Problem Set 3

MaCCS 201 - Fall 2024

2024-11-08

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(broom)
```

Part A: Data Analysis from Blackburn and Neumark (QJE 1992)

Let's take a look at the data from Blackburn and Neumark (QJE 1992). We would like to estimate the model:

$$\log(\text{wage}) = \beta_0 + \text{exper} \cdot \beta_1 + \text{tenure} \cdot \beta_2 + \text{married} \cdot \beta_3 + \text{south} \cdot \beta_4 + \text{urban} \cdot \beta_5 + \text{black} \cdot \beta_6 + \text{educ} \cdot \beta_7 + \text{abil} \cdot \gamma + \epsilon$$

One of the big problems in the wage literature is that we do not observe ability. If ability is not correlated with any of the right hand side variables, we can include it in the disturbance and nothing is lost by not observing it. If, however, it is correlated with one or more of the right hand side variables, OLS is no longer unbiased or consistent. Assume that ability is correlated with education and none of the other right hand side variables.

Q1 Derive the Bias of β_7

Derive the bias of β_7 and show what direction the bias goes in depending on whether the correlation between ability and education is positive or negative.

$$\text{bias}(\hat{\beta}_7) = \gamma \cdot \frac{\text{Cov}(educ, abil)}{\text{Var}(educ)}$$

```
data <- read.csv("newburn.csv")

# Step 1: Check correlation between ability (IQ) and education
cor(data$iq, data$educ, use = "complete.obs")
```

```
## [1] 0.515697
```

The correlation between ability and education is positive.

```
# Step 2: Estimate the model without IQ (omitting ability)
model_without_ability <- lm(lwage ~ exper + tenure + married + south + urban + black + educ, data = data

# Step 3: Estimate the model including IQ as a proxy for ability
model_with_ability <- lm(lwage ~ exper + tenure + married + south + urban + black + educ + iq, data = da

# Step 4: Compare estimates of beta_7 (educ coefficient) in both models
beta7_comparison <- data.frame(
  Model = c("Without Ability", "With Ability"),
  Estimate = c(coef(model_without_ability)["educ"], coef(model_with_ability)["educ"]),
  Std_Error = c(summary(model_without_ability)$coefficients["educ", "Std. Error"],
                summary(model_with_ability)$coefficients["educ", "Std. Error"])
)
print(beta7_comparison)
```

```
##             Model   Estimate   Std_Error
## 1 Without Ability 0.06543073 0.006250395
## 2    With Ability 0.05441062 0.006928489
```

The correlation between ability and education is positive, when ability (IQ) is included, the estimate of β_7 decreases from approximately 0.065 to 0.054. This indicates that omitting ability from the model leads to an upward bias in the estimate of β_7.

Q2 Proxy for Ability with IQ

You showed in the first part that we can derive the sign/direction of the bias. One approach that has been take in the literature is using a "proxy" variable for the unobservable ability. We will use IQ here to proxy for ability. Estimate the model above excluding ability, record your parameter estimates, standard errors and R^2.

```
# Extract parameter estimates and standard errors
model_without_ability_summary <- summary(model_without_ability)
estimates <- coef(model_without_ability_summary)

parameter_estimates <- data.frame(
  Term = rownames(estimates),
  Estimate = estimates[, "Estimate"],
  Std_Error = estimates[, "Std. Error"]
)

# Display results
parameter_estimates
```

```
##                     Term    Estimate   Std_Error
## (Intercept) (Intercept)  5.39549713 0.113225043
## exper             exper  0.01404301 0.003185185
## tenure           tenure  0.01174728 0.002452973
## married         married  0.19941705 0.039050151
## south             south -0.09090365 0.026248508
## urban             urban  0.18391207 0.026958329
## black             black -0.18834991 0.037666636
## educ               educ  0.06543073 0.006250395
```

```
model_without_ability_summary$r.squared
```

```
## [1] 0.2525577
```

Q3 Estimation with IQ as a Proxy

Estimate the model including IQ as a proxy, record your parameter estimates, standard errors and R^2.

