面板数据模型

黄嘉平

2025-04-28

1 面板数据

针对面板数据，首先我们可以考虑混合模型（pooled model）

$$y_{it} = \boldsymbol{X}_{it}\boldsymbol{\beta} + \alpha + u_{it}$$

这里 $\boldsymbol{X}_{it}$ 中不包含常数项，α 是常数项。如果满足 OLS 假设，系数 $\boldsymbol{\beta}$ 的 OLS 估计量是 BLUE 的。也就是说，利用面板数据的优势仅仅体现在样本量的增加。

其次是固定效应模型（仅包含了个体效应）

$$y_{it} = \boldsymbol{X}_{it}\boldsymbol{\beta} + \alpha_i + u_{it}$$

这里 $\boldsymbol{X}_{it}$ 中不包含常数项，α_i 是不可观测随机变量并且可以与 $\boldsymbol{X}_{it}$ 相关。我们可以把 α_i 当作未知参数进行估计。

最后是随机效应模型（仅包含了个体效应）

$$y_{it} = \boldsymbol{X}_{it}\boldsymbol{\beta} + (\alpha + v_i) + u_{it}$$

同样，$\boldsymbol{X}_{it}$ 中不包含常数项，α 是常数项，v_i 是不可观测随机变量，并假设 v_i 和 $\boldsymbol{X}_{it}$ 独立。此时 OLS 估计量虽然满足非偏性，但误差项变成了 $v_i + u_{it}$，因此存在组内自相关，导致 OLS 估计量不是最有效的。通常采用 GLS 对 $\boldsymbol{\beta}$ 和 α 进行估计。

2 固定效应模型

固定效应模型由于其假设较为宽松，往往是实证研究中的首选。估计固定效应模型有以下几种方法。

最小二乘虚拟变量法（least squares dummy variable/LSDV）
对每个个体 i，固定效应模型可以写成

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \iota\alpha_i + \boldsymbol{u}_i$$

令 $\boldsymbol{d}_i$ 为指示是否为个体 i 的虚拟变量，$\boldsymbol{D} = [\boldsymbol{d}_1 \ \ldots \ \boldsymbol{d}_n]$，则固定效应模型可以写成矩阵形式

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}\boldsymbol{\alpha} + \boldsymbol{u}$$

用 OLS 估计此模型可以得到系数和常数项的非偏估计量。这种方法称为 LSDV，对应的系数估计量记作 $\hat{\boldsymbol{\beta}}_{\text{LSDV}}$。

中心化法（demeaning）和组内估计量（within-group estimator）

如果个体（或分组）过多，则虚拟变量也会过多，计算 OLS 估计量的难度也会随之增大。但如果我们应用 FWL 定理，即用 $\boldsymbol{D}$ 分别估计 $\boldsymbol{y}$ 和 $\boldsymbol{X}$ 并获得残差，然后用残差模型进行估计，就能减小计算量。令 $\boldsymbol{M_D} = \boldsymbol{I} - \boldsymbol{D}(\boldsymbol{D}^\top\boldsymbol{D})^{-1}\boldsymbol{D}^\top$，则系数的估计量为

$$\hat{\boldsymbol{\beta}}_{\text{within}} = (\boldsymbol{X}^\top\boldsymbol{M_D}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{M_D}\boldsymbol{y}$$

由于 $\boldsymbol{M_D}\boldsymbol{z}$ 等于针对 $\boldsymbol{z}$ 中的每个观测值计算 $z_{it} - \bar{z}_{i.}$，相当于去除了均值，仅保留组内差异，因此称该估计量为组内估计量。

3 在 R 中拟合面板数据模型

首先安装 plm 包并调入。这里省略了调入后显示的信息。

```
library(tidyverse)
library(plm)
```

我们利用下面的数据展示估计结果。

```
data1 <- tibble(
  group = c(1, 1, 1, 2, 2, 3, 3, 3, 4, 4, 4),
  x = c(0, 8, 17, 10, 16, 4, 11, 5, 18, 5, 2),
  y = c(-5, 23, 44, 29, 26, 17, 17, 31, 50, 26, 17)
)
data1
```

```
## # A tibble: 11 x 3
##    group     x     y
##    <dbl> <dbl> <dbl>
##  1     1     0    -5
##  2     1     8    23
##  3     1    17    44
##  4     2    10    29
##  5     2    16    26
##  6     3     4    17
##  7     3    11    17
##  8     3     5    31
##  9     4    18    50
## 10     4     5    26
## 11     4     2    17
```

