1 Research Methods 2:

1.1 Econometric Methods for Cross Section Data

In cross section data -survey, cross country etc. the time series issues we have covered are no longer relevant, but a number of others are

To remind you:

1.1.1 Problems with OLS

Considering :

$$Y_i = \alpha + \beta X_i + u_i$$

we assume

$$\begin{aligned} E(u_i) &= 0 \\ E(u_i^2) &= \sigma^2 \text{ or } var(u_i) = \sigma^2 \\ E(u_i u_j) &= 0 \text{ or } cov(u_i, u_j) = 0 \end{aligned}$$

We have seen that we have to make very specific assumptions about u_i in order to get OLS estimates with the desirable properties.

If these assumptions don't hold than the OLS estimators are not necessarily BLU.

We can respond to such problems by changing specification and/or changing the method of estimation.

First we consider the problems that might occur and what they imply. In all of these we are basically looking at the residuals to see if they are random.

- The error variances are not constant $\Rightarrow$ heteroscedasticity
- In multivariate analysis two or more of the independent variables are closely correlated $\Rightarrow$ multicollinearity
- The function is non-linear
- There are problems of outliers or extreme values -but what are outliers?
- There are problems of missing variables $\Rightarrow$can lead to missing variable bias

Of course these problems do not have to come separately, nor are they likely to

- After time series, the cross section R^2s will look very low.
- Really want to be able to identify a misleading regression that you may take seriously when you should not.
- The tests in Limdep cover many of the above concerns, but you should always plot the residuals and look at them.

1.2 Heteroscedasticity

When look at survey data across household, lack of independence seems to be a rule rather than an exception, soheteroscedasticity almost always present in survey data

Even when individual behaviour is homogeneous within clusters, heterogeneity between clusters can still lead to heteroscedasticity overall.

In this case

$$Y_i = \alpha + \beta X_i + u_i$$

we assume

$$\begin{aligned} E(u_i) &= 0 \\ E(u_i^2) &= \sigma_i^2 \\ E(u_i u_j) &= 0 \end{aligned}$$

So in this case the errors do not have a common variance. The effect of this will be to leave the OLS estimator of β unbiased, but the estimated standard error will be biased.

Tests:

- Use Breusch Pagan or White test for heteroscedasticity. B-P normalise reiduals from regression by dividing by SER, square them and regress on variables thought ti generate the heteroscedasticity: orig xs, squares, cross products
- Can analyse using quantile regressions: eg median regression which minimises the absloute sum of the errors. Can also use other percentiles
- and look at variation of estimated relations and to investigate the shape of the conditioanl distribution.

Solution:

- Correct the standard errors, using the White Heteroscedastic robust errors available in Limdep
- In practice heterosced correction is usually less important than the correction for intra cluster correlation

1.3 Multicollinearity

- When we have more than one explanatory variable there is a possible problem.
- There may be two or more variables that explain the dependent variable well, but they may be closely correlated.
- This could mean that is difficult to distinguish the individual effects of both variables.

1.3.1 Problem and identification

Have considered perfect multicollinearity in exercise

In practise unlikely to find this extreme , but may get close to it.

There has to be some multicollinearity, so the question is identifying when it is important and so when it is a problem.

- What will tend to get is high $R^2$and F test statistics, but low individual significance of the individual coefficients.
- Parameter estimates become very sensitive to the addition or deletion of observations. So can test by dropping observations and seeing what happens
- Predictions will be worse than those of a model with only a subset of variables
- Standard errors of the regression coefficients will be high. Though in fact this is not necessarily the result of multicollinearity alone.

1.3.2 Cures

- Get more data. But what is important is not the number of observations but the informational content.
- Drop variables:may work in some cases where not interested in individual parameter values, but there is a problem of omitted variable bias.
- Simply present the OLS estimates and the variance-covariance matrix to allow an assessment of multicollinearity.

