Homework 3

Rohan Thakur, Charles Kekeh and Megan Jasek

February 13, 2016

```
# 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

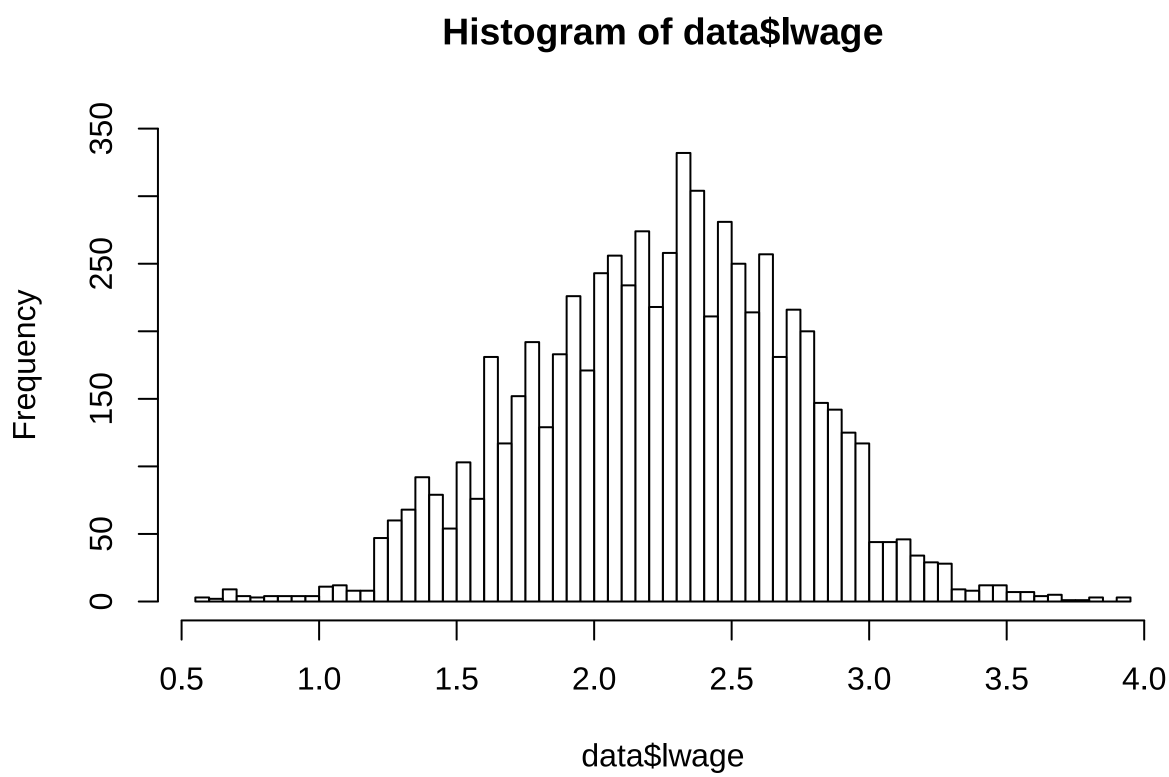
```
summary(data$lwage)
```

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

```
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))
```



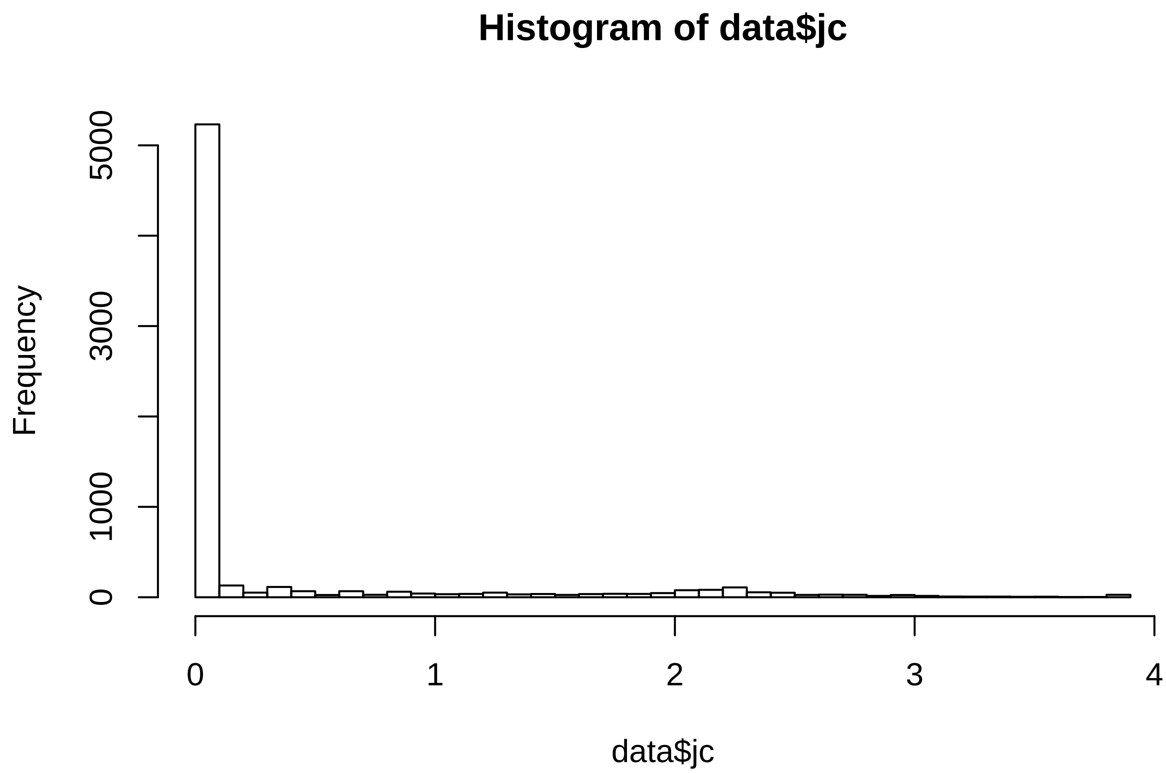

