Linear Regression Project

STAT GR 5205
Author: Caspar Chen, jc5067

Table of Contents

Ⅰ. Introduction

This research aims at studying the the most indicative determinants, interaction terms, and functional expressions of males' wages between the age of 18 and 70 who are full time workers. Moreover, the research answers whether African American males have statistically different wages compared to Caucasian males or all other males.

The original dataset contains 9 variables, and roughly 25000 observations for males between the age of 18 and 70 who are full time workers in total. The response variable (Y) is weekly wages (in dollars). The 8 explanatory variables are years of education (X1) and job experience (X2), working in or near a city (yes, no) (X3), US region (midwest, northeast, south, west) (X4), race (African American, Caucasian, Other) (X5) which is our most interested variable, college graduate degree (yes, no) (X6), commuting distance (X7), and number of employees in a company (X8). So there are four category variables and four continuous variables.

In order to effectively train our model and generalize to other data sets gathered from the same population, we initially split the data set up into a model building data set and a model validation data set so that exploratory data analysis was conducted on the training dataset. The validation data set had 4965 entries, or about 20% of the input dataset. The training data consisted of the remaining 80% of the rows. Since we need to make sure the proportion of the race levels to be the similar for the full data, training data and validation data, a quality control check is done.

Exploratory Data Analysis

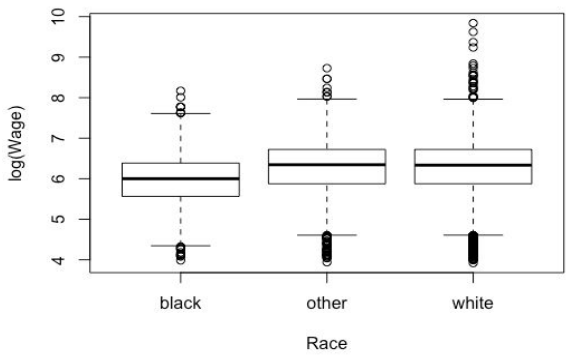


Figure 1.1 boxplot of log(wage) vs Race

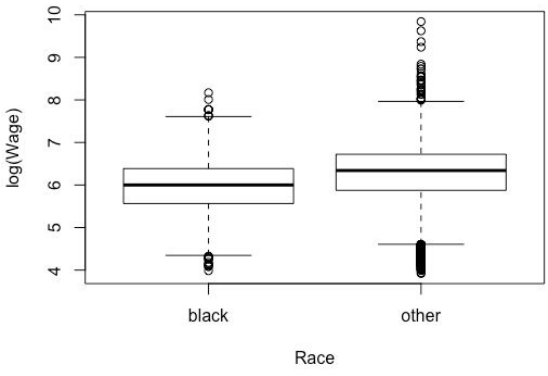


Figure 1.2 boxplot of log(wage) vs Raceblack

From the figure 1.1 and 1.2, a preliminary answer to the research question is that the value of log (Wage) is different between black people and white people, black and all other males.

Summary Statistics

Summary Statistics[1]	Sample size	Mean	Median	Standard Deviation	Min	Max
Wages (Y)	24823	637.82	546.06	451.263	50.39	18777.20

On average, full-time US employees in this dataset earn $637.82 per week, and standard deviation is $451 relative large compared to the mean.

Summary Statistics	0 black/yes	1 other/no	2 white
Race (X5)	1934	4584	18305
City (X3)	18411	6412	
Degree (X6)	20699	4124	

Note that, Race is our most interested variable we need to be careful about. The African American males compose 7.79% males in this dataset. Moreover, people live in city and non-city are quite biased, so as the males who obtained the degree in this dataset.

Ⅱ. Statistical Model

The final regression model for this dataset is chosen as follows:

$$\log(Y) = \beta0 + \beta1*X1 + \beta2*X2 + \beta3*X2^2 + \beta4*X3 + \beta5*X4 + \beta6*X5 + \beta7*X6 + \beta8*X7 + \beta9*X1*X2 + \beta10*X1*X4 + \beta11*X1*X6 + \beta12*X2*X3 + \beta13*X2*X6 + \beta14*X3*X4 + \beta15*X3*X6 + \beta16*X4*X6 + Ɛ$$

In words, the model is expressed as:

$$\log(wages) = \beta o + \beta1*edu + \beta2*exp + \beta3*exp^2 + \beta4*city + \beta5*reg + \beta6*race + \beta7*deg + \beta8*emp + \beta9*edu*exp + \beta10*edu*city + \beta11*edu*deg + \beta12*exp*city + \beta13*exp*deg + \beta14*city*reg + \beta15*city*deg + \beta16*reg*deg + Ɛ,$$

where beta values are described in the following R output.

