# *GR5205 LINEAR REGRESSION*

CAPSTON PROJECT

*Siwei Liu sl4224*

*Dec. 8th, 2018*

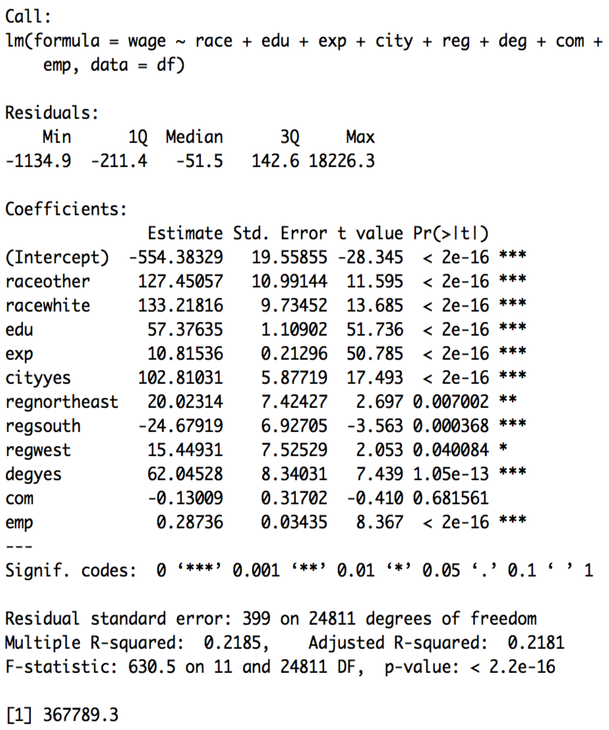
# I. INTRODUCTION

The goal of this project is to find a linear regression model which incorporates all relevant variables, interaction and functional forms of the covariates. And answer the two research questions:

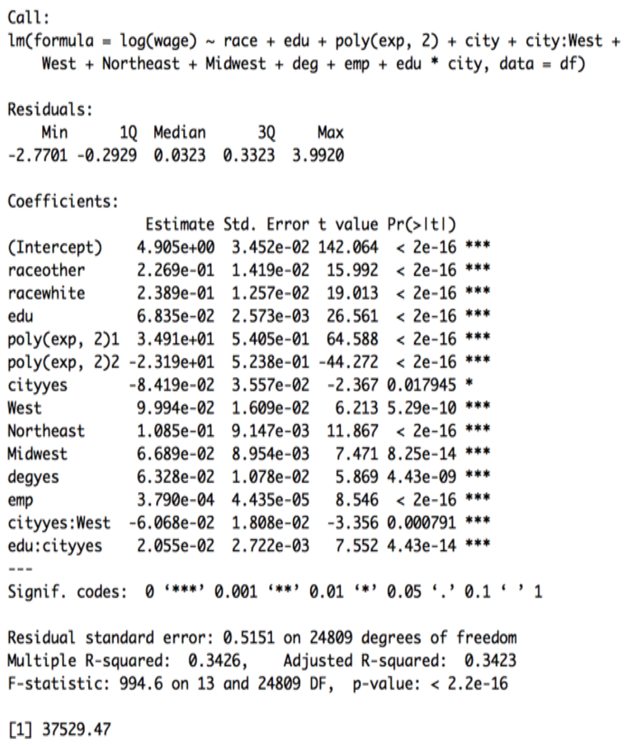
1. Do African American males have statistically different wages compared to Caucasian males?
2. Do African American males have statistically different wages compared to other males?

Data contains around 25,000 records of males between the age of 18 and 70 who are full time workers. Possible independent variables in the model are: years of education, job experience, college graduate, working in or near a city, US region, commuting distance, number of employees in a company and race.

Below are some summary statistics and exploratory data analysis plots that are the most relevant ones in terms of my model building and research questions. The first summary output is the one from the rough model which includes all the potential covariates and where no transformations or interactions are applied. Next to it, is the final model I derived, we can see that the coefficient of determination has increased to a satisfactory number and AIC has dropped to a much smaller order of magnitude in comparison to the rough model. Details will be discussed in the following sections.

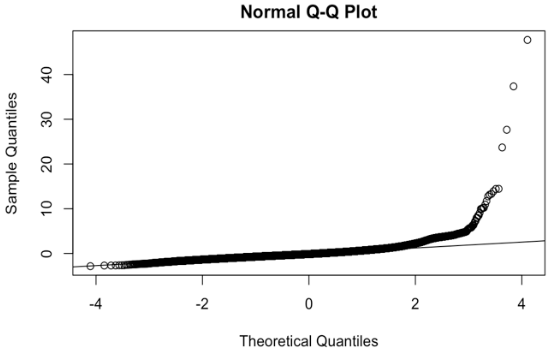


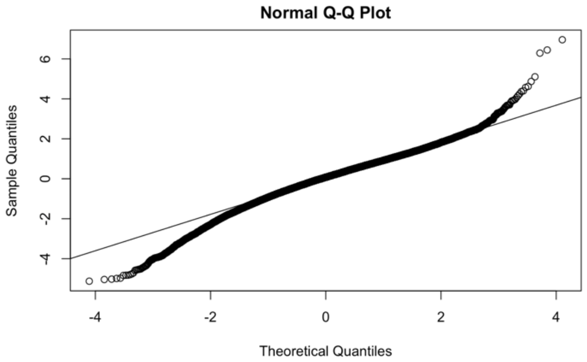
```
Call:
lm(formula = wage ~ race + edu + exp + city + reg + deg + com + 
    emp, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-1134.9  -211.4   -51.5   142.6 18226.3 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -554.38329   19.55855 -28.345  < 2e-16 ***
raceother     127.45057   10.99144  11.595  < 2e-16 ***
racewhite     133.21816    9.73452  13.685  < 2e-16 ***
edu            57.37635    1.10902  51.736  < 2e-16 ***
exp            10.81536    0.21296  50.785  < 2e-16 ***
cityyes       102.81031    5.87719  17.493  < 2e-16 ***
regnortheast   20.02314    7.42427   2.697 0.007002 ** 
regsouth      -24.67919    6.92705  -3.563 0.000368 ***
regwest        15.44931    7.52529   2.053 0.040084 *  
degyes         62.04528    8.34031   7.439 1.05e-13 ***
com            -0.13009    0.31702  -0.410 0.681561    
emp             0.28736    0.03435   8.367  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 399 on 24811 degrees of freedom
Multiple R-squared:  0.2185,	Adjusted R-squared:  0.2181 
F-statistic: 630.5 on 11 and 24811 DF,  p-value: < 2.2e-16

[1] 367789.3
```

