GR5205 Project Report

Name: Chenning Liu cl3769
Instructor: Gabriel Young
Date: Dec 2018

Introduction

The aim of this project is to find the relationship between wage levels and workers' races. The project starts with analysis on whole data, and then constructs several regression models to investigate the influence from relevant factors, including years of education, job experience, college graduate, US region, commuting distance and so on. The final model is selected based on a relative good R^2 and AIC value.

Initially, the quality of data and inner relationships are checked. From Figure 1.1 most wage levels are concentrated within around 2500. Thus, outliers exist in the dataset and the model should be able to deal with this situation.

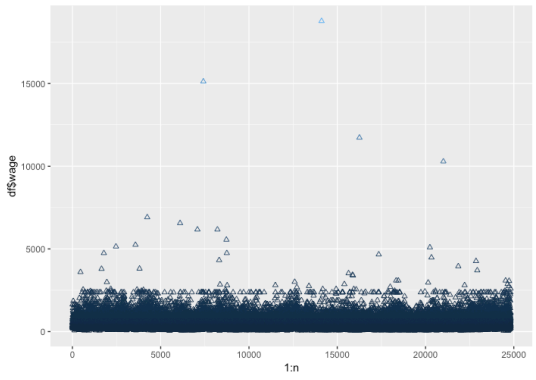


Figure 2.1

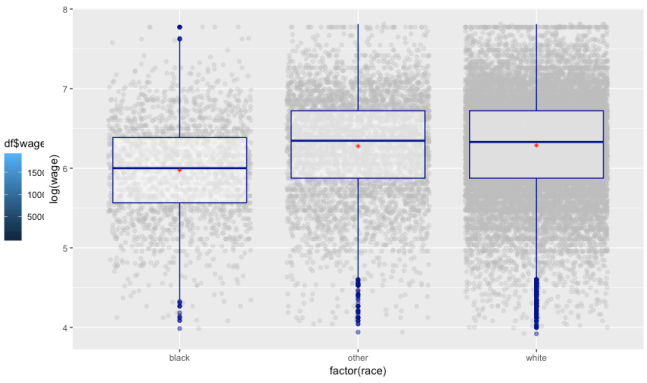


Figure 1.2

From Figure 1.2, the box plot of average wage levels in different races, it indicates that black male workers may have a lower average wage levels. Figure 1.3 reflects that there may exist a quadratic relationship between wage levels and years of experience, while Figure 1.4 shows that there could exist a linear relationship between wage levels and years of education. Figure 1.5 and

1.6 illustrates that workers with a college degree and works in northeast may have a higher wage level. And it also reveals that black male workers have a lower wage level no matter in different regions or with different degrees.

