Applied Economics

Panel Data

Department of Economics
Universidad Carlos III de Madrid

See also Wooldridge (chapter 13), and Stock and Watson (chapter 10)

Panel Data vs Repeated Cross-sections

- In a cross-section each observation can represent an individual, a family, a firm, a state, a country, etc.
- With repeated cross-sections we have more than one period, and observations in each period correspond to different individuals, families, firms, etc.
- Panel data (or longitudinal data) consists of repeated observations on the same cross-section: same individuals, families, firms or states are observed in different periods. In general we consider applications with many cross section observations but not many periods.

Example

- In order to get a panel to characterize a population of interest it is necessary to select randomly a set of individuals in a certain period and collect the information we want. In the next period (month, quarter, year) we need to find the same individuals and collect the information again.
- In the first class we mentioned some examples. Another one for Spain is the “Encuesta de Presupuestos Familiares‘” from INE. The new survey started in January, 2006. The sample size is about 24,000 households per year, and each household is interviewed two consecutive years.

Using panel data

- Panel data allow us to consider some cases of omitted variables that in repeated cross-sections would yield inconsistent OLS estimations. It is the case of omitted variables that differ across units, but are constant over time.
- When using panel data does not seem reasonable to assume that observations are independent: for instance, unobservable factors that affect wages in 2014 will affect wages also in 2015.
- Then, it is necessary to use special models and methods. We start with a simple model with two periods.

Notation

Simple Model with T=2

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \delta_0 d2_t + a_i + u_{it}, t = 1, 2$$

- Subscript i references the individual, firm, etc.; subscript t indicates the period of time. In the simple case of two periods we denote $t = 1$ (for the first period) and $t = 2$ (for the second one).
- $d2$ is a dummy variable taking the value zero at $t = 1$, and one at $t = 2$. It is the same for all units, then it has no subscript i.
- The variable a_i captures all unobserved factors affecting Y_{it} that do not change over time: then it has no subscript t.

Unobserved Heterogeneity

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \delta_0 d2_t + a_i + u_{it}$$

- a_i represents unobservable time-invariant individual variation or heterogeneity (it is called unobserved effect, fixed effect, unobserved heterogeneity, or individual heterogeneity).
- This model is usually called a fixed effects model or unobserved effects model.
- The usual error term is u_{it}, and includes unobserved factors affecting Y_{it} that change over time.
- The term $v_{it} = a_i + u_{it}$ is called the composite error: it has a constant component and a component that changes over time.

OLS with unobserved heterogeneity

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \delta_0 d2_t + v_{it}$$

- We know that to consistently estimate β_1 using OLS, we need to assume that X_{it} is not correlated with v_{it}.
- Even if we assume that $C(X_{it}, u_{it}) = 0$, OLS estimators will be biased and inconsistent if $C(X_{it}, a_i) \neq 0$.
- The resulting bias for omitting a_i in pooled OLS is sometimes called heterogeneity bias. It is a bias from omitting a time-constant variable.

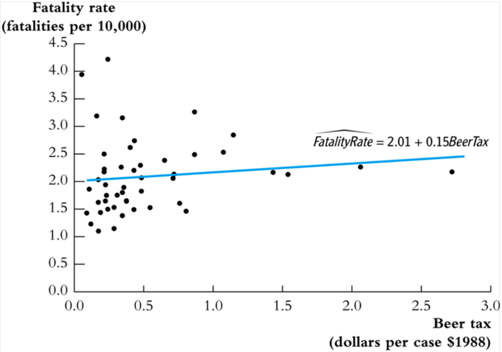
Example: Drunk Driving Laws and Traffic Deaths

- Goal: analyze policies to reduce the number of drunk drivers.
- Data in fatality.gdt in `gretl`: 48 U.S. states, 7 years (1982,...,1988), we use only 1982 and 1988: i represents a state, t a year.
- Variables:

  - Traffic fatality rate: deaths per 10,000 residents (variable TM)
  - Tax on a case of beer in 1988 dollars (variable $Beertax$).

Example: using a single cross-section

Simple model: $MR_i = \beta_0 + \beta_1 BeerTax_i + u_i$



(a) 1982 data

The coefficient in 1982 is not statistically significant. Using pooled cross-sections is positive and significant (see in `gretl`). Does a tax on beer increase the number of traffic fatalities?

Potential explanations

- Many factors affecting traffic fatalities:
  - vintage of autos on the road
  - quality of roads
  - culture of drinking and driving
  - traffic
- If some of these factors are correlated with tax on beer...
- We could try to get data on these factors, but not all of them are easy to get.
- With panel data we can obtain consistent estimators if the unobserved factors that may be correlated with tax on beer, are constant through the period.

