Panel Data Course

Spring 2004

REGRESSION (Revision)

Cross section: $i = 1,\ldots,N$ **(NON-ORDERED)**
Individuals, firms, countries etc.

Time Series: $t = 1,\ldots,T$ **(ORDERED)**

Variables:

y_i (y_t): **DEPENDENT (Endogenous)**

x_{ki} (x_{kt}): **INDEPENDENT (Exogenous),** $k = 1,\ldots,K$

MODEL

$$y_i = \alpha + \sum_{j=1}^{K} \beta_j x_{ji} + \varepsilon_i$$

α **and** β **are parameters.** ε **is a stochastic error**

NOTE: The model can be written in matrix form

$$y = X\beta + \varepsilon, \qquad \text{where } y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}, \beta = \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}, \text{ etc}$$

The simple regression model

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

is used as an example in this course. We always write "*K*" as the number of exogenous variables, however

ASSUMPTIONS

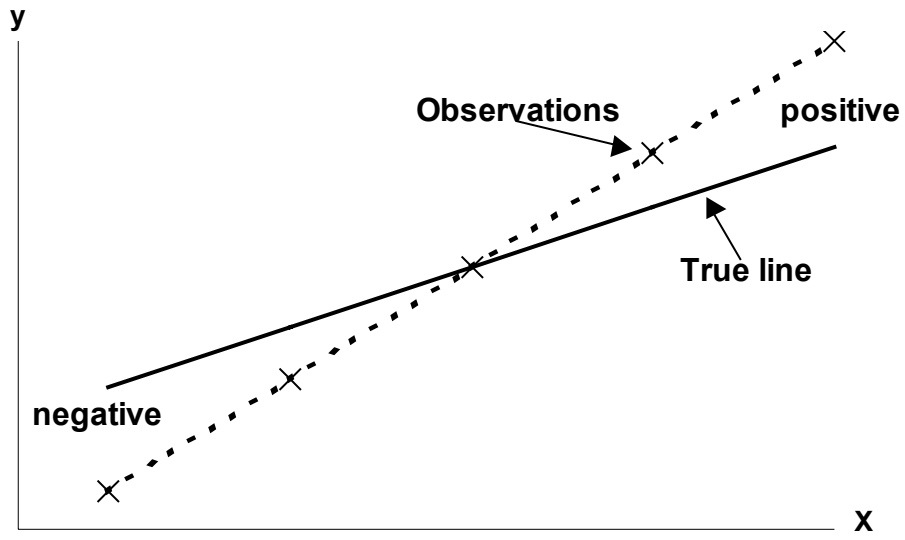
1) Correct Model	$\mathrm{E}(\varepsilon_i) = 0$
2) Exogeneity	$\mathrm{Cor}(\varepsilon_i, x_i) = 0$
3) Homoscedasticity	$\mathrm{Var}(\varepsilon_i) = \sigma^2$, constant
4) Serial independence	$\mathrm{Cor}(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j$
5) Normality	$\varepsilon_i \sim$ Normal
6) No incidental parameters (*K* does not grow with *N*)	

(1), (2) and (6) are needed for CONSISTENCY (*i.e.*, in large samples OLS parameter estimates will be correct "on the average")

(3) and (4) are needed for EFFICIENCY (*i.e.*, in large samples OLS yields "best" estimates, significance tests are correct, *etc.*)

(5) is needed for small sample properties

Problem when $\mathrm{Cor}(\varepsilon, x) \neq 0$



FORMULAE for OLS (Ordinary Least Squares)

$$\hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Residual: $\hat{\varepsilon} = y_i - \hat{\beta} x_i$

Error variance: $\hat{\sigma}^2 = \frac{1}{\nu} \sum \hat{\varepsilon}_i^2$,
where $\nu = N - K - 1$, **the degrees of freedom**

Variance; $\mathrm{Var}(\hat{\beta}) = \dfrac{\sigma^2}{\sum (x_i - \bar{x})^2}$

Standard error; $\mathrm{se}(\hat{\beta}) = \sqrt{\dfrac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}}$

t-value; $t(\hat{\beta}) = \dfrac{\hat{\beta}}{se(\hat{\beta})}$

MATRIX FORMULAE

$$\hat{\beta} = (X'X)^{-1} X'y$$

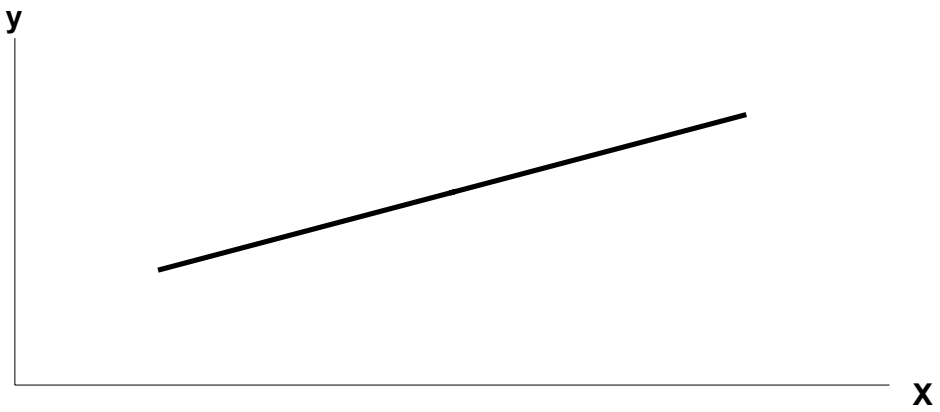
$$\mathrm{Var}(\hat{\beta}) = \sigma^2 (X'X)^{-1}, \textit{ etc}$$

PANEL DATA MODELS

y_{it}, x_{it} $\quad i = 1,\ldots,N, t = 1,\ldots T$

THE POOLED MODEL

$$y_{it} = \alpha + \beta x_{it} + \varepsilon_{it}$$

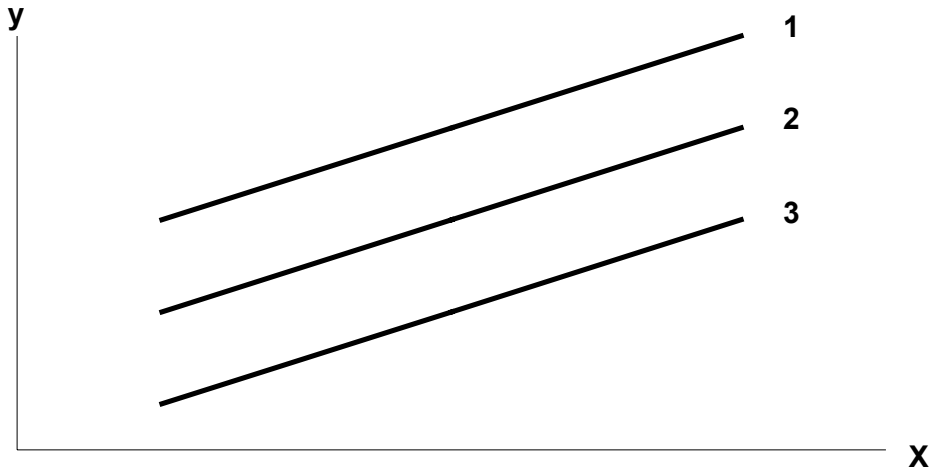


Here we are NOT using any Panel information.

The data are treated as if there was only a single index.

TRADITIONAL PANEL MODEL

$$y_{it} = \alpha_i + \beta x_{it} + \varepsilon_{it}$$



The constant terms, α_i, vary from individual to individual.

This is called INDIVIDUAL (UNOBSERVED) HETEROGENEITY

The slopes are, however, the same for all individuals.

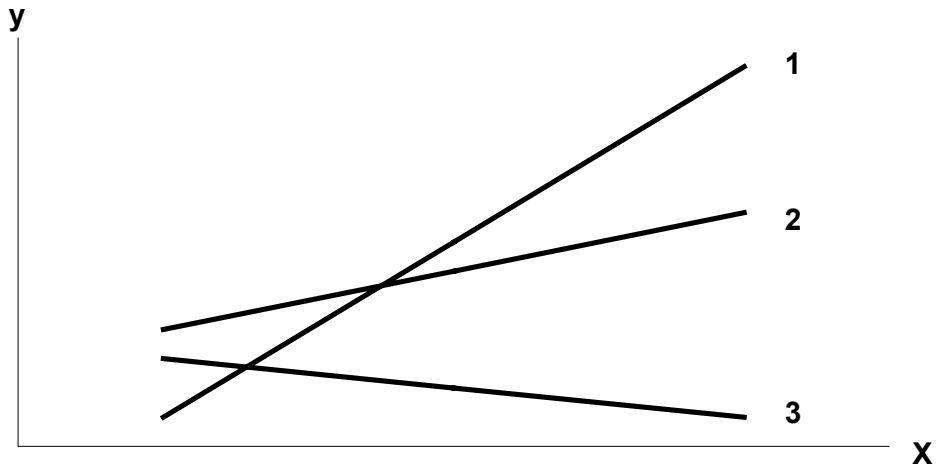
In both the Pooled and Panel Models we assume that the errors are homoscedastic and serially independent both *within* and *between* individuals

$$\text{Var}(\varepsilon_{it}) = \sigma^2$$

$$\text{Cor}(\varepsilon_{it}, \varepsilon_{js}) = 0 \quad \text{when } i \neq j \text{ and/or } t \neq s$$

SUR MODEL
SEEMINGLY UNRELATED REGRESSIONS

$$y_{it} = \alpha_i + \beta_i x_{it} + \varepsilon_{it}$$



The constant terms, α_i, and slopes, β_i, vary from individual to individual.

In SUR models the errors are allowed to be contemporally correlated and heteroscedastic *between* individuals. We still assume serial independence as well as homoscedasticity *within* individuals

$$\text{Var}(\varepsilon_{it}) = \sigma_i^2$$

$$\text{Cor}(\varepsilon_{it}, \varepsilon_{jt}) = \sigma_{ij}$$

$$\text{Cor}(\varepsilon_{it}, \varepsilon_{js}) = 0 \text{ when } t \neq s$$

TWO COMMON SITUATIONS

1) **There are a LARGE number of independent individuals observed for a FEW time periods.**

   $N >> T$

   ***N* is often in the range 500 - 20,000, while *T* lies between 2 and 10. In this case it is not possible to estimate different individual slopes for all the exogenous variables.**

   The PANEL DATA MODEL is most appropriate.

2) **There are some MEDIUM length time series for RELATIVELY FEW, possibly dependent, equations (countries, firms, sectors *etc*)**

   $T > N$

   ***T* is usually in the range 30 - 150, while *N* often lies between 2 and 15.**

   In this case the SUR MODEL is appropriate.

   Efficient (SUR) estimation is used when $T \geq N$

   Equation-by-equation OLS is used if $K \leq T < N$**.**

Panel models are MORE general than Pooled models but LESS general than SUR models.

FIXED EFFECTS MODELS

Here we treat the individual heterogeneity as N parameters that are to be estimated

$$y_{it} = \alpha_i + \beta x_{it} + \varepsilon_{it}$$

$$i = 1,\ldots,N, \quad t = 1,\ldots T$$

N is large (and can often be increased). T is small and fixed.

WHY CAN'T WE USE OLS?