- Rather than the zero restriction try some others across the variables.
- Use extraneous estimates: e.g. from cross section estimates
- Transform data: Using ratios or first differences will get rid of any multicollinearity caused by common trend. But see discussion of dynamic models.
- Principal components and ridge regression

So no easy solution to multicollinearity, but have to be aware of the problem when dropping seemingly insignificant variables from the regression. Consider:

	Xs important	Xs not important
Include Xs	✓	inefficiency
Exclude Xs	specification bias	✓

1.3.3 Functional form

When estimate models often don't have information from theory on functional form.

- Use of linear is for simplicity of estimation
- Note that it is linear in the parameters that allows estimation by OLS not linearity in the variables
- There are methods of estimating non linear relations, but they are not straightforward so we try to find a way of transforming relation to make it linear in the parameters
- There are a wide variety of functional forms that can be approximated using the powers of the independent variables and applying OLS -though problems of multicollinearity possible

1.4 Outliers

1.4.1 Problem

Regression parameters can be influenced by a few extreme values or outliers

- Should be able to spot from a careful analysis of the residuals $\widehat{u}_i$
- In the case of a simple bivariate regression you can simply plot the data.
- Outlier is an observation that is very different: usually generated by some unusual factor
- Least squares estimates are very sensitive to outliers, particularly in small samples

1.4.2 Actions

- Drop the observations with large residuals and reestimate the equation. This should really be a last resort
- The outliers may provide important information. They may not be outliers at all. An example of this is the relation between infant mortality and GDP per capita in Asian countries.
- For cross section should maybe try to get more data rather than drop observations.
- Problem of what is an outlier also relates to leverage: need variation in the data or cant estimate any relationship. Its not always obvious when information on a system becomes an outlier.
- Can treat extreme observations with dummy variables

1.5 Non-normality

We need to assume normality for OLS to be MLE and for testing restrictions

- but it is not common to find variables that are normally distributed even after transformation
- outliers might be important
- need to take care beacuase of possible efficiency losses

1.6 Omitted variable bias

If we miss out an important variable it not only means our model is poorly specified it also means that any estimated parameters are likely to be biased.

- Incorrect omission of variables leads to biased estimates of the parameters that are included
- Incorrect inclusion only produces inefficient estimates, so don't have minimum variance
- So better to include the wrong variables rather than exclude the right ones.

1.7 Limited Dependent Variable Models

An important set of models in cross section work are those which deal with dependent variables that are limited in some way

Have considered how to deal with discrete variables in terms of dummy variables -as explanatory variables.

In some cases we may have a dummy dependent variable.

For example if we want to look at transport mode choice, what determines whether individuals use a car. We have:

$$\begin{aligned} y_i &= 1 \text{ if choose car} \\ y_i &= 1 \text{ otherwise} \end{aligned}$$

Now if we simply estimate an OLS regression

$$y_i = \beta x_i + u_i$$

Then this is called the Linear Probability Model

$$\begin{aligned} E(u_i) &= 0 \\ E(y_i \mid x_i) &= \beta x_i \text{ which can be interpreted in probability terms} \end{aligned}$$

Clearly u_i can only take two values

$$\begin{aligned} \text{when } y_i &= 1 \text{ then } u_i = 1 - \beta x_i \\ \text{when } y_i &= 0 \text{ then } u_i = -\beta x_i \end{aligned}$$

which means the variance

$$var(u_i) = E(u_i u_i)^2 = -\beta x_i (1 - \beta x_i)^2 + (1 - \beta x_i)(\beta x_i)^2 = E(y_i)\left[1 - E(y_i)\right]$$

is not constant and will vary with y. So u is heteroscedastic. We could overcome this problem with WLS but there is a more important problem and readily available alternatives. The problem is that while $E(y_i \mid x_i)$ may be interpreted as a probability it can lie outside 0 and 1.

One alternative is to use linear discriminant analysis rather than OLS. This minimises the ratio

$$\frac{\text{Between group variance}}{\text{Within group variance}}$$

of

$$y_i = \alpha + \beta x_i$$

But as Maddala shows this is very similar to an alternative and better approach.

