HW 5: Writeup

Christopher Rusnak (UNI: cjr2176)

November 28, 2016

I. Introduction

We have been provided data on 24,823 males who work full time in the United States and whose ages range from 18 to 70. Our dataset includes the following covariates:

1. Weekly wages (in US dollars)
2. Number of years of school
3. Indicator of graduating college (*yes* or *no*)
4. Indicator of being employed close to or within a city (*yes* or *no*)
5. US region (*west*, *midwest*, *south*, or *northeast*)
6. race (*black*, *white*, or *other*)
7. Travel distance from residence to employer
8. Total company employees
9. Number of years of employment

There are two goals of this case study:

1. Formulate a linear regression model that includes the most indicative variables, interaction terms, and functional expressions of the input variables.
2. Apply this model to answer the following research questions:
    a. Is there a statistically significant difference between wages for African American men and wages for Caucasian men?
    b. Is there a statistically significant difference between wages for African American men and wages for all other men?

Below are dataset summary statistics for wage and race:

	wage
1	Min. : 50.39
2	1st Qu.: 356.13
3	Median : 546.06
4	Mean : 637.82
5	3rd Qu.: 830.96
6	Max. :18777.20

	race
1	black: 1934
2	other: 4584
3	white:18305

On average, full-time US employees in this dataset earn $637.82 per week, or about $33166.64 per year. The first and third quartiles of weekly wages are within $300 of this average, indicating that there are relatively few outiers or extreme values.

These summary statistics indicate that African American males comprise about 7.8% of full-time employees, and are the smallest group in the dataset. However, they collectively only have 5.8% of the total wages, which is smaller than their proportionality in the dataset. This result may be an indicator of wealth inequality between different races, which will be formally tested in Section III. This point is further exemplified by the following Pareto chart on distribution of weekly wages.

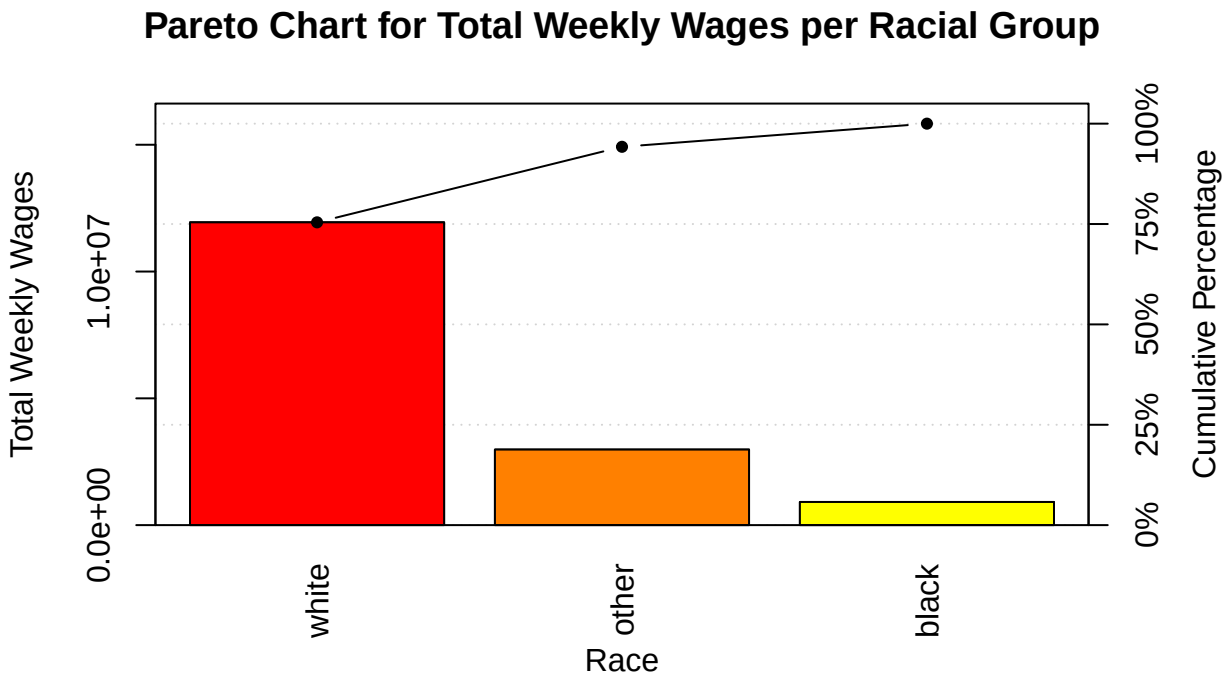

Figure 1: Salary Distribution

To begin to answer the research questions, we look at the distribution of wages for each race. Figure 2 shows overlaid histograms and boxplots of wages for black males and white males. They demonstrate that white males have higher proportions of individuals with higher wages than black males.

Figure 3 shows overlaid histograms and boxplots of wages for black males and all other males. They also show that all other males have higher proportions of individuals with higher wages than black males.

II. Statistical Model

A linear regression model was built to predict weekly wage values. For the final model, the following variables were selected:

1. Number of years of school (*edu*)
2. Number of years of employment (*exp*)
3. Indicator of being employed close to or within a city (*city*)
4. US region(*reg*)
5. Racial group (*race*)
6. Indicator of graduating college (*deg*)
7. Interaction between *race* and *reg*

