The Panel Data Regression Function

FORDHAM
THE JESUIT UNIVERSITY OF NEW YORK
Gabelli School of Business

Panel Data Metrics, Part I

A panel of data consists of a group of cross-sectional units (people, households, firms, states, countries) that are observed over time

- Denote the number of cross-sectional units (individuals) by *N*
- Denote the number of time periods in which we observe them as *T*

Panel Data Metrics, Part II

Different ways of describing panel data sets

- Long and narrow
  - “Long” describes the time dimension, and “narrow” implies a relatively small number of cross sectional units
- Short and wide
  - There are many individuals observed over a relatively short period of time
- Long and wide
  - Both *N* and *T* are relatively large

Panel Data Metrics, Part III

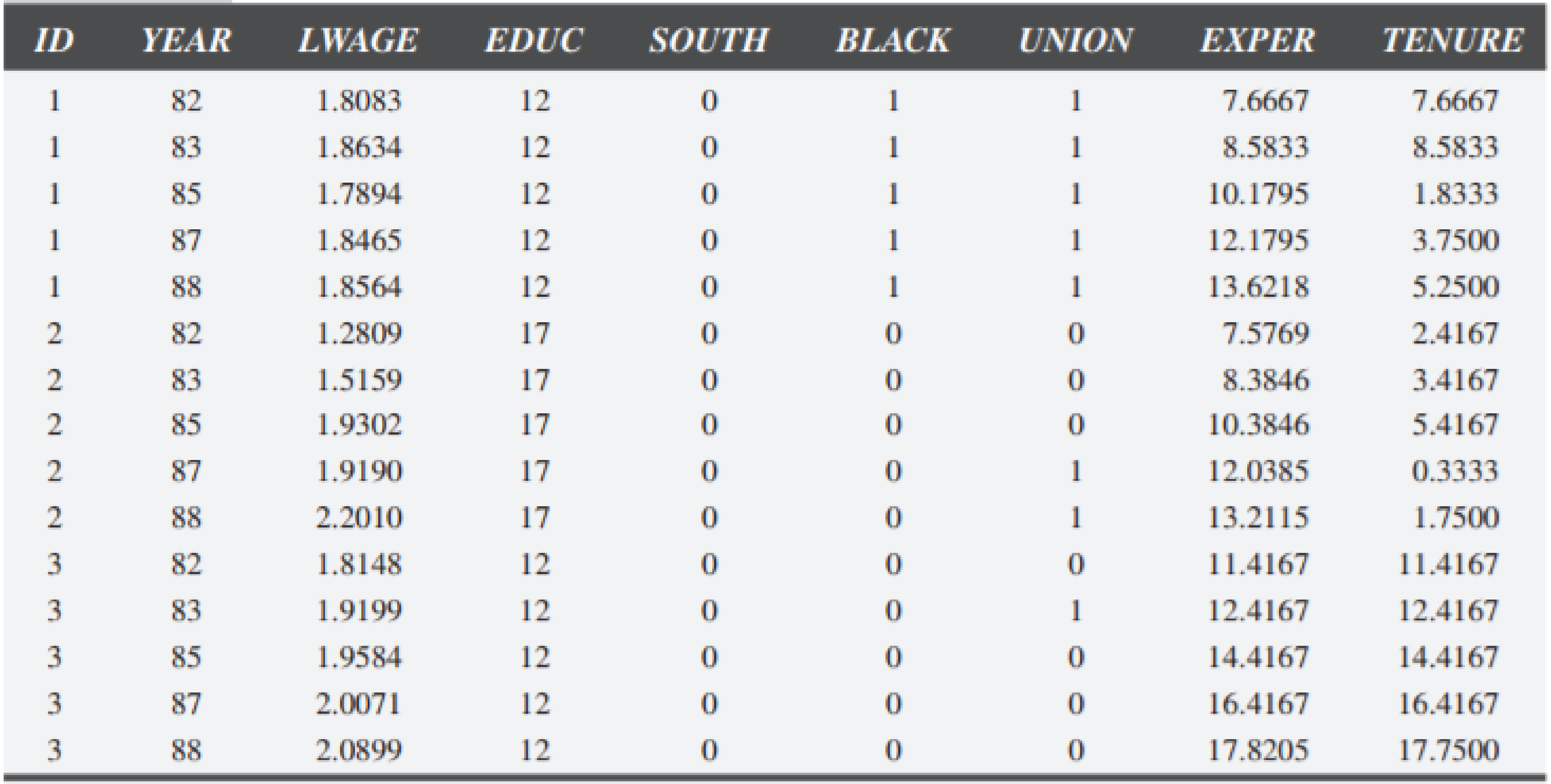
It is possible to have data that combine cross-sectional and time-series data that do not constitute a panel

- We may collect a sample of data on individuals from a population at several points in time, but the individuals are not the same in each time period
  - Such data can be used to analyze a “natural experiment”

Example 1: A Microeconometric Panel

- In microeconomic panels, the individuals are not always interviewed the same number of times, leading to an **unbalanced panel** in which the number of time-series observations is different across individuals.
- In a **balanced panel**, each individual has the same number of observations.
- The data file on the next slide is, however, a balanced panel.

