4장 Panel Data 분석

Definition

- Panel data: Panel data contain observations on multiple phenomena observed over multiple time periods for the same entities (countries, firms, or individuals)
- Example)

country	year	Y	X1	X2	X3
1	2000	6.0	7.8	5.8	1.3
1	2001	4.6	0.6	7.9	7.8
1	2002	9.4	2.1	5.4	1.1
2	2000	9.1	1.3	6.7	4.1
2	2001	8.3	0.9	6.6	5.0
2	2002	0.6	9.8	0.4	7.2
3	2000	9.1	0.2	2.6	6.4
3	2001	4.8	5.9	3.2	6.4
3	2002	9.1	5.2	6.9	2.1

Model

- General linear panel model

$$y_{it} = \beta_{it} x_{it} + \alpha_{it} + u_{it}$$

⇨ Some restriction: $\beta_{it} = \beta$ and $\alpha_{it} = \alpha_i$

$$y_{it} = \beta x_{it} + \alpha_i + u_{it} \quad \text{with} \quad Cov(x_{it}, u_{it}) = 0 \quad \text{and} \quad Cov(\alpha_i, u_{it}) = 0$$

u_{it} : idiosyncratic error, usually $u_{it} \sim ind\ N(0, \sigma_u^2)$

α_i : time-invariant individual-specific component

individual heterogeneity

- Example) $crime_{it} = \beta \cdot unemp_{it} + \alpha_i + u_{it}$

$$HP_{it} = \beta_1 \cdot crime_{it} + \beta_2 \cdot unemp_{it} + \alpha_i + u_{it}$$

Estimation

- POLS (pooled OLS)

Disregard heterogeneity ($\alpha_i = \alpha$), and run OLS

If $Cov(\alpha_i, x_{it}) = 0$, Pols is consistent.

⇨ Even if Pols is consistent, it is still inefficient in case of autocorrelation on u_{it}.

- LSDV (least square w/ dummy variable)

Put α_i as dummy variables, and run OLS

- FD (first difference)

Make the first difference in time (remove α_i from the model), and run OLS

$$\Delta y_{it} = \beta \Delta x_{it} + \Delta u_{it},$$

$$\text{with } \Delta y_{it} = y_{it} - y_{i,t-1}, \quad \Delta x_{it} = x_{it} - x_{i,t-1}$$

Example) $\Delta crime_{it} = \beta \cdot \Delta unemp_{it} + \Delta u_{it}$

Estimation

- FE (fixed effect)

$Cov(\alpha_i, X_{it}) \neq 0$ ⇨ transform the model to remove α_i from the model

Example)

$$crime_{it} - \overline{crime_i} = \beta \cdot (unemp_{it} - \overline{unemp_i}) + (u_{it} - \overline{u_i})$$

$$HP_{it} - \overline{HP_i} = \beta_1 \cdot (crime_{it} - \overline{crime_i}) + \beta_2 \cdot (unemp_{it} - \overline{unemp_i}) + (u_{it} - \overline{u_i})$$

$$\overrightarrow{HP_{it}} = \beta_1 \cdot \overrightarrow{crime_{it}} + \beta_2 \cdot \overrightarrow{unemp_{it}} + \vec{u}_{it}$$

$$\hat{\alpha}_i = \overline{HP_i} - \hat{\beta}_1 \cdot \overline{crime_i} + \hat{\beta}_2 \cdot \overline{unemp_i}$$

Estimation

- FE (fixed effect)

Note 1) FE vs. FD: parameter estimates are identical if time period is two.
FE vs. LSDV

Note 2) All time-constant (invariant) effects are removed

⇨ Fixed-effect models are designed to study the causes of changes within a person (entity). Time invariant characteristic can not cause such a change because it is constant for each person.

⇨ Since it throws away information, FE estimator is inefficient.

Estimation

- RE (random effect) $Cov(\alpha_i, X_{it}) = 0$

Error components model:

$$y_{it} = \beta x_{it} + \alpha_i + u_{it}, \text{ with } \alpha_i \sim N(0, \sigma_\alpha^2). \text{ Thus, } \alpha_i + u_{it} = \eta_{it}$$

Note 1) you can include time invariant variable as a predictor in the RE model.

Note 2) RE estimation is <u>more efficient</u> than FE or Pols.

Note 3) FGLS (feasible GLS) estimation is applied (based on quasi-demeaning framework).