Figure 2: Wage Distributions of Black and White Males

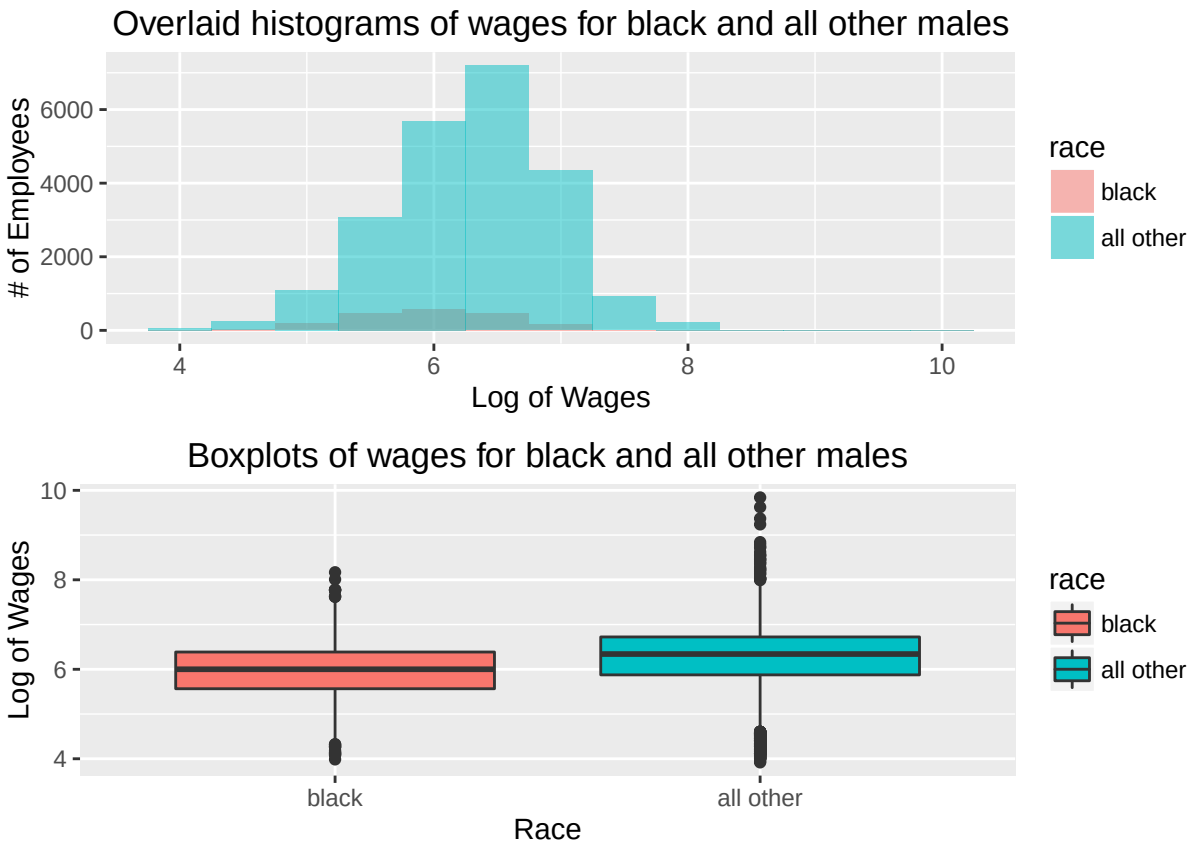

Figure 3: Wage Distributions of Black and all Other Males

8. Interaction between *city* and *reg*
9. Interaction between *edu* and *race*
10. Interaction between *edu* and *city*
11. Interaction between *edu* and *reg*
12. Interaction between *edu* and *deg*

In addition, a natural logarithm transformation was applied to the wages. No functional transformations were applied to any of the covariates. The summary output of the model is as follows:

lm(formula = log(wage) ~ edu + exp + city + reg + race + deg + race * reg + city * reg + edu * race + edu * city + edu * reg + edu * deg, data = train.data)

[1] "F statistic value: 323.983730791891"

[1] "F statistic numerator: 25"

[1] "F statistic denominator: 19832"

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.5921	0.0891	51.52	0.0000
edu	0.0708	0.0064	11.09	0.0000
exp	0.0184	0.0003	57.22	0.0000
cityyes	0.0235	0.0453	0.52	0.6037
regnortheast	-0.0354	0.0745	-0.48	0.6340
regsouth	-0.1787	0.0628	-2.85	0.0044
regwest	-0.0213	0.0807	-0.26	0.7915
raceother	0.2183	0.0840	2.60	0.0094
racewhite	0.2314	0.0743	3.11	0.0019
degyes	0.0693	0.1726	0.40	0.6883
regnortheast:raceother	0.0000	0.0547	0.00	0.9994
regsouth:raceother	-0.0468	0.0460	-1.02	0.3090
regwest:raceother	-0.0506	0.0637	-0.79	0.4274
regnortheast:racewhite	-0.0166	0.0499	-0.33	0.7387
regsouth:racewhite	-0.0265	0.0410	-0.65	0.5173
regwest:racewhite	-0.0472	0.0597	-0.79	0.4288
cityyes:regnortheast	-0.0430	0.0281	-1.53	0.1268
cityyes:regsouth	-0.1132	0.0229	-4.94	0.0000
cityyes:regwest	-0.0836	0.0252	-3.32	0.0009
edu:raceother	0.0035	0.0058	0.59	0.5542
edu:racewhite	0.0026	0.0052	0.51	0.6113
edu:cityyes	0.0159	0.0032	4.95	0.0000
edu:regnortheast	0.0081	0.0042	1.93	0.0540
edu:regsouth	0.0167	0.0038	4.34	0.0000
edu:regwest	0.0096	0.0040	2.37	0.0177
edu:degyes	-0.0019	0.0103	-0.18	0.8534

The final model statistics on the model building dataset are as follows:

- AIC: 31577.94
- R^2: 0.29
- R_a^2: 0.289

III. Research Question

To answer each research question, we turn to the appropriate statistical hypothesis testing procedures.

For the first research question, we want to know if the average weekly wage differs between African American males and Caucasian males. We subset the data to focus on these two groups. We establish the null hypothesis that the average weekly wages are the same per group, against the alternate hypothesis that average weekly wages differ between groups. We perform one-way ANOVA to test this pair of hypotheses, at a significance level of 0.05.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
black	1	56713844.29	56713844.29	275.60	0.0000
Residuals	20237	4164404694.01	205781.72		

Based on the low p-value (1.7e-61) of the F-score (275.6), we therefore reject the null hypothesis at this significance level.

For the second research question, we are testing whether mean weekly salary is the same between African American males and all other males. In this scenario, Caucasians and other racial groups (categorized as "*other*" in the dataset) will now be grouped into a single category of "*all other*". The null hypothesis is that the average weekly wages are the same between African Americans and all others. The alternate hypothesis is that average weekly wages differ between groups. As before, we perform one-way ANOVA to test these hypotheses, setting the significance level to be 0.05.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
black	1	57524328.03	57524328.03	285.72	0.0000
Residuals	24821	4997185009.75	201328.92		

Based on the low p-value (9.7e-64) of the F-score (285.72), we therefore reject the null hypothesis at this significance level.

