Homework #4

Collin Rafferty

10/7/2021

My group thought it would be interesting to examine the average wages of people holding business degrees in the later stages of their career- the people we examined were between the ages of 50 and 65 years. We examined how variables like gender, age, and race affected the average wage someone earned.

```
# Creation of subgroup
load("acs2017_ny_data.RData")
attach(acs2017_ny)
use_varb <- (AGE >= 50) & (AGE <=65) & (LABFORCE == 2) & (WKSWORK2 > 4) & (UHRSWORK >= 40) & (DEGFIELD==
dat_use <- subset(acs2017_ny,use_varb)
detach()
attach(dat_use)

# Testing for obvious errors in the subgroup
summary(AGE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   50.00   52.00   55.00   55.87   59.00   65.00
```

```
summary(DEGFIELD== "Business")
```

```
##    Mode    TRUE
## logical    1780
```

```
summary(female)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3697  1.0000  1.0000
```

```
summary(Hispanic)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.00000 0.05225 0.00000 1.00000
```

```
summary(AfAm)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.00000 0.08034 0.00000 1.00000
```

```
summary(Asian)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.00000 0.07584 0.00000 1.00000
```

```
summary(race_oth)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.00000 0.06798 0.00000 1.00000
```

```r
#Linear Regression Model
model_temp1 <- lm(INCWAGE ~ AGE + female+ Hispanic+ Asian+ AfAm + Asian + Amindian + race_oth)

require(stargazer)
```

```
## Loading required package: 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
```