用 plm 包提供的 plm() 函数可以直接估计固定效应模型的系数。

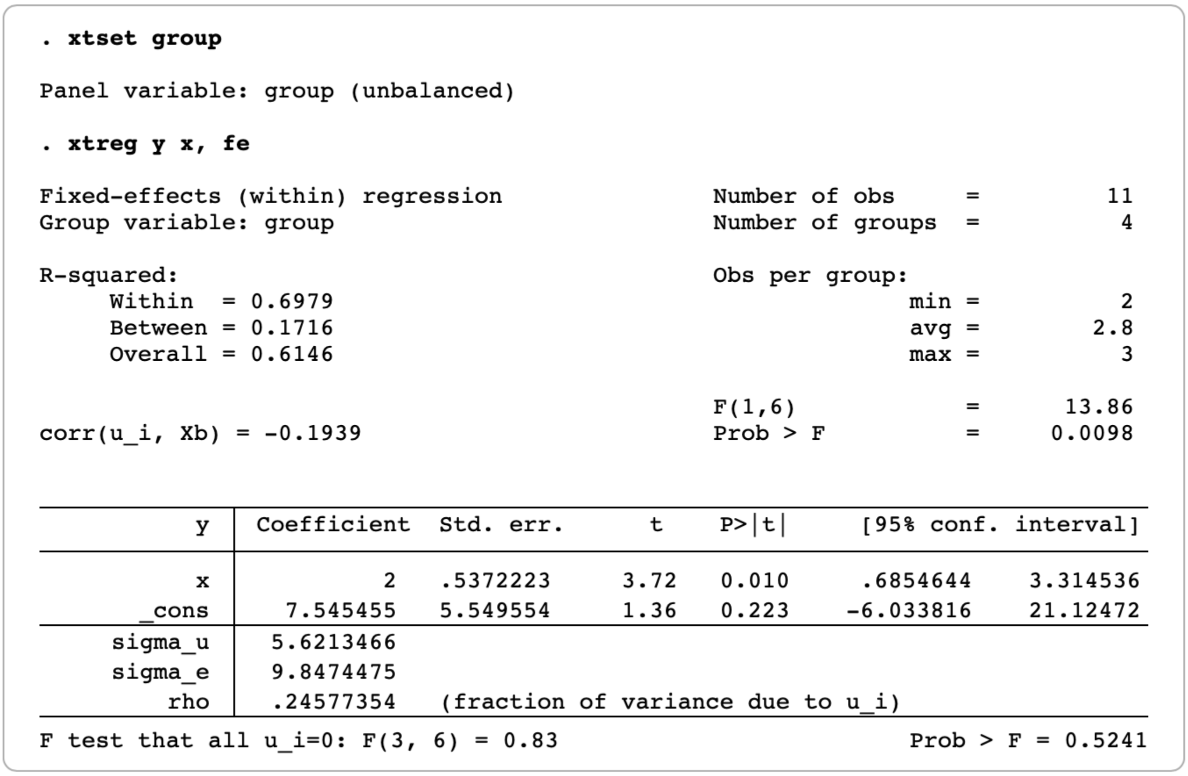
```r
# Within estimator
fe_fit <- plm(
  y ~ x,
  data = data1,
  effect = "individual", # 仅估计个体效应
  model = "within", # 指定固定效应模型
  index = "group" # 指定代表个体的变量为 "group"
)
fe_fit |> summary()
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = y ~ x, data = data1, effect = "individual", model = "within",
##     index = "group")
##
## Unbalanced Panel: n = 4, T = 2-3, N = 11
##
## Residuals:
##       Min.   1st Qu.    Median   3rd Qu.      Max.
## -13.33333  -4.41667   0.66667   4.50000  12.66667
##
## Coefficients:
##   Estimate Std. Error t-value Pr(>|t|)
## x  2.00000    0.53722  3.7229 0.009819 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1925.8
## Residual Sum of Squares: 581.83
## R-Squared:      0.69788
## Adj. R-Squared: 0.49647
## F-statistic: 13.8596 on 1 and 6 DF, p-value: 0.0098192
```

x 的系数估计值为 2.0，标准误为 0.537。

作为对比，在 STATA 中常用的拟合面板数据模型的命令是 xtreg，从 https://www.stata.com/support/faqs/statistics/intercept-in-fixed-effects-model/ 可知估计结果为图 1 所示。其中对 x 系数的估计值和标准误的计算结果和 R 相同，但是它也估计了一个 “常数项”_cons。这是怎么回事呢？

根据网站上的解释可知，STATA 在估计固定效应模型时加入了常数项，即令 $\alpha_i = \alpha + v_i$。这就会形成完美共线性（$\boldsymbol{\iota} = \sum_i \boldsymbol{d}_i$）。通常的解决方法是去掉多余的参数（在实践中多是去掉常数项 α，但理论上去掉任意一个的效果都一样，例如保留 α 但去掉 v_1），但是 STATA 选择了保留常数项，同时



```
. xtset group

Panel variable: group (unbalanced)

. xtreg y x, fe

Fixed-effects (within) regression               Number of obs     =         11
Group variable: group                           Number of groups  =          4

R-squared:                                      Obs per group:
     Within  = 0.6979                                         min =          2
     Between = 0.1716                                         avg =        2.8
     Overall = 0.6146                                         max =          3

                                                F(1,6)            =      13.86
corr(u_i, Xb) = -0.1939                         Prob > F          =     0.0098

------------------------------------------------------------------------------
           y | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
           x |          2   .5372223     3.72   0.010     .6854644    3.314536
       _cons |   7.545455   5.549554     1.36   0.223    -6.033816    21.12472
-------------+----------------------------------------------------------------
     sigma_u |  5.6213466
     sigma_e |  9.8474475
         rho |  .24577354   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(3, 6) = 0.83                        Prob > F = 0.5241
```

图 1: STATA 中 `xtreg` 命令的拟合结果

人为添加了限制条件

$$\sum_{i=1}^{n} v_i = 0$$

这样确实可以解决共线性问题，但是也容易产生不必要的误解。例如很多同学在汇报固定效应模型的估计结果时汇报了这个“常数项”的估计值，但是并没有标注其含义。这个所谓的“常数项”实际上是个体固定效应的均值，即

$$\frac{1}{n}\sum_{i=1}^{n}\alpha_i = \frac{1}{n}\sum_{i=1}^{n}(\alpha + v_i) = \alpha + \frac{1}{n}\sum_{i=1}^{n} v_i = \alpha$$

通常情况下不需要针对固定效应模型汇报常数项的估计值，例如 Greene (2018, p.438) 中的 Table 11.9。在面板数据模型中，人们更想了解的是解释变量的系数，而不是个体效应。我们的目的是剔除不可观测因素产生的个体差异对解释变量系数的影响，因此只要明确是否控制了个体（或时间）效应即可，无需汇报个体效应的估计值。

另一个需要注意的地方是在解释变量中如果包含了不随时间变化的量，例如性别、民族等，则固定效应模型无法估计出它们的系数（因为存在完美共线性），而是将它们的影响都包含在个体效应中。从 Greene (2018) 的 Table 11.9 中也可以观察到这个特点。

虽然 STATA 在固定效应模型中估计“常数项”的设定与常见的理论模型不符，但是毕竟它的用户群体非常庞大，带来的影响不可忽视。R 的开发者有针对性地准备了 `within_intercept()` 函数以计算和 STATA 相同的“常数项”估计值及其标准误。