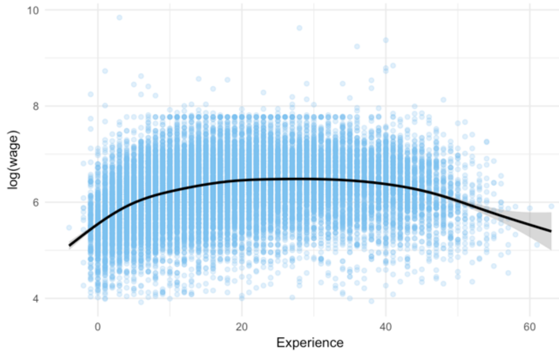


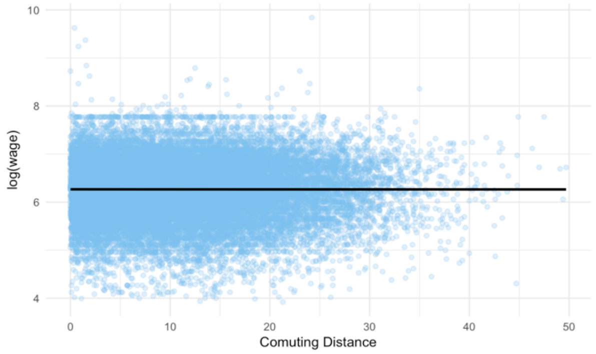
```
Call:
lm(formula = log(wage) ~ race + edu + poly(exp, 2) + city + city:West + 
    West + Northeast + Midwest + deg + emp + edu * city, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.7701 -0.2929  0.0323  0.3323  3.9920 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    4.905e+00  3.452e-02 142.064  < 2e-16 ***
raceother      2.269e-01  1.419e-02  15.992  < 2e-16 ***
racewhite      2.389e-01  1.257e-02  19.013  < 2e-16 ***
edu            6.835e-02  2.573e-03  26.561  < 2e-16 ***
poly(exp, 2)1  3.491e+01  5.405e-01  64.588  < 2e-16 ***
poly(exp, 2)2 -2.319e+01  5.238e-01 -44.272  < 2e-16 ***
cityyes       -8.419e-02  3.557e-02  -2.367 0.017945 *  
West           9.994e-02  1.609e-02   6.213 5.29e-10 ***
Northeast      1.085e-01  9.147e-03  11.867  < 2e-16 ***
Midwest        6.689e-02  8.954e-03   7.471 8.25e-14 ***
degyes         6.328e-02  1.078e-02   5.869 4.43e-09 ***
emp            3.790e-04  4.435e-05   8.546  < 2e-16 ***
cityyes:West  -6.068e-02  1.808e-02  -3.356 0.000791 ***
edu:cityyes    2.055e-02  2.722e-03   7.552 4.43e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5151 on 24809 degrees of freedom
Multiple R-squared:  0.3426,	Adjusted R-squared:  0.3423 
F-statistic: 994.6 on 13 and 24809 DF,  p-value: < 2.2e-16

[1] 37529.47
```

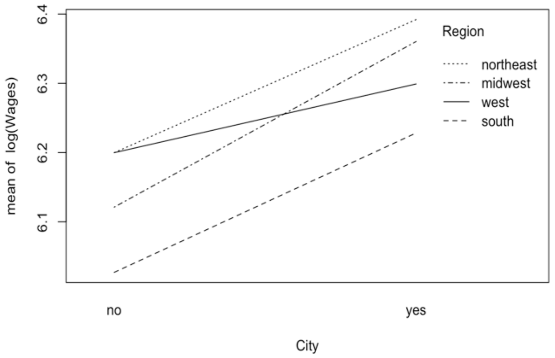
The following six plots are the most important ones that helped me land in my final model. Details of how transformations and interactions are applied to my model are included in the appendix.

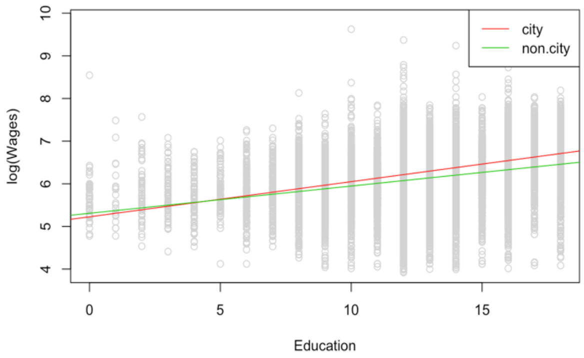












Further discussion of these plots is included in section IV. Appendix.

# II. STATISTICAL MODEL

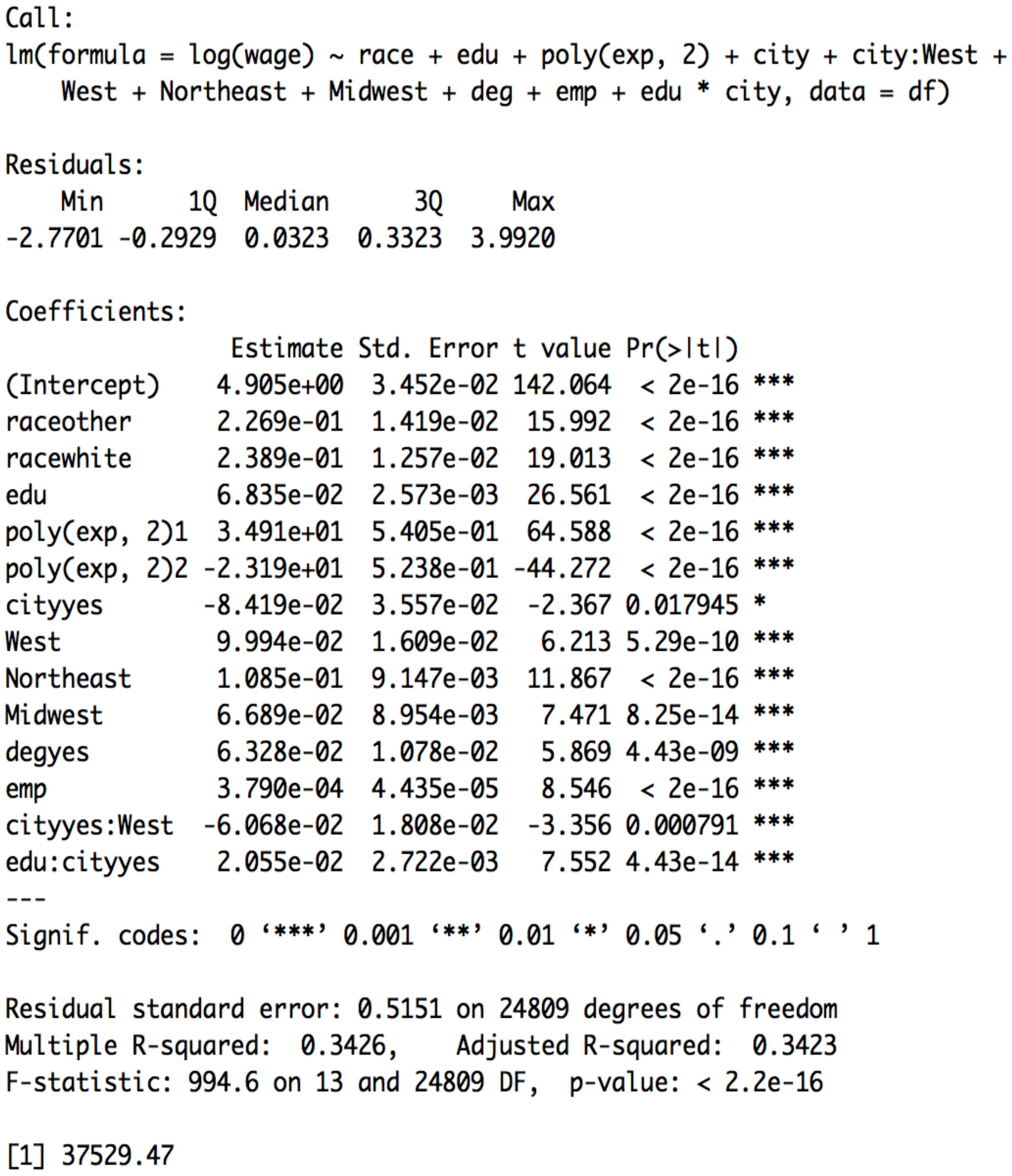
The final statistical model that I build is

$$\begin{aligned} log(wage) = {} & 4.905 + 0.2269raceother + 0.2389racewhite + 0.06835edu \\ & + 0.3491poly(exp, 2)1 - 0.2319poly(exp, 2)2 - 0.08419city \\ & + 0.09994West + 0.1085Northeast + 0.06689Midwest + 0.06328deg \\ & + 3.790 \times 10^{-4}emp - 0.06068city \times West + 0.02055edu \times city \end{aligned}$$

where *raceother, racewhite, city, West, Northeast, Midwest, deg* are all dummy variables, and takes values 0 and 1. *raceother* takes value 1 if that individual male belongs to race other than white and black, and takes value 0 otherwise; *racewhite* takes value 1 if the individual male belongs to race white, and 0 otherwise; *city* takes value 1 if the individual works in or near a city and 0 otherwise; *West, Northeast* and *Midwest* are three dummy variables I created since directly adding the interaction between region and city did not work well. So, I created these three dummy variables to separate the *reg* (region) variable into four parts and only add interaction between *West* and *city* as suggested in the interaction plot. *West* takes value 1 if the individual is from region west, 0 otherwise; *Northeast* takes value 1 if that person if from region northeast, 0 otherwise; *Midwest* takes value 1 if the male is from region midwest and 0 otherwise. Also, it's worth mentioning that since there are strong collinearity between *exp* and *exp*$^2$, I used orthogonal polynomial applying the poly() function.

