STAT UN2103 Homework 6

Jackson Chen jc4697

December 22, 2017

1 Introduction

The purpose of this project is to conduct analysis on data containing a variety demographic factors regarding wages. More specifically, the aim is to determine a model with worker demographics as the covariates and the wage as a response variable. The model is supposed to correlate the input data as much as possible, so that the model would be able to be an effective predictive tool for wages. Creating a more realistic model also allows us to examine the larger research question surrounding this project, which is as follows:

> Do African Americans have statistically different wages compared to Caucasian males? How do their wages also statistically compare all other males?

The covariates included in the data consist of years of education (`edu` or x_1), job experience (`exp` or x_2), working in or near a city (`city` or x_3), US region (`reg` or x_4), race (`race` or x_5), college graduate (`deg` or x_6), and commuting distance (`com` or x_7). The data set also includes data for the response variable wages (`wage` or Y).

In order to effectively train a model and test it, the input data was split into a training and a validation set. The input dataset initially had around $25,000$ data points. The validation data set had 4965 entries, or about 20% of the input dataset. The training set consisted of the remaining 80% of the rows. A quality assurance check was then done on the split sets in order to assure relative heterogeneity of the datasets.

Exploratory data analysis was conducted on the training dataset. The process will be described in the following sub-sections.

1.1 Preliminary Model Investigation

2 Statistical Model

The final statistical model that has been chosen for this dataset is as follows:

$$\log Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_3 + \beta_5 x_5 + \beta_6 x_1 x_2 + \beta_7 x_1 x_4 + \beta_8 x_2 x_3$$

Each x_i covariate is described in the Introduction. These are the following adjustments made to the very standard linear model that led to this final model:

- `exp` or the number of years of experience took the functional form of a quadratic
- The response variable `wage` took on a logarithmic transformation
- There is an interaction between the following variables:
  - Education and experience
  - Region and education
  - City and experience

The following is the `R` output for `summary(model)`

```
Call:
lm(formula = log(wage) ~ edu + poly(exp, degree = 2) + city +
    race + edu * exp + reg * edu + city * exp, data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7917 -0.2909  0.0321  0.3299  3.8230

Coefficients: (1 not defined because of singularities)
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              4.794600   0.041925 114.361  < 2e-16 ***
edu                      0.119859   0.003636  32.964  < 2e-16 ***
poly(exp, degree = 2)1  69.285708   2.370536  29.228  < 2e-16 ***
poly(exp, degree = 2)2 -24.657310   0.572074 -43.102  < 2e-16 ***
cityyes                  0.138095   0.015260   9.049  < 2e-16 ***
raceother                0.243396   0.015808  15.397  < 2e-16 ***
racewhite                0.245805   0.013970  17.596  < 2e-16 ***
exp                            NA         NA      NA       NA
regnortheast            -0.085959   0.054043  -1.591 0.111722
regsouth                -0.265407   0.048429  -5.480 4.30e-08 ***
regwest                 -0.180143   0.051450  -3.501 0.000464 ***
edu:exp                 -0.001829   0.000102 -17.935  < 2e-16 ***
edu:regnortheast         0.009521   0.003999   2.381 0.017282 *
edu:regsouth             0.015051   0.003617   4.161 3.19e-05 ***
edu:regwest              0.012855   0.003822   3.363 0.000772 ***
cityyes:exp              0.001357   0.000667   2.035 0.041834 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.5127 on 19843 degrees of freedom
Multiple R-squared:  0.3488,Adjusted R-squared:  0.3483
F-statistic: 759.1 on 14 and 19843 DF,  p-value: < 2.2e-16
```

The following are the values to important model selection criteria:

$$AIC = 29839.28$$
$$R^2 = 0.3488$$
$$R_a^2 = 0.3483$$

3 Research Question

4 Appendix

This section contains the exact process that led to the determination of the final model described in Section 2. This process can be broken into two steps: The selection, transformation, and interaction of the covariates that created the final model, and the validation of the final model against a range of diagnostic tools.

