Homework Lab#4

Thakur Prasad Ghimire

10/5/2021

load("C:/Homework EcoB2000/Lab#4/acs2017\_ny\_data.RData")  
attach(acs2017\_ny)  
use\_varb <- (AGE >=25) & (AGE <=70) & (LABFORCE ==2) & (WKSWORK2 >4) & (UHRSWORK >=35)  
dat\_use <-subset(acs2017\_ny,use\_varb)  
detach()  
attach(dat\_use)  
library(stargazer)

##   
## Please cite as:

## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

library(AER)

## Loading required package: car

## Loading required package: carData

## 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

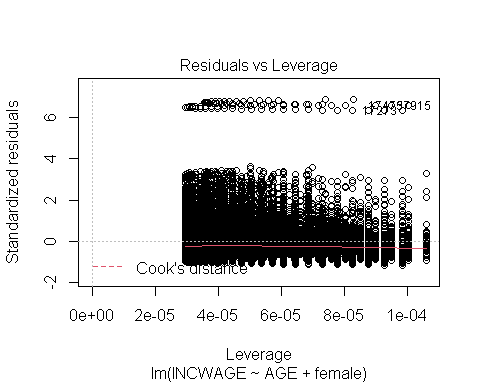
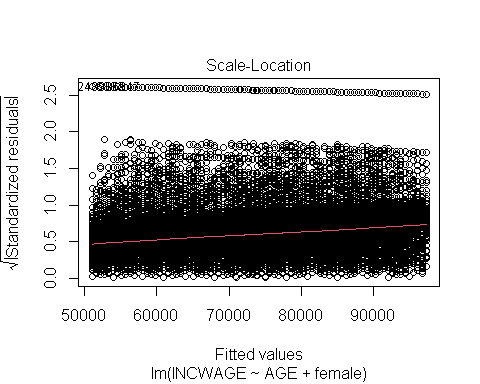
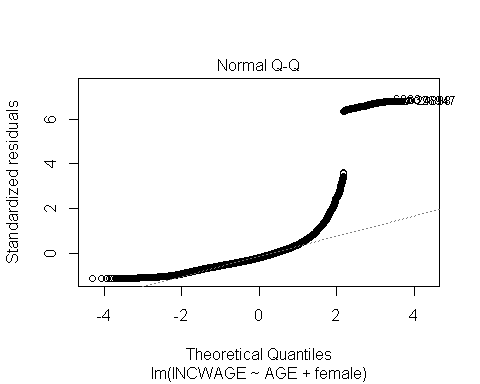
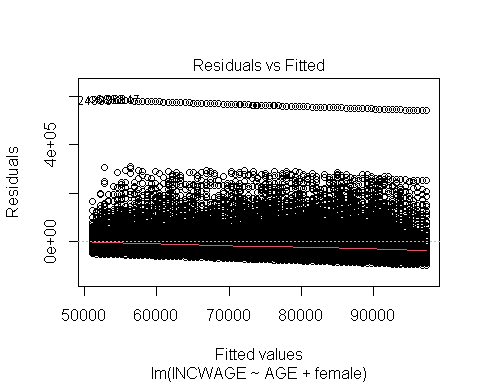
##   
## Attaching package: 'survival'

## The following object is masked from 'dat\_use':  
##   
## veteran

load("C:/Homework EcoB2000/Lab#4/acs2017\_ny\_data.RData")  
age\_gender <-lm(INCWAGE ~ AGE + female)  
summary(age\_gender)

##   
## Call:  
## lm(formula = INCWAGE ~ AGE + female)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -97275 -40016 -18510 12513 586940   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 56373.20 1395.05 40.41 <2e-16 \*\*\*  
## AGE 584.31 28.92 20.20 <2e-16 \*\*\*  
## female -19921.41 689.65 -28.89 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 85420 on 61942 degrees of freedom  
## Multiple R-squared: 0.01984, Adjusted R-squared: 0.01981   
## F-statistic: 626.9 on 2 and 61942 DF, p-value: < 2.2e-16

plot(age\_gender)



stargazer(age\_gender, type ="text")

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## INCWAGE   
## -----------------------------------------------  
## AGE 584.309\*\*\*   
## (28.922)   
##   
## female -19,921.420\*\*\*   
## (689.651)   
##   
## Constant 56,373.200\*\*\*   
## (1,395.051)   
##   
## -----------------------------------------------  
## Observations 61,945   
## R2 0.020   
## Adjusted R2 0.020   
## Residual Std. Error 85,415.510 (df = 61942)   
## F Statistic 626.851\*\*\* (df = 2; 61942)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