In this model, I include one square term on *exp* (experience), and two interactions, which are *city* with *West* and *edu* (education) with *city.* I also deleted one variable *com* (commuting distance). Details of how I finalized my model like this is shown in appendix.

Below is the summary output. We can see the AIC is 37529.47, $R^2$ is 0.3426 and $R_a^2$ is 0.3423. After having the finalized the model, I split the data set into the training data and testing data, where the testing data is around 20% of the whole dataset. From the chart below, we can see that the MSPR is 0.2645, and is pretty close to the MSE. It looks like the model fits out-of-sample data similar to in-sample data.



```
Call:
lm(formula = log(wage) ~ race + edu + poly(exp, 2) + city + city:West +
    West + Northeast + Midwest + deg + emp + edu * city, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.7701 -0.2929  0.0323  0.3323  3.9920 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    4.905e+00  3.452e-02 142.064  < 2e-16 ***
raceother      2.269e-01  1.419e-02  15.992  < 2e-16 ***
racewhite      2.389e-01  1.257e-02  19.013  < 2e-16 ***
edu            6.835e-02  2.573e-03  26.561  < 2e-16 ***
poly(exp, 2)1  3.491e+01  5.405e-01  64.588  < 2e-16 ***
poly(exp, 2)2 -2.319e+01  5.238e-01 -44.272  < 2e-16 ***
cityyes       -8.419e-02  3.557e-02  -2.367 0.017945 *  
West           9.994e-02  1.609e-02   6.213 5.29e-10 ***
Northeast      1.085e-01  9.147e-03  11.867  < 2e-16 ***
Midwest        6.689e-02  8.954e-03   7.471 8.25e-14 ***
degyes         6.328e-02  1.078e-02   5.869 4.43e-09 ***
emp            3.790e-04  4.435e-05   8.546  < 2e-16 ***
cityyes:West  -6.068e-02  1.808e-02  -3.356 0.000791 ***
edu:cityyes    2.055e-02  2.722e-03   7.552 4.43e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5151 on 24809 degrees of freedom
Multiple R-squared:  0.3426,	Adjusted R-squared:  0.3423 
F-statistic: 994.6 on 13 and 24809 DF,  p-value: < 2.2e-16

[1] 37529.47
```

| MSPR | MSE | MSE earlier |
|---|---|---|
| 0.2645 | 0.2654 | 0.2868 |

# III. RESEARCH QUESTION

The test I use to answer the two research questions is t-test.

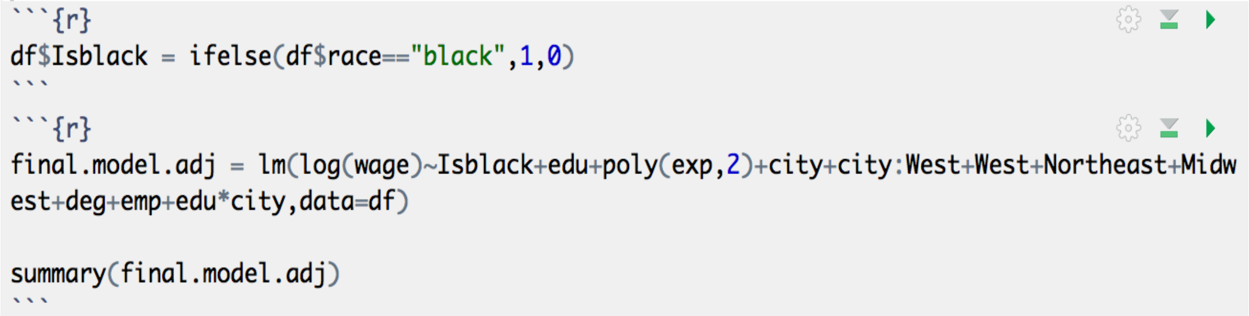
**Question I**:  Do African American males have statistically different wages compared to Caucasian males?

This question can be directly answered according to the summary output in part II. STATISTICAL MODEL. The null hypothesis is that African American males do not have statistically different wages compared to Caucasian males ($H_0$: $wage_{african\ amarican} = wage_{caucasian}$). And the alternative hypothesis is African American males have statistically different wages compared to Caucasian males. ($H_1$: $wage_{african\ amarican} \neq wage_{caucasian}$).

According to the summary output, the variable for race black (African American) is in the intercept, so in this case, when we have the p-value of *racewhite* less than 2e-16, which is significant, we can conclude from the t-test that African American males do have significantly different wages compared to Caucasian males.

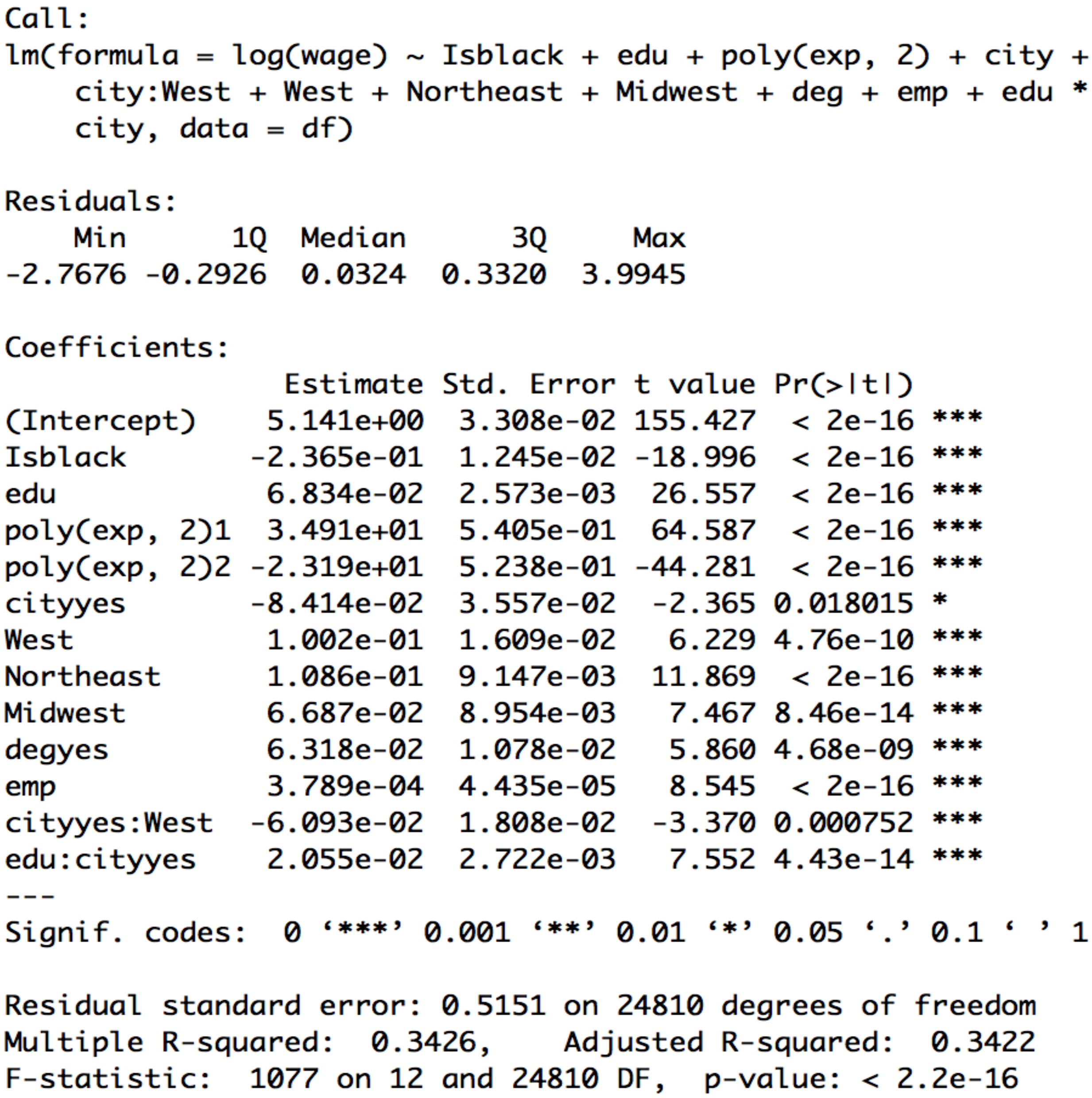
**Question II**:  Do African males have statistically different wages compared to all other males?