```
# Extract parameter estimates and standard errors
model_with_ability_summary <- summary(model_with_ability)
estimates <- coef(model_with_ability_summary)

parameter_estimates <- data.frame(
  Term = rownames(estimates),
  Estimate = estimates[, "Estimate"],
  Std_Error = estimates[, "Std. Error"]
)

# Display results
parameter_estimates
```

```
##                    Term     Estimate   Std_Error
## (Intercept) (Intercept)  5.176439197 0.128000596
## exper             exper  0.014145850 0.003165104
## tenure           tenure  0.011395095 0.002439383
## married         married  0.199764380 0.038802482
## south             south -0.080169470 0.026252920
## urban             urban  0.181946304 0.026792867
## black             black -0.143125310 0.039492453
## educ               educ  0.054410617 0.006928489
## iq                   iq  0.003559103 0.000991808
```

```
model_with_ability_summary$r.squared
```

```
## [1] 0.2628093
```

Q4 Impact on Returns to Schooling

What happens to returns to schooling? Does this result confirm your suspicion of how ability and schooling are expected to be correlated?

The returns to schooling, represented by the coefficient for education (educ), decrease when IQ (as a proxy for ability) is included in the model:

- Without IQ (Omitted Ability): The coefficient on education is approximately 0.065.
- With IQ (Included Ability): The coefficient on education decreases to approximately 0.054.

The result is intuitive, when IQ is omitted, the estimated coefficient for education is higher, suggesting a stronger effect of education on wages. Including IQ reduces this coefficient, implying that part of what was attributed to education was actually due to ability (IQ) rather than the effect of education alone, which means wage and education level are all dependent on ability.

Part B: Data Analysis From David Card (1995)

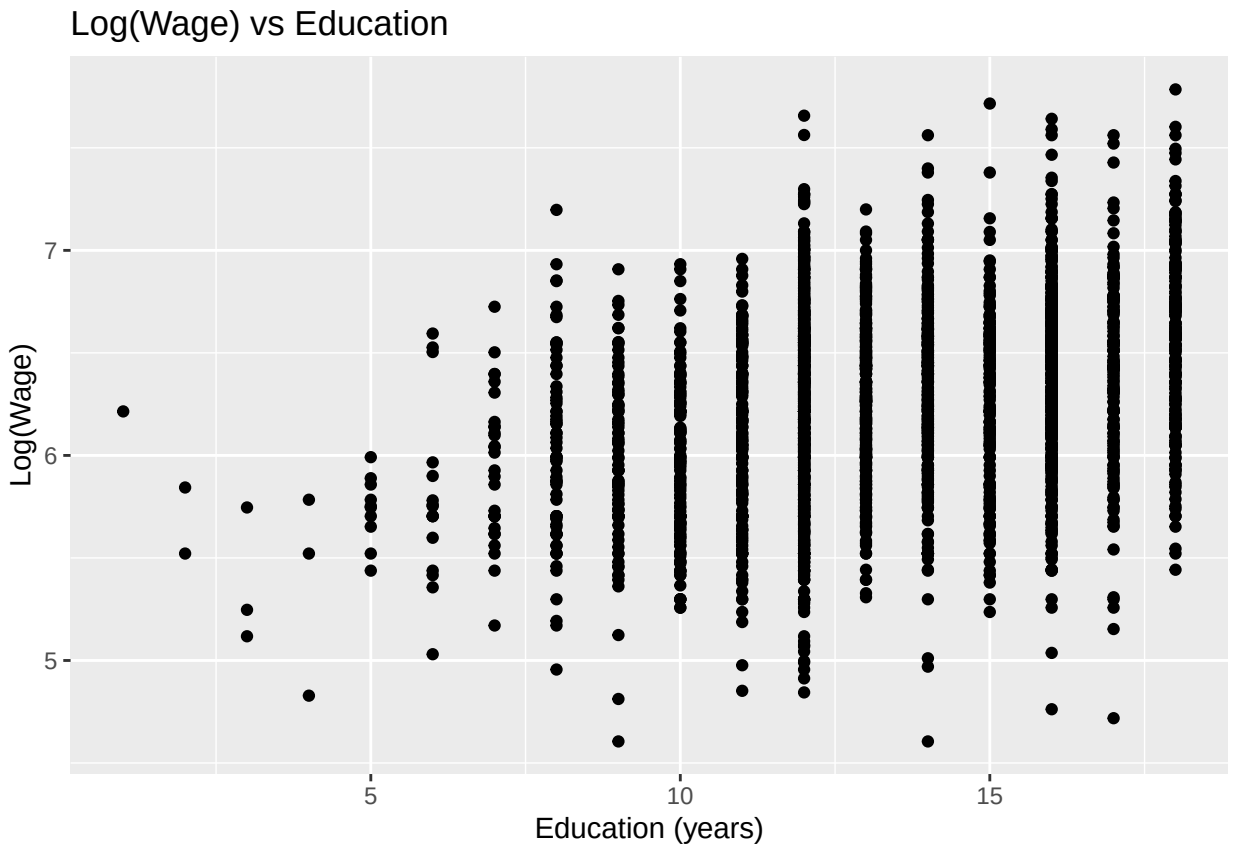
This next problem asks you to try and recreate some of the results in Card (1995), which is on git. Use the dataset card. raw on git.

Q1 Data Import and Visualization

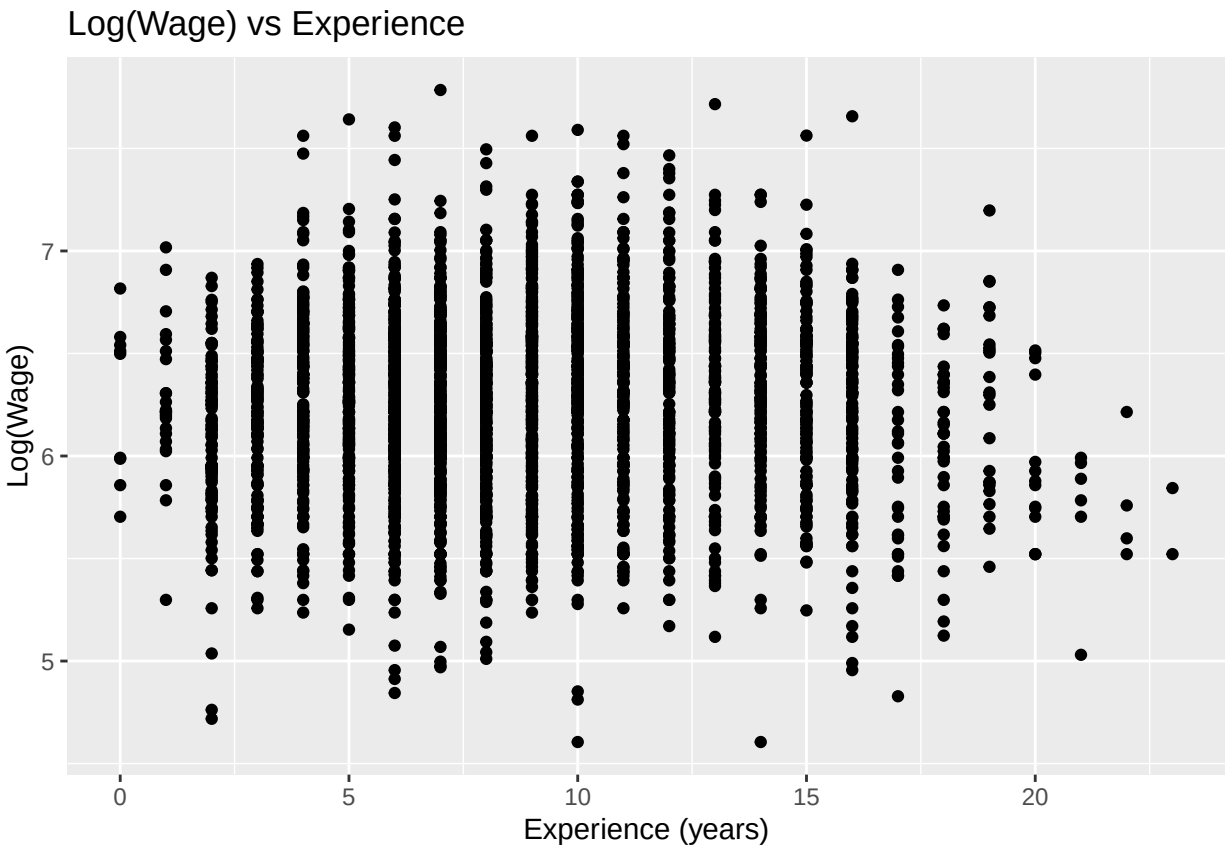
Read the data into R. Plot the series make sure your data are read in correctly.

