Introduction to Microeconometrics
Lecture Notes for Econ 220C

Yixiao Sun

Department of Economics,

University of California, San Diego

Spring 2004

Contents

Preface

The primary goal of Econ 220C is to introduce tools necessary to understand and implement empirical studies in economics focusing on issues other than time-series analysis. This course contains two parts. The first part deals with panel data models: (1) static panel data models (2) dynamic panel data models, and (3) other misc. panel topics. Multiple Equation GMM and Minimum Distance Estimator will be introduced and used to estimate some panel data models. The second part of the course deals with limited-dependent-variable models: (1) discrete choice models; (2) censored and truncated regression models, and (3) sample selection models. While the second part focuses mainly on cross sectional data, it also covers panel Probit/Logit, panel Tobit and panel attrition models.

We will study different issues in the specification, estimation and testing of these models with cross-sectional data and with panel data. The emphasis of the course is on both econometric ideas and econometric techniques. For some of the problem sets you will have to deal with actual data or perform simulation experiments. You should become familiar as soon as possible with some general features of the econometric package that you choose. MATLAB and GAUSS are widely used by econometricians. It seems that more and more people start using MATLAB. STATA seems to have gained increasing popularity in recent years among applied micro economists. SAS is another option.

Chapter 1

Introduction to Panel Data Modeling

1.1 Introduction

Recently empirical research in economics has been enriched by the availability of a wealth of new sources of data: Cross sections of individuals observed over time. The availability of panel data has stimulated a rapid growth in both methodological approaches and applications during the last twenty years.

The basic linear model is:

$$y_{it} = x_{it}\beta + \varepsilon_{it}, i = 1, ..., N, \ t = 1, ...T_i \tag{1.1}$$

If x_{it} contains no lagged dependent variables, the model is a static linear panel data model. Otherwise, it is a dynamic linear panel data model.

Remark 1 *T_i, the number of time periods may differ for each person. If $T_i = T$ for every i, the panel data set is said to be "balanced." Otherwise, it is "unbalanced." An important issue to determine at the outset is whether a panel data set is unbalanced due to endogenous causes, i.e., causes related to the economic mechanism we are trying to model. For example, if y_{it} is earnings and the richer people are more likely to drop out of the sample as time goes by because the value of their time is higher than others', then the data set is endogenously unbalanced. In such a case, though the basic model we are trying to fit is the linear regression (3.99), to take correct account of the fact that in such a case the relevant expression would be the conditional expectation*

$$E(y_{it}|x_{it} \textit{ and individual } i \textit{ stays in the sample at period } t)$$

we would need non-linear sample-selectivity methods. In other words, we would need to model the discrete mechanism characterizing the dummy variable: $d_{it} = 1$ *if individual* i *is in the sample in period* t*;* 0 *otherwise.*

Remark 2 *The set of explanatory variables may include:*

1. *variables that vary across individuals and time periods, e.g., wage, age, and years of experience. Denote them as* x_{it}.

2. *variables that are time-invariant, i.e., vary only across individuals, e.g., race and sex. Denote them as* x_i.

3. *variables that vary only over time but not across individuals, e.g., economy-wide unemployment, minimum-wage level, and other macroeconomic factors. Denote them as* x_t.

1.1.1 Types of Panel Data Sets

- Small T, large N (traditionally considered in panel data econometrics)

In micro panels, N is typically very large (several hundreds or even thousands) while T_i is quite small (ranging from 2 to 10 in most cases, and very rarely exceeding 20). If T is much smaller than N, the usual asymptotics is to let $N \to \infty$ with T fixed. Panel data set with small T dimension is often called traditional panels or micro panels.

- Small N, large T (Seemingly Unrelated Regression Equation (SURE))

  SURE: Consider a system of N equations without any feedback mechanism:

$$\begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_N \end{pmatrix} = \begin{pmatrix} X_1 & 0 & 0 & ... & 0 \\ 0 & X_2 & & & 0 \\ 0 & 0 & X_3 & & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & X_N \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_N \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ ... \\ e_N \end{pmatrix}$$

Let $y_i = (y_{1,i}, y_{2,i}, ..., y_{N,i})'$, $y = (y_1', ..., y_T')$ and define X and e similarly. We can write the above system as

$$y = X\beta + e$$

with

$$E(e|X) = 0.$$

and

$$E(e'e|X) = \Phi.$$

A special case is

$$\Phi = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1N} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2N} \\ \dots & \dots & \dots & \dots \\ \sigma_{N1} & \sigma_{N2} & \dots & \sigma_{NN} \end{pmatrix} \otimes I_T = \Sigma \otimes I_T$$

The GLS estimator of β is

$$\beta_{GLS} = \left(X'\Phi^{-1}X\right)^{-1}\left(X'\Phi^{-1}y\right).$$

which is a BLUE.

Under one of the following two conditions, OLS applied to each equation is equivalent to GLS when $\Phi = \Sigma \otimes I_T$

(i) $\Sigma = I_N$ (ii) $X_1 = X_2 = ... = X_N$)

The proof of this last condition is quite simple and is left as an exercise.

- Large N and T (panel time series, data fields)

For some macro panels and financial panels, both N and T can be large. We need to allow both N and T goes to infinity. This is the so called multidimensional asymptotics. We may let $N \to \infty$ first and then let $T \to \infty$ or let $T \to \infty$ first and then let $N \to \infty$ or let N and T go to ∞ at the same time but control the relative rate of expansions (i.e. $\sqrt{N}/T \to 0$). The first two asymptotics are called sequential asymptotics and the last one is called joint asymptotics.

- Small N and T (hopeless!)

1.1.2 Examples of Panel Data Sets

- Panel Study of Income Dynamics (PSID)

  http://www.isr.umich.edu/src/psid/index.html

- National Longitudinal Surveys of Labor Market Experience (NLS)
  http://www.bls.gov/nls/
- Penn World Table Data/Global Development,
  http://www.nuff.ox.ac.uk/Economics/Growth/summers.htm
- Network Growth Database,
  http://www.worldbank.org/research/growth/GDNdata.htm

1.2 Benefits and Limitations of Using Panel Data

1.2.1 Benefits of Using Panel Data

What is generally referred to as the panel data approach to economics research provides several major advantage over conventional cross sectional or time series data approaches. Both Hsiao (2003) in his seminal monograph and Baltagi (2002) in his excellent book provide extensive summaries.

- more informative data, more variability, more degrees of freedom and more efficiency.
- dynamics of adjustment
- identify and measure effects that can not identify by cross sectional or time series data alone. Repeated observations on the same unit allow identification in the presence of some types of unobservable, specifically, "permanent" unobserved differences across the countries, firms, or individuals that are related to the regressors of interest.
  - A cross section of women with 50% average yearly labor force participation rate. (a) each woman having 50% chance of being in the labor force in any given year (b) 50% of the women work all the time and 50% do not.
  - An estimator may be inconsistent if we only have time series data while it is consistent if both time series and cross sectional observations are available.

1.2.2 Limitations of Using Panel Data

- heterogeneity of the units and presence of unobservable can make estimation complex.
- the time series are generally too short to rely on asymptotics in the T dimension.

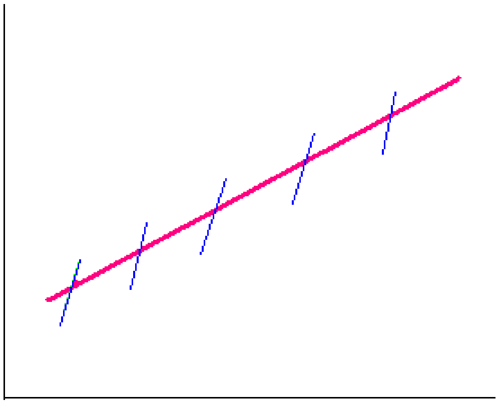

Figure 1.1: We are interested in the slope of the thin lines. If we do not control for heterogeneity in the intercepts, the fitted line will be the thick one. The estimated slope is obviously biased downward.

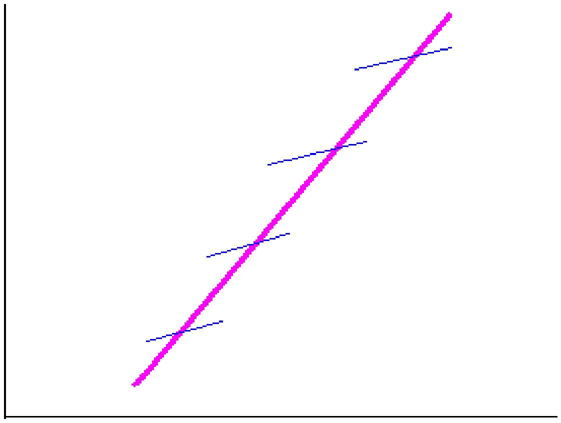

Figure 1.2: We are interested in the slope of the thin lines. If we do not control for heterogeneity in the intercepts, the fitted line will be the thick one. The estimated slope is obviously biased upward.

1.3 The Development of Panel Data Approach

1. First stage (1970s and early 1980s): Static error component models and random coefficient models.
2. Second stage (middle 1980s to the middle 1990s): Dynamic homogeneous panel data model.
3. Third Stage (from the middle 1990s to presents); Dynamic heterogenous model, Panel data with large N and T, Nonstationary panels.

1.4 Unobserved Heterogeneity (an example)

Example: state traffic fatality data

- The data are for 48 states, where each state is observed in T=7 time periods (each of years 1982, . . . , 1988)
- 40,000 highway traffic fatalities each year in the US
- Approx. 1/3 fatal crashes involves a driver who was drinking
- A study estimated that 25% of driver on the road between 1am and 3am have been drinking
- A driver who is legally drunk is 13 times as likely to cause a fatal crash

Objective: the effect of government policies designed to discourage drunk driving on the fatality rate

Indep. Var.: $\Longrightarrow$ Fatality rate: the number of annual traffic death per 10,000 people in a state

Dep. Var. $\Longrightarrow$ Beer tax: the "real" tax on a case of beer, i.e. the beer tax put into 1988 dollars.

Scatterplot[1]:

1982 Estimation:

$$\begin{aligned} Fatalityrate &= 2.01 + 0.15 Beertax \\ & \quad\ (0.15) \quad (0.13) \end{aligned} \tag{1.2}$$

1988 Estimation

$$\begin{aligned} Fatalityrate &= 1.86 + 0.44 Beertax \\ & \quad\ (0.11) \quad (0.13) \end{aligned} \tag{1.3}$$

[1] Figures in this section are reproduced from Stock and Watson (2002)

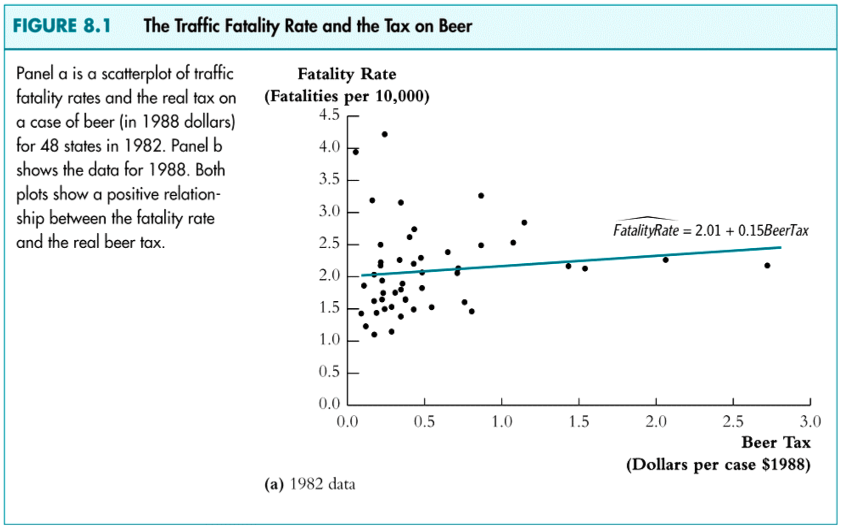

FIGURE 8.1 **The Traffic Fatality Rate and the Tax on Beer**

Panel a is a scatterplot of traffic fatality rates and the real tax on a case of beer (in 1988 dollars) for 48 states in 1982. Panel b shows the data for 1988. Both plots show a positive relationship between the fatality rate and the real beer tax.

(a) 1982 data

Figure 1.3:

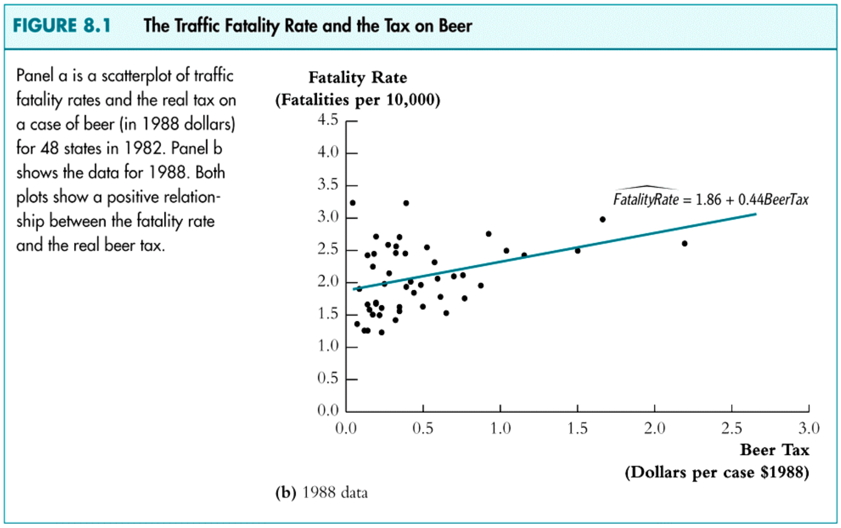

FIGURE 8.1 **The Traffic Fatality Rate and the Tax on Beer**

Panel a is a scatterplot of traffic fatality rates and the real tax on a case of beer (in 1988 dollars) for 48 states in 1982. Panel b shows the data for 1988. Both plots show a positive relationship between the fatality rate and the real beer tax.

(b) 1988 data

Figure 1.4:

- t1982 is not significant at 10% level while t1988 is significant at 1% level
- Higher tax are associate with more, not fewer traffic fatalities???
- Omitted variable bias: quality of the auto, highway conditions, social attitude toward drink and drive
- Solution: Collect all the relevant data and argument the simple regression But: some of these variables are not observable or measurable
- Keep those variables constant across different period ⇒ fixed effect model

Let Z_i be a variable that determines the fatality rate in state I but does not change over time. Let Y = Fatality rate and X = Beertax, then

$$Y_{it} = b_0 + b_1 X_{it} + b_2 Z_i + u_{it} \tag{1.4}$$

When $t = 1982$, we have

$$Y_{i1982} = b_0 + b_1 X_{i1982} + b_2 Z_i + u_{i1982} \tag{1.5}$$

When $t = 1988$, we have

$$Y_{i1988} = b_0 + b_1 X_{i1988} + b_2 Z_i + u_{i1988} \tag{1.6}$$

Subtracting (1.5) from (1.6), we get

$$Y_{i,1988} - Y_{i1982} = b_1(X_{i1988} - X_{i1982}) + (u_{i1988} - u_{i1982}) \tag{1.7}$$

Cultural attitudes toward drinking and driving affect the level of drunk driving and thus the fatality rate. However, if they do not change over time, then they do not produce any change in fatalities in the state. The changes must arise from other sources.

1.5 Unobserved Heterogeneity (another example)

Consider agricultural Cobb-Douglas production function. Let

$$\begin{aligned} Y_{it} &= \text{ log output} \\ X_{it} &= \text{log of a variable input} \\ Z_i &= \text{An input that remains constant over time (soil quality)} \\ u_{it} &= \text{A stochastic input which is outside the farmers' control (rainfall)} \end{aligned}$$

Suppose Z_i is known by the farmer but not by the econometrician. Then the profit maximizing choice of X_{it} will depend on Z_i. Therefore X_{it} will be (positively) correlated with Z_i. A pooled panel regression estimator of b_1 will have a upward bias.

For more examples on unobserved heterogeneity, See Arellano (2003, pages 8-10)

FIGURE 8.2 **Changes in Fatality Rates and Beer Taxes, 1982–1988**

This is a scatterplot of the *change* in the traffic fatality rate and the *change* in real beer taxes between 1982 and 1988 for 48 states. There is a negative relationship between changes in the fatality rate and changes in the beer tax.

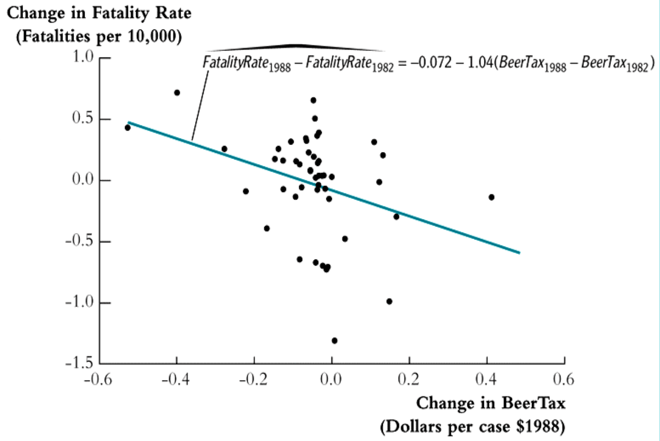

Figure 1.5:

1.6 Clustered Sampling

N and T do not necessarily refer to number of individuals and time periods respectively. Other examples include families and family members, industries and firms. Many types of cross sectional survey data are obtained through “cluster” sampling. Certain geographical units are first selected (e.g. villages), then individuals are sampled within each village. Thus, the village from which the individual observation comes may be thought of as one dimension of the data. Thus panel data methods are of special importance in research in developing countries. A simple model is

$$y_{ci} = \alpha_c + X_{ci}\beta + u_{ci}, \;\; c = 1, ...C, \;\; i = 1, ...I_c$$

where c indexes the cluster and i indexed individuals in the cluster. If we have a large number of clusters and relatively small group sizes ($\max(I_c)$ is small), then we have a traditional linear panel data model)

If you will do research in development and deal with survey data, it is worthwhile reading the book by Deaton (1997).

Bibliography

[1] Arellano, M. Panel Data Econometrics. Oxford University Press, 2003

[2] Baltagi, Badi H., Econometric Analysis of Panel Data, John Wiley & Sons, 2002.

[3] Deaton, Angus, The Analysis of Household Survey, The John Hopkins University Press, 1997.

[4] Hsiao, Cheng, Analysis of Panel Data, Cambridge University Press, 2003.

[5] Stock, J. and M. Watson, Introduction to Econometrics, Addison and Wesley, 2002

Chapter 2

Static Panel Data Models

2.1 The Static Model in Matrix Form

2.1.1 The Model

In matrix notation, the model

$$y_{it} = x_{it}\beta + \varepsilon_{it}, i = 1, ..., N,\ t = 1, ...T_i \tag{2.1}$$

can be written as:

$$y = X\beta + \epsilon$$

where y and ϵ are NT $\times 1$ vectors, and X is an $NT \times k$ matrix. The convention is to stack observations in groups of all time observations for each individual, e.g.,

$$y = (y_{11}, y_{12}, ..., y_{1,T_1}, y_{21}, y_{22}, ..., y_{2T,} ..., y_{N,1}, y_{N,2}, ..., y_{N,T})'$$

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix} \text{ and } X_i = \begin{pmatrix} x_{i1}^{(1)} & x_{i1}^{(2)} & ... & x_{i1}^{(k)} \\ x_{i1}^{(1)} & x_{i1}^{(2)} & ... & x_{i1}^{(k)} \\ ... & ... & ... & ... \\ x_{i,T}^{(1)} & x_{i,T}^{(2)} & ... & x_{i,T}^{(k)} \end{pmatrix} = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iT} \end{pmatrix} \tag{2.2}$$

In this section, we assume x_{it} contains no lagged dependent variables. So we consider only the static model here.

2.1.2 Fixed Effects or Random Effects?

The simple one-factor error component model is:

$$\varepsilon_{it} = \alpha_i + u_{it}$$

In our simple example

$$\alpha_i = b_2 Z_i \tag{2.3}$$

- α_i : individual effect or individual heterogeneity; u_{it}: the idiosyncratic error.
- In methodological papers, one often discuss about whether α_i should be treated as a random effect (random variable) or a fixed effect (parameters to be estimated) .
- The key issue is whether α_i is correlated with x_i or put in a stronger form whether $E(\alpha_i|x_{it}) = 0$. Following Wooldridge (Chapter 10), we treat α_i as a random variable. When $E(\alpha_i|x_{it}) = 0$, we say the model is a random-effects model. Otherwise, it is a fixed-effects model (We will also call the model a fixed-effects model if $\alpha_i's$ are treated as parameters to be estimated). Some authors refer to the random-effects model and fixed-effects model as the uncorrelated effect model and correlated effect model, respectively.
- Error Component Model

2.1.3 Strict Exogeneity Assumption

Assume

$$E\left(y_{it}|x_{i1},x_{i2},...x_{iT},\alpha_i\right) = E\left(y_{it}|x_{it},\alpha_i\right) = x_{it}\beta + \alpha_i \tag{2.4}$$

Compare

$$E\left(y_{it}|x_{i1},x_{i2},...,x_{iT}\right) = E\left(y_{it}|x_{it}\right) = x_{it}\beta \tag{2.5}$$

The assumption in (2.4) is the same as

$$E\left(u_{it}|x_{i1},x_{i2},...,x_{iT},\alpha_i\right) = 0 \tag{2.6}$$

which implies that

$$E\left(u_{it}x_{is}\right) = 0 \text{ for all s and } t. \tag{2.7}$$

This is certainly stronger than zero contemporaneous correlation.

2.1.4 Some Counter Examples

Example 3 *Program Evaluation:*

$$\log(w_{it}) = \theta_t + z_{it}\gamma + \delta_1 prog_{it} + \alpha_i + u_{it} \tag{2.8}$$

(a) Omitted variable bias story. The ability α_i is likely to be correlated with $prog_{it}$
(b) Feedback effect $u_{it} \Rightarrow prog_{it+1}$

Example 4 *Lagged Dependent Variable.*

$$y_{i,t} = \beta y_{i,t-1} + \alpha_i + u_{it} \tag{2.9}$$

(a) $y_{i,t-1}$ is correlated with α_i; u_{it} is correlated with $y_{i,t+s}$ for $s \geq 0$.
(b) This is the topic for the next chapter.

2.2 Estimation: Random-effects Approach

2.2.1 Assumptions

We first assume the following is true
Assumption RE.1(a): $E\left(u_{it}|X_i, \alpha_i\right) = 0$; $E\left(\alpha_i|X_i\right) = E\left(\alpha_i\right) = 0$,
Assumption RE.1(b): α_i is i.i.d. over i, u_{it} is i.i.d. over i and t, is independent of every α_j for all i, j, t

Assumption RE.1(b) is very strong as it rules out cross sectional dependence and the series dependence. While we can easily allow u_{it} to be correlated over time, it is more difficult to allow cross-sectional dependence, especially when N is large for a small T. Panel data models with cross sectional dependence have attracted much attention in recent years. In this course, we main the assumption of cross-section independence.

The presence of the time-invariant random-effect α_i implies the presence of persistent unobserved heterogeneity and the variance-covariance structure:

$$E\varepsilon_{it}\varepsilon_{js} = \begin{cases} \sigma_\alpha^2 + \sigma_u^2 & \text{if } i = j \text{ and } t = s \\ \sigma_\alpha^2 & \text{if } i = j \text{ and } t \neq s \\ 0 & \text{if } i \neq j \end{cases} \tag{2.10}$$

Error-components structures imply serial correlation in the error terms. Hence OLS estimation in such models will not be BLUE and will have a variance-covariance matrix not equal to $\sigma^2(X'X)^{-1}$.

Let $\varepsilon_i = (\varepsilon_{i1}, ..., \varepsilon_{iT})'$ and $\Omega = E\varepsilon_i\varepsilon_i'$ then

$$\begin{aligned}\Omega &= \begin{pmatrix} \sigma_\alpha^2 + \sigma_u^2 & \sigma_\alpha^2 & ... & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_u^2 & \cdots & \vdots \\ \vdots & & \ddots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & & & \sigma_\alpha^2 + \sigma_u^2 \end{pmatrix} && (2.11) \\ &= \sigma_u^2 I_T + \sigma_\alpha^2 J_T && (2.12)\end{aligned}$$

where I_T is $T \times T$ identity matrix and J_T is $T \times T$ is the matrix with unity in every element.

Note that $E\epsilon\epsilon' = I_N \otimes \Omega,$ the GLS estimator is

$$\begin{aligned}\hat{\beta}_{REGLS} &= \left\{X' \left(I_N \otimes \Omega\right)^{-1} X\right\}^{-1} \left\{X' \left(I_N \otimes \Omega\right)^{-1} Y\right\} && (2.13) \\ &= \left\{X' I_N \otimes \Omega^{-1} X\right\}^{-1} \left\{X' I_N \otimes \Omega^{-1} Y\right\} && (2.14)\end{aligned}$$

Now $X'\left(I_N \otimes \Omega^{-1}\right) X$ is

$$\begin{aligned} & X' \begin{pmatrix} \Omega^{-1} & 0 & ... & 0 \\ 0 & \Omega^{-1} & \cdots & \vdots \\ \vdots & & \ddots & 0 \\ 0 & & & \Omega^{-1} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix} && (2.15) \\ &= \left(X_1', X_2', ...X_N'\right) \begin{pmatrix} \Omega^{-1}X_1 \\ \Omega^{-1}X_2 \\ \vdots \\ \Omega^{-1}X_N \end{pmatrix} && (2.16) \\ &= \sum_{i=1}^{N} X_i'\Omega^{-1}X_i && (2.17)\end{aligned}$$

Similarly $X'\left(I_N \otimes \Omega^{-1}\right) Y = \sum_{i=1}^{N} X_i'\Omega^{-1}y_i$.Therefore

$$\hat{\beta}_{REGLS} = \left(\sum_{i=1}^{N} X_i'\Omega^{-1}X_i\right)^{-1} \left(\sum_{i=1}^{N} X_i'\Omega^{-1}y_i\right). \qquad (2.18)$$

For consistency of the GLS estimator, we require:
Assumption RE.2. Rank$\big(EX_i'\Omega^{-1}X_i\big) = k$.

For the efficiency of the GLS estimator, we assume:
Assumption RE.3 $E(\alpha_i^2|X_i) = \sigma_\alpha^2$, and $E(u_i u_i'|X_i) = \sigma_u^2 I_T$.

2.2.2 Asymptotic Inference

- Under assumptions RE.1 and RE.2, the GLS estimator is consistent.
- Under assumptions RE.1, RE.2 and RE.3, GLS estimator is efficient in the class of linear and unbiased estimators.

The estimator in (2.18) is not feasible. As in the typical GMM setup, we replace Ω by I_T to get an initial consistent estimator of β. This estimate is the pooled OLS estimator.

$$\hat{\beta}_{OLS} = \left(\sum_{i=1}^N X_i'X_i\right)^{-1}\left(\sum_{i=1}^N X_i'y_i\right) \tag{2.19}$$

$$= \sum_{i=1}^N \left(\sum_{j=1}^N X_j'X_j\right)^{-1} X_i'X_i \left(X_i'X_i\right)^{-1} X_i'y_i \tag{2.20}$$

$$= \sum_{i=1}^N W_i \left(X_i'X_i\right)^{-1} X_i'y_i = \sum_{i=1}^N W_i \hat{\beta}_{OLS}^{(i)} \tag{2.21}$$

where

$$W_i = \left(\sum_{j=1}^N X_j'X_j\right)^{-1} X_i'X_i, \tag{2.22}$$

and

$$\hat{\beta}_{OLS}^{(i)} = \left(X_i'X_i\right)^{-1} X_i'y_i \tag{2.23}$$

is the OLS estimator using only the time series observations for individual i. With $\hat{\beta}_{OLS}$, we can construct estimate of σ_α^2 and σ_u^2. For example, $\sigma_\varepsilon^2 = \sigma_\alpha^2 + \sigma_u^2$ can be estimated by

$$\widehat{\sigma}_\varepsilon^2 = \frac{1}{NT-k}\sum_{i=1}^N\sum_{t=1}^T (y_{it} - x_{it}\beta)^2. \tag{2.24}$$

Note that $\sigma_\alpha^2 = E\varepsilon_{it}\varepsilon_{is}$ for $t \neq s$, so it can be estimated by

$$\widehat{\sigma}_\alpha^2 = \frac{1}{NT(T-1)/2-k}\sum_{i=1}^N\sum_{t=1}^{T-1}\sum_{s=t+1}^T (y_{it} - x_{it}\widehat{\beta}_{OLS})(y_{is} - x_{is}\widehat{\beta}_{OLS}) \tag{2.25}$$

Plugging in $\widehat{\sigma}_{\varepsilon}^2$ and $\widehat{\sigma}_{\alpha}^2$ into the definition of Ω yields $\widehat{\Omega}$. Using $\widehat{\Omega}$, we get the feasible GLS estimator

$$\hat{\beta}_{REFGLS} = \left(\sum_{i=1}^{N} X_i' \widehat{\Omega}^{-1} X_i \right)^{-1} \left(\sum_{i=1}^{N} X_i' \widehat{\Omega}^{-1} y_i \right). \tag{2.26}$$

As a practical matter, $\widehat{\sigma}_{\alpha}^2$ may not be positive. A negative value of $\widehat{\sigma}_{\alpha}^2$ indicates negative correlation in u_{it}, probably a substantial amount, which means one of our assumptions is violated.

Under assumptions RE.1, RE.2 and RE.3, $\hat{\beta}_{RE}$ is asymptotically equivalent to the infeasible GLS estimator $\hat{\beta}_{GLS}$.

2.2.3 Understanding the GLS Estimator

It may be huge task to invert the matrix Ω or V where

$$V = E\epsilon\epsilon' = I_N \otimes \Omega = \sigma_u^2 \left(I_N \otimes I_T \right) + \sigma_\alpha^2 \left(I_N \otimes J_T \right) \tag{2.27}$$

Fortunately, we have an analytical expression of V^{-1}. Let

$$\begin{aligned} P &= I_N \otimes \overline{J}_T \text{ where } \overline{J}_T = J_T/T \\ Q &= I_{NT} - P \end{aligned} \tag{2.28}$$

then

$$P' = P, P^2 = P, \ Rank(P) = Trace(P) = N, \tag{2.29}$$

and

$$Q' = Q, Q^2 = Q, Rank(Q) = Trace(Q) = NT - N. \tag{2.30}$$

Exercise 5 *Suppose A is a symmetric idempotent matrix. Prove that $a_{ii} \in [0,1]$ where a_{ii} is any diagonal element of matrix A.*

Exercise 6 *Suppose A is a symmetric idempotent matrix. Prove that $Rank(A) = Trace(A)$.*

Note that P is the matrix which averages the observations across time for each individual and Q is the demeaning operator, which removes the "within" means from a vector or matrix. More specifically,

$$PX = \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_N \end{pmatrix} \otimes l_T, \ QX = \begin{pmatrix} X_1 - \bar{X}_1 \\ X_2 - \bar{X}_2 \\ \vdots \\ X_N - \bar{X}_N \end{pmatrix} \text{ and } X_i - \bar{X}_i = \begin{pmatrix} x_{i1} - \frac{1}{T}\sum_{t=1}^T x_{it} \\ x_{i2} - \frac{1}{T}\sum_{t=1}^T x_{it} \\ \vdots \\ x_{iT} - \frac{1}{T}\sum_{t=1}^T x_{it} \end{pmatrix}. \tag{2.31}$$

Then

$$\begin{aligned} V &= \sigma_u^2 \left(I_N \otimes I_T\right) + \sigma_\alpha^2 \left(I_N \otimes J_T\right) \\ &= \sigma_u^2 (P+Q) + T\sigma_\alpha^2 P \\ &= \left(T\sigma_\alpha^2 + \sigma_u^2\right) P + \sigma_u^2 Q \\ &: = \sigma_1^2 P + \sigma_u^2 Q \end{aligned} \tag{2.32}$$

But

$$V^{-1} = \sigma_1^{-2} P + \sigma_u^{-2} Q \tag{2.33}$$

as

$$\left(\sigma_1^{-2} P + \sigma_u^{-2} Q\right)\left(\sigma_1^2 P + \sigma_u^2 Q\right) = P + 0 + 0 + Q = I_{NT}. \tag{2.34}$$

In fact

$$V^r = \sigma_1^{2r} P + \sigma_u^{2r} Q. \tag{2.35}$$

In particular,

$$V^{-1/2} = \sigma_1^{-1} P + \sigma_u^{-1} Q. \tag{2.36}$$

If we premultiply the regression model

$$y = X\beta + \epsilon \tag{2.37}$$

by

$$\sigma_u V^{-1/2} = \left(\sigma_u / \sigma_1\right) P + Q, \tag{2.38}$$

then we have

$$\left(y_{it} - \theta \bar{y}_{i,.}\right) = \left(x_{it} - \theta \bar{x}_{i,.}\right)\beta + \left(\varepsilon_{it} - \theta \bar{\varepsilon}_{i,.}\right) \tag{2.39}$$

where

$$\theta = 1 - \sigma_u / \sigma_1 \tag{2.40}$$

and the error term $(\varepsilon_{it} - \theta\bar{\varepsilon}_{i,.})$ are uncorrelated and have the same variance across all i and t. Note that by definition, the variance of $\varepsilon_{it} - \theta\bar{\varepsilon}_{i,.}$ is σ_u^2. So OLS is BLUE if it is based on the above regression model.

2.2.4 A General FGLS Analysis

If the error term u_{it} are generally heteroscedastic and serially correlated across t.Then

$$\begin{aligned} \Omega &= \begin{pmatrix} \sigma_\alpha^2 + \sigma_u^2 & \sigma_\alpha^2 + \sigma_{u,12}^2 & \cdots & \sigma_\alpha^2 + \sigma_{u,1n}^2 \\ \sigma_\alpha^2 + \sigma_{u,12}^2 & \sigma_\alpha^2 + \sigma_u^2 & \cdots & \vdots \\ \vdots & & \ddots & \sigma_\alpha^2 + \sigma_{u,n-1,n}^2 \\ \sigma_\alpha^2 + \sigma_{u,n1}^2 & & & \sigma_\alpha^2 + \sigma_u^2 \end{pmatrix} \\ &= \Omega_u + \sigma_\alpha^2 J_T \end{aligned} \tag{2.41}$$

In this case, Ω can be estimated by

$$\widehat{\Omega} = \frac{1}{N}\sum_{i=1}^{N}\widehat{\varepsilon}_i\widehat{\varepsilon}_i' \tag{2.42}$$

where $\widehat{\varepsilon}_i$ is the pooled OLS residual.

- The GLS estimator with above variance estimator is efficient regardless of assumption RE.3.
- Other possible restriction on the correlation structure of $\{u_{it}\}$.

2.3 Estimation: Fixed-effects Approach

2.3.1 Assumptions

Again consider the liner unobserved effect model:

$$y_{it} = x_{it}\beta + \alpha_i + u_{it} \tag{2.43}$$

Now we assume that x_{it} and α_i are correlated. In this case, the random effect estimator is biased.

Assumption FE.1: $E(u_{it}|X_i, \alpha_i) = 0$.

- If x_{it} contains some time invariant variables, then we can not identify the effects of these time invariant variables on y_{it}.
- For individuals, factors such as race and gender can not be included in x_{it}
- For firms, industry can not be included in x_{it}
- Only requires that each element of x_{it} varies over time for some cross sectional units.

2.3.2 Estimation Strategy

The idea is to transform the equation to eliminate the unobserved effect α_i. There are several transformation that can be used for this purpose. Recalled that we already used “first difference” for a two-period model. Now we consider fixed effects transformation, also called the within transformation.

- Averaging equation $y_{it} = x_{it}\beta + \alpha_i + u_{it}$ over t to get

$$\bar{y}_i = \bar{x}_i\beta + \alpha_i + \bar{u}_i \tag{2.44}$$

- Subtracting the above equation from $y_{it} = x_{it}\beta + \alpha_i + u_{it}$ to get

$$\begin{aligned} y_{it} - \bar{y}_i &= (x_{it} - \bar{x}_i)\,\beta + (u_{it} - \bar{u}_i) \text{ or} \\ \ddot{y}_{it} &= \ddot{x}_{it}\beta + \ \ddot{u}_{it} \end{aligned} \tag{2.45}$$

In matrix forms, equations (2.44) and (2.45) are nothing but

$$Py = PX\beta + P\epsilon \tag{2.46}$$

and

$$Qy = QX\beta + Qu \tag{2.47}$$

respectively.

Can we use the OLS estimator on (2.45)? Note that

$$E(u_{it} - \bar{u}_i)\,(x_{it} - \bar{x}_i) = 0 \tag{2.48}$$

or the following stronger orthogonal condition:

$$E(u_{it} - \bar{u}_i|\,(x_{is} - \bar{x}_i)) = 0 \text{ for all } t \text{ and } s, \tag{2.49}$$

So the OLS estimator is consistent and unbiased. Note that the above condition will not hold if we only assume that $E(u_{it}|x_{it}, \alpha_i) = 0$.

FE.2 Rank condition: rank$(E(X'QX)) = k$.

$$\begin{aligned} \widehat{\beta}_{FE} &= \left(\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)\,(x_{it} - \bar{x}_i)'\right)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)\,(y_{it} - \bar{y}_i) \\ &= \left(X'QX\right)^{-1}\left(X'QY\right) \end{aligned} \tag{2.50}$$

A computational warning: $NT - k$ or $N(T-1) - k$?

