PANEL DATA ANALYSIS

Ass.Prof. Andriy Stavytskyy

Agenda

- Panel Data
- Panel Data DGP’s
- Fixed Effects
- Random Effects
- The Hausman Test

PANEL DATA

Panel Data

- **Panel Data** is data in which we observe repeated cross-sections of the same individuals.
- Examples:
  - *Annual unemployment rates of each country over several years*
  - *Quarterly sales of individual stores over several quarters*
  - *Wages for the same worker, working at several different jobs*

Panel Data: Motivation – 1

- With cross-sectional data, there is no particular reason to differentiate between omitted variables that are fixed over time and omitted variables that are changing.

- However, when an omitted variable is fixed over time, panel data offers another tool for eliminating the bias.

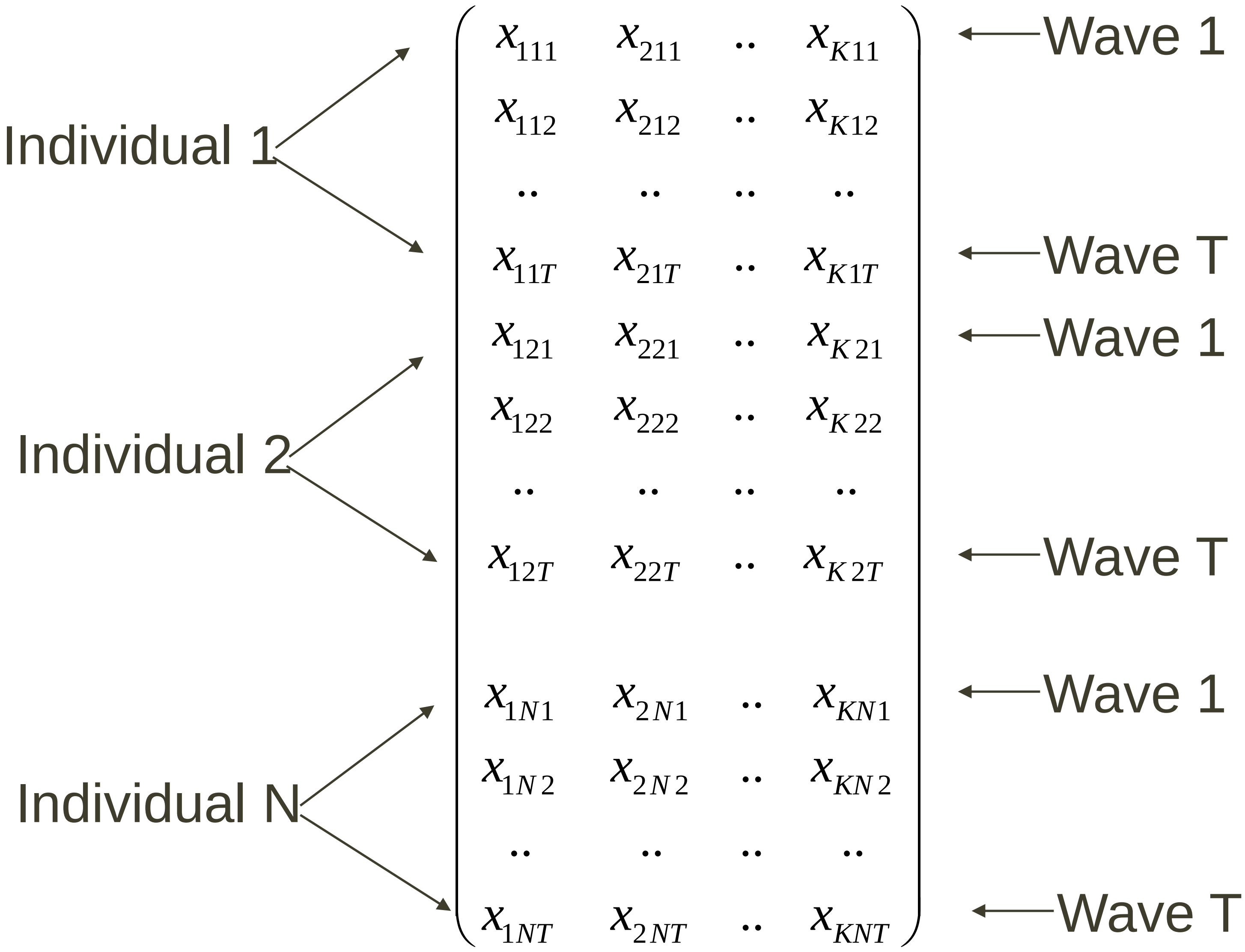
Panel Data: Motivation – 2

- The key feature of panel data is that we observe the same individual in more than one condition.
- Omitted variables that are fixed will take on the same values each time we observe the same individual.

Panel Data: Motivation – 3

- Some of the most valuable data sets in economics are panel data sets.
- Longitudinal surveys return year after year to the same individuals, tracking them over time.

The Basic Data Structure



$$
\begin{array}{l}
\text{Individual 1} \\
\\
\\
\\
\text{Individual 2} \\
\\
\\
\\
\\
\text{Individual N} \\
\\
\\
\\
\end{array}
\left(
\begin{array}{cccc}
x_{111} & x_{211} & .. & x_{K11} \\
x_{112} & x_{212} & .. & x_{K12} \\
.. & .. & .. & .. \\
x_{11T} & x_{21T} & .. & x_{K1T} \\
x_{121} & x_{221} & .. & x_{K21} \\
x_{122} & x_{222} & .. & x_{K22} \\
.. & .. & .. & .. \\
x_{12T} & x_{22T} & .. & x_{K2T} \\
 & & & \\
x_{1N1} & x_{2N1} & .. & x_{KN1} \\
x_{1N2} & x_{2N2} & .. & x_{KN2} \\
.. & .. & .. & .. \\
x_{1NT} & x_{2NT} & .. & x_{KNT}
\end{array}
\right)
\begin{array}{l}
\leftarrow \text{Wave 1} \\
\\
\\
\leftarrow \text{Wave T} \\
\leftarrow \text{Wave 1} \\
\\
\\
\leftarrow \text{Wave T} \\
\\
\leftarrow \text{Wave 1} \\
\\
\\
\leftarrow \text{Wave T}
\end{array}
$$

Formulate an hypothesis

$$y_{it} = f(x_{1it}, x_{2it}, \ldots, x_{kit})$$

Example: Cross-Industry Wage Disparities – 1

- A great puzzle in labor economics is the presence of cross-industry wage disparities.
- Workers of seemingly equivalent ability, in seemingly equivalent occupations, receive different wages in different industries.
- Do high-wage industries actually pay higher wages, or do they attract workers of unobservably higher quality?

Example: Cross-Industry Wage Differentials – 2

- Gibbons and Katz (Review of Economic Studies 1992) exploited panel data to explore these differentials.
- They observed workers in 1984 and 1986.
- They focused on workers who lost their 1984 jobs because of plant closings (on the grounds that plant closings are unlikely to be correlated with an individual worker's abilities). They looked only at workers who were re-employed by 1986.

Example: Cross-Industry Wage Differentials – 3

- Gibbons and Katz estimated wages as

$$\ln w_{it} = \delta_o + \delta_1 X_{1it} + .. + \delta_k X_{kit} + \alpha_1 D_{it}^{Industry\_1}$$
$$.. + \alpha_m D_{it}^{Industry\_m} + \varepsilon_{it}$$

where

- X_{kit} are demographic variables,
- D_{it} are a set of dummy variables for being employed in different industries

Example: Cross-Industry Wage Differentials – 4

Estimating with simple OLS, Gibbons and Katz estimate α's that are very similar to other estimates of cross-industry wage differentials.

$$\boxed{\begin{gathered}\ln w_{it} = \delta_o + \delta_1 X_{1it} + .. + \delta_k X_{kit} + \alpha_1 D_{it}^{Industry\_1} \\ .. + \alpha_m D_{it}^{Industry\_m} + \varepsilon_{it}\end{gathered}}$$

Example: Cross-Industry Wage Differentials – 5

- Gibbons and Katz speculated that any unmeasured ability is fixed over time and equally rewarded in all industries.

$$\ln w_{it} = \delta_o + \delta_1 X_{1it} + .. + \delta_k X_{kit} + \alpha_1 D_{it}^{Industry\_1}$$
$$.. + \alpha_m D_{it}^{Industry\_m} + v_i + \mu_{it}$$

- Differencing the 1986 and 1984 observations eliminated the v_i

Example: Cross-Industry Wage Differentials – 6

Differencing the 1986 and 1984 observations eliminated the v_i

$$\begin{aligned}
\ln w_{i1986} &= \delta_o + \delta_1 X_{1i1986} + .. + \delta_k X_{ki1986} + \alpha_1 D_{i1986}^{Industry\_1} \\
&\quad .. + \alpha_m D_{i1986}^{Industry\_m} + v_i + \mu_{i1986} \\
-\ln w_{i1984} &= \delta_o + \delta_1 X_{1i1984} + .. + \delta_k X_{ki1984} + \alpha_1 D_{i1984}^{Industry\_1} \\
&\quad \underline{.. + \alpha_m D_{i1984}^{Industry\_m} + v_i + \mu_{i1984}} \\
\Delta \ln w_i &= \delta_1 \Delta X_{1i} + .. + \delta_k \Delta X_{ki} + \alpha_1 \Delta D_i^{Industry\_1} \\
&\quad .. + \alpha_m \Delta D_i^{Industry\_m} + \Delta \mu_i
\end{aligned}$$

Example: Cross-Industry Wage Differentials – 7

- The estimated industry coefficients from the differenced equation are about 80% of the estimated industry coefficients from the levels equation.
- Unobserved worker ability appears to explain relatively little of the cross-industry wage differentials.