4.1 Model Selection

The first step in the exploratory data analysis consisted of creating new columns in the data frame to map the categorical string variables `deg, city, reg, race` into numeric categorical variables. The purpose of

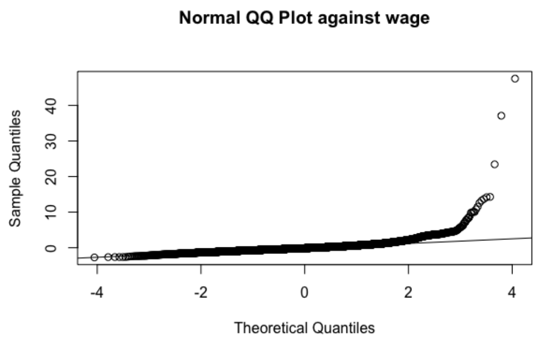

Figure 1: QQ plot on untransformed wage

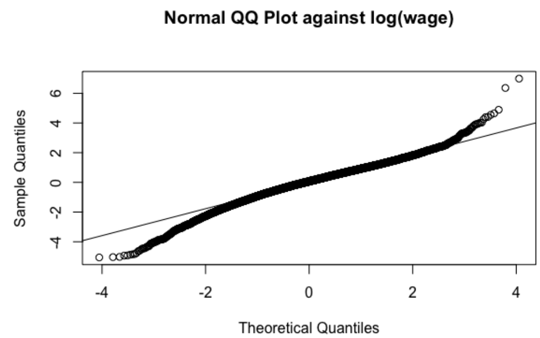

Figure 2: QQ plot on log transformed wage

this step is to satisfy the signature for `R`'s `I()` function, which groups covariates together into an interaction variable.

After the training-validation split was done on the data, a QQ plot was plotted on the untransformed dataset, as shown in Figure 1. However the end of the QQ plot had too extreme of a vertical gradient. Therefore a logarithmic transformation would reduce the extremity at this end. Figure 2 shows the QQ plot on the logarithmically transformed response variable. The overall structure of the points as well as the end behavior are much more normal.

The next step involved verifying whether each of the covariates in the input dataset actually correlated with `wage`. This purpose was this was to determine which variables to investigate further in terms of transformations, functional forms, and interactions. The more formal process for eliminating covariates occurs in the reduction of explanatory variables stage.

From Figure 3, it appears that the commute time has no impact on wages. This is determined visually by examining that the linear smoothing of the graph has no gradient and thus indicates no relationship. We can verify this more numerically by running a marginal t-test on a linear model between the wages and the commute time. The null hypothesis in this test predict that $\beta_1 = 0$, with the commute time as the x_1 term. Running the t-test in `R` calculates the p value to be 0.79, which is greater than the 0.05 significance cutoff, indicating the failure to reject the null hypothesis. The following is the `R` output.

```
Call:
lm(formula = log(data$wage) ~ data$com)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.2634186  0.0079812 784.770   <2e-16 ***
data$com    0.0001497  0.0005631   0.266     0.79
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.6351 on 19856 degrees of freedom
Multiple R-squared:  3.559e-06,Adjusted R-squared:  -4.68e-05
F-statistic: 0.07067 on 1 and 19856 DF,  p-value: 0.7904
```

On the other hand, the other covariates show an existent relationship with wages (or the log of the wage). Figures 4 to 8 provide graphical support behind such claims. Only a select number of covariates were displayed since the graphs behind the other covariates follow a very similar trend. By observing the smoothing functions for all of the aforementioned graphs, one can determine the transformations necessary to construct the most predictive linear model with respect to wages. Most of the smoothing functions indicate a somewhat linear path indicating a strict linear relationship, with one exception.

Figure 4 depicts a downward-facing quadratic relationship between the number of years of experience and the log of the wages. Upon further inspection, Figure 5 represents the relationship more clearly with the inner quartile ranges following the same quadratic trend. Therefore, `exp` requires a quadratic transformation to resolve this violation in linearity.