The individual heterogeneity can be considered as N dummy variables. A regression with $N + K$ variables (so called Least Squares Dummy Variables (LSDV) regression) must therefore be estimated.

There are two problems with LSDV regression.

1) **There are INCIDENTAL PARAMETERS**

   The number of α_i grows as N increases. The usual proof of consistency therefore does not hold for LSDV

2) **Inverting a $N + K$ matrix can be impossible if N is very large. Even when possible it can be impracticable and/or inaccurate.**

WE NEED A "TRICK" TO REMOVE THE INCIDENTAL PARAMETERS!

The original model is

$$y_{it} = \alpha_i + \beta x_{it} + \varepsilon_{it} \qquad \textbf{(1)}$$

Averaging over the *T* observations for each individual yields

$$\bar{y}_{i.} = \alpha_i + \beta \bar{x}_{i.} + \bar{\varepsilon}_{i.} \qquad \textbf{(2)}$$

where the "dot" notation is simply $\bar{y}_{i.} = \frac{1}{T}\sum_t y_{it}$ **, *etc.***

Subtracting (2) from (1) gives

$$(y_{it} - \bar{y}_{i.}) = \beta(x_{it} - \bar{x}_{i.}) + (\varepsilon_{it} - \bar{\varepsilon}_{i.}) \qquad \textbf{(3)}$$

This is called the WITHIN REGRESSION. There are no incidental parameters and the errors still satisfy the usual assumptions. We can therefore use LS on (3) to obtain consistent estimates.

The Within Regression Estimates

To simplify the notation we define

$$\tilde{y}_{it} = y_{it} - \bar{y}_{i.}, \quad \tilde{x}_{it} = x_{it} - \bar{x}_{i.}, \textit{ etc}$$

The within regression can thus be written

$$\tilde{y}_{it} = \beta \tilde{x}_{it} + \tilde{\varepsilon}_{it}$$

The estimates can thus be written

$$\hat{\beta}_w = \frac{\sum\sum \tilde{x}_{it}\tilde{y}_{it}}{\sum\sum \tilde{x}_{it}^2} = \frac{\sum\sum (x_{it} - \bar{x}_{i.})(y_{it} - \bar{y}_{i.})}{\sum\sum (x_{it} - \bar{x}_{i.})^2}$$

and the individual effects can be estimated as

$$\hat{\alpha}_{w,i} = \bar{y}_{i.} - \hat{\beta}_w \bar{x}_{i.}$$

PROPERTIES OF THE WITHIN (FE) ESTIMATES

- $\hat{\beta}_w$ is consistent if either N or T become large.
- $\hat{\alpha}_{w,i}$ is only consistent when T become large.
- The number of degrees of freedom must be adjusted.

  Degrees of freedom $=$ #obs $-$ #pars, *i.e.*

  $$\nu = NT - N - K$$
  $$= N(T-1) - K$$

  Usual OLS programs, that are not explicitly designed for panel data, assume that the degrees of freedom are $NT - K$. Their standard errors, test statistics and P-values must therefore be corrected.
- The parameter estimates from LSDV are the same as from the within regression!

  This is NOT a general result (incidental parameters *do* cause inconsistencies in many models)

THE FIXED EFFECTS MODEL: A SUMMARY

1) **Calculate the within averages:** $\bar{y}_{i.}$ **and** $\bar{x}_{i.}$

2) **Calculate the differences from within averages:**

$$\tilde{y}_{it} = y_{it} - \bar{y}_{i.} \text{ and } \tilde{x}_{it} = x_{it} - \bar{x}_{i.}$$

3) **Regress** $\tilde{y}_{it}$ **on** $\tilde{x}_{it}$ **(without a constant term). This gives** $\hat{\beta}_w$ **and** $se(\hat{\beta}_w)$

4) **Estimate the individual effects (if required):**

$$\hat{\alpha}_{w,i} = \bar{y}_{i.} - \hat{\beta}_w \bar{x}_{i.}$$

5) **If the regression has been performed with an ordinary least squares program, then the degrees of freedom *etc.* must be adjusted.**

$$\nu_a = \nu_u - N$$

$$se_a(\hat{\beta}_w) = \sqrt{\frac{\nu_u}{\nu_a}} \cdot se_u(\hat{\beta}_w)$$

$$t_a(\hat{\beta}_w) = \sqrt{\frac{\nu_a}{\nu_u}} \cdot t_u(\hat{\beta}_w)$$

which is distributed t_{ν_a} **under** H_0

"*a*" denotes ADJUSTED and "*u*" UNADJUSTED

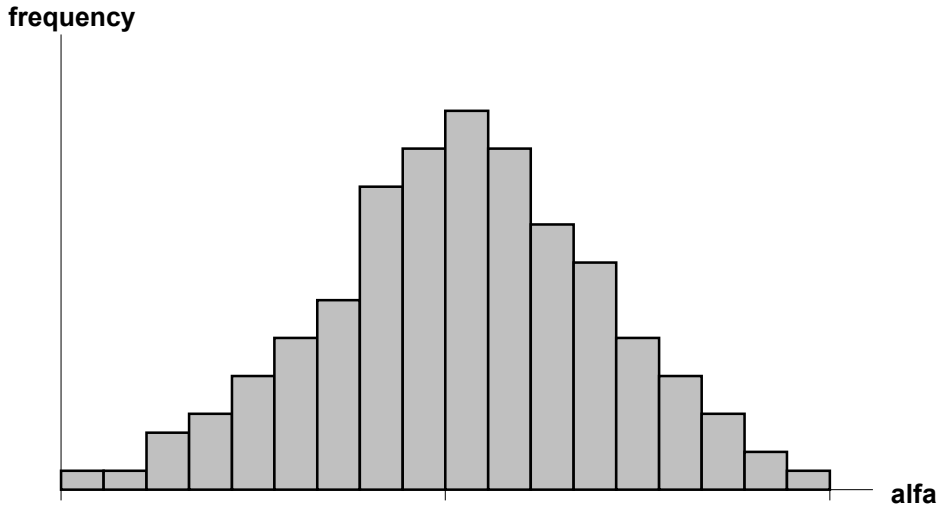
RANDOM EFFECTS MODELS

In Fixed Effects models:

- **We aren't interested in the individual effects**
- **We can't estimate them consistently**

WHY BOTHER WITH THEM?

The individual effects have an "empirical" distribution



which has certain characteristics, *e.g.*

$\mu =$ **average** $\alpha = \frac{1}{N}\sum \alpha_i$

$\sigma_\alpha^2 =$ **variance of** α

We can use these definitions to rewrite the panel data model

$$y_{it} = \mu + \beta x_{it} + (\alpha_i - \mu) + \varepsilon_{it}$$

Defining the new error:

$$u_{it} = (\alpha_i - \mu) + \varepsilon_{it}$$

we can write

$$y_{it} = \mu + \beta x_{it} + u_{it}$$

This is the RANDOM EFFECTS MODEL.

This looks almost the same as the POOLED model, but note two differences

- **The constant term can be interpreted as the average individual effect**
- **The error term now has a special form**

We can obviously estimate the RE model using OLS to obtain estimates of μ and β

When is this consistent?

If consistent, is it efficient?

WHEN IS THE RANDOM EFFECTS MODEL CONSISTENT?

Two conditions must be fulfilled

- $\mathrm{E}(u_{it}) = \mathrm{E}(\alpha_i - \mu) + \mathrm{E}(\varepsilon_{it}) = 0$
- $\mathrm{Cov}(u_{it}, x_{it}) = \mathrm{Cov}(\alpha_i, x_{it}) + \mathrm{Cov}(\varepsilon_{it}, x_{it}) = 0$

The first condition is OK as long as the original errors are unbiased.

The second condition needs x_{it} to be independent of ε_{it} (which has already been assumed) *and* of α_i.

IS IT REASONABLE TO ASSUME THAT THE INDIVIDUAL EFFECTS ARE INDEPENDENT OF THE EXOGENOUS VARIABLES?

EXAMPLE:

y_{it} = # days unemployed year *t*

x_{it} = income

α_i = unmeasured individual propensity to be unemployed (depends on such factors as Education, Health Status *etc.*)

THIS ASSUMPTION MUST BE TESTED!

IS *OLS* EFFICIENT IN THE RANDOM EFFECTS MODEL?

Efficient OLS needs homoscedasticity and serial independence in the errors, u_{it}.

Remember that $u_{it} = (\alpha_i - \mu) + \varepsilon_{it}$ we obtain

$\text{Var}(u_{it}) = \sigma_\alpha^2 + \sigma_\varepsilon^2$	**Assuming that α_i and ε_{it} are independent**
$\text{Cov}(u_{it}, u_{js}) = 0, \quad j \neq i$	**Can be assumed if all individuals are independent**
$\text{Cov}(u_{it}, u_{is}) = \sigma_\alpha^2 \neq 0$	**Since α_i is the same for all t within the same individual**

The last condition violates the "serial independence" assumption.

OLS is thus INEFFICIENT in the random effects model, and yields INCORRECT standard errors and tests.

EFFICIENT ESTIMATION IN THE RANDOM EFFECTS MODEL

The Random Effects Model can be efficiently estimated using GLS (Generalised Least Squares)

1) Define $\theta = 1 - \frac{\sigma_\varepsilon}{\sigma_1}$, **where**

$$\sigma_1^2 = T\sigma_\alpha^2 + \sigma_\varepsilon^2$$

2) Calculate the "pseudo within differences"

$$y_{it}^* = y_{it} - \theta\bar{y}_{i.}, \quad x_{it}^* = x_{it} - \theta\bar{x}_{i.}$$

3) Perform an OLS regression on

$$y_{it}^* = \mu^* + \beta x_{it}^* + u_{it}^*$$

where $\mu^* = (1-\theta)\mu$

and u_{it}^* **satisfies the LS assumptions**

4) The Random Effects estimate of β **is given by**

$$\hat{\beta}_{re} = \frac{\sum\sum(x_{it}^* - \bar{x}_{i.}^*)(y_{it}^* - \bar{y}_{i.}^*)}{\sum\sum(x_{it}^* - \bar{x}_{i.}^*)^2}$$

PROBLEM: θ is not known

Unfortunately σ_ε^2 and σ_α^2 are unknown.

If the errors *u* and ε (or α) where known we could estimate the variances using

$$\hat{\sigma}_1^2 = \frac{T}{N}\sum \bar{u}_{i.}^2 \qquad (4)$$

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N(T-1)}\sum\sum (u_{it} - \bar{u}_{i.})^2 \qquad (5)$$

$$= \frac{1}{N(T-1)}\sum\sum (\varepsilon_{it} - \bar{\varepsilon}_{i.})^2$$

$$\hat{\sigma}_\alpha^2 = \frac{1}{N-1}\sum (\alpha_i - \bar{\alpha})^2 \qquad (6)$$

Since *u*, ε and α are unknown there are a number of suggestions for how they can be estimated.

