STAT 5205 Project: Are the average male wages statistically different across race classes?

Name: Siddhanth Sabharwal, UNI: ss5689

December 08, 2018

I. Introduction

The goal of this analysis is to investigate whether or not the average male wages are statistically difference across three race classes. The three race classes, in alphabetical order, are black, other, and white. In this analysis African American will be used interchangeably with black, and Caucasian will be used interchangeably with white. The data is for males between 18 and 70 who are full time workers. We will come up with a linear regression model to predict weekly wages in dollars. We will then use that model to answer 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 all other males?

First, let us see the dimensions and structure of the input data.

```
## [1] "The dataset salary.txt has 24823 rows and 9 columns."
```

This means we have 24,823 rows of data, 1 response variable, and 8 explanatory variables.

```
##     wage edu exp city       reg  race deg  com emp
## 1 354.94   7  45  yes northeast white  no 24.3 200
## 2 370.37   9   9  yes northeast white  no 26.2 130
## 3 754.94  11  46  yes northeast white  no 26.4 153
## 4 593.54  12  36  yes northeast other  no  9.9  86
## 5 377.23  16  22  yes northeast white yes  7.1 181
```

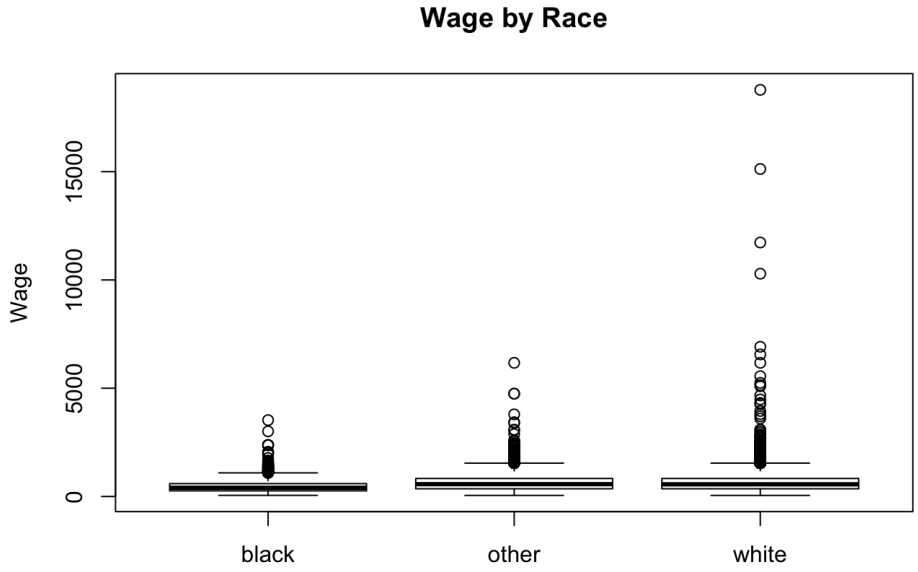
The response variable is wage. Wage stands for “weekly wages (in dollars)”. It is a continuous, numeric variable that is never negative (minimum is 50.39). The remaining variables are the explanatory variables.

Second, let us investigate how we should think of the explanatory variables present in this study.

```
## [1] "edu"  "exp"  "city" "reg"  "race" "deg"  "com"  "emp"
```

Edu stands for “years of education”. It is a discrete, numeric variable as it only takes whole number values (minimum is 0). Exp stands for “years of job experience”. It is a discrete, numeric variable as it only takes integer values (minimum is -4). City stands for “working in or near a city”. It is a factor, non-numeric variable with two levels (no and yes). Reg stands for “US region”. It is a factor, non-numeric variable with four levels (midwest, northeast, south, and west). Race is self-explanatory, it is simply the race of that male. It is a factor, non-numeric variable with three levels (black, other, and white). Deg stands for “college graduate”. It is a factor, non-numeric variable with two levels (no and yes). Com stands for “commuting distance”. It is a continuous, numeric variable that is never negative (minimum is 0). Emp stands for “number of employees”. It is a discrete, numeric variable as it only takes natural number values (minimum is 3).

Third, let us do some exploratory data analysis and generate some basic summary statistics to see if we can immediately spot differences in wages by race.




```
##
##
## Table: Summary statistics of Wage against Race
##
##                race = black   race = other   race = white
## -------------  -------------  -------------  -------------
## Min            53.83          51.44          50.39
## 1st Quartile   261.16         356.13         356.41
## Median         403.61         569.80         563.77
## 3rd Quartile   593.54         830.96         830.96
## Max            1091.22        1540.84        1541.41
## Mean           472.21         649.97         652.27
## SD             309.56         426.27         466.25
```

While the boxplot is quite condensed around the median, we can immediately see that Caucassians have a lot bigger upper outliers than either African Americans or other males. While the median is robust to the presence of outliers, the mean is not. This is providing some indication that the averages may be statistically different. The table of basic summary statistics also indicates differences in the averages between African Americans and the other races. The summary statistics also show that African Americans trail behind the other races when looking at wages across the 1st quartile, Median, and 3rd Quartile - so it is not unreasonable to assume this is translating into the lower across-the-board average we see. However, we need to see if these differences are significant when controlling for the other explanatory variables before we jump to any conclusions, as correlation does not mean causation. There could be something else going on specifically with African Americans that is causing their wages to be lower on average. For example, maybe African Americans live in areas where wages are lower in higher concentration than other races. This could mean that the differences in wages is driven by region, not race. Issues like this are addressed in the Model Selection part of the Appendix.

II. Statistical Model

The final model is log(wage) = 4.775 + 0.0846(edu) + 35.02(exp) - 23.06(exp^2) + 0.1643(I(city=yes)) + 0.044(I(reg=northeast)) - 0.06319(I(reg=south)) - 0.01231(I(reg=west)) + 0.227(I(race=other)) + 0.2392(I(race=white)) + 0.06419(I(deg=yes)) + 0.0003759(emp). I() is an indicator function that takes the value of 1 when the condition inside the parantheses is satisfied, and 0 otherwise. The summary output of this model is below.

```
##
## Call:
## lm(formula = log(wage) ~ edu + poly(exp, degree = 2) + city +
##     reg + race + deg + emp, data = salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7469 -0.2932  0.0332  0.3330  3.9250
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.775e+00  2.342e-02 203.904  < 2e-16 ***
## edu                     8.460e-02  1.453e-03  58.207  < 2e-16 ***
## poly(exp, degree = 2)1  3.502e+01  5.410e-01  64.740  < 2e-16 ***
## poly(exp, degree = 2)2 -2.306e+01  5.242e-01 -43.981  < 2e-16 ***
## cityyes                 1.643e-01  7.598e-03  21.628  < 2e-16 ***
## regnortheast            4.400e-02  9.598e-03   4.584 4.58e-06 ***
## regsouth               -6.319e-02  8.954e-03  -7.058 1.74e-12 ***
## regwest                -1.231e-02  9.730e-03  -1.265    0.206
## raceother               2.270e-01  1.421e-02  15.974  < 2e-16 ***
## racewhite               2.392e-01  1.258e-02  19.011  < 2e-16 ***
## degyes                  6.419e-02  1.079e-02   5.948 2.76e-09 ***
## emp                     3.759e-04  4.440e-05   8.465  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5158 on 24811 degrees of freedom
## Multiple R-squared:  0.3409, Adjusted R-squared:  0.3406
## F-statistic:  1166 on 11 and 24811 DF,  p-value: < 2.2e-16
```

Below you will find some performance statistics related to the final model.

```
## [1] "AIC - 37591.0523 R^2 - 0.3409 (R^2)_adj - 0.3406 MSPR(CV) - 0.259"
```

III. Research Question

Now that we have a good model, let us try to answer the research questions.

