Homework2

Echo Liu

September 16, 2018

Question 1 and 2:Mammal Brain Weight

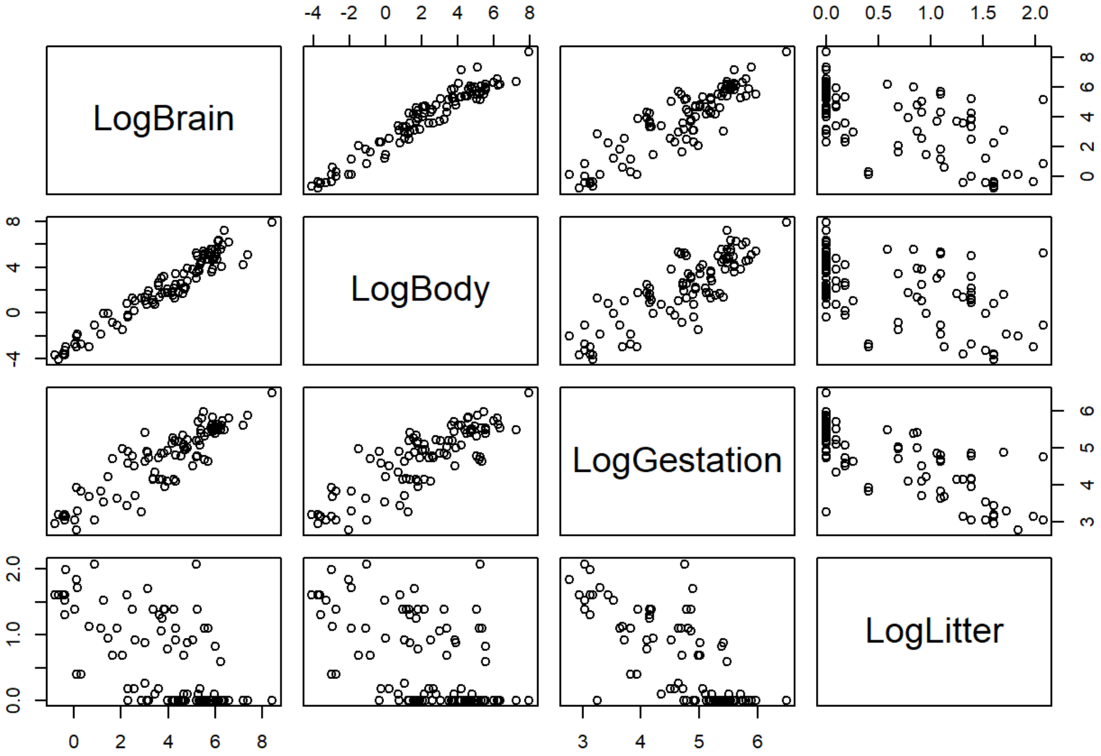
We aim to predict mammal brain weight from predictors such as body weight, gestation and litter size. Let’s first look at the summary of this data set.

```
#read in the mammal brain weight data
Mammal <- read.csv("C:/Users/Echo Liu/Downloads/Duke University/1st semester/702_Modeling_and_Representatio
n/HW2/Ex0912.csv", stringsAsFactors=FALSE)

summary(Mammal)
```

```
##        X            Species              Brain             Body        
##  Min.   : 1.00   Length:96          Min.   :   0.45   Min.   :   0.017  
##  1st Qu.:24.75   Class :character   1st Qu.:  12.60   1st Qu.:   2.075  
##  Median :48.50   Mode  :character   Median :  74.00   Median :   8.900  
##  Mean   :48.50                      Mean   : 218.98   Mean   : 108.328  
##  3rd Qu.:72.25                      3rd Qu.: 260.00   3rd Qu.:  94.750  
##  Max.   :96.00                      Max.   :4480.00   Max.   :2800.000  
##    Gestation         Litter    
##  Min.   : 16.0   Min.   :1.00  
##  1st Qu.: 63.0   1st Qu.:1.00  
##  Median :133.5   Median :1.20  
##  Mean   :151.3   Mean   :2.31  
##  3rd Qu.:226.2   3rd Qu.:3.20  
##  Max.   :655.0   Max.   :8.00
```

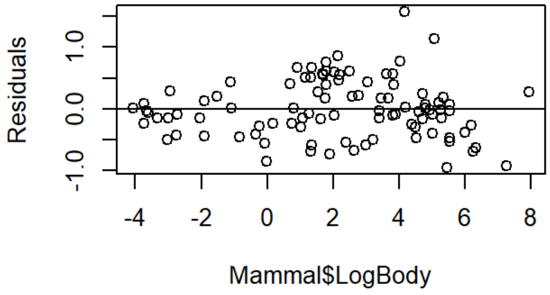
Part a: A Matrix of Scatterplots with all variables transformed to their logarithms

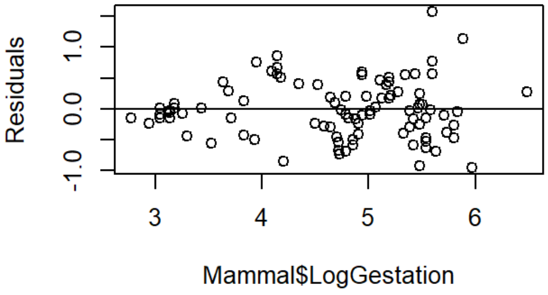
```
#Transform all variables to their logarithims
Mammal$LogBrain <- log(Mammal$Brain)
Mammal$LogBody <- log(Mammal$Body)
Mammal$LogGestation <- log(Mammal$Gestation)
Mammal$LogLitter <- log(Mammal$Litter)

#Using pairs to create a matrix of scatterplots
pairs(Mammal[,c(7,8,9,10)])
```

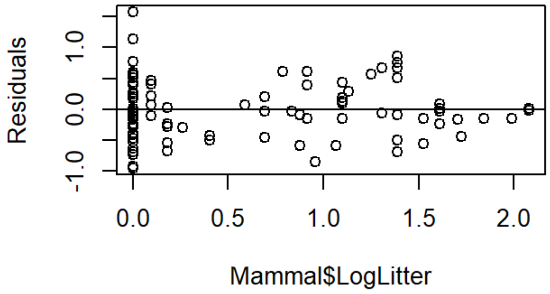


Part b: Fit a Multiple Linear Regression Model (log scale for all variables)

```
regLogMammal = lm(LogBrain~LogBody + LogGestation + LogLitter, data = Mammal)

#to do check of residuals versus each predictor
par(mfrow=c(2,2))
plot(regLogMammal$residuals, x=Mammal$LogBody, ylab="Residuals")
abline(0,0)
plot(regLogMammal$residuals, x=Mammal$LogGestation, ylab="Residuals")
abline(0,0)
plot(regLogMammal$residuals, x=Mammal$LogLitter, ylab="Residuals")
abline(0,0)
```







The horn-shape pattern exists in all three residual plots, violating the constant variance assumption. Therefore, this model is not very reliable and log transformation hasn't solve the fanning problem.