These methods use various residuals instead of the unknown errors:

$\hat{u}_{ols}$ **= RE residuals from the POOLED regression**

$$y_{it} = \mu + \beta x_{it} + u_{it} \qquad \text{\#obs = NT}$$

$\hat{u}_{b}$ **= RE residuals from the BETWEEN regression**

$$\bar{y}_{i.} = \mu + \beta \bar{x}_{i.} + \bar{u}_{i.} \qquad \text{\#obs = N}$$

$\hat{\varepsilon}_{w}$ **= FE residuals from the WITHIN regression**

$$\tilde{y}_{it} = \beta \tilde{x}_{it} + \tilde{\varepsilon}_{it} \qquad \text{\#obs = NT}$$

$\hat{u}_{w}$ **= RE residuals from the WITHIN regression**

$$= \hat{\varepsilon}_{w} + (\hat{\alpha}_{w} - \bar{\hat{\alpha}}_{w})$$

$\hat{u}_{re}$ **= residuals from the RE regression**

$$y^{*}_{it} = \mu^{*} + \beta x^{*}_{it} + u^{*}_{it} \qquad \text{\#obs = NT}$$

SOME DIFFERENT METHODS OF ESTIMATING θ

I WALLACE and HUSSAIN.

Use $\hat{u}_{ols}$ instead of u in (4) and (5)

II AMEMIYA

Use $\hat{u}_w$ in (4) and $\hat{\varepsilon}_w$ in (5)

III SWAMY and ARORA

Use $\hat{u}_b$ in (4) and $\hat{\varepsilon}_w$ in (5)

IV NERLOVE

Use $\hat{\alpha}_w$ in (6) and $\hat{\varepsilon}_w$ in (5)

V MAXIMUM LIKELIHOOD

Start with one of the previous methods, estimate the RE parameters and then use $\hat{u}_{re}$ to calculate a new θ. Iterate.

Different authors suggest different degrees-of-freedom corrections in the variance formulae. For example (5) is often calculated as

$$\hat{\hat{\sigma}}_{\varepsilon}^{2} = \frac{1}{N(T-1)-K}\sum\sum \hat{\varepsilon}_{w,it}^{2}$$

where we have also used the fact that $\bar{\hat{\varepsilon}}_{w} = 0$

PROPERTIES

Research has established the following:

- **There is not much difference between I - V when the Random Effects model is correct.**
- **Only NERLOVE guarantees that $\hat{\sigma}_{\alpha}^{2} > 0$. Many users of the other methods set $\theta = 1$ (Fixed Effects) if a negative value of $\hat{\sigma}_{\alpha}^{2}$ is obtained.**
- **It is difficult to give any general rules as to which method to use. SWAMY/ARORA is probably the most common.**
- **The Random Effects estimates are more efficient than the Fix Effects estimates when the RE model is correct. They are inconsistent, however, when the model is incorrect.**
- **It is important to test which model is correct.**

INDIVIDUAL SPECIFIC VARIABLES

In many cases we have some exogenous variables that vary between individuals, but which do not vary over time within a given individual (*e.g.*, gender, race, nationality).

Denote such an individual specific variable as q_i

In a FIX EFFECT Model we will thus write

$$y_{it} = \alpha_i + \gamma q_i + \beta x_{it} + \varepsilon_{it}$$

The term $(\alpha_i + \gamma q_i)$ does not vary over time, and will thus be removed by the within transformation, *i.e.*,

$$(y_{it} - \bar{y}_{i.}) = \beta(x_{it} - \bar{x}_{i.}) + (\varepsilon_{it} - \bar{\varepsilon}_{i.})$$

The parameters of the individual specific variables (γ) cannot be estimate in the Fix Effects model (that is, we cannot distinguish between observed and unobserved heterogeneity)

If q_i only varies slightly over time, and only for a few individuals, then γ will be estimated with poor precision (for example, Education, Marital Status)

In a RANDOM EFFECTS model we will write

$$y_{it} = \mu + \gamma q_i + \beta x_{it} + u_{it}$$

in which case γ can be estimated (although not when using the NERLOVE method).

For the Random Effects model to be appropriate, however, the observed heterogeneity (q) must be independent of the unobserved heterogeneity (α).

The Random Effects model therefore has the added advantage of allowing us to estimate parameters of which we are probably interested

TESTING

Hypothesis testing is central to statistical inference. In econometric modelling we often distinguish between three types of tests

1) **SPECIFICATION TESTS**

   Is the model correct? (*e.g.*, POOLED, RE, FE, SUR)

2) **MISSPECIFICATION TESTS**

   Are any of the statistical assumptions violated? (*e.g.*, Serial independence, Homoscedasticity)

3) **PARAMETER TESTS**

   Do the parameters have specified values? (*e.g.*, is a parameter "significant")

(1) and (2) are really two aspects of the same question - we are asking "Can the model be estimated efficiently using Least Squares?". They should be answered together

(3) can only be addressed after (1) and (2).

PARAMETER TESTS

There are no new problems with panel data models. The same principals that apply to ordinary regression can be applied here.

The usual way to test hypotheses concerning the parameters in regression models is to use *t*-tests (one parameter) and *F*-tests (several parameters).

These tests can be calculated in two ways, which give identical results in linear models. We use the method of calculation that is easiest.

Sum of Squares Tests

We have

- **an UNRESTRICTED model, where all the parameters are estimated, and**
- **a RESTRICTED model, where the parameters satisfy a number of restrictions.**

The null hypothesis (H_0) is that the RESTRICTED model is true, while the alternative hypothesis (H_1) is that the UNRESTRICTED model holds

The restrictions are usually of the form that certain parameters are zero - *i.e.*, some variables are not important.

If the hypothesis that a single parameter is zero is rejected, then we say that this parameter is significant (more correctly: "the parameter is significantly different from zero at the given significance level ")

The unrestricted model, containing p_u parameters, is estimated using LS. The residual sum of squares is calculated

$$URSS = \sum \hat{\varepsilon}_{u,i}^2$$

The restricted model, which has p_r ($<p_u$) parameters, is also estimated using LS. The residual sum of squares is also calculated here

$$RRSS = \sum \hat{\varepsilon}_{r,i}^2$$

Defining

$$\nu_1 = p_u - p_r = \#\text{ restrictions}$$

$$\nu_2 = \#\text{ observations} - p_u$$

we calculate the test statistic

$$F = \frac{(RRSS - URSS)/\nu_1}{URSS/\nu_2} \sim F_{\nu_1,\nu_2} \text{ under } H_0$$

- **The SS test is derived through a small sample adjustment of the Likelihood Ratio Test**

- **If $\nu_1 = 1$ then the square root of F is t-distributed under the null.**

Standard Error Tests

The *t*-test for a single parameter can also be written

$$t(\hat{\beta}) = \frac{\hat{\beta}}{se(\hat{\beta})} \sim t_{v_1} \quad \text{under } H_0 : \beta = 0$$

- **This is a small sample adjustment of the Wald Test.**

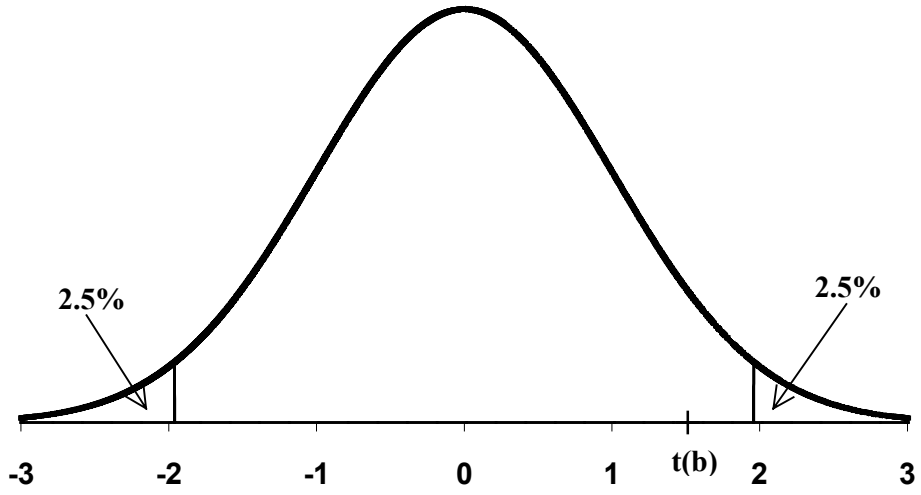
- **The F-test for multiple parameters can also be derived from the equivalent Wald test, but is now expressed in matrix terms.**

In general

- **tests for single parameters are easiest to use in standard error form, while**
- **tests for several parameters are calculated easiest in sum-of-squares form.**

CRITICAL VALUES AND P-VALUES

The traditional way to test a hypothesis is to see whether the t-statistic is greater than the critical value for a given significance level



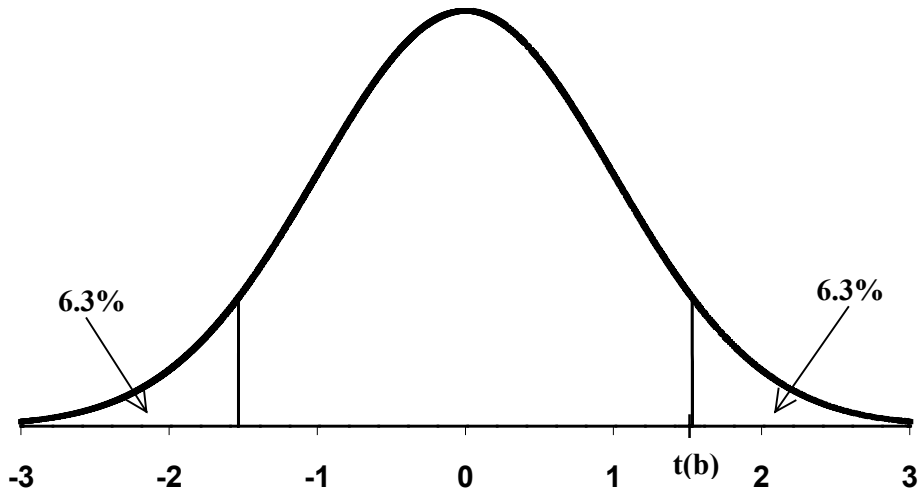
The null hypothesis is rejected if $\left|t(\hat{\beta})\right| > c_\alpha$.

In the above diagram α **= 5%,** c_α **= 1.96 and** t **= 1.51,** ***i.e.*** **the hypothesis is not rejected.**

This method of describing test results is not very informative, however.

The reader is dependent on what significance level the author considers to be interesting. At best the author will use the "star" convention (one, two or three stars to show significance at the 5%, 1% and 0.1% levels)

An alternative, which is gaining more and more popularity, is to quote P-VALUES



The P-Value is $P(t > |t(\hat{\beta})|) + P(t < -|t(\hat{\beta})|)$.

The null hypothesis is rejected if the P-value is less than the significance level.

In the above diagram the P-value is 12.6%, which is greater than 5%. The hypothesis would therefore not be rejected at the 5% level.