First, when asking “Do African American males have statistically different wages compared to Caucasian males?”, we are actually trying to test the following hypothesis: $H_0 : \beta_{racewhite} = 0$ vs. $H_A : \beta_{racewhite} \neq 0$. We can run this test at level of significance, $\alpha = 0.05$. To evaluate this hypothesis we need the T-statistic and p-value related to the indicator variable ‘racewhite’ from the summary above.

```
## [1] "The T-statistic for the hypothesis above is: 19.011. The associated p-value is approximately: 0."
```

Since $p - value \leq \alpha\ (0 \leq 0.05)$, we reject $H_0 : \beta_{racewhite} = 0$ and can conclude that African American males have statistically different wages compared to Caucasian males.

Second, when asking “Do African American males have statistically different wages compared to all other males?”, we can come up with a new variable called raceCom that is “black” when race = “black” and “all_other” when race=“other” or when race=“white”. Then we rerun the model with raceCom instead of race. The hypothesis is : $H_0 : \beta_{allOther} = 0$ vs. $H_A : \beta_{allOther} \neq 0$. We can run this test at level of significance, $\alpha = 0.05$.

```
## [1] "The coefficient for I(race=allOther) is: 0.2368. The T-statistic for the hypothesis above is: 18.9908. Th
e associated p-value is approximately: 0."
```

Since $p - value \leq \alpha\ (0 \leq 0.05)$, we reject $H_0 : \beta_{allOther} = 0$ and can conclude that African American males have statistically different wages compared to all other males.

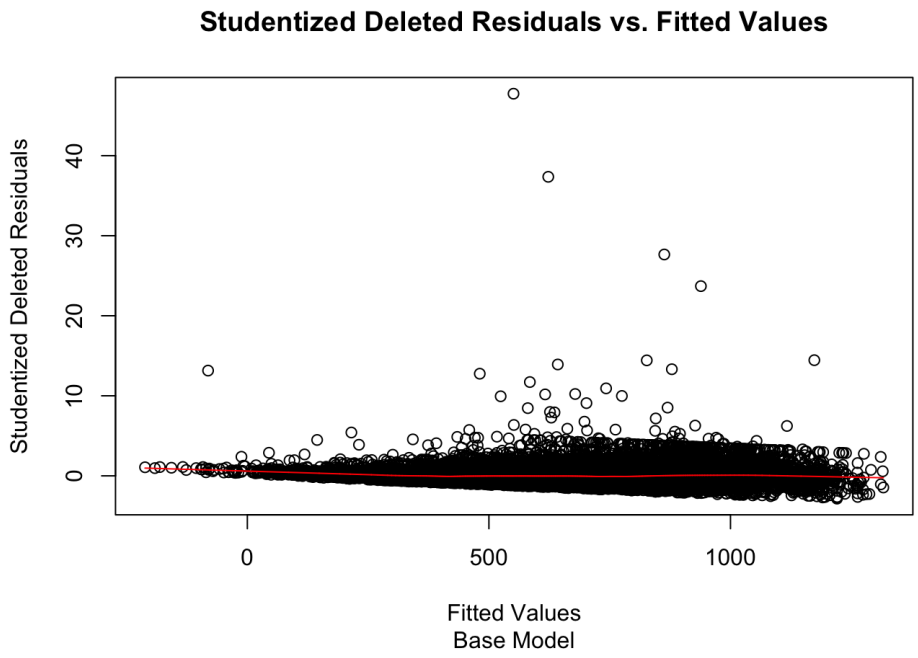
We can conclude that the average male wages are statistically different for the three race classes. Our results show that African Americans earn about 2.42% less than Caucasians, and about 2.40% less than all other races (Caucasian + Other), when we hold everything else constant.

IV. Appendix

a. Model Selection

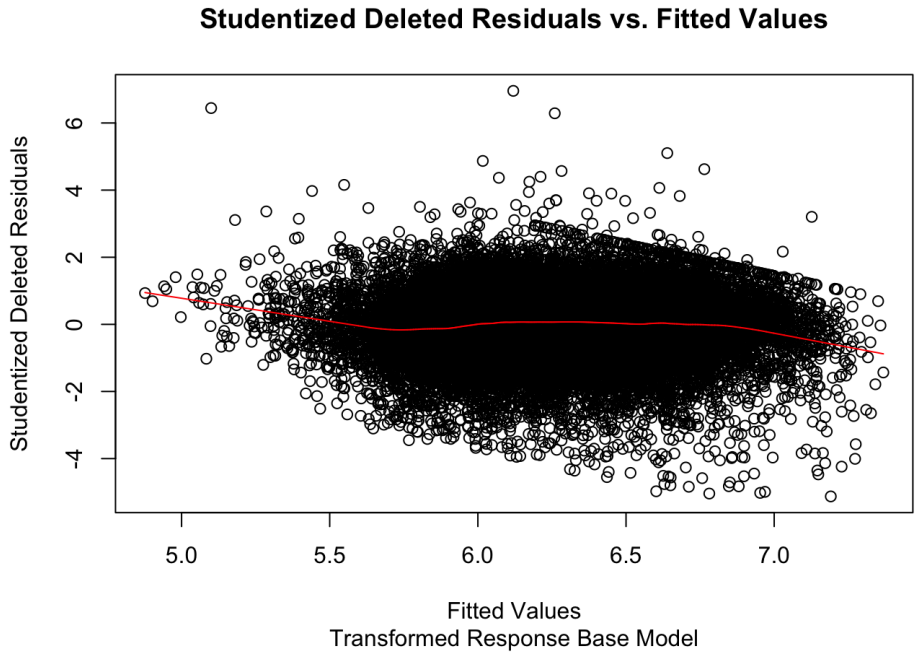
In our final model we need to determine if we need to transform our response variable, if any explanatory variables need to be transformed, if any explanatory variables can be eliminated, and if any interaction effects among explanatory variables need to be added in.