TABLE 15.1 Representative Observations from NLS Panel Data



ID	*YEAR*	*LWAGE*	*EDUC*	*SOUTH*	*BLACK*	*UNION*	*EXPER*	*TENURE*
1	82	1.8083	12	0	1	1	7.6667	7.6667
1	83	1.8634	12	0	1	1	8.5833	8.5833
1	85	1.7894	12	0	1	1	10.1795	1.8333
1	87	1.8465	12	0	1	1	12.1795	3.7500
1	88	1.8564	12	0	1	1	13.6218	5.2500
2	82	1.2809	17	0	0	0	7.5769	2.4167
2	83	1.5159	17	0	0	0	8.3846	3.4167
2	85	1.9302	17	0	0	0	10.3846	5.4167
2	87	1.9190	17	0	0	1	12.0385	0.3333
2	88	2.2010	17	0	0	1	13.2115	1.7500
3	82	1.8148	12	0	0	0	11.4167	11.4167
3	83	1.9199	12	0	0	1	12.4167	12.4167
3	85	1.9584	12	0	0	0	14.4167	14.4167
3	87	2.0071	12	0	0	0	16.4167	16.4167
3	88	2.0899	12	0	0	0	17.8205	17.7500

The Panel Data Regression Function, Part I

- A panel of data consists of a group of cross-sectional units (people, households, firms, states or countries) that are observed over time
- Let w_{1i}, w_{2i}, …, w_{Mi} be observed data on M factors that do not change over time
  - Note that these variables **do not** have a time subscript and are said to be **time-invariant**
- In addition to the observed variables, there will be unobserved, omitted factors in each time period for each individual that will compose the regression’s random error term

The Panel Data Regression Function, Part II

- Consider unobserved and/or unmeasurable, time-invariant individual characteristics; these are denoted as $u_{1i}, u_{2i}, \ldots, u_{Si}$
- Economists say that u_i represents **unobserved heterogeneity**, summarizing the unobserved factors leading to individual differences
- Econometricians call the random error e_{it} that varies across individual and time, an idiosyncratic error
- A third type of random error is time specific, an effect that varies over time but not individual

The Panel Data Regression Function, Part III

- A simple but representative panel data regression model is:
  - (15.1) $y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + (u_i + e_{it})$ $= \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + v_{it}$
- We define the combined:
  - (15.2) $v_{it} = u_i + e_{it}$
- Because the regression error in equation 15.2 has two components, one for the individual and one for the regression, it is often called an error components model

The Panel Data Regression Function, Part IV

- The complicating factor in panel data modeling is that we observe each cross-sectional unit, individual *i*, for more than one time period, *t*
- If individuals are randomly sampled, then observations on the *i*th individual are statistically independent of observations on the *j*th individual; however, using panel data, we must consider dynamic, time-related effects, and model assumptions should take them into account

The Panel Data Regression Function, Part V

- The regression function of interest in a panel data model is: (15.3)

$$E\left[y_{it} \mid \overbrace{x_{2i1,} x_{2i2,} x_{2it,}}^{T\,terms} w_{li}, u_i \right] = E(y_{it} \mid x_{2i,} w_{li,} u_i) = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{li} + u_i$$

- Equation 15.3 says that the population average value of the outcome variable is $\beta_1 + \beta_2 x_{2it} + \alpha_1 w_{li} + u_i$, given (i) the values of x_{2it} in all time periods, past, present, and future; (ii) the observable individual-specific variable w1i; and (iii) the unobservable individual heterogeneity term u_i

The Panel Data Regression Function, Part VI

- Equation 15.3 has interesting features
  1. The model states that, once we have controlled for x_{2it} in all time periods, and the individual-specific factors w_{1i} and u_i, only the current, contemporaneous value of x_{2it} has an effect on the expected outcome
  2. The model conditions on the unobservable time-invariant error u_i

Further Discussion of Unobserved Heterogeneity

- Unobservable individual differences are called unobservable heterogeneity in the economics and econometrics literature.
- When using panel data, it is important to separate this component of the random error term from other components if we can argue that the factors causing the individual differences are unchanging over time.
- The beauty of having panel data is that we can control for the omitted variables bias, caused by time-invariant omitted variables.

The Panel Data Regression Exogeneity Assumption

- A new exogeneity assumption takes into account the presence of the unobserved heterogeneity term:
  - (15.4) $E(e_{it}|x_{2i}, w_{1i}, u_i) = 0$
- The meaning of this strict exogeneity assumption is that, given the values of the explanatory variable x_{2i} in all time periods, given w_{1i}, and given the unobserved heterogeneity term u_i, the best prediction of the idiosyncratic errors is zero

The Panel Data Regression Exogeneity Assumption (cont.)

- Two of the implications of assumption (equation 15.4) are:
  - (15.5a) $\operatorname{cov}(e_{it}, x_{2is}) = 0, and \operatorname{cov}(e_{it}, w_{1i}) = 0$
- The first part, $\operatorname{cov}(e_{it}, x_{2is}) = 0$, is much stronger than the usual sort of exogeneity assumption
  - (15.5b) $\operatorname{cov}(e_{it}, x_{1is} = 1) = E(e_{it}, x_{1i}) = E(e_{it}) = 0$
- Thus, the expected value of the idiosyncratic error is zero

Using OLS To Estimate the Panel Data Regression

- We require:
  - (15.6a) $E(x_{2it}, e_{it}) = 0, \qquad E(w_{1i}, e_{it}) = 0$
  - (15.6b) $E(x_{2it}, u_i) = 0, \qquad E(w_{1i}, u_i) = 0$
  - (15.6c) $E(e_{it}) = E(u_i) = E(v_{it}) = 0$
- Each of the random errors has mean zero; finally, even if equations 15.6 a–c hold, using the OLS estimator will require using a type of robust standard error