To answer this question, I recode the race variable to be two groups, the new variable is *Isblack*, which take the value 1 if the value of the race variable is "black", and take the value 0 otherwise. Then, I replace the *race* variable in my model with *Isblack,* run t-test again.



~~~
```{r}
df$Isblack = ifelse(df$race=="black",1,0)
~~~

```{r}
final.model.adj = lm(log(wage)~Isblack+edu+poly(exp,2)+city+city:West+West+Northeast+Midw
est+deg+emp+edu*city,data=df)

summary(final.model.adj)
```
```

In this question, the null hypothesis is African American males do not have statistically different wages from all other males. The alternative hypothesis is African American males do have different wages from all other males.
```



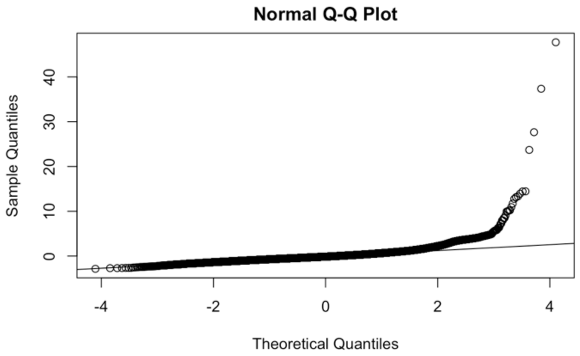
```
Call:
lm(formula = log(wage) ~ Isblack + edu + poly(exp, 2) + city +
    city:West + West + Northeast + Midwest + deg + emp + edu *
    city, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.7676 -0.2926  0.0324  0.3320  3.9945 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    5.141e+00  3.308e-02 155.427  < 2e-16 ***
Isblack       -2.365e-01  1.245e-02 -18.996  < 2e-16 ***
edu            6.834e-02  2.573e-03  26.557  < 2e-16 ***
poly(exp, 2)1  3.491e+01  5.405e-01  64.587  < 2e-16 ***
poly(exp, 2)2 -2.319e+01  5.238e-01 -44.281  < 2e-16 ***
cityyes       -8.414e-02  3.557e-02  -2.365 0.018015 *  
West           1.002e-01  1.609e-02   6.229 4.76e-10 ***
Northeast      1.086e-01  9.147e-03  11.869  < 2e-16 ***
Midwest        6.687e-02  8.954e-03   7.467 8.46e-14 ***
degyes         6.318e-02  1.078e-02   5.860 4.68e-09 ***
emp            3.789e-04  4.435e-05   8.545  < 2e-16 ***
cityyes:West  -6.093e-02  1.808e-02  -3.370 0.000752 ***
edu:cityyes    2.055e-02  2.722e-03   7.552 4.43e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5151 on 24810 degrees of freedom
Multiple R-squared:  0.3426,	Adjusted R-squared:  0.3422 
F-statistic:  1077 on 12 and 24810 DF,  p-value: < 2.2e-16
```

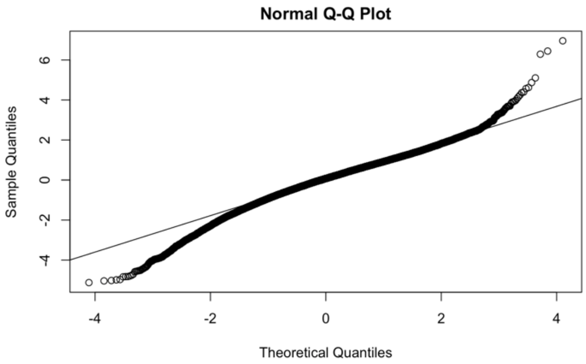
Now, all other males including raceother and racewhite are in the intercept. We can see from the summary output above, Isblack has the p-value less than 2e-16. This means that the t-test shows that African males do have statistically different wages compared to all other males.

# IV. APPENDIX

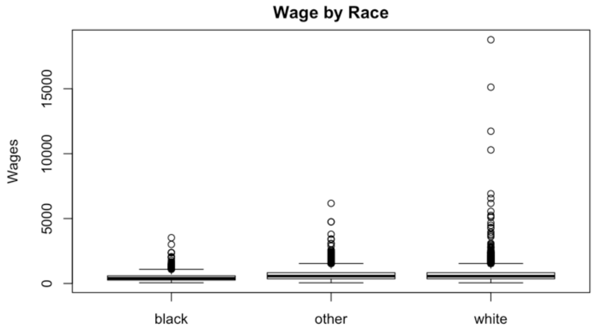
## Model Selection

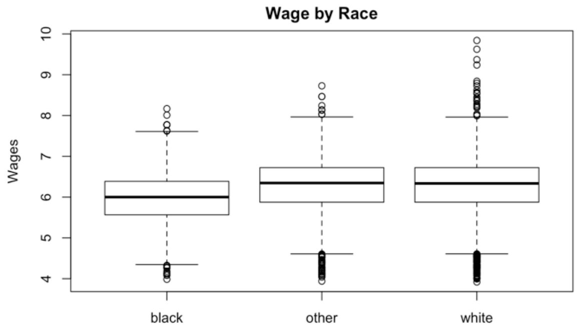
