10.18 Consider the data file *mroz* on working wives. Use the 428 observations on married women who participate in the labor force. In this exercise, we examine the effectiveness of a parent's college education as an instrumental variable.

a. Create two new variables. *MOTHERCOLL* is a dummy variable equaling one if *MOTHEREDUC* > 12, zero otherwise. Similarly, *FATHERCOLL* equals one if *FATHEREDUC* > 12 and zero otherwise. What percentage of parents have some college education in this sample?

```
> # 輸出結果
> cat("=== (a) Percentage of Parents with Some College Education ===\n")
=== (a) Percentage of Parents with Some College Education ===
> cat("Mother: ", mother_pct, "%\n")
Mother:  12.15 %
> cat("Father: ", father_pct, "%\n")
Father:  11.68 %
```

母親：12.15%
父親：11.68%

b. Find the correlations between *EDUC*, *MOTHERCOLL*, and *FATHERCOLL*. Are the magnitudes of these correlations important? Can you make a logical argument why *MOTHERCOLL* and *FATHERCOLL* might be better instruments than *MOTHEREDUC* and *FATHEREDUC*?

```
> print(round(cor_matrix, 3))
             educ MOTHERCOLL FATHERCOLL
educ        1.000      0.359      0.398
MOTHERCOLL  0.359      1.000      0.355
FATHERCOLL  0.398      0.355      1.000
```

mothercoll與fathercoll與子女的EDUC呈現中度正相關，這兩個虛擬變數符合工具變數的相關性條件；相較於連續變數的父母教育年數，「是否上過大學」這類虛擬變數可減少測量誤差

c. Estimate the wage equation in Example 10.5 using *MOTHERCOLL* as the instrumental variable. What is the 95% interval estimate for the coefficient of *EDUC*?

```
Call:
ivreg(formula = log(wage) ~ educ + exper + exper2 | MOTHERCOLL +
    exper + exper2, data = iv_data)

Residuals:
     Min       1Q   Median       3Q      Max
-3.08719 -0.32444  0.04147  0.36634  2.35621

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.1327561  0.4965325  -0.267  0.78932
educ         0.0760180  0.0394077   1.929  0.05440 .
exper        0.0433444  0.0134135   3.231  0.00133 **
exper2      -0.0008711  0.0004017  -2.169  0.03066 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6703 on 424 degrees of freedom
Multiple R-Squared: 0.147,      Adjusted R-squared: 0.1409
Wald test:  8.2 on 3 and 424 DF,  p-value: 2.569e-05
```

```
> # 取得 95% 信賴區間
> confint(iv_model, level = 0.95)
                  2.5 %       97.5 %
(Intercept) -1.105942034  8.404298e-01
educ        -0.001219763  1.532557e-01
exper        0.017054428  6.963439e-02
exper2      -0.001658392 -8.385898e-05
```

95% CI of EDUC

[-0.0012, 0.1533]

d. For the problem in part (c), estimate the first-stage equation. What is the value of the *F*-test statistic for the hypothesis that *MOTHERCOLL* has no effect on *EDUC*? Is *MOTHERCOLL* a strong instrument?

```
Linear hypothesis test:
MOTHERCOLL = 0

Model 1: restricted model
Model 2: educ ~ MOTHERCOLL + exper + exper2

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    425 2219.2
2    424 1929.9  1    289.32 63.563 1.455e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

F = 63.563 > 10

MOTHERColl 為 a strong instrument

e. Estimate the wage equation in Example 10.5 using *MOTHERCOLL* and *FATHERCOLL* as the instrumental variables. What is the 95% interval estimate for the coefficient of *EDUC*? Is it narrower or wider than the one in part (c)?