FORDHAM

THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School
of Business

The Fixed Effects Estimator

FORDHAM THE JESUIT UNIVERSITY OF NEW YORK | Gabelli School of Business

The Fixed Effects Estimator

- We consider estimation procedures that employ a transformation to eliminate the individual heterogeneity from the estimation equation and thus solve the common **endogeneity** problem
- The estimators we will consider are:
  1. The difference estimator
  2. The within estimator
  3. The fixed effects estimator

The Difference Estimator: T = 2

- When we observe each individual in two different time periods, t = 1 and t = 2, the two observations written out as in (15.1) are:
  - (15.7a) $y_{i1} = \beta_1 + \beta_2 x_{2i1} + \alpha_1 w_{1i} + u_i + e_{i1}$
  - (15.7b) $y_{i2} = \beta_1 + \beta_2 x_{2i2} + \alpha_1 w_{1i} + u_i + e_{i2}$
- Subtracting equation 15.7a from equation 15.7b creates a new equation:
  - (15.8) $(y_{i2} - y_{i1}) = \beta_2(x_{2i2} - x_{2i1}) + (e_{i2} - e_{i1})$

The Difference Estimator: T = 2 (cont.)

- Simplifying equation 15.8:
  - (15.9) $\Delta y_i = \Delta\beta_2 x_{i2} + \Delta e_i$
- The OLS estimator of β_2 in equation 15.9 is called the **first-difference** estimator, or simply the **difference estimator**
- It is consistent if:
  - Δe_i has zero mean and is uncorrelated with Δx_{i2}
  - Δx_{i2} takes more than two values

The Within Estimator: T = 2, Part I

- The advantage of the within transformation is that it generalizes nicely to situations when we have more than T = 2 time observations on each individual
- The time average of equations 15.7a and 15.7b are:

$$\frac{1}{2}\sum_{t=1}^{T}(y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + u_i + e_{it})$$

The Within Estimator: T = 2, Part II

- The time-averaged model for i = 1, …, N is:
  - (15.10) $\bar{y}_{i.} = \beta_1 + \beta_2 \bar{x}_{2i.} + \alpha_1 w_{1i} + u_i + \bar{e}_{i.}$
- The within transformation subtracts equation 15.10 from the original observations to obtain:
  - (15.11) $y_{it} - \bar{y}_{.i} = \beta_2 (x_{2it} - \bar{x}_{2i.}) + (e_{it} - \bar{e}_{i.})$
- The within-transformed model is:
  - (15.12) $\tilde{y}_{it.} = \beta_2 \tilde{x}_{2it} + \tilde{e}_{it}$

The Within Estimator: T = 2, Part III

- The OLS estimator of β_2 using equation 15.12 is called the within estimator
- It is a consistent estimator if:
  1. $\tilde{e}_{it}$ has zero mean and is uncorrelated with $\tilde{x}_{2it}$
  2. $\tilde{x}_{2it}$ takes more than two values
- In practice, there is no need to use the difference estimator, which was introduced as a pedagogical device to illustrate that it is possible to eliminate unobserved heterogeneity when panel data are available

The Within Estimator: T > 2

- The advantage of the within transformation and use of the within estimator is that they generalize nicely to situations when we have more than T = 2 time observations on each individual
- Suppose that we have T observations on each individual:

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + u_i + e_{it}, \qquad i = 1, \ldots, N, \qquad t = 1, \ldots, T$$

- Averaging over all time observations:

$$\frac{1}{T}\sum_{t=1}^{T}(y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + u_i + e_{it})$$

The Within Estimator: T > 2 (cont.)

- The time-averaged model, for i = 1,…, N, is:
  - (15.13) $\bar{y}_{i.} = \beta_1 + \beta_2 \bar{x}_{2i.} + \alpha_1 w_{1i} + u_i + \bar{e}_{i.}$
- The within transformation subtracts equation 15.13 from the original observations to obtain:
  - (15.14) $y_{it} - \bar{y}_{i.} = \beta_2 (x_{2it} - \bar{x}_{2i.}) + (e_{it} - \bar{e}_{i.})$
- The within-transformed model is:
  - (15.15) $\tilde{y}_{it.} = \beta_2 \tilde{x}_{2it} + \tilde{e}_{it}$

The Least Squares Dummy Variable Model

- To be as general as possible, we expand our equation of interest to include more variables:
  - (15.16) $y_{it} = \beta_1 + \beta_2 x_{2it} + \cdots + \beta_k x_{Kit} + \alpha_1 w_{1i} + \cdots + \alpha_M w_{Mi} + (u_i + e_{it})$
- Unobserved heterogeneity is also controlled for by including in the panel data regression

$$D_{1i} = \begin{cases} 1 & i=1 \\ 0 & \text{otherwise} \end{cases} \quad D_{2i} = \begin{cases} 1 & i=2 \\ 0 & \text{otherwise} \end{cases}, \cdots, \quad D_{Ni} = \begin{cases} 1 & i=N \\ 0 & \text{otherwise} \end{cases}$$

The Least Squares Dummy Variable Model (cont.)