Estimation

- RE (random effect)

Example)

$$HP_{it} - \theta\overline{HP_i} = \beta_0(1-\theta) + \beta_1 \cdot (crime_{it} - \theta\overline{crime_i}) + \beta_2 \cdot (unemp_{it} - \theta\overline{unemp_i}) + (\eta_{it} - \theta\overline{\eta}_i)$$

$$\text{, where } \theta = 1 - \left(\frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}\right)^{1/2}$$

If $\sigma_\alpha^2 = 0 \Rightarrow \theta = 0$, RE≈OLS

If $T\sigma_\alpha^2 = \infty \Rightarrow \theta = 1$, RE≈FE

If $0 < \theta < 1$, RE may not be OLS or FE

⇨ Estimate θ first, then run OLS.

Test

- Hausman test

$Y_{it} = \beta_0 + \beta_1 \cdot X_{it} + \alpha_i + u_{it}$, where α_i : between-entity error, u_{it} : within-entity error

$H_0 : Cov(\alpha_i, X_{it}) = 0$: use random effect model

$H_1 : Cov(\alpha_i, X_{it}) \neq 0$: use fixed effect model

If $Cov(\alpha_i, X_{it}) = 0$, both $\hat{\beta}_{RE}$ and $\hat{\beta}_{FE}$ are consistent. $se(\hat{\beta}_{RE}) < se(\hat{\beta}_{FE})$ (efficiency)

If $Cov(\alpha_i, X_{it}) \neq 0$, only $\hat{\beta}_{FE}$ is consistent

$$W = \frac{\left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right)^2}{\sqrt{V(\hat{\beta}_{FE}) - V(\hat{\beta}_{RE})}} \sim \chi_1^2$$

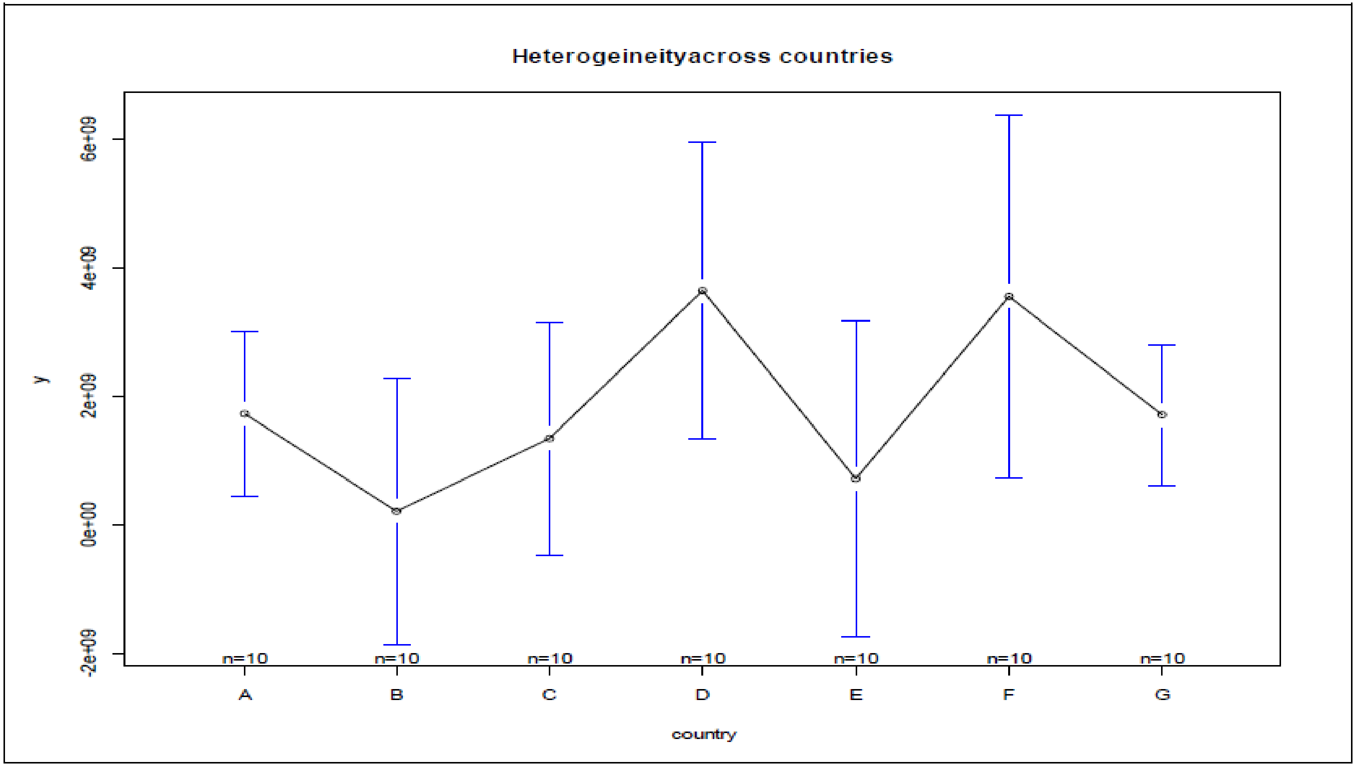
R code example

```
library(foreign)
Panel <-read.dta("http://dss.princeton.edu/training/Panel101.dta")
coplot(y ~ year|country, type="l", data=Panel) # Lines
coplot(y ~ year|country, type="b", data=Panel) # Points and lines
Panel[1:10,]
```

```
   country year           y y_bin         x1         x2          x3   opinion
1        A 1990  1342787840     1 0.27790365 -1.1079559  0.28255358 Str agree
2        A 1991 -1899660544     0 0.32068470 -0.9487200  0.49253848     Disag
3        A 1992   -11234363     0 0.36346573 -0.7894840  0.70252335     Disag
4        A 1993  2645775360     1 0.24614404 -0.8855330 -0.09439092     Disag
5        A 1994  3008334848     1 0.42462304 -0.7297683  0.94613063     Disag
6        A 1995  3229574144     1 0.47721413 -0.7232460  1.02968037 Str agree
7        A 1996  2756754176     1 0.49980500 -0.7815716  1.09228814     Disag
8        A 1997  2771810560     1 0.05162839 -0.7048455  1.41590083 Str agree
9        A 1998  3397338880     1 0.36641079 -0.6983712  1.54872274     Disag
10       A 1999    39770336     1 0.39584252 -0.6431540  1.79419804 Str disag
…
```