First let's start with a model that has no transformations, all the explanatory variables, and no added interaction variables. We will call this model the base model. Let's check our assumption of the errors having constant variance. To check this we will plot the studentized deleted residuals vs. fitted values for the base model. Let's also look at the adjusted R^2 value and the AIC of the model to establish a baseline against which we can judge improvements.




```
## [1] "The base model has an adjusted R^2 value of: 0.2181, and an AIC of 367789.3236."
```

The plot of the studentized deleted residuals vs. fitted values for the base model has a megaphone shape. This indicates that the errors have a non-constant variance, or are heteroscedastic. This violates our assumption of the errors having constant variance. To fix this let's attempt a natural logarithm transformation on our response variable and check the studentized deleted residuals vs. fitted value plot again. We will call this the transformed response base model, since it still includes all the explanatory variables, and we have not added any interaction variables. Let's also look at the adjusted R^2 value and the AIC of this transformed response base model to see if we have made any improvements.

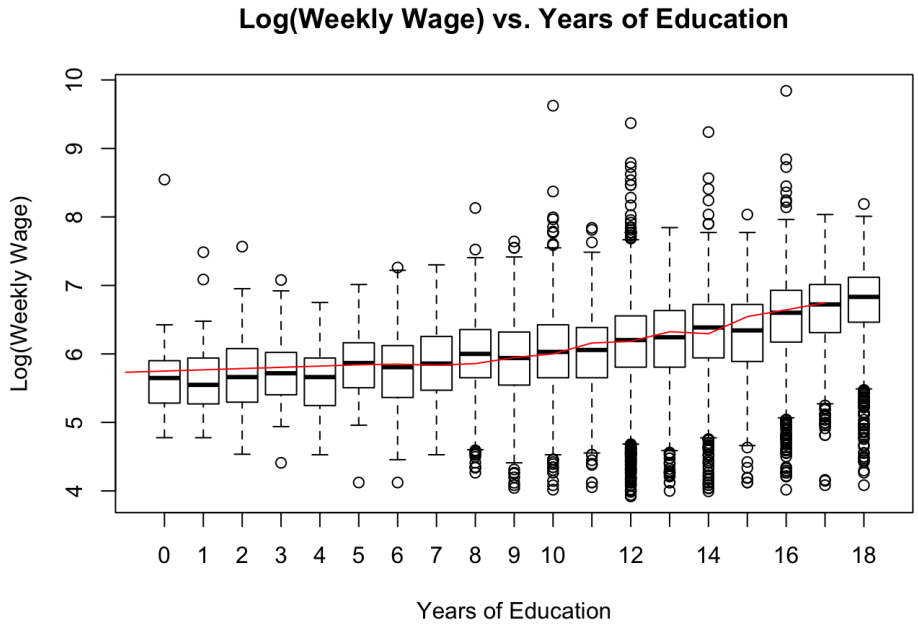



```
## [1] "The transformed response base model has an adjusted R^2 value of: 0.2892, and an AIC of 39454.0771."
```

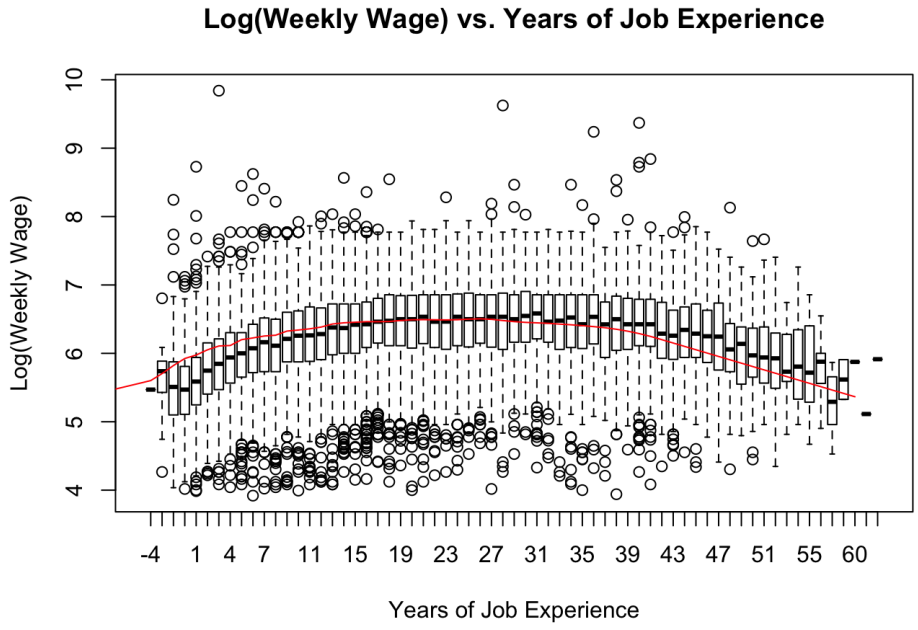
The plot of the studentized deleted residuals vs. fitted values for the transformed response base model has a random shape, with no pattern. This indicates that the errors have a constant variance, or are homoscedastic. This means that by using the natural log of the response variable we have met our assumption of the errors having constant variance. Also note that both the adjusted R^2 and AIC have improved from our base

model. In any further analysis we will use log(wage) as our response variable and in our final model the functional form of the response variable will be log(wage).

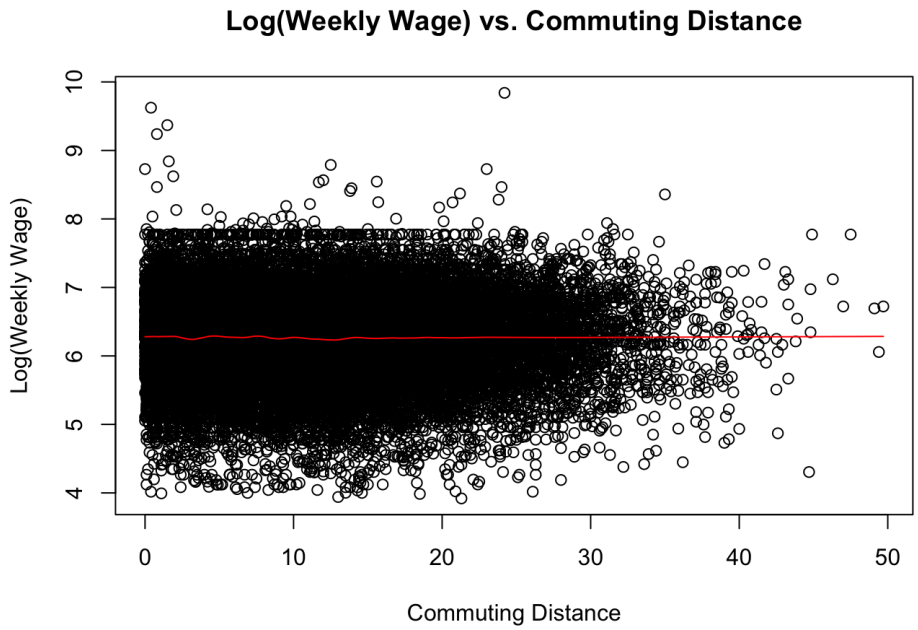
Next, let's see if any of the numeric explanatory variables need to be transformed.



