Homework 3

Rohan Thakur, Charles Kekeh and Megan Jasek

February 13, 2016

```r
# Load the dataframe
load("twoyear.RData")
desc
```

```
##    variable                             label
## 1    female                       =1 if female
## 2   phsrank  % high school rank; 100 = best
## 3        BA         =1 if Bachelor's degree
## 4        AA        =1 if Associate's degree
## 5     black          =1 if African-American
## 6  hispanic                    =1 if Hispanic
## 7        id                         ID Number
## 8     exper  total (actual) work experience
## 9        jc              total 2-year credits
## 10     univ              total 4-year credits
## 11    lwage                   log hourly wage
## 12   stotal   total standardized test score
## 13   smcity          =1 if small city, 1972
## 14  medcity            =1 if med. city, 1972
## 15   submed    =1 if suburb med. city, 1972
## 16   lgcity          =1 if large city, 1972
## 17    sublg   =1 if suburb large city, 1972
## 18  vlgcity     =1 if very large city, 1972
## 19   subvlg =1 if sub. very lge. city, 1972
## 20       ne                  =1 if northeast
## 21       nc              =1 if north central
## 22    south                      =1 if south
## 23  totcoll                         jc + univ
```

Question 1

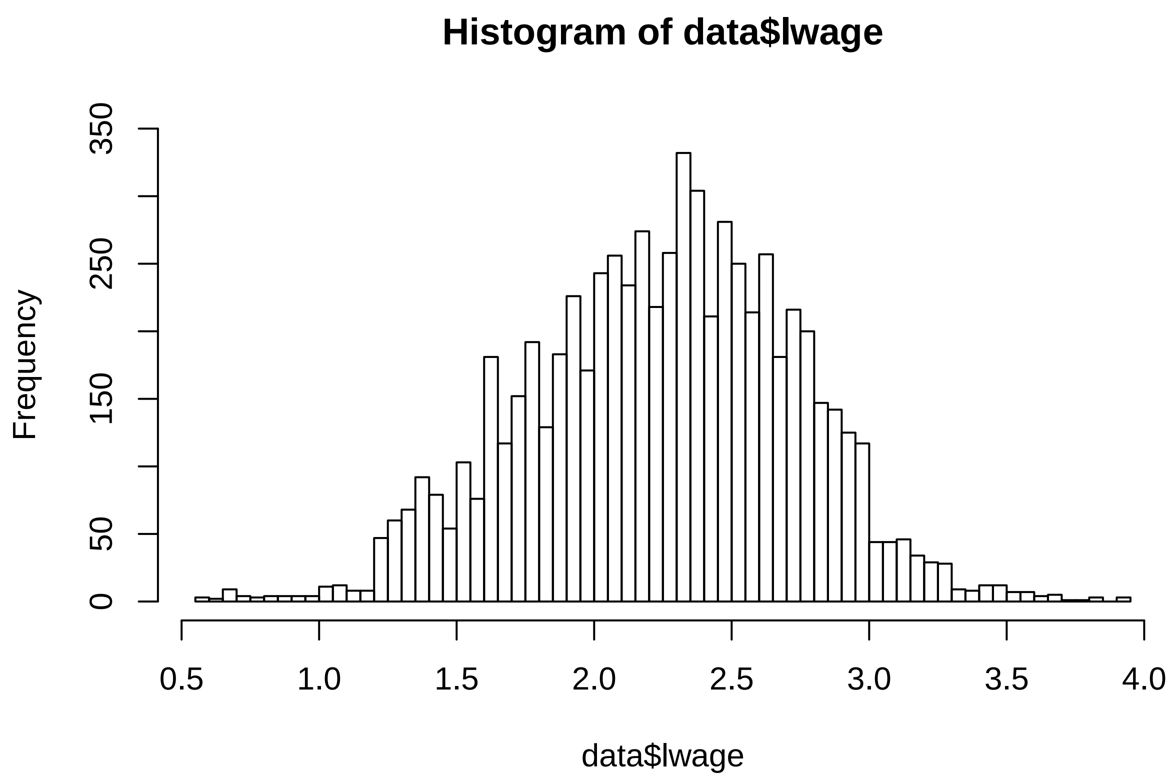
```r
summary(data$lwage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.5555  1.9250  2.2760  2.2480  2.5970  3.9120
```

```r
print(quantile(data$lwage, probs = c(0.01, 0.05, 0.1,
    0.25, 0.5, 0.75, 0.9, 0.95, 0.99, 1)))
```

```
##       1%       5%      10%      25%      50%      75%      90%      95%
## 1.148702 1.398129 1.609438 1.925291 2.276300 2.596916 2.851921 2.995732
##      99%     100%
## 3.325316 3.911953
```

```r
hist(data$lwage, 50, ylim = c(0, 350))
```



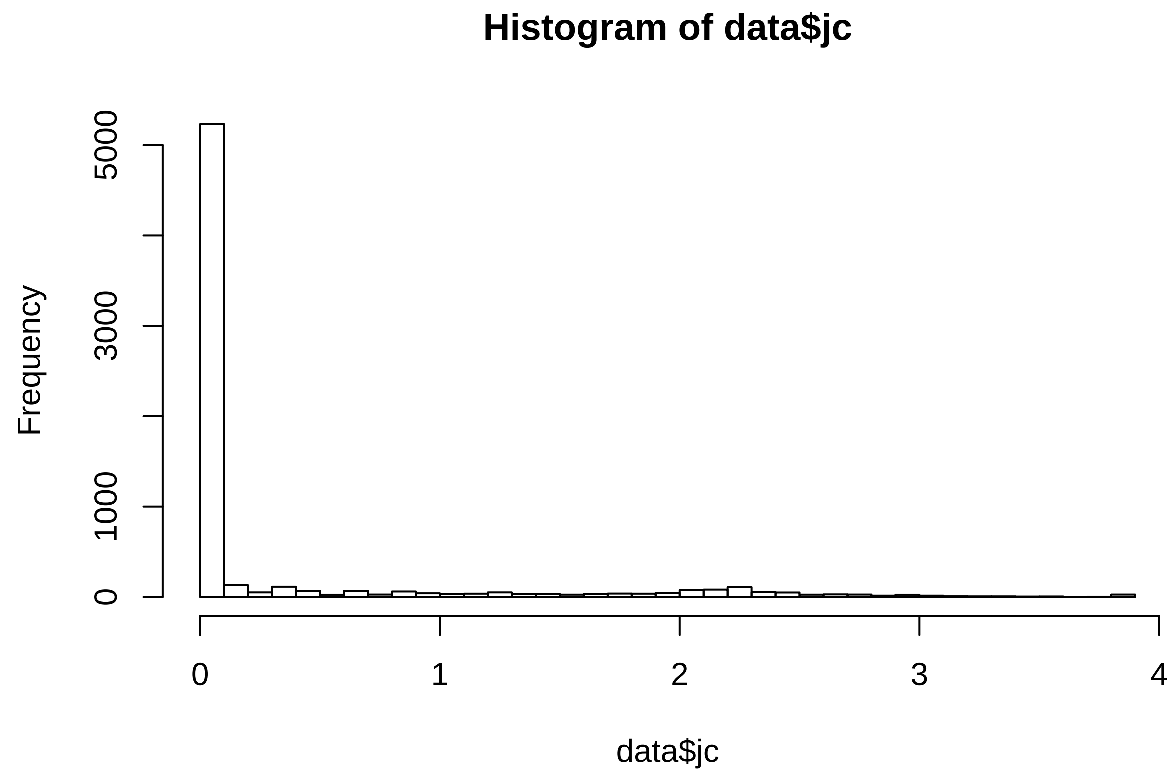

```r
summary(data$jc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3389  0.0000  3.8330
```

```r
print(quantile(data$jc, probs = c(0.01, 0.05, 0.1,
    0.25, 0.5, 0.75, 0.9, 0.95, 0.99, 1)))
```

```
##       1%       5%      10%      25%      50%      75%      90%      95%
## 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 1.766667 2.266667
##      99%     100%
## 3.089665 3.833333
```

```r
hist(data$jc, 50)
```