```
Call:
ivreg(formula = log(wage) ~ educ + exper + exper2 | MOTHERCOLL +
    FATHERCOLL + exper + exper2, data = iv_data2)

Residuals:
     Min       1Q   Median       3Q      Max
-3.07797 -0.32128  0.03418  0.37648  2.36183

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.2790819  0.3922213  -0.712  0.47714
educ         0.0878477  0.0307808   2.854  0.00453 **
exper        0.0426761  0.0132950   3.210  0.00143 **
exper2      -0.0008486  0.0003976  -2.135  0.03337 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6679 on 424 degrees of freedom
Multiple R-Squared: 0.153,      Adjusted R-squared: 0.147
Wald test: 9.724 on 3 and 424 DF,  p-value: 3.224e-06

> # 計算 95% 信賴區間
> confint(iv_model2, level = 0.95)
                  2.5 %        97.5 %
(Intercept) -1.04782153  4.896578e-01
educ         0.02751845  1.481769e-01
exper        0.01661839  6.873386e-02
exper2      -0.00162779 -6.940599e-05
>
```

95% CI of EDUC

= [0.027, 0.148]

CI narrower than part(c), 更有效率

f. For the problem in part (e), estimate the first-stage equation. Test the joint significance of *MOTHERCOLL* and *FATHERCOLL*. Do these instruments seem adequately strong?

```
                                                    < 2.2e-16
> linearHypothesis(first_stage2, c("MOTHERCOLL = 0", "FATHERCOLL = 0"))

Linear hypothesis test:
MOTHERCOLL = 0
FATHERCOLL = 0

Model 1: restricted model
Model 2: educ ~ MOTHERCOLL + FATHERCOLL + exper + exper2

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    425 2219.2
2    423 1748.3  2    470.88 56.963 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

F = 56.963 > 10

兩個工具變數共同具備強解釋力

g. For the IV estimation in part (e), test the validity of the surplus instrument. What do you conclude?

```
Diagnostic tests:
                 df1 df2 statistic p-value
Weak instruments   2 423    56.963  <2e-16 ***
Wu-Hausman         1 423     0.519   0.472
Sargan             1  NA     0.238   0.626
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

Residual standard error: 0.6679 on 424 degrees of freedom
Multiple R-Squared: 0.153,      Adjusted R-squared: 0.147
Wald test: 9.724 on 3 and 424 DF,  p-value: 3.224e-06
```

By sargan test
p-value = 0.626 > 0.05
don't reject H_0, 多出來的工具變數(FATHERColl)是有效的，不違反外生性假設

10.18R

```r
#10.18(a)
mroz_lfp <- subset(mroz, lfp == 1)
# 建立虛擬變數 (dummy variables)
mroz_lfp$MOTHERCOLL <- as.integer(mroz_lfp$mothereduc > 12)
mroz_lfp$FATHERCOLL <- as.integer(mroz_lfp$fathereduc > 12)

# 計算母親與父親有部分大學教育的比例 (忽略 NA)
mother_pct <- round(mean(mroz_lfp$MOTHERCOLL, na.rm = TRUE) * 100, 2)
father_pct <- round(mean(mroz_lfp$FATHERCOLL, na.rm = TRUE) * 100, 2)

# 輸出結果
cat("=== (a) Percentage of Parents with Some College Education ===\n")
cat("Mother: ", mother_pct, "%\n")
cat("Father: ", father_pct, "%\n")

#(b)
# 選取三個變數組成新資料框
corr_data <- mroz_lfp[, c("educ", "MOTHERCOLL", "FATHERCOLL")]

# 計算相關係數矩陣 (排除 NA)
cor_matrix <- cor(corr_data, use = "complete.obs")

# 印出結果
print(round(cor_matrix, 3))

#(c)
# 加入 exper^2, 建立變數
iv_data <- subset(mroz_lfp, !is.na(wage) & !is.na(educ) & !is.na(MOTHERCOLL) & !is.na(exper))
iv_data$exper2 <- iv_data$exper^2

# 工具變數回歸 (educ 為內生變數, MOTHERCOLL 為工具)
iv_model <- ivreg(log(wage) ~ educ + exper + exper2 | MOTHERCOLL + exper + exper2, data = iv_data)

# 顯示結果摘要
summary(iv_model)

# 取得 95% 信賴區間
confint(iv_model, level = 0.95)
```

```r
#(d)
# 第一階段回歸：educ ~ MOTHERCOLL + controls
first_stage <- lm(educ ~ MOTHERCOLL + exper + exper2, data = iv_data)

# 顯示結果摘要 (看 MOTHERCOLL 的係數與 F 值)
summary(first_stage)
linearHypothesis(first_stage, "MOTHERCOLL = 0")

#(e)
# 確保資料不含缺失值
iv_data2 <- subset(mroz_lfp, !is.na(wage) & !is.na(educ) &
                     !is.na(MOTHERCOLL) & !is.na(FATHERCOLL) & !is.na(exper))

# 加入 exper^2
iv_data2$exper2 <- iv_data2$exper^2

# 使用 MOTHERCOLL 和 FATHERCOLL 兩個工具
iv_model2 <- ivreg(log(wage) ~ educ + exper + exper2 |
                     MOTHERCOLL + FATHERCOLL + exper + exper2,
                   data = iv_data2)

# 顯示摘要
summary(iv_model2)

# 計算 95% 信賴區間
confint(iv_model2, level = 0.95)

#(f)
first_stage2 <- lm(educ ~ MOTHERCOLL + FATHERCOLL + exper + exper2, data = iv_data2)

# 查看摘要 (可觀察 R²、係數、顯著性等)
summary(first_stage2)
linearHypothesis(first_stage2, c("MOTHERCOLL = 0", "FATHERCOLL = 0"))

#(g)
summary(iv_model2, diagnostics = TRUE)
```