From above we see the Ordinary Least Squares (OLS) coefficient for the variables age and gender is 584.31 and -19921.41 respectively. Likewise, the wages has y-intercept at 56373.20. All three estimators are statistically significant to the 95% level. This is confirmed when we examine the t and p values for the estimators; the t-values for all variables are above the absolute critical t-value: 1.96 and the p-values are below 0.025. The confidence intervals are:

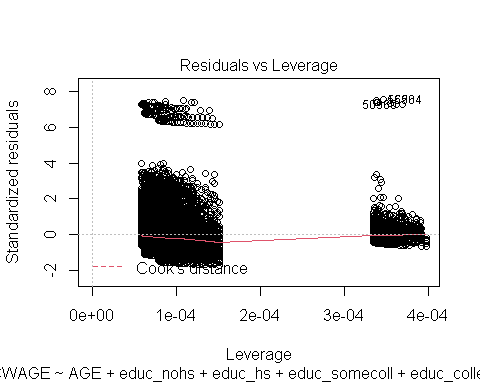
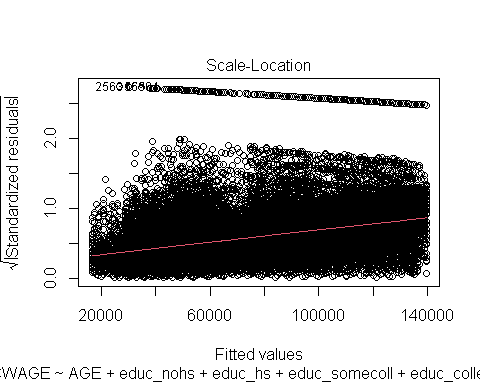
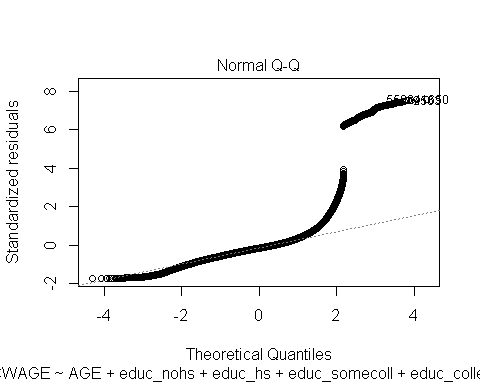
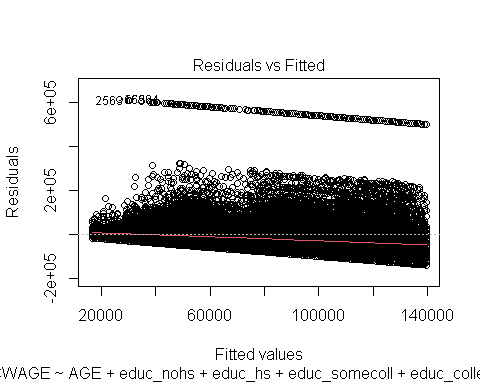
## 2.5 % 97.5 %  
## (Intercept) 53638.8985 59107.5039  
## AGE 527.6217 640.9965  
## female -21273.1315 -18569.6978

Since all of the estimators are within the intervals above can estimate that correlations exist between income, age and gender.

Considering R2 the model does not help to predict the income based on the variables of age and gender which may be because of opposite correlations with income.

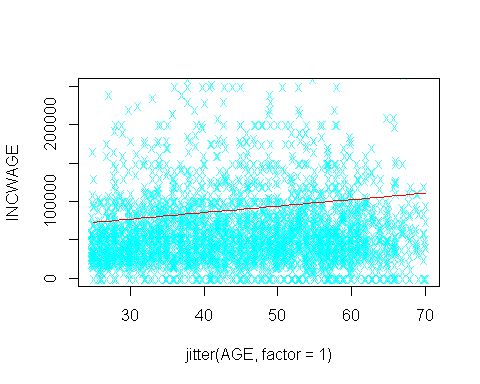
Next, removing the gender variable to see the impact of earning a degree on income, we see that degree holders earn a higher wage regardless of gender.