Why not use GLS on (2.47)? Note that $EQuu'Q' = EQQ' = Q$, which is degenerate. The GLS estimator is

$$\left(X'Q'Q^{-}QX\right)^{-1}\left(X'Q'Q^{-}Qy\right) = \left(X'QX\right)^{-1}\left(X'QY\right) = \widehat{\beta}_{FE} \tag{2.51}$$

where Q^{-} is the generalized inverse of Q satisfying $Q'Q^{-}Q = Q$. So the GLS estimator is the same as the OLS estimator. There is no efficiency gain at all.

- Kruskal theorem:

$$\underset{NT\times NT, NT\times k}{Vx} = \underset{NT\times k, k\times k}{xA} \tag{2.52}$$

  for some matrix A, then OLS=GLS.

In the present context, $V = Q$, $x = QX$, the condition reduces to $QX = QXA$, which is an identity when $A = I_k$

If we estimate β using OLS based on equation (2.46), then we get the between estimator:

$$\widehat{\beta}_{BE} = \left(X'PX\right)^{-1}\left(X'PY\right) \tag{2.53}$$

- $\widehat{\beta}_{BE}$ is inconsistent under the fixed effect assumption because $\bar{x}_i$ and α_i are correlated.
- However, it is consistent under the random effect assumption. It is inefficient because it discards the time series information in the data set.

2.3.3 Asymptotic Inference

We maintain the following assumption:

Assumption FE.3: $E(u_i u_i'|x_i, \alpha_i) = \sigma_u^2 I_T$

$$\widehat{\beta}_{FE} - \beta = \left(X'QX\right)^{-1}\left(X'QU\right) \Rightarrow N(0, \sigma_u^2\left(EX'QX\right)^{-1}) \tag{2.54}$$

Now define the fixed effects residual

$$\begin{aligned}
\widehat{u} &= Qy - QX\left(X'QX\right)^{-1}X'Qy \\
&= QX\beta + Qu - QX\left(X'QX\right)^{-1}X'Q\left(QX\beta + u\right) \\
&= Qu - QX\left(X'QX\right)^{-1}X'Qu \\
&= (I - QX\left(X'QX\right)^{-1}X'Q)Qu
\end{aligned} \tag{2.55}$$

Note

$$\begin{aligned}
\widehat{u}'\widehat{u} &= u'Q(I - QX\left(X'QX\right)^{-1}X'Q)Qu \\
&= u'Qu - u'QX\left(X'QX\right)^{-1}X'Qu
\end{aligned} \tag{2.56}$$

So

$$\begin{aligned} E\widehat{u}'\widehat{u} &= Eu'Qu - Eu'QX\left(X'QX\right)^{-1}X'Qu \\ &= EtrQu * u' - EtrQX\left(X'QX\right)^{-1}X'Qu \times u' \\ &= N(T-1)\sigma_u^2 - tr\left(QX\left(X'QX\right)^{-1}X'Q\right)\sigma_u^2 \\ &= \left(N(T-1)-k\right)\sigma_u^2 \end{aligned}$$

Thus, a unbiased estimate of σ_u^2 is

$$\widehat{\sigma}_u^2 = \frac{SSR}{N(T-1)-k} \tag{2.57}$$

2.3.4 Dummy Variable Regression

Traditional approaches to fixed effects estimation view the $\alpha_i's$ as parameters to be estimated along with β. How would we estimate α_i?

$$\alpha_i = \bar{y}_i - \bar{x}_i\widehat{\beta}_{FE} \tag{2.58}$$

Alternatively, we can use the least squares dummy variable (LSDV) estimation. Define $d_{ij}=1$ if $i=j$ and $d_{ij}=0$ otherwise. Let

$$d_i = \begin{pmatrix} d_{i1} \\ d_{i2} \\ \vdots \\ d_{iN} \end{pmatrix}', \; \alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{pmatrix}. \tag{2.59}$$

Note that $d_i'\alpha = \alpha_i$, so

$$\begin{aligned} y_{it} &= x_{it}\beta + d_i\alpha + u_{it} \\ &= (x_{it}, d_i)\begin{pmatrix} \beta \\ \alpha \end{pmatrix} + u_{it} \\ : &= z_{it}\gamma + u_{it} \end{aligned}$$

Therefore the linear model reduces to the usual form.

Important difference between α and β

- $\widehat{\beta}_{FE}$ is consistent with fixed T as $N \to \infty$
- $\widehat{\alpha}_i$ is a unbiased estimator for α_i but may not be consistent for a fixed T.
- Incidental parameter problem

2.3.5 Robust Variance Matrix Estimator

The Fixed effect estimator is consistent and asymptotically normal under assumptions FE.1 and FE.2. But without FE.3, the variance of $\widehat{\beta}_{FE}$ is not $\sigma_u^2 \left(EX'QX\right)^{-1}$. Note that

$$\widehat{\beta}_{FE} - \beta = \left(X'QX\right)^{-1} X'QU. \tag{2.60}$$

In the presence of possible heteroscedasticity and autocorrelation, the variance of $\widehat{\beta}_{FE}$ is

$$E\left(X'QX\right)^{-1} X'QUU'XQ\left(X'QX\right)^{-1} \tag{2.61}$$

But

$$\begin{aligned} X'QUU'XQ &= \sum_{i=1}^{N} \left(X_i - \bar{X}_i\right)' u_i u_i' \left(X_i - \bar{X}_i\right) && (2.62) \\ X'QX &= \sum_{i=1}^{N} \left(X_i - \bar{X}_i\right)' \left(X_i - \bar{X}_i\right), && (2.63) \end{aligned}$$

As a consequence,

$$\begin{aligned} Var\left(\widehat{\beta}_{FE}\right) &= E\left(\sum_{i=1}^{N} \left(X_i - \bar{X}_i\right)' \left(X_i - \bar{X}_i\right)'\right)^{-1} \sum_{i=1}^{N} \left(X_i - \bar{X}_i\right)' u_i u_i' \left(X_i - \bar{X}_i\right) \\ &\quad \times \left(\sum_{i=1}^{N} \left(X_i - \bar{X}_i\right)' \left(X_i - \bar{X}_i\right)'\right)^{-1} \end{aligned} \tag{2.64}$$

An estimates of the middle term $\sum_{i=1}^{N} \left(X_i - \bar{X}_i\right)' u_i u_i' \left(X_i - \bar{X}_i\right)$ is

$$\sum_{i=1}^{N} \left(X_i - \bar{X}_i\right)' \widehat{u}_i \widehat{u}_i' \left(X_i - \bar{X}_i\right) \tag{2.65}$$

- Need large N to deliver a good variance estimator
- The above variance formula is valid regardless of Assumption FE.3.

2.3.6 Robust Variance Matrix Estimator for Large T and Fixed N

The previous distribution theory for small T and large N allows for arbitrary time series dependence but replied on cross-sectional independence. With large T and fixed N, we can allow for arbitrary cross sectional dependence by relying on sufficiently weak time series dependence.

Note that

$$\widehat{\beta}_{FE} - \beta = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \bar{x}_i) u_{it} \tag{2.66}$$

To estimate the asymptotic variance of $\widehat{\beta}_{FE} - \beta$ for large T, we only need to calculate

$$V = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} \hat{v}_t \hat{v}_s'$$

where

$$\hat{v}_t = \sum_{i=1}^{N} (x_{it} - \bar{x}_i) \hat{u}_{it} \text{ and } \hat{u}_{it} = y_{it} - \hat{\alpha}_i - x_{it}\hat{\beta}_{FE}.$$

But V is nothing but the long run variance of $1/\sqrt{T}\sum_{t=1}^{T} v_t$. Using the HAC/long run variance estimator used in the time series literature, we can estimate V by

$$\frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} \hat{v}_t k(\frac{t-s}{M}) \hat{v}_s'$$

for some kernel function k and bandwidth parameter M. The above estimator is consistent for V for 'mixing' data.

2.4 Estimation: First Differencing Approach

Lagging

$$y_{it} = x_{it}\beta + \alpha_i + u_{it} \tag{2.67}$$

one period and subtracting yields

$$\Delta y_{it} = \Delta x_{it}\beta + \Delta u_{it} \tag{2.68}$$

The first-difference (FD) estimator is the pooled OLS estimator for the above regression

Assumption FD.1: Same as Assumption FE.1

Under the above assumption, we have $E(\Delta u_{it}|\Delta x_{i2}, \Delta x_{i3}, ..., \Delta x_{iT}) = 0$. So the $\widehat{\beta}_{FD}$ is consistent and unbiased. Note that

$$\begin{aligned} \Delta x_{it} \Delta u_{it} &= (x_{it} - x_{it-1})(u_{it} - u_{it-1}) \\ &= x_{it}u_{it} - x_{it-1}u_{it} - x_{it}u_{it-1} + x_{it-1}u_{it-1} \end{aligned} \tag{2.69}$$

So $E\Delta x_{it}\Delta u_{it}$ may not equal to zero if u_{it} is correlated with x_{it-1}, x_{it} or x_{it+1}.

Assumption FD.2: Rank$(\sum_{t=2}^{T} E\Delta x_{it}\Delta x_{it}') = k$

- A computational warning: If you stack the data, the difference across different individuals should be discarded.
- The FD estimator is less efficient than the FE estimator under the FE assumptions

Assumption FD.3: $Ee_i e_i' := (\Delta u_i \Delta u_i' | x_i, \alpha_i) = \sigma_e^2 I_{T-1}$

- Under assumptions FD.1–FD.3, $\widehat{\beta}_{FD}$ is the most efficient estimator.
- $\widehat{Avar}(\widehat{\beta}_{FD}) = \widehat{\sigma}_e^2 \left(\Delta X' \Delta X\right)^{-1}, \widehat{\sigma}_e^2 = 1/(NT - N - k) \sum_{i=1}^{N} \sum_{t=2}^{T} \widehat{e}_{it}^2$
- $\widehat{e}_{it} = \Delta y_{it} - \Delta x_{it}\widehat{\beta}_{FD}$
- Robust Variance Estimator:

$$\widehat{Avar}(\widehat{\beta}_{FD}) = \left(\sum_{i=1}^{N} \Delta X_i' \Delta X_i\right)^{-1} \left(\sum_{i=1}^{N} \Delta X_i' \widehat{e}_i \widehat{e}_i' \Delta X_i\right)^{-1} \left(\sum_{i=1}^{N} \Delta X_i' \Delta X_i\right)^{-1} \tag{2.70}$$

What if **Assumption FE.3**: $E(u_i u_i' | x_i, \alpha_i) = \sigma_u^2 I_T$ holds? In this case, $\widehat{\beta}_{FD}$ is less efficient than $\widehat{\beta}_{FE}$ because var$(\Delta u_{i1}, \Delta u_{i2}, ..., \Delta u_{iT})$ is not a diagonal matrix.

Let

$$D = \begin{pmatrix} -1 & 1 & & & 0 \\ & -1 & 1 & & \\ & & & ... & \\ 0 & & ... & -1 & 1 \end{pmatrix}_{(T-1)\times T}$$

then

$$\Delta X_i = DX_i, \Delta y_i = Dy, \Delta u_i = Du.$$

The variance of Δu_i is then given by

$$\sigma_u^2 DD'.$$

The GLS estimator based on the first-differenced model

$$Dy_i = DX_i + Du_i$$

is

$$\hat{\beta}_{FD,GLS} = \left\{ \sum_{i=1}^{N} X_i' D' \left(DD'\right)^{-1} DX_i \right\}^{-1} \left\{ \sum_{i=1}^{N} X_i' D' \left(DD'\right)^{-1} Dy_i \right\}.$$

Note that $D' \left(DD'\right)^{-1} D$ is a projection matrix projecting to the row space of D. Since $D\ell_T = 0$, ℓ_T is orthogonal to the row space of D. So projecting to the the row space of D is the same as projecting to the space orthogonal to ℓ_T. Therefore

$$D' \left(DD'\right)^{-1} D = I_T - \ell_T(\ell_T' \ell_T)^{-1} \ell_T'$$

As a consequence

$$\hat{\beta}_{FD,GLS} = \hat{\beta}_{FE}.$$

It is worth pointing out that the $\hat{\beta}_{FD,GLS}$ estimator is the OLS estimator based on the transformed model

$$\left(DD'\right)^{-1/2} Dy_i = \left(DD'\right)^{-1/2} DX_i + \left(DD'\right)^{-1/2} Du_i.$$

A natural question is: what $u_i^* = \left(DD'\right)^{-1/2} Du_i$ is for any vector $u_i = (u_{i,1}, ..., u_{i,T})'$? So algebra shows that

$$u_{it}^* = c_t \left[u_{it} - \frac{1}{(T-t)} \left(u_{i,t+1} + ... + u_{iT}\right) \right],$$

where

$$c_t^2 = \frac{T-t}{T-t+1}.$$

We refer to this transformation as *forward orthogonal transformation.* Thus, if $\text{var}(u_i) = \sigma_u^2 I_T$, then $\text{var}(u_i^*) = \sigma_u^2 I_{T-1}$. Therefore, the forward transformation can be regarded an alternative transformation, which in common with first-differencing eliminates the individual effects but in contrast does not introduce serial correlation in the transformed errors. *Forward transformation* turns out to be very useful in dynamic models.

2.5 Comparison: FE and FD Estimators

- They are identical with a balanced set with $T = 2$ for all individuals. In that case, the FD model and the FE model are numerically identical models, since $y_{i2} - y_{i1} = 2\left(y_{i2} - 1/2(y_{i1} + y_{i2})\right).$

- When $T > 2$, the choice between FD and FE hinges on the assumption about the u_{it}
- FD estimator and FE estimator will have different probability limit when the strict exogeneity assumption is violated.
- The correlation between u_{it} and x_{is} leads to inconsistent FD and FE estimators.

2.6 Comparison: RE and FE Estimators

Note that

$$\widehat{\beta}_{RE} = \left(X'V^{-1}X\right)^{-1} XV^{-1}y \tag{2.71}$$

and

$$V^{-1} = \sigma_1^{-2}P + \sigma_u^{-2}Q. \tag{2.72}$$

Therefore

$$\begin{aligned}
\widehat{\beta}_{RE} &= \left(\sigma_1^{-2}X'PX + \sigma_u^{-2}X'QX\right)^{-1}\left(\sigma_1^{-2}XPy + \sigma_u^{-2}XQy\right) \\
&= \left(\sigma_1^{-2}X'PX + \sigma_u^{-2}X'QX\right)^{-1}\sigma_1^{-2}X'PX\left(X'PX\right)^{-1}XPy \\
&\quad + \left(\sigma_1^{-2}X'PX + \sigma_u^{-2}X'QX\right)^{-1}\sigma_u^{-2}X'QX\left(X'QX\right)^{-1}XQy \\
&= W_1\widehat{\beta}_{between} + \left(I - W_1\right)\widehat{\beta}_{within}
\end{aligned} \tag{2.73}$$

- Recall $\sigma_1^2 = T\sigma_\alpha^2 + \sigma_u^2$. If $\sigma_\alpha^2 = 0$, then $\sigma_1^2 = \sigma_u^2$. So $\widehat{\beta}_{RE} = \left(X'X\right)^{-1}\left(X'y\right) = \widehat{\beta}_{POLS}$
- The pooled OLS estimator is a weighted average of the within and between estimators.

$$\begin{aligned}
\widehat{\beta}_{POLS} &= \left(X'X\right)^{-1}\left(X'y\right) = \left(X'\left[P+Q\right]X\right)^{-1}\left(X'\left[P+Q\right]y\right) \\
&= \left(X'PX + X'QX\right)^{-1}\left(X'Py + X'Qy\right) \\
&= \left\{\left(X'PX + X'QX\right)^{-1}X'PX\right\}\left(X'PX\right)^{-1}X'Py \\
&\quad + \left\{\left(X'PX + X'QX\right)^{-1}X'QX\right\}\left(X'QX\right)^{-1}X'Qy
\end{aligned} \tag{2.74}$$

- If $T \to \infty$, $\sigma_u/\sigma_1 \to 0$, then $\widehat{\beta}_{RE} \to \widehat{\beta}_{within} = \widehat{\beta}_{FE}$
- The larger σ_α^2 is, the close $\widehat{\beta}_{RE}$ is to $\widehat{\beta}_{FE}$. If $\sigma_\alpha^2 \to \infty$, $\widehat{\beta}_{RE} \to \widehat{\beta}_{within} = \widehat{\beta}_{FE}$.
- $Var(\widehat{\beta}_{RE}) = \left(\sigma_1^{-2}X'PX + \sigma_u^{-2}X'QX\right)^{-1}$ and $Var(\widehat{\beta}_{within}) = \left(\sigma_u^{-2}X'QX\right)^{-1}$. Hence, $Var(\widehat{\beta}_{RE}) \leq Var(\widehat{\beta}_{within})$.

Another Prospective

$$\begin{aligned} PY &= PX\beta + P\varepsilon \\ QY &= QX\beta + Q\varepsilon \end{aligned}$$

$$(\sigma_1^2 = T\sigma_\alpha^2 + \sigma_u^2)$$

Weight	(1,0)	(0,1)	(1/2,1/2)	$(\sigma_u/\sigma_1, 1)$
Estimate	Between	Within (fixed)	POLS	RE (GLS)
Cov$(\alpha_i, x_i) = 0$	unbiased	unbiased	unbiased	Efficient/unbiased
Cov$(\alpha_i, x_i) \neq 0$	biased	unbiased	biased	biased

2.7 Hausman-Wu Test

One might expect that the random effects estimator is superior to the fixed effects estimator. After all, it is the GLS estimator; moreover, the previous discussion shows that the fixed effects estimator is a limiting case of RE, corresponding to situations where the variation in the individual effects is large. Since the feasible version can actually estimate the variance of the individual effects, this would seem preferable to assuming it is arbitrarily large. However, there is a very strong assumption built in to the random effects estimator: the disturbances, including α_i, are orthogonal to the explanatory variables. In this section, we test the null $H_0 : \alpha_i$ and x_{it} are uncorrelated.

2.7.1 General Principle

Suppose we have two alternative estimators, $\widehat{\beta}_I$ and $\widehat{\beta}_{II}$, for a true parameter vector β. Further suppose that if a particular hypothesis H_0 is correct, both estimators are consistent and asymptotically normal with variance-covariance matrices V_I and V_{II}, and matrix of covariances between the two estimators $V_{I,II}$. Finally, suppose that if the null hypothesis is false the two estimators converge to different answers — for example, one of them might remain consistent while the other one becomes inconsistent, or both of them might become inconsistent but idiosyncratically so. Then the Wu-Hausman quadratic form:

$$m = (\widehat{\beta}_I - \widehat{\beta}_{II})'(V_I + V_{II} - V_{I,II} - V_{II,I})^{-1}(\widehat{\beta}_I - \widehat{\beta}_{II}) \tag{2.75}$$

under H_0 converges in distribution to a $\chi^2(k)$, where k is the number of elements in β. In the case that one of the estimators, say $\widehat{\beta}_I$, is efficient under H_0, it from the Rao-Blackwell theorem follows that $V_{I,II} = V_{II,I} = V_I$. Hence, the variance-covariance expression in the middle of m simplifies to $V_{II} - V_I$.

The intuition behind the Rao-Blackwell is as follows: Suppose we have two consistent estimators $\widehat{\beta}_I$ and $\widehat{\beta}_{II}$ and $\widehat{\beta}_I$ is an efficient estimator. Then the variance var$(a\widehat{\beta}_I + (1-a)\widehat{\beta}_{II})$ is smallest when $a = 1$. But the FOC for the minimization problem $\min_a$var$(a\widehat{\beta}_I + (1-a)\widehat{\beta}_{II})$ is

$$2var(\widehat{\beta}_I) - 2(1-a)var(\widehat{\beta}_{II}) + (2-4a)cov(\widehat{\beta}_I, \widehat{\beta}_{II}) = 0. \tag{2.76}$$

Letting $a = 1$ yields

$$var(\widehat{\beta}_I) = cov(\widehat{\beta}_I, \widehat{\beta}_{II}) \tag{2.77}$$

as desired.

2.7.2 The Hausman-Wu Specification Test

Applying this approach to the linear panel data problem, we can use the m statistic based the $\widehat{\beta}_{RE}$ and $\widehat{\beta}_{FE}$ test the null $H_0 : cov(\alpha_i, x_{it}) = 0$ for all t.

Under assumptions RE.1–RE.3 and H_0 :

- $\widehat{\beta}_{RE}$ is consistent, asymptotically normal and efficient. $\widehat{\beta}_{FE}$ is consistent and asymptotically normal.

In contrast, under assumptions RE.1–RE.3 and H_1 :

- $\widehat{\beta}_{RE}$ is inconsistent while $\widehat{\beta}_{FE}$ is consistent.

$$m = \left(\widehat{\beta}_{RE} - \widehat{\beta}_{FE}\right)' \left[Var(\widehat{\beta}_{RE} - \widehat{\beta}_{FE})\right]^{-1} \left(\widehat{\beta}_{RE} - \widehat{\beta}_{FE}\right)' \tag{2.78}$$

Note

$$\widehat{\beta}_{RE} - \widehat{\beta}_{FE} = \left(X'V^{-1}X\right)^{-1} X'V^{-1}\varepsilon - \left(X'QX\right)^{-1} X'Q\varepsilon \tag{2.79}$$

So, under $H_0 : \widehat{\beta}_{RE} - \widehat{\beta}_{FE} \simeq 0$ and

$$\begin{aligned} cov\left(\widehat{\beta}_{RE}, \widehat{\beta}_{FE}\right) &= E\left(X'V^{-1}X\right)^{-1} X'V^{-1}\varepsilon\,\varepsilon'QX\left(X'QX\right)^{-1} \\ &= E\left(X'V^{-1}X\right)^{-1} X'V^{-1}VQX\left(X'QX\right)^{-1} \\ &= E\left(X'V^{-1}X\right)^{-1} = var(\widehat{\beta}_{RE}) \end{aligned} \tag{2.80}$$

Thus

$$\begin{aligned} Var(\widehat{\beta}_{RE}-\widehat{\beta}_{FE}) &= Var\left(\widehat{\beta}_{FE}\right)-Var\left(\widehat{\beta}_{RE}\right) \\ &= \sigma_u^2\left(X'QX\right)^{-1}-\left(X'V^{-1}X\right)^{-1} \end{aligned} \tag{2.81}$$

which can be estimated by

$$\widehat{\sigma}_u^2\left(X'QX\right)^{-1}-\left(X'\widehat{V}^{-1}X\right)^{-1} \tag{2.82}$$

Alternatively,

$$\widehat{Var}\left(\widehat{\beta}_{RE}\right)=\widehat{\sigma}_u^2\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it}-\theta\bar{x}_{i,.}\right)'\left(x_{it}-\theta\bar{x}_{i,.}\right)\right)^{-1} \tag{2.83}$$

$$\widehat{Var}\left(\widehat{\beta}_{FE}\right)=\widehat{\sigma}_u^2\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it}-\bar{x}_{i,.}\right)'\left(x_{it}-\bar{x}_{i,.}\right)\right)^{-1} \tag{2.84}$$

To ensure the positive definiteness of $Var(\widehat{\beta}_{RE}-\widehat{\beta}_{FE})$, we need to use the same $\widehat{\sigma}_u^2$ in all the places.

2.7.3 Caveats

- Strict exogeneity is maintained under both H_0 and H_1
- RE.3 may not hold. Robust $\Rightarrow$Wald-test.

2.8 Panel Data: An Application

2.8.1 Macro Effects

Before presenting the application, we generalize the previous models to include time effects

$$y_{it}=x_{it}\beta+\alpha_i+\lambda_t+u_{it} \tag{2.85}$$

where λ_t denotes macro or time effects.

Usually, we treat λ_t as parameters to be estimated when T is small. Therefore all the approaches developed before can be applied after we introduce a number of time dummies $ds_{it}=\{t=s\}$. Then the model becomes

$$y_{it}=x_{it}\beta+\alpha_i+\sum_{s=1}^{T}\lambda_s ds_{it}+u_{it}. \tag{2.86}$$

To avoid the dummy variable trap, one dummy variable needs to be dropped from the regression.

2.8.2 Policy Evaluation: Difference in Differences

The evaluation problem arises because one is unable to observe the outcome variable for participants in a particular program had they not participated and the same goes the other way round for members of the control group. For example, we may have the following data:

Individual	X	Participate(Y/N)	Y(Y)	Y(N)	Causal Effect
1	X_1	1	$Y_1(\mathrm{Y})$	?	$Y_1(\mathrm{Y}) - Y_1(\mathrm{N})$
2	X_2	0	?	$Y_2(\mathrm{N})$	$Y_2(\mathrm{Y}) - Y_2(\mathrm{N})$
N	X_N	1	$Y_N(\mathrm{Y})$	?	$Y_N(\mathrm{Y}) - Y_N(\mathrm{N})$

Apparently, we can not measure the causal effect because one of the $Y_i(\mathrm{Y})$ and $Y_i(\mathrm{N})$ is unobservable.

In appropriately defined social experiments, the measurement problem can be overcome by randomly assigning individuals out of a particular group to the treatment. However, often, experimental data are not available and even when they are, side effects occur like people dropping out in a nonrandom way or a change in the behavior of the participants due to external factors or caused by the experiment itself. Other disadvantages of experiments are that they are difficult to extrapolate; they might be expensive to administer and the ethical approval might be doubtful — can you deny someone a promising new treatment which will likely cure him from a life threatening disease?

One way to measure the impact of a treatment in the setting of a natural experiment, is using the difference in difference (DID) estimator. To apply this estimator, longitudinal or repeated cross section data are needed, with at least one wave before and one wave after the program change. Let

$$y_{it} : \textit{outcome variable}$$

$$prog_{it} : \textit{program participation dummy variable},$$

$prog_{it} = 1$ if individual i participates, =0 otherwise

We consider the simple unobserved effect model:

$$y_{it} = b_0 + d_0 d2_t + b_1 prog_{it} + Z_i + u_{it} \tag{2.87}$$

where $d2_t$: time dummy that is 1 in period 2, 0 in period 1. Z_i: time invariant unobserved effect. Note that Z_i is likely to be correlated with $prog_{it}$.

If program participation only occurs in the second period, the OLS estimate of b_1 in the difference equation has a very simple representation.

$$y_{i2} = b_0 + d_0 + b_1 prog_{i2} + Z_i + u_{i2} \tag{2.88}$$

$$y_{i1} = b_0 + Z_i + u_{i1} \tag{2.89}$$

So

$$y_{i2} - y_{i1} = d_0 + b_1 prog_{i2} + u_{i2} - u_{i1} \tag{2.90}$$

Average over all the individuals that participate

$$\overline{\Delta y_{treat}} = d_0 + b_1 + \overline{\Delta u_{treat}} \tag{2.91}$$

Average over all the individuals that do not participate

$$\overline{\Delta y_{control}} = d_0 + \overline{\Delta u_{control}} \tag{2.92}$$

Therefore

$$\widehat{b}_1 = \overline{\Delta y_{treat}} - \overline{\Delta y_{control}} \tag{2.93}$$

What if there are some time varying variables so that

$$y_{it} = b_0 + d_0 \times d2_t + b_1 prog_{it} + Z_i + x_{it}\beta + u_{it} \tag{2.94}$$

Again, let's assume that program participation only occur in the second period, then

$$y_{i2} = b_0 + d_0 + b_1 prog_{i2} + Z_i + x_{i2}\beta + u_{i2} \tag{2.95}$$

$$y_{i1} = b_0 + Z_i + x_{i1}\beta + u_{i1} \tag{2.96}$$

So

$$\Delta y_i = d_0 + b_1 prog_{i2} + (\Delta x_i)\beta + \Delta u_i \tag{2.97}$$

Estimate β first by dummy variable regression and construct the adjusted difference $\Delta\widetilde{y_i} = \Delta y_i - (\Delta x_i)\widehat{\beta}$. Then

$$\widehat{b}_1 = \overline{\Delta\widetilde{y}_{treat}} - \overline{\Delta\widetilde{y}_{control}} \tag{2.98}$$

Example 7 *Consider a simple example to illustrate the basic philosophy behind the difference-in-differences approach. Suppose we are evaluating a program whose purpose is to increase employment. We have a group that participates in the program* *and a comparison group of non-participants. We also have data on the outcome measure for the participants and the comparison group in the time prior to and after the program. The data are summarized in the table below. The number in each cell is the employment rate for each group.*

	Before the program	*After the program*
Program Participants	*14.7%*	*17.6%*
Comparison Group	*16.7%*	*18.4%*

(2.99)

Let us consider different ways to evaluate this program based on the data presented in this table.

Method 1: Suppose that we look simply at the employment rate for participants after the program and compare that to the employment rate for the comparison group after the program. If we do this, we must conclude that the program actually reduces employment since 17.6% - 18.4% = -0.8%. Obviously this is a very unsatisfying result since, just by looking at the table, we can see that it neglects to take into account the fact that participants started off at a much lower level than the comparison group.

Method 2: Another approach to evaluating this program is to conduct a pre-post evaluation. That is, we can look at program participants before and after the program. By doing this, we see a very strong result of the program: 17.6% - 14.7% = 2.9%. Yet this answer is also open to criticism. By looking at the table, we see that the comparison group also improved between the before and after time periods. This leads us to wonder if there is some external force acting on everyone – both the comparison group and participants – that leads to higher employment rates. If that is the case, then some portion of the improvement for participants may be due to this external force rather than the program itself. For example, if the overall employment rate has been rising, then both participants and the comparison group members would see an increase in employment.

Difference-in-Difference: An alternative approach to these two evaluation methods that takes into account all of the information in the table above is the "difference-in-differences" approach. First, compute the difference in employment for participants before and after the program: 17.6% - 14.7% = 2.9%. Second, compute the difference in employment for the comparison group before and after the program: 18.4% - 16.7% = 1.7%. Now, compute the difference between these two differences: 2.9% - 1.7% = 1.2%. By subtracting off the 1.7%, we are removing the increased employment that would have occurred anyway (the benefit of an improving economy, for example) leaving us with an estimate of the increased employment due just to the program itself.

The textbook by Stock and Watson (2002) provides a very nice discussion on Program Evaluation. Although it is a textbook for undergraduate econometrics, graduate students would also benefit from some chapters such as Ch 11: Experiments and Quasi Experiments.

2.9 Chamberlain's Approach

Consider the simple linear panel data model

$$\begin{aligned} y_{it} &= x_{it}\beta + \varepsilon_{it}, i = 1, ..., N,\ t = 1, ...T_i \\ \varepsilon_{it} &= \alpha_i + u_{it} \end{aligned} \tag{2.100}$$

where α_i may be correlated with $x_{it.}$ Chamberlain's approach (1982, 1984) is to replace α_i with its linear projection onto $\{x_{it}\}$ Assume that α_i and $\{x_{it}\}$ have finite second moments. The projection can always be written as

$$\alpha_i = \psi + x_{i1}\lambda_1 + ... + x_{iT}\lambda_T + v_i. \tag{2.101}$$

Plugging this into the original model, we have

$$y_{it} = \psi + x_{i1}\lambda_1 + ... + x_{it}(\beta + \lambda_t) + ... + x_{iT}\lambda_T + r_{it} \tag{2.102}$$

2.9.1 System OLS Estimator

By definition, $E(r_{it}) = 0$ and $Ex_i' r_{it} = 0$ for all t.The above model can be written as

$$\begin{pmatrix} y_{i1} \\ y_{i2} \\ \cdots \\ y_{iT} \end{pmatrix} = \begin{pmatrix} 1 & x_{i1} & x_{i2} & \cdots & x_{iT} & x_{i1} \\ 1 & x_{i1} & x_{i2} & \cdots & x_{iT} & x_{i2} \\ \cdots & \cdots & & & & \\ 1 & x_{i1} & x_{i2} & & x_{iT} & x_{iT} \end{pmatrix} \begin{pmatrix} \psi \\ \lambda_1 \\ \lambda_2 \\ \\ \lambda_T \\ \beta \end{pmatrix} + \begin{pmatrix} r_{i1} \\ r_{i2} \\ \cdots \\ r_{iT} \end{pmatrix} \tag{2.103}$$

or

$$y_i = W_i\theta + r_i \tag{2.104}$$

Since $EW_i' r_i = 0$, system OLS is a way to consistently estimate ξ. The rank condition requires $EW_i' W_i = Tk + k + 1$.

$$\widehat{\theta}_{OLS} = \left(\sum W_i' W_i\right)^{-1} \sum W_i' y_i \tag{2.105}$$

or

$$\widehat{\theta}_{GLS} = \left(\sum W_i'\widehat{\Omega}_r^{-1}W_i\right)^{-1}\sum W_i'\widehat{\Omega}y_i \tag{2.106}$$

where $\widehat{\Omega}_r^{-1} = (1/N\sum \widetilde{r}_i\widetilde{r}_i')^{-1}$ where $\widetilde{r}_i = y_i - W_i\widehat{\theta}_{OLS}$

2.9.2 Minimum Distance Estimator

Chamberlain uses a different approach, known as minimum distance estimator. We start with a brief treatment of classical minimum distance estimation.

Suppose that $S \times 1$ vector of interest θ_0 is known to be related to an $P \times 1$ vector π_0, where $P > S$. In particular, $\pi_0 = h(\theta_0)$ for a known smooth function $h : R^S \rightarrow R^P$. For example,

$$\begin{aligned}
\pi_{0,1} &= \theta_{0,1} + \theta_{0,2} \\
\pi_{0,2} &= \log(\theta_{0,1})\exp(\theta_{0,2}^2) \\
\pi_{0,3} &= \sqrt{\theta_{0,1}} + 2\theta_{0,2}.
\end{aligned}$$

MD estimation of θ_0 entails first estimating π_0 by $\hat{\pi}$, and then choosing an estimate of $\hat{\theta}$ by making the distance between $\hat{\pi}$ and $h(\hat{\theta})$ as small as possible. Assuming that for a $S \times S$ positive definite matrix Ξ_0

$$\sqrt{N}\left(\hat{\pi} - \pi_0\right) \Rightarrow N(0, \Xi_0) \tag{2.107}$$

then an efficient MD estimator solves

$$\min_{\theta} Q_n\left(\theta\right) = \left(\hat{\pi} - h(\theta)\right)'\hat{\Xi}^{-1}\left(\hat{\pi} - h(\theta)\right) \tag{2.108}$$

where plim$\hat{\Xi} = \Xi_0$.

How to show that the MDE is consistent? It suffices to verify the following two conditions:

(i) $Q_n\left(\theta\right) \rightarrow Q\left(\theta\right) = \left(\pi_0 - h(\theta)\right)'\Xi_0^{-1}\left(\pi_0 - h(\theta)\right)$ uniformly.

(ii) identification: there exists a unique $\theta_0 \in \Theta$ such that $\pi_0 = h(\theta_0)$.

Given the consistency of $\hat{\theta}$, we can easily derive the asymptotic distribution of $\sqrt{N}\left(\hat{\theta} - \theta_0\right)$. The FOC is

$$H(\hat{\theta})'\hat{\Xi}^{-1}\left(\hat{\pi} - h(\hat{\theta})\right) = 0 \tag{2.109}$$

where $H\left(\theta\right) = \nabla_\theta h(\theta)$ is the $P \times S$ Jacobian of $h(\theta)$.

$$H\left(\theta\right) = \begin{pmatrix} \frac{\partial h_1(\theta)}{\partial \theta_1} & \frac{\partial h_1(\theta)}{\partial \theta_2} & \cdots & \frac{\partial h_1(\theta)}{\partial \theta_S} \\ \frac{\partial h_2(\theta)}{\partial \theta_1} & \frac{\partial h_2(\theta)}{\partial \theta_2} & \cdots & \frac{\partial h_2(\theta)}{\partial \theta_S} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial h_P(\theta)}{\partial \theta_1} & \frac{\partial h_P(\theta)}{\partial \theta_2} & \cdots & \frac{\partial h_P(\theta)}{\partial \theta_S} \end{pmatrix}_{P\times S} \tag{2.110}$$

Since $h(\theta_0) = \pi_0$, and

$$\sqrt{N}\left(h(\hat{\theta}) - h(\theta_0)\right) = H(\theta_0)\sqrt{N}(\widehat{\theta} - \theta_0) + o_p(1) \tag{2.111}$$

we have

$$\begin{aligned} 0 &= H(\hat{\theta})'\hat{\Xi}^{-1}\left(\sqrt{N}(\hat{\pi} - \pi_0) - \sqrt{N}\left(h(\widehat{\theta}) - h(\theta_0)\right)\right) \\ &= H(\hat{\theta})'\hat{\Xi}^{-1}\left(\sqrt{N}(\hat{\pi} - \pi_0) - H(\theta_0)\sqrt{N}(\widehat{\theta} - \theta_0)\right) + o_p\left(1\right) \\ &= H(\theta_0)'\Xi_0^{-1}\left(\sqrt{N}(\hat{\pi} - \pi_0) - H(\theta_0)\sqrt{N}(\widehat{\theta} - \theta_0)\right) + o_p\left(1\right). \end{aligned} \tag{2.112}$$

This implies

$$\begin{aligned} H(\theta_0)'\Xi_0^{-1}H(\theta_0)\sqrt{N}(\widehat{\theta} - \theta_0) &= H(\theta_0)'\Xi_0^{-1}\left(\sqrt{N}(\hat{\pi} - \pi_0)\right) + o_p\left(1\right) \\ &\Rightarrow N(0, H(\theta_0)'\Xi_0^{-1}H(\theta_0)) \end{aligned} \tag{2.113}$$

Hence,

$$\sqrt{N}(\widehat{\theta} - \theta_0) \Rightarrow N\left(0, \left(H(\theta_0)'\Xi_0^{-1}H(\theta_0)\right)^{-1}\right) \tag{2.114}$$

provided that $H_0 = H(\theta_0)$ has full column rank.

