治理技术专题

定量政治分析方法

Quantitative Analysis II

苏 毓 淞

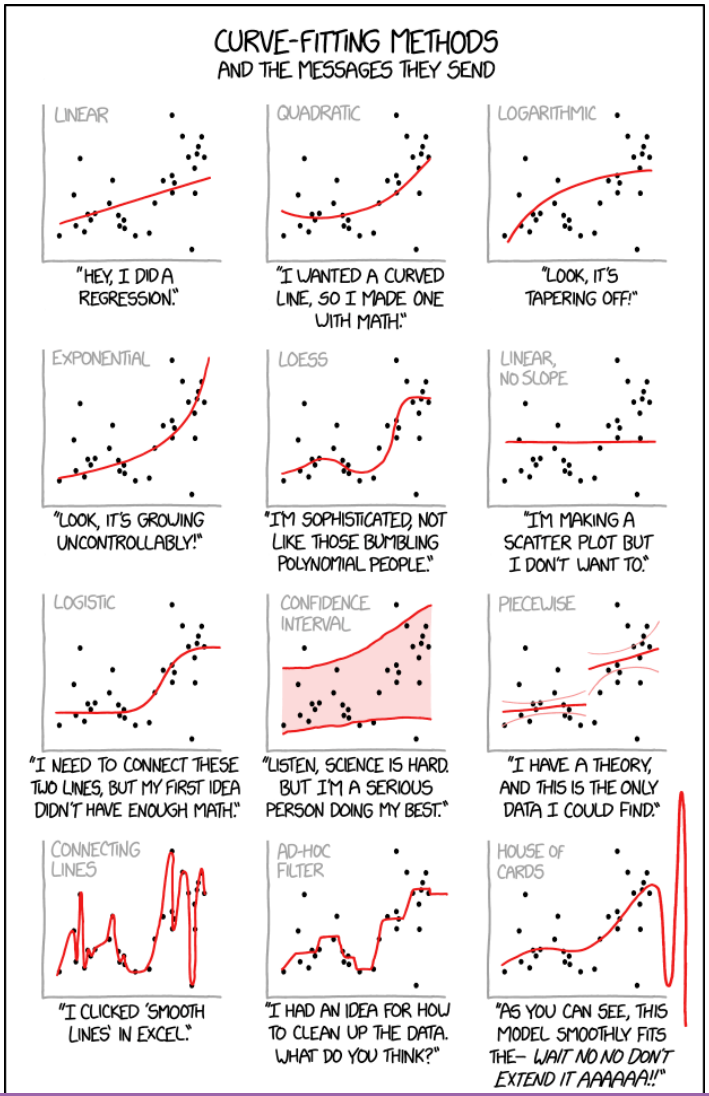
清华大学社会科学院政治学系

第十三讲　时间序列分析 (III)

线性回归模型基本假定

1. 因变量 y 和自变量 x 之间的关系是线性的 (linearity)、相加性的 (additivity)：
2. 余数项 (residual errors) 间是彼此独立的。
3. 余数项具同质性，也就是方差是固定的 (constant）variance)，$var(y_1) = var(y_2) \ldots var(y_n)$。
4. 余数项的分布呈正态分布，$E(\epsilon_i) = 0, var(\epsilon_i) = \sigma^2$，$\epsilon \sim N(0, \sigma^2)$

数据拟合的尝试


CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND
LINEAR
"HEY, I DID A REGRESSION."
QUADRATIC
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."
LOGARITHMIC
"LOOK, IT'S TAPERING OFF!"
EXPONENTIAL
"LOOK, IT'S GROWING UNCONTROLLABLY!"
LOESS
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."
LINEAR, NO SLOPE
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."
LOGISTIC
"I NEED TO CONNECT THESE TWO LINES, BUT MY FIRST IDEA DIDN'T HAVE ENOUGH MATH."
CONFIDENCE INTERVAL
"LISTEN, SCIENCE IS HARD. BUT I'M A SERIOUS PERSON DOING MY BEST."
PIECEWISE
"I HAVE A THEORY, AND THIS IS THE ONLY DATA I COULD FIND."
CONNECTING LINES
"I CLICKED 'SMOOTH LINES' IN EXCEL."
AD-HOC FILTER
"I HAD AN IDEA FOR HOW TO CLEAN UP THE DATA. WHAT DO YOU THINK?"
HOUSE OF CARDS
"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE– WAIT NO NO DON'T EXTEND IT AAAAAA!!"