Unobserved heterogeneity in panel data

- With panel data we can allow the unobserved effects (a_i) to be correlated with the controls.
- Since a_i is the same in the two periods, we can take first differences and get rid of a_i.

$$Y_{i2} - Y_{i1} = (\beta_0 + \beta_1 X_{i2} + \delta_0 + a_i + u_{i2}) - (\beta_0 + \beta_1 X_{i1} + a_i + u_{i1})$$

$$Y_{i2} - Y_{i1} = \delta_0 + \beta_1 (X_{i2} - X_{i1}) + (u_{i2} - u_{i1})$$

That it can be written as:

$$\Delta Y_i = \delta_0 + \beta_1 \Delta X_i + \Delta u_i$$

where Δ denotes the change from $t = 1$ to $t = 2$.

First-Differenced estimator (FD)

$$\Delta Y_i = \delta_0 + \beta_1 \Delta X_i + \Delta u_i$$

- This equation in first differences is just a single cross-sectional equation, but each variable is differenced over time.
- The intercept represents the change in Y when there is no change in X.
- Key assumption: Δu_i not correlated with ΔX_i. If that is true, OLS will be consistent.
- The OLS estimate for β_1 in that equation is called the First-Differenced estimator (FD).

About the FD estimator

- Allows for arbitrary correlation between a_i and the controls.
- Under general conditions $plim(\hat{\beta}) = \beta$ when $N \to \infty$ and T fixed.
- There is no estimation of the fixed-effects a_i.
- All time-invariant controls are dropped when taking first differences.
- Then, ΔX_i needs to vary across units. We may have a problem if ΔX_i does not change much. If that is the case, sometimes it makes sense to take differences between periods far away.

Mortality example: a panel

Simple model

$$MR_{it} = \beta_0 + \beta_1 BeerTax_{it} + \delta_0 d1988_t + a_i + u_{it}, \text{ with } t = 1982, 1988$$

- a_i represents invariant and unobservable variables affecting fatality rate in state i.
- It may include local attitudes towards drunk driving (if they change slowly we can consider them as constant between 1982 and 1988).
- For instance, states with a less favorable attitude towards drunk driving will have on average less traffic fatalities and also probably higher alcohol taxes.
- This would be a case of omitted variable bias: *BeerTax* captures in part the effect of the local attitude towards drunk driving.

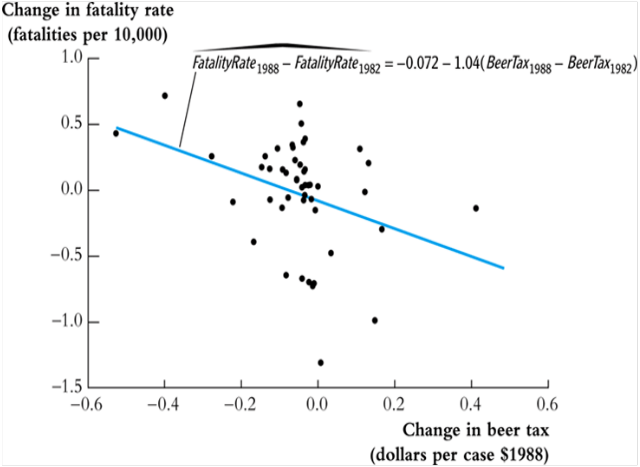
Mortality example: first differences

- The effect of a_i can be eliminated taking first differences.

$$\Delta MR_i = \delta_0 + \beta_1 \Delta BeerTax_i + \Delta u_i$$

- Intuitively:
  - Attitudes towards drunk driving in a state affects the number of drunk drivers and therefore the traffic fatality rate in that state.
  - However, if those attitudes did not change between 1982 and 1988, they could not have affected the change in the number of traffic fatalities in the state.
  - Variation in *MR* in the period has to be explained by other factors. In the previous equation these other factor are changes in beer taxes and changes in u_i.

Mortality example: FD graph



In contrast to the results with pooled cross-sections, using FD we find that an increase in alcohol taxes reduces traffic fatalities.

Mortality example: FD results

- The intercept allows a change in MR in case of no change in $BeerTax$. In this case, the negative estimate could reflect improvements in cars and roads that reduce the number of fatalities.
- An increase in the beer tax of 1$ per case reduces traffic mortality on average by 1.04 deaths per 10,000 residents.
- It is a sizable estimated effect: average fatality rate in the data is 2 deaths per 10,000 people per year. This result suggests that deaths could be reduced by half with a tax increase of 1$ per case of beer.

