Micro-Econometrics: Limited Dependent Variables and Panel Data

Outline of the course

Andrea Ichino
(European University institute, IGIER and CEPR) *

January 13, 2003

The course is intended to introduce students to some standard methods specifically designed for the analysis of particular types of microeconomic data. For each method the general theoretical background will be provided, followed by the critical discussion of one or more applied papers.

Household datasets for different countries in STATA format (plus some documentation) can be downloaded via web for exercises and problem sets. Look at:
http://www.iue.it/LIB−STAGE/DataSets/DATA/courses/ichino/metric_spring/

The final exam will be a take home exercise involving also a part to be answered at the computer with real data.

In addition to these lecture notes, the following textbooks are suggested: J. Wooldridge, Econometric Analysis of Cross Section and Panel Data, The MIT Press, last edition; J. Wooldridge, Introductory Econometrics, South Western College Publishing, last edition; W.H. Greene, Econometric Analysis, last edition. Prentice Hall 1997; ;J. Johnston, J. DiNardo, Econometric Methods, McGraw-Hill, last edition . Additional references are listed at the end of these notes.

*Address correspondence to: Andrea Ichino, European University Institute, 50016 San Domenico di Fiesole, Firenze, Italia, e-mail:ichino@iue.it. The lecture notes for this course can be downloaded from: http://www.iue.it/Personal/Ichino/Welcome.html

Contents

1 Introduction **4**

2 Binary choices **5**
2.1 Theory . . . 5
2.1.1 Basic framework and notation . . . 5
2.1.2 Linear probability model . . . 6
2.1.3 Non-linear probability model and the latent index function . . . 8
2.1.4 Estimation of non-linear probability models . . . 9
2.1.5 Goodness of fit . . . 12
2.1.6 Probit model . . . 13
2.1.7 Logit model . . . 16
2.1.8 Comparison between linear probability, probit and logit models . . . 19
2.1.9 Maximum score estimator . . . 21
2.2 Applications: binary choices models for the identification of social effects . . . 22

3 Multiple choices **27**
3.1 Theory . . . 27
3.1.1 Basic framework and notation . . . 27
3.1.2 The logit model. . . . 28
3.1.3 Independence from Irrelevant Alternatives Property (IIA) . . . 29
3.1.4 Which parameters are identified in the logit model? . . . 30
3.1.5 The multinomial logit model . . . 32
3.1.6 The (pure) Conditional Logit model . . . 38
3.1.7 A test for the IIA hypothesis . . . 41
3.2 Applications of multiple choices models . . . 44

4 Panel data **45**
4.1 Examples . . . 45
4.2 Problems arising in cross sections and solved by panel data . . . 46
4.2.1 Example 1: Production functions and managerial ability . . . 46
4.2.2 Example 2: Returns to schooling, ability and twins . . . 48
4.3 A general framework and more notation . . . 50
4.4 Fixed effects (within) estimators . . . 52
4.4.1 Least squares dummy variable model (LSDV) . . . 52
4.4.2 Analysis of Covariance: using deviations from individual specific means 54
4.4.3 A parenthesis on partitioned regressions . . . 55
4.4.4 Back to the Analysis of Covariance . . . 58
4.4.5 First differences . . . 61
4.4.6 Differences-in-Differences (DD) strategies . . . 62
4.4.7 Fixed effects estimators and measurement error . . . 67
4.4.8 Fixed effects estimators and lagged dependent variables . . . 72

4.4.9 Other pitfalls of fixed effects estimation 79
4.5 Between estimator 80
4.5.1 OLS, "within" and "between" estimators 82
4.6 Random effects estimator 84
4.6.1 GLS estimation of Random Effects models 86
4.6.2 Feasible GLS estimation of random effects models 87
4.6.3 Random effects, within, between and OLS estimators 88
4.7 Mundlak (1978): a reconciliation of fixed and random effects models? 90
4.7.1 A test for random or fixed effects 92
4.7.2 Random effects models and Instrumental Variables 95
4.8 Extensions 100
4.9 Panel data analysis in STATA 100

5 Panel data with discrete dependent variables 101
5.1 The conditional maximum likelihood approach 102
5.2 Fixed effects conditional logit estimation in STATA 105
5.3 Applications 105

6 References 106

1 Introduction

In labor economics and more generally in the analysis micro-economic datasets we have often to deal with phenomena that are intrinsically discrete or that are measured in a discrete fashion.

- End-of-highschool decision: go to college or drop out.
- Female's decision to participate in the labor market.
- Employment or unemployment after training.
- Self-employment of wage work.
- Welfare participation.
- Consumer choices.
- Means of transportation.
- Marriage.
- Crime.
- Voting.
- Locational decisions of firms.
- Entering the EU.
- ...

It is convenient to distinguish between:

- *Binary choices*: the dependent variable can take two values.
- *Multiple choices*: the dependent variable can take more than two values.

2 Binary choices

2.1 Theory

2.1.1 Basic framework and notation

Consider a sample of individuals indexed by $i = \{1, 2, 3, ...N\}$.

For each individual we observe the binary variable:

$$Y = \begin{cases} 1 & \text{with probability} \quad Pr(Y = 1) = P \\ 0 & \text{with probability} \quad Pr(Y = 0) = 1 - P \end{cases} \tag{1}$$

Let X be the row vector of K potential factors (including the constant) that explain which outcome prevails. For individual i we observe the vector X_i.

We denote matrices in bold face characters. So $\mathbf{X}$ is the $N \times K$ matrix of observations on the K explanatory factors for the N individuals.

Our objective is to estimate the effect of the factors X on the probability of observing the outcome $Y = 1$:

$$\gamma = \frac{dP}{dX'}. \tag{2}$$

where γ is a column vector of K marginal effects

Note that:

$$E(Y) = 1P + 0(1 - P) = P \tag{3}$$

2.1.2 Linear probability model

The linear probability model assumes that P is a linear function of X:

$$P = F(X, \beta) = X\beta \tag{4}$$

where β is a column vector of K parameters and X includes a constant term.

Using this assumption and equation 3:

$$\begin{aligned} Y &= E(Y) + (Y - E(Y)) \\ &= P + (Y - E(Y)) \\ &= X\beta + \epsilon \end{aligned} \tag{5}$$

where

$$\epsilon = \begin{cases} 1 - X\beta & \text{with probability} \quad P \\ -X\beta & \text{with probability} \quad 1 - P \end{cases} \tag{6}$$

The marginal effect of X on the P is therefore:

$$\gamma = \frac{dP}{dX'} = \beta \tag{7}$$

which we can estimate using OLS in equation 5.

Advantages:

- Computational simplicity.
- Very little structure or assumptions imposed on the data.

Disadvantages:

i. *Heteroschedasticity*
The mean of the error term is zero by construction:

$$\begin{aligned} E(\epsilon) &= (1 - X\beta)P + (-X\beta)(1 - P) \\ &= (1 - X\beta)X\beta + (-X\beta)(1 - X\beta) = 0. \end{aligned} \tag{8}$$

However, the variance is given by:

$$\begin{aligned} E(\epsilon^2) &= (1 - X\beta)^2 X\beta + (-X\beta)^2(1 - X\beta) \\ &= (1 - X\beta)X\beta \end{aligned} \tag{9}$$

which shows that the error term is heteroschedastic. Observations for which $P_i = X_i\beta$ is close to 1 or 0 have relatively low variance while observations with $P_i = X_i\beta$ close to .5 have relatively high variance.

- Note that it is not advisable to use GLS because of next problem

ii. *Predicted probabilities $\hat{P}_i = X_i\hat{\beta}$ may lie outside the [0,1] range*
This may produce non-sense probabilities for forecasting purposes and negative estimated variances so that GLS cannot be implemented.

iii. *Estimates of β are fairly sensitive to extreme realizations of X*

iv. *Hypothesis testing*

2.1.3 Non-linear probability model and the latent index function

To avoid the problem of out-of-range probabilities in the linear probability model, we can assume that:

$$P = Pr(Y = 1) = F(X\beta) \tag{10}$$

where F is a (symmetric) cumulative distribution such that:

$$\begin{aligned} \lim_{X\beta \to +\infty} F(X\beta) &= 1 \\ \lim_{X\beta \to -\infty} F(X\beta) &= 0 \end{aligned} \tag{11}$$

One way to introduce this assumption is to consider an unobservable index function which determines the value of the binary outcome (note that the choice of the threshold is irrelevant):

$$Y^* = X\beta + \epsilon \tag{12}$$

$$Y = \begin{cases} 1 & \text{if } Y^* \geq 0 \\ 0 & \text{if } Y^* < 0 \end{cases} \tag{13}$$

Assume that ϵ is distributed according to F:

$$Y = \begin{cases} 1 & \text{with} \quad Pr(\epsilon \geq -X\beta) = F(X\beta) \\ 0 & \text{with} \quad Pr(\epsilon < -X\beta) = 1 - F(X\beta) \end{cases} \tag{14}$$

Given these assumptions, the marginal effect of X on P is:

$$\gamma = \frac{dP}{dX'} = F'\beta = f\beta \tag{15}$$

where f is the density function of F. Note that F' and f are scalar functions of $X\beta$. In contrast with the linear probability model, an estimate of β is not enough to estimate the marginal effect: γ has to be evaluated at some realization of X. (See below page 11.)

2.1.4 Estimation of non-linear probability models

Using the Maximum Likelihood approach, the likelihood function is:

$$\begin{aligned} L &= Pr(Y_1 = y_1, Y_2 = y_2, ...Y_N = y_N) && (16) \\ &= \prod_{y_i=0} [1 - F(X_i\beta)] \prod_{y_i=1} F(X_i\beta) \\ &= \prod_{i=1}^{N} [1 - F(X_i\beta)]^{1-y_i} \; F(X_i\beta)^{y_i} && (17) \end{aligned}$$

where $y_i = \{0, 1\}$ is the realization of the binary outcome Y_i.

Taking logs:

$$ln(L) = \sum_{i=1}^{N} [(1 - y_i) \; ln(1 - F(X_i\beta)) + y_i \; ln(F(X_i\beta))] \tag{18}$$

The first order conditions for the maximization are:

$$\frac{\partial ln(L)}{\partial \beta'} = \sum_{i=1}^{N} \left[\frac{y_i f(X_i\beta)}{F(X_i\beta)} + \frac{-(1 - y_i) f(X_i\beta)}{1 - F(X_i\beta)} \right] X_i = 0 \tag{19}$$

The solution of this system gives the vector of ML estimates $\hat{\beta}$.

The asymptotic covariance matrix $\mathbf{V}$ of the β is the inverse of the Hessian:

$$\mathbf{V} = -\mathbf{H}^{-1} = \left(\frac{\partial^2 ln(L)}{\partial \beta \partial \beta'} \right)^{-1} \tag{20}$$

which is a $K \times K$ matrix.

Coefficients, probabilities and marginal effect

In the linear probability model the coefficients β coincide with the marginal effect of the factors X on P.

In the non-linear latent index model the coefficients β represent just the marginal effect of the factors X on the unobservable index Y^*, which may not say much.

We are interested in estimates of:

The probability of the outcome:

$$Prob(Y = 1) = P = F(X\beta) \tag{21}$$

$$\text{Asy. Var } [P] = \left[\frac{\partial F}{\partial \beta'}\right] \ \mathbf{V} \ \left[\frac{\partial F}{\partial \beta'}\right]' = f^2 X \mathbf{V} X' \tag{22}$$

which is a scalar.

The marginal effects:

$$\gamma = \frac{dP}{dX'} = f\beta \tag{23}$$

$$\text{Asy. Var } [\gamma] = \left[\frac{\partial \gamma}{\partial \beta'}\right] \ \mathbf{V} \ \left[\frac{\partial \gamma}{\partial \beta'}\right]' \tag{24}$$

which is a $K \times K$ matrix.

See the rules and notation for matrix differentiation in Greene (1997). Note that f is a function of $X\beta$; hence, to estimate the probability of the outcome and the marginal effects we need an estimate of β and some realization of X.

At which X should we evaluate the estimates of P and γ?

We can compute $\hat{P}$ and $\hat{\gamma}$:

i. for each i and then take the averages over all the observations;

ii. for the sample mean of the observations X_i.

iii. for a particularly relevant observation (median, other percentiles).

iv. for an artificially created individual with values of X defined by us.

Note that solutions 1 and 2 are asymptotically equivalent but may differ in small samples.

Marginal effects of dummy variables

The marginal effect of a dummy should be computed as the difference between the estimated probabilities evaluated at the two values of the dummy (keeping the other X constant).

2.1.5 Goodness of fit

An analog of the R^2 is the log-likelihood ratio index.

$$LRI = 1 - \frac{ln\ L}{ln\ L_0} \tag{25}$$

where $ln\ L_0$ is the value of the log-likelihood computed with only a constant term. This is sometime called Pseudo-R^2.

However, it may be misleading because $LRI = 1$ only when $X_i\beta$ explodes to $+\infty$ or $-\infty$, which may actually be indicative of a flaw of the model.

A model may tell us that an increase in X significantly increases the $Pr(Y = 1)$ and yet have little explanatory power on which y_i is actually going to be equal to 1.

An F test on the significance of the parameters is a better indication of the explanatory power of the model.

Another measure of fit: the % of ones hit with the following prediction rule:

$$\hat{Y}_i = 1 \quad \text{if} \quad \hat{P}_i > P^* \tag{26}$$

with P^* equal for example to 0.5.

This may be misleading as well: it could do worse than the the naive rule

$$\hat{Y}_i = 1 \quad \text{if the proportion of 1 in the sample is } > \ 0.5 \tag{27}$$

2.1.6 Probit model

When F is assumed to be standard normal we obtain the Probit model.

$$\begin{aligned} Pr(Y=1) = F(X\beta) &= \int_{-\infty}^{X\beta} \phi(t)dt \\ &= \Phi(X\beta) \end{aligned} \tag{28}$$

Log-likelihood:

$$ln(L) = \sum_{i=1}^{N} [(1-y_i)\ ln(1-\Phi(X_i\beta)) + y_i\ ln(\Phi(X_i\beta))] \tag{29}$$

First order conditions:

$$\frac{\partial ln(L)}{\partial \beta'} = \sum_{i=1}^{N} \lambda_i X_i = \sum_{i=1}^{N} \left(\frac{q_i \phi(q_i X_i \beta)}{\Phi(q_i X_i \beta)} \right) X_i = 0 \tag{30}$$

where $q_i = 2y_i - 1$.

Hessian:

$$\mathbf{H} = \frac{\partial^2 ln(L)}{\partial \beta \partial \beta'} = \sum_{i=1}^{N} -\lambda_i(\lambda_i + X_i\beta)X_i'X_i \tag{31}$$

Estimated variance of the coefficients:

$$\mathbf{V} = -\mathbf{H}^{-1} \tag{32}$$

Probability of the outcome :

$$Prob(Y = 1) = P = \Phi(X\beta) \tag{33}$$

$$\text{Asy. Var } [P] = \left[\frac{\partial \Phi}{\partial \beta'}\right] \mathbf{V} \left[\frac{\partial \Phi}{\partial \beta'}\right]' = \phi^2 \; X\mathbf{V}X' \tag{34}$$

which is a scalar.

Marginal effect:

$$\gamma = \frac{dP}{dX'} = \phi(X\beta) \; \beta \tag{35}$$

$$\begin{aligned} \text{Asy. Var } [\gamma] &= \left[\frac{\partial \gamma}{\partial \beta'}\right] \mathbf{V} \left[\frac{\partial \gamma}{\partial \beta'}\right]' && (36) \\ &= \phi^2 \; [I - (X\beta)\beta X)] \; \mathbf{V} \; [I - (X\beta)\beta X)]' && (37) \end{aligned}$$

which is a $K \times K$ matrix. Note that, for any z, $\frac{d\phi(z)}{dz} = -z\phi(z)$.

See the rules and notation for matrix differentiation in Greene (1997). In all the expressions above X is a row vector of observations on the explanatory factors. Note that Φ and ϕ are functions of $X\beta$.