Take

$$y_i = \alpha + \beta x_i + u_i$$

Then

$$\begin{aligned} P_i &= \Pr ob(y_i = 1) = \Pr ob(u_i > -(\alpha + \beta x_i) \\ &= 1 - F\left[-\alpha - \beta x_i\right] \end{aligned}$$

where F is the cumulative distribution. Now

$$P_i = F\left[\alpha + \beta x_i\right]$$

as

$$1 - F(-z) = F(z)$$

which we can estimate using maximum likelihood (ML) methods

$$L = \underset{y_i=1}{\Pi} P_i \Pi (1 - P_i)$$

The method we use depends up the assumption we make about the error term. The most common are

Logit: assume logistic distribution for u_i which means

$$P_i = \frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}}$$

or

$$\log\left[\frac{P_i}{1 - P_i}\right] = \alpha + \beta x_i$$

Note the interpretation of the coefficients differs from the LPM

Probit: assume a normal distribution for the u_i which means

$$P_i = \Phi(\alpha + \beta x_i) = \int_{-\infty}^{\frac{\alpha+\beta x_i}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp \frac{(-t^2)}{2} dt$$

These two are now very commonly available in econometrics and statistics packages. For more complex models it is customary to start with the linear probability model to get starting values.

The cumulative normal and logistic distributions are similar, so we would expect similar results. They are not, however, directly comparable and we need to make a constant adjustment. Amemiya suggests

$$1.6\widehat{\beta}_{\Phi} \approx \widehat{\beta}_{Logit}$$

Also

$$\begin{aligned} \widehat{\beta}_{LPM} &\approx 0.4\widehat{\beta}_{\Phi} \text{ except constant} \\ \widehat{\beta}_{LPM} &\approx 0.4\widehat{\beta}_{\Phi} + 0.5 \text{ for constant} \\ \widehat{\beta}_{LPM} &\approx 0.25\widehat{\beta}_{Logit} \text{ except constant} \\ \widehat{\beta}_{LPM} &\approx 0.25\widehat{\beta}_{Logit} + 0.5 \text{ for constant} \end{aligned}$$

This will work for probabilities between 30%and 70%, as over this range the logistic can easily be approximated by a straight line

In practice the LPM model will give acceptable results, but there is the issue of heteroscedasticity and nowadays it is easily to estimate logits and probits.

Note that these models differ form the usual ones in practice in that we cant interpret the coefficients directly -e.g. as elasticities. They are disaggregate models and estimate a probability for each observation, so when trying to forecast we have to aggregate. For the linear regression model

$$y_i = \alpha + \beta x_i \text{ and } \overline{y} = \alpha + \beta \overline{x}$$

but for the logit model:

$$P_i = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} \text{ but } \overline{P} \neq \frac{e^{\alpha+\beta \overline{x}}}{1+e^{\alpha+\beta \overline{x}}}$$

When interpreting the logit/probit results will often see them reported in a table which gives the average or the extreme values of the variables and then use the coefficients to give the probability. For example in mode choice you might indicate what an individual who has a really high probability will look like in terms of the explanatory variables and compare with one who has a very low probability.

Goodness of fit: Cant use conventional R^2 type of measure with limited dependent variable methods. Common to look at measure based on the likelihood ratio

$$\begin{aligned} \lambda &= \frac{L(\beta_0)}{L(\beta_0, ..., \beta_K)} \\ -2\log\lambda &\sim \chi^2_k \end{aligned}$$

Can also use this to test restrictions on subsets of coefficients.

Analogous to an R^2 is

$$\rho^2 = 1 - \frac{L^*(\beta_{0,...,}\ \beta_K)}{L*(\beta_0)}$$

which can be adjusted for degrees of freedom as well. Note that while this will lie between 0 and 1, in contrast to the R^2a perfect fit value is about 0.7 and a range of 0.2 to 0.4 can be considered a good fit. Might also consider the proportion of correct predictions

$$\frac{\text{no. correct predictions } (y_i = 1 \text{ and } P_i > 0.5)}{\text{no. observations}}$$

worth reporting, but has low discriminatory power. Maddala discusses some other measures

