Panel Data Econometrics

Dong Li

Kansas State University

Fall 2009

- Panel Data: **Individuals** can be true individuals or households, firms, states, countries, etc.
- The sample size $N \times T$. N and T may be very different (large N and small T, or small N and large T) - this is important in choosing models.

- Panel data - Cross Sectional Time Series.
- Repeated measurements (biometrics): growth of rat i at time t.
- Longitudinal Data (demography, sociology).
- It has many communalities with spatio-temporal data, multilevel analysis.

Some of the Available Micro Data Sets

U.S. data sets:

- The Panel Study of Income Dynamics (PSID): collected by University of Michigan `www.isr.umich.edu/src/psid/index.html`
- the National Longitudinal Surveys of Labor Market Experience (NLS) from the Center for Human Resource Research at Ohio State University and the Census Bureau. `www.bls.gov/nlshome.htm`
- Medical Expenditure Panel Survey `www.meps.ahrq.gov/mepsweb`
- Longitudinal retirement history supply
- Social Security Administration's Continuous Work History Sample
- Labor Department's continuous wage and benefit history
- Labor Department's continuous longitudinal manpower survey
- Negative income tax experiments
- Current Population Survey

International:

- The German Social-Economic Panel (GSOEP)
- The Belgian Socioeconomic Panel
- The Canadian Survey of Labor Income Dynamics (SLID)
- The French Household Panel
- The Hungarian Household Panel
- The British Household Panel Survey
- The Japanese Panel Survey on Consumers (JPSC)

- Sample size: typically N is large and T is small. But it is not always the case.
- Sampling: often individuals are selected randomly, at least at the beginning of the sample, but time is not.
- Non-independent data:
  - Among data to the same individual: because of unobservable characteristics of each individual.
  - Among individuals: because of unobservable characteristics common to several individuals.
  - between time period: because of dynamic behavior.

Benefits:

1. Controlling for individual heterogeneity.
2. Panel data gives more informative data, more variability, less collinearity among the variables, and more degrees of freedom.
3. Panel data is better able to study the dynamics of adjustment.
4. Panel data is better able to identify and measure effects that are simply not detectable in pure cross-sections or pure time-series data. Such as study of union membership.
5. Panel data models allow us to construct and test more complicated behavioral models than purely cross-section or time-series data. For example, technical efficiency is better studied and modeled with panels.
6. Panel data is usually gathered on micro units, like individuals, firms and households. Many variables can be more accurately measured at the micro level, and biases resulting from aggregation over firms or individuals are eliminated, see Blundell (1988) and Klevmarken (1989).

Limitations:

1. Design and data collection problems. These include problems of coverage (incomplete account of the population of interest), nonresponse (due to lack of cooperation of the respondent or because of interviewer error), recall (respondent not remembering correctly), frequency of interviewing, interview spacing, reference period, the use of bounding and time-in-sample bias, ...
2. Distortions of measurement errors. Measurement errors may arise because of faulty responses due to unclear questions, memory errors, deliberate distortion of responses (e.g. prestige bias), inappropriate informants, misrecording of responses and interviewer effects.
3. Selectivity problems.
    1. Self-selectivity.
    2. Non-response.
    3. Attrition.
4. Short Time Series Dimension. Asymptotics and limited dependent variable models.

- A panel data regression has a double subscript on its variables:

$$y_{it} = \alpha + X_{it}'\beta + u_{it}, \quad i = 1, \ldots, N; \; t = 1, \ldots, T. \tag{1}$$

- i denotes individuals, households, firms, countries, etc., the cross-section dimension.
- t denotes time, the time-series dimension.
- α is a scalar, β is $K \times 1$ and X_{it} is the it-th observation on K explanatory variables.

- Most of the panel data applications utilize a one-way model

$$u_{it} = \mu_i + \nu_{it} \tag{2}$$

  where μ_i denotes the ***unobservable*** individual specific effect and ν_{it} denotes the remainder disturbance.
- For example, in an earnings equation in labor economics, μ_i may include the individual's (time-invariant) unobserved ability.
- A production function utilizing data on firms across time, $\mu_i$may capture the unobservable entrepreneurial or managerial skills of the firm's executives.
- The remainder disturbance ν_{it} varies with individuals and time and can be thought of as the usual disturbance in the regression.

- It is called a balanced panel if every individual has the same time span $t = 1, ..., T$.
- It is called an unbalanced panel if not so (then we have total number of observations $\sum_{i=1}^{N} T_i$).
- We are going to assume we have balanced panel throughout the semester unless otherwise noted.
- The unbalanced panel estimation can be obtained similar to the balanced one. Most statistical software can take care of it "automatically."

- In vector form (1) can be written as

$$y = \alpha \iota_{NT} + X\beta + u = Z\delta + u \tag{3}$$

where y is $NT \times 1$, X is $NT \times K$, $Z = [\iota_{NT}, X]$, $\delta' = (\alpha', \beta')$, and ι_{NT} is a vector of ones of dimension $NT = N \times T$.

- (2) can be written in matrix form as

$$u = Z_\mu \mu + \nu \tag{4}$$

where $u' = (u_{11}, \ldots, u_{1T}, u_{21}, \ldots, u_{2T}, \ldots, u_{N1}, \ldots, u_{NT})$ with the observations stacked such that the slower index is over individuals and the faster index is over time.

- $Z_\mu = I_N \otimes \iota_T$ where I_N is an identity matrix of dimension N, ι_T is a vector of ones of dimension T, and $\otimes$ denotes Kronecker product.
- Z_μ is a selector matrix of ones and zeros, or simply the matrix of individual dummies that one may include in the regression to estimate the μ_i's if they are assumed to be fixed parameters. $\mu' = (\mu_1, \ldots, \mu_N)$ and $\nu' = (\nu_{11}, \ldots, \nu_{1T}, \ldots, \nu_{N1}, \ldots, \nu_{NT})$.

Matrices P and Q

- $Z_\mu' Z_\mu = (I_N \otimes \iota_T')(I_N \otimes \iota_T) = I_N \otimes T = TI_N$.
- The projection matrix,
  $P = Z_\mu (Z_\mu' Z_\mu)^{-1} Z_\mu' = (I_N \otimes \iota_T)\frac{1}{T} I_N^{-1} (I_N \otimes \iota_T') = I_N \otimes \bar{J}_T$, where $\bar{J}_T = J_T/T$ (average matrix) and J_T is a matrix of ones of dimension $T \times T$.
- P is a matrix which averages the observation across time for each individual, and $Q = I_{NT} - P = I_N \otimes (I_T - \bar{J}_T)$ is a matrix which obtains the deviations from individual means.
- For example, Pu has a typical element $\bar{u}_{i\cdot} = \sum_{t=1}^{T} u_{it}/T$ repeated T times for each individual and Qu has a typical element $(u_{it} - \bar{u}_{i\cdot})$.

Matrices P and Q

Properties of P and Q:

- P and Q are symmetric idempotent matrices, i.e., $P' = P$ and $P^2 = P$. This means that the $\text{rank}(P) = \text{tr}(P) = N$ and $\text{rank}(Q) = \text{tr}(Q) = N(T-1)$. This uses the result that rank of an idempotent matrix is equal to its trace, see Graybill (1961, Theorem 1.63).
- P and Q are orthogonal, i.e., $PQ = 0$.
- They sum to the identity matrix $P + Q = I_{NT}$.

Any two of these properties imply the third, see Graybill (1961, Theorem 1.68).

- The μ_i's are assumed to be fixed parameters to be estimated.
- The remainder disturbances are stochastic with $\nu_{it} \sim \text{IID}(0, \sigma_\nu^2)$.
- The X_{it}'s are assumed independent of the ν_{it}'s for all i and t. (We will relax this assumption later.)
- The fixed effects model is an appropriate specification if we are focusing on a specific set of N firms and our inference is restricted to the behavior of these sets of firms. Alternatively, it could be a set of N OECD countries, or N American States. Inference in this case is conditional on the particular N firms, countries, or states that are observed.

One can substitute the disturbances given by (4) into (3) to get

$$y = \alpha \iota_{NT} + X\beta + Z_\mu \mu + \nu = Z\delta + Z_\mu \mu + \nu. \tag{5}$$

Methods to estimate the FE model

Method 1: Brutal force OLS

Brutal force OLS on (5) to get estimates of α, β and μ: Note that Z is $NT \times (K+1)$ and Z_μ, the matrix of individual dummies is $NT \times N$. If N is large, (5) will include too many individual dummies, and the matrix to be inverted by OLS is large and of dimension $(N+K)$. It is not feasible in most statistical software when N is large.

Since α and β are the parameters of interest, one can obtain the estimates from (5) by pre-multiplying the model by Q and performing OLS on the resulting transformed model:

$$Qy = QX\beta + Qv \tag{6}$$

This uses the fact that $QZ_\mu = Q\iota_{NT} = 0$, since $PZ_\mu = Z_\mu$. In other words, the Q matrix wipes out the individual effects. This is a regression of $\tilde{y} = Qy$ with typical element $(y_{it} - \bar{y}_{i\cdot})$ on $\tilde{X} = QX$ with typical element $(X_{it,k} - \bar{X}_{i\cdot,k})$ for the k-th regressor, $k = 1, 2, \ldots, K$. This involves the inversion of a $(K \times K)$ matrix rather than $(N+K) \times (N+K)$ as in (5). The resulting OLS estimator is

$$\tilde{\beta} = (X'QX)^{-1} X'Qy \tag{7}$$

with $\text{Var}(\tilde{\beta}) = \sigma_v^2 (X'QX)^{-1} = \sigma_v^2 (\tilde{X}'\tilde{X})^{-1}$. Notice the variance-covariance matrix.

GLS on (5): $\text{Var}(Qv) = \sigma_v^2 Q$. So the GLS estimator

$$\hat{\beta} = ((QX)'(\sigma_v^2 Q)^{-1}(QX))^{-1}(QX)'(\sigma_v^2 Q)^{-1}(Qy) = (X'QX)^{-1}X'Qy$$

with $\text{Var}(\tilde{\beta}) = \sigma_v^2 (X'QX)^{-1} = \sigma_v^2 (\tilde{X}'\tilde{X})^{-1}$.

Methods to estimate the FE model

While the theory in the above estimation discussions is simple, it may not be practical to implement these three methods.
When NT is large (for example, $N = 1000$ and $T = 10$: it means that Q is a 10000 by 10000 matrix. in GAUSS it takes 8×10^8 bytes or roughly 763MB. Imagine that you need more space to operate on it) pre-multiplying the model by Q in the above methods 2 and 3 is infeasible in most statistical software.

Methods to estimate the FE model

Method 4: Within(demean) transformation in scalar

Consider the within transformation without matrix notation for the simple regression

$$y_{it} = \alpha + \beta x_{it} + \mu_i + v_{it}. \tag{8}$$

Averaging over time for each individual gives the between regression

$$\bar{y}_{i\cdot} = \alpha + \beta \bar{x}_{i\cdot} + \mu_i + \bar{v}_{i\cdot} \tag{9}$$

and the difference between the above two regressions gives the within regression

$$y_{it} - \bar{y}_{i\cdot} = \beta(x_{it} - \bar{x}_{i\cdot}) + (v_{it} - \bar{v}_{i\cdot}) \tag{10}$$

also averaging across all observations gives

$$\bar{y}_{\cdot\cdot} = \alpha + \beta \bar{x}_{\cdot\cdot} + \bar{v}_{\cdot\cdot} \tag{11}$$

where we utilized the restriction that $\sum \mu_i = 0$. This is an arbitrary restriction on the dummy variable coefficients to avoid the dummy variable trap, or perfect multicollinearity.

$\tilde{\beta}$ is obtained from regression (10), $\tilde{\alpha} = \bar{y}_{..} - \tilde{\beta}\bar{x}_{..}$ can be recovered from the last equation and $\tilde{\mu}_i = \bar{y}_{i.} - \tilde{\alpha} - \tilde{\beta}\bar{x}_{i.}$ from (9).