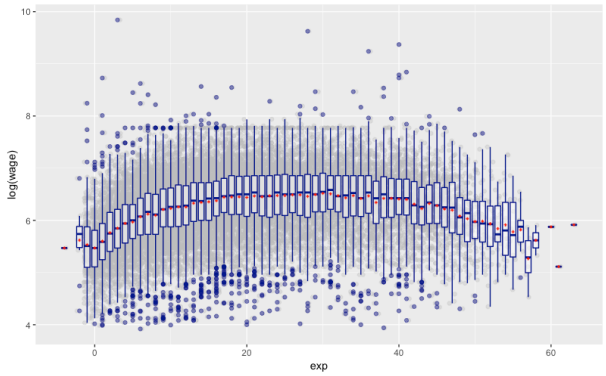


Figure 6.2

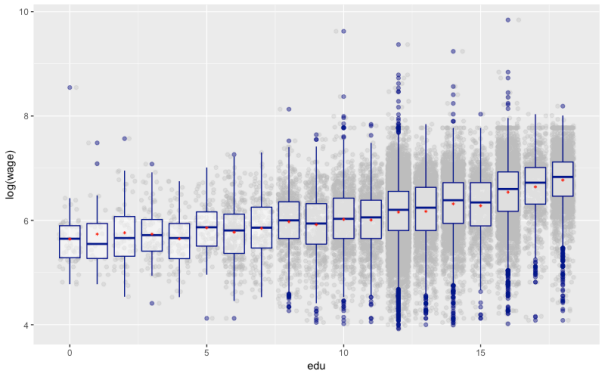


Figure 5.3

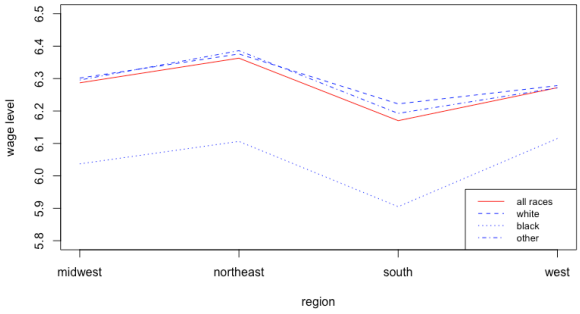


Figure 4.4

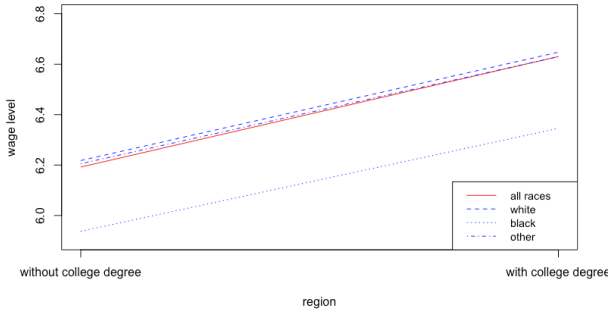


Figure 3.5

From preliminary analysis on the data, African American male workers may have lower wages compared to all other races. The below statistical model tests the statistical relationship and examines the interaction between wage levels and other variables.

Statistical Model

Based on the dataset given, several regression models are tested and checked, and the final selected model is:

$$Y_i = 4.392 + 0.039X_{i,1} + 0.002X_{i,2} + 0.057X_{i,3} - 0.001X_{i,4} + 0.161X_{i,5} + 0.049X_{i,6} - 0.062X_{i,7} + 0.0004X_{i,8} + 0.237X_{i,9} + 0.233X_{i,10}$$

Table 2.1 shows the variables and transformation in the final model:

Y_i	Log of wage (Float)		
X_1	Years of education (Integer)	X_6	Work in northeast region (Boolean)
X_2	Square of years of education(Integer)	X_7	Work in south region (Boolean)
X_3	Job experience (Integer)	X_8	Number of employees in the company (Integer)
X_4	Square of job experience (Integer)	X_9	Caucasian race (Boolean)
X_5	Work in/near a city (Boolean)	X_{10}	Other races (Boolean)

Table 2.1

Figure 2.1 is the R summary output of the final model. Quantities of this model are: AIC $=$ $30026.21, \mathrm{R}^2 = 0.3424, Adjusted\ R^2 = 0.342, MSPR = 0.2651$. All of the variables are significant at $\alpha = 0.05$

```
Call:
lm(formula = log(wage) ~ edu + I(edu^2) + exp + I(exp^2) + city +
    reg.northeast + reg.south + emp + race.white + race.other,
    data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7666 -0.2962  0.0300  0.3334  3.8893

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.392e+00  4.400e-02  99.834  < 2e-16 ***
edu            3.922e-02  6.143e-03   6.385 1.75e-10 ***
I(edu^2)       2.056e-03  2.433e-04   8.452  < 2e-16 ***
exp            5.652e-02  1.015e-03  55.680  < 2e-16 ***
I(exp^2)      -8.701e-04  2.201e-05 -39.534  < 2e-16 ***
city           1.614e-01  8.518e-03  18.950  < 2e-16 ***
reg.northeast  4.853e-02  9.457e-03   5.132 2.90e-07 ***
reg.south     -6.173e-02  8.598e-03  -7.180 7.22e-13 ***
emp            3.959e-04  4.941e-05   8.011 1.20e-15 ***
race.white     2.367e-01  1.401e-02  16.892  < 2e-16 ***
race.other     2.331e-01  1.587e-02  14.692  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5152 on 19847 degrees of freedom
Multiple R-squared:  0.3424,    Adjusted R-squared:  0.342
F-statistic:  1033 on 10 and 19847 DF,  p-value: < 2.2e-16
```

Figure 2.1

Research Questions

The two research questions can be investigated using F test. Thus, to test whether African American males have statistically different wages compared to Caucasian and all other males, here I use 2 hypotheses ($H_0: \beta_9 = 0$ for question 1, $H_0: \beta_9 = \beta_{10} = 0$ for question 2).

