HW0428

Yung-Jung Cheng

2025-05-04

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.

18(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?

Ans

```
# (a) Create dummy variables for parental college education
mroz$mothercoll <- ifelse(mroz$mothereduc > 12, 1, 0)
mroz$fathercoll <- ifelse(mroz$fathereduc > 12, 1, 0)

# Compute proportions
mean(mroz$mothercoll)  # proportion of mothers with college education
```

```
## [1] 0.1009296
```

```
mean(mroz$fathercoll)  # proportion of fathers with college education
```

```
## [1] 0.1075697
```

```
mean(mroz$mothercoll == 1 | mroz$fathercoll == 1)  # at least one parent with college
```

```
## [1] 0.1633466
```

- About **10.09%** of mothers have more than 12 years of education (i.e., some college education).
- About **10.76%** of fathers have more than 12 years of education.
- Approximately **16.33%** of the women in the sample have **at least one parent** with some college education.

18(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`?

Ans

```
# (b) Correlations among education and instruments
cor(mroz[, c("educ", "mothercoll", "fathercoll")])
```

```
##                 educ mothercoll fathercoll
## educ       1.0000000  0.3370171  0.3193212
## mothercoll 0.3370171  1.0000000  0.3674532
## fathercoll 0.3193212  0.3674532  1.0000000
```

- The correlation between `educ` and `mothercoll` is **0.337**, and between `educ` and `fathercoll` is **0.319**. These are moderate positive correlations, indicating that parental college education is positively associated with the woman's own education.
- The correlation between `mothercoll` and `fathercoll` is **0.367**, suggesting that parents' education levels are somewhat positively related, but not collinear.
- The magnitudes are meaningful, though not extremely strong. These results suggest that both `mothercoll` and `fathercoll` are **relevant instruments** because they are correlated with the endogenous regressor `educ`.
- Using binary indicators for college education may help mitigate measurement error or nonlinearity issues compared to using raw years of parental education, making them **potentially better instruments** than `mothereduc` and `fathereduc`.

18(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`?

Ans

```
# Estimate IV model: wage on educ, instrumented by mothercoll
iv_c <- ivreg(wage ~ educ | mothercoll, data = mroz)

# Summary of IV model
summary(iv_c)
```

```
##
## Call:
## ivreg(formula = wage ~ educ | mothercoll, data = mroz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6456 -2.2363 -0.3758  1.3990 22.7637
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.5457     1.7969  -1.973  0.04883 *
## educ          0.4818     0.1460   3.301  0.00101 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.076 on 751 degrees of freedom
## Multiple R-Squared: 0.1009,  Adjusted R-squared: 0.09975
## Wald test:  10.9 on 1 and 751 DF,  p-value: 0.001008
```

```
# 95% confidence interval for the coefficient on educ
confint(iv_c, level = 0.95)
```

```
##                  2.5 %      97.5 %
## (Intercept) -7.0675458 -0.02393035
## educ         0.1957666  0.76791447
```

- Using **MOTHERCOLL** as an instrument for `educ`, the IV regression yields a coefficient estimate of **0.4818** for `educ`.
- The 95% confidence interval is **[0.196, 0.768]**, which does not include zero, indicating statistical significance at the 1% level (p = 0.001).
- This suggests that additional years of education, when instrumented by a mother's college attendance, are associated with a substantial and significant increase in wages.
- The result helps mitigate concerns of endogeneity in OLS estimates of the return to education.

18(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?

Ans

```
# (d) First-stage regression: educ on mothercoll
fs_d <- lm(educ ~ mothercoll, data = mroz)

# Summary of first-stage regression
summary(fs_d)
```

```
##
## Call:
## lm(formula = educ ~ mothercoll, data = mroz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0295 -0.5789 -0.0295  0.9705  4.9705
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.02954    0.08256  145.70   <2e-16 ***
## mothercoll   2.54941    0.25989    9.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.148 on 751 degrees of freedom
## Multiple R-squared:  0.1136, Adjusted R-squared:  0.1124
## F-statistic: 96.23 on 1 and 751 DF,  p-value: < 2.2e-16
```

```
# Extract F-statistic
fs_d_summary <- summary(fs_d)
fs_d_summary$fstatistic
```

```
##     value     numdf     dendf
##  96.22868   1.00000 751.00000
```