Possible drawbacks for the FE model

1. The fixed effects (FE) least squares, also known as least squares dummy variables (LSDV) suffers from a large loss of degrees of freedom.
2. We are estimating $(N-1)$ extra parameters, and too many dummies may aggravate the problem of multicollinearity among the regressors.
3. In addition, the FE estimator cannot estimate the effect of any time invariant variable like sex, race, religion, schooling, or union participation. These time invariant variables are wiped out by the Q transformation, the deviation-from-mean transformation.
4. If the FE model is the true model, LSDV is BLUE as long as ν is a standard classical disturbance with mean 0 and variance covariance matrix $\sigma_\nu^2 I_{NT}$. Note that as $T \to \infty$, the FE estimator is consistent. However, if T is fixed and $N \to \infty$ as typical in short labor panels, then only the FE estimator of β is consistent, the FE estimators of the individual effects $(\alpha + \mu_i)$ are not consistent since the number of these parameters increase as N increases.

A few comments

1. Testing for fixed effects. One could test the joint significance of these dummies, i.e., H_0: $\mu_1 = ... = \mu_{N-1} = 0$, by performing an F test. This is a simple F test with the RRSS being that of OLS on the pooled model and the URSS being that of the LSDV regression.

$$F = \frac{(RRSS - URSS)/(N-1)}{URSS/(NT-N-K)} \overset{H_0}{\sim} F_{N-1,N(T-1)-K} \tag{12}$$

   This test is available after FE regression in Stata.
2. Computational Warning. One computational caution for those using the within regression given by (10). The s^2 of this regression as obtained from a typical regression package divides the residual sums of squares by $NT-K$ since the intercept and the dummies are not included. The proper s^2, say s^{*2} from the LSDV regression would divide the same residual sums of squares by $N(T-1)-K$. Therefore, one has to adjust the variances obtained from the within regression (10) by multiplying the variance-covariance matrix by (s^{*2}/s^2) or simply by multiplying by $[NT-K]/[N(T-1)-K]$.

Method 5: First Difference

First difference the model (1) to get

$$\Delta y_{it} = \Delta X_{it}'\beta + \Delta u_{it} = \Delta X_{it}'\beta + \Delta v_{it}, \quad i = 1, \ldots, N;\ t = 2, \ldots, T. \tag{13}$$

where $\Delta y_{it} = y_{it} - y_{i,t-1}$. The original model can be fixed effects or random effects. But FD is mostly used in the fixed effects model. It offers another method to remove the individual effects.

Some potential drawbacks with first difference method

1. Any time-invariant regressor would result in a column of 0s and we cannot estimate its effect.
2. Some time variant variables may result in a constant term. For example, variable age in yearly data would generate $\Delta age = 1$.
3. If the temporal variation of x_{it} is small, i.e., Δx_{it} is small, we get to estimate the coefficients with low precision in the differenced model.

We will discuss the first difference model with more details in the dynamic models.

Assumptions

1. $\mu_i \sim \text{IID}(0, \sigma_\mu^2)$;
2. $v_{it} \sim \text{IID}(0, \sigma_v^2)$;
3. μ_i's are independent of the v_{it}'s;
4. In addition, the X_{it}'s are independent of the μ_i's and v_{it}'s for all i and t. This is the cost of random effects model compared to the fixed effects model.

The random effects model is an appropriate specification if we are drawing N individuals randomly from a large population. This is usually the case for household panel studies. Care is taken in the design of the panel to make it 'representative' of the population we are trying to make inference about. In this case, N is usually large and a fixed effects model would lead to an enormous loss of degrees of freedom. The individual effect is characterized as random and inference pertains to the population from which this sample was randomly drawn.

Estimation

From (4), one can compute the variance-covariance matrix

$$\Omega = E(uu') = Z_\mu E(\mu\mu')Z'_\mu + E(\nu\nu') \tag{14}$$

$$= \sigma^2_\mu (I_N \otimes J_T) + \sigma^2_\nu (I_N \otimes I_T) \tag{15}$$

This implies a homoskedastic variance $\text{Var}(u_{it}) = \sigma^2_\mu + \sigma^2_\nu$ for all i and t, and an equi-correlated block-diagonal covariance matrix which exhibits serial correlation over time only between the disturbances of the same individual.

$$\text{Cov}(u_{it}, u_{js}) = \begin{cases} \sigma^2_\mu + \sigma^2_\nu, & \text{for } i = j,\ t = s; \\ \sigma^2_\mu, & \text{for } i = j,\ t \neq s; \\ 0, & \text{for } i \neq j. \end{cases}$$

The variance-covariance matrix $\mathrm{E}(uu')$ is given by

$$\Omega = \begin{bmatrix} \begin{pmatrix} \sigma_u^2 & \sigma_\mu^2 & \dots & \sigma_\mu^2 \\ \sigma_\mu^2 & \sigma_u^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \sigma_\mu^2 \\ \sigma_\mu^2 & \dots & \sigma_\mu^2 & \sigma_u^2 \end{pmatrix} & & 0 & 0 \\ & & & \\ & 0 & (\dots) & 0 \\ & & 0 & (\dots) \end{bmatrix}$$

where $\sigma_u^2 = \mathrm{Var}(u_{it}) = \sigma_\mu^2 + \sigma_v^2$.

The correlation coefficient between u_{it} and u_{js} is

$$\text{Corr}(u_{it}, u_{js}) = \begin{cases} 1, & \text{for } i = j,\ t = s; \\ \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_\nu^2}, & \text{for } i = j,\ t \neq s; \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

- To obtain the GLS estimator of the regression coefficients, we need Ω^{-1}.
- This is a huge matrix for typical panels and is of dimension $(NT \times NT)$. Impractical to invert the matrix using brutal force.
- Results in Wansbeek and Kapteyn (1982b,1983) provide the derivations of Ω^{-1} and $\Omega^{-1/2}$.

- Define $E_T = I_T - \bar{J}_T$. Replace J_T by $T\bar{J}_T$, and I_T by $(E_T + \bar{J}_T)$. Then

$$\begin{aligned}\Omega &= T\sigma_\mu^2(I_N \otimes \bar{J}_T) + \sigma_\nu^2(I_N \otimes E_T) + \sigma_\nu^2(I_N \otimes \bar{J}_T) \\ &= (T\sigma_\mu^2 + \sigma_\nu^2)(I_N \otimes \bar{J}_T) + \sigma_\nu^2(I_N \otimes E_T) \\ &= \sigma_1^2 P + \sigma_\nu^2 Q \qquad (17)\end{aligned}$$

  where $\sigma_1^2 = T\sigma_\mu^2 + \sigma_\nu^2$.

- This is the spectral decomposition representation of Ω, with σ_1^2 being the first unique characteristic root of Ω of multiplicity N and σ_ν^2 is the second unique characteristic root of Ω of multiplicity $N(T-1)$.
- It is easy to verify, using the properties of P and Q, that

$$\Omega^{-1} = \frac{1}{\sigma_1^2} P + \frac{1}{\sigma_\nu^2} Q \qquad (18)$$

  and

$$\Omega^{-1/2} = \frac{1}{\sigma_1} P + \frac{1}{\sigma_\nu} Q. \qquad (19)$$

- Now we can obtain GLS estimator without actually inverting $NT \times NT$ matrix Ω

$$\delta = (Z'\Omega^{-1}Z)^{-1}Z'\Omega^{-1}y. \tag{20}$$

  GLS is the BLUE.

- Fuller and Battese (1973, 1974) suggested pre-multiplying the regression equation given in (3) by $\sigma_v \Omega^{-\frac{1}{2}} = Q + (\sigma_v/\sigma_1)P$ and performing OLS on the resulting transformed regression. This was particularly helpful in the old days when OLS was a big deal in computation.
- In this case, $y^* = \sigma_v \Omega^{-\frac{1}{2}} y$ has a typical element $y_{it} - \theta \bar{y}_{i\cdot}$ where $\theta = 1 - (\sigma_v/\sigma_1)$. This transformed regression only requires inversion of a matrix of dimension (K+1) and can be easily implemented using any regression package. Notice that you have to transform every variable, including the intercept. Also when you run the transformed regression, make sure no constant is included.

- But the above method is not feasible yet since we do not know σ_1 or σ_v.
- The Best Quadratic Unbiased (BQU) estimators of the variance components arise naturally from the spectral decomposition of Ω:

$$\hat{\sigma}_1^2 = \frac{u'Pu}{\text{tr}(P)} \text{ where } \text{tr}P = N. \tag{21}$$

$$\hat{\sigma}_v^2 = \frac{u'Qu}{\text{tr}(Q)} \text{ where } \text{tr}Q = N(T-1). \tag{22}$$

We can show that the above estimators are unbiased.

$$\begin{aligned} \mathrm{E}(u'Qu) &= \mathrm{E}(\mathrm{tr}(u'Qu)) = \mathrm{E}(\mathrm{tr}(uu'Q)) = \mathrm{tr}(\mathrm{E}(uu'Q)) = \mathrm{tr}(\mathrm{E}(uu')Q) \\ &= \mathrm{tr}(\Omega Q) = \mathrm{tr}[(\sigma_1^2 P + \sigma_v^2 Q)Q] = \mathrm{tr}(\sigma_v^2 Q) = \sigma_v^2 \mathrm{tr}\, Q \end{aligned}$$

So

$$\mathrm{E}(\hat{\sigma}_v^2) = \mathrm{E}\left(\frac{u'Qu}{\mathrm{tr}(Q)}\right) = \frac{\sigma_v^2 \mathrm{tr}(Q)}{\mathrm{tr}(Q)} = \sigma_v^2.$$

Similar results can be obtained for $\hat{\sigma}_1^2$.
The true disturbances u are not known and therefore (21) and (22) are not feasible. How to estimate u ($\hat{u}$) to make the GLS feasible (FGLS)?

Ways to estimate the error components:

1. Wallace and Hussain (1969) suggest substituting OLS residuals $\hat{u}_{OLS}$ instead of the true u's, because the OLS estimates are still unbiased and consistent, though no longer efficient.
2. Amemiya (1971) shows that the Wallace and Hussain estimators of the variance components have a different asymptotic distribution from that knowing the true disturbances. He suggests using the LSDV residuals instead of the OLS residuals.
3. Swamy and Arora (1972) suggest running two regressions to get estimates of the variance components from the corresponding mean square errors of these regressions. The first regression is the Within regression which yields $\hat{\hat{\sigma}}_v^2 = [y'Qy - y'QX(X'QX)^{-1}X'Qy]/[N(T-1)-K]$. The second regression is the Between regression which yields $\hat{\hat{\sigma}}_1^2 = [y'Py - y'PZ(Z'PZ)^{-1}Z'Py]/(N-K-1)$.
4. Nerlove (1971) suggests $\hat{\sigma}_\mu^2 = \sum_{i=1}^{N}(\hat{\mu}_i - \bar{\hat{\mu}})^2/(N-1)$ where $\hat{\mu}_i$ are the dummy coefficients estimates from the LSDV regression. $\hat{\sigma}_v^2$ is estimated from the within residual sums of squares divided by NT without correction for degrees of freedom.

Some further discussions on the Swamy-Arora method: Note that stacking the two transformed regressions, the between and the within regression, yields

$$\begin{pmatrix} Qy \\ Py \end{pmatrix} = \begin{pmatrix} QZ \\ PZ \end{pmatrix} \delta + \begin{pmatrix} Qu \\ Pu \end{pmatrix} \tag{23}$$

and the transformed error has mean 0 and variance-covariance matrix given by

$$\begin{pmatrix} \sigma_v^2 Q & 0 \\ 0 & \sigma_1^2 P \end{pmatrix}.$$

- OLS on this system of $2NT$ observations yields OLS on the pooled model (3). Also, GLS on this system yields GLS on (3). (This means the two transformations have not lost any information here.)
- One can show that the GLS random effects estimator is a weighted average of the within estimator and the between estimator: $\hat{\beta}_{GLS} = W\hat{\beta}_{Between} + (I - W)\hat{\beta}_{Within}$.

A significant “drawback" of random effects model is that it is assumed that the individual effects are not correlated with x_{it}, which is questionable in many applications.