KKV 的样本增量建议

- 时间序列横截面 (Time Series Cross Sectional, TSCS) 数据：不同的单元有至少 1 个时间段以上的观测点。
- 这类重复性观测数据的优势：
  - 数量上：增加样本量，解决自由度的问题。
  - 质量上：让我们可以回答不同单元在时间上的变化。
  - 思考：这样的数据结构对吗？我们关心的推论主体是什么？

TSCS 举例

- 假设我们有一份 TSCS 数据，观测到 40% 的拉丁裔美国公民参与选举投票，这表示两种可能：
  - 在特定的选举中，特定的拉丁裔公民的参与投票的概率为 40%。
  - 在任何选举中，40% 的拉丁裔美国公民参与选举投票，60% 的拉丁裔美国公民不参与选举投票。
- 如果我们没有逐年的观测选民，我们无法从数据中分别出以上两种情况。

重复性数据

- 当时间 $T >$ 单元样本量 N，称之为时间序列横截面数据 (TSCS Data)。
- 当单元样本量 N 时间 > 时间 T，称之为面板数据 (Panel Data)

$$y_{it} = \alpha + \beta x_{it} + \epsilon_{it}$$

- 如果将所有数据 pool 在一起，我们就有 $N \times T$ 的样本量，但是 . . .