```r
summary(data$jc)
```

```
##    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)
```



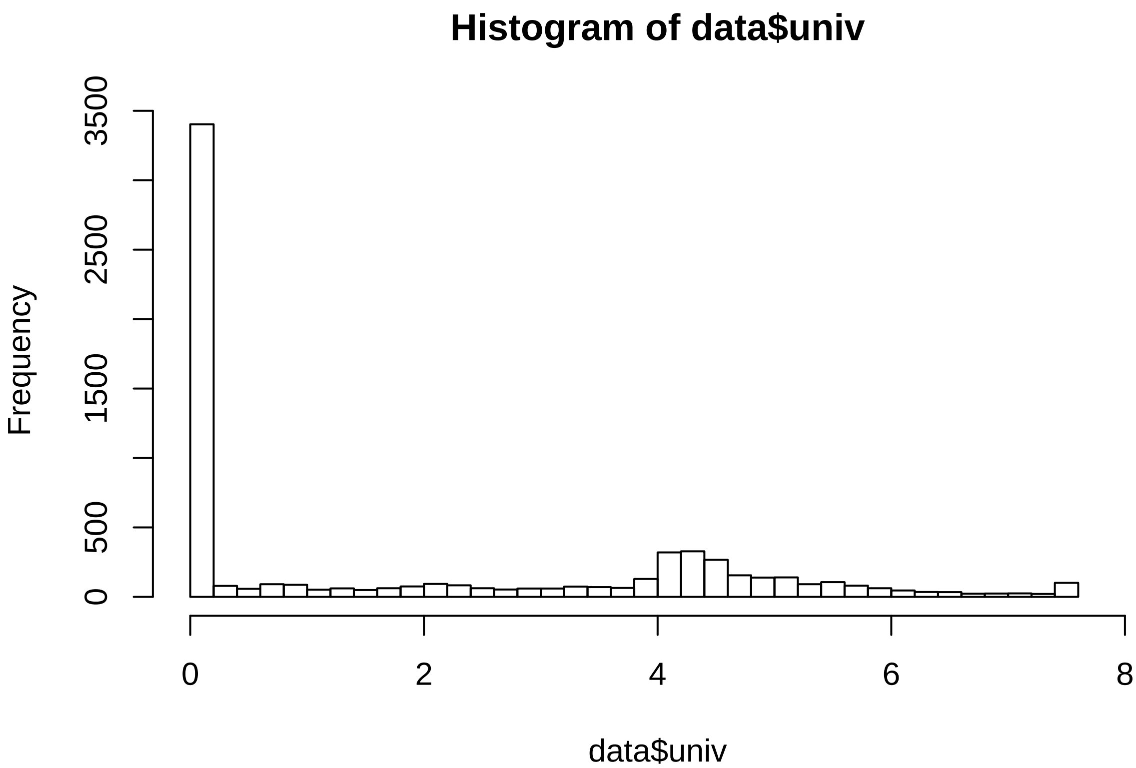

```r
summary(data$univ)
```

```
##    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))
```



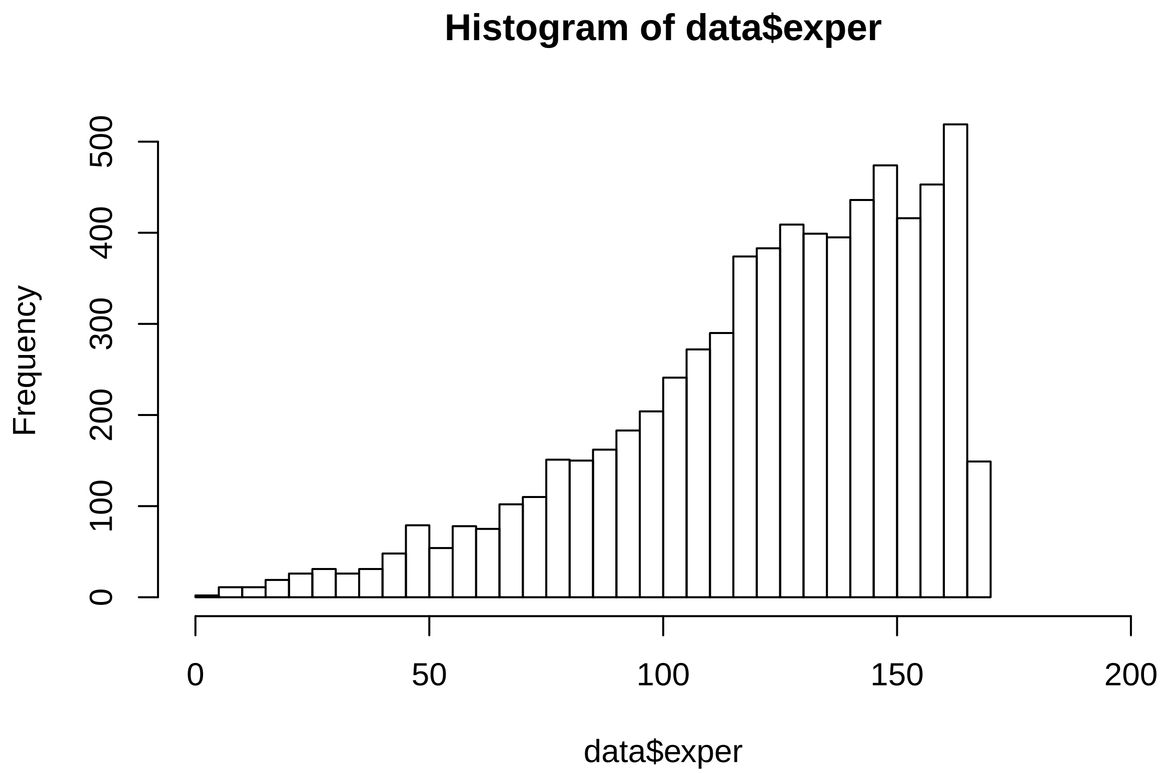