SPECIFICATION TESTS

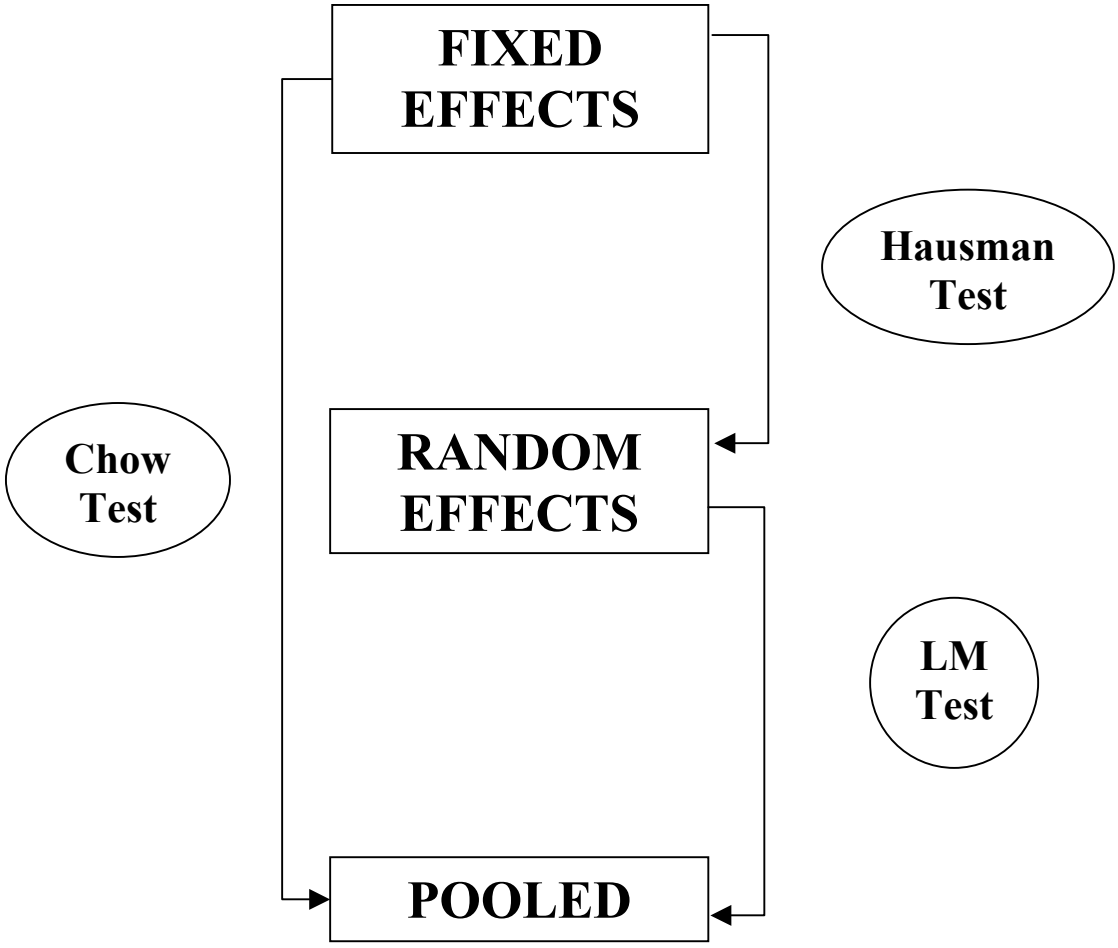
So far we have considered four models for panel data:

$y_{it} = \alpha_i + \beta_i x_{it} + \varepsilon_{it}$	**SUR**
$\tilde{y}_{it} = \beta \tilde{x}_{it} + \tilde{\varepsilon}_{it}$	**FIXED EFFECTS**
$y^*_{it} = \mu^* + \beta x^*_{it} + u^*_{it}$	**RANDOM EFFECTS**
$y_{it} = \alpha + \beta x_{it} + \varepsilon_{it}$	**POOLED**

The error terms in these models satisfy the OLS assumptions IF the respective model is correct (this is why we have expressed the FE and RE models in their transformed form).

We will not be considering the SUR model in detail, since this is not possible to estimate if *T* is small.

TESTS FOR VARIOUS MODELS



A TEST FOR UNOBSERVED HETEROGENEITY

The CHOW TEST of the POOLED MODEL against the FIX EFFECTS MODEL.

H_0: POOLED MODEL (Restricted)

H_1: FIX EFFECTS MODEL (Unrestricted)

The URSS is calculated using the residuals from the Within Regression ($\tilde{\varepsilon}_w$).The number of parameters is $p_u = N + K$.

The RRSS is calculated using the residuals from the Pooled Regression ($\hat{\varepsilon}_{ols}$).The number of parameters is $p_r = K + 1$.

The number of observations is *NT* in both cases.

The Sum-of-Squares test of H_0 is thus

$$CHOW = \frac{(RRSS - URSS)/(N-1)}{URSS/(NT-N-K)}$$

which is distributed $F_{N-1,NT-N-K}$ under H_0.

This test is called a CHOW test because of its similarity to the well known CHOW test for parameter stability.

INDIVIDUAL SPECIFIC VARIABLES

If there are p_q individual specific variables in the model, then these are INCLUDED in the POOLED model, but EXCLUDED from the FIXED EFFECTS model.

This is reasonable, since we want to test for unobserved heterogeneity, not observed heterogeneity!

In this case we must use $p_r = K + 1 + p_q$ and $\nu_1 = N - p_q - 1$ in the Chow test.

RANDOM or FIX EFFECTS? The HAUSMAN TEST

The Hausman test is a general test procedure which is used when we want to test the validity of an assumption that is necessary for efficient estimation.

For the test to work we need two estimation methods:

METHOD 1, called $\hat{\beta}_a$, is both consistent and efficient under H_0, but is inconsistent under H_1.

METHOD 2, called $\hat{\beta}_b$, is consistent under both H_0 and H_1, but is inefficient under H_0.

If there is only one parameter to be tested, then the test statistic is very simple

$$h = \frac{(\hat{\beta}_b - \hat{\beta}_a)^2}{s_b^2 - s_a^2} \sim \chi_1^2 \quad \text{under } H_0$$

where s_a and s_b are the standard errors of the parameter estimates.

- **Although $\sigma_b^2 > \sigma_a^2$ under the null, this relation need not hold in small samples for the standard errors. If $s_b^2 < s_a^2$ the test is not applicable.**
- **If there are $J > 1$ parameters to be compared, the Hausman test statistic must be expressed in matrix terms and is distributed χ_J^2.**
- **There often exists an "omitted variables" version of the Hausman test, which has the same asymptotic properties and which is never negative**

The Hausman test for RE vs. FE

In the case of testing for random effects we have the following situation

H_0: Random Effects model $[\text{Cor}(\alpha_i, x_{it}) = 0]$

H_1: Fix Effects model $[\text{Cor}(\alpha_i, x_{it}) \neq 0]$

Our estimates satisfy the Hausman conditions

$\hat{\beta}_{re}$ **is consistent and efficient under H_0, but inconsistent under H_1**

$\hat{\beta}_{w}$ **is consistent under H_0 and H_1, but inefficient under H_0**

The Hausman test can now be calculated in matrix terms or through an omitted variables procedure

OMITTED VARIABLES VERSION

Define as before

$\tilde{x}_{it} = x_{it} - \bar{x}_{i.}$ **from the FE model, and**

$y^*_{it} = y_{it} - \theta\bar{y}_{i.}$, $x^*_{it} = x_{it} - \theta\bar{x}_{i.}$ **from the RE model.**

We now estimate the RE regression with the within regressors as extra variables

$$y^*_{it} = \mu^* + \beta x^*_{it} + \gamma\tilde{x}_{it} + w_{it}$$

The alternative Hausman test is a simple *F*-test that γ is zero.

This is appropriate since

$$\mathbf{H_0}: \gamma = 0 \quad \Leftrightarrow \quad \mathbf{H_0}: \mathrm{Cor}(\alpha_i, x_{it}) = 0$$

If there are individual specific variables we simply test $\mathbf{H_0}: \gamma = 0$ in the regression

$$y^*_{it} = \mu^* + \phi^* q_i + \beta x^*_{it} + \gamma\tilde{x}_{it} + w_{it}$$

Note now that we must assume that the q_i are independent of the α_i (this is not testable).

POOLED or RANDOM EFFECTS? The BREUSCH-PAGAN LM TEST

The RE model reduces to the POOLED model if the variance of the individual effects becomes zero. The hypothesis we wish to test is thus

H_0: $\sigma_\alpha^2 = 0$

H_1: $\sigma_\alpha^2 > 0$

LM tests are useful when it is easy to estimate the model under the null (here the POOLED model) and more complicated under the alternative (here the RE model)

The Breusch-Pagan statistic is calculated using the OLS residuals from the pooled model ($e = \hat{\varepsilon}_{ols}$)

$$LM = \frac{NT}{2(T-1)}\left(\frac{T^2\sum \bar{e}_{i.}^2}{\sum\sum e_{it}^2} - 1\right)^2 \sim \chi_1^2 \text{ under } H_0$$

- **Unfortunately, the Breusch-Pagan test is two-sided against the alternative $\sigma_\alpha^2 \neq 0$, in spite of the fact that we know that variances cannot be negative.**

- **An improvement suggested by HONDA is to use the one-sided test**

$$HONDA = \sqrt{\frac{NT}{2(T-1)}} \cdot \left(\frac{T^2 \sum \bar{e}_{i.}^2}{\sum\sum e_{it}^2} - 1 \right) \sim N(0,1) \text{ under } H_0$$

  A one-sided P-value is calculated; $P(x > HONDA)$

- **Another problem is that LM tests often have low power**

  Experiments have shown that in many cases it is better to use the CHOW test for FE against POOLED even if we suspect that RE is the correct alternative

SUR MODELS

SUR models can easily be tested against FE and POOLED models using Chow tests, if we assume homoscedasticity and independence between and within individuals.

SUR models can be estimated when there is heteroscedasticity and/or correlation between individuals. In this case we must adapt the Chow tests.

Testing SUR against RE always needs generalised Chow tests.

MISSPECIFICATION TESTS

When *T* is small it is very difficult to investigate the time series properties of panel data model. This is quite possible when *T* gets larger, however.

Misspecification testing should be performed on the most general model being considered

SMALL *T*

Autocorrelation can only be tested with great difficulty

Heteroscedasticity can be tested, but it is difficult to distinguish within individual from between individual differences.

TEST FOR HETEROSCEDASTICITY

The proposed test is the Bickel version of the Breusch-Pagan test. This tests for both within and between heteroscedasticity, and is performed in three steps

1) **Estimate the within regression. Obtain the residuals ($\hat{\tilde{\varepsilon}}_{it}$) and the predictions ($\hat{\tilde{y}}_{it} = y_{it} - \hat{\tilde{\varepsilon}}_{it}$)**
2) **Regress the squared residuals on powers of the predictions**

$$\hat{\tilde{\varepsilon}}_{it}^2 = \gamma_0 + \gamma_1 \hat{\tilde{y}}_{it} + \cdots + \gamma_p \hat{\tilde{y}}_{it}^p + w_{it}$$

3) **Test $\gamma_1 = \cdots = \gamma_p = 0$ with an *F*-test.**

With *N* being at least fairly large we can choose *p* to be of a reasonable size (values of *p* between 5 and 10 could be appropriate)

MEDIUM AND LARGE *T*

TEST FOR HETEROSCEDASTICITY

The proposed test is a small sample adjustment to Bartlett's test. We assume homoscedasticity within individuals and test for heteroscedasticity between individuals.