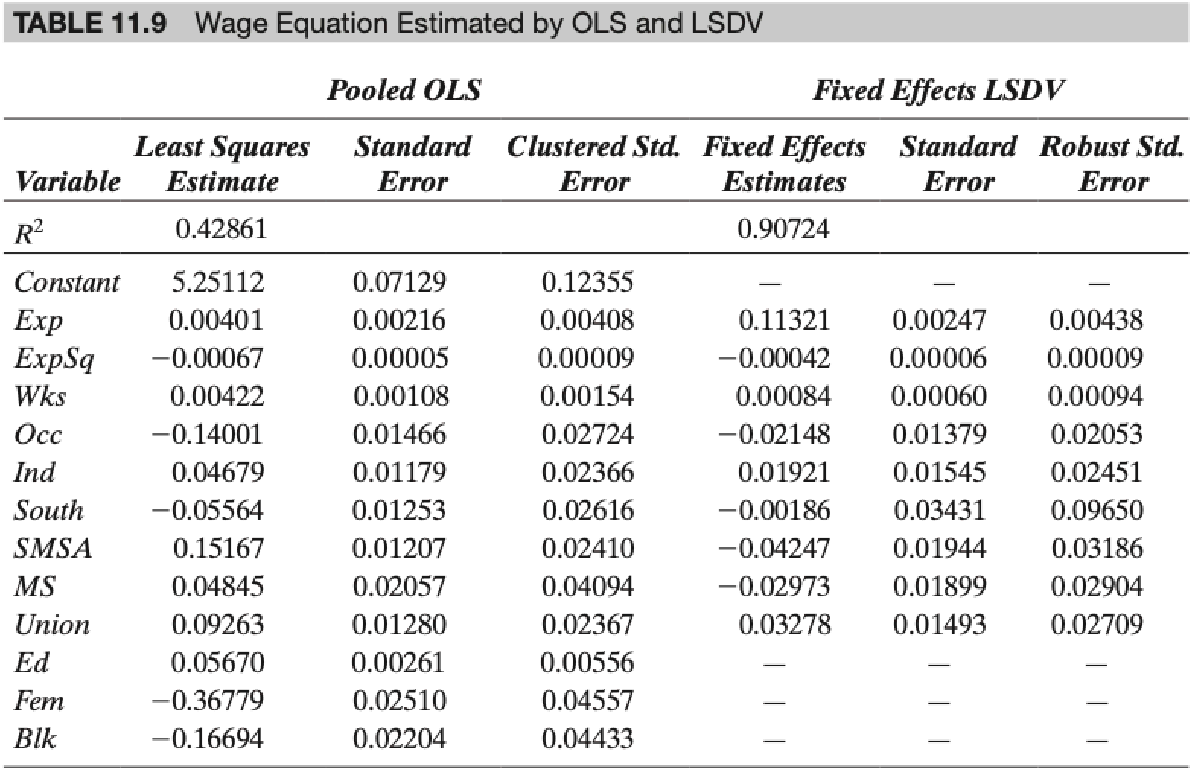


TABLE 11.9 Wage Equation Estimated by OLS and LSDV

	Pooled OLS			*Fixed Effects LSDV*		
Variable	*Least Squares Estimate*	*Standard Error*	*Clustered Std. Error*	*Fixed Effects Estimates*	*Standard Error*	*Robust Std. Error*
R^2	0.42861			0.90724		
Constant	5.25112	0.07129	0.12355	—	—	—
Exp	0.00401	0.00216	0.00408	0.11321	0.00247	0.00438
ExpSq	−0.00067	0.00005	0.00009	−0.00042	0.00006	0.00009
Wks	0.00422	0.00108	0.00154	0.00084	0.00060	0.00094
Occ	−0.14001	0.01466	0.02724	−0.02148	0.01379	0.02053
Ind	0.04679	0.01179	0.02366	0.01921	0.01545	0.02451
South	−0.05564	0.01253	0.02616	−0.00186	0.03431	0.09650
SMSA	0.15167	0.01207	0.02410	−0.04247	0.01944	0.03186
MS	0.04845	0.02057	0.04094	−0.02973	0.01899	0.02904
Union	0.09263	0.01280	0.02367	0.03278	0.01493	0.02709
Ed	0.05670	0.00261	0.00556	—	—	—
Fem	−0.36779	0.02510	0.04557	—	—	—
Blk	−0.16694	0.02204	0.04433	—	—	—

图 2: Greene (2018, p.438) 中的 Table 11.9

```
fe_fit |> fixef() # 从固定效应估计中提取个体效应的估计值
```

```
##       1       2       3       4
##  4.0000  1.5000  8.3333 14.3333
```

```
fe_fit |> within_intercept() # 计算 STATA 中的“常数项”估计值和标准误
```

```
## (overall_intercept)
##            7.545455
## attr(,"se")
## [1] 5.549554
```

```
fe_fit |>
  within_intercept(return.model = TRUE) |>
  summary() # 添加 `return.model = TRUE` 参数以展示全部估计结果
```

```
## Pooling Model
##
## Call:
## plm(formula = form, data = data, model = "pooling")
##
## Unbalanced Panel: n = 4, T = 2-3, N = 11
##
```

```
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -13.33333  -4.41667   0.66667   4.50000  12.66667
##
## Coefficients:
##             Estimate Std. Error t-value Pr(>|t|)
## (Intercept)  7.54545    5.54955  1.3597  0.20703
## x            2.00000    0.53722  3.7229  0.00475 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1925.8
## Residual Sum of Squares: 581.83
## R-Squared:      0.69788
## Adj. R-Squared: 0.66431
## F-statistic: 13.8596 on 1 and 9 DF, p-value: 0.0047496
```

我们也可以手动添加虚拟变量并计算 LSDV 估计值，以及手动进行中心化后计算 OLS 估计值。从下面的回归结果可以看出，这两个估计值都和组内估计值相同。但是，通过手动中心化计算出的系数标准误是错误的。

```r
# LSDV estimator
dummy_fit <- lm(y ~ x + factor(group) - 1, data = data1)
dummy_fit |> summary()
```

```
##
## Call:
## lm(formula = y ~ x + factor(group) - 1, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3333  -4.4167   0.6667   4.5000  12.6667
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## x                2.0000     0.5372   3.723  0.00982 **
## factor(group)1   4.0000     7.2365   0.553  0.60040
## factor(group)2   1.5000     9.8621   0.152  0.88410
## factor(group)3   8.3333     6.7195   1.240  0.26121
## factor(group)4  14.3333     7.2365   1.981  0.09493 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.847 on 6 degrees of freedom
## Multiple R-squared:  0.9354, Adjusted R-squared:  0.8816
## F-statistic: 17.38 on 5 and 6 DF,  p-value: 0.00164
```

```
# Demeaned estimator
data2 <- data1 |>
  group_by(group) |>
  mutate(
    x_demeaned = x - mean(x),
    y_demeaned = y - mean(y)
  )
data2
```

```
## # A tibble: 11 x 5
## # Groups:   group [4]
##    group     x     y x_demeaned y_demeaned
##    <dbl> <dbl> <dbl>      <dbl>      <dbl>
##  1     1     0    -5     -8.33      -25.7
##  2     1     8    23     -0.333       2.33
##  3     1    17    44      8.67       23.3
##  4     2    10    29     -3           1.5
##  5     2    16    26      3          -1.5
##  6     3     4    17     -2.67       -4.67
##  7     3    11    17      4.33       -4.67
##  8     3     5    31     -1.67        9.33
##  9     4    18    50      9.67       19
## 10     4     5    26     -3.33      -5
## 11     4     2    17     -6.33      -14
```

```
dm_fit1 <- lm(y_demeaned ~ x_demeaned - 1, data = data2)
dm_fit1 |> summary()
```

```
##
## Call:
## lm(formula = y_demeaned ~ x_demeaned - 1, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3333  -4.4167   0.6667   4.5000  12.6667
##
## Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
## x_demeaned   2.0000     0.4161   4.806 0.000717 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 7.628 on 10 degrees of freedom
## Multiple R-squared:  0.6979, Adjusted R-squared:  0.6677 
## F-statistic:  23.1 on 1 and 10 DF,  p-value: 0.0007169
```