PANEL DATA DGP

A Panel Data DGP (data gathering panel) – 1

$$Y_{it} = \beta_{0i} + \beta_1 X_{1it} + \beta_2 X_{2i} + \beta_3 X_{3t} + .. + \beta_K X_{Kit} + \varepsilon_{it}$$

$$i = 1...n; \quad t = 1...T$$

$$E(\varepsilon_{it}) = 0$$

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

$E(\varepsilon_{it}\varepsilon_{i't'}) = 0$ if $i \neq i'$ OR $t \neq t'$

$E(X_{jit}\varepsilon_{it}) = 0$ for all j, i, t

Panel Data DGPs

- Notice that when we have panel data, we index observations with both **i** and **t**.
- Pay close attention to the subscripts on variables.
- Some variables vary only across time or across individual.

A Panel Data DGP (data gathering panel) – 2

$$Y_{it} = \beta_{0i} + \beta_1 X_{1it} + \beta_2 X_{2i} + \beta_3 X_{3t} + .. + \beta_K X_{Kit} + \varepsilon_{it}$$

$$i = 1...n; \quad t = 1...T$$

For example,

X_{2i} varies only by individual, and is fixed over time.

X_{2i} might be a variable such as race or gender.

X_{3t} varies only by time, and is fixed across i.

X_{3t} might be national unemployment.

X_{1it} varies across BOTH individual and time.

X_{1it} might refer to wages.

A Panel Data DGP (data gathering panel) – 3

$$Y_{it} = \beta_{0i} + \beta_1 X_{1it} + \beta_2 X_{2i} + \beta_3 X_{3t} + .. + \beta_K X_{Kit} + \varepsilon_{it}$$

$$i = 1...n; \quad t = 1...T$$

One of the key features of the DGP is that we allow each individual i to have a distinct intercept β_{0i}. This intercept includes ALL aspects of unobserved heterogeneity that are fixed over the length of the panel.

A Panel Data DGP (data gathering panel) – 4

- In this DGP, the β_{0i} are fixed across samples.
- The unmeasured heterogeneity is the same in every sample.
- It is suitable for panels of states or countries, where the same individuals would be selected in each sample.

A Panel Data DGP (data gathering panel) – 5

- With longitudinal data on individual workers or consumers, we draw a different set of individuals from the population each time we collect a sample.
- Each individual has his/her own set of fixed omitted variables.
- We cannot fix each individual intercept.

Another Panel Data DGP – 1

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2i} + \beta_3 X_{3t} + .. + \beta_K X_{Kit} + v_i + \mu_{it}$$

$$i = 1...n; \quad t = 1...T$$

$E(\mu_{it}) = 0$ $\qquad$ $Var(\mu_{it}) = \sigma_\mu^2$

$E(\mu_{it}\mu_{i't'}) = 0$ if $i \neq i'$ OR $t \neq t'$ $\qquad$ $E(v_i) = 0$

$E(v_i v_{i'}) = 0$ for $i \neq i'$ $\qquad$ $Var(v_i) = \sigma_v^2$

$E(\mu_{it} v_{i'}) = 0$ for all i, i', t

$E(x_{jit}\mu_{it}) = 0$ for all j, i, t

EITHER $E(X_{jit} v_i) = 0$ for all j, i, t

OR $E(X_{jit} v_i) \neq 0$ for at least some j, i, t

Another Panel Data DGP – 2

- In this DGP, we return to a model with a single intercept for all data points, β_0
- However, we break the error term into two components:

$$\varepsilon_{it} = v_i + \mu_{it}$$

- When we draw an individual *i*, we draw one v_i that is fixed for that individual in all time periods
- v_i includes all fixed omitted variables.

Comparison of DGP's

- In the first DGP, the unobserved heterogeneity is absorbed into the individual-specific intercept β_{0i}
- This DGP is called the "Distinct Intercepts" DGP.

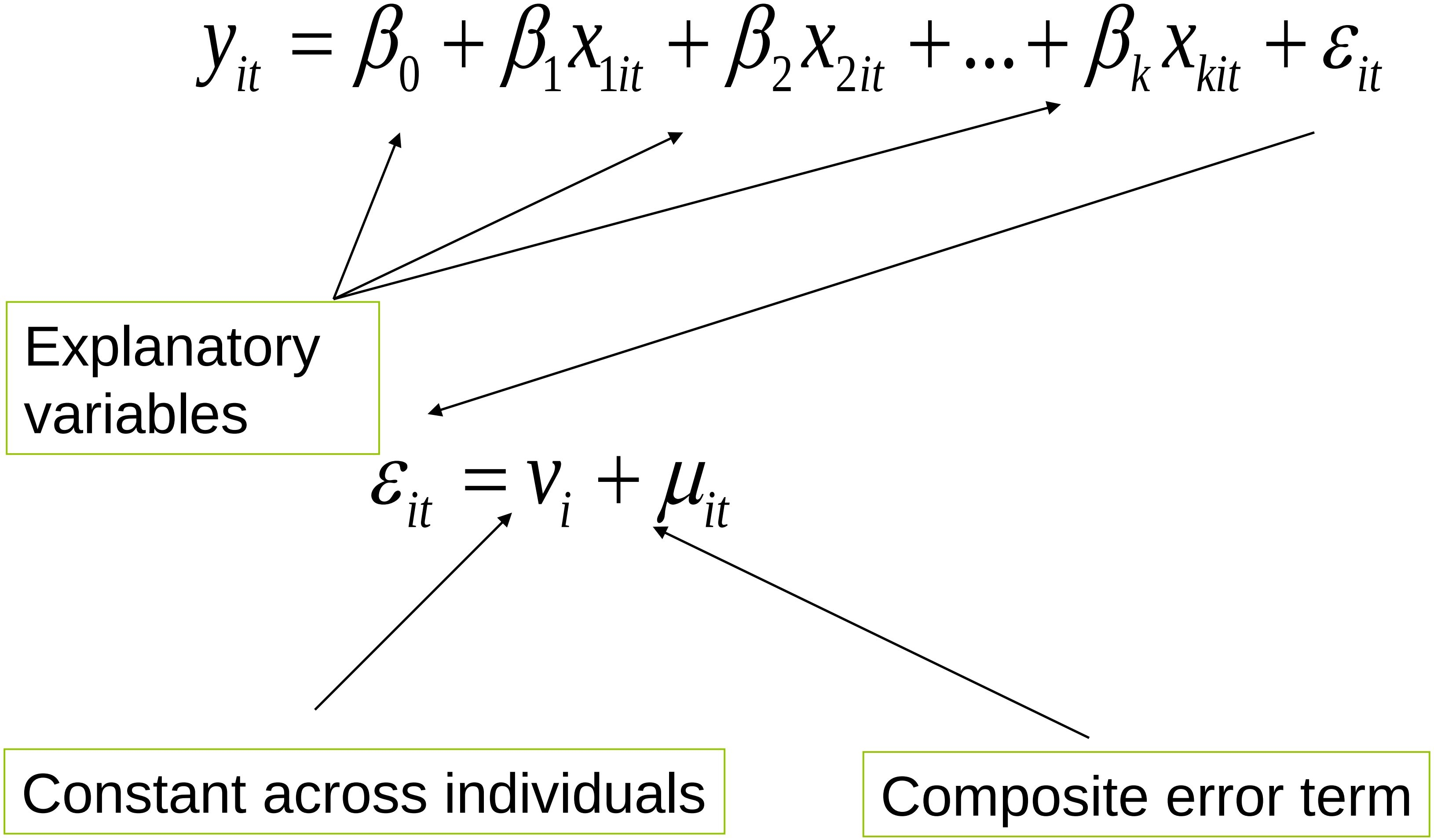
- In the second DGP, the unobserved heterogeneity is absorbed into the individual fixed component of the error term, v_i
- This DGP is an "Error Components Model."