Note that $\Omega = \mathrm{E}(uu') = \sigma_1^2 P + \sigma_v^2 Q$. So the loglikelihood function is

$$\begin{aligned}
\log L &= -\frac{NT}{2}\ln(2\pi) - \frac{1}{2}\ln|\Omega| - \frac{1}{2}u'\Omega^{-1}u \\
&= -\frac{NT}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma_1^2)^N(\sigma_v^2)^{N(T-1)} - \frac{1}{2}u'\left(\frac{1}{\sigma_1^2}P + \frac{1}{\sigma_v^2}Q\right)u \\
&\quad \text{(by noting that } |A| = \lambda_1\lambda_2...\lambda_n)
\end{aligned}$$

- Specify the i and t variables: `tsset id year` or `xtset id year`
- One benefit is that afterwards you can use the lag operator. In a panel setup, lag will be within each individual.

Sample Stata program.

- If the (unobserved) individual effects μ_i's are random, the error terms u_{it} are autocorrelated (but homoscedastic). The OLS estimators are still unbiased, but not efficient. The OLS standard errors are misleading.
- If the (unobserved) individual effects μ_i's are fixed and we run OLS without μ in the regression, it causes the omitted variable problem (omission of relevant variables). The OLS estimators are biased.
  If the true model is $y = X\beta + Z\gamma + u$ but we estimate $y = X\beta + u$ instead ($\tilde{\beta} = (X'X)^{-1}X'y$), we have

$$\begin{aligned} \tilde{\beta} &= (X'X)^{-1}X'(X\beta + Z\gamma + u) \\ &= \beta + (X'X)^{-1}X'Z\gamma + (X'X)^{-1}X'u \end{aligned}$$

  So

$$\mathrm{E}(\tilde{\beta}) = \beta + \mathrm{E}[(X'X)^{-1}X'Z]\gamma + 0 \tag{24}$$

  The above expectation is not β unless the second term is zero.

Appendix 2A: Trace of a (Square) Matrix

The trace of a square $n \times n$ matrix A, denoted $\text{tr}(A)$, is the sum of its diagonal elements: $\text{tr}(A) = \sum_{i=1}^{n} a_{ii}$.

We have

$$\begin{aligned} \text{tr}(A+B) &= \text{tr}A + \text{tr}B && (25) \\ \text{tr}(kA) &= k\,\text{tr}A && (26) \\ \text{tr}A' &= \text{tr}A && (27) \\ \text{tr}(AB) &= \text{tr}(BA) && (28) \end{aligned}$$

Appendix 2B: The Kronecker Product

Let A be an $m \times n$ matrix and B a $p \times q$ matrix. The $mp \times nq$ matrix defined by

$$\begin{bmatrix} a_{11}B & \dots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \dots & a_{mn}B \end{bmatrix}$$

is called the Kronecker product of A and B and written $A \otimes B$.

Some properties of the Kronecker product:

$$\begin{aligned} A \otimes B \otimes C &= (A \otimes B) \otimes C = A \otimes (B \otimes C) \\ (A+B) \otimes (C+D) &= A \otimes C + A \otimes D + B \otimes C + B \otimes D \text{ if } A+B \text{ and } C+D \text{ exist} \\ (A \otimes B)(C \otimes D) &= AC \otimes BD \text{ if } AC \text{ and } BD \text{ exist} \\ (A \otimes B)' &= A' \otimes B' \\ (A \otimes B)^{-1} &= A^{-1} \otimes B^{-1} \text{ if } A \text{ and } B \text{ are non-singular.} \end{aligned}$$

Appendix 2C: Eigenvector and Eigenvalue

A scalar (real number) λ is said to be an eigenvalue of an $n \times n$ matrix A if there exists an $n \times 1$ non-null vector x such that $Ax = \lambda x$. x is called the eigenvector associated with λ. Eigenvalues are solutions to $|A - \lambda I| = 0$.

Two important properties:

$$\text{tr}A = \lambda_1 + \lambda_2 + ... + \lambda_n$$

and

$$|A| = \lambda_1 \lambda_2 ... \lambda_n.$$

Appendix 2D: Matrix Calculus

Suppose that we have the vectors $\mathbf{a}$, $\mathbf{x}$, and the matrix $\mathbf{B}$ defined as

$$\mathbf{a}=\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \qquad \mathbf{x}=\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \qquad \mathbf{B}=\begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix}$$

Then $L=\mathbf{a}'\mathbf{x}=\sum_{i=1}^{3} a_i x_i$ is said to be a linear form in $\mathbf{x}$ and $Q=\mathbf{x}'\mathbf{B}\mathbf{x}=\sum_{i=1}^{3}\sum_{j=1}^{3} b_{ij}x_i x_j$ is said to be a quadratic form in $\mathbf{x}$. If the dimension of the vectors and matrix is more than 3, the following rules still apply.

Note that $\partial L/\partial x_i = a_i$. We shall denote the vector of partial derivatives

$$\begin{pmatrix} \frac{\partial L}{\partial x_1} \\ \frac{\partial L}{\partial x_2} \\ \frac{\partial L}{\partial x_3} \end{pmatrix}$$

by $\partial L/\partial \mathbf{x}$. Thus we have $\partial L/\partial \mathbf{x}=\mathbf{a}$. We also have

$$\begin{aligned} \frac{\partial Q}{\partial \mathbf{x}} &= \mathbf{B}\mathbf{x}+\mathbf{B}'\mathbf{x} \\ &= 2\mathbf{B}\mathbf{x} \text{ if } \mathbf{B} \text{ is symmetric} \end{aligned}$$

Appendix 2E: Partitioned Regression Consider a partition regression $y = Z\beta + W\gamma + e = X\delta + e$ where all are matrices. To estimate β without estimating γ.

Recall that the normal equation for OLS is $X'X\hat{\delta} = X'y$, which is equivalent to

$$\begin{pmatrix} Z' \\ W' \end{pmatrix} \begin{pmatrix} Z & W \end{pmatrix} \hat{\delta} = \begin{pmatrix} Z'y \\ W'y \end{pmatrix} \tag{29}$$

and

$$\begin{pmatrix} Z'Z & Z'W \\ W'Z & W'W \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = \begin{pmatrix} Z'y \\ W'y \end{pmatrix} \tag{30}$$

and

$$Z'Z\hat{\beta} + Z'W\hat{\gamma} = Z'y \tag{31}$$

$$W'Z\hat{\beta} + W'W\hat{\gamma} = W'y \tag{32}$$

From (32) one can find $\hat{\gamma} = (W'W)^{-1}W'(y - Z\hat{\beta})$. Plug this equation into (31) one can obtain

$$Z'Z\hat{\beta} + Z'W(W'W)^{-1}W'(y - Z\hat{\beta}) = Z'y \tag{33}$$

$$Z'Z\hat{\beta} - Z'W(W'W)^{-1}W'Z\hat{\beta} = Z'y - Z'W(W'W)^{-1}W'y \tag{34}$$

$$Z'[I - W(W'W)^{-1}W']Z\hat{\beta} = Z'[I - W(W'W)^{-1}W']y \tag{35}$$

Define $P_W = W(W'W)^{-1}W'$ and $\bar{P}_W = I - P_W$. P_W is the projection matrix and $\bar{P}_W$ is the residual projection. It is easy to verify that both P_W and $\bar{P}_W$ are symmetric and idempotent. The above equation becomes

$$(Z'\bar{P}_W Z)\hat{\beta} = Z'\bar{P}_W y \tag{36}$$

$$\hat{\beta} = (Z'\bar{P}_W Z)^{-1}Z'\bar{P}_W y \tag{37}$$

which is also the OLS estimator from the following regression $\bar{P}_W y = \bar{P}_W Z\beta + \bar{P}_W e$.

Consider two regressions:

- Regress y on Z and W to obtain $\hat{\beta}$;
- regress y on W and obtain the residuals; regress Z on W and obtain the residuals; regress the first set of residuals on the second set of residuals and obtain the coefficients $\tilde{\beta}$.

The Frisch-Waugh-Lovell Theorem (FWL) states: $\hat{\beta}$ and $\tilde{\beta}$ are identical; the residuals from the partitioned regression are identical to the residuals from the original regression.

References

- Harville, D.A., *Matrix Algebra from a Statistician's Perspective*, Springer, 1997
- Magnus, J.R. and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Wiley, 1988.
- Berck, P. and K. Sydsaeter, *Economists's Mathematical Manual*, Second Ed, Springer-Verlag, 1993.

- The first question we want to ask is whether we can pool the different individuals, firms, states, regions, countries together.
- For the unrestricted model, we have a regression equation for each region given by

$$y_i = Z_i \delta_i + u_i \text{ for } i = 1, 2, ..., N \tag{38}$$

where $y_i' = (y_{i1}, ..., y_{iT})$, $Z_i = (\iota_T, X_i)$ and X_i is $T \times K$. δ_i' is $1 \times (K+1)$ and u_i is $T \times 1$.

- Note that δ_i is different for every i. This unrestricted model can be written as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} Z_1 & 0 & \dots & 0 \\ 0 & Z_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & Z_N \end{pmatrix} \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_N \end{pmatrix} + u \tag{39}$$

- We want to test the hypothesis $H_0 : \delta_i = \delta$ for all i, so that under H_0 we can write the restricted model as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_N \end{pmatrix} \delta + u \tag{40}$$

- Under assumption $u \sim N(0, \sigma^2 I_{NT})$ the test becomes a Chow test (F test). You can get the Unrestricted RSS from the separate regressions in (38) and add them up. You can obtain the Restricted RSS from (40).
- If $u \sim N(0, \sigma^2 \Omega)$ the test becomes a Chow test (F test) for a GLS model.

- You want to test the null hypothesis $H_0 : \mu_1 = ... = \mu_{N-1} = 0$.
- The unrestricted model is

$$y_{it} = \alpha + X_{it}'\beta + \mu_i + \nu_{it}, \quad i = 1, \ldots, N; \; t = 1, \ldots, T. \tag{41}$$

- The restricted model is

$$y_{it} = \alpha + X_{it}'\beta + \nu_{it}, \quad i = 1, \ldots, N; \; t = 1, \ldots, T. \tag{42}$$

- The F test is

$$F = \frac{(RRSS - URSS)/(N-1)}{URSS/(NT - N - K)} \overset{H_0}{\sim} F_{N-1, N(T-1)-K} \tag{43}$$

- Stata actually reports this F test after fixed effects estimation.

- Consider

$$y_{it} = \alpha + X_{it}'\beta + \mu_i + \nu_{it}, \quad i = 1, \ldots, N;\ t = 1, \ldots, T. \tag{44}$$

  where $\mu_i \sim$ IID $N(0, \sigma_\mu^2)$. We derived the loglikelihood function in Chapter 2. The Lagrange Multiplier (LM) test to test $H_0 : \sigma_\mu^2 = 0$, also called the Breusch-Pagan test, can be constructed based on the logl.
- In Stata the command `xttest0` after random effects estimation will give the test result.
- Honda (1985) modified the test to account for the fact that the alternative is one-sided $\sigma_\mu^2 > 0$.
- Moulton and Randolph (1989) and Honda (1991) further standardize the above test by subtracting its mean and then divided by its standard deviation.

Some comparisons between the random effects model and the fixed effects model:

- The fixed effects model loses large degrees of freedom $(N-1)$.
- The fixed effects model allows the individual effects to be correlated with the regressors.
- If you want to estimate time-invariant variables in the panel data model, you have to use the random effects model.
- If we want to make inference only this set of cross-sectional units then we should treat μ_i's as fixed. On the other hand, if we want to make inference about the population from which the cross sectional data came, we should treat μ_i as random. In most of the applied econometric work, the latter is the case.

The Hausman Test

The Hausman test is based on the following idea: If H_0 is true, FE and RE estimators are close to each other; otherwise they are very different from each other.

Table: The Hausman Test

	Under H_0	H_1
RE	consistent and efficient	inconsistent
FE	consistent but inefficient	consistent
H_0: There is no correlation between μ_i's and regressors.		
H_1: There may exist correlation between μ_i's and regressors.		

This same idea leads to the Hausman test for OLS vs. IV.