##   
## Call:  
## lm(formula = INCWAGE ~ AGE + educ\_nohs + educ\_hs + educ\_somecoll +   
## educ\_college + educ\_advdeg)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -139520 -34236 -12350 11330 611330   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 80677.28 1443.19 55.90 <2e-16 \*\*\*  
## AGE 840.61 27.75 30.30 <2e-16 \*\*\*  
## educ\_nohs -85109.74 1647.70 -51.65 <2e-16 \*\*\*  
## educ\_hs -72879.58 952.35 -76.53 <2e-16 \*\*\*  
## educ\_somecoll -62905.54 1009.73 -62.30 <2e-16 \*\*\*  
## educ\_college -28637.00 963.59 -29.72 <2e-16 \*\*\*  
## educ\_advdeg NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 80990 on 61939 degrees of freedom  
## Multiple R-squared: 0.1188, Adjusted R-squared: 0.1187   
## F-statistic: 1670 on 5 and 61939 DF, p-value: < 2.2e-16



##   
## ================================================  
## Dependent variable:   
## ----------------------------  
## INCWAGE   
## ------------------------------------------------  
## AGE 840.608\*\*\*   
## (27.746)   
##   
## educ\_nohs -85,109.740\*\*\*   
## (1,647.698)   
##   
## educ\_hs -72,879.580\*\*\*   
## (952.348)   
##   
## educ\_somecoll -62,905.540\*\*\*   
## (1,009.726)   
##   
## educ\_college -28,637.000\*\*\*   
## (963.586)   
##   
## educ\_advdeg   
##   
##   
## Constant 80,677.280\*\*\*   
## (1,443.191)   
##   
## ------------------------------------------------  
## Observations 61,945   
## R2 0.119   
## Adjusted R2 0.119   
## Residual Std. Error 80,992.050 (df = 61939)   
## F Statistic 1,669.641\*\*\* (df = 5; 61939)  
## ================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

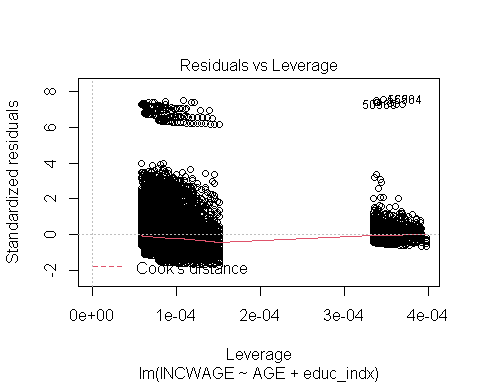
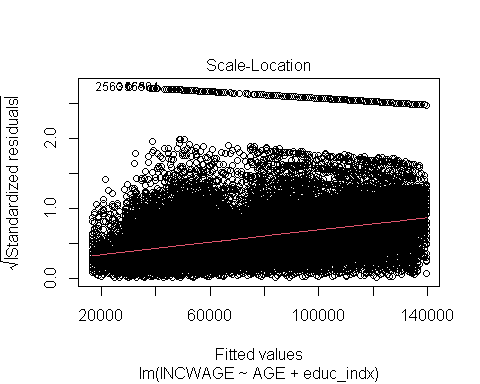
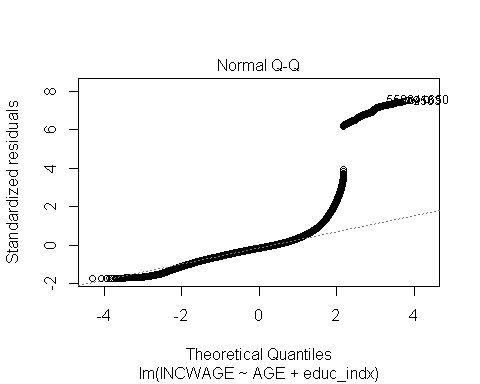
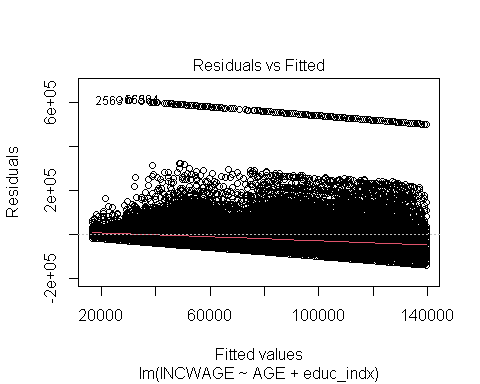
## Warning in predict.lm(age\_degree, newdata = predicteddeg): prediction from a  
## rank-deficient fit may be misleading