The Error Components DGP

- If $E(X_{jit}v_i) = 0$, then the unobserved heterogeneity is uncorrelated with the explanators.
- OLS is unbiased and consistent.

- If $E(X_{jit}v_i) \neq 0$, then the unobserved heterogeneity IS correlated with the explanators.
- OLS is BIASED and INCONSISTENT.
- Using panel data, we can create a consistent estimator: Fixed Effects.

Develop an error components model



$$y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_k x_{kit} + \varepsilon_{it}$$

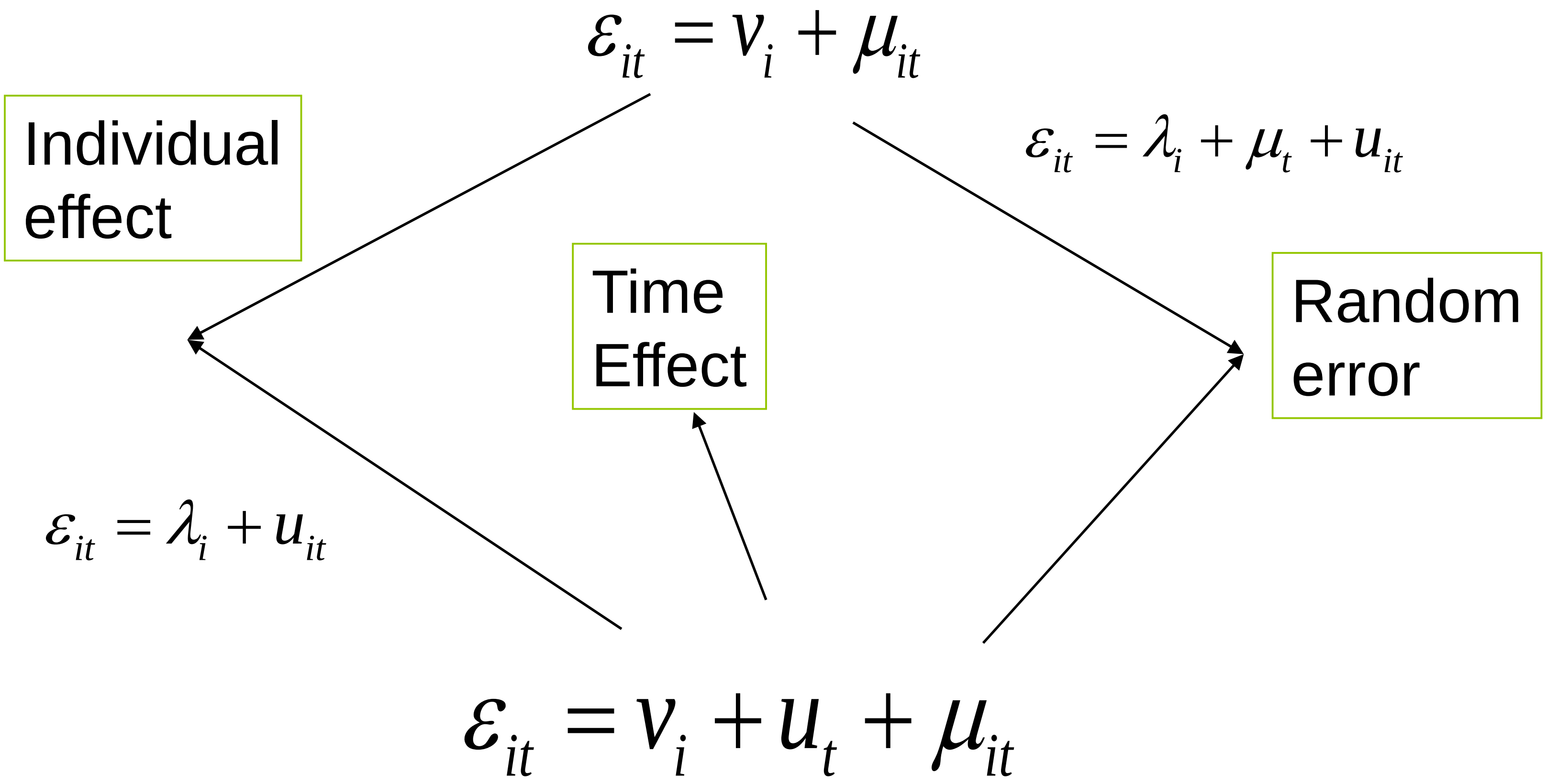
Explanatory variables

$$\varepsilon_{it} = v_i + \mu_{it}$$

Constant across individuals

Composite error term

One-way or two-way error components?



$$\varepsilon_{it} = v_i + \mu_{it}$$

Individual effect

Time Effect

Random error

$$\varepsilon_{it} = \lambda_i + \mu_t + u_{it}$$

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

$$\varepsilon_{it} = v_i + u_t + \mu_{it}$$

Treatment of individual effects

Restrict to one-way model. Then two options for treatment of individual effects:

- *Fixed effects – assume v_i are constants*
- *Random effects – assume v_i are drawn independently from some probability distribution*

The Fixed Effects Model

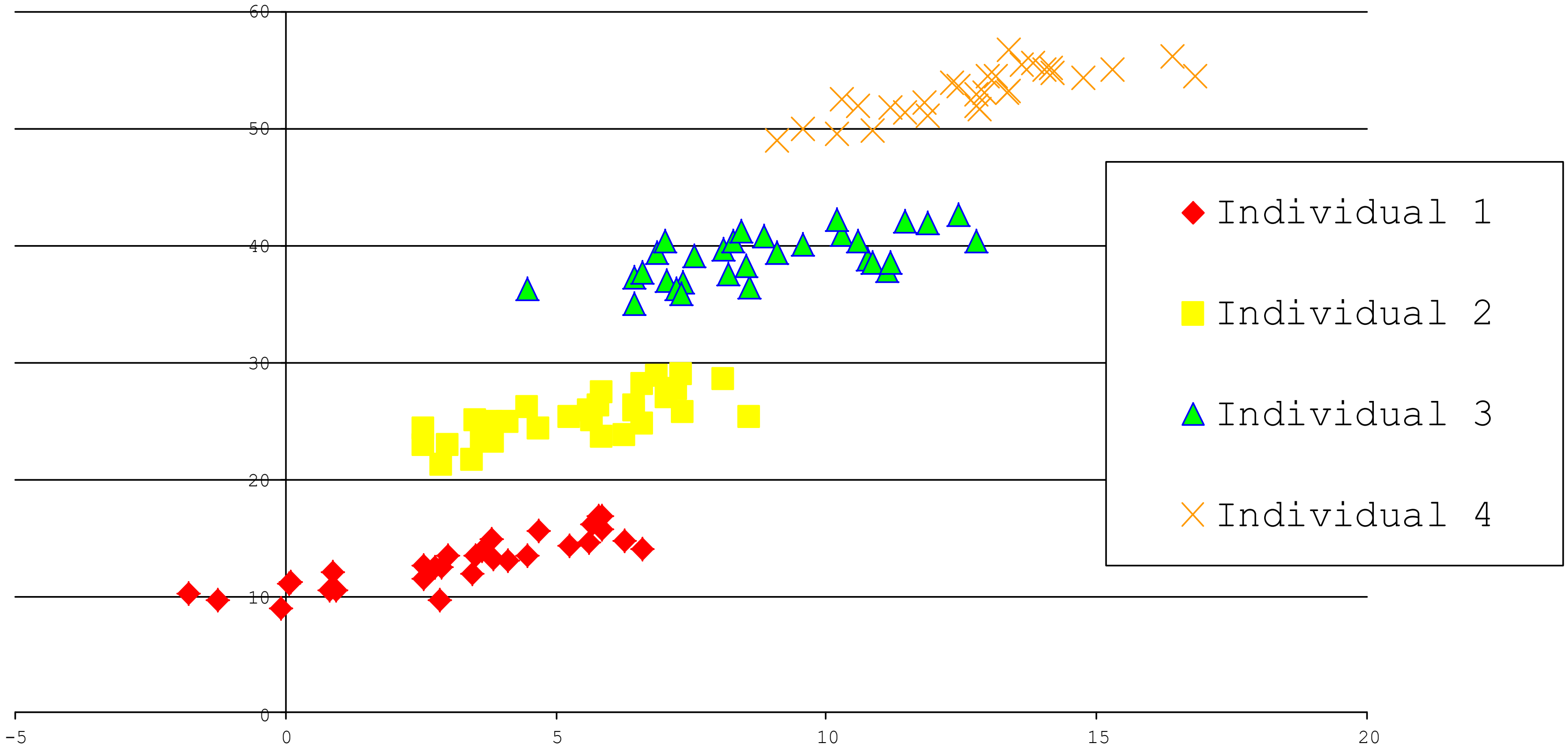
Treat v_i as a constant for each individual

$$y_{it} = \left(\beta_0 + v_i\right) + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_k x_{kit} + \mu_{it}$$