```r
summary(data$exper)
```

```
##    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))
```



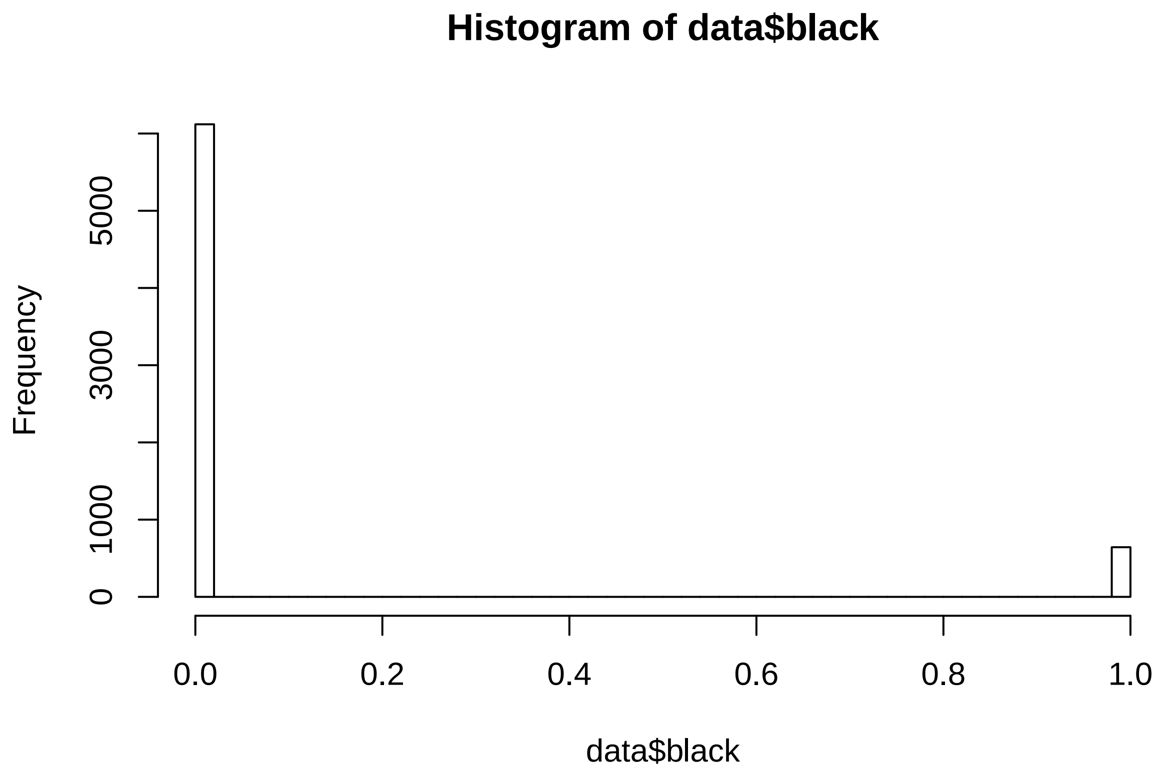

```r
summary(data$black)
```

```
##    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)
```



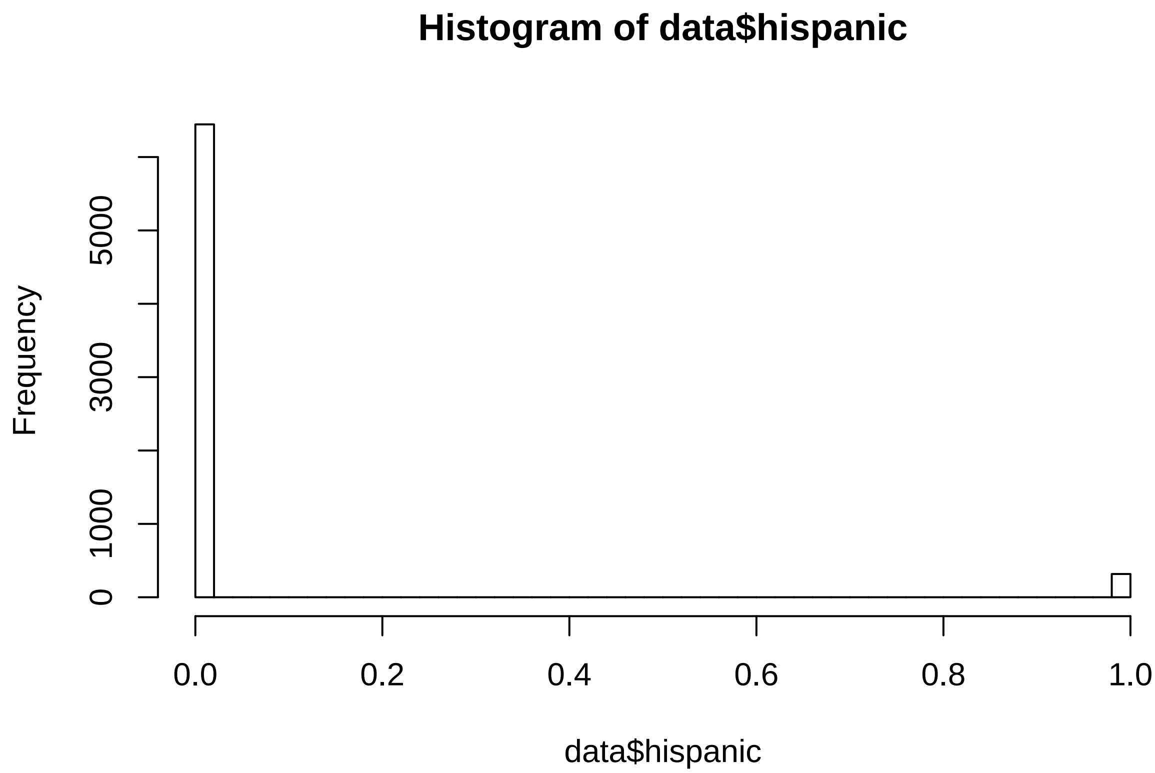

```r
summary(data$hispanic)
```

```
##    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)
```



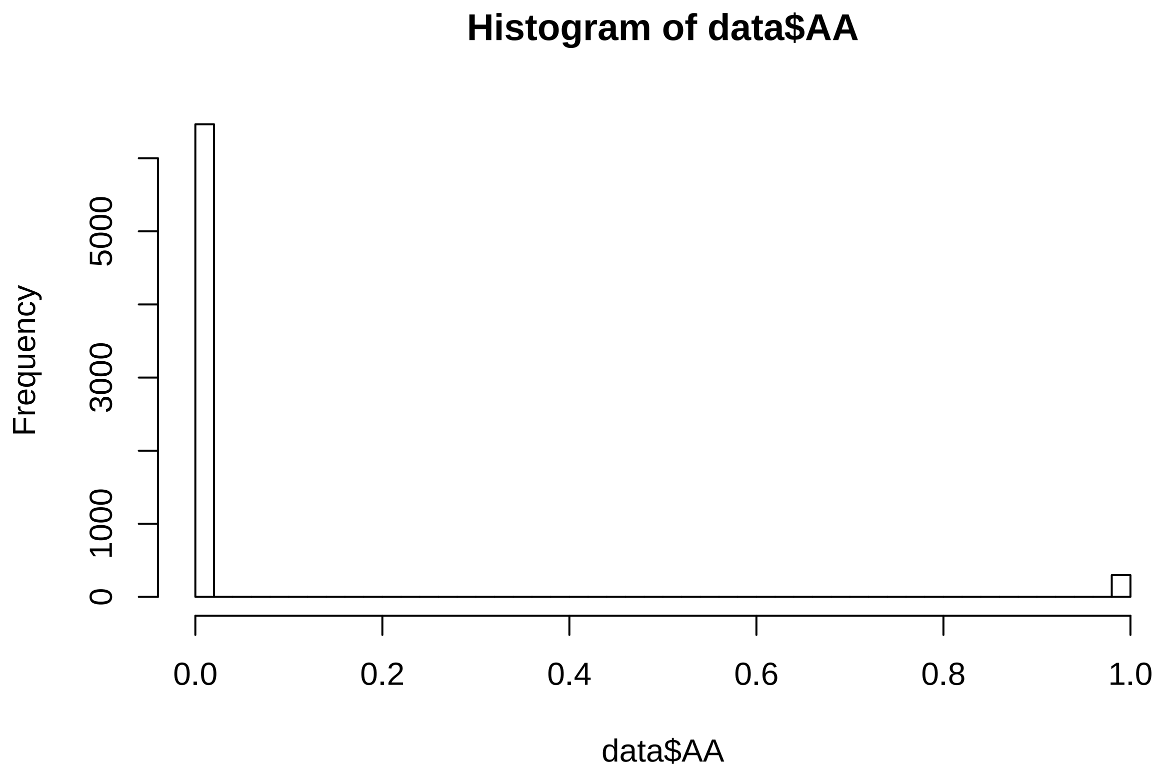