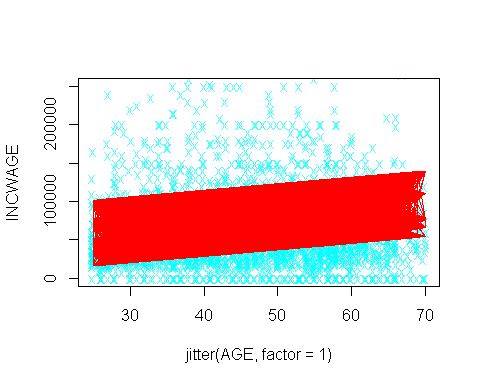
In this regression, education level has been taken as individual dummy variable. The slope coefficients for education levels are negative. Whereas the degree of slope reduces with the increase in education attainment. We can conclude that as the level of education increases, there is greater chance for higher income.

The regression here includes educational levels as variable while excluding No High School as a variable.

##   
## Call:  
## lm(formula = INCWAGE ~ AGE + educ\_indx)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -139520 -34236 -12350 11330 611330   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4432.46 1971.15 -2.249 0.0245 \*   
## AGE 840.61 27.75 30.296 < 2e-16 \*\*\*  
## educ\_indxHS 12230.16 1606.03 7.615 2.67e-14 \*\*\*  
## educ\_indxSmColl 22204.20 1642.23 13.521 < 2e-16 \*\*\*  
## educ\_indxBach 56472.74 1616.80 34.929 < 2e-16 \*\*\*  
## educ\_indxAdv 85109.74 1647.70 51.654 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 80990 on 61939 degrees of freedom  
## Multiple R-squared: 0.1188, Adjusted R-squared: 0.1187   
## F-statistic: 1670 on 5 and 61939 DF, p-value: < 2.2e-16



##   
## ================================================  
## Dependent variable:   
## ----------------------------  
## INCWAGE   
## ------------------------------------------------  
## AGE 840.608\*\*\*   
## (27.746)   
##   
## educ\_indxHS 12,230.160\*\*\*   
## (1,606.032)   
##   
## educ\_indxSmColl 22,204.200\*\*\*   
## (1,642.235)   
##   
## educ\_indxBach 56,472.740\*\*\*   
## (1,616.799)   
##   
## educ\_indxAdv 85,109.740\*\*\*   
## (1,647.698)   
##   
## Constant -4,432.460\*\*   
## (1,971.150)   
##   
## ------------------------------------------------  
## Observations 61,945   
## R2 0.119   
## Adjusted R2 0.119   
## Residual Std. Error 80,992.050 (df = 61939)   
## F Statistic 1,669.641\*\*\* (df = 5; 61939)  
## ================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01



Slope coefficients were all estimated as positive with t-values and p-values having statistical significance. Referring to the confidence intervals we see that:

## 2.5 % 97.5 %  
## (Intercept) -8295.9183 -569.0013  
## AGE 786.2256 894.9909  
## educ\_indxHS 9082.3299 15377.9814  
## educ\_indxSmColl 18985.4178 25422.9856  
## educ\_indxBach 53303.8061 59641.6664  
## educ\_indxAdv 81880.2469 88339.2323

We can be 95% confident that there is evidence to reject the H0 of no relationship since there is no zero value. All of the absolute t-values exceed the critical value of 1.96 state it statistically significant.

We have assumed that the variables are homoskedastic that is the variance in the errors of the dependent variable remains constant. It is assumed that other influences such as career choices, location and other lifestyle factors, the variance of errors in income will not remain constant while the independent variables change. If we adjust for heteroskedasticity, we observe the following:

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 56373.201 1182.709 47.664 < 2.2e-16 \*\*\*  
## AGE 584.309 26.528 22.026 < 2.2e-16 \*\*\*  
## female -19921.415 661.420 -30.119 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