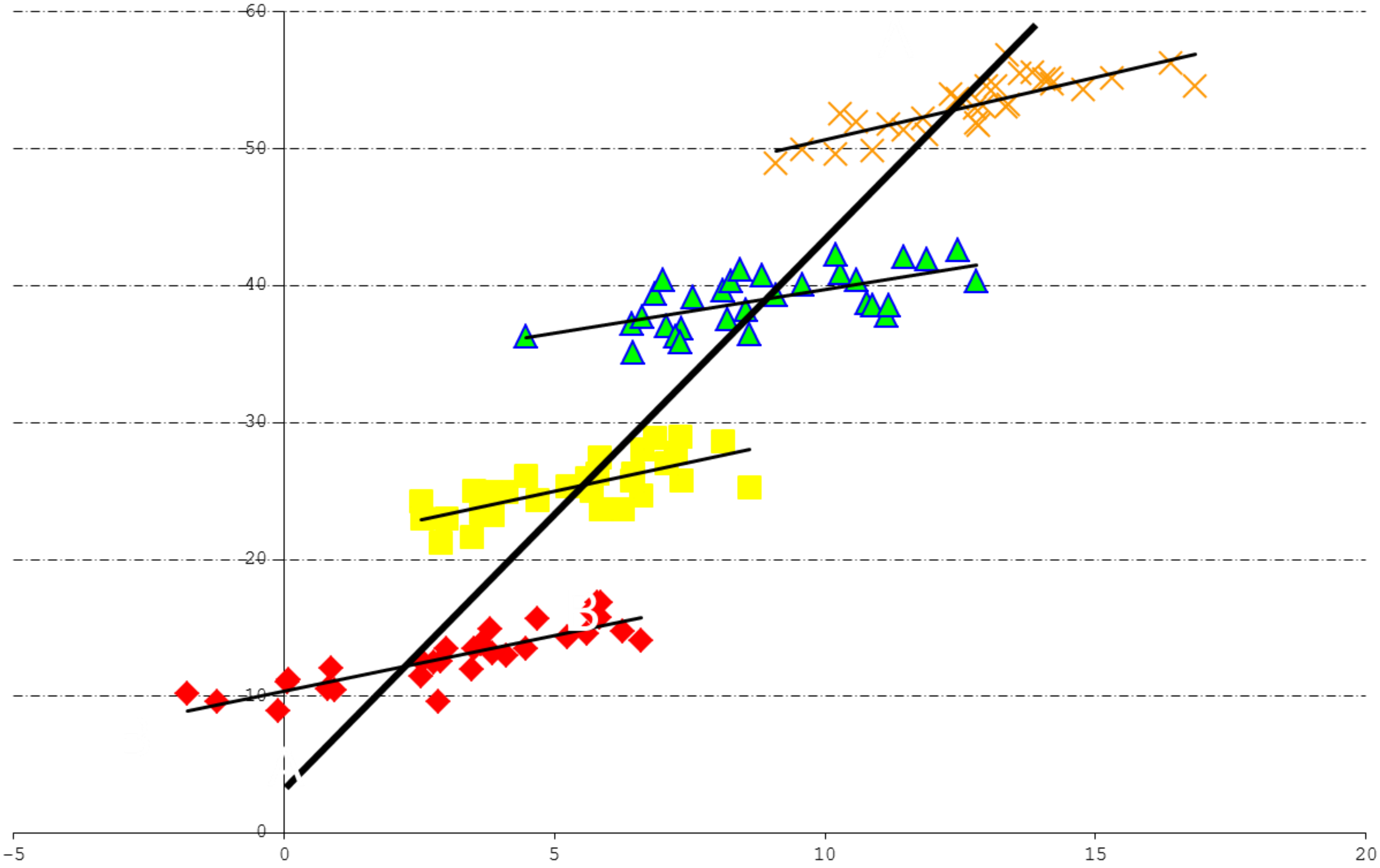
v_i now part of constant – but varies by individual

Graph

Different Constant for Each Individual



Note that the slope is the same for each individual. Only the constant varies



FIXED EFFECTS

Fixed Effects

- The Fixed Effects Estimator used with EITHER the distinct intercepts DGP OR the error components DGP
- ***Basic Idea:*** estimate a separate intercept for each individual with dummy variables (least squares dummy variable estimator - LCDV).

Least Squares Dummy Variable Estimator – 1

- We have already seen that we can use dummy variables to estimate separate intercepts for different groups.
- With panel data we have multiple observations for each individual. We can group these observations.

Least Squares Dummy Variable Estimator – 2

The LSDV estimator is conceptually quite simple:

- Create a set of n dummy variables, D_j, such that $D_j = 1$ if $i = j$, $D_j = 0$ otherwise.
- Regress Y_{it} against all the dummies, X_t , and X_{it} variables (you must omit X_i variables and the constant).

Least Squares Dummy Variable Estimator – 3

In practice the tricky parts are:

- *Creating the dummy variables*
- *Entering the regression into the computer*
- *Reporting results*

Example – 1

- Suppose, we have a longitudinal dataset with 300 workers over 10 years.
- n = 300
- We must create 300 dummy variables and then specify a regression with 300+ explanators.
- How do we do this in our software package?

Example – 2

- Our regression output includes 300 intercepts. Usually, we are not interested in the intercepts themselves.
- In reporting your regression output, it is preferable to note that you have included “individual fixed effects.” Then omit the dummy variable coefficients from your table of results.

Example – 3

- At some point, **n** becomes too large for the computer to handle easily.
- Modern computers can implement LSDV for ever larger data sets, but eventually LSDV becomes computationally intractable.

Solution: Fixed Effects estimator

- The initial insight for the Fixed Effects estimator: if we DIFFERENCE observations for the same individual, the v_i cancels out.

$$\begin{array}{rl} Y_{it} & = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2i} + v_i + \mu_{it} \\ -Y_{it'} & = \beta_0 + \beta_1 X_{1it'} + \beta_2 X_{2i} + v_i + \mu_{it'} \\ \hline \end{array}$$

$$(Y_{it} - Y_{it'}) = 0 + \beta_1 (X_{it} - X_{it'}) + 0 + 0 + \mu_{it} - \mu_{it'}$$

Fixed Effects estimator – 1

- When we difference, the heterogeneity term v_i drops out.
- In the distinct intercepts model, the β_{0i} would drop out.
- OLS would be a consistent estimator of β_1

Fixed Effects estimator – 2

- If T = 2, then we have only 2 observations for each individual.
- Differencing the 2 observations is efficient.
- If T > 2, then differencing any 2 observations ignores valuable information in the other observations for each individual.

Fixed Effects estimator – 3

We can use all the observations for each individual if we subtract the individual-specific mean from each observation.

Fixed Effects estimator – 4

$$\begin{array}{rl} Y_{it} & = \beta_0 + \beta_1 X_{it} + \beta_2 X_i + v_i + \mu_{it} \\ -\bar{Y}_i & = \beta_0 + \beta_1 \bar{X}_i + \beta_2 X_i + v_i + \bar{\mu}_i \\ \hline \end{array}$$

$$(Y_{it} - \bar{Y}_i) = 0 + \beta_1 (X_{it} - \bar{X}_i) + 0 + 0 + \mu_{it} - \bar{\mu}_i$$

where $\bar{Y}_i = \frac{1}{T} \sum_{t=1}^{T} Y_{it}$

Note: $\frac{1}{n} \Sigma v_i = \frac{1}{n} \bullet n \bullet v_i = v_i$

Fixed Effects estimator – 5

Fixed Effects:

1) Construct $y_{it}^{FE} = Y_{it} - \bar{Y}_i$

$$x_{it}^{FE} = X_{it} - \bar{X}_i$$

2) Regress $y_{it}^{FE} = \beta_1 x_{it}^{FE} + \eta_{it}$

Analysis – 1

- The ***Fixed Effects*** (FE) and DVLS estimators provide exactly identical estimates.
- Demeaning each observation by the individual-specific mean eliminates the need to create **n *dummy variables***.
- FE is computationally much simpler.

