ECON207 Session 10

Generalized Least Squares / Intro to Panel Data Models

Anthony Tay

This Version: 29 Oct 2024

Session 9 GLS / Panel Data Models

- OLS with Correlated Errors
    - Clustered Standard Errors
- Generalized Least Squares
    - Panel Data (Random Effects Models)
- Panel Data Models (Fixed Effects Model)

Correlated Errors

So far we have worked with regression model:

$$y = X\beta + \epsilon\,,\;\; E(\epsilon \mid X) = 0\,,\;\; E(\epsilon\epsilon^{\mathrm{T}} \mid X) = \sigma^2\Omega$$

where Ω is I_n or $\mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$

In this chapter, we consider the case where Ω is not diagonal

i.e., there is correlation between ϵ_i and ϵ_j where $i \neq j$

Correlated Errors

E.g. Effect of in-store promotion on customer spending

$$Spend_{i,s} = \beta_0 + \beta_1 Promo_{i,s} + \beta_2 Ctrl_{i,s} + \epsilon_{i,s}\ ,\ s = 1, \ldots, S\ ,\ i = 1, \ldots, n_s$$

$$\begin{bmatrix} Spend_{1,1} \\ \vdots \\ Spend_{n_1,1} \\ \hdashline Spend_{1,2} \\ \vdots \\ Spend_{n_2,2} \\ \hdashline \vdots \\ \hdashline Spend_{1,S} \\ \vdots \\ Spend_{n_S,S} \end{bmatrix} = \begin{bmatrix} 1 & Promo_{1,1} & Ctrl_{1,1} \\ \vdots & \vdots & \vdots \\ 1 & Promo_{n_1,1} & Ctrl_{n_1,1} \\ \hdashline 1 & Promo_{1,2} & Ctrl_{1,2} \\ \vdots & \vdots & \vdots \\ 1 & Promo_{n_2,2} & Ctrl_{n_2,2} \\ \hdashline & & \vdots \\ \hdashline 1 & Promo_{1,S} & Ctrl_{1,S} \\ \vdots & \vdots & \vdots \\ 1 & Promo_{n_S,S} & Ctrl_{n_S,S} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{1,1} \\ \vdots \\ \epsilon_{n_1,1} \\ \hdashline \epsilon_{1,2} \\ \vdots \\ \epsilon_{n_2,2} \\ \hdashline \vdots \\ \hdashline \epsilon_{1,S} \\ \vdots \\ \epsilon_{n_S,S} \end{bmatrix}$$

Correlated Errors

$$\begin{bmatrix} \underset{n_1\times 1}{y_1} \\ \underset{n_2\times 1}{y_2} \\ \vdots \\ \underset{n_S\times 1}{y_S} \end{bmatrix} = \begin{bmatrix} \underset{n_1\times 3}{X_1} \\ \underset{n_2\times 3}{X_2} \\ \vdots \\ \underset{n_S\times 3}{X_S} \end{bmatrix} \underset{3\times 1}{\beta} + \begin{bmatrix} \underset{n_1\times 1}{\epsilon_1} \\ \underset{n_2\times 1}{\epsilon_2} \\ \vdots \\ \underset{n_S\times 1}{\epsilon_S} \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} \underset{n_1\times 1}{y_1} \\ \underset{n_2\times 1}{y_2} \\ \vdots \\ \underset{n_S\times 1}{y_S} \end{bmatrix} = \begin{bmatrix} \underset{n_1\times 1}{X_1\beta} \\ \underset{n_2\times 1}{X_2\beta} \\ \vdots \\ \underset{n_S\times 1}{X_S\beta} \end{bmatrix} + \begin{bmatrix} \underset{n_1\times 1}{\epsilon_1} \\ \underset{n_2\times 1}{\epsilon_2} \\ \vdots \\ \underset{n_S\times 1}{\epsilon_S} \end{bmatrix}$$

or

$$y = X\beta + \varepsilon$$

where

$$y \text{ is } n \times 1\,,\ X \text{ is } n \times 3\,,\ \beta \text{ is } 3 \times 1\,,\ \epsilon \text{ is } n \times 1\,,\ \text{ and } n = n_1 + n_2 + \cdots + n_S$$

Correlated Errors

Suppose there are omitted store-specific factors that may affect customer behavior

- different store environments
- specific stores attracting specific kinds of customers

Then for any specific store $s = 1, \ldots, S$, the noise terms for customers i and j of that store, $\epsilon_{i,s}$ and $\epsilon_{j,s}$, are correlated, i.e.,

$$Var(\epsilon_s \mid X) = E(\epsilon_s \epsilon_s^{\mathrm{T}} \mid X) = \underset{n_s \times n_s}{\Omega} \text{ is not diagonal } \; s = 1, \ldots, S$$

We might still assume that noise terms associated with customers from different stores are uncorrelated

$$E(\epsilon_s \epsilon_{s'}^{\mathrm{T}} \mid X) \; = \underset{n_s \times n_{s'}}{0} \,, \; s, s' = 1, \ldots, S \,, \; s \neq s'$$

Correlated Errors

For the regression $y = X\beta + \varepsilon$, we have

$$Var(\varepsilon \mid X) = Var\left(\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_S \end{bmatrix}\right) = \begin{bmatrix} \Omega_1 & 0 & \dots & 0 \\ 0 & \Omega_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Omega_S \end{bmatrix} = \sigma^2\Omega \neq \sigma^2 I_n$$

As long as $E(\varepsilon \mid X) = 0$

- $\hat{\beta}^{ols}$ is linear, unbiased and consistent
- but no longer the "most efficient"
- $Var(\beta^{ols}) \neq \sigma^2(X^{\mathrm{T}}X)^{-1}$

Correlated Errors

Solution

- Stick with OLS, calculate appropriate variance-covariance matrix?
- More efficient estimator?