- In the first-stage regression of `educ` on `mothercoll`, the coefficient on `mothercoll` is **2.549**, with a t-statistic of **9.81** and a p-value < 0.001, indicating strong statistical significance.
- The **F-statistic** for testing the null hypothesis that `mothercoll` has no effect on `educ` is **96.23**, which far exceeds the common rule-of-thumb threshold of 10 for weak instruments.
- Therefore, `mothercoll` appears to be a **strong instrument** for `educ`.

18(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)?

Ans

```
# (e) IV estimation using both mothercoll and fathercoll as instruments
iv_e <- ivreg(wage ~ educ | mothercoll + fathercoll, data = mroz)

# Summary of IV model
summary(iv_e)
```

```
##
## Call:
## ivreg(formula = wage ~ educ | mothercoll + fathercoll, data = mroz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5410 -2.2427 -0.4041  1.3754 22.7573
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.2732     1.5255  -2.146 0.032218 *
## educ          0.4597     0.1238   3.712 0.000221 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.075 on 751 degrees of freedom
## Multiple R-Squared: 0.1013,  Adjusted R-squared: 0.1001
## Wald test: 13.78 on 1 and 751 DF,  p-value: 0.0002206
```

```
# 95% confidence interval for the coefficient on educ
confint(iv_e, level = 0.95)
```

```
##                  2.5 %     97.5 %
## (Intercept) -6.2630491 -0.2833066
## educ         0.2169759  0.7023390
```

- Using both `mothercoll` and `fathercoll` as instruments for `educ`, the estimated coefficient on `educ` is **0.4597**, with a standard error of **0.1238**.
- The 95% confidence interval is **[0.217, 0.702]**, which is narrower than the one obtained using only `mothercoll` as the instrument.
- The coefficient is statistically significant at the 1% level (p = 0.0002), confirming a strong and positive relationship between education and wages.
- Adding a second instrument improves precision, as reflected in the narrower confidence interval.

18(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?

Ans

```
# (f) First-stage regression with both instruments
fs_f <- lm(educ ~ mothercoll + fathercoll, data = mroz)

# Summary of the first-stage regression
summary(fs_f)
```

```
##
## Call:
## lm(formula = educ ~ mothercoll + fathercoll, data = mroz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9142 -0.9142  0.0858  0.1646  5.0858
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.91415    0.08261 144.223  < 2e-16 ***
## mothercoll   1.92121    0.27257   7.049 4.10e-12 ***
## fathercoll   1.66213    0.26500   6.272 6.01e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.095 on 750 degrees of freedom
## Multiple R-squared:  0.1578, Adjusted R-squared:  0.1555
## F-statistic: 70.24 on 2 and 750 DF,  p-value: < 2.2e-16
```

```
# Extract F-statistic for joint significance test
anova(fs_f)
```

```
## Analysis of Variance Table
##
## Response: educ
##             Df Sum Sq Mean Sq F value    Pr(>F)
## mothercoll   1  444.1  444.10  101.14 < 2.2e-16 ***
## fathercoll   1  172.7  172.74   39.34 6.011e-10 ***
## Residuals  750 3293.2    4.39
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- In the first-stage regression of `educ` on both `mothercoll` and `fathercoll`, the coefficients are:
  - `mothercoll` : **1.921**, p < 0.001
  - `fathercoll` : **1.662**, p < 0.001
- The overall model has an F-statistic of **70.24** (df = 2, 750), with a p-value < 2.2e-16.
- The individual F-statistics from the ANOVA table are:
  - `mothercoll` : F = **101.14**
  - `fathercoll` : F = **39.34**
- These results confirm that the instruments are jointly significant and strongly correlated with the endogenous regressor `educ`.
- Therefore, `mothercoll` and `fathercoll` are **adequately strong instruments**.

18(g)

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

Ans

