Homework 4

Fareha Sameen

10/14/2020

```
Fareha Sameen
Homework #4
Group members: Neshma, Hertz
```

For this analysis we will be using the subgroup of people whose ages are in the range of 25 to 55. This subgroup is ideal because this group is most likely part of the labor force and work full time. This allows us to exclude people who are unemployed with high qualifications.

```
attach(acs2017_ny)
use_varb <- (AGE >= 25) & (AGE <= 55) & (LABFORCE == 2) & (WKSWORK2 > 4) & (UHRSWORK >= 35)
dat_use <- subset(acs2017_ny,use_varb) #
detach()
attach(dat_use)
```

Then, we try linear regression with the dat we have. In this, we set the wage as dependent and a dummy.

```
model_temp1 <- lm(INCWAGE ~ AGE + female + AfAm + Asian + Amindian + race_oth + Hispanic + educ_hs + edu
summary(model_temp1)
require(stargazer)
stargazer(model_temp1, type = "text")
```

The linear regression gives us the following data:

```
Call:
lm(formula = INCWAGE ~ AGE + female + AfAm + Asian + Amindian +
    race_oth + Hispanic + educ_hs + educ_somecoll + educ_college +
    educ_advdeg)

Residuals:
    Min      1Q  Median      3Q     Max
-148088  -33205  -10708   13053  625543

Coefficients:
               Estimate Std. Error t value
(Intercept)    -7096.25    2446.71  -2.900
AGE             1316.69      39.66  33.199
female        -24939.46     720.43 -34.617
AfAm          -11934.26    1130.37 -10.558
Asian            566.53    1369.83   0.414
Amindian       -8858.57    6077.71  -1.458
race_oth       -7526.49    1272.49  -5.915
Hispanic       -4224.82    1183.47  -3.570
educ_hs        10592.37    1814.71   5.837
educ_somecoll  22461.39    1857.67  12.091
educ_college   57155.71    1830.96  31.216
```

```
educ_advdeg    82766.43    1878.64  44.057
              Pr(>|t|)
(Intercept)   0.003730 **
AGE            < 2e-16 ***
female         < 2e-16 ***
AfAm           < 2e-16 ***
Asian         0.679188
Amindian      0.144971
race_oth      3.35e-09 ***
Hispanic      0.000358 ***
educ_hs       5.35e-09 ***
educ_somecoll  < 2e-16 ***
educ_college   < 2e-16 ***
educ_advdeg    < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 76760 on 46959 degrees of freedom
Multiple R-squared:   0.15, Adjusted R-squared:  0.1498
F-statistic: 753.6 on 11 and 46959 DF,  p-value: < 2.2e-16
```

```
================================================
                        Dependent variable:
                    ----------------------------
                               INCWAGE
------------------------------------------------
AGE                          1,316.691***
                               (39.661)

female                      -24,939.460***
                               (720.433)

AfAm                        -11,934.250***
                              (1,130.372)

Asian                          566.528
                             (1,369.834)

Amindian                      -8,858.569
                             (6,077.710)

race_oth                     -7,526.487***
                             (1,272.485)

Hispanic                     -4,224.816***
                             (1,183.469)

educ_hs                      10,592.370***
                             (1,814.709)

educ_somecoll                22,461.390***
                             (1,857.674)
```

```
educ_college                57,155.710***
                             (1,830.963)

educ_advdeg                 82,766.430***
                             (1,878.638)

Constant                    -7,096.252***
                             (2,446.712)

-----------------------------------------------
Observations                    46,971
R2                               0.150
Adjusted R2                      0.150
Residual Std. Error   76,755.980 (df = 46959)
F Statistic         753.551*** (df = 11; 46959)
===============================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

par(mfrow=c(2,2)) plot(model_temp1,col="red",pch=16,cex=1,lwd=1,lty=2)

Then, we try the regression with a different variable, the incwage of Amindian in the data. it gives us

nAmindian<-as.numeric(as.character(dat_use$INCWAGE)) par(mfrow=c(2,2)) Wage_Amindian<-lm(INCWAGE~Amindian) plot(Wage_Amindian,col="green",pch=14,cex=1,lwd=1,lty=2) summary(Wage_Amindian)

```
Call:
lm(formula = INCWAGE ~ Amindian)

Residuals:
   Min     1Q Median     3Q    Max
-72553 -40553 -20553  12447 587481

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  72552.5      384.7 188.586   <2e-16
Amindian    -22033.3     6571.2  -3.353    8e-04

(Intercept) ***
Amindian    ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 83240 on 46969 degrees of freedom
Multiple R-squared:  0.0002393, Adjusted R-squared:  0.000218
F-statistic: 11.24 on 1 and 46969 DF,  p-value: 8e-04
```

We performed the regression once again with another variable because the p value of Amindian is 0.14449

nHispanic<-as.numeric(as.character(dat_use$INCWAGE)) par(mfrow=c(2,2)) Wage_Hispanic<-lm(INCWAGE~Hispanic) plot(Wage_Hispanic,col="purple",pch=14,cex=1,lwd=1,lty=2) summary(Wage_Hispanic)

Call: lm(formula = INCWAGE ~ Hispanic)

Residuals: Min 1Q Median 3Q Max -75702 -39702 -18702 12168 585168

Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 75701.7 412.5 183.50 <2e-16 Hispanic -22869.5 1098.6 -20.82 <2e-16

(Intercept) ***Hispanic*** — Signif. codes:
0 ‘***’ 0.001 ‘**’ *0.01* ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 82860 on 46969 degrees of freedom Multiple R-squared: 0.009142, Adjusted R-squared: 0.00912 F-statistic: 433.3 on 1 and 46969 DF, p-value: < 2.2e-16

The last thing we did was use this code to get a regression line in a plot with all the data points to give us an idea about the relationship between the dependent variables and the independent variable we performed the regression on. It shows us how one variable changes due to a change in the other. The plot shows a positive correlation bewteen age and inctotal. “‘

require(AER) NNobs <- length(INCTOT) set.seed(12345) graph_obs <- (runif(NNobs) < 0.1) dat_graph <-subset(dat_use,graph_obs)
plot(INCTOT ~ jitter(AGE, factor = 2), pch = 16, col = rgb(0.5, 0.5, 0.5, alpha = 0.2), data = dat_graph)

plot(INCTOT ~ jitter(AGE, factor = 2), pch = 16, col = rgb(0.5, 0.5, 0.5, alpha = 0.2), ylim = c(0,150000), data = dat_graph)

to_be_predicted2 <- data.frame(AGE = 25:55, female = 1, AfAm = 0, Asian = 0, Amindian = 1, race_oth = 1, Hispanic = 1, educ_hs = 0, educ_somecoll = 0, educ_college = 1, educ_advdeg = 0) to_be_predicted2$yhat <- predict(model_temp1, newdata = to_be_predicted2)

lines(yhat ~ AGE, data = to_be_predicted2)