Under the null hypothesis, $N\left(\hat{\pi} - h(\widehat{\theta})\right)'\hat{\Xi}^{-1}\left(\hat{\pi} - h(\widehat{\theta})\right) \Rightarrow \chi^2_{P-S}$. To show this, note that $\sqrt{N}\left(\hat{\pi} - h(\widehat{\theta})\right)$ is

$$\begin{aligned} &\sqrt{N}(\hat{\pi} - \pi_0) - H(\theta_0)\sqrt{N}(\widehat{\theta} - \theta_0) + o_p\left(1\right) \\ =\ &\left(I_P - H(\theta_0)\left[H(\theta_0)'\Xi_0^{-1}H(\theta_0)\right]^{-1}H(\theta_0)'\Xi_0^{-1}\right)\sqrt{N}(\hat{\pi} - \pi_0) + o_p\left(1\right) \end{aligned} \tag{2.115}$$

Therefore, up to $o_p\left(1\right)$,

$$\begin{aligned} \Xi_0^{-1/2}\sqrt{N}\left(\hat{\pi} - h(\hat{\theta})\right) &= \left(I_s - \Xi_0^{-1/2}H_0\left[H_0'\Xi_0^{-1}H_0\right]^{-1}H_0'\Xi_0^{-1/2}\right)\Xi_0^{-1/2}\sqrt{N}(\hat{\pi} - \pi_0) \\ &= M_0\mathcal{L} \end{aligned} \tag{2.116}$$

But M_0 is idempotent with rank $P-S$ ($\Xi_0^{-1/2}H_0$ is a $P \times S$ matrix), $\mathcal{L} \Rightarrow N(0,I)$. As a consequence

$$N\left(\hat{\pi}-h(\hat{\theta})\right)\hat{\Xi}^{-1}\left(\hat{\pi}-h(\hat{\theta})\right) \Rightarrow \chi^2_{P-S}. \tag{2.117}$$

Exercise 8 *Suppose A is symmetric and idempotent matrix and $u \sim N(0,I)$. Prove that $u'Au \sim \chi^2_k$ where $k = rank(A)$.*

Testing restriction on θ_0 is also straightforward. Suppose $\theta_0 = d(\eta_0)$ for some function d: $R^D \rightarrow R^S$ where $D < S$. then

$$\pi_0 = h(\theta_0) = h\left(d(\eta_0)\right) := g(\eta_0). \tag{2.118}$$

This η_0 can be estimate using the MD estimator, i.e.

$$\hat{\eta} = \arg\min\left(\hat{\pi}-g(\eta)\right)'\hat{\Xi}^{-1}\left(\hat{\pi}-g(\eta)\right) \tag{2.119}$$

Then it can be shown that

$$N\left(\hat{\pi}-g(\hat{\eta})\right)'\hat{\Xi}^{-1}\left(\hat{\pi}-g(\hat{\eta})\right) - N\left(\hat{\pi}-h(\hat{\theta})\right)'\hat{\Xi}^{-1}\left(\hat{\pi}-h(\hat{\theta})\right) \Rightarrow \chi^2_{S-D} \tag{2.120}$$

We now use the MD estimator to deal with Chamberlain's problem. The model is

$$\begin{aligned} y_{it} &= \psi + x_{i1}\lambda_1 + ... + x_{it}(\beta+\lambda_t) + ... + x_{iT}\lambda_T + r_{it} \\ &= \pi_{t0} + x_i\pi_t + r_{it} \\ &= \pi_{t0} + x_{i1}\pi_{t1} + x_{i2}\pi_{t2} + ... + x_{iT}\pi_{tT} + r_{it} \end{aligned} \tag{2.121}$$

where $\pi_{t0} = \psi$, the vector $\pi_t = (\lambda_1', \lambda_2', ..., (\beta+\lambda_t)', ..., \lambda_T')'$ is $kT \times 1$ and $\pi = (\pi_{10}, \pi_1, \pi_{20,}\pi_2, ..., \pi_{T0}, \pi_T)'$ is $(kT+1)\,T \times 1$.

We can write $\pi = H\theta$ for a $(kT+1)\,T \times (1+Tk+k)$ matrix H. For example, when $T = 2$

$$H = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & I_k & 0 & I_k \\ 0 & 0 & I_k & 0 \\ 1 & 0 & 0 & 0 \\ 0 & I_k & 0 & 0 \\ 0 & 0 & I_k & I_k \end{pmatrix}, \theta = \begin{pmatrix} \psi \\ \lambda_1 \\ \lambda_2 \\ \beta \end{pmatrix} \tag{2.122}$$

Written in matrix form, the model becomes

$$\begin{aligned} y_i &= \begin{pmatrix} 1, x_i & 0 & 0 & 0 \\ 0 & 1, x_i & 0 & 0 \\ 0 & 0 & ... & ... \\ 0 & 0 & ... & 1, x_i \end{pmatrix} \pi + r_i \\ &= \tilde{X}_i \pi + r_i \end{aligned}$$

where $X_i = I_T \otimes (1, x_i)$ is a $T \times \left(T^2 k + T\right)$ matrix.

To estimate π, we can simply run OLS based on the above model, i.e.

$$\hat{\pi} = \left(\sum_{i=1}^{N} \tilde{X}_i' \tilde{X}_i \right)^{-1} \left(\sum_{i=1}^{N} \tilde{X}_i' y_i \right)$$

The above OLS estimator is the same as the OLS estimator obtained period by period. (OLS and SUR are identical because the same regressors appear in each equation). The asymptotic variance of $\hat{\pi}$ can be estimated by

$$\left(\frac{1}{N} \sum_{i=1}^{N} \tilde{X}_i' \tilde{X}_i \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \tilde{X}_i' \widehat{v}_i \widehat{v}_i' \tilde{X}_i \right) \left(\frac{1}{N} \sum_{i=1}^{N} \tilde{X}_i' \tilde{X}_i \right)^{-1} \qquad (2.123)$$

where $\widehat{v}_i$ is vector of OLS residuals.

To test $\lambda_1 = \lambda_2 = ... = \lambda_T = 0$, we can use (2.120).

Example 9 *Chamberlain's approach with $T = 3$ periods and scaler x_{it}.*

In a simple regression setting $y = x_{it}\beta + \epsilon_{it}$ with

$$\epsilon_{it} = \alpha_i + u_{it}, u_{it} \sim iid(0, \sigma_u^2), u_{it} \perp\perp \alpha_i, \qquad (2.124)$$

we have:

$$\alpha_i = x_{i1}\lambda_1 + x_{i2}\lambda_2 + x_{i3}\lambda_3 + v_i \qquad (2.125)$$

Then:

$$\begin{aligned} y_{i1} &= x_{i1}\beta + x_{i1}\lambda_1 + x_{i2}\lambda_2 + x_{i3}\lambda_3 + v_i + u_{i1} \\ y_{i2} &= x_{i2}\beta + x_{i1}\lambda_1 + x_{i2}\lambda_2 + x_{i3}\lambda_3 + v_i + u_{i2} \\ y_{i3} &= x_{i3}\beta + x_{i1}\lambda_1 + x_{i2}\lambda_2 + x_{i3}\lambda_3 + v_i + u_{i3} \qquad (2.126) \end{aligned}$$

For a factor structure, we get:

$$\epsilon_{it} = \delta_t \alpha_i + u_{it}, u_{it} \sim iid(0, \sigma_u^2), u_{it} \perp\perp \alpha_i, \qquad (2.127)$$

Then:

$$\begin{aligned} y_{i1} &= x_{i1}\beta + (x_{i1}\lambda_1 + x_{i2}\lambda_2 + x_{i3}\lambda_3)\,\delta_1 + v_i + u_{i1} \\ y_{i2} &= x_{i2}\beta + (x_{i1}\lambda_1 + x_{i2}\lambda_2 + x_{i3}\lambda_3)\,\delta_2 + v_i + u_{i2} \\ y_{i3} &= x_{i3}\beta + (x_{i1}\lambda_1 + x_{i2}\lambda_2 + x_{i3}\lambda_3)\,\delta_3 + v_i + u_{i3} \end{aligned} \tag{2.128}$$

we can identify:

$$\begin{pmatrix} \beta + \lambda_1\delta_1 & \lambda_2\delta_1 & \lambda_3\delta_1 \\ \lambda_1\delta_2 & \beta + \lambda_2\delta_2 & \lambda_3\delta_2 \\ \lambda_1\delta_3 & \lambda_2\delta_3 & \beta + \lambda_3\delta_3 \end{pmatrix} \tag{2.129}$$

we can estimate the model by MDE.

2.10 Hausman and Taylor's Approach

2.10.1 The Idea

$$y_{it} = z_i\gamma + x_{it}\beta + \alpha_i + u_{it} \tag{2.130}$$

Assume that $Ez_i'\alpha_i = 0$. The between equation is

$$\bar{y}_i, . = z_i\gamma + \bar{x}_{i,.}\beta + \alpha_i + \bar{u}_i, . \tag{2.131}$$

premultiplying by z_i' gives

$$z_i'\bar{y}_i, . = z_i'z_i\gamma + z_i'\bar{x}_{i,.}\beta + z_i'\alpha_i + z_i'\bar{u}_i, . \tag{2.132}$$

Take expectation on both sides to obtain

$$Ez_i'\bar{y}_i, . = Ez_i'z_i\gamma + Ez_i'\bar{x}_{i,.}\beta \tag{2.133}$$

It follows by usual analogy principle that

$$\widehat{\gamma} = \left(\frac{1}{N}\sum_{i=1}^{N} z_i'z_i\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} z_i'\left(\bar{y}_i, . - \bar{x}_{i,.}\widehat{\beta}_{FE}\right). \tag{2.134}$$

Alternatively,

$$z_i\gamma + \alpha_i + \bar{u}_i = \bar{y}_i, . - \bar{x}_{i,.}\beta. \tag{2.135}$$

But $Ez_i'\alpha_i = 0$, so we can omit the regressor α_i and estimate γ consistently using the following OLS regression

$$\bar{y}_i, . - \bar{x}_{i,.}\widehat{\beta}_{FE} = z_i\gamma + \bar{u}_i + error \tag{2.136}$$

This yields exactly the same estimator as above.

2.10.2 A General Approach

Hausman and Taylor (1981) partitioned z_i and x_{it} as $z_i = (z_{1,i}, z_{2,i}), x_{it} = (x_{1,it}, x_{2,it})$ where z_{1i} is $1 \times j_1$ and z_{2i} is $1 \times j_2$, $x_{1,it}$ is $1 \times k_1$ and $x_{2,it}$ is $1 \times k_2$. HT assume that

$$Ez_{1i}'\alpha_i = 0 \text{ and } \ Ex_{1,it}'\alpha_i = 0 \tag{2.137}$$

we still maintain the assumption that z_i and x_{it} are uncorrelated with u_{is} for all t and s.

We proceed to obtain initial estimates of β, $\sigma_u^2, \sigma_\alpha^2$ and γ. The initial estimate of β is the within estimate $\widehat{\beta}_{within}$. With $\widehat{\beta}_{within}$, we can construct the within residual

$$u_{it}^{within} = y_{it} - \bar{y}_i - (x_{it} - \bar{x}_{i,.})\,\widehat{\beta}_{within} \tag{2.138}$$

and estimate the variance of u_{it} by

$$\widehat{\sigma}_u^2 = \frac{\sum_{i=1}^N \sum_{t=1}^T \left(u_{it}^{within}\right)^2}{N(T-1)-K} \tag{2.139}$$

where $K = k_1 + k_2$ is the number of regressor in the within regression.

To obtain initial estimates of γ and σ_α^2, we consider the following cross sectional regression:

$$d_i = z_i\gamma + c + \tilde{u}_i \tag{2.140}$$

where

$$d_i = \bar{y}_{i,.} - \bar{x}_{i,.}\widehat{\beta}_{within} \text{ and } c \text{ is the intercept.} \tag{2.141}$$

Note that z_{1i} is uncorrelated with α_i and u_{it} but z_{2i} may be correlated with α_i. To estimate γ, we have to find instruments for z_{2i}. By assumption, $(\bar{x}_{1,i.}, z_{1i})$ are valid instruments. Using these instruments, we can obtain a consistent IV estimate of γ, denoted as $\widehat{\gamma}_w$. Let

$$s^2 = \frac{\sum_{i=1}^N (d_i - z_i\widehat{\gamma}_w)^2}{N} \tag{2.142}$$

Then the variance σ_α^2 of α_i can be estimated by

$$\widehat{\sigma}_\alpha^2 = s^2 - \widehat{\sigma}_u^2/T. \tag{2.143}$$

With the initial estimates of $\widehat{\sigma}_u^2$ and $\widehat{\sigma}_\alpha^2$, we can make the GLS transformation:

$$\begin{aligned} y_{it} - \theta\bar{y}_i &= (x_{it} - \theta\bar{x}_i)\beta + (1-\theta)z_i\gamma + \varepsilon_{it} - \theta\bar{\varepsilon}_{i,.} \\ \varepsilon_{it} &= \alpha_i + u_{it} \end{aligned} \tag{2.144}$$

where

$$\theta = 1 - \sqrt{\frac{\sigma_u^2}{T\sigma_\alpha^2 + \sigma_u^2}} \tag{2.145}$$

If all regressors are uncorrelated with α_i and u_{it}, then we can get the GLS estimate by OLS regression based on (2.144). However, by assumption, some of the regressors are endogenous. Write the model in (2.144) for all time periods as

$$y - \theta\bar{y}. = (1-\theta)Z\gamma + (X - \theta\bar{X}_.)\beta + \varepsilon - \theta\bar{\varepsilon}_. \tag{2.146}$$

where

$$Z = (Z_1, Z_2),\ \ X = (X_1, X_2) \tag{2.147}$$

We now proceed to find instruments for the model in (2.146). Since x_{it} is uncorrelated with u_{is}, it follows that

$$E\underset{NT\times k}{(QX)'}\underset{NT\times 1}{(\varepsilon - \theta\bar{\varepsilon}_.)} = E\,(QX)'\,Q\varepsilon = 0. \tag{2.148}$$

Therefore QX_1 and QX_2 can be used as instruments. In addition, since $Ez_{i1}'\alpha_i = 0$ and $Ex_{it1}'\alpha_i = 0$, we have

$$E\underset{NT\times j_1}{Z_1'}\underset{NT\times 1}{(\varepsilon - \theta\bar{\varepsilon}_.)} \ = \ 0, \tag{2.149}$$

$$E\underset{NT\times k_1}{(PX_1)'}\underset{NT\times 1}{(\varepsilon - \theta\bar{\varepsilon}_.)} \ = \ 0, \tag{2.150}$$

Therefore PX_1 and Z_1 can also be used as instruments. Collecting QX_1, QX_2, PX_1 and Z_1, we obtain the instrument list suggested by Hausman and Taylor (1981):

$$A_{HT} = (QX_1, QX_2, PX_1, Z_1) \tag{2.151}$$

Amemiya and MaCurdy (1986) recommended additional instruments by observing that (2.150) uses only the fact the means of X_1 are uncorrelated with α and u, *i.e.*

$$E\left(\frac{1}{T}\sum_{t=1}^{T} x_{1,it}'\right)\alpha_i = 0 \text{ and } E\left(\frac{1}{T}\sum_{t=1}^{T} x_{1,it}'\right)u_{is} = 0 \tag{2.152}$$

But the assumption (2.137) contains more information, as it implies

$$E\left(x_{1,it} - \bar{x}_{1,i.}\right)'\alpha_i = 0 \text{ and } E\left(x_{1,it} - \bar{x}_{1,i.}\right)'u_{is} = 0 \tag{2.153}$$

for all t and s. Therefore

$$(QX_1)^o = \begin{pmatrix} X_{1,11} - \bar{X}_{1,1.} & X_{1,12} - \bar{X}_{1,1.} & X_{1,13} - \bar{X}_{1,1.} & ... & X_{1,1T} - \bar{X}_{1,1.} \\ ... & ... & ... & ... & ... \\ X_{1,i1} - \bar{X}_{1,i.} & X_{1,i2} - \bar{X}_{1,i.} & X_{1,i3} - \bar{X}_{1,i.} & ... & X_{1,iT} - \bar{X}_{1,i.} \\ ... & ... & ... & ... & ... \\ X_{1,N1} - \bar{X}_{1,N.} & X_{1,N2} - \bar{X}_{1,N.} & X_{1,N3} - \bar{X}_{1,N.} & ... & X_{1,NT} - \bar{X}_{1,N.} \end{pmatrix} \otimes l_T \tag{2.154}$$

are also valid instruments. The instrument list suggested by Amemiya and MaCurdy (1986) is

$$A_{AM} = (QX_1, QX_2, PX_1, (QX_1)^o, Z_1) \tag{2.155}$$

Bruesch, Mizon and Schmidt (1987) extend the AM treatment of the QX_1 variables to the QX_2 variables and obtain the following instrument set

$$A_{BMS} = (QX_1, QX_2, PX_1, (QX_1)^o, (QX_2)^o, Z_1) \tag{2.156}$$

In general, A_{BMS} delivers estimators that are more efficient than A_{AM}, which in turn delivers more efficient estimators than A_{HT}. Note that the potential efficiency gain from the BMS procedure depends on whether the $(QX_2)^o$ are legitimate instruments. The $(QX_2)^o$ are valid instruments if the variables in X_2 are correlated with the individual effect only through a time invariant component. If these were true, QX_2 would not contain this component, and using the deviations separately for each period would be legitimate.

2.11 Maximum Likelihood Panel Data Estimators

Considering the panel data models from a more general viewpoint, we can form different maximum likelihood estimators of the parameters of interest. A word on notation. In this section, the function f is a generic density function, which may change from one equality to another equality.

Assume

$$\epsilon_{it} = \alpha_i + U_{it} \tag{2.157}$$

Write: $Z_i = (y_{i1}, ..., y_{iT}, x_{i1}, ..., x_{iT}), i = 1, ..., N$ Then Z_i is an iid random vector with distribution depending on:

$$\theta = (\beta, \alpha_1, ..., \alpha_i, ..., \alpha_N) = (\beta, \alpha) \tag{2.158}$$

(Note that we treat α_i's as parameters here.) We then have the likelihood function:

$$\pounds(\theta) = \prod_{i=1}^{N} f(Z_i|\theta) = \prod_{i=1}^{N} f(Z_i|\beta, \alpha) \tag{2.159}$$

Then $\theta = argmax\pounds(\theta)$ yields the $\widehat{\theta}_{ML}$, which generally possesses all the attractive features of the ML estimator. However, in this case we do not have: $\widehat{\alpha}_{ML} \rightarrow \alpha_i$ as $N \rightarrow \infty$ for fixed T. This is because, in most cases T is fixed and small, so that each $\widehat{\alpha}_{ML}$ is based on only a small number of observations. In general, $\widehat{\beta}_{ML} \rightarrow \tilde{\beta} \neq \beta$ as $N \rightarrow \infty$ because of this; unlike in linear models, in general, roots of these equations interconnected so that inconsistency in the estimation of one parameter affects other estimates. The joint system of equations:

$$\frac{\partial \pounds(\theta)}{\partial \beta} = 0, \ \frac{\partial \pounds(\theta)}{\partial \alpha_i} = 0, \ i = 1, 2, ..., N \tag{2.160}$$

These set of likelihood equations can be solved using three distinct concepts:
(1) Marginal Likelihood;
(2) Conditional Likelihood; and
(3) Integrated Likelihood.
These concepts are discussed in the following subsections.

2.11.1 Marginal Likelihood (or Ancillary Likelihood)

Find (if possible) $g(y, x)$ independent of α_i's, i.e. find some statistic $S_i = S(y_i, x_i)$ such that the marginal likelihood of S_i is independent of the α_i's. Then:

$$f(S_i|\beta, \alpha) = f(S_i|\beta) \tag{2.161}$$

The corresponding likelihood function is :

$$\pounds_M(\beta) = \prod_{i=1}^{N} f(S_i|\beta, \alpha) = \prod_{i=1}^{N} f(S_i|\beta) \tag{2.162}$$

Then we can form the ML estimators for β (the parameters of interest) without worrying about the α_i's.

$$\widehat{\beta} = \arg\max \pounds_M(\beta) \tag{2.163}$$

We say that S_i is "ancillary for α given β with respect to original model". An example of this is the within estimator.

$$y_{it} = x_{it}\beta + \alpha_i + u_{it}, u_{it} \sim iidN(0, \sigma_u^2) \tag{2.164}$$

Consider for example with $T = 2$, the statistic: $S_i = y_{i2} - y_{i1}$. S_i is called an ancillary statistic, whose distribution is independent is α_i

$$S_i|x \sim N((x_{i2} - x_{i1})\beta, 2\sigma_u^2). \tag{2.165}$$

Thus an example of the marginal likelihood estimator is the first difference estimator, which is almost identical to the "Within" estimator. Here, the within estimate would also be a marginal likelihood estimator.

2.11.2 Conditional Likelihood

Under this method, we find s, a sufficient statistic for α_i such that $f(y_i|s_i)$ is independent of α_i. In other words, find $s_i = s(y_i, x_i)$ so that:

$$f(Z_i|\beta, \alpha_i, s_i) = f(Z_i|\beta, s_i) \tag{2.166}$$

By cleverly choosing s_i, we can later throw away s_i. In the panel data case we have: e.g. $s_i = y_{i1} + y_{i2}$. We have:

$$y_{i1} + y_{i2} \sim N(\alpha_i + (x_{i1} + x_{i2})\beta, 2\sigma_u^2 + 4\sigma_\alpha^2). \tag{2.167}$$

Transform observations:

$$\begin{pmatrix} y_{i1} \\ y_{i2} \end{pmatrix} \rightarrow \begin{pmatrix} y_{i2} - y_{i1} \\ y_{i2} + y_{i1} \end{pmatrix} \tag{2.168}$$

Note that

$$Cov(y_{i2} - y_{i1}, y_{i2} + y_{i1}|X_i) = Cov(u_{i2} - u_{i1}, u_{i2} + u_{i1} + \alpha_i) = 0. \tag{2.169}$$

$$\Rightarrow f(y_{i1}, y_{i2}) = \mathfrak{f}(y_{i2} - y_{i1}, y_{i1} + y_{i2}) = f(y_{i2} - y_{i1}|X_i)f(y_{i1} + y_{i2}|X_i) \tag{2.170}$$

But since $s_i = y_{i2} + y_{i1}$, we have:

$$f(y_{i2} - y_{i1}, y_{i2} + y_{i1}|X_i, s_i) = f(y_{i2} - y_{i1}|X_i) \tag{2.171}$$

Thus here we get that the conditional likelihood function is the same as in previous case; i.e. the first difference estimator or the "Within estimator" is also an example of a conditional maximum likelihood estimator.

2.11.3 Integrated L.F. or Random Effects Estimator

In this method, we pick a density for α_i (note that the other methods did not require this):

$$pdf\ of\ \alpha_i \equiv g(\alpha_i|x). \tag{2.172}$$

For each person then:

$$g(y_i|X_i,\beta) = \int g(y_i|X_i,\beta,\alpha)g(\alpha|X_i)d\alpha \tag{2.173}$$

We get the corresponding integrated likelihood function independent of the α_i's:

$$\pounds_I(\beta) = \prod_{i=1}^{N} g(y_i|X_i,\beta) \tag{2.174}$$

e.g.

$$\alpha_i = Z_i\lambda + a_i, a_i \sim N(0,\sigma_a^2) \tag{2.175}$$

Mundlak's Point: Mundlak pointed out that if: $Z_i = \bar{X}_i$. Then we get in a regression setting, that:

$$\widehat{\beta}_{Marginal} = \widehat{\beta}_{Conditional} = \widehat{\beta}_{Integrated} = \widehat{\beta}_{Within} = \widehat{\beta}_{MLE} \tag{2.176}$$

The intuition for this can be illustrated as follows. Suppose that $\alpha_i = \bar{X}_i\alpha + a_i$ Then:

$$y_i = X_i\beta + \bar{X}_i\alpha + \iota a_i + u_i. \tag{2.177}$$

Now what is the random effect estimator? Note that we can write:

$$y_i = (X_i - \bar{X}_i)\beta + \bar{X}_i(\alpha+\beta) + \iota a_i + u_i \tag{2.178}$$

Thus intuitively, we see that you get info only on β from within. To show this mathematically, write:

$$F = I - \theta\iota\iota'/T \tag{2.179}$$

Then we get the GLS transformation as:

$$Fy_i = F(X_i - \bar{X}_i)\beta + F\bar{X}_i(\varphi+\beta) + F\left(\iota a_i + u_i\right) \tag{2.180}$$

so that the GLS estimator here is:

$$\widehat{\beta}_{GLS} = \left(\sum_{i=1}^{N}(X_i - \bar{X}_i)'F'F(X_i - \bar{X}_i)\right)^{-1}\left(\sum_{i=1}^{N}(X_i - \bar{X}_i)'F'F(y_i)\right) \tag{2.181}$$

Now:

$$F(X_i - \bar{X}_i) = \left(I - \theta\iota\iota'/T\right)(X_i - \bar{X}_i) = X_i - \bar{X}_i \tag{2.182}$$

So

$$\widehat{\beta}_{GLS} = \left(\sum_{i=1}^{N}(X_i - \bar{X}_i)'(X_i - \bar{X}_i)\right)^{-1}\left(\sum_{i=1}^{N}(X_i - \bar{X}_i)'y_i)\right) = \beta_{within} \tag{2.183}$$

Thus Mundlak's point was that the within estimator is the GLS estimator always if:

$$\alpha_i = \bar{X}_i\alpha + a_i. \tag{2.184}$$

2.12 Appendix: Static Panel Data Models in Stata

2.12.1 Introduction

There are two kinds of information in cross-sectional time-series data: the cross-sectional information reflected in the differences between subjects, and the time-series or within-subject information reflected in the changes within subjects over time. Panel data regression techniques allow you to take advantage of these different types of information.

While it is possible to use ordinary multiple regression techniques on panel data, they may not be optimal. The estimates of coefficients derived from regression may be subject to omitted variable bias - a problem that arises when there is some unknown variable or variables that cannot be controlled for that affect the dependent variable. With panel data, it is possible to control for some types of omitted variables even without observing them, by observing changes in the dependent variable over time. This controls for omitted variables that differ between cases but are constant over time. It is also possible to use panel data to control for omitted variables that vary over time but are constant between cases.

2.12.2 Using Panel Data in Stata

A panel dataset should have data on N cases, over T time periods, for a total of $N \times T$ observations. Data like this is said to be in long form. In some cases your data may come in what is called the wide form, with only one observation per case and variables for each different value at each different time period. To analyze data like this in Stata using commands for panel data analysis, you need to first convert it to long form. This can be done using Stata's reshape command. For assistance in using reshape, see Stata's online help.

Stata provides a number of tools for analyzing panel data. The commands all begin with the prefix *xt* and include *xtreg*, *xtprobit*, *xtsum* and *xttab*- panel data versions of the familiar *reg*, *probit*, *sum* and *tab* commands.

To use these commands, first tell Stata that your dataset is panel data. You need to have a variable that identifies the case element of your panel (for example, a country or person identifier) and also a time variable that is in Stata date format.

Sort your data by the panel variable and then by the date variable within the panel variable. Then you need to issue the *tsset* command to identify the panel and date variables. If your panel variable is called panelvar and your date variable is called datevar, the commands needed are:

```
. sort panelvar datevar
. tsset panelvar datevar
```

If you prefer to use menus, use the command under Statistics > Time Series > Setup and Utilities > Declare Data to be Time Series.

2.12.3 Fixed Effects

Fixed effects regression is the model to use when you want to control for omitted variables that differ between cases but are constant over time. It lets you use the changes in the variables over time to estimate the effects of the independent variables on your dependent variable, and is the main technique used for analysis of panel data.

The command for a linear regression on panel data with fixed effects in Stata is *xtreg* with the fe option, used like this:

```
xtreg dependentvar independentvar1 independentvar2 independentvar3 ... ,
fe
```

If you prefer to use the menus, the command is under Statistics > Cross-sectional time series > Linear models > Linear regression.

This is equivalent to generating dummy variables for each of your cases and including them in a standard linear regression to control for these fixed "case effects". It works best when you have relatively fewer cases and more time periods, as each dummy variable removes one degree of freedom from your model.

2.12.4 Between Effects

Regression with between effects is the model to use when you want to control for omitted variables that change over time but are constant between cases. It allows you to use the variation between cases to estimate the effect of the omitted independent variables on your dependent variable.

The command for a linear regression on panel data with between effects in Stata is *xtreg* with the *be* option.

Running *xtreg* with between effects is equivalent to taking the mean of each variable for each case across time and running a regression on the collapsed dataset of means. As this results in loss of information, between effects are not used much in practice. Researchers who want to look at time effects without considering panel effects generally will use a set of time dummy variables, which is the same as running time fixed effects.

The between effects estimator is mostly important because it is used to produce the random effects estimator.

2.12.5 Random Effects

If you have reason to believe that the unobserved omitted variables are uncorrelated with the regressors, then you can use a random-effects model. Stata's random-effects estimator is a weighted average of fixed and between effects.

The command for a linear regression on panel data with random effects in Stata is *xtreg* with the *re* option.

2.12.6 Choosing Between Fixed and Random Effects

The generally accepted way of choosing between fixed and random effects is running a Hausman test.

Statistically, fixed effects are always a reasonable thing to do with panel data (they always give consistent results) but they may not be the most efficient model to run. Random effects will give you better P-values as they are a more efficient estimator, so you should run random effects if it is statistically justifiable to do so.

The Hausman test checks a more efficient model against a less efficient but consistent model to make sure that the more efficient model also gives consistent results.

To run a Hausman test comparing fixed with random effects in Stata, you need to first estimate the fixed effects model, save the coefficients so that you can compare them with the results of the next model, estimate the random effects model, and then do the comparison.

```
. xtreg dependentvar independentvar1 independentvar2 independentvar3 ... , fe
. estimates store fixed
. xtreg dependentvar independentvar1 independentvar2 independentvar3 ... , re
. estimates store random
. hausman fixed random
```

The hausman test tests the null hypothesis that the coefficients estimated by the efficient random effects estimator are the same as the ones estimated by the consistent fixed effects estimator. If they are insignificant P-value, then it is safe to use random effects. If you get a significant P-value, however, you should use fixed effects.

Bibliography

[1] Amemiya, Takeshi, and Thomas E. Macurdy (1986). "Instrumental Variable Estimation of an Error-Components Model." *Econometrica* 54: 869-881.

[2] Breusch, Trevor S., Grayham E. Mizon, and Peter Schmidt (1986). "Efficient Estimation Using Panel Data." *Econometrica* 57: 695-700.

[3] Chamberlain, Gary (1982). "Multivariate Regression Models for Panel Data." *Journal of Econometrics* 18: 5-45.

[4] Chamberlain, Gary (1984). "Panel Data," in Griliches and Intriligator (eds.), *Handbook of Econometrics*, Volume 2. Amsterdam: North-Holland.

[5] Hausman, J. (1978). "Specification Tests in Econometrics," *Econometrica* 46:1251-1272

[6] Hausman, Jerry A., and William E. Taylor (1981). "Panel Data and Unobservable Individual Effects." *Econometrica* 49: 1377-1398.

[7] Im, K.S. , S. C. Ahn, P. Schmidt and J. Wooldridge (1999). " Efficient Estimation of Panel Data Models with Strictly Exogenous Explanatory Variables," *Journal of Econometrics*, 93: 177-201

[8] Maddala, G.S. (1971). "The Use of Variance Component Models in Pooling Cross Section and Time Series Data," *Econometrica* 39: 341-358.

[9] Mundlak, Y. (1978). "On the Pooling of Time Series and Cross Section Data", *Econometrica,* Vol. 46, pp. 69-85.

[10] Stock, J. and M. Watson (2002). *Introduction to Econometrics*, Addison and Wesley.

Chapter 3

Dynamic Panel Data Models

3.1 Models with Sequentially Exogenous Variables

Consider the linear panel data model below:

$$\begin{aligned} y_{it} &= x_{it}\beta + \varepsilon_{it}, i = 1, ..., N,\ t = 1, ...T_i \\ \varepsilon_{it} &= \alpha_i + u_{it} \end{aligned} \tag{3.1}$$

In addition to allowing α_i and x_{it} to be arbitrarily correlated, we now allow u_{it} to be correlated with the future value of x_{it}, i.e $(x_{i,t+1}, x_{i,t+2,}..., x_{i,T})$.

In a dynamic panel data model where $x_{it} = y_{i,t-1,}$ u_{it} is obviously correlated with $y_{i,t}, y_{i,t+1}, ..., y_{i,T}$. Because we have, using

$$y_{it} = y_{i,t-1}\beta + \varepsilon_{it}, \tag{3.2}$$

$$y_{it} = \sum_{s=0}^{t-1} \beta^s \varepsilon_{i,t-s} + \beta^t y_{i0} = \frac{1-\beta^t}{1-\beta}\alpha_i + \sum_{s=0}^{t-1} \beta^s u_{i,t-s} + \beta^t y_{i0}. \tag{3.3}$$

Following Chamberlain, we introduce **sequential moment restrictions:**

$$E(u_{it}|x_{i,t}, x_{i,t-1}, ..., x_{i,1}, \alpha_i) = 0, \tag{3.4}$$

which implies

$$E(y_{it}|x_{i,t}, x_{i,t-1}, ..., x_{i,1}, \alpha_i) = x_{i,t}\beta + \alpha_i. \tag{3.5}$$

Example 10 *Suppose*

$$y_{it} = \rho_1 y_{i,t-1} + z_{it}\beta + \varepsilon_{it}, \tag{3.6}$$

then the sequential moment restriction becomes

$$E(u_{it}|z_{it}, y_{i,t-1}, z_{i,t-1}, y_{i,t-2}, ..., z_{i,1}, y_{i0}, \alpha_i) = 0 \tag{3.7}$$

So u_{it} is correlated with future values of $y_{i,t-1}$ and is allowed to be correlated with future value of z_{it}. If

$$E(z_{it}u_{is}) = 0 \text{ for all } t \text{ and } s, \tag{3.8}$$

then we have additional moment conditions.

Example 11 *Rational expectation models: suppose u_{it} is the forecasting error and I_{it} is information set at time t, then $E(u_{it}|I_{it-1}) = 0$.*

Example 12 *Euler Equation:*

$$E\left(\frac{U_c^t(c_t)}{U_c^{t-1}(c_{t-1})}r_t|I_{t-1}\right) = 1. \tag{3.9}$$

Suppose

$$U_t = \frac{c_t^{1-\gamma}-1}{1-\gamma}, U_c^t = c_t^{-\gamma}. \tag{3.10}$$

Then

$$E\left[\left(\frac{c_t}{c_{t-1}}\right)^{-\gamma} r_t|I_{t-1}\right] = 0, \tag{3.11}$$

or:

$$E\left(c_t^{-\gamma}r_t|I_{t-1}\right) = 0. \tag{3.12}$$

so the moment conditions are

$$Ec_t^{-\gamma}r_tZ_{t-1} = 0 \tag{3.13}$$

where $Z_{t-1} \in I_{t-1}$.

