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.
The percentage of mother with college education : 12.15%

The percentage of father with college education : 11.69%

```
> mean(MOTHERCOLL)
[1] 0.1214953
> mean(FATHERCOLL)
[1] 0.1168224
```

b.
① It's important! Since we want the IVs are strongly correlated to educ and have less correlation between those IVs.

② Although both mothereduc and fathereduc have higher correlation with educ, the correlation between mothereduc and fathereduc are much higher than the correlation between Mothercoll and Fathercoll.

Thus, we prefer using MOTHercoll and Fathercoll.

```
> corMatrix1
                 educ MOTHERCOLL FATHERCOLL
educ        1.0000000  0.3594705  0.3984962
MOTHERCOLL  0.3594705  1.0000000  0.3545709
FATHERCOLL  0.3984962  0.3545709  1.0000000
> corMatrix2
                 educ motheredu c fathereduc
educ        1.0000000  0.3870198  0.4154030
mothereduc  0.3870198  1.0000000  0.5540632
fathereduc  0.4154030  0.5540632  1.0000000
```

c.

```
> summary(modC)

Call:
ivreg(formula = log(wage) ~ exper + I(exper^2) + educ | exper +
    I(exper^2) + MOTHERCOLL, data = theData)

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
exper        0.0433444  0.0134135   3.231  0.00133 **
I(exper^2)  -0.0008711  0.0004017  -2.169  0.03066 *
educ         0.0760180  0.0394077   1.929  0.05440 .
---
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

> confint(modC, level=0.95)
                   2.5 %        97.5 %
(Intercept) -1.105942034  8.404298e-01
exper        0.017054428  6.963439e-02
I(exper^2)  -0.001658392 -8.385898e-05
educ        -0.001219763  1.532557e-01
```

The interval :

[-0.00122, 0.153]

d.
Since the value of F test = 63.21602 > 10, Mother Coll is a strong IV.

```
> F_test
    value     numdf     dendf
 63.21602   1.00000 426.00000
```

e. Narrower than part (c)

The interval is [0.0275, 0.1481]

```
> summary(modE)

Call:
ivreg(formula = log(wage) ~ exper + I(exper^2) + educ | exper +
    I(exper^2) + MOTHERCOLL + FATHERCOLL, data = theData)

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
exper        0.0426761  0.0132950   3.210  0.00143 **
I(exper^2)  -0.0008486  0.0003976  -2.135  0.03337 *
educ         0.0878477  0.0307808   2.854  0.00453 **
---
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

> confint(modE, level=0.95)
                  2.5 %        97.5 %
(Intercept) -1.04782153  4.896578e-01
exper        0.01661839  6.873386e-02
I(exper^2)  -0.00162779 -6.940599e-05
educ         0.02751845  1.481769e-01
```

f.

```
> summary(modIV2)

Call:
lm(formula = educ ~ MOTHERCOLL + FATHERCOLL, data = theData)

Residuals:
    Min      1Q  Median      3Q     Max
-7.1897 -0.1897 -0.1897  0.8103  4.8103

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.1897     0.1076 113.310  < 2e-16 ***
MOTHERCOLL    1.7436     0.3215   5.423 9.84e-08 ***
FATHERCOLL    2.2031     0.3270   6.737 5.27e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.032 on 425 degrees of freedom
Multiple R-squared:  0.2132,    Adjusted R-squared:  0.2095
F-statistic:  57.6 on 2 and 425 DF,  p-value: < 2.2e-16

> anova(modIV2)
Analysis of Variance Table

Response: educ
            Df  Sum Sq Mean Sq F value    Pr(>F)
MOTHERCOLL   1  288.18 288.184  69.803 9.387e-16 ***
FATHERCOLL   1  187.39 187.394  45.390 5.266e-11 ***
Residuals  425 1754.62   4.129
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The F value of MOTHERCOLL = 69.803

The F value of FATHERCOLL = 45.39

Both of them greater than 10, that is, their are valid.

g.

```
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
```

According to Sargan test, its p-value = 0.626 < 0.238 = test_statistic, we fail to reject null hypothesis, IV is valid.

```r
theData<-mroz[which(mroz$LFP==1), ]
mothereduc<-c(theData$MOTHEREDUC)
fathereduc<-c(theData$FATHEREDUC)
educ<-c(theData$EDUC)
wage<-c(theData$WAGE)
exper<-c(theData$EXPER)

#a
MOTHERCOLL<-c()
FATHERCOLL<-c()

for( i in c(1:428)){}
mean(MOTHERCOLL)
mean(FATHERCOLL)
theData$MOTHERCOLL<-MOTHERCOLL
theData$FATHERCOLL<-FATHERCOLL

#b
dataB1<-data.frame(educ, MOTHERCOLL, FATHERCOLL)
corMatrix1<-cor(dataB1)
dataB2<-data.frame(educ, mothereduc, fathereduc)
corMatrix2<-cor(dataB2)

