

Problem Set 3 - Answer Key

MaCCS 201 - Fall 2025

2025-11-12

Tentative Due Date: November 8

Please submit markdown file named [last_name]_[first_name]_ps1.Rmd or a pdf with all code and answers.

##Part 1: Blackburn and Neumark (1992)

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

$$\log(wage) = \beta_0 + exper \cdot \beta_1 + tenure \cdot \beta_2 + married \cdot \beta_3 + south \cdot \beta_4 + urban \cdot \beta_5 + black \cdot \beta_6 + educ \cdot \beta_7 + abil \cdot \gamma + \varepsilon \quad (1)$$

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.

- 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.

The estimated coefficient on education (β_7) is given by the OVB formula. If you leave out ability, you will estimate $\beta_7 + \gamma \cdot \delta_1$. δ_1 here is the coefficient of education on ability in the “weird regression” from class. We can sign that one. We would think that $\gamma > 0$ as we might think that higher ability (whatever that means) might lead to higher wages (all else equal). We also might think that higher ability might “go with” more years of education. Also positive. Hence the product is positive, and hence the bias is positive!

- You showed in the first part that we can derive the sign/direction of the bias. One approach that has been taken 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 .

```
newburn <- read.csv("newburn.csv")
mod_1 <- lm(lwage ~ exper + tenure + married + south + urban + black + educ, data=newburn)
summary(mod_1)

##
## Call:
## lm(formula = lwage ~ exper + tenure + married + south + urban +
##     black + educ, data = newburn)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -10.00000 -1.00000  0.00000  1.00000 10.00000
```

```

## -1.98069 -0.21996 0.00707 0.24288 1.22822
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.395497  0.113225 47.653 < 2e-16 ***
## exper       0.014043  0.003185  4.409 1.16e-05 ***
## tenure      0.011747  0.002453  4.789 1.95e-06 ***
## married     0.199417  0.039050  5.107 3.98e-07 ***
## south       -0.090904  0.026249 -3.463 0.000558 ***
## urban        0.183912  0.026958  6.822 1.62e-11 ***
## black        -0.188350  0.037667 -5.000 6.84e-07 ***
## educ         0.065431  0.006250 10.468 < 2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3655 on 927 degrees of freedom
## Multiple R-squared: 0.2526, Adjusted R-squared: 0.2469
## F-statistic: 44.75 on 7 and 927 DF, p-value: < 2.2e-16

```

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

```

mod_2 <- lm(lwage ~ exper + tenure + married + south + urban + black + educ + iq, data=newburn)
summary(mod_2)

##
## Call:
## lm(formula = lwage ~ exper + tenure + married + south + urban +
##     black + educ + iq, data = newburn)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.01203 -0.22244  0.01017  0.22951  1.27478
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.1764392  0.1280006 40.441 < 2e-16 ***
## exper       0.0141458  0.0031651  4.469 8.82e-06 ***
## tenure      0.0113951  0.0024394  4.671 3.44e-06 ***
## married     0.1997644  0.0388025  5.148 3.21e-07 ***
## south       -0.0801695  0.0262529 -3.054 0.002325 **
## urban        0.1819463  0.0267929  6.791 1.99e-11 ***
## black        -0.1431253  0.0394925 -3.624 0.000306 ***
## educ         0.0544106  0.0069285  7.853 1.12e-14 ***
## iq           0.0035591  0.0009918  3.589 0.000350 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3632 on 926 degrees of freedom
## Multiple R-squared: 0.2628, Adjusted R-squared: 0.2564
## F-statistic: 41.27 on 8 and 926 DF, p-value: < 2.2e-16

```

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

Yes! What we see is that the coefficient without including IQ is larger than when we include it. It is always so nice to see that this stuff really works!

Part 2: Card (1995)

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

```
card <- read.csv("card.csv")
print(dfSummary(card))
```



```

## 32   exper      Mean (sd) : 8.9 (4.1)          24 distinct values : 301
##           [integer] min < med < max:           : (100
##           0 < 8 < 23           : : .
##           IQR (CV) : 5 (0.5)           . : : : . : .
##           :
##           :
## 33   lwage      Mean (sd) : 6.3 (0.4)          755 distinct values : 301
##           [numeric] min < med < max:           : (100
##           4.6 < 6.3 < 7.8           : : : .
##           IQR (CV) : 0.6 (0.1)           : : : .
##           :
##           :
## 34   expersq     Mean (sd) : 95.6 (84.6)         24 distinct values : 301
##           [integer] min < med < max:           : (100
##           0 < 64 < 529           : : .
##           IQR (CV) : 85 (0.9)           : : .
##           :
##           :
## -----