3.2 Properties of FE and FD Estimators under SeqEx

3.2.1 Inconsistency of the FE Estimator

$$\begin{aligned} plim_{N\to\infty}\left(\widehat{\beta}_{FE}\right) &= \beta + plim_{N\to\infty}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)'(x_{it}-\bar{x}_i)\right)^{-1} \\ &\quad \times\left(\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)'(u_{it}-\bar{u}_i)\right) \\ &= \beta + \left(T^{-1}\sum_{t=1}^{T}E(x_{it}-\bar{x}_i)'(x_{it}-\bar{x}_i)\right)^{-1} \\ &\quad \times T^{-1}\sum_{t=1}^{T}E(x_{it}-\bar{x}_i)'(u_{it}-\bar{u}_i) \end{aligned} \tag{3.14}$$

Now

$$\begin{aligned} & T^{-1}\sum_{t=1}^{T} E\left(x_{it}-\bar{x}_i\right)'\left(u_{it}-\bar{u}_i\right) \\ = \; & T^{-1}\sum_{t=1}^{T} E\left(x_{it}'u_{it}-x_{it}'\bar{u}_i-\bar{x}_i'u_{it}+\bar{x}_i'\bar{u}_i\right) \\ = \; & -E\bar{x}_i'\bar{u}_i \end{aligned} \tag{3.15}$$

Assume that x_{it} is stationary and weakly dependent, then

$$T^{-1}\sum_{t=1}^{T} E\left(x_{it}-\bar{x}_i\right)'\left(x_{it}-\bar{x}_i\right) = T^{-1}\sum_{t=1}^{T} Ex_{it}x_{it}' - E\bar{x}_i\bar{x}_i' = A \in (0,\infty) \tag{3.16}$$

and

$$E\bar{x}_i^{(j)}\bar{u}_i \leq \left\{E\left(\bar{x}_i^{(j)}\right)^2\right\}^{1/2}\left\{E\bar{u}_i^2\right\}^{1/2} = O(1/T) \tag{3.17}$$

provided that $E\bar{x}_i^{(j)} = 0$. Therefore

$$plim_{N\to\infty}\left(\widehat{\beta}_{FE}\right) = \beta + O(1/T). \tag{3.18}$$

3.2.2 Inconsistency of the FD Estimator

$$\begin{aligned} & plim_{N\to\infty}\left(\widehat{\beta}_{FD}\right) \\ = \; & \beta + plim_{N\to\infty}\left(\sum_{i=1}^{N}\sum_{t=2}^{T}\Delta x_{it}'\Delta x_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\Delta x_{it}'\Delta u_{it}\right) \\ = \; & \beta + \left(1/(T-1)\sum_{t=2}^{T} E\Delta x_{it}'\Delta x_{it}\right)^{-1} 1/(T-1)\sum_{t=2}^{T} E\Delta x_{it}'\Delta u_{it} \end{aligned} \tag{3.19}$$

Now

$$\begin{aligned} & \frac{1}{T-1}\sum_{t=2}^{T} E\Delta x_{it}'\Delta u_{it} = \frac{1}{T-1}E\sum_{t=2}^{T}\left(x_{i,t}-x_{t,t-1}\right)'\left(u_{i,t}-u_{i,t-1}\right) \\ = \; & \frac{1}{T-1}\sum_{t=2}^{T}\left(Ex_{i,t}'u_{i,t}-Ex_{i,t}'u_{i,t-1}-Ex_{t,t-1}'u_{i,t}+Ex_{t,t-1}'u_{i,t-1}\right) \\ = \; & -Ex_{i,t}'u_{i,t-1} = O(1). \end{aligned}$$

Assume that x_{it} is stationary and weakly dependent, then

$$1/(T-1)\sum_{t=2}^{T} E\Delta x_{it}'\Delta x_{it} \in (0,\infty) \tag{3.20}$$

Hence

$$plim_{N\to\infty}\left(\widehat{\beta}_{FD}\right) = \beta + O(1). \tag{3.21}$$

Example 13 *Dynamic Panel Data Model: $y_{it} = y_{i,t-1}\beta+\alpha_i+u_{it}$. Add the condition that $|\beta| < 1$ so that y_{it} is (weakly) stationary (we will investigate the consequences of relaxing this condition later on). Assume u_{it} is iid over i and t. The fixed effects estimator (LSDV) estimator is*

$$\begin{aligned}\widehat{\beta} &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}-\bar{y}_i)(y_{it-1}-\bar{y}_{i,-1})}{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it-1}-\bar{y}_{i,-1})^2} \\ &= \beta + \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(u_{it}-\bar{u}_i)(y_{it-1}-\bar{y}_{i,-1})/NT}{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it-1}-\bar{y}_{i,-1})^2/NT} \end{aligned} \tag{3.22}$$

See Nickell (1981) for an explicit derivation. The probability limit of the second term of (3.22) gives the asymptotic bias of the LSDV estimator of the autocorrelation coefficient.

Note that

$$y_{i,t-1} = \sum_{s=0}^{\infty}\beta^s\varepsilon_{i,t-1-s} = \frac{1}{1-\beta}\alpha_i + \sum_{s=0}^{\infty}\beta^s u_{i,t-1-s}, \tag{3.23}$$

the numerator of the second term is, as N tends to infinity,

$$\begin{aligned} & -E\bar{y}_{i,-1}\bar{u}_i \\ &= E\left(\frac{1}{1-\beta}\alpha_i + \frac{1}{T}\sum_{t=1}^{T}\sum_{s=0}^{\infty}\beta^s u_{i,t-1-s}\right)\left(\frac{1}{T}\sum_{i=1}^{T}u_{it}\right) \\ &= E\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{s=0}^{\infty}\beta^s u_{i,t-1-s}\right)\left(\frac{1}{T}\sum_{\tau=1}^{T}u_{i\tau}\right) \\ &= E\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{q=-\infty}^{t-1}\beta^{t-1-q}u_{i,q}\right)\left(\frac{1}{T}\sum_{\tau=1}^{T}u_{i\tau}\right) \\ &= E\left(\frac{1}{T}\sum_{q=-\infty}^{T-1}\sum_{t=q+1}^{T}\beta^{t-1-q}u_{i,q}\right)\left(\frac{1}{T}\sum_{\tau=1}^{T}u_{i\tau}\right) \end{aligned}$$

$$
\begin{aligned}
&= E\left(\frac{1}{T}\sum_{q=-\infty}^{T-1}\frac{\beta^{-q+T}-1}{\beta-1}u_{i,q}\right)\left(\frac{1}{T}\sum_{\tau=1}^{T}u_{i\tau}\right)\\
&= \frac{1}{T^2}\sum_{q=1}^{T-1}\frac{\beta^{-q+T}-1}{\beta-1}\sigma_u^2 = -\frac{1}{T^2}\frac{T-1-\beta T+\beta^T}{(\beta-1)^2}\sigma_u^2 \qquad (3.24)
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&T^{-1}\sum_{t=1}^{T}E\left(y_{it-1}-\bar{y}_{i,-1}\right)^2\\
&= T^{-1}\sum_{t=1}^{T}E\left(\sum_{s=0}^{\infty}\beta^s u_{i,t-1-s}-\frac{1}{T}\sum_{\tau=1}^{T}\sum_{s=0}^{\infty}\beta^s u_{i,\tau-1-s}\right)^2. \qquad (3.25)
\end{aligned}
$$

But

$$
\begin{aligned}
&E\left(\sum_{s=0}^{\infty}\beta^s u_{i,t-1-s}-\frac{1}{T}\sum_{\tau=1}^{T}\sum_{s=0}^{\infty}\beta^s u_{i,\tau-1-s}\right)^2\\
&= \frac{\sigma_u^2}{1-\beta^2}-\frac{2\sigma_u^2}{T(1-\beta^2)}\left(\frac{1-\beta^t}{1-\beta}+\beta\frac{1-\beta^{T-t}}{1-\beta}\right)\\
&\quad +\frac{\sigma_u^2}{T(1-\beta)^2}\left(1-\frac{2\beta(1-\beta^T)}{T(1-\beta^2)}\right) \qquad (3.26)
\end{aligned}
$$

So

$$
\begin{aligned}
&T^{-1}\sum_{t=1}^{T}E\left(y_{it-1}-\bar{y}_{i,-1}\right)^2\\
&= -\frac{\sigma_u^2}{T^2}\frac{\beta^2T^2+T\beta^2-2\beta T^2+T^2-T+2\beta-2\beta^{T+1}}{(\beta-1)^2\left(\beta^2-1\right)} \qquad (3.27)
\end{aligned}
$$

The asymptotic bias of the Fixed effect estimator or LSDV estimator is

$$
\frac{\left(T-1-\beta T+\beta^T\right)\left(\beta^2-1\right)}{\beta^2T^2+T\beta^2-2\beta T^2+T^2-T+2\beta-2\beta^{T+1}} \qquad (3.28)
$$

When $T=2$, *the bias reduces to* $-(\frac{1}{2}\beta+\frac{1}{2})$. *When* $T=3$, *the bias reduces to* $-1/2(\beta+1)(\beta+2)$. *When* T *is large, the dominating bias term is*

$$
\frac{-(\beta-1)\left(\beta^2-1\right)}{\beta^2-2\beta+1}\frac{1}{T}(1+o(1)) = -\frac{(\beta+1)}{T} \qquad (3.29)
$$

- *For small T and $\beta > 0$, we can see that the bias is always negative.*
- *the bias does not tend to zero as β tends to zero.*
- *the smaller T is, the larger the bias.*
- *when T is large, the right-hand side variables become asymptotically uncorrelated; the bias tends to zero as T tends to infinity.*
- *Note that we have assumed that $y_{i0} = \sum_{s=0}^{\infty} \beta^s \varepsilon_{i,-s}$. So the DGP for y_{i0} is the same as any other y_{it}, for $t > 0$. Sometimes, we assume that y_{i0} is a fixed constant. In this case, the exact expression for asymptotic bias will be different. See, for example, the asymptotic bias of the MLE in the next section, where we assume that y_{i0} is not random.*

Exercise 14 *Calculate the asymptotic bias of the FD estimator. Compare it with that of the FE estimator.*

3.2.3 Inconsistency of the ML estimator

Consider the following statistical model:

$$\begin{aligned} y_{it} &= \lambda y_{it-1} + x_{it}\beta + \varepsilon_{it}, i = 1, ..., N, \ \ t = 1, ...T_i \\ \varepsilon_{it} &= \alpha_i + u_{it} \end{aligned} \tag{3.30}$$

where $u_{it} \sim iidN(0, \sigma^2)$, x_{it} is strictly exogeneous and the parameters is

$$\theta = (\lambda, \beta, \alpha_1, ..., \alpha_N, \sigma^2). \tag{3.31}$$

Note that we treat α_i as parameters to be estimated here.

The model can be written in the following form:

$$y_{it} | \{y_{it-1}, ..., y_{i0}, x_i\} \sim N(\lambda y_{it-1} + x_{it}\beta + \alpha_i, \sigma^2) \tag{3.32}$$

Then the density of $\{y\}$ is

$$\begin{aligned} p_\theta\left(y|x, y_0\right) &= \prod_{i=1}^{N} f_\theta\left(y_{i1}, y_{i2}, ..., y_{iT}|x_i, y_{i0}\right) \\ &= \prod_{i=1}^{N} f_\theta(y_{i1}|x_i, y_{i0}) f_\theta\left(y_{i2}|x_{i,}y_{i0}, y_{i1}\right) \times ... \times f_\theta\left(y_{iT}|x_i, y_{i0}, y_{i1}, ..., y_{i,T-1}\right) \end{aligned} \tag{3.33}$$

where

$$f_\theta\left(y_{it}|x_{i,}y_{i0},y_{i1},...,y_{i,t-1}\right)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2\sigma^2}\left(y_{it}-\lambda y_{it-1}-x_{it}\beta-\alpha_i\right)^2\right]. \quad (3.34)$$

Hence

$$\begin{aligned} p_\theta\left(y|x,y_0\right) &= \prod_{i=1}^{N}\prod_{t=1}^{T}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2\sigma^2}\left(y_{it}-\lambda y_{it-1}-x_{it}\beta-\alpha_i\right)^2\right] \\ &= \left(2\pi\right)^{-n/2}\sigma^{-n}\exp\left[-\frac{1}{2\sigma^2}\left(y-X\gamma\right)'\left(y-X\gamma\right)\right] \quad (3.35)\end{aligned}$$

with $n = NT$ with y defined as before and

$$X=\begin{pmatrix} x_{11} & y_{10} & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ x_{1T} & y_{1T-1} & 1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{N1} & y_{N0} & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{NT} & y_{NT} & 0 & 0 & \cdots & 1 \end{pmatrix},\gamma=\begin{pmatrix}\beta \\ \lambda \\ \alpha_1 \\ \alpha_2 \\ ... \\ \alpha_N\end{pmatrix} \quad (3.36)$$

The ML estimator is

$$\hat{\gamma}_{ML}=\left(X'X\right)^{-1}X'y,\ \hat{\sigma}^2_{ML}=e'e/n \quad (3.37)$$

where $e = y - X\hat{\gamma}_{ML}$. It can be shown that $\hat{\gamma}_{ML}$ is the same as the fixed effects estimator or the $LSDV$ estimator except that the $\hat{\gamma}_{ML}$ is conditional on $\{y_{i0}\}_{i=1}^N$.

Consider $T = 2$ and x is null. Then $\hat{\lambda}_{ML}$ can be written as

$$\hat{\lambda}_{ML}=\lambda+\frac{1/N\sum_{i=1}^N\left(y_{i1}-y_{i0}\right)\left(u_{i2}-u_{i1}\right)}{1/N\sum_{i=1}^N\left(y_{i1}-y_{i0}\right)^2}. \quad (3.38)$$

Consider the special case $y_{i0} = 0$ for all i and suppose $\sum\alpha_i^2/N$ converges to a limit as $N\to\infty$. Then

$$\begin{aligned}\hat{\lambda}_{ML} &\to \lambda+\lim\frac{1/N\sum_{i=1}^N\left(\alpha_i+u_{i1}\right)\left(u_{i2}-u_{i1}\right)}{1/N\sum_{i=1}^N\left(\alpha_i+u_{i1}\right)^2} \\ &= \lambda-\frac{\sigma^2}{\lim\sum\alpha_i^2/N+\sigma^2}. \quad (3.39)\end{aligned}$$

If $\alpha_i = 0$ for all i, then $\hat{\lambda}_{ML}$ converges to $\lambda - 1$. The bias can be very serious.

The usual argument for the consistency of maximum likelihood estimator does not apply here because the dimension of the parameter space increases with the sample size. The dimension of parameter space is $k + 2 + N$, and we are taking a limit as $N \to \infty$ for fixed T. It is not surprising that we do not obtain a consistent estimator for α_i. If λ is known, the ML estimate for α_i is based on only two observations. One might, however, have hoped that the ML estimator of a parameter like λ, which is common to all of the individuals, would be consistent as the number of individuals increases. That this is not true in general is known as the incidental parameters problem (Neyman and Scott (1948) and Lancaster (1998)).

3.3 FD+IV Estimator (Anderson and Hsiao)

Consider the first-differenced model:

$$\Delta y_{it} = \Delta x_{it}\beta + \Delta u_{it} \tag{3.40}$$

Under the sequential exogeneity assumption

$$E(x_{is}' u_{it}) = 0 \text{ for all } s = 1, 2, ..., t, \tag{3.41}$$

we have

$$E\Delta x_{is}' \Delta u_{it} = 0 \text{ for all } s = 1, ..., t-1. \tag{3.42}$$

So at time t, we can use $\Delta x_{i,t-1}$ as the potential instrument for Δx_{it} (Anderson and Hsiao (1982)).

Different Choices:

- Estimate $\Delta y_{it} = \Delta x_{it}\beta + \Delta u_{it}$ by pooled 2SLS using $\Delta x_{i,t-1}$ as instruments. When $T = 3$, this reduce to a cross-sectional 2SLS: $x_{i2} - x_{i1}$ is used as instruments for $x_{i3} - x_{i2}$.

- Rather than use lagged Δx_{it} as instruments, we can use lagged x_{it} as instruments. For example, choosing $x_{i,t-1}, x_{i,t-2}$ as instruments is as least as efficient as the procedure than uses $\Delta x_{i,t-1}$ as instruments. The former also gives k overidentifying restrictions that can be used to test the sequential exogeneity. It has been found that the estimator resulting from instrumenting using differences has a singularity point and very large variances over a significant range of parameter values. Instrumenting using levels does not lead to the singularity problem, and results in much smaller variances, and so is preferable.

Problem of Weak Instruments

- When $T = 2$, β may be poorly identified, as the correlation between x_{i1} and $\Delta x_{i,2}$ may be small.
- Even when T is large, we may still the weak instrument problem.

Example 15 *Let*

$$y_{it} = \rho y_{i,t-1} + \alpha_i + u_{it}, \tag{3.43}$$

then the simplest IV estimators are

$$\widehat{\beta}_{IV,1} = \frac{\sum_{i=1}^{N}\sum_{t=3}^{T}(y_{i,t} - y_{i,t-1})(y_{i,t-2} - y_{i,t-3})}{\sum_{i=1}^{N}\sum_{t=3}^{T}(y_{i,t-1} - y_{i,t-2})(y_{i,t-2} - y_{i,t-3})} \tag{3.44}$$

$$\widehat{\beta}_{IV,2} = \frac{\sum_{i=1}^{N}\sum_{t=3}^{T}(y_{i,t} - y_{i,t-1})y_{i,t-2}}{\sum_{i=1}^{N}\sum_{t=3}^{T}(y_{i,t-1} - y_{i,t-2})y_{i,t-2}} \tag{3.45}$$

If the true $\alpha_i = 0$ and the true $\rho = 1$. First differencing yields:

$$\Delta y_{it} = \rho \Delta y_{i,t-1} + \Delta u_{it}. \tag{3.46}$$

Notice that $\Delta y_{i,t-s}$ is uncorrelated with $\Delta y_{i,t-1}$ for $s = 1, 2, ..., t-2$.

3.4 Panel GMM estimator (Arellano and Bond)

3.4.1 The GMM Estimator: Definition

The Anderson-Hsiao instrumental variables estimator may be consistent, but it is not efficient because it does not take into account all the available moment restrictions. Arellano and Bond (1991) argue that a more efficient estimator results from the use of additional instruments whose validity is based on orthogonality between lagged values of x_{it} and the errors.

Consider

$$\Delta y_i = \Delta X_i \beta + \Delta u_i \tag{3.47}$$

At $t = 2$, we have

$$y_{i2} - y_{i1} = (x_{i2} - x_{i1})\,\beta + u_{i2} - u_{i1} \tag{3.48}$$

and x_{i1} is a valid instrument for $(x_{i2} - x_{i1})$, since tbey are likely to be correlated, and is not correlated $u_{i2} - u_{i1}$. At $t = 3$,

$$y_{i3} - y_{i2} = (x_{i3} - x_{i2})\,\beta + u_{i3} - u_{i2}. \tag{3.49}$$

Here x_{i1} and x_{i2} are both valid instruments: neither is correlated with $u_{i3} - u_{i2}$. Proceeding in this manner, we can see that at t, the valid instrument set is $x^o_{i,t-1}$ where

$$x^o_{i,t-1} = (x_{i,1}, x_{i,2}, ..., x_{i,t-1}) \tag{3.50}$$

Define

$$Z_i = \begin{pmatrix} x^o_{i1} & 0 & \cdots & 0 \\ 0 & x^o_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x^o_{i,T-1} \end{pmatrix} \tag{3.51}$$

which is a $(T-1) \times k\,(1+2+...+T-1)$ matrix or a $(T-1) \times k(T(T-1)/2)$ matrix. Then $EZ_i'\Delta u_i = 0$ because

$$Z_i'\Delta u_i = \begin{pmatrix} (x^o_{i1})'\,\Delta u_2 \\ (x^o_{i2})'\,\Delta u_3 \\ \vdots \\ x^o_{i,T-1}\Delta u_T \end{pmatrix} \tag{3.52}$$

The moment condition $EZ_i'\Delta u_i = 0$ is not enough to identify β. We need the following rank condition:

Assumption GMM: Rank$(EZ_i'\Delta X_i) = k$

Note that ΔX_i is a $(T-1) \times k$ matrix and $Z_i'\Delta X_i$ is a $Tk(T-1)/2 \times k$ matrix. $Z_i'\Delta X_i$ is

$$\begin{pmatrix} (x^o_{i1})' & 0 & \cdots & 0 \\ 0 & (x^o_{i2})' & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \left(x^o_{i,T-1}\right)' \end{pmatrix} \begin{pmatrix} \Delta x_{i2} \\ \Delta x_{i3} \\ \vdots \\ \Delta x_{iT} \end{pmatrix} = \begin{pmatrix} (x^o_{i1})'\,\Delta x_{i2} \\ (x^o_{i2})'\,\Delta x_{i3} \\ \vdots \\ \left(x^o_{i,T-1}\right)'\Delta x_{iT} \end{pmatrix} \tag{3.53}$$

Under the assumption $EZ_i'\Delta u_i = 0$ and **Rank**$(EZ_i'\Delta X_i) = k$, β is the unique vector that solves

$$EZ_i'(\Delta y_i - \Delta X_i\beta) = 0. \tag{3.54}$$

To estimate β, we solve the sample analogue:

$$\frac{1}{N}\sum_{i=1}^{N} Z_i'(\Delta y_i - \Delta X_i \beta) = 0. \tag{3.55}$$

In general, we have $k(T(T-1)/2) > k$ (the only exception is $T = 2$), so the above equation will not have a solution. Instead, we choose $\widehat{\beta}$ to make the vector as "small" as possible. Let W be a $k(T(T-1)/2) \times k(T(T-1)/2)$ symmetric and positive semi-definite matrix, a GMM estimator of β is

$$\widehat{\beta}_{GMM} = \arg\min \left(\sum_{i=1}^{N} Z_i'(\Delta y_i - \Delta X_i\beta)\right)' W \left(\sum_{i=1}^{N} Z_i'(\Delta y_i - \Delta X_i\beta)\right) \tag{3.56}$$

The solution is

$$\begin{aligned} \widehat{\beta}_{GMM} &= \left[\left(\frac{1}{N}\sum_{i=1}^{N}\Delta X_i' Z_i\right) W \left(\frac{1}{N}\sum_{i=1}^{N} Z_i'\Delta X_i\right)\right]^{-1} && (3.57)\\ &\times \left(\frac{1}{N}\sum_{i=1}^{N}\Delta X_i' Z_i\right) W \left(\frac{1}{N}\sum_{i=1}^{N} Z_i'\Delta y_i\right) && (3.58) \end{aligned}$$

Let

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_N \end{pmatrix}, \ \Delta X = \begin{pmatrix} \Delta X_1 \\ \Delta X_2 \\ \vdots \\ \Delta X_N \end{pmatrix}, \Delta y = \begin{pmatrix} \Delta y_1 \\ \Delta y_2 \\ \vdots \\ \Delta y_N \end{pmatrix} \tag{3.59}$$

Then

$$\widehat{\beta}_{GMM} = \left(\Delta X' Z W Z' \Delta X\right)^{-1} \Delta X' Z W Z' \Delta y \tag{3.60}$$

3.4.2 The GMM Estimator: Asymptotics

Under the orthogonality and rank conditions, we can show that $\widehat{\beta}_{GMM}$ is consistent. Note

$$\begin{aligned} \widehat{\beta}_{GMM} - \beta &= \left[\left(\frac{1}{N}\sum_{i=1}^{N}\Delta X_i' Z_i\right) W \left(\frac{1}{N}\sum_{i=1}^{N} Z_i'\Delta X_i\right)\right]^{-1} \\ &\times \left(\frac{1}{N}\sum_{i=1}^{N}\Delta X_i' Z_i\right) W \left(\frac{1}{N}\sum_{i=1}^{N} Z_i'\Delta u_i\right). && (3.61) \end{aligned}$$

But

$$\left(\frac{1}{N}\sum_{i=1}^{N} Z_i'\Delta X_i\right) \to EZ_i'\Delta X_i = C \tag{3.62}$$

with rank$(C) = k$, and

$$\left(\frac{1}{N}\sum_{i=1}^{N} Z_i'\Delta u_i\right) \to EZ_i'\Delta u_i = 0. \tag{3.63}$$

Hence

$$\widehat{\beta}_{GMM} - \beta \to \left(C'WC\right)^{-1} CW0 = 0, \tag{3.64}$$

since $rank(CWC') = k$.

We can also show that $\widehat{\beta}_{GMM}$ is asymptotically normal:

$$\sqrt{N}(\widehat{\beta}_{GMM} - \beta) \Rightarrow N(0, V_\beta) \tag{3.65}$$

where

$$V_\beta = \left(C'WC\right)^{-1} C'W\Lambda WC \left(C'WC\right)^{-1} \tag{3.66}$$

This follows easily from the fact that

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} Z_i'\Delta u_i \Rightarrow N(0, \Lambda), \tag{3.67}$$

provided that some moments conditions hold.

3.4.3 Selection of the Weighting Matrix

Let

$$W = \left(\frac{1}{N}\sum_{i=1}^{N} Z_i'Z_i\right)^{-1} = \left(Z'Z/N\right)^{-1} \tag{3.68}$$

Then

$$\begin{aligned}\widehat{\beta}_{GMM} &= \left(\Delta X'ZWZ'\Delta X\right)^{-1} \Delta X'ZWZ'\Delta y \\ &= \left(\Delta X'Z\left(Z'Z\right)^{-1} Z'\Delta X\right)^{-1} \Delta X'Z\left(Z'Z\right)^{-1} Z'\Delta y \end{aligned} \tag{3.69}$$

This is the Pooled 2SLS estimator.

Let $W = \left(\frac{1}{N}\sum_{i=1}^{N} Z_i' G Z_i\right)^{-1}$ where

$$G = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & & 0 & 0 \\ 0 & -1 & 2 & \ddots & 0 & 0 \\ & & \ddots & \ddots & \ddots & \\ 0 & 0 & 0 & \ddots & 2 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 2 \end{pmatrix} \tag{3.70}$$

then the estimator is the pooled 3SLS estimator.

Let $W = \Lambda^{-1}$, then $\sqrt{N}(\widehat{\beta}_{GMM} - \beta)$ is asymptotically normal with variance $\left(C'\Lambda^{-1}C\right)^{-1}$. Λ^{-1} is the optimal weighting matrix in the sense that

$$\left(C'\Lambda^{-1}C\right)^{-1} \leq \left(C'WC\right)^{-1} C'W\Lambda WC \left(C'WC\right)^{-1} \tag{3.71}$$

A Feasible Procedure

- let $\widehat{\beta}$ be an initial consistent estimator if β, for example, $\widehat{\beta} = \widehat{\beta}_{2SLS}$ estimator
- Define $\Delta\widetilde{u}_i = \Delta y_i - \Delta x_i \widehat{\beta}_{2SLS}$.
- Construct a consistent estimator of $\widehat{\Lambda} = N^{-1}\sum_{i=1}^{N} Z_i' \Delta\widetilde{u}_i \Delta\widetilde{u}_i' Z_i$ and choose $\widehat{W} = \widehat{\Lambda}^{-1}$
- Use $\widehat{W}$ to construct the GMM estimator.

$\widehat{W}$ is a consistent estimator of Λ^{-1} under general conditions. It is easy to see that the consistency and asymptotic normality of $\widehat{\beta}_{GMM}$ remain valid with $W = \widehat{\Lambda}^{-1}$.

The 3SLS estimator is asymptotically equivalent to the GMM estimator if

$$EZ_i'\Delta u_i \Delta u_i' Z_i = EZ_i' G Z_i. \tag{3.72}$$

Note that a typical block of $Z_i'\Delta u_i \Delta u_i' Z_i$ is $\Delta u_{it}\Delta u_{is} x_{i,t-1}^{o} x_{i,s-1}^{o\prime}$. So

$$E\Delta u_{it}\Delta u_{is} x_{i,t-1}^{o} x_{i,s-1}^{o\prime} = E\Delta u_{it}\Delta u_{is} E x_{i,t-1}^{o} x_{i,s-1}^{o\prime}$$

i.e. $EZ_i'\Delta u_i \Delta u_i' Z_i = EZ_i' G Z_i$ if

$$E\left(u_{it}^2 | x_{i,t-1}^{o}\right) = \sigma_t^2$$

for some conditional variance σ_t^2 which is independent of $x_{i,t-1}^{o}$ and satisfies $E\sigma_t^2 = \sigma^2$.

3.4.4 Inference Based on the Optimal GMM Estimator

Test the null $H_0 : R\beta = r$

- Wald test
- LR-like test

$$Q_r = 1/N \left(\sum_{i=1}^{N} Z_i'(\Delta y_i - \Delta X_i \widehat{\beta}_r) \right)' \widehat{W} \left(\sum_{i=1}^{N} Z_i'(\Delta y_i - \Delta X_i \widehat{\beta}_r) \right) \tag{3.73}$$

$$Q_{ur} = 1/N \left(\sum_{i=1}^{N} Z_i'(\Delta y_i - \Delta X_i \widehat{\beta}_{\underline{ur}}) \right)' \widehat{W} \left(\sum_{i=1}^{N} Z_i'(\Delta y_i - \Delta X_i \widehat{\beta}_{\underline{ur}}) \right) \tag{3.74}$$

  Then $Q_r - Q_{ur} \Rightarrow \chi^2_q$, where q is the number of independent restrictions.

- Testing Overidentification: $\underline{Q_{ur} \Rightarrow \chi^2_{Tk(T-1)/2-k}.}$

3.5 Models with Other Types of Indep. Variables

3.5.1 Strictly and Sequentially Exogeneous Variables

Suppose

$$\begin{aligned} y_{it} &= x_{it}\beta + w_{it}\delta + \varepsilon_{it}, i = 1, ..., N, \ t = 1, ...T_i \\ \varepsilon_{it} &= \alpha_i + u_{it} \end{aligned} \tag{3.75}$$

where x_{it} is sequentially exogenous while w_{it} is strictly exogeneous. Then the instruments matrix can be expanded into

$$Z_i = \begin{pmatrix} (x_{i1}^o, w_{iT}^o) & 0 & \cdots & 0 \\ 0 & (x_{i2}^o, w_{iT}^o) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \left(x_{i,T-1}^o, w_{iT}^o\right) \end{pmatrix} \tag{3.76}$$

where $w_{iT}^0 = (w_{i1,}w_{i2}, ..., w_{iT})$. Now, $Z_i'\Delta u_i$ becomes

$$Z_i' \Delta u_i = \begin{pmatrix} (x_{i1}^o)' \, \Delta u_{i2} \\ (w_{iT}^o)' \, \Delta u_{i2} \\ (x_{i2}^o)' \, \Delta u_{i3} \\ (w_{iT}^o)' \, \Delta u_{i3} \\ \vdots \\ \left(x_{i,T-1}^o\right)' \, \Delta u_{iT} \\ (w_{iT}^o)' \, \Delta u_{iT} \end{pmatrix} \tag{3.77}$$

3.5.2 SeqEx Variables that Are Uncorrelated with α_i

Write $x_{it} = (x_{1,it}, x_{2,it})$. Assume both $x_{1,it}$ and $x_{2,it}$ are sequentially exogeneous and $cov(x_{1,it}, \alpha_i) = 0$. Give the assumptions on $x_{1,it}$, observations on $x_{1,it}$ up to and including $t = s$ are valid instruments for the levels equation at $t = s$.

To combine the moment conditions for both the first-differenced equations and levels equations, we stack the equations in differences and levels. Let

$$\varepsilon_i^+ = \begin{pmatrix} u_{i2} - u_{i1} \\ u_{i3} - u_{i2} \\ \cdots \\ u_{iT} - u_{iT-1} \\ \varepsilon_{i1} \\ \varepsilon_{i2} \\ \cdots \\ \varepsilon_{iT} \end{pmatrix}, \; y_i^+ = \begin{pmatrix} y_{i2} - y_{i1} \\ y_{i3} - y_{i2} \\ \cdots \\ y_{iT} - y_{iT-1} \\ y_{i1} \\ y_{i2} \\ \cdots \\ y_{iT} \end{pmatrix}, \; X_i^+ = \begin{pmatrix} \Delta x_{i2}^{(1)} & \Delta x_{i2}^{(2)} & ... & \Delta x_{i2}^{(k)} \\ \Delta x_{i3}^{(1)} & \Delta x_{i3}^{(2)} & & \Delta x_{i3}^{(k)} \\ & & & \\ \Delta x_{iT}^{(1)} & \Delta x_{iT}^{(2)} & & \Delta x_{iT}^{(k)} \\ x_{i1}^{(1)} & x_{i1}^{(2)} & ... & x_{i1}^{(k)} \\ x_{i1}^{(1)} & x_{i1}^{(2)} & ... & x_{i1}^{(k)} \\ ... & ... & ... & ... \\ x_{i,T}^{(1)} & x_{i,T}^{(2)} & ... & x_{i,T}^{(k)} \end{pmatrix} \tag{3.78}$$

where $x_{it}^{(k)}$ is the value of k-th regressor for individual i observed at time t. In view of the first differenced and level equations, we have

$$y_i^+ = X_i^+ \beta + \varepsilon_i^+.$$

Denote

$$Z_i^+ = \begin{pmatrix} (x_{2,i1}^o, w_{iT}^o) & 0 & \cdots & 0 & 0 & 0 & & 0 \\ 0 & \left(x_{2,i2}^o, w_{iT}^o\right) & \cdots & 0 & 0 & 0 & & 0 \\ \vdots & \vdots & \ddots & \vdots & & & & \\ 0 & 0 & \cdots & (x_{2,i,T-1}^o, w_{iT}^o) & 0 & & & \\ 0 & 0 & & & x_{1,i1}^o & 0 & & \\ 0 & 0 & & & 0 & x_{1,i2}^o & 0 & \\ \cdots & \cdots & & & & & \ddots & 0 \\ 0 & 0 & & & & & 0 & x_{1,iT}^o \end{pmatrix} \tag{3.79}$$

then $EZ_i^{+\prime}\varepsilon_i^+ = 0$ as

$$\left(Z_i^+\right)' \varepsilon_i^+ = \begin{pmatrix} \left(x_{2,i1}^o\right)' \Delta u_{i2} \\ \left(w_{iT}^o\right)' \Delta u_{i2} \\ \left(x_{2,i2}^o\right)' \Delta u_{i3} \\ \left(w_{iT}^o\right)' \Delta u_{i3} \\ \vdots \\ \left(x_{2,i,T-1}^o\right)' \Delta u_{iT} \\ \left(w_{iT}^o\right)' \Delta u_{iT} \\ \left(x_{1,i1}^o\right)' \varepsilon_{i1} \\ \left(x_{1,i2}^o\right)' \varepsilon_{i2} \\ \vdots \\ \left(x_{1,iT}^o\right)' \varepsilon_{iT} \end{pmatrix}. \tag{3.80}$$

So we can use the GMM approach as before.

3.5.3 Strictly Exogeneous Variables that Are Uncorrelated with α_i

We now consider the case $x_{it} = (x_{1,it}, x_{2,it})$ where $cov(x_{1,it}, \alpha_i) = 0$ and $x_{1,it}$ are strictly exogeneous. In this case, the observations on $x_{1,it}$ for all periods become valid instruments in the level equations. Using the notation ε_i^+, y_i^+ and x_i^+ as before and defining

$$Z_i^+ = \begin{pmatrix} (x_{2,i1}^o, w_{iT}^o) & 0 & \cdots & 0 & 0 & 0 & & 0 \\ 0 & \left(x_{2,i2}^o, w_{iT}^o\right) & \cdots & 0 & 0 & 0 & & 0 \\ \vdots & \vdots & \ddots & \vdots & & & & \\ 0 & 0 & \cdots & (x_{2,i,T-1}^o, w_{iT}^o) & 0 & & & \\ 0 & 0 & & & x_{1,iT}^o & 0 & & \\ 0 & 0 & & & 0 & x_{1,iT}^o & 0 & \\ \cdots & \cdots & & & & & \ddots & 0 \\ 0 & 0 & & & & & 0 & x_{1,iT}^o \end{pmatrix}, \tag{3.81}$$

we have $E\left(Z_i^+\right)' \varepsilon_i^+ = 0$. Again, we can use the GMM approach as before.

To sum up, we have considered the following cases:

	α_i	u_{is}	Levels/Differences
x_{it}	$\text{cov}(x_{it}, \alpha_i) \neq 0$	sequential exogeneity	Differences
x_{it}	$\text{cov}(x_{it}, \alpha_i) \neq 0$	strict exogeneity	Differences
x_{it}	$\text{cov}(x_{it}, \alpha_i) = 0$	sequential exogeneity	Levels
x_{it}	$\text{cov}(x_{it}, \alpha_i) = 0$	strict exogeneity	Levels

3.6 Number of Moment Restrictions

There is a large literature on weak instruments, although there is no consensus on the definition of weak instruments. All researchers seem to agree that when the instruments are weakly correlated with the regression, the problem of weak instruments is present. In this case, the GMM estimator may not be consistent. More recently, many papers find that using too many overindentifying restrictions leads to poor finite sample properties. In practice, it may be better to use a couple of lags (say 3) rather lags back to $t = 1$. A rigorous study of the weak instruments problem is beyond the scope of this course.

3.7 Testing for individual effects (Optional)

A simple case: Suppose

$$y_{it} = \beta y_{it-1} + \varepsilon_{it} \tag{3.82}$$

$$\varepsilon_{it} = \alpha_i + u_{it} \tag{3.83}$$

and $T = 3$. Under the null hypothesis of no individual effects, we have

$$E(y_{i1}\varepsilon_{i2}) = 0, E(y_{i1}\varepsilon_{i3}) = 0, E(y_{i2}\varepsilon_{i3}) = 0 \tag{3.84}$$

or equivalently

$$\begin{aligned} E(y_{i1}(\varepsilon_{i3} - \varepsilon_{i2}) &= 0 \\ E(y_{i1}\varepsilon_{i2}) &= 0 \\ E(y_{i2}\varepsilon_{i3}) &= 0 \end{aligned} \tag{3.85}$$

the first condition holds regardless of the individual effects. So we can use the first condition to identify β and see to what extent the second and third conditions are violated.