- (15.17) $y_{it} = \beta_{11} D_{1i} + \beta_{12} D_{2i} + \cdots + \beta_{1N} D_{Ni} + \beta_2 x_{2it} + \cdots + \beta_K x_{Kit} + e_{it}$
  - This is called the fixed effects model, or sometimes the least squares dummy variable model
- Because the fixed effects estimator is simply an OLS estimator, it has the usual OLS estimator variances and covariances

(15.18) $$\hat{\sigma}_e^2 = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} \hat{e}_{it}^2}{NT - N - K_s}$$

The Least Squares Dummy Variable Model: Testing for Unobserved Heterogeneity

- Testing for individual differences in the fixed effects model is a test of the joint hypothesis
  - (15.19) $H_0: \beta_{11} = \beta_{12} = \beta_{13} = \cdots = \beta_{1,N-1} = \beta_{1N}$ and

    H_1: the $\beta_{1\mathrm{i}}$ are not all equal
- The “restricted model” is:

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \cdots + \beta_K x_{Kit} + e_{it}$$

- The F-test statistic is:
  - (15.20) $$F = \frac{(SSE_R - SSE_U)/(N-1)}{SSE_U/(NT - N - K_s)}$$

FORDHAM

THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

Panel Data
Regression Error Assumptions

FORDHAM THE JESUIT UNIVERSITY OF NEW YORK | Gabelli School of Business

Panel Data
Regression Error Assumptions

- There are two random errors in the panel data model we have been using
  1. u_i accounts for time invariant unobserved heterogeneity across individuals
  2. e_{it} is the “usual” regression error that varies across individuals and time
- With the more complete model specification, assumption 15.4 becomes 15.21: $E(e_{it}|X_i, w_i, u_i) = 0$

Panel Data
Regression Error Assumptions (cont.)

- If the explanatory variables X_i and w_i carry no information about random error component u_i, then its best prediction is zero, meaning that:
  - (15.22) $E(u_i|X_i, w_i) = 0$
- Using the law of iterated expectations, it follows that:
  - (15.23) $E(u_i) = 0, \mathrm{cov}(u_i, x_{kit}) = E(u_i x_{kit}) = 0, \mathrm{cov}(u_i, w_{mi}) = E(u_i w_{mi}) = 0$

Panel Data Regression Error Assumptions: Conditional Homoskedasticity

- The usual homoskedasticity assumption for the idiosyncratic error e_{it} is that the conditional and unconditional variances are constant
  - (15.24a) $\text{var}(e_{it}|X_i, w_i, u_i) = \sigma_e^2$
- It follows that:
  - (15.24b) $\text{var}(e_{it}) = E(e_{it}^2) = \sigma_e^2$
- Similarly:
  - (15.25) $\text{var}(u_i) = E(u_i^2) = \sigma_u^2$
- Combining the two homoskedasticity assumptions and the statistical independence of u_i and e_{it}, we have:
  - (15.26) $\text{var}(v_{it}) = \text{E}\left(v_{it}^2\right) = \sigma_v^2 = \sigma_u^2 + \sigma_e^2$

Panel Data Regression Error Assumptions: Conditionally Uncorrelated

- Find the covariance between the combined random errors in any two time periods

(15.27) $\text{cov}(v_{it}, v_{is}) = E(v_{it}v_{is}) = E[(u_i +$

$$\rho = corr(v_{it}, v_{is}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$$

OLS Estimation With Cluster-Robust Standard Errors

- In that case, var$(e_{it}) = \sigma_{it}^2$:
  - (15.29) $\operatorname{var}(v_{it}) = \sigma_u^2 + \sigma_{it}^2 = \psi_{it}^2$
- The covariance between the random errors v_{it} and v_{is} is:
  - (15.30) $\operatorname{cov}(v_{it}, v_{is}) = E(v_{it}v_{is}) = E[(u_i + e_{it})(u_i e_{is})]$ $= E(u_i^2) + E(e_{it}e_{is}) = \sigma_u^2 + \operatorname{cov}(e_{it}, e_{is})$
- Assume $\operatorname{cov}(e_{it}, e_{is}) = \sigma_{its}$, then:
  - (15.31) $\operatorname{cov}(v_{it}, v_{is}) = \sigma_u^2 + \sigma_{its} = \psi_{its}$

OLS Estimation With Cluster-Robust Standard Errors: Two Important Notes

1. Cluster-robust standard errors can be used in many contexts other than with panel data.
   - Any data containing **groups** of observations can be treated as clusters if there are within-group correlations but no across-group correlations.
2. Using cluster-robust standard errors is not always appropriate.
   - The number of individuals N must be large relative to T so that the panel is “short and wide.”

Fixed Effects Estimation With Cluster-Robust Standard Errors

Consider now the fixed effects estimation procedure that employs the "within" transformation shown in equation 15.14.

- The within transformation removes the unobserved heterogeneity so that only the idiosyncratic error e_{it} remains.
- It is possible that, within the cluster of observations defining each individual cross-sectional unit, there remains serial correlation and/or heteroskedasticity.

FORDHAM

THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

The Random Effects Estimator

The Random Effects Estimator, Part I

- Panel data applications fall into one of two types
  1. The first type of application is when the unobserved heterogeneity term u_i is correlated with one or more of the explanatory variables
  2. The second type of application is when the unobserved heterogeneity term u_i is not correlated with any of the explanatory variables
- The panel data regression model (equation 15.1) with unobserved heterogeneity is sometimes called the random effects model

The Random Effects Estimator, Part II

- The minimum variance, efficient, estimator for the model is a GLS estimator
- The FGLS estimator is called the random effects estimator
- The transformed model, using K = 2 and M = 1 in equation 15.16, is:
  - (15.32) $y_{it}^{*} = \beta_1 x_{1it}^{*} + \beta_2 x_{2it}^{*} + \alpha_1 w_{1i}^{*} + v_{it}^{*}$
- Transformed variables are:
  - (15.33) $y_{it}^{*} = y_{\mathrm{it}} - \alpha \bar{y}_{i.}, \quad x_{1it}^{*} = 1 - \alpha, \quad x_{2it}^{*} = x_{2it} - \alpha \bar{x}_{2i.}, \quad w_{1i}^{*} = w_{1i}(1-\alpha)$

