assignment-3

Sidra Effendi

13/03/2021

```
earnings_and_height <- read_dta("Earnings_and_Height.dta")
```

Q.1

```
# linear model for earnings and education
m1 <- lm(earnings ~ educ, data=earnings_and_height)
# Adjust standard errors
cov1         <- vcovHC(m1, type = "HC1")
r_se1 <- sqrt(diag(cov1)) # Robust SEs
```

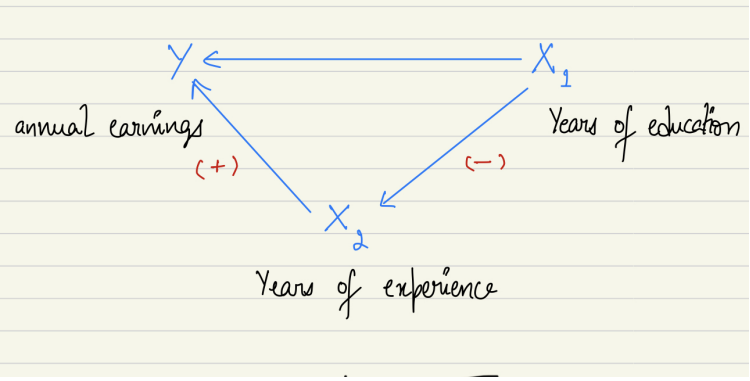
From the column (1) in the table at the bottom of the document we can see that with one year increase in education there is an expected $3,953.76 increase in annual earnings on average. This is statistically significant at 99% confidence interval because the p-value is less than 0.01.

Q.2

In order to bias our regression estimates, years of experience must be :

1. correlated with the annual earnings
2. correlated with the years of education

It is given that years_experience = age in years - years of education - 6. To find the nature of relationship between years of experience and years of education we can take into account that the age of a person keeps increasing every year and for a certain age we expect the years of education will increase which results in low value for years of experience. After a certain point, lets assume 16 years old (taking child labor laws into consideration), it is plausible that more individuals will leave education and find some employment, but the more an individual spends years acquiring education the less will be their years od experience. So, we can conclude the relationship between years of education and years of experience is negative. The relationship between the years of experience and the average annual earnings on the other hand, is positive as mentioned in the question.



$$Bias = \gamma_1 * \beta_2 = (-) * (+) = (-)$$

This would create a negative bias in our coefficient for years of education, meaning that the coefficient on years of education(3,953.76) will become more positive if we controlled for years of experience.