The first rough model included all the potential independent variables, i.e. *wage = race+edu+exp+city+reg+deg+com+emp*. However, after looking at the QQplot, I realized there should be a transformation on the response variable, so I took the logarithm of wage, and drew the QQplot again. It looked much nicer.





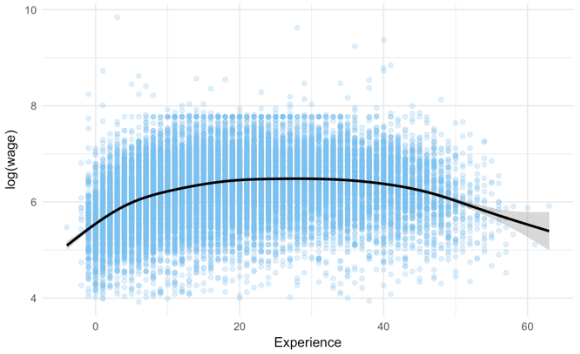
Also, after the transformation, the boxplot looks a lot more reasonable too.

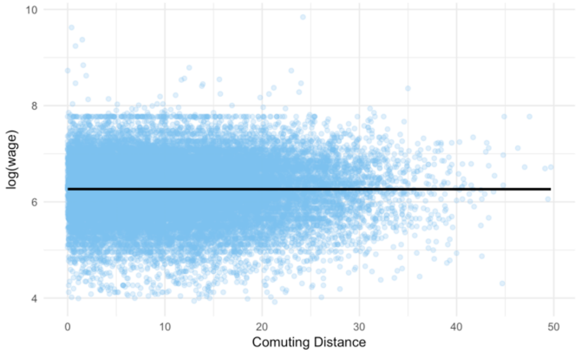




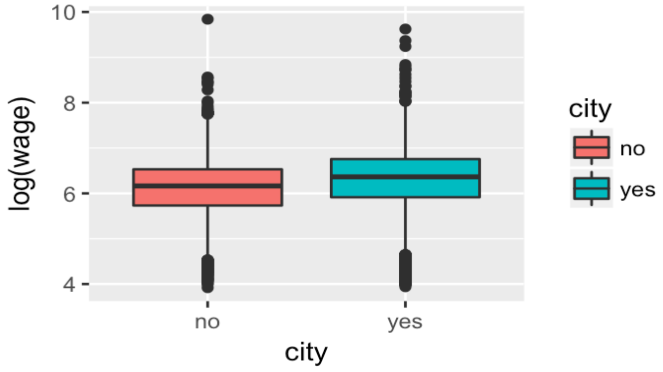
If we look at the summary output, in fact, after just take the logarithm of the wage, the model has already had a great improvement. $R^2$ has increased from around 22% to around 29%. AIC has also dropped to an entirely different order of magnitude, from 367789.3 to 39454.08.

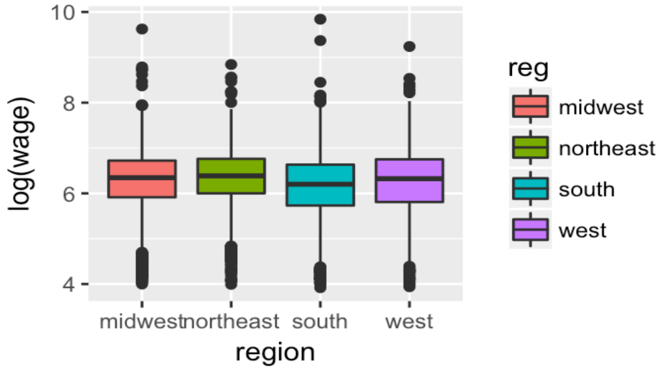
After taking the logarithm, I drew log(wage) with other potential independent variables using ggplot. It is obvious that there is a curvature pattern between log(wage) and experience. Also, according to the plot, commuting distance doesn't seem to affect the response variable, since no matter what the commuting distance is, the log(wage) is almost a straight line.

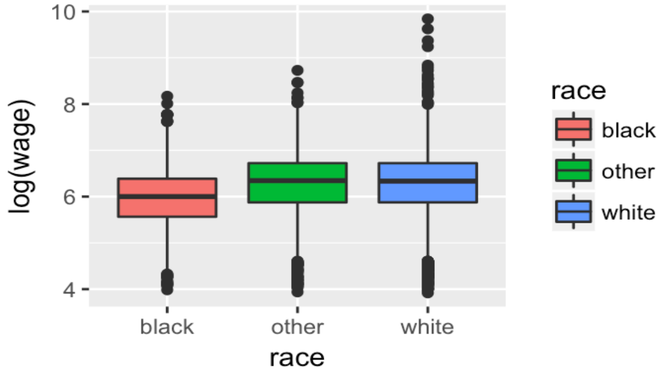


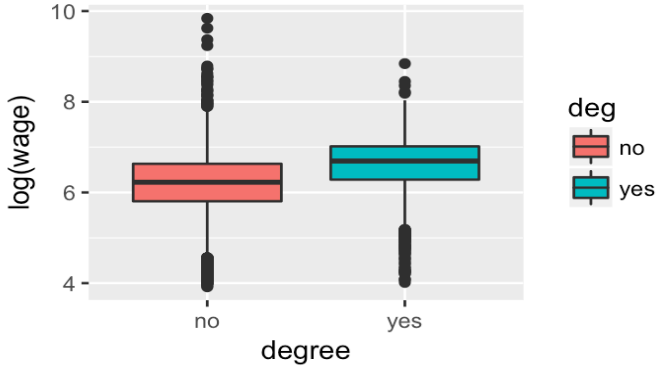


Other potential covariates seem to all, more or less have some impact on log(wage) according to the EDA plots.

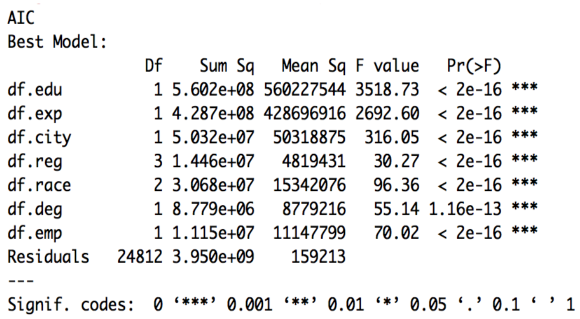




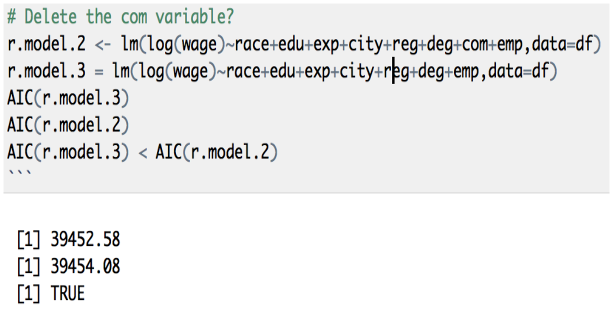


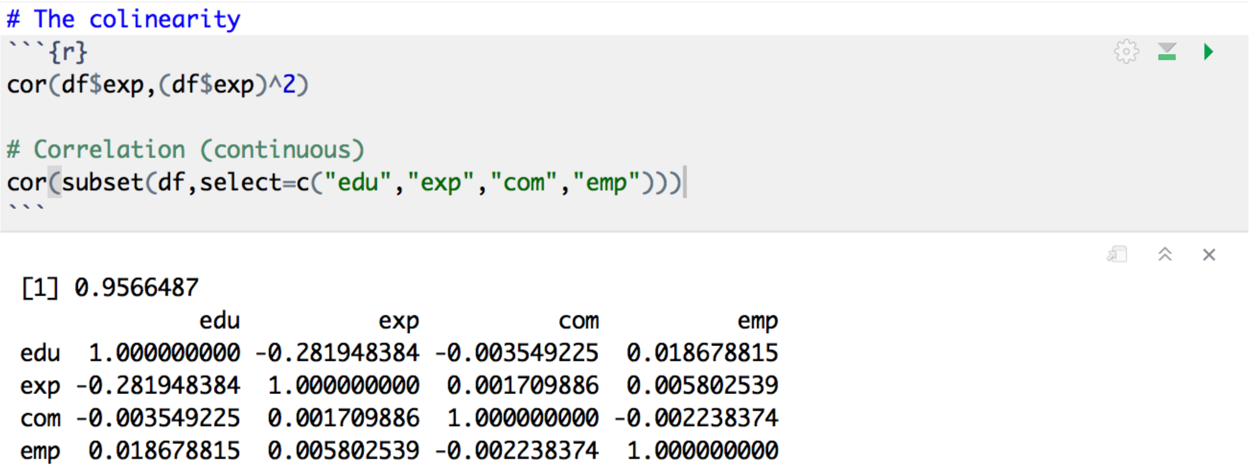


Also, I ran the *bestglm* function not to decide but as a supportive information to my decision of whether the *com* (commuting distance) variable should indeed be deleted from the model, as suggested by the Exploratory Data Analysis plot.



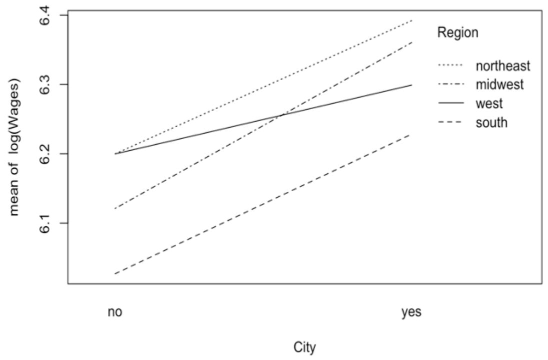
```
AIC
Best Model:
             Df    Sum Sq   Mean Sq F value    Pr(>F)    
df.edu        1 5.602e+08 560227544 3518.73  < 2e-16 ***
df.exp        1 4.287e+08 428696916 2692.60  < 2e-16 ***
df.city       1 5.032e+07  50318875  316.05  < 2e-16 ***
df.reg        3 1.446e+07   4819431   30.27  < 2e-16 ***
df.race       2 3.068e+07  15342076   96.36  < 2e-16 ***
df.deg        1 8.779e+06   8779216   55.14 1.16e-13 ***
df.emp        1 1.115e+07  11147799   70.02  < 2e-16 ***
Residuals 24812 3.950e+09    159213                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