In order to estimate the probability of the outcome and the marginal effects we need the Maximum Likelihood estimate of β and some realization of X chosen by us (see page 11: a specific individual i, the sample mean, etc...)

Unit variance and homoschedasticity assumptions for ϵ

To obtain the probit specification we have assumed that the distribution F is a standard normal and therefore $\sigma_\epsilon = 1$.

If F is such that $\sigma_\epsilon \neq 1$:

$$\begin{aligned} Pr(Y_i = 1) &= Pr(\epsilon_i \geq -X_i\beta) \\ &= Pr(\frac{\epsilon_i}{\sigma_\epsilon} \geq -X_i\frac{\beta}{\sigma_\epsilon}) \\ &= \phi(X_i\frac{\beta}{\sigma_\epsilon}) \end{aligned} \tag{38}$$

given that $\frac{\epsilon}{\sigma_\epsilon}$ is now distributed as a standard normal and everything else follows as before.

Hence, the assumption of unit variance is equivalent to say that:

- we cannot identify the variance of ϵ;
- we can only identify $\frac{\beta}{\sigma_\epsilon}$;
- the absolute size of the estimated coefficients in a probit does not say much;
- the comparison between estimated coefficients may say more;
- in any case it is better to look at the marginal effects and not at the estimated coefficients;
- since we are interested in marginal effects also heteroschedasticity is less problematic than one may think, if, for example, it takes the form $\sigma_\epsilon = \sigma g(X_i)$.

2.1.7 Logit model

When F is assumed to be logistic we obtain the Logit model.

$$\begin{aligned} Pr(Y=1) = F(X\beta) &= \frac{e^{X\beta}}{1+e^{X\beta}} \\ &= \Lambda(X\beta) \end{aligned} \tag{39}$$

Note that in this case:

$$\begin{aligned} F'(X\beta) = f(X\beta) &= \frac{e^{X\beta}}{(1+e^{X\beta})^2} \\ &= \Lambda(X\beta)[1-\Lambda(X\beta)] \end{aligned} \tag{40}$$

where $[1-\Lambda(X\beta)] = \frac{1}{(1+e^{X\beta})}$

Log-likelihood:

$$ln(L) = \sum_{i=1}^{N}[(1-y_i)\ ln(1-\Lambda(X_i\beta)) + y_i\ ln(\Lambda(X_i\beta))] \tag{41}$$

First order conditions:

$$\frac{\partial ln(L)}{\partial \beta'} = \sum_{i=1}^{N}(y_i - \Lambda(X_i\beta))X_i = 0 \tag{42}$$

Hessian:

$$\mathbf{H} = \frac{\partial^2 ln(L)}{\partial\beta\partial\beta'} = \sum_{i=1}^{N}\Lambda_i(1-\Lambda_i)X_i'X_i \tag{43}$$

Estimated variance of the coefficients:

$$\mathbf{V} = -\mathbf{H}^{-1} \tag{44}$$

Probability of the outcome:

$$Prob(Y = 1) = P = \Lambda(X\beta) = \frac{e^{X\beta}}{1 + e^{X\beta}} \tag{45}$$

$$\text{Asy. Var } [P] = \left[\frac{\partial \Lambda}{\partial \beta'}\right] \mathbf{V} \left[\frac{\partial \Lambda}{\partial \beta'}\right]' = [\Lambda(1-\Lambda)]^2 \, X\mathbf{V}X' \tag{46}$$

which is a scalar.

Marginal effect:

$$\gamma = \frac{dP}{dX} = [\Lambda(1-\Lambda)] \, \beta \tag{47}$$

$$\begin{aligned} \text{Asy. Var } [\gamma] &= \left[\frac{\partial \gamma}{\partial \beta'}\right] \mathbf{V} \left[\frac{\partial \gamma}{\partial \beta'}\right]' && (48) \\ &= [\Lambda(1-\Lambda)]^2 \, [I + (1-2\Lambda)\beta X)] \, \mathbf{V} \, [I + (1-2\Lambda)X'\beta')]' && (49) \end{aligned}$$

which is a $K \times K$ matrix.

See the rules and notation for matrix differentiation in Greene (1997). In all the expressions above X is a row vector of observations on the explanatory factors. Note that Λ is a function of $X\beta$.

In order to estimate the probability of the outcome and the marginal effects we need the Maximum Likelihood estimate of β and some realization of X chosen by us (see page 11: a specific individual i, the sample mean, etc...)

Effects on odds ratios in the Logit model

The Logit model is convenient for a presentation of results in terms of the effects of X on the odds of the outcome $Y = 1$:

$$\Omega(Y = 1|X) = \frac{P}{1 - P} = \frac{\Lambda}{1 - \Lambda} = e^{X\beta} \tag{50}$$

Given two realizations of X, say X_1 and X_0, we can define the odds ratio

$$\frac{\Omega(Y = 1|X_1)}{\Omega(Y = 1|X_0)} = e^{(X_1 - X_0)\beta} \tag{51}$$

This statistics tells us how the odds of observing $Y = 1$ change when X changes from X_0 to X_1.

Stata offers the possibility to display the estimated coefficients in this odds ratio format. For example for the variable j:

$$e^{\beta_j} \tag{52}$$

tells us how the odds of observing $Y = 1$ change when X_j changes by one unit.

- If $e^{\beta_j} > 1$, the variable j increases the odds of observing $Y = 1$.
- If $e^{\beta_j} < 1$, the variable j decreases the odds of observing $Y = 1$.

This way of presenting results is particularly convenient for dummy explanatory variables.

2.1.8 Comparison between linear probability, probit and logit models

The estimated coefficients will clearly differ, but the marginal effects should be fairly similar in general.

Logistic distribution has fatter tails.

We should expect greater differences in case of very few or very large observations with $Y = 1$.

Choice most often based on practical considerations.

See Table 19.2 in Greene (1997) for a comparison of results.

Analysis of proportions Data

Using equation 51, we observe also that the log of the odds Ω is:

$$ln(\Omega) = ln\left(\frac{P}{1-P}\right) = ln\left(\frac{\Lambda}{1-\Lambda}\right) = X\beta \tag{53}$$

This suggests a convenient way to estimate the determinants of dependent variables which are expressed as proportions:

- Proportion of votes for a political party in different elections.
- Proportion of unemployed workers in different regions.
- Proportion of individuals committing crime in different cities.
- ...

In other words, this is convenient when we do not observe the individual outcomes Y_i but only the proportion P_j of outcomes equal to 1 among the individuals in group j.

2.1.9 Maximum score estimator

In section 2.1.5 we have discussed the lack of satisfactory measures of the goodness of fit for ML estimates of binary choices models.

The problem is that ML estimators are not meant to maximize the goodness of fit (which is done, for example, by OLS) .

The Maximum Score estimator for binary choices models is instead based on a fitting rule.

$$\text{Max}_\beta S_{N\alpha}(\beta) = \frac{1}{N}\sum_{i=1}^{N}[Z_i - (1-2\alpha)]sgn(X_i\beta) \tag{54}$$

where:

- α is a preset quintile;
- $Z_i = 2Y_i - 1$ so that $Z = -1$ if $Y = 0$ and $Z = 1$ if $Y = 1$;

If α is set to 0.5, the maximum score estimator chooses β to maximize the number of times that the prediction has the same sign as Z.

In other words, given a prediction rule based on a given percentile, it maximizes the number of correct predictions.

Bootstrapping is used to get an indication of the variability of the estimator. (see Greene).

2.2 Applications: binary choices models for the identification of social effects

The reflection problem (Manski, 1993)

Let each member of the population be characterized by:

y : a scalar outcome (e.g. crime);

z : a vector of individual attributes directly affecting y (e.g. family income, age, education, employment status);

x : a vector of attributes characterizing the reference group (e.g. neighborhood indicator, efficiency of the local public employment office, quality of schools).

We are interested in answering the following questions:

- Does the propensity to commit crime depend on the average crime rate in the neighborhood?
- Does the propensity to commit crime depend on the average individual attributes of the people living in the neighborhood like age, education, family income, or unemployment rate?
- Does the propensity to commit crime depend on exogenous characteristics of the neighborhood like the efficiency of the local public employment office or the quality of schools?

A formal characterization of these questions.

We focus on the linear case in order to understand the nature of the problem. Consider the following model:

$$E(y|x,z) = \alpha + \beta E(y|x) + E(z|x)'\gamma + x'\delta + z'\eta \tag{55}$$

- if $\beta \neq 0$ the model expresses an *endogenous social effect*: the individual's propensity to behave in some way changes with the average behavior of a given reference group;

- if $\gamma \neq 0$ the model expresses an *exogenous social effect*: individuals in the same reference group behave similarly because they have similar personal exogenous attributes, (e.g. sorting on the basis of z);

- if $\delta \neq 0$ individuals in the same reference group behave similarly because they face a similar environment (e.g. local attributes);

- if $\eta \neq 0$ individual characteristics are relevant for the outcome.

Can we identify the parameters?

Problems in the identification of β and γ

If $\beta \neq 1$ we can integrate both sides of 55 with respect to z in order to solve for $E(y|x)$:

$$\begin{aligned} E(y|x) &= \alpha + \beta E(y|x) + E(z|x)'(\gamma + \eta) + x'\delta \\ &= \frac{\alpha}{1-\beta} + E(z|x)'\frac{\gamma+\eta}{1-\beta} + x'\frac{\delta}{1-\beta} \end{aligned} \tag{56}$$

If we plug this back into 55 we get the reduced form:

$$E(y|x,z) = \frac{\alpha}{1-\beta} + E(z|x)'\frac{\gamma+\eta\beta}{1-\beta} + x'\frac{\delta}{1-\beta} + z'\eta \tag{57}$$

which shows that the structural parameters cannot be identified.

If we estimate the reduced form 57 we can only say that:

- if the coefficient on $E(z|x)' \neq 0$,
- if the regressors $[1, E(z|x), x, z]$ are linearly independent in the population,

at least one of the social effects is present, but we cannot determine which one.

A tautological case: z is a function of x

Suppose that:

- z is family income and x is an indicator function for "uptown" and "downtown";
- family income is a function $z = z(x)$ of the neighborhood: individuals work where they live and downtown firms are less productive.

It follows that

$$E(y|x, z) = E(y|x) \tag{58}$$

And therefore equation 55 becomes:

$$E(y|x) = \alpha + \beta E(y|x) + E(z|x)'\gamma + x'\delta + z'\eta \tag{59}$$

which makes sense only with $\beta = 1$ and $\alpha = \gamma = \delta = \eta = 0$.

The model is just a tautology.

In fact, there is no real endogenous social effect. There are only different types of individuals sorted in different groups.

These groups are taken as the reference groups which should originate the social effects.

The pure endogenous social effect model

Empirical studies of endogenous social effects often assume implicitly or explicitly that $\gamma = \delta = 0$ which means:

- no exogenous social effect;
- no effect of local attributes.

The reduced form becomes in this case:

$$E(y|x,z) \;=\; \frac{\alpha}{1-\beta} + E(z|x)'\frac{\eta\beta}{1-\beta} + z'\eta \tag{60}$$

and β is identified as long as $[1, E(z|x), z]$ are linearly independent in the population.

However this is not really a solution: we are assuming away the problem!

In the applications that follows we see some recent alternative solutions.

3 Multiple choices

3.1 Theory

3.1.1 Basic framework and notation

- $i = \{1, 2, 3...N\}$
  denotes a set of decision makers.

- $j = \{0, 1, 2, 3...H\}$
  denotes a *finite* set of *mutually exclusive* and *exhaustive* possible choices.

- $U_{ij} = X_{ij}\beta_j + \epsilon_{ij}$
  is the utility of the decision maker i if the choice is j; it is a function of:
  - a systematic component $X_{ij}\beta_j$ where
    - X_{ij} is a row vector of observed characteristics of the decision maker and of the choices and
    - β_j is a column vector of unknown parameters which may change across choices;
  - a random unobservable component ϵ_{ij}.

- Y_i is the indicator function that denotes which option has been chosen by the decision maker:

$$Y_i = j \quad \text{if } i \text{ chooses } j \tag{61}$$

Decision makers are assumed to maximize utility, and therefore:

$$Y_i = j \quad \text{if} \quad U_{ij} > U_{is} \quad \text{for all } s \neq j \text{ in the choice set} \tag{62}$$

Since we observe only the systematic component of utility, we cannot predict with certainty the choice of each decision maker. We can only try to assess the probability that the decision maker will choose each alternative.

3.1.2 The logit model.

$$\begin{aligned} P_{ij} &= Pr(Y_i = j) \\ &= Pr(U_{ij} > U_{is} \ , \forall \ s \neq j) \\ &= Pr(X_{ij}\beta_j + \epsilon_{ij} > X_{is}\beta_s + \epsilon_{is} \ , \forall \ s \neq j) \\ &= Pr(\epsilon_{is} - \epsilon_{ij} < X_{ij}\beta_j - X_{is}\beta_s \ , \forall \ s \neq j) \end{aligned} \tag{63}$$

If each ϵ_{ij} is distributed independently according to the *extreme value* cumulative distribution

$$exp(-e^{-\epsilon_{ij}}) \tag{64}$$

then, using 63, the probability that the alternative j is chosen is given by the logit distribution (see Train 1986, p53):

$$P_{ij} = \frac{e^{X_{ij}\beta_j}}{\sum_{s=0}^{H} e^{X_{is}\beta_s}} \tag{65}$$

Note that:

- $0 \leq P_{ij} \leq 1$;
- $\sum_{j=0}^{H} P_{ij} = 1$;
- the logit probabilities exhibit the Independence from Irrelevant Alternatives Property (IIA).

3.1.3 Independence from Irrelevant Alternatives Property (IIA)

This property implies that the odds of two alternatives j and s do not depend on the other existing alternatives:

$$\frac{P_{ij}}{P_{is}} = \frac{e^{X_{ij}\beta_j}}{e^{X_{is}\beta_s}} \tag{66}$$

which depends only on i and j.

This property may not be desirable. Consider the following classic example:

- Initially there are only two options: $j =$ "car"; $s =$ "red bus".
- Suppose $\frac{P_{ij}}{P_{is}} = 1$.
- A new option is added: $t =$ "blue bus".
- Suppose that decision makers who choose a bus are indifferent with respect to the color: then we would expect the model to predict: $P_{ij} = 0.5$ and $P_{is} = P_{it} = 0.25$.
- However, the logit model would continue to imply $\frac{P_{ij}}{P_{is}} = 1$.
- In order for this to be compatible with $P_{is} = P_{it}$, the estimated probabilities must be $P_{ij} = P_{is} = P_{it} = \frac{1}{3}$, which is clearly unsatisfactory.

In the context of this example the property is undesirable. In other contexts it may instead be desirable. Examples ...

The validity of the IIA hypothesis can be tested (see below).

3.1.4 Which parameters are identified in the logit model?

Consider as an example the (binary for simplicity) problem of consumer i who has to choose between Japanese ($j = 0$) or European ($j = 1$) cars.

The vector of attributes X_{ij} includes:

- factors Z_{ij} that change across both individuals and choices (e.g. the price or the number of dealers of each car in the city where i lives);
- factors W_i that change only across individuals (e.g. sex, age or income of the consumer);
- a choice specific constants α_j capturing factors that change across choices but not across individuals.

The vector of parameters to be estimated is $\beta_j' = \{\alpha_j, \gamma, \delta\}$, which differs across choices because there is a different constant for each choice. The parameters γ and δ are instead assumed to be identical across choices.

Under these assumptions the probability of the European choice would be:

$$\begin{aligned} P_{i1} = Pr(Y_i = 1) &= \frac{e^{\alpha_1 + Z_{i1}\gamma + W_i\delta}}{e^{\alpha_0 + Z_{i0}\gamma + W_i\delta} + e^{\alpha_1 + Z_{i1}\gamma + W_i\delta}} \\ &= \frac{1}{1 + e^{(\alpha_0 - \alpha_1) + (Z_{i0} - Z_{i1})\gamma}} \end{aligned} \tag{67}$$

This example highlights some identification problems in the logit model:

- if δ is identical across choices, this model cannot identify the effect of the decision maker's attributes ($W_i\delta$ cancels out);
- the model cannot identify the choice-specific constants but only the difference between them $\alpha_0 - \alpha_1$;
- the model can identify the effects γ of the choice-specific attributes also if they are identical across choices.

