Econometrics

These slides follow very closely
William H. Green, Econometric Analysis, 6th Edition
All graphics and formulae are taken out of this book.

Content

Introduction I

- Panel data (longitudinal data) refers to a cross-section repeatedly sampled over time, but where the same individual has been followed throughout the period of the sample.
  - Individuals: person, household, plant, firm, municipality, state or a country.
  - Time: five years intervals, annual, quarters, weeks, days or just an observation time.
  - We cannot assume that the observations are independently distributed across time (e.g. unobserved factors that affects a person's wage in 1990 will also affect that person's wage in 1991)
- Independently pooled cross-section data (repeated cross-sections): obtained by sampling randomly from a large population at different points in time. Therefore they consist of independently sampled observations. This rules out correlation in the error terms across different observations.

Introduction II

- Examples for observation units of Panel Data
  - Firm or company data
  - Longitudinal data on patterns of individual behavior over the life-cycle.
  - Comparative country-specific macroeconomic data over time.

- Examples of Panel Data Sets
  - Panel Study of Income Dynamics (PSID)
  - National Longitudinal Surveys of Labor Market Experience (NLS)
  - German Socioeconomic Panel (GSOEP)
  - The British Household Panel Survey (BHPS)
  - Swedish Farm Economic Survey (JEU)
  - ...

Introduction III

- In microeconomic data, n (the number of individuals, firms...) is typically large, while T is small.
- In aggregate data, longer T is more common.
- In the opposite case, say $n = 5$ countries and $T = 40$ years, the topic becomes **multiple time series**.

Introduction IV

- If the time periods for which we have the data are the same for all n individuals we have a balanced panel, otherwise an unbalanced panel.
- Analyzing unbalanced data typically raises few additional issues compared with analysis of balanced data. However, if the panel is unbalanced for reasons that are not entirely random, then we may need to take this into account when estimating the model (e.g., with a sample selection model).
- Rotating panel: Rotation of samples should avoid that interviewees change their answering behavior during time or refuse to answer at all; also dropped out observation units may be replaced.

New Opportunities I

When we have a dataset with both time series and a cross-section dimension, this opens up new opportunities in our research:

- Large sample size than single cross-section, and so more precise estimates (i.e. lower standard errors).
- Increased degrees of freedom.
- Reduces collinearity among explanatory variables.
- Panel data enable you to solve an omitted variables problem.
- More variability, e.g. less aggregation over firm and individuals
- Better able to study dynamics of adjustment in unemployment, income mobility, ...
- Panel data provide better prediction of individual's behavior

New Opportunities II

- Possibility of identifying and measuring effects that are not detectable in pure cross-section (CS) or time-series (TS) data. Control for unobservable individual heterogeneity and dynamics not possible in TS ($N = 1$) and CS ($T = 1$). Example: married woman labour-force participation of 50% interpreted as 50% chance of being in labour force in any given year, or alternatively 50% always work and 50% never.
- Microdynamic and macrodynamic effects cannot be estimated using CS data.
- Multicollinearity problem in single TS data. Insufficient information to obtain unconditional estimates of lag coefficients.

New Difficulties

- Complicated survey design, stratification
- Changing structure of population (use of rotating panel data)
- Incomplete coverage of the population of interest
- Data collection and management problem
- Distortions of measurement errors due to faulty response, unclear questions, ...
- Non-response (partial or complete) due to lack of cooperation
- Attrition problem, non-response over time is increasing
- Short time-series dimension, increased N costly, increased T deteriorates attrition
- New estimation problems
- Imputations of unit non-response/missing.

9.2.1 General modeling framework for analyzing panel data

$$
\begin{aligned}
y_{it} &= x_{it}'\beta + z_i'\alpha + \varepsilon_{it} \qquad t = 1, ..., T_i; i = 1, ..., n \\
&= x_{it}'\beta + c_i + \varepsilon_{it}
\end{aligned}
$$

x_{it} contains K regressors, not including a constant

The heterogeneity or individual effect is $z_i'\alpha$ where z_i contains a constant term and a set of individual or group specific variables, which may be observed (race, sex, etc.) or unobserved such as family specific characteristics, individual heterogeneity in skill or preferences.