Mortality example: some comments

- The regression in first differences controls for all invariant and unobserved factors, for instance local attitudes towards drunk driving.
- Note that we need variation in *ImpCerv* to be able to estimate the effect of taxes on traffic deaths.
- There may be factors affecting traffic security, correlated with alcohol taxes, that are changing over time. If we omit those variables we may have a bias when using FD. We do not discuss this problem in this course.

First Differences with more periods

- It is possible to apply the same strategy with $T > 2$: we can take differences from adjacent periods: $t - (t-1)$
- For simplicity assume $T = 3$:

$$Y_{it} = \delta_1 + \delta_2 d2_t + \delta_3 d3_t + \beta_1 X_{it} + a_i + u_{it}, \text{ for } t = 1, 2, 3$$

- We are interested in β_1. If a_i is correlated with X_{it}, OLS will be inconsistent.
- Key assumption: $C(X_{it}, u_{is}) = 0$ for all t, s: once we eliminate the effect of a_i, X_{it} is exogenous.

Getting rid of a_i

$$\Delta Y_{it} = \delta_2 \Delta d2_t + \delta_3 \Delta d3_t + \beta_1 \Delta X_{it} + \Delta u_{it}, \text{ with } t = 2, 3$$

First Differences with more periods (cont.)

- Key assumption for OLS to be consistent: ΔX_{it} not correlated wtih Δu_{it} for $t = 2, 3$.
- It can be shown that the previous equation is equivalent to one including a constant and only one time dummy. Then, in practice we estimate:

FD with $T = 3$

$$\Delta Y_{it} = \alpha_0 + \alpha_1 d3_t + \beta_1 \Delta X_{it} + \Delta u_{it}, \text{ for } t = 2, 3$$

- With $T > 3$ we do the same: we will have $T - 1$ observations per unit, and we will include $T - 2$ time dummies.

Another option when $T > 2$

- There is another method of getting rid of the individual heterogeneity, that under some assumptions (that we do not discuss here) is better than FD.
- Let's start with the simple model:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + a_i + u_{it},\ t = 1, 2, \ldots, T$$

- For each i, we average this equation over time:

$$\bar{Y}_i = \beta_0 + \beta_1 \bar{X}_i + a_i + \bar{u}_i, \text{ where } \bar{Y}_i = T^{-1} \sum_{t=1}^{T} Y_{it}.$$

- We subtract the second equation from the first one for each t:

$$Y_{it} - \bar{Y}_i = \beta_1 (X_{it} - \bar{X}_i) + u_{it} - \bar{u}_i,\ t = 1, 2, \ldots, T$$

Time-demeaned data

- It can be expressed as:

Time-demeaned data

$$\ddot{Y}_{it} = \beta_1 \ddot{X}_{it} + \ddot{u}_{it},\ t = 1, 2, \ldots, T$$

- where $\ddot{Y}_{it} = Y_{it} - \bar{Y}_i$ represents deviations from the mean (called time-demeaned data).
- The effect a_i has disappeared, then we can estimate the equation by pooled OLS.
- The OLS estimator based on the time-demeaned variables is called the fixed effects estimator or the within estimator.

Fixed Effects

- In applied economics a traditional view of the fixed effects model is to assume that the unobserved effect a_i is a parameter to be estimated for each i. Under this view, a_i is the intercept for unit i that is to be estimated along with β_1.
- The way to do it is to include a dummy variable for each unit i.
- Each dummy variable is called a fixed effect. Note that it is not a fixed parameter, but it is constant for each unit in the period.
- It can be shown that this model gives us exactly the same estimates that the ones obtained with the time-demeaned data. Therefore, the within estimator can be obtained including unit dummies as well as time-demeaning the data.

Fixed Effects or Within estimator characteristics

- The fixed effects or within estimator is consistent if u_{it} is uncorrelated with X_{it} for all t.
- Like the FD estimator, the within estimator allows for any type of correlation between a_i and the controls. Also in this case all invariant controls are dropped.
- In the usual case of large N and small T, the choice between FD or within estimators is based on the relative efficiency of the estimators since both will be consistent.
- In most cases with more than two periods the within estimator is more efficient.