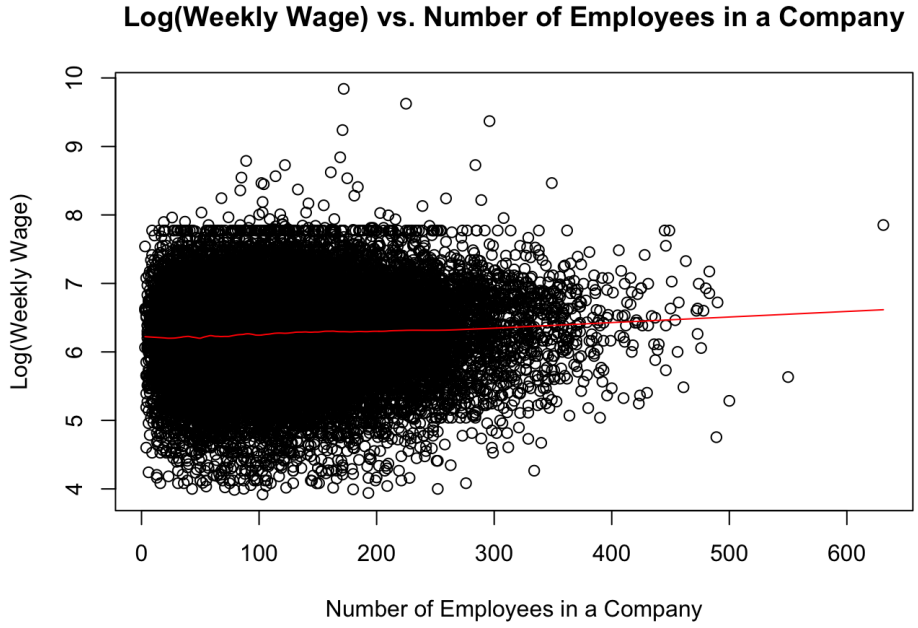
The smoother has a clear positive slope, and we can see there is an increasing relationship between years of education and log(weekly wage). There do not seem to be diminishing returns with this variable, so I believe leaving years of education untransformed is appropriate.



The smoother clearly has a curvilinear shape, with what seems like 1 inflection point around the 25 years of job experience mark. I believe adding in a second order years of job experience covariate could be appropriate.



The smoother seems to be horizontal, or have a slope of 0. This indicates that we should not transform commuting distance. In fact, this variable is a candidate for elimination



While number of employees in a company is a discrete variable in the sense that you cannot have 100.5 employees in a company, there are far too many unique values for that variable for a multiple-value boxplot to be appropriate. The slope of the smoother seems slightly positive. I believe we can leave number of employees in a company as is.

The plots of Log(Wage) vs. the numeric variables in our data has given us 4 insights:
1. Leave edu as is.
2. Add a second order term for exp.
3. Consider eliminating com.
4. Leave emp as is.

Let's compare two models, one where we incorporate all the insights above. We will call this candidate model 1. Let's also create a model where we do not eliminate com, we will call this candidate model 2. Let's also look at the adjusted R^2 value and the AIC of these 2 candidate models to see if we have made any improvements.

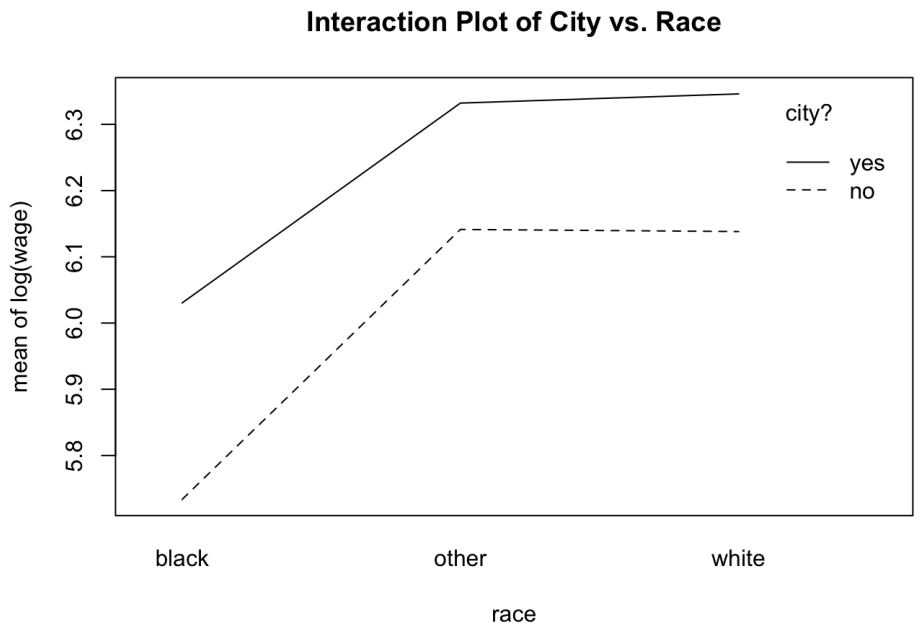
```
## [1] "Candidate model 1 has an adjusted R^2 value of: 0.3406, and an AIC of 37591.0523."
```

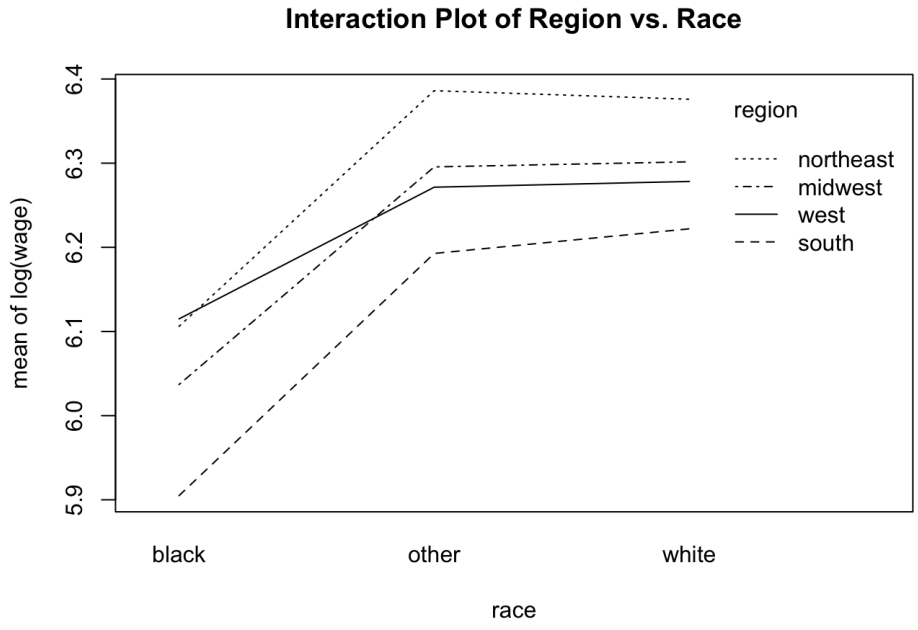
```
## [1] "Candidate model 2 has an adjusted R^2 value of: 0.3406, and an AIC of 37591.9949."
```