In order to understand the implications of these findings it is better to focus separately on:

i. models with only individual-specific attributes;

ii. models with only choice specific attributes.

Actual applications may of course jointly consider both types of attributes.

3.1.5 The multinomial logit model

This is the conventional name for a multiple choice problem in which the representative utility of each choice depends only on the attributes of the decision maker:

$$U_{ij} = X_i\beta_j + \epsilon_{ij} \tag{68}$$

Note that, to achieve identification, the attributes are allowed to have different effects on the utility of the different choices. This assumption is also reasonable from an economic point of view.

In this case, the probability of a choice becomes:

$$\begin{aligned} P_{ij} &= \frac{e^{X_i\beta_j}}{\sum_{s=0}^{H} e^{X_i\beta_s}} \\ &= \frac{1}{\sum_{s=0}^{H} e^{X_i(\beta_s-\beta_j)}} \end{aligned} \tag{69}$$

which shows that only differences between parameters can be identified.

It is therefore convenient to impose the normalization with respect to a reference choice, for example $j = 0$ (but any other would do equally well).

Taking $j = 0$ as the reference choice means to impose the normalization $\beta_0 = 0$, which implies $e^{X_i\beta_0} = 1$ and therefore:

$$\begin{aligned} P_{ij} &= Pr(Y_i = j) \\ &= \frac{e^{X_i\beta_j}}{1 + \sum_{s=1}^{H} e^{X_i\beta_s}} \end{aligned} \tag{70}$$

$$\begin{aligned} P_{i0} &= Pr(Y_i = 0) \\ &= \frac{1}{1 + \sum_{s=1}^{H} e^{X_i\beta_s}} \end{aligned} \tag{71}$$

Note that if $H = 1$ we obtain the standard binary choice logit model described in section 2.1.7.

If the matrix X_i includes a vector of ones, the model estimates also H choice specific (normalized) constants.

Estimation of the parameters

The log-likelihood function of the Multinomial logit model is

$$ln(L) = \sum_{i=1}^{N} \sum_{j=0}^{H} d_{ij} ln(P_{ij}) \tag{72}$$

where $d_{ij} = 1$ if i chooses j.

The first order conditions for the maximization of the likelihood are

$$\frac{\partial ln(L)}{\partial \beta_j} = \sum_{i=1}^{N} (d_{ij} - P_{ij}) X_i = 0 \tag{73}$$

Note that this is a system of $K \times H$ equations.

The second derivatives matrix is composed by H^2 blocks each with dimension $K \times K$.

The "main diagonal" blocks have the form

$$\frac{\partial^2 ln(L)}{\partial \beta_j \partial \beta_j'} = -\sum_{i=1}^{N} P_{ij}(1 - P_{ij}) X_i' X_i \tag{74}$$

The "off main diagonal" blocks (for $j \neq s$) have the form

$$\frac{\partial^2 ln(L)}{\partial \beta_j \partial \beta_s'} = \sum_{i=1}^{N} P_{ij} P_{is} X_i' X_i \tag{75}$$

Interpretation of the parameters

The parameters β_j should be interpreted carefully.

Note that:

$$ln\frac{P_{ij}}{P_{i0}} = X_i\beta_j \tag{76}$$

The coefficient β_j measures the impact of the attributes X_i on the log-odds that the decision maker chooses j instead of 0.

Note also that:

$$ln\frac{P_{ij}}{P_{is}} = X_i(\beta_j - \beta_s) \tag{77}$$

The *difference* between the coefficients β_j and β_s measure the impact of the attributes X_i on the log-odds that the decision maker chooses j instead of s.

Because of the IIA property the odds concerning any couple of choices are independent from all the other choices.

If only choice specific constants are included, their maximum likelihood estimate is the proportion of individuals that make each choice.

Interpretation of marginal effects

The marginal effects of the individual attributes X_i on the probability of a choice $Y_i = j$ are even more difficult to interpret.

$$\gamma_j = \frac{\partial P_j}{\partial X_i} = P_j(\beta_j - \sum_{s=0}^{H} P_s \beta_s) = P_j(\beta_j - \bar{\beta}) \tag{78}$$

Hence, the effect of X_i on P_j (the generic probability of a j choice) depends on the parameters concerning all the choices, not just on the parameters concerning choice j.

The problem is that when X_i changes all the probabilities of all the choices are contemporaneously affected.

- Consider the car example with three choices: Japanese ($j = 0$) , European ($j = 1$) and American ($j = 2$).
- Suppose X_i is the age of the buyer and that $\beta_2 > \beta_1 > 0$.
- This implies that older workers tend to buy more European and more American cars than Japanese cars. Moreover, older workers tend to buy more American cars than European cars.
- However, if β_2 is much larger than β_1 it may happen that the probability of a European choice decreases, when age X_i increases.

Note also that the marginal effects are a function of the explanatory factors X (which are in $P_j = \frac{e^{X\beta_j}}{\sum_{s=0}^{H} e^{X\beta_s}}$), and therefore have to be computed at some reference value of X (the mean, the median, a particular i ...)

Effects on odds ratios

As in the binary case, results can be expressed in the form of odds ratios, or exponentiated form.

The odds of a choice j instead of 0, given X_i, are:

$$\Omega(Y_i = j; Y_i = 0|X) = \frac{P_{ij}}{P_{i0}} = e^{X_i \beta_j} \tag{79}$$

Given two realizations of X_i, say X_1 and X_0, we can define the odds ratio

$$\frac{\Omega(Y_i = j; Y_i = 0|X_1)}{\Omega(Y_i = j; Y_i = 0|X_0)} = e^{(X_1 - X_0)\beta_j} \tag{80}$$

This statistics tells us how the odds of observing $Y = j$ instead of $Y = 0$ change when X_i changes from X_0 to X_1.

Stata offers the possibility to display the estimated coefficients in the odds ratio format:

$$e^{\beta_j} \tag{81}$$

which tells us how the odds change when the individual attributes change by one unit.

- If $e^{\beta_j} > 1$, X_i increases the odds of observing $Y = j$ as opposed to $Y = 0$.
- If $e^{\beta_j} < 1$, the variable j decreases the odds of observing $Y = j$ as opposed to $Y = 0$.

3.1.6 The (pure) Conditional Logit model

This is the conventional name for a multiple choice problem in which the representative utility of each choice depends on choice specific attributes:

$$U_{ij} \;=\; X_{ij}\beta + \epsilon_{ij} \tag{82}$$

The probability of a choice would be:

$$P_{ij} = \frac{e^{X_{ij}\beta}}{\sum_{s=0}^{H} e^{X_{ij}\beta}} \tag{83}$$

and in this case the coefficients β are identified even if they are identical across choices. Marginal effects can be characterized and interpreted more easily.

Note that the name Conditional Logit model is also used for the general situation in which both individual-specific and choice- specific attributes are considered.

Marginal effects

Consider the car example and suppose that X_{ij} is the number of dealers in the city of each buyer, for each type of car

The marginal effect of an increase in the number of dealers of car j on the probability that car j is bought is:

$$\gamma_{jj} = \frac{\partial P_j}{\partial X_{ij}} = P_j(1 - P_j)\beta_{dealer} \tag{84}$$

The marginal effect of an increase in the number of dealers of car s on the probability that car j is bought is:

$$\gamma_{js} = \frac{\partial P_j}{\partial X_{is}} = -P_j P_s \beta_{dealer} \tag{85}$$

The usual odds ratios (exponentiated) representation of coefficients is also possible

Estimation of the parameters

The log-likelihood function of the (pure) Conditional logit model is

$$ln(L) = \sum_{i=1}^{N} \sum_{j=0}^{H} d_{ij} ln(P_{ij}) \tag{86}$$

where $d_{ij} = 1$ if i chooses j.

The first order conditions for the maximization of the likelihood are

$$\frac{\partial ln(L)}{\partial \beta} = \sum_{i=1}^{N} \sum_{j=0}^{H} d_{ij}(X_{ij} - \bar{X}_i) = 0 \tag{87}$$

where $\bar{X}_i = \sum_{j=0}^{H} P_{ij} X_{ij}$ Note that this is a system of K conditions.

The second derivatives matrix is:

$$\frac{\partial^2 ln(L)}{\partial \beta \partial \beta'} = \sum_{i=1}^{N} \sum_{j=0}^{H} P_{ij}(X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)' \tag{88}$$

3.1.7 A test for the IIA hypothesis

Hausman and McFadden (1984) suggest that if a subset of the choice set is really irrelevant, omitting it from the model should not change the parameter estimates systematically.

Consider a choice set $A = \{B, C\}$ where B and C are subsets of A. We want to test whether the presence of the choices in C are irrelevant for the odds between the choices in B.

The statistic for the "Hausman's specification test":

$$HM = (\hat{\beta}_B - \hat{\beta}_A)'[\hat{V}_B - \hat{V}_A]^{-1}(\hat{\beta}_B - \hat{\beta}_A) \tag{89}$$

where

- $\hat{\beta}_B$ and $\hat{\beta}_A$ are the ML estimates of the parameters of the restricted and unrestricted models;
- $\hat{V}_B$ are $\hat{V}_A$ are the ML estimates of the asymptotic covariance matrices of the restricted and unrestricted models.
- Both estimates are consistent under the null and $\hat{\beta}_A$ is more efficient.

The statistic HM is distributed as a chi-squared with degrees of freedom equal to the number of parameters.

An example

Consider the car example: we would like to know whether the odds of the choice "European versus Japanese" are really independent from the the presence of the "American" choice.

The explanatory factors are the number of dealers for each type of car and each consumer (*dealer*), the age of the consumer (*age*) and the choice specific constants.

The procedure for the test is as follows.

i. Estimate the un-restricted model with all the three choices and all the observations.

ii. Estimate the restricted model with only two choices (European and Japanese), dropping the observations for consumers who choose American cars. Also the constant for the American choice has to be dropped, because it cannot be estimated in the restricted model

iii. Construct the test statistics using only the parameters estimated for both model

iv. Therefore, note that the test cannot involve a comparison of the estimates of the American-specific constant; moreover, the rows and columns corresponding to this parameter in the asymptotic covariance matrix of the un-restricted model should be dropped.

v. If the test statistic is "greater" than the preferred critical value it means that there is a statistically significant difference between the estimates of the restricted and the unrestricted model.

vi. Hence, the evidence would not support the IIA property.

Problems of this test

Three types of problems arise with this test.

i. HM is not bounded to be positive in finite sample because the difference between the two covariance matrices may not be positive semi-definite.

   - Hausman and McFadden (1984) suggest that this supports the null.

ii. Only a subset of the parameters is identified in the restricted model.

iii. It is not obvious how to select the choices to be included in the restricted subset B and the choices to be tested for irrelevance and included in C.

Alternative tests are available (see the survey in Brooks, Fry and Harris, 1998). They are of two kinds:

- Other (non-Hausman type) tests based on partitions of the choice set:
  - for example, McFadden, Train and Tye (1981) propose a likelihood ratio test based on the comparison between the un-restricted and the restricted model:

$$MTT = -2[logL(\hat{\beta}_A) - logL(\hat{\beta}_B)] \tag{90}$$

  - note that these tests solve the problem 1 above, but do not solve the other two problems.
- Tests designed against specific alternatives, such as Nested Logit Models, which solve the other two problems and offer more power at the cost of a loss of generality (See Hausman and McFadden, 1984).

3.2 Applications of multiple choices models

4 Panel data

4.1 Examples

The standard situation: a sample of individuals observed for several time periods.

- $i = \{1, 2, 3...N\}$
  individuals (workers, firms ...)
- $t = \{1, 2, 3...T\}$
  time periods for which we have observations on the individuals.

Other apparently different but in fact formally similar situations:

- Siblings (or twins) in families:
  - $i = \{1, 2, 3...N\}$: siblings
  - $j = \{1, 2, 3...J\}$: families
- Workers in different geographic areas and different language groups:
  - $i = \{1, 2, 3...N\}$: individuals
  - $j = \{1, 2, 3...J\}$: geographic areas
  - $k = \{1, 2, 3...K\}$: language groups
- Workers in branches of the same firm and in different years:
  - $i = \{1, 2, 3...N\}$: individuals
  - $j = \{1, 2, 3...J\}$: branches
  - $t = \{1, 2, 3...T\}$: time periods

4.2 Problems arising in cross sections and solved by panel data

4.2.1 Example 1: Production functions and managerial ability

We would like to estimate the following linear approximation to a production function (see Mundlack, 1961)

$$y_{it} = \beta_1 + \beta_2 l_{it} + \beta_3 m_i + \epsilon_{it} \tag{91}$$

where

- i is a firm;
- t is time;
- $y = log(Y)$ is the log of output;
- $l = log(L)$ is the log of labor;
- $m = log(M)$ is the log of managerial ability: *unobservable*;
- ϵ_{it} is an i.i.d. error term such that $E\{\epsilon_{it}\} = 0$.

Suppose we have information only on a cross section of firms for a given t, so that we can only estimate

$$y_i = \beta_1 + \beta_2 l_i + \eta_i \tag{92}$$

where $\eta_i = \beta_3 m_i + \epsilon_i$.

Given 92 we have that:

$$\begin{aligned} E(y_i|l) &= \beta_1 + \beta_2 l_i + E(\eta_i|l) \\ E(y_i|l) &= \beta_1 + \beta_2 l_i + \beta_3 E(m_i|l) \end{aligned} \tag{93}$$

Suppose that:

$$E(m_i|l) \;=\; \lambda_1 + \lambda_2 l_i \tag{94}$$

then, substituting 94 in 93 gives:

$$E(y_i|l) \;=\; (\beta_1 + \beta_3\lambda_1) + (\beta_2 + \beta_3\lambda_2)l_i \tag{95}$$

If we estimate using OLS the regression of y on l we obtain:

$$\hat{Y}_i = b_1 + b_2 l_i \tag{96}$$

but, given 95 and the OLS properties, b_2 *is a biased estimate of the causal effect of* l *on* y because:

$$E(b_2) = \beta_2 + \beta_3\lambda_2 \tag{97}$$

Note that that true causal effect of l on y is β_2. Assuming that $\beta_3 > 0$ (which is reasonable) OLS:

- over-estimates labor productivity β_2 if managerial quality is positively correlated with the quantity of labour $\lambda_2 > 0$;
- under-estimates labor productivity β_2 if managerial quality is negatively correlated with the quantity of labour $\lambda_2 < 0$;

Panel data can solve this problem as long as managerial quality can be assumed not to change over time.

4.2.2 Example 2: Returns to schooling, ability and twins

We would like to estimate the returns to schooling in the following model (see Ashenfelter and Krueger, 1994):

$$y_{ij} = \alpha + \beta S_{ij} + \gamma X_j + \mu A_j + \epsilon_{ij} \tag{98}$$

where

- i is a twin and suppose for simplicity that $i = 1, 2$;
- j is a family;
- y_{ij}: log of the wage rate;
- X_j: family income;
- A_j: genetic and cultural ability of family members (*nature and nurture*);
- S_{ij} years of schooling of each twin;
- ϵ_{ij} is an i.i.d. error term such that $E\{\epsilon_{ij}\} = 0$.

Suppose we have information only on one twin per family for whom we observe only earnings and years of schooling. Then the model becomes

$$y_{1j} = \alpha + \beta S_{1j} + \eta_{1j} \tag{99}$$

where $\eta_{1j} = \gamma X_j + \mu A_j + \epsilon_{ij}$

Following the same steps as in example 1, the OLS estimates of the return to schooling β in a regression of y on S is biased:

$$E(b) = \beta + \gamma\lambda_2 + \mu\delta_2 \tag{100}$$

where we are assuming that

$$\begin{aligned} E(X_j|S) &= \lambda_1 + \lambda_2 S_j \\ E(A_j|S) &= \delta_1 + \delta_2 S_j \end{aligned} \tag{101}$$

and the bias is positive since γ, λ_2, μ and δ_2 are likely to be positive.