10.20 The CAPM [see Exercises 10.14 and 2.16] says that the risk premium on security j is related to the risk premium on the market portfolio. That is

$$r_j - r_f = \alpha_j + \beta_j(r_m - r_f)$$

where r_j and r_f are the returns to security j and the risk-free rate, respectively, r_m is the return on the market portfolio, and β_j is the jth security's "*beta*" value. We measure the market portfolio using the Standard & Poor's value weighted index, and the risk-free rate by the 30-day LIBOR monthly rate of return. As noted in Exercise 10.14, if the market return is measured with error, then we face an errors-in-variables, or measurement error, problem.

a. Use the observations on Microsoft in the data file *capm5* to estimate the CAPM model using OLS. How would you classify the Microsoft stock over this period? Risky or relatively safe, relative to the market portfolio?

```
Call:
lm(formula = Rj_minus_Rf ~ Rm_minus_Rf, data = capm5)

Residuals:
     Min       1Q   Median       3Q      Max
-0.27424 -0.04744 -0.00820  0.03869  0.35801

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.003250   0.006036   0.538    0.591
Rm_minus_Rf 1.201840   0.122152   9.839   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08083 on 178 degrees of freedom
Multiple R-squared:  0.3523,    Adjusted R-squared:  0.3486
F-statistic:  96.8 on 1 and 178 DF,  p-value: < 2.2e-16

>
```

$r_{MSFT} - r_f = 0.0033 + 1.2018(r_m - r_f)$

$\because \beta = 1.20 > 1$ $\therefore$ 波動大於市場

risky relative to market portfolio

b. It has been suggested that it is possible to construct an IV by ranking the values of the explanatory variable and using the rank as the IV, that is, we sort $(r_m - r_f)$ from smallest to largest, and assign the values *RANK* $= 1, 2, \ldots, 180$. Does this variable potentially satisfy the conditions IV1–IV3? Create *RANK* and obtain the first-stage regression results. Is the coefficient of *RANK* very significant? What is the R^2 of the first-stage regression? Can *RANK* be regarded as a strong IV?

```
Call:
lm(formula = Rm_minus_Rf ~ RANK, data = capm5)

Residuals:
      Min        1Q    Median        3Q       Max
-0.110497 -0.006308  0.001497  0.009433  0.029513

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.903e-02  2.195e-03   -36.0   <2e-16 ***
RANK         9.067e-04  2.104e-05    43.1   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01467 on 178 degrees of freedom
Multiple R-squared:  0.9126,    Adjusted R-squared:  0.9121
F-statistic:  1858 on 1 and 178 DF,  p-value: < 2.2e-16

>
```