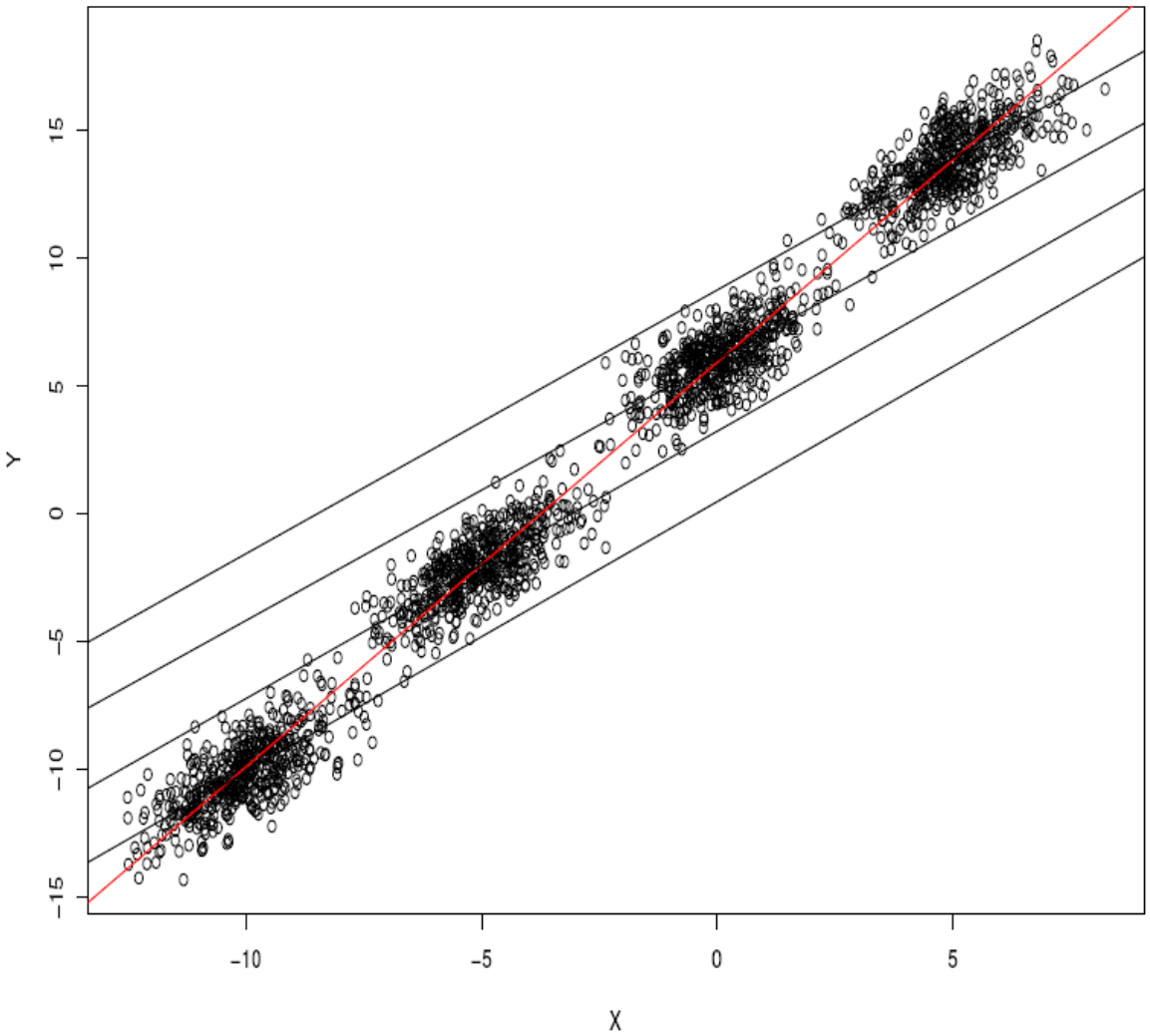
未观测到的国家效应

- 探讨政府支出与其开放程度的关系：

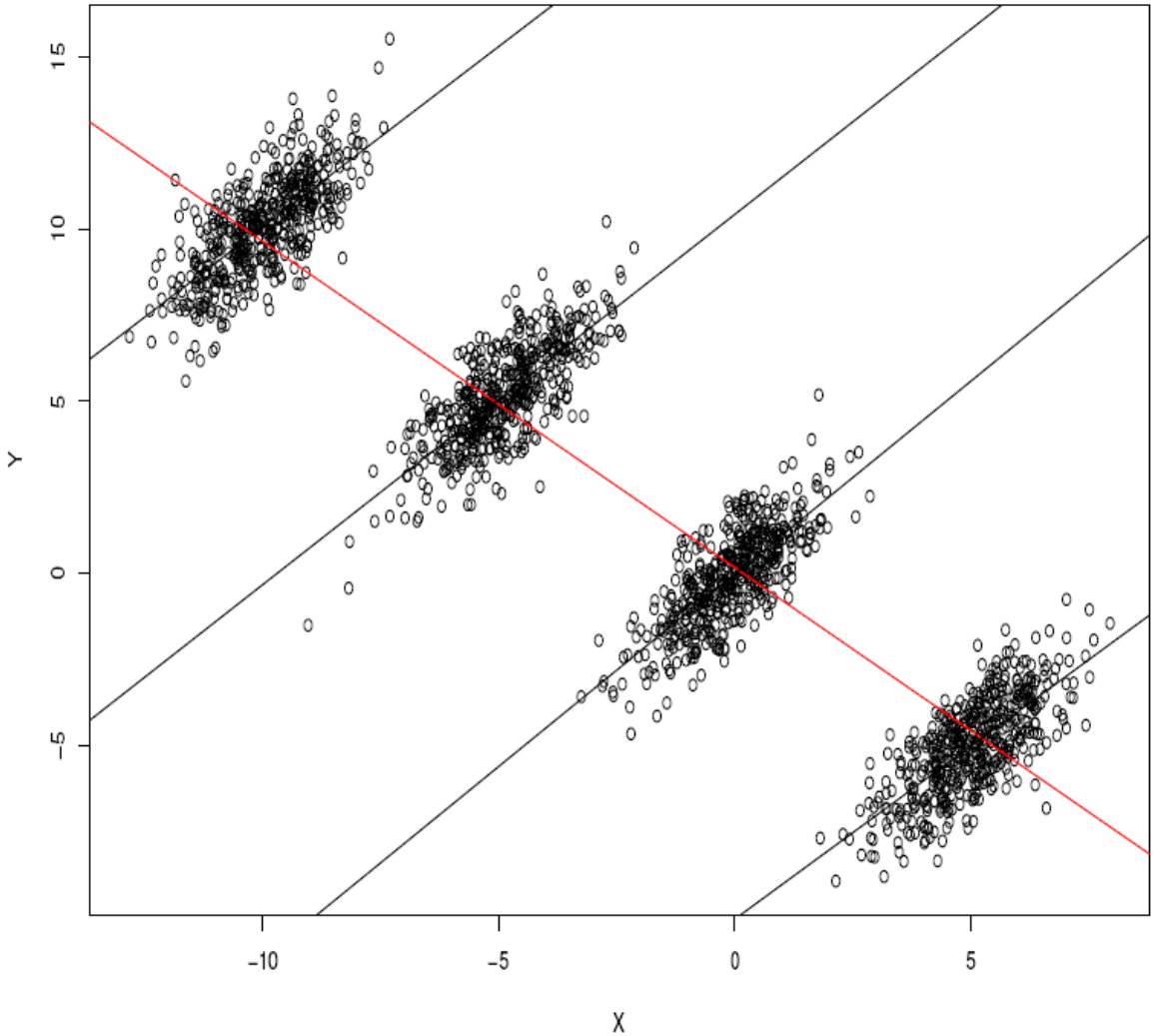
$$\text{政府支出}_{it} = \beta_0 + \beta_1 \text{开放程度}_{it} + \beta_2 Z_{it} + \epsilon_{it}$$