At this stage of the analysis, we have determined that the model will have a logarithmically transformed response variable `wage`, a quadratically transformed covariate `exp`, and linear relationships for all other covariates. Mathematically speaking, the equation is:

$$\log Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_3 + \beta_5 x_4 + \beta_6 x_5 + \beta_7 x_6 + \beta_8 x_7$$

The last important component of the model now needs to be examined: interactions between covariates. This analysis will consist of three parts:

1. Interactions among categorical variables
2. Interactions between categorical and continuous variables
3. Interactions among continuous variables

Testing for collinearity among categorical variables requires testing every pair of two variables. Since there are 4 categorical variables, this results in 6 tests. Each test consists of generating an interaction graph between the two chosen variables and the response variable log(Wage). These graphs are displayed in Figures 9 to 14. Each test also consists of running a Type III Sum of Squares ANOVA test to numerically check for multilinearity.

Region and city are correlated. Figure 9 show that the lines for each region intersect, indicating that the worker's region influences the rate at which living in a city increases ones wages. Numerically, we can use a marginal t-test to ensure that a correlation between `reg` and `city` is statistically significant.

The null hypothesis in this situation states that there exists no relationship between region and city. The alternate hypothesis states the contrary. Using `R`, we compute the p value to be 0.0021 which is less than the required 0.05 significance. Therefore we reject the null hypothesis and accept H_A, which states that a statistically significant relationship between region and city exists. The following is the sample `R` output. Note how the final variable is combined into an interaction variable, allowing the marginal t-test to be run.

```
Call:
lm(formula = log(data$wage) ~ city + reg + I(city_num * reg_num),
    data = data)

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            6.01866    0.04720 127.515  < 2e-16 ***
cityyes                0.27351    0.02907   9.409  < 2e-16 ***
regnortheast           0.07304    0.01695   4.309 1.65e-05 ***
regsouth              -0.16686    0.01797  -9.284  < 2e-16 ***
regwest               -0.10474    0.02949  -3.552 0.000383 ***
I(city_num * reg_num)  0.03174    0.01030   3.082 0.002057 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.6256 on 19852 degrees of freedom
Multiple R-squared:   0.03,Adjusted R-squared:  0.02976
F-statistic: 122.8 on 5 and 19852 DF,  p-value: < 2.2e-16
```