• The response variable Y wage took on a logarithmic transformation.

• There are 8 interactions between the following variables: Education and experience - Education and city -Education and degree - Experience and city - Experience and degree - City and region - City and degree - Region and degree

[1] Figure 1.3 Summary Statistics

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

Residuals:
    Min      1Q  Median      3Q     Max
-2.7911 -0.2891  0.0307  0.3290  3.8347

Coefficients: (1 not defined because of singularities)
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          4.702e+00  4.676e-02 100.541  < 2e-16 ***
edu                  1.206e-01  4.471e-03  26.983  < 2e-16 ***
poly(exp, 2)1        7.109e+01  2.552e+00  27.858  < 2e-16 ***
poly(exp, 2)2       -2.471e+01  5.733e-01 -43.099  < 2e-16 ***
cityyes              5.673e-02  4.873e-02   1.164 0.244300
regnortheast         6.241e-02  2.401e-02   2.599 0.009357 **
regsouth            -7.898e-03  1.847e-02  -0.428 0.668871
regwest              4.991e-02  2.055e-02   2.429 0.015141 *
raceother            2.430e-01  1.577e-02  15.413  < 2e-16 ***
racewhite            2.453e-01  1.393e-02  17.608  < 2e-16 ***
degyes               2.487e-01  1.677e-01   1.483 0.138115
emp                  4.074e-04  4.904e-05   8.308  < 2e-16 ***
exp                         NA         NA      NA       NA
edu:exp             -1.848e-03  1.198e-04 -15.425  < 2e-16 ***
edu:cityyes          1.256e-02  3.702e-03   3.393 0.000693 ***
edu:degyes          -2.184e-02  9.940e-03  -2.197 0.028025 *
degyes:exp           1.672e-03  1.068e-03   1.566 0.117440
cityyes:regnortheast -4.182e-02  2.668e-02  -1.568 0.117010
cityyes:regsouth     -1.059e-01  2.170e-02  -4.879 1.08e-06 ***
cityyes:regwest      -9.435e-02  2.390e-02  -3.948 7.90e-05 ***
cityyes:degyes        7.256e-02  3.028e-02   2.396 0.016584 *
regnortheast:degyes   2.418e-02  2.833e-02   0.853 0.393450
regsouth:degyes       7.293e-02  2.719e-02   2.682 0.007314 **
regwest:degyes        4.619e-02  2.897e-02   1.594 0.110940
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5111 on 19835 degrees of freedom
Multiple R-squared:  0.353,     Adjusted R-squared:  0.3523
F-statistic: 491.9 on 22 and 19835 DF,  p-value: < 2.2e-16
```

Figure 2.1 final model R output

The following are the model selection criteria:

AIC: 29725.7

R^2: 0.353

R_a^2: 0.3523

MSPR :0.2623

III. Research Question

1. Do African American males have statistically different wages compared to Caucasian males?
Null hypothesis(H_0): African American males do not have statistically different wages compared to Caucasian males (i.e $\beta 6 = 0$).
Alternative hypothesis(H_A): African American males have statistically different wages compared to Caucasian males ($\beta 6 \neq 0$).
With our final model r output in Figure 2.1, we have *t* value of race white is 17.608 and the *p* value is less than 2×10^{-16}. Thus at any significance level, we reject the null hypothesis, and in favor of the alternative. This shows that the 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?
Null hypothesis(H_0): African American males do not have statistically different wages compared to all other males (i.e $\beta 6 = 0$.)
Alternative hypothesis(H_A): African American males have statistically different wages compared to all other males ($\beta 6 \neq 0$).
In this case, we combine the two categories of whites and other males into one category: raceother.