R code example

```
library(gplots)
plotmeans(y ~ country, main="Heterogeineityacross countries", data=Panel)
# plotmeansdraw a 95% confidence interval around the means
```



R code example

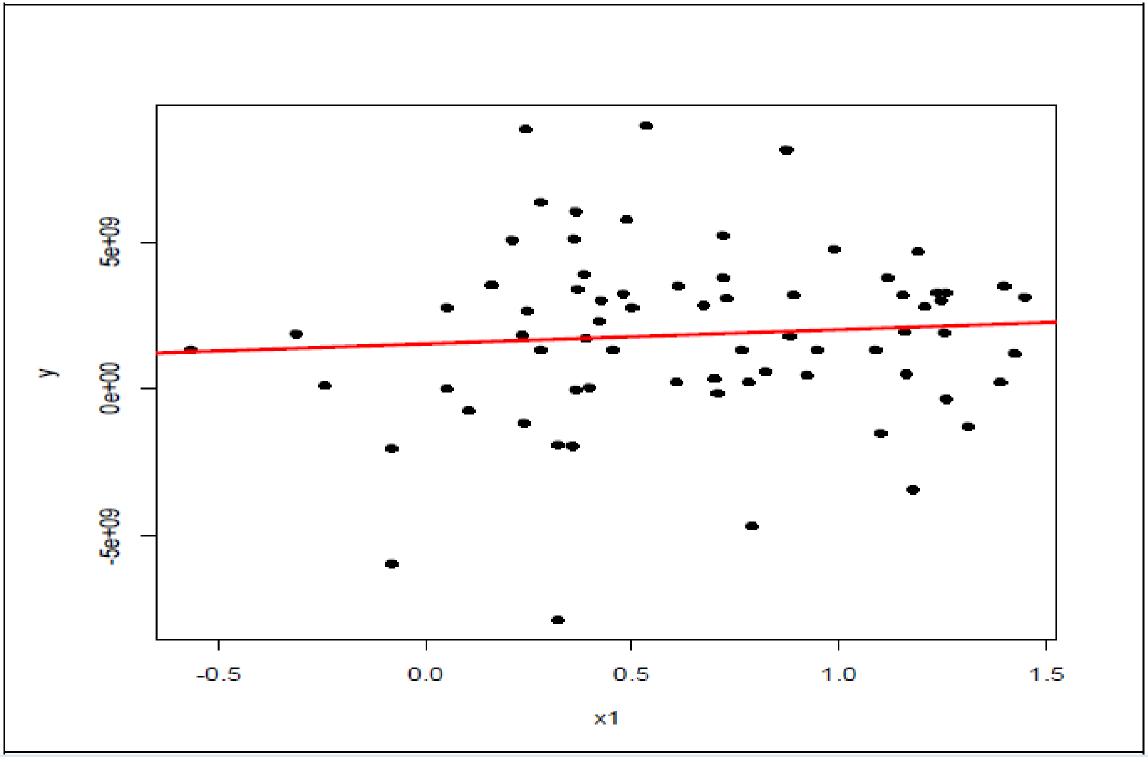
```
# POLS
ols<-lm(y ~ x1, data=Panel)
summary(ols)
```

```
Coefficients:
              Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   1.524e+09  6.211e+08    2.454     0.0167 *
x1            4.950e+08  7.789e+08    0.636     0.5272
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.028e+09 on 68 degrees of freedom
Multiple R-squared:  0.005905,  Adjusted R-squared:  -0.008714
F-statistic: 0.4039 on 1 and 68 DF,  p-value: 0.5272
```