- A "mechanic" way to decide the model choice between the fixed effects and the random effects. Compare the difference between the random effects estimates and the fixed effects estimates. If the difference is significant, we go for the fixed effects. If not, go for the random effects (because it is more efficient).
- H_0 : μ_i's are uncorrelated with X'_{it} versus H_1 : μ_i's are correlated with X'_{it}. The fixed effects model is unbiased and consistent under H_0 and H_1. It is efficient under H_1 but inefficient under H_0. The random effects model is biased and inconsistent under H_1 but consistent and efficient under H_0.
- Define $\hat{q} = \hat{\beta}_{FE} - \hat{\beta}_{RE}$. It turns out that $\text{Var}(\hat{q}) = \text{Var}(\hat{\beta}_{FE}) - \text{Var}(\hat{\beta}_{RE})$ under H_0:

$$\begin{aligned}
\text{Var}(\hat{q}) &= \text{Var}(\hat{\beta}_{FE} - \hat{\beta}_{RE}) \\
&= \text{Var}(\hat{\beta}_{FE}) + \text{Var}(\hat{\beta}_{RE}) - 2\text{Cov}(\hat{\beta}_{FE}, \hat{\beta}_{RE}) \\
&= \text{Var}(\hat{\beta}_{FE}) + \text{Var}(\hat{\beta}_{RE}) - 2(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}E(uu')QX(X'QX)^{-1} \\
&= \text{Var}(\hat{\beta}_{FE}) + \text{Var}(\hat{\beta}_{RE}) - 2(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\Omega QX(X'QX)^{-1} \\
&= \text{Var}(\hat{\beta}_{FE}) - \text{Var}(\hat{\beta}_{RE})
\end{aligned}$$

The test statistic

$$m = \hat{q}' [\text{Var}(\hat{q})]^{-1} \hat{q} \tag{45}$$

and under H_0 it is asymptotically distributed as χ^2_K, where K denotes the dimension of slope vector.

```
xtreg lgaspcar lincomep lrpmg lcarpcap, re
est store random
xtreg lgaspcar lincomep lrpmg lcarpcap, fe
est store fixed
hausman fixed random
```

- Note the order of the two estimates: `hausman consistent efficient`
- A pitfall of the Hausman test is that it may be negative in finite samples.

Considered the two-way models:

$$u_{it} = \mu_i + \lambda_t + \nu_{it}, \;\; i = 1, \ldots, N; \; t = 1, \ldots, T \tag{46}$$

where μ_i denotes the *unobservable* individual effect, λ_t denotes the *unobservable* time-effect and ν_{it} is the remainder stochastic disturbance term.
Note that λ_t is individual-invariant and it accounts for any time-specific effect that is not included in the regression.

In vector form, (46) can be written as

$$u = Z_\mu \mu + Z_\lambda \lambda + \nu \tag{47}$$

where Z_μ, μ and ν were defined earlier. $Z_\lambda = \iota_N \otimes I_T$ is the matrix of time-dummies that one may include in the regression to estimate the λ_t's if they are fixed parameters, and $\lambda' = (\lambda_1, \ldots, \lambda_T)$.
Note that $Z_\lambda Z_\lambda' = J_N \otimes I_T$ and the projection on Z_λ is $Z_\lambda (Z_\lambda' Z_\lambda)^{-1} Z_\lambda' = \bar{J}_N \otimes I_T$. This last matrix averages over individuals, i.e., $(\bar{J}_N \otimes I_T)u$ has a typical element $\bar{u}_{.t} = \sum_{i=1}^{N} u_{it}/N$.

If the μ_i's and λ_t's are assumed to be fixed parameters to be estimated and the remainder disturbances stochastic with $\nu_{it} \sim \text{IID}(0, \sigma_\nu^2)$, then (46) represents a two-way fixed effects model.

- The X_{it}'s are assumed independent of the ν_{it}'s for all i and t.
- Inference in this case is conditional on the particular N individuals and over the specific time-periods observed.
- If N or T is large, there will be too many dummy variables in the regression $(N-1)+(T-1)$ of them, and this causes an enormous loss in degrees of freedom.
- In addition, this attenuates the problem of multicollinearity among the regressors.

- Rather than invert a large $(N+T+K-1)$ matrix, one can obtain the fixed effects estimates of β by performing the following Within transformation given by Wallace and Hussain (1969):

$$Q = E_N \otimes E_T = I_N \otimes I_T - I_N \otimes \bar{J}_T - \bar{J}_N \otimes I_T + \bar{J}_N \otimes \bar{J}_T \tag{48}$$

  where $E_N = I_N - \bar{J}_N$ and $E_T = I_T - \bar{J}_T$.
- This transformation removes the μ_i and λ_t effects.
- $\tilde{u} = Q \cdot u$ has a typical element $\tilde{u}_{it} = (u_{it} - \bar{u}_{i\cdot} - \bar{u}_{\cdot t} + \bar{u}_{\cdot\cdot})$ where $\bar{u}_{\cdot\cdot} = \sum_i \sum_t u_{it}/NT$.
- The LSDV regression

$$(y_{it} - \bar{y}_{i\cdot} - \bar{y}_{\cdot t} + \bar{y}_{\cdot\cdot}) = (x_{it} - \bar{x}_{i\cdot} - \bar{x}_{\cdot t} + \bar{x}_{\cdot\cdot})\beta + (v_{it} - \bar{v}_{i\cdot} - \bar{v}_{\cdot t} + \bar{v}_{\cdot\cdot}) \tag{49}$$

- One can test the presence of fixed effects by a usual F test. Three different hypotheses can be considered.
-

$$y_{it} = \alpha + X_{it}'\beta + \mu_i + \lambda_t + \nu_{it}, \quad i = 1,\ldots,N;\ t = 1,\ldots,T. \tag{50}$$

$$y_{it} = \alpha + X_{it}'\beta + \lambda_t + \nu_{it}, \quad i = 1,\ldots,N;\ t = 1,\ldots,T. \tag{51}$$

$$y_{it} = \alpha + X_{it}'\beta + \mu_i + \nu_{it}, \quad i = 1,\ldots,N;\ t = 1,\ldots,T. \tag{52}$$

$$y_{it} = \alpha + X_{it}'\beta + \nu_{it}, \quad i = 1,\ldots,N;\ t = 1,\ldots,T. \tag{53}$$

1. H_0 : Both the individual effects and the time effects are zero.
   $H_0 : \mu_1 = \ldots = \mu_{N-1} = 0, \lambda_1 = \lambda_{T-1} = 0$. The restricted model is (53).
   $F(N+T-2,(N-1)(T-1)-K)$.
2. H_0 : The individual effects are zero with presence of the time effects.
   $H_0 : \mu_1 = \ldots = \mu_{N-1} = 0$, given $\lambda_t \neq 0$. The restricted model is (51).
   $F(N-1,(N-1)(T-1)-K)$.
3. H_0 : The time effects are zero with presence of the individual effects.
   $H_0 : \lambda_1 = \lambda_{T-1} = 0$, given $\mu_i \neq 0$. The restricted model is (52).
   $F(T-1,(N-1)(T-1)-K)$.

- If $\mu_i \sim \text{IID}(0, \sigma_\mu^2)$, $\lambda_t \sim \text{IID}(0, \sigma_\lambda^2)$, and $v_{it} \sim \text{IID}(0, \sigma_v^2)$ and independent of each other, then this is the two-way random effects model.
- In addition, X_{it} is independent of μ_i, λ_t and v_{it} for all i and t.
- The variance-covariance matrix

$$\begin{aligned} \Omega &= \text{E}(uu') = Z_\mu \text{E}(\mu\mu') Z_\mu' + Z_\lambda \text{E}(\lambda\lambda') Z_\lambda' + \sigma_v^2 I_{NT} && (54) \\ &= \sigma_\mu^2 (I_N \otimes J_T) + \sigma_\lambda^2 (J_N \otimes I_T) + \sigma_v^2 (I_N \otimes I_T). && (55) \end{aligned}$$

- The disturbances

$$\begin{aligned} \text{Cov}(u_{it}, u_{js}) &= \sigma_\mu^2 + \sigma_\lambda^2 + \sigma_v^2 && \text{if } i = j, t = s \\ &= \sigma_\mu^2 && \text{if } i = j, t \neq s \\ &= \sigma_\lambda^2 && \text{if } i \neq j, t = s \\ &= 0 && \text{if } i \neq j, t \neq s \end{aligned}$$

- Similar to the one way model, we can use FGLS and MLE to estimate the model.

- The random time effects are typically difficult to justify.
- In practice for two way models usually we add time dummies (fixed time effects) to the regression and then consider the fixed effects and/or the random effects for the individual effects.
- You can test if the time dummies are significant or not in the regression.

Treatment Effects and Selection Bias

- We observe N units, indexed by $i = 1, \ldots, N$, viewed as drawn randomly from a large population.
- $D_i = 0$ if unit i does not receive the treatment; $D_i = 1$ if unit i receives the treatment.
- The **potential** outcomes

$$\begin{cases} Y_{0i}, & \text{if } D_i = 0; \\ Y_{1i}, & \text{if } D_i = 1. \end{cases}$$

  $Y_{1i} - Y_{0i}$ is the unit-level causal effect (which may be heterogeneous).
- Covariates X_i (not affected by treatment). But let's forget about X_i for a minute.
- The **observed** outcome

$$Y_i = Y_i(D_i) = \begin{cases} Y_{0i}, & \text{if } D_i = 0; \\ Y_{1i}, & \text{if } D_i = 1. \end{cases} = Y_{0i} + (Y_{1i} - Y_{0i}) D_i. \tag{56}$$

- But we never observe both potential outcomes for any one person.
- A naive comparison of observed difference is not right.
-

$$\underbrace{\mathrm{E}(Y_i|D_i=1)-\mathrm{E}(Y_i|D_i=0)}_{\text{Observed difference in average outcome}} = \underbrace{\mathrm{E}(Y_{1i}|D_i=1)-\mathrm{E}(Y_{0i}|D_i=1)}_{\text{average treatment effect on the treated}} + \underbrace{\mathrm{E}(Y_{0i}|D_i=1)-\mathrm{E}(Y_{0i}|D_i=0)}_{\text{selection bias}}.$$

- The first term on the RHS is the (causal) average treatment effect on the treated. It is the average difference between the outcome of the treated, $\mathrm{E}(Y_{1i}|D_i=1)$, and what would have happened to them had they not been treated, $\mathrm{E}(Y_{0i}|D_i=1)$.
- The observed difference in outcome (the LHS) adds to this causal effect a term called *selection bias*. This term is the difference in average Y_{0i} between those who were treated and were not treated.
- The goal of most empirical economic research is to "correct" selection bias, and therefore to find out the causal effect of a variable like D_i.

- Random assignment of D_i removes the selection bias because random assignment makes D_i independent of potential outcomes.
- *Unconfounded assignment*: The assignment probabilities do not depend on the potential outcomes, or $D_i \perp (Y_{0i}, Y_{1i})|X_i$.
-

$$\begin{aligned} \mathrm{E}(Y_i|D_i=1)-\mathrm{E}(Y_i|D_i=0) &= \mathrm{E}(Y_{1i}|D_i=1)-\mathrm{E}(Y_{0i}|D_i=0) \\ &= \mathrm{E}(Y_{1i}|D_i=1)-\mathrm{E}(Y_{0i}|D_i=1) \\ &= \mathrm{E}(Y_{1i}-Y_{0i}|D_i=1) \\ &= \mathrm{E}(Y_{1i}-Y_{0i}). \end{aligned}$$

Example – The evaluation of government-subsidized training programs:

- These are programs that provide a combination of classroom instruction and on-the-job training for groups of disadvantaged workers.
- The idea is to increase employment and earnings.
- Studies based on non-experimental comparisons of participants and non-participants often show that after training, the trainees earn less than plausible comparison groups (Ashenfelter and Card, 1985; Lalonde 1995).
- Selection bias is a concern since subsidized training programs are meant to serve people with low earnings potential. Therefore simple comparisons of program participants with non-participants often show lower earnings for the participants.
- However evidence from randomized evaluations of training programs show positive effects (Lalonde, 1986; Orr, et al. 1996).
- field experiments ...