```
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              4.702e+00  4.676e-02 100.543  < 2e-16 ***
edu                      1.206e-01  4.470e-03  26.983  < 2e-16 ***
poly(exp, 2)1            7.110e+01  2.552e+00  27.860  < 2e-16 ***
poly(exp, 2)2           -2.471e+01  5.733e-01 -43.103  < 2e-16 ***
cityyes                  5.675e-02  4.872e-02   1.165 0.244185
regnortheast             6.246e-02  2.401e-02   2.601 0.009293 **
regsouth                -7.871e-03  1.846e-02  -0.426 0.669919
regwest                  4.998e-02  2.054e-02   2.433 0.014993 *
raceother                2.449e-01  1.381e-02  17.737  < 2e-16 ***
degyes                   2.485e-01  1.677e-01   1.482 0.138291
emp                      4.074e-04  4.904e-05   8.307  < 2e-16 ***
exp                             NA         NA      NA       NA
edu:exp                 -1.848e-03  1.198e-04 -15.427  < 2e-16 ***
edu:cityyes              1.256e-02  3.701e-03   3.393 0.000692 ***
edu:degyes              -2.183e-02  9.939e-03  -2.197 0.028058 *
degyes:exp               1.672e-03  1.068e-03   1.565 0.117534
cityyes:regnortheast    -4.186e-02  2.668e-02  -1.569 0.116662
cityyes:regsouth        -1.059e-01  2.170e-02  -4.880 1.07e-06 ***
cityyes:regwest         -9.441e-02  2.389e-02  -3.951 7.81e-05 ***
cityyes:degyes           7.259e-02  3.028e-02   2.397 0.016531 *
regnortheast:degyes      2.417e-02  2.833e-02   0.853 0.393644
regsouth:degyes          7.294e-02  2.719e-02   2.683 0.007305 **
regwest:degyes           4.617e-02  2.897e-02   1.594 0.111043
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5111 on 19836 degrees of freedom
Multiple R-squared:  0.353,     Adjusted R-squared:  0.3523
F-statistic: 515.4 on 21 and 19836 DF,  p-value: < 2.2e-16