Stack the following equation

$$\begin{aligned} y_{1,3} - y_{1,2} &= (y_{1,2} - y_{1,1})\beta + \varepsilon_{1,3} - \varepsilon_{1,2} \\ y_{1,2} &= y_{1,1}\beta + \varepsilon_{1,2} \\ y_{1,3} &= y_{1,2}\beta + \varepsilon_{1,3} \\ \\ &\dots \\ y_{N,3} - y_{N,2} &= (y_{N,2} - y_{N,1})\beta + \varepsilon_{N,3} - \varepsilon_{N,2} \\ y_{N,2} &= y_{N,1}\beta + \varepsilon_{N,2} \\ y_{N,3} &= y_{N,2}\beta + \varepsilon_{N,3} \end{aligned} \tag{3.86}$$

we can write

$$y^* = Y^*\delta + \varepsilon^* \tag{3.87}$$

Let W_i be the instruments for each set of equations. Denote

$$W = \begin{pmatrix} W_1 & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & W_N \end{pmatrix} \tag{3.88}$$

where

$$W_i = diag(W_{i1}, W_{i2}, W_{i3}) = diag(y_{i1}, y_{i1}, y_{i2}) \tag{3.89}$$

Premultiplying $y^* = Y^*\delta + u^*$ by W' yields

$$\begin{pmatrix} W_{i1}'\left(y_{i3} - y_{i2}\right) \\ W_{i2}'y_{i2} \\ W_{i3}'y_{i3} \end{pmatrix} = \begin{pmatrix} W_{i1}'\left(y_{i2} - y_{i1}\right) \\ W_{i2}'y_{i1} \\ W_{i3}'y_{i2} \end{pmatrix} \beta + v_i \tag{3.90}$$

where

$$v_i = \begin{pmatrix} v_{i1} \\ v_{i2} \\ v_{i3} \end{pmatrix} = \begin{pmatrix} W_{i1}'\left(\varepsilon_{i3} - \varepsilon_{i2}\right) \\ W_{i2}'\varepsilon_{i2} \\ W_{i3}'\varepsilon_{i3} \end{pmatrix} \text{ and } v = \begin{pmatrix} v_1 \\ ... \\ v_N \end{pmatrix} = W'\varepsilon^* \tag{3.91}$$

Let $\Omega = var(v) = W'\left(I_N \otimes \Sigma\right)W$ where $\Sigma = var(v_i)$. Then Σ can be estimated by

$$\widehat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} \widehat{v}_i \widehat{v}_i' \tag{3.92}$$

where $\widehat{v}_i$ consists of 2SLS residuals on each equation separately. For example, $\{\widehat{v}_{i1}\}_{i=1}^{N}$ is the residual on the first equation

$$y_{i,3} - y_{i,2} = (y_{i,2} - y_{i,1})\beta + \varepsilon_{i,3} - \varepsilon_{i,2} \tag{3.93}$$

Let SSQ be the weighted sum of the squared transformed residuals

$$SSQ = \left(y^* - Y^*\widehat{\delta}\right)' W\widehat{\Omega}^{-1}W'\left(y^* - Y^*\widehat{\delta}\right)'/N \tag{3.94}$$

where

$$\widehat{\delta} = [Y^{*'}W\widehat{\Omega}^{-1}W'Y^*]^{-1}Y^{*'}W\widehat{\Omega}^{-1}W'y^* \tag{3.95}$$

Compute

$$L = SSQ_R - SSW \tag{3.96}$$

where SSQ_R is the SSR when imposing the full set of orthogonality conditions implied by the null hypothesis and SSW is the SSR when imposing only those restrictions needed for the first differenced version. Specifically, SSW is the weighted sum of the squared residuals based on only the differenced equations:

$$\begin{aligned} y_{1,3} - y_{1,2} &= (y_{1,2} - y_{1,1})\beta + \varepsilon_{1,3} - \varepsilon_{1,2} \\ &\quad ... \\ y_{N,3} - y_{N,2} &= (y_{N,2} - y_{N,1})\beta + \varepsilon_{N,3} - \varepsilon_{N,2} \end{aligned} \tag{3.97}$$

As $N \to \infty$, L is asymptotically χ^2 with 2 degrees of freedom. The test can be generalized for panel AR(p) models, See Holtz-Eakin (1988).

3.8 Initialization and Maximum Likelihood Estimator

The GMM estimator in the previous sections is consistent and asymptotically normal regardless of how the process is initialized. The GMM estimator is robust at the cost of ignoring information in the first observation. In the time series context, whether the first observation is used in the estimation does not matter for robustness and asymptotic efficiency. With short panels, the situation is fundamentally different. As shown by Blundell and Bond (1998) and Hahn (1999), imposing restrictions on the initial condition can greatly improve the efficiency of GMM over certain parts of the parameter space. In this section, we will not discuss how to incorporate the information in the first observation in the GMM framework. Instead, we discuss the problem of initialization in the MLE framework.

Consider the standard dynamic linear panel data model

$$\begin{aligned} y_{it} &= \mu + y_{it-1}\beta + x_{it}\gamma + \varepsilon_{it}, i = 1, ..., N,\ t = 1, ...T_i \\ \varepsilon_{it} &= \alpha_i + u_{it} \end{aligned} \tag{3.98}$$

where α_i and u_{it} are normally distributed and

$$\begin{aligned} E(\alpha_i) &= E(u_{it}) = 0, \\ E\alpha_i x_{it} &= 0,\ E(u_{it}x_{is}) = 0 \text{ for all } t \text{ and } s, \\ E\alpha_i\alpha_j &= \sigma_\alpha^2 1\{i = j\},\ Eu_{it}u_{js} = \sigma_u^2 1\{i = j, t = s\}. \end{aligned}$$

To use information in the first observation y_{i0}, we need to specify how y_{i0} is generated. We assume that

$$y_{i0} = \delta_0 + \delta_1\alpha_i + v_i$$

where $v_i \sim iidN(0, \sigma_v^2)$ and is independent of α_i and $\{u_{it}\}_{t=1}^T$. Some special cases of this specification are:

a. $\delta_1 = 0 : y_{i0}$ is random but uncorrelated with α_i

b. $\delta_1 = 0$ and $\sigma_v^2 = 0 : y_{i0}$ is a fixed constant.

c. $\delta_0 = 0, \delta_1 = 1/(1-\beta)$ and $\sigma_v^2 = \sigma^2/(1-\beta^2) : y_{i0}$ follows the stationary and unconditional distribution of the process.

We will not impose any of the restrictions above.

The likelihood function for the observations $\{y_{i0}, y_{i1}, ..., y_{iT}\}_{i=1}^T$ is:

$$L = \prod_{i=1}^N f(y_{i1}, ..., y_{iT}) = \prod_{i=1}^N f(y_{i1}, ..., y_{iT}|y_{i0}) \prod_{i=1}^N f(y_{i0}) \tag{3.99}$$

with

$$\prod_{i=1}^N f(y_{i0}) = (2\pi)^{-N2} \left|\sigma_0^2\right|^{-N/2} \exp \sum_{i=1}^N \left(-\frac{(y_{i0} - \delta_0)^2}{2\sigma_0^2} \right) \tag{3.100}$$

where

$$\sigma_0^2 = \delta_1^2 \sigma_\alpha^2 + \sigma_v^2$$

Next, conditional on y_{i0},

$$\alpha_i \sim N(\phi(y_{i0} - \delta_0), \sigma_\alpha^2 - \phi^2 \sigma_0^2) \text{ where } \phi = \frac{\delta_1 \sigma_\alpha^2}{\sigma_0^2}$$

Therefore, conditional on y_{i0}, $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, ..., \varepsilon_{iT})'$ has mean $\phi(y_{i0} - \delta_0)$ and variance Ω where

$$\begin{aligned} \Omega &= \left(\sigma_\alpha^2 - \phi^2 \sigma_0^2\right) J_T + \sigma_u^2 I_T \\ : &= \sigma_{\alpha|0}^2 J_T + \sigma_u^2 I_T \end{aligned}$$

So $\prod_{i=1}^N f(\varepsilon_{i1}, ..., \varepsilon_{iT}|y_{i0})$ is

$$(2\pi)^{-NT/2} |\Omega|^{-N/2} \exp \sum_{i=1}^N \left(-\frac{1}{2} \left[\varepsilon_i - \phi(y_{i0} - \delta_0)\right]' \Omega^{-1} \left[\varepsilon_i - \phi(y_{i0} - \delta_0)\right] \right)$$

Using

$$\prod_{i=1}^N f(y_{i1}, ..., y_{iT}|y_{i0}) = \prod_{i=1}^N f(\varepsilon_{i1}, ..., \varepsilon_{iT}|y_{i0})$$

and combining (3.99) and (3.100) gives

$$\begin{aligned}
& L(\sigma_0^2, \sigma_u^2, \sigma_\alpha^2, \delta_0, \mu, \beta, \gamma, \phi) \\
= \; & (2\pi)^{-N/2} \left|\sigma_0^2\right|^{-N/2} \exp \sum_{i=1}^{N} \left(-\frac{(y_{i0} - \delta_0)^2}{2\sigma_0^2} \right) \times \\
& (2\pi)^{-NT/2} \left|\Omega\right|^{-N/2} \exp \left(-\sum_{i=1}^{N} \frac{1}{2} e_i' \Omega^{-1} e_i \right)
\end{aligned}$$

where

$$e_i = y_i - \mu - \beta y_{i,-1} - x_i \gamma - \phi(y_{i0} - \delta_0).$$

Remark 16 *If $\delta_1 = 0$, then $\phi = 0$. In this case, the ML estimates of μ, β, γ are the random effects estimator (or the GLS estimator) if the quasi-demeaning uses the MLE of θ. Therefore the random effect estimator is consistent when y_{i0} is a fixed constant or y_{i0} is random but uncorrelated with α_i.*

Remark 17 *The random effects estimator is inconsistent when $\phi \neq 0$. The consistency of the random effects estimator thus depends crucially on the initialization of the process.*

Remark 18 *The conditional maximum likelihood estimator that maximizes the condition likelihood function defined by*

$$\begin{aligned}
& L_C \left(\sigma_u^2, \sigma_{\alpha|0}^2, \delta_0, \mu, \beta, \gamma, \phi \right) \\
= \; & \prod_{i=1}^{N} f(y_{i1}, ..., y_{iT} | y_{i0}) \\
= \; & (2\pi)^{-NT/2} \left|\Omega\right|^{-N/2} \exp \left(-\sum_{i=1}^{N} \frac{1}{2} e_i' \Omega^{-1} e_i \right)
\end{aligned}$$

is consistent but asymptotically less efficient than the maximum likelihood estimator that maximizes $L\left(\sigma_0^2, \sigma_u^2, \sigma_\alpha^2, \delta_0, \mu, \beta, \gamma, \phi\right)$.

Remark 19 *If the model is correctly specified, MLE will be more efficient than the GMM estimator based on the first differenced equation. To improve the efficiency of the GMM estimator, Ahn and Schmidt (1995), Arellano and Bover (1995) and Blundell and Bond (1998) proposed an additional set of moment conditions. See Ahn and Schmidt (1999) for a survey on the GMM approach applied to the dynamic panel context.*

Bibliography

[1] Ahn, S. C., and Schmidt, P. (1995). "Efficient estimation of models for dynamic panel data", Journal of Econometrics, 68, 5-28.

[2] Ahn, S. C., and Schmidt, P. (1999). "Estimation of linear panel data models using GMM," Generalized Method of Moments Estimation, edited by L. Mátyás.

[3] Anderson, T.W. and Cheng Hsiao (1981). "Estimation of dynamic models with error components," Journal of the American Statistical Association, 589-606.

[4] Anderson, T.W., and Cheng Hsiao (1982). "Formulation and Estimation of Dynamic Models Using Panel Data," Journal of Econometrics, 18, 47-82.

[5] Arellano, M. (1989). "A Note on the Anderson-Hsiao Estimator for Panel Data," Economic Letters, 31, 337-341.

[6] Arellano, M., and Bond, S. (1991). "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations," Review of Economic Studies, 58, 277-297.

[7] Arellano, M., and Bover, O. (1995). "Another look at the instrumental variable estimation of error-components models," Journal of Econometrics, 68, 29-52.

[8] Blundell, R., and Bond, S. (1998). "Initial conditions and moment restrictions in dynamic panel data models," Journal of Econometrics, 87, 115-143.

[9] Holtz-Eakin, D. (1988). "Testing for individual effects in autoregressive models, Journal of Econometrics, 39, 297-307.

[10] Hsiao, Cheng, (1986). Analysis of Panel Data, (New York: Cambridge University Press).

[11] Kiviet, Jan F., (1995). "On Bias, Inconsistency, and Efficiency of Various Estimators in Dynamic Panel Data Models," Journal of Econometrics, 68, 53-78.

[12] Nickell, S. (1981). "Biases in Dynamic Models with Fixed Effects," Econometrica, 49, 1981, 1417-1426

[13] Neyman, J. and Scott, E. L. (1948). "Consistent Estimation from Partially Consistent Observations," Econometrica, 16, 1-32.

[14] Lancaster, T. (2000). "The Incidental Parameter Problem Since 1948," Journal of Econometrics, 95, 391-413.

Chapter 4

Binary Choice Models

Many economic variables are observed as the result of individuals' choices between a limited number of alternatives. In this section, we shall assume that only two alternatives are available, e.g. purchase/not purchase a car, apply for/not apply for a job, obtain/not obtain a loan, travel to work by own car/public transport. These are examples of genuine qualitative choices. Since there are two alternatives, we call it a binary choice. We represent the outcome of the choice by a binary variable.

4.1 Linear Probability Model

4.1.1 Introduction and estimation

In the setting where y_i takes only two values (say zero and one), we have

$$E(y_i|x_i) := g(x_i\beta) = \Pr(y_i = 1|x_i). \tag{4.1}$$

Suppose we would use a linear regression model to explain y_i :

$$y_i = x_i\beta + \epsilon_i. \tag{4.2}$$

Because y_i can take only two values, the error term ϵ_i, for a given value of x_i, can only take two values:

$$\epsilon_i = \begin{cases} 1 - x_i\beta \;\; if \;\, y_i = 1, \\ \;\; -x_i\beta \quad if \;\, y_i = 0, \end{cases} \tag{4.3}$$

In fact, y_i is a Bernoulli random variable. Its p.d.f. is given by:

$$f(y_i; X, \beta) = (Pr(y_i = 1|x_i))^{y_i}(Pr(y_i = 0|x_i))^{1-y_i}. \tag{4.4}$$

The variance of a Bernoulli random variable is given by

$$Var(y_i|x_i) = Pr(y_i = 1|x_i) \times (1 - Pr(y_i = 1|x_i)) = Var(\epsilon_i|x_i). \tag{4.5}$$

But $Pr(y_i = 1|x_i) = x_i\beta$, so

$$Var(\epsilon_i|x_i) = x_i\beta(1 - x_i\beta). \tag{4.6}$$

Clearly heteroskedastic! Given (4.2) and (4.6), we know that the OLS estimator of β is unbiased and consistent. Because of heteroskedasticity, we have to use the robust variance estimator and the robust *t-statistic* to do inference.

Since we know the form of heterogeneity, we can use WLS to obtain more efficient estimates by regressing $y_i/\widehat{\sigma}_i$ on $x_i/\widehat{\sigma}_i$, where

$$\widehat{\sigma}_i = x_i\widehat{\beta}_{OLS}(1 - x_i\widehat{\beta}_{OLS}). \tag{4.7}$$

4.1.2 Pros and Cons

- Simple to calculate.
$$\beta_j = \partial P(y_i = 1|x_i)/\partial x_{ij}. \tag{4.8}$$
- provide good estimates of the partial effects near the center of the distribution of x.
- But $x_i\widehat{\beta}$ may be outside of the interval $[0, 1]$. The partial effect may not be reliable for extreme values of x.

4.2 Probit and Logit

4.2.1 Introduction

To overcome the problems with the linear model, there exist a class of binary choice models designed to model the 'choice' between two discrete alternatives. Essentially these models describe the probability that $y_i = 1$ directly. Typically we choose a cumulative density function $F(x_i\beta)$ for $g(x_i\beta)$, since cdf's by nature are restricted to lie between zero and one. We let

$$\Pr(y_i = 1|x_i) = F(x_i\beta). \tag{4.9}$$

Common choices are:

$$F(w) = \Phi(w) = \int_{-\infty}^{w} \phi(z)dz \text{ standard normal distribution } \Rightarrow \text{Probit Model.} \tag{4.10}$$

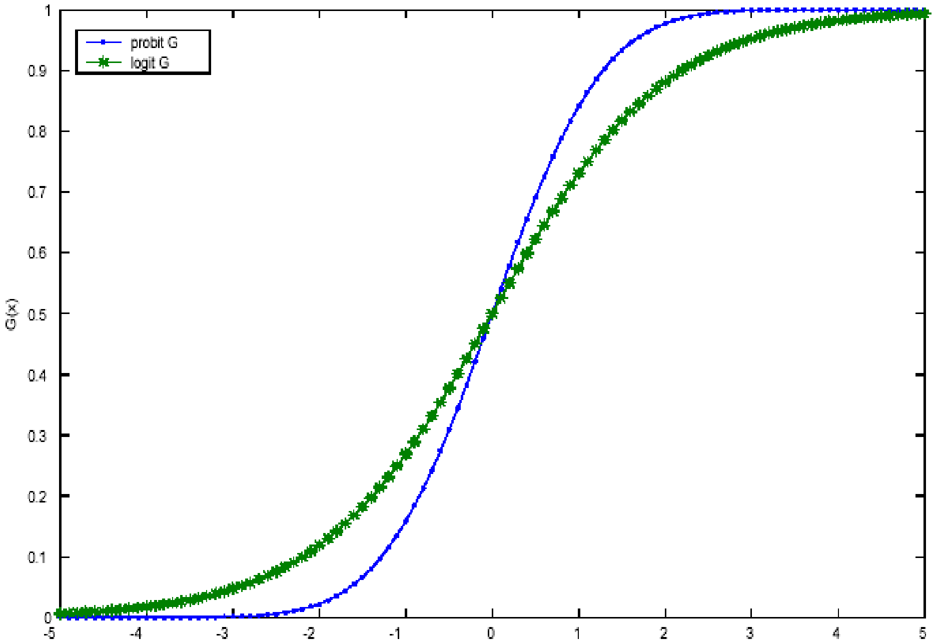

Figure 4.1: Logit vs Probit functions

$$F(w) = L(w) = \frac{\exp(w)}{1+\exp(w)} \text{ logistic distribution } \Rightarrow \text{Logit Model.} \tag{4.11}$$

For the latter one, we have

$$f(w) = F'(w) = \frac{\exp(w)}{(1+\exp(w))^2}. \tag{4.12}$$

The probit model, which uses the normal distribution, may be justified by appealing to a central limit theorem, while the logit model can be justified by the fact that it is similar to a normal distribution but has a much simpler form. The difference between the logit and normal distributions is that the logit has slightly heavier tails. The standard normal has mean zero and variance 1 while the logit has mean zero and variance equal to $\pi^2/3$.

Often the binary choice model is derived from underlying behavioral assumptions: a woman will choose to work for pay if the utility she derives from working is larger than the utility not to work for pay. This leads to a latent variable representation of the model.

Assuming a linear additive relationship, we obtain the utility difference, denoted by y_i^* :

$$y_i^* = x_i\beta + \epsilon_i \tag{4.13}$$

Because utility y_i^*, is unobserved, it is referred to as a latent variable. We assume that an individual chooses to work if the utility difference exceeds a certain threshold level, which can be set equal to zero without loss of generality (assuming our model contains an intercept):

$$y_i = 1\left\{y_i^* > 0\right\}. \tag{4.14}$$

Consequently, we have

$$\begin{aligned} \Pr(y_i &= 1|x_i) = Pr(y_i^* > 0|x_i) = Pr(\epsilon_i > -x_i\beta|x_i) \\ &= 1 - F(-x_i\beta/\sigma_\varepsilon) = F(x_i\beta/\sigma_\varepsilon) \end{aligned} \tag{4.15}$$

provided that the distribution F is symmetric, where $F(x)$ is the c.d.f. of $\epsilon_i/\sigma_\varepsilon$.

In limited dependent variable models we typically lack identification for some unknown parameter(s). This is due to the fact that we observe a limited set of the latent variable: $y_i = \tau(y_i^*)$.

First, in the binary choice example, σ_ε is not identified. Observing y_i, we only know whether y_i^* exceeds the threshold or not, there is no way to find the scale of y_i^*. In the sequel, therefore, we will set $\sigma_\varepsilon = 1$.

Second, setting the threshold for y^* at 0 is likewise innocent if the model contains a constant term. (In general, unless there is some compelling reason, binomial probability models should not be estimated without constant terms.)

Remark 20 *Marginal Effect*

$$\partial \Pr(y = 1|x)/\partial x_{ij} = f(x_i\beta)\beta_j \tag{4.16}$$

Remark 21

$$\frac{\partial p(x)/\partial x_{ij}}{\partial p(x)/\partial x_{ih}} = \beta_j/\beta_h. \tag{4.17}$$

Remark 22 *For discrete x_k, the marginal effect is*

$$\begin{aligned} &F\left(\beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + ... + \left(C_k + 1\right)\beta_k\right) \\ &-F\left(\beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + ... + C_k\beta_k\right). \end{aligned} \tag{4.18}$$

Remark 23 *Apart from the sign of the coefficients, the coefficients in these binary choice models are not easily interpretable. Except maybe in the logit model, where one can consider the β's to represent the marginal effect of x_{ik} on the log of the odds:* $\log(\text{“}odds\text{”}) = x_i\beta$*, where*

$$\text{“}odds\text{”} = \frac{P(y_i = 1|x_i)}{1 - P(y_i = 1|x_i)} \tag{4.19}$$

4.2.2 Estimation

If we assume that $F(\cdot)$ is known, then the optimal parametric estimator for this problem will be ML:

$$\begin{aligned}
&\sum_{i=1}^{n} \log(Pr(y_i = 1|x_i))^{y_i}(Pr(y_i = 0|x_i))^{1-y_i} \\
=& \sum_{i=1}^{n} y_i \log F(x_i\beta) + (1 - y_i)(\log(1 - F(x_i\beta))). \qquad (4.20)
\end{aligned}$$

The score is

$$\begin{aligned}
s_i(\beta) &= \frac{y_i}{F(x_i\beta)} f(x_i\beta)x_i' - \frac{(1 - y_i)}{1 - F(x_i\beta)} f(x_i\beta)x_i' \\
&= \left(\frac{y_i}{F(x_i\beta)} - \frac{(1 - y_i)}{1 - F(x_i\beta)}\right) f(x_i\beta)x_i' \\
&= \frac{y_i - F(x_i\beta)}{F(x_i\beta)(1 - F(x_i\beta))} f(x_i\beta)x_i', \qquad (4.21)
\end{aligned}$$

and the expected Hessian is

$$E\frac{\partial s_i(\beta)}{\partial \beta} = -f^2(x_i\beta)\frac{x_i'x_i}{F(x_i\beta)(1 - F(x_i\beta))}, \qquad (4.22)$$

which is negative semidefinite.

Computational notice: in the probit setting, our probabilities are one dimensional integrals, whereas in the logit setting our probabilities have simple expressions like

$$\frac{exp(x_i\beta)}{1 + exp(x_i\beta)}, \qquad (4.23)$$

and the first order condition is

$$\sum_{i=1}^{n} (y_i - F(x_i\beta)) x_i = 0. \qquad (4.24)$$

Problem 24 *What's going to happen if $y_i = 0$ or 1 for all i?*

The asymptotic variance of $\widehat{\beta}_{MLE}$ is

$$\left(\sum_{i=1}^{N} f^2(x_i\beta)\frac{x_i'x_i}{F(x_i\beta)(1 - F(x_i\beta))}\right)^{-1} \qquad (4.25)$$

which can be written in a familiar form $(X'\Lambda X)^{-1}$ with

$$\begin{aligned} \Lambda &= diag(f^2(x_i\beta)/(F(x_i\beta)(1-F(x_i\beta))), \\ \text{and } X' &= (x_1', x_2', ..., x_k'). \end{aligned} \tag{4.26}$$

For testing about the coefficients, the full menu of procedures is available (LR, LM and Wald):

- The model $P(y=1|x,z) = F(x\beta + z\gamma)$.
- The null $H_0 : \gamma = 0$.
- The Tests: Wald test, LR test, $2(L_{ur} - L_r)$ and LM

4.2.3 Report the Results for Probit and Logit

1. Percentage of correct prediction

2. Weighted average of correct prediction when $y_i = 1$ and $y_i = 0$

3. **Pesudo R**$^2 = 1 - \mathcal{L}_{ur}/\mathcal{L}_0$ (**McFadden**) where $\mathcal{L} = \ln L$.

The most basic way to describe how successful the model is at fitting the data is to report the value of $\ln L$ at $\hat{\beta}$. Since the hypothesis that all other slopes in the model are zero is also interesting, $\ln L$ computed with only a constant term, $\mathcal{L}_0$, should also be reported. Comparing $\mathcal{L}_0$ to $\mathcal{L}_{ur}$ gives us an idea of how much the likelihood improves on adding the explanatory variables.

4. **Pesudo R**$^2 = 1 - SSR_{ur}/SSR_0$ (**McFadden**), where

$$SSR_{ur} = \sum_{i=1}^{N} \left[y_i - g(x_i\widehat{\beta}_{ur}) \right]^2 \tag{4.27}$$

5. Partial effects: $\Delta P(y=1|x) \approx f(x\widehat{\beta})\widehat{\beta}_j \Delta x_j$ for small x_j. Typically, we report at the point $\bar{x}\widehat{\beta}$.

6. Compare logit and probit: for probit $f(\bar{x}\widehat{\beta}) = f(0) = 0.4$; for logit, $f(0) = 0.25$. The logit estimates can be expected to be larger by a factor 0.4/0.25=1.6. To compare with LPM, logit estimates should be divided by 4 while probit estimates should be divided by 2.5.

7. Variance of the partial effect.

Exercise 25 *Derive the asymptotic distributions of the predicted probability and marginal effects.*

4.3 Optimization Methods

This section describes numerical procedures that are used to maximize a likelihood function. Analogous procedures apply when maximizing other functions such as a GMM objective function or any other objective function for a extreme estimator.

The goal is to find the value of β that maximizes $l\left(\beta\right)$, the log-likelihood function. Numerically, we start from any initial point and try to find the direction and the size of the step to take to increase the objective function. We iterate the process until no further increase can be found. That is why many programs ask you for the initial value(s).

Suppose the current value of β is β_j, which is attained after j steps from the starting values. The question is: what is the best step we can take next, that is, what is the best value for β_{j+1}?

Let g_j and H_j denote the gradiant and Hessian of $l\left(\beta\right)$ at β_j, i.e.

$$g_j = \frac{\partial l(\beta)}{\partial \beta}|_{\beta=\beta_j},\ H_j = \frac{\partial^2 l(\beta)}{\partial \beta \partial \beta'}|_{\beta=\beta_j}. \tag{4.28}$$

As we will see the gradient tells us which direction to go and the Hessian helps us to know how far to go. Depending on how the Hessian is estimated and how the step size is determined, we have different algorithms.

4.3.1 Newton–Raphson Algorithm

The first step of the Newton–Raphson algorithm is to approximate the log-likelihood locally by a quadratic function:

$$l(\beta) \approx l(\beta_j) + g_j(\beta - \beta_j) + \frac{1}{2}(\beta - \beta_j)' H_j (\beta - \beta_j). \tag{4.29}$$

We then pick up β to maximize the quadratic function. The FOC is

$$g_j + H_j(\beta - \beta_j) = 0. \tag{4.30}$$

So the maximizing point is

$$\beta = \beta_j - H_j^{-1} g_j. \tag{4.31}$$

We take the above β as our β_{j+1}.

This formula is intuitively meaningful. Consider $k = 1$. Each step of β is the slope of the log-likelihood function divided by its curvature. Suppose the function is locally concave. If the slope is positive, we increase β; If the slope is negative, we decrease β. The curvature determines how large a step is made. If the curvature is great, meaning that the slope changes quickly, then the maximum is likely to be close, and so a small step is taken.

If $l\left(\beta\right)$ were exactly quadratic in β, then the NR procedure would reach the maximum in one step from any starting value. This can verified without much difficulty. However, the log-likelihood function is typically nonquadratic. The step $-H_j^{-1}g_j$ may not be the best and we may want to adjust the step by defining

$$\beta_{j+1} = \beta_j - \lambda H_j^{-1} g_j, \tag{4.32}$$

where λ determines the size of step. To ensure that the objective function increases when we take the next step, we do a 'line search'. We search λ over some interval, say $[0, 2]$, so that $l(\beta_{j+1})$ achieves its maximum at $\lambda^* \in [0, M]$. Mathematically,

$$\lambda^* = \arg \max_{\lambda \in [0,M]} l(\beta_j - \lambda H_j^{-1} g_j). \tag{4.33}$$

Equivalently,

$$\lambda^* = \mu^{-1} \text{ and } \mu = \arg \max_{\mu \in (0,1]} l(\beta_j - \mu^{-1} H_j^{-1} g_j). \tag{4.34}$$

The latter forms may be preferred because μ is constrained to $(0, 1]$ without loss of generality.

The NR procedure has two drawbacks. First, calculation of the Hessian is usually computation-intensive. Procedures that avoid calculating the Hessian at every iteration can be much faster. Second, the NR procedure does not guarantee an increase in each step if the log-likelihood function is not globally concave. Other approaches use approximations to the Hessian that address these two issues. The methods differ in the form of the approximation. Each procedure defines a step as

$$\beta_{j+1} = \beta_j - \lambda Q_j^{-1} g_j \tag{4.35}$$

where Q_j is a $k \times k$ matrix. For NR, $Q_j = H_j^{-1}$. Other procedures use Q_j's that are easier to calculate than the Hessian and are necessarily positive definite, so as to guarantee an increase at each iteration even in convex regions of the log-likelihood function.

4.3.2 BHHH Algorithm

In the context of MLE, Berndt, Hall, Hall, Hausman propose using the outer product in place of Hessian. Note that the gradiant g_j can be written as

$$g(\beta) = \sum_{i=1}^{N} s_i(\beta) \tag{4.36}$$

where

$$s_i(\beta) = \frac{\partial l_i(\beta)}{\partial \beta} \tag{4.37}$$

and the Hessian can be written as

$$H(\beta) = \sum_{i=1}^{N} \frac{\partial^2 l_i(\beta)}{\partial \beta \partial \beta'}. \tag{4.38}$$

From the information equality:

$$E\frac{\partial^2 l_i(\beta)}{\partial \beta \partial \beta'} = -E\frac{\partial l_i(\beta)}{\partial \beta}\left(\frac{\partial l_i(\beta)}{\partial \beta}\right)' \tag{4.39}$$

we may estimate $H(\beta)$ by

$$Q(\beta) = H(\beta) = -\sum_{i=1}^{N} s_i(\beta)s_i'(\beta). \tag{4.40}$$

The resulting iterative procedure

$$\beta_{j+1} = \beta_j - \lambda Q_j^{-1} g_j \tag{4.41}$$

is called BHHH algorithm.

There are two advantages to the BHHH procedure over NR:

1. Q_j is far faster to calculate than H_j. The scores must be calculated to obtain the gradient for the NR procedure anyway, and so calculating Q_j as the average outer product of the scores takes hardly any extra computer time. In contrast, calculating H_j requires calculating the second derivatives of the log-likelihood function.

2. Q_j is necessarily negative (semi)definite. The BHHH procedure is therefore guaranteed to provide an increase in $l(\beta)$ in each iteration, even in convex portions of the function.

4.3.3 The Gauss-Newton Algorithm

The Gauss-Newton algorithm is the same as the Newton-Raphson algorithm except that we replace the Hessian by its expected value, i.e.

$$Q = E(H(\beta)|x). \tag{4.42}$$

To illustrate the difference, consider a nonlinear regression $y_i = m(x_i, \beta) + u_i$ with $E(u_i|x_i) = 0$. Under certain regularity conditions, β can be consistently estimated by

$$\hat{\beta}_{NLS} = \arg\min f(x, y; \beta) = \arg\min \sum_{i=1}^{N} (y_i - m(x_i, \beta))^2. \tag{4.43}$$

Note that

$$\frac{\partial f}{\partial \beta} = -\sum_{i=1}^{N}(y_i - m(x_i,\beta))\frac{\partial m(x_i,\beta)}{\partial \beta} \tag{4.44}$$

and

$$\frac{\partial^2 f}{\partial \beta \partial \beta'} = \sum_{i=1}^{N}\frac{\partial m(x_i,\beta)}{\partial \beta}\left(\frac{\partial m(x_i,\beta)}{\partial \beta}\right)' - \sum_{i=1}^{N}(y_i - m(x_i,\beta))\frac{\partial m^2(x_i,\beta)}{\partial \beta \partial \beta'}. \tag{4.45}$$

As a consequence,

$$E\frac{\partial^2 f}{\partial \beta \partial \beta'} = \sum_{i=1}^{N}\frac{\partial m(x_i,\beta)}{\partial \beta}\left(\frac{\partial m(x_i,\beta)}{\partial \beta}\right)'. \tag{4.46}$$

The Newton-Raphson uses the rhs of (4.45) as $Q(\beta)$ while Gauss-Newton uses the rhs of (4.46) as $Q(\beta)$.

4.4 Specification Issues in Binary Response Models

4.4.1 Neglected Heterogeneity

The structural model of interest is

$$P(y = 1|x, c) = \Phi\left(x\beta + \gamma c\right) \tag{4.47}$$

where x is $1 \times k$ with $x_1 = 1$ and c is a 1×1 random variable. The underlying behavior model is

$$y^* = x\beta + \gamma c + e, \tag{4.48}$$

where $e|(x, c) = N(0, 1)$ and c is independent of x and $c \sim N(0, \tau^2)$. Therefore

$$P(y = 1|x) = \Phi\left(x\beta/\sigma\right) \tag{4.49}$$

where

$$\sigma^2 = \gamma^2\tau^2 + 1. \tag{4.50}$$

If c is neglected in the probit regression, then

$$p\lim \widehat{\beta}_j = \beta_j/\sigma \tag{4.51}$$

⇒ attenuation bias.

In nonlinear models, we usually want to estimate the partial effect instead of the parameters. For the purpose of obtaining the directions of effect or the relative effects of the explanatory variables, estimating β_j/σ is just as good as estimating β_j.

For continuous variable x_j, we would like to estimate

$$\partial P(y=1|x,c)/\partial x_j = \beta_j \phi(x\beta + \gamma c). \tag{4.52}$$

The partial effect evaluated at $c = 0$ is simply $\beta_j \phi(x\beta)$. But the probit regression gives:

$$\partial P(y=1|x)/\partial x_j = \beta_j/\sigma \phi(x\beta/\sigma). \tag{4.53}$$

This means that, if we are interested in the partial effects evaluated at $c = 0$, then the probit of y on x does not do the trick.

However, we can show that

$$\partial P(y=1|x)/\partial x_j = E\partial P(y=1|x,c)/\partial x_j = \beta_j E\phi(x\beta + \gamma c). \tag{4.54}$$

In other words, probit of y on x consistently estimates the average partial effect.

Claim 26 *If η is a normal random variable with zero and variance τ^2 and ξ is a constant, then*

$$E_\eta \phi(\xi + \eta) = \frac{1}{\sqrt{1+\tau^2}} \phi\left(\frac{\xi}{\sqrt{1+\tau^2}}\right) \tag{4.55}$$

Proof. By definition, $E_\eta \phi(\xi + \eta)$ is

$$\begin{aligned}
&\int \frac{1}{2\pi\tau} \exp\left(-\frac{(\xi+\eta)^2}{2} - \frac{\eta^2}{2\tau^2}\right) d\eta \\
&= \frac{1}{2\pi\tau} \exp\left(-\frac{\xi^2}{2}\right) \int \exp\left(-\frac{\eta^2}{2}(1+\tau^{-2}) - \frac{2\xi\eta}{2}\right) d\eta \\
&= \frac{1}{2\pi\tau} \exp\left(-\frac{\xi^2}{2}\right) \int \exp\left(-\frac{(1+\tau^{-2})}{2}\left(\eta + \frac{\xi}{(1+\tau^{-2})}\right)^2 + \frac{\xi^2}{2(1+\tau^{-2})}\right) d\eta \\
&= \frac{1}{\sqrt{2\pi}\tau\sqrt{1+\tau^{-2}}} \exp\left(-\frac{\xi^2}{2} + \frac{\xi^2}{2(1+\tau^{-2})}\right) \\
&= \frac{1}{\sqrt{2\pi}\sqrt{1+\tau^2}} \exp\left(-\frac{1}{2}\frac{\xi^2}{\tau^2+1}\right) = \frac{1}{\sqrt{1+\tau^2}} \phi\left(\frac{\xi}{\sqrt{1+\tau^2}}\right)
\end{aligned} \tag{4.56}$$

■

Remark 27 *The above results depend crucially on the probit structure and the normality assumption.*

Remark 28 *The above results also depend on the assumption that c is independent of x. Otherwise, when $c|x \sim N(x\delta, \sigma_c^2)$, then the probit of y on x gives consistent estimates of $(\beta + \gamma\delta)/\sqrt{(1+\gamma^2\sigma_c^2)}$.*

4.4.2 Continuous Endogenous Explanatory Variables

We now study what can be done to account for endogeneity in probit models. Consider:

$$\begin{aligned} y_1^* &= z_1\delta_1 + \alpha_1 y_2 + u_1 \\ y_2 &= z_1\delta_{21} + z_2\delta_{22} + v_2 = z\delta_2 + v_2 \\ y_1 &= 1\left\{y_1^* > 0\right\} \end{aligned} \tag{4.57}$$

where (u_1, v_2) are bivariate normal distributions and is independent of z. Assume that $\mathrm{var}(u_1) = 1$. Write

$$u_1 = \theta_1 v_2 + e_1 \tag{4.58}$$

where

$$\theta_1 = cov(v_2, u_1)/var(v_2) := \rho_1/\tau_2^2, \tag{4.59}$$

and

$$e_1 \sim N(0, 1-\rho_1^2),\ \ \rho_1 = corr(v_2, u_1). \tag{4.60}$$

Then, we have

$$y_1^* = z_1\delta_1 + \alpha_1 y_2 + \theta_1 v_2 + e_1, \tag{4.61}$$

and

$$P(y = 1|z, y_2, v_2) = \Phi((z_1\delta_1 + \alpha_1 y_2 + \theta_1 v_2)/(1-\rho_1^2)^{1/2}). \tag{4.62}$$

But we do not know v_2 and have to estimate it.