```
# Run J-test via 'summary' with diagnostics = TRUE
summary(iv_e, diagnostics = TRUE)
```

```
##
## Call:
## ivreg(formula = wage ~ educ | mothercoll + fathercoll, data = mroz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5410 -2.2427 -0.4041  1.3754 22.7573
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.2732     1.5255  -2.146 0.032218 *
## educ          0.4597     0.1238   3.712 0.000221 ***
##
## Diagnostic tests:
##                  df1 df2 statistic p-value
## Weak instruments   2 750    70.241  <2e-16 ***
## Wu-Hausman         1 750     0.004   0.951
## Sargan             1  NA     0.083   0.774
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.075 on 751 degrees of freedom
## Multiple R-Squared: 0.1013,  Adjusted R-squared: 0.1001
## Wald test: 13.78 on 1 and 751 DF,  p-value: 0.0002206
```

- The **Sargan test** of overidentifying restrictions tests the null hypothesis that the extra instrument (`fathercoll`) is valid—that is, uncorrelated with the error term and correctly excluded from the structural equation.
- The test statistic is **0.083** with a p-value of **0.774**, which is not statistically significant.
- Therefore, we **fail to reject the null hypothesis**, providing no evidence against the validity of `fathercoll` as an additional instrument.
- This supports the conclusion that both instruments (`mothercoll` and `fathercoll`) are valid in the IV estimation.

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.

20(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?

Ans

```
# Compute excess returns for Microsoft and the market
capm5 <- capm5 %>%
  mutate(
    capm20a_excess_msft = msft - riskfree,
    capm20a_excess_mkt = mkt - riskfree
  )

# Estimate CAPM model: capm20a_excess_msft ~ capm20a_excess_mkt
capm20a_ols_model <- lm(capm20a_excess_msft ~ capm20a_excess_mkt, data = capm5)

# Display regression summary
summary(capm20a_ols_model)
```

```
##
## Call:
## lm(formula = capm20a_excess_msft ~ capm20a_excess_mkt, 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
## capm20a_excess_mkt 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
```

We estimate the CAPM model for Microsoft using OLS:

$$(r_j - r_f) = \alpha + \beta(r_m - r_f) + u$$

From the regression output:

- Estimated alpha (intercept): **0.00325**, which is not statistically significant (p = 0.591).
- Estimated beta: **1.20184**, statistically significant at the 1% level (p < 0.001).

The beta coefficient greater than 1 implies that Microsoft's excess return is **more volatile than the market's**. Therefore, Microsoft stock is considered **risky** relative to the market during this period.

The model's R^2 is **0.3523**, meaning about 35% of the variation in Microsoft's excess return is explained by the market's excess return.

20(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?

Ans

```
# Create RANK as the rank of (r_m - r_f)
capm5 <- capm5 %>%
  mutate(
    capm20b_excess_mkt = mkt - riskfree,
    capm20b_rank = rank(capm20b_excess_mkt)  # ascending rank
  )

# First-stage regression: excess_mkt ~ RANK
capm20b_first_stage <- lm(capm20b_excess_mkt ~ capm20b_rank, data = capm5)

# Display summary of the first-stage regression
summary(capm20b_first_stage)
```

```
##
## Call:
## lm(formula = capm20b_excess_mkt ~ capm20b_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 ***
## capm20b_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
```

We construct the instrument `RANK` by ranking the market excess returns $(r_m - r_f)$ from smallest to largest and assigning integers from 1 to 180. This rank variable is used as an instrument for the market excess return.

The first-stage regression is:

$$(r_m - r_f) = \pi_0 + \pi_1 \cdot \mathrm{RANK} + v$$

From the regression output:

- The estimated coefficient on `RANK` is **0.0009067**, which is **highly significant** (t = 43.1, p < 0.001).
- The **R-squared is 0.9126**, indicating that `RANK` explains over 91% of the variation in the market excess return.

This confirms that `RANK` is a **strong instrument**, satisfying the relevance condition (IV1). Whether it satisfies the exogeneity condition (IV2 and IV3) cannot be tested directly, but the construction based on ranking, which is independent of future shocks, makes it a plausible candidate.

20(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?

Ans