Another variant on these models is the Tobit model which deals with the situation when the observed value is either 0 or some positive number. For example if we are looking at what determines smoking we have 0 if the person does not smoke an the number of cigarettes when they do. So

$$y_i^* = \beta x_i + u_i$$

but observe y_i^* only if it is greater than 0

$$\begin{aligned} y_i &= y_i^* = \beta x_i + u_i \text{ if } y_i^* > 0 \text{ and } u_i \sim IN(0, \sigma^2) \\ y_i &= 0 \qquad\qquad\qquad\qquad \text{if } y_i^* \leq 0 \end{aligned}$$

Can estimate using MLE

$$L = \prod_{y_i=0} \frac{1}{\sigma} f\left(\frac{y_i - \beta x_i}{\sigma}\right) \prod_{y_i=0} F\left(\frac{-\beta x_i}{\sigma}\right)$$

1.8 Sample Selection Models

A useful approach for analysing the growth of companies is an empirical analysis within the framework of testing Gibrat's law (Dunne and Hughes, 1994 and Sutton, (1997), Caves (1998) provide reviews). This approach was used in the 1970s to analyses the reasons for an observed inexorable rise in concentration of manufacturing industry,. There was a concern that this would continue and lead to increasing monopoly power Hannah and Kay (1977). In fact as this problem was identified things were changing and there was a steady rise in the share of smaller firms in total output taking place. Gibrat's law stated that the probability distribution of growth rates was the same for all sizes of firms.

$$\frac{S_{it}}{S_{it-1}} = \varepsilon_{it}$$

If this held it suggested that the growth rate of companies would be random across size classes, a feature that would lead to the inexorable rise of concentration within industry and the economy. This can be tested by writing it as:

$$log S_{it} = \alpha \ + \ \beta \log S_{it-1} + \varepsilon_{it}$$

and testing if $\beta = 1$. If $\beta < 1$ smaller firms are growing faster than the larger firms and if $\beta > 1$ the larger firms are growing faster than the smaller firms. This can also be reparameterised as a growth rate equation

$$\Delta \log S_{it} = \alpha + (\beta - 1) \log S_{it-1} + \varepsilon_t$$

in this case the test is for the coefficient on $\log S_{it-1}$to be zero.

Another way of interpreting these regressions is to consider the model in log deviations form. define

$$\begin{aligned} y_{it} &= \log S_{it} - \log S_t \\ \log S_t &= N^{-1} \sum_{i=1}^{N} \log S_{it} \end{aligned}$$

then

$$y_{it} = \beta y_{it-1} + \varepsilon_{it}.$$

Squaring, summing over i and dividing by N, and taking expected values, noting that ε_{it} is independent of y_{it-1} gives

$$E(\frac{\sum y_{it}^2}{N}) = \beta^2 E(\frac{\sum y_{it-1}^2}{N}) + E(\frac{\sum \varepsilon_{it}^2}{N})$$

which gives the relationship determining the evolution of the variance of log firm size:

$$\sigma_t^2 = \beta^2 \sigma_{t-1}^2 + \sigma_\varepsilon^2.$$

This implies

$$1 = \beta^2 \frac{\sigma_{t-1}^2}{\sigma_t^2} + \frac{\sigma_\varepsilon^2}{\sigma_t^2}$$

or

$$\beta^2 \frac{\sigma_{t-1}^2}{\sigma_t^2} = 1 - \frac{\sigma_\varepsilon^2}{\sigma_t^2}$$

now the right hand side of this equation is the formula for the R^2 of the cross-section regression, so

$$\frac{\sigma_{t-1}^2}{\sigma_t^2} = \frac{R^2}{\beta^2}$$

This means that the evolution of the variance of log size, a measure of concentration, is determined by the ratio of the R^2 to β^2. Whether the variance increases or decreases depends both on β and the size of the stochastic shocks.

