.806 (0.031)
$\sigma$	1.810 (0.062)	
$\gamma$		
$\kappa$		
Log-likelihood	-1597.059	-1633.032

Table show estimated coefficients and parameters with standard errors in parentheses.  $N = 1445$ . \*: dummy (0/1) variable.