Regression Analysis:

- Assume that the treatment effect is the same for everyone, i.e, $Y_{1i} - Y_{0i} = \rho$, not ρ_i.
- Rewrite equation (56) as

$$\begin{array}{lccc} Y_i = & \alpha & + & \rho \ D_i \ + & e_i, \\ & \| & & \| & \| \\ & \mathrm{E}(Y_{0i}) & & (Y_{1i} - Y_{0i}) & (Y_{0i} - \mathrm{E}(Y_{0i})) \end{array} \tag{57}$$

- which implies

$$\mathrm{E}(Y_i | D_i = 1) = \alpha + \rho + \mathrm{E}(e_i | D_i = 1)$$

and

$$\mathrm{E}(Y_i | D_i = 0) = \alpha + \mathrm{E}(e_i | D_i = 0).$$

- So $\mathrm{E}(Y_i|D_i=1) - \mathrm{E}(Y_i|D_i=0) = \rho + (\mathrm{E}(e_i|D_i=1) - \mathrm{E}(e_i|D_i=0))$, the treatment effect plus the selection bias.
- The selection bias, $\mathrm{E}(e_i|D_i=1) - \mathrm{E}(e_i|D_i=0) = \mathrm{E}(Y_{0i}|D_i=1) - \mathrm{E}(Y_{0i}|D_i=0)$, is the difference in no treatment outcomes between treated and control.
- We can add covariates $Y_i = \alpha + \rho D_i + X_i'\beta + e_i$.

- Propensity score: the conditional probability of receiving the treatment. $e(x) = \Pr(D_i = 1 | X_i = x) = \mathrm{E}(D_i | X_i = x)$.
- Population average treatment effect (PATE): $\mathrm{E}(Y_{1i}) - \mathrm{E}(Y_{0i})$.
- Population average treatment effect on the treated (PATT): $\mathrm{E}(Y_{1i} - Y_{0i} | D = 1)$.
- Sample average treatment effect (SATE): $\frac{1}{N}\sum_{i=1}^{N}(Y_{1i} - Y_{0i})$.
- Sample average treatment effect on the treated (SATT) ...

Estimations:

- Regression estimators.
- Matching estimators.
- Propensity estimators.
- Mixed estimators.

When allow for heterogenous treatment effects, we have the local treatment effects. LATE, LATT, ...

- One of the oldest questions in labor economics is the connection between union membership and wages.
- Do workers whose wages are set by collective bargaining earn more because of union, or would they earn more anyway? (Perhaps because they are more experienced or skilled).
- Let Y_{it} equal the (log) earnings of worker i at time t and let D_{it} denote his union status.
- The observed Y_{it} is either Y_{0it} or Y_{1it}, depending on union status.
- Suppose further that

$$\mathrm{E}(Y_{0it}|A_i, X_{it}, t, D_{it}) = \mathrm{E}(Y_{0it}|A_i, X_{it}, t),$$

i.e., union status is as good as randomly assigned conditional on unobserved worker ability, A_i, and other observed covariates X_{it}, like age and schooling.

- The key to FE estimation is the assumption that the unobserved A_i appears without a time subscript in a linear model for $\mathrm{E}(Y_{0it}|A_i, X_{it}, t)$:

$$\mathrm{E}(Y_{0it}|A_i, X_{it}, t) = \alpha + \lambda_t + A_i'\gamma + X_{it}'\delta. \tag{58}$$

- Finally, we assume that the causal effect of union membership is additive and constant:

$$\mathrm{E}(Y_{1it}|A_i, X_{it}, t) = \mathrm{E}(Y_{0it}|A_i, X_{it}, t) + \rho. \tag{59}$$

  ρ_i would allow for **local** treatment effect.
- So

$$\mathrm{E}(Y_{it}|A_i, X_{it}, t, D_{it}) = \alpha + \lambda_t + \rho D_{it} + A_i'\gamma + X_{it}'\delta. \tag{60}$$

- Equation (60) implies

$$Y_{it} = \alpha_i + \lambda_t + \rho D_{it} + X_{it}'\delta + \nu_{it}, \tag{61}$$

  where $\alpha_i = \alpha + A_i'\gamma$.
- This is the FE model, which can be estimated by within or first difference.

Table: Union on (log) wage from Freeman (1984)

Data	CS	FE
May CPS, 1974-75	0.19	0.09
National Longitudinal Survey of Young Men, 1970-78	0.28	0.19
Michigan PSID, 1970-79	0.23	0.14
QES, 1973-77	0.14	0.16

- Freeman (1984) uses four data sets to estimate union wage effects under the assumption that selection into union status is based on unobserved-but-fixed individual characteristics.
- The cross section estimates are typically higher than the FE.
- This may indicate positive selection bias in the cross-section estimates, but not the only explanation for the lower FE estimates.

- Although they control for a certain type of omitted variable, FE estimates are notoriously susceptible to attenuation bias from measurement error.
- On one hand, economic variables like union status tend to be persistent (a worker who is a union member this year is most likely a union member next year).
- On the other hand, measurement error often changes from year-to-year (union status may be misreported or miscoded this year but not next year).
- Therefore, while union status may be misreported or miscoded for only a few workers in any single year, the observed year-to-year changes in union status may be mostly noise.
- In other words, there is more measurement error in the regressors in within or first-difference than in the levels of the regressors.
- This fact may account for smaller FE estimates.

- A variant on the measurement-error problem arises from that fact that the differencing and deviations-from-means estimators used to control for FE typically remove both good and bad variation.
- An example is the use of twins to estimate the causal effect of schooling on wages.
- Although there is no time dimension to this problem, the basic idea is the same as the union problem discussed above: twins have similar but largely unobserved family and genetic backgrounds.
- We can therefore include a family FE.
- Ashenfelter and Krueger (1994) and Ashenfelter and Rouse (1998) estimate the returns to schooling using samples of twins, controlling for family FE.
- Surprisingly, the with-family estimates are larger than OLS.
- Bound and Solon (1999) point out that there are small differences between twins, with first-borns typically having higher birth weight and higher IQ scores (here differences in birth timing are measured in minutes).
- While these within-twin differences are not large, neither is the difference in their schooling.

- What should be done about measurement error and related problems in FE?
- A possible solution for measurement error is instrumental variables.
- Ashenfelter and Krueger (1994) use cross-sibling reports to construct instruments for schooling differences across twins.
- A second approach is to bring in external information on the extent of measurement error and adjust naive estimates accordingly.
- In a study of union wage effects, Card (1996) uses external information from a separate validation survey to adjust panel-data estimates for measurement error in reported union status.
- But data from multiple reports and repeated measures of the sort used by Ashenfelter and Rouse (1994) and Card (1996) are unusual.
- At a minimum, therefore, it is important to avoid overly strong claims when interpreting FE estimates.

- Often the regressor of interest varies only at a more aggregate level such as state or cohort. For example, state policies regarding health care benefits for pregnant workers or minimum wages change across states but not within states.
- The source of omitted variables bias when evaluating these policies must therefore be unobserved variables at the state and year level. Consider the following example.
- In a competitive labor market, increases in the minimum wage move up a downward-sloping demand curve. Higher minimums therefore reduce employment, perhaps hurting the very workers minimum-wage policies were designed to help.
- Card and Krueger (1994) use a dramatic change in the New Jersey state minimum wage to see if this is true.

- On April 1, 1992, New Jersey raised the state minimum from \$4.25 to \$5.05.
- Card and Krueger collected data on employment at fast food restaurants in New Jersey in February 1992 and again in November 1992.
- Card and Krueger collected data from the same type of restaurants in eastern Pennsylvania, just across the Delaware river.
- The minimum wage in Pennsylvania stayed at \$4.25 throughout this period.
- They used their data set to compute DID estimates of the effects of the New Jersey minimum wage increase. That is, they compared the change in employment in New Jersey to the change in employment in Pennsylvania around the time New Jersey raised its minimum.
- DID is a version of FE estimation using aggregate data.

- Y_{1ist} = fast food employment at restaurant i and period t in state s if there is a high state minimum wage.
- Y_{0ist} = fast food employment at restaurant i and period t in state s if there is a low state minimum wage.
- These are potential outcomes – in practice, we only get to see one or the other.
- The heart of the DID is an additive structure for potential outcomes in the no-treatment state. Specifically, we assume that

$$\mathrm{E}(Y_{0ist}|s,t) = \gamma_s + \lambda_t. \tag{62}$$

- Let D_{st} be a dummy for high-minimum-wage states.

- Assuming that $\mathrm{E}(Y_{1ist} - Y_{0ist}|s,t)$ is a constant (β), we have:

$$Y_{ist} = \gamma_s + \lambda_t + \beta D_{st} + \varepsilon_{ist} \tag{63}$$

where $\mathrm{E}(\varepsilon_{ist}) = 0$.

- So

$$\begin{aligned}
\mathrm{E}(Y_{ist}|s=PA, t=Nov) - \mathrm{E}(Y_{ist}|s=PA, t=Feb) &= \lambda_{Nov} - \lambda_{Feb}, \\
\mathrm{E}(Y_{ist}|s=NJ, t=Nov) - \mathrm{E}(Y_{ist}|s=NJ, t=Feb) &= \lambda_{Nov} - \lambda_{Feb} + \beta.
\end{aligned}$$

- The population difference-in-differences,

$$\begin{aligned}
&[\mathrm{E}(Y_{ist}|s=NJ, t=Nov) - \mathrm{E}(Y_{ist}|s=NJ, t=Feb)] \\
&-[\mathrm{E}(Y_{ist}|s=PA, t=Nov) - \mathrm{E}(Y_{ist}|s=PA, t=Feb)] \\
&= \beta,
\end{aligned}$$

is the causal effect of the policy.

Table: NJ Minimum Wage Increase from Card and Krueger (1994)

		PA (i)	NJ (ii)	$NJ - PA$ (iii)
1.	FTE employment before	23.33	20.44	−2.89
		(1.35)	(0.51)	(1.44)
2.	FTE employment after	21.17	21.03	−0.14
		(0.94)	(0.52)	(1.07)
3.	Change in mean FTE	−2.16	0.59	2.76
		(1.25)	(0.54)	(1.36)

- Employment in Pennsylvania falls by November.
- Employment in New Jersey increases slightly.
- These two changes produce a positive difference-in-differences, the opposite of what we might expect if a higher minimum wage pushes businesses up the labor demand curve.

- How convincing is this evidence against the standard labor-demand story?
- The key identifying assumption here is that employment trends would be the same in both states in the absence of treatment. Treatment induces a deviation from this common trend.
- Although the treatment and control states can differ, this difference in captured by the state fixed effect.
- Card and Krueger (2000) obtained administrative payroll data for restaurants in New Jersey and Pennsylvania for a number of years.
- Pennsylvania may not provide a very good measure of counterfactual employment rates in New Jersey in the absence of a policy change, and vice versa.

Regression DID

- We can run regressions to estimate DID.
- Let NJ_s be a dummy for restaurants in New Jersey and d_t be a time-dummy that switches on for observations obtained in November (after the increase):

$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \beta (NJ_s \cdot d_t) + \varepsilon_{ist} \tag{64}$$

- β, the coefficient for the interaction term, is the treatment effect.
- This can be easily extended to more states and more periods.
- Another advantage of the regression framework is it allows continuous treatment, i.e., the intensity of the treatment can be continuous, not just 0/1.
- A third advantage is one can easily add other covariates in the regression.
- It is important to pick the control group in DID.