9.2.2 Model Structures

1. Pooled regression: If z_i contains only a constant term, i.e. $y_{it} = x_{it}'\beta + \alpha + \varepsilon_{it}$, then OLS provides consistent and efficient estimates of the common α and the slope vector β.
2. Fixed effects: If z_i is unobserved and **correlated** with x_{it}, then OLS of β is biased and inconsistent as a consequence of an omitted variable. However, in this case $y_{it} = x_{it}'\beta + \alpha_i + \varepsilon_{it}$, embodies all the observable effects and specifies an estimable conditional mean. This fixed effects approach takes α_i to be a group-specific constant term in the regression model.
3. Random effects: If the unobserved individual heterogeneity can be assumed to be **uncorrelated** with the included variables, then the model may be formulated as $y_{it} = x_{it}'\beta + \alpha + u_i + \varepsilon_{it}$.
4. Random parameters: $y_{it} = x_{it}'(\beta + h_i) + \alpha + u_i + \varepsilon_{it}$

9.3 The Pooled Regression Model I

$$y_{it} = \alpha + x_{it}'\beta + \varepsilon_{it}, \qquad i = 1, ..., n, t = 1, ..., T_i,$$

$$\begin{aligned} E[\varepsilon_{i,t}|x_{i,1}, x_{i,2}, ..., x_{i,T_i}] &= 0, \\ Var[\varepsilon_{i,t}|x_{i,1}, x_{i,2}, ..., x_{i,T_i}] &= \sigma_\varepsilon^2, \\ Cov[\varepsilon_{i,t}, \varepsilon_{j,s}|x_{i,1}, x_{i,2}, ..., x_{i,T_i}] &= 0 \text{ if } i \neq j \text{ or } t \neq s. \end{aligned}$$

OLS is the consistent and efficient estimator and inference can reliably be done.

9.3 The Pooled Regression Model II

What happens if instead of α we have u_i? If we estimate the model using OLS then u_i will go into the error term: $w_{it}^{OLS} \equiv u_i + \varepsilon_{it}$.

1. u_i is **uncorrelated** with x_{it}

   $$Cov(w_{it}^{OLS}, w_{i,t-s}^{OLS}) = E[(u_i + \varepsilon_{i,t})(u_i + \varepsilon_{i,t-s})] = \sigma_u^2$$

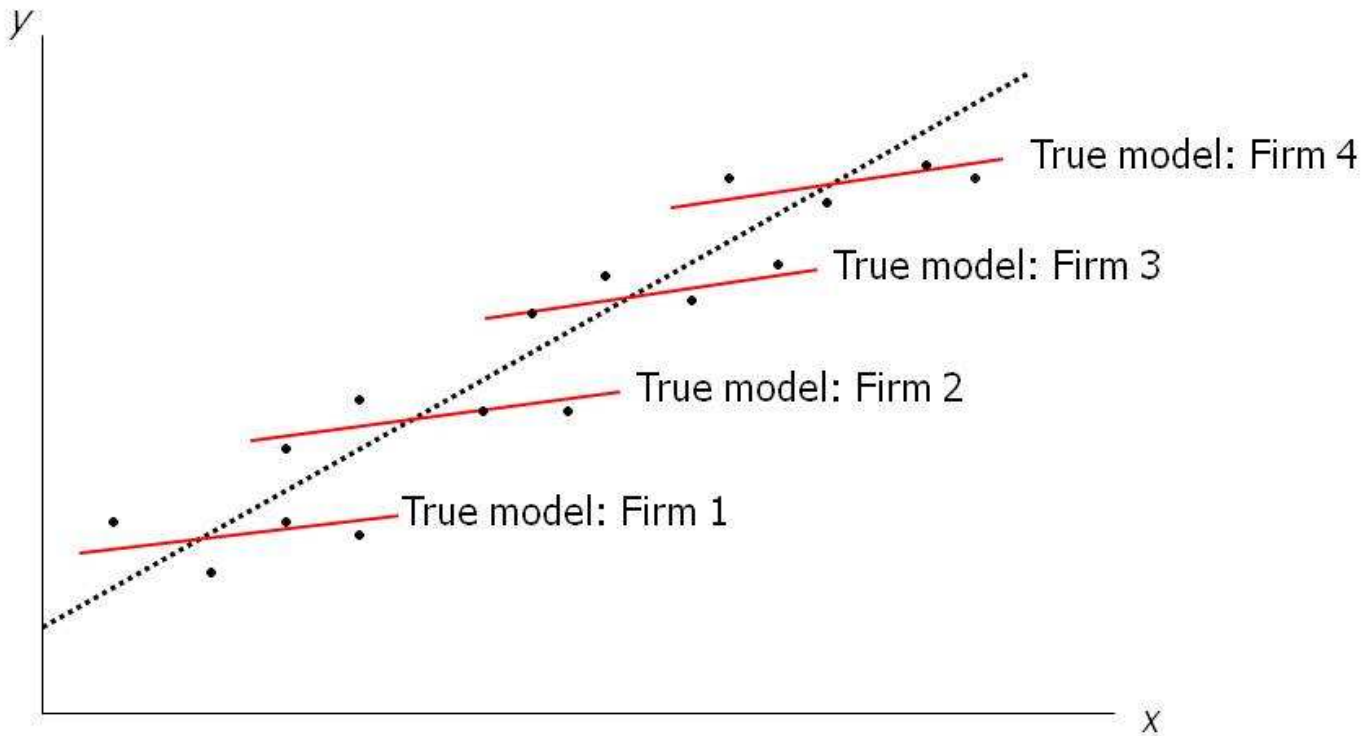
   Hence, w_{it}^{OLS} is serially correlated. OLS is consistent but wrong standard errors.$\Rightarrow$ FGLS.