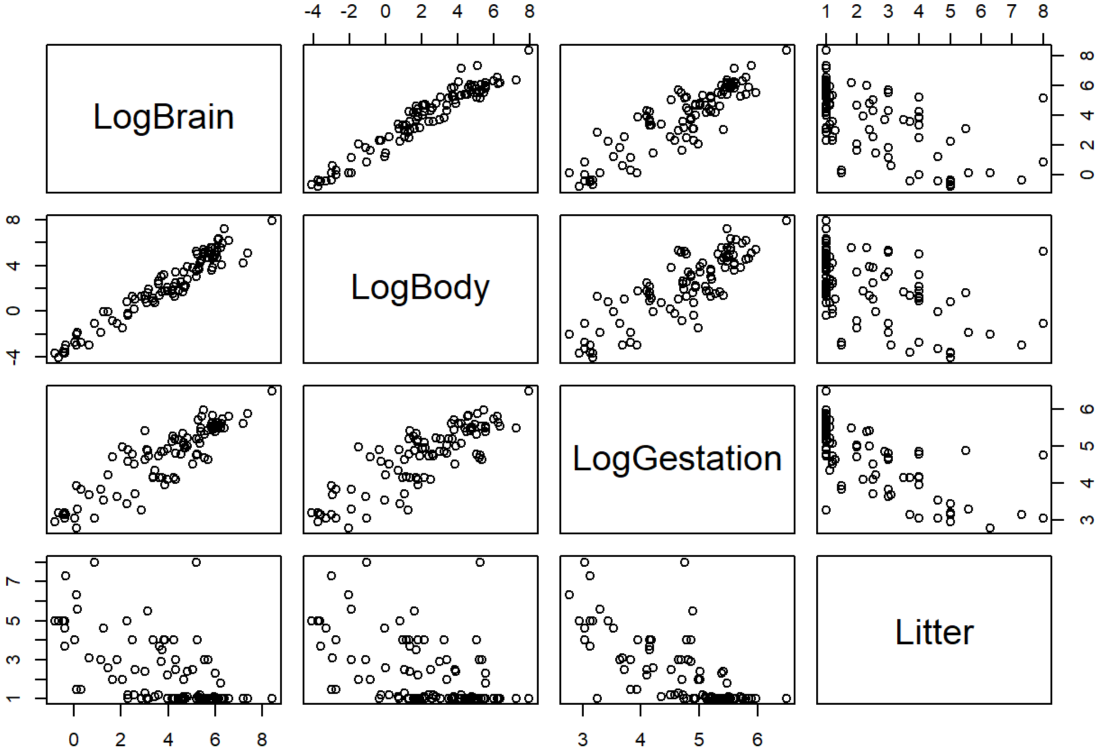
```
#Table of the regression output
summary(regLogMammal)
```

```
##
## Call:
## lm(formula = LogBrain ~ LogBody + LogGestation + LogLitter, data = Mammal)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95415 -0.29639 -0.03105  0.28111  1.57491
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.85482    0.66167   1.292  0.19962
## LogBody       0.57507    0.03259  17.647  < 2e-16 ***
## LogGestation  0.41794    0.14078   2.969  0.00381 **
## LogLitter    -0.31007    0.11593  -2.675  0.00885 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4748 on 92 degrees of freedom
## Multiple R-squared:  0.9537, Adjusted R-squared:  0.9522
## F-statistic: 631.6 on 3 and 92 DF,  p-value: < 2.2e-16
```

```
#Associated 95% Confidence Interval
confint(regLogMammal)
```

```
##                   2.5 %      97.5 %
## (Intercept)  -0.4593167  2.16896055
## LogBody       0.5103490  0.63979373
## LogGestation  0.1383359  0.69754827
## LogLitter    -0.5403124 -0.07982996
```

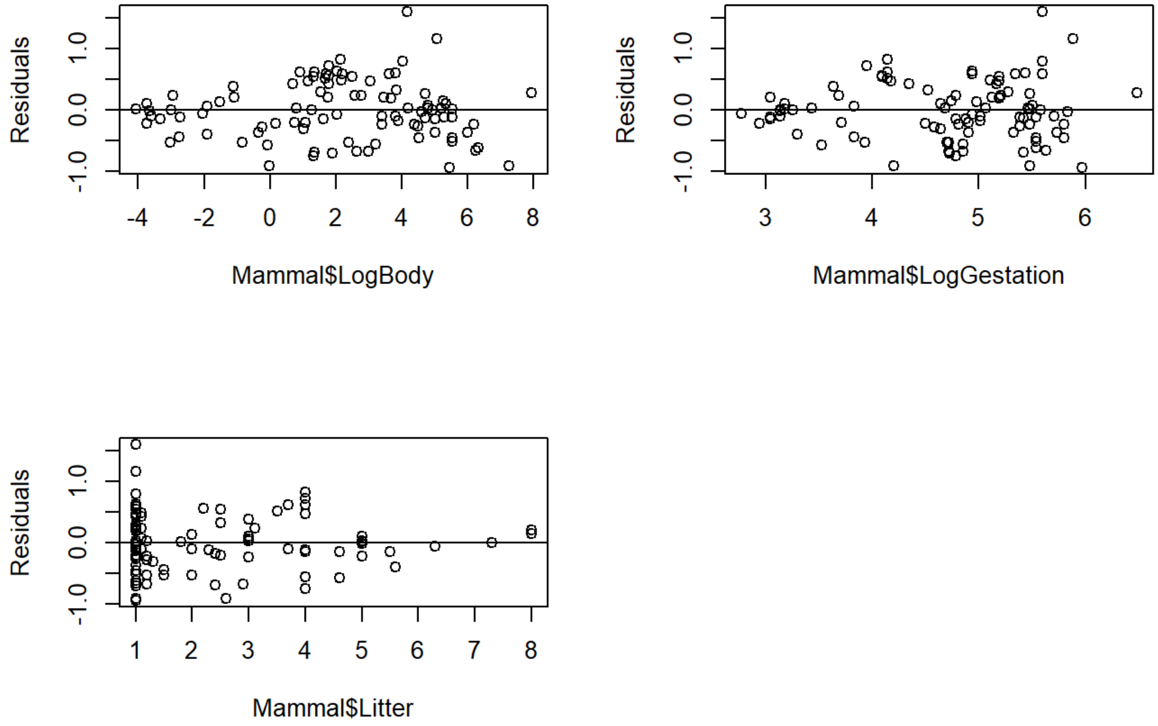
Part c: A Matrix of Scatterplots with Litter Size on Its Natural Scale



The relationship between log brain weight and litter size now seems to be quadratic instead of linear. However, the relationship between log brain weight and log litter size seems to be more linear.

Part d: Fit a Multiple Linear Regression Model (natural scale for litter size)

```
regLogMammal_nonLogLitter = lm(LogBrain~LogBody + LogGestation + Litter, data = Mammal)

#to do check of residuals versus each predictor
par(mfrow=c(2,2))
plot(regLogMammal_nonLogLitter$residuals,x=Mammal$LogBody,ylab="Residuals")
abline(0,0)
plot(regLogMammal_nonLogLitter$residuals,x=Mammal$LogGestation,ylab="Residuals")
abline(0,0)
plot(regLogMammal_nonLogLitter$residuals,x=Mammal$Litter,ylab="Residuals")
abline(0,0)
```



The horn-shape pattern exists in all three residual plots, violating the constant variance assumption. Therefore, the second model is also showing a poor fit.

```
#Table of the regression output
summary(regLogMammal_nonLogLitter)
```

```
## 
## Call:
## lm(formula = LogBrain ~ LogBody + LogGestation + Litter, data = Mammal)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.93895 -0.27922 -0.00929  0.28646  1.59743 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   0.82338    0.66206   1.244  0.21678    
## LogBody       0.57455    0.03264  17.601  < 2e-16 ***
## LogGestation  0.43964    0.13698   3.210  0.00183 ** 
## Litter       -0.11038    0.04227  -2.611  0.01053 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.4756 on 92 degrees of freedom
## Multiple R-squared:  0.9535, Adjusted R-squared:  0.952 
## F-statistic: 629.4 on 3 and 92 DF,  p-value: < 2.2e-16
```

```
#Associated 95% Confidence Interval on log scale
confint(regLogMammal_nonLogLitter)
```

```
##                   2.5 %      97.5 %
## (Intercept)  -0.4915254  2.13829063
## LogBody       0.5097143  0.63937813
## LogGestation  0.1675856  0.71169994
## Litter       -0.1943220 -0.02643223
```

```
#Exponentiate to get interpretation as ratio of medians
exp(confint(regLogMammal_nonLogLitter))
```

```
##                  2.5 %   97.5 %
## (Intercept)  0.6116926 8.484921
## LogBody      1.6648155 1.895302
## LogGestation 1.1824465 2.037452
## Litter       0.8233927 0.973914
```