> AIC(final.model2) < AIC(final.model)
[1] TRUE
```

Figure 2.2 final model2 with binary race category R output

With our final model R output in Figure 2.2, we have even bigger *t* value of raceother: 17.737, and the *p* value is less than 2×10^{-16}. Thus at any significance level, we reject the null hypothesis, and in favor of the alternative. This shows that the African American males have statistically different wages compared to all other males.
In summary, both of these findings indicate that, as a whole, African American males earn less money per week than their Caucasians did, as well as males from all other races for this dataset.

IV. Appendix

a. Model Selection

First, we can include every variable without transformations for a try, and it isn't fitting the data very well and the qq-plot shows a large deviation from normality. The coefficient of determination is only 0.2125 and the AIC is 294768.2, extremely large.
While after the log-transformation, the graph looks pretty normal, the coefficient of determination is 0.2899 and AIC is 31552.36, which is better.

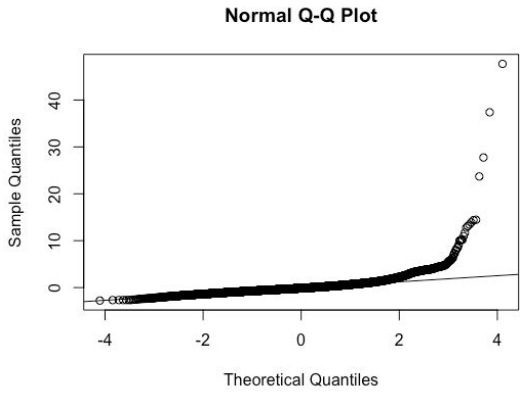


Figure 4.1 boxplot of wage (Y)

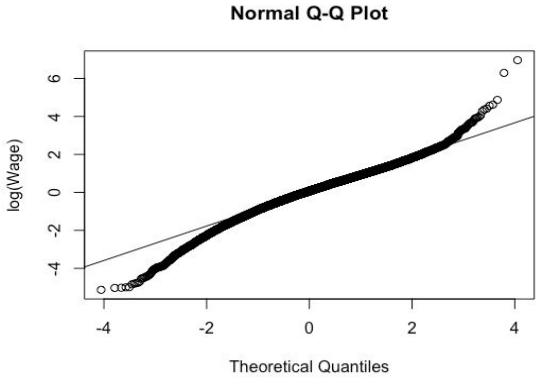


Figure 4.2 boxplot of log(wage) logY

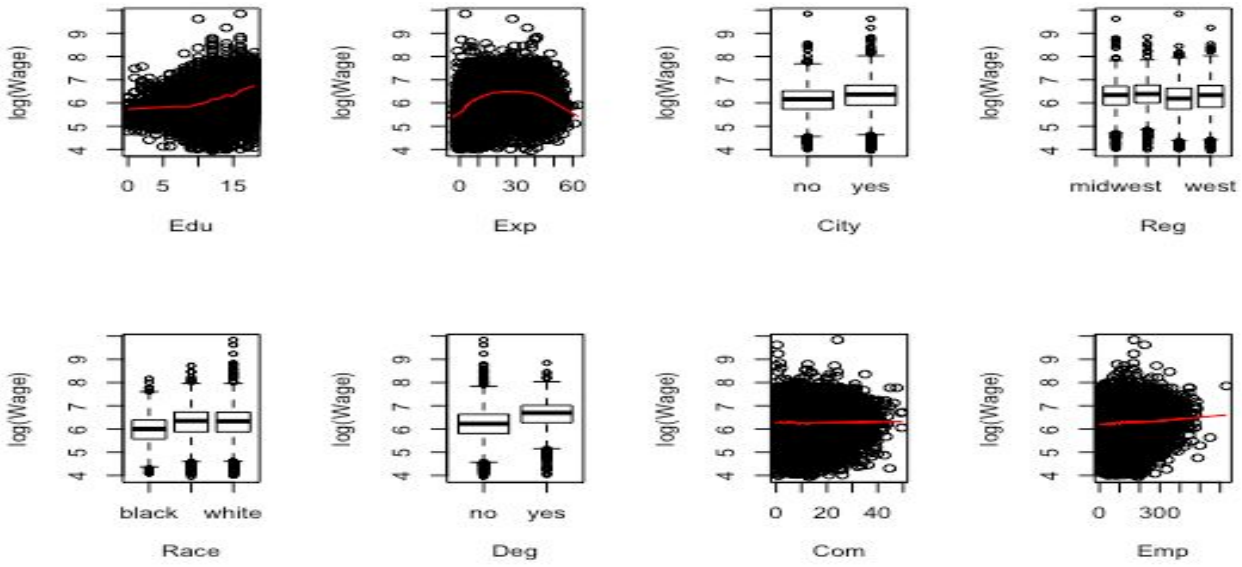


Figure 4.3 boxplot & scatterplot of log(wage) vs all 8 x variables

Second, using Exploratory Data Analysis, scatterplot with smoothers, between response variable and every single variable to decide whether there is a linear relationship to include the x variables.

From Figure 4.3,

- it appears that the x variable com do not have an effect on wages. This is determined by examining the linear smoothing of the graph. If the lines are parallel-ish, then log(wage) does not depend on the x variable.