R code example

```
yhat <-ols$fitted
plot(Panel$x1, Panel$y, pch=19, xlab="x1", ylab="y")
abline(lm(Panel$y~Panel$x1),lwd=3, col="red")
```



R code example

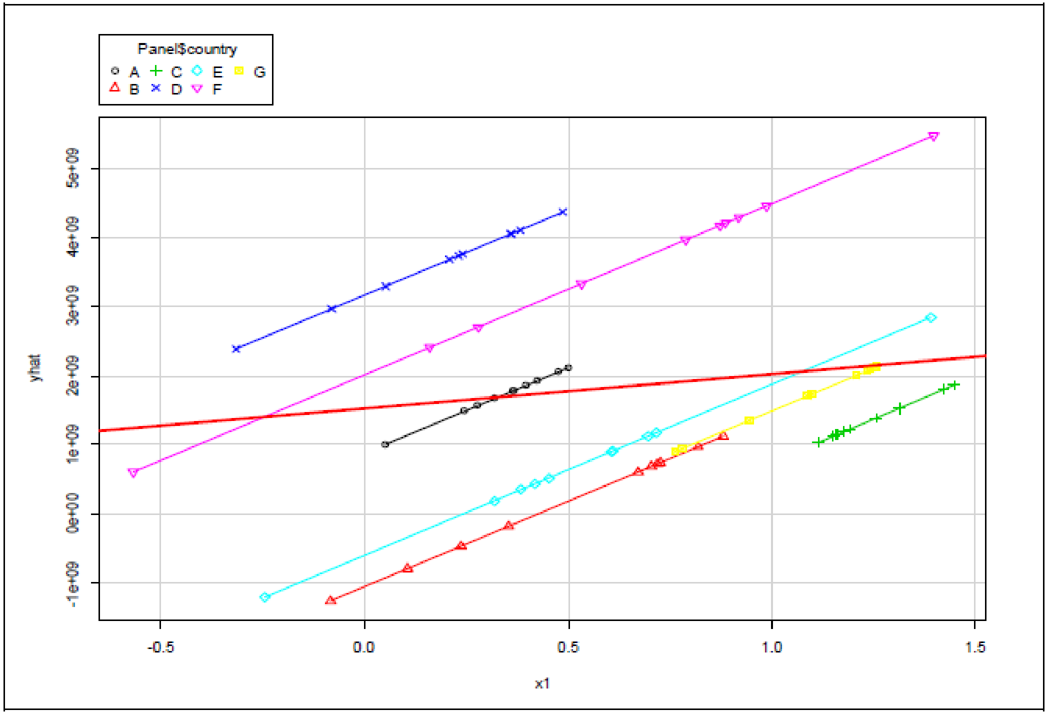
```
#LSDV
fixed.dum <-lm(y ~ x1 + factor(country) -1, data=Panel)
summary(fixed.dum)
```

```
lm(formula = y ~ x1 + factor(country) - 1, data = Panel)
Coefficients:
                    Estimate   Std. Error   t value   Pr(>|t|)
x1                  2.476e+09  1.107e+09    2.237     0.02889 *
factor(country)A    8.805e+08  9.618e+08    0.916     0.36347
factor(country)B   -1.058e+09  1.051e+09   -1.006     0.31811
factor(country)C   -1.723e+09  1.632e+09   -1.056     0.29508
factor(country)D    3.163e+09  9.095e+08    3.478     0.00093 ***
factor(country)E   -6.026e+08  1.064e+09   -0.566     0.57329
factor(country)F    2.011e+09  1.123e+09    1.791     0.07821 .
factor(country)G   -9.847e+08  1.493e+09   -0.660     0.51190
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.796e+09 on 62 degrees of freedom
Multiple R-squared:  0.4402,    Adjusted R-squared:  0.368
F-statistic: 6.095 on 8 and 62 DF,  p-value: 8.892e-06
```