Mortality example: within estimator

$$\ddot{MR}_{it} = \beta_1 \ddot{BeerTax}_{it} + \ddot{u}_{it}, \text{ with } t = 1982, 1988$$

- In more detail in the problem set, very briefly here.
- First we need to tell `gretl` that we have panel data, using the unit and the time variables (`setobs` state year −−panel−vars).
- Then we use the command `panel` (if we want with the option −−time-dummies).
- We get: $\hat{\beta}_1 = -1.04097$ with a standard error of 0.3475.
- Note that we get the same result if we include state dummies.

Random Effects (*RE*)

- The *FE* method uses a transformation of the model to eliminate unobserved heterogeneity that gives us a consistent estimator without imposing any additional assumption on a_i.
- The *RE* estimator is useful if we assume that unobserved heterogeneity is not correlated with the controls.
- If we include the correct variables in our model, we can think that unobserved heterogeneity only leads to autocorrelation in the error term, but not to correlation between the error term and the controls.
- If this additional assumption is true, the *RE* estimator is going to be consistent, and more efficient than the *FE* estimator, as it exploits more information.

Random Effects Model

- A model with unobserved heterogeneity

$$Y_{it} = \beta_0 + \beta_1 X_{it1} + \beta_2 X_{it2} + ... + \beta_k X_{itk} + a_i + u_{it},\ t = 1, 2, ..., T$$

- Note that the X can include time dummies.
- Suppose that a_i is not correlated with the X in each period. In this case, if we transform the model to eliminate a_i we will obtain inefficient estimators.
- This model is called a random effects model when we assume that unobserved heterogeneity a_i is not correlated with any of the controls: $Cov(X_{itj}, a_i) = 0, t = 1, 2, ..., T; j = 1, 2, ..., k$

OLS under Random Effects

Let's use the simplest model:

$$Y_{it} = \beta_0 + \beta_1 X_{it1} + a_i + u_{it},\ t = 1, 2, ..., T$$

- Note that if $Cov(X_{it1}, a_i) = 0 \Rightarrow \hat{\beta_1}^{OLS}$ using a single cross-section will be consistent: no need to use panel data.
- But using a single cross-section implies not considering useful information. So we can use pooled OLS and we will get consistent estimators.
- But, the error term $v_{i,t} = a_i + u_{it}$, is serially correlated due to a_i.
- Because under OLS the standard errors are calculated without considering this autocorrelation, OLS estimators are not asymptotically efficient.

The Random Effects Estimator

- A method called Feasible Generalized Least Squares (FGLS) gives asymptotically efficient estimates (we don't see the details here).
- The *GLS* transformation is the following:

$$y_{it} - \lambda \overline{y}_i = \beta_0 (1 - \lambda) + \beta_1 \left(X_{it1} - \lambda \overline{X}_{i1} \right) + (v_{it} - \lambda \overline{v}_i)$$

  where $\overline{y}_i$ denotes averages over time, and λ lies between zero and one and depends on the variances of a_i and u_{it}, and T:
  - if u_{it} is large relative to a_i, then λ will be close to 0 and *RE* will be similar to *OLS*
  - if u_{it} is small relative to a_i, then λ will be close to 1 and *RE* will be similar to *FE*
- GLS is the OLS estimator of the transformed equation (straightforward in `gretl`).

Fixed Effects (*FE*) vs. Random Effects (*RE*)

- If we are primarily interested in the effect of a time-constant variable in a panel data study, FE is practically useless.
- Without using IV, not an easy task, RE is probably our only choice.
- We need to add variables that capture the part of a_i correlated with X, for instance including dummy variables for groups (if we have many observations within each group). For example, if we have panel data at the student level, we might include school dummy variables.
- Including dummy variables for groups controls for a certain amount of heterogeneity that might be correlated with the invariant elements of X.

A Test for *FE* vs. *RE*

- If time-invariant unobserved heterogeneity is correlated with the controls, then *FE* is consistent while *RE* is not
- If random effects assumptions are true, both *FE* and *RE* are consistent, but *RE* will be more efficient than *FE*
- We can test whether unobservable heterogeneity is correlated with the controls using a test proposed by Hausman:
    - H_0 : random effects assumption is true $\Rightarrow$ insignificant differences between *RE* and *FE*
    - H_1 : random effects assumption is false $\Rightarrow$ differences between *RE* and *FE* significant

The Hausman Test

- The Hausman test is based on the difference between the FE and the RE estimators.
- A statistically significant difference suggests the null is false, and then is evidence against the RE assumption.
- We don't derive the test, but we know that under the null, $H \xrightarrow{a} \chi^2_K$ where K is the number of regressors
- Straightforward to implement it in `gretl`.