- 有可能各国政府的支出因为国情不同而有不同，称之为横截面异质性 (Cross-sectional heterogeneity)
- 如果这些单元（国家）特殊性原因与其他变量相关，我们就面临遗漏变量偏差 (omitted variable bias)，甚至我们会得到偏大的标准误。
- 如果我们可以搜集到这些特殊原因（变量），我们可以将这些变量放入回归右项，但是数据搜集可能费力费时。
- 或者，简单的方法就是在回归右项加入单元（国家）的虚拟变量。

未观测到的国家效应



未观测到的国家效应



时间效应

- 但是数据可能还存在时间效应，例如 1973-74 年，所有 OECD 经济体因为石油危机造成经济衰退，影响政府支出。
- 我们可以加入时间虚拟变量来捕捉时间效应。
- 现在自由度变成 $NT - k - N - T$。
- $k =$ 自变量个数

最小二阶方程虚拟变量回归, LSDV

- LSDV, Least Sqaures Dummy Variables model
- 即在回归中加入虚拟变量 Dummy Variables，捕捉单元效应和时间效应，也就是“固定效应”(fixed effects)。
- `lm(y x + factor(country) + factor(year))`
- 跑完回归后检验所有“固定效应”（虚拟变量的系数）是否 $= 0$
- 如果 F-test 显著，则拒绝 $H_0 : \beta_1 = \beta_2 = \ldots = \beta_j = 0$，因此，固定效应不为 0，使用固定效应模型是合适的。

LSDV, F-test

- 余数平方和，sum of residual squares，SSR

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- F-test 统计量

$$F = (k-1)\frac{SSR_p - SSR_f}{\frac{SSR_f}{df_f}}$$

k：虚拟变量数目; SSR_p: 全池模型余数平方和; SSR_f: 固定效应模型余数平方和；df_f: 固定效应模型自由度

- `pf(Fstats, df1=k-1, df2=`df_f`, lower.tail=FALSE)`

案例：OECD14 国的 GDP 与政党关系

- Garrett (1998) *Partisan Politics in the Global Economy*。
- 研究问题：劳工集中化程度（工会能力）和左派政府如何影响经济发展？
- 检验当政府与劳工同步时，是否会增进经济增长
- 14 个 OECD 国家，1966–1990 年期间 ($N = 14, T = 25$)
- 因变量：GDP
- 自变量：OILD（对石油依赖程度），LEFTLAB（左派政党占据内阁比例），CORP（劳工集中化程度），CLINT（CORP 和 LEFTCAB 的交叉项），DEMAND（所有 OECD 国家的经济增长率）