first-stage regression

$r_m - r_f = -0.0790 + 0.000907\,\text{Rank}$

t value $= 43.1$ 很 significant

$R^2 = 0.9126$

Rank 是 strong IV

c. Compute the first-stage residuals, $\hat{v}$, and add them to the CAPM model. Estimate the resulting augmented equation by OLS and test the significance of $\hat{v}$ at the 1% level of significance. Can we conclude that the market return is exogenous?

```
Call:
lm(formula = Rj_minus_Rf ~ Rm_minus_Rf + vhat, data = capm5)

Residuals:
      Min        1Q    Median        3Q       Max
-0.27140 -0.04213 -0.00911  0.03423  0.34887

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.003018   0.005984   0.504   0.6146
Rm_minus_Rf  1.278318   0.126749  10.085   <2e-16 ***
vhat        -0.874599   0.428626  -2.040   0.0428 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08012 on 177 degrees of freedom
Multiple R-squared:  0.3672,    Adjusted R-squared:  0.36
F-statistic: 51.34 on 2 and 177 DF,  p-value: < 2.2e-16
```

$\hat{v}$ p-value = 0.0428 > 0.01

don't reject H_0, the market return is exogenous

d. Use *RANK* as an IV and estimate the CAPM model by IV/2SLS. Compare this IV estimate to the OLS estimate in part (a). Does the IV estimate agree with your expectations?

```
Call:
ivreg(formula = Rj_minus_Rf ~ Rm_minus_Rf | RANK, data = capm5)

Residuals:
      Min        1Q    Median        3Q       Max
-0.271625 -0.049675 -0.009693  0.037683  0.355579

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.003018   0.006044   0.499    0.618
Rm_minus_Rf 1.278318   0.128011   9.986   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08092 on 178 degrees of freedom
Multiple R-Squared: 0.3508,     Adjusted R-squared: 0.3472
Wald test: 99.72 on 1 and 178 DF,  p-value: < 2.2e-16
```

$\hat{\beta}_{IV} = 1.278 > 1.202 = \hat{\beta}_{OLS}$

OLS可能存在低估风险的偏误，但差異不大且不顯著，OLS結果大致可靠

e. Create a new variable $POS = 1$ if the market return $(r_m - r_f)$ is positive, and zero otherwise. Obtain the first-stage regression results using both *RANK* and *POS* as instrumental variables. Test the joint significance of the IV. Can we conclude that we have adequately strong IV? What is the R^2 of the first-stage regression?

```
Call:
lm(formula = Rm_minus_Rf ~ RANK + POS, data = capm5)

Residuals:
      Min        1Q    Median        3Q       Max
-0.109182 -0.006732  0.002858  0.008936  0.026652

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0804216  0.0022622  -35.55   <2e-16 ***
RANK         0.0009819  0.0000400   24.55   <2e-16 ***
POS         -0.0092762  0.0042156   -2.20   0.0291 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01451 on 177 degrees of freedom
Multiple R-squared:  0.9149,    Adjusted R-squared:  0.9139
F-statistic: 951.3 on 2 and 177 DF,  p-value: < 2.2e-16
```

```
Linear hypothesis test:
RANK = 0
POS = 0

Model 1: restricted model
Model 2: Rm_minus_Rf ~ RANK + POS

  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    179 0.43784
2    177 0.03727  2   0.40057 951.26 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

joint IV, $F = 95.26 > 10$, $R^2 = 0.915$

Rank and POS are strong IV

f. Carry out the Hausman test for endogeneity using the residuals from the first-stage equation in (e). Can we conclude that the market return is exogenous at the 1% level of significance?