OLS estimator

$$\hat{\beta}^{ols} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = \left\{ \begin{bmatrix} X_1^{\mathrm{T}} & \dots & X_S^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_S \end{bmatrix} \right\}^{-1} \begin{bmatrix} X_1^{\mathrm{T}} & \dots & X_S^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_S \end{bmatrix}$$

$$= \left(\sum_{s=1}^{S} X_s^{\mathrm{T}} X_s \right)^{-1} \sum_{s=1}^{S} X_s^{\mathrm{T}} y_s$$

Correlated Errors

For variance-covariance matrix of $\hat{\beta}^{ols}$, we have

$$\hat{\beta}^{ols} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = \beta + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\varepsilon$$

$$\begin{aligned} Var(\hat{\beta}^{ols} \mid X) &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}E(\varepsilon\varepsilon^{\mathrm{T}} \mid X)X(X^{\mathrm{T}}X)^{-1} \\ &= (X^{\mathrm{T}}X)^{-1}\begin{bmatrix} X_1^{\mathrm{T}} & \dots & X_S^{\mathrm{T}} \end{bmatrix} E\left(\begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_S \end{bmatrix}\begin{bmatrix} \epsilon_1^{\mathrm{T}} & \dots & \epsilon_S^{\mathrm{T}} \end{bmatrix}\right)\begin{bmatrix} X_1 \\ \vdots \\ X_S^{\mathrm{T}} \end{bmatrix}(X^{\mathrm{T}}X)^{-1} \end{aligned}$$

"(One-way) cluster-robust variance-covariance matrix"

$$Var_{CSE}(\hat{\beta}^{ols}) = (X^{\mathrm{T}}X)^{-1}\sum_{s=1}^{S} X_s^{\mathrm{T}}\hat{\epsilon}_s\hat{\epsilon}_s^{\mathrm{T}}X_s(X^{\mathrm{T}}X)^{-1}$$

Correlated Errors

- Heteroskedasticity-Robust (HC) Var-Cov matrices are special cases
  - every observation is its own cluster
- Cluster-Robust Var-Cov matrices allow for heteroskedasticity
- Can cluster over two variables C_s and C_p
  - $s = 1, \ldots, S$ and $p = 1, \ldots, P$
  - Let (s, p) represent a group of error terms associated with $(C_s = s, C_p = p)$

Correlated Errors

	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)	(3,1)	(3,2)	(3,3)
(1,1)	*, @	*	*	@			@		
(1,2)	*	*, @	*		@			@	
(1,3)	*	*	*, @			@			@
(2,1)	@			*, @	*	*	@		
(2,2)		@		*	*, @	*		@	
(2,3)			@	*	*	*, @			@
(3,1)	@			@			*, @	*	*
(3,2)		@			@		*	*, @	*
(3,3)			@			@	*	*	*, @

$$Var_{CSE(S,P)}(\hat{\beta}^{ols}) = Var_{CSE(S)}(\hat{\beta}^{ols}) + Var_{CSE(P)}(\hat{\beta}^{ols}) - Var_{CSE(S\&P)}(\hat{\beta}^{ols})$$

Correlated Errors

Two-way clustering often done in context of panel data

- the same cross-section of individual entities observed over time
- data in `promo2.csv` has panel structure
- clustered over individuals and over time

Two-way clustering can also be done on non-panel data

- as long as there are two variable to be grouped over
- application involving loans, bank_id, region_id, interest rates, borrower risk
- cluster over bank_id and region_id

Correlated Errors

Example: Effect of Discounts on Sales (Dataset: promo2.csv)

```
library(tidyverse); library(fpp3); library(ggfortify); library(patchwork)
library(fixest); library(plm); library(sandwich); library(lmtest)

df <- read_csv("data\\promo2.csv", show_col_types = FALSE)
glimpse(df)
```

```
Rows: 80,820
Columns: 9
$ store_id      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ date          <date> 2024-03-28, 2024-03-29, 2024-03-30, 2024-03-31, 2024-04~
$ sales         <dbl> 51.31, 55.42, 89.81, 91.00, 56.58, 66.70, 52.16, 78.27, ~
$ promo         <dbl> 0.05441892, 0.04934865, 0.06047297, 0.06102703, 0.060283~
$ avail_daytime <dbl> 0.7559122, 0.8209459, 0.8158784, 0.7685811, 0.8015203, 0~
$ holiday       <dbl> 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0,~
$ precpt        <dbl> 1.6999, 3.0190, 2.0942, 1.5618, 3.5386, 3.1459, 1.7165, ~
$ temp          <dbl> 15.48, 15.08, 15.91, 16.13, 15.37, 15.69, 16.11, 16.08, ~
$ log_sales     <dbl> 3.957188, 4.032824, 4.508769, 4.521789, 4.053175, 4.2150~
```

Correlated Errors

One quarter of data (90 days)

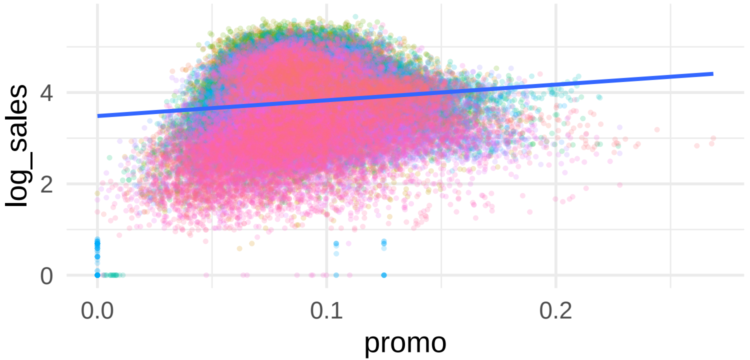
- `log_sales`: total daily (log) sales
- `promo`: average discount over all products
- `avail_daytime`: proportion of hours from 0600hrs to 2200hrs where stock is available
- `precpt`, `temp`: average precipitation and temperature

Relationship of interest is `log_sales` vs `promo`

```
p1 <- df %>% ggplot(aes(x=promo, y=log_sales, color=factor(store_id))) +
  geom_point(alpha=0.2, size=0.3) +
  geom_smooth(aes(color=NULL), method = "lm", se=FALSE, linewidth=0.7) +
  theme_minimal() + theme(legend.position = "none")
```