```
summary(data$AA)
```

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

```
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
```

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



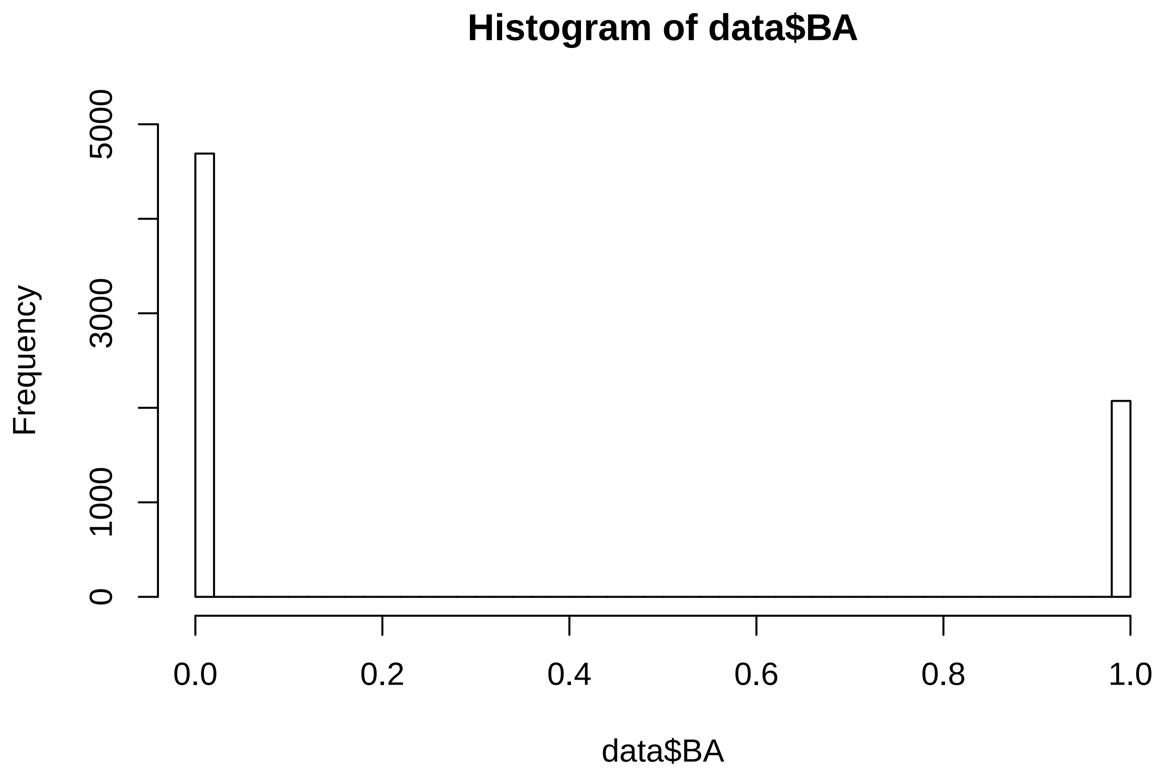