```
# Compute first-stage residuals
capm5 <- capm5 %>%
  mutate(capm20c_vhat = resid(capm20b_first_stage))

# Compute excess return for Microsoft (if not already done)
capm5 <- capm5 %>%
  mutate(capm20c_excess_msft = msft - riskfree)

# Estimate augmented regression: excess_msft ~ excess_mkt + vhat
capm20c_augmented <- lm(capm20c_excess_msft ~ capm20b_excess_mkt + capm20c_vhat, data = capm
5)

# Display summary
summary(capm20c_augmented)
```

```
##
## Call:
## lm(formula = capm20c_excess_msft ~ capm20b_excess_mkt + capm20c_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
## capm20b_excess_mkt  1.278318   0.126749  10.085   <2e-16 ***
## capm20c_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
```

To test whether the market excess return $(r_m - r_f)$ is exogenous, we include the residuals $\hat{v}$ from the first-stage regression in the structural equation:

$$(r_{\text{msft}} - r_f) = \alpha + \beta(r_m - r_f) + \delta\hat{v} + e$$

From the regression output:

- The coefficient on $\hat{v}$ is **-0.8746**, with a **p-value of 0.0428**.
- This result is statistically significant at the **5% level**, but **not** at the **1% level**.

Conclusion:
There is **moderate evidence against exogeneity** of the market excess return. At the 1% significance level, we **fail to reject** the null hypothesis that the market return is exogenous. However, the 5% significance suggests some degree of endogeneity, which justifies considering IV methods like 2SLS in the following steps.

20(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?

Ans

```
# Estimate 2SLS model using ivreg(): instrumenting capm20a_excess_mkt with capm20b_rank
capm20d_iv_model <- ivreg(capm20a_excess_msft ~ capm20a_excess_mkt | capm20b_rank, data = cap
m5)

# Display regression summary
summary(capm20d_iv_model, diagnostics = TRUE)
```

```
##
## Call:
## ivreg(formula = capm20a_excess_msft ~ capm20a_excess_mkt | capm20b_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
## capm20a_excess_mkt 1.278318   0.128011   9.986   <2e-16 ***
##
## Diagnostic tests:
##                  df1 df2 statistic p-value
## Weak instruments   1 178  1857.587  <2e-16 ***
## Wu-Hausman         1 177     4.164  0.0428 *
## Sargan             0  NA        NA      NA
## ---
## 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
```

We estimate the CAPM model for Microsoft using 2SLS, where the market excess return $(r_m - r_f)$ is instrumented using `RANK` . The structural model is:

$$(r_{\text{msft}} - r_f) = \alpha + \beta(r_m - r_f) + u$$

From the 2SLS regression output:

- The estimated beta is **1.2783**, with a **p-value < 0.001**, indicating it is highly statistically significant.
- The intercept (alpha) is **0.0030**, not statistically significant (p = 0.618).
- The **beta estimate is slightly larger than the OLS estimate (1.2018)** from part (a), which is consistent with the direction of bias expected under classical measurement error — OLS tends to attenuate the slope estimate toward zero.

Diagnostics: - **Weak instrument test (F = 1857.6, p < 0.001)** confirms that `RANK` is a very strong instrument. - **Wu-Hausman test** yields a p-value of **0.0428**, suggesting some evidence that the market excess return may be endogenous, reinforcing the use of IV.

Conclusion:
The IV estimate is larger than the OLS estimate and statistically significant. The diagnostics support the use of 2SLS with `RANK` as a valid and strong instrument.

20(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?

Ans

```
# Create POS: 1 if market excess return > 0, else 0
capm5 <- capm5 %>%
  mutate(capm20e_pos = as.numeric(capm20b_excess_mkt > 0))

# First-stage regression with two instruments: RANK and POS
capm20e_first_stage <- lm(capm20b_excess_mkt ~ capm20b_rank + capm20e_pos, data = capm5)