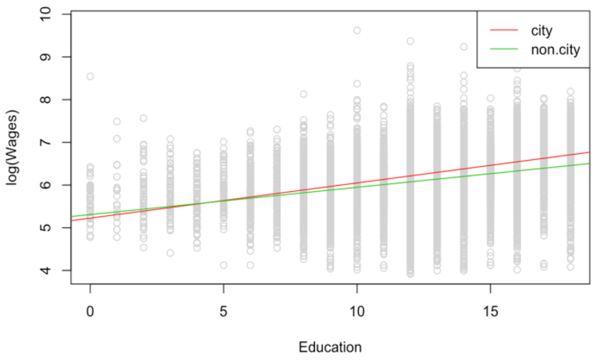


```
# Delete the com variable?
r.model.2 <- lm(log(wage)~race+edu+exp+city+reg+deg+com+emp,data=df)
r.model.3 = lm(log(wage)~race+edu+exp+city+reg+deg+emp,data=df)
AIC(r.model.3)
AIC(r.model.2)
AIC(r.model.3) < AIC(r.model.2)
```
```

```
[1] 39452.58
[1] 39454.08
[1] TRUE
```

We can see from the above result from the *bestglm* model, it supports the decision to delete the *com* (commuting distance) variable as well. Also, after deleting it, the AIC of the model decreased.

So, up to this point, according to the QQplot, the EDA plots and the result from the *bestglm* function, I decide to first take the logarithm of the response variable (wage), then square *exp* (experience), and delete *com* (commuting distance). However, before writing down the modified model, one thing that need to be considered is the collinearity between the covariates, here, for example, the square transformation I am going to applied on *exp* is probably going to cause this issue.



```
# The colinearity
```{r}
cor(df$exp,(df$exp)^2)

# Correlation (continuous)
cor(subset(df,select=c("edu","exp","com","emp")))
```
```

```
[1] 0.9566487
             edu          exp          com          emp
edu  1.000000000 -0.281948384 -0.003549225  0.018678815
exp -0.281948384  1.000000000  0.001709886  0.005802539
com -0.003549225  0.001709886  1.000000000 -0.002238374
emp  0.018678815  0.005802539 -0.002238374  1.000000000
```

According to the result above, the *exp* and *exp* squared definitely have a strong collinearity, so I used the *poly()* function in R using orthogonal polynomial to solve this issue. Also, I calculated the correlation among the other continuous variables, fortunately, there appears to be no strong correlations among these variables. After solving the collinearity problem, I modified my model as described above. Let's call this modified model, *my.model.1*.

*my.model.1: log(wage)=race+edu+poly(exp,2)+city+reg+deg+emp*
```

If we look at the summary output and AIC of *my.model.1*, $R^2$ has increased from around 29% to 34%! And the AIC has decreased from 39454.08 to 37591.05.

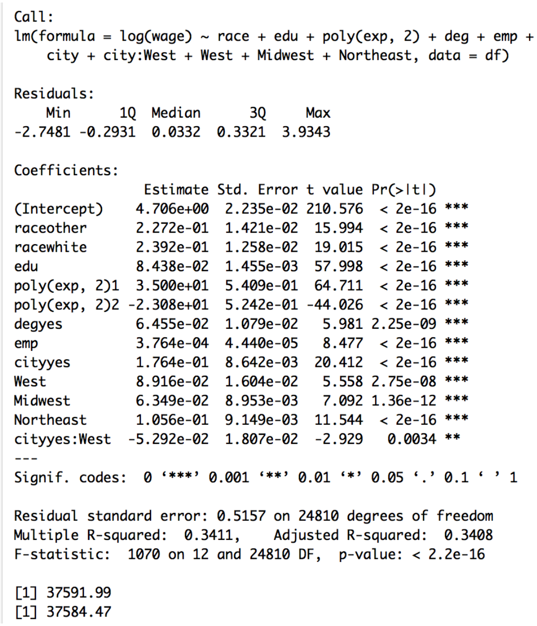
Next, let's explore further to see if we need to add some interactions to the model.



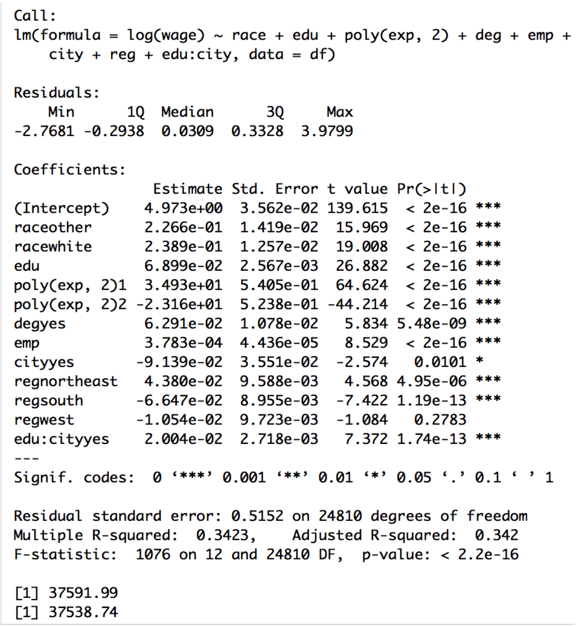


We can clearly see from the two plots that there is interaction between city and region west, also there is interaction between city and education. It doesn't seem to produce satisfactory results if we just add interaction between city and the whole region variable. So, I created 3 dummy variables to separate the four regions. And I only added interaction between *city* and *West* instead of city and the whole region. After adding this interaction, the AIC has dropped from 37592 to 37584.

Also, when I added the interaction between city and education, the AIC dropped from 37592 to 37539. As shown in the charts below.



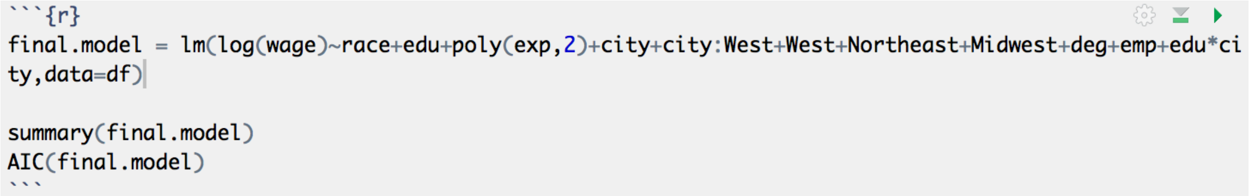
```
Call:
lm(formula = log(wage) ~ race + edu + poly(exp, 2) + deg + emp +
    city + city:West + West + Midwest + Northeast, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7481 -0.2931  0.0332  0.3321  3.9343

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.706e+00  2.235e-02 210.576  < 2e-16 ***
raceother      2.272e-01  1.421e-02  15.994  < 2e-16 ***
racewhite      2.392e-01  1.258e-02  19.015  < 2e-16 ***
edu            8.438e-02  1.455e-03  57.998  < 2e-16 ***
poly(exp, 2)1  3.500e+01  5.409e-01  64.711  < 2e-16 ***
poly(exp, 2)2 -2.308e+01  5.242e-01 -44.026  < 2e-16 ***
degyes         6.455e-02  1.079e-02   5.981 2.25e-09 ***
emp            3.764e-04  4.440e-05   8.477  < 2e-16 ***
cityyes        1.764e-01  8.642e-03  20.412  < 2e-16 ***
West           8.916e-02  1.604e-02   5.558 2.75e-08 ***
Midwest        6.349e-02  8.953e-03   7.092 1.36e-12 ***
Northeast      1.056e-01  9.149e-03  11.544  < 2e-16 ***
cityyes:West  -5.292e-02  1.807e-02  -2.929   0.0034 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5157 on 24810 degrees of freedom
Multiple R-squared:  0.3411,	Adjusted R-squared:  0.3408
F-statistic:  1070 on 12 and 24810 DF,  p-value: < 2.2e-16

[1] 37591.99
[1] 37584.47
```