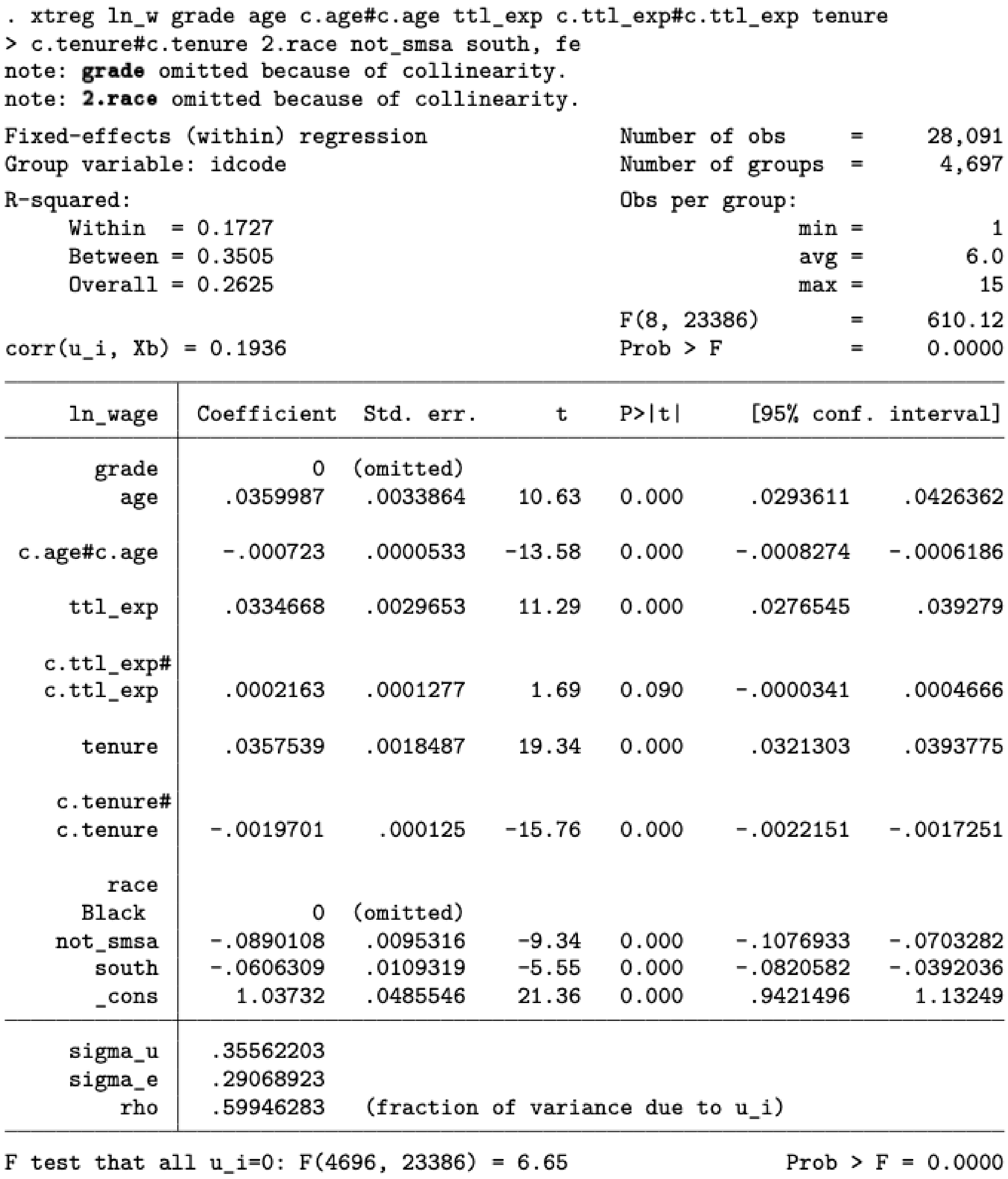
4 利用真实数据进行比较

在 *Stata 18 Longitudinal-Data/Panel-Data Reference Manual* 中，作者利用 nlswork 数据（National Longitudinal Survey of Young Women）展示了 xtreg 的各项功能。在 R 中，该数据由 sampleSelection 包提供。

```r
library(sampleSelection)
data("nlswork") # 调用 `nlswork` 数据

nls_fefit <- plm(
  ln_wage ~ grade + age + I(age^2) + ttl_exp + I(ttl_exp^2) + tenure +
    I(tenure^2) + factor(race == 2) + factor(not_smsa) + factor(south),
  data = nlswork,
  effect = "individual",
  model = "within",
  index = c("idcode", "year")
)
nls_fefit |> summary()
```

```
## Oneway (individual) effect Within Model
## 
## Call:
## plm(formula = ln_wage ~ grade + age + I(age^2) + ttl_exp + I(ttl_exp^2) + 
##     tenure + I(tenure^2) + factor(race == 2) + factor(not_smsa) + 
##     factor(south), data = nlswork, effect = "individual", model = "within", 
##     index = c("idcode", "year"))
## 
## Unbalanced Panel: n = 4697, T = 1-15, N = 28091
## 
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max. 
## -1.9180562 -0.1163833  0.0031468  0.1273779  3.0278371 
## 
## Coefficients:
##                       Estimate  Std. Error  t-value  Pr(>|t|)    
## age                 3.5999e-02  3.3864e-03  10.6304 < 2.2e-16 ***
## I(age^2)           -7.2299e-04  5.3258e-05 -13.5753 < 2.2e-16 ***
## ttl_exp             3.3467e-02  2.9653e-03  11.2860 < 2.2e-16 ***
```



```
. xtreg ln_w grade age c.age#c.age ttl_exp c.ttl_exp#c.ttl_exp tenure
> c.tenure#c.tenure 2.race not_smsa south, fe
note: grade omitted because of collinearity.
note: 2.race omitted because of collinearity.

Fixed-effects (within) regression               Number of obs     =     28,091
Group variable: idcode                          Number of groups  =      4,697

R-squared:                                      Obs per group:
     Within  = 0.1727                                         min =          1
     Between = 0.3505                                         avg =        6.0
     Overall = 0.2625                                         max =         15

                                                F(8, 23386)       =     610.12
corr(u_i, Xb) = 0.1936                          Prob > F          =     0.0000

------------------------------------------------------------------------------
     ln_wage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       grade |          0  (omitted)
         age |   .0359987   .0033864    10.63   0.000     .0293611    .0426362
             |
 c.age#c.age |   -.000723   .0000533   -13.58   0.000    -.0008274   -.0006186
             |
     ttl_exp |   .0334668   .0029653    11.29   0.000     .0276545     .039279
             |
   c.ttl_exp#|
   c.ttl_exp |   .0002163   .0001277     1.69   0.090    -.0000341    .0004666
             |
      tenure |   .0357539   .0018487    19.34   0.000     .0321303    .0393775
             |
    c.tenure#|
    c.tenure |  -.0019701    .000125   -15.76   0.000    -.0022151   -.0017251
             |
        race |
      Black  |          0  (omitted)
    not_smsa |  -.0890108   .0095316    -9.34   0.000    -.1076933   -.0703282