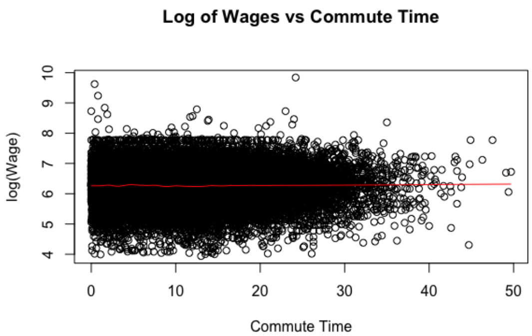

Figure 3: Scatter plot on commute time

Figure 4: Scatter plot on years of experience

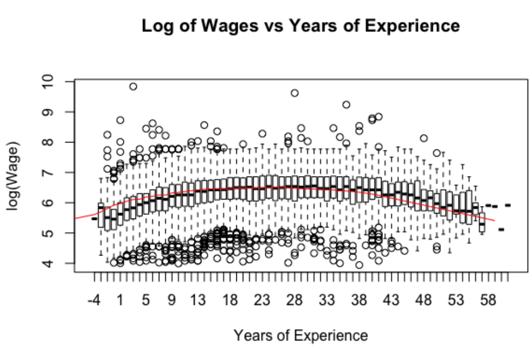

Figure 5: Box plot on years of experience

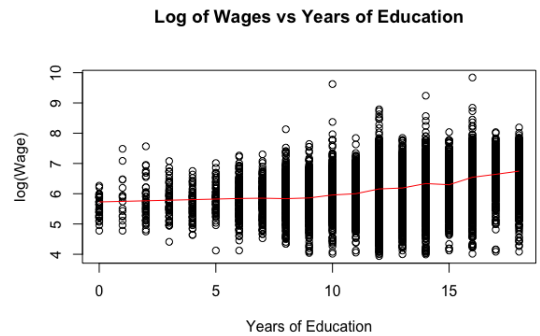

Figure 6: Scatter plot on number of employees

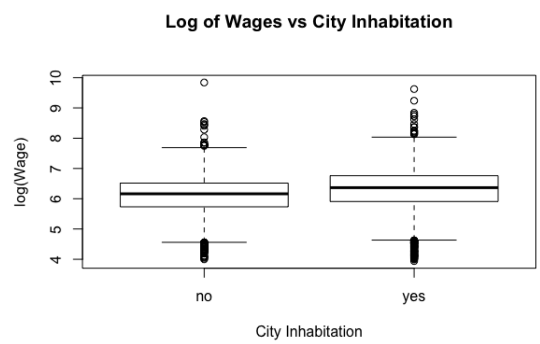

Figure 7: Scatter plot on number of employees

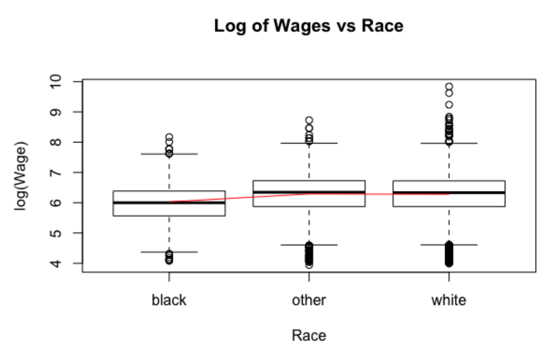

Figure 8: Scatter plot on number of employees

According to Figure 10, city and degree are also correlated. The slopes, although less noticeable compared to the previous case, are different. To verify this numerically, an ANOVA test will be run between city and degree, and wages. In this situation, we will examine the SSR values to determine correlation rather than the marginal t-test.