#c
modIV<-lm(educ~MOTHERCOLL, data=theData)
modC<-ivreg(log(wage)~exper+I(exper^2)+educ | exper+I(exper^2)+MOTHERCOLL, data=theData)
summary(modC)
confint(modC, level=0.95)

#d
F_test<-summary(modIV)$fstatistic
F_test
```

```r
#e
modIV2<-lm(educ~MOTHERCOLL+FATHERCOLL, data=theData)
modE<-ivreg(log(wage)~exper+I(exper^2)+educ | exper+I(exper^2)+MOTHERCOLL+FATHERCOLL, data=theData)
summary(modE)
confint(modE, level=0.95)

#f
summary(modIV2)
anova(modIV2)

#g
summary(modE, 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?

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. Since $\beta \approx 1.2 > 1$, it is risky.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.003250   0.006036   0.538    0.591
x           1.201840   0.122152   9.839   <2e-16
```

b. since the regression model is $r_j - r_f = \alpha + \beta(r_m - r_f)$, RANK is not on the right hand side

$\Rightarrow$ IV1 holds

Since the correlation between RANK and the residual is 7.467×10^{-17}, which is very small.

$\Rightarrow$ IV2 holds

Since the correlation between RANK and $r_m - r_f$ is 0.9552, which is very high.

$\Rightarrow$ IV3 holds

And the $R^2 = 0.91255$

Thus, RANK is a strong IV.

```
> cor(RANK, v_head, use="complete.obs")
[1] 7.467451e-17
> cor(RANK, theDataB$x, use="complete.obs")
[1] 0.9552779
> summary(modB)$r.squared
[1] 0.9125559
```

c. Since the p-value is 0.0428, which is greater than 0.01,

we fail to reject Null Hypothesis

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.003018   0.005984   0.504   0.6146
x            1.278318   0.126749  10.085   <2e-16
v_head      -0.874599   0.428626  -2.040   0.0428
```

d. Their estimations are similar.

```
> summary(modD)

Call:
lm(formula = y ~ x_head, data = theDataB)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.27247 -0.04073 -0.00825  0.03585  0.34577 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.003018   0.005984   0.504    0.615    
x_head      1.278318   0.126739  10.086   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08011 on 178 degrees of freedom
Multiple R-squared:  0.3637,	Adjusted R-squared:  0.3601 
F-statistic: 101.7 on 1 and 178 DF,  p-value: < 2.2e-16
```

```
> summary(modA)

Call:
lm(formula = y ~ x, data = theDataA)

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    
x           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
```

e. Estimation:

R^2: 0.915

```
> summary(modE)

Call:
lm(formula = x ~ RANK + POS, data = theDataB)

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
```

Since F value > 10 and both coefficient are significant
(p-value of RANK < 0.001
p-value of POS < 0.05)

$\Rightarrow$ jointly strong instruments.

f.

Since p-value is 0.0287 > 0.01, we fail to reject Null hypothesis.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.003004   0.005972   0.503   0.6157
x            1.283118   0.126344  10.156   <2e-16
v_head2     -0.954918   0.433062  -2.205   0.0287
```

g. Two estimation are similar:

```
> summary(modG)

Call:
lm(formula = y ~ x_head2, data = theDataB)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.27220 -0.04109 -0.00810  0.03396  0.34635 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.003004   0.005966   0.503    0.615    
x_head2     1.283118   0.126212  10.166   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07988 on 178 degrees of freedom
Multiple R-squared:  0.3673,	Adjusted R-squared:  0.3638 
F-statistic: 103.4 on 1 and 178 DF,  p-value: < 2.2e-16
```

```
> summary(modA)

Call:
lm(formula = y ~ x, data = theDataA)

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    
x           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
```

h. Since the p-value is 0.448 < 0.05, we failto reject null hypothesis.

$\Rightarrow$ IV are valid.

```
> p_value
[1] 0.4489907
```

```r
msft<-c(capm5$msft)
mkt<-c(capm5$mkt)
rf<-c(capm5$riskfree)

#a
y<-msft-rf
x<-mkt-rf
theDataA<-data.frame(x, y)
modA<-lm(y~x, data=theDataA)
summary(modA)

#b
theDataB<-theDataA[order(theDataA$x), ]
RANK<-c(1:180)
theDataB$RANK<-RANK
modB<-lm(x~RANK, data=theDataB)
a<-summary(modB)$coef[, 1]
v_head<-summary(modB)$resid
cor(RANK, v_head, use="complete.obs")
cor(RANK, theDataB$x, use="complete.obs")
summary(modB)$r.squared

#c
theDataB$v_head<-v_head
modC<-lm(y~x+v_head, data=theDataB)
summary(modC)

#d
x_head<-a[1]+a[2]*RANK
theDataB$x_head<-x_head
modD<-lm(y~x_head, data=theDataB)
summary(modD)
```