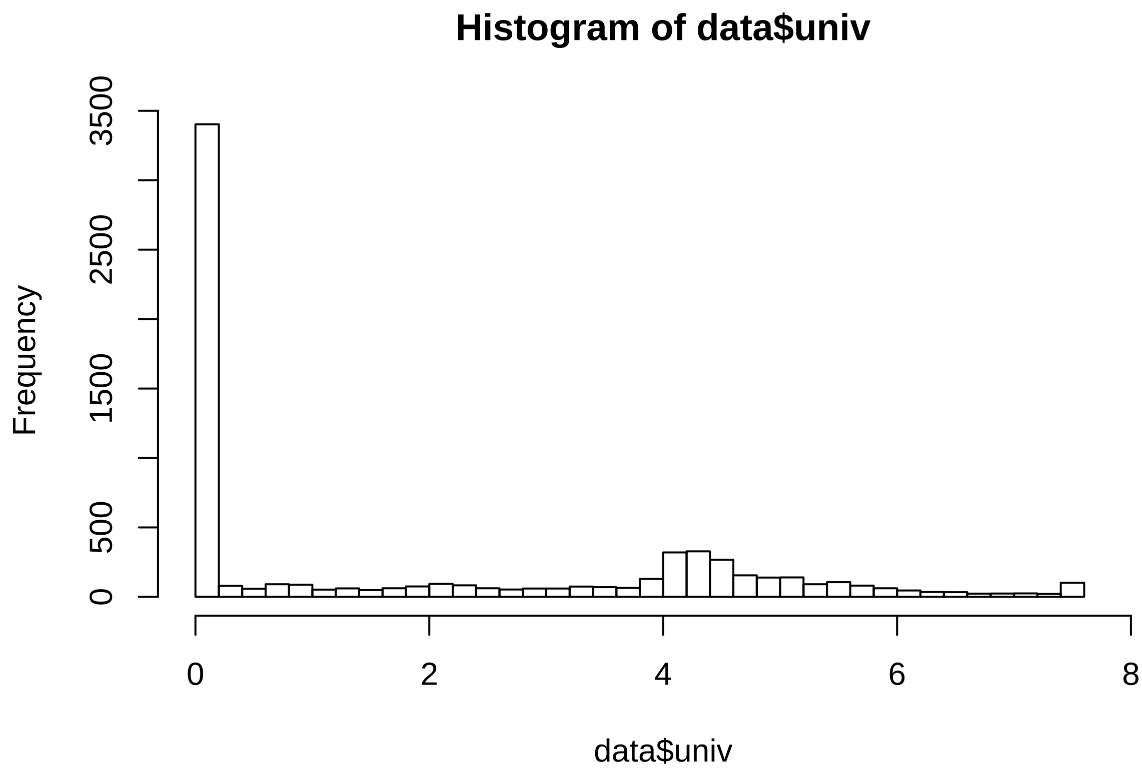

```r
summary(data$univ)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.200   1.926   4.200   7.500
```

```r
print(quantile(data$univ, probs = c(0.01, 0.05, 0.1,
    0.25, 0.5, 0.75, 0.9, 0.95, 0.99, 1)))
```

```
##        1%        5%       10%       25%       50%       75%       90%
## 0.0000000 0.0000000 0.0000000 0.0000000 0.1999997 4.1999998 5.1777687
##       95%       99%      100%
## 5.9099934 7.5000000 7.5000000
```

```r
hist(data$univ, 50, xlim = c(0, 8))
```



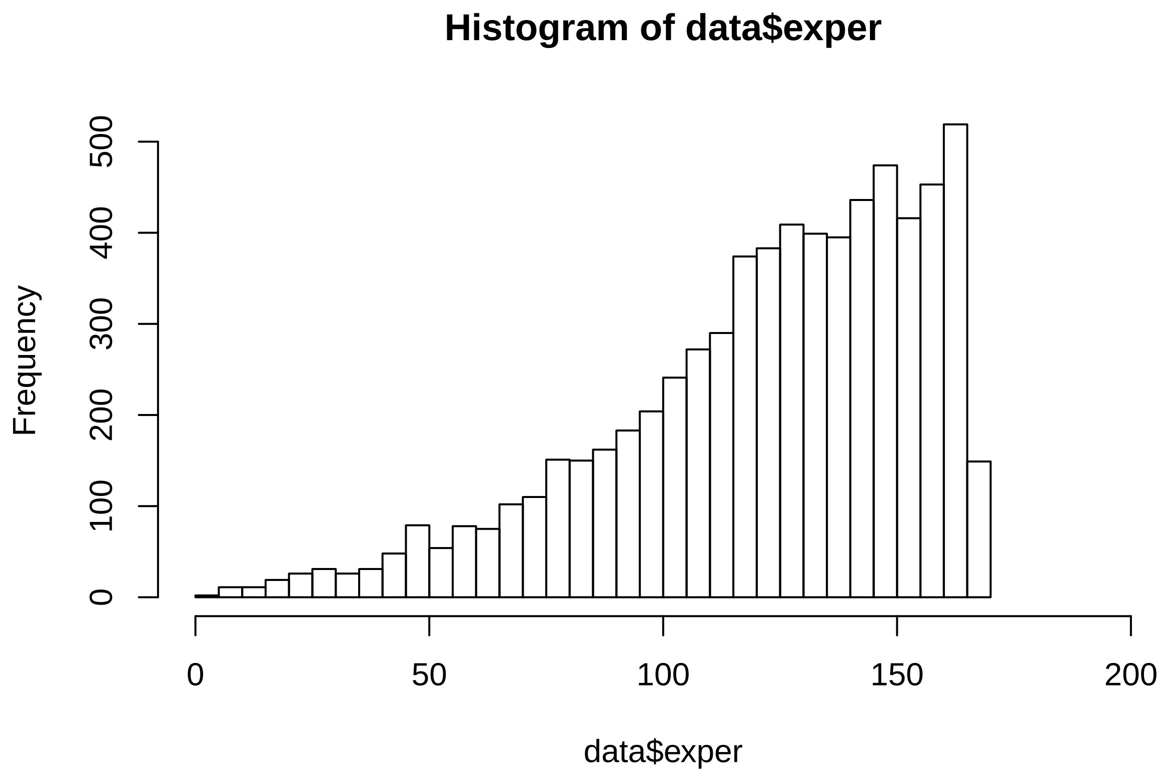

```r
summary(data$exper)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     3.0   104.0   129.0   122.4   149.0   166.0
```

```r
print(quantile(data$exper, probs = c(0.01, 0.05, 0.1,
    0.25, 0.5, 0.75, 0.9, 0.95, 0.99, 1)))
```

```
##   1%   5%  10%  25%  50%  75%  90%  95%  99% 100%
##   25   56   74  104  129  149  160  163  166  166
```

```r
hist(data$exper, 50, xlim = c(0, 200))
```



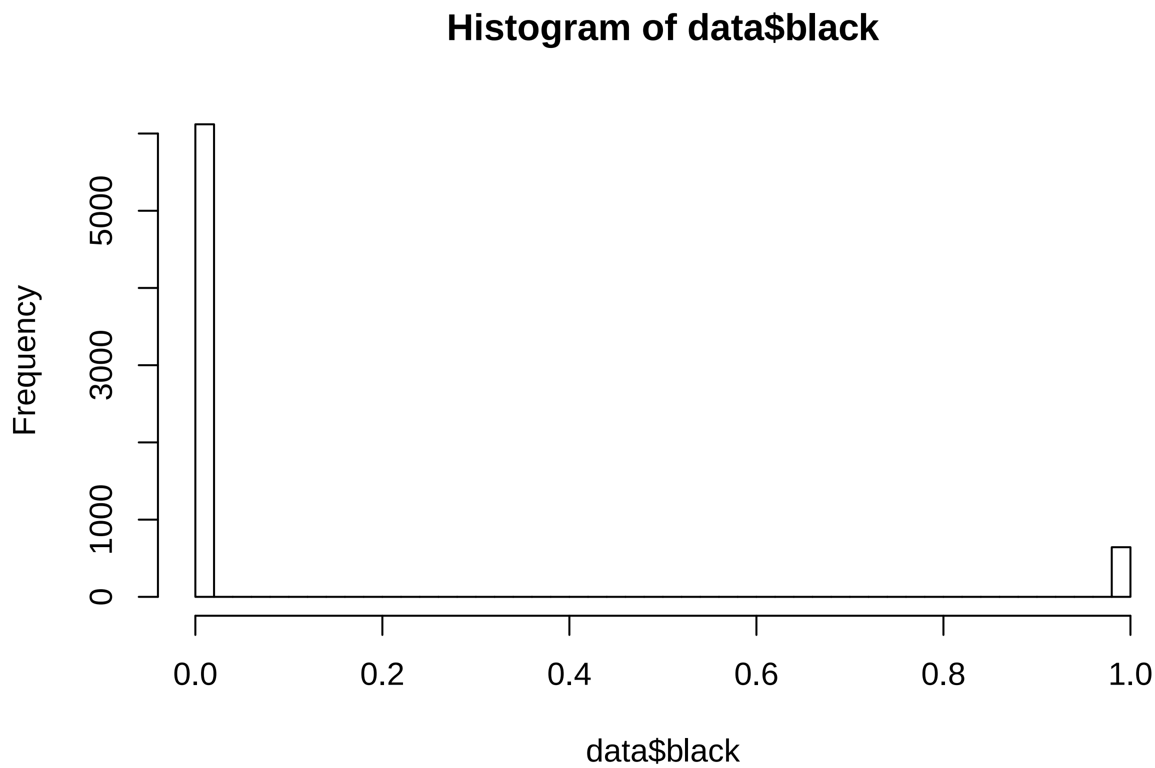

```r
summary(data$black)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.00000 0.09508 0.00000 1.00000
```

```r
print(quantile(data$black, probs = c(0.01, 0.05, 0.1,
    0.25, 0.5, 0.75, 0.9, 0.95, 0.99, 1)))
```

```
##   1%   5%  10%  25%  50%  75%  90%  95%  99% 100%
##    0    0    0    0    0    0    0    1    1    1
```

```r
hist(data$black, 50)
```