Part e: Interpretation of Coefficients in Part d

1. Coefficients from the regression and confidence interval

- LogBody: Slope estimation for LogBody is 0.57507 (95% CI:0.5097143 to 0.63937813). A doubling of body weight is associated with a multiplicative change of $2^{0.57507}$ (CI: $2^{0.51}$ to $2^{0.64}$) in the median of brain wieght.
- LogGestation: Slope estimation for LogGestation is 0.43964 (95% CI: 0.1675856 to 0.71169994). A doubling of gestation is associated with a multiplicative change of $2^{0.0.43964}$ (CI:$2^{0.17}$ to $2^{0.71}$) in the median of brain wieght.
- Litter: Slope estimation for Litter Size is -0.11038 (95% CI: -0.1943220 to -0.02643223) An increase in litter size of 1 unit is associated with a multiplicative change of $e^{-0.11038}$ (CI: $e^{-0.19}$ to $e^{-0.026}$) in median of brain weight.

Part f: Based on the quality of residual plots and the value of R^2

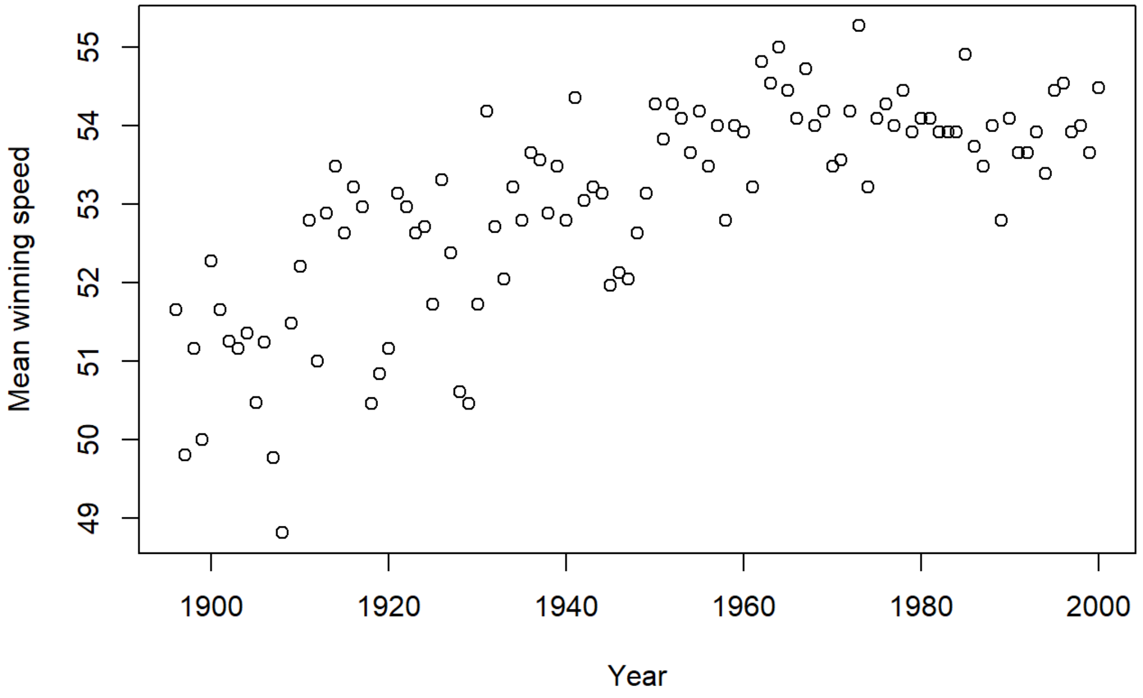
Value of R square in Part B is 0.9537 which is 0.0002 higher than the value in part D. However, both methods exhibit horn-shape patterns in the residual plots , so they’re not reliable models. Therefore, I’m ambivalent between the two. While in terms of interpretability, it seems that model in part D is easier to interpretate compared to model in part B since Litter size isn’t transformed to its logarithm and its coefficient is relatively easier and meaningful to interpret.

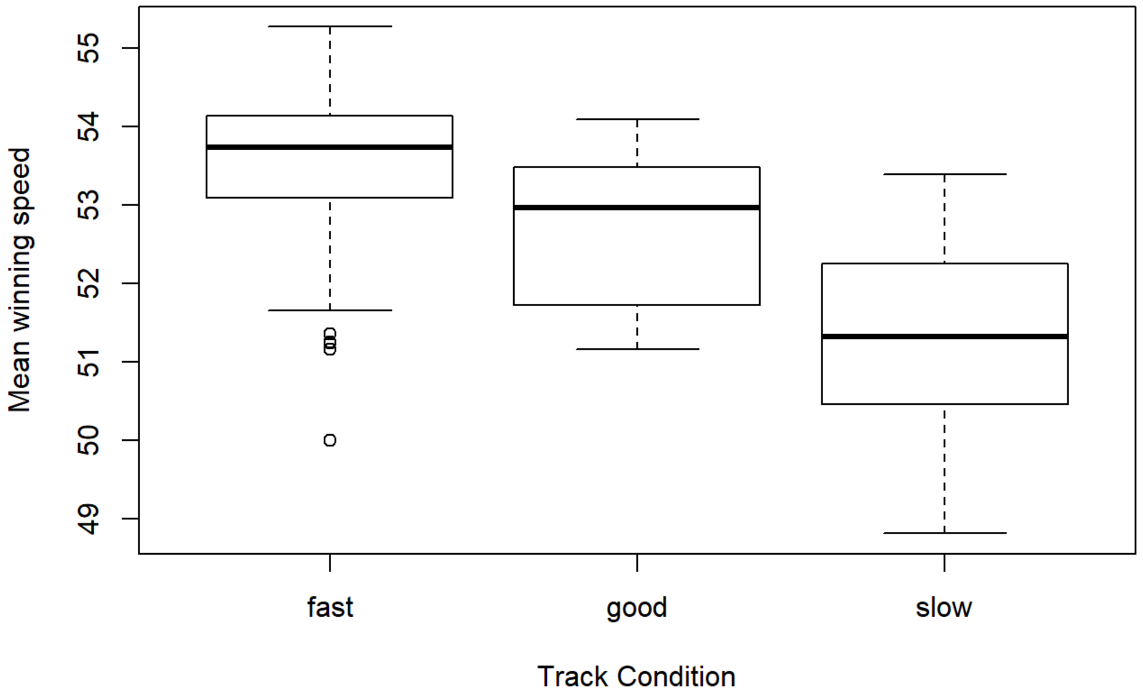
Question 3:Kentucky Derby Data Analysis Writeup