We can improve the situation by extending the available information on the observed twin. For example, if we obtain information on family income X_j than the model would be

$$y_{1j} = \alpha + \beta S_{1j} + \gamma X_j + u_{1j} \tag{102}$$

where $u_{1j} = \mu A_j + \epsilon_{ij}$, and the bias would decrease to

$$E(b) = \beta + \mu\delta_2 \tag{103}$$

But there are variables like *ability* that are not *fully* observable. In this case, to solve the problem we need a panel data structure.

As long as we can assume that genetic and cultural ability (*nature and nurture*) is constant across twins of the same family, data on more than one twin per family would allow us to eliminate the bias $\mu\delta_2$ due to unobservable ability, because this variable changes only across families but is fixed within families.

4.3 A general framework and more notation

Consider the following model:

$$Y_{it} = \alpha_i + X_{it}\beta + \epsilon_{it} \tag{104}$$

where:

- Y_{it} is an outcome for individual i at time t.
- α_i is a time invariant individual effect. Note that it measure the effect of all the factors that are specific to individual i but constant over time.
- X_{it} is a row vector of observations on K explanatory factors for individual i at time t, *not including the constant term.*
- β is a column vector of K parameters.
- ϵ_{it} is an i.i.d. error term such that $E\{\epsilon_{it}\} = 0$.

Note that we could write the model as

$$Y_{it} = (\alpha + \nu_i) + X_{it}\beta + \epsilon_{it} \tag{105}$$

allowing for a general constant term. But clearly the parameters α and ν_i would not be uniquely identified. A normalization is needed and the standard one is to assume $\alpha = 0$ and $\nu_i = \alpha_i$.

But other normalizations would do like for example

- $\nu_1 = 0$ which would allow us to identify the general constant and $N - 1$ individual-specific fixed effects.
- $\sum \nu_i = 0$ which is the normalization assumed by STATA in the command XTREG, FE (so a general constant term a is estimated) for reasons to be explained below.

In matrix form the model can be written as:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_N \end{bmatrix} = \begin{bmatrix} \mathbf{i} & 0 & . & . & . & 0 \\ 0 & \mathbf{i} & . & . & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & . & . & . & \mathbf{i} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ . \\ . \\ . \\ \alpha_N \end{bmatrix} + \begin{bmatrix} X_1 \\ X_2 \\ . \\ . \\ . \\ X_N \end{bmatrix} \beta + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_N \end{bmatrix} \tag{106}$$

where:

- Y_i and X_i are the T time observations on the outcome and on the K explanatory factors for individual i.
- β is a column vector of K parameters.
- α_i is the time invariant individual fixed effect.
- ϵ_i is the the vector of T disturbances for individual i.
- $\mathbf{i}$ is a T dimensional column vector with all elements equal to 1.

We are primarily interested in obtaining estimates of the parameters β which represent the causal effect of X on Y.

It is useful to draw a distinction between the model described above and the estimators we analyze below.

4.4 Fixed effects (within) estimators

4.4.1 Least squares dummy variable model (LSDV)

A first way to proceed is to estimate with OLS a model in which we include a dummy variable for each individual in the sample. The estimated model for individual i at time t would be:

$$Y_{it} = d_1\alpha_1 + d_2\alpha_2 + \ \ldots \ \ d_N\alpha_N + X_{it}\beta + \epsilon_{it} \tag{107}$$

where

- $d_j(i) = 1$ if $j = i$;
- $d_j(i) = 0$ if $j \neq i$.

In compact matrix format, we can write the model as:

$$Y = \mathbf{D}\alpha + \mathbf{X}\beta + \epsilon \tag{108}$$

where:

- Y is the NT column vector of the observations on the outcome;
- $\mathbf{D}$ is the $NT \times N$ matrix of the observations on the dummies;
- α is the N column vector of the individual-specific fixed effects;
- $\mathbf{X}$ is the $NT \times K$ matrix of the observations on the explanatory factors;
- β is the K column vector of the parameters of primary interest;
- ϵ is the NT column vector of disturbances.

Note that this is a correctly specified regression with $N + K$ regressors. OLS applied to 108 would give unbiased estimates of the parameters of interest: i.e. if b_{LSDV} indicates the OLS estimate of β in 108 we have that:

$$E(b_{LSDV}) = \beta \tag{109}$$

Disadvantages of this procedure

- It may be computationally unfeasible if the number of time invariant fixed effects to be estimated is too large.

Advantages of this procedure

- If the computer is powerful enough, it is a very simple way to estimate the parameter of interest.

Examples

Bertrand, Luttmer and Mullianathan (1998) estimate a regression with the following form (eq. 3 in their paper):

$$W_{ijk} = (CA_{jk} * \bar{W}_k)\alpha + X_i\beta + \gamma_k + \delta_j + CA_{jk}\theta + \epsilon_{ijk} \tag{110}$$

in which they include 42 language group fixed effects γ_k and 1196 local area fixed effects δ_j; the parameter of interest is α which is identified controlling for these fixed effects.

To see the link with the standard panel setup described above, you can consider their dataset as a panel of "area - language" cells observed over different individuals.

4.4.2 Analysis of Covariance: using deviations from individual specific means

If N is too large the LSDV estimator is not feasible and we need a trick to construct a feasible estimator for the parameters β.

The trick is offered by the results concerning "partioned regressions", "projection matrices" and "partialling out matrices".

The intuition is the following. Given a regression like 108:

$$Y = \mathbf{D}\alpha + \mathbf{X}\beta + \epsilon$$

unbiased estimates of β can be obtained with this procedure.

- Regress Y on $\mathbf{D}$ and retrieve the estimated residuals Y^*.
- Regress $\mathbf{X}$ on $\mathbf{D}$ and retrieve the estimated residuals $\mathbf{X}^*$.
- Regress Y^* on $\mathbf{X}^*$ to obtain an estimate of β; This estimate is numerically equivalent to the LSDV estimate of 108.

In our panel setup in which $\mathbf{D}$ is a matrix of individual specific dummies:

- the elements of Y^* are the deviations of each element of Y with respect to the correspondent individual specific mean;
- the elements of $\mathbf{X}^*$ are the deviations of each element of $\mathbf{X}$ with respect to the correspondent individual specific mean.

So to obtain an estimate of β when the LSDV model is unfeasible, we can compute the deviations of Y and $\mathbf{X}$ with respect to their individual specific means and then regress the deviation of Y on the deviations of $\mathbf{X}$.

4.4.3 A parenthesis on partitioned regressions

To understand the "mechanics" of this procedure, consider again equation 108:

$$Y = \mathbf{D}\alpha + \mathbf{X}\beta + \epsilon$$

The matrix

$$\mathbf{H} = \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}' \tag{111}$$

is called "projection matrix" because if you premultiply any vector Z by H, from a graphical point of view the result is the projection of the vector Z on $\mathbf{D}$. Numerically it gives the least square prediction of Z given $\mathbf{D}$ (see graphical interpretation of OLS).

If we premultiply Y by $\mathbf{H}$ we obtain the least square prediction:

$$\hat{Y} = \mathbf{H}Y = \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'Y = \mathbf{D}b_{YD} \tag{112}$$

where b_{YD} is the OLS estimate of the coefficients of the regression of Y on $\mathbf{D}$.

If we premultiply $\mathbf{X}$ by $\mathbf{H}$ we obtain the least square prediction:

$$\hat{\mathbf{X}} = \mathbf{H}\mathbf{X} = \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\mathbf{X} = \mathbf{D}b_{XD} \tag{113}$$

where b_{XD} is the OLS estimate of the coefficients of the regression of $\mathbf{X}$ on $\mathbf{D}$.

Note that $\mathbf{H}$ is an idempotent matrix.

The matrix

$$\mathbf{M} = I - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}' \tag{114}$$

is called "partialling out matrix"; if you premultiply any vector Z by $\mathbf{M}$ you obtain the least square estimated residuals of the regression of Z on $\mathbf{D}$ (see graphical analysis).

If we premultiply Y by $\mathbf{M}$ we obtain the residuals:

$$Y^* = \mathbf{M}Y = Y - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'Y = Y - \mathbf{D}b_{YD} \tag{115}$$

estimated from the regression of Y on $\mathbf{D}$.

If we premultiply $\mathbf{X}$ by $\mathbf{M}$ we obtain the residuals:

$$\mathbf{X}^* = \mathbf{M}\mathbf{X} = \mathbf{X} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\mathbf{X} = \mathbf{X} - \mathbf{D}b_{XD} \tag{116}$$

estimated from the regression of $\mathbf{X}$ on $\mathbf{D}$.

If we premultiply ϵ by $\mathbf{M}$ we obtain the residuals:

$$\epsilon^* = \mathbf{M}\epsilon = \epsilon - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\epsilon = \epsilon - \mathbf{D}b_{\epsilon D} \tag{117}$$

estimated from the regression of ϵ on $\mathbf{D}$. Note that $E(\epsilon^*) = 0$ if $E(\epsilon) = 0$.

If we premultiply $\mathbf{D}$ by $\mathbf{M}$ we obtain:

$$\mathbf{M}\mathbf{D} = \mathbf{D} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\mathbf{D} = 0 \tag{118}$$

Note that also $\mathbf{M}$ is an idempotent matrix.

If we premultiply by $\mathbf{M}$ the entire equation 108 we obtain

$$\mathbf{M}Y = \mathbf{MD}\alpha + \mathbf{MX}\beta + \mathbf{M}\epsilon \tag{119}$$

$$Y^* = \mathbf{X}^*\beta + \epsilon^* \tag{120}$$

which explains why $\mathbf{M}$ is called "partialling out" matrix.

Equation 120 is a well behaved equation that can be estimated with OLS to obtain an unbiased and consistent estimate of β without having to directly estimate α.

Note that equation 120 is a regression of the component of Y which is orthogonal to $\mathbf{D}$ on the component of $\mathbf{X}$ which is orthogonal to $\mathbf{D}$.

This is in fact what partial regression coefficients capture.

4.4.4 Back to the Analysis of Covariance

$\mathbf{D}$ is an $NT \times N$ matrix of dummies with the following form

$$\mathbf{D} = \begin{bmatrix} \mathbf{i} & 0 & . & . & . & 0 \\ 0 & \mathbf{i} & . & . & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & . & . & . & \mathbf{i} \end{bmatrix} \tag{121}$$

where $\mathbf{i}$ is an T column vector with elements equal to 1.

Given this particular form, partialling out D implies taking away from each variable its individual specific mean.

To see this note that the partialling out matrix takes the followint form:

$$\mathbf{M} = I - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}' \tag{122}$$

$$= \begin{bmatrix} \bar{\mathbf{M}} & 0 & . & . & . & 0 \\ 0 & \bar{\mathbf{M}} & . & . & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & . & . & . & \bar{\mathbf{M}} \end{bmatrix} \tag{123}$$

where $\bar{M}$ is a $T \times T$ matrix equal to:

$$\bar{\mathbf{M}} = I_T - \frac{1}{T}\mathbf{i}\mathbf{i}' \tag{124}$$

If we premultiply any T vector Z by $\bar{\mathbf{M}}$ we obtain the vector

$$\bar{\mathbf{M}}Z = Z - \bar{Z}\mathbf{i} \tag{125}$$

where $\bar{Z}$ is the mean of the T elements of Z

Therefore, our partitioned regression is:

$$\begin{aligned} \mathbf{M}Y &= \mathbf{MD}\alpha + \mathbf{MX}\beta + \mathbf{M}\epsilon \\ Y^* &= \mathbf{X}^*\beta + \epsilon^* \end{aligned}$$

is equivalent to the following regression

$$[Y_{it} - \bar{Y}_{i.}] = [X_{it} - \bar{X}_{i.}]\beta + [\epsilon_{it} - \bar{\epsilon}_{i.}] \tag{126}$$

where

- $\bar{Y}_{i.}$ is the mean of the T observations on the outcome for individual i;
- $\bar{X}_{i.}$ is the K row vector of the means of the T observations on the explanatory factors X for individual i;

Exercise: verify the above procedure for the case $i = \{1, 2\}$ and $t = \{1, 2\}$.

OLS estimation of 126 gives the fixed effect estimator b_{FE} of β, which can be written in matrix form as:

$$b_{FE} = [\mathbf{X}'\mathbf{MX}]^{-1}[\mathbf{X}'\mathbf{M}Y] \tag{127}$$

which is unbiased and consisitent:

$$E(b_{FE}) = \beta \tag{128}$$

Note that b_{FE} is numerically equal to b_{LSDV} and is also called *within* estimator, to be distinguished from the *between* estimator which will be discussed below. Another name for this way to proceed is "Analysis of Covariance".

Estimates a_i of the individual fixed effects can be obtained as estimates of the mean residual for each individual:

$$a_i = \bar{Y}_{i.} - \bar{X}_{i.} b_{FE} \tag{129}$$

The estimator of the covariance matrix for b_{FE} is:

$$\widehat{COV}(b_{FE}) = s^2[\mathbf{X}'\mathbf{M}\mathbf{X}]^{-1} \tag{130}$$

where

$$s^2 = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - a_i - X_{it} b_{FE})^2}{NT - N - K} \tag{131}$$

where

$$e_{it} = (Y_{it} - a_i - X_{it} b_{FE}) \tag{132}$$

is the estimated ith residual.

The estimator of the covariance matrix for a_i is:

$$\widehat{VAR}(a_i) = \frac{s^2}{T} + \bar{X}_i \widehat{COV}[b_{FE}] \bar{X}_i' \tag{133}$$

Note that the STATA command XTREG, FE computes the fixed effect (within) estimate b_{FE} of β. However the individual fixed effects are estimated as deviations from a common mean, as in the model

$$Y_{it} = \alpha + \nu_i + X_{it}\beta + \epsilon_{it} \tag{134}$$

with the constraint $\sum \nu_i = 0$.

One of the advantages of this choice is that it simplifies the computation of predicted values of the outcome.

4.4.5 First differences

Consider the standard model and assume that we have only two observarvations for each i

$$Y_{i1} = \alpha_i + X_{i1}\beta + \epsilon_{i1} \tag{135}$$

$$Y_{i2} = \alpha_i + X_{i2}\beta + \epsilon_{i2} \tag{136}$$

and assume that we have only two time observations.

If we subtract 136 from 135 we obtain the equation in first difference:

$$Y_{i1} - Y_{i2} = [X_{i1} - X_{i2}]\beta + \epsilon_{i1} - \epsilon_{i2} \tag{137}$$

which, given our assumptions, can be estimated with OLS. The estimate of β is numerically equal to the fixed effect (within) estimator b_{FE}.

Example 1: Ashenfelter and Krueger (1994)

$$y_{ij} = \alpha + \beta S_{ij} + \gamma X_j + \mu A_j + \epsilon_{ij} \tag{138}$$

$$y_{i1} - y_{i2} = \beta[S_{i1} - S_{i2}] + \epsilon_{i1} - \epsilon_{i2} \tag{139}$$

Example 2: Ichino and Maggi (1999)

$$S_{it} = \alpha_i + \delta t X_i + \beta \bar{S}_{it} + \sum_j \zeta_j D_{ijt} + \gamma Z_{it} + \epsilon_{it}, \tag{140}$$

$$S_{it} - S_{it-1} = \delta X_i + \beta(\bar{S}_{it} - \bar{S}_{it-1}) + \sum_j (D_{ijt} - D_{ijt-1})\zeta_j + \gamma(Z_{it} - Z_{it-1}) + \epsilon_{it} - \epsilon_{it-1}. \tag{141}$$

Note in this example:

- time invariant observable characteristics have time varying coefficients and therefore do not cancel out in the first difference equation.
- the first differences of the branch fixed effects are variables that take values $\{-1, 0, 1\}$.