```
# Read the Card (1995) dataset
data_card <- read.csv("card.csv")

ggplot(data_card, aes(x = educ, y = lwage)) +
  geom_point() +
  labs(title = "Log(Wage) vs Education",
       x = "Education (years)",
       y = "Log(Wage)")
```




```
ggplot(data_card, aes(x = exper, y = lwage)) +
  geom_point() +
  labs(title = "Log(Wage) vs Experience",
       x = "Experience (years)",
       y = "Log(Wage)")
```



Q2 Log(Wage) Regression via Least Squares

Estimate a $\log(wage)$ **regression via Least Squares with** $educ$, $exper$, $exper^2$, $black$, $south$, $smsa$, $reg661$ **through** $reg668$ **and** $smsa66$ **on the right hand side. Check your results against Table2, column 5.**

```
# Create an experience squared variable for the regression
data_card <- data_card %>%
  mutate(exper2 = exper^2)

# Run the OLS regression for log(wage) with the specified predictors
model_log_wage <- lm(lwage ~ educ + exper + exper2 + black + south + smsa +
                     reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
                     reg667 + reg668 + smsa66, data = data_card)

# Display the summary of the regression results
summary(model_log_wage)
```

```
##
## Call:
## lm(formula = lwage ~ educ + exper + exper2 + black + south +
##     smsa + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
##     reg667 + reg668 + smsa66, data = data_card)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.62326 -0.22141  0.02001  0.23932  1.33340
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.7393766  0.0715282  66.259  < 2e-16 ***
## educ         0.0746933  0.0034983  21.351  < 2e-16 ***
## exper        0.0848320  0.0066242  12.806  < 2e-16 ***
## exper2      -0.0022870  0.0003166  -7.223 6.41e-13 ***
## black       -0.1990123  0.0182483 -10.906  < 2e-16 ***
## south       -0.1479550  0.0259799  -5.695 1.35e-08 ***
## smsa         0.1363845  0.0201005   6.785 1.39e-11 ***
## reg661      -0.1185698  0.0388301  -3.054 0.002281 **
## reg662      -0.0222026  0.0282575  -0.786 0.432092
## reg663       0.0259703  0.0273644   0.949 0.342670
## reg664      -0.0634942  0.0356803  -1.780 0.075254 .
## reg665       0.0094551  0.0361174   0.262 0.793503
## reg666       0.0219476  0.0400984   0.547 0.584182
## reg667      -0.0005887  0.0393793  -0.015 0.988073
## reg668      -0.1750058  0.0463394  -3.777 0.000162 ***
## smsa66       0.0262417  0.0194477   1.349 0.177327
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3723 on 2994 degrees of freedom
## Multiple R-squared:  0.2998, Adjusted R-squared:  0.2963
## F-statistic: 85.48 on 15 and 2994 DF,  p-value: < 2.2e-16
```

Variable	Estimate (Table 2, Col 5)	Estimate (My Output)
Education	0.073	0.0747
Experience	0.085	0.0848
Experience-Squared	-0.229	-0.0023
Black	-0.189	-0.199
South	-0.146	-0.148
SMSA	0.138	0.136
R-squared	0.304	0.2998

Q3 Reduced Form Equation for Education

Estimate a reduced form equation for educ containing all of the explanatory variables and the dummy variable *nearc4*. **Is the partial correlation between** *nearc4* **and** *educ* **statistically significant?**