Correlated Errors

```
p1
```



Correlated Errors

Estimated equation:

$$log\_sales = \beta_0 + \beta_1 promo + \beta_2 avail\_daytime + \beta_3 holiday + \beta_4 precpt + \beta_5 temp + \epsilon$$

```
m1 <- feols(log_sales ~ promo + avail_daytime + holiday + precpt + temp, data = df)
```

Correlated Errors

```
s1 <- summary(m1)
s2 <- summary(m1, vcov = "HC1")
s3 <- summary(m1, vcov = ~ store_id)
s4 <- summary(m1, vcov = ~ date)
s5 <- summary(m1, vcov = ~ store_id + date)
results <- as.data.frame(s1$coeftable[,c("Estimate", "Std. Error")]) %>%
  full_join(as.data.frame(s2$coeftable[,c("Estimate", "Std. Error")]), by="Estimate") %>%
  full_join(as.data.frame(s3$coeftable[,c("Estimate", "Std. Error")]), by="Estimate") %>%
  full_join(as.data.frame(s4$coeftable[,c("Estimate", "Std. Error")]), by="Estimate") %>%
  full_join(as.data.frame(s5$coeftable[,c("Estimate", "Std. Error")]), by="Estimate") %>%
  as.matrix()

colnames(results) <- c("Estimate", "SE(iid)", "HC1", "CSE(store_id)", "CSE(date)", "CSE(2-way)")
rownames(results) <- rownames(s1$coeftable)
```

Correlated Errors

```
results %>% round(4)
```

```
              Estimate SE(iid)    HC1 CSE(store_id) CSE(date) CSE(2-way)
(Intercept)     2.3942  0.0285 0.0364        0.1517    0.1703     0.2256
promo           3.8704  0.0967 0.1204        0.7494    0.7656     1.0671
avail_daytime   2.3786  0.0339 0.0456        0.1794    0.1868     0.2556
holiday         0.2760  0.0053 0.0052        0.0051    0.0431     0.0431
precpt          0.0058  0.0007 0.0007        0.0037    0.0075     0.0083
temp           -0.0422  0.0008 0.0009        0.0031    0.0050     0.0058
```

Generalized Least Squares

Suppose that

$$y = X\beta + \epsilon\,,\;\; E(\epsilon \mid X) = 0\,,\;\; E(\epsilon\epsilon^{\mathrm{T}} \mid X) = \sigma^2\Omega$$

where Ω is an $n \times n$ positive-definite matrix not equal to I_n

- Assume that σ^2 is unknown
- Assume (for the moment) that Ω is **known**

OLS estimator is unbiased / consistent but not efficient

GLS estimator provides a more efficient estimator

Generalized Least Squares

Because Ω is positive definite, we can find non-singular $n \times n$ matrix P such that

$$P\Omega P^{\mathrm{T}} = I_n$$

- Since Ω is known, P is known
- Furthermore $P^{-1} = P^{\mathrm{T}}$, so we have $P^{\mathrm{T}}P = \Omega^{-1}$

This comes from eigendecomposition of Ω

Pre-multiply regression equation by P

$$Py = PX\beta + P\epsilon \quad \text{or} \quad y^* = X^*\beta + \epsilon^*$$

Generalized Least Squares Theory

The noise term in the modified equation satisfies

$$E(\epsilon^* \mid X) = E(P\epsilon \mid X) = PE(\epsilon \mid X) = 0$$

$$E(\epsilon^* {\epsilon^*}^{\mathrm{T}} \mid X) = E(P\epsilon\epsilon^{\mathrm{T}} P^{\mathrm{T}} \mid X) = PE(\epsilon\epsilon^{\mathrm{T}} \mid X)P^{\mathrm{T}} = \sigma^2 P\Omega P^{\mathrm{T}} = \sigma^2 I_n$$

Since $y^* = X^*\beta + \epsilon^*$ satisfies required conditions for OLS to give BLU Estimators

$$\hat{\beta}^{gls} = ({X^*}^{\mathrm{T}} X^*)^{-1} {X^*}^{\mathrm{T}} y^* = (X^{\mathrm{T}} P^{\mathrm{T}} P X)^{-1} X^{\mathrm{T}} P^{\mathrm{T}} P y = (X^{\mathrm{T}} \Omega^{-1} X)^{-1} X^{\mathrm{T}} \Omega^{-1} y$$

is BLU. We have

$$Var(\beta^{gls} \mid X) = \sigma^2 ({X^*}^{\mathrm{T}} X^*)^{-1} = \sigma^2 (X^{\mathrm{T}} P^{\mathrm{T}} P X)^{-1} = \sigma^2 (X^{\mathrm{T}} \Omega^{-1} X)^{-1}$$

Generalized Least Squares Theory

Remarks

- An unbiased estimator for σ^2 is

$$\widehat{\sigma_*^2} = \frac{\widehat{\epsilon^*}^{\mathrm{T}} \widehat{\epsilon^*}}{n-k-1} \quad \text{where} \quad \widehat{\epsilon^*} = y^* - X^* \hat{\beta}^{gls}$$

- Weighted Least Squares is a special case where

$$\Omega = \mathrm{diag}(\eta_1, \eta_2, \ldots, \eta_n)$$

Appropriate P is

$$P = \mathrm{diag}(\eta_1^{-1/2}, \eta_2^{-1/2}, \ldots, \eta_n^{-1/2})$$

Generalized Least Squares Theory

- GLS is equivalent to

$$\hat{\beta}^{gls} = \underset{\hat{\beta}}{\arg\min}(y - X\hat{\beta})^{\mathrm{T}}\Omega^{-1}(y - X\hat{\beta})$$

since

$$\begin{aligned}\hat{\beta}^{gls} &= \underset{\hat{\beta}}{\arg\min}(y^* - X^*\hat{\beta})^{\mathrm{T}}(y^* - X^*\hat{\beta}) = \underset{\hat{\beta}}{\arg\min}(Py - PX\hat{\beta})^{\mathrm{T}}(Py - PX\hat{\beta}) \\ &= \underset{\hat{\beta}}{\arg\min}(P(y - X\hat{\beta}))^{\mathrm{T}}P(y - X\hat{\beta}) = \underset{\hat{\beta}}{\arg\min}(y - X\hat{\beta})^{\mathrm{T}}P^{\mathrm{T}}P(y - X\hat{\beta}) \\ &= \underset{\hat{\beta}}{\arg\min}(y - X\hat{\beta})^{\mathrm{T}}\Omega^{-1}(y - X\hat{\beta})\end{aligned}$$