1) **Estimate the within regression and obtain the residuals ($\hat{\varepsilon}_{it}$)**

2) **Calculate the total residual variance**

$$s^2 = \frac{1}{NT-N-K}\sum\sum \hat{\varepsilon}_{it}^2$$

Remembering that $\bar{\hat{\varepsilon}}_{i.} = 0$, calculate the within individual variances

$$s_i^2 = \frac{1}{T-1}\sum_t \hat{\varepsilon}_{it}^2$$

3) **Calculate the Bartlett statistic**

$$B = \frac{(T-1)[N \ln s^2 - \sum \ln s_i^2]}{1+\{(N+1)/3(T-1)\}} \sim \chi^2_{N-1} \text{ under } H_0$$

Bickel's test can also be used if we suspect heteroscedasticity within individuals.

TESTS FOR AUTOCORRELATION

The first order within individual autocorrelation coefficient is calculated from the within regression residuals

$$r = \frac{\sum_{i=1}^{N}\sum_{t=2}^{T} \hat{\varepsilon}_{it}\hat{\varepsilon}_{i,t-1}}{\sum_{i=1}^{N}\sum_{t=2}^{T} \hat{\varepsilon}_{it}^2}$$

The simplest test is the LM test due to Breusch and Godfrey

$$LM = \sqrt{\frac{NT^2}{T-1}} \cdot r \sim N(0,1) \text{ under } H_0$$

The autocorrelation coefficient is known to have a slow convergence to normality, however, so a superior alternative is probably a test due to Fisher

$$z = \frac{\sqrt{NT-N-K}}{2} \cdot \ln\frac{1+r}{1-r} \sim N(0,1) \text{ under } H_0$$

ROBUST STANDARD ERRORS

If we discover (or even suspect) heteroscedasticity or serial autocorrelation we must decide what to do.

One approach is to try and model these variances and/or correlations. This can be difficult even for large *T*, and is generally impossible for small T.

An alternative approach is to accept the usual estimates, but to calculate their so called Robust Standard Errors.

- **If we only suspect heteroscedasticity then we can use WHITE'S ROBUST ERRORS.**
- **If we suspect heteroscedasticity and/or within individual autocorrelation we can use ARELLANO'S ROBUST ERRORS**

WHITE'S method is a standard approach performed by most econometric software.

The robust variances estimate is given for a fixed effects model with one exogenous variable as

$$\widehat{\mathrm{Var}}(\hat{\beta}) = \frac{\sum\sum \hat{\tilde{\varepsilon}}_{it}^2 \tilde{x}_{it}^2}{(\sum\sum \tilde{x}_{it}^2)^2}$$

where the residuals and variables are from the within regression.

In the general case with *K* exogenous variables the variance-covariance matrix is given by

$$\widehat{\mathrm{Var}}(\hat{\beta}) = (\tilde{X}'\tilde{X})^{-1}(\sum\sum \hat{\tilde{\varepsilon}}_{it}^2 \tilde{X}_{it}' \tilde{X}_{it})(\tilde{X}'\tilde{X})^{-1}$$

where $\tilde{X}$ is the $(NT \times K)$ "difference-from-mean" matrix of all the exogenous variables and $\tilde{X}_{it}$ is the $(1 \times K)$ row vector of variables for a given observation.

ARELLANO'S method is not standard.

For one variable we obtain

$$\widehat{\mathrm{Var}}(\hat{\beta}) = \frac{\sum_i (\sum_t \tilde{x}_{it} \hat{\tilde{\varepsilon}}_{it})^2}{(\sum\sum \tilde{x}_{it}^2)^2}$$

while in the general case we have

$$\widehat{\mathrm{Var}}(\hat{\beta}) = (\tilde{X}'\tilde{X})^{-1} (\sum \tilde{X}_i' \hat{\tilde{\varepsilon}}_i \hat{\tilde{\varepsilon}}_i' \tilde{X}_i) (\tilde{X}'\tilde{X})^{-1}$$

where $\tilde{X}_i$ is the $(T \times K)$ "difference-from-mean" matrix of exogenous variables, and $\hat{\tilde{\varepsilon}}_i$ is the $(T \times 1)$ vector of residuals, for the i^{th} individual

STRATEGY

1) **Test for Heteroscedasticity and Serial Correlation in the most general model available (SUR if possible, FE otherwise)**

2) **If there is no violation of the assumptions we test**

   a) **RE vs. FE (Hausman)**

   b) **POOLED vs. FE (Chow)**

   If (b) not significant $\Rightarrow$ Use POOLED model

   If (b) significant but (a) is not $\Rightarrow$ Use RE model

   If both tests significant $\Rightarrow$ Use FE model

   If *T* is large we can also test the FE model against SUR using a generalised Chow test or Wald test of $\beta_i = \beta \ \forall i$.

3) **If the assumptions are violated**

   a) **For small *T* estimate the FE model with Arellano standard errors**

   b) **For medium *T* use the following strategy in the FE model. Test against pooled after**

      i) **Adjusting for autocorrelation by making the model dynamic**

      ii) **Adjusting for heteroscedasticity between individuals by using weighted least squares**

   c) **For large *T* estimate the SUR model. Test against FE and pooled after making the model dynamic**

      Estimating SUR with the restriction $\beta_i = \beta \ \forall i$ is sometimes called Park's model.

ONE-WAY PANEL MODEL

We have written the one-way panel model as

$$y_{it} = \alpha_i + \beta x_{it} + \varepsilon_{it} \qquad \textbf{(1)}$$

This is often rewritten as

$$y_{it} = \alpha + \mu_i + \beta x_{it} + \varepsilon_{it} \qquad \textbf{(2a)}$$

$$\left.\begin{array}{ll} \textbf{where} & \alpha = \dfrac{1}{N}\sum \alpha_i \\ \textbf{and} & \mu_i = \alpha_i - \alpha \end{array}\right\} \Leftrightarrow \sum \mu_i = 0 \qquad \textbf{(2b)}$$

α is the AVERAGE individual effect, while μ_i is the individual DEVIATION FROM AVERAGE.

(2) seems to be just a complicated way of writing (1).

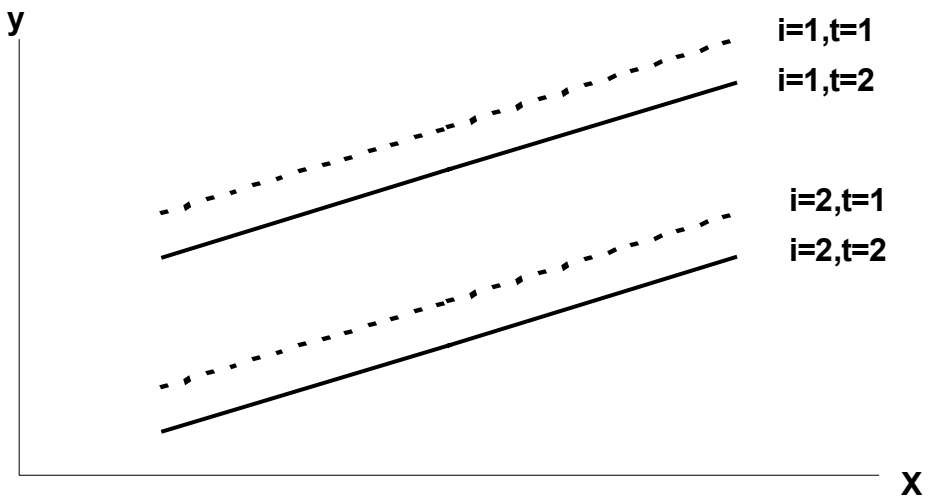
BUT it has the advantage that it can easily be extended to two-way models.

TWO-WAY PANEL MODELS

In the one-way model we assume that there exists an unobserved individual heterogeneity, but that the model is homogeneous over time.

Is it reasonable to assume that all time heterogeneity can be captured using observed explanatory variables?

Assume that the individual and time effects are additive, i.e. there is no interaction,



This is the Two-Way Panel Model

The Two-Way Panel Model is written

$$y_{it} = \alpha + \mu_i + \lambda_t + \beta x_{it} + \varepsilon_{it} \qquad \textbf{(3a)}$$

$$\text{where} \quad \sum_i \mu_i = 0, \quad \sum_t \lambda_t = 0 \qquad \textbf{(3b)}$$

We can define the individual/time effect as

$$\alpha_{it} = \alpha + \mu_i + \lambda_t \qquad \textbf{(4)}$$

Using the usual "dot" notation we obtain

$$\alpha = \bar{\alpha}_{\bullet\bullet} \equiv \frac{1}{NT} \sum_i \sum_t \alpha_{it} \qquad \textbf{average effect} \qquad \textbf{(5a)}$$

$$\alpha + \mu_i = \bar{\alpha}_{i\bullet} \equiv \frac{1}{T} \sum_t \alpha_{it} \qquad \textbf{individual effect} \qquad \textbf{(5b)}$$

$$\alpha + \lambda_t = \bar{\alpha}_{\bullet t} \equiv \frac{1}{N} \sum_i \alpha_{it} \qquad \textbf{time effect} \qquad \textbf{(5c)}$$

Note that some programmes report the individual effects as $\bar{\alpha}_{i\bullet}$, while others report μ_i

Note also that we can substitute (5) into (4) to obtain

$$\alpha_{it} - \bar{\alpha}_{i\bullet} - \bar{\alpha}_{\bullet t} + \bar{\alpha}_{\bullet\bullet} = 0 \qquad \textbf{(6)}$$

THE TWO-WAY MODEL WITH FIXED EFFECTS

The Two-Way model (3) has incidental parameters as either *N* or *T* go to infinity.

We need a new "within" transformation to remove these. We can see from (6) how this can be done

$$\tilde{\tilde{y}}_{it} = y_{it} - \bar{y}_{i\bullet} - \bar{y}_{\bullet t} + \bar{y}_{\bullet\bullet}$$

The Two-Way Within Model can thus be written

$$\tilde{\tilde{y}}_{it} = \beta \tilde{\tilde{x}}_{it} + \tilde{\tilde{\varepsilon}}_{it} \quad \Rightarrow \quad \text{OLS estimates } \hat{\beta}_w$$

The average, individual and time effects can now be estimated

$$\hat{\alpha}_w = \bar{y}_{\bullet\bullet} - \hat{\beta}_w \bar{x}_{\bullet\bullet}$$

$$\hat{\tilde{\alpha}}_{w,i\bullet} = \bar{y}_{i\bullet} - \hat{\beta}_w \bar{x}_{i\bullet}$$

$$\hat{\tilde{\alpha}}_{w,\bullet t} = \bar{y}_{\bullet t} - \hat{\beta}_w \bar{x}_{\bullet t}$$

- $\hat{\alpha}_w$ and $\hat{\beta}_w$ are consistent as either N or $T \to \infty$

  $\hat{\bar{\alpha}}_{w,i\bullet}$ is only T-consistent

  $\hat{\bar{\alpha}}_{w,\bullet t}$ is only N-consistent

- The Two-Way within transformation removes both *observed* and *unobserved* heterogeneity, for both *individual* and *time* effects.

  A dummy for an "oil-shock" or a "flu epidemic" will disappear in a FE estimation

- If T is small then the 2-Way FE model can easily be estimated using a 1-Way program. We write

$$y_{it} = \alpha + \mu_i + \beta x_{it} + \sum_{s=1}^{T-1} \lambda_s D_{st} + \varepsilon_{it}, \qquad (7)$$

  where D_s are dummies for year s. We can simply treat these dummies as explanatory variables.