       south |  -.0606309   .0109319    -5.55   0.000    -.0820582   -.0392036
       _cons |    1.03732   .0485546    21.36   0.000     .9421496     1.13249
-------------+----------------------------------------------------------------
     sigma_u |  .35562203
     sigma_e |  .29068923
         rho |  .59946283   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(4696, 23386) = 6.65                 Prob > F = 0.0000
```

图 3: STATA 中对 `nlswork` 数据的固定效应估计结果

```
## I(ttl_exp^2)       2.1627e-04  1.2774e-04   1.6931   0.09046 .
## tenure             3.5754e-02  1.8487e-03  19.3401 < 2.2e-16 ***
## I(tenure^2)       -1.9701e-03  1.2499e-04 -15.7619 < 2.2e-16 ***
## factor(not_smsa)1 -8.9011e-02  9.5316e-03  -9.3385 < 2.2e-16 ***
## factor(south)1    -6.0631e-02  1.0932e-02  -5.5462 2.951e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    2388.6
## Residual Sum of Squares: 1976.1
## R-Squared:      0.17267
## Adj. R-Squared: 0.006261
## F-statistic: 610.122 on 8 and 23386 DF, p-value: < 2.22e-16
```

注意以上结果中并没有变量 `grade` 和 `factor(race == 2)` 的系数，这是完美共线性造成的，但是 R 并没有给出提示。

```
nls_fefit |>
  within_intercept(return.model = TRUE) |>
  summary()
```

```
## Pooling Model
##
## Call:
## plm(formula = form, data = data, model = "pooling")
##
## Unbalanced Panel: n = 4697, T = 1-15, N = 28091
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -1.9180562 -0.1163833  0.0031468  0.1273779  3.0278371
##
## Coefficients:
##                      Estimate  Std. Error  t-value  Pr(>|t|)
## (Intercept)        1.0373e+00  4.8555e-02  21.3640 < 2.2e-16 ***
## age                3.5999e-02  3.3864e-03  10.6304 < 2.2e-16 ***
## I.age.2.          -7.2299e-04  5.3258e-05 -13.5753 < 2.2e-16 ***
## ttl_exp            3.3467e-02  2.9653e-03  11.2860 < 2.2e-16 ***
## I.ttl_exp.2.       2.1627e-04  1.2774e-04   1.6931   0.09046 .
## tenure             3.5754e-02  1.8487e-03  19.3401 < 2.2e-16 ***
## I.tenure.2.       -1.9701e-03  1.2499e-04 -15.7619 < 2.2e-16 ***
## factor.not_smsa.1 -8.9011e-02  9.5316e-03  -9.3385 < 2.2e-16 ***
## factor.south.1    -6.0631e-02  1.0932e-02  -5.5462 2.945e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    2388.6
## Residual Sum of Squares: 1976.1
## R-Squared:      0.17267
## Adj. R-Squared: 0.17244
## F-statistic: 610.122 on 8 and 28082 DF, p-value: < 2.22e-16
```

5 双向固定效应和非平衡面板数据

双向固定效应（two way fixed effects）模型就是在个体固定效应 α_i 的基础上，再加入时间固定效应 δ_t，记作

$$y_{it} = \boldsymbol{X}_{it}\boldsymbol{\beta} + \alpha_i + \delta_t + u_{it}$$

双向固定效应模型也可以通过虚拟变量法估计，但需要注意的是，n 个个体虚拟变量之和等于 T 个时间虚拟变量之和，因此为了避免完美共线性，我们通常加入 n 个个体虚拟变量和 $T-1$ 个时间虚拟变量。此时估计出的 $\hat{\alpha}_i$ 依然是个体固定效应，但 $\hat{\delta}_t$ 则是 t 时点和基准时点（也就是缺失的时间虚拟变量对应的时间点）之间的差值。由此，我们称此模型为非对称模型。

对称模型指的是下面包含共通常数项的模型

$$y_{it} = \boldsymbol{X}_{it}\boldsymbol{\beta} + \alpha + v_i + \delta_t + u_{it}$$

为了避免完美共线性，我们需要引入两个附加条件，即

$$\sum_{i=1}^{n} v_i = 0, \quad \sum_{t=1}^{T} \delta_t = 0$$

此时，我们可以利用中心化法估计斜率、常数项（固定效应的均值）、以及个体和时间固定效应（差值）。首先对任意变量 z 定义

$$\bar{z}_{i.} = \frac{1}{T}\sum_{t=1}^{T} z_{it}, \quad \bar{z}_{.t} = \frac{1}{n}\sum_{i=1}^{n} z_{it}, \quad \bar{\bar{z}} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} z_{it}$$