Research Question 1:

Assume null hypothesis $H_0: \beta_9 = 0$ and alternative hypothesis $H_A: \beta_9 \neq 0$. Then, the full model is:

$$Y_i = 4.392 + 0.039X_{i,1} + 0.002X_{i,2} + 0.057X_{i,3} - 0.001X_{i,4} + 0.161X_{i,5} + 0.049X_{i,6} - 0.062X_{i,7} + 0.0004X_{i,8} + 0.237X_{i,9} + 0.233X_{i,10}$$

Under null hypothesis, the reduced model is:

$$Y_i = 4.621 + 0.037X_{i,1} + 0.002X_{i,2} + 0.057X_{i,3} - 0.001X_{i,4} + 0.151X_{i,5} + 0.046X_{i,6} - 0.086X_{i,7} + 0.0004X_{i,8} + 0.018X_{i,10}$$

Table 3.1 is the R ANOVA output. The p-value of this test is less than 0.05, thus we reject the null hypothesis $H_0: \beta_9 = 0$. Hence, African American males have statistically different wages compared to Caucasian males.

	Res.Df	Rss	Df	Sum of Sq	F	Pr(>F)
1	19848	5343.3				
2	19847	5267.6	1	75.733	285.35	<2.2e-16

Table 3.1

Research Question 2:

Assume null hypothesis $H_0: \beta_9 = \beta_{10} = 0$ and alternative hypothesis

H_A: at least one of β_9 *and* $\beta_{10} \neq 0$. Then, the full model is same as research question 1:

$$Y_i = 4.392 + 0.039X_{i,1} + 0.002X_{i,2} + 0.057X_{i,3} - 0.001X_{i,4} + 0.161X_{i,5} + 0.049X_{i,6} - 0.062X_{i,7} + 0.0004X_{i,8} + 0.237X_{i,9} + 0.233X_{i,10}$$

Under null hypothesis, the reduced model is:

$$Y_i = 4.624 + 0.037X_{i,1} + 0.002X_{i,2} + 0.057X_{i,3} - 0.001X_{i,4} + 0.151X_{i,5} + 0.046X_{i,6} - 0.086X_{i,7} + 0.0004X_{i,8}$$

Table 3.2 is the R ANOVA output. The p-value of this test is less than 0.05, thus we reject the null hypothesis $H_0: \beta_9 = \beta_{10} = 0$. Hence, African American males have statistically different wages compared to all other males.

	Res.Df	Rss	Df	Sum of Sq	F	Pr(>F)
1	19848	5343.3				
2	19847	5267.6	1	75.733	285.35	<2.2e-16

Table 3.2

Appendix

Model Selection

To find a well-explained model, this project started with the original rough model. Table 4.1 below shows the models in model selection procedure.

Initially (from model #0 to #1), the transformation of response variable is tested. This evidently helps reduce AIC and we see great improvement in R square. Then Boolean values are tested in region and race (model #2 to #3), and it shows that working in northeast and south are significant (hence mid-west is abandoned). From previous data analysis, years of experience, years of education and number of employees may have a non-linear relationship with wage level. Thus, quadratic terms are added (model #6) and tested. The R summary output shows that square number of employees does not statistically influence wage level and hence, the final model only includes the first power of it (model #7).

Model#	Response	Variables	R^2	Adj-R^2	AIC
0	wage	edu, exp, city, reg, race, deg, com, emp	0.2185	0.2181	367789.3
1	Log(wage)	edu, exp, city, reg, race, deg, com, emp	0.2895	0.2892	39454.08
2	Log(wage)	edu, exp, city, region (midwest, northeast, south), com, emp, race (white, other)	0.2891	0.2888	39466.02
3	Log(wage)	edu, exp, city, region (northeast, south), com, emp, race (white, other)	0.2895	0.2892	39454.08
4	Log(wage)	edu, exp, deg, city, region (northeast, south), com, emp, race (white, other)	0.2895	0.2892	39452.1
5	Log(wage)	edu, exp, deg, city, region (northeast, south), emp, race (white, other)	0.2895	0.2892	39450.61
6	Log(wage)	edu, edu^2, exp, exp^2, deg, city, region (northeast, south), emp, emp^2, race (white, other)	0.3427	0.3424	37523.79
7	Log(wage)	edu, edu^2, exp, exp^2, deg, city, region	0.3426	0.3424	37523.15

Table 4.1

		(northeast, south), emp, race (white, other)			

Best subset procedure is also applied to help find significant variables. The procedure is based on Mallow's Cp criterion and Figure 4.1 shows that college degree, working in mid-west and commuting are not significant and hence should be removed. However, with quadratic terms included, the college degree becomes significant and it is kept in the final model.