We aim to predict average winning speed from predictors such as year and track condition. Let’s first look at the summary of this data set.

```
##        X            Year         Winner           Condition
##  Min.   :  1   Min.   :1896   Length:105         Length:105
##  1st Qu.: 27   1st Qu.:1922   Class :character   Class :character
##  Median : 53   Median :1948   Mode  :character   Mode  :character
##  Mean   : 53   Mean   :1948
##  3rd Qu.: 79   3rd Qu.:1974
##  Max.   :105   Max.   :2000
##      Speed
##  Min.   :48.82
##  1st Qu.:52.28
##  Median :53.40
##  Mean   :53.04
##  3rd Qu.:54.01
##  Max.   :55.28
```

1. Exploratory Data Analysis

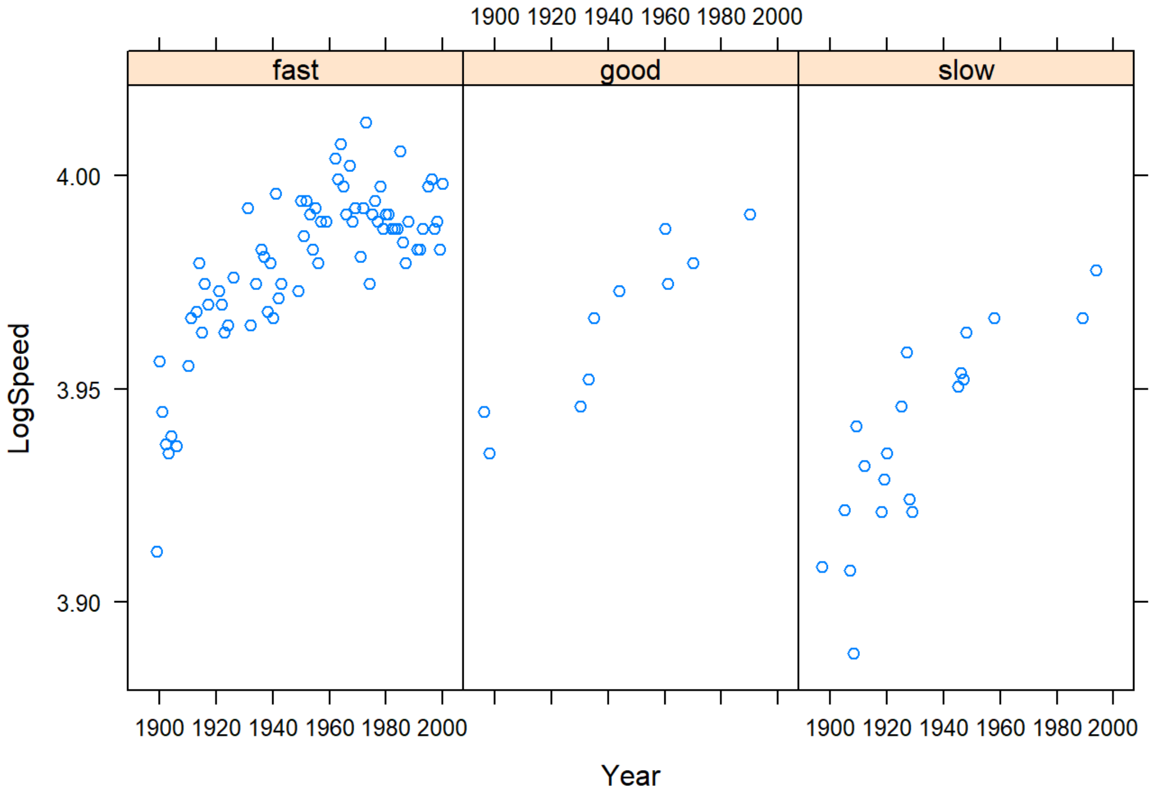




The scatter plot of mean winning speed versus year is not a straight line and the variability decreases as the mean of the response variable increases. There seems to be a quadratic relationship with year exhibited in the first plot. And we should try both a transformation of $log(speed)$ and $Year^2$.

The boxplot of mean winning speed versus track condition violates the constant variance assumption, so indeed we need some transformations.

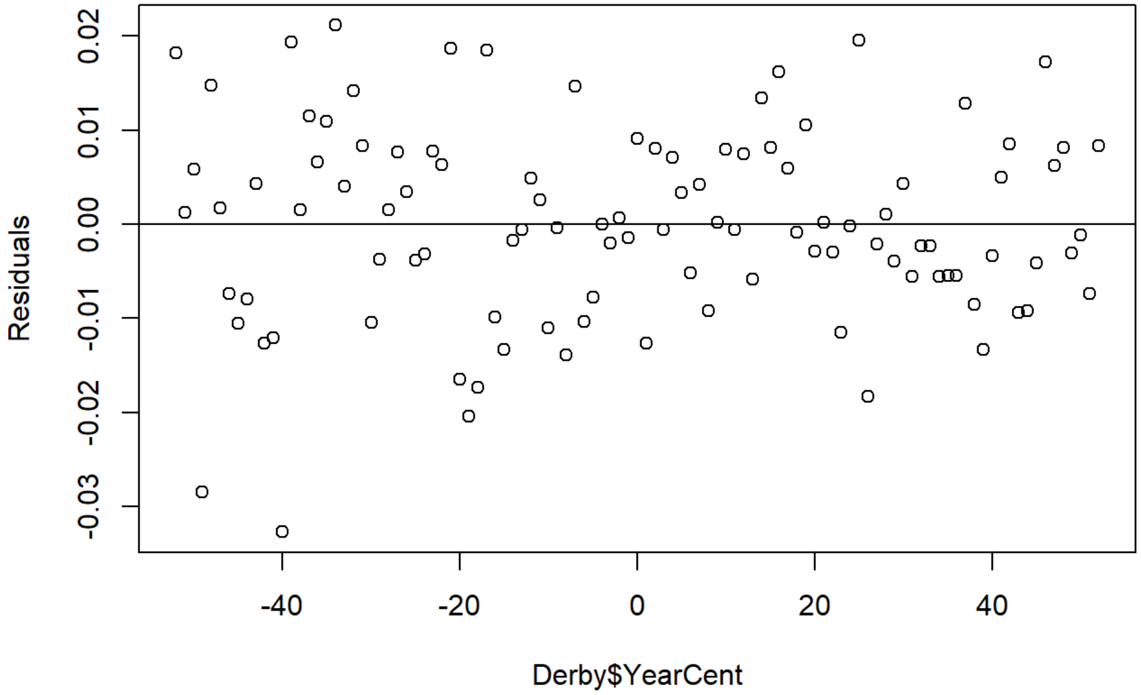
```
##Transformations
#make log of speed
Derby$LogSpeed <- log(Derby$Speed)

#it helps interpretation to mean-centering Year
Derby$YearCent <- Derby$Year - mean(Derby$Year)

#make year squared
Derby$Year2 <- Derby$YearCent^2

#See whether there is interaction effects using trellis plots.
library(lattice)
xyplot(LogSpeed~Year|Condition,data = Derby)
```