```r
summary(data$BA)
```

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

```r
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
```

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



Basic structure of the data

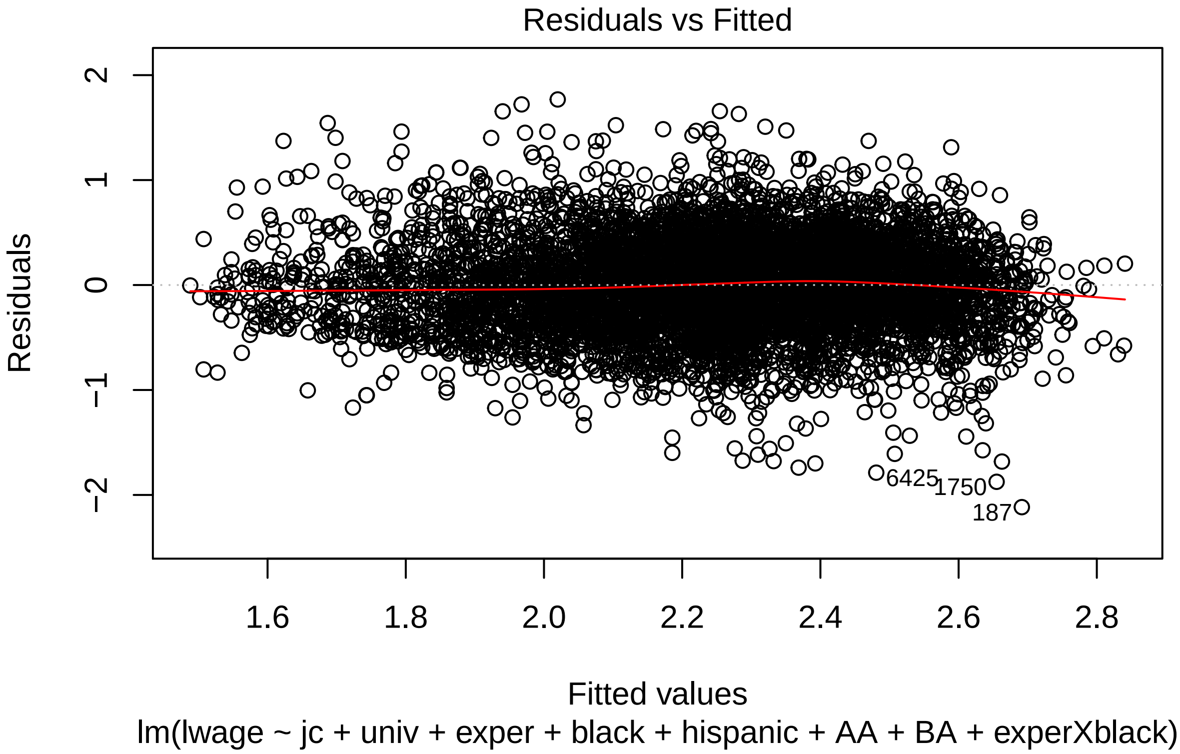
There are no missing values in the data.
lwage variable has a normal-like distribution.
jc variable has values from 0 to about 4 and is heavily positively skewed with a majority of values at or near 0.
univ variable has values from 0 to 7.5 and is heavily positively skewed with a majority of values at 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.

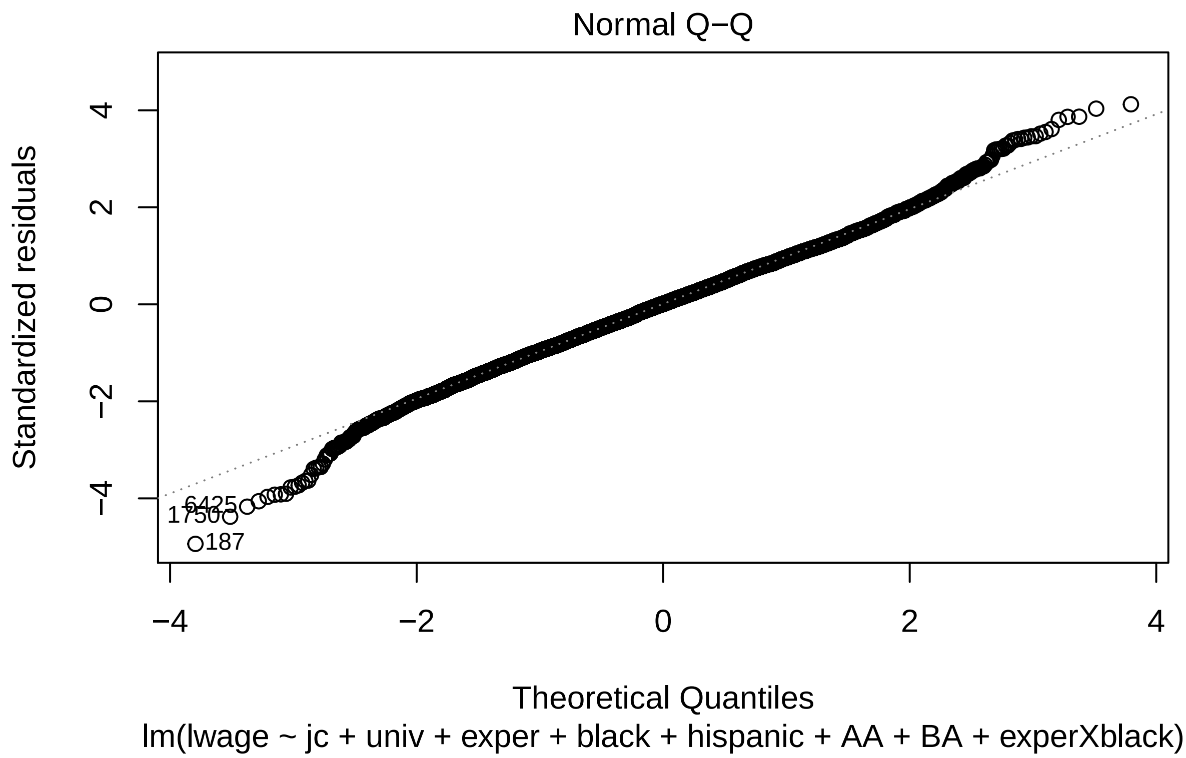
Question 2

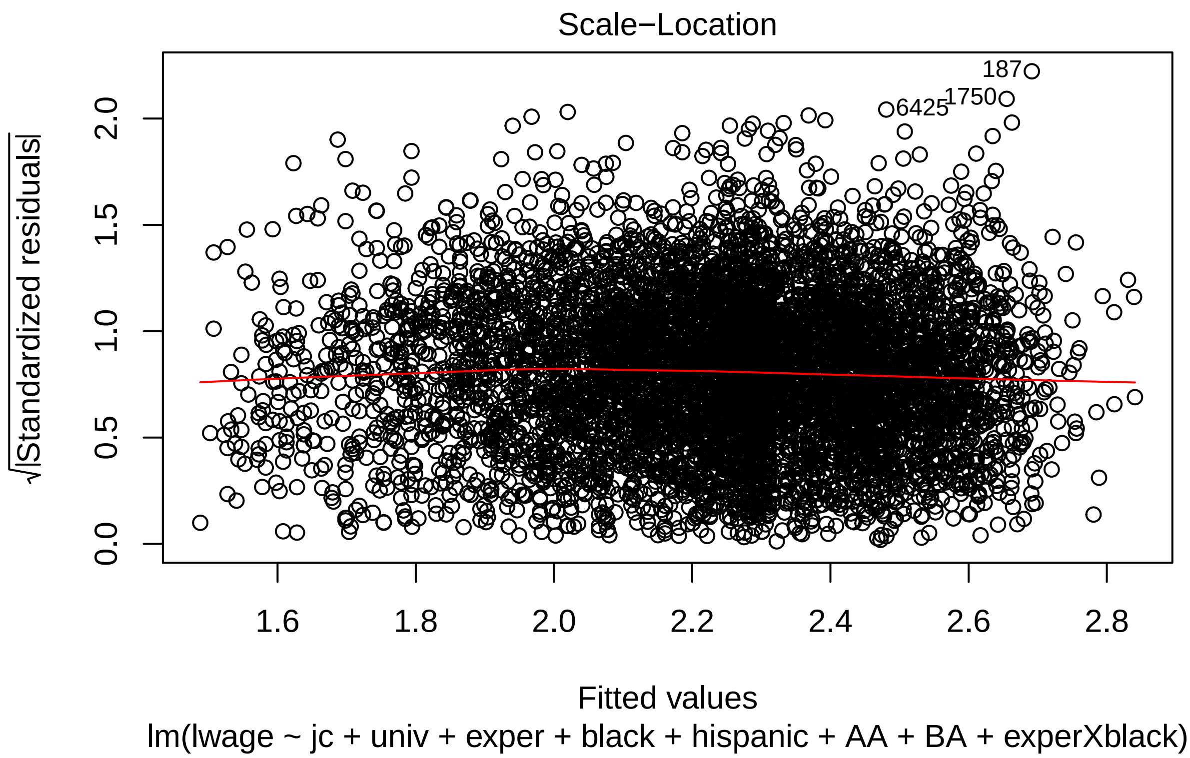
```
# Create the experXblack variable by multiplying
# the exper and black variables.
data$experXblack = data$exper * data$black