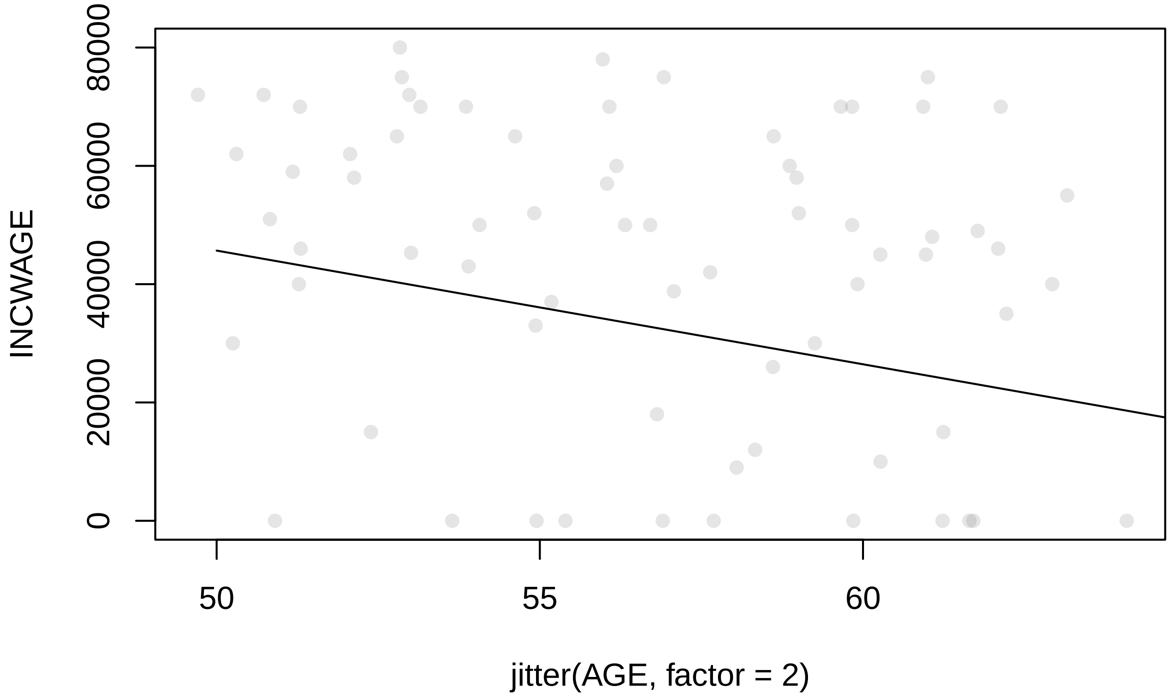
```r
stargazer(model_temp1, type = "text")
```

```
##
## ===============================================
##                          Dependent variable:
##                      ---------------------------
##                                INCWAGE
## -----------------------------------------------
## AGE                        -1,921.712**
##                              (815.338)
##
## female                    -44,799.390***
##                             (6,945.342)
##
## Hispanic                    -24,211.530
##                            (15,865.810)
##
## Asian                       -20,077.600
##                            (16,073.700)
##
## AfAm                      -54,185.080***
##                            (12,264.970)
##
## Amindian                    -80,997.750
##                            (98,352.980)
##
## race_oth                  -49,207.280***
##                            (17,348.090)
##
## Constant                  264,957.000***
##                            (46,096.370)
##
## -----------------------------------------------
## Observations                       1,780
## R2                                 0.054
## Adjusted R2                        0.050
## Residual Std. Error   138,941.500 (df = 1772)
## F Statistic          14.321*** (df = 7; 1772)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

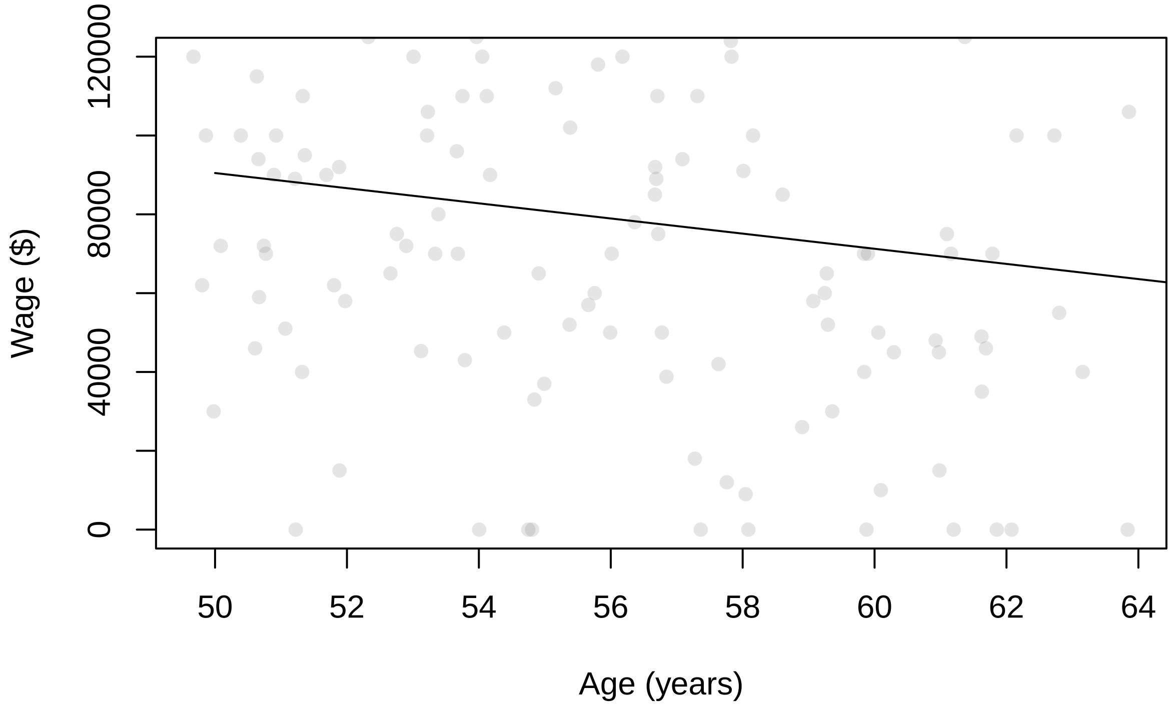
```r
#Confidence interval calculations
AGEl <--301.099-163.689
AGEr <--301.099+163.689
femalel <--22291.850-1413.806
femaler <--22291.850+1413.806
Hispanicl <--24464.040-2653.239
Hispanicr <--24464.040+2653.239
Asianl <- -9671.716-3178.407
Asianr <- -9671.716+3178.407
AfAml <--26408.360-2305.169
AfAmr <--26408.360+2305.169
Amindianl <--13521.030-13546.530
Amindianr <--13521.030+13546.530
race_othtl <--12728.930-2994.731
race_othtr <--12728.930+2994.731

require(AER)
```

```
## Loading required package: 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
```

```r
# Plotting the subset
NNobs <- length(INCWAGE)
set.seed(12345)
graph_obs <- (runif(NNobs) < 0.1)
dat_graph <-subset(dat_use,graph_obs)

plot(INCWAGE ~ jitter(AGE, factor = 2), pch = 16, col = rgb(0.5, 0.5, 0.5, alpha = 0.2), ylim = c(0,800

# Changing line to fit regression
to_be_predicted2 <- data.frame(AGE = 50:65, female = 1, AfAm = 1, Asian = 0, Amindian = 0, race_oth = 0
to_be_predicted2$yhat <- predict(model_temp1, newdata = to_be_predicted2)

lines(yhat ~ AGE, data = to_be_predicted2)
```



```
# Plotting different X variables

plot(INCWAGE ~ jitter(AGE, factor = 2), pch = 16, col = rgb(0.5, 0.5, 0.5, alpha = 0.2), ylim = c(0,120