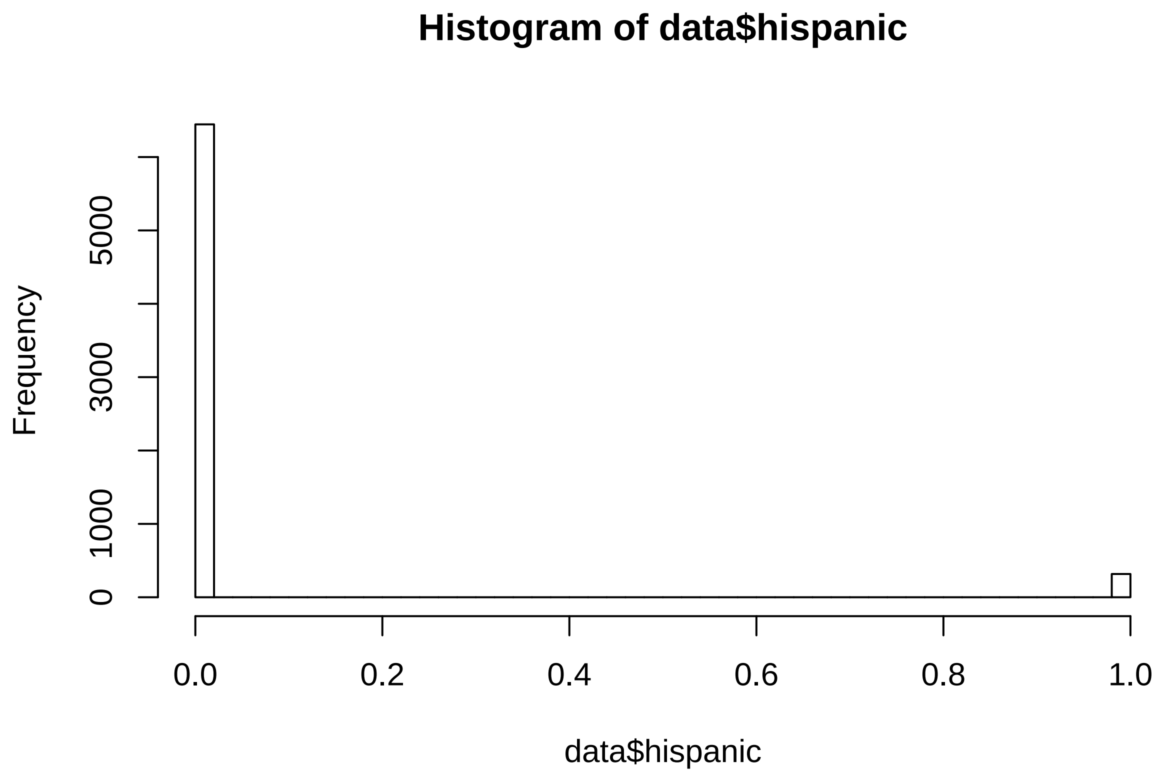

```r
summary(data$hispanic)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.00000 0.04687 0.00000 1.00000
```

```r
print(quantile(data$hispanic, probs = c(0.01, 0.05,
    0.1, 0.25, 0.5, 0.75, 0.9, 0.95, 0.99, 1)))
```

```
##   1%   5%  10%  25%  50%  75%  90%  95%  99% 100%
##    0    0    0    0    0    0    0    0    1    1
```

```r
hist(data$hispanic, 50)
```



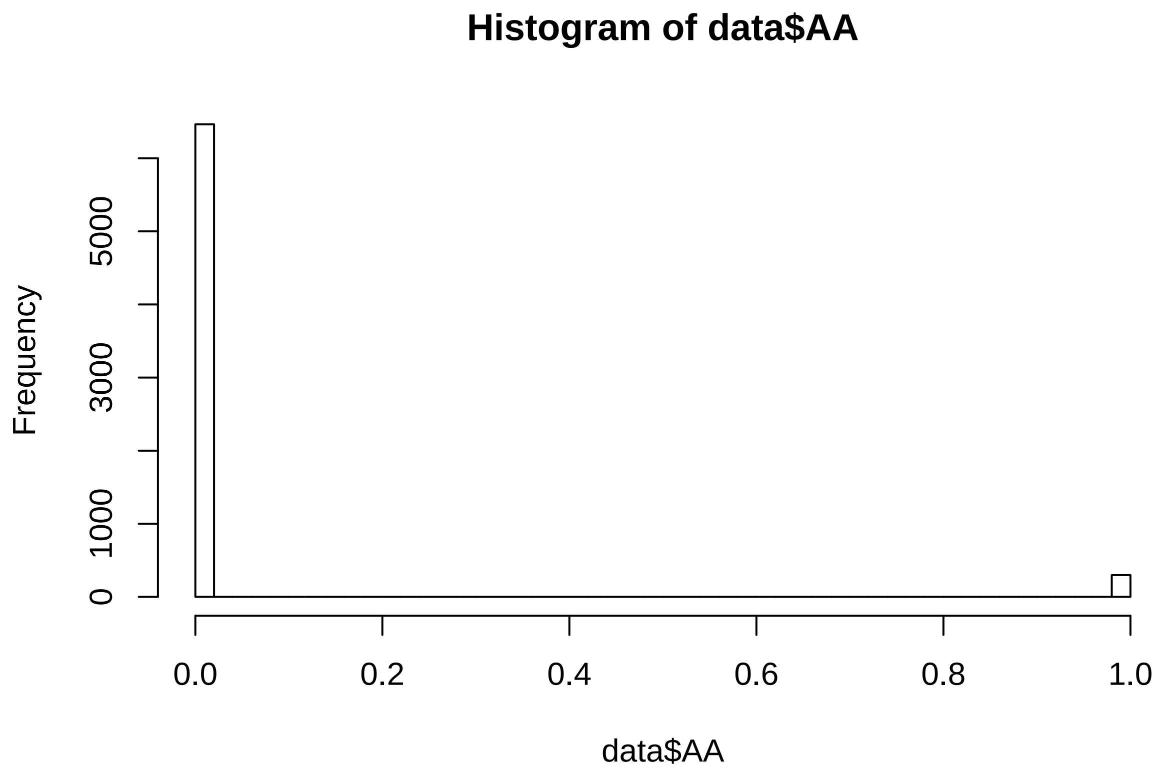

```r
summary(data$AA)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.00000 0.04406 0.00000 1.00000
```

```r
print(quantile(data$AA, probs = c(0.01, 0.05, 0.1,
    0.25, 0.5, 0.75, 0.9, 0.95, 0.99, 1)))
```

```
##   1%   5%  10%  25%  50%  75%  90%  95%  99% 100%
##    0    0    0    0    0    0    0    0    1    1
```

```r
hist(data$AA, 50)
```



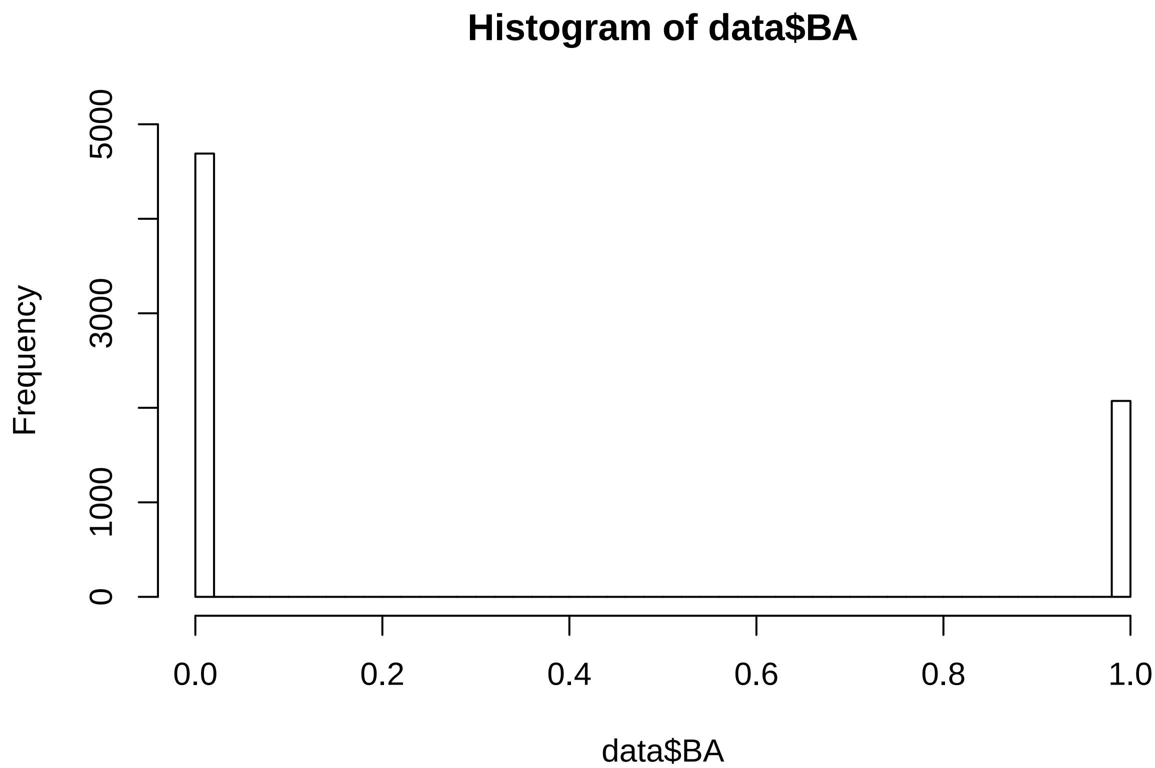

```
summary(data$BA)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3065  1.0000  1.0000
```

```
print(quantile(data$BA, probs = c(0.01, 0.05, 0.1,
    0.25, 0.5, 0.75, 0.9, 0.95, 0.99, 1)))
```

```
##   1%   5%  10%  25%  50%  75%  90%  95%  99% 100%
##    0    0    0    0    0    1    1    1    1    1
```

```
hist(data$BA, 50, ylim = c(0, 5000))
```