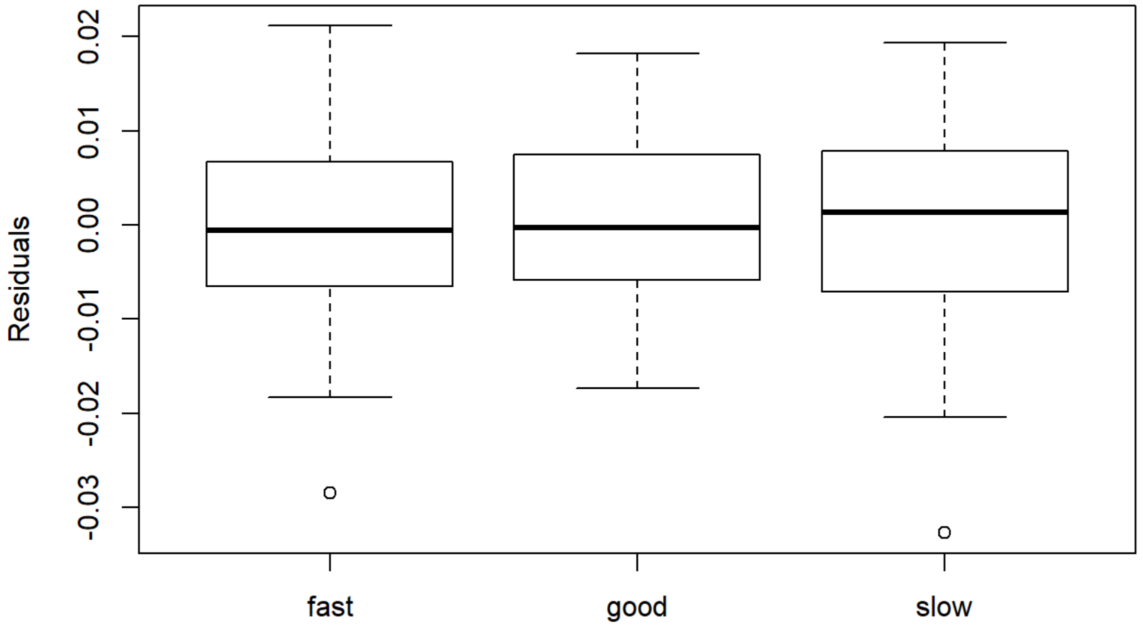
Slope in each plot is similar, so no strong evidence of interaction between LogSpeed and year.

2. Fitting the Multiple Regression

```
#Create series of indicator variable for condition
n <- nrow(Derby)
Derby$slow <- rep(0,n)
Derby$slow[Derby$Condition == "slow"] = 1

Derby$good <- rep(0,n)
Derby$good[Derby$Condition == "good"] = 1

Derby$fast <- rep(0,n)
Derby$fast[Derby$Condition == "fast"] = 1

#here is the regression
regDerby <- lm(LogSpeed~YearCent + Year2 + good + fast, data = Derby)

#Check of Residuals versus Each Predictor
plot(regDerby$residuals, x=Derby$YearCent, ylab = "Residuals")
abline(0,0)
```



```
boxplot(regDerby$residuals~Derby$Condition, ylab = "Residuals")
```

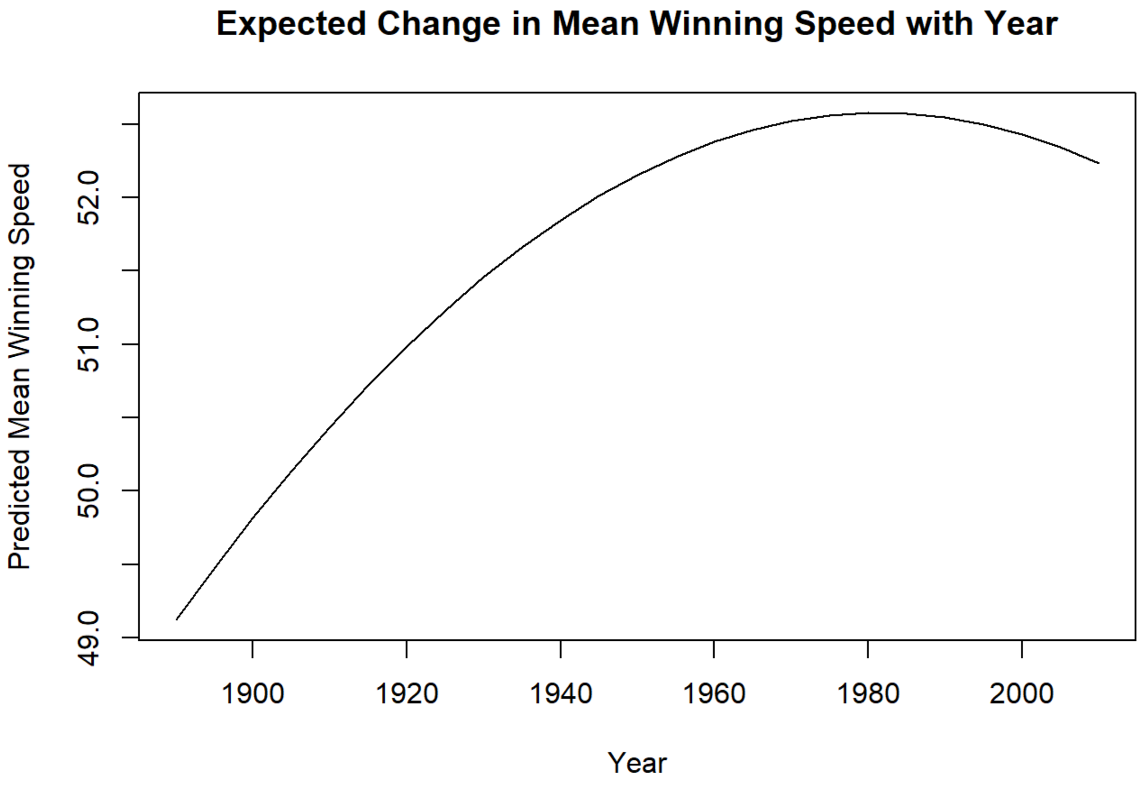


No real patterns show up in the scatter plot and dots are randomly scattered above and below x axis except for two outliers in the lower left corner. Constant-variance and normality assumptions are satisfied in this case.

3. Summary and Interpretation

```
##
## Call:
## lm(formula = LogSpeed ~ YearCent + Year2 + good + fast, data = Derby)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.032690 -0.005823 -0.000398  0.007473  0.021123
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  3.954e+00  2.686e-03 1472.060  < 2e-16 ***
## YearCent     5.118e-04  3.528e-05   14.505  < 2e-16 ***
## Year2       -8.118e-06  1.248e-06   -6.504 3.11e-09 ***
## good         2.087e-02  4.085e-03    5.108 1.56e-06 ***
## fast         3.084e-02  2.752e-03   11.207  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0105 on 100 degrees of freedom
## Multiple R-squared:  0.8358, Adjusted R-squared:  0.8292
## F-statistic: 127.3 on 4 and 100 DF,  p-value: < 2.2e-16
```

R-square for this model is 0.8358, which shows that it’s a really good prediction model. Residual SE is 0.015. All predictors have p-value less than 0.05, so they’re all pretty significant. Therefore, I choose this model for prediction.



It’s hard to interpret quadratic effect in words, thus I include a plot showing the relationship between mean winning speed and year in “slow” condition above.

```
##                  2.5 %     97.5 %
## (Intercept) 51.8777438 52.4336457
## YearCent     1.0004419  1.0005820
## Year2        0.9999894  0.9999944
## good         1.0128435  1.0293935
## fast         1.0257020  1.0369620
```

In 1948 (YearCent=0), the mean winning speed with “slow” condition is $e^{3.954} \approx 52.143$ (95% CI: 51.8777438, 52.4336457).

The exponentiated confidence interval shows that if we change condition of the track from “slow” to “good”, we have 95% confidence that the mean winning speed would increase by (1.0128435, 1.0293935).

Brief Conclusion:

The model I chose for Kentucky Derby data set is a multiple regression model with log(Speed) on Year(mean-centered), Year squared. I chose the log transformation of Speed and quadratic effects on Year in the innitial exploratory data analysis process. I’m fairly satisfied with the quality of the fit. The only limitation might be that there are more relevant predictors existed, which could further boost r-square of the model. In addition, quadratic effect for year might imply that the winning speed eventually will decrease and even be negative in the far future, which makes no sense.