R code example

```
yhat<-fixed.dum$fitted
library(car)
scatterplot(yhat~Panel$x1|Panel$country, boxplots=FALSE, xlab="x1", ylab="yhat",smooth=FALSE)
abline(lm(Panel$y~Panel$x1),lwd=3, col="red")
```



R code example

```
#POLS vs. LSDV
library(apsrtable)
apsrtable(ols,fixed.dum, model.names= c("OLS", "OLS_DUM")) # Displays a table in Latex form
```

```
%                & OLS                  & OLS_DUM              \\
(Intercept)      & 1524319070.05 ^*     &                      \\
                 & (621072623.86)       &                      \\
x1               & 494988913.90         & 2475617827.10 ^*\\
                 & (778861260.95)       & (1106675593.60) \\
```

R code example

```
# FE
library(plm)
fixed <-plm(y ~ x1, data=Panel, index=c("country", "year"), model="within")
summary(fixed)
```

```
plm(formula = y ~ x1, data = Panel, model = "within", index = c("country", "year"))
Balanced Panel: n=7, T=10, N=70

      Estimate     Std. Error    t-value   Pr(>|t|)
x1    2475617827   1106675594    2.237     0.02889 *
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Total Sum of Squares:    5.2364e+20
Residual Sum of Squares: 4.8454e+20
R-Squared:        0.074684
Adj. R-Squared: 0.066148
F-statistic: 5.00411 on 1 and 62 DF, p-value: 0.028892
```

R code example

```
fixef(fixed)          # Display the fixed effects (constants for each country)

        A           B           C          D          E          F          G
880542404 -1057858363 -1722810755 3162826897 -602622000 2010731793 -984717493

pFtest(fixed, ols)       # Testing for fixed effects,   null: OLS better than fixed

F test for individual effects
data:   y ~ x1
F = 2.9655, df1 = 6, df2 = 62, p-value = 0.01307
alternative hypothesis: significant effects
```

R code example

```
#RE
random <-plm(y ~ x1, data=Panel, index=c("country", "year"), model="random")
summary(random)
```

```
Oneway (individual) effect Random Effect Model
   (Swamy-Arora's transformation)

plm(formula = y ~ x1, data = Panel, model = "random", index = c("country", "year"))
Balanced Panel: n=7, T=10, N=70

                   var       std.dev      share
idiosyncratic   7.815e+18   2.796e+09    0.873
individual      1.133e+18   1.065e+09    0.127
theta:          0.3611
                Estimate     Std. Error    t-value   Pr(>|t|)
(Intercept)     1037014284   790626206     1.3116    0.1941
x1              1247001782   902145601     1.3823    0.1714

Total Sum of Squares:    5.6595e+20
Residual Sum of Squares: 5.5048e+20
R-Squared:         0.02733
Adj. R-Squared: 0.026549
F-statistic: 1.91065 on 1 and 68 DF, p-value: 0.17141
```

R code example

```
# Setting as panel data (an alternative way to run the above model)
Panel.set<-plm.data(Panel, index = c("country", "year"))
# Random effects using panel setting (same output as above)
random.set<-plm(y ~ x1, data = Panel.set, model="random")
summary(random.set)
```

#FE or RE?