Basic structure of the data

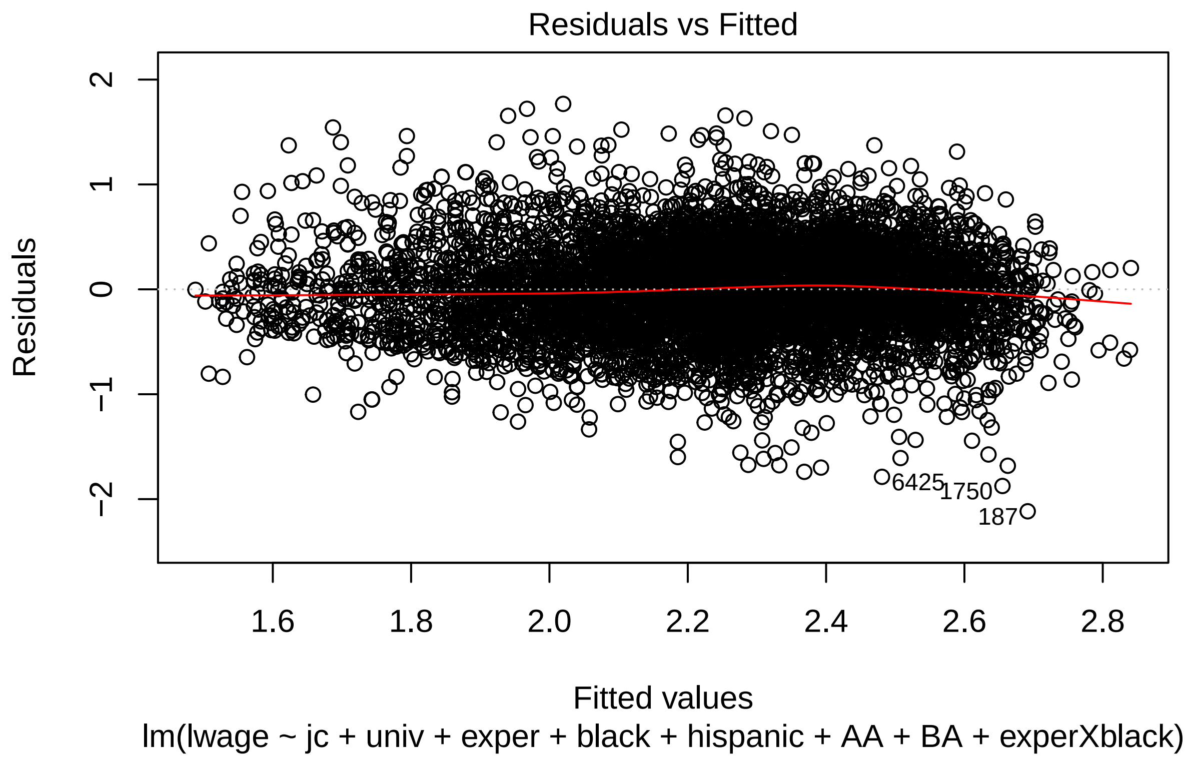
There are no missing values in the data.
lwage variable has a normal-like distribution.
jc variable has values from 0 to about 4 and is heavily positively skewed with a majority of values at or near 0.
univ variable has values from 0 to 7.5 and is heavily positively skewed with a majority of values at or near 0.
exper variable has values from 0 to 166 and is negatively skewed with a hill-climb distribution from 0 to about 500.
black, hispanic, AA, BA variables are binary with values of 0 or 1.

Question 2

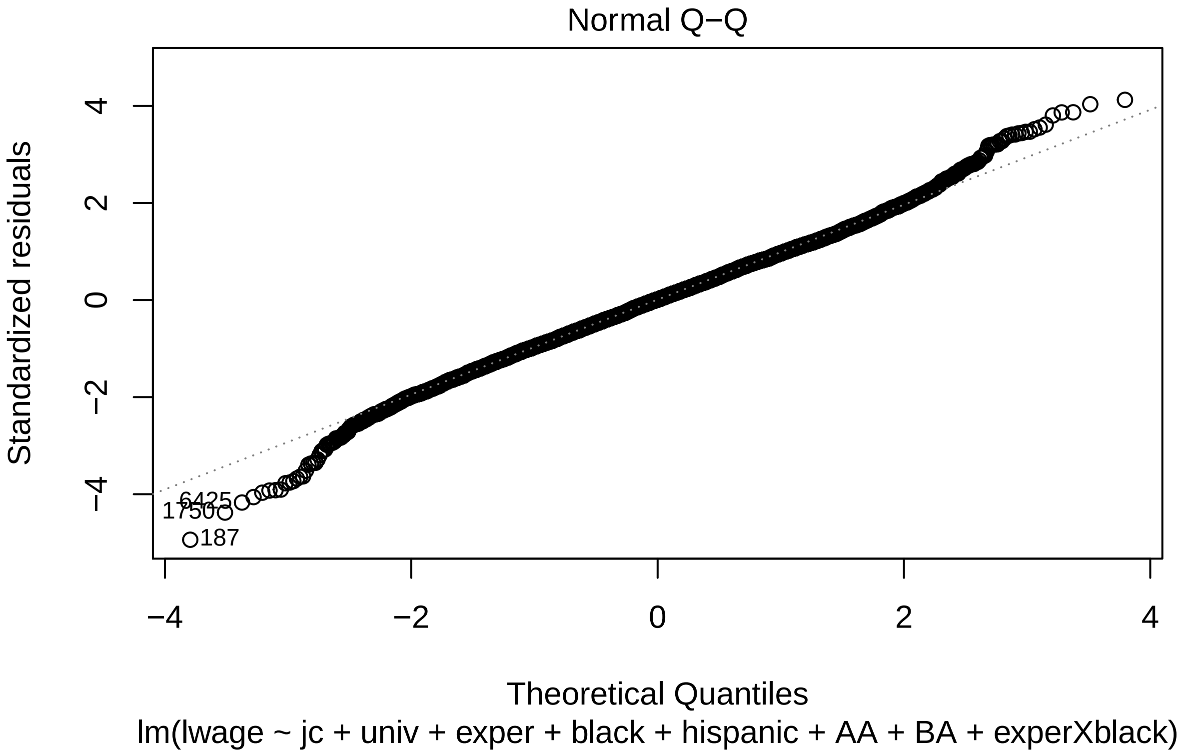
```r
# Create the experXblack variable by multiplying
# the exper and black variables.
data$experXblack = data$exper * data$black

# Run the requested OLS regression.
ols.lwage.8ind = lm(lwage ~ jc + univ + exper + black +
    hispanic + AA + BA + experXblack, data = data)
summary(ols.lwage.8ind)
```