An Example: Firm Innovation

- strong consensus on the role of innovation on economic growth
- innovation determinants much less studied: geographical market range (GMR)
    - successfully innovative firms are in a position to conquer new markets for their products
    - firms with a large presence in distant markets more likely to incur in innovation costs to keep market shares
    - unobservable factors: high quality management

GMR

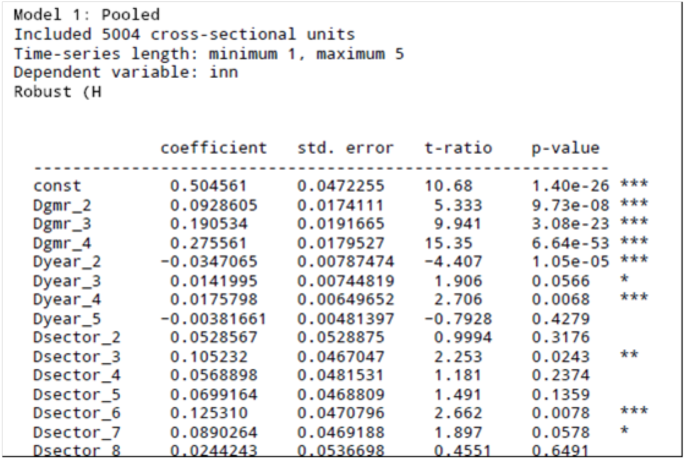
- Consider the unobserved effects model

$$innovation_{it} = \beta_{1j} GMR_{it}(j) + \beta_{2k} Dsize_{it}(k) + \beta_{3l} Dsector(l) + a_i + u_{it}$$

- $GMR(j)$: range of GMR: 1: local, 2: national, 3: european, 4: international
- If *GMR* affects innovation positively, β_{1j} would increase with j
- We can use the "Panel de Innovación Tecnológica" (Panel of Technological Innovation), or PITEC. This unbalanced panel includes firm level detailed information on innovation from 2003 to 2008.

Pooled OLS

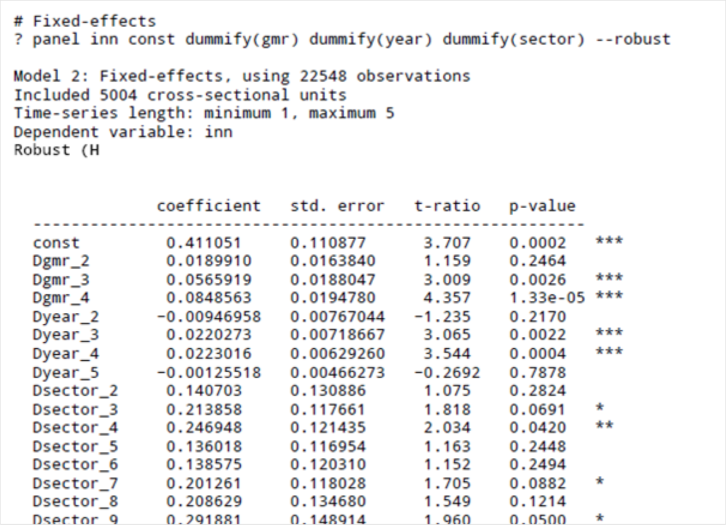
ols inn const dummify(gmr) dummify(year) dummify(sector) −−robust



```
Model 1: Pooled
Included 5004 cross-sectional units
Time-series length: minimum 1, maximum 5
Dependent variable: inn
Robust (H


              coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const        0.504561     0.0472255    10.68     1.40e-26 ***
  Dgmr_2       0.0928605    0.0174111     5.333    9.73e-08 ***
  Dgmr_3       0.190534     0.0191665     9.941    3.08e-23 ***
  Dgmr_4       0.275561     0.0179527    15.35     6.64e-53 ***
  Dyear_2     -0.0347065    0.00787474   -4.407    1.05e-05 ***
  Dyear_3      0.0141995    0.00744819    1.906    0.0566   *
  Dyear_4      0.0175798    0.00649652    2.706    0.0068   ***
  Dyear_5     -0.00381661   0.00481397   -0.7928   0.4279
  Dsector_2    0.0528567    0.0528875     0.9994   0.3176
  Dsector_3    0.105232     0.0467047     2.253    0.0243   **
  Dsector_4    0.0568898    0.0481531     1.181    0.2374
  Dsector_5    0.0699164    0.0468809     1.491    0.1359
  Dsector_6    0.125310     0.0470796     2.662    0.0078   ***
  Dsector_7    0.0890264    0.0469188     1.897    0.0578   *
  Dsector_8    0.0244243    0.0536698     0.4551   0.6491
```