案例：OECD14 国的 GDP 与政党关系, 全池化模型 (pooled model)

```
> print(summary(mp), digits=4)
Pooling Model

Call:
plm(formula = gdp ~ oild + demand + corp + leftlab + clint, data = dat,
    model = "pooling", index = "country")

Balanced Panel: n = 14, T = 25, N = 350

...

Coefficients:
             Estimate Std. Error t-value Pr(>|t|)
(Intercept)   5.919865   0.735638   8.047 1.39e-14 ***
oild        -15.232100   4.572497  -3.331 0.000958 ***
demand        0.004998   0.000999   5.003 9.03e-07 ***
corp         -1.139716   0.304399  -3.744 0.000212 ***
leftlab      -1.483548   0.384465  -3.859 0.000136 ***
clint         0.454718   0.123378   3.686 0.000265 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    2065
Residual Sum of Squares: 1774
R-Squared:      0.141
```

案例：OECD14 国的 GDP 与政党关系，固定效应模型

加入 $N-1$ 个国家虚拟变量：

```
> print(summary(mf), digits=4)
Oneway (individual) effect Within Model

Call:
plm(formula = gdp ~ oild + demand + corp + leftlab + clint, data = dat,
    model = "within", index = "country")

Balanced Panel: n = 14, T = 25, N = 350

...

Coefficients:
          Estimate Std. Error t-value Pr(>|t|)
oild    -25.598084   5.946569  -4.305 2.21e-05 ***
demand    0.008495   0.001129   7.524 5.05e-13 ***
corp     -0.250064   0.665419  -0.376  0.70731
leftlab  -1.172257   0.446878  -2.623  0.00911 **
clint     0.503091   0.159668   3.151  0.00178 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    1793
Residual Sum of Squares: 1378
R-Squared:      0.2315
Adj. R-Squared: 0.1897
F-statistic: 19.943 on 5 and 331 DF, p-value: < 2.2e-16
```

案例：OECD14 国的 GDP 与政党关系

检验虚拟变量系数是否 $= 0$:

```
> # F-test manually
> # H0: u=0
> df1 <- df.residual(mp)-df.residual(mf) # number of dummies K-1
> df2 <- df.residual(mf)                 # degree of freedom of fixed effect model (N-K-1)
> ssrp <- sum(resid(mp)^2)               # sum of residual square of pooling model
> ssrf <- sum(resid(mf)^2)               # sum of residual square of fixed effect model
> fQuant <- df2*(ssrp-ssrf)/ssrf/df1
> pf(fQuant, df1, df2, lower.tail=FALSE)
[1] 1.318267e-12
> # significant means reject H0, ie u#0, so fixed effect model is appropriate
> # F-test in a function
> pFtest(mf, mp)

        F test for individual effects

data:  gdp ~ oild + demand + corp + leftlab + clint
F = 7.3089, df1 = 13, df2 = 331, p-value = 1.318e-12
alternative hypothesis: significant effects
```

随机效应 (random effect)

- 固定效应模型是合适的前提是我们相信单元效应是固定的。
- 例如，我们估计到瑞典的固定效应为 1.2，如果我们收集到新数据，我们同样也会估计到 1.2 的固定效应。
- 但是我们有理由相信这个理解是过分天真的，各个单元的效应应该是还来自于来自于随机部分（随机效应）。

随机效应 (random effect) vs 固定效应 (fixed effect)

- 当 T 足够大时，两个模型估计的结果是一样的。
- 但是如果 T 不够大，而 N 很大，则结果会差异很大。
- 如果自变量中存在着不随时间改变且单元特定的变量，则固定效应就无法使用，例如一国的国土面积。
- 但是如果单元效应与自变量互相关，则随机效应模型便不合适。关于此点，可以用 Hausman 检验

$$H = (\hat{\beta}_{fe} - \hat{\beta}_{re})' \left[\text{var}(\beta_{fe}) - \text{var}(\beta_{re})\right]^{-1} (\hat{\beta}_{fe} - \hat{\beta}_{re})$$