# Changing line to fit regression
to_be_predicted3 <- data.frame(AGE = 50:65, female = 0, AfAm = 1, Asian = 0, Amindian = 0, race_oth = 0
to_be_predicted3$yhat <- predict(model_temp1, newdata = to_be_predicted3)

lines(yhat ~ AGE, data = to_be_predicted3)
```



The summary the model depicts gives quite revealing insights into the relationship between wage and the variables outlined in the model above. To begin with the variable of age, the null hypothesis associated with this variable(and all the others in the model) states the coefficient associated with the variable of age would be zero. The alternative hypothesis associated with age(and all the other variables in the model) states the coefficient would not be equal to zero. Or, put more simply, there is a relationship between age(the independent variable) and wage(the dependent variable). The value of the coefficient for age is about $-301, which means an increase in age of a year will cause a decrease to wages of $301. The coefficient is a non-zero number, so we reject the null. The P-value for age was 0.06586, which was not statistically significant but did weakly support the rejection of the null. The t-stat for age was -1.839458, which confirms the rejection of the null because it is a non-zero number. The confidence interval for age states we 95% confident that age's effect on the population wage lies between $-464.788 and $-137.41.

Similarly, the coefficients for all the other variables tested in the model were non-zero numbers, so we can reject the null hypotheses for all variables because they all, in varying degrees, have an effect on wage. The P-values for all of the variables, excluding American Indian, were below p<.01, which shows a highly statistically significant impact of these variables on wages. The p-value of American Indian was 0.31823, however, which was not statistically significant. The t-stat calculations are shown above, and all are non-zero numbers confirming a relationship exists between wage and each of these variables. The t-stat for the variable of female was the largest at about -16, while the t-stat for American Indian was the smallest at -1. This shines some interesting insight on the data: out of any of the variables test, the average wage for the variable of female has the most evidence for being significantly different than the average wage. The confidence intervals for all of the variables tested in the model are shown above. We 95% confident the true wage for each of the variables lies between these values.

The constant coefficient in the table would be the expected mean value of wage if all of the X variables were zero. It is saying that a person would make about $116,000 if they were male, zero years old, and not any of the races tested in this model. This does not make any sense.

```r
#Model without heteroskedasticity-consistent standard errors
summary(model_temp1)
```

```
##
## Call:
## lm(formula = INCWAGE ~ AGE + female + Hispanic + Asian + AfAm +
##     Asian + Amindian + race_oth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -168871  -74282  -33615   21787  546100
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 264957.0    46096.4   5.748 1.06e-08 ***
## AGE          -1921.7      815.3  -2.357  0.01853 *
## female      -44799.4     6945.3  -6.450 1.44e-10 ***
## Hispanic    -24211.5    15865.8  -1.526  0.12718
## Asian       -20077.6    16073.7  -1.249  0.21179
## AfAm        -54185.1    12265.0  -4.418 1.06e-05 ***
## Amindian    -80997.7    98353.0  -0.824  0.41031
## race_oth    -49207.3    17348.1  -2.836  0.00461 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 138900 on 1772 degrees of freedom
## Multiple R-squared:  0.05354,    Adjusted R-squared:  0.0498
## F-statistic: 14.32 on 7 and 1772 DF,  p-value: < 2.2e-16
```

```
#Model with heteroskedasticity-consistent standard errors
summary(coeftest(model_temp1,vcovHC))
```

```
##     Estimate        Std. Error         t value          Pr(>|t|)
##  Min.   :-80998   Min.   :  820.1   Min.   :-9.775   Min.   :0.0000000
##  1st Qu.:-50452   1st Qu.: 6118.6   1st Qu.:-7.375   1st Qu.:0.0000000
##  Median :-34505   Median :11706.1   Median :-2.989   Median :0.0001432
##  Mean   : -1305   Mean   :14167.0   Mean   :-3.538   Mean   :0.0404405
##  3rd Qu.:-15539   3rd Qu.:14981.7   3rd Qu.:-1.552   3rd Qu.:0.0394008
##  Max.   :264957   Max.   :46738.2   Max.   : 5.669   Max.   :0.2041026
```

Heteroskedasticity-consistent standard errors would affect the model by changing the standard errors of the variables test, and because of that, also their t-stats and p-values. It is important to include these errors in the model, however, because homoscedasticity can't always be counted on to be present in the model.