- From the log(wage) vs covariate exp plot, there seems a high order relationship. So we apply a polynomial function to it. This looks quadratic to me, so I set the degree to 2.

Now, the rough model 3 becomes: $\log(wages) = \beta o + \beta 1*edu + \beta 2*exp + \beta 3*exp^2 + \beta 4*city + \beta 5*reg + \beta 6*race + \beta 7*deg + \beta 8*emp + \mathcal{E}$

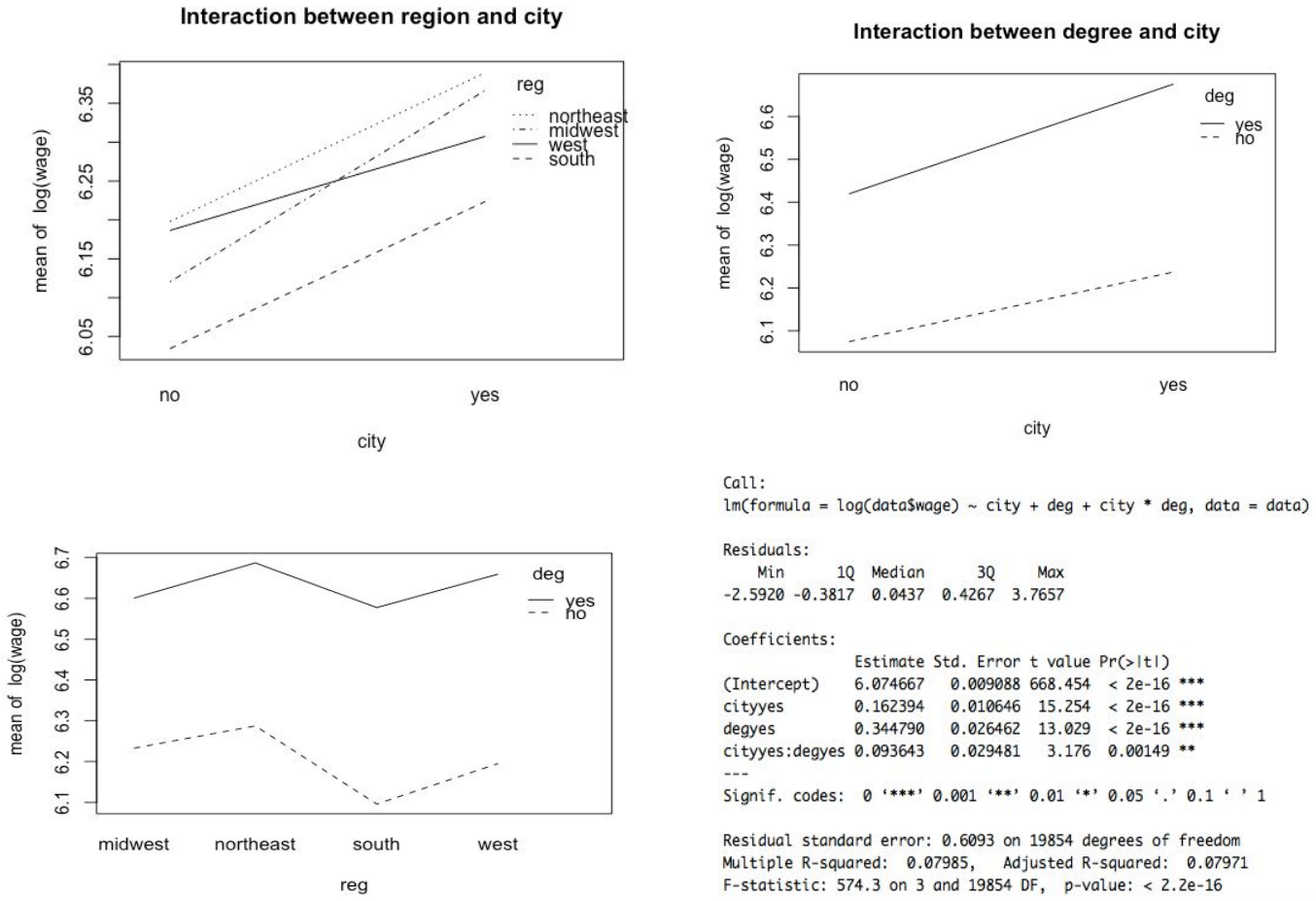
```
## Call:
## lm(formula = log(wage) ~ edu + poly(exp, degree = 2) + city +
##         reg + race + deg + emp, data = train.data)
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    4.761e+00  2.620e-02 181.745  < 2e-16 ***
edu                            8.548e-02  1.626e-03  52.578  < 2e-16 ***
poly(exp, degree = 2)1         3.133e+01  5.405e-01  57.957  < 2e-16 ***
poly(exp, degree = 2)2        -2.049e+01  5.237e-01 -39.133  < 2e-16 ***
cityyes                        1.656e-01  8.514e-03  19.455  < 2e-16 ***
regnortheast                   4.073e-02  1.077e-02   3.781 0.000157 ***
regsouth                      -6.669e-02  1.000e-02  -6.667 2.68e-11 ***
regwest                       -1.192e-02  1.088e-02  -1.095 0.273334
raceother                      2.356e-01  1.589e-02  14.833  < 2e-16 ***
racewhite                      2.398e-01  1.403e-02  17.087  < 2e-16 ***
degyes                         5.713e-02  1.209e-02   4.726 2.31e-06 ***
emp                            3.991e-04  4.947e-05   8.066 7.65e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.5158 on 19846 degrees of freedom
Multiple R-squared:  0.3408,      Adjusted R-squared:  0.3404
F-statistic: 932.6 on 11 and 19846 DF,  p-value: < 2.2e-16
> AIC(r.model.3)
[1] 30076.33
```

Figure 4.4 rough model 3 R output

This turns out that for our rough model 3, $R^2$to 34.08%, R_a^2 34.04%, and AIC improves to 30076.33.

Third, we need to conduct research on Interaction between covariates.

- Interactions among categorical variables




