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?

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*?

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*?

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?

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)?

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?

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

(a) 母親有 college degree 的比例: 12.15 %

父親有 college degree 的比例: 11.68 %

(b)

	educ	MOTHERCOLL	FATHERCOLL
educ	1.0000000	0.3594705	0.3984962
MOTHERCOLL	0.3594705	1.0000000	0.3545709
FATHERCOLL	0.3984962	0.3545709	1.0000000

① IV必須與自變數具相關性，而兩者相關性中等，故可以使用。

② 使用 dummy variables 變數較單純，不會讓其它家庭背景因素干擾

(c) first stage:

$$educ_i = \pi_0 + \pi_1 \underline{MOTHERCOLL_i} + \pi_2 exper_i + \pi_3 \cdot expersq_i + v_i$$

先求出各係數 π_i，再將各 x_i 代入得到 $\widehat{educ_i}$

second stage:

$$wage_i = \beta_0 + \beta_1 \widehat{educ_i} + \beta_2 exper_i + \beta_3 expersq_i + u_i$$

result:

$$wage = -1.3891 + \underbrace{0.4069}_{\beta_1} educ + 0.0417\, exper - 0.0005\, expersq$$

β_1 的 95% C.I. $[-0.0012198, 0.1532537]$

(d) $F = 63.563 > 10$，顯示 MOTHERCOLL 是一個強的工具變數

(通常設定 $F=10$ 為門檻)

(e) 如果用 Mothercoll，Fathercoll 兩個 IV，估計 $\widehat{educ}$

second stage：

$$wage = -1.8725 + \underbrace{0.4460}_{\beta_1} educ + 0.0395\, exper - 0.00047\, expersq$$

95% β_1 的 C.I. $[0.02735, 0.14835]$

(f)

first stage:

$$\widehat{educ}_i = 11.8903 + 1.7499\, Mothercoll + 2.1866\, Fathercoll + 0.0491\, exper - 0.0014\, expersq$$

(g) H_0 : Mothercoll $= 0$, Fathercoll $= 0$

H_a : 其中1個或以上 $\neq 0$

$F = 56.96 > 10 \longrightarrow$ strong instrument

(h) surplus instrument $\longrightarrow$ Fathercoll

Sargan test of overidentifying restriction

$\begin{cases} H_0: \text{所有 IV 都是外生的} \\ H_a: \text{not 所有 IV 都是外生的} \end{cases}$

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?

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?

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?

$R_i = R_f + \beta(R_m - R_f)$

a. $r_{MSFT} - r_f = \alpha + \beta(r_m - r_f) + \varepsilon$

$\Rightarrow r_{MSFT} - r_f = 0.00325 + 1.02184(r_m - r_f) + \varepsilon$

$\beta = 1.20 > 1 \rightarrow$ relative risky stock

b. ∵ $R_m - R_f$ 可能產生測量誤差的問題 ∴用Rank作為IV

IV1: Rank要和 $(R_m - R_f)$ 相關

IV2: Rank 不可和 ε 相關

IV3: Rank 不能影響 $R_i - R_f$，只能透過 $R_m - R_f$ 影響

```
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 ***
```

→ significant

R^2

```
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
```

$F = 1858 > 10 \rightarrow$ strong IV.

c. 檢查 $r_m - r_f$ 是否具有內生性

$$r_i - r_f = \alpha + \beta(r_m - r_f) + \delta \hat{v} + \varepsilon$$

if δ significant → Exdogenous

δ not significant → Endogenous

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.003018   0.005984   0.504   0.6146
excess_mkt   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
```

$\hat{v}$ p-value = 0.0428 > 0.01 ∴ we can't reject exdogenous → regard as exdogenous.

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?

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?

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?

d. use RANK as IV. $excess\_mkt_t = \pi_0 + \pi_1 \cdot RANK_t + v_t$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.003018   0.006044   0.499    0.618
excess_mkt  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
```

OLS : $\beta = 1.2018$ (If $r_m - r_f$ has measurement error → β would be lower.)