  (Note that λ is now defined as the difference from year T, not difference from average.)

ONE and TWO-WAY MODELS with FIXED and RANDOM EFFECTS

A One-Way model has either fixed or random effects. Let

$\{\mu_F\}$, $\{\mu_R\}$, $\{\lambda_F\}$ and $\{\lambda_R\}$

denote the One-Way fixed and random models for individual and time effects.

In a Two-Way model both the individual effects and the time effects can be fixed or random. Let

$\{\mu_F, \lambda_F\}$ denote the fully FE model

$\{\mu_R, \lambda_F\}$ and $\{\mu_F, \lambda_R\}$ denote mixed FE/RE model

$\{\mu_R, \lambda_R\}$ denote the fully RE model

- **Estimation of the One-Way and fully FE Two-Way models have been described earlier**

THE FULLY RANDOM EFFECTS TWO-WAY MODEL

The model can be written

$y_{it} = \alpha + \beta x_{it} + u_{it}$, **with**

$u_{it} = \mu_i + \lambda_t + \varepsilon_{it}$,

where μ, λ, ε **and** x **are independent**

OLS will be consistent but inefficient. The efficient estimate is obtained by regressing y^{**} **on** x^{**}, **where**

$$y_{it}^{**} = y_{it} - \theta_1 \bar{y}_{i\cdot} - \theta_2 \bar{y}_{\cdot t} + \theta_3 \bar{y}_{\cdot\cdot}$$

and where

$$\theta_1 = 1 - \frac{\sigma_\varepsilon}{\sigma_1} \text{ with } \sigma_1^2 = T\sigma_\mu^2 + \sigma_\varepsilon^2, \qquad \textbf{(8a)}$$

$$\theta_2 = 1 - \frac{\sigma_\varepsilon}{\sigma_2} \text{ with } \sigma_2^2 = N\sigma_\lambda^2 + \sigma_\varepsilon^2 \text{ and} \qquad \textbf{(8b)}$$

$$\theta_3 = \theta_1 + \theta_2 + \frac{\sigma_\varepsilon}{\sigma_3} - 1 \text{ with } \sigma_3^2 = \sigma_1^2 + \sigma_2^2 - \sigma_\varepsilon^2 \qquad \textbf{(8c)}$$

PROBLEM: The θ are unknown.

If the two-way errors u and ε where known then

$$\hat{\sigma}_1^2 = \frac{T}{N-1}\sum_i \bar{u}_{i\bullet}^2 \qquad \textbf{(9a)}$$

$$\hat{\sigma}_2^2 = \frac{N}{T-1}\sum_t \bar{u}_{\bullet t}^2 \qquad \textbf{(9b)}$$

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{(N-1)(T-1)}\sum_i\sum_t \varepsilon_{it}^2 \qquad \textbf{(9c)}$$

Similar alternatives as in One-Way RE are available

- **WALLACE uses OLS residuals**
- **AMEMIYA uses within residuals**
- **SWAMY/ARORA uses the between individual and between time residuals for (9a) and (9b) and within residuals for (9c)**
- **NERLOVE estimates σ_μ^2 and σ_λ^2 Error! Not a valid link. directly from the FE model, and uses within residuals for (9c).**
- **+ more complicated alternatives**
- **It is common to adjust the denominators of (9) for degrees-of-freedom**

- **Negative estimates of σ_μ^2 and σ_λ^2 Error! Not a valid link. are possible for all methods except NERLOVE. A common procedure is to use max($\hat{\sigma}^2$,0).**

MIXED FE/RE TWO-WAY MODELS

If the number of time periods (or individuals) is small then a mixed model can be estimated by using RE on a One-Way model with dummies (as in (7))

Otherwise we proceed as follows

1) Adjust the model for the fixed effects

2) Adjust for the random effects

3) Regress adjusted *y* on adjusted *x* (no constant)

Step	Action	$\{\mu_R, \lambda_F\}$	$\{\mu_F, \lambda_R\}$
1)	Within transformation	$\ddot{\tilde{y}}_{it} = y_{it} - \bar{y}_{\bullet t}$	$\tilde{y}_{it} = y_{it} - \bar{y}_{i\bullet}$
2)	RE transformation	$\ddot{\tilde{y}}^*_{it} = \ddot{\tilde{y}}_{it} - \theta_1 \bar{\ddot{\tilde{y}}}_{i\bullet}$	$\tilde{y}^*_{it} = \tilde{y}_{it} - \theta_2 \bar{\tilde{y}}_{\bullet t}$
	Theta estimation	θ_1 from (8a) and (9a)	θ_2 from (8b) and (9b)
3)	RE regression (no constant)	$\ddot{\tilde{y}}^*_{it}$ on $\ddot{\tilde{x}}^*_{it}$	$\tilde{y}^*_{it}$ on $\tilde{x}^*_{it}$

Note that $\bar{\ddot{\tilde{y}}}_{i\bullet} = \bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}$ **and** $\bar{\tilde{y}}_{\bullet t} = \bar{y}_{\bullet t} - \bar{y}_{\bullet\bullet}$

PARAMETER TESTS

There are 9 different models when we allow for the possibility of both individual and time effects

$\{\mu_F, \lambda_F\}$	$\{\mu_R, \lambda_F\}$	$\{\lambda_F\}$
$\{\mu_F, \lambda_R\}$	$\{\mu_R, \lambda_R\}$	$\{\lambda_R\}$
$\{\mu_F\}$	$\{\mu_R\}$	**POOLED**

We must therefore choose:

- **The level (2-way, 1-way, pooled).**
  - **CHOW tests for FE**
  - **LM tests for RE**
- **The type of effects (FE, RE)**
  - **HAUSMAN tests for given level**

This is a "chicken-egg" problem. But

- **LM tests have poor power in small samples and are complicated to adjust.**
- **Chow tests have good power even in RE models**

TWO-WAY CHOW TESTS

The best power is obtained by always testing against the unrestricted two-way model

Model		# Parameters (p)	RSS
IT	$\{\mu_F, \lambda_F\}$	$N+T+K-1$	RSS_{IT}
I	$\{\mu_F\}$	$N+K$	RSS_I
T	$\{\lambda_F\}$	$T+K$	RSS_T
0	POOLED	K	RSS_0

We perform three Chow tests, for $m = 0, T, I$

$$CHOW_m = \frac{(RSS_m - RSS_{IT})/(N+T+K-1-p_m)}{RSS_{IT}/[(N-1)(T-1)-K]}$$

Reject 0/*IT* ⇒something	Reject *T*/*IT* ⇒individual	Reject *I*/*IT* ⇒time	Conclusion
YES	YES	YES	2-way
YES	YES	NO	Individual
YES	NO	YES	Time
NO	NO	NO	Pooled
YES	NO	NO	?? (2-way)
NO	YES	NO	? (Individ.)
NO	NO	YES	? (Time)
NO	YES	YES	?? (2-way)

TWO-WAY HAUSMAN TESTS

The best power is obtained by always testing against the fully FE model The omitted variables variant of the tests are as follow

1) To test $\{\mu_R, \lambda_R\}$ against $\{\mu_F, \lambda_F\}$

Regress y^{**} on x^{**}, $\tilde{\tilde{x}}$. Test the coefficients of $\tilde{\tilde{x}}$

2) To test $\{\mu_R, \lambda_F\}$ against $\{\mu_F, \lambda_F\}$

Regress $\ddot{\tilde{y}}^*$ on $\ddot{\tilde{x}}^*$, $\tilde{\tilde{x}}$ (no constant). Test $\tilde{\tilde{x}}$

3) To test $\{\mu_F, \lambda_R\}$ against $\{\mu_F, \lambda_F\}$

Regress $\tilde{y}^*$ on $\tilde{x}^*$, $\tilde{\tilde{x}}$ (no constant). Test $\tilde{\tilde{x}}$

Reject (1) $\Rightarrow$some FE	Reject (2) $\Rightarrow$ind. FE	Reject (3) $\Rightarrow$time FE	Conclusion
YES	YES	YES	$\{\mu_F, \lambda_F\}$
YES	YES	NO	$\{\mu_F, \lambda_R\}$
YES	NO	YES	$\{\mu_R, \lambda_F\}$
NO	NO	NO	$\{\mu_R, \lambda_R\}$
YES	NO	NO	?? $\{\mu_F, \lambda_F\}$
NO	YES	NO	? $\{\mu_F, \lambda_R\}$
NO	NO	YES	? $\{\mu_R, \lambda_F\}$
NO	YES	YES	?? $\{\mu_F, \lambda_F\}$

STRATEGY

1) **Test for Heteroscedasticity and Serial Correlation in Two-Way FE model**

2) **If there is no violation of the assumptions we**
   a) **First choose level with CHOW tests**
   b) **Then decide RE/FE with Hausman tests**

3) **If the assumptions are violated: see p. 53.**

INCOMPLETE PANELS

Panel data studies where all individuals are observed at each time period are called COMPLETE.

INCOMPLETE surveys are those with missing data. These can occur for several reasons

1) We can plan our survey so that it is incomplete. We have DETERMINISTIC missing data

2) The missing data is unplanned, but the selection rule is independent of the data (observed and unobserved). We have RANDOMLY missing data.

3) There is a correlation between the selection rule and the data. There is a SELECTION BIAS

Complete surveys are BALANCED, *i.e.* each individual and each time period is observed equally often (*N* and *T* respectively).

Stochastic missing data is UNBALANCED, while deterministic missing data can be either.

- **DETERMINISTIC and RANDOM missing values are methodologically equivalent**
- **UNBALANCED models without selection bias only cause technical problems**
- **The data for unbalanced panels is written**

  $\{y_{it}, x_{it}\}$ **for** $i = 1,\ldots,N, \quad t = 1,\ldots,T_i$

- **SELECTION BIAS is a serious problem that needs complicated estimation methods in panel data models**
- **Missing data is often caused by ATTRITION; the tendency of individuals to drop out of surveys that stretch over many periods. We often suspect that the causes of attrition are correlated with the data.**

UNBALANCED PANELS

Assumption: There is no selection bias

ONE-WAY FIXED EFFECTS

- **The individual means are redefined:**

$$\bar{y}_{i\cdot} = \frac{1}{T_i}\sum_t y_{it}$$

- **As in the balanced model we regress $\tilde{y}$ on $\tilde{x}$**

ONE-WAY RANDOM EFFECTS

- **The GLS transformation is now: $y_{it}^* = y_{it} - \theta_i \bar{y}_{i\cdot}$**

  with $\theta_i = 1 - \frac{\sigma_\varepsilon}{\omega_i}$ and $\omega_i^2 = T_i\sigma_\mu^2 + \sigma_\varepsilon^2$

- **The variances can be estimated consistently as**

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N(\bar{T}-1)-K}\sum\sum \hat{\tilde{\varepsilon}}_{it}^2 \text{, where } \bar{T} = \frac{1}{N}\sum T_i$$

$$\hat{\sigma}_b^2 = \frac{RSS}{N-K} \text{ from regressing } \sqrt{T_i}\,\bar{y}_{i\cdot} \text{ on } \sqrt{T_i}\,\bar{x}_{i\cdot}$$

$$\hat{\sigma}_\mu^2 = \frac{\hat{\sigma}_b^2 - \hat{\sigma}_\mu^2}{\bar{T}}$$

- **The estimates are obtained by regressing y^* on x^***

<u>TESTING</u> Chow and Hausman (omitted variables) tests as before. LM tests must be adjusted slightly

<u>TWO-WAY MODELS</u> are messy, but not difficult

<u>SELECTION BIAS</u> models are difficult to estimate (we need numerical integration). Some simple specification tests exist, however

- **Hausman-type tests are available if we estimate the model in the full (unbalanced) sample and a balanced sub-sample. The two methods should give the same results if there is no selection bias**
- **Omitted variable tests can be used with such extra variables as**
  - **# times i^{th} individual is in sample**
  - **dummy for whether i^{th} individual is in the whole sample**
  - **dummy for whether i^{th} individual was present in the previous period**

ROTATING PANELS

Surveyors are wary of designs where individuals have to answer questions many times over a long period of time. This often leads to a large degree of attrition, which can very well include selection bias

One method of avoiding this is to introduce a deterministic attrition. By only interviewing each individual a few times we hope to reduce the stochastic attrition.