We can introduce the heteroskedasticity either through μ_i or through v_{it}.
Case 1: $\mu_i \sim (0, w_i^2)$**,** $v_{it} \sim$ **IID**$(0, \sigma_v^2)$ **for** $i = 1, ..., N$

$$E(uu') = \Omega = \text{diag}(w_i^2) \otimes J_T + \text{diag}(\sigma_v^2) \otimes I_T \tag{65}$$

We can use technique similar to the RE discussed earlier.
Note that

$$\text{Var}(u_{it}) = \sigma_i^2 = w_i^2 + \sigma_v^2 \tag{66}$$

Follow these steps to estimate the model:

1. Run the within regression to get $\hat{\sigma}_v^2$.
2. Run the OLS regression to get $\hat{u}_{it}$. Estimate

$$\hat{\sigma}_i^2 = \sum_{t=1}^{T} \frac{(\hat{u}_{it} - \bar{\hat{u}}_{i\cdot})^2}{T-1} \quad \text{for } i = 1, ..., N \tag{67}$$

3. Obtain $\hat{w}_i^2 = \hat{\sigma}_i^2 - \hat{\sigma}_v^2$.
4. Obtain $\hat{\tau}_i^2 = T\hat{w}_i^2 + \hat{\sigma}_v^2$, $\hat{\theta}_i = 1 - (\hat{\sigma}_v / \hat{\tau}_i)$.
5. Transform every variable including the intercept

$$\hat{y}_{it}^* = y_{it} - \hat{\theta}_i \bar{y}_{i\cdot} \tag{68}$$

6. Run the transformed regression (without constant).

This model requires large T and small N, which is not the case for most labor applications.

Case 2: $\mu_i \sim$ **IID**$(0, \sigma_\mu^2)$**,** $v_{it} \sim (0, w_i^2)$ **for** $i = 1, ..., N$

In this case

$$E(uu') = \Omega = \text{diag}(\sigma_\mu^2) \otimes J_T + \text{diag}(w_i^2) \otimes I_T \tag{69}$$

Note that

$$\text{Var}(u_{it}) = \sigma_i^2 = \sigma_\mu^2 + w_i^2 \tag{70}$$

Follow these steps to estimate the model:

1. Run the within regression to get $\tilde{u}_{it}$. Estimate

$$\hat{w}_i^2 = \sum_{t=1}^{T} \frac{(\tilde{u}_{it} - \bar{\tilde{u}}_{i\cdot})^2}{T-1} \quad \text{for } i = 1, ..., N \tag{71}$$

2. Run the OLS regression to get $\hat{u}_{it}$. Estimate

$$\hat{\sigma}_i^2 = \sum_{t=1}^{T} \frac{(\hat{u}_{it} - \bar{\hat{u}}_{i\cdot})^2}{T-1} \quad \text{for } i = 1, ..., N \tag{72}$$

3. Here you have N estimates of σ_μ^2. Simply get the average $\hat{\sigma}_\mu^2 = \sum_{i=1}^{N} (\hat{\sigma}_i^2 - \hat{w}_i^2)/N$.
4. Obtain $\hat{\tau}_i^2 = T\hat{\sigma}_\mu^2 + \hat{w}_i^2$, $\hat{\theta}_i = 1 - (\hat{w}_i/\hat{\tau}_i)$.
5. Transform every variable including the intercept

- In practice we do not like the feasible GLS style estimations in Case 1 and Case 2.
- Instead, we would like to use some sort of heteroskedasticity-robust standard errors which do not specify the heteroskedastic form.

- Consider the AR(1) process in the error term v:

$$v_{it} = \rho v_{i,t-1} + \epsilon_{it} \tag{74}$$

where $|\rho| < 1$ and $\epsilon_{it} \sim \text{IID}(0, \sigma_\epsilon^2)$. $\text{Var}(v_{it}) = \sigma_\epsilon^2/(1-\rho^2)$.

- In a pure time series setup, we would use the following Prais-Winsten transformation:

$$y_{it}^* = \begin{cases} \sqrt{1-\rho^2}\, y_{it}, & \text{if } t = 1; \\ y_{it} - \rho y_{i,t-1}, & \text{if } t > 1. \end{cases}$$

- Similar to the Prais-Winsten method in pure time series model (If we omit the first observation, it would be the Cochrane-Orcutt procedure):

Here goes the panel procedure:

1. Run the within estimation to obtain $\tilde{v}_{it}$. Estimate $\hat{\rho} = \sum_i \sum_t \tilde{v}_{it}\tilde{v}_{i,t-1} / \sum_i \sum_t \tilde{v}_{i,t-1}^2$. If you want to allow different ρ for different individual, you can use the individual formula.
2. ...

- In Stata command `xtregar` can estimate the AR(1) model.
- It is also possible to consider the situation where ρ is different for each i, though it is not very popular in practice.

- Remember that in RE models OLS estimators are still unbiased and consistent. So the only issue is to correct the standard errors.
- In dynamic models typically the robust variance is constructed in GMM fashion.
- In EViews one can use different combinations of individual and time effects, together with various "robust" variances: Ordinary, White cross-sectional, White (diagonal) method, cross-sectional weight (PCSE), period weight (PCSE), White period, cross-sectional SUR (PCSE), and period SUR (PCSE).

In LIMDEP you have different combinations of heteroskedasticity and autocorrelation.
Define:

1. groupwise heteroskedasticity: $E(\varepsilon_{it}^2 = \sigma_{ii})$.
2. cross group correlation: $Cov(\varepsilon_{it}, \varepsilon_{jt}) = \sigma_{ij}$.
3. within group autocorrelation: $\varepsilon_{it} = \rho \varepsilon_{i,t-1} + u_{i,t-1}$.

Table: LIMDEP 9.0 Robust Variances

Heteroskedasticity	
S0	homoskedastic and uncorrelated (OLS std err)
S1	groupwise heteroskedasticity
S2	groupwise heteroskedasticity and cross group correlation
Autocorrelation	
R0	no autocorrelation
R1	common autoregressive coefficient, ρ
R2	group specific autoregressive coefficient, ρ_i

LIMDEP can estimate models with nine combinations of the above models.

Stata:

- xtpcse can produce the nine combinations in LIMDEP discussed above.
  correlation(independent, ar1, psar1).
  hetonly, independent, and none (default = both heteroskedasticity and cross group correlation).
- xtreg allows corr(independent, ar1, psar1) option.
- xtgls allows nine combinations.
- xtgee is a version of Generalized Estimating Equations (GEE) for panel:

$$g\{\mathrm{E}(y_{it})\} = \mathbf{x}_{it}'\beta, \text{ where } y \sim F \text{ with parameters } \theta_{it}. \tag{75}$$

  $g()$ is the called the link function and F the distributional family.
  The link function can be cloglog, identity, log, logit, negative binomial, odds power, power, probit, and reciprocal.
  The distributional family can be Bernoulli/binomial, gamma, normal/Gaussian, inverse Gaussian, negative binomial, and Poisson.
  xtgee allows the within-group correlation structure to be independent, exchangeable, autoregressive, stationary, non-stationary, unstructured, and user-specified.

- If one or more of the right-hand-side variables are correlated with the error terms, there is endogeneity in the equation. The usual least square method does not work because of this endogeneity.
- You have two choices: the system method (3SLS, GMM, and FIML et al.) or the single equation method (IV, 2SLS, GMM, and LIML et al.).
- We will discuss the single equation method here. The system approach, which is not popular in recent applied research, can be found in the textbook.
- Note that in the first section the regressors are correlated with the error term v_{it}. In the second section the regressors are correlated with the individual random effects μ_i.
- In the first section you have FE and RE models. In the second section we will only discuss the RE model since in FE this would not be a problem.

- By endogeneity we mean the correlation of the right hand side regressors and the disturbances.
- This may be due to the omission of relevant variables, measurement error, sample selectivity, self-selection or other reasons.
- Endogeneity causes inconsistency of the usual OLS estimates and typically requires instrumental variable methods like 2SLS/GMM to obtain consistent parameter estimates.
- Assume you are familiar with the identification and estimation of a single equation and a system of simultaneous equations.

- Consider the following first structural equation of a simultaneous equation model

$$y = Z\delta + u \tag{76}$$

where $Z = [Y, X_1]$ and $\delta' = (\gamma', \beta')$. As in the standard simultaneous equation literature, Y is the set of g RHS endogenous variables, and X_1 is the set of k_1 included exogenous variables. Let $X = [X_1, X_2]$ be the set of all exogenous variables in the system. This equation is identified with $k_2 \geq g$.

- We will focus on the one-way error component model

$$u = Z_\mu \mu + \nu \tag{77}$$

where $Z_\mu = (I_N \otimes \iota_T)$ and $\mu' = (\mu_1, \ldots, \mu_N)$ and $\nu' = (\nu_{11}, .., \nu_{NT})$ are random vectors with zero means and covariance matrix

$$E\begin{pmatrix} \mu \\ \nu \end{pmatrix}(\mu', \nu') = \begin{bmatrix} \sigma_\mu^2 I_N & 0 \\ 0 & \sigma_\nu^2 I_{NT} \end{bmatrix}. \tag{78}$$

- One can transform (76) by $Q = I_{NT} - P$ with $P = I_N \otimes \bar{J}_T$, to get

$$Qy = QZ\delta + Qu. \tag{79}$$

- Let $\tilde{y} = Qy$ and $\tilde{Z} = QZ$. Performing 2SLS on (79) with $\tilde{X} = QX$ as the set of instruments, one gets the Within 2SLS

$$\tilde{\delta}_{W2SLS} = (\tilde{Z}' P_{\tilde{X}} \tilde{Z})^{-1} \tilde{Z}' P_{\tilde{X}} \tilde{y}. \tag{80}$$

- Similarly, if we let $\bar{y} = Py$ and $\bar{Z} = PZ$, we can transform (76) by P and perform 2SLS with $\bar{X} = PX$ as the set of instruments. In this case, we get the Between 2SLS estimator of δ

$$\hat{\delta}_{B2SLS} = (\bar{Z}' P_{\bar{X}} \bar{Z})^{-1} \bar{Z}' P_{\bar{X}} \bar{y} \tag{81}$$

- The Error Component Two Stage Least Squares (EC2SLS) estimator of δ:

$$\hat{\delta}_{EC2SLS} = \left[\frac{\tilde{Z}' P_{\tilde{X}} \tilde{Z}}{\sigma_v^2} + \frac{\bar{Z}' P_{\bar{X}} \bar{Z}}{\sigma_1^2} \right]^{-1} \left[\frac{\tilde{Z}' P_{\tilde{X}} \tilde{y}}{\sigma_v^2} + \frac{\bar{Z}' P_{\bar{X}} \bar{y}}{\sigma_1^2} \right] \tag{82}$$

  which is a weighted average of $\tilde{\delta}_{W2SLS}$ and $\hat{\delta}_{EC2SLS}$, can be derived from GLS on

$$\begin{pmatrix} \tilde{X}' \tilde{y} \\ \bar{X}' \bar{y} \end{pmatrix} = \begin{pmatrix} \tilde{X}' \tilde{Z} \\ \bar{X}' \bar{Z} \end{pmatrix} \delta + \begin{pmatrix} \tilde{X}' \tilde{u} \\ \bar{X}' \bar{u} \end{pmatrix}. \tag{83}$$

- The EC2SLS is just the typical 2SLS with a more complicated Ω, the variance-covariance matrix for the error term.
- In Stata the command `xtivreg` can produce the above estimators.

- Endogeneity through the unobserved individual effects.
- Examples where μ_i and the explanatory variables may be correlated: an earnings equation where the unobserved individual ability may be correlated with schooling and experience.

Motivation: not exactly a RE model

- Mundlak (1978) considered the one-way error component regression model with the additional auxiliary regression:

$$\mu_i = \bar{X}_{i.}'\pi + \epsilon_i \tag{84}$$

where $\epsilon_i \sim IIN(0, \sigma_\epsilon^2)$.

- In other words, Mundlak assumed that the individual effects are a linear function of the averages of *all* the explanatory variables across time. These effects are uncorrelated with the explanatory variables if and only if $\pi = 0$.
- Mundlak (1978) assumed, without loss of generality, that the X's are deviations from their sample mean.
- In vector form, one can write (84) as

$$\mu = Z_\mu' X\pi/T + \epsilon \tag{85}$$

where $\mu' = (\mu_1, .., \mu_N)$, $Z_\mu = I_N \otimes \iota_T$ and $\epsilon' = (\epsilon_1, ..., \epsilon_N)$.