There are a number of econometric issues that arise. Growth persistence from previous periods can lead to serial correlation in the errors, but this is unlikely given the period over which we take the growth rates. Heteroscedasticity can be present if the variance of the growth decreases with size, as seems to be the case. This can lead to unbiased but inconsistent estimates and to be safe we will be using heteroscedastistic robust standard error estimates. Outliers may have an effect on the estimates, but it is not clear what an outlier is in this case. Given the focus on the top 100 companies we would certainly be loathe to drop any companies from the sample, but have to be aware that the results may be sensitive to extreme values. Finally a particularly important concern is the issue of sample selection. The way the model is formulated it is only possible to include companies that survive over the whole period. However, if the non surviving companies share certain characteristics, such as they are slow growing

then this can obviously bias the estimation results. More formally, what we have is:

$$\begin{aligned} logS_{it} &= \alpha + \beta \log S_{it-1} + \varepsilon_{it} \text{ if } S_{it} > 0 \\ &= 0 \text{ otherwise} \end{aligned}$$

thus

$$E(\log S_{it} \mid \log S_{it-1}, S_{it} > 0) = \alpha + \beta \log S_{it-1} + E(\varepsilon_{it} \mid S_{it} > 0)$$

with

$$\varepsilon_t \sim N(0, \sigma^2)$$

This can be written as

$$E(\log S_{it} \mid \log S_{it-1}, S_{it} > 0) = \alpha + \beta \log S_{it-1} + \sigma\lambda_i$$

where

$$\lambda_i = \frac{f(V_i)}{1 - F(V_i)} \text{ and } V_i = [\frac{\alpha + \beta \log S_{it-1}}{\sigma}]$$

with $f(..)$ the density function for the standard normal and $F(..)$ the distribution function for the standard normal. If there is sample selection bias and we were to estimate a simple OLS regression omitting $\sigma\lambda_i$ giving biased and inconsistent estimators.

For the two stage procedure let

$$\begin{aligned} d_i &= 1 \text{ when } S_{it} > 0 \\ d_i &= 0 \text{ otherwise} \end{aligned}$$

Then we can set up a likelihood function

$$\begin{aligned} L &= \Pi_{i=1}^{N} [\Pr(\varepsilon_i \leqslant -V_i]^{1-d_i} [\Pr(\varepsilon_i \leqslant -V_i]^{d_i} \\ &= \Pi\ F\left[\frac{V_i}{\sigma}\right]^{d_i} \left\{1 - F\left[\frac{V_i}{\sigma}\right]\right\}^{1-d_i} \end{aligned}$$

As $F(-t) = 1 - F(t)$ this is the likelihood function for the probit estimation on d_i and $E(d_i) = V_i/\sigma$. So we estimate a probit:

$$\Pr(d_i = 1) = P(V_i)$$

compute V_i and

$$\lambda_i = \left[\frac{f(V_i)}{1 - F(V_i)}\right]$$

For the second stage we use the consistent estimator of λ_i, $\widehat{\lambda}_i$ to estimate

$$E(S_{it} \mid S_{it-1}, S_{it} > 0) = \alpha + \beta S_{it-1} + \sigma\widehat{\lambda}_i$$

giving a consistent estimator of β.

It is also possible to use a maximum likelihood method , that uses this consistent estimator as a starting value to search for a solution on the highly non-linear likelihood function;

$$L = \Pi_{d_i=0} F\left(-V_i, \sigma^2\right) \ \Pi_{d_i=1}\left(S_{it} - V_i, \sigma^2\right)$$

now as $1 - F(-V_i, \sigma^2) = 1 - F(V_i, \sigma^2)$ which we call $1 - F_i$

$$L = \sum_{d_i=0} \ln(1 - F_i) - \frac{N-S}{2} ln\sigma^2 - \frac{1}{2\sigma^2} \sum_{d_i=1} \left(S_{it} - V_i, \sigma^2\right)$$

which can be solved using an iterative process such as Newton Raphson.

These models have a wide useage in applied work, especially in labour economics.

1.9 Econometric Methods and Surveys

When analysing survey data using statistic/econometrics can distinguish two approaches, descriptive and modelling, and the distinction can be important.