The most common deterministic design is the method of ROTATING PANELS.

	Period 1	Period 2	Period 3	Period 4
Wave 1	N	$N/2$		
Wave 2		$N/2$	$N/2$	
Wave 3			$N/2$	$N/2$
Wave 4				$N/2$

DYNAMIC MODELS

Dynamic models include lagged values of the endogenous variable on the RHS (they can also include lagged exogenous variables)

$$y_{it} = \alpha_i + \delta y_{i,t-1} + \beta x_{it} + \varepsilon_{it}$$

ε_{it} is assumed independent of $x_{it}, y_{i,t-1}$

$\Rightarrow$ ε_{it} and y_{it} are dependent

$\Rightarrow$ ε_{it} is correlated with $\bar{y}_{i.}$ (at least for small *T*)

$\Rightarrow$ ε_{it} is also correlated with $\tilde{y}_{it}$

$\Rightarrow$ $\hat{\beta}_w$ is only *T*-consistent, not *N*-consistent

The standard way of estimating models with correlation between the errors and the RHS variables is to use the INSTRUMENTAL VARIABLES method

INSTRUMENTAL VARIABLES (IV)

Consider a simple linear regression

$$y = \alpha + \beta x + \varepsilon$$

where $\mathrm{E}(\varepsilon x) \neq 0$.

A variable *z* is called an *instrumental variable* if

$\mathrm{E}(\varepsilon z) = 0$ **and** $\mathrm{Cov}(x, z) \neq 0$

The IV estimate of β is given by

$$\hat{\beta}_{IV} = \frac{\sum (y - \bar{y})(z - \bar{z})}{\sum (x - \bar{x})(z - \bar{z})}$$

- **In matrix terms $\hat{\beta}_{IV} = (Z'X)^{-1} Z'y$ if there are as many instruments as RHS variables.**
- **If there are more instruments than RHS variables then $\hat{\beta}_{IV} = (\hat{X}'X)^{-1} \hat{X}'y$, where $\hat{X}$ are the fitted values from the regression of *X* on *Z*.**
- **If a variable is exogenous it will be its own instrument**

IV ESTIMATION OF DYNAMIC MODELS

The problem with the IV method is how to find good instruments.

In the panel data model life is made easier if we get rid of the incidental parameters using a different transformation

Since

$y_{it} = \alpha_i + \delta y_{i,t-1} + \beta x_{it} + \varepsilon_{it}$ **and**

$y_{i,t-1} = \alpha_i + \delta y_{i,t-2} + \beta x_{i,t-1} + \varepsilon_{i,t-1}$

we can remove the individual heterogeneity by taking first differences

$\Delta y_{it} = \delta(\Delta y_{i,t-1}) + \beta(\Delta x_{it}) + \Delta\varepsilon_{it}$

- $\Delta\varepsilon_{it} = \varepsilon_{it} - \varepsilon_{i,t-1}$ **is correlated with** $\Delta y_{i,t-1} = y_{i,t-1} - y_{i,t-2}$
- $\Delta\varepsilon_{it} = \varepsilon_{it} - \varepsilon_{i,t-1}$ **will be uncorrelated with** $y_{i,t-2}$ **as long as there is no autocorrelation in** ε
- $\Delta y_{i,t-1} = y_{i,t-1} - y_{i,t-2}$ **is obviously correlated with** $y_{i,t-2}$**, which should therefore make a good IV.**

Using $y_{i,t-2}$ as an instrument for $\Delta y_{i,t-1}$ in the IV regression of Δy_{it} on $\Delta y_{i,t-1}, \Delta x_{it}$ will therefore give consistent, but not efficient, estimates (as long as there is no autocorrelation in the errors).

Better estimates can be obtained by

- **Using more instruments ($y_{i,t-3}, y_{i,t-4}$, *etc*)**
- **Taking account of the fact that $\Delta\varepsilon_{it} = \varepsilon_{it} - \varepsilon_{i,t-1}$ (this implies using GIV)**
- **Taking account of the structure of the random effects**

The dynamic modelling of panel data has an extensive recent literature. The estimation methods can be quite complicated, however.

LIMITED DEPENDENT VARIABLES

Limited Dependent Variables Models are those where the dependent variable is not completely continuous

y variable	Example	Models
Binary Choice	sick/not sick	Logit, Probit
Multiple Choice		
Categorical	type of absence	Multinomial Logit/Probit
Hierarchical	routes within means of transport	Nested Logit
Ordinal	school grades	Ordered logit/probit
Count models (Cardinal)	# days sick	Poisson, Neg. Binomial
Multivariate Choice	sick/not sick and work/unemployed	Bivariate Probit
Limited Continuous Variables		
Truncated	wages (subpop)	Trunc. Reg.
Censored	wages (whole pop)	Tobit
Sample Selection	hours worked given employment	Heckit

BINARY CHOICE

Each individual can "choose" between two events

Examples: employed/unemployed,
improve/not improve

Data: dependent variable $y = \begin{cases} 0 & \text{event 0} \\ 1 & \text{event 1} \end{cases}$

explanatory variables x**:** $k \times 1$

Model: $P(y = 1) = F(\beta' x)$

- $\beta' x = \sum \beta_j x_j$, **where** x_1 **is usually the constant**

- **Note that** y **is a binomially distributed**

- **for given** x **with** $P = F(\beta' x)$

  $\Rightarrow \mathrm{E}(y \mid x) = P$ **and** $\mathrm{Var}(y \mid x) = P(1 - P)$

MARGINAL EFFECTS

The change in the dependent variable (y) for a given change in an explanatory variable (x_j) is called the MARGINAL EFFECT of that variable.

If x_j is continuous: $ME_c = \dfrac{\partial \mathrm{E}(y)}{\partial x_j}$

If x_j is a dummy:

$ME_d = \mathrm{E}(y \mid x_j = 1) - \mathrm{E}(y \mid x_j = 0)$

In the LINEAR model $ME_c = ME_d = \beta$

In NONLINEAR models

- $ME \neq \beta$
- ***ME* is a function of *x***
- $ME_c \neq ME_d$**, but often** $ME_c \approx ME_d$
- **We are not always so interested in** β

LINEAR PROBABILITY MODEL (LPM)

The simplest binary choice model is to assume that F is linear

$$F(\beta' x) = \beta' x \;\Rightarrow\; \mathrm{E}(y \mid x) = \beta' x$$

$$\Rightarrow y_i = \beta' x_i + \varepsilon_i$$

$\Rightarrow$ We can fit a linear regression, treating y as an ordinary variable.

A technical problem is that ε is heteroscedastic, since

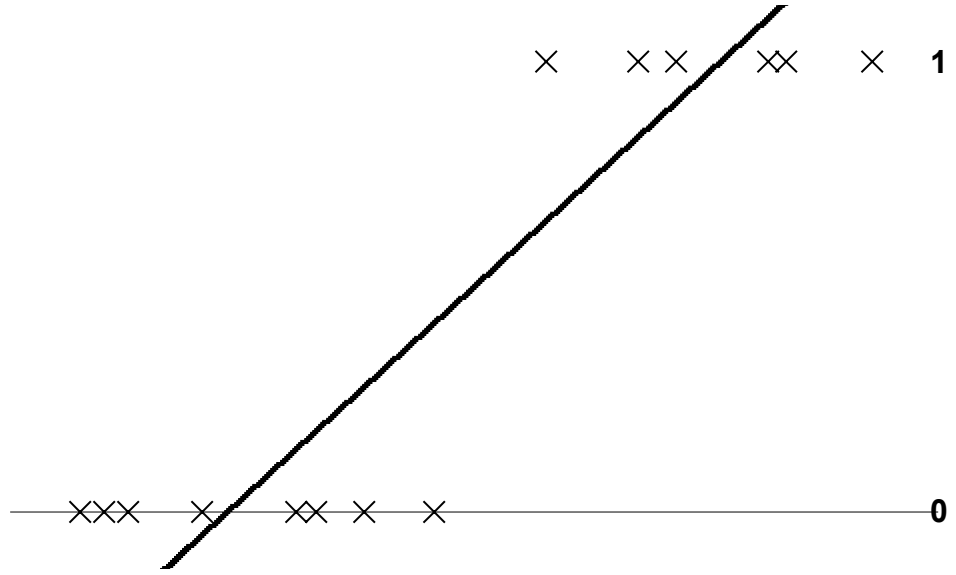
$$\mathrm{Var}(\varepsilon_i) = P(1-P) = (\beta' x_i)(1-\beta' x_i) \neq \text{constant}$$

LPM can be estimated using

- **OLS, with White's robust variance estimates**
- **WLS**
- **ML**

A more serious problem is the assumption behind the model.

Plotting *y* against *x* in a model with only one explanatory variable yields



The OLS line estimates $P(y=1)<0$ for some values of *x*, and $P(y=1)>1$ for some others!!

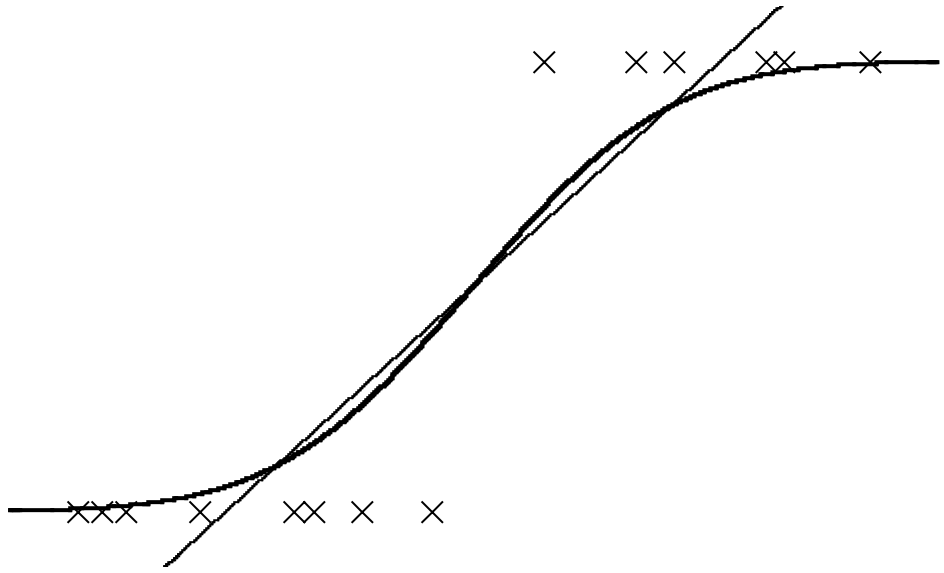
PROBIT and LOGIT MODELS

Two commonly used functions that always lie in the interval (0,1) are

PROBIT: $F(\beta' x) = \Phi(\beta' x)$**, the standard normal distribution function**

LOGIT: $F(\beta' x) = \dfrac{e^{\beta' x}}{1 + e^{\beta' x}}$

The LOGIT function was first proposed as an approximation to the PROBIT. In most cases they give very similar results.