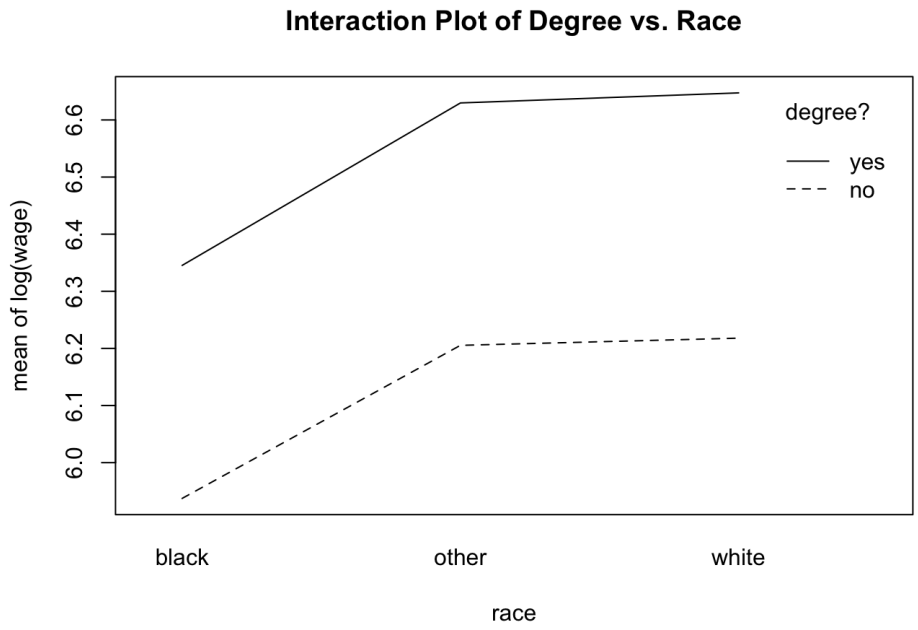
```
## [1] "The t-value for the coefficient 'com' from candidate model 2 is: -1.0281, and the associated p-value is:
0.3039."
```

Both candidate model 1 and 2 are improvements on the transformed response base model, but candidate model 2 includes the variable commuting distance (com), whereas candidate model 1 does not. We can see that the adjusted R^2 and AIC are basically the same for both models, so including the variable com is not really improving our model's performance. Furthermore, the coefficient for com from candidate model 2 is not statistically different from 0 for any reasonable level of significance (p-value = 0.3), so we can safely eliminate the variable com and proceed with candidate model 1.

Next, let us investigate if any categorical, interaction effects between race and the other variables needs to be included.







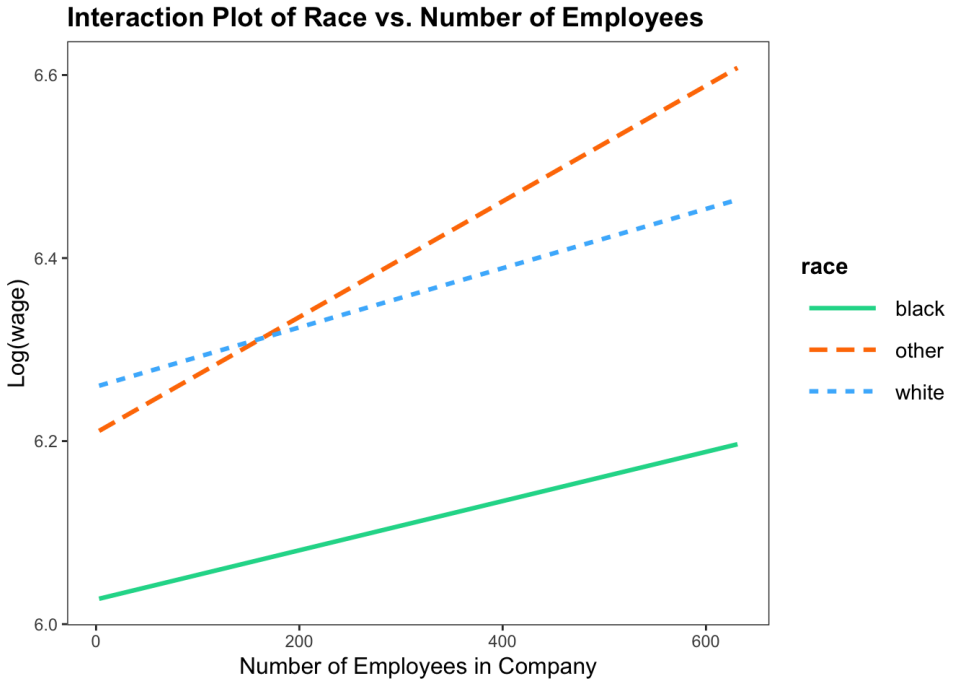
The only interaction plot that suggests any sign of interaction is the one between region and race. Let's add a race*reg interaction effect to candidate model 1 and see if we get any improvements in performance. We will call this model candidate model 3.

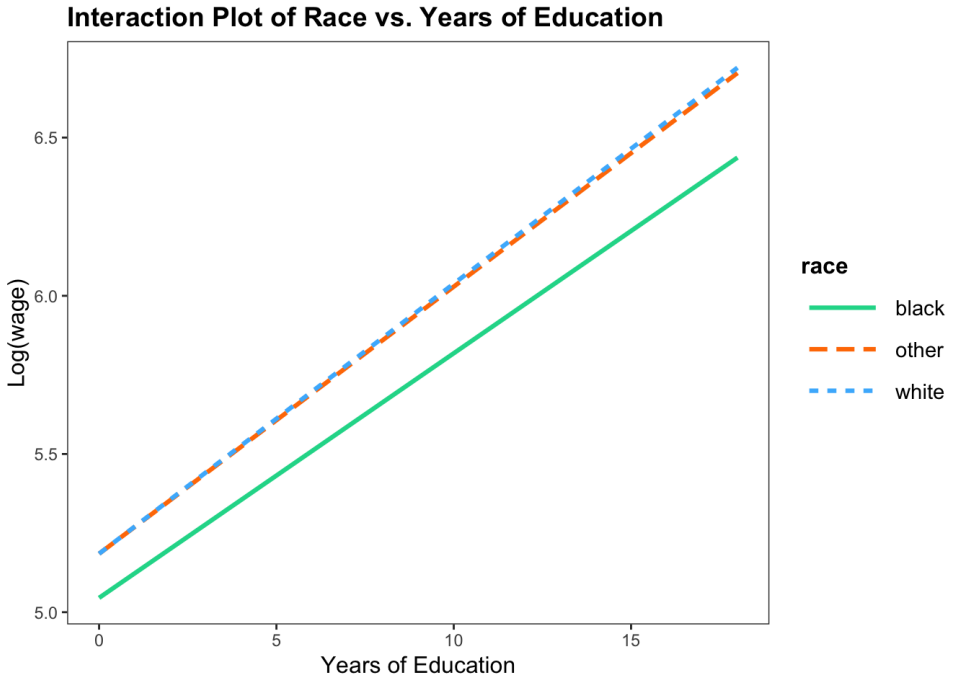
```
## [1] "Candidate model 3 has an adjusted R^2 value of: 0.3405, and an AIC of 37600.8293."
```

```
##                            Estimate Std. Error    t value  Pr(>|t|)
## regnortheast:raceother -0.017270645 0.04671133 -0.3697314 0.7115858
## regsouth:raceother     -0.018991970 0.03913592 -0.4852823 0.6274804
## regwest:raceother      -0.063437440 0.05454793 -1.1629670 0.2448541
## regnortheast:racewhite -0.022543264 0.04275288 -0.5272923 0.5979954
## regsouth:racewhite     -0.005438356 0.03482882 -0.1561453 0.8759198
## regwest:racewhite      -0.051310849 0.05103491 -1.0054068 0.3147108
```