```r
# Fit the reduced form regression model for education
# educ is the dependent variable, and we include nearc4 and other explanatory variables
reduced_form_model <- lm(educ ~ exper + exper2 + black + south + smsa +
                         reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
                         reg667 + reg668 + smsa66 + nearc4, data = data_card)

# Display the summary of the reduced form model
summary(reduced_form_model)
```

```
##
## Call:
## lm(formula = educ ~ exper + exper2 + black + south + smsa + reg661 +
##     reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668 +
```

```
##     smsa66 + nearc4, data = data_card)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.545 -1.370 -0.091  1.278  6.239
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.8485239  0.2111222  79.805  < 2e-16 ***
## exper       -0.4125334  0.0336996 -12.241  < 2e-16 ***
## exper2       0.0008686  0.0016504   0.526 0.598728
## black       -0.9355287  0.0937348  -9.981  < 2e-16 ***
## south       -0.0516126  0.1354284  -0.381 0.703152
## smsa         0.4021825  0.1048112   3.837 0.000127 ***
## reg661      -0.2102710  0.2024568  -1.039 0.299076
## reg662      -0.2889073  0.1473395  -1.961 0.049992 *
## reg663      -0.2382099  0.1426357  -1.670 0.095012 .
## reg664      -0.0930890  0.1859827  -0.501 0.616742
## reg665      -0.4828875  0.1881872  -2.566 0.010336 *
## reg666      -0.5130857  0.2096352  -2.448 0.014442 *
## reg667      -0.4270887  0.2056208  -2.077 0.037880 *
## reg668       0.3136204  0.2416739   1.298 0.194490
## smsa66       0.0254805  0.1057692   0.241 0.809644
## nearc4       0.3198989  0.0878638   3.641 0.000276 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.941 on 2994 degrees of freedom
## Multiple R-squared:  0.4771, Adjusted R-squared:  0.4745
## F-statistic: 182.1 on 15 and 2994 DF,  p-value: < 2.2e-16
```

The partial correlation between *nearc*4 and *educ* is significantly 0.3199***.

Q4 Instrumental Variables Estimation of Log(Wage)

Estimate the log(wage) equation by instrumental variables, using *nearc*4 **as an instrument for** *educ*.

```
library(AER)
```

```
## Loading required package: car
## Loading required package: carData
##
## Attaching package: 'car'
## The following object is masked from 'package:dplyr':
##
##     recode
## The following object is masked from 'package:purrr':
##
##     some
## Loading required package: lmtest
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## Loading required package: survival
```

```
# Instrumental Variables (IV) estimation for log(wage) with nearc4 as an instrument for educ
iv_model <- ivreg(lwage ~ educ + exper + exper2 + black + south + smsa +
                  reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
                  reg667 + reg668 + smsa66 |
                  exper + exper2 + black + south + smsa + reg661 + reg662 +
                  reg663 + reg664 + reg665 + reg666 + reg667 + reg668 +
                  smsa66 + nearc4, data = data_card)

# Display the summary of the IV regression results
summary(iv_model)
```

```
##
## Call:
## ivreg(formula = lwage ~ educ + exper + exper2 + black + south +
##     smsa + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
##     reg667 + reg668 + smsa66 | exper + exper2 + black + south +
##     smsa + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
##     reg667 + reg668 + smsa66 + nearc4, data = data_card)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83164 -0.24075  0.02428  0.25208  1.42760
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.7739651  0.9349470   4.037 5.56e-05 ***
## educ         0.1315038  0.0549637   2.393  0.01679 *
## exper        0.1082711  0.0236586   4.576 4.92e-06 ***
## exper2      -0.0023349  0.0003335  -7.001 3.12e-12 ***
## black       -0.1467757  0.0538999  -2.723  0.00650 **
## south       -0.1446715  0.0272846  -5.302 1.23e-07 ***
## smsa         0.1118083  0.0316620   3.531  0.00042 ***
## reg661      -0.1078142  0.0418137  -2.578  0.00997 **
## reg662      -0.0070465  0.0329073  -0.214  0.83046
## reg663       0.0404445  0.0317806   1.273  0.20325
## reg664      -0.0579172  0.0376059  -1.540  0.12364
## reg665       0.0384577  0.0469387   0.819  0.41267
## reg666       0.0550887  0.0526597   1.046  0.29559
## reg667       0.0267580  0.0488287   0.548  0.58373
## reg668      -0.1908912  0.0507113  -3.764  0.00017 ***
## smsa66       0.0185311  0.0216086   0.858  0.39119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3883 on 2994 degrees of freedom
```