然后计算中心化后的变量值

$$y_{it}^* = y_{it} - \bar{y}_{i.} - \bar{y}_{.t} + \bar{\bar{y}}$$
$$\boldsymbol{X}_{it}^* = \boldsymbol{X}_{it} - \bar{\boldsymbol{X}}_{i.} - \bar{\boldsymbol{X}}_{.t} + \bar{\bar{\boldsymbol{X}}}$$

我们可以用下列模型的 OLS 估计量估计斜率 $\boldsymbol{\beta}$

$$y_{it}^* = \boldsymbol{X}_{ij}^*\boldsymbol{\beta} + \mathrm{error}_{it}$$

并利用估计值计算常数项和固定效应

$$\hat{\alpha} = \bar{\bar{y}} - \bar{\bar{\boldsymbol{X}}}\hat{\boldsymbol{\beta}}$$
$$\hat{v}_i = (\bar{y}_{i.} - \bar{\bar{y}}) - (\bar{\boldsymbol{X}}_{i.} - \bar{\bar{\boldsymbol{X}}})\hat{\boldsymbol{\beta}}$$
$$\hat{\delta}_t = (\bar{y}_{.t} - \bar{\bar{y}}) - (\bar{\boldsymbol{X}}_{.t} - \bar{\bar{\boldsymbol{X}}})\hat{\boldsymbol{\beta}}$$

这里将这种方法称为对称中心化法。

下面利用 `plm` 包中的 `Grunfeld` 数据集对以上方法进行验证。可以用 `?Grunfeld` 命令了解该数据集的具体内容。

```r
data("Grunfeld") # 调用 `Grunfeld` 数据
gf <- as_tibble(Grunfeld)

# plm
gf_twfefit <- plm(
  inv ~ value + capital,
  data = gf,
  effect = "twoways",
  model = "within",
  index = c("firm", "year")
)
gf_twfefit |> summary()
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = inv ~ value + capital, data = gf, effect = "twoways",
##     model = "within", index = c("firm", "year"))
##
## Balanced Panel: n = 10, T = 20, N = 200
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -162.6094  -19.4710   -1.2669   19.1277  211.8420
##
## Coefficients:
##          Estimate Std. Error t-value  Pr(>|t|)
## value    0.117716   0.013751  8.5604 6.653e-15 ***
## capital 0.357916   0.022719 15.7540 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1615600
## Residual Sum of Squares: 452150
## R-Squared:      0.72015
## Adj. R-Squared: 0.67047
## F-statistic: 217.442 on 2 and 169 DF, p-value: < 2.22e-16
```

```r
# LSDV
gf_dummyfit <- lm(
  inv ~ value + capital + factor(firm) + factor(year) - 1,
  data = gf
)
gf_dummyfit |> summary()
```

```
##
## Call:
## lm(formula = inv ~ value + capital + factor(firm) + factor(year) -
##     1, data = gf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -162.609  -19.471   -1.267   19.128  211.842
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## value                 0.11772    0.01375   8.560 6.65e-15 ***
## capital               0.35792    0.02272  15.754  < 2e-16 ***
## factor(firm)1       -86.90023   56.04663  -1.550 0.122893
## factor(firm)2       120.15401   29.16688   4.120 5.93e-05 ***
## factor(firm)3      -222.13103   28.59744  -7.768 7.37e-13 ***
## factor(firm)4         8.45361   20.41784   0.414 0.679377
## factor(firm)5       -92.33883   20.91106  -4.416 1.79e-05 ***
## factor(firm)6        15.98841   19.88487   0.804 0.422498
## factor(firm)7       -35.43362   20.17003  -1.757 0.080772 .
## factor(firm)8       -19.40972   20.49076  -0.947 0.344868
## factor(firm)9       -56.68267   19.81211  -2.861 0.004756 **
## factor(firm)10       39.93689   20.40337   1.957 0.051951 .
## factor(year)1936    -19.19741   23.67586  -0.811 0.418596
## factor(year)1937    -40.69001   24.69541  -1.648 0.101277
## factor(year)1938    -39.22640   23.23594  -1.688 0.093221 .
## factor(year)1939    -69.47029   23.65607  -2.937 0.003780 **
## factor(year)1940    -44.23508   23.80979  -1.858 0.064930 .
## factor(year)1941    -18.80446   23.69400  -0.794 0.428519
## factor(year)1942    -21.13979   23.38163  -0.904 0.367219
## factor(year)1943    -42.97762   23.55287  -1.825 0.069808 .
## factor(year)1944    -43.09877   23.61020  -1.825 0.069701 .
## factor(year)1945    -55.68304   23.89562  -2.330 0.020974 *
## factor(year)1946    -31.16928   24.11598  -1.292 0.197957
## factor(year)1947    -39.39224   23.78368  -1.656 0.099522 .
## factor(year)1948  -43.71651   23.96965  -1.824 0.069945 .
## factor(year)1949  -73.49510   24.18292  -3.039 0.002750 **
## factor(year)1950  -75.89611   24.34553  -3.117 0.002144 **
## factor(year)1951  -62.48091   24.86425  -2.513 0.012911 *
## factor(year)1952  -64.63234   25.34950  -2.550 0.011672 *
## factor(year)1953  -67.71797   26.61108  -2.545 0.011832 *
## factor(year)1954  -93.52622   27.10786  -3.450 0.000708 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 51.72 on 169 degrees of freedom
## Multiple R-squared:  0.9668, Adjusted R-squared:  0.9607
## F-statistic: 158.8 on 31 and 169 DF,  p-value: < 2.2e-16
```

```
gf_twfefit |>
  fixef(effect = "individual") |>
  summary()
```

```
##       Estimate
## 1   -86.900230
## 2   120.154010
## 3  -222.131030
## 4     8.453612
## 5   -92.338825
## 6    15.988413
## 7   -35.433620
## 8   -19.409715
## 9   -56.682674
## 10   39.936893
```

```
gf_twfefit |>
  fixef(effect = "time", type = "dfirst") |>
  summary()
```

```
##       Estimate
## 1936 -19.19741
## 1937 -40.69001
## 1938 -39.22640
## 1939 -69.47029
## 1940 -44.23508
## 1941 -18.80446
## 1942 -21.13979
## 1943 -42.97762
## 1944 -43.09877
## 1945 -55.68304
## 1946 -31.16928
## 1947 -39.39224
## 1948 -43.71651
## 1949 -73.49510
## 1950 -75.89611
## 1951 -62.48091
## 1952 -64.63234
## 1953 -67.71797
## 1954 -93.52622
```

```
# Demeaned
gf <- gf |>
  group_by(firm) |>
  mutate(
    inv_imean = mean(inv),
    value_imean = mean(value),
    capital_imean = mean(capital)
  )
gf <- gf |>
  group_by(year) |>
  mutate(
    inv_tmean = mean(inv),
    value_tmean = mean(value),
    capital_tmean = mean(capital)
  )
gf <- gf |>
  ungroup() |>
  mutate(
    inv_demeaned = inv - inv_imean - inv_tmean + mean(inv),
    value_demeaned = value - value_imean - value_tmean + mean(value),
    capital_demeaned = capital - capital_imean - capital_tmean + mean(capital)
  )

gf_dmfit <- lm(
  inv_demeaned ~ value_demeaned + capital_demeaned - 1,
  data = gf
)
gf_dmfit |> summary()
```

```
##
## Call:
## lm(formula = inv_demeaned ~ value_demeaned + capital_demeaned -
##     1, data = gf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -162.609  -19.471   -1.267   19.128  211.842
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## value_demeaned    0.11772    0.01270   9.266   <2e-16 ***
## capital_demeaned  0.35792    0.02099  17.052   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47.79 on 198 degrees of freedom
## Multiple R-squared:  0.7201, Adjusted R-squared:  0.7173
## F-statistic: 254.8 on 2 and 198 DF,  p-value: < 2.2e-16
```