4.4.6 Differences-in-Differences (DD) strategies

The DD strategies offer simple ways to estimate causal effects in panel data when certain groups of observations are exposed to the causing variable and other not.

This approach is particularly well suited to estimating the effect of sharp changes in the economic environment or changes in government policy.

A good example is Card (1990) which examines the effect of immigration on the employment of natives using the "natural experiement" generated by the sudden large-scale migration from Cuba to Miami known as the "Mariel Boatlift".

Card asks whether the Mariel immigration (an increase of 7% of the Miami labor force between May and September 1980) reduced the employment or the wages of non-immigrants labor groups.

The identification strategy is based on the comparison between what happened in Miami and what happened in other *comparable* US cities, assumed to be representative of what would have happened in Miami absent the Mariel immigration (see Figure 1 in WP version of Card, 1990).

Another example is Card and Sullivan (1988) who use a DD estimator to evaluate the effect of a training program on the probability of employment after training.

Consider the following framework:

- i denotes workers in a city;
- *In the absence of immigration*:
  - $Y_i = Y_{0i} = 1$ if worker i is unemployed;
  - $Y_i = Y_{0i} = 0$ if worker i is employed.
- *In the presence of immigration*:
  - $Y_i = Y_{1i} = 1$ if worker i is unemployed;
  - $Y_i = Y_{1i} = 0$ if worker i is employed.

Note that only one of these outcomes is actually observed for each individual, but to understand this approach is useful to think in terms of "counterfactuals" and to consider that all outcomes exist although only one is observed.

The unemployment rate in city c at time t is:

- $E(Y_{0i}|c,t)$ in the absence of immigration;
- $E(Y_{1i}|c,t)$ in the presence of immigration.

The DD approach assumes that:

$$E(Y_{0i}|c,t) = \beta_t + \gamma_c \tag{142}$$

$$E(Y_{1i}|c,t) = \beta_t + \gamma_c + \delta = E(Y_{0i}|c,t) + \delta \tag{143}$$

Hence, unemployment in a city is determined only by:

- a time fixed effect β_t equal for all cities;
- a city fixed effect γ_c equal for all time periods;
- the causal effect of immigration which appears only if the city is exposed to an immigration wave.

Suppose that we have two cities:

- $c = M$ which has been exposed to migration (Miami);
- $c = L$ which has not been exposed to migration (Los Angeles);

and two time periods

- $t = 79$: before the migration wave;
- $t = 81$: after the migration wave.

The sample statistics that we can use are the ones which correspond to the following population parameters:

- $E(Y_i|c = M, t = 79) = E(Y_{0i}|c, t) = \beta_{79} + \gamma_M$
- $E(Y_i|c = M, t = 81) = E(Y_{1i}|c, t) = \beta_{81} + \gamma_M + \delta$
- $E(Y_i|c = L, t = 79) = E(Y_{0i}|c, t) = \beta_{79} + \gamma_L$
- $E(Y_i|c = L, t = 81) = E(Y_{0i}|c, t) = \beta_{81} + \gamma_L$

The crucial role of the assumptions 142 and 143 is to ensure that unemployment growth would have been the same in both cities

- if both of them were not not exposed to migration:

$$\begin{aligned} E(Y_{0i}|c = M, t = 81) - E(Y_{0i}|c = M, t = 79) &= \beta_{81} - \beta_{79} \\ E(Y_{0i}|c = L, t = 81) - E(Y_{0i}|c = L, t = 79) &= \beta_{81} - \beta_{79} \end{aligned}$$

- if both of them were exposed to migration:

$$\begin{aligned} E(Y_{1i}|c = M, t = 81) - E(Y_{0i}|c = M, t = 79) &= \beta_{81} - \beta_{79} + \delta \\ E(Y_{1i}|c = L, t = 81) - E(Y_{0i}|c = L, t = 79) &= \beta_{81} - \beta_{79} + \delta \end{aligned}$$

In other words, controlling for city fixed effects, if the cities have the same migration history, they also have the same changes in unemployment rates.

This is the crucial identifying assumption and is *non-testable* because the migration history is not the same in the two cities.

If this assumption holds, the difference between the unemployment changes in the two cities (the *difference-in-difference*) measures the causal effect of migration on unemployment:

$$\begin{aligned} [E(Y_{1i}|c = M, t = 81) \quad &- \quad E(Y_{0i}|c = M, t = 79)] - \\ [E(Y_{0i}|c = L, t = 81) \quad &- \quad E(Y_{0i}|c = L, t = 79)] = \\ [\beta_{81} - \beta_{79} + \delta] \quad &- \quad [\beta_{81} - \beta_{79}] = \\ \delta & \end{aligned} \tag{144}$$

Note that by taking the difference-in-differences we control for city fixed effects and time fixed effects.

If data on individuals are available, using a more standard regression framework, the difference-in-difference estimator can be obtained from an estimate of the following equation based on the pooled observations for all workers, in all cities and all years:

$$Y_{ict} = \beta_t + \gamma_c + \delta D_{ict} + \epsilon_{ict} \tag{145}$$

where:

- $D_{ict} = 1$ if $c = M$ and $t = 81$ (0 otherwise)
- $E(\epsilon_{ict}|c, t) = 0$ and the these disturbances are i.i.d.
- $c = \{M, L\}$
- $t = \{79, 81\}$

It is easy to check that this model generates the same conditional expectations described above.

This regression framework shows that the DD estimator can also be computed controlling for individual characteristics, by including a vector X_{ict} of these characteristics in the regression 145:

$$Y_{ict} = X_{ict}\alpha + \beta_t + \gamma_c + \delta D_{ict} + \epsilon_{ict} \tag{146}$$

The DD approach rests on the assumption that time differences in the outcomes are identical across cities if the treatment histories are the same.

This assumption can be more easily considered plausible when we control for X as in equation 146.

However, this assumption cannot be tested and evidence on the trends in outcomes before and after the event of interest may help making it more plausible.

4.4.7 Fixed effects estimators and measurement error

Consider the twins' model:

$$y_{ij} = \alpha_j + \beta S_{ij} + \epsilon_{ij} \tag{147}$$

where $i = \{1, 2\}$ denotes a twin, j denotes a family, α_j includes all the family specific effects which are fixed across twins in the same family and S_{ij} is schooling. For simplicity we omit other covariates.

Suppose that S_{ij} is the true number of years of schooling, but because of measurement error we observe

$$\tilde{S}_{ij} = S_{ij} + \mu_{ij} \tag{148}$$

where μ_{ij} is a classical error of measurement, assumed to be i.i.d. and uncorrelated with all the true S_{ij}.

$$COV(S, \mu) = 0 \tag{149}$$

The estimated equation is:

$$y_{ij} = \alpha_j + \beta \tilde{S}_{ij} - \beta \mu_{ij} + \epsilon_{ij} \tag{150}$$

Classical measurement error in a non-panel framework

As a reference benchmark to understand what happens in the case of panel data, let's look at what happens in the classical case with no individual specific fixed effects: i.e. $\alpha_j = \alpha$ for all j.

So the estimated model is

$$Y_{ij} = \alpha + \beta \tilde{S}_{ij} + \eta_{ij} \tag{151}$$

where $\eta_{ij} = -\beta\mu_{ij} + \epsilon_{ij}$

The error term in 151 is clearly correlated with the regressor $\tilde{S}_{ij}$. So the OLS estimate of β is biased in the following way:

$$\begin{aligned} E(b_{OLS}) &= \frac{Cov(Y, \tilde{S})}{Var(\tilde{S})} \\ &= \beta + \frac{Cov(\eta, \tilde{S})}{Var(\tilde{S})} \\ &= \beta - \beta\frac{Cov(\mu, \tilde{S})}{Var(\tilde{S})} \\ &= \beta - \beta\frac{Var(\mu)}{Var(S) + Var(\mu)} \\ &= \beta\left(1 - \frac{Var(\mu)}{Var(S) + Var(\mu)}\right) = \beta\left(1 - \frac{Var(\mu)}{Var(\tilde{S})}\right) \end{aligned} \tag{152}$$

Because of measurement error, OLS underestimates the true parameter.

The attenuation bias is larger the larger the "(un)-reliability ratio", i.e. the ratio between the variance of the noise and the variance of the signal. Note that $0 < \frac{Var(\mu)}{Var(\tilde{S})} < 1$.

Measurement error in panel data

Going back to a panel data framework, the presence of unobservable individual specific fixed effects (i.e $\alpha_j \neq \alpha$) combined with measurement error of the observed regressors X causes OLS to be biased for two reasons:

- the omitted variable bias due to the fact that OLS does not control for the individual specific fixed effects (see section 4.2).
- the attenuation bias caused by measurement error (see equation 152);

If we use a fixed effect (within) estimator,

- we eliminate the bias due to the omission of the fixed effects;
- but the measurement error bias can be larger or smaller and under plausible assumptions will be larger.

To see this, consider for example the true model in first differences:

$$y_{1j} - y_{2j} = \beta[S_{1j} - S_{2j}] + \epsilon_{1j} - \epsilon_{2j} \tag{153}$$

However, the model that we can actually estimate is:

$$\begin{aligned} Y_{1j} - Y_{2j} &= \beta[\tilde{S}_{1j} - \tilde{S}_{2j}] - \beta[\mu_{1j} - \mu_{2j}] + \epsilon_{1j} - \epsilon_{2j} \\ Y_{1j} - Y_{2j} &= \beta[\tilde{S}_{1j} - \tilde{S}_{2j}] + \phi_j \end{aligned} \tag{154}$$

where $\phi_j = -\beta[\mu_{1j} - \mu_{2j}] + \epsilon_{1j} - \epsilon_{2j}$

Note that in equation 154 we have one observation per family and the error term ϕ is correlated with the regressor.

The OLS estimate of 154 gives the fixed effect (within) estimator but is biased by the existence of measurement error. Note that the bias can be easily computed using the formula 152 for the standard (non-panel) case:

$$E(b_{FE}) = \beta \left(1 - \frac{Var(\mu_{1j} - \mu_{2j})}{Var(S_{1j} - S_{2j}) + Var(\mu_{1j} - \mu_{2j})}\right) = \beta \left(1 - \frac{Var(\mu_{1j} - \mu_{2j})}{Var(\tilde{S}_{1j} - \tilde{S}_{2j})}\right) \tag{155}$$

In order to simplify and interpret this expression we have to make some assumptions on the correlation structure of these variables.

- The measurement error of each twin is uncorrelated with his/her own true schooling:

$$Cov(\mu_{ij}, S_{ij}) = 0 \tag{156}$$

  so that

$$Var(\tilde{S}_{ij}) = Var(S_{ij}) + Var(\mu_{ij}) \tag{157}$$

- The measurement errors have the same variance:

$$Var(\mu_{1j}) = Var(\mu_{2j}) = Var(\mu) \tag{158}$$

- The measurement errors of the two twins can be correlated, so that:

$$Var(\mu_{1j} - \mu_{2j}) = 2Var(\mu) - 2Cov(\mu_1, \mu_2) \tag{159}$$

- The true schooling levels have the same variance:

$$Var(S_{1j}) = Var(S_{2j}) = Var(S) \tag{160}$$

- The measured schooling levels of the two twins may be correlated because true schooling levels are correlated and because measurement errors are correlated; hence:

$$\begin{aligned} Var(\tilde{S}_{1j} - \tilde{S}_{2j}) &= Var(\tilde{S}_{1j}) + Var(\tilde{S}_{2j}) - 2Cov(\tilde{S}_1, \tilde{S}_2) \\ &= 2Var(S) + 2Var(\mu) - 2Cov(\tilde{S}_1, \tilde{S}_2) \end{aligned} \tag{161}$$

Using these assumptions, we can rewrite equation 155 as:

$$E(b_{FE}) = \beta \left(1 - \frac{Var(\mu)[1 - Corr(\mu_1, \mu_2)]}{[Var(S) + Var(\mu)][1 - Corr(\tilde{S}_1, \tilde{S}_2)]}\right) \tag{162}$$

where:

$$Corr(\mu_1, \mu_2) = \frac{Cov(\mu_1, \mu_2)}{Var(\mu)} = \frac{Cov(\mu_1, \mu_2)}{Var(\mu_1)^{1/2} Var(\mu_2)^{1/2}} \tag{163}$$

$$Corr(\tilde{S}_1, \tilde{S}_2) = \frac{Cov(\tilde{S}_1, \tilde{S}_2)}{Var(\tilde{S})} = \frac{Cov(\tilde{S}_1, \tilde{S}_2)}{Var(\tilde{S}_1)^{1/2} Var(\tilde{S}_2)^{1/2}} \tag{164}$$

This result shows that in panel data the bias due to measurement error can be larger or smaller than in the standard non-panel case (see Griliches and Hausman, 1986).

- If $Corr(\mu_1, \mu_2) < Corr(\tilde{S}_1, \tilde{S}_2)$ the bias is larger. In particular, in the classical case in which measurement errors are uncorrelated ($Corr(\mu_1, \mu_2) = 0$), an error that would cause a small bias in the cross-sectional case may have very big effects in the panel case.
  - The intuition is that, in relative terms, the variance of the signal is reduced by first differencing $\tilde{S}$, while the variance of the noise is unchanged because the errors are independent.
- If $Corr(\mu_1, \mu_2) > Corr(\tilde{S}_1, \tilde{S}_2)$ instead the bias may be smaller in panel data.
  - Less likely to happen; shows usefulness of validation studies.

Ashenfelter and Krueger provide a clever (but not general ...) solution to the probelm of measurement error in panel data. Note that they assume $Corr(\mu_1, \mu_2) = 0$. (See slides from paper.)

4.4.8 Fixed effects estimators and lagged dependent variables

Fixed effect (within) estimation becomes problematic when the model includes lagged dependent variables as in

$$Y_{it} = \alpha_i + Y_{it-1}\rho + X_{it}\beta + \epsilon_{it} \tag{165}$$

where $E(\epsilon_{it}) = 0$.

Note that a similar specification can be obtained also from the following different initial assumptions:

$$Y_{it} = a_i + \tilde{X}_{it}\beta + u_{it} \tag{166}$$

$$u_{it} = \rho u_{it-1} + \epsilon_{it} \tag{167}$$

because substitution of 167 in 166 gives:

$$Y_{it} = a_i(1-\rho) + \rho Y_{it-1} + (\tilde{X}_{it} - \rho\tilde{X}_{it-1})\beta + \epsilon_{it} \tag{168}$$

$$Y_{it} = \alpha_i + Y_{it-1}\rho + X_{it}\beta + \epsilon_{it} \tag{169}$$

which is equal to 165 if $\alpha_i = a_i(1-\rho)$ and $X_{it} = (\tilde{X}_{it} - \rho\tilde{X}_{it-1})$.

The problem in the estimation of 165 is that Y_{it-1} is a predetermined variable but not a strictly exogenous variable.

With weakly exogenous (predetermined) regressors, the standard techniques to deal with the unobservable heterogeneity represented by the α_i fail.

Both the LSDV estimator and the Analyis of Covariance estimator of ρ and β will be biased and inconsistent.

LSDV and lagged dependent variables

To understand the reason of the bias let's define the $N + K + 1$ row vector:

$$Z_{it} = [D_{it} \quad Y_{it-1} \quad X_{it}] \tag{170}$$

of the observed fixed effect dummies, lagged dependent variable and explanatory factors for individual i at time t and the $N + K + 1$ column vector

$$\gamma = \begin{bmatrix} \alpha \\ \rho \\ \beta \end{bmatrix} \tag{171}$$

where α is the column vector of the N individual specific fixed effects α_i.

