Faculty of Law, Business and Social Sciences Graduate School

Social Science Statistics II

Module II Project 2009/2010

This assessed project constitutes 100% of marks for Social Science Statistics II.

Submission

- The project is due **Thursday 15th April 2010** (two copies to be handed in at the LBSS Faculty Office). The main text should be presented in hard copy **and should not exceed 30 pages in total (including tables and appendices)**. Font size should not be less than 12pt text, 10pt in tables. Margins should not be less than 2cm. The main text should be 1.5 spacing, whereas text in tables and figures should be single spaced. Both copies of the assignment should be spiral bound (either wire comb binding or standard plastic comb binding).

- *You must also submit the Word file of your project on CD, with your amended data set, output file and syntax file. These files will* ***only be consulted if there is doubt over authorship*** *(i.e. suspected plagiarism) so you should* ***include in your hard copy anything you want the marker to see***. NB use CD & file names that clearly identify you (e.g. your matriculation number).

- Late projects may be accepted, but will be penalised accordingly. Extensions to the deadline will only be considered in cases with genuine extenuating circumstances and supporting evidence (**please consult the Graduate School Handbook** before making enquiries – the handbook explains the procedures and criteria for extensions and contains a copy of the application form).

- Students may co-operate to some extent, but the final submission (particularly the interpretation of results) should be your own work. Tutors can be consulted for advice in principle (e.g. “how do I run a Chow Test?”), but not regarding specific actions on your project (e.g. “What Chow Tests do you think I should run on this data, and can you show me how I do it?”).

Project Brief

You have two options:

Either Using *good modelling practice,* construct a regression model of risk pricing in the mortgage market using the dataset provided (**riskpricing.sav**). As well as conducting all the diagnostic and refinement processes on an ordinary least squares model of interest rates, you will also need to construct a simple logit model of at least one mortgage type to see whether lenders allocate risky borrowers to particular product groups, rather than charging them a higher rate of interest (no need to do the full range of diagnostic checks on the logit model(s)).

Background to Risk Pricing Project:

One of the criticisms levelled at banks in the aftermath of the financial crisis was their failure to adequately price risk. This failure supposedly occurred at various levels. One aspect relatively unexplored is the pricing of individual mortgages. Did banks charge interest rates to higher risks? If they did not, then low risk borrowers are effectively subsidising high risk borrowers, and the disincentives to high risk applying for mortgage finance are weak. Anecdotal evidence suggests that UK lenders have been reluctant to charge higher interest rates because this would imply that poorer households would pay more for the same dwelling than a wealthy neighbour—those who are most able to pay, pay least. Because income tends to fall along race and ethnic lines, lenders might further worry that they be accused of discrimination, and there may be additional economic reasons arising from problems of adverse selection (see Pryce, G. (2003) 'Worst of the good and best of the bad: adverse selection consequences of risk pricing', *Journal of Property Investment and Finance*, 21(6)). Whatever their motives, and whatever the anecdotal evidence suggests, we still do not know the extent to which banks avoid pricing risk.

Your task is to analyse a sample of 1,000 mortgage transactions to see if there is any pattern of risk pricing behaviour. Many of the variables in the data are categorical variables, so you will need to convert these to dummy variables in order to enter them in the regression. You will also need to present an intuitive ‘theory’ of mortgage pricing in terms of its main causes (the working paper by Cairns and Pryce 2008 summarises the main factors that drive the risk of default). If your theory (backed up by tests for omitted variables) leads you to believe that there are important variables not included in the data you might want to include a brief discussion of what these might be. Your model should include financial ratios to measure the risk associated with each borrower, such as the LTV (= loan to value ratio = mortgage amount / price of property); and LTI (= loan to income ratio).

Once you have run the various diagnostic checks, variable transformations and regression refinements, you should then construct a separate logit model of at least one mortgage product to see if there is evidence that banks allocate risky borrowers to particular products rather than/in addition to charging higher interest rates.

Or Using *good modelling practice,* construct a OLS regression model on a subject that interests you/relevant to your work, and at least one simple logit model.