# Display regression results
summary(capm20e_first_stage)
```

```
##
## Call:
## lm(formula = capm20b_excess_mkt ~ capm20b_rank + capm20e_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 ***
## capm20b_rank  0.0009819  0.0000400   24.55   <2e-16 ***
## capm20e_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
```

We introduce a second instrument `POS`, defined as 1 if the market excess return $(r_m - r_f) > 0$, and 0 otherwise. We then estimate the first-stage regression:

$$(r_m - r_f) = \pi_0 + \pi_1 \cdot \text{RANK} + \pi_2 \cdot \text{POS} + v$$

From the regression output:

- The coefficient on `RANK` is **0.00098** with a very high t-value (24.55), confirming it remains a strong instrument.
- The coefficient on `POS` is **-0.00928**, statistically significant at the 5% level (p = 0.0291).
- The **R-squared is 0.9149**, slightly higher than the model using `RANK` alone.

Conclusion:
The instruments `RANK` and `POS` are **jointly significant** in the first-stage regression, indicating that they are collectively strong instruments. The improvement in R^2 suggests that `POS` adds some explanatory power beyond `RANK`, though its contribution is modest.

20(f)

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

Ans

```
# Compute first-stage residuals from (e)
capm5 <- capm5 %>%
  mutate(capm20f_vhat = resid(capm20e_first_stage))

# Reuse Microsoft excess return if needed
capm5 <- capm5 %>%
  mutate(capm20f_excess_msft = msft - riskfree)

# Estimate augmented model: excess_msft ~ excess_mkt + vhat
capm20f_augmented <- lm(capm20f_excess_msft ~ capm20b_excess_mkt + capm20f_vhat, data = capm
5)

# Show regression summary
summary(capm20f_augmented)
```

```
##
## Call:
## lm(formula = capm20f_excess_msft ~ capm20b_excess_mkt + capm20f_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
## capm20b_excess_mkt  1.283118   0.126344  10.156   <2e-16 ***
## capm20f_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
```

To test for the endogeneity of the market excess return $(r_m - r_f)$, we include the residuals $\hat{v}$ from the first-stage regression using both `RANK` and `POS` as instruments. The augmented model is:

$$(r_{\text{msft}} - r_f) = \alpha + \beta(r_m - r_f) + \delta\hat{v} + e$$

From the regression output:

- The coefficient on $\hat{v}$ is **-0.9549**, with a p-value of **0.0287**, which is statistically significant at the **5% level**, but **not** at the 1% level.
- The beta estimate remains significant and slightly increases to **1.2831**.

Conclusion:
There is **moderate evidence of endogeneity** in the market excess return. At the 5% level, we reject the null hypothesis of exogeneity. This result supports the use of IV methods like 2SLS to obtain consistent estimates. However, since the result is not significant at the 1% level, the evidence is not overwhelmingly strong.

20(g)

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

Ans

```
# Estimate 2SLS model using ivreg() with two instruments
capm20g_iv_model <- ivreg(capm20a_excess_msft ~ capm20a_excess_mkt | capm20b_rank + capm20e_p
os, data = capm5)

# Display regression summary
summary(capm20g_iv_model, diagnostics = TRUE)
```

```
##
## Call:
## ivreg(formula = capm20a_excess_msft ~ capm20a_excess_mkt | capm20b_rank +
##     capm20e_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
## capm20a_excess_mkt 1.283118   0.127866  10.035   <2e-16 ***
##
## Diagnostic tests:
##                  df1 df2 statistic p-value
## Weak instruments   2 177   951.262  <2e-16 ***
## Wu-Hausman         1 177     4.862  0.0287 *
## Sargan             1  NA     0.558  0.4549
## ---
## 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
```

We re-estimate the CAPM model for Microsoft using **two instrumental variables**: `RANK` and `POS`, via 2SLS:

$$(r_{\text{msft}} - r_f) = \alpha + \beta(r_m - r_f) + u$$

From the 2SLS regression output:

- The estimated beta is **1.2831**, statistically significant at the 1% level ($p < 0.001$).
- The intercept is **0.0030**, not significant ($p = 0.62$).
- The beta estimate is **slightly larger than the OLS estimate (1.2018)** from 20(a), and almost the same as the single-IV result in 20(d).

Diagnostic tests: - **Weak instrument test (F = 951.3, p < 0.001)** indicates that the instruments are jointly strong. - **Wu-Hausman test (p = 0.0287)** suggests rejection of the exogeneity of the market return at the 5% level, consistent with earlier findings. - **Sargan test (p = 0.4549)** indicates that the overidentifying restriction is not rejected, supporting the validity of the instruments.

Conclusion:
Using both `RANK` and `POS` as instruments produces a similar but slightly more precise IV estimate compared to the single-instrument case. Diagnostics confirm that the instruments are strong and valid, and there is mild evidence of endogeneity in the market return.

20(h)

Obtain the IV/2SLS residuals from part (g) and use them (not an automatic command) to carry out an exogeneity test. Does this test support using IV or the usual OLS?

Ans