```
Call:
lm(formula = log(data$wage) ~ city + deg + city * deg, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5920 -0.3817  0.0437  0.4267  3.7657

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.074667   0.009088 668.454  < 2e-16 ***
cityyes      0.162394   0.010646  15.254  < 2e-16 ***
degyes       0.344790   0.026462  13.029  < 2e-16 ***
cityyes:degyes 0.093643   0.029481   3.176  0.00149 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6093 on 19854 degrees of freedom
Multiple R-squared:  0.07985,   Adjusted R-squared:  0.07971
F-statistic: 574.3 on 3 and 19854 DF,  p-value: < 2.2e-16
```

Figure 4.4 Interaction between Categorical variables and r output

There exists interaction between city and reg[west], deg and city, reg and deg. Since both the slope of city and region, city and degree are different, thus they are also correlated. From the interaction plot between reg, deg, and log(wage) in Figure 4.4, it shows that the degree slopes are different per region change.

- Interactions between categorical and continuous variables
  We exclude the interaction between covariates and race, since it will simplify the answer for the research question and we can focus mainly on linear effect of race. Also, we only include the statistically significant figure as shown below.

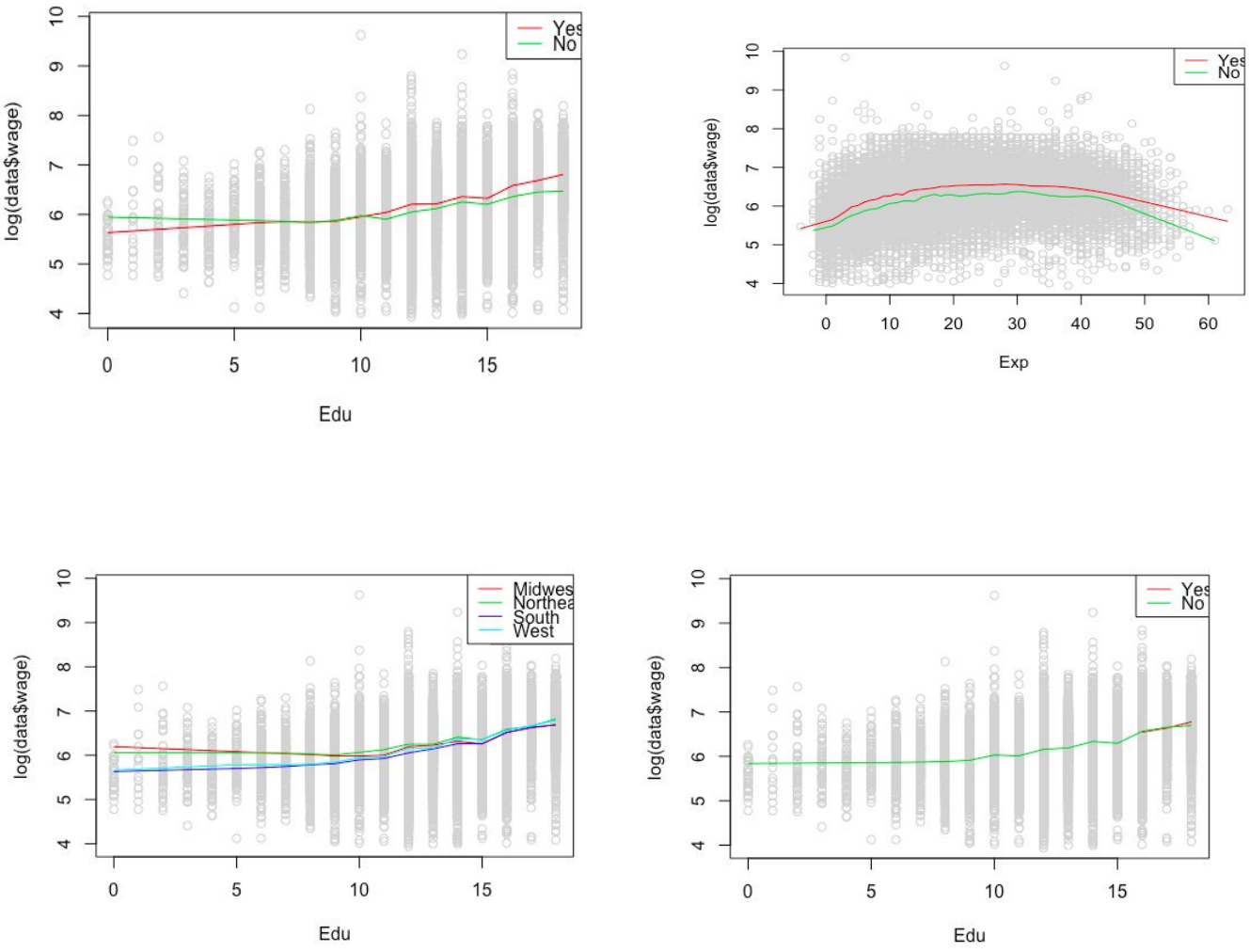


Figure 4.5 Interaction between Categorical variable and Continuous variables

From the Figure 4.5, there are interactions between city and edu, city and exp, reg and exp, reg and edu, reg and exp, edu and deg.

- Interactions among continuous variables

To determine which continuous variables are correlated, use Pearson correlation as below.