The Random Effects Estimator, Part III

- The transformation parameter α is between 0 and 1, 0 < α < 1, and is given by:
  - (15.34) $\alpha = 1 - \frac{\sigma_e}{\sqrt{T\sigma_u^2 + \sigma_e^2}}$
- The variables $\bar{y}_{i.}$ and $\bar{x}_{2i.}$ are the individual time-averaged means (equation 15.13)
- w_{1i}^* is a fraction of w_{1i}
- A key feature of the random effects model is that time-invariant variables are not eliminated

Testing for Random Effects, Part I

- We can test for the presence of heterogeneity by testing the null hypothesis H_0:$\sigma^2_u = 0$ against the alternative hypothesis H_1:$\sigma^2_u > 0$
- If the null hypothesis is rejected, we conclude that there are random individual differences among sample members and that the random effects model might be appropriate
- if we fail to reject the null hypothesis, then we have no evidence to conclude that random effects are present

Testing for Random Effects, Part II

- The Lagrange multiplier (LM) principle for test construction is very convenient in this case
- If the null hypothesis is true, then $u_i = 0$, and the random effects model reduces to the usual linear regression model $$y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + e_{it}$$
- The test statistic is based on the OLS residuals

$$\hat{e}_{it} = y_{it} - b_1 - b_2 x_{2it} - a_1 w_{1i}$$

Testing for Random Effects, Part III

- The test statistic for balanced panels is:
  - (15.35) $$LM = \sqrt{\frac{NT}{2(T-1)}}\left\{\frac{\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\hat{e}_{it}\right)^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{e}_{it}^2} - 1\right\}$$
- The numerator of the first term in curly brackets differs from the denominator
- If the sum of the cross-product terms is not significant, the first term in the curly brackets is not significantly different from one, and the term in the curly brackets is not significantly different from zero

Testing for Random Effects, Part IV

- If the null hypothesis H_0:$\sigma^2_u = 0$ is true, then LM ~ $N(0, 1)$ in large samples
- Thus, we reject H_0 at significance level α and accept the alternative H_1:$\sigma^2_u > 0$ if LM > $z_{(1 – \alpha)}$, where $z_{(1 – \alpha)}$ is the $100_{(1 – \alpha)}$ percentile of the standard normal distribution
- This critical value is 1.645 if α = 0.05 and 2.326 if α = 0.01
- Rejecting the null hypothesis leads us to conclude that random effects are present

A Hausman Test for Endogeneity in the Random Effects Model, Part I

- The problem of endogenous regressors is common in random effects models because the individual-specific error component u_i may well be correlated with some of the explanatory variables
  - Such a correlation will cause the random effects estimator to be inconsistent
- The ability to test whether the random effect u_i is correlated with some of the explanatory variables is important

A Hausman Test for Endogeneity in the Random Effects Model, Part II

- To check for any correlation between the error component u_i and the regressors in a random effects model, we can use a Hausman test
- The test compares the coefficient estimates from the random effects model to those from the fixed effects model
  - The Hausman test can be carried out for specific coefficients, using a *t*-test, or jointly, using a chi-square test

A Hausman Test for Endogeneity in the Random Effects Model, Part III

- Let the parameter of interest be β_k
- Denote the fixed effects estimate as $b_{FE,k}$ and the random effects estimate as $b_{RE,k}$
- The *t*-statistic for testing that there is no difference between the estimators is:
  - (15.36) $t = \dfrac{b_{FE,k} - b_{RE,k}}{\left[\widehat{\mathrm{var}}\left(b_{FE,k}\right) - \widehat{\mathrm{var}}\left(b_{RE,k}\right)\right]^{1/2}} = \dfrac{b_{FE,k} - b_{RE,k}}{\left[se\left(b_{FE,k}\right)^2 - se\left(b_{RE,k}\right)^2\right]^{1/2}}$

A Hausman Test for Endogeneity in the Random Effects Model, Part IV

- We expect to find $\widehat{\mathrm{var}}\left(b_{FE,k}\right) - \widehat{\mathrm{var}}\left(b_{RE,k}\right)$ > 0, which is necessary for a valid test
- A second interesting feature of this test statistic is that:
  - (15.37) $\mathrm{var}\left(b_{FE,k} - b_{RE,k}\right) = \mathrm{var}\left(b_{FE,k}\right) + \mathrm{var}\left(b_{RE,k}\right) - 2\mathrm{cov}\left(b_{FE,k}, b_{RE,k}\right) = \mathrm{var}\left(b_{FE,k}\right) - \mathrm{var}\left(b_{RE,k}\right)$
- The unexpected result in the last line occurs because Hausman proved that, in this particular case, $\mathrm{cov}\left(b_{FE,k}, b_{RE,k}\right) = \mathrm{var}\left(b_{RE,k}\right)$

A Regression-Based Hausman Test, Part I

- A “regression-based” Hausman test is sometimes called the **Mundlak approach**
- Consider the general model in equation 15.16 with K = 3 and M = 2:

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \beta_3 x_{3it} + \alpha_1 w_{1i} + \alpha_1 w_{1i} + u_i + e_{it}$$

- Consider:
  - (15.38) $u_i = \gamma_1 + \gamma_2 \bar{x}_{2i.} + \gamma_3 \bar{x}_{3i.} + c_i$

  where E(c_i| X_i) = 0

A Regression-Based Hausman Test, Part II

- Specify the panel data model:
  - (15.39) $y_{it} = \beta_1 + \beta_2 x_{2it} + \beta_3 x_{3it} + \alpha_1 w_{1i} + \alpha_2 w_{2i} + u_i + e_{it} = \delta_1 + \beta_2 x_{2it} + \beta_3 x_{3it} + \alpha_1 w_{1i} + \alpha_2 w_{2i} + \gamma_2 \bar{x}_{2i\bullet} + \gamma_3 \bar{x}_{3i\bullet} + (c_i +$