If you choose this second option, you must be confident that your chosen data is adequate (e.g. 150 or more observations, at least 10 to 15 explanatory variables, a continuous **and appropriate** dependent variable, sufficient continuous (appropriate) explanatory variables). You will also need a binary variable(s) that can be used as the dependent variable to build a separate logit model(s). It would be advisable to check with me first that the data is suitable. You can also select from the datasets supplied in SSSI provided they meet the above requirements and **provided you did not use the same dataset in the regression section of the SSSI assignment**.

Marking:

The project will be marked primarily on model construction: i.e. the extent to which you have followed good modelling practice in using diagnostic testing etc. The recommended "general to specific" strategy was outlined in lecture 7 and is summarised below. What you must demonstrate is that you know how to use these tests and the meaning of the output, the consequences of the tests pointing to failed OLS assumptions, and the appropriate way of dealing with such failures. You must also demonstrate an ability to combine the information from diagnostic testing in a way that results in a robust yet meaningful model.

Marks will be given for the amount of appropriate work done (e.g. using a variety of tests with discussions of their pros and cons rather than just one test). You can pass with using only one or two tests for each failure of OLS assumptions provided these are appropriately chosen and applied and understanding is demonstrated (particularly of the meaning of coefficients and other output for the final model). Very good marks will be given to students who demonstrate that: (i) they have not only an understanding of the available tests and solutions, but also of their limitations; (ii) they have read widely re statistical analysis and used a range of tests and diagnostic methods for each violation of OLS assumptions; (iii) they appreciate the implications of their model; and (iv) they have been innovative in the way they have developed or applied the model (e.g. added relevant data, or used the model to predict). Note that you need to convince the marker that you understand what you are doing – we can only mark what you've written down in the hard copy of the assignment you submit.

Style of Presentation

The project should be presented as *a technical report to an academic audience* with full explanation of all presented results. Three key questions/themes should run through the report: (i) what is the real-world meaning/implications/usefulness of the model? (ii) how do you know that the model is correctly specified? and (iii) how generalisable is the model to other samples? **The report should also include a one page non-technical Executive Summary, which should be understandable to an intelligent lay person.**

NB: because this project is to be presented as a technical report rather than in the style of a journal article, you should focus on explaining the details of diagnostic tests and the modelling process, and less discussion of the existing literature or theory related to the dependent variable you are trying to explain (**i.e. I am not really interested in your knowledge of risk pricing and financial economics, but I am interested in whether you can build a decent econometric model**). Note that presentation of test results should be done without including the surplus information that comes with SPSS output. In other words, you should try to present your results in as concise a way as possible – pages and pages of unedited and unexplained SPSS output will not be well received, and projects that exceed the page limit will be penalised.

Summary of Good Modelling Practice:

(i) Theory

- Always start with theory where possible.
- Try to consider all possible determinants of the dependent variable
- Try to identify specific hypotheses you want to test

(ii) Anticipated Regression Model

- identify the regression model that follows from your theory and that will allow you to test the hypotheses you are most interested in.

(iii) Data Collection & Coding

- make sure the data you collect, the way you collect it (i.e. unbiased sampling, large *n*, precise measurement) & the coding will allow you to build your general model and test specific hypotheses.
- if you are using secondary data, be aware of the sampling and coding issues associated with the data.

(iv) General Model

- attempt your first regression model using the data available:
    - start with all available variables and all available observations
    - make obvious modifications before starting the diagnostic/refinement process

(v) Diagnostic Checks and Refinement

- Examine Residual plots
    - scatter plots of residuals on y & xs
    - should be "spherical"
    - normal probability plots
    - outliers (use Cook's distances etc.)

- Heteroscedasticity
    - Test using Koenker B-P etc.
    - If heterosk. exists, use White's SEs & avoid Chow's 2nd Test

- Wrong signs and Mispecification
    - t-tests & multicollinearity tests
    - RAMSEY reset test.
    - Non-linear Transformations
    - interactions

- Low Adjusted R^2
    - Transform variables
    - drop irrelevant variables
    - get data on new variables

- F-Tests
    - structural stability (Chow)
    - linear restrictions

- Multicolinearity
    - check VIF, eigenvalues, Condition indices etc.
    - present joint hypothesis tets.