```
> cor(data.frame(log(data$wage), data$edu, data$exp, data$emp))
```

	log.data.wage.	data.edu	data.exp	data.emp
log.data.wage.	1.00000000	0.36972348	0.232751149	0.059898053
data.edu	0.36972348	1.00000000	-0.279363622	0.020344053
data.exp	0.23275115	-0.27936362	1.000000000	-0.003872245
data.emp	0.05989805	0.02034405	-0.003872245	1.000000000

Figure 4.6 Pearson correlation table between continuous variables

Since there is a very weak negative correlation between exp and emp and very weak correlation between edu and emp in the training dataset, so we include one more interaction in this step, which are edu and exp.

Now, the final model becomes: $\log(wages) = \beta o + \beta 1*edu + \beta 2*exp + \beta 3*exp^2 + \beta 4*city + \beta 5*reg + \beta 6*race + \beta 7*deg + \beta 8*emp + \beta 9*edu*exp + \beta 10*edu*city + \beta 11*edu*deg + \beta 12*exp*city + \beta 13*exp*deg + \beta 14*city*reg + \beta 15*city* deg + \beta 16*reg*deg + \varepsilon$

i.e.,

$\log Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X2^2 + \beta 4X3 + \beta 5X4 + \beta 6X5 + \beta 7X6 + \beta 8X7 + \beta 9X1X2 + \beta 10X1X4 + \beta 11X1X6 + \beta 12X2X3 + \beta 13X2X6 + \beta 14X3X4 + \beta 15X3X6 + \beta 16X4X6 + \varepsilon$

b. Diagnostics and Model Validation

.i Diagnostic plots on the final model for the train data

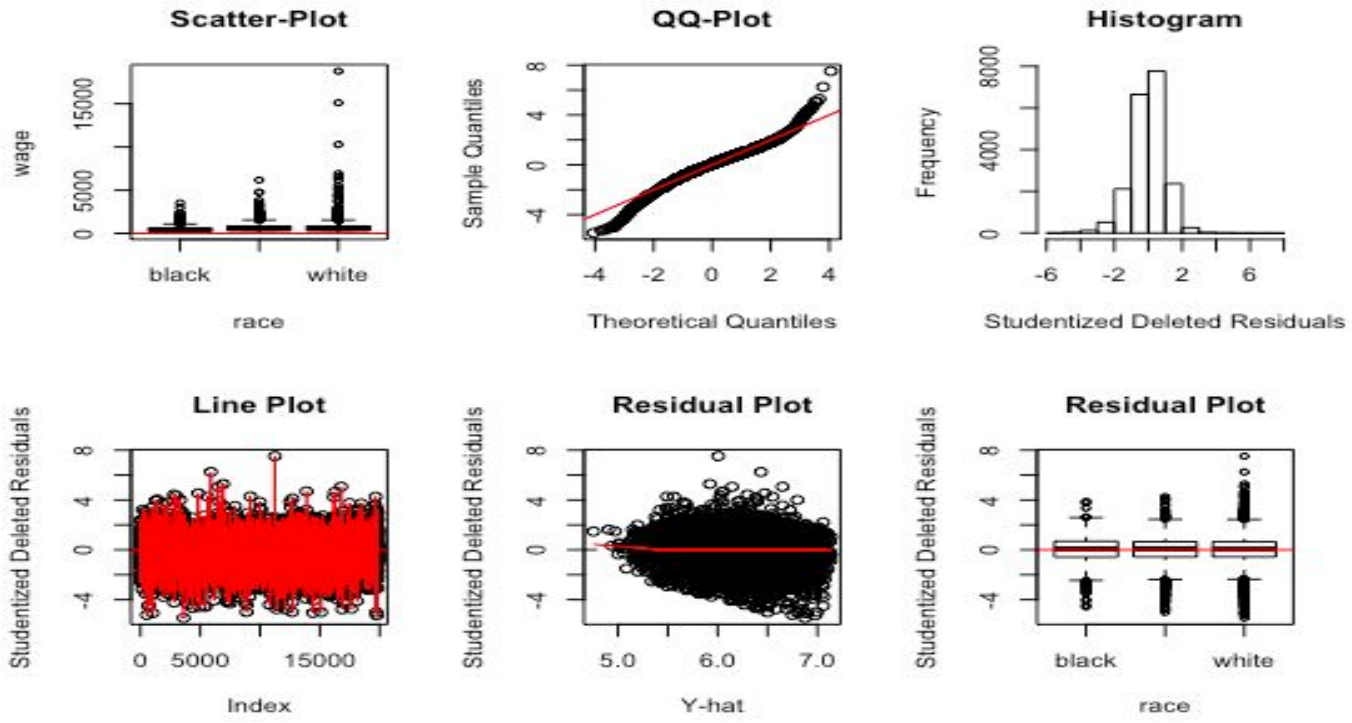


Figure 4.6 Diagnostic plots

According to the plots above, from both the histogram and qq-plot, we could find that our model has normality; from line plot and residual plot, we notice homoscedasticity (constant variance), and independence of the errors.

ii. MSPR and compare it to the computed MSE of the final model