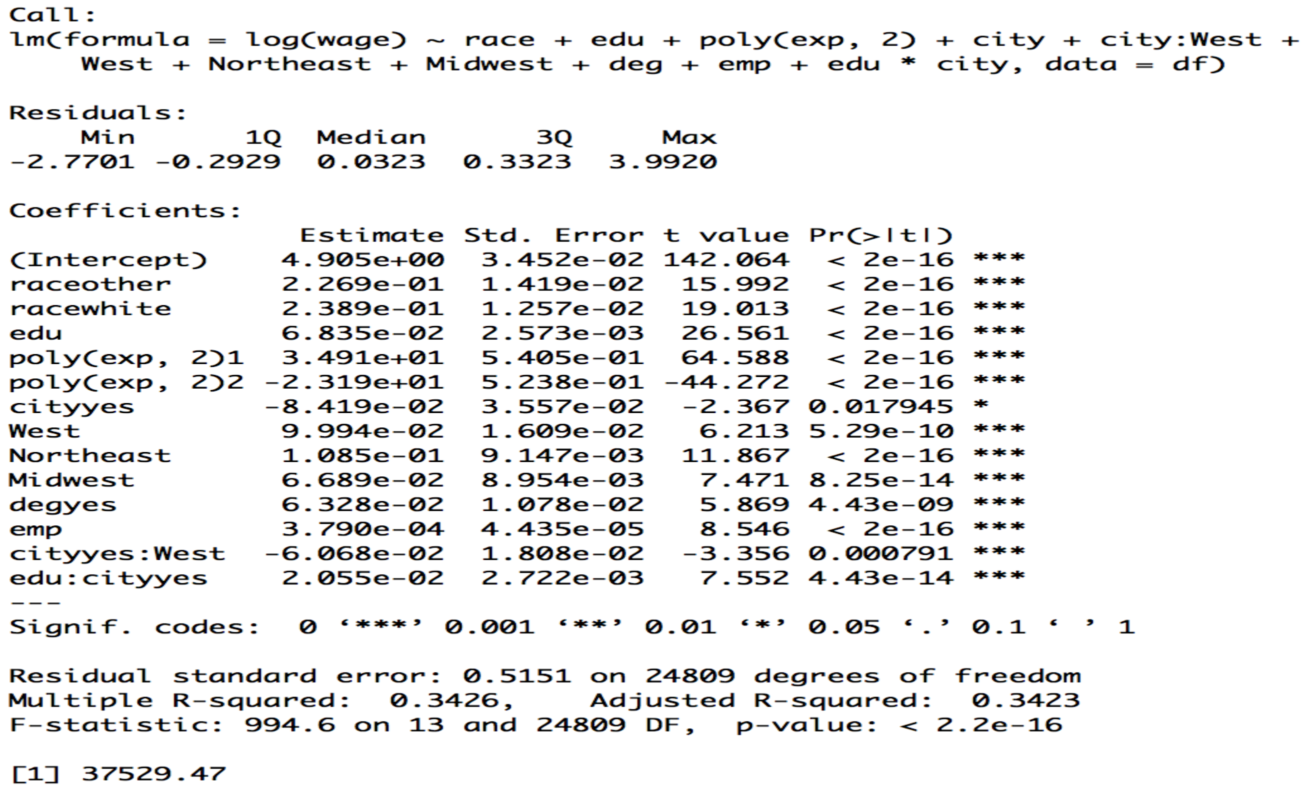
```
Call:
lm(formula = log(wage) ~ race + edu + poly(exp, 2) + deg + emp +
    city + reg + edu:city, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7681 -0.2938  0.0309  0.3328  3.9799

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.973e+00  3.562e-02 139.615  < 2e-16 ***
raceother      2.266e-01  1.419e-02  15.969  < 2e-16 ***
racewhite      2.389e-01  1.257e-02  19.008  < 2e-16 ***
edu            6.899e-02  2.567e-03  26.882  < 2e-16 ***
poly(exp, 2)1  3.493e+01  5.405e-01  64.624  < 2e-16 ***
poly(exp, 2)2 -2.316e+01  5.238e-01 -44.214  < 2e-16 ***
degyes         6.291e-02  1.078e-02   5.834 5.48e-09 ***
emp            3.783e-04  4.436e-05   8.529  < 2e-16 ***
cityyes       -9.139e-02  3.551e-02  -2.574   0.0101 *
regnortheast   4.380e-02  9.588e-03   4.568 4.95e-06 ***
regsouth      -6.647e-02  8.955e-03  -7.422 1.19e-13 ***
regwest       -1.054e-02  9.723e-03  -1.084   0.2783
edu:cityyes    2.004e-02  2.718e-03   7.372 1.74e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5152 on 24810 degrees of freedom
Multiple R-squared:  0.3423,	Adjusted R-squared:  0.342
F-statistic:  1076 on 12 and 24810 DF,  p-value: < 2.2e-16

[1] 37591.99
[1] 37538.74
```

I also noticed that it appears in the graph that there are interactions between race and region west, also, education and degree. But when I added them in the model, the AIC did not drop, so I decided not to include these interactions in my final model.

So, my final model and its summary statistics is as below:



~~~
```{r}
final.model = lm(log(wage)~race+edu+poly(exp,2)+city+city:West+West+Northeast+Midwest+deg+emp+edu*ci
ty,data=df)

summary(final.model)
AIC(final.model)
```
~~~

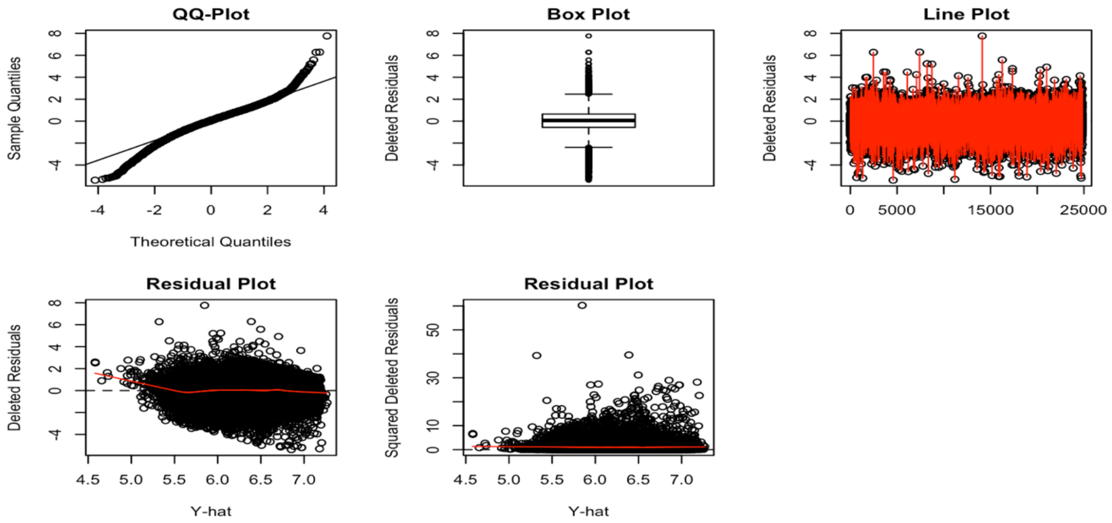


```
Call:
lm(formula = log(wage) ~ race + edu + poly(exp, 2) + city + city:West +
    West + Northeast + Midwest + deg + emp + edu * city, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.7701 -0.2929  0.0323  0.3323  3.9920 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     4.905e+00  3.452e-02 142.064  < 2e-16 ***
raceother       2.269e-01  1.419e-02  15.992  < 2e-16 ***
racewhite       2.389e-01  1.257e-02  19.013  < 2e-16 ***
edu             6.835e-02  2.573e-03  26.561  < 2e-16 ***
poly(exp, 2)1   3.491e+01  5.405e-01  64.588  < 2e-16 ***
poly(exp, 2)2  -2.319e+01  5.238e-01 -44.272  < 2e-16 ***
cityyes        -8.419e-02  3.557e-02  -2.367 0.017945 *  
West            9.994e-02  1.609e-02   6.213 5.29e-10 ***
Northeast       1.085e-01  9.147e-03  11.867  < 2e-16 ***
Midwest         6.689e-02  8.954e-03   7.471 8.25e-14 ***
degyes          6.328e-02  1.078e-02   5.869 4.43e-09 ***
emp             3.790e-04  4.435e-05   8.546  < 2e-16 ***
cityyes:West   -6.068e-02  1.808e-02  -3.356 0.000791 ***
edu:cityyes     2.055e-02  2.722e-03   7.552 4.43e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5151 on 24809 degrees of freedom
Multiple R-squared:  0.3426,	Adjusted R-squared:  0.3423 
F-statistic: 994.6 on 13 and 24809 DF,  p-value: < 2.2e-16

[1] 37529.47
```

The coefficient of determination is around 34% and the AIC value has dropped from initially 367789.3 (the rough model which includes all the potential covariates) to 37529.47. It is pretty satisfying.