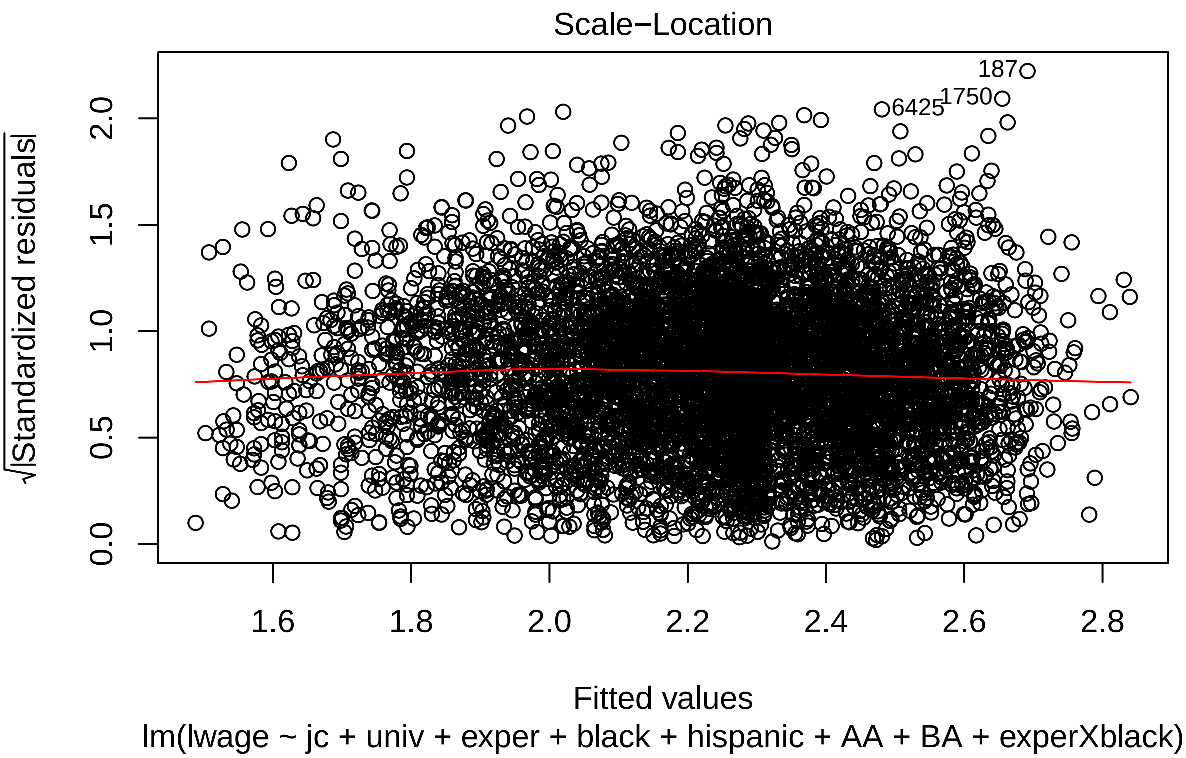
```
##
## Call:
## lm(formula = lwage ~ jc + univ + exper + black + hispanic + AA +
##     BA + experXblack, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.11612 -0.27836  0.00432  0.28676  1.76811
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.4773315  0.0223780  66.017  < 2e-16 ***
## jc           0.0637926  0.0079034   8.072 8.15e-16 ***
## univ         0.0732806  0.0031486  23.274  < 2e-16 ***
## exper        0.0050234  0.0001667  30.141  < 2e-16 ***
## black        0.0331709  0.0613984   0.540   0.5890
## hispanic    -0.0193629  0.0248914  -0.778   0.4367
## AA          -0.0077759  0.0295497  -0.263   0.7924
## BA           0.0176735  0.0156553   1.129   0.2590
## experXblack -0.0012679  0.0004991  -2.541   0.0111 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4287 on 6754 degrees of freedom
## Multiple R-squared:  0.2282, Adjusted R-squared:  0.2272
## F-statistic: 249.6 on 8 and 6754 DF,  p-value: < 2.2e-16
```

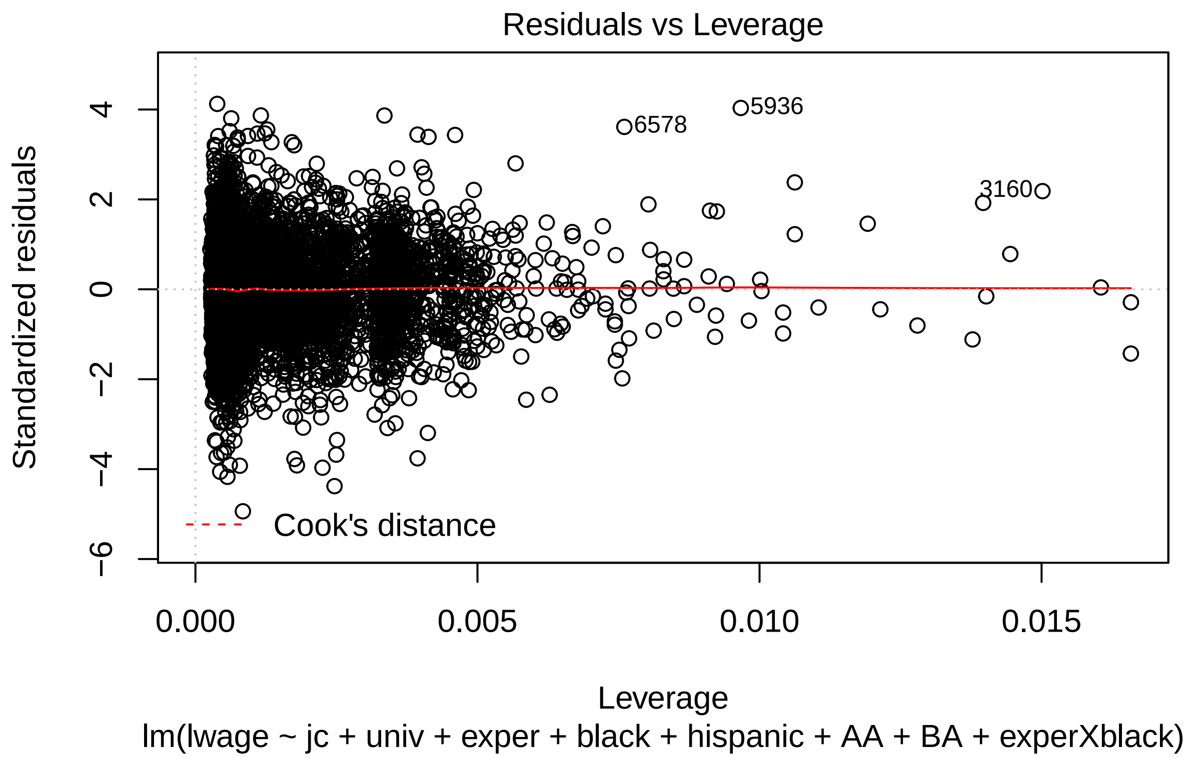
```r
# Print the diagnostic plots
plot(ols.lwage.8ind)
```



Residuals vs Fitted
Residuals
2
1
0
−1
−2
6425
1750
187
1.6
1.8
2.0
2.2
2.4
2.6
2.8
Fitted values
lm(lwage ~ jc + univ + exper + black + hispanic + AA + BA + experXblack)


Normal Q−Q
Standardized residuals
4
2
0
−2
−4
6425
1750
187
−4
−2
0
2
4
Theoretical Quantiles
lm(lwage ~ jc + univ + exper + black + hispanic + AA + BA + experXblack)


Scale−Location
187
6425
1750
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
1.6
1.8
2.0
2.2
2.4
2.6
2.8
Fitted values
lm(lwage ~ jc + univ + exper + black + hispanic + AA + BA + experXblack)




```r
# Print the B_hat4 and B_hat8 coefficients
print(ols.lwage.8ind$coefficients["black"])
```

```
##      black
## 0.03317088
```

```r
print(ols.lwage.8ind$coefficients["experXblack"])
```

```
##  experXblack
## -0.001267898
```

Interpret the coefficients $\hat{\beta}4$ and $\hat{\beta}8$

$\hat{\beta}4$ is the estimate for the black variable coefficient.
$\hat{\beta}8$ is the estimate for the experXblack variable.

The β_4 coefficient captures the effect of being black on the log of wage, holding all other variables in the model fix and assuming zero-employment experience. It provides an indication of how much percentage point change to expect in the wage of an individual when they move from the baseline (non-black) to being black. Hence the coefficient captures the difference at the intercept of the log(wage) vs experience plot between black and white respondants.

The β_8 coefficient captures the effect of being black on the impact of experience over the years on wage. It is better explained in terms of derivatives with:

$$\frac{\delta log(wage)}{\delta exper} = \beta_3 + \beta_8 * black$$

In the previous formulation, we can see that the β_8 coefficient captures the impact of being back on the slope of the log(wage) vs experience curve. In other words, the coefficient describes how much more or less experience impacts the log of wage for black people over the years, vs the baseline (non-black people). And because the outcome variable is the log of wage, the impact above can actually be formulated in terms of impact of ethnicity on percentage changes on the actual wage of individuals.

Question 3

```r
# Show the summary of the model again
summary(ols.lwage.8ind)
```

```
##
## Call:
## lm(formula = lwage ~ jc + univ + exper + black + hispanic + AA +
##     BA + experXblack, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.11612 -0.27836  0.00432  0.28676  1.76811
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.4773315  0.0223780  66.017  < 2e-16 ***
## jc           0.0637926  0.0079034   8.072 8.15e-16 ***
## univ         0.0732806  0.0031486  23.274  < 2e-16 ***
## exper        0.0050234  0.0001667  30.141  < 2e-16 ***
## black        0.0331709  0.0613984   0.540   0.5890
## hispanic    -0.0193629  0.0248914  -0.778   0.4367
## AA          -0.0077759  0.0295497  -0.263   0.7924
## BA           0.0176735  0.0156553   1.129   0.2590
## experXblack -0.0012679  0.0004991  -2.541   0.0111 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4287 on 6754 degrees of freedom
## Multiple R-squared:  0.2282, Adjusted R-squared:  0.2272
## F-statistic: 249.6 on 8 and 6754 DF,  p-value: < 2.2e-16
```

```r
# Print the univ coefficient
print(ols.lwage.8ind$coefficients["univ"])
```

```
##       univ
## 0.07328063
```

Test that the return to university education is 7%.