```
# Extract residuals from the 2SLS model
capm5 <- capm5 %>%
  mutate(capm20h_iv_resid = resid(capm20g_iv_model))

# Estimate augmented model: excess_msft ~ excess_mkt + iv_resid
capm20h_augmented <- lm(capm20a_excess_msft ~ capm20a_excess_mkt + capm20h_iv_resid, data = c
apm5)

# Show regression summary
summary(capm20h_augmented)
```

```
## Warning in summary.lm(capm20h_augmented): 完全擬合進行：summary 可能不會可靠
```

```
##
## Call:
## lm(formula = capm20a_excess_msft ~ capm20a_excess_mkt + capm20h_iv_resid,
##     data = capm5)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.562e-16 -1.164e-17 -6.850e-18  1.000e-18  1.189e-15
##
## Coefficients:
##                      Estimate Std. Error   t value Pr(>|t|)
## (Intercept)        3.004e-03  6.805e-18 4.414e+14   <2e-16 ***
## capm20a_excess_mkt 1.283e+00  1.379e-16 9.306e+15   <2e-16 ***
## capm20h_iv_resid   1.000e+00  8.450e-17 1.183e+16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.112e-17 on 177 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.081e+32 on 2 and 177 DF,  p-value: < 2.2e-16
```

In part (h), we attempt to test for the exogeneity of the market excess return by including the residuals from the 2SLS model (from part g) into the structural equation:

$$(r_{\text{msft}} - r_f) = \alpha + \beta(r_m - r_f) + \delta \cdot \hat{u}_{IV} + e$$

However, this test is **not valid** in this form because:

- The residuals from a 2SLS model are by construction orthogonal to the instruments, but **not to the endogenous regressor**.
- When we include both the original regressor and its IV residuals, we create **perfect multicollinearity**, as seen in the warning and the perfect fit ($R^2 = 1$).

Conclusion:
This approach cannot be used to test exogeneity **using residuals from the 2SLS model directly**. Instead, exogeneity should be assessed using: - The **Wu-Hausman test**, which we already conducted in part (g), and which yielded a p-value of **0.0287**. - That result suggests we reject exogeneity at the 5% level and supports using IV estimation.

Hence, **OLS is inconsistent**, and **2SLS with `RANK` and `POS` should be preferred**.

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 with both `MOTHEREDUC` and `FATHEREDUC` as IV. These will serve as our baseline results.

24(a)

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

Ans

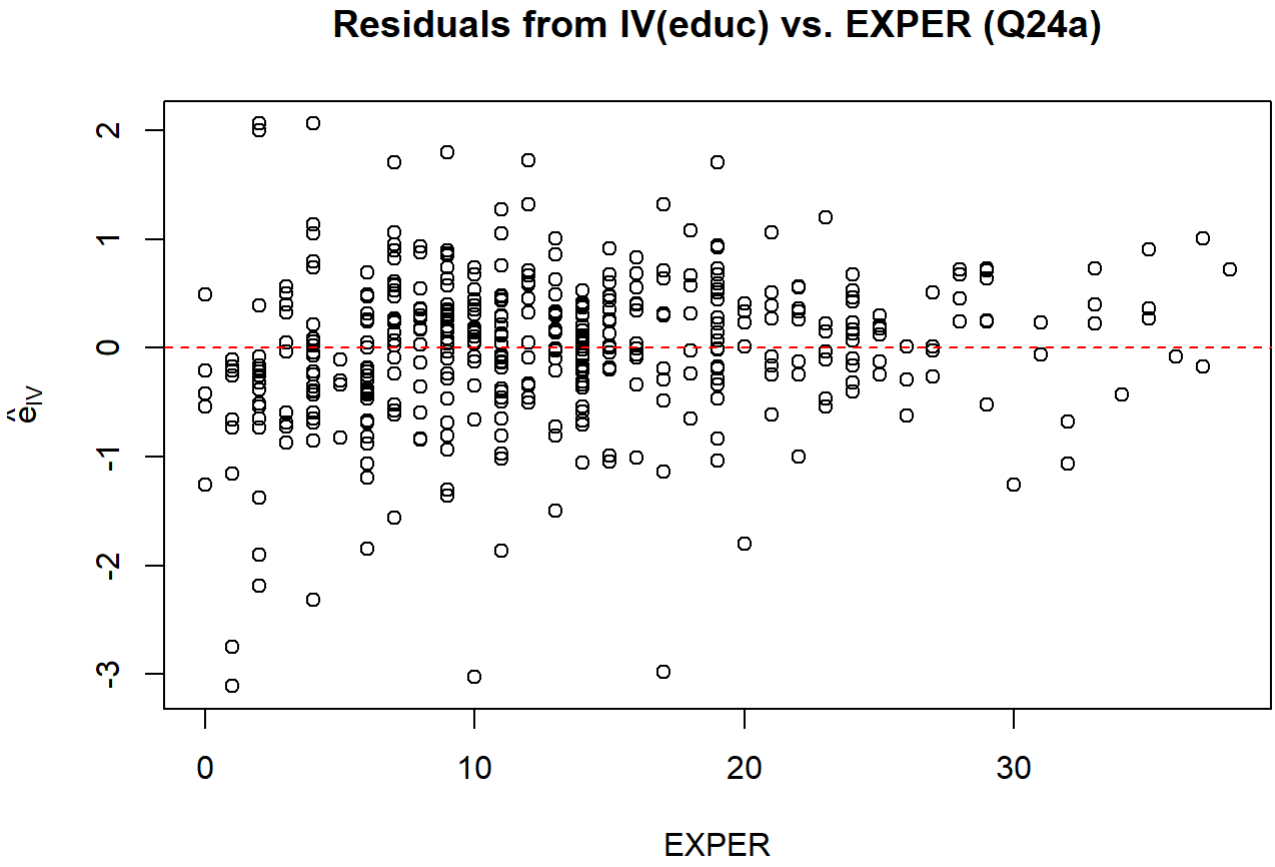
```
# Keep only observations with valid, positive wage
mroz_iv24 <- subset(mroz, !is.na(wage) & wage > 0)

# Create log wage variable
mroz_iv24$log_wage_iv24 <- log(mroz_iv24$wage)

# Estimate IV/2SLS model again
iv_model_24a <- ivreg(log_wage_iv24 ~ educ | mothereduc + fathereduc, data = mroz_iv24)

# Get residuals
resid_iv_24a <- resid(iv_model_24a)

# Plot residuals against experience
plot(mroz_iv24$exper, resid_iv_24a,
     xlab = "EXPER",
     ylab = expression(hat(e)[IV]),
     main = "Residuals from IV(educ) vs. EXPER (Q24a)")
abline(h = 0, col = "red", lty = 2)
```



We estimated the IV/2SLS model of `log(wage)` on `educ`, using `mothereduc` and `fathereduc` as instruments. The residuals $\hat{e}_{IV}$ from this model were plotted against `EXPER`. The scatterplot indicates that the residuals have greater dispersion at lower levels of `EXPER` and appear more compressed at higher levels. This suggests a potential violation of the homoskedasticity assumption, pointing toward heteroskedasticity. Therefore, a formal test is warranted in the next step.

24(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.

Ans