```
## Multiple R-Squared: 0.2382,  Adjusted R-squared: 0.2343
## Wald test: 51.01 on 15 and 2994 DF,  p-value: < 2.2e-16
```

Compare the 95% confidence interval for the return to education to that obtained from the Least Squares regression above.

```
# Extract the 95% confidence interval for the return to education (educ coefficient)
confint(model_log_wage, level = 0.95)["educ", ]
```

```
##      2.5 %     97.5 %
## 0.06783385 0.08155266
```

```
confint(iv_model, level = 0.95)["educ", ]
```

```
##      2.5 %     97.5 %
## 0.02377702 0.23923066
```

- **Significance**: Both intervals are above zero, indicating that education has a statistically significant positive effect on log(wage) in both models.
- **Potential Endogeneity in OLS Estimate**: The difference in intervals may suggest that the OLS estimate of the return to education is biased.
- **Wider Confidence Interval in IV Model**: The IV estimate might be less biased, but the larger confidence interval suggests that the instrumented estimate for the return to education is less precise.

Q5 Multiple Instruments for Education

Now use multiple instruments. Use *nearc2* **and** *nearc4* **as instruments for educ. Comment on the significance of the partial correlations of both instruments in the reduced form. Show your standard errors from the second stage and compare them to the correct standard errors.**

```
# Run the reduced form regression for educ with both nearc2 and nearc4 as instruments
reduced_form_multiple_instruments <- lm(educ ~ exper + exper2 + black + south + smsa +
                                          reg661 + reg662 + reg663 + reg664 + reg665 +
                                          reg666 + reg667 + reg668 + smsa66 + nearc2 + nearc4,
                                          data = data_card)

# Display the summary of the reduced form model
summary(reduced_form_multiple_instruments)
```

```
##
## Call:
## lm(formula = educ ~ exper + exper2 + black + south + smsa + reg661 +
##     reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668 +
##     smsa66 + nearc2 + nearc4, data = data_card)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5851 -1.3845 -0.0823  1.2765  6.2930
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.677e+01  2.163e-01  77.528  < 2e-16 ***
## exper       -4.123e-01  3.369e-02 -12.237  < 2e-16 ***
## exper2       8.479e-04  1.650e-03   0.514 0.607379
## black       -9.452e-01  9.391e-02 -10.065  < 2e-16 ***
## south       -4.191e-02  1.355e-01  -0.309 0.757162
```

```
## smsa         4.014e-01  1.048e-01   3.830 0.000131 ***
## reg661      -1.688e-01  2.041e-01  -0.827 0.408286
## reg662      -2.690e-01  1.478e-01  -1.820 0.068884 .
## reg663      -1.902e-01  1.458e-01  -1.305 0.192022
## reg664      -3.772e-02  1.892e-01  -0.199 0.841990
## reg665      -4.371e-01  1.903e-01  -2.297 0.021703 *
## reg666      -5.022e-01  2.097e-01  -2.395 0.016679 *
## reg667      -3.775e-01  2.079e-01  -1.816 0.069511 .
## reg668       3.820e-01  2.454e-01   1.557 0.119683
## smsa66       7.825e-05  1.069e-01   0.001 0.999416
## nearc2       1.230e-01  7.743e-02   1.589 0.112256
## nearc4       3.206e-01  8.784e-02   3.650 0.000267 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.94 on 2993 degrees of freedom
## Multiple R-squared:  0.4776, Adjusted R-squared:  0.4748
## F-statistic:   171 on 16 and 2993 DF,  p-value: < 2.2e-16
```