Null Hypothesis: H0: $\beta 2 = 0.07$.
Alternate Hypothesis: H1: $\beta 2 \neq 0.07$.
Using the linear model and summary statistics, we compute the pvalue for the test above as:

```
pvalue = 2 * (1 - pt((ols.lwage.8ind$coefficients["univ"] -
    0.07)/coef(summary(ols.lwage.8ind))[, "Std. Error"]["univ"],
    df = summary(ols.lwage.8ind)$df[1]))
pvalue
```

```
##      univ
## 0.3246284
```

Based on the p-value, the test is not significant at the 0.05% significance level. Therefore, we can't reject the null hypothesis that the return to university education is 7%.

Question 4

Test that the return to junior college education is equal for black and non-black

```
# Create the jc times black interaction variable
data$jcXblack = data$jc * data$black
# Re-run the regression with the new interaction
# variable added
ols.lwage.jcblack.diff = lm(lwage ~ jc + univ + exper +
    black + jcXblack + hispanic + AA + BA + experXblack,
    data = data)
# Show the summary of the model
summary(ols.lwage.jcblack.diff)
```

```
##
## Call:
## lm(formula = lwage ~ jc + univ + exper + black + jcXblack + hispanic +
##     AA + BA + experXblack, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.11547 -0.27839  0.00394  0.28669  1.76883
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.4767425  0.0223831  65.976  < 2e-16 ***
## jc           0.0659081  0.0081083   8.128 5.13e-16 ***
## univ         0.0733407  0.0031490  23.290  < 2e-16 ***
## exper        0.0050222  0.0001667  30.134  < 2e-16 ***
## black        0.0428709  0.0619565   0.692    0.489
## jcXblack    -0.0337383  0.0289025  -1.167    0.243
## hispanic    -0.0194598  0.0248909  -0.782    0.434
## AA          -0.0087614  0.0295610  -0.296    0.767
```

```
## BA           0.0174258  0.0156563   1.113    0.266
## experXblack -0.0012865  0.0004993  -2.577    0.010 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4287 on 6753 degrees of freedom
## Multiple R-squared:  0.2283, Adjusted R-squared:  0.2273
## F-statistic:   222 on 9 and 6753 DF,  p-value: < 2.2e-16
```

Create the interaction variable jcXblack and add it to the model as follows:

$$lwage = \beta0+\beta1jc+\beta2univ+\beta3exper+\beta4black+\beta5jcXblack+\beta6hispanic+\beta7AA+\beta8BA+\beta9experXblack+\epsilon$$

The coefficient for the jcXblack interaction variable ($\beta5$) now represents the difference in return to junior college between black and non-black, so we can set up our test as follows:
Null Hypothesis: H0: $\beta5 = 0$.
Alternate Hypothesis: H1: $\beta5 \neq 0$.
Based on the p-value of 0.243 for $\beta5$, the test is not significant at the 0.05% significance level. Therefore, we can't reject the null hypothesis that $\beta5 = 0$. Said another way we cannot reject the null hypothesis that the return to junior college education is equal for black and non-black.

Or alternatively, intuitively, we can see from the linear model, we derive: $\frac{\delta log(wage)}{\delta jc} = \beta_1$
The model is specified in a way that the return to junior college education, is β_1, and is independent of ethnicity. Therefore without additional computation, we can immediately answer that the return on junior college education is the same for all ethnicities.

Question 5

Test whether the return to university education is equal to the return to 1 year of working experience.

Original model:

$$lwage = \beta0 + \beta1jc + \beta2univ + \beta3exper + \beta4black + \beta5hispanic + \beta6AA + \beta7BA + \beta8experXblack + \epsilon$$

Convert the experience variable from months to years by creating a new variable experYr that divides the original variable exper by 12. Replace the exper variable in the original model with this variable.

$$lwage = \beta0 + \beta1jc + \beta2univ + \beta3experYr + \beta4black + \beta5hispanic + \beta6AA + \beta7BA + \beta8experXblack + \epsilon$$

We would like to know if the $\beta2$ and $\beta3$ coefficients are the same or, equivalently, if their difference is 0. We can define a variable θ such that $\theta = \beta2 - \beta3$ and rewrite our model like this:

$$lwage = \beta0+\beta1jc+(\theta+\beta3)univ+\beta3experYr+\beta4black+\beta5hispanic+\beta6AA+\beta7BA+\beta8experXblack+\epsilon$$

Rewrite the model to get θ by itself as a coefficient:

$$lwage = \beta0+\beta1jc+\theta univ+\beta3(univ+experYr)+\beta4black+\beta5hispanic+\beta6AA+\beta7BA+\beta8experXblack+\epsilon$$

Now our null hypothesis is $H0 : \theta = 0$.
Alternate Hypothesis: H1: $\theta \neq 0$.

```r
# Convert the exper variable from months to years
# by dividing it by 12.
data$experYr = data$exper/12
# Create a variable that is the sum of the univ and
# experYr variables
data$univ_plus_experYr = data$univ + data$experYr
# Rerun the regression with the new variables.
ols.lwage.univ.experYr = lm(lwage ~ jc + univ + univ_plus_experYr +
    black + hispanic + AA + BA + experXblack, data = data)
# Display a summary of the new model
summary(ols.lwage.univ.experYr)
```

```
##
## Call:
## lm(formula = lwage ~ jc + univ + univ_plus_experYr + black +
##     hispanic + AA + BA + experXblack, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.11612 -0.27836  0.00432  0.28676  1.76811
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.4773315  0.0223780  66.017  < 2e-16 ***
## jc                 0.0637926  0.0079034   8.072 8.15e-16 ***
## univ               0.0129997  0.0035721   3.639 0.000276 ***
## univ_plus_experYr  0.0602810  0.0020000  30.141  < 2e-16 ***
## black              0.0331709  0.0613984   0.540 0.589038
## hispanic          -0.0193629  0.0248914  -0.778 0.436659
## AA                -0.0077759  0.0295497  -0.263 0.792446
## BA                 0.0176735  0.0156553   1.129 0.258972
## experXblack       -0.0012679  0.0004991  -2.541 0.011088 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4287 on 6754 degrees of freedom
## Multiple R-squared:  0.2282, Adjusted R-squared:  0.2272
## F-statistic: 249.6 on 8 and 6754 DF,  p-value: < 2.2e-16
```

Based on the very low p-value (0.000276) for θ, the test is significant at the 0.05% significance level. And even though the value of θ is close to 0 at 0.0129997, we can reject the null hypothesis that $\theta = 0$.

Question 6

```r
# Show the summary of the model again
summary(ols.lwage.8ind)
```

```
##
## Call:
## lm(formula = lwage ~ jc + univ + exper + black + hispanic + AA +
```

```
##     BA + experXblack, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.11612 -0.27836  0.00432  0.28676  1.76811
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.4773315  0.0223780  66.017  < 2e-16 ***
## jc           0.0637926  0.0079034   8.072 8.15e-16 ***
## univ         0.0732806  0.0031486  23.274  < 2e-16 ***
## exper        0.0050234  0.0001667  30.141  < 2e-16 ***
## black        0.0331709  0.0613984   0.540   0.5890
## hispanic    -0.0193629  0.0248914  -0.778   0.4367
## AA          -0.0077759  0.0295497  -0.263   0.7924
## BA           0.0176735  0.0156553   1.129   0.2590
## experXblack -0.0012679  0.0004991  -2.541   0.0111 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4287 on 6754 degrees of freedom
## Multiple R-squared:  0.2282, Adjusted R-squared:  0.2272
## F-statistic: 249.6 on 8 and 6754 DF,  p-value: < 2.2e-16
```

Test the overall significance of this regression.

We are testing the overall significance of the original model as stated below:

$$lwage = \beta 0 + \beta 1jc + \beta 2univ + \beta 3exper + \beta 4black + \beta 5hispanic + \beta 6AA + \beta 7BA + \beta 8experXblack + \epsilon$$

Here is the output from the summary of the original model.
Residual standard error: 0.4287 on 6754 degrees of freedom
Multiple R-squared: 0.2282, Adjusted R-squared: 0.2272
F-statistic: 249.6 on 8 and 6754 DF, p-value: < 2.2e-16

1. Our model null hypothesis is that there is no relationship among any of the independent variables and lwage variable. We are able to reject the null hypothesis since our p-value of the f-statistic of the model is significant at < 2.2e-16.
2. Practical significance: we have an R-squared value of 0.2282, indicating that 22.82% of the variation in lwage is explained by our model.