2. u_i is **correlated** with x_{it} (**omitted variable problem**)

   $$\begin{aligned} b &= \beta + (X'X)^{-1}X'(u + \varepsilon) \\ plim\ b &= \beta + plim(X'X)^{-1}X'(u + \varepsilon), \end{aligned}$$

   which shows that the OLS estimator is inconsistent unless $cov(X, u) = 0$.

Heterogeneity bias as a result of pooled regression



9.3.2 Robust covariance matrix estimation I

In a longitudinal data set, omitted components left out of the model will carry across all periods and cause autocorrelation that may be more serious than heteroscedasticity between individuals. Although we estimate the pooled regression model, we may want to compute a robust covariance matrix.

Stack the T_i observations for individual i in a single equation,

$$\begin{aligned} y_i &= X_i\beta + w_i \qquad\qquad i = 1, ..., n \\ Var[w_i|X_i] &= \sigma_\varepsilon^2 I_{T_i} + \Sigma_i = \Omega_i \end{aligned}$$

9.3.2 Robust covariance matrix estimation I

$$
\begin{aligned}
b &= (X'X)^{-1}X'y \\
&= \left[\sum_{i=1}^{n} X_i'X_i\right]^{-1} \sum_{i=1}^{n} X_i'(X_i\beta + w_i) \\
&= \beta + \left[\sum_{i=1}^{n} X_i'X_i\right]^{-1} \sum_{i=1}^{n} X_i'w_i
\end{aligned}
$$

We obtain an estimate of the covariance matrix that is **robust to heteroscedasticity and autocorrelation** using the sandwich formula:

$$
Est.Asy.Var[b] = \left[\sum_{i=1}^{n} X_i'X_i\right]^{-1} \left[\sum_{i=1}^{n} X_i'\hat{w}_i\hat{w}_i'X_i\right] \left[\sum_{i=1}^{n} X_i'X_i\right]^{-1},
$$

where $\hat{w}_i$ is the vector of T_i residuals for individual i.

9.3.6 The Within- and Between-Groups Estimators I

The pooled regression model can be formulated in three ways:

Original equation

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

In group means

$$\bar{y}_{i\cdot} = \alpha + \bar{x}_{i\cdot}'\beta + \bar{\varepsilon}_{i\cdot},$$

In terms of deviations from the group means

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

(Assumption: no time-invariant variables.)

9.4 The Fixed Effects Model ('Within' Estimator)

$$y_{it} = x'_{it}\beta + \alpha_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T; i = 1, 2, ..., n$$

Differences across groups can be captured in differences in the constant term.

Major shortcoming: The coefficients of the time-invariant variables cannot be estimated.