# 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
```

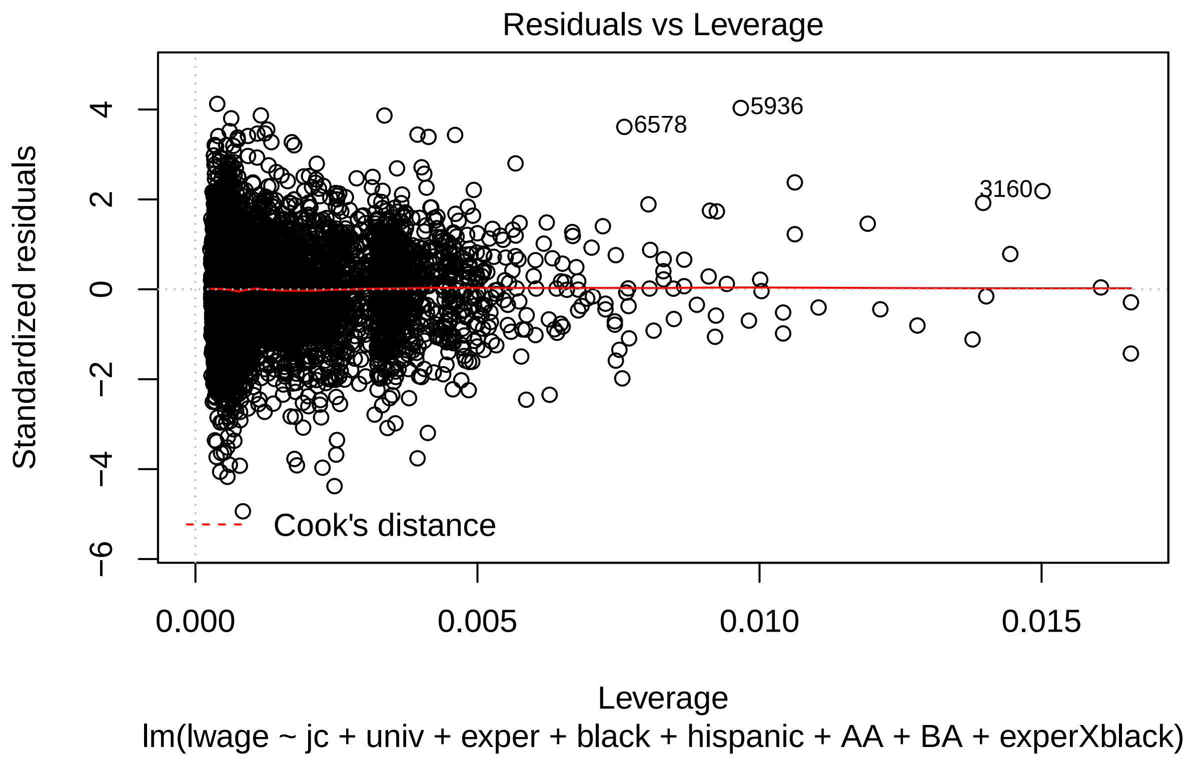
```r
# Print the diagnostic plots
plot(ols.lwage.8ind)
```


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


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



Scale−Location
√|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
187
1750
6425
Fitted values
lm(lwage ~ jc + univ + exper + black + hispanic + AA + BA + experXblack)




```
# Print the B_hat4 and B_hat8 coefficients
print(ols.lwage.8ind$coefficients[5])
```

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

```
print(ols.lwage.8ind$coefficients[9])
```

```
##  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.
Do we talk about:
zero-conditional mean seems to be met
homoskedasticity seems to be met
assuming random sample
assuming linear relationship

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[3])
```

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

```r
(0.0733 - 0.07)/(0.0031)
```

```
## [1] 1.064516
```

```r
2 * (1 - 0.8554)
```

```
## [1] 0.2892
```

Test that the return to university education is 7%.

Null Hypothesis: H0: $\beta 2 = 0.07$.
Alternate Hypothesis: H1: $\beta 2 \neq 0.07$.

Formula for t-statistic $= (\beta 2 - H0)/(se) = (.0733 - .07)/(.0031) = 1.064516$
p-value $= 2 * (1 - .8554) = 0.2892$
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

Question 5

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

Original model:

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

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.

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

We would like to know if the $\beta 2$ and $\beta 3$ coefficients are the same or, equivalently, if their difference is 0. We can define a variable θ such that $\theta = \beta 2 - \beta 3$ and rewrite our model like this:

$lwage = \beta 0 + \beta 1 jc + (\theta + \beta 3) univ + \beta 3 experYr + \beta 4 black + \beta 5 hispanic + \beta 6 AA + \beta 7 BA + \beta 8 experXblack + \epsilon$

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

$lwage = \beta 0 + \beta 1 jc + \theta univ + \beta 3 (univ + experYr) + \beta 4 black + \beta 5 hispanic + \beta 6 AA + \beta 7 BA + \beta 8 experXblack + \epsilon$

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
print(sqrt(0.2282))
```

```
## [1] 0.4777028
```

Test the overall signiffcance of this regression.

Here is the output from the summary of our 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. An R value of 0.478 indicates a ?? effect size.
   ??which regression model are we supposed to be using here, the one with univPlusexperYr or the first one??