```r
#e
POS<-c()
for(i in c(1:180)){}
theDataB$POS<-POS
modE<-lm(x~RANK+POS, data=theDataB)
summary(modE)
a2<-summary(modE)$coef[, 1]
#test
eqE<-ivreg(y~x | RANK+POS, data=theDataB)
summary(eqE, diagnostics=TRUE)

#f
v_head2<-summary(modE)$resid
theDataB$v_head2<-v_head2
modF<-lm(y~x+v_head2, data=theDataB)
summary(modF)

#g
x_head2<-a2[1]+a2[2]*RANK+a2[3]*POS
theDataB$x_head2<-x_head2
modG<-lm(y~x_head2, data=theDataB)
summary(modG)

#h
e<-summary(modG)$resid
theDataB$e<-e
modH<-lm(e~RANK+POS, data=theDataB)
summary(modH)
test_t<-length(e)*summary(modH)$r.squared
p_value<-1-pchisq(test_t, 1)
p_value
```

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.

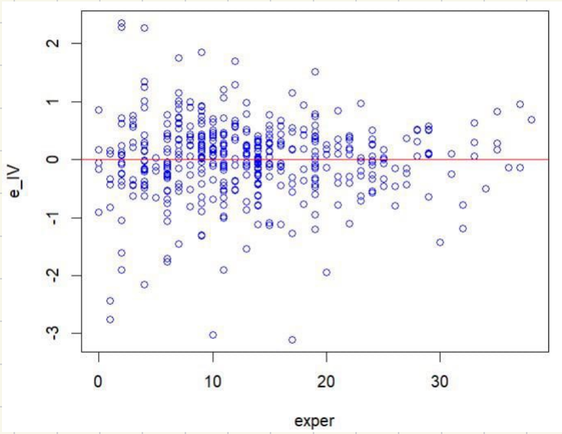
a. Calculate the IV/2SLS residuals, $\hat{e}_{IV}$. Plot them versus *EXPER*. Do the residuals exhibit a pattern consistent with homoskedasticity? NO

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. The residuals become smaller ⇒ Does not exhibit a pattern consistent with homoskedasticity.



b. Since the p-value is 0.006 < 0.05, we reject null hypothesis ⇒ heteroskedasticity.

```
> cat(test, p)
7.438552 0.006384122
```

c. larger

Robust:

```
> vcv

t test of coefficients:

               Estimate  Std. Error t value Pr(>|t|)
(Intercept)  0.04810030  0.42979772  0.1119 0.910945
exper        0.04417039  0.01554638  2.8412 0.004711 **
I(exper^2)  -0.00089897  0.00043008 -2.0902 0.037193 *
educ         0.06139663  0.03333859  1.8416 0.066231 .
```

part (a):

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0481003  0.4003281   0.120  0.90442
exper        0.0441704  0.0134325   3.288  0.00109
I(exper^2)  -0.0008990  0.0004017  -2.238  0.02574
educ         0.0613966  0.0314367   1.953  0.05147
```

The interval:

```
> lower
[1] -0.004132416
> upper
[1] 0.1269257
```

```
theData<-mroz[which(mroz$LFP==1), ]
mothereduc<-c(theData$MOTHEREDUC)
fathereduc<-c(theData$FATHEREDUC)
educ<-c(theData$EDUC)
wage<-c(theData$WAGE)
exper<-c(theData$EXPER)

#a
#first
modIV<-ivreg(log(wage)~exper+I(exper^2)+educ | exper+I(exper^2)+mothereduc+fathereduc, data=theData)
e_IV<-summary(modIV)$resid
#plot
plot(exper, e_IV, col="blue")
abline(h=0, col="red")

#b
r_2<-e_IV^2
dataB<-data.frame(r_2, exper)
modB<-lm(r_2~exper, dataB)
R<-summary(modB)$r.squared
test<-R*428
p<-1-pchisq(test, 1)
cat(test, p)
```

```
#c
#cov<-hccm(modIV, type="hc1")
vcv<-coeftest(modIV, vcov=vcovHC(modIV, type="HC1"))
b_robust<-vcv["educ", 1]
se1_robust<-vcv["educ", 2]
t<-qt(0.975, 425)
upper<-b_robust+t*se1_robust
lower<-b_robust-t*se1_robust
lower
upper

#d
boot_fn<-function(data, indices){
  d<-data[indices, ]
  model<-ivreg(log(wage)~educ+exper+I(exper^2) | mothereduc+fathereduc+exper+I(exper^2), data=d)
  return(coef(model)["educ"])
}
set.seed(123)
boot_results<-boot(data=theData, statistic=boot_fn, R=200)
boot_se<-sd(boot_results$t)
educ_est<-mean(boot_results$t)
educ_ci_boot<-c(educ_est-t*boot_se, educ_est+t*boot_se)
cat(boot_se)
cat(educ_ci_boot)
```