Application of GLS to Heteroskedasticity

Suppose

$$y = X\beta + \epsilon\ ,\ \ \epsilon \,|\, X \sim (0_n\ ,\ \sigma^2\Omega)$$

Under heteroskedasticity only, we have

$$\Omega = \begin{bmatrix} \eta_1 & 0 & \dots & 0 \\ 0 & \eta_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \eta_n \end{bmatrix} \quad \text{and} \quad P = \begin{bmatrix} \eta_1^{-1/2} & 0 & \dots & 0 \\ 0 & \eta_2^{-1/2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \eta_n^{-1/2} \end{bmatrix}$$

Straightforward to verify that

$$P\Omega P^{\mathrm{T}} = I_n \quad \text{and} \quad P^{\mathrm{T}}P = \Omega^{-1}$$

Application of GLS to Heteroskedasticity

Furthermore

$$Py = \begin{bmatrix} \eta_1^{-1/2} & 0 & \dots & 0 \\ 0 & \eta_2^{-1/2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \eta_n^{-1/2} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} y_1/\eta_1^{1/2} \\ y_2/\eta_2^{1/2} \\ \vdots \\ y_n/\eta_n^{1/2} \end{bmatrix}$$

$$PX = \begin{bmatrix} \eta_1^{-1/2} & 0 & \dots & 0 \\ 0 & \eta_2^{-1/2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \eta_n^{-1/2} \end{bmatrix} \begin{bmatrix} X_{1*} \\ X_{2*} \\ \vdots \\ X_{n*} \end{bmatrix} = \begin{bmatrix} (1/\eta_1^{1/2})X_{1*} \\ (1/\eta_2^{1/2})X_{2*} \\ \vdots \\ (1/\eta_n^{1/2})X_{n*} \end{bmatrix}$$

Application of GLS to Autocorrelated Errors

If

$$y_t = X_{t*}\beta + u_t\,,\;\; u_t = \rho u_{t-1} + \epsilon_t\,,\; \epsilon \overset{iid}{\sim} (0,\sigma^2)\,,\; |\rho| < 1\,,\; t = 1, \dots, T$$

the model can be written as $y = X\beta + u$ where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix},\; X = \begin{bmatrix} X_{1*} \\ X_{2*} \\ \vdots \\ X_{T*} \end{bmatrix},\; u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{bmatrix},\; u \sim (0, \Omega)\,,$$

$$\Omega = \frac{\sigma^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \dots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \dots & 1 \end{bmatrix}$$

Application of GLS to Autocorrelated Errors

The appropriate transformation matrix is

$$P = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \dots & 0 & 0 \\ -\rho & 1 & 0 & \dots & 0 & 0 \\ 0 & -\rho & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -\rho & 1 \end{bmatrix}$$

It is left as an exercise to

- show that $P\Omega P^{\mathrm{T}} = \sigma^2 I_n$
- show what the transformation Py and PX is

Application of GLS to Panel Data Structures

Suppose we have

$$Y_{i,t} = \beta_0 + \beta_1 X_{i,t,1} + \cdots + \beta_K X_{i,t,K} + \underbrace{\alpha_i + u_{i,t}}_{\epsilon_{i,t}}, \;\; i = 1, \ldots, N \,; \;\; t = 1, \ldots, T$$

Assume α_i and $u_{i,t}$ independent random variables with

- α_i zero-mean with variance σ_α^2 (think of $E(\alpha_i)$ as having been absorbed into β_0)
- $u_{i,t}$ zero-mean with variance σ_u^2
  - $Var(\epsilon_{i,t}) = \sigma_\alpha^2 + \sigma_u^2$
  - $Cov(\epsilon_{i,t}, \epsilon_{i,s}) = E((\alpha_i + u_{i,t})(\alpha_i + u_{i,s})) = \sigma_\alpha^2$ for $s \neq t$
  - $Cov(\epsilon_{i,t}, \epsilon_{j,s}) = E((\alpha_i + u_{i,t})(\alpha_j + u_{j,s})) = 0$ for all $i \neq j$, all s, t

Application of GLS to Panel Data Structures

Setting up using matrix algebra

For any given individual $i = 1, \dots, N$, let

$$\underset{T\times 1}{y_i} = \begin{bmatrix} Y_{i,1} \\ Y_{i,2} \\ \vdots \\ Y_{i,T} \end{bmatrix}, \quad \underset{T\times(K+1)}{X_i} = \begin{bmatrix} 1 & X_{i,1,1} & \dots & X_{i,1,K} \\ 1 & X_{i,2,1} & \dots & X_{i,2,K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{i,T,1} & \dots & X_{i,T,K} \end{bmatrix}, \quad \underset{T\times 1}{\epsilon_i} = \begin{bmatrix} \epsilon_{i,1} \\ \epsilon_{i,2} \\ \vdots \\ \epsilon_{i,T} \end{bmatrix}$$

The regression for the ith individual is

$$y_i = X_i\beta + \epsilon_i$$

Application of GLS to Panel Data Structures

Stacking up the samples for each individual, we get the "pooled regression model"

$$y = X\beta + \epsilon$$

where $\underset{NT\times 1}{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad \underset{NT\times(K+1)}{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}, \quad \underset{NT\times 1}{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}$