Candidate model 3 has about the same adjusted R^2 as compared to candidate model 1, but it has a higher AIC. Also note that none of reg-race coefficients are statistically different from 0 (minimum p-value is 0.244). Thus we can safely not include the reg*race interaction effect.

Let's also investigate if an interaction effects should be added between race and years of education, and between race and number of employees in a company. The model that includes an emp-race interaction on top of candidate model 1 will be called candidate model 4. The model that includes an edu-race interaction on top of candidate model 1 will be called candidate model 5.





It seems that it might be useful to include the interaction between number of employees in a company and race. Let's see how candidate model 4 performs, to see if we are improving on top of candidate model 1.

```
## [1] "Candidate model 4 has an adjusted R^2 value of: 0.3407, and an AIC of 37587.4894."
```

We see a small reduction in AIC and a small increase in adjusted R^2 as compared to candidate model 1, so we should consider including the interaction emp*race in our final model.

Lastly, let us look for any sign of multicollinearity in our numeric explanatory variables for candidate model 4 to see if we should be centering or eliminating any variables.

```
##                 edu poly(exp, degree = 2)                city
##            1.674991              1.064556            1.032007
##                 reg                  race                 deg
##            1.019798              3.809903            1.505831
##                 emp             race:emp
##           12.918193              7.749248
```

```
## [1] "The means of the variance-inflation factors for candidate model 4 is: 3.85."
```

Candidate model 4 suffers from multicollinearity, notice the generalized variance inflation factor for the variable 'emp' is 12.92. Let's see the variance inflation factors for candidate model 1, which does not include the race*emp interaction effect.

```
# variance-inflation factors candidate model 1
gvif.canMod1 <- (vif(canMod1)[,3])^2
meanGvif.canMod1 <- mean(gvif.canMod1)
print(gvif.canMod1)
```

```
##                  edu poly(exp, degree = 2)                 city
##             1.674850              1.064451             1.032001
##                  reg                  race                  deg
##             1.019755              1.020842             1.505743
##                  emp
##             1.000846
```

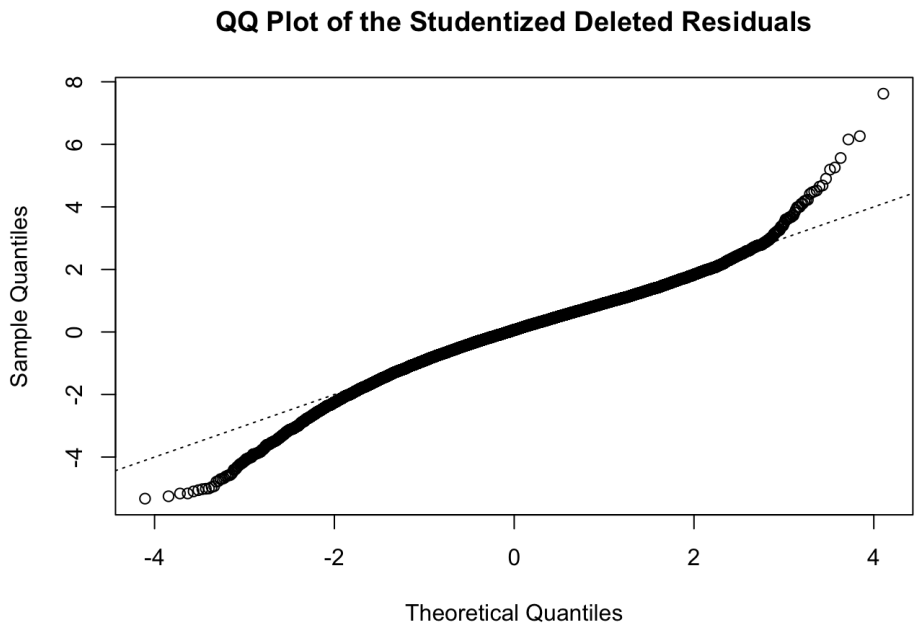
```
print(paste0("The means of the variance-inflation factors for candidate model 1 is: ", round(x = meanGvif.canMod
1, digits = 2), "."))
```

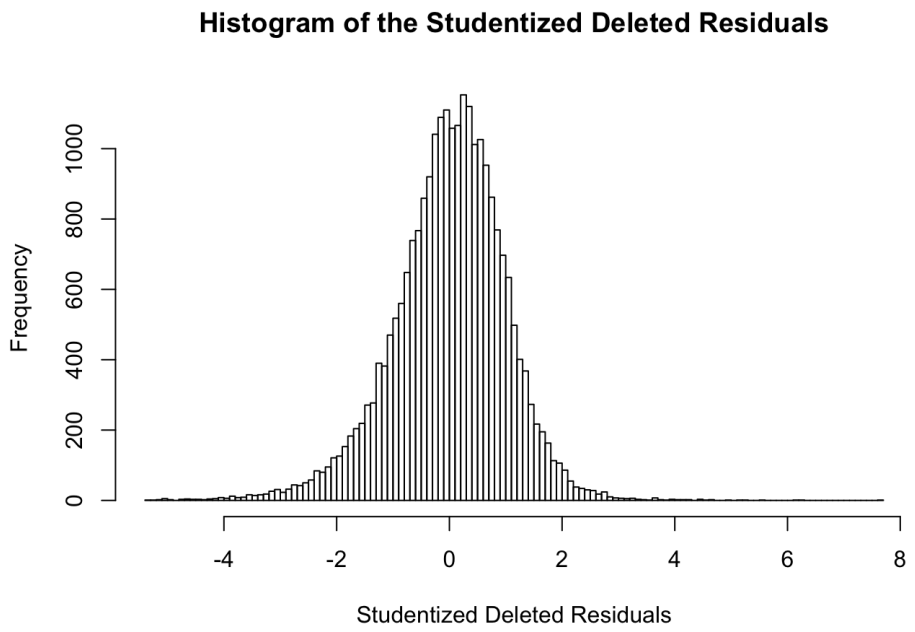
```
## [1] "The means of the variance-inflation factors for candidate model 1 is: 1.19."
```

