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.

```
#a
#create MOTHERCOLL if MOTHEREDUC > 12, zero otherwise.
#FATHERCOLL equals one if FATHEREDUC > 12 and zero otherwise.
married_data$mothercoll <- ifelse(married_data$mothereduc > 12, 1, 0)
married_data$fathercoll <- ifelse(married_data$fathereduc > 12, 1, 0)

mother_college_pct <- mean(married_data$mothercoll) * 100
father_college_pct <- mean(married_data$fathercoll) * 100

cat("Percentage of wife's mothers with some college:", round(mother_college_pct, 2), "%\n")
cat("Percentage of wife's fathers with some college:", round(father_college_pct, 2), "%\n")
```

Percentage of wife's mothers with some college: 12.15 %

Percentage of wife's fathers with some college: 11.68 %

b. correlations between EDUC, MOTHERCOLL, and FATHERCOLL are important. One of the rules of a good IV is the explanatory variable and IV is correlated.

MOTHERCOLL and FATHERCOLL can be better IVs than MOTHEREDUC and FATHEREDUC for two reasons. First, there might exist measurement errors in original Ivs because they are not directly measured but are often constructed or self-reported. Second, binary variables MOTHERCOLL and FATHERCOLL can become more simpler and explainable, for simply distinguish whether he/she has a high degree of education.

```
> cor(married_data$educ, married_data$mothercoll, use = "complete.obs")#0.3594
[1] 0.3594705
> cor(married_data$educ, married_data$fathercoll, use = "complete.obs")#0.3984
[1] 0.3984962
> cor(married_data$mothercoll, married_data$fathercoll, use = "complete.obs")#0.3545
[1] 0.3545709
```

c.

```
> cat("95% CI for educ coefficient: [", round(educ_lb, 5), ",", round(educ_ub, 5), "]\n")
95% CI for educ coefficient: [ -0.00144 , 0.15348 ]
```

d. f-statistic = 63.563, p-value < 0.05

```
                  df1 df2 statistic  p-value
Weak instruments   1 424    63.563 1.46e-14 ***
```

reject H0, MOTHERCOLL is a strong instrument

e.

```
> cat("95% CI for educ coefficient: [", round(educ_lb, 5), ",", round(educ_ub, 5), "]\n")
95% CI for educ coefficient: [ 0.02735 , 0.14835 ]
```

```
> if(vcoc_mroz2[2, 2] > vcoc_mroz[2, 2]){
+   print("wider than part c")
+ }else{
+   print("narrower than part c")
+ }
[1] "narrower than part c"
```

f. F-statistic = 56.963, p-value < 0.05

```
                  df1 df2 statistic p-value
Weak instruments   2 423    56.963  <2e-16 ***
```

Reject H0, these Ivs seems adequately strong with.

g. Chi-square statistic = 0.238, p-value = 0.626 > 0.05
Non-reject H0, the extra instruments are valid

	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

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?

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?

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.

a. $\beta > 1$：風險大（報酬對市場反應強）

```
> cat("βj:", round(ols$coefficients[2], 2))
βj: 1.2
```

b. yes, it's potential.

Rank is correlated with (rm-rf).

Rank is not place in the origin model.

Rank may have no correlation with error.

```
library(AER)
capm5$rank = rank(capm5$mkt - capm5$riskfree)
first_stage = lm(x ~ capm5$rank)
summary(first_stage)
#p-value < 2e-16, sinificant
#Multiple R-squared:  0.9126
#yes, strong IV
```

c. it's statistically significant that market return is exogenous

```
fsresidual = first_stage$residuals
exo_test = lm(y ~ x + fsresidual)
summary(exo_test)
#fsresidual p-value = 0.0428 > 0.01
#non-reject H0
#x is exogenous
```

d. different coefficient with OLS

```
> if(capm5_iv $coefficients[2] == ols$coefficients[2]){
+   print("same coef with OLS")
+ }else{
+   print("different coef with OLS")
+ }
[1] "different coef with OLS"
```