Question 7

```
# Define a square term for the exper variable
data$experXexper = data$exper * data$exper
# Add the new variable to the regression
ols.lwage.9ind = lm(lwage ~ jc + univ + exper + black +
    hispanic + AA + BA + experXblack + experXexper,
```

```r
    data = data)
# Show the summary of the model
summary(ols.lwage.9ind)
```

```
##
## Call:
## lm(formula = lwage ~ jc + univ + exper + black + hispanic + AA +
##     BA + experXblack + experXexper, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.11982 -0.27743  0.00475  0.28741  1.77397
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.510e+00  4.427e-02  34.108  < 2e-16 ***
## jc           6.417e-02  7.916e-03   8.106 6.14e-16 ***
## univ         7.382e-02  3.211e-03  22.992  < 2e-16 ***
## exper        4.301e-03  8.588e-04   5.008 5.64e-07 ***
## black        2.994e-02  6.152e-02   0.487   0.6265
## hispanic    -1.932e-02  2.489e-02  -0.776   0.4378
## AA          -7.539e-03  2.955e-02  -0.255   0.7986
## BA           1.797e-02  1.566e-02   1.147   0.2513
## experXblack -1.239e-03  5.002e-04  -2.477   0.0133 *
## experXexper  3.379e-06  3.939e-06   0.858   0.3911
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4287 on 6753 degrees of freedom
## Multiple R-squared:  0.2282, Adjusted R-squared:  0.2272
## F-statistic: 221.9 on 9 and 6753 DF,  p-value: < 2.2e-16
```

Estimated return to work experience in this model

$lwage = \beta0 + \beta1 jc + \beta2 univ + \beta3 exper + \beta4 black + \beta5 hispanic + \beta6 AA + \beta7 BA + \beta8 exper X black + \beta9 exper X exper$

We obtain the return on a year of experience by evaluating:

$$\frac{\delta lwage}{\delta exper} = \beta3 + \beta8 black + 2 * \beta9 exper$$

$\beta3 = 4.301e - 03 = .004301$, $\beta8 = .001239$, $\beta9 = 000003379$
Substituting these values in the the equation above we get:

$$\frac{\delta lwage}{\delta exper} = (.004301 - .001239 * black + .000006758 * exper)$$

Thus for blacks, the return to one year of working experience is:

```r
.004301 - .001239 + .000006758
```

```
## [1] 0.003068758
```

We interpret is as 3/10th of a percent of increase in wage per year of experience.

And for non-blacks, that return is:

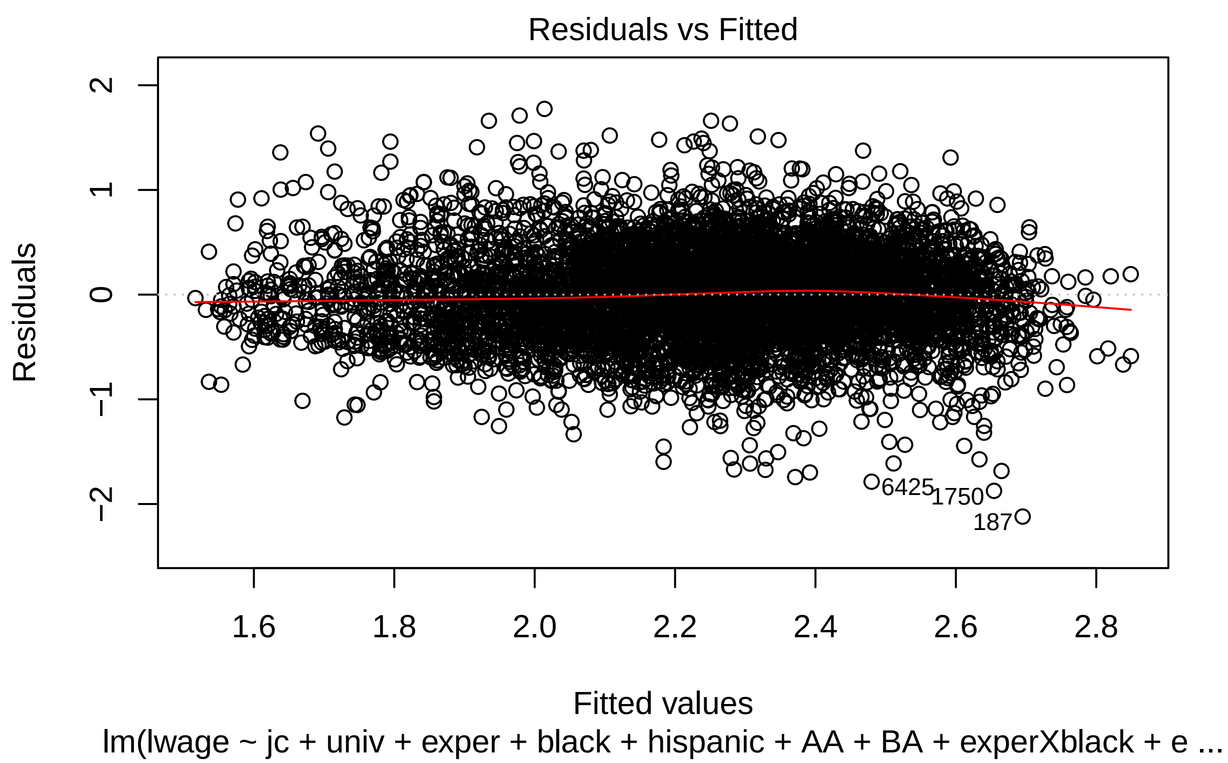
```
.004301 + .000006758
```

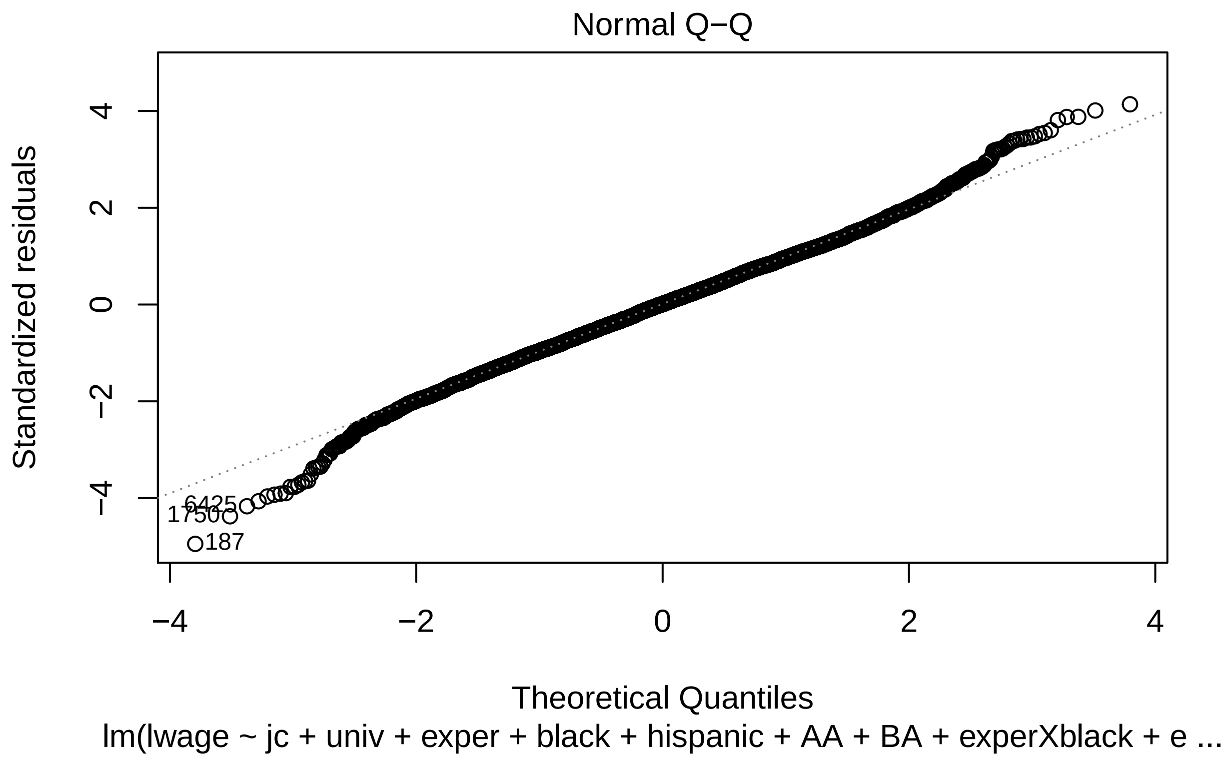
```
## [1] 0.004307758
```

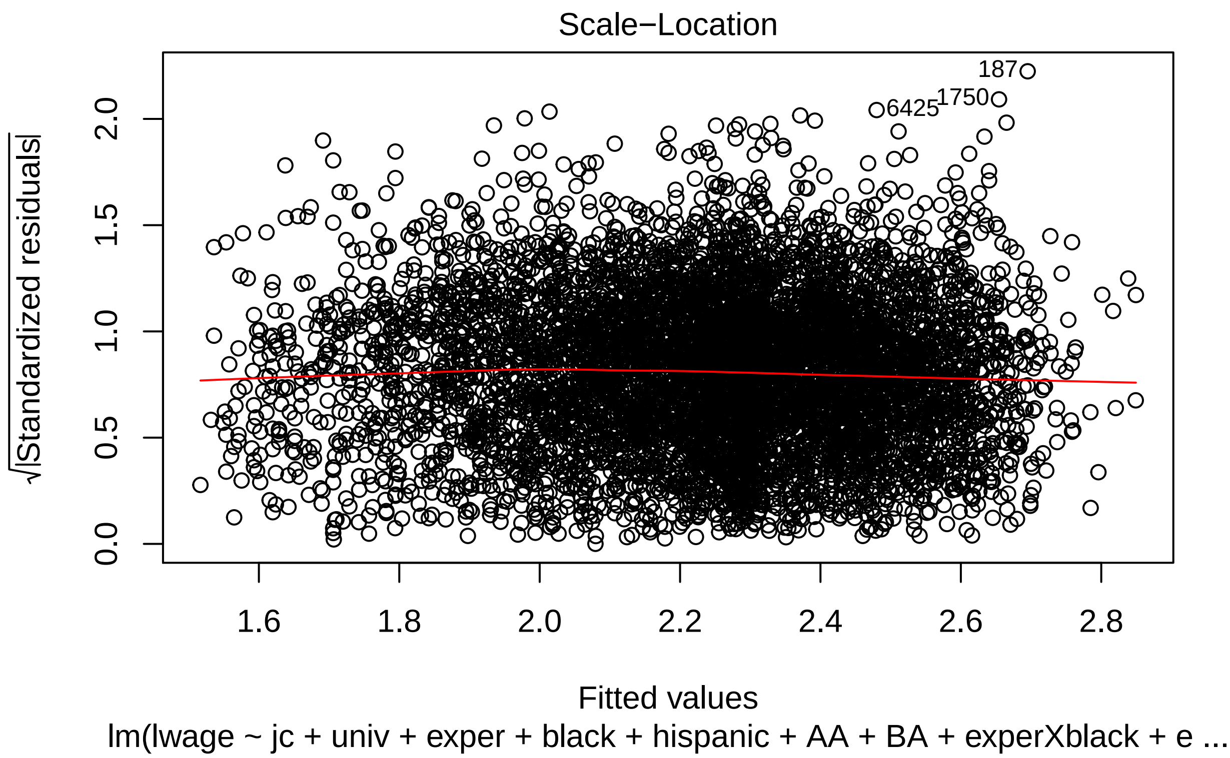
We interpret is as 4.3/10th of a percent of increase in wage per year of experience.

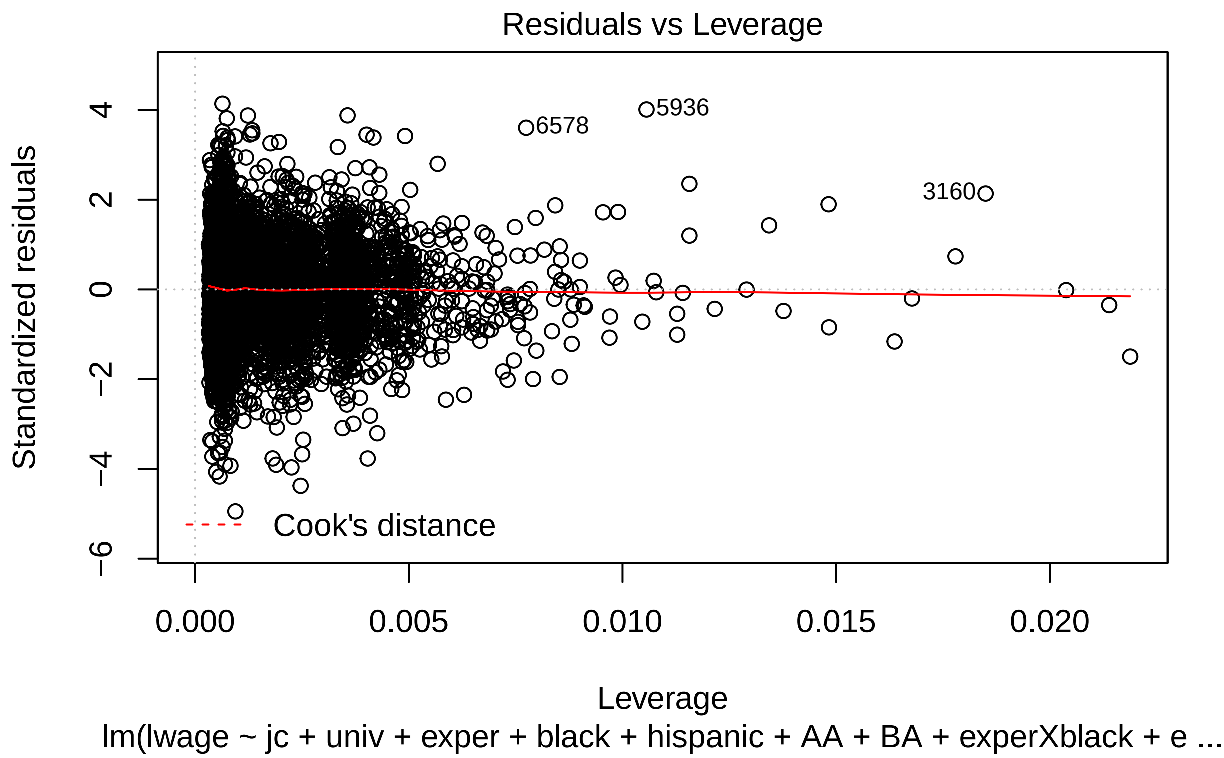
Question 8

```
# Plot the graphs from the model
plot(ols.lwage.9ind)
```




Normal Q−Q
Standardized residuals
4
2
0
−2
−4
6425
1750
187
−4
−2
0
2
4
Theoretical Quantiles
lm(lwage ~ jc + univ + exper + black + hispanic + AA + BA + experXblack + e ...


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
187
6425
1750
1.6
1.8
2.0
2.2
2.4
2.6
2.8
Fitted values
lm(lwage ~ jc + univ + exper + black + hispanic + AA + BA + experXblack + e ...




```r
# Show the summary of the model
summary(ols.lwage.9ind)
```

```
##
## Call:
## lm(formula = lwage ~ jc + univ + exper + black + hispanic + AA +
##     BA + experXblack + experXexper, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.11982 -0.27743  0.00475  0.28741  1.77397
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.510e+00  4.427e-02  34.108  < 2e-16 ***
## jc           6.417e-02  7.916e-03   8.106 6.14e-16 ***
## univ         7.382e-02  3.211e-03  22.992  < 2e-16 ***
## exper        4.301e-03  8.588e-04   5.008 5.64e-07 ***
## black        2.994e-02  6.152e-02   0.487   0.6265
## hispanic    -1.932e-02  2.489e-02  -0.776   0.4378
## AA          -7.539e-03  2.955e-02  -0.255   0.7986
## BA           1.797e-02  1.566e-02   1.147   0.2513
## experXblack -1.239e-03  5.002e-04  -2.477   0.0133 *
## experXexper 3.379e-06  3.939e-06   0.858   0.3911
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.4287 on 6753 degrees of freedom
## Multiple R-squared:  0.2282, Adjusted R-squared:  0.2272
## F-statistic: 221.9 on 9 and 6753 DF,  p-value: < 2.2e-16
```

```r
# Use the robust standard errors
coeftest(ols.lwage.9ind, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                 Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  1.5101e+00  4.3591e-02 34.6427 < 2.2e-16 ***
## jc           6.4168e-02  7.6224e-03  8.4183 < 2.2e-16 ***
## univ         7.3819e-02  3.4501e-03 21.3963 < 2.2e-16 ***
## exper        4.3008e-03  8.4541e-04  5.0873 3.731e-07 ***
## black        2.9937e-02  6.8436e-02  0.4374   0.66180
## hispanic    -1.9317e-02  2.4985e-02 -0.7731   0.43947
## AA          -7.5392e-03  2.7481e-02 -0.2743   0.78383
## BA           1.7967e-02  1.6579e-02  1.0837   0.27853
## experXblack -1.2388e-03  5.3539e-04 -2.3139   0.02071 *
## experXexper  3.3790e-06  3.8745e-06  0.8721   0.38318
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Homoskedasticity analysis:

We are testing homoskedasticity of the model with the quadratic experience term, expressed as:

$$lwage = \beta0+\beta1jc+\beta2univ+\beta3exper+\beta4black+\beta5hispanic+\beta6AA+\beta7BA+\beta8experXblack+\beta9experXexper+\epsilon$$

We observe a small amount of heteroskedasticity from the plot:
1 - We can see from the residuals vs fitted plot that the variance band changes very slightly as we move to higher fitted values.
2 - The same story is told by the scale-location plot where we see that the smoothing line is not quite completely horizontal. Therefore we conclude a very small amount of heteroskedasticity and decide to use heteroskedasticity-robust methods for coefficient estimation.
3 - We do not look at the Breusch Pagan test since we have a large number of observations, therefore we know almost certainly that we will obtain significance.

The implication of heteroskedastcity (even small) in the data is that the standard error of the univ coefficient ($\beta2$) may be biased. Biased standard errors can impact the outcomes of statistical tests. Therefore, it can affect the testing of no effect of university education on salary, which is the t-test on the coefficient β_2

The $\beta2$ coefficient from the robust method was essentially unchanged, going from a value of $\beta2 = 7.382e-02$ using the non-robust estimation to $\beta2 = 7.3819e-02$ with robust estimation. The standard error of the $\beta2$ coefficient was changed, going from 3.211e-03 using the non-robust estimation to 3.4501e-03 using the robust estimation. However, even using the robust estimation, the p-value for our $\beta2$ coefficient remains significant at the 0.05, and we confirm that we must reject the null hypothesis that the coefficient value is null, with the meaning of that null hypothesis being that there is no relationship between time at university and wages.