We will run the ANOVA test with `city` as x_1 and `deg` as x_2, and then another ANOVA test with `city` as x_2 and `deg` as x_1. These are the respective `R` outputs.

```
Analysis of Variance Table

Response: log(data$wage)
             Df Sum Sq Mean Sq F value    Pr(>F)
data$city     1  153.9  153.91  414.42 < 2.2e-16 ***
data$deg      1  481.9  481.89 1297.52 < 2.2e-16 ***
Residuals 19855 7374.0    0.37
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Analysis of Variance Table

Response: log(data$wage)
             Df Sum Sq Mean Sq F value    Pr(>F)
data$deg      1  521.0  520.96 1402.74 < 2.2e-16 ***
data$city     1  114.8  114.84  309.21 < 2.2e-16 ***
Residuals 19855 7374.0    0.37
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

If the two variables are perfectfully uncorrelated, then the following two statements are true:

$$SSR(x_1|x_2) = SSR(x_1)$$
$$SSR(x_2|x_1) = SSR(x_2)$$

However, from the above `R` output, the sum of squares values in the ANOVA table change significantly for both `city` and `deg` between the two commutative models. For example, assuming that `city` is x_1, $SSR(x_1) = 153.9$ while $SSR(x_1|x_2) = 114.8 \implies SSR(x_1) \neq SSR(x_1|x_2)$. Therefore `deg` and `city` are correlated.

Both the SSR method and marginal t-test method are rigorous methods to verify statistical significance between two variables. Future examples will use the marginal t-test for the sake of simplicity. Furthermore, future correlations will not go as in depth as the previous two cases because the process is algorithmically the same.

The final correlated case between two categorical variables is between region and degree, since Figure 12 demonstrates that the degree slopes are different per region. The p value from the t-test is 0.0031, rejecting the null hypothesis and accepting the alternate hypothesis that region and degree are related. On the other hand, the other three pairings: `city` vs `race`, `region` vs `race`, and `degree` vs `race` show no interaction because the results from their marginal t-tests failed to reject the null hypotheses.

The next batch of interactions to check for is between categorical and continuous variables. Only the graphs of correlated variables will be shown (Figures 15 to 20). Uncorrelated variables have graphs with slopes that do not change between different the categories, and have p values greater than 0.05 when running the marginal t-test.