Any time invariant variables in x_{it} will mimic the individual specific constant term, e.g.

$$lnWage_{it} = x'_{it}\beta + [\beta_{10}Ed_i + \beta_{11}Fem_i + \beta_{12}Blk_i + c_i] + \varepsilon_{it}$$

The fixed effects formulation of the model will absorb the last four terms in the regression in α_i.

9.4.1 Least squares estimation I

Let y_i and X_i be the T observations for the ith unit, ι be a $T \times 1$ column of ones, and let ε_i be the associated $T \times 1$ vector of disturbances. Then

$$
\begin{aligned}
y_i &= X_i\beta + \iota\alpha_i + \varepsilon_i, \qquad i = 1, ..., n \\
\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} &= \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \beta + \begin{pmatrix} \iota & 0 & \cdots & 0 \\ 0 & \iota & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \iota \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \\
y &= \begin{pmatrix} X & d_1 & d_2 & \cdots & d_n \end{pmatrix} \begin{pmatrix} \beta \\ \alpha \end{pmatrix} + \varepsilon \\
y &= X\beta + D\alpha + \varepsilon
\end{aligned}
$$

Least squares dummy variables model (LSDV).

9.4.1 Least squares estimation II

$$b = (X'M_D X)^{-1}(X'M_D y)$$

$$M_D = I - D(D'D)^{-1}D'$$

$$M_D = \begin{pmatrix} M^0 & 0 & 0 & \cdots & 0 \\ 0 & M^0 & 0 & \cdots & 0 \\ & & \cdots & & \\ 0 & 0 & 0 & \cdots & M^0 \end{pmatrix}$$

$M^0 = I_T - \frac{1}{T}\iota\iota'$, Therefore $M^0 x_i = x_i - \bar{x}_i.\iota$ and $M^0 y_i = y_i - \bar{y}_i.\iota$ (time-demeaned data, within transformation).

Therefore the least squares regression of $M_D y$ on $M_D X$ is equivalent to a regression of $(y_{it} - \bar{y}_i.)$ on $(x_{it} - \bar{x}_i.)$. Hence, $b = b^{within}$.

9.4.1 Least squares estimation III

The dummy variable coefficients can be recovered from the other normal equation in the partitioned regression:

$$D'Da + D'Xb = D'y,\ a = (D'D)^{-1}D'(y - XB).$$

This implies that for each i, $a_i = \bar{y}_{i.} - \bar{x}_{i.}'b$.

Estimated standard errors:

$$\begin{aligned} Est.Asy.Var[b] &= s^2 \left(X'M_D X\right)^{-1} \\ s^2 &= \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it} - x_{it}'b - a_i)^2}{nT - n - K} \\ &= \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(M_D y - M_D Xb)'(M_D y - M_D Xb)}{nT - n - K} \\ Est.Asy.Var[a_i] &= \frac{s^2}{T} + \bar{x}_{i.}'(Est.Asy.Var[b])\bar{x}_{i.} \end{aligned}$$

9.4.3 Testing the significance of the group effects

$$
\begin{aligned}
H_0 \quad &: \quad \alpha_1 = \alpha_2 = ... = \alpha_n \\
F(n-1, nT-n-K) \quad &= \quad \frac{(R^2_{LSDV} - R^2_{Pooled})/(n-1)}{(1-R^2_{LSDV})/(nT-n-K)}
\end{aligned}
$$

9.4.4 Fixed time and group effects

The LSDV approach can be extended to include a time-specific effect as well by inclusion of additional $T-1$ dummy variables:

$$y_{it} = x'_{it}\beta + \alpha_i + \delta_t + \varepsilon_{it}$$

Note for interpretation: The time effects are **contrasts**.

9.3.5 The First Differencing Estimator

Instead of using Dummies (or time-demeaning the data) we now difference the data:

$$\begin{aligned} y_{it} &= x_{it}\beta + \alpha_i + \varepsilon_{it} \qquad t = 1, ..., T; i = 1, ..., n \\ y_{it} - y_{i,t-1} &= (x_{it} - x_{i,t-1})\beta + (\alpha_i - \alpha_i) + (\varepsilon_{it} - \varepsilon_{i,t-1}) \\ \Delta y_{it} &= \Delta x_{it}\beta + \Delta \varepsilon_{it} \end{aligned}$$

Clearly this removes the individual effect, and so we can obtain consistent estimates of β by estimating the equation in first differences by OLS.