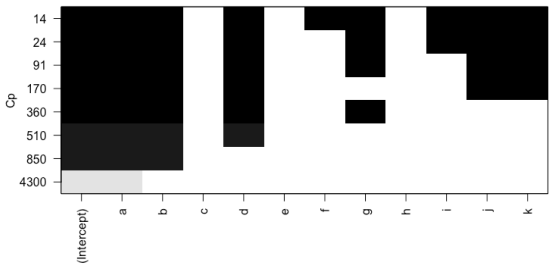


Figure 4.1

Diagnostics and Model Validation:

Model #7 is checked using diagnostic plots and validation procedure before arriving to the final model. The model building data set is randomly selected from 80% of the whole data set. And the rest 20% is model validation data set. Table 4.2 shows the MSE and MSPR of model #7. This illustrates that the MSE drops as a more explanative model is used. Also, the value of MSE and MSPR are very close to each other, and hence MSE from the training dataset is a valid indicator of the predictive ability of the final fitted model.

MSE earlier (model #0)	MSE (model #7)	MSPR (model #7)
0.2868	0.2654	0.2651

Table 4.2

Figure 4.2 shows the residual plot and figure 4.3 shows the normal Q-Q plot of model #7. The residual plot shows that residuals are normally distributed, and the normal Q-Q plot shows that the model is valid.

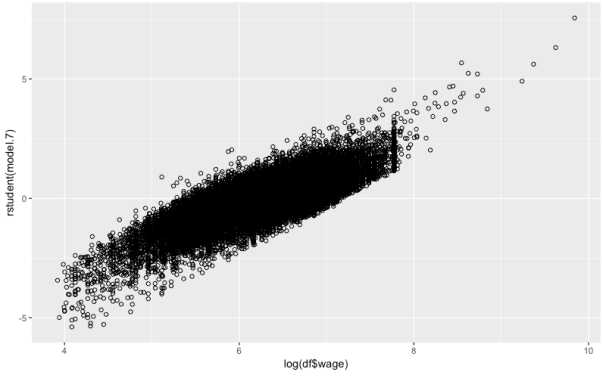


Figure 4.2

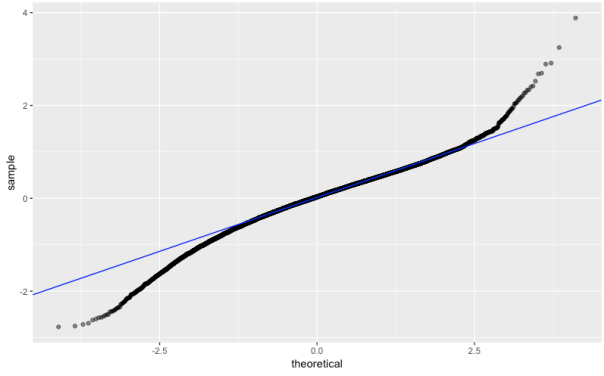


Figure 4.3

Figure 4.4 is the plot of y hat values and deleted residuals. The red line is centered around 0. Again, the boxplot (Figure 4.5) of shows that residuals are appropriately distributed with median around 0. These two figures shows that the model is valid.

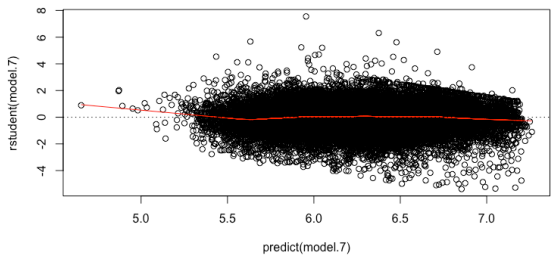


Figure 4.4

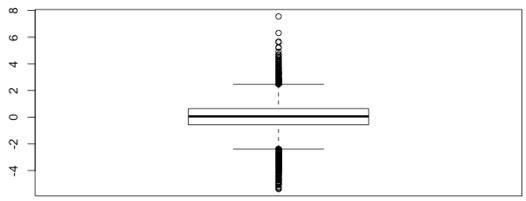
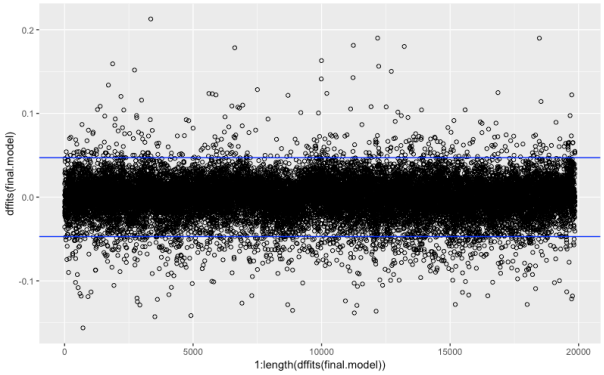