Pooled OLS estimator:

$$\hat{\beta}^{pols} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$$

Application of GLS to Panel Data Structures

Then $\underset{(T\times T)}{Var(\epsilon_i \mid X)} = E(\epsilon_i\epsilon_i^{\mathrm{T}} \mid X)$

$$= \begin{bmatrix} \sigma_u^2+\sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_u^2+\sigma_\alpha^2 & \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_u^2+\sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & \sigma_u^2+\sigma_\alpha^2 \end{bmatrix} = \sigma_u^2 \underbrace{\left(I_T + \frac{\sigma_\alpha^2}{\sigma_u^2} i_T i_T^{\mathrm{T}}\right)}_{\text{"}\Omega\text{"}}$$

$$\underset{(NT\times NT)}{Var(\epsilon \mid X)} = E(\epsilon\epsilon^{\mathrm{T}} \mid X) = \begin{bmatrix} E(\epsilon_1\epsilon_1^{\mathrm{T}}) & E(\epsilon_1\epsilon_2^{\mathrm{T}}) & \dots & E(\epsilon_1\epsilon_N^{\mathrm{T}}) \\ E(\epsilon_2\epsilon_1^{\mathrm{T}}) & E(\epsilon_2\epsilon_2^{\mathrm{T}}) & \dots & E(\epsilon_2\epsilon_N^{\mathrm{T}}) \\ \vdots & \vdots & \ddots & \vdots \\ E(\epsilon_N\epsilon_1^{\mathrm{T}}) & E(\epsilon_N\epsilon_2^{\mathrm{T}}) & \dots & E(\epsilon_N\epsilon_N^{\mathrm{T}}) \end{bmatrix} = \sigma_u^2 \begin{bmatrix} \Omega & 0 & \dots & 0 \\ 0 & \Omega & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Omega \end{bmatrix}$$

Application of GLS to Panel Data Structures

What is the appropriate P? Consider sample for just the ith individual:

$$y_i = X_i\beta + \epsilon_i\,,\; Var(\epsilon_i \mid X) = \sigma_u^2 \left(I_T + \frac{\sigma_\alpha^2}{\sigma_u^2} i_T i_T^{\mathrm{T}} \right) = \sigma_u^2 \Omega$$

Let

$$P = M_0 + \psi(I_T - M_0)$$

where $M_0 = I_T - i_T(i_T^{\mathrm{T}} i_T)^{-1} i_T^{\mathrm{T}}$ and $\psi = \dfrac{\sigma_u}{\sqrt{\sigma_u^2 + T\sigma_\alpha^2}}$

Then $E(P\epsilon_i\epsilon_i^{\mathrm{T}}P^{\mathrm{T}}) = \sigma_u^2 I_T$

Application of GLS to Panel Data Structures

Proof: M_0 and $I - M_0$ are both symmetric, idempotent, and $M_0(I - M_0) = (I - M_0)M_0 = 0$

Furthermore, for any non-zero scalar ψ, we have

$$(M_0 + \psi(I_T - M_0))\left(M_0 - \frac{1}{\psi^2}(I - M_0)\right)(M_0 + \psi(I_T - M_0))^{\mathrm{T}} = \; I_T \;\; \text{(exercise!)}$$

Finally, we have

$$\begin{aligned}
\Omega = I_T + \frac{\sigma_\alpha^2}{\sigma_u^2} i_T i_T^{\mathrm{T}} &= I_T + \frac{T\sigma_\alpha^2}{\sigma_u^2} i_T (i_T^{\mathrm{T}} i_T)^{-1} i_T^{\mathrm{T}} \\
&= I_T - i_T (i_T^{\mathrm{T}} i_T)^{-1} i_T^{\mathrm{T}} + i_T (i_T^{\mathrm{T}} i_T)^{-1} i_T^{\mathrm{T}} + \frac{T\sigma_\alpha^2}{\sigma_u^2} i_T (i_T^{\mathrm{T}} i_T)^{-1} i_T^{\mathrm{T}} \\
&= M_0 + \left(1 + \frac{T\sigma_\alpha^2}{\sigma_u^2}\right) i_T (i_T^{\mathrm{T}} i_T)^{-1} i_T^{\mathrm{T}} \\
&= M_0 + \left(\frac{\sigma_u^2 + T\sigma_\alpha^2}{\sigma_u^2}\right) i_T (i_T^{\mathrm{T}} i_T)^{-1} i_T^{\mathrm{T}} = M_0 + \frac{1}{\psi^2}(I_T - M_0) \;\; \text{where} \;\; \psi = \sqrt{\frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}}
\end{aligned}$$

Application of GLS to Panel Data Structures

It follows that if $Var(\epsilon_i) = \sigma_u^2 \Omega$ where

$$\Omega = I_T + \frac{\sigma_\alpha^2}{\sigma_u^2} i_T i_T^{\mathrm{T}} = M_0 + \frac{1}{\psi^2}(I_T - M_0) \quad \text{where} \quad \psi = \sqrt{\frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}}$$

Then

$$Var(P\epsilon_i) = P\,Var(\epsilon_i)P^T = \sigma_u^2 P \Omega P^{\mathrm{T}} = \sigma_u^2 I_T$$

Apply the transformation to $y_i = X_i\beta + \epsilon_i$ for every i and then pool the transformed regressions, which gives

$$\begin{bmatrix} Py_1 \\ Py_2 \\ \vdots \\ Py_N \end{bmatrix} = \begin{bmatrix} PX_1 \\ PX_2 \\ \vdots \\ PX_N \end{bmatrix} \beta + \begin{bmatrix} P\epsilon_1 \\ P\epsilon_2 \\ \vdots \\ P\epsilon_N \end{bmatrix} \quad \text{or} \quad y^* = X^*\beta + \epsilon^*, \;\; Var(\epsilon^* \mid X) = \sigma_u^2 I_{NT}$$

Application of GLS to Panel Data Structures

What is this transformation?

$$\begin{aligned} P y_i &= (M_0 + \psi(I_T - M_0))y_i \\ &= \left(I_T - i_T(i_T^{\mathrm{T}} i_T)^{-1} i_T^{\mathrm{T}} + \psi i_T(i_T^{\mathrm{T}} i_T)^{-1} i_T^{\mathrm{T}}\right) y_i \\ &= \left(I_T - (1-\psi) i_T(i_T^{\mathrm{T}} i_T)^{-1} i_T^{\mathrm{T}}\right) y_i \\ &= y_i - \lambda\, i_T\, \overline{Y_i} \end{aligned}$$

where $\lambda = 1 - \psi = 1 - \sqrt{\dfrac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}}$