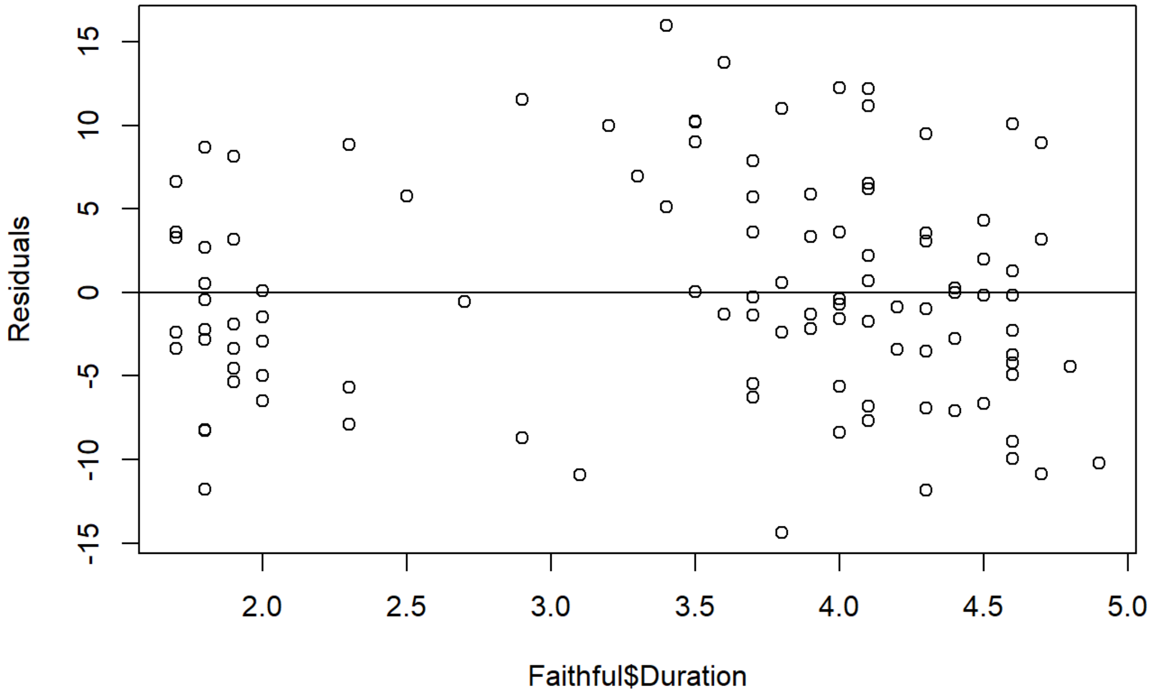
Question 4: Old Faithful

We aim to determine if “Date” variable is a useful predictor for Old Faithful eruption intervals, which means we need to compare fit of the model with and without “Date”. Let’s first look at the summary of this data set.

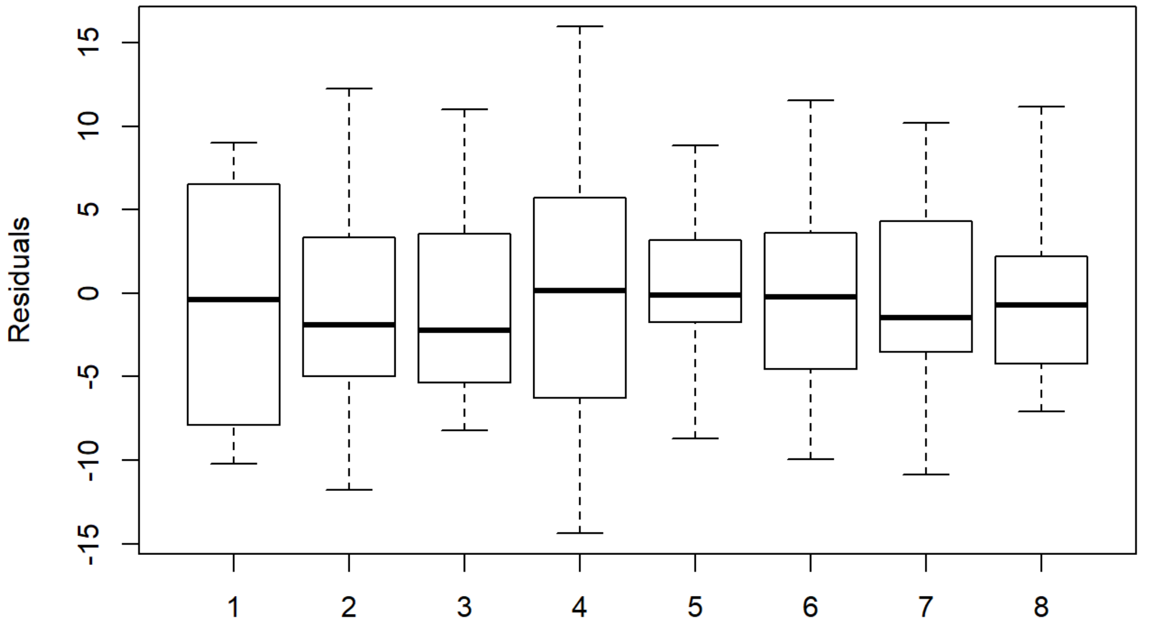
```
##        X              Date          Interval       Duration    
##  Min.   :  1.0   Min.   :1.000   Min.   :42.0   Min.   :1.700  
##  1st Qu.: 27.5   1st Qu.:3.000   1st Qu.:59.0   1st Qu.:2.300  
##  Median : 54.0   Median :5.000   Median :75.0   Median :3.800  
##  Mean   : 54.0   Mean   :4.514   Mean   :71.0   Mean   :3.461  
##  3rd Qu.: 80.5   3rd Qu.:6.000   3rd Qu.:80.5   3rd Qu.:4.300  
##  Max.   :107.0   Max.   :8.000   Max.   :95.0   Max.   :4.900
```

```
#Create series of indicator variables for Date
m=nrow(Faithful)

Faithful$Day1 <- rep(0,m)
Faithful$Day1[Faithful$Date == 1] = 1

Faithful$Day2 <- rep(0,m)
Faithful$Day2[Faithful$Date == 2] = 1

Faithful$Day3 <- rep(0,m)
Faithful$Day3[Faithful$Date == 3] = 1

Faithful$Day4 <- rep(0,m)
Faithful$Day4[Faithful$Date == 4] = 1

Faithful$Day5 <- rep(0,m)
Faithful$Day5[Faithful$Date == 5] = 1

Faithful$Day6 <- rep(0,m)
Faithful$Day6[Faithful$Date == 6] = 1

Faithful$Day7 <- rep(0,m)
Faithful$Day7[Faithful$Date == 7] = 1

Faithful$Day8 <- rep(0,m)
Faithful$Day8[Faithful$Date == 8] = 1
```

```
#Fit a regression with duration and day treated as a factor
regFaithfulDum <- lm(Interval~Duration + as.factor(Date), data = Faithful)

#Check of Residuals versus Each Predictor
plot(regFaithfulDum$residuals, x=Faithful$Duration, ylab= "Residuals")
abline(0,0)
```



```
boxplot(regFaithfulDum$residuals~Faithful$Date, ylab = "Residuals")
```