Both of these findings indicate that, as a whole, African American males earn less money per week than their Caucasian counterparts, as well as males from all other races. For further investigation, it would be helpful to have additional demographic and employment information about the employees, such as job title, sector (information technology, construction, etc.), employer name, employer location, number of years working for employer, and others. For a more comprehensive study, it would be worth including data from female full-time employees, as well as providing more specific race information about employees categorized as "other".

IV. Appendix

a. Model Selection

With a pseudo-random seed of 0, the dataset was partitioned into two subsets: approximately 80% was used for model training, while the remaining 20% was applied for model validation. To determine which covariates to include in the model, each independent variable was plotted against the corresponding wage values. To increase the spread of the data in the plots, we applied the natural logarithm transformation to the wage values. In addition, we also computed the correlation between the natural log of the wages and each of the numeric variables (*com*, *emp*, *exp*, and *edu*). As a rule of thumb, variables whose correlation coefficients were below 15% and did not have a clear relationship (linear or non-linear) with the log of the wages in their plots were not included in any of our models. Commuting distance and the number of employees had low correlation coefficients (0.189% and 5.99%, respectively), and this finding was supported by their scatter plots. As such, these two variables were removed from consideration. Figures 4 and 5 display the scatter plots and box plots that we produced in our exploratory data analysis.

Regarding interaction effects, we examined two-way interaction plots for combinations of covariates for which it would be reasonable to hypothesize that together have an effect on wages. For example, while living in or near a major city may be correlated with higher wages, how high those wages are may also depend on the region in which that city is located. We produced these plots for the following combinations of variables:

1. *edu* vs. *race*

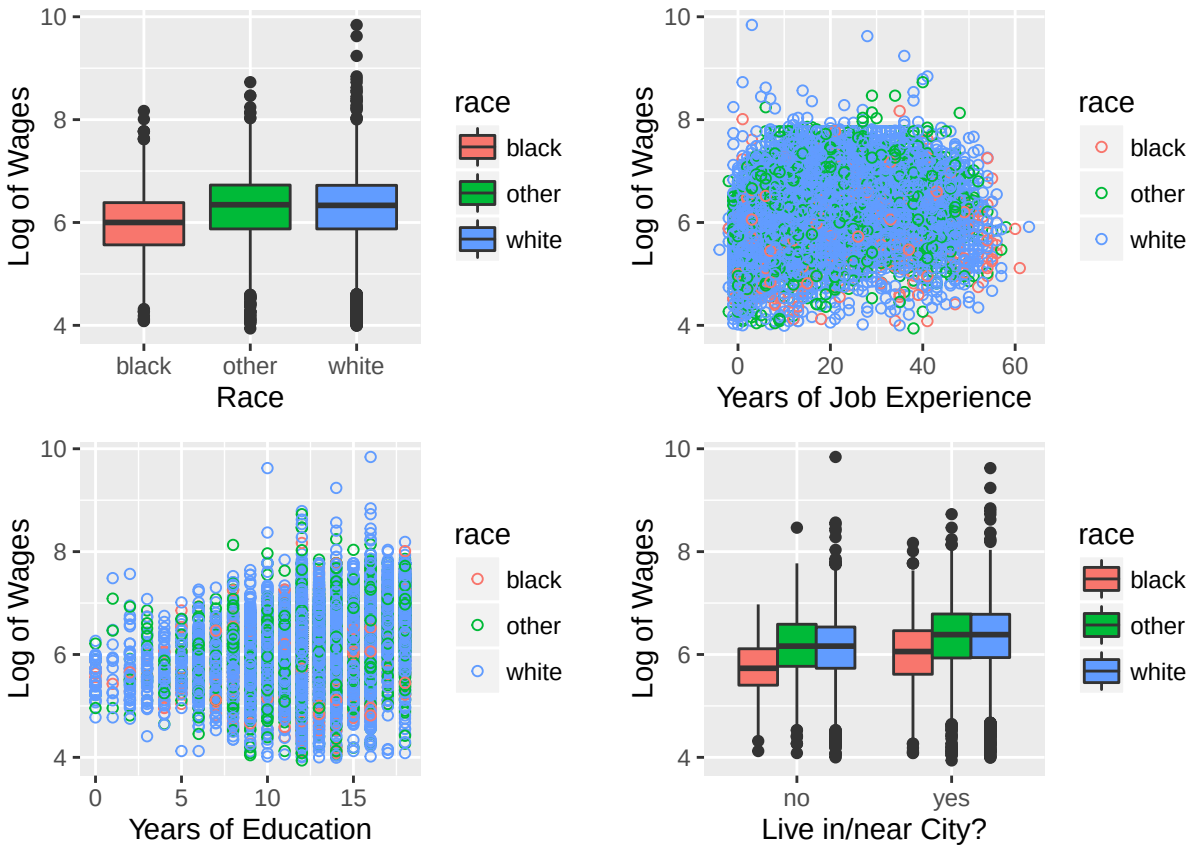

Figure 4: Basic Scatter Plots and Box Plots

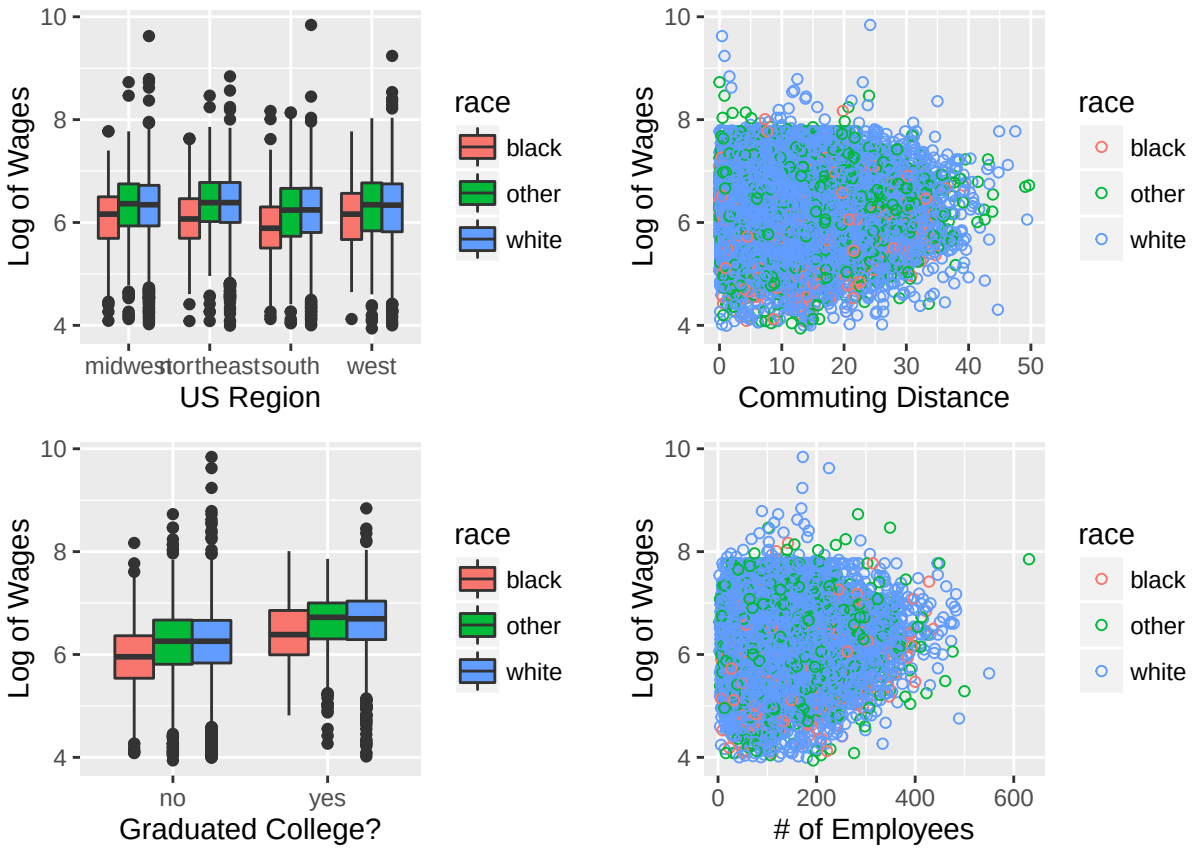

Figure 5: Basic Scatter Plots and Box Plots

2. *edu* vs. *city*
3. *edu* vs. *reg*
4. *edu* vs. *deg*
5. *race* vs. *reg*
6. *race* vs. *city*
7. *race* vs. *deg*
8. *city* vs. *reg*
9. *city* vs. *deg*
10. *reg* vs. *deg*

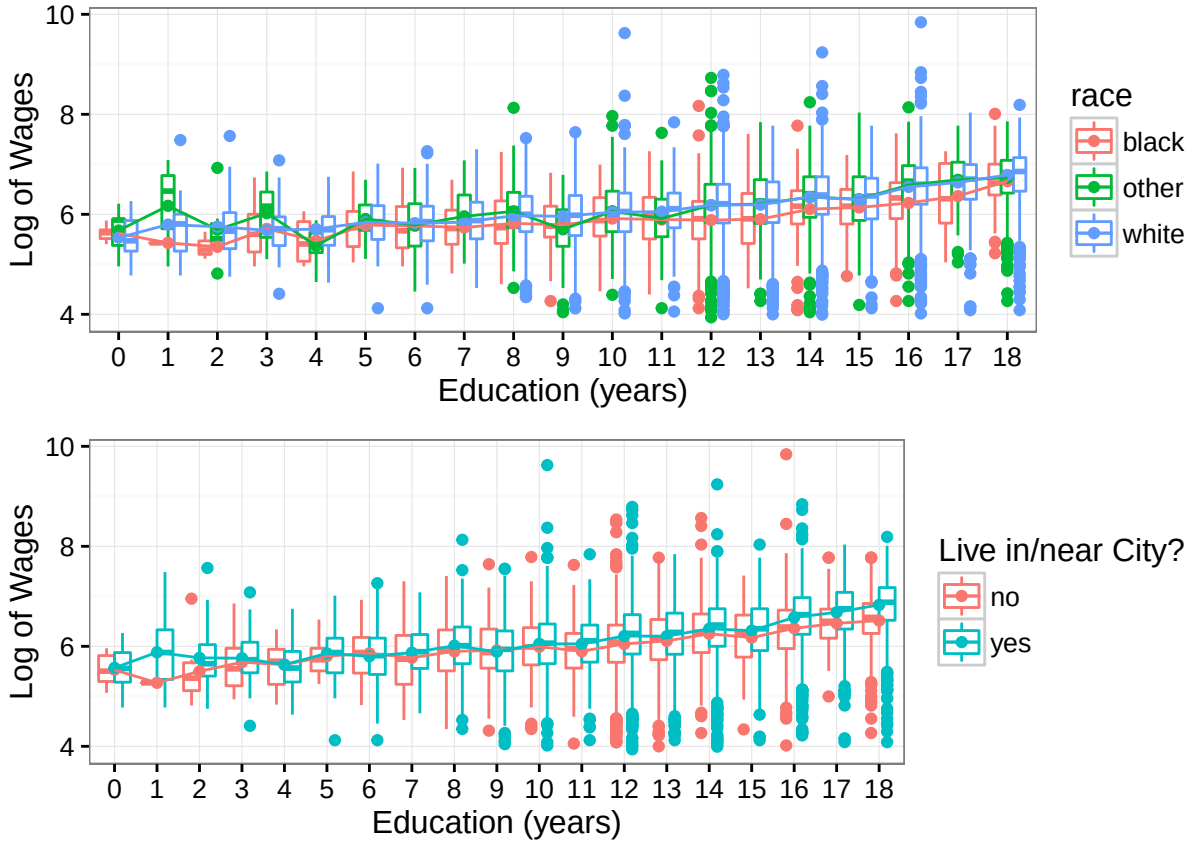

Figure 6: Interaction Plots

Figures 6 through 10 show these interaction plots. We determined that, of the above list of possible interaction terms, the combinations (6), (7), (9), and (10) do not have any apparent interaction effects, and so they were not considered in any of our models.