```
Call:
lm(formula = Rj_minus_Rf ~ Rm_minus_Rf + vhat, data = capm5)

Residuals:
      Min        1Q    Median        3Q       Max
-0.27132 -0.04261 -0.00812  0.03343  0.34867

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.003004   0.005972   0.503   0.6157
Rm_minus_Rf  1.283118   0.126344  10.156   <2e-16 ***
vhat        -0.954918   0.433062  -2.205   0.0287 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07996 on 177 degrees of freedom
Multiple R-squared:  0.3696,    Adjusted R-squared:  0.3625
F-statistic: 51.88 on 2 and 177 DF,  p-value: < 2.2e-16

>
```

$\hat{v}$ p-value $= 0.0287 > 0.01$

don't reject H_0

the market return is exogenous

g. Obtain the IV/2SLS estimates of the CAPM model using *RANK* and *POS* as instrumental variables. Compare this IV estimate to the OLS estimate in part (a). Does the IV estimate agree with your expectations?

```
Call:
ivreg(formula = Rj_minus_Rf ~ Rm_minus_Rf | RANK + POS, data = capm5)

Residuals:
      Min        1Q    Median        3Q       Max
-0.27168 -0.04960 -0.00983  0.03762  0.35543

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.003004   0.006044   0.497     0.62
Rm_minus_Rf 1.283118   0.127866  10.035   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08093 on 178 degrees of freedom
Multiple R-Squared: 0.3507,     Adjusted R-squared: 0.347
Wald test: 100.7 on 1 and 178 DF,  p-value: < 2.2e-16

>
```

$\hat{\beta}_{IV_2} = 1.283 > \hat{\beta}_{OLS}$

$\hat{\beta}_{IV_2}$ is significant

與預期一致

h. Obtain the IV/2SLS residuals from part (g) and use them (not an automatic command) to carry out a Sargan test for the validity of the surplus IV at the 5% level of significance.

```
> cat("Sargan Test Statistic =", round(sargan_stat, 4), "\n")
Sargan Test Statistic = 0.5585
> cat("p-value =", round(p_value, 4), "\n")
p-value = 0.4549
>
```

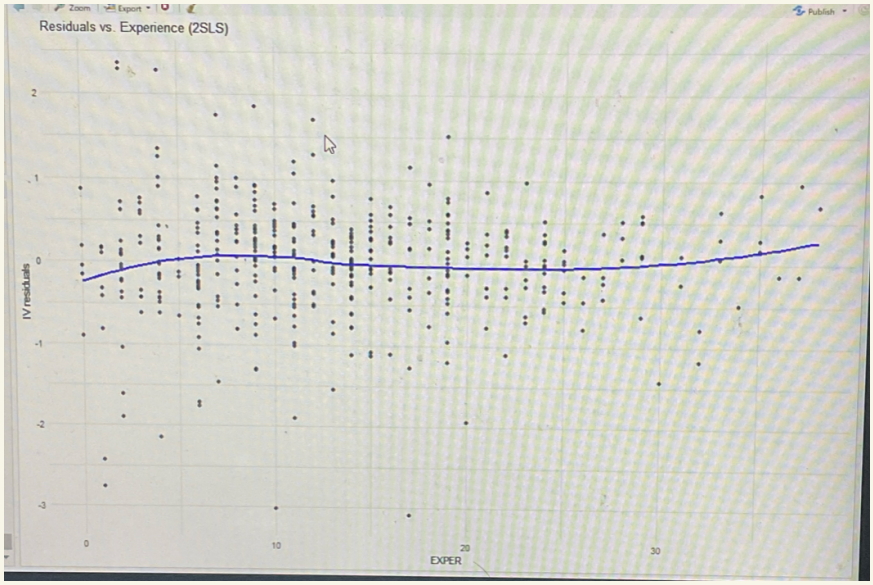
p-value $= 0.4549 > 0.05$

don't reject H_0

Rank and POS are valid IV (overidentification is valid)

10.24 Consider the data file *mroz* on working wives. Use the 428 observations on married women who participate in the labor force. In this exercise, we examine the effectiveness of alternative standard errors for the IV estimator. Estimate the model in Example 10.5 using IV/2SLS using both *MOTHEREDUC* and *FATHEREDUC* as IV. These will serve as our baseline results.

a. Calculate the IV/2SLS residuals, $\hat{e}_{IV}$. Plot them versus *EXPER*. Do the residuals exhibit a pattern consistent with homoskedasticity?



在EXPER小的地方殘差擾散較大，呈現漏斗形

→ heteroskedasticity

b. Regress $\hat{e}_{IV}^2$ against a constant and *EXPER*. Apply the NR^2 test from Chapter 8 to test for the presence of heteroskedasticity.