- Hausman 统计量服从 χ^2 分布

$$H \sim \chi^2_k$$

- 自由度 k 是回归系数的数目

随机效应 (random effect)

```
> mr <- plm(gdp ~ oild + demand + corp + leftlab + clint, data=dat, index = "country", model = "random")
> summary(mr)
Oneway (individual) effect Random Effect Model   (Swamy-Arora's transformation)
...
Balanced Panel: n = 14, T = 25, N = 350
Effects:
                 var std.dev share
idiosyncratic 4.1639  2.0406  0.81
individual    0.9746  0.9872  0.19
theta: 0.618
...
Coefficients:
               Estimate  Std. Error z-value  Pr(>|z|)
(Intercept)   5.1983897   1.1118857  4.6753 2.935e-06 ***
oild        -20.4460163   5.3942568 -3.7903 0.0001504 ***
demand        0.0075601   0.0010875  6.9519 3.604e-12 ***
corp         -1.2100372   0.4209980 -2.8742 0.0040504 **
leftlab      -1.2560966   0.4275844 -2.9377 0.0033070 **
clint         0.4653267   0.1481581  3.1407 0.0016852 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    1833.1
Residual Sum of Squares: 1462.5
R-Squared:      0.20219
Adj. R-Squared: 0.19059
Chisq: 87.1789 on 5 DF, p-value: < 2.22e-16
```

随机效应 (random effect) 检验

```
> # Hausmen test manually
> # H0 Re is appropriate, fe=re, re is more efficient
> # Ha Re is inappropriate, fe#re, re is biased/inconsistent
> mfB <- coef(mf)[1:5]
> mrB <- coef(mr)[2:6]
> mfV <- vcov(mf)[1:5,1:5]
> mrV <- vcov(mr)[2:6,2:6]
> hQuant <- t(mfB-mrB)%*%solve(mfV-mrV)%*%(mfB-mrB)
> hQuant
         [,1]
[1,] 15.39224
> pchisq(hQuant, 5, lower.tail=FALSE)
            [,1]
[1,] 0.008811462
> phtest(mf, mr)

        Hausman Test

data:  gdp ~ oild + demand + corp + leftlab + clint
chisq = 15.392, df = 5, p-value = 0.008811
alternative hypothesis: one model is inconsistent
```

检验结果统计显著：拒绝随机效应是合适的。

Panel Corrected Standard Errors (PCSEs)

- 同期相关性（Contemporaneous correlation）：余数在同一时期互为相关。

$$\mathrm{E}(\epsilon_{it}, \epsilon_{js}) = \begin{cases} \sigma_i^2 & \text{if } i = j \text{ and } s = t \\ \sigma_{ij} & \text{if } i \neq j \text{ and } s = t \\ 0 & \text{otherwise} \end{cases}$$

- 面板异质性（Panel heteroskedasticity）：余数在不同单元各有特殊性

$$\mathrm{E}(\epsilon_{it}, \epsilon_{js}) = \begin{cases} \sigma_i^2 & \text{if } i = j \text{ and } s = t \\ 0 & \text{otherwise} \end{cases}$$

- 序相关（Serial correlation）：各单元的余数在时间上互为相关

$$\epsilon_{it} = \rho \epsilon_{i,t-1} + \nu_{it}$$

Panel Corrected Standard Errors (PCSEs)

- 如果有以上问题（除了序相关之外），用 OLS 跑 TSCS 数据，系数是无偏的，但是系数的标准误是错的。
- Beck and Katz（1995）建议使用 PCSEs。
- PCSEs 就是将同期相关性和面板异质性的余数结构 Ω 考虑进去，重新调整 OLS 产生的余数结果。
- PCSE 计算公式：

$$(X'X)^{-1}X'\hat{\Omega}X(X'X)^{-1}$$

$$\hat{\Omega} = \hat{\Sigma} \otimes I_T$$

$$\hat{\Sigma} = \frac{\sum_{jt} \epsilon'_{jt}\epsilon_{jt}}{T}$$

T：时间个数

Panel Corrected Standard Errors (PCSEs) 实作

```
> mOLS <- lm(gdp ~ oild + demand + corp + leftlab + clint, data=dat)
> summary(mOLS)

Call:
lm(formula = gdp ~ oild + demand + corp + leftlab + clint, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-7.0546 -1.4306 -0.0494  1.3033  9.3270

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.919865   0.735638   8.047 1.39e-14 ***
oild        -15.232100   4.572497  -3.331 0.000958 ***
demand        0.004998   0.000999   5.003 9.03e-07 ***
corp         -1.139716   0.304399  -3.744 0.000212 ***
leftlab      -1.483548   0.384465  -3.859 0.000136 ***
clint         0.454718   0.123378   3.686 0.000265 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.271 on 344 degrees of freedom
Multiple R-squared:  0.141,     Adjusted R-squared:  0.1286
F-statistic:  11.3 on 5 and 344 DF,  p-value: 4.208e-10
```