Figure 4.5

Influential Observations

The influential observations for the statistical calculation is checked by DFFITS and Cook's distance. DFFITS reflects the how much does the influential observations may affect predicted y value. The deletion of influential observation's deletion from the dataset would noticeably change the result of calculation. Cook's Distance is commonly used to estimate of the influence of a data point when performing a least-squares regression analysis.

Figure 4.6 and figure 4.7 are DFFITS and Cook's Distance. DFFITS figure shows that most of the data points are within the blue line $\pm 2\sqrt{p/n}$. A simple operational guideline for Cook's Distance of $D_i > 1$ is suggested, and from the figure, most of the data points are close to 0, and there are very few high influential points using Cook's Distance. This again illustrates that the model is a good fit for the data.



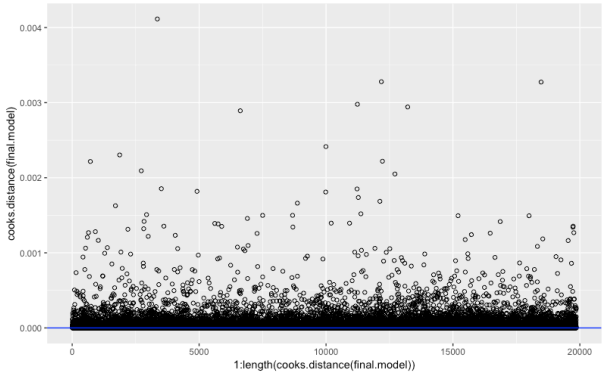
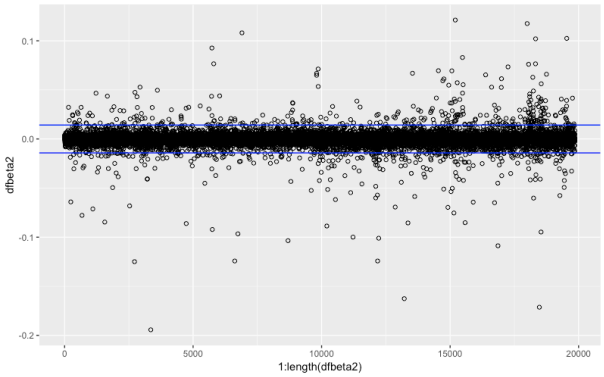


Figure 4.6

Figure 4.7

DFBETAS are also used test how much does the influential observations may affect our Betas. Figure 4.8 (race.other) and figure 4.9 (race.white) are plots for DFBETAS. From the two graphs, most of the data points are included by blue lines and hence it can be concluded that the dataset is valid to construct a regression model, and the final model is a good fit for the data.



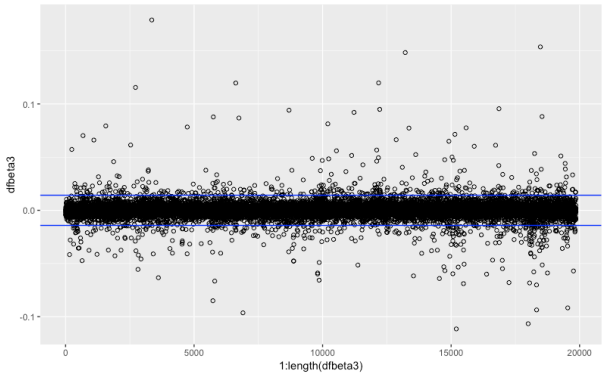


Figure 4.8

Figure 4.9

Conclusion

After performing model construction, selection, diagnostics and validation, Model #7 is selected as the final model. And it is used to test on the two research questions. From diagnostics and validation, this model is valid and convincing. However, interactions between variables are not checked because this might lead to more complicated selection, and usually it is hard to explain their relationship both in the model and in reality. Interactions, including city and degree, race and education, could be further investigated and contribute to a better model.