Analysis – 2

- ***Fixed Effects*** discards all variation between individuals. Fixed Effects uses only variation over time within an individual.
- ***Fixed Effects*** discards a great deal of variation in the explanators (all variation between individuals).
- ***Fixed Effects*** is not efficient if

$$E(X_{it}v_i) = 0$$

Is OLS consistent and efficient?

$$Y_{it} = \beta_0 + \beta_1 X_{it} + v_i + \mu_{it}$$

$E(\mu_{it}) = 0$ $\qquad$ $Var(\mu_{it}) = \sigma_{\mu}^2$

$E(\mu_{it}\mu_{i't'}) = 0$ if $i \neq i'$ OR $t \neq t'$

$E(v_i) = 0$ $\qquad$ $Var(v_i) = \sigma_{v}^2$

$E(v_i v_{i'}) = 0$ for $i \neq i'$ $\qquad$ $E(x_{it} v_i) = 0$ for all i, t

$E(\mu_{it} v_{i'}) = 0$ for all i, i', t $\qquad$ $E(X_{it} \mu_{it}) = 0$ for all i, t

Answer – 1

- Because X is uncorrelated with either *v* or *μ*, OLS is consistent in the uncorrelated version of the error components DGP.
- The error terms are homoskedatic.

$$Var(\varepsilon_{it}) = Var(v_i + \mu_{it}) = \sigma_v^2 + \sigma_\mu^2$$

Answer – 2

However, the covariance between disturbances for a given individual is

$$Cov(\varepsilon_{it}, \varepsilon_{it'}) = E([v_i + \mu_{it}] \bullet [v_i + \mu_{it'}])$$

$$= E(v_i^2) + 2 \bullet E(v_i \mu_{it}) + E(\mu_{it} \mu_{it'})$$

$$= E(v_i^2) = \sigma_v^2$$

$$Corr(\varepsilon_{it}, \varepsilon_{it'}) = \frac{Cov(\varepsilon_{it}, \varepsilon_{it'})}{Var(\varepsilon_{it})} = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\mu^2}$$

Answer – 3

In the presence of serial correlation, OLS is inefficient.

Fixed Effects (GLS Estimation)

- The fixed effects estimator can also be written in GLS form, which brings out its relationship to the RE estimator.

$$\hat{\beta}_{FE} = \left[\sum_{\mathrm{i}=1}^{\mathrm{T}} (X_i^{'} M X_i) \right]^{-1} \sum_{i=1}^{T} X_i^{'} M y_i \text{ where } M = I_T - \frac{1}{T} ee'$$

- The FE estimator uses *M* as the weighting matrix rather than *Ω*.

RANDOM EFFECTS

Random Effects – 1

- When unobserved heterogeneity is uncorrelated with explanators, panel data techniques are not needed to produce a consistent estimator.
- However, we do need to correct for serial correlation between observations of the same individual.

The Random Effects Model

Original equation

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_k x_{kit} + \varepsilon_{it}$$

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_k x_{kit} + v_i + \mu_{it}$$

Remember $\varepsilon_{it} = v_i + \mu_{it}$

v_i now part of error term

- This approach might be appropriate if observations are representative of a sample rather than the whole population.

Random Effects – 2

- When $E(X_{it}v_i) \neq 0$, panel data provides a valuable tool for eliminating omitted variables bias. We use Fixed Effects to gain the benefits of panel data.

- When $E(X_{it}v_i) = 0$, panel data does not offer special benefits. We use Random Effects to overcome the serial correlation of panel data.

Random Effects – 3

The key idea of random effects:

- *Estimate s_v^2 and s_m^2*
- *Use these estimates to construct efficient weights of panel data observations*

Random Effects – 5

1) Estimate the regression using Fixed Effects.

2) Construct Fixed Effects residuals, $\tilde{u}_{it}$

3) Estimate σ_{μ}^{2} :

$$s_{\mu}^{2} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\tilde{u}_{it} - \frac{1}{T}\sum_{\tau=1}^{T}\tilde{u}_{i\tau}\right)}{n(T-k-1)}$$

4) Estimate the regression using OLS

5) Estimate s^{2} as usual

6) Because $\sigma^{2} = \sigma_{\mu}^{2} + \sigma_{v}^{2}$:

$$s_{v}^{2} = s^{2} - s_{\mu}^{2}$$

Random Effects - 6

- Once we have estimates of $\sigma_v{}^2$ and $\sigma_\mu{}^2$, we can re-weight the observations optimally.
- These calculations are complicated, but most computer packages can implement them.

Random Effects (GLS Estimation)

- The Random Effects estimator has the standard generalised least squares form summed over all individuals in the dataset:

$$\hat{\beta}_{RE} = \left[\sum_{i=1}^{N} (X_i^{'} \Omega^{-1} X_i) \right]^{-1} \sum_{i=1}^{N} X_i^{'} \Omega^{-1} y_i$$

where, given Ω from the previous slide, it can be shown that:

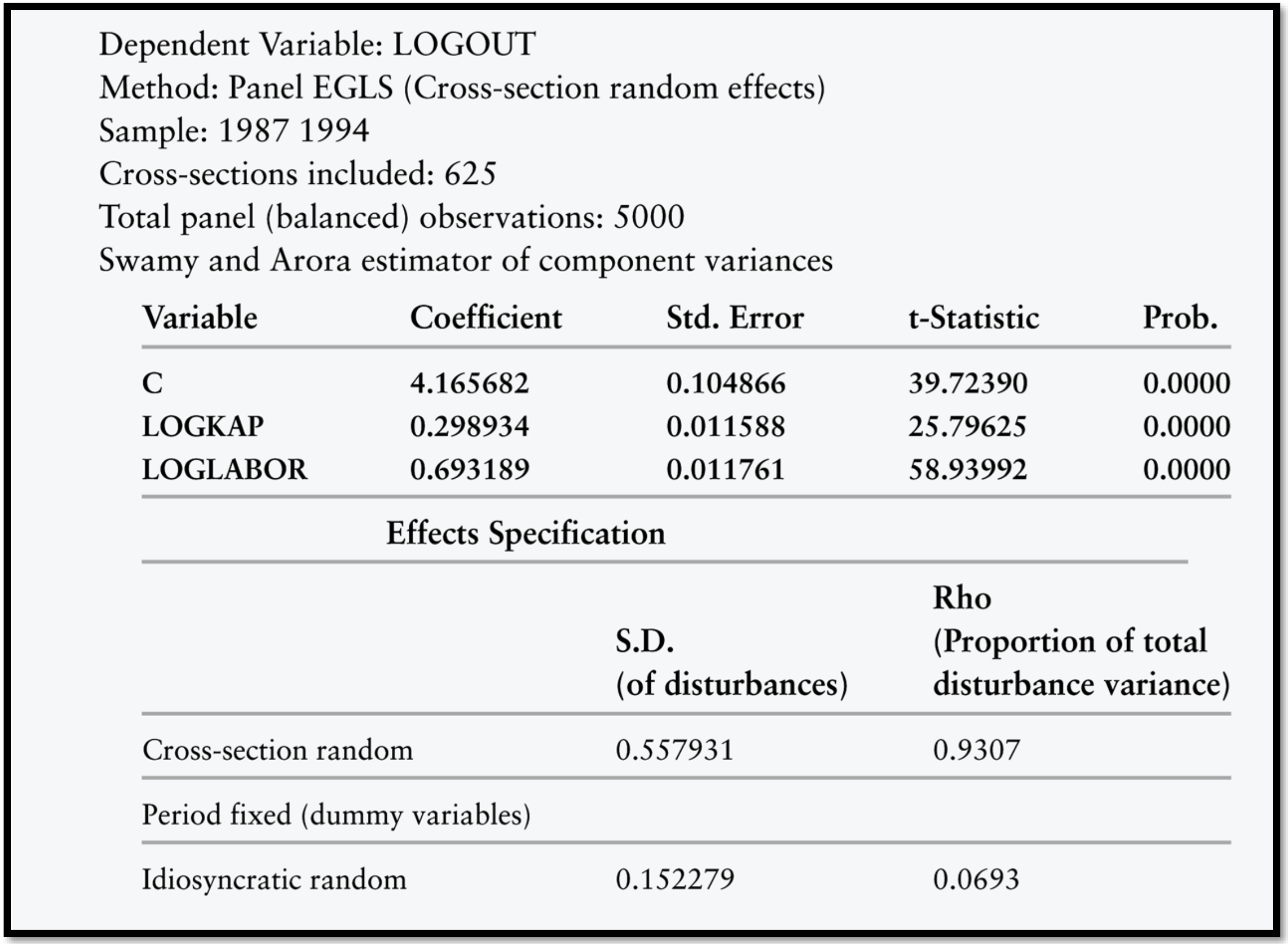
$$\Omega^{-1/2} = \frac{1}{\sigma_u} \left(I_T - \frac{\theta}{T} ee' \right), \; \theta = 1 - \frac{\sigma_u}{\sqrt{T\sigma_\lambda^2 + \sigma_u^2}}$$

Example: Cobb–Douglas production function – 1

- We have data from 625 French firms from 16 countries for 8 years.
- We wish to estimate a Cobb–Douglas production function:
- Taking logs: $Q_i = \beta_0 L_i^{\beta 1} K_i^{\beta 2} \varepsilon_i$
- We estimate using random effects.

$$\ln(Q_i) = \ln(\beta_0) + \beta_1 \ln(L_i) + \beta_2 \ln(K_i) + \ln(\varepsilon_i)$$