The graphs of the collinear variables all contain regression lines that intersect, indicating that the slopes are influenced by the categorical variable. We will run the marginal t-test on all of these pairs of variables to ensure that their relationship is statistically significant. The null hypotheses for each of these pair of

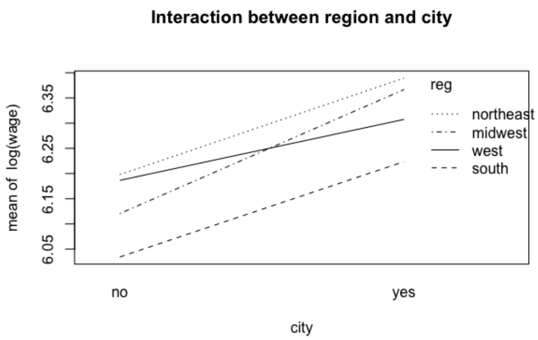

Figure 9: Regional influence on city's relationship with wage

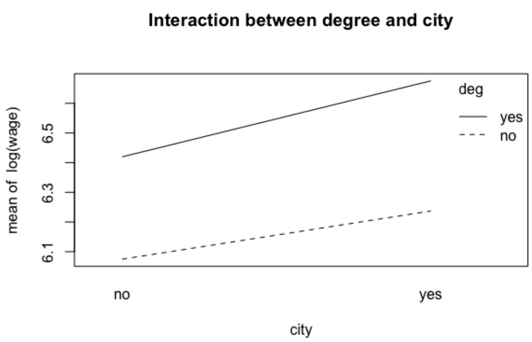

Figure 10: College degree's influence on city's relationship with wage

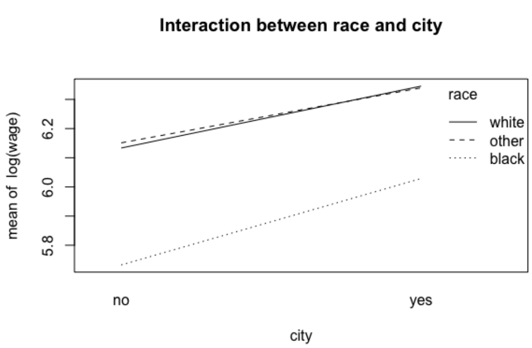

Figure 11: Racial influence on city's relationship with wage

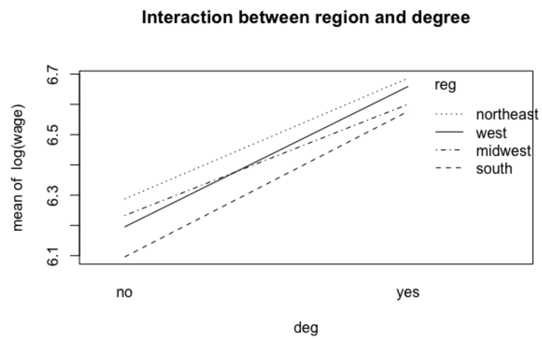

Figure 12: Regional influence on degree's relationship with wage

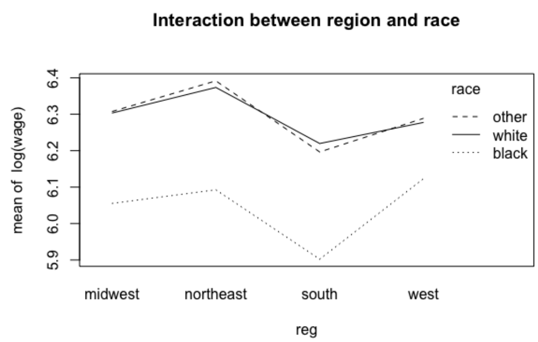

Figure 13: Racial influence on region's relationship with wage

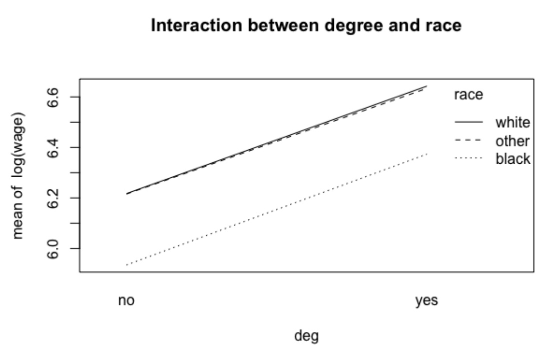

Figure 14: Racial influence on degree's relationship with wage

tests state that the relationship between the two given variables is not statistically significant. All of the respective p-values for these t-tests are shown in Table 1.

Categorical	Continuous	p-value
Race	Education	0.0171
Race	Experience	0.00177
City	Experience	0.0219
Degree	Education	3.35e-05
Degree	Experience	5.14e-09
Region	Education	0.00571

Table 1: p values for correlated categorical and continuous variables.

Since every p value in this table is less than 0.05 significance, the null hypothesis would be rejected in every associated marginal t-test. Therefore the alternate hypothesis is accepted. H_A states that the variables are statistically related. Therefore we must include interactions between these variables in our final model to predict wages.

One can verify that these correlated variables are practical in the context of the real world. The relationships between degree and education and between degree and experience are very strongly correlated due to the very small p value. This makes sense since a certain amount of years of education directly leads to a degree, and having a degree can guarantee job experience.

The last category of correlated variables consist of correlated continuous variables. To determine which continuous variables are correlated, use `R` to populate a correlation table between the covariates in question. Table 2 contains the correlation coefficients between each pair of continuous variable.

	log(wage)	**education**	**experience**	**(experience)^2**	**commute**	**employees**