- We can get

$$y = X\beta + PX\pi + (Z_\mu \epsilon + \nu) \tag{86}$$

where $P = I_N \otimes J_T$. Using the fact that the ϵ's and the ν's are uncorrelated, the new error in (86) has zero mean and variance covariance matrix

$$V = E(Z_\mu \epsilon + \nu)(Z_\mu \epsilon + \nu)' = \sigma_\epsilon^2 (I_N \otimes J_T) + \sigma_\nu^2 I_{NT} \tag{87}$$

- Using partitioned inverse, one can verify that GLS on (86) yields

$$\hat{\beta}_{GLS} = \tilde{\beta}_{Within} = (X'QX)^{-1} X'Qy \tag{88}$$

and

$$\hat{\pi}_{GLS} = \hat{\beta}_{Between} - \tilde{\beta}_{Within} = (X'PX)^{-1} X'Py - (X'QX)^{-1} X'Qy. \tag{89}$$

- Therefore, Mundlak (1978) showed that the BLUE estimator becomes the fixed effects Within estimator once these fixed effects are modeled as a linear function of all the X_{it}'s. The random effects estimator on the other hand is biased because it ignores the relationship.
- Note that Hausman's test based on the between minus Within estimators is basically a test for $H_0 : \pi = 0$.
- Mundlak's (1978) formulation assumes that *all* the explanatory variables are related to theindividual effects. The random effects model on the other hand assumes no correlation between the explanatory variables and the individual effects. The random effects model generates the GLS estimator, whereas Mundlak's formulation produces the within estimator.
- Instead of this 'all or nothing' correlation among the X's and the μ_i's, Hausman and Taylor (1981) consider a model where *some* of the explanatory variables are related to the μ_i's.

- Hausman and Taylor consider the following model:

$$y_{it} = X_{it}'\beta + Z_i'\gamma + \mu_i + \nu_{it} \tag{90}$$

  where the Z_i's are cross-sectional time-invariant variables.
- Hausman and Taylor split X and Z into two sets of variables: $X = [X_1, X_2]$ and $Z = [Z_1, Z_2]$ where X_1 is $n \times k_1$, X_2 is $n \times k_2$, Z_1 is $n \times g_1$, Z_2 is $n \times g_2$ and $n = NT$.
- X_1 and Z_1 are assumed exogenous in the sense that they are not correlated with μ_i and ν_{it} while X_2 and Z_2 are endogenous because they are correlated with the μ_i's, but not with the ν_{it}'s.
- The within transformation would sweep the μ_i's and remove the bias, but in the process it would also remove the Z_i's and hence the within estimator will not give an estimate of the γ's. To get around that, Hausman and Taylor suggest pre-multiplying the model by $\Omega^{-1/2}$ and using the following set of instruments: $A_0 = [Q, X_1, Z_1]$, where $Q = I - P$ and $P = (I_N \otimes \bar{J}_T)$.

- Breusch, Mizon, and Schmidt (1989), hereafter BMS, show that this set of instruments yields the same projection and is therefore equivalent to another set namely, $A_1 = [QX_1, QX_2, Z_1, PX_1]$. The latter set of instruments A_1 is feasible, whereas A_0 is not because it is $NT \times NT$.
- $A_1 = [QX_1, QX_2, Z_1, PX_1]$ are instrumental variables for $[X_1, X_2, Z_1, Z_2]$. Why QX_2 can be instrument for X_2? Because $\text{Cov}(QX_2, Z_\mu \mu) = 0$ though $\text{Cov}(X_2, \mu) \neq 0$.
- The order condition for identification gives the result that the number of X_1's (k_1) must be at least as large as the number of Z_2's (g_2). X_1 is used twice, once as averages and another time as deviations from averages. This is an advantage of panel data allowing instruments from *within* the model.
- Note that the within transformation wipes out the Z_i's and does not allow the estimation of the γ's.

Hausman-Taylor procedure:

1. In order to get consistent estimates of the γ's, HT propose obtaining the within residuals and averaging them over time

$$\hat{d}_i = \bar{y}_{i.} - \bar{X}'_{i.}\tilde{\beta}_W \tag{91}$$

2. Then, running 2SLS of $\hat{d}_i$ on $Z = [Z_1, Z_2]$ with the set of instruments $A = [Z_1, X_1]$ yields

$$\hat{\gamma}_{2SLS} = (Z' P_A Z)^{-1} Z' P_A \hat{d} \tag{92}$$

where $P_A = A(A'A)^{-1}A'$. It is clear that the order condition has to hold $(k_1 \geq g_2)$ for $(Z' P_A Z)$ to be non-singular.

3. Next, the variance-components estimates are obtained as follows:

$$\tilde{\sigma}^2_v = \tilde{y}' \bar{P}_{\tilde{X}} \tilde{y} / N(T-1) \tag{93}$$

where $\tilde{y} = Qy$, $\tilde{X} = QX$, $\bar{P}_A = I - P_A$ and

$$\tilde{\sigma}^2_1 = \frac{(y_{it} - X_{it}\tilde{\beta}_W - Z_i\hat{\gamma}_{2SLS})' P (y_{it} - X_{it}\tilde{\beta}_W - Z_i\hat{\gamma}_{2SLS})}{N} \tag{94}$$

This last estimate is based upon an NT vector of residuals.

4. Once the variance components estimates are obtained, the model is transformed using $\hat{\Omega}^{-\frac{1}{2}}$ as follows:

$$\hat{\Omega}^{-\frac{1}{2}} y = \hat{\Omega}^{-\frac{1}{2}} X\beta + \hat{\Omega}^{-\frac{1}{2}} Z\gamma + \hat{\Omega}^{-\frac{1}{2}} u \tag{95}$$

The HT estimator is basically 2SLS on the above regression using $A_{HT} = [\tilde{X}_1, \tilde{X}_2, Z_1, \bar{X}_1]$ as a set of instruments.

Comments:

1. If $k_1 < g_2$, then the equation is under-identified. In this case $\hat{\beta}_{HT} = \tilde{\beta}_W$ and $\hat{\gamma}_{HT}$ does not exist.

2. If $k_1 = g_2$, then the equation is just-identified. In this case, $\hat{\beta}_{HT} = \tilde{\beta}_W$ and $\hat{\gamma}_{HT} = \hat{\gamma}_{2SLS}$ given by (92).

3. If $k_1 > g_2$, then the equation is over identified and the HT estimator obtained from (95) is more efficient than the within estimator.

- Amemiya and MaCurdy (1986), hereafter AM, suggest a more efficient set of instruments $A_2 = [QX_1, QX_2, X_1^*, Z_1]$ where $X_1^* = X_1^0 \otimes \iota_T$ and

$$X_1^0 = \begin{bmatrix} X_{11} & X_{12} & .. & X_{1T} \\ \vdots & \vdots & .. & \vdots \\ X_{N1} & X_{N2} & .. & X_{NT} \end{bmatrix} \tag{96}$$

  is an $N \times k_1 T$ matrix.
- So X_1 is used $(T+1)$ times, once as X_1 and T times as X_1^*.
- The order condition for identification is now more likely to be satisfied $(Tk_1 > g_2)$.
- However, this set of instruments require a stronger exogeneity assumption than that of Hausman and Taylor (1981).
- The latter requires only uncorrelatedness of the mean of X_1 from the μ_i's, while Amemiya and MaCurdy (1986) require uncorrelatedness at each point in time.
- Breusch, Mizon, and Schmidt (1989) suggest yet a more efficient set of instruments $A_3 = [QX_1, QX_2, PX_1, (QX_1)^*, (QX_2)^*, Z_1]$.

- In Stata use `xtivreg` for FE & RE IV panel regressions. Options include `fd`, `fe`, `re`, and `re ec2sls`.
- `xthtaylor` can be used for RE Hausman-Taylor procedure. You can request the Amemiya and MaCurdy (1986) procedure using the option `amacurdy`.

- $$y_{it} = \delta y_{i,t-1} + x'_{it}\beta + u_{it} \quad i = 1, ..., N; t = 1, ..., T \tag{97}$$

  where δ is a scalar, x'_{it} is $1 \times K$ and β is $K \times 1$. We will assume that the u_{it}'s follow a one-way error component model

$$u_{it} = \mu_i + \nu_{it}. \tag{98}$$

- Since y_{it} is a function of μ_i, $y_{i,t-1}$ is also a function of μ_i. Therefore, $y_{i,t-1}$, a right hand regressor, is correlated with the error term. This renders the OLS estimator biased and inconsistent even if the ν_{it}'s are not serially correlated.
- For the fixed effects (FE) estimator, the within transformation wipes out the μ_i's, but $(y_{i,t-1} - \bar{y}_{i\cdot-1})$ where $\bar{y}_{i\cdot-1} = \sum_{t=2}^{T} y_{i,t-1}/(T-1)$ will still be correlated with $(\nu_{it} - \bar{\nu}_{i\cdot})$ even if the ν_{it}'s are not serially correlated.
- This is because $y_{i,t-1}$ is correlated with $\bar{\nu}_{i\cdot}$ by construction. The latter average contains $\nu_{i,t-1}$ which is obviously correlated with $y_{i,t-1}$.
- In fact, the within estimator will be biased of $O(1/T)$ and its consistency will depend upon T being large.

- Therefore, for the typical labor panel where N is large and T is fixed, the within estimator is biased and inconsistent.
- Only if $T \to \infty$ will the within estimator of δ and β be consistent for the dynamic error component model.
- The same problem occurs with the random effects GLS estimator. In order to apply GLS, quasi-demeaning is performed, and $(y_{i,t-1} - \theta \bar{y}_{i.,-1})$ will be correlated with $(u_{i,t} - \theta \bar{u}_{i.,-1})$.
- An alternative transformation that wipes out the individual effects, yet does not create the above problem is the first difference (FD) transformation.
- Anderson and Hsiao (1981) suggested first differencing the model to get rid of the μ_i's and then using $\Delta y_{i,t-2} = (y_{i,t-2} - y_{i,t-3})$ or simply $y_{i,t-2}$ as an instrument for $\Delta y_{i,t-1} = (y_{i,t-1} - y_{i,t-2})$.
- These instruments will not be correlated with $\Delta v_{it} = v_{i,t} - v_{i,t-1}$, as long as the v_{it}'s themselves are not serially correlated. This instrumental variable (IV) estimation method leads to consistent but not necessarily efficient estimates of the parameters in the model because it does not make use of all the available moment conditions.

- Arellano (1989) finds that for simple dynamic error components models the estimator that uses differences $\Delta y_{i,t-2}$ rather than levels $y_{i,t-2}$ for instruments has a singularity point and very large variances over a significant range of parameter values.

- In contrast, the estimator that uses instruments in levels, i.e., $y_{i,t-2}$, has no singularities and much smaller variances and is therefore recommended.

- Arellano and Bond (1991) argue that additional instruments can be obtained in a dynamic panel data model if one utilizes the orthogonality conditions that exist between lagged values of y_{it} and the disturbances v_{it}.
- Consider a simple autoregressive model with no regressors:

$$y_{it} = \delta y_{i,t-1} + u_{it} \tag{99}$$

where $u_{it} = \mu_i + v_{it}$ with $\mu_i \sim IID(0, \sigma_\mu^2)$ and $v_{it} \sim IID(0, \sigma_v^2)$, independent of each other.

- In order to get a consistent estimate of δ as $N \to \infty$ with T fixed, we first difference to eliminate the individual effects

$$y_{it} - y_{i,t-1} = \delta(y_{i,t-1} - y_{i,t-2}) + (v_{it} - v_{i,t-1}) \tag{100}$$

and note that $(v_{it} - v_{i,t-1})$ is MA(1) with unit root.