Hausman test (H0: random effect model, H1: fixed effect model)

```
phtest(fixed, random)
```

```
Hausman Test

data:   y ~ x1
chisq = 3.674, df = 1, p-value = 0.05527
alternative hypothesis: one model is inconsistent
```

R code example

```
# Test for time fixed effect
library(plm)
fixed <-plm(y ~ x1, data=Panel, index=c("country", "year"), model="within")
fixed.time<-plm(y ~ x1 + factor(year), data=Panel, index=c("country", "year"), model="within")
summary(fixed.time)
```

	Estimate	Std. Error	t-value	Pr(>|t|)	
x1	1389050354	1319849567	1.0524	0.29738	
factor(year)1991	296381559	1503368528	0.1971	0.84447	
factor(year)1992	145369666	1547226548	0.0940	0.92550	
factor(year)1993	2874386795	1503862554	1.9113	0.06138	.
factor(year)1994	2848156288	1661498927	1.7142	0.09233	.
factor(year)1995	973941306	1567245748	0.6214	0.53698	
factor(year)1996	1672812557	1631539254	1.0253	0.30988	
factor(year)1997	2991770063	1627062032	1.8388	0.07156	.
factor(year)1998	367463593	1587924445	0.2314	0.81789	
factor(year)1999	1258751933	1512397632	0.8323	0.40898	

R code example

```
# Testing time-fixed effects. The null is that no time-fixed effects needed
pFtest(fixed.time, fixed)
```

```
F test for individual effects
data:   y ~ x1 + factor(year)
F = 1.209, df1 = 9, df2 = 53, p-value = 0.3094
alternative hypothesis: significant effects
```

```
plmtest(fixed, c("time"), type=("bp"))
```

```
Lagrange Multiplier Test - time effects (Breusch-Pagan)
data:   y ~ x1
chisq = 0.16532, df = 1, p-value = 0.6843
alternative hypothesis: significant effects
```

R code example

 # Testing for random effects: Breusch-Pagan Lagrange multiplier (LM)

```
# Regular OLS (pooling model) using plm
pool <-plm(y ~ x1, data=Panel, index=c("country", "year"), model="pooling")
summary(pool)
```

```
plm(formula = y ~ x1, data = Panel, model = "pooling", index = c("country", "year"))
Balanced Panel: n=7, T=10, N=70

              Estimate    Std. Error   t-value   Pr(>|t|)
(Intercept)   1524319070  621072624    2.4543    0.01668 *
x1            494988914   778861261    0.6355    0.52722
Total Sum of Squares:      6.2729e+20
Residual Sum of Squares: 6.2359e+20
R-Squared:        0.0059046
Adj. R-Squared:   0.0057359
F-statistic: 0.403897 on 1 and 68 DF, p-value: 0.52722
```

R code example

```
# Breusch-Pagan Lagrange Multiplier for random effects. Null is no panel effect (i.e. OLS better).
plmtest(pool, type=c("bp"))
```

```
Lagrange Multiplier Test - (Breusch-Pagan)
data:   y ~ x1
chisq = 2.6692, df = 1, p-value = 0.1023
alternative hypothesis: significant effects
```

```
#The LM test helps you decide between a random effects regression and a simple OLS regression.
# The null hypothesis in the LM test is that variances across entities is zero. This is, no evidence of
# significant difference across units (i.e. no panel effect)
```

R code example

 # Testing for cross-sectional dependence/contemporaneous correlation: using Breusch-Pagan LM test of independence and Pasaran CD tes

```
fixed <-plm(y ~ x1, data=Panel, index=c("country", "year"), model="within")
pcdtest(fixed, test = c("lm"))
```

```
Breusch-Pagan LM test for cross-sectional dependence in panels
data:    formula
chisq = 28.914, df = 21, p-value = 0.1161
alternative hypothesis: cross-sectional dependence
```

```
pcdtest(fixed, test = c("cd"))
```

```
Pesaran CD test for cross-sectional dependence in panels
data:    formula
z = 1.1554, p-value = 0.2479
alternative hypothesis: cross-sectional dependence
```

```
# The null hypothesis in the B-P/LM and Pasaran CD tests of independence is that residuals across
# entities are not correlated. B-P/LM and Pasaran CD (cross-sectional dependence) tests are used to
# test whether the residuals are correlated across entities*. Cross-sectional dependence can lead to
# bias in tests results (also called contemporaneous correlation).
```

R code example

 # Testing for serial correlation

```
pbgtest(fixed)
```

```
        Breusch-Godfrey/Wooldridge test for serial correlation in panel models

data:  y ~ x1
chisq = 14.137, df = 10, p-value = 0.1668
alternative hypothesis: serial correlation in idiosyncratic errors
```

```
# The null is that there is not serial correlation
```