Using this notation the model can be written as

$$Y_{it} = Z_{it}\gamma + \epsilon_{it}. \tag{172}$$

Denoting with $\mathbf{Z}$ the matrix of the observations on Z_{it}, the LSDV estimator is biased because:

$$E(Y|\mathbf{Z}) = \mathbf{Z}\gamma + E(\epsilon|\mathbf{Z}) \neq \mathbf{Z}\gamma \tag{173}$$

Note that not only the estimate of ρ but also the estimates of β and α are biased by the failure of the orthogonality condition

$$E(\epsilon|\mathbf{Z}) = 0 \tag{174}$$

The orthogonality condition $E(\epsilon|\mathbf{Z}) = 0$ fails because, while $E(\epsilon|\mathbf{D}) = E(\epsilon|\mathbf{X}) = 0$, we have instead that

$$E(\epsilon|Y) \neq 0 \tag{175}$$

It is important to understand that 175 requires every element of ϵ to be uncorrelated with every element of Y, i.e.:

$$E(\epsilon_{it}|Y_{is}) = 0 \qquad \text{for all } s \text{ and all } t \tag{176}$$

We can only say that

$$E(\epsilon_{it}|Y_{is}) = 0 \qquad \text{for } s < t \tag{177}$$

but

$$E(\epsilon_{it}|Y_{is}) \neq 0 \qquad \text{for } s \geq t \tag{178}$$

Analysis of Covariance and lagged dependent variables

The Analysis of Covariance approach does not solve the problem because when we partial out the fixed effects and we consider the model in deviation from the individual specific means we obtain:

$$[Y_{it} - \bar{Y}_{i.}] = [Y_{it-1} - \bar{Y}_{i.-1}]\rho + [X_{it} - \bar{X}_{i.}]\beta + [\epsilon_{it} - \bar{\epsilon}_{i.}] \tag{179}$$

and the orthogonality condition required for an unbiased estimate of β and ρ does not hold:

$$E([Y_{is-1} - \bar{Y}_{i.-1}][\epsilon_{it} - \bar{\epsilon}_{i.}]) \neq 0 \qquad \text{for all } s \text{ and all } t \tag{180}$$

Note that in this case the orthogonality condition fails because

$$E(\bar{Y}_{i.-1}\ \epsilon_{it}) \neq 0 \tag{181}$$

$$E(Y_{is-1}\ \bar{\epsilon}_{i.}) \neq 0 \tag{182}$$

Nickell (1981) gives analytical expressions for the bias due to lagged dependent variables and shows that:

- the bias goes to zero when T goes to infinity, but
- the bias *does not* go to zero when N goes to infinity;

Since the typical panel data has a large N but a small T this result is disturbing.

We give an example of the inconsistency of fixed effect estimators in a simple case with $T = 3$.

An example for T = 3

We have 3 time observations for each individual and we include one lag of the dependent variable as explanatory factor. We cannot use, for the estimation, the first observation on each individual, and we can only focus on:

$$Y_{i2} = \alpha_i + Y_{i1}\rho + \epsilon_{i2} \tag{183}$$
$$Y_{i3} = \alpha_i + Y_{i2}\rho + \epsilon_{i3} \tag{184}$$

where $E(\epsilon_{it}) = 0$ and for simplicity we have omitted all the other exogenous explanatory factors X.

We know that to obtain the FE (within) estimator we can equivalently use the LSDV model, the analysis of covariance, or first differencing. The third approach is numerically equivalent to the others because we have effectively only two time observations. It is easier to compute the bias using this third approach:

$$Y_{i3} - Y_{i2} = (Y_{i2} - Y_{i1})\rho + \epsilon_{i3} - \epsilon_{i2} \tag{185}$$

The probability limit for $N \to \infty$ of the OLS estimator of ρ using 185 is:

$$\begin{aligned} \underset{N\to\infty}{P\lim}\, \hat{\rho}_{FE} &= \frac{\frac{1}{N}\sum_{i=1}^{N}(Y_{i3} - Y_{i2})(Y_{i2} - Y_{i1})}{\frac{1}{N}\sum_{i=1}^{N}(Y_{i2} - Y_{i1})^2} \\ &= \frac{E[(Y_{i3} - Y_{i2})(Y_{i2} - Y_{i1})]}{E[(Y_{i2} - Y_{i1})^2]} \end{aligned} \tag{186}$$

Substituting the expectation of 185 in 186 we obtain:

$$\begin{aligned} \underset{N\to\infty}{P\lim}\, \hat{\rho}_{FE} &= \rho + \frac{E[(\epsilon_{i3} - \epsilon_{i2})(Y_{i2} - Y_{i1})]}{E[(Y_{i2} - Y_{i1})^2]} \\ &= \rho - \frac{E(\epsilon_{i2}^2)}{E[(Y_{i2} - Y_{i1})^2]} \end{aligned} \tag{187}$$

And the bias does not go away even if the number of individuals goes to infinity.

A solution based on first differencing and instrumental variables

Going back to the general model

$$Y_{it} = \alpha_i + Y_{it-1}\rho + X_{it/}\beta + \epsilon_{it} \tag{188}$$

we have seen that the LSDV approach, the Analysis of Covariance do not allow to control for the α_i and to fix at the same time the failure of the orthogonality conditions.

However, suppose that $T > 2$, and consider the model in first differences

$$Y_{it} - Y_{it-1} = (Y_{it-1} - Y_{it-2})\rho + (X_{it} - X_{i-1})\beta + \epsilon_{it} - \epsilon_{it-1} \tag{189}$$

with this transformation:

- we have eliminated the α_i;
- variables like Y_{it-2}, $(Y_{it-2} - Y_{it-3})$, X_{it-1} and $(X_{it-1} - X_{it-2})$ appear to be valid instruments.

More generally, if only one lag of the dependent variable is included in the model, the following orthogonality conditions hold:

$$E[y_{is}(\epsilon_{it} - \epsilon_{it-1})] = 0 \qquad \text{for all } s < t-1 \tag{190}$$

while if the X are truly exogenous

$$E[x_{is}(\epsilon_{it} - \epsilon_{it-1})] = 0 \qquad \text{for all } s \text{ and all } t \tag{191}$$

If more lags of the dependent variable are included in the model, we have of course to go further back with lags in order to find valid instruments.

Hence,

- if we have observations on a sufficient number of lags,
- using as instruments the appropriate lags of the dependent and independent variables (either in levels or in differences) ,

we can estimate the model in first differences (equation 189) obtaining consistent estimates of β and ρ.

For alternative solution based on a similar intuition see also Holz-Eakin, Newey and Rosen (1988), Arellano and Bover (1990) and Keane and Runkle (1992).

Arellano and Bond (1991) provide three specification tests for the validity of the instruments in the procedure described above. Note, for example, that the lagged variables could not be valid as instruments if the error term ϵ_{it} is autocorrelated.

4.4.9 Other pitfalls of fixed effects estimation

- **Waste of "between" information.**
  Fixed effect (within) estimators ignore the information offered by the comparison between individuals
  - It is conceivable that, under appropriate assumptions, an estimator capable to exploit also the "between individuals" variation would be more efficient.
- **Loss of degrees of freedom.**
  Due to the estimation of the fixed effects and particularly relevant when the N dimension is large.
  - Further (and related to above) loss of efficiency.
- **Effect of time invariant explanatory factors.**
  The transformations which deliver the fixed effect estimator, eliminate all the time invariant explanatory factors. Therefores the effect of these factors on the outcome cannot be estimated.
  - We can estimate the change of the effect of time invariant variables but only if we are willing to assume that this effect is time varying (see Maggi and Ichino, 1999)
- **Out of sample predictions.**
  The individual effects are not assumed to have a distribution but are instead treated as fixed and estimable parameters, which may lead to difficulties when making out of sample predictions.

In order to overcome the pittfalls of fixed effect (within) estimation, other approaches have been proposed in the literature.

However, the pittfals of fixed effect estimation are overcome by these approaches at the cost of assumptions which are sometimes even more unpleasent in particular from the viewpoint of labor economics.

4.5 Between estimator

At the opposite extreme of the fixed effect (within) analysis, the basic model

$$Y_{it} = \alpha_i + X_{it}\beta + \epsilon_{it} \tag{192}$$

can be transformed to fully exploit the variability "between individuals", ignoring completely the variability "within individuals".

Let $\alpha_i = \alpha + \nu_i$. Whatever the properties of α_i and ϵ_{it}, if 192 is the true model also the following must be true.

$$\begin{aligned} \bar{Y}_{i.} &= \alpha + \bar{X}_{i.}\beta + \nu_i + \bar{\epsilon}_{i.} \\ \bar{Y}_{i.} &= \alpha + \bar{X}_{i.}\beta + \eta_{i.} \end{aligned} \tag{193}$$

where:

- $\bar{Y}_{i.} = \frac{1}{T}\sum_{t=1}^{T} Y_{it}$ is the mean of the outcomes observed for each individual over time;
- $\bar{X}_{i.} = \frac{1}{T}\sum_{t=1}^{T} X_{it}$ is the row vector of the means of the explanatory factors observed for each individual over time;
- $\bar{\epsilon}_{i.} = \frac{1}{T}\sum_{t=1}^{T} \epsilon_{it}$ is the mean of the residuals for each individual over time;
- $\eta_i = \nu_i + \bar{\epsilon}_{i.}$ is the composite error term of the model.

Note that 193 involves N observations on the means for each individual.

To estimate with OLS a model like 193, we need the standard assumptions on ϵ_{it}, but we do not need to assume that the individual specific effects are fixed and estimable, as in the case of fixed effect estimation.

However, these individual effects are now included in the error term, and therefore we have to assume that the individual specific effects are uncorrelated with the explanatory factors:

$$COV(\eta_i, \bar{X}_{i.}) = COV(\nu_i, \bar{X}_{i.}) = 0 \tag{194}$$

Under this assumption, OLS applied to equation 193 gives an unbiased and consistent estimate of β:

$$E(b_{BE}) = \beta \tag{195}$$

b_{BE} is usually called between estimator.

Note that, since b_{BE} ignores the information offered by the within variability, it will not in general be efficient.

It seems then natural to search for transformations of the data that could exploit both the "within" and the "between" variability in order to gain efficiency with respect to both the within and the between estimators.

4.5.1 OLS, "within" and "between" estimators

In search for an estimator that exploits both the within and the between variability of the data we begin by observing that OLS can be expressed as a weighted average of the within and the between estimators.

Let's add to our notation the following definitions

- Overall means of the outcomes and of the explanatory factors:

$$\bar{\bar{Y}} = \sum_{i=1}^{n} \sum_{t=1}^{T} Y_{it} \tag{196}$$

$$\bar{\bar{X}} = \sum_{i=1}^{n} \sum_{t=1}^{T} X_{it} \tag{197}$$

- Moment matrices of the overall sums of squares and cross products:

$$S_{xx}^{0} = \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{\bar{x}})(x_{it} - \bar{\bar{x}})' \tag{198}$$

$$S_{xy}^{0} = \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{\bar{x}})(y_{it} - \bar{\bar{y}}). \tag{199}$$

- Moment matrices of the "within" sums of squares and cross products:

$$S_{xx}^{w} = \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x}_{i.})(x_{it} - \bar{x}_{i.})' \tag{200}$$

$$S_{xy}^{w} = \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x}_{i.})(y_{it} - \bar{y}_{i.}). \tag{201}$$

- Moment matrices of the "between" sums of squares and cross products:

$$S_{xx}^{b} = \sum_{i=1}^{n} T(\bar{x}_{i.} - \bar{\bar{x}})(\bar{x}_{i.} - \bar{\bar{x}})' \tag{202}$$

$$S_{xy}^{b} = \sum_{i=1}^{n} T(\bar{x}_{i.} - \bar{\bar{x}})(\bar{y}_{i.} - \bar{\bar{y}}). \tag{203}$$

It is easy to verify that:

$$S_{xx}^O = S_{xx}^w + S_{xx}^b \tag{204}$$

$$S_{xy}^O = S_{xy}^w + S_{xy}^b. \tag{205}$$

Then, the three estimators of β that we have examined so far are:

- Fixed effect (within) estimator (which so far we indicated as b_{FE}):

$$b^w = [S_{xx}^w]^{-1} S_{xy}^w \tag{206}$$

- Between estimator:

$$b^b = [S_{xx}^b]^{-1} S_{xy}^b \tag{207}$$

- OLS estimator:

$$b^O = [S_{xx}^O]^{-1} S_{xy}^O = [S_{xx}^w + S_{xx}^b]^{-1} [S_{xy}^w + S_{xy}^b]. \tag{208}$$

Note that we can write:

$$S_{xy}^w = S_{xx}^w b^w \tag{209}$$

$$S_{xy}^b = S_{xx}^b b^b \tag{210}$$

Substituting 209 and 210 in the OLS estimator 208 we obtain:

$$b^O = F^w b^w + F^b b^b, \tag{211}$$

where

$$F^w = [S_{xx}^w + S_{xx}^b]^{-1} S_{xx}^w = I - F^b. \tag{212}$$

which shows that the OLS estimator can be interpreted as weighted average of the within and between estimators with weights that depend on the "within" versus "between" variability of the explanatory factors.

However:

- in general this is not the most efficient way to exploit jointly the within and between variability;
- it leads to biased and inconsistent estimates of the true causal effect β if the individual specific effects are correlated with the regressors.

4.6 Random effects estimator

Starting again from our basic model:

$$\begin{aligned} Y_{it} &= \alpha + X_{it}\beta + \nu_i + \epsilon_{it} \\ Y_{it} &= \alpha + X_{it}\beta + w_{it} \end{aligned} \tag{213}$$

in order to exploit efficiently the between and within variation we have to:

- abandon the assumption that the individual effects are fixed and estimable;
- assume that they measure our individual specific ignorance which should be treated similarly to our general ignorance ϵ_{it};
- assume that the composite error term is uncorrelated with the regressors;
- explicit carefully the covariance structure of the two types of ignorance.

A starting set of assumptions on the covariance structure is the following:

$$E[\varepsilon_{it}|\mathbf{X}] = E[\nu_i|\mathbf{X}] = 0, \tag{214}$$

$$E[\varepsilon_{it}^2|\mathbf{X}] = \sigma_\varepsilon^2, \tag{215}$$

$$E[\nu_i^2|\mathbf{X}] = \sigma_\nu^2, \tag{216}$$

$$E[\varepsilon_{it}\nu_j|\mathbf{X}] = 0 \qquad \text{for all}\, i, t, \text{and}\, j, \tag{217}$$

$$E[\varepsilon_{it}\varepsilon_{js}|\mathbf{X}] = 0 \qquad \text{if}\, t \neq s \text{or}\, i \neq j, \tag{218}$$

$$E[\nu_i\nu_j|\mathbf{X}] = 0 \qquad \text{if}\, i \neq j. \tag{219}$$

In terms of the composite error term

$$w_{it} = \nu_i + \varepsilon_{it} \tag{220}$$

these assumptions imply:

$$E[w_{it}^2|\mathbf{X}] = \sigma_\varepsilon^2 + \sigma_\nu^2, \tag{221}$$

$$E[w_{it}w_{is}|\mathbf{X}] = \sigma_\nu^2 \qquad \text{for } t \neq s. \tag{222}$$

For the T observations on individual i let:

- $w_i = [w_{i1}, w_{i2}, \ldots, w_{iT}]'$
- $\mathbf{\Omega} = E(w_i w_i')$

so that

$$\mathbf{\Omega} = \begin{bmatrix} \sigma_\varepsilon^2 + \sigma_\nu^2 & \sigma_\nu^2 & \sigma_\nu^2 & \ldots & \sigma_\nu^2 \\ \sigma_\nu^2 & \sigma_\varepsilon^2 + \sigma_\nu^2 & \sigma_\nu^2 & \ldots & \sigma_\nu^2 \\ & \vdots & & & \\ \sigma_\nu^2 & \sigma_\nu^2 & \sigma_\nu^2 & & \sigma_\varepsilon^2 + \sigma_\nu^2 \end{bmatrix} = \sigma_\epsilon^2 \mathbf{I} + \sigma_\nu^2 \mathbf{i}\mathbf{i}' \tag{223}$$

where $\mathbf{i}$ is a T column vector of 1s.