Assumption: $E[x_{it}\varepsilon_{is}] = 0$ for $s = t, t-1$.
Otherwise there will be endogeneity bias if we use OLS.

First Differencing versus Fixed Effect I

FE and FD are two alternative ways of obtaining estimates in the presence of fixed effects. Which method should we use?

- When $T = 2$ (i.e. only two time periods), FE and FD are exactly equivalent.

- When $T \geq 3$, FE and FD are not the same. Under the null hypothesis that the model is correctly specified, FE and FD differ only because of sampling error.
  Hence, if FE and FD are significantly different - so that the differences in the estimates cannot be attributed to sampling error - we would worry about the validity of the strict exogeneity assumption, necessary for FE: $E[x_{it}\varepsilon_{is}] = 0$ for $s = 1, 2, .., T$.

First Differencing versus Fixed Effect II

- If $\varepsilon_{i,t}$ is a random walk ($\varepsilon_{i,t} = \varepsilon_{i,t-1} + \xi_{i,t}$), then $\Delta\varepsilon_{i,t}$ is serially uncorrelated and so the FD estimator will be more efficient than the FE estimator.

- Under 'classical' assumptions, i.e. $\varepsilon_{it} \sim iid(0, \sigma_\varepsilon^2)$, the FE estimator will be more efficient than the FD estimator as in this case the FD residual $\Delta\varepsilon_{it}$ exhibits negative serial correlation.

9.5 Random Effects

$$y_{it} = x_{it}'\beta + (\alpha + u_i) + \varepsilon_{it} \qquad t = 1, ..., T; i = 1, ..., n$$

$$\begin{aligned}
E[\varepsilon_{it}|X] &= E[u_i|X] = 0 \\
E[\varepsilon_{it}^2|X] &= \sigma_\varepsilon^2 \\
E[u_i^2|X] &= \sigma_u^2 \\
E[\varepsilon_{it} u_j|X] &= 0 \text{ for all } i, t, \text{ and } j, \\
E[\varepsilon_{it}\varepsilon_{js}|X] &= 0 \text{..if } t \neq s \text{or } i \neq j, \\
E[u_i u_j|X] &= 0 \text{..if } i \neq j.
\end{aligned}$$

Hence,

- u_i uncorrelated with x_{it}: $E[x_{it} u_i] = 0$ and
- Strict exogeneity holds: $E[x_{it}\varepsilon_{is}] = 0$ for $s = 1, 2, ..., T$.

9.5 Random Effects I

View again the model in blocks of T observations for group i:

$$\begin{aligned}
\eta_{it} &= \varepsilon_{it} + u_i \\
E[\eta_{it}^2|X] &= \sigma_\varepsilon^2 + \sigma_u^2 \\
E[\eta_{it}\eta_{is}|X] &= \sigma_u^2, \qquad t \neq s \\
E[\eta_{it}\eta_{js}|X] &= 0 \qquad \text{for all } t \text{ and } s \text{ if } i \neq j.
\end{aligned}$$

Hence, the residuals are **serially correlated**.

$$\begin{aligned}
\Sigma &= E[\eta_i \eta_i'|X] \\
\Sigma &= \begin{pmatrix}
\sigma_\varepsilon^2 + \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\
\sigma_u^2 & \sigma_\varepsilon^2 + \sigma_u^2 & \cdots & \sigma_u^2 & \sigma_u^2 \\
 & \cdots & & & \\
\sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_\varepsilon^2 + \sigma_u^2
\end{pmatrix} = \sigma_\varepsilon^2 I_T + \sigma_u^2 \iota_T \iota_T'
\end{aligned}$$

9.5 Random Effects II

The disturbance covariance matrix for the full nT observations is

$$\Omega = \begin{pmatrix} \Sigma & 0 & 0 & \cdots & 0 \\ 0 & \Sigma & 0 & \cdots & 0 \\ & \cdots & & & \\ 0 & 0 & 0 & \cdots & \Sigma \end{pmatrix} = I_n \otimes \Sigma$$