```
library(AER)

iv_model_multiple_instruments <- ivreg(lwage ~ educ + exper + exper2 + black + south + smsa +
                                       reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
                                       reg667 + reg668 + smsa66 |
                                       exper + exper2 + black + south + smsa + reg661 + reg662 +
                                       reg663 + reg664 + reg665 + reg666 + reg667 + reg668 +
                                       smsa66 + nearc2 + nearc4, data = data_card)

summary(iv_model_multiple_instruments)
```

```
##
## Call:
## ivreg(formula = lwage ~ educ + exper + exper2 + black + south +
##     smsa + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
##     reg667 + reg668 + smsa66 | exper + exper2 + black + south +
##     smsa + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
##     reg667 + reg668 + smsa66 + nearc2 + nearc4, data = data_card)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.93841 -0.25068  0.01932  0.26519  1.46998
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.3396868  0.8945377   3.733 0.000192 ***
## educ         0.1570594  0.0525782   2.987 0.002839 **
## exper        0.1188149  0.0228061   5.210 2.02e-07 ***
## exper2      -0.0023565  0.0003475  -6.781 1.43e-11 ***
## black       -0.1232778  0.0521500  -2.364 0.018147 *
## south       -0.1431945  0.0284448  -5.034 5.08e-07 ***
## smsa         0.1007530  0.0315193   3.197 0.001405 **
## reg661      -0.1029760  0.0434224  -2.371 0.017779 *
## reg662      -0.0002286  0.0337943  -0.007 0.994602
## reg663       0.0469556  0.0326490   1.438 0.150484
## reg664      -0.0554084  0.0391828  -1.414 0.157437
```

```
## reg665       0.0515041  0.0475678   1.083 0.279005
## reg666       0.0699968  0.0533049   1.313 0.189237
## reg667       0.0390596  0.0497499   0.785 0.432446
## reg668      -0.1980371  0.0525350  -3.770 0.000167 ***
## smsa66       0.0150626  0.0223360   0.674 0.500132
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4053 on 2994 degrees of freedom
## Multiple R-Squared: 0.1702,  Adjusted R-squared: 0.166
## Wald test: 47.07 on 15 and 2994 DF,  p-value: < 2.2e-16
```

```
library(sandwich)

# Calculate robust (heteroskedasticity-consistent) standard errors
robust_se <- sqrt(diag(vcovHC(iv_model_multiple_instruments, type = "HC1")))

# Display the coefficients along with both regular and robust standard errors
results <- cbind(
  Estimate = coef(iv_model_multiple_instruments),
  Std_Error = summary(iv_model_multiple_instruments)$coefficients[, "Std. Error"],
  Robust_SE = robust_se
)
results
```

```
##                  Estimate    Std_Error    Robust_SE
## (Intercept)  3.3396868121 0.8945377471 0.8932944001
## educ         0.1570593700 0.0525782417 0.0525525557
## exper        0.1188148807 0.0228060685 0.0229515635
## exper2      -0.0023564836 0.0003475175 0.0003683661
## black       -0.1232777953 0.0521500372 0.0516278294
## south       -0.1431944615 0.0284447849 0.0302678138
## smsa         0.1007530001 0.0315193428 0.0314457744
## reg661      -0.1029759964 0.0434223690 0.0426891448
## reg662      -0.0002286491 0.0337942719 0.0346150992
## reg663       0.0469556243 0.0326490358 0.0336146699
## reg664      -0.0554083884 0.0391828388 0.0410017705
## reg665       0.0515041450 0.0475677852 0.0507625378
## reg666       0.0699968047 0.0533049273 0.0535957697
## reg667       0.0390595603 0.0497498738 0.0515681885
## reg668      -0.1980370807 0.0525349918 0.0523728697
## smsa66       0.0150625816 0.0223359739 0.0211683384
```