Q.3

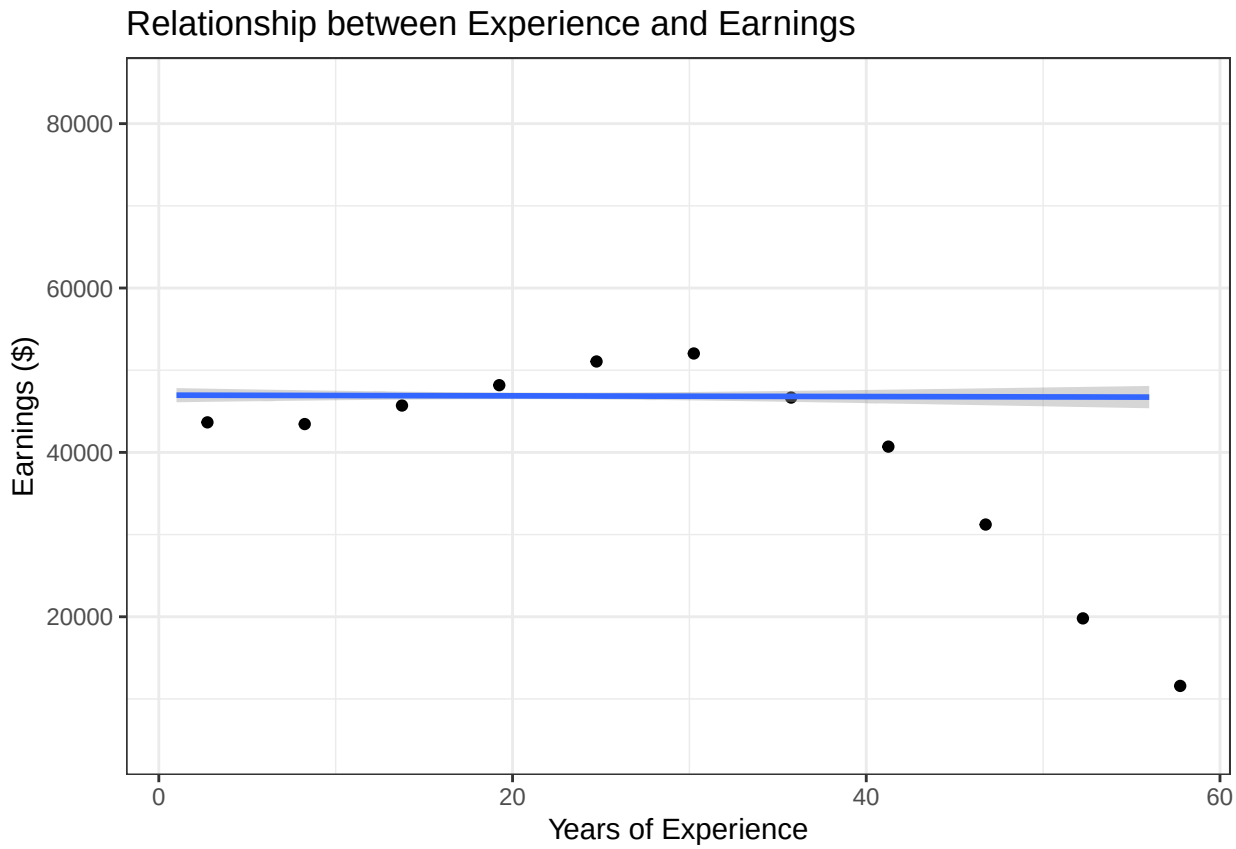
```
earnings_and_height<- earnings_and_height %>% mutate(exp=age-educ-6)
m2 <- lm(earnings ~ educ + exp, data=earnings_and_height)
# Adjust standard errors
cov2 <- vcovHC(m2, type = "HC1")
r_se2 <- sqrt(diag(cov2)) # Robust SEs
```

Q.4

Observing the coefficient value of education when we control for years of experience versus when we don't, we can see that the average predicted annual earnings for one more year of education increases in the former case. The result is consistent with my findings from the OVB triangle, which means when we omit years of experience when predicting the average earnings, we are underestimating the impact of years of education on earnings. Therefore omitting years of experience from the regression model introduces a negative bias in it.

Q.5

```
earnings_and_height %>%
ggplot()+ aes(x=exp,y=earnings)+ geom_point(alpha=0)+ stat_summary_bin(fun='mean',bins=10,geom='point')+
```




```
earnings_and_height %>%
ggplot()+ aes(x=exp,y=earnings)+ geom_point(alpha=0)+ stat_summary_bin(fun='mean',bins=10,geom='point')+
```

We can see that a regression model which includes quadratic term fits our data perfectly because it successfully explains the curve in output value. The curve translates to the fact that more years of experience do not necessarily mean an individual's annual earnings will increase. The yearly earnings increase with the increase in years of experience on average but the upward trend peaks at about 30 years of experience. Beyond that, the annual earnings for an individual on average starts to decrease, which might be because the individual's age is increasing, and he/she can no more handle the workload a high-income position brings.

Q.6

```
# add quadratics term to the dataset
earnings_and_height<- earnings_and_height %>% mutate(exp2=exp^2)
# regress annual earnings against experience and its square
m3 <- lm(earnings ~ educ + exp + exp2,data=earnings_and_height)
# Adjust standard errors
cov3  <- vcovHC(m3, type = "HC1")
r_se3 <- sqrt(diag(cov3)) # Robust SEs
```

(a)

From the table generated at the end of the document,column(3), we note that the coefficient on the quadratic term is -26.42.