Example: Cobb–Douglas production function – 2



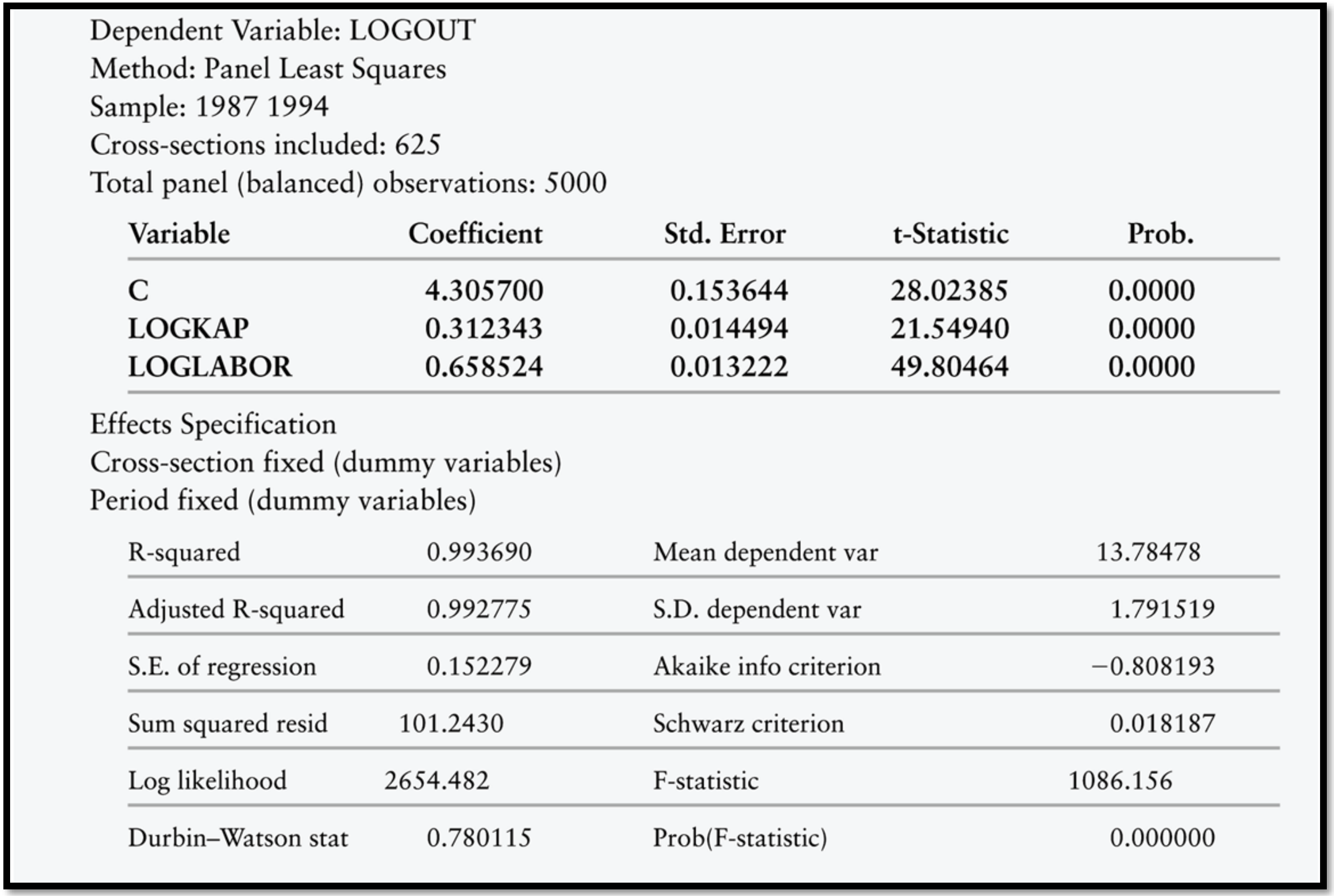
Dependent Variable: LOGOUT
Method: Panel EGLS (Cross-section random effects)
Sample: 1987 1994
Cross-sections included: 625
Total panel (balanced) observations: 5000
Swamy and Arora estimator of component variances

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.165682	0.104866	39.72390	0.0000
LOGKAP	0.298934	0.011588	25.79625	0.0000
LOGLABOR	0.693189	0.011761	58.93992	0.0000

Effects Specification

	S.D. (of disturbances)	Rho (Proportion of total disturbance variance)
Cross-section random	0.557931	0.9307
Period fixed (dummy variables)		
Idiosyncratic random	0.152279	0.0693

Example: Cobb–Douglas production function – 2



```
Dependent Variable: LOGOUT
Method: Panel Least Squares
Sample: 1987 1994
Cross-sections included: 625
Total panel (balanced) observations: 5000

    Variable           Coefficient      Std. Error       t-Statistic      Prob.
    ---------------------------------------------------------------------------
    C                    4.305700        0.153644         28.02385       0.0000
    LOGKAP               0.312343        0.014494         21.54940       0.0000
    LOGLABOR             0.658524        0.013222         49.80464       0.0000
    ---------------------------------------------------------------------------
Effects Specification
Cross-section fixed (dummy variables)
Period fixed (dummy variables)

    R-squared            0.993690      Mean dependent var          13.78478
    ---------------------------------------------------------------------------
    Adjusted R-squared   0.992775      S.D. dependent var           1.791519
    ---------------------------------------------------------------------------
    S.E. of regression   0.152279      Akaike info criterion       −0.808193
    ---------------------------------------------------------------------------
    Sum squared resid   101.2430       Schwarz criterion            0.018187
    ---------------------------------------------------------------------------
    Log likelihood     2654.482        F-statistic               1086.156
    ---------------------------------------------------------------------------
    Durbin–Watson stat   0.780115      Prob(F-statistic)            0.000000
```

Example: Cobb–Douglas production function – 3

- We arrive at similar estimates using either random effects or fixed effects.
- Because only fixed effects controls for unobserved heterogeneity that is correlated with the explanators, the similarity between the two estimates suggests that unobserved heterogeneity is not creating a large bias in this sample.

Example: Cobb–Douglas production function – 4

- The fixed effects estimator discards all variation between firms, and must use 624 more degrees of freedom than random effects.
- The RE estimator provides more precise estimates
  - Moving from RE to FE increases the s.e. on capital from 0.0116 to 0.0145
  - The s.e. on labor moves from 0.0118 to 0.0132

Example: Cobb–Douglas production function – 5

- We would prefer to use RE instead of FE, but RE might be inconsistent if $E(X_{it}v_i) \neq 0$
- We need a test to help determine whether it is safe to use RE.

For and against random effects:

- Random effects are efficient
- Why should we assume one set of unobservables fixed and the other random?
- Sample information more common than that from the entire population?
- Can deal with regressors that are fixed across individuals

- Likely to be correlation between the unobserved effects and the explanatory variables. These are assumed to be zero in the random effects model, but in many cases we might expect them to be non-zero. This implies inconsistency due to omitted-variables in the RE model. In this situation, fixed effects is inefficient, but still consistent.

THE HAUSMAN TEST

The Hausman Test – 1

- Hausman's specification test for error components DGPs provides guidance on whether $E(X_{it}v_i) \neq 0$
- The key idea: if $E(X_{it}v_i) \neq 0$, then the inconsistent RE estimator and the consistent FE estimator converge to different estimates.

The Hausman Test – 2

- If $E(X_{it}v_i) = 0$, then the unobserved heterogeneity is uncorrelated with X and does not create a bias.
- RE and FE are both consistent.
- For two consistent estimators to provide significantly different estimates would be surprising.

The Hausman Test – 3

- We know the FE estimator is consistent even when $E(X_{it}v_i) \neq 0$
- The problem with FE is its inefficiency.
- FE is not as precise as RE.
- Although FE is imprecise, it may provide a good enough estimate to detect a large bias in RE.

The Hausman Test – 4

- If FE is very imprecise, then the Hausman test has very weak power and cannot rule out even large biases.
- If FE is very precise, then the Hausman test has very good power, but we gain little benefit from switching to the more efficient RE.

The Hausman Test – 5

- If FE is somewhat precise, then the Hausman test can warn us away from using RE in the presence of a large bias, but there is still room for substantial efficiency gains in switching to RE.