Application of GLS to Panel Data Structures

i.e., $Py_i = y_i^* = \begin{bmatrix} Y_{i,1} - \lambda\overline{Y_i} \\ Y_{i,2} - \lambda\overline{Y_i} \\ \vdots \\ Y_{i,T} - \lambda\overline{Y_i} \end{bmatrix}$ where $\lambda = 1 - \dfrac{\sigma_u}{\sqrt{\sigma_u^2 + T\sigma_\alpha^2}}$

Same for other regressors $X_i^* = PX_i$

- To operationalize this theory, we have to estimate σ_u^2 and σ_α^2
- We can obtain these from the Pooled OLS residuals (details omitted)
- Assumptions on Ω more restrictive than in pooled OLS with Clustered S.E.
- This GLS estimator is called the **random effects estimator** for panel data

Application of GLS to Panel Data Structures

Other versions:

- Time effects only

$$Y_{i,t} = \beta_0 + \beta_1 X_{i,t,1} + \cdots + \beta_K X_{i,t,K} + \underbrace{\tau_t + u_{i,t}}_{\epsilon_{i,t}}, \; i = 1, \dots, N; \; t = 1, \dots, T$$

- Two-way effects

$$Y_{i,t} = \beta_0 + \beta_1 X_{i,t,1} + \cdots + \beta_K X_{i,t,K} + \underbrace{\alpha_i + \tau_t + u_{i,t}}_{\epsilon_{i,t}}, \; i = 1, \dots, N; \; t = 1, \dots, T$$

Application of GLS to Panel Data Structures

Effect: individual (store)

```
mdl_re <- plm(log_sales ~ promo + avail_daytime + holiday + precpt + temp,
              data = df, model="random", effect="individual",
              index=c("store_id", "date"))
summary(mdl_re)$coefficients %>% round(4)
```

```
              Estimate Std. Error  z-value Pr(>|z|)
(Intercept)     2.8086     0.0180 155.9733        0
promo           1.9368     0.0371  52.1864        0
avail_daytime   0.3044     0.0090  33.9659        0
holiday         0.2518     0.0013 191.6816        0
precpt          0.0046     0.0002  23.5133        0
temp            0.0204     0.0002  91.0024        0
```

Application of GLS to Panel Data Structures

Effect: time (date)

```
mdl_re <- plm(log_sales ~ promo + avail_daytime + holiday + precpt + temp,
              data = df, model="random", effect="time",
              index=c("store_id", "date"))
summary(mdl_re)$coefficients %>% round(4)
```

```
              Estimate Std. Error   z-value Pr(>|z|)
(Intercept)     4.7382     0.0409  115.7298        0
promo           2.5932     0.0939   27.6044        0
avail_daytime   2.3372     0.0359   65.0588        0
holiday         0.2766     0.0191   14.4555        0
precpt          0.0187     0.0007   27.0225        0
temp           -0.1409     0.0012 -116.6439        0
```

Application of GLS to Panel Data Structures

Effect: two-ways

```
mdl_re <- plm(log_sales ~ promo + avail_daytime + holiday + precpt + temp,
              data = df, model="random", effect="twoways",
              index=c("store_id", "date"))
summary(mdl_re)$coefficients %>% round(4)
```

```
              Estimate Std. Error  z-value Pr(>|z|)
(Intercept)     3.2293     0.0240 134.3754   0.0000
promo           1.7901     0.0356  50.2840   0.0000
avail_daytime   0.3428     0.0090  37.9095   0.0000
holiday         0.2520     0.0189  13.3284   0.0000
precpt          0.0051     0.0002  25.9518   0.0000
temp            0.0011     0.0005   2.2254   0.0261
```

Fixed Effects Estimator for Panel Data

One benefit of having a dataset with panel data structure is that it can help deal with the endogeneity problem. Suppose again that

$$\begin{aligned} Y_{i,t} &= \beta_0 + \beta_1 X_{i,t,1} + \cdots + \beta_K X_{i,t,K} + \alpha_i + u_{i,t} \\ &= \beta_0 + \beta_1 X_{i,t,1} + \cdots + \beta_K X_{i,t,K} + \epsilon_{i,t}\ , \ \ i = 1, \ldots, N\ ;\ \ t = 1, \ldots, T \end{aligned}$$

Now suppose α_i is some **unobserved** variable or combination of variables, a.k.a. "time invariant individual effects" that is correlated with some of the regressors

- α_i is subsumed into error term
- if α_i is correlated with included regressors, $\hat{\beta}_1^{ols}$ will be inconsistent
- in returns to schooling application, α_i might include ability

Fixed Effects Estimator for Panel Data

The panel data structure allows us to remove the individual effects

For example, if we "time-demean" the sample, we get

$$
\begin{aligned}
Y_{i,t} &= \beta_0 + \beta_1 X_{i,t,1} + \cdots + \beta_K X_{i,t,K} + \alpha_i + u_{i,t} \\
\overline{Y_i} &= \beta_0 + \beta_1 \overline{X_{i,1}} + \cdots + +\beta_1 \overline{X_{i,K}} + \alpha_i + \overline{u_i} \\
Y_{i,t} - \overline{Y_i} &= \beta_1 (X_{i,t,1} - \overline{X_{i,1}}) + \cdots + \beta_K (X_{i,t,K} - \overline{X_{i,K}}) + u_{i,t} - \overline{u_i} \\
\ddot{Y}_{i,t} &= \beta_1 \ddot{X}_{i,t,1} + \cdots + \beta_K \ddot{X}_{i,t,K} + \ddot{u}_{i,t}
\end{aligned}
$$