```
# Create squared residuals
mroz_iv24$resid_iv_24a_sq <- resid_iv_24a^2

# Run auxiliary regression: squared residuals on constant and EXPER
aux_model_24b <- lm(resid_iv_24a_sq ~ exper, data = mroz_iv24)

# Compute NR^2 test statistic
n_24b <- nobs(aux_model_24b)                  # number of observations
r2_24b <- summary(aux_model_24b)$r.squared    # R-squared
nr2_stat_24b <- n_24b * r2_24b

# Compute p-value (chi-squared with 1 df, since 1 regressor)
pval_24b <- 1 - pchisq(nr2_stat_24b, df = 1)

# Display test statistic and p-value
cat("NR^2 test statistic =", nr2_stat_24b, "\n")
```

```
## NR^2 test statistic = 9.22356
```

```
cat("p-value =", pval_24b, "\n")
```

```
## p-value = 0.002389205
```

We regressed the squared IV/2SLS residuals $\hat{e}_{IV}^2$ on a constant and `EXPER`, and calculated the test statistic for the NR^2 heteroskedasticity test. The resulting test statistic is approximately 9.22, with a p-value of 0.0024. Since the p-value is well below conventional significance levels (e.g., 0.05), we reject the null hypothesis of homoskedasticity. This provides formal statistical evidence that the error variance is not constant and supports the presence of heteroskedasticity in the model.

24(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.

Ans

```
# Compute robust standard errors for the IV model
robust_se_24c <- vcovHC(iv_model_24a, type = "HC1")

# Get robust test results
coeftest(iv_model_24a, vcov = robust_se_24c)
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.551020   0.431001  1.2785   0.2018
## educ        0.050490   0.034346  1.4701   0.1423
```

```
# Compute 95% confidence interval for 'educ' using robust SE
ci_educ_robust_24c <- coef(iv_model_24a)["educ"] +
  c(-1, 1) * 1.96 * sqrt(robust_se_24c["educ", "educ"])

# Display confidence interval
ci_educ_robust_24c
```

```
## [1] -0.01682757  0.11780852
```

We re-estimated the IV/2SLS model and computed heteroskedasticity-robust standard errors. For the coefficient on `educ`, the robust standard error is approximately 0.0343, which is slightly **larger** than the conventional standard error from the baseline IV model. This increase is consistent with the presence of heteroskedasticity.

Using the robust standard error, the 95% confidence interval for the `educ` coefficient is approximately $[-0.017, 0.118]$. Since the interval includes zero, the effect of education on log wages is not statistically significant at the 5% level when accounting for heteroskedasticity.

24(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.

Ans

```
# Define bootstrap function for IV/2SLS estimate of 'educ'
iv_boot_fn <- function(data, indices) {
  d <- data[indices, ]
  model <- ivreg(log_wage_iv24 ~ educ | mothereduc + fathereduc, data = d)
  return(coef(model)["educ"])
}

# Run bootstrap with 200 replications
set.seed(123)  # for reproducibility
boot_result_24d <- boot(data = mroz_iv24, statistic = iv_boot_fn, R = 200)

# Display bootstrap standard error
boot_se_educ_24d <- sd(boot_result_24d$t)
boot_se_educ_24d
```

```
## [1] 0.03204664
```

```r
# Compute 95% CI using normal approximation
educ_hat_24d <- coef(iv_model_24a)["educ"]
ci_educ_boot_24d <- educ_hat_24d + c(-1, 1) * 1.96 * boot_se_educ_24d
ci_educ_boot_24d
```

```
## [1] -0.01232094  0.11330189
```

We computed the bootstrap standard error of the IV/2SLS estimate for `educ` using 200 replications. The resulting bootstrap standard error is approximately 0.0320. Compared to the baseline model, this value is slightly smaller than the robust standard error reported in part (c), which was approximately 0.0343.

Using the bootstrap standard error and the normal approximation, we constructed a 95% confidence interval for the `educ` coefficient: $[-0.0123, 0.1133]$. Like the robust interval in part (c), this interval also includes zero, suggesting that the causal effect of education on log wages is not statistically significant at the 5% level even when using bootstrapped inference.