- For the first period we observe this relationship, i.e., $t=3$, we have

$$y_{i3}-y_{i2}=\delta(y_{i2}-y_{i1})+(v_{i3}-v_{i2}) \tag{101}$$

- In this case, y_{i1} is a valid instrument, since it is highly correlated with $(y_{i2}-y_{i1})$ and not correlated with $(v_{i3}-v_{i2})$ as long as the v_{it}'s are not serially correlated.
- For $t=4$, the second period we observe

$$y_{i4}-y_{i3}=\delta(y_{i3}-y_{i2})+(v_{i4}-v_{i3}) \tag{102}$$

- In this case, y_{i2} as well as y_{i1} are valid instruments for $(y_{i3}-y_{i2})$, since both y_{i2} and y_{i1} are not correlated with $(v_{i4}-v_{i3})$.
- One can continue in this fashion, so that for period T, the set of valid instruments becomes $(y_{i1}, y_{i2}, \ldots, y_{i,T-2})$.
- This instrumental variable procedure still does not account for the differenced error term. In fact

$$E(\Delta v \Delta v')=\sigma_v^2(I_N \otimes G) \tag{103}$$

where $\Delta v_i = (v_{i3} - v_{i2}, \ldots, v_{iT} - v_{i,T-1})$ and

$$G = \begin{pmatrix} 2 & -1 & 0 & \ldots & 0 & 0 \\ -1 & 2 & -1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 2 & -1 \\ 0 & 0 & 0 & \ldots & -1 & 2 \end{pmatrix}$$

is $(T-2) \times (T-2)$, since Δv is MA(1) with unit root.

- Define

$$W_i = \begin{bmatrix} [y_{i1}] & & & 0 \\ & [y_{i1}, y_{i2}] & & \\ & & \ddots & \\ 0 & & & [y_{i1}, \ldots, y_{i,T-2}] \end{bmatrix} \tag{104}$$

- Then, the matrix of instruments is $W = [W_1', \ldots, W_N']'$ and the moment equations described above are given by $\mathrm{E}(W_i' \Delta v_i) = 0$.
- Pre-multiplying the differenced equation (100) in vector form by W', one gets

$$W' \Delta y = W'(\Delta y_{-1})\delta + W' \Delta v \tag{105}$$

- Performing GLS on the above regression one gets the Arellano and Bond (1991) one-step consistent estimator

$$\begin{aligned}\hat{\delta}_1 &= [(\Delta y_{-1})' W (W'(I_N \otimes G) W)^{-1} W' (\Delta y_{-1})]^{-1} \\ &\quad [(\Delta y_{-1})' W (W'(I_N \otimes G) W)^{-1} W' (\Delta y)] \end{aligned} \tag{106}$$

- The optimal Generalized Method of Moments (GMM) estimator of δ_1 (Hansen, 1982) for $N \to \infty$ and T fixed using only the above moment restrictions yields the same expression except that $W'(I_N \otimes G)W = \Sigma_{i=1}^N W_i' G W_i$ is replaced by $V_N = \Sigma_{i=1}^N W_i' (\Delta v)(\Delta v)' W_i$.
- This GMM estimator requires no knowledge concerning the initial conditions or the distributions of v_i and μ_i.
- To make this estimator feasible, Δv is replaced by differenced residuals obtained from the preliminary consistent estimator $\hat{\delta}_1$.
- The resulting estimator is the two-step Arellano and Bond (1991) GMM estimator:

$$\hat{\delta}_2 = [(\Delta y_{-1})' W \hat{V}_N^{-1} W' (\Delta y_{-1})]^{-1} [(\Delta y_{-1})' W \hat{V}_N^{-1} W' (\Delta y)] \tag{107}$$

- A consistent estimate of asymptotic $\text{Var}(\hat{\delta}_2)$ is given by the first term in the above equation,

$$\widehat{\text{Var}}(\hat{\delta}_2) = [(\Delta y_{-1})' W \hat{V}_N^{-1} W' (\Delta y_{-1})]^{-1} \tag{108}$$

  Note that $\hat{\delta}_1$ and $\hat{\delta}_2$ are asymptotically equivalent if the v_{it}'s are $IID(0, \sigma_v^2)$.

What if we have exogenous variables?

- If there are additional *strictly exogenous* regressors x_{it}'s with $\text{E}(x_{it} v_{is}) = 0$ for all $t, s = 1, 2, ..., T$, but where *all* the x_{it}'s are correlated with μ_i, then all the x_{it}'s are valid instruments for the first differenced equation. Therefore, $[x_{i1}', x_{i2}', ..., x_{iT}']$ should be added to each diagonal element of W_i in (104):

$$W_i = \begin{bmatrix} [y_{i1}, x_{i1}', x_{i2}', ..., x_{iT}'] & & & 0 \\ & [y_{i1}, y_{i2}, x_{i1}', x_{i2}', ..., x_{iT}'] & & \\ & & \ddots & \\ 0 & & & [y_{i1}, ..., y_{i,T-2}, x_{i1}', x_{i} \end{bmatrix} \tag{109}$$

In this case, (105) becomes

$$W'\Delta y = W'(\Delta y_{-1})\delta + W'(\Delta X)\beta + W'\Delta v \tag{110}$$

where ΔX is the stacked $N(T-2)\times K$ matrix of observations on Δx_{it}. One-step and two-step estimators of (δ, β') can be obtained from

$$\begin{pmatrix} \hat{\delta} \\ \hat{\beta} \end{pmatrix} = ([\Delta y_{-1}\Delta X]' W \hat{V}_N^{-1} W'[\Delta y_{-1}\Delta X])^{-1}([\Delta y_{-1}\Delta X]' W \hat{V}_N^{-1} W'\Delta y). \tag{111}$$

- If x_{it}'s are *predetermined* rather than *strictly exogenous* with $\mathrm{E}(x_{it}v_{is}) \neq 0$ for $s > t$, and zero otherwise, then only $[x_{i1}', x_{i2}', ..., x_{i,(s-1)}]$ are valid instruments for the differenced equation at period s. This can be illustrated as follows: For $t = 3$, the first differenced equation becomes

$$y_{i3} - y_{i2} = \delta(y_{i2} - y_{i1}) + (x_{i3}' - x_{i2}')\beta + (v_{i3} - v_{i2}) \tag{112}$$

For this equation, x_{i1}' and x_{i2}' are valid instruments, since both are not correlated with $(v_{i3} - v_{i2})$. For $t = 4$, the next period we observe this relationship

$$y_{i4} - y_{i3} = \delta(y_{i3} - y_{i2}) + (x_{i4}' - x_{i3}')\beta + (v_{i4} - v_{i3}) \tag{113}$$

and we have additional instruments since now x'_{i1}, x'_{i2} and x'_{i3} are not correlated with $(v_{i4} - v_{i3})$. Continuing in this fashion, we get

$$W_i = \begin{bmatrix} [y_{i1}, x'_{i1}, x'_{i2}] & & 0 & \\ & [y_{i1}, y_{i2}, x'_{i1}, x'_{i2}, x'_{i3}] & & \\ & & \ddots & \\ 0 & & & [y_{i1}, \ldots, y_{i,T-2}, x'_{i1}, \ldots, x'_{i,T-1}] \end{bmatrix}$$

and one and two-step estimators are again given by (110) with this choice of W_i.

- There are further extensions to this model, mainly with additional moment conditions. Such as the system Arellano-Bover and Blundell-Bond estimators.
- In Stata and LIMDEP, the dynamic panel data models are available.
- `xtabond`, `xtdpd`, and `xtdpdsys` in Stata.

- Logit, Probit, Tobit, Count models.
- Within or first difference transformation cannot remove the individual effects.
- Estimation is more complicated. It typically involves numerical integration (sometimes multi-dimensional integration) and numerical optimization.
- The interpretation is similar to that in cross sectional data.
- This is one of the most rapidly developed fields.

Papers and program. GeoDa.

Material here is largely from Levitt and List (2008).
Use experiments to induce necessary variation to test economic theories and eliminate unwanted sources of variation that confound interpretation.
This is not to say that there aren't problems in experiments:

- Randomization bias: the experimental sample is different from the population of interest because of randomization.
- Attrition bias: there are systematic differences between the treatment and control groups because of differential losses of participants.
- Generating misleading inference out of sample due to the increased scrutiny in the experiment. The John Henry effect and the Hawthorn effect. Subjects (in the control group) adopt a competitive attitude toward the experimental group, thereby negating their status as controls. A short-term improvement caused by realizing being observed.
- Substitution bias: control group members seek available substitutes for treatment.

The British electricity pricing 1966–1972:

- The experiment included six Area Boards in UK, which included 3,420 residential customers who purchased 3,000+ kWh yearly.
- The experiment divided customers into four pricing schemes: i) Seasonal: 150% of normal rate for Dec.-Feb.; 70% of normal for the rest of the year; ii) Seasonal Time-of-Day: 300% of normal rate for 8:00-13:00 and 16:30-19:30 from Dec.-Feb.; 40% of normal otherwise; iii) Load: Subjects set a target yearly total, receiving a standard rate for that total and paying 60% of the standard rate until the target was reached and 100-200% thereafter; iv) Control: Subjects received block rates, price falling toward a final rate as consumption increased.
- All treatment schemes were found to increase the annual energy sold, though the difference between the Load and Control schemes were not statistically significant. The seasonal scheme, together with restricted hour rates, was the most effective in increasing daytime energy sold, while the Seasonal Time-of-Day scheme was the most effective at diverting consumption away from peak times.

Ausubel (1999) direct mail solicitations of credit card applications, adverse selection, level and duration of a teaser introductory interest rate. Randomization.

- The less attractive credit card offers attract customers with inferior observable characteristics, as measured by income and past credit histories. This is consistent with economic theory, since these are the consumers with the worst outside options.
- Even more interesting from an economic perspective is the strong evidence of adverse selection on unobservable dimensions. Even controlling for detailed information that the credit card issuer knows about the consumers at the time of the solicitation, customers responding to the inferior offers are far more likely to subsequently default.

Karlan and Zinman (2007) pursue a similar question using a South African lender's direct mailing.

- They randomize the interest rate that consumers receive in their mail solicitation, and the rate the consumer will be charged for their next loan if he or she successfully pays off the first loan.
- They incorporate an additional element: half of the consumers who respond to the initial offer are randomized into receiving a lower interest rate.
- This two-step determination of interest rates help distinguish a moral hazard effect of higher interest rates (the higher rate makes it more difficult for a given consumer to pay back) from an adverse selection effect (the consumers who accept higher interest rates are drawn from a pool that is less likely to pay back).

- moral hazard: outcomes for consumers who responded to the high interest rate offer and received that rate versus consumers who responded to the high interest rate but were ex post randomized into receiving a lower interest rate.
- adverse selection: outcomes for the consumers who responded to a low interest rate offer relative to those who responded to the high interest rate offer, but were randomized ex post into receiving the low interest rate.
- In a follow-up paper they find that consumers are more responsive to loan maturity than to interest rate, which is consistent with the borrowers being liquidity constrained.

Anderson and Simester (2003) \$9 endings on prices.

- A retail catalog merchant.
- Randomly selected customers receive one of three catalog versions showing different prices for the same product. For example, a dress may be offered to all consumers, but at prices of \$34, \$39, and \$44 in each catalog version.
- They find a positive effect of a price ending in \$9 on quantity demanded, large enough that a price of \$39 actually produced higher quantities than a price of \$34.

Ashraf, Berry, and Shapiro (2007): price on demand and utilization.

- door-to-door salespeople in Zambia.
- Clorin, a product used to purify water in the home.
- Two-step process: A consumer is quoted a randomly determined price for Clorin by a salesperson. Among those who agree to purchase the product at that price, some are randomly allowed to purchase the good at a lower price.
- Two weeks after the purchase, a follow-up survey was done to ask the consumer about their use of the product, and the household's water supply was tested chemically.
- The quantity falls with price.
- Those who are willing to pay more appear to value the good more highly – higher utilization rate after purchase.
- In general, they do not find much difference in utilization between consumers who are willing to pay a high price and are charged that high price versus consumers willing to pay a high price, but who are subsequently randomized into receiving a lower price.
- Those who are given Clorin for free may be less likely to use it than those who are required to pay a positive price.

Gneezy and Rustichini (2000) fines in a day care.

- After observing the frequency with which parents arrived late to pick up their children for four weeks, a small monetary fine is introduced at random to a subset of the day-care centers.
- The result was an **increase** in the number of late-arriving parents, and even after the fine was removed, late arrivals did not return to their original levels.
- Simple deterrence theory would predict that adding a monetary fine on top of any informal sanctions would reduce rather than increase late pick ups.
- Charging a small fine weakens the social sanctions. Once late arrivals are priced, there is less need to feel guilty about being late since the day care provider is compensated, presumably at a level commensurate with the day care provider's loss since it is the provider that set the price.

Rand Health Insurance Experiment.