We produced a total of our feasible candidate linear regression models, henceforth labeled as *Model 1*, *Model 2*, *Model 3*, and *Model 4*.

Model 1 the simplest model and contained no interaction terms or variable transformations. It included the following covariates: *edu*,*exp*,*city*,*reg*,*race*, and *deg*. These are all of the covariates that we deemed to have a significant relationship with wages.

We applied a Box-Cox Transformation on the dependent variable (*wage*), resulting in a power transformation value of 0.14. As such, we decided to apply the natural logarithm transformation to the wage variable.

Model 2 is the same as the previous model, but has the aforementioned transformation of the dependent variable. Both of these models serve as effective baselines.

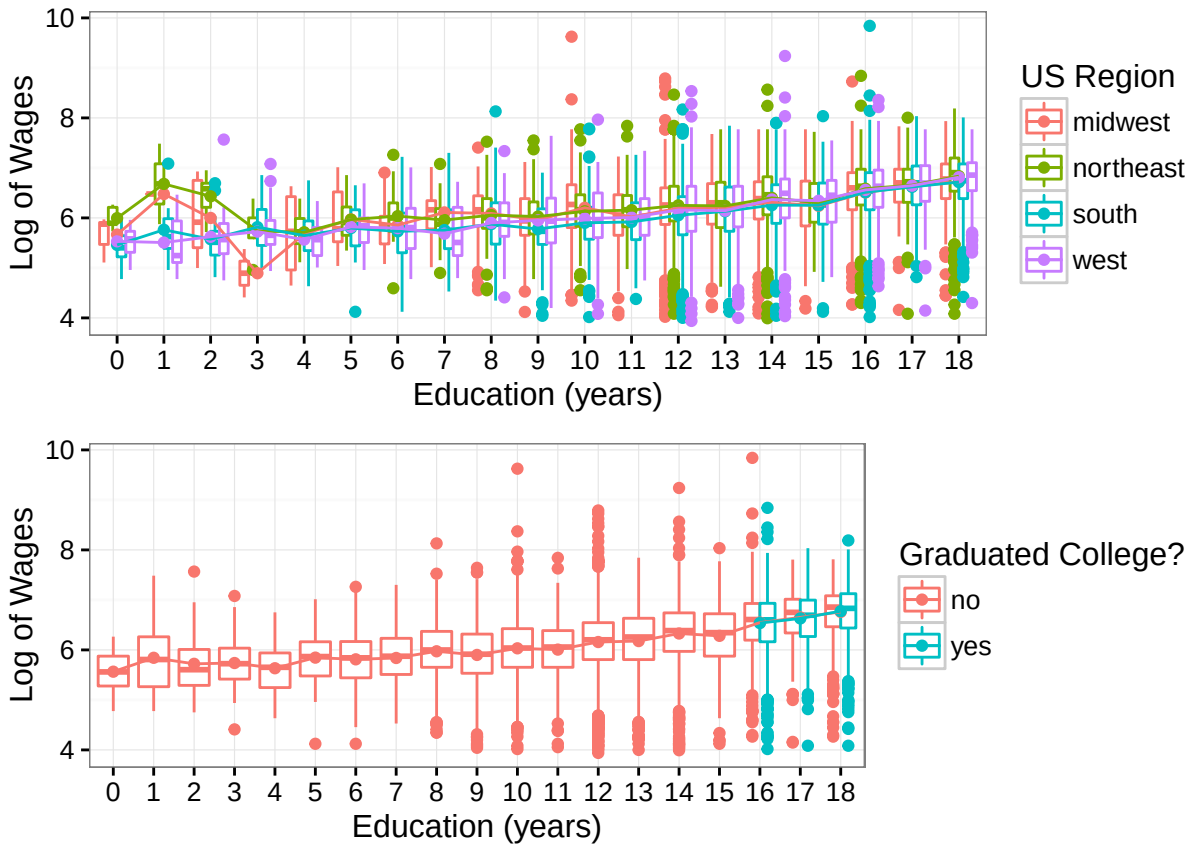

Figure 7: Interaction Plots

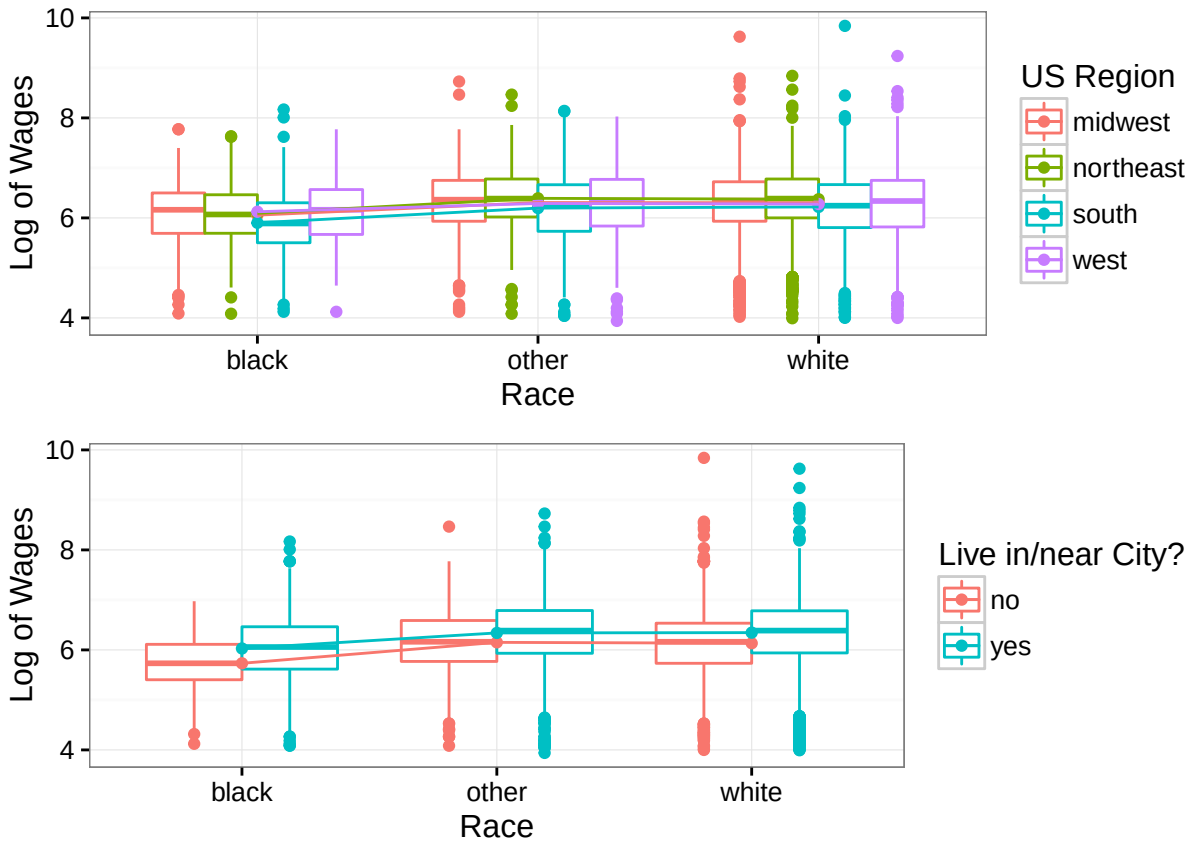

Figure 8: Interaction Plots

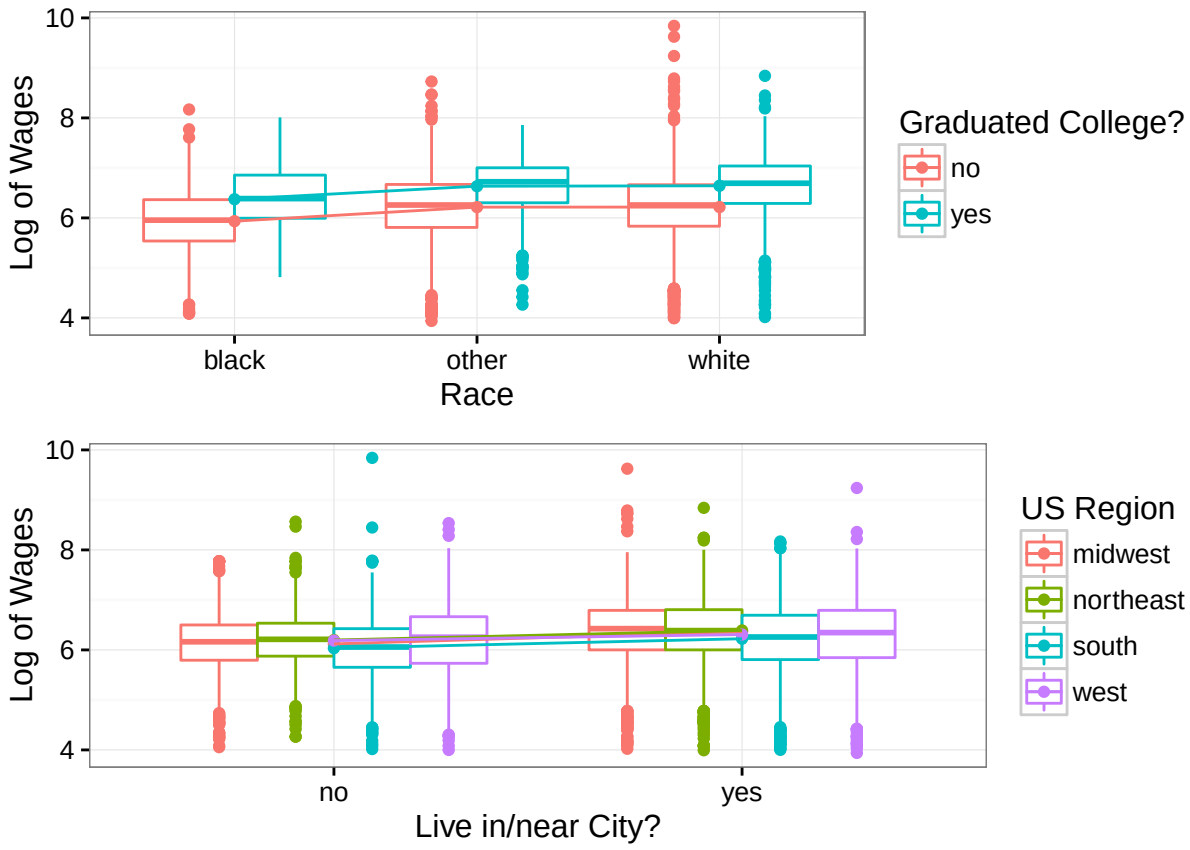

Figure 9: Interaction Plots

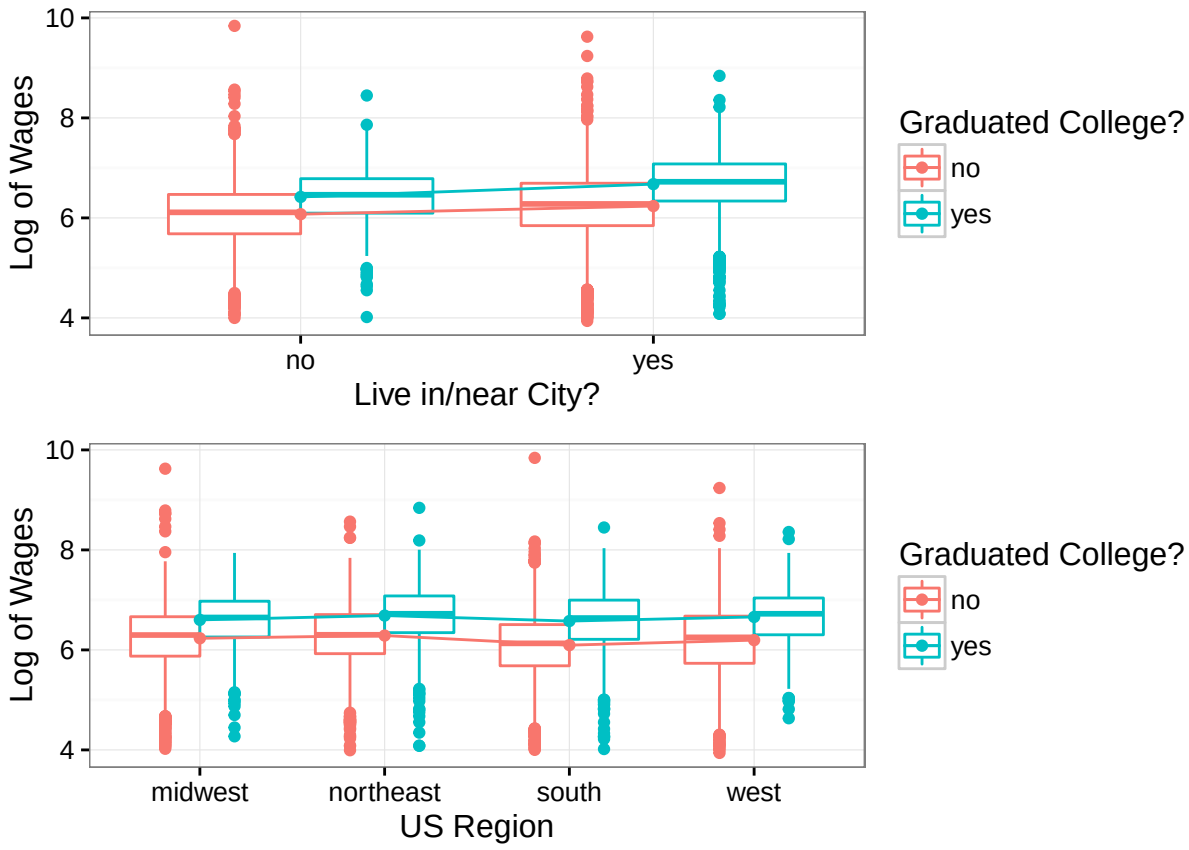

Figure 10: Interaction Plots

Model 3 adds to *Model 2* the interaction terms with respect to the region covariate (*race & reg* and *city & reg*). *Model 4* adds all of the interaction terms (*race & reg*, *city & reg*, *edu & race*, *edu & city*, *edu & reg*, and *edu & deg*). For each model, we calculated the *AIC*, *BIC*, R^2, and adjusted $R^2(R_a^2)$ on the model building dataset. These evaluation statistics are provided below:

	AIC	BIC	R_2	adj_R_2
Model 1	294827.50	294914.36	0.21	0.21
Model 2	31617.78	31704.64	0.29	0.29