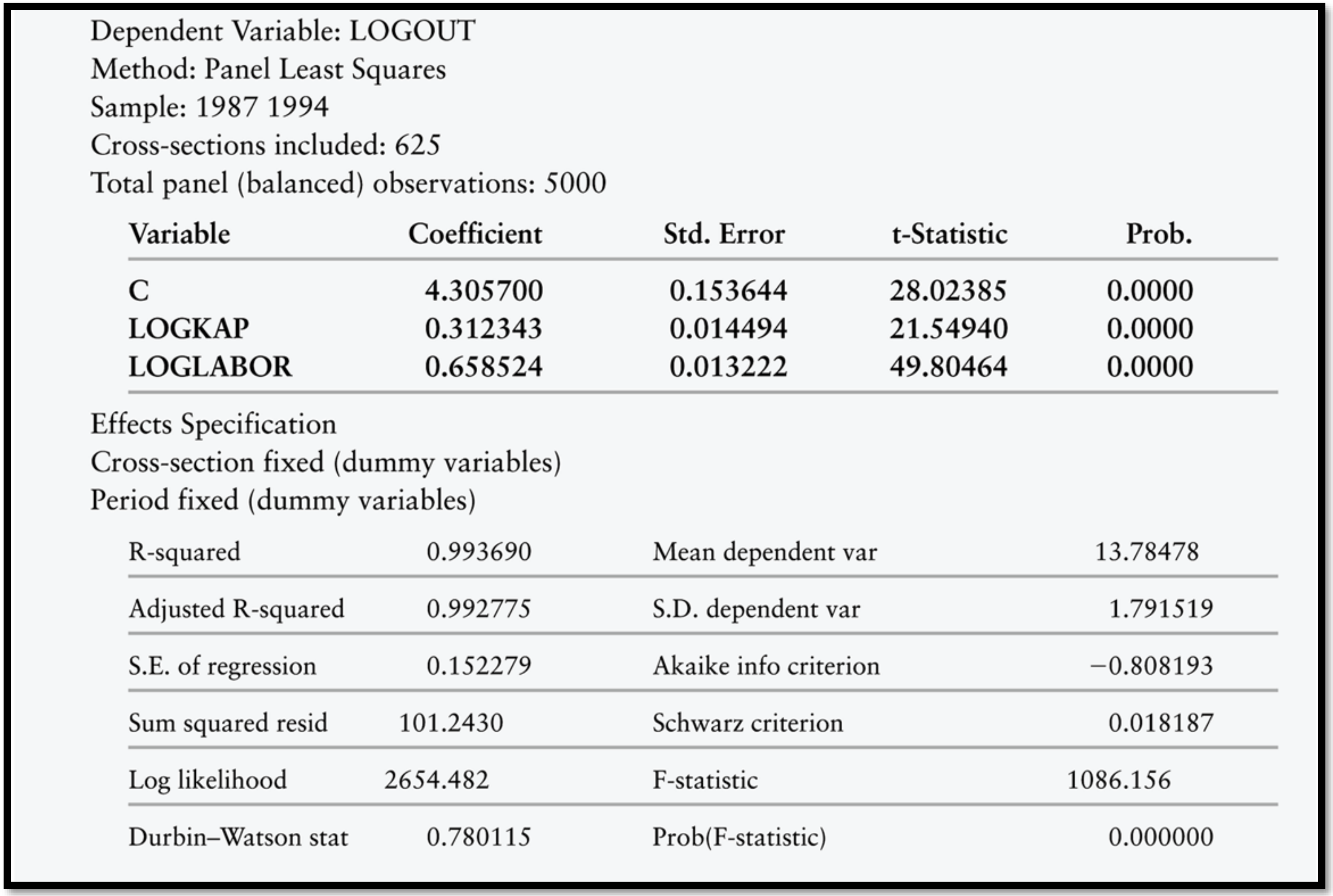
The Hausman Test: Calcus

- A test for the independence of the v_i and the x_{kit}. The covariance of an efficient estimator with its difference from an inefficient estimator should be zero. Thus, under the null hypothesis we test:

$$\mathrm{W}=(\beta_{\mathrm{RE}}-\beta_{FE})'\hat{\Sigma}^{-1}(\beta_{\mathrm{RE}}-\beta_{FE}) \sim \chi^2(k)$$

- If *W* is significant, we should not use the random effects estimator.

Example: Cobb–Douglas production function with fixed effects



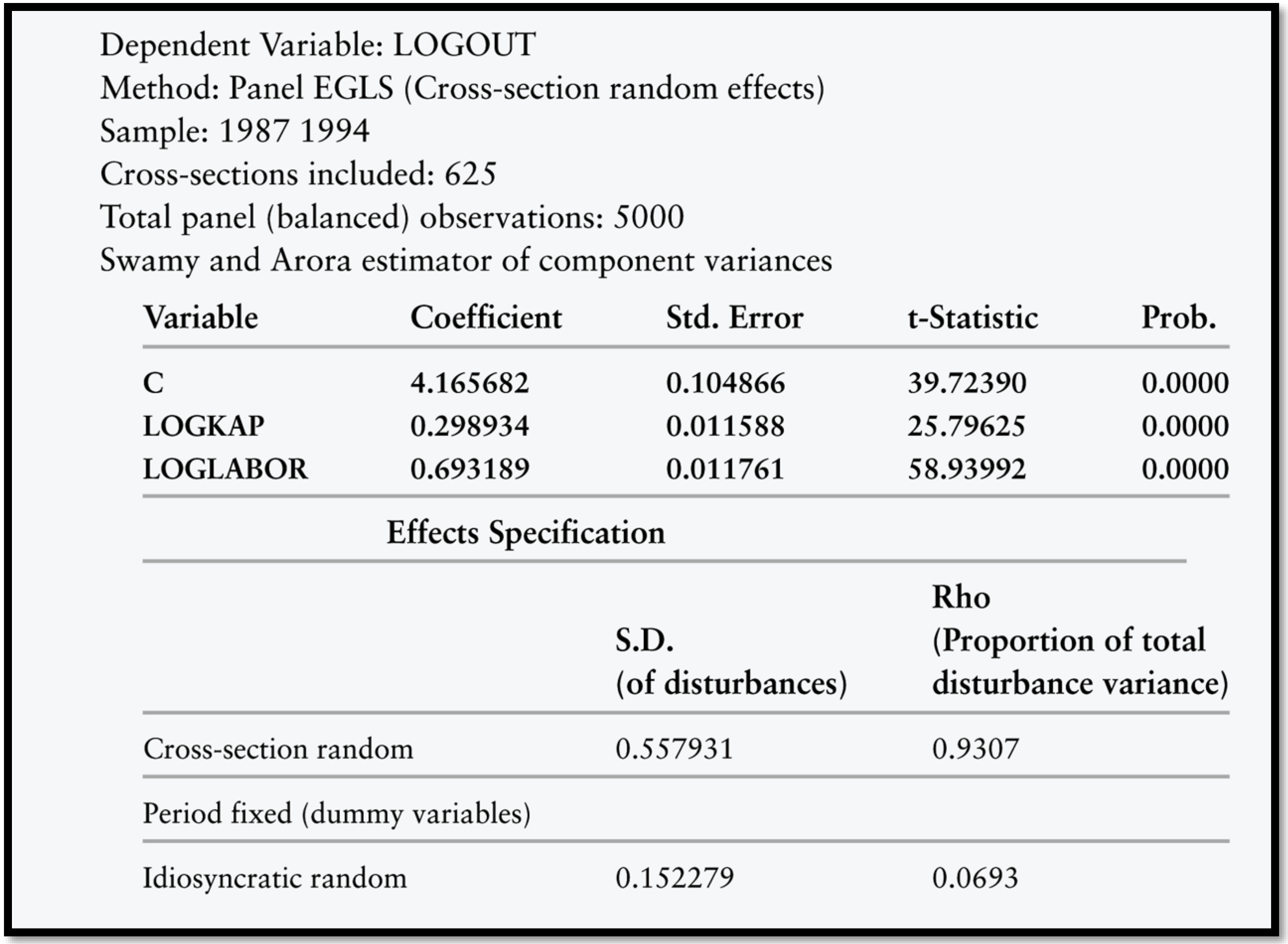
Dependent Variable: LOGOUT
Method: Panel Least Squares
Sample: 1987 1994
Cross-sections included: 625
Total panel (balanced) observations: 5000

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.305700	0.153644	28.02385	0.0000
LOGKAP	0.312343	0.014494	21.54940	0.0000
LOGLABOR	0.658524	0.013222	49.80464	0.0000

Effects Specification
Cross-section fixed (dummy variables)
Period fixed (dummy variables)

R-squared	0.993690	Mean dependent var	13.78478
Adjusted R-squared	0.992775	S.D. dependent var	1.791519
S.E. of regression	0.152279	Akaike info criterion	−0.808193
Sum squared resid	101.2430	Schwarz criterion	0.018187
Log likelihood	2654.482	F-statistic	1086.156
Durbin–Watson stat	0.780115	Prob(F-statistic)	0.000000