A Two-step Approach

(a) run the OLS regression $y_2 = z_1\delta_{21} + z_2\delta_{22} + v_2 = z\delta_2 + v_2$ to get $\widehat{v}_2$

(b) run the probit regression using $\widehat{v}_2$ in place of v_2 to get estimates of

$$\delta_1/(1-\rho_1^2)^{1/2}, \alpha_1/(1-\rho_1^2)^{1/2} \text{ and } \theta_1/(1-\rho_1^2)^{1/2} \tag{4.63}$$

The average partial effect:

$$E_{v_2}\frac{\partial P(y = 1|z, y_2, v_2)}{\partial z_{1j}} = E_{v_2}\phi\left(z_1\delta_{1\rho} + \alpha_{1\rho} y_2 + \theta_{1\rho} v_2\right)\delta_{1j\rho} \tag{4.64}$$

where

$$\delta_{1\rho} = \delta_1/(1-\rho_1^2)^{1/2},\ \alpha_{1\rho} = \alpha_1/(1-\rho_1^2)^{1/2} \text{ and } \theta_{1\rho} = \theta_1/(1-\rho_1^2)^{1/2}. \tag{4.65}$$

But, under the normality assumption,

$$E_{v_2}\phi\left(z_1\delta_{1\rho}+\alpha_{1\rho}y_2+\theta_{1\rho}v_2\right)=\frac{1}{\sqrt{(1+\tau_2^2\theta_{1\rho}^2)}}\phi\left(\frac{z_1\delta_{1\rho}+\alpha_{1\rho}y_2}{\sqrt{(1+\tau_2^2\theta_{1\rho}^2)}}\right).$$

Therefore, we can estimate the APE by

$$\frac{1}{\sqrt{(1+\widehat{\tau}_2^2\widehat{\theta}_{1\rho}^2)}}\phi\left(\frac{z_1\widehat{\delta}_{1\rho}+\widehat{\alpha}_{1\rho}y_2}{\sqrt{(1+\tau_2^2\widehat{\theta}_{1\rho}^2)}}\right). \tag{4.66}$$

Unfortunately, the asymptotic variance of APE is difficult to compute. An alternative estimate is

$$\frac{1}{n}\sum_{i=1}^{n}\Phi\left(z_1\widehat{\delta}_{1\rho}+\widehat{\alpha}_{1\rho}y_2+\widehat{\theta}_{1\rho}v_{2i}\right). \tag{4.67}$$

Conditional Likelihood Approach

The likelihood function conditional on z is

$$\begin{aligned} f(y_1,y_2|z) &= f(y_1|y_2,z)f(y_2|z) \\ &= f(y_1|y_2,z)\phi\left(\frac{y_2-z\delta_2}{\tau_2}\right). \end{aligned} \tag{4.68}$$

To find $f(y_1,y_2|z)$, we first figure out $P(y_1=1|y_2,z)$:

$$\begin{aligned} P(y_1 &= 1|y_2,z)=P(z_1\delta_1+\alpha_1y_2+u_1>0|y_2,z) \\ &= P(z_1\delta_1+\alpha_1y_2+u_1>0|y_2,z) \\ &= \Phi\left(\frac{z_1\delta_1+\alpha_1y_2+\theta_1\left(y_2-z\delta_2\right)}{\sqrt{1-\rho_1^2}}\right) \\ &= \Phi\left(\frac{z_1\delta_1+\alpha_1y_2+\rho_1/\tau_2\left(y_2-z\delta_2\right)}{\sqrt{1-\rho_1^2}}\right) \\ : &= \Phi(w) \end{aligned} \tag{4.69}$$

Here we have used the fact that, given y_2 and z, $u_1\sim N(\theta_1\left(y_2-z\delta_2\right),1-\rho_1^2)$, where

$$\theta_1=cov(u_1,v_2)/var(v_2)\text{ and }\rho_1=\frac{cov(u_1,v_2)}{\sqrt{var(v_2)}}. \tag{4.70}$$

Therefore the conditional likelihood is

$$\{\Phi(w)\}^{y_1}(1-\Phi(w))^{1-y_1}\phi\left(\frac{y_2-z\delta_2}{\tau_2}\right). \tag{4.71}$$

We can then maximize the the sum of the log-likelihood wrt $\delta_1, \alpha_1, \rho_1, \delta_2$ and τ_2^2.

Remark 29 *1. The conditional MLE is more efficient than the two step procedure but computationally more demanding.*

2. Test $H_0 : \rho_1 = 0$ is straightforward$\Rightarrow$ either t test or LR test.

3. It is easy to abuse the two-step procedure.

4.5 Grouped Data

Sometimes, the data do not record the individual sample values but consist instead of counts or proportions for a number of population strata. Suppose that there are S such strata, let n_s be the sample size from stratum s, and let π_s and p_s denote, respectively, the probability of "success" in stratum s and the fraction of "successes" in the sample drawn from stratum s.

Under stratified random sampling, the sample frequencies of success are independent across strata and have a Bi(n_s;s) distribution. We may also write

$$p_s = \pi_s + U_s; s = 1, 2, ..., S \tag{4.72}$$

where U_s is a random error with mean zero and variance equal to $\pi_s(1-\pi_s)/n_s$.

Given a parametric model $\pi_s = F(x_s\beta)$ for π_s, the above observations suggest two alternative methods for estimating β. The first consists in maximizing the binomial likelihood

$$L(\beta) = c + \sum_{s=1}^{S} n_s(p_s(\log(F(x_s\beta) + (1-p_s)\log(1-F(x_s\beta)). \tag{4.73}$$

The second method is based instead on the regression specification. Since $\pi_s = F(x_s\beta)$, we have

$$F^{-1}(\pi_s) = x_s\beta \tag{4.74}$$

So

$$F^{-1}(p_s) = x_s\beta + v_s \tag{4.75}$$

where v_s approximating error, $v_s = F^{-1}(p_s) - F^{-1}(\pi_s)$.

The model parameter may then be estimated by a feasible WLS regression of the transformed sample proportion on x_s with weights equal to the reciprocal of some consistent estimate of var(v_s).

4.6 Semiparametric Estimation [Optional]

In the nonparametric approach, conditional response probabilities are left completely unspecified, except for the assumption that $P(y = 1|x)$ is a smooth function of x. In the semiparametric approach, conditional response probabilities are specified instead as $P(y = 1|x) = F(x\beta)$, where F is a monotonic function with values in the unit interval, but one seeks to estimate the parameter without making any assumption about the precise shape of F. One attractive feature of this approach is that, by retaining the single-index assumption, it avoids the curse of dimensionality problem that plagues fully nonparametric estimation.

This section discusses two examples. The first is based on quantile restrictions, the second is a semiparametric ML estimator based on nonparametric estimation of the conditional distribution of the unobservables. Both examples are easier to motivate by assuming that the data $(x_1, y_1), ..., (x_n, y_n)$ are a sample from the behavior model $y_i = 1\{\epsilon_i > -x_i\beta\}$.

4.6.1 The Maximum Score Estimator

The maximum score (MS) estimator of Manski (1975, 1985) is an M-estimator based on the criterion function

$$\min \sum_{i=1}^{n} |y_i - 1\{x_i\beta > 0\}| \tag{4.76}$$

over β such that $||\beta|| = 1$. Note that

$$|y_i - 1\{x_i\beta > 0\}| = \begin{cases} 1 \text{ when } y_i = 0, x_i\beta > 0 \\ 0 \text{ when } y_i = 1, x_i\beta > 0 \\ 0 \text{ when } y_i = 0, x_i\beta < 0 \\ 1 \text{ when } y_i = 1, x_i\beta < 0 \end{cases} \tag{4.77}$$

The arg min problems is equivalent to the arg max problem below,

$$\max \sum_{i=1}^{n} y_i 1\{x_i\beta > 0\} + (1 - y_i)1\{x_i\beta < 0\} \tag{4.78}$$

Thus, the MS estimate is the parameter value which maximizes a measure of concordance between the data and the predictions based on $x_i\beta$, namely the frequency of cases for which either $x_i\beta > 0$ and $y = 1$, or $x_i\beta < 0$ and $y = 0$.

If β_0 denotes the target parameter, then the MS estimator is consistent for β_0 whenever the conditional median of the latent response y^* is unique and equal to

$x\beta$ or, equivalently, the conditional median $Med(\epsilon_i|x_i) = 0$. A sufficient condition is that the latent regression errors is continuous and strictly increasing at the origin. For the condition $Med(\epsilon_i|x_i) = 0$ to hold, ϵ_i needs not be independent of x_i. Only its conditional median is required not to depend on x_i. In particular, ϵ_i may well be heteroskedastic without the consistency of the MS estimator being affected.

The MS criterion Q_n, being a step function, is neither continuous nor differentiable. This has a number of implications. Theoretically, it can be shown that the sequence $\hat{\beta}_n$ of MS estimators is not $\sqrt{n}$ -consistent and is not asymptotically normal. Kim and Pollard (1991) show that $n^{1/3}(\hat{\beta} - \beta_0)$ converges in distribution to a random variable that maximizes a particular Gaussian process. This result cannot be used for inference, however, since the properties of the limiting distribution are largely unknown. Inference about the MS estimator, therefore, is usually based on the bootstrap or subsampling.

To overcome the difficulties with the MS estimator, Horowitz (1992) suggests smoothing the MS criterion in order to make it continuous and differentiable. The suggested objective function is

$$\max \sum_{i=1}^{n} \left[2\{y_i = 1\} - 1\right] K\left(\frac{x_i\beta}{\sigma}\right). \tag{4.79}$$

Note that when K is the sign function, the arg max problem is equivalent to the arg max problem considered by Manski. This can be seen by observing

$$\begin{aligned} & \left[2\{y_i = 1\} - 1\right] K\left(\frac{x_i\beta}{\sigma}\right) \\ = & \begin{cases} -1 \;\; if \;\; y_i = 0, x_i\beta > 0 \\ \;\;1 \;\;\; if \;\; y_i = 1, x_i\beta > 0 \\ \;\;1 \;\;\; if \, y_i = 0, x_i\beta < 0 \\ -1 \;\; if \;\; y_i = 1, x_i\beta < 0 \end{cases} \end{aligned}$$

So

$$\begin{aligned} & \left\{\left[2\{y_i = 1\} - 1\right] K\left(\frac{x_i\beta}{\sigma}\right) + 1\right\} /2 \\ = \;\; & y_i 1\left\{x_i\beta > 0\right\} + (1 - y_i) 1\left\{x_i\beta < 0\right\} \end{aligned} \tag{4.80}$$

Horowitz chose $\sigma \to 0$ at a rate that depends on the smoothness of the underlying $F := E(y = 1|x)$ and showed that the rate of convergence can be very close to $n^{-1/2}$. The smoothed maximu score estimator is asymptotically normal.

4.6.2 Semiparametric ML Estimator

A different approach to consistent estimation of β_0 under weak distributional assumptions has been proposed by Klein and Spady (1993). They notice that if we knew the distribution $F(\cdot)$ of the latent errors in the behavior model then the ML estimate of β_0 would be obtained by maximizing the sample log-likelihood

$$l_n(\beta) = \sum_{i=1}^{n} y_i \log F(x_i\beta) + (1-y_i)(\log(1-F(x_i\beta))) \tag{4.81}$$

The basic idea is to replace the function $F(x_i\beta)$, which cannot be computed without knowledge of F, by a function $F^*(x_i\beta)$ which can be estimated, and then maximize with respect to the resulting feasible pseudo log-likelihood.

To construct the function F^*, notice that, by Bayes rule

$$\begin{aligned} F(x_i\beta) &= P\{\epsilon_i > -x_i\beta | x_i\} = P\{\epsilon_i > -x_i\beta | x_i\beta\} \\ &= \frac{P\{\epsilon_i > -x_i\beta\} f_c(x_i\beta | \epsilon_i > -x_i\beta)}{f(x_i\beta)} \end{aligned} \tag{4.82}$$

where f_c and f denotes, respectively, the conditional and unconditional densities of $x_i\beta$. The event $\{\epsilon_i > -x_i\beta_0\}$ is equivalent to $\{y_i = 1\}$. Thus, define

$$\begin{aligned} F^*(x_i\beta) &= \frac{P\{\epsilon_i > -x_i\beta_0\} f_c(x_i\beta | \epsilon_i > -x_i\beta_0)}{f(x_i\beta)} \\ &= \frac{P\{y_i = 1\} f_c(x_i\beta | y_i = 1)}{f(x_i\beta)} \end{aligned} \tag{4.83}$$

and note that $F^*(x_i\beta_0) = F(x_i\beta_0)$.

The resulting pseudo log-likelihood is given by

$$l_n^*(\beta) = \sum_{i=1}^{n} y_i \log F^*(x_i\beta) + (1-y_i)(\log(1-F^*(x_i\beta))). \tag{4.84}$$

The pseudo log-likelihood $l_n^*(\beta)$ cannot be employed directly since $F^*(x_i\beta)$ is unknown. Notice, however, that $P\{y_i = 1\}$ can be estimated consistently by $\sum\{y_i = 1\}/n$ while both the conditional and unconditional densities f_c and f can be estimated nonparametrically as smooth functions of $x_i\beta$.

$$\begin{aligned} \widehat{f}(x\beta) &= \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{(x-x_i)\beta}{h}\right) \\ \widehat{f}_c(x\beta) &= \frac{1}{\sum\{y_i = 1\}h} \sum_{\{i:\, y_i = 1\}} K\left(\frac{(x-x_i)\beta}{h}\right) \end{aligned} \tag{4.85}$$

so

$$F^*(x_i\beta) = \left[\frac{1}{nh}\sum_{\{i:y_i=1\}} K\left(\frac{(x-x_i)\beta}{h}\right)\right]\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{(x-x_i)\beta}{h}\right)\right]^{-1} \tag{4.86}$$

Klein and Spady (1993) show that the resulting estimator is $\sqrt{n}$ -consistent for β_0 and asymptotically normal, and provides a consistent estimator of its asymptotic variance. The key steps in the consistency proof involve showing that $l_n^*(\beta)$ behaves asymptotically as $l_n(\beta)$, and that $l_n(\beta)$ converges, uniformly in β, to a nonstochastic function which has a unique global maximum at $\beta = \beta_0$.

For more thorough introduction on semiparametric estimation, please refer to the book by Pagan and Ullah (1999).

4.7 Panel Logit and Probit Models

4.7.1 Pooled Probit and Logit

Suppose the model is

$$P(y_{it} = 1|x_{it}) = G(x_{it}\beta) \tag{4.87}$$

where G is a known function and x_{it} can contain a variety of factors, including time dummies, time constant or lagged dependent variables.

Partial Likelihood Estimation

The partial log-likelihood is

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\log f(y_{it}|x_{it},\beta) = \sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}\log G(x_{it}\beta) + (1-y_{it})\log(1-G(x_{it}\beta)) \tag{4.88}$$

Note we do not assume that $\prod_{t=1}^{T} f(y_{it}|x_{it})$ is the conditional likelihood of the vector $y_i = (y_{i1}, y_{i2}, ..., y_{iT})$ given some set of conditional variables. For the behavior model $y_{it} = 1\{\varepsilon_{it} > -x_{it}\beta\}$, ε_{it} may be serially correlated for each i. We can assume that $\varepsilon_i = (\varepsilon_{i1}, ..., \varepsilon_{iT})'$ has multivariate normal distribution with variance matrix Σ_ε and construct the joint probability density of y_i given x_i. But this is very complicated and estimation is very computationally intensive. In addition, ε_{it} may be correlated with past and future value of x_{it}. For example, let y_{it} indicate whether a person was arrested for a crime in year t and x_{it} measure the amount of time the person has spent in prison prior to the current year. An arrest this year, $y_{it} = 1$ certainly has an effect on the expected future values of x_{it}.

To understand the partial log-likelihood, let's assume that we only observe $\{y_{i\tau}, x_{i\tau}\}$ for a specific τ and pretend that we do not have the observations for other periods. In other words, we only have cross sectional observations. In this case, the log-likelihood for $\{y_{1\tau}, y_{2,\tau}, ..., y_{N,\tau}\}$ conditional on $\{x_{1\tau}, x_{2,\tau}, ..., x_{N,\tau}\}$ is

$$\sum_{i=1}^{N} y_{i\tau} \log G(x_{i\tau}\beta) + (1 - y_{i\tau}) \log(1 - G(x_{i\tau}\beta)) \tag{4.89}$$

It follows from the usual argument that

$$\hat{\beta}_{\tau} = \arg\max \sum_{i=1}^{N} y_{i\tau} \log G(x_{i\tau}\beta) + (1 - y_{i\tau}) \log(1 - G(x_{i\tau}\beta)) \tag{4.90}$$

is consistent and asymptotically normal. The partial likelihood is just a way to combine the $\beta_{\tau}'s$. Of course, we could just take simple or weighted average of $\hat{\beta}_{\tau}'s$ to get our final estimator. But it is typically in the literature to pool the objective functions and define $\hat{\beta}$ as in (4.88). In fact, the so-defined $\hat{\beta}$ is a weighted average of $\hat{\beta}_{\tau}'s$ with weights depending on $asymvar(\hat{\beta}_{\tau})$.

We now show that the partial MLE will be asymptotically normal. We usually proceed as follows: The FOC is

$$\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(\hat{\beta}) = 0 \tag{4.91}$$

where $s_{it}(\beta) = \nabla_{\beta} \log f(y_{it}|x_{it}, \beta)$. A Taylor expansion of the above FOC gives

$$0 = \sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(\beta_0) - \sum_{i=1}^{N}\sum_{t=1}^{T} H_{it}(\tilde{\beta}) \left(\hat{\beta} - \beta_0\right) \tag{4.92}$$

where $H_{it} = -\nabla_{\beta'}\nabla_{\beta} \log f(y_{it}|x_{it}, \beta)$. Therefore

$$\sqrt{N}\left(\hat{\beta} - \beta_0\right) = \left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} H_{it}(\tilde{\beta})\right)^{-1} \left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(\beta_0)\right) \tag{4.93}$$

Now, under mild regularity conditions,

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(\beta_0) \Rightarrow N(0, B) \tag{4.94}$$

where

$$B = \lim_N var \left(\frac{1}{\sqrt{N}} \sum_{i=1}^N \sum_{t=1}^T s_{it}(\beta_0) \right) \tag{4.95}$$

and

$$\frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T H_{it}(\tilde{\beta}) \to E \sum_{t=1}^T H_{it}(\beta_0) := A \tag{4.96}$$

So

$$\sqrt{N} \left(\hat{\beta} - \beta_0 \right) \Rightarrow N(0, A^{-1}BA^{-1}). \tag{4.97}$$

It remains to estimate A and B. It is easy to see that A can be estimated by

$$\hat{A} = \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T H_{it}(\hat{\beta}) \tag{4.98}$$

or

$$\hat{A} = \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T s_{it}(\hat{\beta}) s_{it}(\hat{\beta})' \tag{4.99}$$

Due to cross sectional independence, $var \left(\frac{1}{\sqrt{N}} \sum_{i=1}^N \sum_{t=1}^T s_{it}(\beta_0) \right)$ is

$$\frac{1}{N} \sum_{i=1}^N E \left(\sum_{t=1}^T s_{it}(\beta_0) \right) \left(\sum_{t=1}^T s_{it}(\beta_0) \right)' \tag{4.100}$$

So B can be estimated by

$$\begin{aligned} & \frac{1}{N} \sum_{i=1}^N \left(\sum_{t=1}^T s_{it}(\hat{\beta}) \right) \left(\sum_{t=1}^T s_{it}(\hat{\beta}) \right)' \\ = & \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T s_{it} \left(\hat{\beta} \right) s_{it} \left(\hat{\beta} \right)' + \frac{1}{N} \sum_{i=1}^N \sum_{t \neq \tau}^T s_{it} \left(\hat{\beta} \right) s_{i\tau} \left(\hat{\beta} \right)' \end{aligned} \tag{4.101}$$

where the second term in the above expression accounts for possible serial correlation in the score.

For the probit model, a simple, general estimator of the asymptotic variance is

$$\left(\sum_{i=1}^N \sum_{t=1}^T A_{it} \left(\hat{\beta} \right) \right)^{-1} \left(\sum_{i=1}^N s_i(\hat{\beta}) s_i \left(\hat{\beta} \right)' \right) \left(\sum_{i=1}^N \sum_{t=1}^T A_{it} \left(\hat{\beta} \right) \right)^{-1} \tag{4.102}$$

where

$$A_{it}\left(\hat{\beta}\right)=\frac{\phi^2\left(x_{it}\hat{\beta}\right)x_{it}'x_{it}}{\Phi\left(x_{it}\hat{\beta}\right)\left[1-\Phi\left(x_{it}\hat{\beta}\right)\right]}, \tag{4.103}$$

and

$$s_i\left(\hat{\beta}\right)=\sum_{t=1}^{T}s_{it}\left(\hat{\beta}\right)=\sum_{t=1}^{T}\frac{\phi\left(x_{it}\hat{\beta}\right)x_{it}'\left[y_{it}-\Phi(x_{it}\hat{\beta})\right]}{\Phi\left(x_{it}\hat{\beta}\right)\left[1-\Phi\left(x_{it}\hat{\beta}\right)\right]}. \tag{4.104}$$

Dynamically Complete Models

Definition: $f_t\left(y_t|x_t,\beta\right)$ is dynamically complete if

$$f_t\left(y_t|x_t,\beta\right)=f_t\left(y_t|x_t,y_{t-1},x_{t-1},y_{t-2},...,y_1,x_1,\beta\right). \tag{4.105}$$

Under this condition

$$Es_{it}\left(\beta_0\right)s_{i\tau}\left(\beta_0\right)'=0 \text{ for } t\neq\tau. \tag{4.106}$$

This is because

$$E\left(s_{it}\left(\beta_0\right)|x_{it}\right)=0 \tag{4.107}$$

and

$$E\left\{s_{it}\left(\beta_0\right)|x_{it},y_{it-1},x_{it-1},y_{it-2},...,y_{i1},x_{i1}\right\}=0. \tag{4.108}$$

Now, for $\tau<t$

$$\begin{aligned}&Es_{it}\left(\beta_0\right)s_{i\tau}\left(\beta_0\right)'\\=\;&EE\left(s_{it}\left(\beta_0\right)s_{i\tau}\left(\beta_0\right)'|x_{it},y_{it-1},x_{it-1},y_{it-2},...,y_{i1},x_{i1}\right)\\=\;&E\left[E\left(s_{it}\left(\beta_0\right)|x_{it},y_{it-1},x_{it-1},y_{it-2},...,y_{i1},x_{i1}\right)s_{i\tau}\left(\beta_0\right)'\right]\\=\;&0\end{aligned} \tag{4.109}$$

because $s_{i\tau}\left(\beta_0\right)$ depends only on $x_{i\tau}$ and $y_{i\tau}$.

4.7.2 Unobserved-effect Probit Model under Strict Exogeneity

The main assumption is

$$P\left(y_{it}|x_i,\alpha_i\right)=P\left(y_{it}|x_{it},\alpha_i\right)=\Phi\left(x_{it}\beta+\alpha_i\right), \tag{4.110}$$

which implies that x_{it} is strictly exogeneous conditional on α_i. This rules out lagged dependent variables, as well as explanatory variables whose future movements depend on current or past values of y.

In addition, we assume that

$$y_{it} \text{ are independent across } t \text{ conditional on } x_i \text{ and } \alpha_i. \tag{4.111}$$

Under assumptions (4.110) and (4.111), the DGP may be described as :

$$\begin{aligned} y_{it} &= \{y_{it}^* > 0\}, \\ y_{it}^* &= x_{it}\beta + \alpha_i + u_{it}, \end{aligned} \tag{4.112}$$

where x_{it} is strongly exogenous and u_{it} is iid N(0,1) across i and t. In this case, the density of $y_{i1}, y_{i2}, ..., y_{iT}$ conditional on x_i and α_i is

$$\prod_{t=1}^{T} f\left(y_{it}|x_{it}, \alpha_i; \beta\right) = \prod_{t=1}^{T} \Phi\left(x_{it}\beta + \alpha_i\right)^{y_{it}} \left(1 - \Phi\left(x_{it}\beta + \alpha_i\right)\right)^{1-y_{it}}. \tag{4.113}$$

The fixed effects probit treats α_i as parameters to be estimated. Unfortunately, in addition to being computationally difficult, estimation of the α_i along with β introduces an incidental parameter problem. In the present case, fixed effects estimate of β is inconsistent for a fixed T.

To sum up, probit does not allow the fixed effect treatment at all. Random effects model is feasible but has been difficult because of multidimensional integration. To prevent contamination of β, we need to integrate the random effects α out. For MLE we must assume a particular distribution for α, say

$$g(\alpha) = 1/\sigma_\alpha \phi\left(\alpha/\sigma_\alpha\right) \tag{4.114}$$

depending on parameters σ_α. Note that the above distribution is the distribution conditional on x_i. This distributional assumption implies that α_i is independent of x_i. Given the above assumptions, we can maximize the following conditional log-likelihood function with respect to both β and σ_α:

$$\sum_{i=1}^{N} \log \int_{-\infty}^{\infty} \prod_{t=1}^{T} f\left(y_{it}|x_{it}, \alpha; \beta\right) 1/\sigma_\alpha \phi\left(\alpha/\sigma_\alpha\right) d\alpha. \tag{4.115}$$

Since β and σ_α can be consistently estimated, we can estimate the partial effect at $\alpha = 0$ and the APE, viz.

$$\beta_j/\sqrt{1+\sigma_\alpha^2}\phi\left(x_i\beta/\sqrt{1+\sigma_\alpha^2}\right). \tag{4.116}$$

Assumptions (4.111) and (4.114) are very strong, and it is possible to relax them. Consider relaxing Assumption (4.111). Using (4.110) and (4.114), we have

$$P\left(y_{it} = 1|x_i\right) = P\left(y_{it} = 1|x_{it}\right) = \Phi\left(x_{it}\beta/\sqrt{1+\sigma_\alpha^2}\right). \tag{4.117}$$

Therefore, as in the previous section, we can estimate $\beta/\sqrt{1+\sigma_\alpha^2}$ from pooled probit of y_{it} on x_{it}. If α_i is truly present or u_{it} is autocorrelated, then y_{it} will not be independent across t. Robust inference is needed to account for the serial correlation, as discussed in the previous section.

Note that the pooled probit likelihood can be obtained by the following 'wrong' operations:

$$\int_{-\infty}^{\infty}\left\{\prod_{t=1}^{T} f\left(y_{it}|x_{it},\alpha;\beta\right)1/\sigma_\alpha\phi\left(\alpha/\sigma_\alpha\right)\right\}d\alpha$$

$$\overset{WRONG}{=}\prod_{t=1}^{T}\left\{\int_{-\infty}^{\infty} f\left(y_{it}|x_{it},\alpha;\beta\right)1/\sigma_\alpha\phi\left(\alpha/\sigma_\alpha\right)d\alpha\right\} \tag{4.118}$$

$$=\prod_{t=1}^{T}\left[\Phi\left(x_{it}\beta/\sqrt{1+\sigma_\alpha^2}\right)\right]^{y_{it}}\left[1-\Phi\left(x_{it}\beta/\sqrt{1+\sigma_\alpha^2}\right)\right]^{1-y_{it}} \tag{4.119}$$

This opertation is wrong because the first equality does not hold. However, this operation can give us some intuition. The partial likelihood approach ignores the possible series correlation.

To allow (flexible) correlation between x_i and α_i, we may follow Chamberlain (1980), but we now need the true regression function and a distributional assumption on the α equation error term. Specifically, we need to assume that

$$\alpha_i|x_i \sim N(\psi+\bar{x}_i\xi,\sigma_a^2) \tag{4.120}$$

or

$$\alpha_i=\psi+\bar{x}_i\xi+a_i \tag{4.121}$$

with $a_i \sim N(0,\sigma_a^2)$. As in the linear model, we can not estimate the effect of time invariant variables. This is because they are indistinguishable from the effect $\bar{x}_i\xi$.

If assumptions (4.110), (4.111) and (4.114) hold, then the latent structure is

$$y_{it}^*=x_{it}\beta+\psi+\bar{x}_i\xi+a_i+u_{it} \tag{4.122}$$

where $u_{it}\sim iidN(0,1)$. Now the parameters β,ψ,ξ and σ_a can be estimated as before, i.e. by maximizing (4.115) with x_{it} properly defined.

Given estimates of ψ,ξ, we can estimate $E\left(\alpha_i\right)$ by

$$\widehat{\psi}+\bar{x}\widehat{\xi} \tag{4.123}$$

where $\bar{x}$ is the cross sectional average of $\bar{x}_i$. Therefore, for any given vector x_{it}, we can estimate the response probability at $E\alpha_i$ by

$$\Phi\left(x_{it}\beta+\widehat{\psi}+\bar{x}\widehat{\xi}\right). \tag{4.124}$$

Even if we drop the independence assumption, we can still estimate the scaled version of β, ψ, ξ. Using (4.110) and (4.121), we have

$$\begin{aligned} P\left(y_{it}=1|x_i\right) &= \Phi\left(\left(x_{it}\beta+\psi+\bar{x}_i\xi\right)/\sqrt{1+\sigma_a^2}\right) && (4.125) \\ &: = \Phi\left(x_{it}\beta_a+\psi_a+\bar{x}_i\xi_a\right) && (4.126)\end{aligned}$$

The average response probability is

$$\begin{aligned} & E_\alpha P\left(y_{it}=1|x_{it}=x^0,\alpha_i\right)=E_\alpha\Phi\left(x^0\beta+\alpha_i\right) \\ = \; & E_\alpha\Phi\left(x^0\beta+\psi+\bar{x}_i\xi+a_i\right)=\Phi\left[\left(x^0\beta+\psi+\bar{x}_i\xi\right)/\sqrt{1+\sigma_a^2}\right] \qquad (4.127)\end{aligned}$$

which can be estimated by

$$\frac{1}{N}\sum_{i=1}^{N}\Phi\left(x^0\hat{\beta}_a+\hat{\psi}_a+\bar{x}_i\hat{\xi}_a\right) \qquad (4.128)$$

4.7.3 Unobserved-effect Logit Model under Strict Exogeneity

We consider the same model as in the previous section:

$$\begin{aligned} y_{it} &= \left\{y_{it}^*>0\right\}, \\ y_{it}^* &= x_{it}\beta+\alpha_i+u_{it}, \qquad (4.129)\end{aligned}$$

where x_{it} is strongly exogenous and u_{it} is iid logistic across i and t.

The problem with the logit model is: Integrate $P\left(y_{it}=1|x_i,\alpha_i\right)=\Lambda\left(x_{it}\beta+\alpha_i\right)$ with respect to the normal density (or other popular continuous density) yields no simple function forms. However, fixed effects logit is possible. The idea is to find the joint distribution of y_i conditional on x_i and α_i and $n_i=\sum_{t=1}^{T}y_{it}$. It turns out that this conditional joint density does not depend on α_i so that it is also the distribution of y_i given x_i and n_i.The idea is in essence that n_i is a sufficient statistic for α_i : given $\{x_{it}\}_{t=1}^{T}$, the likelihood of $\{y_{it}\}_{t=1}^{T}$ does not depend on α_i when conditioned on n_i. This is the same as the linear case but in the current situation the conditional likelihood is more complicated.

First, consider the $T=2$ case, where n_i takes values $\{0,1,2\}$. Intuitively, the conditional distribution of $\{y_{i1},y_{i2}\}$ given n_i can not be informative for β when $n_i=0$ or 2 as these values completely determine the outcome on y_i. However, for

$n_i = 1$,

$$\begin{aligned} P(y_{i2} &= 1|x_i, \alpha, n_i = 1) = \frac{P(y_{i2} = 1, n_i = 1|x_i, \alpha)}{P(n_i = 1|x_i, \alpha)} \\ &= \frac{P(y_{i2} = 1|x_i, \alpha) P(y_{i1} = 0|x_i, \alpha)}{P(y_{i1} = 0, y_{i2} = 1|x_i, \alpha) + P(y_{i1} = 1, y_{i2} = 0|x_i, \alpha)} \\ &= \frac{\Lambda(x_{i2}\beta + \alpha_i)\left(1 - \Lambda(x_{i1}\beta + \alpha_i)\right)}{\left(1 - \Lambda(x_{i1}\beta + \alpha_i)\right)\Lambda(x_{i2}\beta + \alpha_i) + \Lambda(x_{i1}\beta + \alpha_i)\left(1 - \Lambda(x_{i2}\beta + \alpha_i)\right)} \\ &= \frac{\exp\left(x_{i2}\beta + \alpha_i\right)}{\exp\left(x_{i2}\beta + \alpha_i\right) + \exp\left(x_{i1}\beta + \alpha_i\right)} = \frac{\exp((x_{i2} - x_{i1})\,\beta)}{1 + \exp((x_{i2} - x_{i1})\,\beta)} \\ &= \Lambda\left((x_{i2} - x_{i1})\,\beta\right) \end{aligned}$$

Similarly,

$$P(y_{i1} = 1|x_i, \alpha, n_i = 1) = 1 - \Lambda\left((x_{i2} - x_{i1})\,\beta\right). \tag{4.130}$$

The conditional likelihood for observation i is

$$\{n_i = 1\}\, w_i \log \Lambda\left((x_{i2} - x_{i1})\,\beta\right) + (1 - w_i) \log\left[1 - \Lambda\left((x_{i2} - x_{i1})\,\beta\right)\right] \tag{4.131}$$

where

$$w_i = \{y_{i1} = 0, y_{i2} = 1\}. \tag{4.132}$$

The above likelihood approach is equivalent to a standard cross-sectional logit of w_i on $x_{i2} - x_{i1}$ using the observations for which $n_i = 1$.

For a general T, the log-likelihood is more complicated, but it is tractable. First,

$$\begin{aligned} & P\left(y_{i1} = y_1, ..., y_{it} = y_T|x_i, c_i, n_i = n\right) \\ = \; & \frac{P\left(y_{i1} = y_1, ..., y_{it} = y_T|x_i, c_i\right)}{P(n_i = n)} \\ = \; & \frac{\prod_{t=1}^{T} P\left(y_{it} = y_t|x_i, c_i\right)}{\sum' P\left(y_{i1} = y_1, ..., y_{it} = y_T|x_i, c_i, n_i = n\right)} \\ = \; & \frac{\exp \sum_{t=1}^{T} y_{it}(x_{it}\beta)}{\sum_{a \in R_i} \exp \sum_{t=1}^{T} a_t(x_{it}\beta)} \end{aligned} \tag{4.133}$$

where

$$R_i = \left\{ a \in R^T : a_t \in \{0, 1\}, \sum_{t=1}^{T} a_t = n_i \right\}. \tag{4.134}$$

The log-likelihood summed over i can be used to obtain a $\sqrt{N}-$asymptotically normal estimator of β, and all inference follows from conditional MLE theory.

Remark 30 *The log-odds ratio depends on α, which is not known.*

Remark 31 *We can not estimate the average partial effect because we do not know the distribution of α_i. Even worse, the mean of α_i may be nonzero.*

Remark 32 *The consistency replies on the independency assumption.*

4.7.4 Dynamic Unobserved Effect Model

The model:

$$P\left(y_{it}=1|y_{i,t-1,...,}y_{i,0},z_i,\alpha_i\right)=G\left(z_{it}\delta+\rho y_{i,t-1}+\alpha_i\right) \tag{4.135}$$

where we have assumed that z_{it} is strictly exogeneous. The joint density is

$$\begin{aligned} & f(y_{i1},y_{i2},...,y_{iT}|y_{i0,}z_i,\alpha_i;\beta) \\ = & \prod_{t=1}^{T}P\left(y_{it}|y_{i,t-1,...,}y_{i,0},z_i,\alpha_i;\beta\right) \quad (4.136) \\ = & \prod_{t=1}^{T}G\left(z_{it}\delta+\rho y_{i,t-1}+\alpha_i\right)^{y_{it}}\left(1-G\left(z_{it}\delta+\rho y_{i,t-1}+\alpha_i\right)\right)^{1-y_{it}} \quad (4.137) \end{aligned}$$

With fixed T asymptotics, this density will not deliver a consistent estimator of β, due to the incidental parameter problem. To avoid the the incidental parameter problem, we again make distributional assumptions on $\alpha_i's$ and integrate them out.

$$f(y_{i1},y_{i2},...,y_{iT}|y_{i0,}z_i;\theta)=\int_{-\infty}^{\infty}f(y_{i1},y_{i2},...,y_{iT}|y_{i0,}z_i,\alpha;\beta)h(\alpha|y_{i0},z_i;\gamma)d\alpha \tag{4.138}$$

When $G=\Phi$, it is convenient to assume that $\alpha_i=\psi+y_{i0}\xi_0+\bar{z}_i\xi+a_i$ where $a_i\sim N(0,\sigma_a^2)$ and is independent of (y_{i0},z_i). In this case, we have

$$y_{it}=\{\psi+z_{it}\delta+\rho y_{i,t-1}+y_{i0}\xi_0+\bar{z}_i\xi+a_i+e_{it}>0\} \tag{4.139}$$

Therefore, the density of $y_{i1},y_{i2},...,y_{iT}$ given (y_{i0},z_i) is

$$\sum_{i=1}^{N}\log\int_{-\infty}^{\infty}\prod_{t=1}^{T}f\left(y_{it}|y_{i0,}z_i,a;\beta\right)1/\sigma_a\phi\left(\alpha/\sigma_a\right)da \tag{4.140}$$

where

$$f\left(y_{it}|y_{i0,}z_i,a;\beta\right)=\Phi\left(x_{it}\beta\right)^{y_{it}}\left(1-\Phi\left(x_{it}\beta\right)\right)^{1-y_{it}} \tag{4.141}$$

and $x_{it}=\left(1,z_{it},y_{it-1},y_{i0,}\bar{z}_i\right).$

For more details such as how to initialize the process differently, see Ch 7.4 of Hsiao (2003).