Hence, the Random Effects estimator is a GLS estimator that takes this covariance matrix into account:

$$\begin{aligned} \hat{\beta} &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y \\ &= \left(\sum_{i=1}^{n} X_i'\Sigma^{-1}X_i\right)^{-1}\left(\sum_{i=1}^{n} X_i'\Sigma^{-1}y_i\right) \end{aligned}$$

9.5.1 Generalized Least Squares

Let us transform the data and use OLS with the transformed data. Therefore we will require $\Omega^{-1/2} = [I_n \otimes \Sigma]^{-1/2}$. What is $\Sigma^{-1/2}$?

$$\begin{aligned} \Sigma &= \sigma_\varepsilon^2 I_T + \sigma_u^2 \iota_T \iota_T' \\ \Sigma^{-1/2} &= \frac{1}{\sigma_\varepsilon}\left[I - \frac{\theta}{T}\iota_T \iota_T'\right] \qquad \theta = 1 - \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + T\sigma_u^2}} \end{aligned}$$

$$\Sigma^{-1/2} y_i = \frac{1}{\sigma_\varepsilon}\begin{pmatrix} y_{i1} - \theta \bar{y}_{i\cdot} \\ y_{i2} - \theta \bar{y}_{i\cdot} \\ \vdots \\ y_{iT} - \theta \bar{y}_{i\cdot} \end{pmatrix}$$

and likewise for the rows of X_i.

Note: Similarity of this procedure to the computation in the LSDV model which uses $\theta = 1$.

9.5.2 Feasible Generalized Least Squares when Σ is unknown

$$\begin{aligned}
\hat{\sigma}_{\varepsilon} &= s^2_{LSDV} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(e_{it} - \bar{e}_{i\cdot}]^2}{nT - n - K} \\
s^2_{Pooled} &= \frac{e'e}{nT - K - 1} \\
\hat{\sigma}^2_u &= s^2_{Pooled} - s^2_{LSDV}
\end{aligned}$$

9.5.3 Testing for random effects

Breusch and Pagan Lagrange multiplier test for the random effects model based on OLS residuals:

$$
\begin{aligned}
H_0 &: \sigma_u^2 = 0 \qquad \text{(Pooled OLS model)} \\
H_1 &: \sigma_u^2 \neq 0 \qquad \text{(RE model)} \\
LM &= \frac{nT}{2(T-1)} \left(\frac{\sum_{i=1}^{n} (T\bar{e}_{i\cdot})^2}{\sum_{i=1}^{n} \sum_{t=1}^{T} e_{it}^2} - 1 \right)^2 \\
&= \frac{nT}{2(T-1)} \left(\frac{\sum_{i=1}^{n} [\sum_{t=1}^{T} e_{it}]^2}{\sum_{i=1}^{n} \sum_{t=1}^{T} e_{it}^2} - 1 \right)^2 \\
LM &\sim \chi^2(df = 1)
\end{aligned}
$$

9.5.4 Hausman's specification test for the random effects model I

Usually applied to test for fixed versus random effects models

- Tests for orthogonality of the common effects and the regressors
- Compares directly the random effects estimator, $\hat{\beta}_{RE}$, to the fixed effects estimator, $\hat{\beta}_{FE}$
- In the presence of a correlation between the individual effects and the regressors the GLS estimates are inconsistent, while the OLS fixed effects results are consistent
- If there is no correlation between the fixed effects and the regressors both estimators are consistent, but the OLS fixed effects estimator is inefficient

9.5.4 Hausman's specification test for the random effects model II

Construct $q = b_{FE} - \hat{\beta}_{RE}$ and $Var(q) = Var(b_{FE}) - Var(\hat{\beta}_{RE})$

$$
\begin{aligned}
H_0 &: \quad q = 0 \qquad \text{random effects model} \\
H_1 &: \quad q \neq 0 \qquad \text{fixed effects model} \\
H &= \quad q' Var(q) q \sim \chi^2(df = K)
\end{aligned}
$$

Under H_0 both estimators are consistent, only RE estimator is efficient.
Under H_1 only FE consistent, RE inconsistent.

Summary of specification tests