注意对称中心化法对斜率标准误的估计是错误的。

下面计算常数项和固定效应。

```
gf_dmfit_fixef <- list(
  alpha = mean(gf$inv) - c(mean(gf$value), mean(gf$capital)) %*% coef(gf_dmfit),
  v = (unique(gf$inv_imean) - unique(mean(gf$inv))) -
    t(rbind(
      unique(gf$value_imean) - mean(gf$value),
      unique(gf$capital_imean) - mean(gf$capital)
    )) %*% coef(gf_dmfit),
  delta = (unique(gf$inv_tmean) - unique(mean(gf$inv))) -
    t(rbind(
      unique(gf$value_tmean) - mean(gf$value),
      unique(gf$capital_tmean) - mean(gf$capital)
    )) %*% coef(gf_dmfit)
)

gf_dmfit_fixef$v[, 1] + gf_dmfit_fixef$alpha[1] + gf_dmfit_fixef$delta[1, 1]
```

```
##  [1]  -86.900230  120.154010 -222.131030    8.453612  -92.338825   15.988413
##  [7]  -35.433620  -19.409715  -56.682674   39.936893
```

```
# individual fixed effects, equivalent to
gf_twfefit |>
  fixef(effect = "individual") |>
  summary()
```

```
##        Estimate
## 1    -86.900230
## 2    120.154010
## 3  -222.131030
## 4      8.453612
## 5    -92.338825
## 6     15.988413
## 7    -35.433620
## 8    -19.409715
## 9    -56.682674
## 10    39.936893
```

```
gf_dmfit_fixef$delta[, 1] + gf_dmfit_fixef$alpha[1] + gf_dmfit_fixef$v[1, 1]
```

```
##  [1]  -86.90023 -106.09764 -127.59024 -126.12663 -156.37052 -131.13531
##  [7] -105.70469 -108.04002 -129.87785 -129.99900 -142.58327 -118.06951
## [13] -126.29247 -130.61674 -160.39533 -162.79634 -149.38114 -151.53257
## [19] -154.61820 -180.42645
```

```
# time fixed effects, equivalent to
gf_twfefit |>
  fixef(effect = "time", type = "level") |>
  summary()
```

```
##         Estimate
## 1935  -86.90023
## 1936 -106.09764
## 1937 -127.59024
## 1938 -126.12663
## 1939 -156.37052
## 1940 -131.13531
## 1941 -105.70469
## 1942 -108.04002
## 1943 -129.87785
## 1944 -129.99900
## 1945 -142.58327
## 1946 -118.06951
## 1947 -126.29247
## 1948 -130.61674
## 1949 -160.39533
## 1950 -162.79634
## 1951 -149.38114
## 1952 -151.53257
## 1953 -154.61820
## 1954 -180.42645
```