Hence the covariance matrix of the error term w_{it} in the basic model 213 is:

$$V = \mathbf{I} \otimes \mathbf{\Omega} = \begin{bmatrix} \Omega & 0 & 0 & \cdots & 0 \\ 0 & \Omega & 0 & \cdots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \cdots & \Omega \end{bmatrix}. \tag{224}$$

which clearly implies that OLS estimates of 213 are inefficient and the method of Generalised Least Squares (GLS) is necessary for efficiency.

Note that this covariance structure implies that:

- the error terms for different units i are uncorrelated;
- the error terms of the same unit i in two different periods t and s are correlated independently of the distance between t and s.

This covariance structure makes probably more sense if i indicates families and t individuals within families.

When t is really time, a decreasing correlation across time (but within individuals) would probably make more sense.

4.6.1 GLS estimation of Random Effects models

Assuming that $\mathbf{V}$ is known, the GLS estimation of the basic model

$$Y_{it} = \alpha + X_{it}\beta + [\nu_i + \epsilon_{it}] \tag{225}$$

implies the estimation of the transformed model

$$\mathbf{V}^{-\frac{1}{2}} Y_{it} = \alpha + \mathbf{V}^{-\frac{1}{2}} X_{it}\beta + \mathbf{V}^{-\frac{1}{2}} w_{it} \tag{226}$$

where

$$\mathbf{V}^{-\frac{1}{2}} = \mathbf{I} \otimes \mathbf{\Omega}^{-\frac{1}{2}} \tag{227}$$

and

$$\mathbf{\Omega}^{-\frac{1}{2}} = I - \frac{\theta}{T} ii' \tag{228}$$

and

$$\theta = 1 - \frac{\sigma_\varepsilon}{\sqrt{T\,\sigma_\nu^2 + \sigma_\varepsilon^2}} \tag{229}$$

This implies that the transformation for each individual-time observation is

$$Y_{it} - \theta \bar{Y}_{i.} = (1-\theta)\alpha + (X_{it} - \theta \bar{X}_{i.})\beta + [(1-\theta)\nu_i + (\epsilon_{it} - \theta \bar{\epsilon}_{i.})] \tag{230}$$

The transformed model can be estimated with OLS to obtain an efficient and consistent estimate of β under the assumptions 214–219.

Note that:

- if $\sigma_v^2 = 0 \rightarrow \theta = 0$
  in which case the random effect estimator is identical to the OLS estimator on the pooled individual-time observations, because there is no individual heterogeneity ($\nu_i = 0$).
- if $\sigma_\epsilon^2 = 0 \rightarrow \theta = 1$
  in which case the only existing ignorance would be the individual–specific one captured by ν_i and the random effect estimator would be identical to the fixed effect estimator.

4.6.2 Feasible GLS estimation of random effects models

Since the covariance matrix $\mathbf{V}$ is usually not known, a feasible GLS procedure has to be adopted.

A standard procedure (see Greene or the STATA manual) is the folliwing:

- use the fixed effects specification to get b_{FE} which is a consistent but inefficient estimate of β;
- use b_{FE} to get the fixed effect (within) estimated residuals and then compute a consistent estimate of σ_ϵ^2;
- use similarly the between specification to get a consistent estimate of $T\sigma_\nu^2 + \sigma_\epsilon^2$;
- use the estimates obtained above to compute an estimate of $\theta = 1 - \frac{\sigma_\varepsilon}{\sqrt{T\,\sigma_\nu^2+\sigma_\varepsilon^2}}$;
- Using the estimated θ, apply the GLS transformation to the data and estimate equation 230 to get the consistent and efficient estimate b_{RE} of β.

4.6.3 Random effects, within, between and OLS estimators

Like the OLS estimator, also the random effect estimator can be interpreted as a weighted average of the within and between estimator.

Using the same notation introduced in section 4.5.1 it can be shown (see Maddala, 1971 and Hausman and Taylor, 1981) that:

$$b_{RE} = \tilde{F}^w b^w + (I - \tilde{F}^w) b^b \tag{231}$$

where

$$\tilde{F}^w = [S^w_{xx} + \lambda S^b_{xx}]^{-1} S^w_{xx}, \tag{232}$$

$$\lambda = \frac{\sigma^2_\varepsilon}{\sigma^2_\varepsilon + T\sigma^2_\nu} = (1-\theta)^2. \tag{233}$$

Note again that:

- If $\sigma^2_v = 0 \rightarrow \lambda = 1$:
  the random effect estimator is identical to the OLS estimator because there is no individual heterogeneity; in this case the OLS estimator is the most efficient and in particular more efficient than the within estimator because it uses both the within and the between information.

- If $\sigma^2_v > 0 \rightarrow \lambda < 1$:
  the OLS estimator would put too much weight on the between information, because it imposes $\lambda = 1$ while the best estimator (GLS) uses the correct $\lambda < 1$.

- If $\sigma^2_\epsilon = 0 \rightarrow \lambda = 0$:
  the only uncertainty is generated by the individual heterogeneity. In this case the best (GLS) estimator coincides wiht the within estimator while again OLS would put too much weight on the between information.

So the random effect estimator seems preferable because

- it uses efficiently the between and within information;
- it coincides with the within or OLS estimator when the efficient use of the information requires to do so;
- it allows for the estimation of the effects of time invariant explanatory factors;
- it can be used more convincingly for out of sample predictions.

However, the random effect estimator is consistent only when the individual specific effects are not correlated with the explanatory factors: i.e. it requires

$$COV(\varepsilon_{it}\mathbf{X}) = COV(\nu_i\mathbf{X}) = 0 \tag{234}$$

This is a crucial assumption but:

- it is hard to find it convincing in most labor applications;
- it should anyway be tested before accepting a random effect specification;
- if rejected, appropriate solutions (IV estimation) should be adopted in order to maintain a random effect specification.

We will consider below how to test for the orthogonality condition and what to do in case of a rejection,

Before, however, we want to explore an attempt, proposed by Mundlack (1978), to reconcile the fixed effects and random effects models.

4.7 Mundlak (1978): a reconciliation of fixed and random effects models?

Mundlack (1978) suggests that:

- the distinction between random and fixed effects models is "arbitrary and unnecessary".
- "when the model is properly specified the GLS-random effect estimator is identical to the fixed effect (within) estimator; thus there is in fact only one estimator."
- "The whole literature which has been based on an imaginary difference between the two estimators, starting with Balestra and Nerlove (1966), is based on an incorrect specification which ignores the correlation between the effects and the explanatory variables in the regression."

Given the basic model

$$Y_{it} = X_{it}\beta + \alpha_i + \epsilon_{it} \tag{235}$$

where α_i is potentially correlated with X_{it}. The starting point of Mundlack's contribution is to take an explicit account of such relationship assuming that

$$\alpha_i = \bar{X}_{i.}\delta + u_i \tag{236}$$

where a crucial assumption is that:

$$E(u_i|X) = 0 \tag{237}$$

As long as $\delta \neq 0$

$$E(\alpha_i|X) = \bar{X}_{i.}\delta \neq 0 \tag{238}$$

and the GLS random effect estimation of 235 would give biased and inconsistent estimates of β.

Note that this is an attempt to model explicitly the individual heterogeneity. We have seen another, but different, attempt to do so in Ashenfelter and Kruger (1994).

Substituting 236 into 235 we obtain

$$Y_{it} = X_{it}\beta + \bar{X}_{i.}\delta + [u_i + \epsilon_{it}] \tag{239}$$

and now note that this is a well specified random effect model, because

$$E(u_i + \epsilon_{it}|X) = 0 \tag{240}$$

Equation 239 can be rearranged to obtain

$$Y_{it} = (X_{it} - \bar{X}_{i.})\beta + \bar{X}_{i.}(\beta + \delta) + [u_i + \epsilon_{it}] \tag{241}$$

$$Y_{it} = (X_{it} - \bar{X}_{i.})\phi + \bar{X}_{i.}\psi + [u_i + \epsilon_{it}] \tag{242}$$

Mundlack (1978) shows that if we apply the GLS transformation to 242 we obtain that:

$$\hat{\phi}_{GLS} = b_{FE} \quad \rightarrow \quad E(b_{FE}) = E(\hat{\phi}_{GLS}) = \beta \tag{243}$$

$$\hat{\psi}_{GLS} = b_{BE} \quad \rightarrow \quad E(b_{BE}) = E(\hat{\psi}_{GLS}) = \beta + \delta \tag{244}$$

which shows that, if 239 is the specification that takes into account correctly the correlation between individual effects and regressors,

- the fixed effect estimator coincides with the GLS estimator and is unbiased for β;
- the between estimator is unbiased for β only if the orthogonality condition holds (i.e. $\delta = 0$).

This conclusion, however,

- is crucially based on the linear specification 236 of the correlation between α_i and X;
- disregards efficiency considerations in the comparison between random effects and fixed effects estimators.

4.7.1 A test for random or fixed effects

Since the random effect estimator appears superior when the individual specific effects are orthogonal to the regressors, it would be nice to test this orthogonality condition.

The framework leads naturally to a Hausman's specification test (see Hausman, 1978 and Hausman and Taylor, 1981).

Given the random effect specification

$$Y_{it} = \alpha + X_{it/}\beta + [\nu_i + \epsilon_{it}] \tag{245}$$

and assuming that the orthogonality condition holds for the ϵ_{it}, the null hypothesis that we want to test is:

$$H_o : \qquad E(\nu_i|X_{it}) = 0 \tag{246}$$

while the alternative hypothesis is:

$$H_1 : \qquad E(\nu_i|X_{it}) \neq 0 \tag{247}$$

- Under H_o:
  - the fixed effect estimator b_{FE} is consistent but inefficient;
  - the random effect estimator b_{RE} is consistent *and* efficient;
- Under H_1:
  - the fixed effect estimator b_{FE} remains consisitent;
  - the random effect estimator b_{RE} becomes inconsistent;

Therefore, under the null hypothesis the two estimators should not differ and this observation provides the basis for the test.

The test statistic that follows from this intuition is:

$$W = [b_{FE} - b_{RE}]'[Var(b_{FE} - b_{RE})]^{-1}[b_{FE} - b_{RE}] \tag{248}$$

but the problem is how to compute the covariance matrix of the difference between the two estimators: $Var(b_{FE} - b_{RE})$.

The crucial result of Hausman (1978) is to show that in general the covariance of an efficient estimator with its difference from an inefficient estimator is zero, which in our case implies:

$$Cov(b_{RE}, [b_{FE} - b_{RE}]) = Cov(b_{RE}, b_{FE}) - Var(b_{RE}) = 0 \tag{249}$$

Using this result we can write:

$$\begin{aligned} Var(b_{FE} - b_{RE}) &= Var(b_{FE}) + Var(b_{RE}) - Cov(b_{RE}, b_{FE}) - Cov(b_{RE}, b_{FE})' \\ Var(b_{FE} - b_{RE}) &= Var(b_{FE}) - Var(b_{RE}) \end{aligned} \tag{250}$$

Therefore the test statistic can be written as:

$$W = [b_{FE} - b_{RE}]' \, [Var(b_{FE}) - Var(b_{RE})]^{-1} \, [b_{FE} - b_{RE}] \tag{251}$$

which is asymptotically distributed as a chi-squared with K degrees of freedom (i.e. the dimension of the vector of parameters β to be estimated).

If W is "greater" than the preferred critical value it means that there is a statistically significant difference between the two estimators,

Since only b_{FE} is consistent we have to conclude that b_{RE} is inconsistent because the orthogonality condition fails.

Hausman and Taylor (1981) describe two other tests based on the same intuition but applied to the comparison of

- the random effect and the between estimator,
- the within and the between estimators,

and show that both these tests are identical to the one described above in contrast with what was previously thought.

Hausman (1978) proposes also a convenient regression format for the test.

Consider the regression:

$$Y_{it} - \theta Y_{it} = (1-\theta)\alpha + (X_{it} - \theta \bar{X}_{i.})\beta + (X_{it} - \bar{X}_{i.})\delta + u_{it} \tag{252}$$

where $\theta = 1 - \frac{\sigma_\varepsilon}{\sqrt{T\,\sigma_\nu^2 + \sigma_\varepsilon^2}}$.

Note that this regressions amounts to estimate with OLS the "GLS-transformed" data adding also the "within-transformed" explanatory factors (i.e. the simple deviations from the individual specific means) to the regression.

The Hausman test described above for $H_o : E(\nu_i | X_{it}) = 0$ is equivalent to a test that the parameters δ are equal to 0 in the auxiliary regression 252.

4.7.2 Random effects models and Instrumental Variables

Consider the model

$$Y_{it} = X_{it}\beta + Z_i\gamma + \alpha_i + \epsilon_{it} \tag{253}$$

in which we would like to consider the individual effects α_i as random but the Hausman test rejects the null hypothesis that they are uncorrelated with the regressors.

In this situation there are two solutions:

- We could abandon the random effect specification for a fixed effects specification,
  - but we would run into the problems of fixed effect estimation and in particular we would not be able to estimate γ.
- We could search for instruments , i.e. variables satisfying the following two conditions:
  - they should not be correlated with the individual specific effects α_i;
  - they should be correlated with the regressors X_{it} andZ_i.

Hausman and Taylor (1981) (HT), Amemiya and MacCurdy (1986) (AM) and Breush, Mizon and Schmid (1989) (BMS) show how the panel structure of the data could be exploited to find (or better construct) these instruments.

Cornwell and Rupert (1988) compare the efficiency of the three procedures.

The intuition of HT is that the panel structure of the data may offer good instruments if we are ready to assume that a sufficient number of explanatory factors are asymptotically uncorrelated with the individual specific effects.

Given the model:

$$Y_{it} = X_{it}\beta + Z_i\gamma + \alpha_i + \epsilon_{it} \tag{254}$$

assume that:

- $X_{it} = [X_{1it} : X_{2it}]$
  denotes a row of the $NT \times K$ matrix of time varying explanatory factors which can be divided in two sub-matrices:
  - X_{1it} denotes the K_1 exogenous time varying factors for which:

$$\underset{N\to\infty}{P\lim} \frac{1}{N} X_{1it}\alpha_i = 0 \tag{255}$$

  - X_{2it} denotes the K_2 correlated time varying factors for which:

$$\underset{N\to\infty}{P\lim} \frac{1}{N} X_{2it}\alpha_i \neq 0 \tag{256}$$

- $Z_i = [X_{1i} : X_{2i}]$
  denotes a row of the $NT \times G$ matrix of time invariant explanatory factors which can be divided in two sub-matrices:
  - Z_{1i} denotes the G_1 exogenous time invariant factors for which:

$$\underset{N\to\infty}{P\lim} \frac{1}{N} Z_{1i}\alpha_i = 0 \tag{257}$$

  - Z_{2i} denotes the G_2 correlated time invariant factors for which:

$$\underset{N\to\infty}{P\lim} \frac{1}{N} Z_{2i}\alpha_i \neq 0 \tag{258}$$

Consider the GLS transformation of the model:

$$Y_{it} - \theta \bar{Y}_{i.} = (X_{it} - \theta \bar{X}_{i.})\beta + (1-\theta) Z_i \gamma + [(1-\theta)\alpha_i + (\epsilon_{it} - \theta \bar{\epsilon}_{i.})] \tag{259}$$

where $\theta = 1 - \frac{\sigma_\varepsilon}{\sqrt{T\,\sigma_\nu^2 + \sigma_\varepsilon^2}}$.

There are $K + G$ regressors of which $K_2 + G_2$ are correlated with the error term. So in order to achieve identification we need at least the same number of instruments. HT propose the following list of instruments:

- $(X_{1it} - \bar{X}_{1i.})$
  K_1 instruments given by the fixed effect transformation of the exogenous time varying regressors;
- $(X_{2it} - \bar{X}_{2i.})$
  K_2 instruments given by the fixed effect transformation of the correlated time varying regressors; note that by construction these instruments are orthogonal to the α_i because they are the estimated residuals of an OLS regressions of X_{2it} on the α_i (see section 4.4.3).
- $\bar{X}_{1i.}$
  K_1 instruments given by the averages for each individual across time of the exogenous time varying regressors.
- Z_1
  G_1 instruments corresponding to the exogenous time invariant regressors.