log(wage)	1.000000000	3.697235e-01	2.327511e-01	-2.920242e-01	1.886612e-03	0.059898053
education	0.369723477	1.000000e+00	-2.793636e-01	-1.530700e-01	-2.564785e-05	0.020344053
experience	0.232751149	-2.793636e-01	1.000000e+00	5.146722e-16	7.603960e-03	-0.003872245
(experience)^2	-0.292024228	-1.530700e-01	5.146722e-16	1.000000e+00	-4.578743e-03	-0.017773812
commute	0.001886612	-2.564785e-05	7.603960e-03	-4.578743e-03	1.000000e+00	-0.003173171
employees	0.059898053	2.034405e-02	-3.872245e-03	-1.777381e-02	-3.173171e-03	1.000000000

Table 2: Correlation coefficients between every pair of continuous variable.

In the correlation table, only correlation values above 0.05 or below -0.05 will be considered correlated. Only one pair of continuous variables satisfies this requirement: education and experience. We can verify that this is the case by running a marginal t-test. The null hypothesis states that education and experience are not statistically correlated. The p value calculated by `R` is $2 * 10^{-6}$, which is less than the 0.05 significance threshold. Therefore we reject the null hypothesis and accept that `edu` and `exp` are statistically related.

At this stage, we have discovered all of the interactions between the covariates as well as any transformations that are needed to satisfy linearity. Even though we found instances of interactions with race, these interactions will not be included in the model to simplify the answer for the research question. Thus, The mathematical model at this state can be summed up with:

$$\begin{aligned}\log Y =& \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_3 \\ &+ \beta_5 x_4 + \beta_6 x_5 + \beta_7 x_6 + \beta_8 x_7 \\ &+ \beta_9 x_1 x_2 \\ &+ \beta_{10} x_1 x_6 + \beta_{11} x_2 x_6 + \beta_{12} x_1 x_4 + \beta_{13} x_2 x_3 \\ &+ \beta_{14} x_3 x_4 + \beta_{15} x_3 x_6 + \beta_{16} x_4 x_6\end{aligned}$$

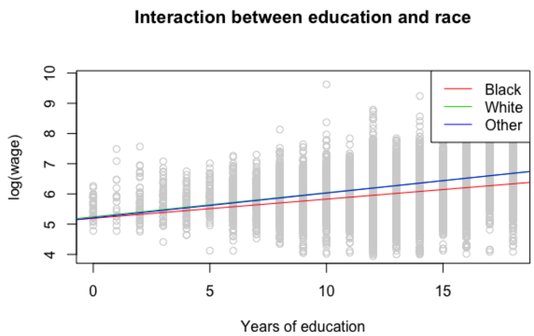

Figure 15: Racial influence on education's relationship with wage

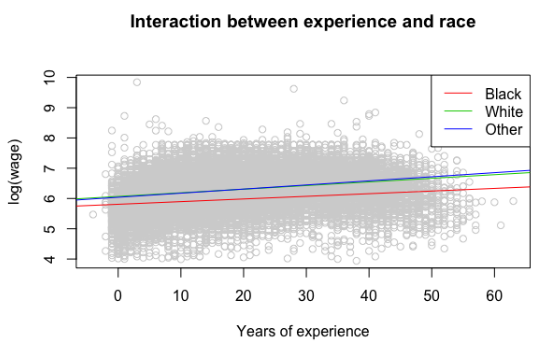

Figure 16: Racial influence on experience's relationship with wage

Figure 17: City influence on experience's relationship with wage

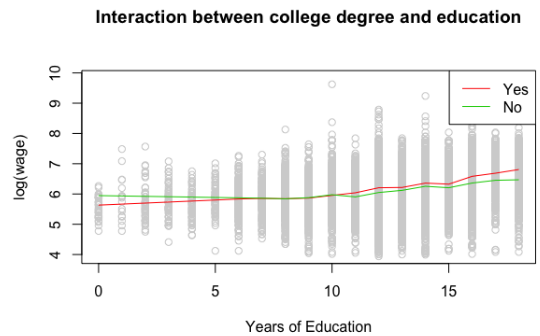

Figure 18: Degree influence on education's relationship with wage

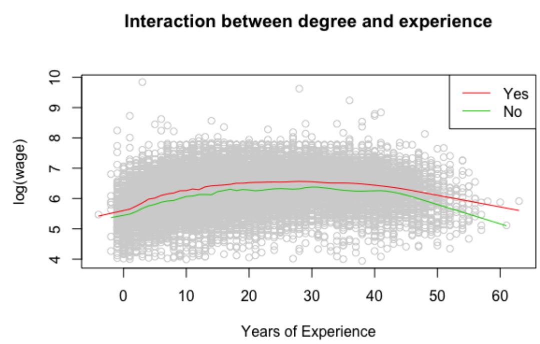

Figure 19: Degree influence on experience's relationship with wage

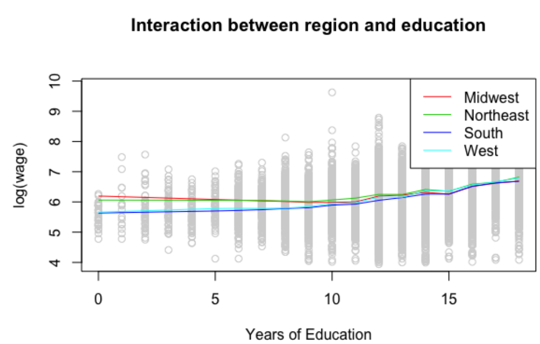