A Regression-Based Hausman Test, Part III

- The null hypothesis is that there is no endogeneity arising from a correlation between the unobserved heterogeneity and the explanatory variables
- Equation 15.39 can be estimated by OLS, with cluster-robust standard errors, or by random effects, which should be more efficient
- Both OLS and random effects estimation of equation 15.39 yield fixed effects estimates of β_2 and β_3

The Hausman–Taylor Estimator, Part I

- The Hausman–Taylor estimator is an instrumental variables estimator applied to the random effects model to overcome the problem of inconsistency caused by correlation between the random effects and some of the explanatory variables
- Consider the regression model:
  - (15.40) $y_{it} = \beta_1 + \beta_2 x_{it,exog} + \beta_3 x_{it,endog} + \beta_3 w_{i,exog} + \beta_4 w_{i,endog} + u_i + e_{it}$

The Hausman–Taylor Estimator, Part II

- We have divided the explanatory variables into four categories:
  1. $x_{\text{it,exog}}$: exogenous variables that vary over time and individuals
  2. $x_{\text{it,endog}}$: endogenous variables that vary over time and individuals
  3. $w_{\text{i,exog}}$: time-invariant exogenous variables
  4. $w_{\text{i,endog}}$: time-invariant endogenous variables

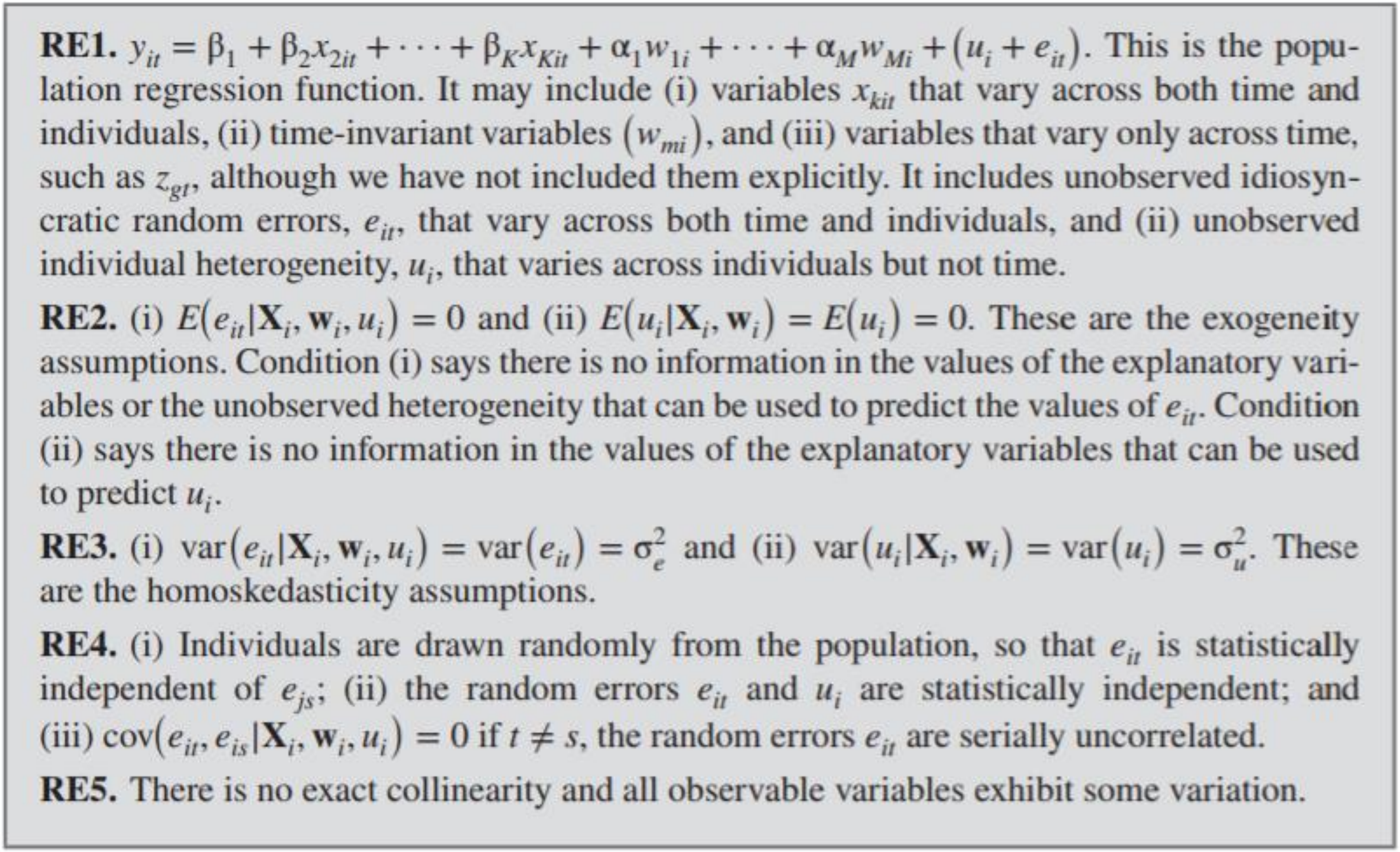
The Hausman–Taylor Estimator, Part III

- A slightly modify instrument set that can be shown to yield the same results:

$$y_{it}^* = \beta_1 + \beta_2 x_{it,exog}^* + \beta_3 x_{it,endog}^* + \beta_3 w_{i,exog}^* + \beta_4 w_{i,endog}^* + v_{it}^*$$