(vi) Specific Model

- should be "well behaved"
    - stable
    - passes general misspecification tests if possible
    - e.g. RESET test

- coefficients should be meaningful
    - do the coefficients make sense?
    - How do they relate to your theory/intuition?

  - ➢ Alternative explanations/interpretations

(vii) Revise Theory?

- Do your empirical results mean that you need to modify your initial theory, hypotheses and Anticipated Regression Model?

- Often, it is only when you start the empirical process that you really grasp the key aspects or limitations of your theory

(iix) Present the Final model (to an academic audience: e.g. journal article)

- you should present your (revised) theory first

- then the (revised) anticipated regression model

- then discuss the data and measurement of (revised) anticipated variables

- then present a selection of regression models

- present a series of "preferred" regressions which might vary by:

  - ➢ selection of regressors
  - ➢ measurement of dependent variable
  - ➢ and/or sample selection

- present the selection of regressions in columns all in a single table rather than as separate tables -- this will assist comparison

- only present statistics that you explain/discuss in your text

  - ➢ always present sample size, Adjusted R^2, t values on individual coefficients or SEs or Sig.

- then offer a full discussion

  - ➢ I.e. of the different regressions and statistics that you have presented and discuss any relevant elements of the refinement process

- this discussion should lead you to select a final "preferred" model(s) (if there is one) on the basis of the diagnostics, intuition and relevance to the theory

  - ➢ it is a good idea to present this in a separate table in more detail -- e.g. with confidence intervals for the coefficients

- you should comment on the limitations of you model

  - ➢ given the data and the anticipated effect of measurement problems, omitted variables, bias in sample, insufficient sample size etc.

- Then present the results of your specific hypothesis tests

  - these should be run on your final preferred model(s) and include a full discussion of their meaning and the limitations implied by the inadequacies of your model.

- If you are presenting to a non-academic audience, you will have to select which of the above are likely to be most meaningful/important to them.

  - Whether or not you present the results of the diagnostics, you MUST construct your model using them otherwise:

    - how do you know that you have specified it correctly?
    - How do you know that it can be generalised beyond your little sample!?

Statistical Reading:

There is no set statistical reading for this assignment. To pass, you could probably get away with reading only my overheads and lab notes. However, I would strongly recommend that you read more than these. In particular, I would recommend that you purchase your own copy of *A guide to econometrics* by Peter Kennedy since this will be helpful for the assignment but also serve as a useful reference tool when you read empirical material in future or if you go on to do any of your own statistical analysis.

In addition there are a wide range of texts in the library that you could consult (see the list below for a sample). These are written in varying degrees of technical detail, but most provide useful commentary even if you don't understand the maths. It's probably worth browsing through a few to find a style you feel comfortable with. I list a large number of possibilities below, but the best recent one I've come across is:

- □ **Wooldridge, Introductory Econometrics: A Modern Approach**

which is by an eminent statistician, very clear and at a fairly accessible level. Others include:

- □ Basic econometrics Damodar N. Gujarati
- □ Introduction to econometrics Christopher Dougherty
- □ Using econometrics a practical guide A.H. Studenmund
- □ Undergraduate econometrics R. Carter Hill, William E. Griffi
- □ A companion to theoretical econometrics edited by Badi H. Ba
- □ Applied econometrics for health economists a practical guide
- □ A guide to econometrics Peter Kennedy
- □ Econometric methods Jack Johnston, John DiNardo
- □ Econometric analysis William H. Greene
- □ Statistics and econometric models Christian Gourieroux
- □ Econometric methods and applications G.S. Maddala
- □ A dictionary of econometrics Adrian C. Darnell.
- □ Essentials of econometrics Damodar Gujarati
- □ Undergraduate econometrics R. Carter Hill, William E.
- □ Introduction to econometrics G. S. Maddala
- □ Using econometrics : a practical guide ; A.H. Studenmund
- □ Econometrics Jon Stewart
- □ A course in econometrics Arthur S. Goldberger
- □ Understanding econometrics Jon Stewart

Dr Gwilym Pryce
15th February 2010