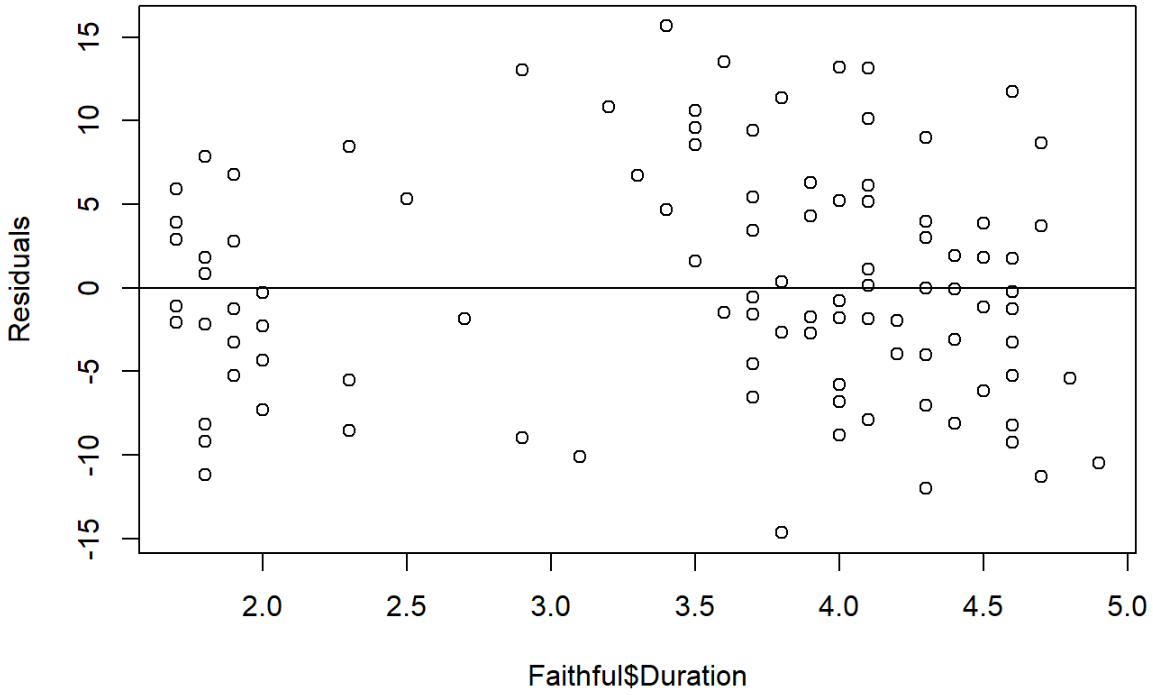
No real patterns show up in the scatter plot and dots are randomly scattered above and below x axis. Box plot doesn’t show any significant pattern. Constant-variance and normality assumptions are satisfied in this case.

```
## 
## Call:
## lm(formula = Interval ~ Duration + as.factor(Date), data = Faithful)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -14.3886  -4.7332  -0.5622   3.9759  15.9639 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)        32.8770     3.0672  10.719   <2e-16 ***
## Duration           10.8813     0.6622  16.431   <2e-16 ***
## as.factor(Date)2    1.3275     2.7173   0.489    0.626    
## as.factor(Date)3    0.7825     2.6994   0.290    0.773    
## as.factor(Date)4    0.1625     2.6461   0.061    0.951    
## as.factor(Date)5    0.2463     2.6459   0.093    0.926    
## as.factor(Date)6    1.9918     2.6580   0.749    0.455    
## as.factor(Date)7   -0.1700     2.7020  -0.063    0.950    
## as.factor(Date)8   -0.6944     2.6957  -0.258    0.797    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6.866 on 98 degrees of freedom
## Multiple R-squared:  0.7408, Adjusted R-squared:  0.7196 
## F-statistic:    35 on 8 and 98 DF,  p-value: < 2.2e-16
```

```
##                       2.5 %    97.5 %
## (Intercept)      26.790153 38.963809
## Duration          9.567135 12.195545
## as.factor(Date)2 -4.064928  6.719930
## as.factor(Date)3 -4.574299  6.139349
## as.factor(Date)4 -5.088529  5.413622
## as.factor(Date)5 -5.004338  5.496977
## as.factor(Date)6 -3.282858  7.266529
## as.factor(Date)7 -5.532115  5.192078
## as.factor(Date)8 -6.043893  4.655112
```

F statistic in this model is 35 on 8 and 98 DF and its p-value is less than 2.2e-16. Confidence interval for all indicator variables include 0, which means that those variables should be ruled out in the prediction model.

```
#Fit the regression of interval on duration alone
regFaithful <- lm(Interval~Duration, data = Faithful)

#Check of Residuals versus Predictor
plot(regFaithful$residuals, x=Faithful$Duration, ylab= "Residuals")
abline(0,0)
```



No real patterns show up in the scatter plot and dots are randomly scattered above and below x axis. Constant-variance and normality assumptions are satisfied in this case.

```
##
## Call:
## lm(formula = Interval ~ Duration, data = Faithful)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.644  -4.440  -1.088   4.467  15.652
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.8282     2.2618   14.96   <2e-16 ***
## Duration     10.7410     0.6263   17.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.683 on 105 degrees of freedom
## Multiple R-squared:  0.7369, Adjusted R-squared:  0.7344
## F-statistic: 294.1 on 1 and 105 DF,  p-value: < 2.2e-16
```

F statistic in this model is 294.1 on 1 and 105 DF and its p-values is less than 2.2e-16. Duration which is the only predictor in this model is quite significant in terms of its p-value. Deleting date variable doesn't change p-value of the model very much. Thus, we could delete date.

```
#nested f test to see if a whole set of dummy variables (date) are significant
anova(regFaithfulDum, regFaithful)
```

```
## Analysis of Variance Table
## 
## Model 1: Interval ~ Duration + as.factor(Date)
## Model 2: Interval ~ Duration
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     98 4620.2                           
## 2    105 4689.0 -7   -68.853 0.2086 0.9828
```

Again, when we use nested f test to see if whole set of dummy variables are significant, F statistic in this case is 0.2086. P-value is 0.9828, which is way bigger than 0.05. This is too big a p-value, so we can conclude that there is no effect of Date on Interval. We can thus rule out the effect of date to interval of geyser eruptions.

Question 5: Wage and Races

We aim to decide whether and to what extent black males were paid less than non black males holding all other variables constant, which means we need to decide whether “Black” is an important predictor for estimating wage. Let’s first look at the summary of this data set.