Example: Cobb–Douglas production function with random effects



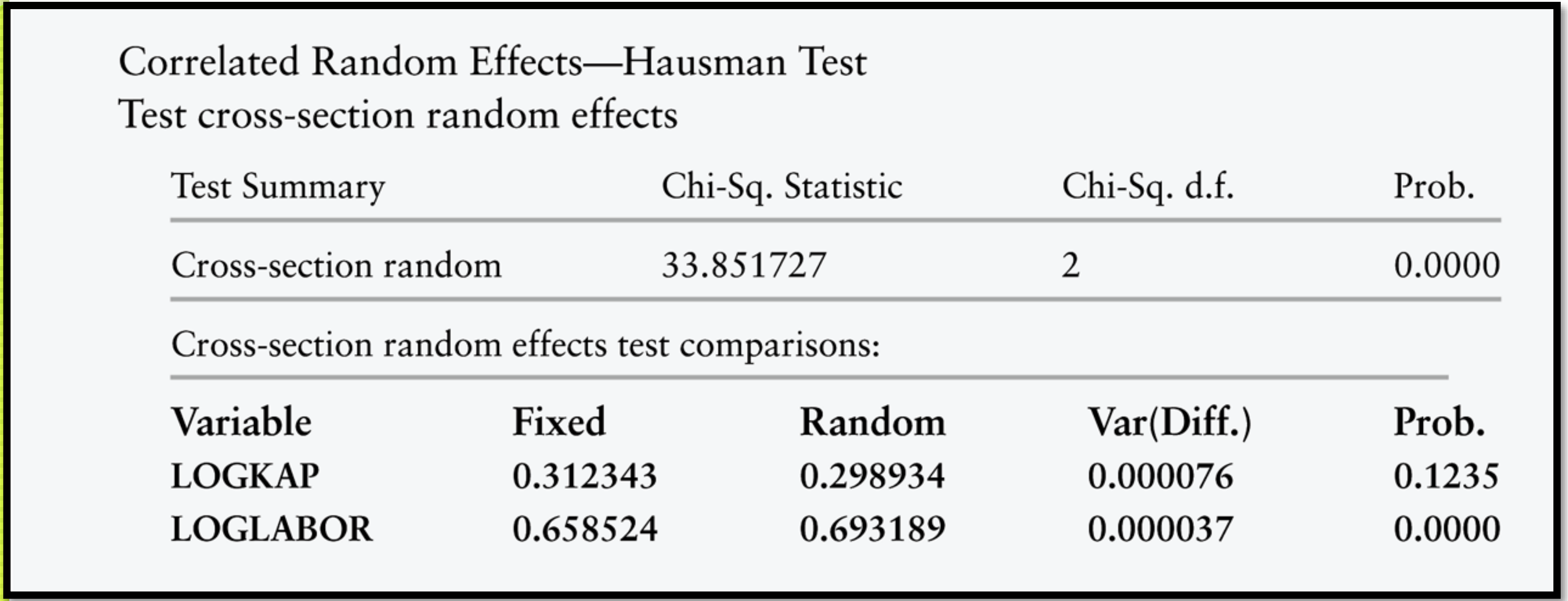
Dependent Variable: LOGOUT
Method: Panel EGLS (Cross-section random effects)
Sample: 1987 1994
Cross-sections included: 625
Total panel (balanced) observations: 5000
Swamy and Arora estimator of component variances

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.165682	0.104866	39.72390	0.0000
LOGKAP	0.298934	0.011588	25.79625	0.0000
LOGLABOR	0.693189	0.011761	58.93992	0.0000

Effects Specification

	S.D. (of disturbances)	Rho (Proportion of total disturbance variance)
Cross-section random	0.557931	0.9307
Period fixed (dummy variables)		
Idiosyncratic random	0.152279	0.0693

Example: The Hausman Test



Correlated Random Effects—Hausman Test
Test cross-section random effects

Test Summary	Chi-Sq. Statistic	Chi-Sq. d.f.	Prob.
Cross-section random	33.851727	2	0.0000

Cross-section random effects test comparisons:

Variable	**Fixed**	**Random**	**Var(Diff.)**	**Prob.**
LOGKAP	0.312343	0.298934	0.000076	0.1235
LOGLABOR	0.658524	0.693189	0.000037	0.0000

Example: Analysis

With the French manufacturing firms, FE is precise enough to reject the null even though the two estimates are fairly close.

The Hausman Test: notes – 1

- Fixed effects exacerbates measurement error bias.
- There is likely to be less variation in X within the experience of a single individual than across several individuals.
- Small measurement errors can become large relative to the within-variation in X.

The Hausman Test: notes – 2

- The Hausman Test warns us that RE and FE provide significantly different estimates.
- This difference could arise because of omitted variables bias in RE, caused by $E(X_{it}v_i) \neq 0$
- This difference could ALSO arise because of measurement error biases in FE.

REVIEW

Problem

- Potential unobserved heterogeneity is a form of omitted variables bias.
- "***Unobserved heterogeneity***" refers to omitted variables that are fixed for an individual (at least over a long period of time).
- A person's upbringing, family characteristics, innate ability, and demographics (except age) do not change.

Data

- Panel Data is data in which we observe repeated cross-sections of the same individuals.
- The key feature of panel data is that we observe the same individual in more than one condition.
- Omitted variables that are fixed will take on the same values each time we observe the same individual.

3 different DGP's for panel data – 1

- In the distinct intercept DGP, across samples we would observe the same individuals with the same unobserved heterogeneity.
- Each *i* has its own intercept, β_{0i}, that is fixed across samples.

Model #1

$$Y_{it} = \beta_{0i} + \beta_1 X_{1it} + .. + \beta_K X_{Kit} + \varepsilon_{it}$$

$$i = 1...n; \quad t = 1...T$$

$$E(\varepsilon_{it}) = 0$$

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

$E(\varepsilon_{it}\varepsilon_{i't'}) = 0$ if $i \neq i'$ OR $t \neq t'$

$E(X_{jit}\varepsilon_{it}) = 0$ for all j, i, t

3 different DGP's for panel data – 2

- Error components DGP's are suitable when we would draw different individuals across samples.
- When each *i* is drawn, its unobserved heterogeneity is captured in a v_i term.
- We learned two error components DGP, depending on whether the v_i is correlated with the X_{kit} 's.

Models #2 and #3

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2i} + \beta_3 X_{3t} + .. + \beta_K X_{Kit} + v_i + \mu_{it}$$

$$i = 1...n; \quad t = 1...T$$

$E(\mu_{it}) = 0$ $\qquad$ $Var(\mu_{it}) = \sigma_\mu^2$

$E(\mu_{it}\mu_{i't'}) = 0$ if $i \neq i'$ OR $t \neq t'$ $\qquad$ $E(v_i) = 0$

$E(v_i v_{i'}) = 0$ for $i \neq i'$ $\qquad$ $Var(v_i) = \sigma_v^2$

$E(\mu_{it} v_{i'}) = 0$ for all i, i', t

$E(x_{jit}\mu_{it}) = 0$ for all j, i, t

EITHER $E(X_{jit} v_i) = 0$ for all j, i, t

OR $E(X_{jit} v_i) \neq 0$ for at least some j, i, t

Problem and solution – 1

- If $E(X_{it}v_i) \neq 0$, OLS would be inconsistent
- By estimating a separate intercept for each individual, we can control for the v_i
- We learned two equivalent strategies: DVLS and FE.

Problem and solution – 2

- The simplest way to estimate separate intercepts for each individual is to use dummy variables (least squares dummy variable estimator).

- Fixed effects estimator:
  - Construct

$$y_{it}^{FE} = Y_{it} - \bar{Y}_i$$

$$x_{it}^{FE} = X_{it} - \bar{X}_i$$

  - Regress

$$y_{it}^{FE} = \beta_1 x_{it}^{FE} + \eta_{it}$$

Problem and solution – 3

- Fixed Effects (however estimated) discards all variation between individuals.
- Fixed Effects uses only variation over time within an individual.
- Because X is uncorrelated with either v or μ, OLS is consistent in the uncorrelated version of the error components DGP.

Remember!

- When unobserved heterogeneity is uncorrelated with explanators, panel data techniques are not needed to produce a consistent estimator.
- However, we do need to correct for serial correlation between observations of the same individual.

Fixed vs Random effects

- When $E(X_{it}v_i) \neq 0$, panel data provides a valuable tool for eliminating omitted variables bias.
- We use Fixed Effects to gain the benefits of panel data.

- When $E(,X_{it}v_i) = 0$ panel data is less convenient than an equal-sized cross-sectional data set.
- We use Random Effects to overcome the serial correlation of panel data.

The Hausman Test

- Hausman’s specification test for error components DGPs provides guidance on whether $E(X_{it}v_i)\neq 0$
- The key idea: if $E(X_{it}v_i)\neq 0$, then the inconsistent RE estimator and the consistent FE estimator converge to different estimates.

QUESTIONS?

THANK YOU FOR YOUR ATTENTION!