```
> round(c(MSPR=MSPR,MSE=MSE,MSEearler=MSE.earler),4)
   MSPR     MSE MSEearler
 0.2623  0.2612  0.2613
```

Figure 4.7 Mean Square Prediction Error with train and test data

It looks like our final model fits out-of-sample similar to in-sample.

iii. Influential Observations

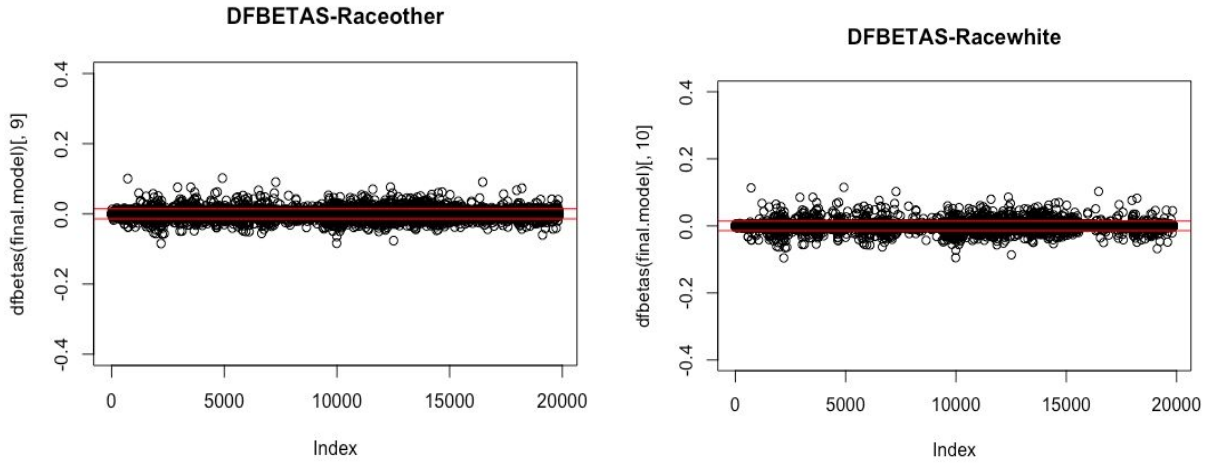


Figure 4.7 $(DFBETAS)_{race}$ plots

DFBETAS – Measure of how much an observation has affected the estimate of a regression coefficient (there is one DFBETA for each regression coefficient, including the intercept). Values larger than 2/sqrt(n) in absolute value are considered highly influential. Thus, there are a large number of influential observations for the two race related beta coefficients given this data set, and they are staying outside of the two red horizontal lines.

Iv. Multicollinearity

Variance Inflation Factor (VIF) – Analyzing the magnitude of multicollinearity by considering the size of the VIF. A rule of thumb is that if VIF >10, then multicollinearity is high and severe. From the table, we can tell all interaction terms have severe multicollinearity, which make sense since they are highly correlated.

```
> vif(df.vif)
                     edu                       X1                       X2                      emp
              258.478780               108.535694                 1.263270                 1.001035
                df.reg_num             df.city_num              df.deg_num             df.race_num
                31.534903               131.761255               300.915777                 1.005594
                 edu...exp        edu...df.city_num        edu...df.deg_num        exp...df.deg_num
                25.829409                42.968619               231.694979                54.833465
df.city_num...df.reg_num df.city_num...df.deg_num df.reg_num...df.deg_num
                18.105519                64.864557                29.733088
```

Figure 4.8 Variance Inflation Factor Table