LATENT VARIABLE INTERPRETATION

There is an alternative interpretation to these binary choice models, which is in some ways more attractive

Assume that there is a "true", but unobservable, variable y^*, *e.g.* the *propensity* to be sick. There is also an observed variable *y*, the *incidence* of being sick

The latent variable is explained in an ordinary linear regression

$$y^* = \beta' x + \varepsilon, \qquad \textbf{(1)}$$

and the observed variable is given by

$$y = \begin{cases} 0 & y^* < 0 \\ 1 & y^* \geq 0 \end{cases} \qquad \textbf{(2)}$$

PROBIT $\Leftrightarrow$ ε is normally distributed

LOGIT $\Leftrightarrow$ ε is logistically distributed

IDENTIFICATION PROBLEMS

Multiplying (1) by a positive constant
$\Rightarrow$ the sign of y^* is unchanged
$\Rightarrow$ y is unchanged
$\Rightarrow$ β is unidentified

However

- **The sign of β_j and the ratio β_j / β_ℓ are identified,.**
- **The marginal effects are also identified**
- **The probit model is normalised by letting ε be *standard* normally distributed ($\sigma^2 = 1$)**
- **Imposing the logistic distribution normalises β**
- **These normalised parameters are related**

  $\beta_{logit} \approx 1.6\beta_{probit} \approx 4\beta_{LPM}$, except for the constant term where $\beta_{logit} \approx 1.6\beta_{probit} \approx 4\beta_{LPM} - 2$.

ML ESTIMATION

The nonlinear probit and logit models are estimated using maximum likelihood

$$\text{Likelihood} = \text{Joint Probability of sample}$$

$$= \prod_{y_i=1} P(y_i = 1) \prod_{y_i=0} P(y_i = 0)$$

$$= \prod_{i} [F(\beta' x_i)]^{y_i} [1 - F(\beta' x_i)]^{1-y_i}$$

and thus

$$\text{log likelihood} = \sum_{i} \left\{ y_i \ln F(\beta' x_i) + (1 - y_i) \ln[1 - F(\beta' x_i)] \right\}$$

This is easy to maximise iteratively for LOGIT and PROBIT models

PREDICTIONS

What do we mean by a prediction from a binary choice model? The intuitive $F(\hat{\beta}'x)$ is a prediction of $P(y=1)$, not of y.

The standard definition is

$$\hat{y} = \begin{cases} 0 & \hat{y}^* < 0 \\ 1 & \hat{y}^* \geq 0 \end{cases}$$

where $\hat{y}^* = \hat{\beta}'x$. Note that

$$\hat{y}^* \geq 0 \Leftrightarrow F(\hat{y}^*) \geq 0.5$$

This rule seems reasonable if $P_{obs} \approx 0.5$, where P_{obs} is the observed proportion of ones amongst the y's. It will however lead to nearly all the prediction being zeroes or ones if P_{obs} is small (large)

An alternative definition is therefore

$$\hat{\hat{y}} = \begin{cases} 0 & F(\hat{y}^*) < P_{obs} \\ 1 & F(\hat{y}^*) \geq P_{obs} \end{cases}$$

The proportion $\hat{\hat{y}}$'s that equal 1 is now P_{obs}

GOODNESS-OF-FIT

How do we judge if our binary choice model is a good one? There is no obvious equivalent to the usual R^2 measure.

There are four common measures

1) $\text{Cor}(y, \hat{y})$

2) Effron's $R^2 = 1 - \dfrac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}$

3) McFadden's $R^2 = 1 - \dfrac{\ln L}{\ln L_0}$,

where L is the likelihood from the estimated model and L_0 is from the model with only a constant

4) $\dfrac{\#\text{ correct predictions } (\hat{y}_i = y_i)}{N}$

The first two measures reduce to the ordinary R^2 measure in a linear model.

We can of course replace $\hat{y}$ with $\hat{\hat{y}}$ in all except (3)

RESULTS IN LIMDEP

LIMDEP includes the following output when using the PROBIT/LOGIT commands

1) LPM (start values)

2) PROBIT/LOGIT model

3) Measure of fit (4)

4) LogL and LogL0 $\Rightarrow$ Measure of fit (3)

One can request the marginal effects

1) $ME_c(\bar{x})$; evaluated at the *average* of the *x*'s

2) $ME_c(\bar{x}_s)$; evaluated at strata averages

Note that $\overline{ME_c(x_i)}$ and $\overline{ME_d(x_i)}$ must be explicitly calculated

One can also save

1) The predictions $\hat{y}$

2) The residuals $y - \bar{y}$

3) The probabilities $F(\hat{y}^*) \Rightarrow \hat{\hat{y}}$

PANEL DATA LOGIT/PROBIT

A panel data binary choice model can be written

$$y_{it}^{*} = \alpha + \mu_i + \beta' x_{it} + \varepsilon_{it},$$

where the observed variable is given as usual by

$$y = \begin{cases} 0 & y^{*} < 0 \\ 1 & y^{*} \geq 0 \end{cases}$$

There are two problem here

1) **FIXED EFFECTS: The incidental parameters cannot be swept away by a simple transformation of the data**

2) **RANDOM EFFECTS: Maximising the likelihood involves numerical integration over *T*-dimensions**

Fixed Effect Panel Logit - Chamberlain's Approach

Chamberlain has shown that the incidental parameters are removed if the likelihood is conditioned on $\sum_t y_{it}$.

The number of observations and regressors available are substantially reduced in the FE logit model

- **Individuals that have the same *y* for all time periods don't contribute to the likelihood and can be removed.**
- **As for all fixed effects models we cannot include individual specific regressors.**
- **As for all unbalanced panels we can remove all individuals that are only observed once**

Other "problems" are

- **There is no obvious way of estimating the individual effects or the marginal effects**
- **It would, be possible to estimate the marginal effects for the "average" person, *i.e.*, when $\mu_i = 0$. LIMDEP doesn't do this, however.**
- **No "conditional ML" Probit is available**

Random Effect Panel Probit and Logit

The full RE model is not possible to estimate. The usual approach is to impose the "equi-correlation" restriction

$$\text{Cor}(u_{it}, u_{is}) = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_\varepsilon^2} \equiv \rho$$

The Probit model also assumes that $\varepsilon \sim N(0,1)$

The Logit model assumes that ε is logistic distr.

Both assume that μ_i is normally distributed

The marginal effects are difficult to estimate since $\text{E}(y)$ is now highly nonlinear. Replacing μ_i with its expected value (zero) in $\text{E}(y)$ leads to the usual probit/logit formulae, however.

TESTS

Testing FE vs. RE vs. POOLED can be performed with Hausman tests. In the Probit model a Wald test is also available for RE vs. POOLED ($\rho = 0$).

LRT or *F*-tests are used for parameter testing.

OTHER LIMDEP MODELS

Multinomial Logit (MNL)

Each individual "chooses" between alternatives $0,1,\ldots,J$. Thus $y_i = j$ if alternative j is chosen

The explanatory variables x_{ij} are of two types

z_{ij}: the choice specific *attributes*
w_i: the individual specific *characteristics*

The multinomial logit model is written

$$P(y_i = j) = \frac{e^{\beta' z_{ij} + \gamma_j' w_i}}{e^{\beta' z_{i0}} + \sum_{\ell=1}^{J} e^{\beta' z_{i\ell} + \gamma_\ell' w_i}}$$

where γ_0 is normalised to 0.

Note that the attributes have choice independent parameters, while the characteristics *can* have choice dependent parameters.

A property of the MNL model is the so called Independence of Irrelevant Alternatives (IIA). When choosing between alternatives 1 and 2 it does not matter if alternative 3 exists or not.

The multinomial Probit model (MNP) does not impose IIA automatically. Estimating MNP needs Monte Carlo integration, however.

In LIMDEP we estimate MNL models using

- **LOGIT if there are no attributes ($z_{ij} = 0$)**
- **NLOGIT if the characteristics have $\gamma_j = \gamma$**
- **NLOGIT + choice dummies/interactions for the general model**

MNP models can also be estimated in LIMDEP

Ordered Logit/Probit

In these models we assume that there is a strict ranking between the alternatives (the classic example is school grades).

The model is given by

$$P(y_i = j) = F(\kappa_{j-1} < y_i^* \leq \kappa_j)$$

where the κ's are to be estimated. F is logistic or normal depending on whether an Ordered Logit or Ordered Probit is used.

Random Effect versions of the ordered models are also available in LIMDEP

Count Models

If we are modelling, for example, the number of sick days it may not seem unreasonable to assume that

y ~ Poisson(λ), where $\ln\lambda = \beta' x$

A problem with a Poisson regression is that it forces the mean and variance of y (given x) to be equal. An alternative which allows for "overdispersion" is the Negative Binomial regression

In many cases the Poisson and Neg.Bin. models underestimate the number of zeroes. This may be due to fact that there are two processes at work

1) **A binary choice model, which determines if we report sick**
2) **A count model, which determines how many days we are absent if we report sick**

Such models are called Zero-Inflated.

LIMDEP estimates Poisson, NegBin, ZIP, ZINB, Fixed and Random Effect Poisson and NegBin., and some sample selection count models

Truncated and Censored Models

In binary choice models we only observe, for example, if consumption occurs. In a censored model we observe the amount of consumption, but only if it is non-negative. The model is

$$y_i^* = \beta' x_i + \varepsilon_i$$

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* > L_i \\ L_i & \text{if } y_i^* \le L_i \end{cases}$$

This model is *left censored* if x_i is observed for all observations and *left truncated* if x_i is only observed when $y_i = y_i^*$. Right censoring and truncation are also possible.

The limit, L_i, can be a constant or a variable. If y is observed consumption, then y_i^* is the *propensity to consume* and the limit value is zero.

Censored and Truncated models usually assume that ε is normally distributed. In this case the censored model is usually called a TOBIT model

Random Effects, nested, bivariate and sample selection forms of the TOBIT model are available in LIMDEP.

Sample Selection

In many situations (more than we like to think) the data we have available has not been obtained randomly from the population of interest.

In addition there may well be a correlation between the mechanism that determines what data is observed and the process we are interested in.

For example, physicians may tend to "deselect" patients considered too ill to take part in a clinical trial. Too few of these patients will therefore participate, which will obviously bias our results.

Sample selection models consist of two parts

1) A selection model (probit, logit, *etc*)

2) An explanatory model (regression, tobit, *etc*).

These models are very easily estimated in LIMDEP.

Note the main problem consists of formulating a reasonable selection model!