Figure 20: Regional influence on education's relationship with wage

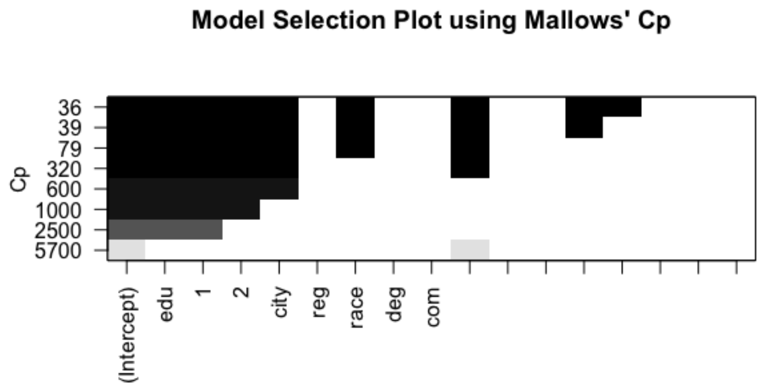

Figure 21: Model Selection Plot using Mallows' C_p

Now that the preliminary model investigation is complete and that all of the functional forms and interactions have been included in the model, we now need to reduce the number of explanatory variables. In this study, we will use the best subsets procedure and optimize for Mallows' C_p.

After using the `leap` library to generate model selection plots (Figures 21 and **??**) the final model is:

$$\log Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_3 + \beta_5 x_5 + \beta_6 x_1 x_2 + \beta_7 x_1 x_4 + \beta_8 x_2 x_3 \quad (1)$$

4.2 Diagnostics and Model Validation

Since the final model has been chosen, the final step is to run diagnostics to validate the model's stability and reasonableness.

We will conduct residual diagnostics to ensure that the linearity of the response function, the normality of the errors, and homoscedasticity are all maintained. The diagnostic plots are in Figures 22 to 25.

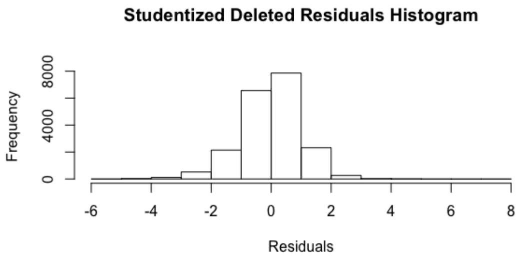

Figure 22: Studentized Deleted Residual Histogram

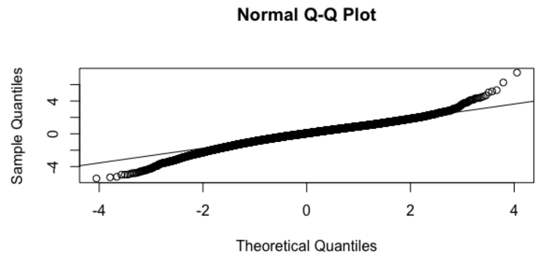

Figure 24: Studentized Deleted Residual QQ Plot

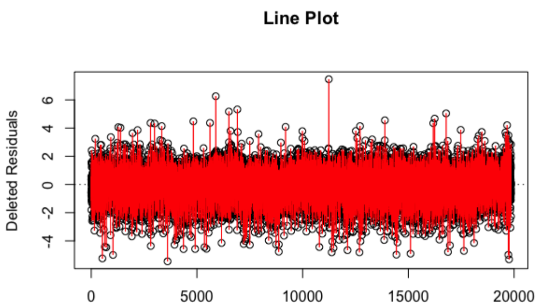

Figure 23: Studentized Deleted Residual Line Plot

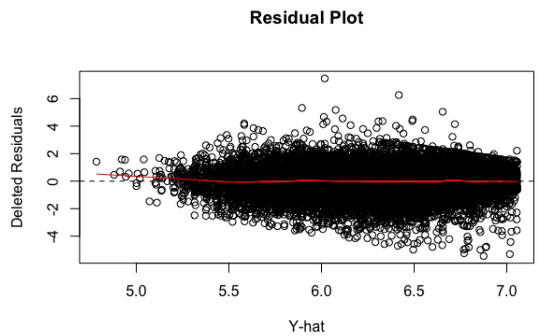

Figure 25: Studentized Deleted Residual Scatter Plot

Figure 22 indicates a histogram with a relatively normal distribution with a very small right skew. The QQ plot in Figure 24 has slightly odd end behavior but is otherwise normal. The Line Plot in Figure 23 is seemingly random without much extreme outliars, as it is supposed to be. And finally, the residual plot in Figure 25 has a linear, zero gradient smoothing function, which is correct. In all, normality may be slightly skewed but the other properties related to linear regression models are satisfied.

For model validation, we need to run the model on the unseen test dataset and measure the mean square prediction error. The relative proximity of MSPE with the mean square error will indicate how well the model responds to new data, and how strong its predictive capabilities are.

According to R, the values are as follows:

$$MSE = 0.2626702$$
$$MSPE = 0.2621056$$

These values are very closely aligned, indicating that our regression function is plausible, and has the ability to generalize inferences well.