```
model_temp1 <- lm(INCWAGE ~ AGE + female + AfAm + Asian + Amindian + Hispanic + race_oth )
summary(model_temp1)
```

```
##
## Call:
## lm(formula = INCWAGE ~ AGE + female + AfAm + Asian + Amindian +
##     Hispanic + race_oth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -168871  -74282  -33615   21787  546100
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 264957.0    46096.4   5.748 1.06e-08 ***
## AGE          -1921.7      815.3  -2.357  0.01853 *
## female      -44799.4     6945.3  -6.450 1.44e-10 ***
## AfAm        -54185.1    12265.0  -4.418 1.06e-05 ***
## Asian       -20077.6    16073.7  -1.249  0.21179
## Amindian    -80997.7    98353.0  -0.824  0.41031
## Hispanic    -24211.5    15865.8  -1.526  0.12718
## race_oth    -49207.3    17348.1  -2.836  0.00461 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 138900 on 1772 degrees of freedom
## Multiple R-squared:  0.05354,    Adjusted R-squared:  0.0498
## F-statistic: 14.32 on 7 and 1772 DF,  p-value: < 2.2e-16
```

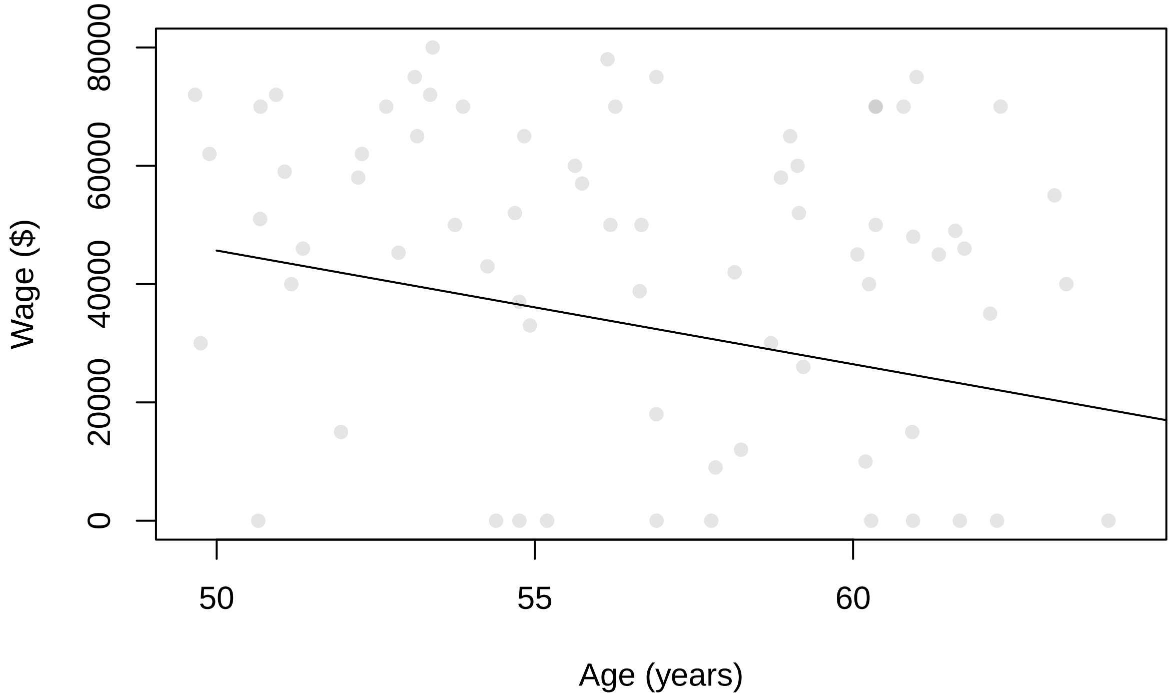
```
# Effect of log on the coefficients of the model
detach(dat_use)
dat_noZeroWage <- subset(dat_use,(INCWAGE > 0))
attach(dat_noZeroWage)
```