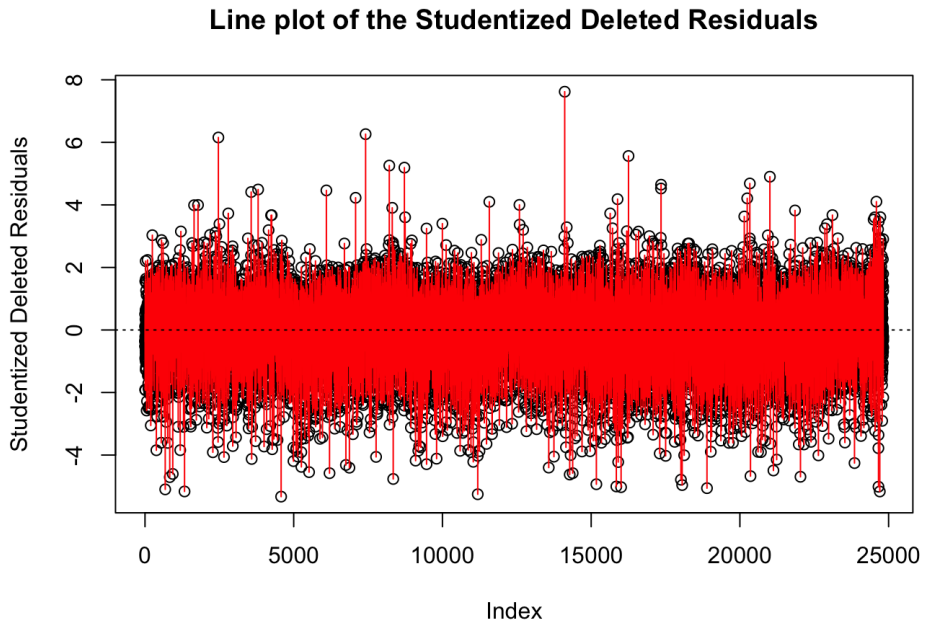
The mean is not much greater than 1 for generalized variance-inflation factors for candidate model 1, notice the generalized variance inflation factor for the variable 'emp' is 1.00. The benefit of choosing candidate model 4 over candidate model 1 is you get a decrease in AIC of 4 (which is a decrease of less than 0.0001%) and you get an increase in the adjusted R^2 of about 0.02%. However, the cost of choosing candidate model 4 over candidate model 1 is the huge amount multicollinearity introduced into the model, which is making the standard error of the 'emp' coefficient quite large. I believe, the cost of including the race*emp interaction outweights the benefit of including it, so we should go back to candidate model 1.

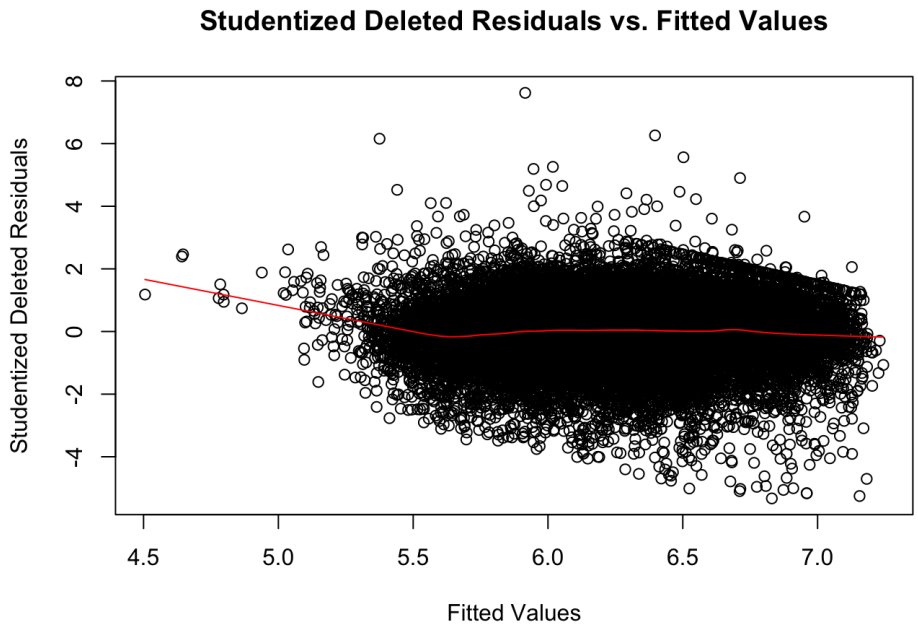
b. Diagnostics and Model Validation

The final model is candidate model 1. Let's run all the diagnostics on this model to make sure we are not violating any assumptions of the linear regression model.



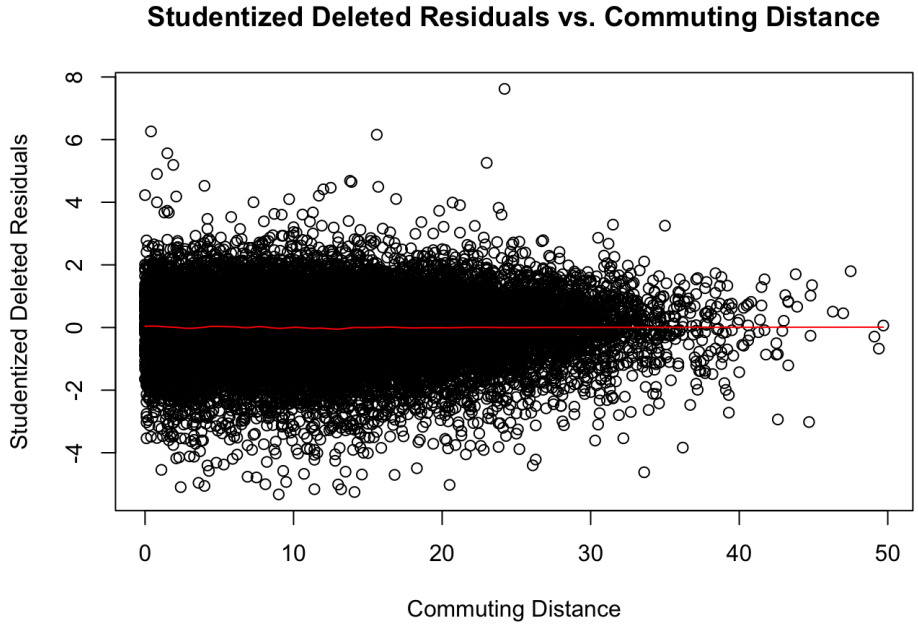

Histogram of the Studentized Deleted Residuals
Frequency
0
200
400
600
800
1000
-4
-2
0
2
4
6
8
Studentized Deleted Residuals


Line plot of the Studentized Deleted Residuals
Studentized Deleted Residuals
-4
-2
0
2
4
6
8
0
5000
10000
15000
20000
25000
Index



The normal probability plot and histogram of the studentized deleted residuals indicate a slight negative skew, but not enough to indicate a departure from normality. The line plot indicates that the errors are independent and identically distributed, there seems to be no pattern in that plot. The plot of the studentized deleted residuals against the fitted values indicates we have constant variance, or homoscedasticity.