e.

```
capm5$pos = ifelse(capm5$mkt - capm5$riskfree > 0, 1, 0)
first_stage_2IV = lm(x ~ capm5$rank + capm5$pos)
summary(first_stage_2IV)#Multiple R-squared:  0.9149
capm5_2iv = ivreg(y ~ x | capm5$rank + capm5$pos)
summary(capm5_2iv, diagnostics = TRUE)
#Weak instruments p-value <2e-16 strong IV
#Sargan p-value = 0.4549 instruments are valid
```

f.

```
re = first_stage_2IV$residuals
exo_test = lm(y ~ x + re)
summary(exo_test)
# re p-value 0.0287 < 0.05,endogenous
```

g. different coefficient with OLS

```
> if(capm5_2iv $coefficients[2] == ols$coefficients[2]){
+   print("same coef with OLS")
+ }else{
+   print("different coef with OLS")
+ }
[1] "different coef with OLS"
```

h. valid

```
re = resid(capm5_2iv)
sargan = lm(re ~ capm5$rank + capm5$pos)
R2 = summary(sargan)$r.squared
n = nrow(capm5)
S = n * R2
p_value = 1 - pchisq(S, df = 1)
#0.45488 non-reject H0
#valid IV
```

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.

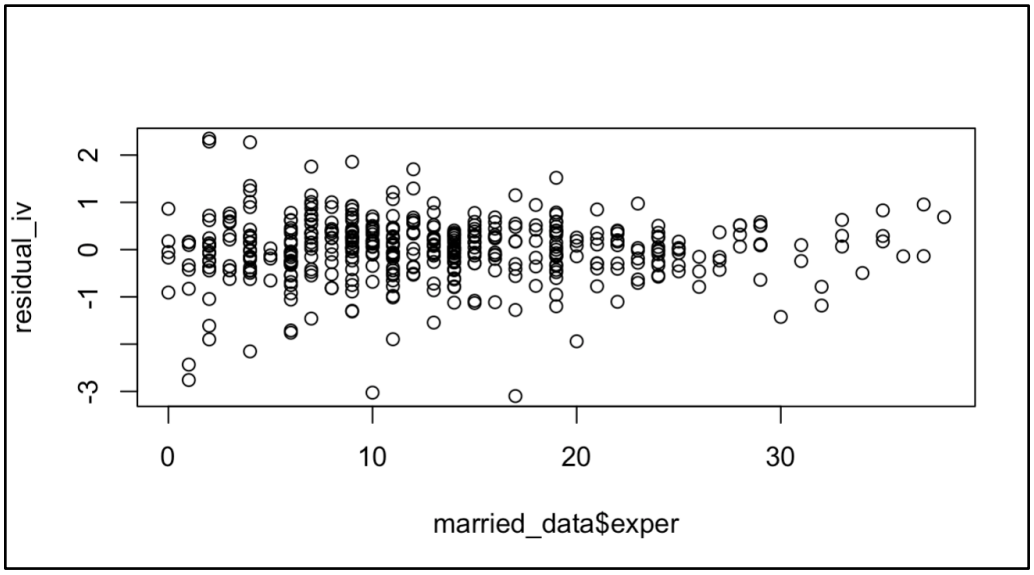
a. Calculate the IV/2SLS residuals, $\hat{e}_{IV}$. Plot them versus *EXPER*. Do the residuals exhibit a pattern consistent with homoskedasticity?

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. it seems to have heteroskedasticity



b. Reject H0, heteroskedasticity

```
re2 = residual_iv^2
model = lm(re2~exper, data = married_data)
R2 = summary(model)$r.squared
N = nrow(married_data)
NR2 = N*R2 #7.4385
p_value = 1 - pchisq(NR2, df = 1) #0.00638
#reject H0, heteroskedasticity
```

c.

```
+   print("robust SE is smaller than baseline")
+ }
[1] "robust SE is smaller than baseline"
> cat("95% CI for educ (robust SE): [", round(lb, 5), ",", round(ub, 5), "]\n")
95% CI for educ (robust SE): [ -0.00039 , 0.12319 ]
```

d. Interval estimate [0.0037, 0.1286]

```
[1] "wider than baseline"
> if(boot_se > (vcov_mroz_robust[2, 2]^0.5)){
+   print("wider than robust")
+ }else{
+   print("narrower than robust")
+ }
[1] "wider than robust"
> ci_boot
       2.5%       97.5%
0.003650778 0.128587054
```