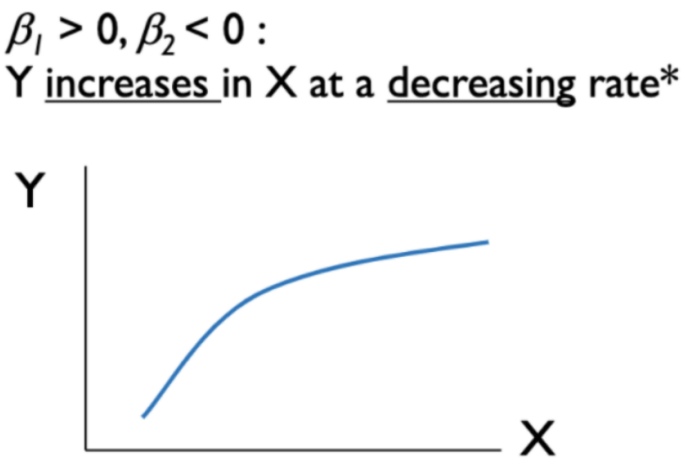
(b)

$$H_0 : \beta_2 = 0$$
$$H_A : \beta_2 \neq 0$$

p-value < 0.01, implies reject the null. So, in statistical sense we reject the hypothesis that the relationship between years of experience and annual earnings is linear with 99% confidence.

(c)

The coefficient on years of experience is positive, but the coefficient is negative for its quadratic terms, therefore we predict that the annual earnings increases in years of experience at a decreasing rate keeping the years of education constant. The shape of our graph will be like this:



(i)

$$\hat{earnings} = -29,659.57 + 4,284.32 * educ + 1,568.57 * exp - 26.42 * exp^2$$
$$\hat{earnings} = -29,659.57 + 4,284.32 * educ + 1,568.57 * 10 - 26.42 * (10)^2$$
$$\hat{earnings} = -29,659.57 + 4,284.32 * educ + 1,568.57 * 20 - 26.42 * (20)^2$$
$$Change\ in\ annual\ earnings = 7759.7$$

(ii)

$$\hat{earnings} = -29,659.57 + 4,284.32 * educ + 1,568.57 * exp - 26.42 * exp^2$$
$$\hat{earnings} = -29,659.57 + 4,284.32 * educ + 1,568.57 * 30 - 26.42 * (30)^2$$
$$\hat{earnings} = -29,659.57 + 4,284.32 * educ + 1,568.57 * 40 - 26.42 * (40)^2$$
$$Change\ in\ annual\ earnings = -2808.3$$

When the years of experience increase from 10 to 20 years there is an increase in the expected average annual earnings, but when the years of experience is from 30 to 40 year there is a decline in the expected yearly earnings. A similar trend was visualized in Q.5

Q.7

```r
earnings_and_height$married <- as.factor(earnings_and_height$married)
earnings_and_height$race <- as.factor(earnings_and_height$race)
earnings_and_height$cworker <- as.factor(earnings_and_height$cworker)
m4 <- lm(earnings ~ educ + exp + exp2 + height + married + race + cworker, data=earnings_and_height)

# Adjust standard errors
cov4 <- vcovHC(m4, type = "HC1")
r_se4 <- sqrt(diag(cov4)) # Robust SEs
```

(a)

With one more year of education we see on average the predicted annual income increase by $3,953.76, $4,347.50, $4,284.32 and $4,051.18 respectively from column 1 to 4 of the regression table. All of these values are significant as the p-value for all the coefficients is < 0.01. The coefficient for years of education is highest when we control for years of experience, because controlling for years of experience omitts bias. When we control for (years of experience)^2, we explain the non-linear nature of our prediction without adding a large no.on variables and when controlling for other variables (height, marital status, racial category and type of worker), the value of coefficient on years of education increases as compared to what we observe in column 1 but not as much as is in column 2 or 3. This signals that imperfect multi-collinearity might be introduced in the last regression which reduces the variance in years of education because when interpreting coefficient on years of education we hold other variables constant and this results in the increase in standard errors.

Since, all values are significant depending on what our requirement is we can pick a value from the column which suits our situation.