Panel Corrected Standard Errors (PCSEs) 实作

```
> library(pcse)
> mPCSE <- pcse(mOLS, dat$country, dat$year)
> summary(mPCSE)

 Results:

                 Estimate        PCSE   t value     Pr(>|t|)
(Intercept)   5.919865231 0.583394944 10.147269 2.447368e-21
oild        -15.232100343 5.228693724 -2.913175 3.811626e-03
demand        0.004997676 0.001539438  3.246429 1.283547e-03
corp         -1.139715542 0.223408756 -5.101481 5.582131e-07
leftlab      -1.483548515 0.275584685 -5.383276 1.356887e-07
clint         0.454718261 0.083952631  5.416367 1.144746e-07

 ---------------------------------------------

# Valid Obs = 350; # Missing Obs = 0; Degrees of Freedom = 344.
```

PCSEs 潜在的问题

- 忽略序相关的问题，因此数据必须预先解决序相关的问题。
- 建议使用滞后变量解决序相关问题，但是可能有其他问题。
  - 面板不平衡的问题：例：每个国家在时间维度上观测的次数不一致
  - 面板丢失的问题：例：某些国家其中几年数据缺失
  - ……

Panel Corrected Standard Errors (PCSEs) 加序相关实作

```
> library(panelAR)
> mPCSEAR1 <- panelAR(gdp ~ oild + demand + corp + leftlab + clint, data=dat, panelVar = "country",
timeVar = "year", autoCorr = "ar1",  panelCorrMethod = "pcse")
> summary(mPCSEAR1)

Panel Regression with AR(1) Prais-Winsten correction and panel-corrected standard errors

Balanced Panel Design:
 Total obs.:       350 Avg obs. per panel 25
 Number of panels: 14  Max obs. per panel 25
 Number of times:  25  Min obs. per panel 25

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.814019   0.807692   7.198 3.86e-12 ***
oild        -13.772264   6.587739  -2.091 0.037298 *
demand        0.006081   0.001641   3.705 0.000247 ***
corp         -1.177445   0.293402  -4.013 7.36e-05 ***
leftlab      -1.467760   0.362348  -4.051 6.32e-05 ***
clint         0.448846   0.111223   4.036 6.72e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-squared: 0.1516
Wald statistic: 31.5475, Pr(>Chisq(5)): 0
>
```

固定和随机效应中的序相关问题

- Arellano Robust Standard Error
- Kiefer Robust Standard Error
- 两者都假设 T 是固定的，而 $N \to \infty$。
- MCMC 模拟下，Arellano 多数情况下优于 Kiefer
- 事实上 Arellano Robust Standard Error 可以透过 LSDV+PCSEs 获得。

Arellano Robust Standard Errors 实作

```
> mPCSEArellano <- panelAR(gdp ~ oild + demand + corp + leftlab + clint + factor(country), data=dat, panel
> summary(mPCSEArellano)

Panel Regression with no autocorrelation and panel-corrected standard errors

Balanced Panel Design:
 Total obs.:       350 Avg obs. per panel 25
 Number of panels: 14  Max obs. per panel 25
 Number of times:  25  Min obs. per panel 25

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          3.374094   1.212579   2.783 0.005702 **
oild               -25.598084   7.590455  -3.372 0.000833 ***
demand               0.008495   0.001847   4.600 6.02e-06 ***
corp                -0.250064   0.578110  -0.433 0.665620
leftlab             -1.172257   0.348437  -3.364 0.000857 ***
clint                0.503091   0.144974   3.470 0.000589 ***
...
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-squared: 0.3326
Wald statistic: 118.9057, Pr(>Chisq(18)): 0
>
```