到目前为止，我们并没有对面板数据是否平衡（balanced）做出限定。实际上，单向固定效应的各种估计方法对平衡和非平衡（unbalanced）面板都适用。在平衡面板上，双向固定效应的各种估计方法都适用。但是在面对非平衡面板时，对称中心化法的估计结果是错误的 (Greene, 2018, p.439)。下面我们利用 `Grunfeld` 数据集人为生成一组非平衡面板，并观察双向固定效应的各种估计方法的结果是否一致。

```
# Twoway fixed effects with unbalanced panel
gfmod <- Grunfeld |>
  as_tibble() |>
  filter(!(firm <= 5 & year >= 1950)) # 删除前五个个体的最后五年间的观测值

# plm
gfmod_twfefit <- plm(
  inv ~ value + capital,
  data = gfmod,
  effect = "twoways",
  model = "within",
  index = c("firm", "year")
)
gfmod_twfefit |> summary()
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = inv ~ value + capital, data = gfmod, effect = "twoways",
##     model = "within", index = c("firm", "year"))
##
## Unbalanced Panel: n = 10, T = 15-20, N = 175
##
## Residuals:
##        Min.    1st Qu.     Median    3rd Qu.       Max.
## -119.13395  -13.93565   -0.08944   14.06739  124.77893
##
## Coefficients:
##         Estimate Std. Error t-value  Pr(>|t|)
## value   0.073884   0.011730  6.2986 3.432e-09 ***
## capital 0.171824   0.033600  5.1138 9.898e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    251990
## Residual Sum of Squares: 184780
## R-Squared:      0.26673
## Adj. R-Squared: 0.11396
## F-statistic: 26.19 on 2 and 144 DF, p-value: 1.9912e-10
```

```r
# LSDV
gfmod_dummyfit <- lm(
  inv ~ value + capital + factor(firm) + factor(year) - 1,
  data = gfmod
)
gfmod_dummyfit |> summary()
```

```
##
## Call:
## lm(formula = inv ~ value + capital + factor(firm) + factor(year) -
##     1, data = gfmod)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -119.134  -13.936   -0.089   14.067  124.779
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## value               0.07388    0.01173   6.299 3.43e-09 ***
## capital             0.17182    0.03360   5.114 9.90e-07 ***
## factor(firm)1     109.62932   48.85397   2.244   0.0264 *
## factor(firm)2     180.14198   24.78267   7.269 2.12e-11 ***
## factor(firm)3    -102.04683   24.83181  -4.110 6.63e-05 ***
## factor(firm)4       3.78330   14.92123   0.254   0.8002
## factor(firm)5     -28.40067   16.78703  -1.692   0.0928 .
## factor(firm)6       7.31342   13.86989   0.527   0.5988
## factor(firm)7     -16.74409   14.57220  -1.149   0.2524
## factor(firm)8     -20.55076   14.27899  -1.439   0.1523
## factor(firm)9     -33.10625   14.38572  -2.301   0.0228 *
## factor(firm)10     -2.33382   15.44229  -0.151   0.8801
## factor(year)1936   -0.67478   16.63674  -0.041   0.9677
## factor(year)1937   -4.87747   17.92216  -0.272   0.7859
## factor(year)1938  -18.75742   16.37974  -1.145   0.2540
## factor(year)1939  -35.78164   17.11351  -2.091   0.0383 *
## factor(year)1940   -7.57898   17.34460  -0.437   0.6628
## factor(year)1941   19.18409   17.33539   1.107   0.2703
## factor(year)1942   12.79402   16.98636   0.753   0.4526
## factor(year)1943   -2.21032   17.40735  -0.127   0.8991
## factor(year)1944   -1.20499   17.49622  -0.069   0.9452
## factor(year)1945   -7.51637   18.01262  -0.417   0.6771
## factor(year)1946   22.48397   18.48193   1.217   0.2258
## factor(year)1947   15.94741   18.51279   0.861   0.3904
## factor(year)1948   17.94048   19.12532   0.938   0.3498
## factor(year)1949   -4.35775   19.84946  -0.220   0.8265
## factor(year)1950  -11.94365   22.17065  -0.539   0.5909
## factor(year)1951   -1.16605   22.43732  -0.052   0.9586
## factor(year)1952   -0.10544   22.89025  -0.005   0.9963
## factor(year)1953   -0.09368   23.54584  -0.004   0.9968
## factor(year)1954   -8.93206   24.02582  -0.372   0.7106
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.82 on 144 degrees of freedom
## Multiple R-squared:  0.9697, Adjusted R-squared:  0.9632
## F-statistic: 148.7 on 31 and 144 DF,  p-value: < 2.2e-16
```

```
# Demeaned
gfmod <- gfmod |>
  group_by(firm) |>
  mutate(
    inv_imean = mean(inv),
    value_imean = mean(value),
    capital_imean = mean(capital)
  )
gfmod <- gfmod |>
  group_by(year) |>
  mutate(
    inv_tmean = mean(inv),
    value_tmean = mean(value),
    capital_tmean = mean(capital)
  )
gfmod <- gfmod |>
  ungroup() |>
  mutate(
    inv_demeaned = inv - inv_imean - inv_tmean + mean(inv),
    value_demeaned = value - value_imean - value_tmean + mean(value),
    capital_demeaned = capital - capital_imean - capital_tmean + mean(capital)
  )

gfmod_dmfit <- lm(
  inv_demeaned ~ value_demeaned + capital_demeaned - 1,
  data = gfmod
)
gfmod_dmfit |> summary()
```

```
##
## Call:
## lm(formula = inv_demeaned ~ value_demeaned + capital_demeaned -
##     1, data = gfmod)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -136.414  -14.162    0.959   15.730  125.631
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## value_demeaned   0.090916   0.007235  12.567  < 2e-16 ***
## capital_demeaned 0.191935   0.028890   6.644 3.82e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.14 on 173 degrees of freedom
## Multiple R-squared:  0.5358, Adjusted R-squared:  0.5304
## F-statistic: 99.83 on 2 and 173 DF,  p-value: < 2.2e-16
```

可见，对称中心化法对斜率的估计结果与其他两种方法不同。Wansbeek & Kapteyn (1989) 给出了非平衡面板双向固定效应模型的正确处理方法。

在实践中，几乎任何一个计量软件都会自动处理非平衡面板数据，因此通常不必为这个问题担心。同学们可以各自在 STATA 或其他计量软件中进行验证。

6 课后练习

利用 Cornwell & Rupert (1989) 中的数据（即 `plm` 包中的 `Wages` 数据集）分别用 `plm` 函数、虚拟变量法、对称中心化法估计下面的双向固定效应模型。

$$\ln Wage_{it} = \beta Wks_{it} + \alpha_i + \delta_t + u_{it}$$

根据 Greene (2018) 中的 Example 11.8 可知，三种方法估计出的 $\hat{\beta}$ 都应是 0.00095，你是否正确估计了该参数?

接下来将 `Wages` 数据中前 300 个个体的最后三个时间点上的观测值删除，从而获得了一组非平衡面板数据（其中前 300 个个体各有 4 个时间点，后 295 个个体各有 7 个时间点的观测值）。再次用 `plm` 函数、虚拟变量法、对称中心化法估计双向固定效应模型。此时，`plm` 法和虚拟变量法估计出的 $\hat{\beta}$ 应是 0.00050，而中心化法给出的（错误）估计值为 0.00283。你是否获得了相同的结果?

注意：`Wages` 数据并没有包含个体和时间变量，需要首先运行下面的命令进行转换

```
data(Wages)
Wag <- pdata.frame(Wages, index = 595) |> as_tibble()
```

于是，在 `Wag` 中新出现了两个变量 `id` 和 `time`，分别保存了个体和时间信息。

参考文献

Cornwell, C., & Rupert, P. (1988). Efficient Estimation With Panel Data: an Empirical Comparison of Instrumental Variables Estimators. *Journal of Applied Econometrics*, *3*, 149–155.

Gould, W. (n.d.). *Interpreting the intercept in the fixed-effects model*. StataCorp. Retrieved April 22, 2025, from https://www.stata.com/support/faqs/statistics/intercept-in-fixed-effects-model/

Greene, W. H. (2018). *Econometric Analysis* (8th ed.). Pearson.

StataCorp. (2023). *Stata 18 Longitudinal-Data/Panel-Data Reference Manual*. Stata Press.

Wansbeek, T. J., & Kapteyn, A. (1989). Estimation of the Error Components Model with Incomplete Panels. *Journal of Econometrics*, *41*, 341–361.