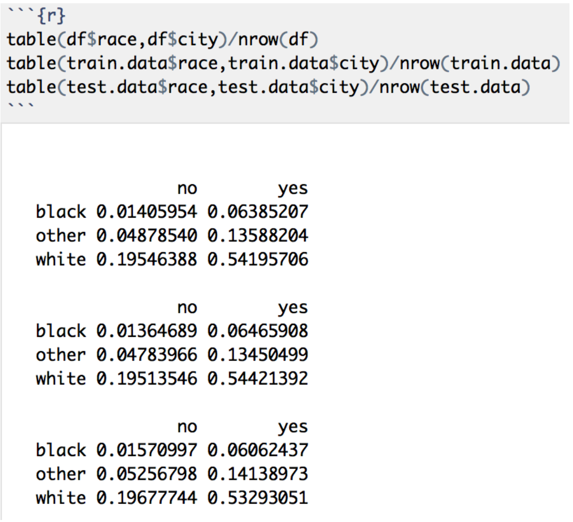
## Diagnostics and Model validation

i) . Several diagnostic residual plots on the final model are included below (student deleted residual)



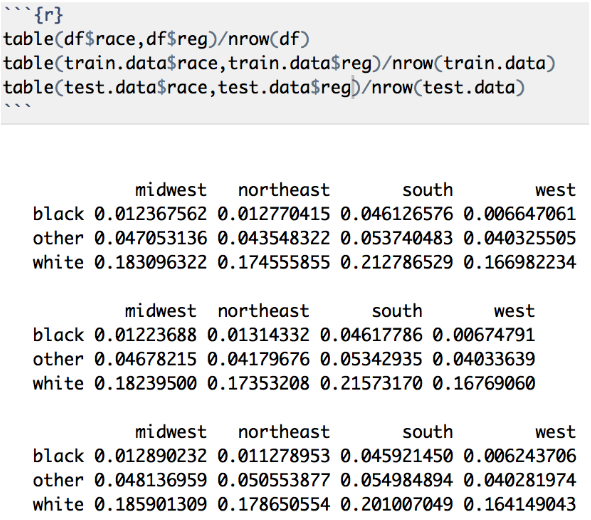
The boxplot looks symmetric, the line plot has no specific pattern, the two residual plots looks good too.

ii)In the model validation process, I split the dataset into the training data and testing data, where the testing data is around 20% of the whole dataset. Before doing the model validation, let’s do some quality control check.



~~~
```{r}
table(df$race,df$city)/nrow(df)
table(train.data$race,train.data$city)/nrow(train.data)
table(test.data$race,test.data$city)/nrow(test.data)
```
~~~

```
              no        yes
black 0.01405954 0.06385207
other 0.04878540 0.13588204
white 0.19546388 0.54195706

              no        yes
black 0.01364689 0.06465908
other 0.04783966 0.13450499
white 0.19513546 0.54421392

              no        yes
black 0.01570997 0.06062437
other 0.05256798 0.14138973
white 0.19677744 0.53293051
```



~~~
```{r}
table(df$race,df$reg)/nrow(df)
table(train.data$race,train.data$reg)/nrow(train.data)
table(test.data$race,test.data$reg)/nrow(test.data)
```
~~~

```
          midwest   northeast       south        west
black 0.012367562 0.012770415 0.046126576 0.006647061
other 0.047053136 0.043548322 0.053740483 0.040325505
white 0.183096322 0.174555855 0.212786529 0.166982234

         midwest  northeast      south       west
black 0.01223688 0.01314332 0.04617786 0.00674791
other 0.04678215 0.04179676 0.05342935 0.04033639
white 0.18239500 0.17353208 0.21573170 0.16769060

          midwest   northeast       south        west
black 0.012890232 0.011278953 0.045921450 0.006243706
other 0.048136959 0.050553877 0.054984894 0.040281974
white 0.185901309 0.178650554 0.201007049 0.164149043
```

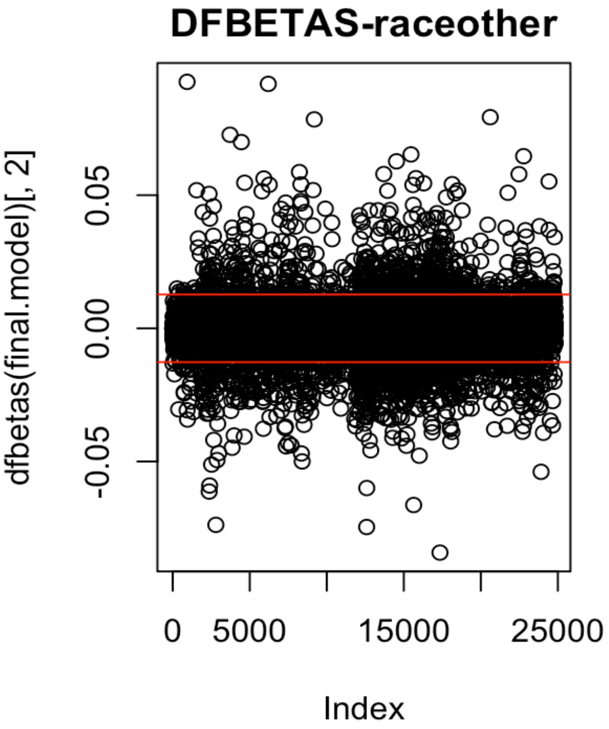
Ideally, we want the proportion of the race levels to be similar for the full data, training data and testing data. From the above table we can see, it is indeed true.

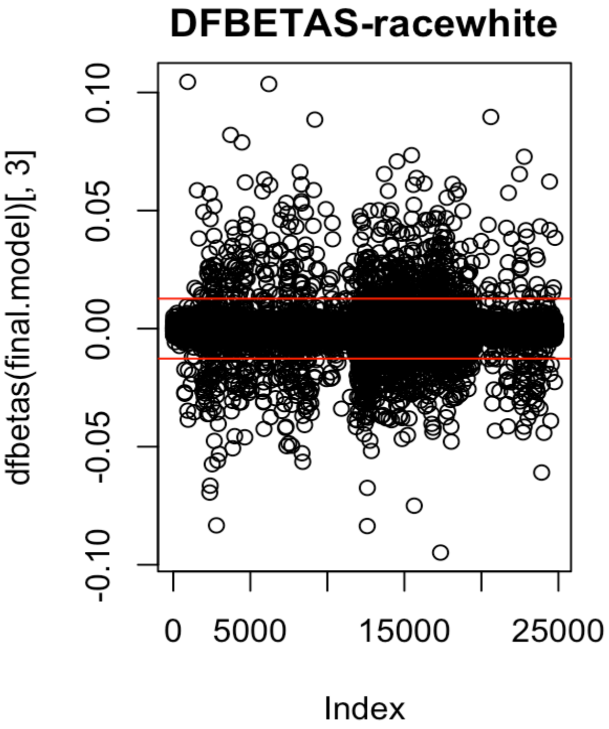
I trained my model with the training set, and then computed the MSPR using my final model trained from the training set, on the test set.

| MSPR | MSE | MSE earlier |
|---|---|---|
| 0.2645 | 0.2654 | 0.2868 |

The chart above shows that the difference of MSPR and MSE is very small, which implies that the model fits out-of-sample data similar to in-sample data. Also the MSE has decreased compared to the MSE we had earlier with our rough model (r.model.2: *log(wage) = race + edu + exp + city + reg + deg + com + emp*).

iii) Influential observations with *race* variable (since for the research question, we are only interested in testing the slopes related to *race*)





We can see from the above plots, there are a lot of influential observations and outliers. However, in this case, I still decide to run the ordinary least squares, because **t -test** we are going to do, is **robust.** Also, we can expect to have a large number of outliers since we have large number of observations (around 25,000).