From a model validation perspective, lets plot the studentized deleted residuals against the omitted explanatory variable 'com', or commuting distance.



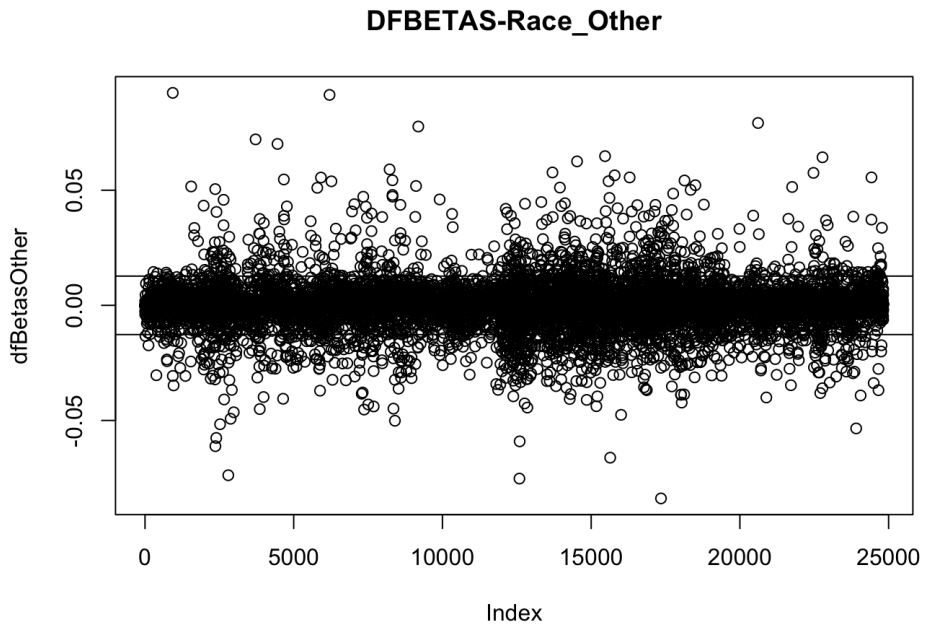
The plot doesn't seem to have any pattern, so we have validated our decision to remove the variable 'com'.

Next, let us run a 20-fold CV and compute the mean square prediction error, or MSPR(CV). We want to make sure that MSPR(CV) is close to the MSE of our final model.

```
## [1] "The MSE from our final model is: 0.2659. The MSPR(CV) for our 20-fold CV is: 0.259."
```

The MSPR(CV) shows that our final model fits out-of-sample data points just as well as in-sample data points. This shows that our model will generalize well to other data sets gathered from the same population.

Let's also look at any influential observations in this data. First, lets look at the influential observations in relation to the 'race-other' indicator variable.



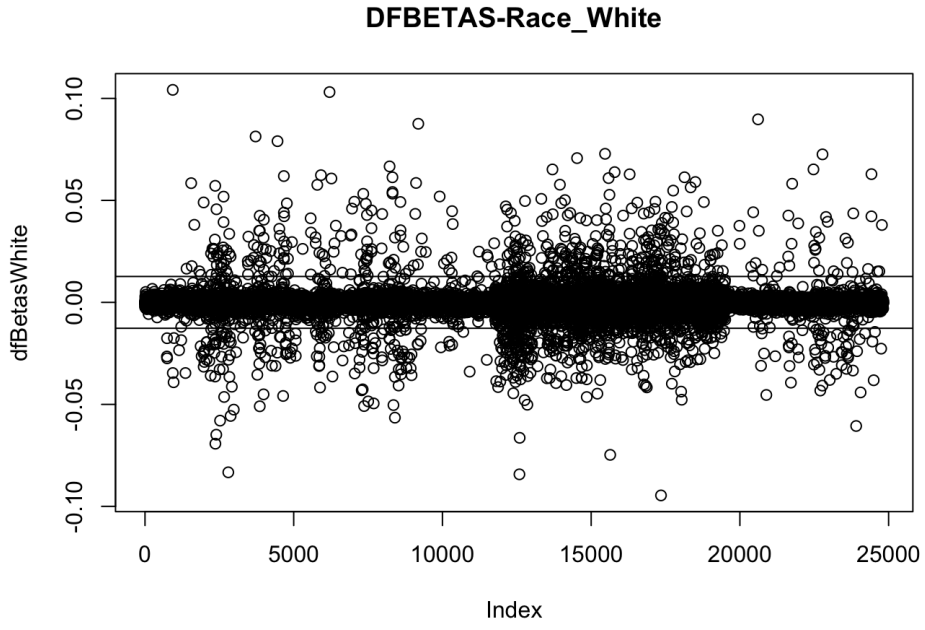
Seems like there a large number of influential observations in relation to their effect on the Race-Other indicator variable. However, we need to judge how influential these observations actually are by looking at the Race-Other coefficient and standard error when those observations are included and when they are not.

```
## [1] "The Race-Other coefficient when the influential observations are included is: 0.227, the standard error f
or that coefficient is: 0.0142."
```

```
## [1] "The Race-Other coefficient when the influential observations are not included is: 0.2537, the standard er
ror for that coefficient is: 0.0171."
```

We can see that the point estimates and standard errors for the race-other slope are very close to each other, regardless of whether or not we include the influential observations. Thus we can safely include these influential obbservations in the data as they are not changing the slope and standard error by much.

Second, lets look at the influential observations in relation to the 'race-white' indicator variable.



Seems like there a large number of influential observations in relation to their effect on the Race-White indicator variable. However, we need to judge how influential these observations actually are by looking at the Race-White coefficient and standard error when those observations are included and when they are not.

```
## [1] "The Race-White coefficient when the influential observations are included is: 0.2392, the standard error for that coefficient is: 0.0126."
```

```
## [1] "The Race-White coefficient when the influential observations are not included is: 0.2353, the standard er ror for that coefficient is: 0.0172."
```

We can see that the point estimates and standard errors for the race-white slope are very close to each other, regardless of whether or not we include the influential observations. Thus we can safely include these influential obbservations in the data as they are not changing the slope and standard error by much.