Bibliography

[1] Kim J. and D. Pollard (1990). “Cube Root Asymptotics,” Annals of Statistics, 18,191-219.

[2] Klein, R., and R. Spady (1993): “An Efficient Semiparametric Estimator for Discrete Choice Models”, Econometrica, 61, 387-421

[3] Horowitz J. (1992) “A Smoothed Maximum Score Estimator for the Binary Response Model,” Econometrica, 60(3), May 1992, pp. 505-31.

[4] Pagan, A., and A. Ullah (1999). “Nonparametric Econometrics”, Cambridge University Press.

[5] D. L. McFadden, “Econometric Analysis of Qualitative Response Models,” Handbook of Econometrics, II, Ch. 24. http://www.elsevier.nl/hes/books/02/02/024/c0202024.htm

[6] Manski C. (1975). “Maximum Score Estimation of the Stochastic Utility Model of Choice,” Journal of Econometrics, 3(3), pp. 205-28.

[7] Manski, C. (1985). “Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator,” Journal of Econometrics 27, pp. 313-333.

Chapter 5

Multinomial Response Models

So far we talked about 0/1 decisions. What if there are more response categories? An important distinction is between ordered categorical data, where the response categories possess a natural ordering (e.g. low income, mid-level income or high income, bond rating A, B, C,...), and unordered categorical data, where the response categories are mere labels totally devoid of structure (e.g. traveling by bus, by train or by car). Different models are used in the two cases.

5.1 Probabilistic Choice Model for Unordered Response

Probabilistic choice models are based on utility maximization. We assume that the utility of individual i from alternative j is given by

$$y_{ij}^{*} = x_{ij}\beta + a_{ij}, \ j = 0, 1, 2, ..., J. \tag{5.1}$$

where x_{ij} is the vector of values of attributes of j-th choice as perceived by the i-th individual and a_{ij} is a $(0, \sigma^2)$ random variable which is used to capture factors unobservable to the researcher. Individuals choose the option that maximizes his utility. Let y_i denote the choice of individual i, then

$$y_i = \arg\max_{j} \{y_{i0}^{*}, y_{i1}^{*}, ..., y_{iJ}^{*}\} \tag{5.2}$$

As an example, consider a person who can take a car, a bus or a subway to work. The researcher observes the time and cost that the person would incur under each mode. However, the researcher realizes that there are factors other than time and cost that affect the person's utility and hence his choice. The researcher specifies

$$\begin{aligned} y_{ic}^* &= T_{ic}\beta_1 + M_{ic}\beta_2 + a_{ic} \qquad (5.3) \\ y_{ib}^* &= T_{ib}\beta_1 + M_{ib}\beta_2 + a_{ib} \qquad (5.4) \\ y_{is}^* &= T_{is}\beta_1 + M_{is}\beta_2 + a_{is} \qquad (5.5) \end{aligned}$$

where T_{ic} and M_{ic} are the time and cost (in money) that the person incurs traveling to work by car, T_{ib} and M_{ib}, T_{is} and M_{is} are defined analogously for bus and subway.

The probability that the person chooses bus instead of car and subway is the probability that

$$\beta_1 T_{ib} + \beta_2 M_{ib} + a_{ib} > \beta_1 T_{ic} + \beta_2 M_{ic} + a_{ic}$$

and

$$\beta_1 T_{ib} + \beta_2 M_{ib} + a_{ib} > \beta_1 T_{is} + \beta_2 M_{is} + a_{is}$$

Remark 33 *Can we include a constant in the utility specification so that*

$$y_{ij}^* = \alpha + x_{ij}\beta + a_{ij}, \; j = 0, 1, 2, ..., J? \qquad (5.6)$$

The answer is no. The presence of α changes all the utilities $(y_{i0}, y_{i1}, ..., y_{iJ})$ by the same amount. The rank of the utilities and thus individual's choice does not depend on α. Therefore the intercept α is not identified.

Remark 34 *Can we include an alternative-specific constant in the utility specification so that*

$$y_{ij}^* = \alpha_j + x_{ij}\beta + a_{ij}, \; j = 0, 1, 2, ..., J? \qquad (5.7)$$

The answer is yes. But we can not identify all $\alpha_j^{\prime}s$ and have to normalize one α, say α_0, to be zero.

Remark 35 *Since*

$$\begin{aligned} \arg\max_j \{y_{ij}^*\} &= \arg\max_j \{y_{ij}^* - y_{i0}^*\} \qquad (5.8) \\ &= \arg\max_j \{(x_{ij} - x_{i0})\beta + a_{ij} - a_{i0}\} \qquad (5.9) \end{aligned}$$

we can not include a variable in x_{ij} if it is constant across different alternatives. For example, we can not include age in y_{ij}^* by simply letting

$$\begin{aligned} y_{ic}^* &= T_{ic}\beta_1 + M_{ic}\beta_2 + Age_i\beta_3 + a_{ic} \qquad (5.10) \\ y_{ib}^* &= T_{ib}\beta_1 + M_{ib}\beta_2 + Age_i\beta_3 + a_{ib} \qquad (5.11) \\ y_{is}^* &= T_{is}\beta_1 + M_{is}\beta_2 + Age_i\beta_3 + a_{is} \qquad (5.12) \end{aligned}$$

because in this specification age does not affect one's decision. If we believe that Age actually plays a role, then we have to allow the coefficient associated with Age to change with the alternative. In general, we can assume

$$y_{ij}^{*} = x_{ij}\beta + z_i\gamma_j + a_{ij},\ j = 0, 1, 2, ..., J. \tag{5.13}$$

where z_i is the individual-specific variable. In the above specification, we can not identify all γ's and need to normalize one γ, say γ_0, to be zero.

Remark 36 *A variant of the utility specification is to allow β to depend on individual-specific characteristics. For example*

$$\beta_k^{(i)} = \beta_k + w_i\theta_k + \sigma_k u_k^{(i)}$$

This specification is used widely in the empirical IO literature. We will not discuss this extension in this class

Remark 37 *Another variant of the utility specification is to allow a_{ij} to be heteroskedastic. In this case, var(a_{ij}) $= \sigma_j^2$.*

5.2 Conditional and Multinomial Logit Model

5.2.1 The model

Assume that a_{ij} are independently distributed with CDF

$$F(a) = \exp\left(-\exp(-a)\right), \tag{5.14}$$

the type I extreme value distribution, then

$$f(a) = F'(a) = \exp\left(-a - \exp(-a)\right), \tag{5.15}$$

In this case, we can show that

$$P\left(y_i = j|x\right) = \frac{\exp(v_{ij})}{\sum_{h=0}^{J}\exp\left(v_{ij}\right)} \text{ where } v_{ij} = x_{ij}\beta + z_i\gamma_j \tag{5.16}$$

Proof.

$$\begin{aligned} P\left(y_i = j|x\right) &= P\left(y_{ij}^{*} > y_{i,-j}^{*}\right) \\ &= P\left(v_{ij} + a_{ij} > v_{ik} + a_{ik}, \text{for all}\, k \neq j\right) \end{aligned} \tag{5.17}$$

So $P\left(y_i = j|x\right)$ is

$$\begin{aligned}
& P\left(a_{ij} + v_{ij} - v_{ik} > a_{ik}, \text{for all}\, k \neq j\right) \\
= & \int_R \prod_{k\neq j} F(a_{ij} + v_{ij} - v_{ik}) f(a_{ij}) da_{ij} \\
= & \int_R \prod_{k\neq j} \exp\left(-\exp(-a_{ij} - (v_{ij} - v_{ik}))\right) \exp\left\{-a_{ij} - \exp(-a_{ij})\right\} da_{ij} \\
= & \int_R \exp\left(-\sum_{k\neq j} \exp(-\xi - (v_{ij} - v_{ik}))\right) \exp\left[-\xi - \exp(-\xi)\right] d\xi \qquad (5.18) \\
= & \int_R \exp\left(-\exp(-\xi) \sum_{k\neq j} \exp(v_{ik} - v_{ij})\right) \exp(\left[-\xi - \exp(-\xi)\right] d\xi \\
= & -\int_{-\infty}^{\infty} \exp\left(-\exp(-\xi)\eta\right) \exp\left[-\exp(-\xi)\right] d\exp(-\xi)
\end{aligned}$$

where $\eta = \sum_{k\neq j} \exp(v_{ik} - v_{ij})$. Let $\lambda = \exp(-\xi)$, then the above probability becomes

$$\begin{aligned}
= & \int_0^{\infty} \exp\left(-\lambda\eta\right) \exp\left[-\lambda\right] d\lambda \\
= & \int_0^{\infty} \exp\left(-\lambda(\eta + 1\right) d\lambda \\
= & -\frac{1}{\eta + 1} \exp\left(-\lambda(\eta + 1\right) |_0^{\infty} \qquad (5.19) \\
= & \frac{1}{\eta + 1}
\end{aligned}$$

Therefore

$$P\left(y_i = j|x\right) = \frac{1}{1 + \sum_{k\neq j} \exp(v_{ik} - v_{ij})} = \frac{\exp(v_{ij})}{\sum_{h=0}^{J} \exp\left(v_{ih}\right)} \qquad (5.20)$$

which completes the proof. ■

Remark 38 *When $v_{ij} = x_{ij}\beta$, we have*

$$P\left(y_i = j|x\right) = \frac{\exp(x_{ij}\beta)}{\sum_{h=0}^{J} \exp\left(x_{ih}\beta\right)}, j = 0, 1, ..., J \qquad (5.21)$$

The above probabilities constitute what is usually called the conditional logit model.

Remark 39 *When $v_{ij} = z_i \gamma_j$, we have*

$$P\left(y_i = j|x\right) = \frac{\exp(z_i \gamma_j)}{\sum_{h=0}^{J} \exp\left(z_i \gamma_h\right)}, j = 0, 1, ..., J \tag{5.22}$$

The above probabilities constitute what is usually called the multinomial logit model.

Remark 40 *The difference between the conditional logit and multinomial logit models:*

- *In the MNL model, the conditioning variables do not change across alternative: for each i, z_i contains variables specific to the individual but not to the alternatives. The model is appropriate for problems where characteristics of the alternatives are not important.*
- *The CL model is intended specifically for problems where the individual choice are at least made based on the observable attributes of each alternative.*
- *Define a set of dummies to indicate the alternatives:*

$$D0_j = \begin{cases} 1 & if\ j{=}0 \\ 0 & otherwise \end{cases}, D1_j = \begin{cases} 1 & if\ j{=}1 \\ 0 & otherwise \end{cases}, ...$$

and let $x_{ij} = (z_i \times D0_j, z_i \times D1_j, ..., z_i \times DJ_j)$, then

$$P\left(y_i = j|x\right) = \frac{\exp(z_i \gamma_j)}{\sum_{h=0}^{J} \exp\left(z_i \gamma_h\right)} = \frac{\exp(x_{ij} \gamma)}{\sum_{h=0}^{J} \exp\left(x_{ij} \gamma\right)}$$

where $\gamma = (\gamma_0, \gamma_1, ..., \gamma_J)'$. Therefore CL model contains MNL model as a special case.

5.2.2 Estimation

Given the probabilities $P\left(y_i = j|x\right)$, we can estimate the logit model by MLE. The log-likelihood function is

$$\ln L = \sum_{i=1}^{n} \sum_{j=0}^{J} 1\left\{y_i = i\right\} \ln P(y_i = j|x_{ij})$$

5.2.3 The limitation of the model

Note that

$$P\left(y_i = j|x\right)/P\left(y_i = h|x\right) = \exp\left[\left(x_{ij} - x_{ih}\right)\beta\right] \tag{5.23}$$

so the relative probabilities for any two alternatives depends only on the attributes of those two alternatives. This is called the independence from irrelevant alternatives. This assumption is not plausible in many applications.

Consider for example the choice between a blue bus and a car. The IIA assumption implies that, if a new alternative, say a red bus, is introduced, all of the existing probabilities are reduced by the same proportion, irrespective of the new choice's degree of similarity to any of the existing ones.

Suppose that a person is indifferent between car and bus, that is,

$$\Pr(C|C, B) = \Pr(B|C, B) = 0.5 \tag{5.24}$$

Since the person is indifferent between car and bus, it would be reasonable to assume that

$$Pr(C|C, B, R) = 0.5 \tag{5.25}$$

In this case, however, the logit link implies that

$$\frac{\Pr(C|C, B, R)}{\Pr(B|C, B, R)} = \frac{\Pr(C|C, B)}{\Pr(B|C, B)} = 1 \tag{5.26}$$

and

$$\frac{\Pr(C|C, B, R)}{\Pr(R|C, B, R)} = \frac{\Pr(C|C, R)}{\Pr(R|C, R)} = 1 \tag{5.27}$$

so

$$\Pr(C|C, B, R) = \Pr(B|C, B, R) = \Pr(R|C, B, R) = \frac{1}{3} \tag{5.28}$$

which is less than 0.5.

The IIA problem arises because we assume that a_{ij} are independent across j. If we two alteratives are close substitute, we expect that the random utilities are correlated. We can test whether some alternatives are potentially correlated by using a typical Hausman test (Hausman and McFadden, 1984). Under H_0 : IIA, one can estimate a subset of the β_j parameters consistently but inefficiently by dropping the individuals who choose the potentially correlated alternatives. These $\beta_j's$ can then be compared to those estimated using the whole data set with all options. Of course, if IIA is violated, the latter will be inconsistent. In absence of some natural grouping of the alternatives, the choice of the subset to leave out is arbitrary and, hence, so is the test.

5.3 Multinomial Probit Model

Multivariate probit allows for a full correlation structure with $a_i \sim N(0, \Sigma)$ and requires J dimensional numerical integration. One has to impose normalization and identification restrictions on the $J(J+1)$ free elements σ of the $m \times m$ matrix Σ.

Consider the case $J = 2$, the choice of the first alternative $P[y_i = 0|x_i]$ corresponds to the joint occurrence of

$$\eta_{01} := a_{i0} - a_{i1} > -(x_{i0} - x_{i1})\beta \tag{5.29}$$

and

$$\eta_{02} := a_{i0} - a_{i2} > -(x_{i0} - x_{i2})\beta. \tag{5.30}$$

One can then derive the variance-covariance of the joint normal pdf of η_{01} and η_{02}, the 2×2 matrix $\tilde{\Sigma}$, from the original σ elements. Finally,

$$\begin{aligned} P(y_i &= 0|x_i) \qquad (5.31) \\ &= \int_{-(x_{i0}-x_{i1})\beta}^{\infty} \int_{-(x_{i0}-x_{i2})\beta}^{\infty} \frac{1}{2\pi} \left|\tilde{\Sigma}\right|^{-1/2} \exp\left(-1/2\left(\eta_{01}, \eta_{02}\right) \tilde{\Sigma}^{-1} \left(\eta_{01}, \eta_{02}\right)'\right) d\eta_{02} d\eta_{01} \end{aligned}$$

Alternatively, the independence assumption of CL can be relaxed using the generalized extreme value (GEV) models. The GEV distribution generalizes the independent univariate extreme value cdfs to allow for a_i correlation across choices:

$$F(a_{i0,}a_{i1}, a_{i2}, ..., a_{iJ}) = \exp[-G(\exp(-a_{i0}), ..., \exp(-a_{iJ}))] \tag{5.32}$$

for some function G. The GEV approach has been widely used in the context of the nested logit model. See Train (2003).

5.4 Nested Logit Model

Example 41 *Choice of house: choose the neighborhood and select a specific house within a chosen neighborhood. Choose to travel by plane, then choose among the airlines.*

In the presence of a nested structure, we assume that the utility from house j in neighborhood k looks as follows:

$$V_{kj} = x_{kj}\beta + z_k\alpha + a_{kj}, \tag{5.33}$$

where z_k are characteristics of neighborhoods and x_{kj} are house-specific characteristics. To facilitate estimation when the number of choices is very large but the decision problem has a tree structure, we use $p_{kj} = p_k p_{j|k}$, whereas it turns out $p_{j|k}$ only involves β but not α. Under the assumption that a_{kj} has iid type I extreme value distribution, we have

$$p(j|k) = \frac{\exp\left(x_{kj}\beta + z_k\alpha\right)}{\sum_{h=1}^{N_k}\exp\left(x_{kh}\beta + z_k\alpha\right)} = \frac{\exp\left(x_{kj}\beta\right)}{\sum_{h=1}^{N_k}\exp\left(x_{kj}\beta\right)} \tag{5.34}$$

where N_k is the number of house in neighborhood k. Similarly

$$p_k = \frac{\sum_{j=1}^{N_k}\exp(x_{kj}\beta + z_k\alpha)}{\sum_{m=1}^{C}\sum_{j=1}^{N_m}\exp(x_{mj}\beta + z_m\alpha)} \tag{5.35}$$

$$= \frac{\exp(I_k + z_k\alpha)}{\sum_{m=1}^{C}\exp(I_m + z_m\alpha)} \tag{5.36}$$

where $I_k = \log\sum_{h=1}^{N_k}\exp\left(x_{kh}\beta\right)$ is the so-called inclusive value (the total contribution of each house in a neighborhood). The expression for p_k may be derived from $p_{j|k}$ and p_{kj} where

$$p_{kj} = \frac{\exp\left(x_{kj}\beta + z_k\alpha\right)}{\sum_{m=1}^{C}\sum_{j=1}^{N_m}\exp(x_{mj}\beta + z_m\alpha)}, \tag{5.37}$$

which is obvious if we think each individual has $k \times j$ options.

One can therefore first estimate β off the choice within neighborhoods (based on $p(j|k)$) and then use the $\widehat{\beta}$ to impute $\hat{I}_k$ and estimate α by maximizing a likelihood consisting of p_k. This sequential estimation provides consistent estimates and can be applied in all problems in which the number of choice is very large but the decision process has a tree structure.

The extension of this model to cases involving several branches of a tree is obvious. See Maddala (Ch.3).

As multinomial/conditional logit model, the above nest logit model suffers from the IIA property. One way to avoid the problem is to assume a variance component structure for the random utility:

$$a_{kj} = \epsilon_k + \lambda_k\epsilon_{kj} \tag{5.38}$$

for some $\lambda_k \in [0, 1]$, where $\epsilon_{kj} \sim$ type I extreme value, $\epsilon_k \sim C(\lambda_k)$ and the $C(\lambda)$ distribution is defined to be the unique distribution for which v and e are independent, $v \sim C(\lambda)$, and $e \sim$ type I extreme value, implies that $v + \lambda e \sim$ type I extreme value; See Cardell (1997). In the variance component specification, ϵ_k is a common component for all houses in neighborhood k and ϵ_{kj} is a random/unobservable component for house j in neighborhood k. λ_k are parameters to be estimated. It measures the degree of independence in the unobserved utility among the alternatives in the same neighborhood. A higher value of λ means greater independence and less correlation. The correlation between a_{kj_1} and a_{kj_2} goes to zero as λ approaches one and goes to one as λ approaches zero. This can be easily seen by noting that the variance of a $C(\lambda)$ random variable is proportional to $(1-\lambda^2)$.

Given the variance component structure, the marginal and conditional probabilities can be written as

$$p_k = \frac{\exp(\lambda_k I_k + z_k\alpha)}{\sum_{m=1}^{C}\exp(\lambda_m I_m + z_m\alpha)} \tag{5.39}$$

$$p(j|k) = \frac{\exp\left(x_{kj}\beta/\lambda_k\right)}{\sum_{h=1}^{N_k}\exp\left(x_{kh}\beta/\lambda_k\right)} \tag{5.40}$$

where

$$I_k = \ln\sum_{h=1}^{N_k}\exp\left(x_{kh}\beta/\lambda_k\right) \tag{5.41}$$

The above probability can be derived by using the following lemma:

Lemma 42 *If ε_j is iid type I extreme value, j=0,1,...,J. Then for any constants k_j, $t_j = \max_{0\le\ell\le j}(k_\ell+\varepsilon_\ell) - \log\sum_{\ell=0}^{j}\exp(k_\ell)$ is type I extreme value*

Note that if we choose the k-th neighborhood, the utility derived from the houses in this neighborhood is

$$\begin{aligned} V_k &= \max_j\{x_{kj}\beta + \lambda_k\epsilon_{kj}\} + z_k\alpha + \epsilon_k \\ &= \lambda_k \max_j\left\{\frac{x_{kj}\beta}{\lambda_k} + \epsilon_{kj}\right\} + z_k\alpha + \epsilon_k \\ &= \lambda_k\left[\max_j\left\{\frac{x_{kj}\beta}{\lambda_k} + \epsilon_{kj}\right\} - \ln\sum_{h=1}^{N_k}\exp\left(x_{kh}\beta/\lambda_k\right)\right] \\ &\quad +\lambda_k\ln\sum_{h=1}^{N_k}\exp\left(x_{kh}\beta/\lambda_k\right) + z_k\alpha + \epsilon_k \qquad (5.42) \\ &= \lambda_k I_k + z_k\alpha + \eta_k, \end{aligned}$$

where

$$\eta_k = \epsilon_k + \lambda_k\left[\max_j\left\{\frac{x_{kj}\beta}{\lambda_k} + \epsilon_{kj}\right\} - \ln\sum_{h=1}^{N_k}\exp\left(x_{kh}\beta/\lambda_k\right)\right]$$

is a type I extreme value random variable. It now follows from (5.42) that equation (5.39) holds.

5.5 Ordered Probit and Logit Model

Example 43 *Ratings, opinion surveys, attained education level. '0' <'1' <'2' but '1'–'0'≠ '2' − '1'.*

Let y be an ordered response taking on the values $\{0, 1, 2, ...J\}$ for some known integer J. The ordered probit model can be derived from a latent variable model. Assume that a latent variable y^* defined by

$$y^* = x\beta + e, e|x \sim N(0, 1) \tag{5.43}$$

Let $c_1 < c_2 < ... < c_J$ be unknown cut points and define

$$y = \begin{cases} 0 & if\ y^* \leq c_1 \\ 1 & if\ c_1 < y^* \leq c_2 \\ ... & ... \\ J & if\ y^* > c_J \end{cases} \tag{5.44}$$

Given the standard normal assumption, we can compute each response probability:

$P(y = 0|x) = P(y^* < c_1) = P(x\beta + e < c_1) = \Phi(c_1 - x\beta)$

$P(y = 1|x) = P(c_1 < y^* \leq c_2) = P(c_1 < x\beta + e \leq c_2) = \Phi(c_2 - x\beta) - \Phi(c_1 - x\beta)$

...

$P(y = J|x) = P(c_J < y^*) = P(c_J < x\beta + e) = 1 - \Phi(c_J - x\beta)$

The parameters c and β can be estimated by maximum likelihood. For each i, the likelihood is

$$\begin{aligned} l_i(c, \beta) &= \{y_i = 0\} \log \Phi(c_1 - x_i\beta) + \{y_i = 1\} \log [\Phi(c_2 - x_i\beta) - \Phi(c_1 - x_i\beta)] \\ &\quad + ... + \{y_i = J\} \log [1 - \Phi(c_J - x_i\beta)] \end{aligned} \tag{5.45}$$

Other distribution functions can be used in place of Φ. Replacing Φ with the logit function, Λ, gives the ordered logit model:

The focus of interest is: $\partial P(y = 1|x)/\partial x_j$ and $c_i's$. Interpreting the coefficients based on their sign is not obvious in the ordered response model. See the textbook by Wooldridge (2002).

5.6 Poisson Regression Models

Consider the case when the discrete response Y_i is a non-negative integer. The typical situation is when Y_i counts the number of times that a certain event occurs during a specified time period. For example, Y_i may record the number of withdrawals from an ATM during a week, or the number of job offers received by an unemployed person during a month, or the number of patents applied for by a firm during a year.

The basic model for this kind of data is the Poisson distribution with natural parameter (equal to the mean and the variance of the distribution):

$$\Pr(Y_i = y|x_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}, y_i = 0, 1, 2, ... \tag{5.46}$$

the most common formulation for λ_i is the log-linear model

$$\ln \lambda_i = x_i\beta. \tag{5.47}$$

In this case

$$E(y_i|x_i) = var(y_i|x_i) = \lambda_i = \exp(x_i\beta). \tag{5.48}$$

The log-likelihood function is

$$\begin{aligned} \ln L &= \sum_{i=1}^{n} (-\lambda_i + y_i \ln \lambda_i - \ln y_i!) && (5.49) \\ &= \sum_{i=1}^{n} \left[-\exp(x_i\beta) + y_i x_i\beta - \ln y_i!\right]. && (5.50) \end{aligned}$$

The score is

$$\begin{aligned} \frac{\partial \ln L}{\partial \beta} &= \sum_{i=1}^{n} -x_i' \exp(x_i\beta) + x_i' y_i && (5.51) \\ &= \sum_{i=1}^{n} x_i' (y_i - \lambda_i) = 0, && (5.52) \end{aligned}$$

and the Hessian is

$$\frac{\partial^2 \ln L}{\partial \beta \partial \beta'} = -\sum_{i=1}^{n} \lambda_i x_i' x_i. \tag{5.53}$$

Using MLE theory, we can show that $\hat{\beta}_{ML}$ is asymptotically normal with variance

$$\left(-\sum_{i=1}^{n} \exp(x_i\widehat{\beta}) x_i' x_i\right)^{-1}, \tag{5.54}$$

which can be estimated by

$$\left(-\sum_{i=1}^{n} \lambda_i x_i' x_i\right)^{-1}. \tag{5.55}$$

Bibliography

[1] Cardell, N Scott (1997), "Variance Components Structures for the Extreme-Value and Logistic Distributions with Application to Models of Heterogeneity." *Econometric Theory*, 13(2)185-213

[2] Hausman, Jerry and Daniel McFadden (1984), "Specification Tests for the Multinomial Logit Model," *Econometrica*, 52 (September), 1219–1240.

[3] Kenneth Train, (2002), Discrete Choice Methods with Simulation, Cambridge University Press.

[4] Maddala, G.S., (1987), Limited Dependent and Qualitative Variables in Econometrics.

Chapter 6

Truncation and Censoring Models

6.1 Truncated Regression Model

6.1.1 The Model

Suppose that $\{y_i^*, x_i\}$ is iid and

$$y_i^* = x_i\beta + \varepsilon_i, \varepsilon_i|x_i \sim N(0, \sigma^2) \tag{6.1}$$

We only observe the y_i^* satisfying $y_i^* > c$ where c is a known constant, i.e.

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* > c \\ \text{no observation} & \text{if } y_i^* \leq c \end{cases} \tag{6.2}$$

6.1.2 Moments of Truncated Normal Variables

If a continuous random variable y has density function $f(y)$. The truncated variable has density

$$f(y|y > c) = \frac{f(y)}{\int_c^\infty f(y)dy} \tag{6.3}$$

To derive $E\left(y|x, y > c\right)$, we need the following fact: if $z \sim N(0,1)$ then

$$\begin{aligned} E\left(z|z > c\right) &= \int_c^\infty \frac{z\phi(z)}{1 - \Phi(c)}dz = \frac{1}{\sqrt{2\pi}}\int_c^\infty \frac{\exp(-z^2/2)}{1 - \Phi(c)}dz^2/2 \\ &= -\frac{1}{\sqrt{2\pi}}\frac{\exp(-z^2/2)}{1 - \Phi(c)}|_c^\infty = \frac{\phi(c)}{1 - \Phi(c)} := \lambda(-c) \end{aligned} \tag{6.4}$$

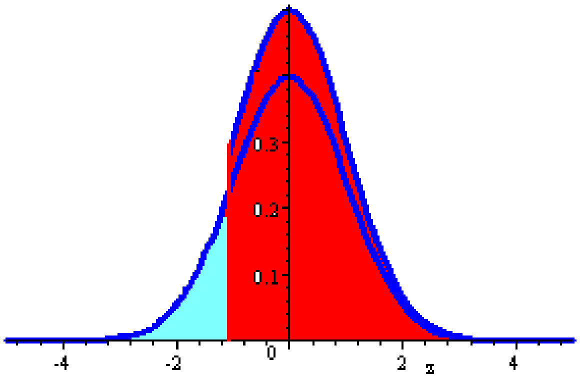

Figure 6.1: Truncated PDF

where

$$\lambda(c) = \frac{\phi(c)}{\Phi(c)}$$

is the so called Miller ratio (obviously $E\left(z|z<c\right) = \lambda(c)$). Therefore, if $u \sim N(0, \sigma^2)$,

$$E(u|u > c) = \frac{\sigma\phi(c/\sigma)}{1-\Phi(c/\sigma)} = \sigma\lambda(-\frac{c}{\sigma}). \tag{6.5}$$

Now we calculate $var\left(z|z>c\right)$. Note that

$$\begin{aligned} E\left(z^2|z>c\right) &= \int_c^\infty z^2\frac{\phi(z)}{1-\Phi(c)}dz = \frac{1}{\sqrt{2\pi}}\int_c^\infty \frac{z^2\exp(-z^2/2)}{1-\Phi(c)}dz \\ &= -\frac{1}{\sqrt{2\pi}}\frac{1}{1-\Phi(c)}\int_c^\infty zd\exp(-z^2/2) \\ &= \frac{1}{\sqrt{2\pi}}\frac{1}{1-\Phi(c)}c\exp(-c^2/2) + \frac{1}{\sqrt{2\pi}}\frac{1}{1-\Phi(c)}\int_c^\infty \exp(-z^2/2)dz \\ &= \frac{c\phi(c)}{1-\Phi(c)} + 1 = 1 + c\lambda(-c). \end{aligned} \tag{6.6}$$

So

$$\begin{aligned} var\left(z|z>c\right) &= \frac{c\phi(c)}{1-\Phi(c)} + 1 - \left(\frac{\phi(c)}{1-\Phi(c)}\right)^2 \\ &= 1 - \frac{\phi(c)}{1-\Phi(c)}\left(\frac{\phi(c)}{1-\Phi(c)} - c\right) \\ &= 1 - \lambda(-c)\left[\lambda(-c) - c\right] \end{aligned}$$

Therefore, if $u \sim N(0, \sigma^2)$, then

$$\begin{aligned} var(u|u\ >\ c) &= \sigma^2 var(\frac{u}{\sigma}|\frac{u}{\sigma} > \frac{c}{\sigma}) \\ &= \sigma^2 \left[1 - \frac{\phi(c/\sigma)}{1-\Phi(c/\sigma)} \left(\frac{\phi(c/\sigma)}{1-\Phi(c/\sigma)} - \frac{c}{\sigma}\right)\right] \qquad (6.7) \\ &= \sigma^2 - \sigma^2 \lambda(-\frac{c}{\sigma}) \left[\lambda(-\frac{c}{\sigma}) - \frac{c}{\sigma}\right]. \qquad (6.8) \end{aligned}$$

From the above analyses,

$$\begin{aligned} E\left(y_i|y_i > c\right) &= x_i\beta + E\left(\varepsilon_i|x_i\beta + \varepsilon_i > c\right) \\ &= x_i\beta + E\left(\varepsilon_i|\varepsilon_i > c - x_i\beta\right) \\ &= x_i\beta + \lambda\left[\left(x_i\beta - c\right)/\sigma\right] \qquad (6.9) \end{aligned}$$

and

$$\begin{aligned} & var\left(y_i|y_i > c\right) = var\left(\varepsilon_i|\varepsilon_i > c - x_i\beta\right) \\ = \ & \sigma^2 \left[1 - \frac{\phi\left[\left(c - x_i\beta\right)/\sigma\right]}{1 - \Phi\left[\left(c - x_i\beta\right)/\sigma\right]} \left(\frac{\phi\left[\left(c - x_i\beta\right)/\sigma\right]}{1 - \Phi\left[\left(c - x_i\beta\right)/\sigma\right]} - \frac{c - x_i\beta}{\sigma}\right)\right] \qquad (6.10) \end{aligned}$$

The above formulae show that the OLS estimate is not consistent because the second term in (6.9) is correlated with x_i. Since the functional form of this term is known, we can avoid the sample selection bias by using NLS, but MLE is preferred because it is asymptotically efficient.

In passing, we note that the sample selection bias does not arise if selection is based on regressors, not on the dependent variable.

6.1.3 Maximum Likelihood Estimation

The density is

$$f(y_i) = \frac{\sigma^{-1}\phi\left(\frac{y_i - x_i\beta}{\sigma}\right)}{1 - \Phi\left(\frac{c - x_i\beta}{\sigma}\right)}. \qquad (6.11)$$

The likelihood function is

$$\left\{-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2}\left(\frac{y_i - x_i\beta}{\sigma}\right)^2\right\} - \log\left[1 - \Phi\left(\frac{c - x_i\beta}{\sigma}\right)\right]. \qquad (6.12)$$

The usual asymptotics for MLE applies.

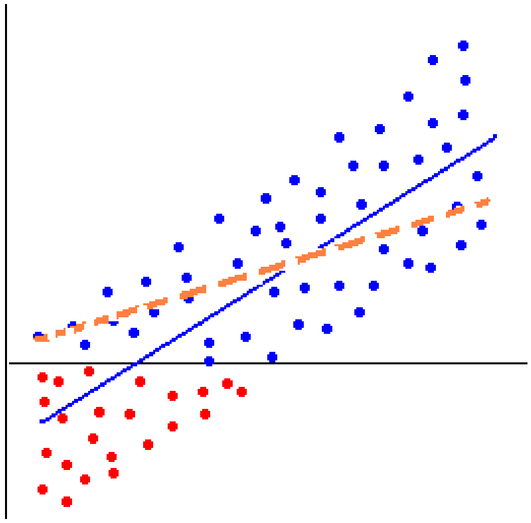

Figure 6.2: Inconsistency of the OLS estimator

6.2 Censored Regression (Tobit) Model

The general model is

$$y_i^* = x_i\beta + u_i, u_i|x_i \sim N(0, \sigma^2) \tag{6.13}$$

and $y_i = \max(0, y_i^*) = \max(0, x_i\beta + u_i)$.