Model 3	31607.22	31765.14	0.29	0.29
Model 4	31577.94	31791.15	0.29	0.29

Model 4 had the highest values of R^2 and R_a^2, along with the lowest value of *AIC*. Therefore, it is our choice for the final model.

b. **Diagnostics and Model Validation**

The following are diagnostic plots for the final model on the model building dataset. Figure 11 consists of the QQ plot of studentized deleted residuals; the histogram of studentized deleted residuals; a line plot of studentized deleted residuals; and a plot of studentized deleted residuals against predicted values.

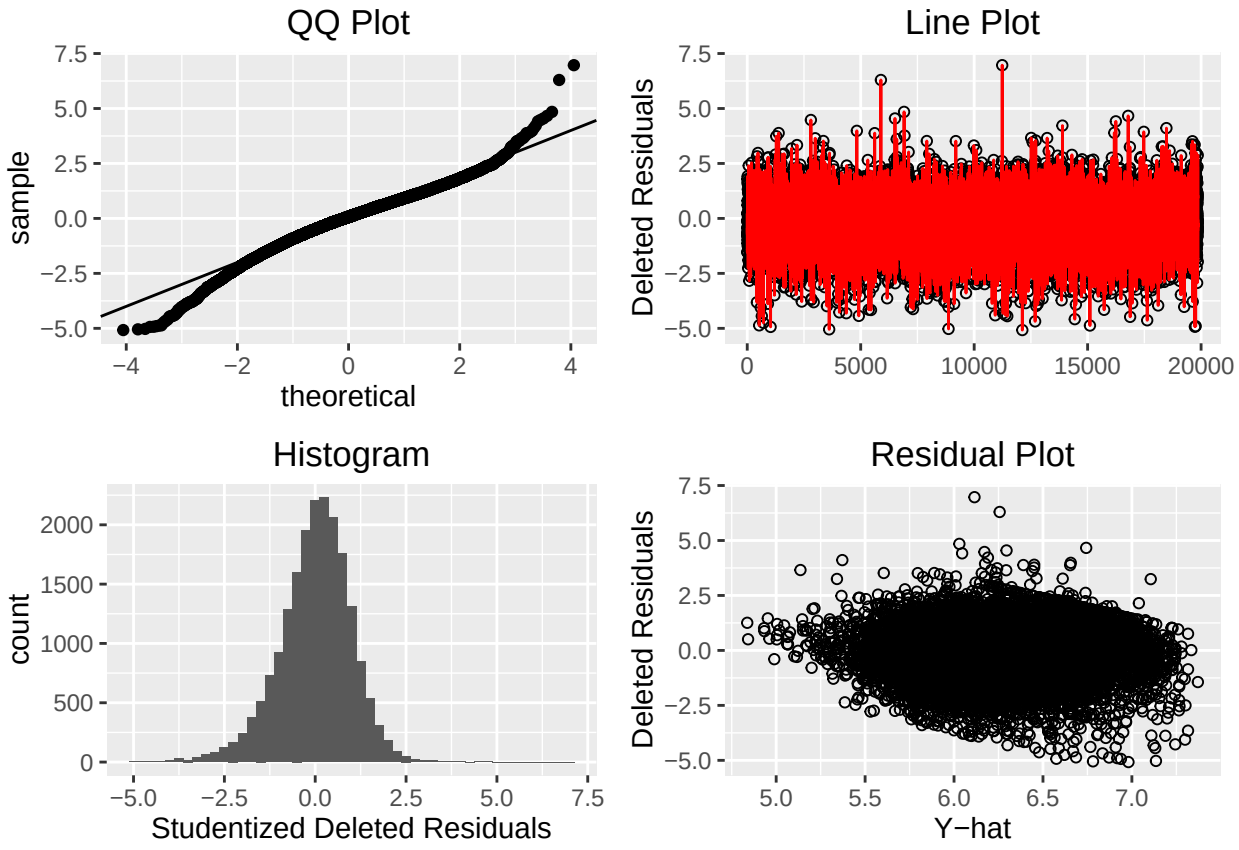

Figure 11: Diagnostic Plots: Training Set

We evaluate this model on the validation dataset and compute the mean square prediction error (*MSPE*) and mean square error (*MSE*).