```

- b) 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.

```
mod_3 <- lm(lwage ~ educ + exper + exper^2 + black + south + smsa + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66, data = card)
```

```

## 
## Call:
## lm(formula = lwage ~ educ + exper + exper^2 + black + south +
##     smsa + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
##     reg667 + reg668 + smsa66, data = card)
## 
## Residuals:
##    Min      1Q  Median      3Q     Max
## -1.72549 -0.22918  0.01839  0.24091  1.30683
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 4.925618  0.067287 73.203 < 2e-16 ***
## educ        0.074458  0.003528 21.105 < 2e-16 ***
## exper       0.039638  0.002194 18.068 < 2e-16 ***
## black       -0.197546  0.018402 -10.735 < 2e-16 ***
## south       -0.153710  0.026189 -5.869 4.85e-09 ***
## smsa        0.140926  0.020262  6.955 4.30e-12 ***
## reg661      -0.119396  0.039160 -3.049 0.002317 ** 
## reg662      -0.025996  0.028493 -0.912 0.361641  
## reg663      0.022845  0.027594  0.828 0.407794  
## reg664      -0.063640  0.035984 -1.769 0.077066 .  
## reg665      0.008056  0.036424  0.221 0.824975  
## reg666      0.017514  0.040435  0.433 0.664940  
## reg667      0.003354  0.039710  0.084 0.932686  
## reg668      -0.172533  0.046732 -3.692 0.000227 *** 
## smsa66      0.024335  0.019611  1.241 0.214754  
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
```

```
## Residual standard error: 0.3754 on 2995 degrees of freedom
## Multiple R-squared:  0.2876, Adjusted R-squared:  0.2843
## F-statistic: 86.38 on 14 and 2995 DF,  p-value: < 2.2e-16
```

c) 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?

```
mod_4 <- lm(educ ~ exper + exper^2 + black + south + smsa + reg661 + reg662 + reg663 + reg664 + reg665  
summary(mod_4)
```

```

## 
## Call:
## lm(formula = educ ~ exper + exper^2 + black + south + smsa +
##      reg661 + reg662 + reg663 + reg664 + reg665 + reg666 + reg667 +
##      reg668 + smsa66 + nearc4, data = card)
## 
## Residuals:
##    Min      1Q  Median      3Q     Max 
## -7.5165 -1.3757 -0.0982  1.2867  6.2210 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 16.779396  0.165267 101.529 < 2e-16 ***
## exper       -0.395405  0.008743 -45.224 < 2e-16 ***
## black        -0.936158  0.093716 -9.989 < 2e-16 ***
## south        -0.049436  0.135349 -0.365 0.714953    
## smsa         0.400515  0.104751  3.824 0.000134 ***  
## reg661      -0.209961  0.202432 -1.037 0.299728    
## reg662      -0.287477  0.147297 -1.952 0.051069 .  
## reg663      -0.237056  0.142602 -1.662 0.096544 .  
## reg664      -0.093045  0.185960 -0.500 0.616864    
## reg665      -0.482419  0.188162 -2.564 0.010400 *  
## reg666      -0.511497  0.209588 -2.440 0.014725 *  
## reg667      -0.428663  0.205574 -2.085 0.037136 *  
## reg668       0.312685  0.241638  1.294 0.195757    
## smsa66       0.026284  0.105745  0.249 0.803716    
## nearc4       0.319703  0.087852  3.639 0.000278 *** 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```
##  
## Residual standard error: 1.94 on 2995 degrees of freedom  
## Multiple R-squared:  0.4771, Adjusted R-squared:  0.4746
```

d) Estimate the $\log(wage)$ equation by instrumental variables, using *nearc4* as an instrument for *educ*. Compute the 95% confidence interval for the parameter estimate of the variable obtained from the Instrumental Variables (IV) estimation.

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

```
summary(mod_5)

## 
## Call:
## iv_robust(formula = lwage ~ educ + exper + exper^2 + black +
##           south + smsa + reg661 + reg662 + reg663 + reg664 + reg665 +
##           reg666 + reg667 + reg668 + smsa66 | exper + exper^2 + black +
```

```

##      south + smsa + reg661 + reg662 + reg663 + reg664 + reg665 +
##      reg666 + reg667 + reg668 + smsa66 + nearc4, data = card,
##
## Standard error type: classical
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
## (Intercept) 3.93214  0.94124  4.1776 3.030e-05  2.08659  5.77768 2995
## educ        0.13315  0.05558  2.3959 1.664e-02  0.02418  0.24212 2995
## exper       0.06288  0.02208  2.8482 4.426e-03  0.01959  0.10616 2995
## black       -0.14354  0.05453 -2.6325 8.520e-03 -0.25046 -0.03663 2995
## south       -0.15044  0.02755 -5.4616 5.106e-08 -0.20445 -0.09643 2995
## smsa        0.11563  0.03193  3.6213 2.980e-04  0.05302  0.17824 2995
## reg661     -0.10830  0.04225 -2.5633 1.042e-02 -0.19114 -0.02546 2995
## reg662     -0.01042  0.03322 -0.3136 7.538e-01 -0.07555  0.05472 2995
## reg663      0.03773  0.03209  1.1759 2.397e-01 -0.02518  0.10065 2995
## reg664     -0.05788  0.03800 -1.5231 1.278e-01 -0.13239  0.01663 2995
## reg665      0.03799  0.04743  0.8011 4.231e-01 -0.05500  0.13098 2995
## reg666      0.05166  0.05317  0.9717 3.313e-01 -0.05259  0.15591 2995
## reg667      0.03170  0.04939  0.6417 5.211e-01 -0.06515  0.12854 2995
## reg668     -0.18889  0.05123 -3.6871 2.309e-04 -0.28935 -0.08844 2995
## smsa66     0.01633  0.02185  0.7472 4.550e-01 -0.02651  0.05917 2995
##
## Multiple R-squared:  0.2218 ,   Adjusted R-squared:  0.2182
## F-statistic: 50.36 on 14 and 2995 DF,  p-value: < 2.2e-16

```

The CI is [0.02418; 0.24212], which includes the point estimate from before (0.074).