IV(2SLS) : $\beta = 1.2783$ → larger

when there is measure error → attenuation error (向零偏誤)

∴ β will be underestimated. Therefore, IV method estimates larger β.

e. First Stage: $r_m - r_f = \pi_0 + \pi_1 Rank + \pi_2 POS + v$

$R^2 = 0.9149$

$F = 951.3 > 10$ → strong IV

Yes, the instrument Rank and POS are jointly strong, as F-statistic = 951.3.

The R^2 of 1st stage is 0.9149, indicating that the IVs explain over 91% of variation of $R_m - R_f$

f. Hausman-type test

① use IV regression: $excess\_mkt = \pi_0 + \pi_1 RANK + \pi_2 POS + \hat{v}$

② get the residual $\hat{v}$

③ Add in CAPM Regression: $excess\_msft = \alpha + \beta\, excess\_market + \delta \hat{v} + \varepsilon$

H_0: excess_mkt is exogeneity ($\delta = 0$)

H_a: excess_mkt is endogeneity ($\delta \neq 0$)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.003004   0.005972   0.503   0.6157
excess_mkt   1.283118   0.126344  10.156   <2e-16 ***
vhat2       -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
```

> 0.01

∴ do not reject H_0, there is no significant evidence shows that excess_mkt is endogenous.

⇒ excess_mkt is exdogenous

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?

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.

g.

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

	β
OLS:	1.2018
IV(RANK):	1.2783
IV(RANK, POS):	1.2831 → 比OLS大，合理，∵OLS通常会低估β

h. Sargan Test (Over Identified Test)

$\hat{u} = \gamma_0 + \gamma_1 \cdot RANK + \gamma_2 POS + e$

$S = nR^2$, $S \sim \chi^2_{g-k}$

p-value $= 0.4549 > 0.05$ ⇒ do not reject H_0, IV is exdogenous.

V **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.

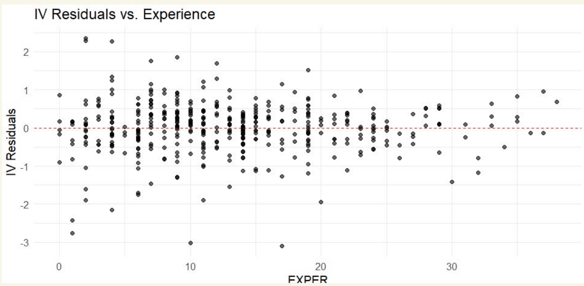
a. Calculate the IV/2SLS residuals, $\hat{e}_{IV}$. Plot them versus *EXPER*. Do the residuals exhibit a pattern consistent with homoskedasticity?

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.

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.

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.

a.



→ heteroskedasticity

b.

$nR^2 = 7.4386$

p-value $= 0.00638 < 0.05$

$\therefore$ reject H_0 → heteroskedasticity

c.

→ 考慮到heteroskedasticity之後的SE

	Coefficient	Baseline_SE	Robust_SE	Ratio
(Intercept)	(Intercept)	0.4003280773	0.4297977164	1.073614
educ	educ	0.0314366956	0.0333385883	1.060499
exper	exper	0.0134324755	0.0155463781	1.157373
expersq	expersq	0.0004016856	0.0004300837	1.070697

95% C.I. for EDUC = educ_coef ± 1.96 educ_robust_se

95% C.I. for educ. = [−0.0039, 0.1267]

d. Bootstrap是不以normal distribution or σ = constant 為assumption的估計方式，重新抽樣 (隨机抽放(放回) 200次，並用這200個sample做OLS，求出SE

Bootstrap SE for educ = 0.03235，介於 Baseline，Robust SE
代表存在heteroskedasticity，
但中等。

95% C.I. = [-0.0020, 0.1248] → 包含0，代表5% sig. level之下
not significant

好處是不用依賴model assumption