```
> cat("NR² 統計量 =", round(NR2, 4), "\n")
NR² 統計量 = 7.4386
> cat("p-value =", round(p_value, 4), "\n")
p-value = 0.0064
> # 結論
> if (p_value < 0.05) {
+   cat("結論：拒絕同變異性假設，存在異變異性。\n")
+ } else {
+   cat("結論：無法拒絕同變異性假設，殘差可能為同變異。\n")
+ }
結論：拒絕同變異性假設，存在異變異性。
```

p-value = 0.0064 < 0.05 = α

reject H_0

exist heteroskedasticity

c. Obtain the IV/2SLS estimates with the software option for Heteroskedasticity Robust Standard Errors. Are the robust standard errors larger or smaller than those for the baseline model? Compute the 95% interval estimate for the coefficient of *EDUC* using the robust standard error.

```
教育變數 EDUC 的 95% 信賴區間為：
> cat("[", round(lower_bound, 4), ",", round(upper_bound, 4), "]\n")
[ -0.0046 , 0.1274 ]
> # Baseline model summary (未使用 robust SE)
> baseline_summary <- summary(iv_model4)
> # 比較兩個模型中的 EDUC 標準誤
> baseline_se <- baseline_summary$coefficients["educ", "Std. Error"]
> robust_se <- robust_summary$coefficients["educ", "Std. Error"]
> cat("EDUC baseline SE:", round(baseline_se, 4), "\n")
EDUC baseline SE: 0.0314
> cat("EDUC robust SE:", round(robust_se, 4), "\n")
EDUC robust SE: 0.0337
>
```

95% CI of EDUC

= [-0.0046, 0.1274]

d. Obtain the IV/2SLS estimates with the software option for Bootstrap standard errors, using $B = 200$ bootstrap replications. Are the bootstrap standard errors larger or smaller than those for the baseline model? How do they compare to the heteroskedasticity robust standard errors in (c)? Compute the 95% interval estimate for the coefficient of *EDUC* using the bootstrap standard error.

```
> cat("Bootstrap 標準誤 (EDUC) :", round(boot_se, 4), "\n")
Bootstrap 標準誤 (EDUC) : 0.0323
> cat("Bootstrap 估計值 (EDUC) :", round(boot_coef, 4), "\n")
Bootstrap 估計值 (EDUC) : 0.0641
> cat("95% Bootstrap CI for EDUC: [", round(boot_ci_lower, 4), ",", round(boot_ci_upper, 4), "]\n"
95% Bootstrap CI for EDUC: [ 7e-04 , 0.1275 ]
```

```
> cat("Baseline SE for EDUC:", round(baseline_se, 4), "\n")
Baseline SE for EDUC: 0.0314
> cat("Robust SE for EDUC:", round(robust_se, 4), "\n")
Robust SE for EDUC: 0.0337
> cat("Bootstrap SE for EDUC:", round(boot_se, 4), "\n")
Bootstrap SE for EDUC: 0.0323
>
```

Baseline SE 最小，假設同變異數

Robust SE 最大，模型殘差存在異變異數

Bootstrap SE 位於兩者之間，不依賴對誤差結構的明確假設