(b)

The R^2 value from the column 1 to 4 is increasing. This is because the addition of more variables into the regression model means greater share of variance in annual earning can now be explained. When regressing on years of education only 15% of variance in annual earnings could be explained, but when other regressors are added we can now explain 33% variance in annual earnings.

(c)

$$H_0 : \beta_{fed\_gov\ emp} = 0\ and\ \beta_{st.\_gov\ emp} = 0\ and\ \beta_{local\_gov\ emp} = 0\ and\ \beta_{incorporated\ business} = 0\ and\ \beta_{self\_emp} = 0$$

$$H_A : \textit{ all or atlest one of these is not equal to } 0$$

```r
linearHypothesis(m4,
 c("cworker2=0","cworker3=0","cworker4=0","cworker5=0","cworker6=0"),
 white.adjust = "hc1", test = "F")
```

```
## Linear hypothesis test
##
## Hypothesis:
## cworker2 = 0
## cworker3 = 0
## cworker4 = 0
## cworker5 = 0
## cworker6 = 0
```

```
##
## Model 1: restricted model
## Model 2: earnings ~ educ + exp + exp2 + height + married + race + cworker
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1  17861
## 2  17856  5 34.793 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-stat = 34.79

p-value < 0.000

Restrictions = 5

We can reject the null because p-value < 0.05. Therefore, the type of worker variables is jointly significant or the type of worker has significant impact on the expected annual earnings of an individual.

(d)

$$H_0 : \beta_{non-hispanic\ black} = \beta_{hispanic}$$
$$H_A : \beta_{non-hispanic\ black} \neq \beta_{hispanic}$$

```
linearHypothesis(m4, c("race2=race3"), white.adjust = "hc1", test = "F")
```

```
## Linear hypothesis test
##
## Hypothesis:
## race2 - race3 = 0
##
## Model 1: restricted model
## Model 2: earnings ~ educ + exp + exp2 + height + married + race + cworker
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1  17857
## 2  17856  1 21.609 3.367e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-stat = 21.609

p-value < 0.000

Restrictions = 1

Yes, we can reject the null because p-value < 0.05. Therefore, there is a difference in the relationship between being non-Hispanic Black and earnings versus the relationship between being Hispanic and earnings, controlling for the other variables included in column 4

```
stargazer(m1,m2,m3,m4, se=list(r_se1,r_se2,r_se3,r_se4), type = "text", header=TRUE, omit.stat = c("f",
```

```
##
## Table 1: Results of Regression of annual earnings on years of education and other control variables
## ================================================================
##                                  earnings
##                  (1)           (2)           (3)           (4)
## ----------------------------------------------------------------
## educ         3,953.76***   4,347.50***   4,284.32***   4,051.18***
##                (67.67)       (69.36)       (69.68)       (73.08)
##
## exp                         326.54***     1,568.57***   1,191.84***
##                              (18.27)       (70.80)       (66.30)
##
## exp2                                       -26.42***     -19.47***
##                                             (1.46)        (1.36)
##
## height                                                    378.09***
##                                                           (43.10)
##
## married1                                                 19,914.95***
##                                                           (341.74)
##
## race2                                                    -6,533.77***
##                                                           (503.63)
##
## race3                                                    -2,935.29***
##                                                           (669.10)
##
## race4                                                    -3,433.22***
##                                                           (918.73)
##
## cworker2                                                  9,107.24***
##                                                           (845.16)
##
## cworker3                                                    583.77
##                                                           (727.88)
##
## cworker4                                                  1,387.37**
##                                                           (543.82)
##
## cworker5                                                  6,281.47***
##                                                           (1,262.07)
##
## cworker6                                                 -3,305.72***
##                                                           (675.77)
##
## Constant     -6,648.03*** -18,960.23*** -29,659.57*** -59,947.31***
##                (918.88)     (1,104.10)    (1,251.53)    (3,093.13)
##
## ----------------------------------------------------------------
## Observations    17,870        17,870        17,870        17,870
## R2               0.15          0.17          0.18          0.33
```

```
## Adjusted R2      0.15         0.17          0.18          0.33
## ================================================================
## Note:                                  *p<0.1; **p<0.05; ***p<0.01
##                              Robust standard errors in parantheses
```