Question 7

```r
data$experXexper = data$exper * data$exper
ols.lwage.9ind = lm(lwage ~ jc + univ + exper + black +
    hispanic + AA + BA + experXblack + experXexper,
    data = data)
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

$lwage = \beta 0+\beta 1jc+\beta 2univ+\beta 3exper+\beta 4black+\beta 5hispanic+\beta 6AA+\beta 7BA+\beta 8experXblack+\beta 9experXexper$

$$\triangle lwage/\triangle exper = \beta 3 + \beta 8black + 2\beta 9exper$$

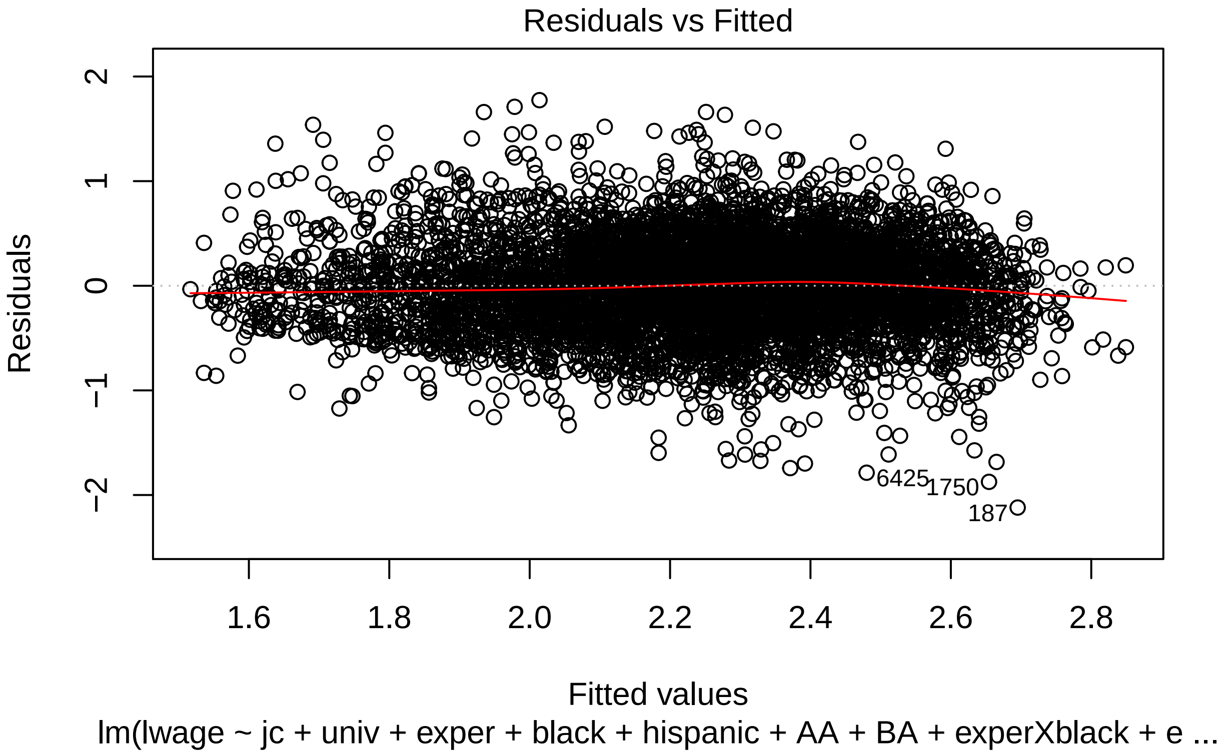
$$= (.004301 - .001239 * black + 2 * .000003379 * exper)$$

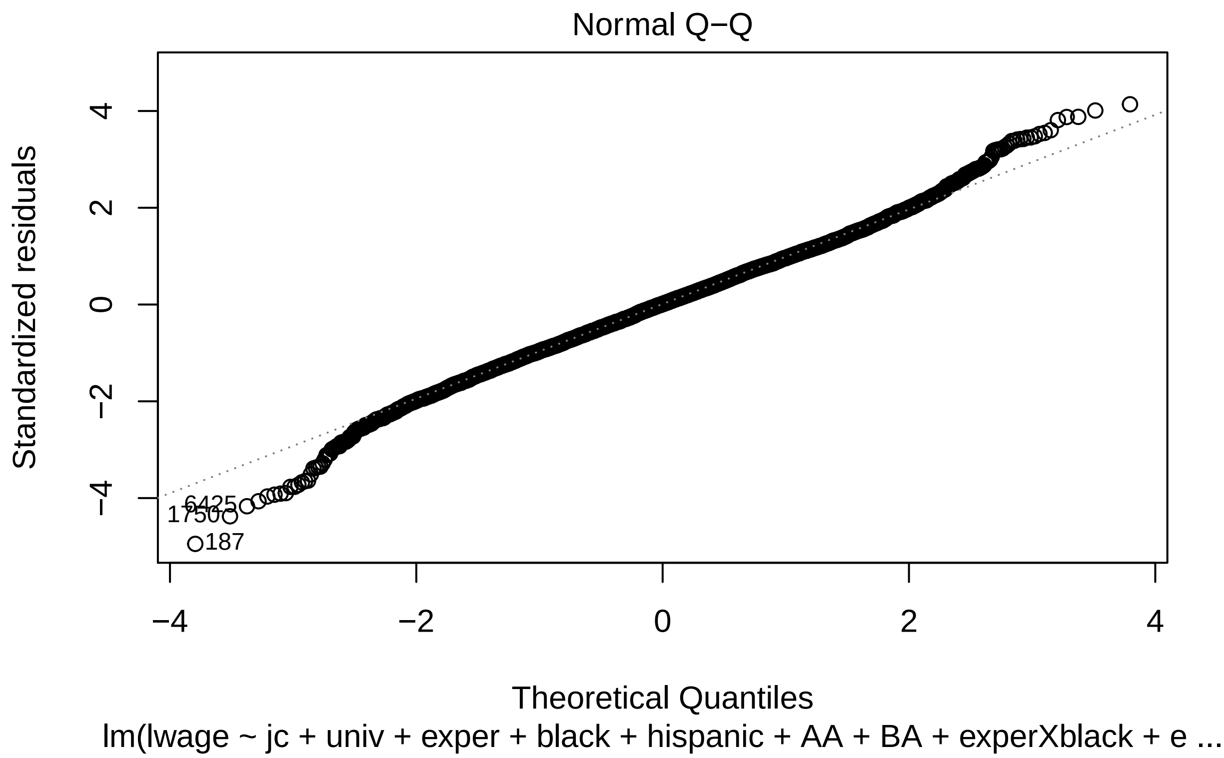
Now convert the log wage back to wage by exponentiating. This gives us a return to work experience:

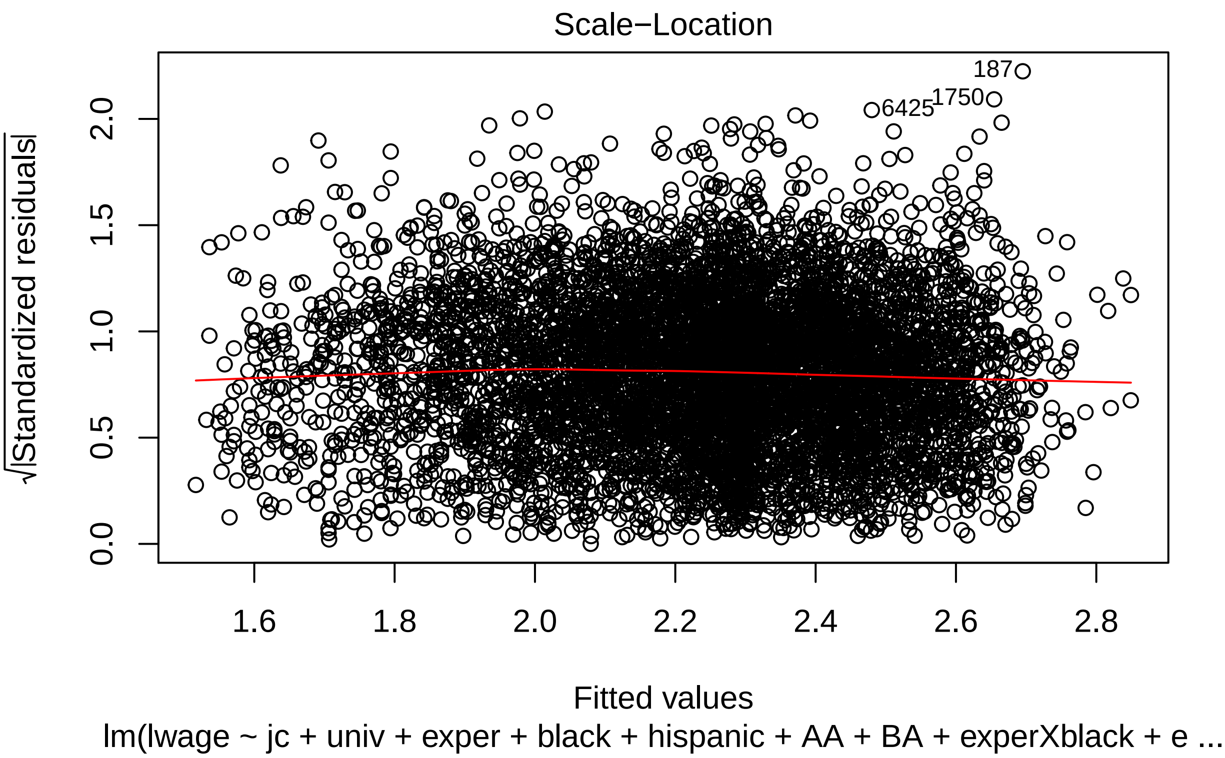
$$= e^{(.004301-.001239*black+2*.000003379*exper)}$$

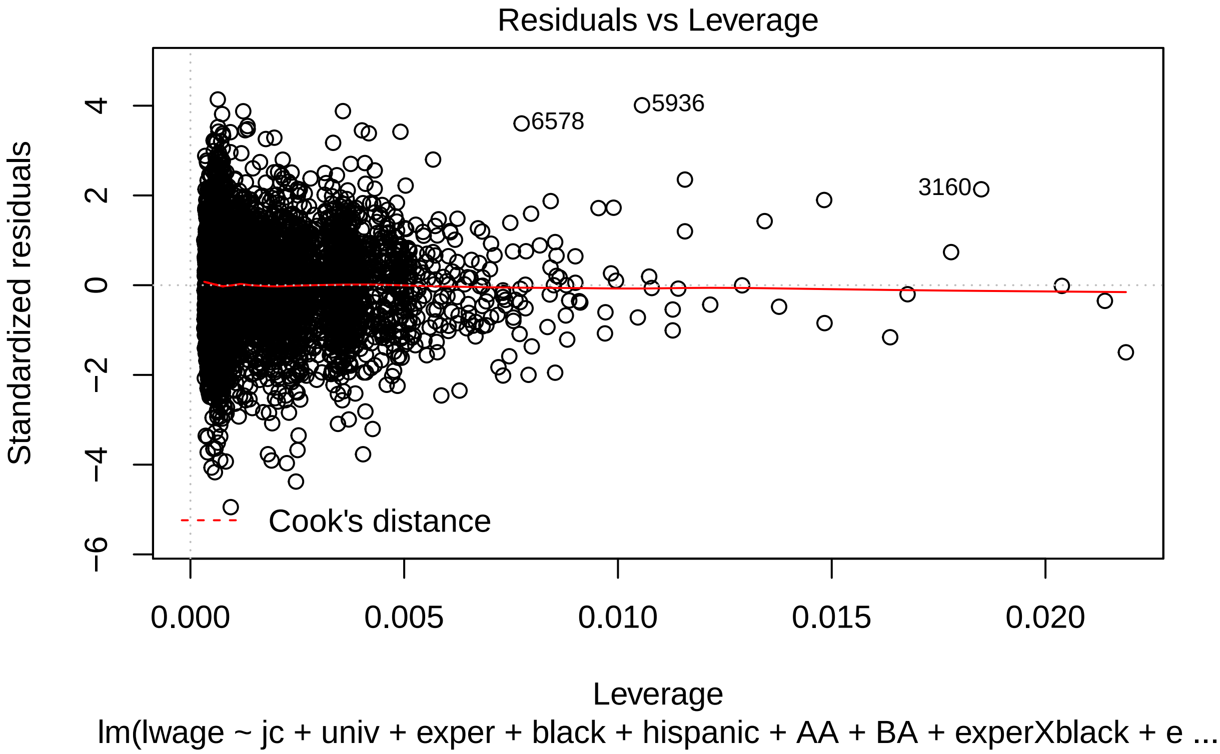
Question 8

```r
# Based on the violation of homoskedasticity, we
# must run robust standard errors. coeftest(model,
# vcov=vcovHC) waldtest(model, vcov=vcovHC)
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
187
6425
1750
0.0
0.5
1.0
1.5
2.0
√|Standardized residuals|
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 ...



Homoskedasticity analysis:

The assumption of homoskedasticity holds:
1 - We can see from the residuals vs fitted plot that the variance band is about the same 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 almost completely horizontal, which is what we get if homoskedasticity is met.
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 homoskedasticity in the data is that the standard error of the univ coefficient ($\beta 2$) is unbiased. Unbiased standard errors will not impact the outcomes of statistical tests. Therefore, it does not affect the testing of no effect of university education on salary change.