Descriptive:

Modelling:

This is particularly the case when dealing with the problem of surveys that have different probabilities of selection

Question is whether need to take account of weights in regression and it depends on what you are doing

1.9.1 Survey designs and regression

consider N_s a population of households and n_s sample of households

$$w_{is} = \left(\frac{N_s}{n_s}\right)$$

and so

$$\overline{x}_w = \frac{\sum_{s=1}^{S} \sum_{i=1}^{n_s} w_{is} x_{is}}{\sum_{s=1}^{S} \sum_{i=1}^{n_s} w_{is}} = \frac{\sum_{s=1}^{S} N_s \overline{x}_s}{\sum_{s=1}^{S} N_s} = \sum_{s=1}^{S} \frac{N_s}{N} \overline{x}_s = \overline{x}$$

Within each sector

$$y_s = \alpha_s + \beta_s x_s + u_s$$

Population weighted average is

$$\beta = \frac{1}{N}\sum_{s=1}^{S} N_s \beta_s$$

So if estimate the regression for each sector the population weighted estimate is

$$\widehat{\beta} = \sum_{s=1}^{S} \frac{N_s}{N}\widehat{\beta}_s$$

such regressions are regularly used when sectors are broad, but when there are few households in sector the parameters will be estimated imprecisely.

Even then worth considering as the variance is

$$var\left(\widehat{\beta}\right) = \sum_{s=1}^{S}\left(\frac{N_s}{N}\right)^2 var\left(\widehat{\beta}_s\right) = \sum_{s=1}^{S}\left(\frac{N_s}{N}\right)^2 \frac{\sigma_s^2}{\sum_{i=1}^{n_s} x_i^2}$$

where σ_s^2 is residual variance in stratum.

because the population fractions are squared (and is a fraction) β will be more precisely estimated than the individual β_ss

$$var\left(\widehat{\beta}\right) < var\left(\widehat{\beta}_s\right)$$

Common for researchers to estimte on all observations at once, either using the inflation factors to calciulate a weighted least squares or ignoring them.

In general OLS estimates will not yield any parameters of interest

- if all the β_s are the same then the OLS estimate will be consistent for the common β
- Even if the structure of the explanatory variables in each stratum is the same sample weighted average will be inconsistent unless the sample is a simple random sample with equal probabilities of acceptance
- It is a problem of population heterogeneity rather than sample design
- This mirrors inconsistency of unweighted mean for population mean

So:

- If population is homogeneous, meaning the coefficients are identical for each stratum both weighted and unweighted are consistent and unweighted is prefered as more efficient (Gauss Markov)

- If population not homogeneous both estimates are inconsistent
- So in neither case is there an argument for weighting

Weighted estimates justified:

- If have many strata and suspect heterogeneity but it is not systematically linked to other varaibles. Weighted estimates consistent if variation in parameters is random and independent of xs and if number of strata is large enough for weights mean to be zero.
- If consider regression as descriptive rather than structural. Are summarising characteristics of the population by the regression, heterogeneity and all

But if trying to estimate behavioral relation and if this is different in different parts of the population -weighting is at best useless.

1.9.2 Recommendations for practice

- If regressions primarily descriptive -exploring association by looking at mean of one variable conditional on others, then use weights and correct the standard errors for the design
- If modelling and concerned with heterogeneity and its interaction with sample design the more complex standard errors can get explicit formulas or use bootstrap, but program bootstrap to reflect sample design
- In practice it is clustering that has the largest effect on standard errors, conventional formulae overstate precision by ignoring dependence of observations within same cluster/psu. This is true for both descriptive and structural estimation

1.9.3 Dealing with heterogeneity and design

- One extreme is standard approach for modeller to assume homogenous behaviour across the subunits and pool the data ignoring weights.

  in fact wise to calculate both weighted and unweighted and comper or test differences using auxilliary regression with

$$Y_i = \alpha + \beta X_i + \gamma W X_i + v_i$$

and testing $\gamma = 0$

- Other extreme is consider behaviour to differ across subunits abd estimate separate regressions for each and then combine results using population wieghts> when distribution between groups are of interest can test for differences using covariance analysis.
- When many sectors can assume intersectoral heterogeneity random variation in parameters: weighted and unweighted resids will be heteroscedastic and dependent and neither weighted nor unweighted will be consistent
  - when explanatory variables differ within clusters or their are unequal numbers of observations in each cluster, then although OLS is inefficient the efficiency loss is typically small
  - but may still be large enough to justify correction
- Could estimate OLS and use residuals to correct for clusters using GLS type estimator