The *MSPE* on the model validation dataset is 148045.52 while the *MSE* on the model building dataset is 171123.09. The *MSPE* is relatively close to the *MSE*, indicating that the model has reasonable predictive power.

c. **Influential Observations and Collinearity**

For our purposes, we are interested in constructing an inferential model (as opposed to predictive) that allows us to better understand the relationship between various demographic factors and salary for employees. Outliers, extreme values, and other influential observations would have a strong impact in the interpretability of that model. We would be most interested in using the DFFITS measure for identifying these observations, because that metric takes into account the impact of each observation on each fitted value the entire model. Figure 12 plots this measure for each data point in the model building dataset.

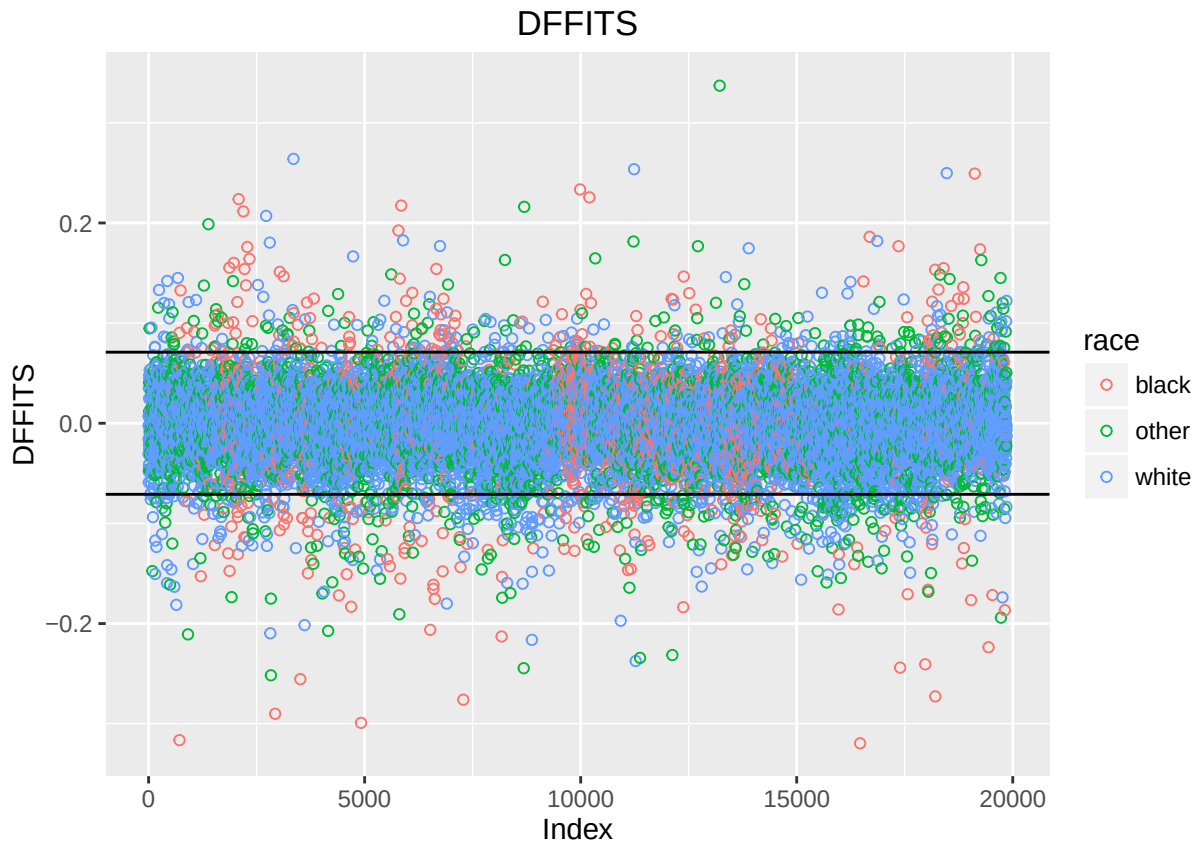

Figure 12: DFFITS Measure on Training Dataset

Due to the size of the dataset, there are many influential observations. It is worth noting that many of the data points deemed influential by this criteria are for African American male employees. This is partially due to their small proportion in the dataset. This finding shows the cruciality of their presence, as the model would otherwise have difficulty predicting their wages. It is likely that these employees have wages that are either above or below the expected value for African American male employees.

We also assess whether of not multicollinearity is present in our model. The generalized variance-inflation factors (*GVIF*) for our model are shown in the table below:

The variables representing interaction terms have high *GVIF* scores due to them being composed of two separate covariates. Each interaction term is collinear with each of its component main effects, and vice versa. The only covariate that has a low *GVIF* value is *exp*, which is not included in any interaction term.

	GVIF	Df	GVIF^(1/(2*Df))
edu	23.88	1.00	4.89
exp	1.11	1.00	1.05
city	27.13	1.00	5.21
reg	77160.33	3.00	6.52
race	688.88	2.00	5.12
deg	285.11	1.00	16.89
reg:race	29361.39	6.00	2.36
city:reg	143.11	3.00	2.29
edu:race	811.37	2.00	5.34
edu:city	28.56	1.00	5.34
edu:reg	13139.94	3.00	4.86
edu:deg	288.94	1.00	17.00