6.2.1 Derivation of Expected Values

It is easy to see that

$$\begin{aligned} E(y|x) &= p(y=0) * 0 + p(y>0)E(y|x, y>0) \\ &= p(y>0)E(y|x, y>0), \end{aligned} \tag{6.14}$$

and

$$p(y>0) = p\left((x_i\beta + u_i) > 0\right) = \Phi\left(x\beta/\sigma\right). \tag{6.15}$$

Hence, in order to find $E(y|x)$, we only need to compute $E(y|x, y>0)$:

$$\begin{aligned} E(y|x, y &> 0) = E\left(x\beta + u|x, y>0\right) \\ &= x\beta + E\left(u|u > -x\beta\right) = x\beta + \frac{\sigma\phi(x\beta/\sigma)}{\Phi(x\beta/\sigma)} && (6.16) \\ &= x\beta + \sigma\lambda(x\beta/\sigma) && (6.17) \end{aligned}$$

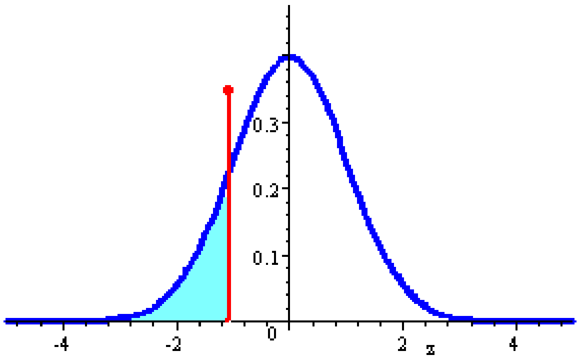

Figure 6.3: Censored PDF: Censoring Point c

As a consequence

$$E\left(y|x\right)=\Phi\left(\frac{x\beta}{\sigma}\right)x\beta+\sigma\phi\left(\frac{x\beta}{\sigma}\right) \tag{6.18}$$

and

$$\begin{aligned}\frac{\partial E\left(y|x\right)}{\partial x_j} &= \Phi\left(\frac{x\beta}{\sigma}\right)\beta_j+\phi\left(\frac{x\beta}{\sigma}\right)x\beta\frac{\beta_j}{\sigma}-\frac{\beta_j}{\sigma}x\beta\phi\left(\frac{x\beta}{\sigma}\right)\\ &= \Phi\left(\frac{x\beta}{\sigma}\right)\beta_j=p(y>0)\beta_j \qquad (6.19)\end{aligned}$$

One may also want to calculate

$$\frac{\partial E(y|x,y>0)}{\partial x_j}=\beta_j+\beta_j\frac{d\lambda}{dc}\left(\frac{x\beta}{\sigma}\right). \tag{6.20}$$

By differentiating $\lambda\left(c\right)$, we have

$$\begin{aligned}\frac{d\lambda}{dc} &= \frac{\phi'\left(c\right)}{\Phi(c)}-\frac{\phi^2\left(c\right)}{\Phi^2(c)}=-\frac{c\phi\left(c\right)}{\Phi(c)}-\frac{\phi^2\left(c\right)}{\Phi^2(c)}\\ &= -\lambda(c)\left[c+\lambda(c)\right]. \qquad (6.21)\end{aligned}$$

So

$$\begin{aligned}\frac{\partial E(y|x,y>0)}{\partial x_j} &= \beta_j-\beta_j\lambda(\frac{x\beta}{\sigma})\left[\frac{x\beta}{\sigma}+\lambda(\frac{x\beta}{\sigma})\right]\\ &= \beta_j\left(1-\lambda(\frac{x\beta}{\sigma})\left[\frac{x\beta}{\sigma}+\lambda(\frac{x\beta}{\sigma})\right]\right). \qquad (6.22)\end{aligned}$$

Note that

$$1-\lambda(\frac{x\beta}{\sigma})\left[\frac{x\beta}{\sigma}+\lambda(\frac{x\beta}{\sigma})\right] \tag{6.23}$$

is strictly between zero and one. Hence

$$\left|\frac{\partial E(y|x,y>0)}{\partial x_j}\right| \leq |\beta_j|.$$

6.2.2 Inconsistency of the OLS Estimator

From equation (6.16), we have

$$y_i = x_i\beta + \sigma\lambda\left(x_i\beta/\sigma\right) + e_i \tag{6.24}$$

with

$$E\left(e_i|x_i, y_i>0\right) = 0. \tag{6.25}$$

This implies that if we run the OLS of y_i on x_i using the sample for which $y_i > 0$, we effectively omit the variable λ. Due to the omitted variable bias, the OLS estimator is inconsistent. This is effectively a truncation regression with omitted variables

Even if we use all the data, the OLS estimator is still inconsistent because

$$E\left(y|x\right) = \Phi\left(\frac{x\beta}{\sigma}\right)x\beta + \sigma\phi\left(\frac{x\beta}{\sigma}\right) \tag{6.26}$$

6.2.3 Estimation and Inference with Censored Tobit

Let $\{x_i, y_i\}$ be a random sample following the censored Tobit model:

$$y_i^* = x_i\beta + u_i, u_i|x_i \sim N(0,\sigma^2), y_i = \max(0, y_i^*) \tag{6.27}$$

then the density of y_i given x_i is

$$f(y_i|x_i) = \frac{1}{\sigma}\phi\left(\frac{y_i - x_i\beta}{\sigma}\right)^{\{y_i>0\}}\left[1-\Phi\left(\frac{x_i\beta}{\sigma}\right)\right]^{\{y_i=0\}} \tag{6.28}$$

Let $\theta = \left(\beta', \sigma^2\right)'$, then the log-likelihood is

$$\begin{aligned} l_i\left(\theta\right) &= \{y_i=0\}\log\left[1-\Phi\left(\frac{x_i\beta}{\sigma}\right)\right] + \{y_i>0\}\log\left[\frac{1}{\sigma}\phi\left(\frac{y_i-x_i\beta}{\sigma}\right)\right] \qquad (6.29) \\ &= \{y_i=0\}\log\left[1-\Phi\left(\frac{x_i\beta}{\sigma}\right)\right] - \{y_i>0\}\left[\left(\frac{\left(y_i-x_i\beta\right)^2}{2\sigma^2}\right)+\frac{\log\left(\sigma^2\right)}{2}\right], \end{aligned}$$

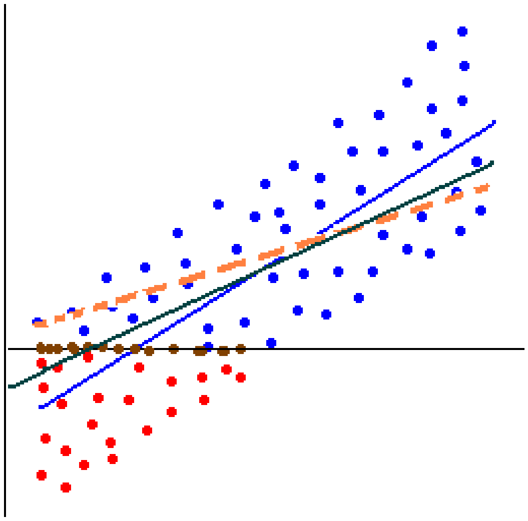

Figure 6.4: Inconsistency of the OLS estimator

which has a single maximum, but two step procedures have been devised by Heckman and Amemiya.

The two step procedure of Heckman starts with a probit on $y_i > 0$ or not. This delivers consistent β/σ. In the second step, bring in the continuous information and consider

$$E\left(y|x\right) = \Phi\left(\frac{x\beta}{\sigma}\right)x\beta + \sigma\phi\left(\frac{x\beta}{\sigma}\right). \tag{6.30}$$

Use the first-step $\widehat{\beta/\sigma}$ to predict $\Phi_i = \Phi\left(x_i\frac{\widehat{\beta}}{\sigma}\right)$ and $\phi_i = \phi\left(x_i\frac{\widehat{\beta}}{\sigma}\right)$ and estimate

$$y_i = \left(\Phi_i x_i\right)\beta + \sigma\phi_i + e_i \tag{6.31}$$

for a new set of β and σ.

Testing is easily carried out in a standard MLE framework: t-test, LR test, Wald test.

6.3 Specification Issues in Tobit Models

6.3.1 Neglect Heterogeneity

Suppose the model is

$$y = \max(0, x\beta + \gamma q + u),\ u|x,q \sim N(0,\sigma^2) \tag{6.32}$$

where q is assumed to be independent of x and has normal(0,τ^2) distribution. Then

$$y = \max(0, x\beta + v),\ \ v|x, q \sim N(0, \sigma^2 + \gamma^2\tau^2). \tag{6.33}$$

Therefore heterogeneity that is independent of x has no important consequence in data censoring examples.

Now suppose we want to estimate APE.

$$\begin{aligned} E_q \frac{\partial E(y|x,q)}{\partial x_j} &= E_q \left\{ \Phi\left(\frac{x\beta + q\gamma}{\sigma}\right) \beta_j \right\} \\ &= \Phi\left(\frac{x\beta}{\sqrt{\sigma^2 + \gamma^2\tau^2}}\right) \beta_j \end{aligned} \tag{6.34}$$

which is exactly the estimated partial effect from Tobit y_i on x_i. In other words, we can estimate the desired quantities — the APE's— by ignoring the heterogeneity.

6.3.2 Endogenous Explanatory Variables

Suppose the model is

$$\begin{aligned} y_1 &= \max(0, z_1\delta + \alpha_1 y_2 + u_1) \\ y_2 &= z\delta_2 + v_2 = z_1\delta_{21} + z_2\delta_{22} + v_2 \end{aligned} \tag{6.35}$$

where (u_1, v_2) are zero mean normally distributed, independent of z. For identification, we need the usual rank condition $\delta_{22} \neq 0$ and $E(z'z)$ is assumed to have full rank, as always.

Under the normality assumption, we have

$$u_1 = \theta_1 v_2 + e_1 \tag{6.36}$$

where

$$\theta_1 = \eta_1/\tau_2^2, \eta_1 = cov(u_1, v_2), \tau_2^2 = var(v_2) \tag{6.37}$$

and $e_1 \sim N(0, \tau_1^2)$ and is independent of (z, v_2). Plugging

$$u_1 = \theta_1 v_2 + e_1 \tag{6.38}$$

into

$$y_1 = \max(0, z_1\delta + \alpha_1 y_2 + u_1) \tag{6.39}$$

gives

$$y_1 = \max(0, z_1\delta + \alpha_1 y_2 + \theta_1 v_2 + e_1) \tag{6.40}$$

The Smith-Blundell procedure

(1) Run OLS of y_2 on z and get the residual $\hat{v}_2 = y_2 - z\hat{\delta}_2$.

(2) Estimate a standard Tobit of y_1 on z_1, y_2 and $\hat{v}_2$ to get consistent estimates of $\delta_1, \alpha_1, \theta_1$ and τ_1^2.

The usual t-statistic on $\hat{v}_2$ provides a simple test of the null $H_0 : \theta_1 = 0$, which says that y_2 is exogeneous. Note that when we compute the asymptotic variance, we need to account for the fact that this is a two step procedure.

A full MLE approach avoid the two-step estimation problem:

$$f(y_1, y_2|z) = f(y_1|y_2, z)f(y_2|z) \tag{6.41}$$

The density $f(y_2|z)$ is normal $\left(z\delta_2, \tau_2^2\right)$ and y_1 given (y_2, z) follows a **censored Tobit** with mean

$$z_1\delta_1 + \alpha_1 y_2 + \eta_1/\tau_2^2\,(y_2 - z\delta_2) \tag{6.42}$$

and variance $\tau_1^2 = \sigma_1^2 - \eta_1^2/\tau_2^2$ where $\sigma_1^2 = var(u_1)$. So $f(y_1, y_2|z)$ is

$$\left\{1 - \Phi\left(\frac{z_1\delta_1 + \alpha_1 y_2 + \eta_1/\tau_2^2\,(y_2 - z\delta_2)}{\tau_1}\right)\right\}^{\{y_1=0\}} \frac{1}{\sqrt{2\pi\tau_2^2}} \exp\left(-\frac{(y_2 - z\delta_2)^2}{2}\right)$$
$$\times \left\{\frac{1}{\sqrt{2\pi\tau_1^2}} \exp\left(-\frac{(y_1 - z_1\delta_1 - \alpha_1 y_2 - \eta_1/\tau_2^2\,(y_2 - z\delta_2))^2}{2}\right)\right\}^{\{y_1>0\}}$$

Once the MLE has been obtained, we can easily test the null hypothesis of exogeneity of y_2 using the t-statistic for $\hat{\theta}_1$.

Exercise 44 *What if* $y_2 = \{z\delta_2 + v_2 > 0\}$?

6.4 Panel Tobit Models

6.4.1 Pooled Tobit

A panel data model is

$$y_{it} = \max(0, x_{it}\beta + u_{it}); \;\; u_{it}|x_{it} \sim N(0, \sigma^2) \tag{6.43}$$

- The relationship between u_{it} and x_{is} is unspecified.
- u_{it} may be serially dependent.

The pooled ML estimator maximized the partial log-likelihood function

$$\sum_{i=1}^{N}\sum_{t=1}^{T} l_{it}\left(\beta,\sigma^2\right) \tag{6.44}$$

Computationally, we just apply Tobit to the data set as if it were on long cross section of size NT. However, without further assumption, a robust variance estimator is needed to account serial correlation in the scores. LR statistics based on the pooled Tobit is not generally valid except in the case of dynamically complete models.

6.4.2 Unobserved-effect Model under Strict Exogeneity

The model:

$$y_{it} = \max\left(0, x_{it}\beta + \alpha_i + u_{it}\right); \left(u_{it}|x_i,\alpha_i\right) \sim N(0,\sigma^2) \tag{6.45}$$

To allow for correlation between the individual specific effects and the explanatory variables, we follow Mundlak (1978) and Chamberlain (1980), explicitly model this correlation by assuming a specific parameterization of the individual specific effects as a function of the explanatory variables and random individual specific effects. This is often referred to as the conditional mean independence assumption (Wooldridge, 1995). A convenient and often made choice is to model the individual specific effects as a linear combination of the averages over time of the explanatory variables plus random individual specific effects. i.e.

$$\alpha_i|x_i \sim N(\psi + \bar{x}_i\xi, \sigma_a^2). \tag{6.46}$$

Under this assumption, our model becomes.

$$\begin{aligned} y_{it} &= \max\left(0, x_{it}\beta + \psi + \bar{x}_i\xi + a_i + u_{it}\right); \\ u_{it}|x_i,\alpha_i &\sim N(0,\sigma^2) \text{ and } a_i|x_i \sim N(0,\sigma_a^2) \end{aligned} \tag{6.47}$$

In addition, we assume that $u_{i1},...,u_{iT}$ are independent given x_i and a_i. Under the above assumptions, we have random effects Tobit model. The density for individual i is

$$f(y_i|x_i,\alpha) = C\int\prod_{t=1}^{T}\left[\frac{1}{\sigma}\phi\left(\frac{y_{it}-x_{it}\beta-\psi-\bar{x}_i\xi-a}{\sigma}\right)\right]^{\{y_{it}>0\}} \tag{6.48}$$

$$\times\left[1-\Phi\left(\frac{y_{it}-x_{it}\beta-\psi-\bar{x}_i\xi-a}{\sigma}\right)\right]^{\{y_{it}=0\}}\frac{1}{\sigma_a}\phi\left(\frac{a}{\sigma_a}\right)da(6.49)$$

for some constant C.

For corner solution applications, we can estimate either partial effect evaluated at $E\left(c\right)$ or APE's. As before, it is convenient to define

$$m(z,\sigma^2)=\Phi\left(z/\sigma\right)z+\sigma\phi\left(z/\sigma\right) \tag{6.50}$$

so that

$$E(y_t|x,\alpha)=m(x_t\beta+\alpha,\sigma_u). \tag{6.51}$$

A consistent estimator of the above quantity is $m\left(x_t\hat{\beta}+\hat{\psi}+\bar{x}_i\hat{\xi},\widehat{\sigma}_u^2\right)$. Estimating APE is also relatively easy. We need to calculate

$$\begin{aligned} Em(x^o\beta+\alpha_i,\sigma_u^2) &= EE\left[m(x^o\beta+\psi+\bar{x}_i\xi+a_i,\sigma_u^2)|x_i\right] \\ &= Em(x^o\beta+\psi+\bar{x}_i\xi,\sigma_u^2+\sigma_a^2) \end{aligned} \tag{6.52}$$

which can be estimated by

$$\frac{1}{N}\sum_{i=1}^{N}m(x^o\widehat{\beta}+\widehat{\psi}+\bar{x}_i\widehat{\xi},\widehat{\sigma}_u^2+\widehat{\sigma}_a^2) \tag{6.53}$$

Problem 45 *Does the assumption of conditional independency ($u_{i1},...,u_{iT}$ are independent given x_i and a_i) indispensable?*

6.4.3 Dynamic Unobserved Effects Tobit Model

The model:

$$y_{it}=\max\left(0,z_{it}\delta+\rho_1 y_{it-1}+c_i+u_{it}\right); \tag{6.54}$$

$$u_{it}|\left(z_{it},y_{it-1,}y_{it-2},...y_{i0},c_i\right)\sim N(0,\sigma_u^2) \tag{6.55}$$

The joint density of the model is

$$\int_{-\infty}^{\infty}\prod_{t=1}^{T}f(y_t|y_{t-1},...,y_1,y_0,z,c;\theta)h(c|y_0,z;\gamma)dc \tag{6.56}$$

A natural specification for $h(c|y_0,z;\gamma)$ is $N(\psi+\xi_0 y_0+\bar{z}\xi,\sigma_a^2)$.

Bibliography

[1] Chamberlain, Gary (1982). “Multivariate Regression Models for Panel Data.” *Journal of Econometrics* 18: 5-45.

[2] Chamberlain, Gary (1984). “Panel Data,” in Griliches and Intriligator (eds.), *Handbook of Econometrics*, Volume 2. Amsterdam: North-Holland.

[3] Heckman, J. (1976). “The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models”, *Annals of Economic and Social Measurement*, 5, 475-492.

[4] Mundlak, Y. (1978). “On the Pooling of Time Series and Cross Section Data”, *Econometrica,* Vol. 46, pp. 69-85.

[5] Smith and Blundell (1986). “An Exogeneity Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply,” *Econometrica* 54, 679-685.

[6] Wooldridge, J. (1995). “Selection Corrections for Panel Data Models Under Conditional Mean Independence Assumptions,” *Journal of Econometrics* 68, 115-132.

Chapter 7

Sample Selection and Attrition

The topic of sample selection or incidental truncation, has been the focus of an enormous volume of empirical and theoretical literature. It involves features of both truncated and censored models.

Example 46 *(Truncation based on Wealth): We are interested in estimating the effect of worker eligibility in a particular pension plan on family wealth*

$$wealth = \beta_1 + \beta plan + \beta_2 educ + \beta_3 age + \beta_4 income + u \tag{7.1}$$

However, we can only sample people with a net wealth less than $300,000, so the sample is selected on the basis of wealth. (People with net wealth more than $300,000 may not be willing to be interviewed because their time is very valuable)

Example 47 *(Wage Offer Function) Consider estimating a wage offer equation for people of working age. By definition, this equation is suppose to represent ALL people of working age. But we can only observe the wage offer for working people. We thus effectively select the sample on this basis.*

The first example can be analyzed using the truncation regression studied before. We now proceed to analyze the second example.

7.1 A Probit Selection Equation

7.1.1 Heckit Two-step Estimator

Interest lies in estimating $E(w_i^o|x_i)$ where w_i^o is the hourly wage offer for a randomly drawn individual i. If w_i^o is observed for everyone of working age, we would proceed in a standard regression framework. However, a potential sample selection problem arises because w_i^o is observed only for people who work.

Suppose we want to estimate a wage regression. The true model is

$$logw_i^o = x_{i1}\beta_1 + u_{i1} \tag{7.2}$$

where w_i^o is wage, x_{i1} is a vector of individual characteristics, with β being the associated vector of coefficients. Due the sample selection problem, the assumption of the classical regression model, namely $E[u_{i1}|x_{i1}] = 0$, is unlikely to hold. This is because a person who chooses to work may be particularly diligent or have other characteristics that make him more desirable as a worker, and $E[u_{i1}|x_{i1}, worker]$ may well be positive.

We now model the decision to work by a simple rule. We assume that everyone of working age has a reservation wage w_i^r. The person chooses to work only if

$$w_i^o \geq w_i^r. \tag{7.3}$$

We parametrize the reservation wage by

$$w_i^r = \exp\left(x_{i2}\beta_2 + \gamma_2 a_i + u_{i2}\right) \tag{7.4}$$

where x_{i2} contains variables that determine the marginal utility of leisure and income and a_i is the non-income wage of person i. We assume that (u_{i1}, u_{i2}) is independent of (x_{i1}, x_{i2}, a_i). Person i decides to work if

$$x_{i1}\beta_1 + u_{i1} > x_{i2}\beta_2 + \gamma_2 a_i + u_{i2} \tag{7.5}$$

or

$$x_i\delta + v_i > 0,\ x_i = (x_{i1}, x_{i2}, a_i), v_i = u_{i1} - u_{i2} \tag{7.6}$$

Remark 48 *If w_i^r is observed and is exogeneous, and x_{i1} is always available, then we would be in the censored regression framework.*

Remark 49 *If w_i^r is observed and is exogeneous, and x_{i1} is available only when w_i^o is available, then we would be in the truncated regression framework*

Remark 50 *Since w_i^r is not observed, we need a new framework.*

Let $y_1 = \log w^o$ and y_2 be the binary labor force participation indicator, then

$$y_1 = x_1\beta_1 + u_1 \tag{7.7}$$

and

$$y_2 = 1\left\{x\delta_2 + v_2 > 0\right\} \tag{7.8}$$

we discuss the estimation of the model under the following set of assumptions:

Assumption A:

(a) (x, y_2) is always observed, y_1 is observed only when $y_2 = 1$;

(b) (u_1, v_2) is independent of x with zero mean

(c) $v_2 \sim N(0,1)$ and

(d) $E(u_1|v_2) = \gamma_1 v_2$.

Amemiya (1985) calls the above model the Type II Tobit Model.

Note that

$$\begin{aligned} E\left(y_1|x, v_2\right) &= E\left(x_1\beta_1 + u_1|x, v_2\right) \\ &= x_1\beta_1 + \gamma_1 v_2 \end{aligned} \tag{7.9}$$

So when $\gamma_1 = 0$, we have

$$E\left(y_1|x, v_2\right) = x_1\beta_1. \tag{7.10}$$

Because y_2 is a function of (x, v_2), we obtain

$$E\left(y_1|y_2\right) = x_1\beta_1. \tag{7.11}$$

In other words, if $\gamma_1 = 0$, there is no sample selection problem, and β_1 can consistently estimated by OLS using the selected sample.

What if $\gamma_1 \neq 0$? We hope to calculate $E\left(y_1|x, y_2 = 1\right)$. Since $E(u_1|v_2) = \gamma_1 v_2$, we can write

$$u_1 = \gamma_1 v_2 + \eta \text{ with } E\left(\eta|v_2\right) = 0. \tag{7.12}$$

Therefore,

$$\begin{aligned} E\left(y_1|x, y_2 = 1\right) &= E\left(x_1\beta_1 + u_1|x, y_2 = 1\right) \\ &= x_1\beta_1 + E\left(u_1|x, x\delta_2 + v_2 > 0\right) \\ &= x_1\beta_1 + E\left(\gamma_1 v_2 + \eta|x, x\delta_2 + v_2 > 0\right) \\ &= x_1\beta_1 + \gamma_1 E\left(v_2|v_2 > -x\delta_2\right) \\ &= x_1\beta + \gamma_1\lambda\left(x\delta_2\right) \end{aligned} \tag{7.13}$$

The above equation makes it clear that an OLS regression of y_1 on x_1 using the selected sample omits the term $\gamma_1\lambda\left(x\delta_2\right)$ and generally leads to inconsistent estimate of β_1.

Following Heckman (1979), we can consistently estimate β_1 and γ_1 using the selected sample by regressing y_{i1} on $x_{i1}, \lambda\left(x_i\delta_2\right)$. The problem is that δ_2 is unknown. Fortunately, δ_2 can be consistently estimated using probit based on y_{i2}. This two step procedure is sometimes called Heckit.

To estimate the asymptotic variance of $\hat{\beta}$, we have to make a correction for the fact that we're not using $\lambda\left(x\delta_2\right)$ but only $\lambda\left(x\hat{\delta}_2\right)$ so that the error term contains the following:

$$\gamma_1\left[\lambda(x\delta_2)-\lambda\left(x\hat{\delta}_2\right)\right]=\gamma_1\frac{\partial\lambda(z)}{\partial z}x\left((\delta_2-\hat{\delta}_2\right) \tag{7.14}$$

evaluated at $z=x\delta_2$. More specifically, the second step regression is

$$y_{1i}=x_{1i}\beta+\gamma_1\lambda\left(x_i\hat{\delta}_2\right)+e_i$$

where

$$e_i=u_{1i}-E(u_{1i}|v_{2i}>-x_i\delta_2)+\gamma_1\left[\lambda(x_i\delta_2)-\lambda\left(x_i\hat{\delta}_2\right)\right]$$

To calculate the variance of e_i, we first compute

$$\begin{aligned}var(u_i|v_{2i} &> -x_i\delta_2)=\gamma_1^2var\left(v_{2i}|v_{2i}>-x_i\delta_2\right)\\ &= \gamma_1^2(1-\lambda(x_i\delta_2)\left[\lambda(x_i\delta_2)+x_i\delta_2\right].\end{aligned}$$

Then

$$\begin{aligned}var(e_i) &= var(u_1|v_2>-x_i\delta_2)+\gamma_1^2\left(\lambda'(x_i\delta_2)^2 asyvar\left(x_i\left(\delta_2-\hat{\delta}_2\right)\right)\\ &= \gamma_1^2(1-\lambda(x_i\delta_2)\left[\lambda(x_i\delta_2)+x_i\delta_2\right]+\gamma_1^2\left(\lambda'(x_i\delta_2)^2 asyvar\left(x_i\left(\delta_2-\hat{\delta}_2\right)\right)\end{aligned}$$

Given this, the second step regression becomes a linear regression with heteroscedastic errors ($var(e_i)$). The variance of $\hat{\beta}$ can then be estimated using robust variance.

Remark 51 *We can use the Heckit to test $H_0:\gamma_1=0$.*

7.1.2 Partial Maximum Likelihood Estimation

To get more efficient estimate, we use MLE. Assume (u_1,v_2) are bivariate normal with mean zero and variance-covariance matrix

$$\begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{21} & 1\end{pmatrix}.$$

The pdf of (y_1,y_2) can be written as

$$f(y_1,y_2|x)=f(y_2|y_1,x)f(y_1|x).$$

Obviously

$$f(y_1|x)=\frac{1}{\sigma_1}\phi\left(\frac{(y_1-x_1\beta)}{\sigma_1}\right)$$

Note that

$$v_2 = \frac{\sigma_{12}}{\sigma_1^2} u_1 + \xi, \tag{7.15}$$

where

$$E\left(\xi|u_1\right) = 0, var(\xi|u_1) = 1 - \alpha_2^2 var(u_1) = 1 - \sigma_{12}^2\sigma_1^{-2}.$$

Therefore, conditional on u_1, v_2 is normal with mean $\sigma_{12}\sigma_1^{-2}u_1$ and variance $1-\sigma_{12}^2\sigma_1^2$. The probability of $y_2 = 1$ conditional on y_1 can then be written as

$$\Phi\left(\frac{x_2\delta_2 + \sigma_{12}\sigma_1^{-2}(y_1 - x_1\beta)}{\sqrt{1-\sigma_{12}^2\sigma_1^{-2}}}\right)$$

In view of the above analysis, $f(y_1, y_2|x)$ is

$$\left\{\frac{1}{\sigma_1}\phi\left(\frac{y_1 - x_1\beta}{\sigma_1}\right)\Phi\left(\frac{x_2\delta_2 + \sigma_{12}\sigma_1^{-2}(y_1 - x_1\beta)}{\sqrt{1-\sigma_{12}^2\sigma_1^2}}\right)\right\}^{y_2}\left[1 - \Phi\left(x_2\delta_2\right)\right]^{1-y_2}$$

Example 52 *Consider the hours labor-supply regression with wages on the RHS. First, you need to correct the hours equation for sample selection into labor force (only observe h for those who work). This correction comes from a comparison of behavior equations governing reservation wages w_i^r and market wages w_i^o which leads to a 0/1 participation estimation depending on $x_i\delta$, where x_i is the collection of RHS variables from both w_i^r and market wages w_i^o. Second, we need to instrument for w_i^o which is likely endogenous. The first stage regression where you predict w_i also needs to have a selection correction in it. Finally, you can estimate*

$$h_i = \alpha\widehat{w}_i + x_1\beta_1 + \gamma_1\lambda(x_i\delta_2) + \epsilon_i. \tag{7.16}$$

There is serious need for exclusion restrictions: you need an exclusion restriction for running IV for w_i (that is a variable predicting wages but not hours) and you need another exclusion restriction to identify the selection correction in the first-stage wage equation (that is you need a variable affecting participation, but not wages).

7.1.3 Endogenous Explanatory Variables

The model:

$$\begin{aligned} y_1 &= z_1\delta_1 + \alpha_1 y_2 + u_1 \\ y_2 &= z\delta_2 + v_2 \\ y_3 &= 1\left\{z\delta_3 + v_3\right\} \end{aligned} \tag{7.17}$$

In the wage estimation context, y_2 may be years of schooling.

Assumptions B:

(a) (z, y_3) is always observed, (y_1, y_2) is observed when $y_3 = 1$

(b) (u_1, v_3) is independent of z

(c) $v_3 \sim N(0, 1)$

(d) $E(u_1|v_3) = \gamma_1 v_3$

(e) $E(z' v_2) = 0$ and writing $z\delta_2 = z_1\delta_{21} + z_2\delta_{22}, \delta_{22} \neq 0.$

To derive an estimating equation, we write

$$y_1 = z_1\delta_1 + \alpha_1 y_2 + g(z, y_3) + e_1 \tag{7.18}$$

where $g(z, y_3) = E(u_1|z, y_3) = \gamma_1 \lambda (z\delta_3)$ and $e_1 = u_1 - E(u_1|z, y_3)$. By definition, $E(e_1|z, y_3) = 0$.

Now we have the following procedure

(a) Obtain $\hat{\delta}_3$ from probit of y_3 on z using all observations. Obtain the estimated inverse Miller ratio $\hat{\lambda}_{i3} = \lambda\left(z_i\hat{\delta}_3\right)$

(b) Using the selected subsample, estimate the following equation

$$y_{i1} = z_{i1}\delta_1 + \alpha_1 y_{2i} + \gamma_1 \hat{\lambda}_{i3} + error_i \tag{7.19}$$

by 2SLS using instrument $\left(z_i, \hat{\lambda}_{i3}\right)$.

7.2 A Tobit Selection Equation

The model:

$$\begin{aligned} y_1 &= x_1\beta_1 + u_1 \\ y_2 &= \max(0, x\delta_2 + v_2) \end{aligned} \tag{7.20}$$

A familiar example occurs when y_1 is the log of the hourly wage offered and y_2 is hours of labor supply. The model is sometimes called the type III Tobit model

Assumptions C

(a) (x, y_2) is always observed, but y_1 is observed only when $y_2 > 0$

(b) (u_1, v_2) is independent of x

(c) $v_2 \sim N(0, \tau_2^2)$

(d) $E(u_1|v_2) = \gamma_1 v_2$

To estimate the model, we hope to derive

$$\begin{aligned} E\left(y_1 | y_2 > 0\right) &= x_1\beta_1 + E\left(u_1 | v_2\right) \\ &= x_1\beta_1 + \gamma_1 v_2 \end{aligned} \tag{7.21}$$

From this, we naturally propose the following two step procedure

(a) Estimate $y_2 = \max(0, x\delta_2 + v_2)$ by standard Tobit using the whole sample. For $y_{i2} > 0$, define

$$\hat{v}_{i2} = y_{i2} - x_i\hat{\delta}_2 \tag{7.22}$$

(b) Using observations for which $y_{i2} > 0$, estimate β_1 and γ_1 by the OLS regression

$$y_{i1} \text{ on } x_{i1} \text{ and } \hat{v}_{i2}. \tag{7.23}$$

This regression produces consistent, $\sqrt{N}$ asymptotically normal estimators of β_1 and γ_1 under assumptions C.

For partial MLE, we assume that (u_1, v_2) is jointly normal. The partial log-likelihood function for individual i is

$$\left(\frac{1}{\tau_2}\phi\left(\frac{y_2 - x\delta_2}{\tau_2}\right)\right)^{\{y_2>0\}} \left[1 - \Phi\left(\frac{x\delta_2}{\tau_2}\right)\right]^{\{y_2=0\}}$$
$$\times \left\{\frac{1}{\sqrt{\sigma_1^2 - \sigma_{12}^2\tau_2^{-2}}}\phi\left(\frac{y_1 - x_1\beta_1 - \sigma_{12}\tau_2^{-2}(y_2 - x\delta_2)}{\sqrt{\sigma_1^2 - \sigma_{12}^2\tau_2^{-2}}}\right)\right\}^{\{y_2>0\}}$$

7.3 Sample Selection and Panel Attrition

In our treatment of panel data, we have assumed that the data set is balanced. Often, some time periods are missing for some units in the population of interest, and we are left with the unbalanced panel. If the unbalanced problem exogenously, then the unbalanced panels are fairly easy to deal with. A more complicated problem arises when attrition from a panel is due to units electing to drop out. If this decision is based on factors that are systematically related to the response variable, a sample selection problem can result.

7.3.1 Fixed Effects Estimation with Unbalanced Panels

We investigate the case that the usual fixed effects estimator on the unbalanced panel is consistent. The model is

$$y_{it} = x_{it}\beta + \alpha_i + u_{it} \tag{7.24}$$

where x_{it} is $1 \times k$ and α_i is allowed to be correlated with x_{it} so that x_{it} contains only time varying variables.

Let $t = 1$ as the first time period for which data on any one in the population are available and $t = T$ as the last possible period. For a random draw i from the population, let

$$s_i = (s_{i1}, s_{i2}, ..., s_{iT})' \tag{7.25}$$

denotes the $T \times 1$ vector of selection indicators: $s_{it} = 1$ if (x_{it}, y_{it}) is observed, and zero otherwise.

We can treat $\{(x_i, y_i, s_i)\}$ as a random sample from the population. The fixed effect estimator is

$$\begin{aligned}\hat{\beta} &= \left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{x}_{it}'\ddot{x}_{it}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{x}_{it}'\ddot{y}_{it}\right) \\ &= \beta + \left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{x}_{it}'\ddot{x}_{it}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{x}_{it}'\ddot{u}_{it}\right) \qquad (7.26)\end{aligned}$$

where

$$\ddot{x}_{it} = x_{it} - \frac{1}{T_i}\sum_{r=1}^{T} s_{ir}x_{ir},\ \ddot{y}_{it} = y_{it} - \frac{1}{T_i}\sum_{r=1}^{T} s_{ir}y_{ir}, T_i = \sum_{r=1}^{T} s_{ir} \tag{7.27}$$

Assumption (unbalanced):

(a) $E(u_{it}|x_i, s_i, \alpha_i) = 0$ for $t = 1, 2, ..., T$ (Strong Exogeneity)

(b) $\sum_{t=1}^{T} Es_{it}\ddot{x}_{it}'\ddot{x}_{it}$ is non-singular

(c) $E(u_i u_i'|x_i, s_i, c_i) = \sigma_u^2 I_T$

Remark 53 *Under the above assumptions (a) and (b), $\hat{\beta}$ is consistent.*

Remark 54 *In the case of a randomly rotating panel, and in other case where selection is entirely random, s_i is independent of (u_i, x_i, c_i), in which case, Assumption (a) follows from the usual fixed effect assumption $E(u_{it}|x_i, \alpha_i) = 0$*

Remark 55 *Assumption (a) holds under much weaker conditions. It does not assume any relationship between s_i and $(x_i,\ \alpha_i)$. In particular, if the selection in all periods is correlated with α_i or x_i, but u_{it} is mean independent of s_i given $(x_i,\ \alpha_i)$ for all t, then FE is consistent and asymptotically normal.*

With assumption (c), the asymptotically variance can be shown to be

$$\sigma_u^2 \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \ddot{x}_{it}' \ddot{x}_{it} \right)^{-1} \tag{7.28}$$

where σ_u^2 can be consistently estimated by

$$\widehat{\sigma}_u^2 = \frac{1}{\sum_{i=1}^{N} (T_i - 1)} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \hat{u}_{it}^2 \tag{7.29}$$

A correction for the degree of freedom can also be implemented.

7.3.2 Testing the Presence of Sample Selection Bias

Idea: Add lagged selection indicator $s_{i,t-1}$ to equation and estimate the model by fixed effects on the unbalanced panel. Test the significance of $s_{i,t-1}$.

Putting $s_{i,t-1}$ doe not work if $s_{i,t-1} = s_{it}$ because there is no variation in $s_{i,t-1}$ in the selected sample.

7.3.3 Attrition

At $t = 1$ a random sample is obtained from from the relevant population. In $t = 2$ and beyond, some people drop out of the sample for reasons that may not entirely random. We assume that once a person is out, she or he is out forever. Therefore, if $s_{it} = 1$ then $s_{ir} = 1$ for all $r < t$.

As before, we may add $s_{i,t+1}$ to the regression to test whether there is a sample selection problem.

The sequential nature of the attrition makes FD a natural choice to remove the unobserved effect:

$$\Delta y_{it} = \Delta x_{it} \beta + \Delta u_{it} \tag{7.30}$$

conditional on $s_{i,t-1} = 1$, write a reduced selection equation for $t \geq 2$ as

$$s_{it} = \{ w_{it} \delta_t + v_{it} > 0 \}, v_{it} | \{ w_{it}, s_{it-1} = 1 \} \sim N(0,1) \tag{7.31}$$

If x_{it} is strictly exogenous and selection does not depend on Δx_{it} once w_{it} is controlled for, a reasonable assumption is that

$$E (\Delta u_{it} | \Delta x_{it}, w_{it}, v_{it}, s_{i,t-1}) = E (\Delta u_{it} | v_{it}) = \rho_t v_{it} \tag{7.32}$$

Then

$$E (\Delta y_{it} | \Delta x_{it}, w_{it}, v_{it}, s_{i,t-1}) = \Delta x_{it} \beta + \rho_t \lambda (w_{it} \delta_t) \tag{7.33}$$

A Two-step Procedure

(a) Run $T-1$ cross section probit to get estimate of $\widehat{\delta}_2, \widehat{\delta}_3, ..., \widehat{\delta}_T$, construct $\widehat{\lambda}_{it} = \lambda\left(w_{it}\widehat{\delta}_t\right)$

(b) Run pooled OLS regression of Δy_{it} on Δx_{it}, $d2_t\widehat{\lambda}_{it}$, $d3_t\widehat{\lambda}_{it}, ..., d2_T\widehat{\lambda}_{it}$

Potential Problems:

(a) x_{it} is assumed to be strictly exogenous

(b) x_{it} does not affect the attrition once w_{it} is controlled for.

Let z_{it} be a vector of variables such that z_{it} is redundant in the selection equation and that z_{it} is exogenous so that

$$E\left(\Delta u_{it}|z_{it}, w_{it}, v_{it}, s_{i,t-1}\right) = E\left(\Delta u_{it}|v_{it}\right) \tag{7.34}$$

For example, z_{it} should contain x_{ir} for $r < t$. Then we can estimate the equation

$$\Delta y_{it} = \Delta x_{it}\beta + \rho_1 d2_t\widehat{\lambda}_{it} + \rho_2 d3_t\widehat{\lambda}_{it} + ... + \rho_T d2_T\widehat{\lambda}_{it} \tag{7.35}$$

by instrumental variable with instruments $\left(z_{it}, d2_t\widehat{\lambda}_{it}, ..., d2_T\widehat{\lambda}_{it}\right)$ using the selected sample.

The attrition bias can be tested by a joint test: $H_0 : \rho_2 = \rho_3 = ... = \rho_T = 0$.

The method for attrition and selection just described apply only to linear models.

Bibliography

[1] Amemiya. T. (1985). Advanced Econometrics, Cambridge, Harvard University Press.

[2] Heckman, J. (1979). “Sample selection bias as a specification error,” Econometrica, 47, 153–161.

[3] Heckman, J. (1990). “Varieties of Sample Selection”, American Economic Review, 80, 313-318.

[4] Kyriazidou, E. (1997). “Estimation of a Panel Data Sample Selection Model,” Econometrica, 65(6), 1335-1364.

[5] Manski, C. (1989). “Anatomy of the Selection Problem”, Journal of Human Resources, 24, 343-360.

[6] Newey,W., J. Powell, and J. Walker (1990). “Semiparametric Estimation of Selection Models,” American Economic Review, 80, 313-318.