Fixed Effects

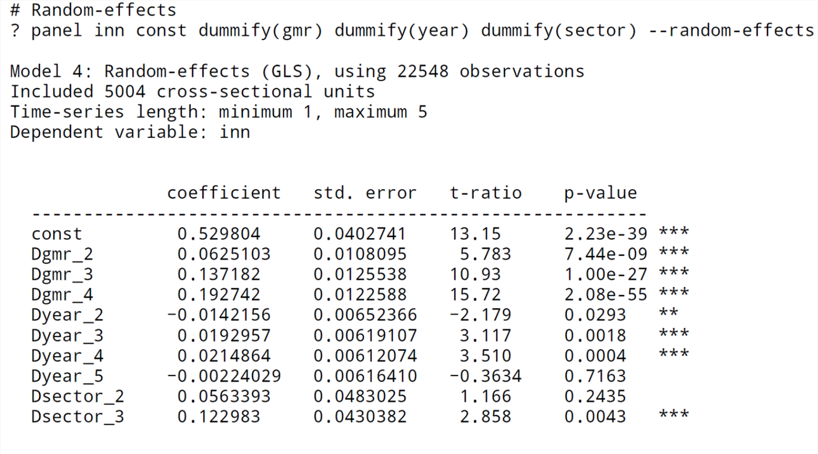


```
# Fixed-effects
? panel inn const dummify(gmr) dummify(year) dummify(sector) --robust

Model 2: Fixed-effects, using 22548 observations
Included 5004 cross-sectional units
Time-series length: minimum 1, maximum 5
Dependent variable: inn
Robust (H


               coefficient   std. error   t-ratio   p-value
  ----------------------------------------------------------
  const         0.411051      0.110877      3.707     0.0002   ***
  Dgmr_2        0.0189910     0.0163840     1.159     0.2464
  Dgmr_3        0.0565919     0.0188047     3.009     0.0026   ***
  Dgmr_4        0.0848563     0.0194780     4.357     1.33e-05 ***
  Dyear_2      -0.00946958    0.00767044   -1.235     0.2170
  Dyear_3       0.0220273     0.00718667    3.065     0.0022   ***
  Dyear_4       0.0223016     0.00629260    3.544     0.0004   ***
  Dyear_5      -0.00125518    0.00466273   -0.2692    0.7878
  Dsector_2     0.140703      0.130886      1.075     0.2824
  Dsector_3     0.213858      0.117661      1.818     0.0691   *
  Dsector_4     0.246948      0.121435      2.034     0.0420   **
  Dsector_5     0.136018      0.116954      1.163     0.2448
  Dsector_6     0.138575      0.120310      1.152     0.2494
  Dsector_7     0.201261      0.118028      1.705     0.0882   *
  Dsector_8     0.208629      0.134680      1.549     0.1214
  Dsector_9     0.291881      0.148914      1.960     0.0500   *
```

Random Effects



```
# Random-effects
? panel inn const dummify(gmr) dummify(year) dummify(sector) --random-effects

Model 4: Random-effects (GLS), using 22548 observations
Included 5004 cross-sectional units
Time-series length: minimum 1, maximum 5
Dependent variable: inn


               coefficient   std. error   t-ratio    p-value
  -----------------------------------------------------------
  const          0.529804    0.0402741    13.15      2.23e-39 ***
  Dgmr_2         0.0625103   0.0108095     5.783     7.44e-09 ***
  Dgmr_3         0.137182    0.0125538    10.93      1.00e-27 ***
  Dgmr_4         0.192742    0.0122588    15.72      2.08e-55 ***
  Dyear_2       -0.0142156   0.00652366   -2.179     0.0293   **
  Dyear_3        0.0192957   0.00619107    3.117     0.0018   ***
  Dyear_4        0.0214864   0.00612074    3.510     0.0004   ***
  Dyear_5       -0.00224029  0.00616410   -0.3634    0.7163
  Dsector_2      0.0563393   0.0483025     1.166     0.2435
  Dsector_3      0.122983    0.0430382     2.858     0.0043   ***
```

λ and the Hausman Test

- 'Within' variance = 0.0837569 'Between' variance = 0.130407
- Hausman test - Null hypothesis: GLS estimates are consistent
  - Asymptotic test statistic: Chi-square(29) = 374.727 with p-value = 9.52973e-62