The unobserved individual effect has been removed

OLS on the time-demeaned equation gives the **Fixed Effect Estimator**

Fixed Effects Estimator for Panel Data

For every i, take $\ddot{y}_i = M_0 y_i$, $\ddot{X}_i = M_0 X_i$, $\ddot{u}_i = M_0 u_i$

$$\underset{(NT\times 1)}{\ddot{y}} = \begin{bmatrix} \ddot{y}_1 \\ \ddot{y}_2 \\ \vdots \\ \ddot{y}_N \end{bmatrix} \quad \underset{(NT\times K)}{\ddot{X}} = \begin{bmatrix} \ddot{X}_1 \\ \ddot{X}_2 \\ \vdots \\ \ddot{X}_N \end{bmatrix} \quad \underset{(NT\times 1)}{\ddot{u}} = \begin{bmatrix} \ddot{u}_1 \\ \ddot{u}_2 \\ \vdots \\ \ddot{u}_N \end{bmatrix}$$

where $M_0 = I_T - i_T(i_T^{\mathrm{T}} i_T) i_T^{\mathrm{T}}$

The transformed model is

$$\ddot{y} = \ddot{X}\beta + \ddot{u}$$

Estimate by OLS

$$\hat{\beta}^{fe} = (\ddot{X}^{\mathrm{T}}\ddot{X})^{-1}\ddot{X}^{\mathrm{T}}\ddot{y}$$

Fixed Effects Estimator for Panel Data

Remarks:

- Also called "Within Estimator"
- Cannot include any time invariant variables such as race (will get removed along with unobserved individual effect)
- Identical to "Least Squares Dummy Variable" Models (LSDV)

$$Y_{i,t} = \theta_1 d_{1,i,t} + \theta_2 d_{2,i,t} + \cdots + \theta_N d_{N,i.t} + \beta_1 X_{1,i,t} + \ldots \beta_K X_{K,i,t} + u_{i,t}$$

  - proof omitted
  - Intuition: individual dummies capture individual time effects
- Often N is very large, so direct OLS on the LSDV model may be infeasible

First Difference Estimator for Panel Data Structure

Remarks (continued):

- Alternative to FE estimator: First Difference estimator

$$
\begin{aligned}
Y_{i,t} &= \beta_1 X_{1,i,t} + \ldots \beta_K X_{K,i,t} + \alpha_i + u_{i,t} \\
Y_{i,t-1} &= \beta_1 X_{1,i,t-1} + \ldots \beta_K X_{K,i,t-1} + \alpha_i + u_{i,t-1} \\
\Delta Y_{i,t} &= \beta_1 \Delta X_{1,i,t} + \ldots \beta_K \Delta X_{K,i,t} + \Delta u_{i,t}
\end{aligned}
$$

  Estimate by OLS

  - This approach also removes the unobserved individual effects
  - If $T = 2$, then FD = FE
  - If there are missing observations, FD can make many observations become unusable

Example 1

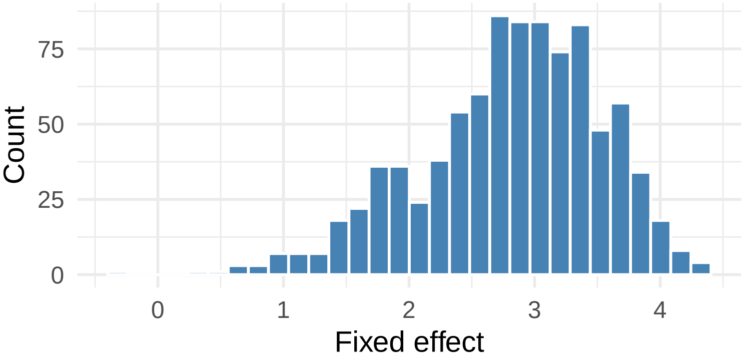
Effect: Individual (store_id)

```
mdl_fe <- plm(log_sales ~ promo + avail_daytime + holiday + precpt + temp,
              data = df, model="within", effect="individual",
              index=c("store_id", "date"))
summary(mdl_fe)$coefficients %>% round(4)
```

```
              Estimate Std. Error  t-value Pr(>|t|)
promo           1.9266     0.0369  52.2078        0
avail_daytime   0.3008     0.0089  33.7797        0
holiday         0.2518     0.0013 192.9478        0
precpt          0.0047     0.0002  23.7723        0
temp            0.0206     0.0002  92.1585        0
```

Example 1

```
fe <- fixef(mdl_fe)
fe_df <- data.frame(id = names(fe), effect = fe)
fe_df %>% ggplot(aes(x=effect)) +
  geom_histogram(color="white", fill="steelblue", bins=30) +
  labs(x="Fixed effect", y="Count") + theme_minimal()
```



Example 1

Effect: Time (date)

```
mdl_fe <- plm(log_sales ~ promo + avail_daytime + holiday + precpt + temp,
              data = df, model="within", effect="time",
              index=c("store_id", "date"))
summary(mdl_fe)$coefficients %>% round(4)
```

```
              Estimate Std. Error   t-value Pr(>|t|)
promo           2.2542     0.0927   24.3051        0
avail_daytime   2.2029     0.0358   61.5719        0
precpt          0.0222     0.0007   32.3479        0
temp           -0.1664     0.0013 -130.0195        0
```

Example 1

Effect: Two-way

```
mdl_fe <- plm(log_sales ~ promo + avail_daytime + holiday + precpt + temp,
              data = df, model="within", effect="twoway",
              index=c("store_id", "date"))
summary(mdl_fe)$coefficients %>% round(4)
```

```
              Estimate Std. Error t-value Pr(>|t|)
promo           1.7805     0.0354 50.2691   0.0000
avail_daytime   0.3392     0.0090 37.7182   0.0000
precpt          0.0052     0.0002 26.1550   0.0000
temp            0.0011     0.0005  2.0914   0.0365
```

Example 2

Data `jtrain` from `wooldridge` library

- Several firms followed over three years (87, 88, 89)
- will use
  - `grant` (whether job training grant was given to firm that year)
  - `grant_1` (whether training grant was given in previous year)
  - `lscrap` (logged scrap rates)
  - `fcode` firm id
  - `year`, `d87`, `d88`, `d89` year and year dummies