- e) (Bonus) 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.

```
mod_6 <- lm(educ ~ exper + exper^2 + black + south + smsa + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66 + nearc4 + nearc2, data = card)
```

```

##
## Call:
## lm(formula = educ ~ exper + exper^2 + black + south + smsa +
##     reg661 + reg662 + reg663 + reg664 + reg665 + reg666 + reg667 +
##     reg668 + smsa66 + nearc4 + nearc2, data = card)
##
## Residuals:
##      Min      1Q Median      3Q      Max
## -7.5573 -1.3897 -0.0904  1.2736  6.2761
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 16.7053892  0.1716330  97.332 < 2e-16 ***
## exper       -0.3955715  0.0087416 -45.252 < 2e-16 ***
## black       -0.9458118  0.0938876 -10.074 < 2e-16 ***
## south       -0.0397620  0.1354504  -0.294 0.769119    
## smsa        0.3997409  0.1047249   3.817 0.000138 *** 
## reg661     -0.1683746  0.2040566  -0.825 0.409360    
## reg662     -0.2675837  0.1477874  -1.811 0.070303 .  
## reg663     -0.1889623  0.1457270  -1.297 0.194839    
## reg664     -0.0375312  0.1891509  -0.198 0.842730    

```

```

## reg665      -0.4365643  0.1903039  -2.294 0.021857 *
## reg666      -0.5006480  0.2096451  -2.388 0.016999 *
## reg667      -0.3789416  0.2078784  -1.823 0.068418 .
## reg668      0.3812661  0.2453828   1.554 0.120347
## smsa66      0.0007981  0.1069222   0.007 0.994045
## nearc4      0.3203922  0.0878309   3.648 0.000269 ***
## nearc2      0.1233125  0.0774137   1.593 0.111288
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.94 on 2994 degrees of freedom
## Multiple R-squared:  0.4775, Adjusted R-squared:  0.4749
## F-statistic: 182.4 on 15 and 2994 DF,  p-value: < 2.2e-16
mod_7 <- iv_robust(lwage ~ educ + exper + exper^2 + black + south + smsa + reg661 + reg662 + reg663 + :
summary(mod_7)

##
## Call:
## iv_robust(formula = lwage ~ educ + exper + exper^2 + black +
##           south + smsa + reg661 + reg662 + reg663 + reg664 + reg665 +
##           reg666 + reg667 + reg668 + smsa66 | exper + exper^2 + black +
##           south + smsa + reg661 + reg662 + reg663 + reg664 + reg665 +
##           reg666 + reg667 + reg668 + smsa66 + nearc4 + nearc2, data = card,
##
## Standard error type: classical
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
## (Intercept) 3.521945  0.89800  3.9220 8.978e-05  1.76120 5.28269 2995
## educ        0.157386  0.05302  2.9687 3.015e-03  0.05344 0.26134 2995
## exper       0.072474  0.02107  3.4393 5.913e-04  0.03116 0.11379 2995
## black       -0.121243  0.05261 -2.3044 2.127e-02 -0.22440 -0.01808 2995
## south       -0.149092  0.02865 -5.2032 2.090e-07 -0.20528 -0.09291 2995
## smsa         0.105187  0.03171  3.3171 9.204e-04  0.04301 0.16736 2995
## reg661      -0.103720  0.04378 -2.3693 1.788e-02 -0.18955 -0.01789 2995
## reg662      -0.003987  0.03404 -0.1171 9.068e-01 -0.07073 0.06275 2995
## reg663       0.043880  0.03289  1.3341 1.823e-01 -0.02061 0.10837 2995
## reg664      -0.055503  0.03950 -1.4050 1.601e-01 -0.13296 0.02196 2995
## reg665       0.050353  0.04795  1.0502 2.937e-01 -0.04366 0.14436 2995
## reg666       0.065760  0.05369  1.2247 2.208e-01 -0.03952 0.17104 2995
## reg667       0.043397  0.05020  0.8645 3.874e-01 -0.05503 0.14182 2995
## reg668      -0.195648  0.05295 -3.6947 2.240e-04 -0.29948 -0.09182 2995
## smsa66      0.013020  0.02253  0.5779 5.634e-01 -0.03116 0.05720 2995
##
## Multiple R-squared:  0.1562 ,   Adjusted R-squared:  0.1523
## F-statistic: 46.69 on 14 and 2995 DF,  p-value: < 2.2e-16

```