The order condition for identification is:

$$\begin{aligned} K_1 + K_2 + K_1 + G_1 &\geq K_1 + K_2 + G_1 + G_2 \\ K_1 &\geq G_2 \end{aligned} \tag{260}$$

If this condition is satisfied, IV estimation or 259 using the instruments described above, provides consistent estimates of all the parameters β and γ.

AM propose instead the following list of instruments for the same transformed model 259:

- K_1 instruments obtained from $(X_{1it} - \bar{X}_{1i.})$;
- K_2 instruments obtained from $(X_{2it} - \bar{X}_{2i.})$;
- TK_1 instruments obtained from $\tilde{X}_{1i}$ where $\tilde{X}_{1i}$ denotes the $NT \times KT$ matrix where each column contains values of X_{it} for a single time period; For example the column t of $\tilde{X}_{1i}$ is $[X_{11t} \ldots X_{11t}, \ldots, X_{1Nt} \ldots X_{1Nt}]$;
- G_1 instruments obtained from Z_1.

The order condition for identification is in this case:

$$\begin{aligned} K_1 + K_2 + TK_1 + G_1 &\geq K_1 + K_2 + G_1 + G_2 \\ TK_1 &\geq G_2 \end{aligned} \tag{261}$$

which is less restrictive than the HT condition.

Note that for these instruments to be valid the exogenous factors X_1 have to be uncorrelated *at each point in time* with the indiviudal effects. However, it is hard to imagine situations in which the HT instruments are valid and these are not.

BMS propose a third possible list of instruments for the same transformed model 259:

- K_1 instruments obtained from $(X_{1it} - \bar{X}_{1i.})$;
- K_1 instruments obtained from $\bar{X}_{1i.}$;
- TK_1 instruments obtained from $(\tilde{X}_{1it} - \bar{X}_{1i.})$;
- TK_2 instruments obtained from $(\tilde{X}_{2it} - \bar{X}_{2i.})$;
- G_1 instruments obtained from Z_1

where note that $(\tilde{X}_{1it} - \bar{X}_{1i.})$ provide only $T-1$ linearly indepedend instruments, and similarly $(\tilde{X}_{2it} - \bar{X}_{2i.})$.

The order condition for identification is in this case:

$$\begin{aligned} K_1 + K_2 + K_1 + (T-1)K_1 + (T-2)K_2 + G_1 &\geq K_1 + K_2 + G_1 + G_2 \\ TK_1 + (T-1)K_2 &\geq G_2 \end{aligned} \tag{262}$$

which is the least restrictive condition, but requires stronger exogeneity assumptions (see BMS).

Cornwell and Rupert (1988) compare the gain in efficiency delivered by the AM and BMS procedures with respect to the HT procedure and conclude, on the basis of a "returns to schooling" example, that efficiency gains are limited to the coefficients of time-invariant endogenous variables.

4.8 Extensions

The analysis described so far can be extended to deal with:

- fixed time effect;
- unbalanced panel data;

These extensions can be found in Greene (1987), Hsiao (1989) and in several of the articles quoted in the reading list.

4.9 Panel data analysis in STATA

See copies from STATA manuals

5 Panel data with discrete dependent variables

Consider the following problem (Card and Sullivan, 1988):

- $i = \{1 \dots N\}$ denote a sample of individuals;
- $t = \{1 \dots T\}$ time periods in which each individual is observed;
- for each individual in each period we observe the employment status

$$Y = \begin{cases} 1 & \text{if } i \text{ is employed in period } t \\ 0 & \text{if } i \text{ is unemployed in period } t \end{cases} \tag{263}$$

- we also observe a (row) vector of K explanatory factors X_{it};
- we assume that the observed binary outcome Y_{it} are independent conditional on X_{it} and on an unobservable individual time invariant effect α_i;
- the probability that individual i is employed in period t is assumed to be logistic:

$$Pr(Y_{it} = 1 | X_{it}, \alpha_i) = \frac{e^{\alpha_i + X_{it}\beta}}{1 + e^{\alpha_i + X_{it}\beta}} \tag{264}$$

The problem is to estimate how the explanatory factors affect the probability of employment controlling for the unobservable heterogeneity.

Card and Sullivan are in particular interested in evaluating how participation into a training program affects the probability of employment after the program.

If the outcome where not discrete we would use one of the techniques discussed above, but we cannot do so.

5.1 The conditional maximum likelihood approach

The solution proposed by Chamberlain (1980) consists in maximizing a conditional version of the likelihood function.

The intuition is that the α_i disappear from the likelihood if the likelihood of a given employment sequence (i.e. of a given individual) is calculated conditioning on the total number of periods of employment for that individual.

To understand this approach consider the simplest case of $T = 2$ and, to simplify the notation, let's indicate with $Pr(\{0, 1\})$ the probability of the sequence $\{0, 1\}$ conditional on α_i and X_i.

Consider first the case of a number of unemployment periods equal to 1. Using this notation we can write

$$Pr(Y_{i2} = 1|X_i, \alpha_i, \sum_{t=1}^{2} Y_{it} = 1) = \tag{265}$$

$$Pr(\{0,1\})|\{0,1\} \text{ or } \{1,0\}) =$$

$$\frac{Pr(\{0,1\})}{Pr(\{0,1\}) + Pr(\{1,0\})} =$$

$$\frac{Pr(Y_{i1} = 0|X_i, \alpha_i)Pr(Y_{i2} = 1|X_i, \alpha_i)}{Pr(Y_{i1} = 0|X_i, \alpha_i)Pr(Y_{i2} = 1|X_i, \alpha_i) + Pr(Y_{i1} = 1|X_i, \alpha_i)Pr(Y_{i2} = 0|X_i, \alpha_i)} =$$

Substituting the corresponding Logit probabilities from equation 264

$$\frac{\frac{1}{1+e^{(\alpha_i+x_{i1}\beta)}}\frac{e^{(\alpha_i+x_{i2}\beta)}}{1+e^{(\alpha_i+x_{i2}\beta)}}}{\frac{1}{1+e^{(\alpha_i+x_{i1}\beta)}}\frac{e^{(\alpha_i+x_{i2}\beta)}}{1+e^{(\alpha_i+x_{i2}\beta)}} + \frac{e^{(\alpha_i+x_{i1}\beta)}}{1+e^{(\alpha_i+x_{i1}\beta)}}\frac{1}{1+e^{(\alpha_i+x_{i2}\beta)}}} =$$

$$\frac{e^{(\alpha_i+x_{i2}\beta)}}{e^{(\alpha_i+x_{i1}\beta)} + e^{(\alpha_i+x_{i2}\beta)}}$$

where the individual effects cancel out and we obtain

$$\frac{e^{(x_{i2}\beta)}}{e^{(x_{i1}\beta)} + e^{(x_{i2}\beta)}}$$

$$\frac{e^{(x_{i2}-x_{i1})\beta}}{1 + e^{(x_{i2}-x_{i1})\beta}}$$

This finding suggests that we can write the likelihood in terms of the probabilities of the possibile sequences of outcomes for each individual.

For $T = 2$ there are four possible sequences. The probability of the first one is the probability we have just calculated:

$$Pr(\{0,1\}|X_i, \alpha_i, \sum_{t=1}^{2} Y_{it} = 1) = \frac{e^{(x_{i2}-x_{i1})\beta}}{1+e^{(x_{i2}-x_{i1})\beta}} = Pr(\{0,1\}|.) \tag{266}$$

Similarly for the second sequence:

$$Pr(\{1,0\}|X_i, \alpha_i, \sum_{t=1}^{2} Y_{it} = 1) = \frac{e^{(x_{i1}-x_{i2})\beta}}{1+e^{(x_{i1}-x_{i2})\beta}} = Pr(\{1,0\}|.) \tag{267}$$

There are two other possible sequences: $\{0,0\}$ and $\{1,1\}$. But note that:

$$Pr(\{1,1\}|X_i, \alpha_i, \sum_{t=1}^{2} Y_{it} = 2) = 1 = Pr(\{1,1\}|.) \tag{268}$$

$$Pr(\{0,0\}|X_i, \alpha_i, \sum_{t=1}^{2} Y_{it} = 0) = 1 = Pr(\{0,0\}|.) \tag{269}$$

These two sequences do not contribute to the likelihood because they are independent of the parameters.

So the likelihood can be written as

$$\begin{aligned} L &= \Pi_{i=1}^{N} Pr(\{0,1\}|.)^{W_{01}} Pr(\{1,0\}|.)^{W_{10}} Pr(\{0,0\}|.)^{W_{00}} Pr(\{1,1\}|.)^{W_{11}} \\ &= \Pi_{i=1}^{N} \left(\frac{e^{(x_{i1}-x_{i2})\beta}}{1+e^{(x_{i1}-x_{i2})\beta}}\right)^{W_{10}} \left(\frac{e^{(x_{i2}-x_{i1})\beta}}{1+e^{(x_{i2}-x_{i1})\beta}}\right)^{W_{01}} 1^{W_{00}} 1^{W_{11}} \end{aligned} \tag{270}$$

where:

- $W_{01} = 1$ for individuals whose sequence is $\{0,1\}$;
- $W_{10} = 1$ for individuals whose sequence is $\{1,0\}$;
- $W_{00} = 1$ for individuals whose sequence is $\{0,0\}$;
- $W_{11} = 1$ for individuals whose sequence is $\{0,0\}$.

Note that:

- the individual fixed effects are eliminated from this likelihood thanks to a transformation that is analogous to first differencing in linear panel models.

- The individuals for which the outcome is always 0 or always 1 do not contribute to the likelihood. In other words the information that they provide is not used to estimate β, which in some occasions may be unsatisfactory.
  - these individuals are unaffected by the explanatory factors;
  - if 99% of the sample is in this situation, we may still estimate a significant β out of the 1% of the sample which changed outcome during the observation period;
  - no weight would be given to the fact that for the vast majority of the sample the explanatory factors do not affect the outcome.

- This conditional likelihood approach cannot be adopted in the presence of lagged dependent variables, which is the problem addressed specifically by Card and Sullivan (1988)

5.2 Fixed effects conditional logit estimation in STATA

See copies from STATA manuals

5.3 Applications

6 References

i. **Discrete dependent variables.**

Bertrand Marianne, Erzo F.P. Luttmer and Sendhil Mullainathan, "Network Effects and Welfare Cultures", NBER Working Paper No. 6832, 1998.

Brooks R. D., T. Fry and M. N. Harris, " The size and power properties of combining choice set partition tests for the IIA property in the Logit model" http://www.monash.edu.au/

Case Anne and Lawrence Katz, The Company You Keep: the Effect of Family and Neighborhood on Disadvantaged Youth, NBER Working Paper No. 3705, 1991.

Foley M., "Labor Market Dynamics in Russia", Yale University, Center Discussion Paper N. 780.

Greene W.H., Econometric Analysis, 3d. edition. Prentice Hall 1987, Chapter 19.

Ichino A., and Maggi G., "Group Interactions and Individual Background", EUI, Mimeo, 1998.

Hausman D. and D. McFadden "Specification Tests for the Multinomial Logit Model", Econometrica, 1984.

Manski Charles F., Identification of Endogenous Social Effects: The Reflection Problem, *Review of Economic Studies*, 60, 1993, pp. 531-542.

McFadden D., K. Train and W. Tye, "An Application of Diagnostic Tests for the Indepedence of Irrelevant Alternatives Property of the Multinomial LOgit Model" Transportation Research Record, 637, 39-46

Merz M. and Schimmelpfenning A, "Career Choices of German High School Graduates: Evidence from the German Socio-Economic Panel", mimeo, 1998

Medoff J. and Abraham K. " Experience, Performance and Earnings", Quarterly Journal of Economics, 1980.

P. Schmidt and R. Strauss "The Prediction of Occupation Using Multiple Logit Models", International Economic Review, 1975a.

P. Schmidt and R. Strauss "Estimation of Models with Jointly Dependent Qualitative Variables: A Simultaneous Logit Approach", Econometrica, 1975b.

ii. Pindyck R. S. and Rubinfeld D. L., "Econometric Models and Economic Forecast", 4th. edition, McGraw-Hill, 1998, Chapter 11.

iii. K. Train "Qualitative Choice Analysis", MIT Press, 1986.

iv. **Panel data**

Angrist J. and Krueger A. " Empirical Strategies in Labor Economics ", Forthcoming in *Handbook of Labor Economics*, North Holland, 1998.

Amemiya, T. and MacCurdy T " Instrumental-Variables Estimation of an Error-Components Model " Econometrica, 54, 869-880, 1986.

Arellano M. and Bond S. " Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment equations ", Review of Economic Studies, 58, 277-297, (1991).

Arellano M. and Bover O. " Another Look at the Instrumental-Variable Estimation of Error-Components Models" Discussion Paper 7, LSE, Center for economic Performance.

O. Ashenfelter and A. Krueger "Estimates of the Economic Return to Schooling from a New Sample of Twins", American Economic Review , 1994.

Balestra, P. and Nerlove, M. " Pooling Cross Section and Time Series Data in the Estimation of a Dynamic Model : The Demand for Natural Gas." Econometrica, 34, 585-612, 1966.

Bound J. and G. Solon., "Double Trouble: on the Values of Twins-Based Estimation of the Return to Schooling" NBER WP.6721, 1998.

Breush, T., Mizon G. and Schmidt, P. " Efficient Estimation Using Panel Data " Econometrica, 51, 1635-1659, 1989.

Chamberlain G. "Panel Data", Ch. 22 in Handbook of Econometrics. North Holland, 1990.

Chamberlain G., Analysis of Covariance with Qualitative Data, *Review of Economic Studies*, 1980.

Card D. " The Impact of the Mariel Boatlift on the Miami Labor Market ", Industrial and Labor Relations Review, 43, 1990, 245-57.

D. Card and D. Sullivan "Measuring the Effect of Subsidized Training Programs on Movements In and Out of Employment", Econometrica, 1988.

Cecchetti S. " The frequency of Price Adjustments. A Study of the Newsstand Prices of Magazines", Journal of Econometrics, 1986.

Cornwell C. and Rupert P., "Efficient Estimation with Panel Data: an Empirical Comparison of Instrumental Variables Estimators" Journal of Applied Econometrics, 1988, 3, 149-155.

Greene W.H., Econometric Analysis, 3d. edition. Prentice Hall 1987, Chapter 19.

Griliches Z., "Sibling Models and data in Economics: Beginnings of a Survey" Journal of Econometrics, 1986, 31, 1, 93-118.

Griliches Z. and Hausman J., "Errors in Variables in Panel Data" Journal of Political Economy, 1979, 87, 5, s37-63.

Hausman J., "Specification Test in Econometrics ", Econometrica, 1978.

Hausman J., Bronwyn H. Hall and Zvi Griliches, "Econometric Models for Count Data With and Application to the Patents-R&D Relationship", Econometrica, 1984.

Hausman, J. and Taylor, W. " Panel Data and Unobservable Individual effects" Econometrica, 49, 1377-1398, 1981.

Holz-Eakin, D., Newey W. and Rosen, H. " Estimating Vector Autoregression with Panel Data " Econometrica, 56, 1371-1395, 1988.

Hsiao C. "Analysis of Panel data" Cambridge University Press, 1989.

Keane, P. and Runkle, D. " On the estimation of Panel-Data Models with Serial Correlation When Instruments Are Not Strictly Exogenous " Journal of Business and Economic Statistics, 10, 1-26, 1992

Maddala, G. S. "The Use of Variance Components Models in Pooling Cross Section and Time Series data ", Econometrica, 39, 341-358, 1971.

Y. Mundlak "Empirical Production Function Free of Management Bias ", Journal of Farm Economics , 1961.

Y. Mundlak "On the Pooling of Time Series and Cross Section Data", Econometrica, 1978.

Nickell S., Biases in Dynamic Models , *Econometrica*, 49, 6, 1981, pp. 1417-1426.