- Where, for example, y_{it}^* = y$_{\text{it}}$ − $\hat{a}\bar{y}_i$, and $\hat{a}=1-\hat{\sigma}_e \Big/ \sqrt{T\hat{\sigma}_u^2+\hat{\sigma}_e^2}$

- The estimate $\hat{\sigma}_e^2$ is obtained from fixed effects residuals; an auxiliary instrumental variables regression is needed to find $\hat{\sigma}_u^2$



RE1. $y_{it} = \beta_1 + \beta_2 x_{2it} + \cdots + \beta_K x_{Kit} + \alpha_1 w_{1i} + \cdots + \alpha_M w_{Mi} + (u_i + e_{it})$. This is the population regression function. It may include (i) variables x_{kit} that vary across both time and individuals, (ii) time-invariant variables (w_{mi}), and (iii) variables that vary only across time, such as z_{gt}, although we have not included them explicitly. It includes unobserved idiosyncratic random errors, e_{it}, that vary across both time and individuals, and (ii) unobserved individual heterogeneity, u_i, that varies across individuals but not time.

RE2. (i) $E(e_{it}|\mathbf{X}_i, \mathbf{w}_i, u_i) = 0$ and (ii) $E(u_i|\mathbf{X}_i, \mathbf{w}_i) = E(u_i) = 0$. These are the exogeneity assumptions. Condition (i) says there is no information in the values of the explanatory variables or the unobserved heterogeneity that can be used to predict the values of e_{it}. Condition (ii) says there is no information in the values of the explanatory variables that can be used to predict u_i.

RE3. (i) $\text{var}(e_{it}|\mathbf{X}_i, \mathbf{w}_i, u_i) = \text{var}(e_{it}) = \sigma_e^2$ and (ii) $\text{var}(u_i|\mathbf{X}_i, \mathbf{w}_i) = \text{var}(u_i) = \sigma_u^2$. These are the homoskedasticity assumptions.

RE4. (i) Individuals are drawn randomly from the population, so that e_{it} is statistically independent of e_{js}; (ii) the random errors e_{it} and u_i are statistically independent; and (iii) $\text{cov}(e_{it}, e_{is}|\mathbf{X}_i, \mathbf{w}_i, u_i) = 0$ if $t \neq s$, the random errors e_{it} are serially uncorrelated.

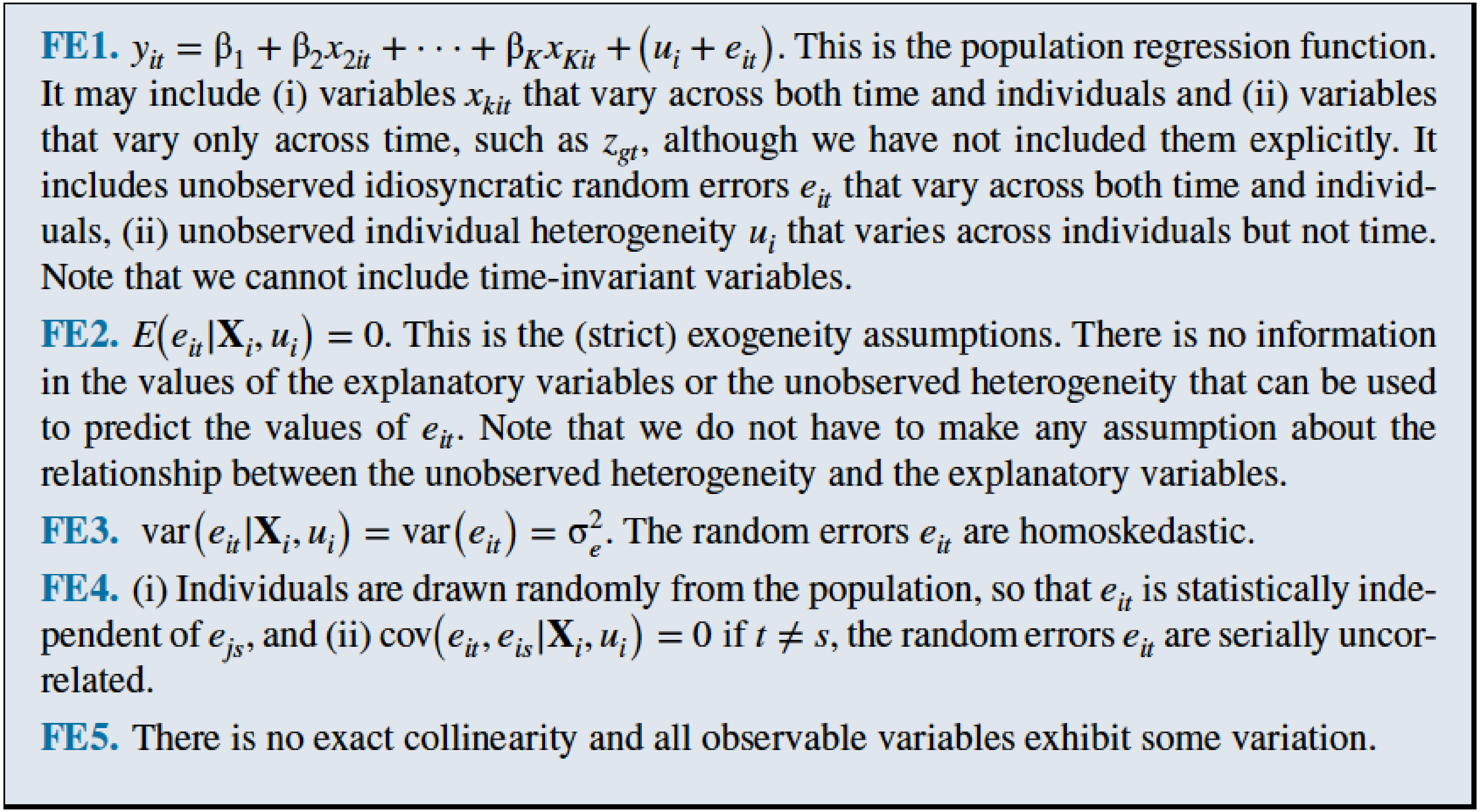
RE5. There is no exact collinearity and all observable variables exhibit some variation.