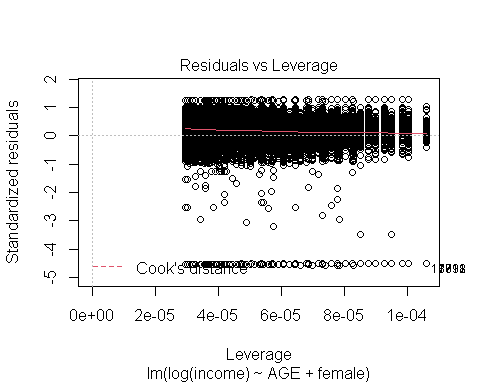
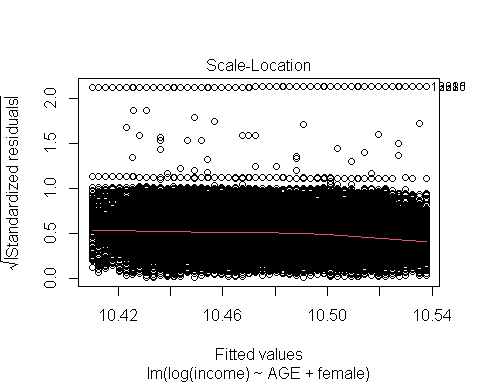
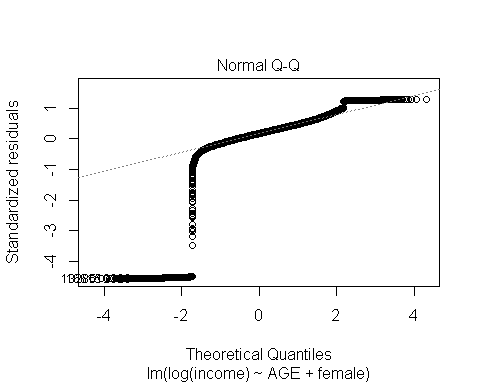
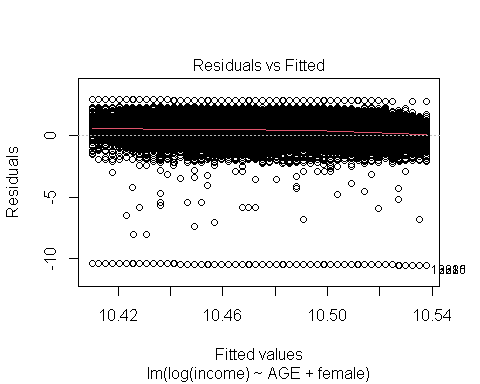
##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4432.460 1447.636 -3.0619 0.002201 \*\*   
## AGE 840.608 26.379 31.8662 < 2.2e-16 \*\*\*  
## educ\_indxHS 12230.156 856.811 14.2740 < 2.2e-16 \*\*\*  
## educ\_indxSmColl 22204.202 898.104 24.7234 < 2.2e-16 \*\*\*  
## educ\_indxBach 56472.736 1077.609 52.4056 < 2.2e-16 \*\*\*  
## educ\_indxAdv 85109.740 1375.636 61.8694 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Comparing the results of the two cases, the values of coefficient estimates have remained the same. In the first regression, standard error increased for the intercept and age coefficients, the t-values increased and the p-values decreased. For the female coefficient, the standard error decreased as shown by the negative slope t-value and p-value.

The confidence interval would fall within the 95 percent interval because of the estimator being consistent.

Taking a log of the y variable we see the percentage change of the dependent variable with the changes in the independent variables.

##   
## Call:  
## lm(formula = log(income) ~ AGE + female)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.5378 -0.0887 0.3982 0.8764 2.9560   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.592407 0.037770 280.447 < 2e-16 \*\*\*  
## AGE -0.002605 0.000783 -3.326 0.00088 \*\*\*  
## female 0.010488 0.018672 0.562 0.57432   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.313 on 61942 degrees of freedom  
## Multiple R-squared: 0.0001843, Adjusted R-squared: 0.000152   
## F-statistic: 5.708 on 2 and 61942 DF, p-value: 0.00332



##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## log(income)   
## -----------------------------------------------  
## AGE -0.003\*\*\*   
## (0.001)   
##   
## female 0.010   
## (0.019)   
##   
## Constant 10.592\*\*\*   
## (0.038)   
##   
## -----------------------------------------------  
## Observations 61,945   
## R2 0.0002   
## Adjusted R2 0.0002   
## Residual Std. Error 2.313 (df = 61942)   
## F Statistic 5.708\*\*\* (df = 2; 61942)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Taking a log-linear model and comparing the coefficients to previous coefficients, we see the values are much smaller.

## (Intercept) AGE female   
## 10.592407234 -0.002604686 0.010487952

## (Intercept) AGE female   
## 56373.2012 584.3091 -19921.4146