```
## The following object is masked from package:survival:
##
##     veteran
```

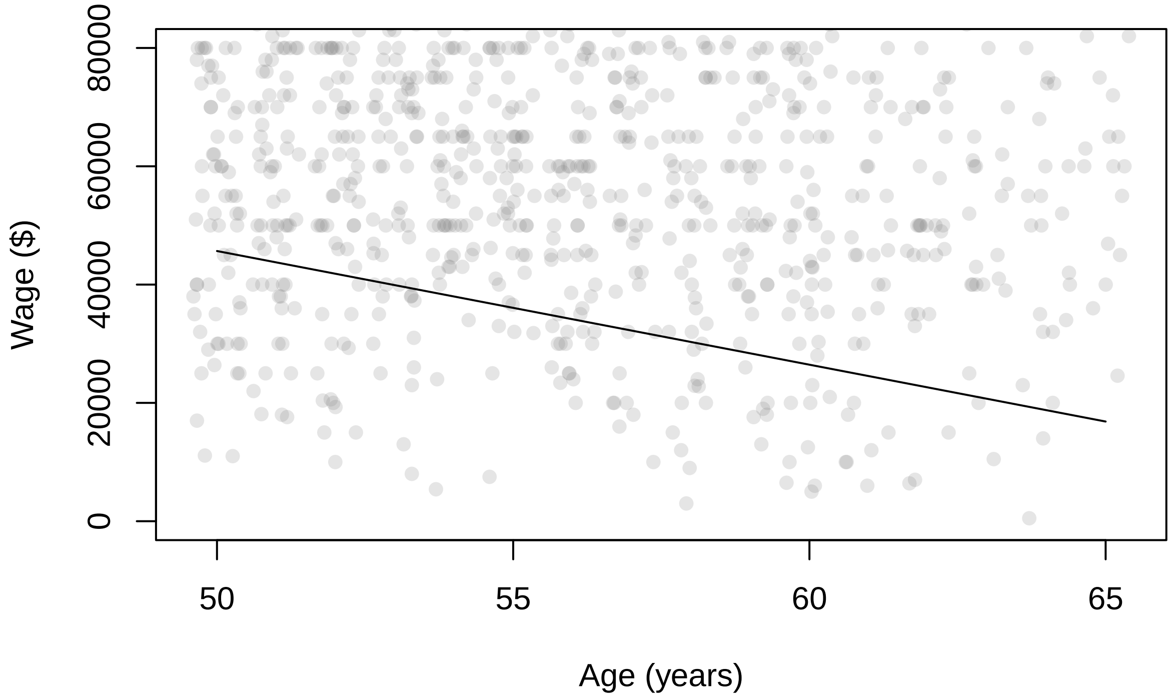
```
model_temp2 <- lm(log(INCWAGE) ~ AGE + female+ Hispanic+ Asian+ AfAm + Asian + Amindian + race_oth)
summary(model_temp2)
```

```
##
## Call:
## lm(formula = log(INCWAGE) ~ AGE + female + Hispanic + Asian +
##     AfAm + Asian + Amindian + race_oth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2933 -0.4669 -0.0364  0.4476  2.2677
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.844124   0.269345  47.687  < 2e-16 ***
## AGE         -0.020878   0.004762  -4.385 1.23e-05 ***
## female      -0.318217   0.040619  -7.834 8.30e-15 ***
## Hispanic    -0.249305   0.093704  -2.661  0.00788 **
## Asian       -0.152245   0.094730  -1.607  0.10821
## AfAm        -0.341461   0.071374  -4.784 1.87e-06 ***
## Amindian    -0.447236   0.561614  -0.796  0.42595
## race_oth    -0.466100   0.101646  -4.586 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7933 on 1678 degrees of freedom
## Multiple R-squared:  0.09494,    Adjusted R-squared:  0.09117
## F-statistic: 25.15 on 7 and 1678 DF,  p-value: < 2.2e-16
```

```
plot(INCWAGE ~ jitter(AGE, factor = 2), pch = 16, col = rgb(0.5, 0.5, 0.5, alpha = 0.2), ylim = c(0,8000

to_be_predicted2 <- data.frame(AGE = 50:65, female = 1, AfAm = 1, Asian = 0, Amindian = 0, race_oth = 0
to_be_predicted2$yhat <- predict(model_temp1, newdata = to_be_predicted2)

lines(yhat ~ AGE, data = to_be_predicted2)
```



```
plot(INCWAGE ~ jitter(AGE, factor = 2), pch = 16, col = rgb(0.5, 0.5, 0.5, alpha = 0.2), ylim = c(0,8000

to_be_predicted2 <- data.frame(AGE = 50:65, female = 1, AfAm = 1, Asian = 0, Amindian = 0, race_oth = 0
to_be_predicted2$yhat <- predict(model_temp1, newdata = to_be_predicted2)

lines(yhat ~ AGE, data = to_be_predicted2)
```



By comparing the two plots, we can see how the variable of wage would be affected by a logarithmic transformation. This transformation makes the data more normal(less skewed). It also increases the linearity between wage and the independent variables tested in the model.

Often the wage disparity between whites and minorities groups is discussed, but much less often, the wage disparity that exists among minorities is discussed. We believe our results help to shed light on this topic.