Example 2

```
library(plm); library(lmtest); library(sandwich); library(wooldridge)
dat <- jtrain

# Pooled OLS
pool_mdl <- plm(lscrap~d88+d89+grant+grant_1, data=dat, model="pooling",
                index=c("fcode", "year"))
summary(pool_mdl)$coefficients %>% round(4)
```

```
            Estimate Std. Error t-value Pr(>|t|)
(Intercept)   0.5974     0.2031  2.9421   0.0038
d88          -0.2394     0.3109 -0.7700   0.4424
d89          -0.4965     0.3379 -1.4693   0.1437
grant         0.2000     0.3383  0.5913   0.5552
grant_1       0.0489     0.4361  0.1122   0.9108
```

Example 2

```
# Pooled OLS with Panel-Robust SE
pool_Var <- vcovHC(pool_mdl, method="arellano", type="HC1", cluster="group")
coeftest(pool_mdl, vcov=pool_Var) %>% round(4)
```

```
t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.5974     0.2184  2.7357   0.0069 **
d88          -0.2394     0.1251 -1.9135   0.0575 .
d89          -0.4965     0.2317 -2.1427   0.0337 *
grant         0.2000     0.3206  0.6239   0.5336
grant_1       0.0489     0.4691  0.1043   0.9171
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example 2

```r
# Random Effects
re_mdl <- plm(lscrap~d88+d89+grant+grant_1, data=dat, model="random",
              index=c("fcode", "year"))
summary(re_mdl)$coefficients %>% round(4)
```

```
            Estimate Std. Error z-value Pr(>|z|)
(Intercept)   0.5974     0.2033  2.9389   0.0033
d88          -0.0935     0.1090 -0.8584   0.3907
d89          -0.2714     0.1315 -2.0643   0.0390
grant        -0.2144     0.1476 -1.4529   0.1463
grant_1      -0.3729     0.2051 -1.8182   0.0690
```

- Results very different from Pooled OLS (which doesn't account for individual fixed effects in any way

Example 2

```r
# Fixed Effects
fe_mdl <- plm(lscrap~d88+d89+grant+grant_1, data=dat, model="within",
              index=c("fcode", "year"))
summary(fe_mdl)$coefficients %>% round(4)
```

```
        Estimate Std. Error t-value Pr(>|t|)
d88      -0.0802     0.1095 -0.7327   0.4654
d89      -0.2472     0.1332 -1.8556   0.0663
grant    -0.2523     0.1506 -1.6751   0.0969
grant_1  -0.4216     0.2102 -2.0057   0.0475
```

Example 2

```
fixef(fe_mdl) %>% round(4)
```

```
 410523  410538  410563  410565  410566  410567  410577  410592  410593  410596
-2.8258  1.0794  1.8915  1.6178  1.7956 -0.5462  0.5973  3.3008  0.1091  1.9360
 410606  410626  410629  410653  410665  410685  418011  418021  418035  418045
-0.7028 -0.1176 -0.2795  0.1821 -2.9002 -1.6470  1.6986  0.3338  1.8337  0.8023
 418051  418054  418065  418076  418083  418091  418097  418107  418118  418125
-0.3945  0.5385  0.5564 -0.2259  0.9310  0.7076  0.6739 -0.1636 -0.6648 -0.2674
 418140  418163  418168  418177  418237  419198  419201  419242  419268  419272
 1.7944  2.3388 -2.6982  3.2422 -1.1806  1.9810  0.8023  1.8211  0.3174  3.2038
 419289  419297  419307  419339  419343  419357  419378  419381  419388  419409
 0.4258 -1.2542  0.1002 -0.5640  1.5580  0.1932  0.7627 -0.4722  2.2829  1.0000
 419432  419459  419482  419483
 1.7180  0.6233  1.1006  3.3144
```

Example 2

LSDV gives the same result as FE (only part of output shown...)

```
# LSDV
lsdv_mdl <- lm(lscrap~d88+d89+grant+grant_1+factor(fcode), data=dat)
summary(lsdv_mdl)$coefficients %>% round(4)
```

```
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          -2.8258     0.2962 -9.5407   0.0000
d88                  -0.0802     0.1095 -0.7327   0.4654
d89                  -0.2472     0.1332 -1.8556   0.0663
grant                -0.2523     0.1506 -1.6751   0.0969
grant_1              -0.4216     0.2102 -2.0057   0.0475
factor(fcode)410538   3.9053     0.4064  9.6092   0.0000
factor(fcode)410563   4.7173     0.4064 11.6074   0.0000
factor(fcode)410565   4.4437     0.4064 10.9340   0.0000
factor(fcode)410566   4.6214     0.4064 11.3715   0.0000
factor(fcode)410567   2.2796     0.4064  5.6091   0.0000
factor(fcode)410577   3.4231     0.4064  8.4230   0.0000
factor(fcode)410592   6.1266     0.4064 15.0751   0.0000
factor(fcode)410593   2.9350     0.4064  7.2217   0.0000
factor(fcode)410596   4.7618     0.4064 11.7169   0.0000
```

Example 2

```
# First difference
fd_mdl <- plm(lscrap~d88+d89+grant+grant_1, data=dat, model="fd",
              index=c("fcode", "year"))
summary(fd_mdl)$coefficients %>% round(4)
```

```
            Estimate Std. Error t-value Pr(>|t|)
(Intercept)  -0.1387     0.0752 -1.8450   0.0679
d88           0.0481     0.0627  0.7669   0.4449
grant        -0.2228     0.1307 -1.7040   0.0914
grant_1      -0.3512     0.2351 -1.4941   0.1382
```

Panel Data

- National Longitudinal Surveys of Labor Market Experience (NLS)
  - *Data on attitudes/behaviors/events related to schooling, employment, marriage, fertility, health, ... Follows various cohorts ('older men', 'mature women', 'young men',... annual/biennially since mid-1960s)*
- Panel Study of Income Dynamics (PSID)
  - *Data on employment, income, housing, travel, ... Follows 6000 families, 15000 individuals (and their descendants), still continuing*
- Singapore Life Panel
- German Social Economics Panel, British Household Panel Survey, ...