Summarizing Panel Data Assumptions: Random Effects Estimation Notes

1. Under the assumptions RE1–RE5 the random effects (GLS) estimator is BLUE, assuming σ_e^2 and σ_u^2 are known.
2. Implementation of the random effects estimator requires the variance parameters to be estimated.
3. If the random errors are either heteroskedastic (RE3 fails) and/or serially correlated (RE4 (iii) fails), then the random effects estimator is consistent and asymptotically normal, but the usual standard errors are incorrect.
4. Under RE1–RE5 the pooled OLS estimator is consistent and asymptotically normal.

Summarizing Panel Data Assumptions: Random Effects Estimation Notes (cont.)

5. Under RE1–RE5 the random effects, FGLS, estimator is more efficient asymptotically than the pooled OLS estimator with corrected, cluster-robust, standard errors.
6. The random effects estimator is more efficient in large samples than the fixed effects estimator for the coefficients of the variables that vary across individuals and time, x_{kit}.
7. The fixed effects estimator is, consistent for the coefficients of the variables that vary across individuals and time, x_{kit}, even if RE2 (ii) fails, and E $(u_i | X_i, w_i) \neq 0$.



FE1. $y_{it} = \beta_1 + \beta_2 x_{2it} + \cdots + \beta_K x_{Kit} + \left(u_i + e_{it}\right)$. This is the population regression function. It may include (i) variables x_{kit} that vary across both time and individuals and (ii) variables that vary only across time, such as z_{gt}, although we have not included them explicitly. It includes unobserved idiosyncratic random errors e_{it} that vary across both time and individuals, (ii) unobserved individual heterogeneity u_i that varies across individuals but not time. Note that we cannot include time-invariant variables.

FE2. $E\left(e_{it}|\mathbf{X}_i, u_i\right) = 0$. This is the (strict) exogeneity assumptions. There is no information in the values of the explanatory variables or the unobserved heterogeneity that can be used to predict the values of e_{it}. Note that we do not have to make any assumption about the relationship between the unobserved heterogeneity and the explanatory variables.

FE3. $\text{var}\left(e_{it}|\mathbf{X}_i, u_i\right) = \text{var}\left(e_{it}\right) = \sigma_e^2$. The random errors e_{it} are homoskedastic.

FE4. (i) Individuals are drawn randomly from the population, so that e_{it} is statistically independent of e_{js}, and (ii) $\text{cov}\left(e_{it}, e_{is}|\mathbf{X}_i, u_i\right) = 0$ if $t \neq s$, the random errors e_{it} are serially uncorrelated.

FE5. There is no exact collinearity and all observable variables exhibit some variation.

Summarizing Panel Data Assumptions: Fixed Effects Estimation Notes

1. Under FE1–FE5 the fixed effects estimator is BLUE.
2. The fixed effects estimator is consistent and asymptotically normal if N grows large and T is fixed.
3. If the random errors are either heteroskedastic (FE3 fails) and/or serially correlated (FE4 (ii) fails), then the fixed effects estimator is consistent and asymptotically normal, but the usual standard errors are incorrect.

Summarizing and Extending Panel Data Model Estimation

1. While we have not discussed it, panel data methods have been extended to unbalanced panels.
2. In addition to unobserved heterogeneity associated with individuals, there can also be unobserved heterogeneity associated with time.
3. When T = 2, first-difference estimation is perfectly equivalent to fixed effects estimation. When T > 2, the first-difference random errors $\Delta v_{it} = \Delta e_{it}$ are serially correlated, unless the idiosyncratic random errors e_{it} follow a random walk.
4. Dynamic panel data models that include a lagged dependent variable on the right-hand side have an endogeneity problem.

Summarizing and Extending Panel Data Model Estimation (cont.)

5. While we have focused on endogeneity resulting from the unobserved heterogeneity term, there can be endogeneity caused by simultaneous equations.
6. In this edition, we have chosen to omit the section on “sets of regression equations” and “seemingly unrelated regressions.” These topics arise when T is large and N is small so that each cross-sectional unit, perhaps a firm, is modeled with its own equation.
7. Unobserved heterogeneity can affect slope coefficients, that is, it is possible that each individual’s response β_{ki} to a change in x_k is different.
8. The panel data methods we have discussed can be used with linear probability models with the usual caveats.

FORDHAM

THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

Key Words

- Balanced panel
- Cluster-robust standard errors
- Deviations about the individual mean
- Difference estimator
- Endogeneity
- Error components model
- Fixed effects estimator
- Fixed effects model
- Hausman test
- Hausman–Taylor estimator
- Heterogeneity
- Instrumental variables
- Least squares dummy variable model
- LM test
- Pooled least squares
- Pooled model
- Random effects estimator
- Random effects model
- Time-invariant variables
- Time-varying variables
- Unbalanced panel
- Within estimator

Copyright

FORDHAM

THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School
of Business