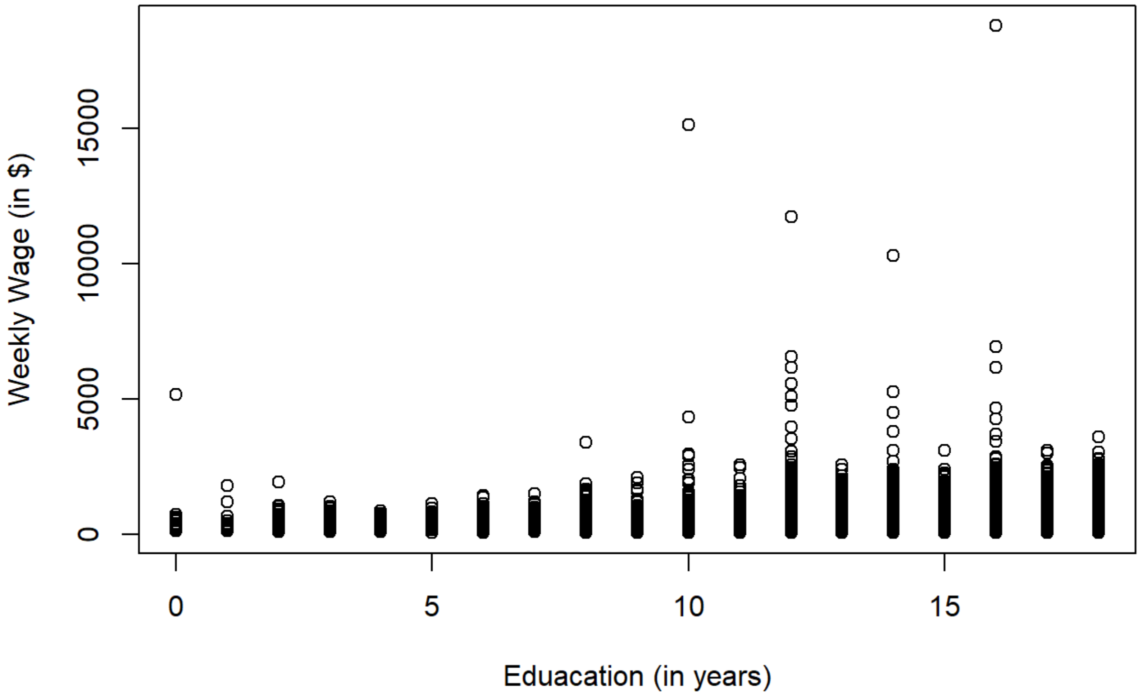
```
##        X              Wage            Education       Experience   
##  Min.   :    1   Min.   :   50.39   Min.   : 0.00   Min.   :-4.00  
##  1st Qu.: 6408   1st Qu.:  356.13   1st Qu.:12.00   1st Qu.: 9.00  
##  Median :12816   Median :  567.23   Median :12.00   Median :16.00  
##  Mean   :12816   Mean   :  640.16   Mean   :13.08   Mean   :18.59  
##  3rd Qu.:19224   3rd Qu.:  826.21   3rd Qu.:16.00   3rd Qu.:27.00  
##  Max.   :25631   Max.   :18777.20   Max.   :18.00   Max.   :63.00  
##     Black               SMSA              Region         
##  Length:25631       Length:25631       Length:25631      
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##                                                          
##                                                          
## 
```

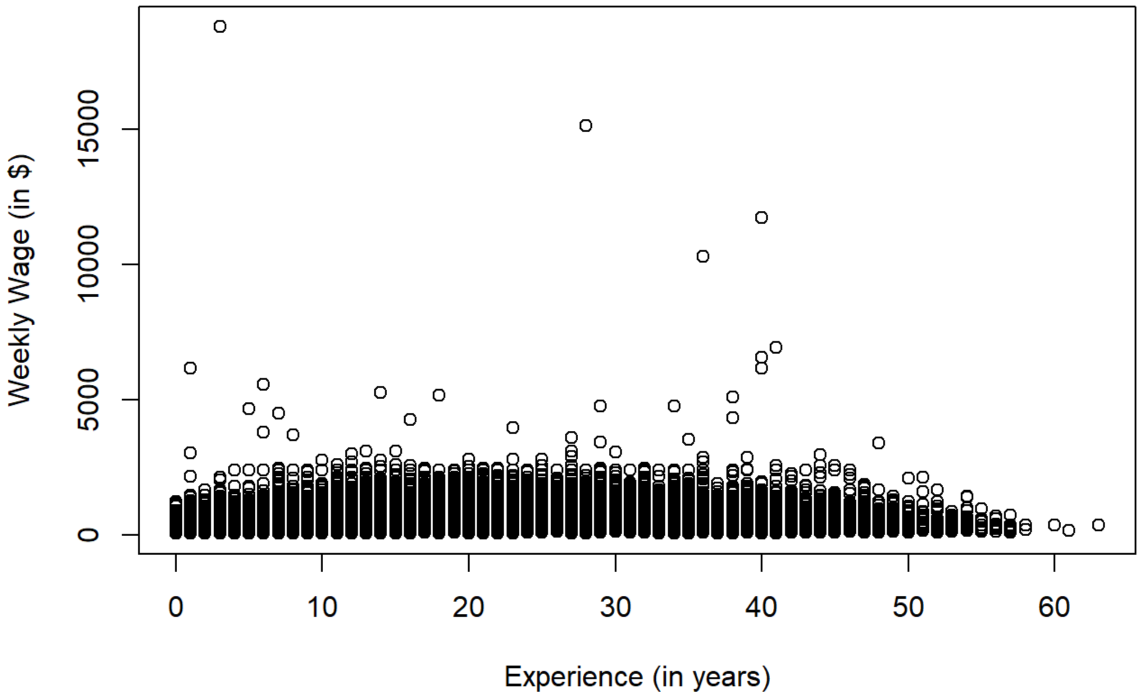
It seems that there are negative entries for the year of experience and it doesn’t make sense to include those data. Therefore, we choose to drop them before we make further exploration.

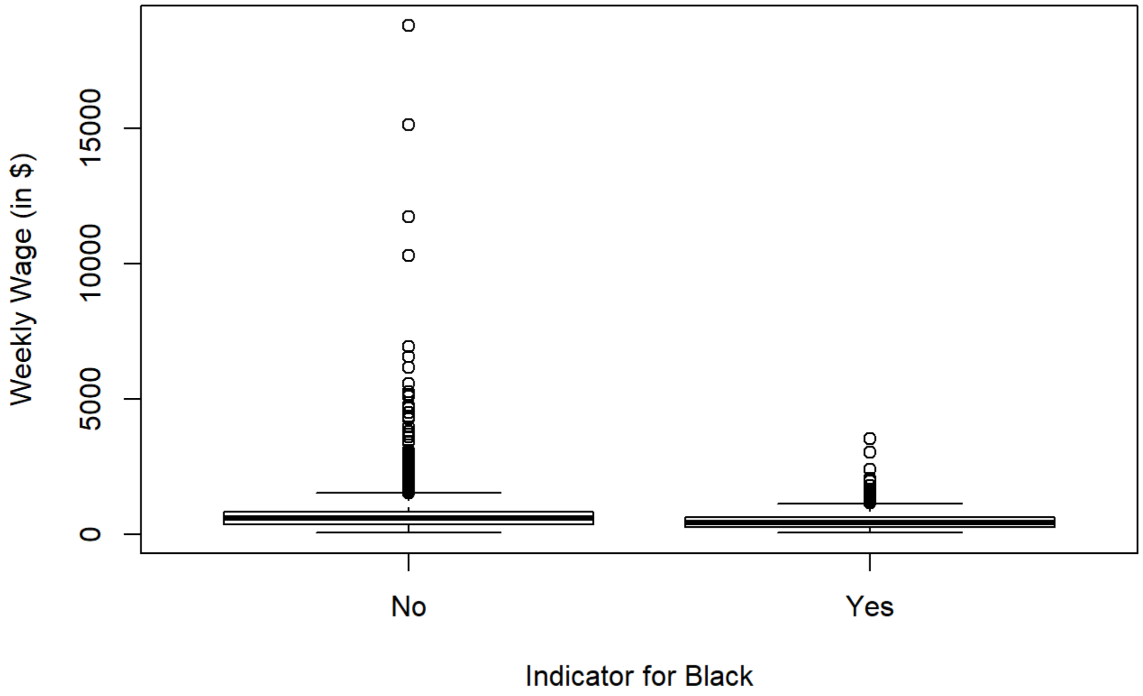
```
weeklywage <- weeklywage[weeklywage$Experience >= 0,]
weeklywage <- weeklywage[weeklywage$Education >= 0,]
```

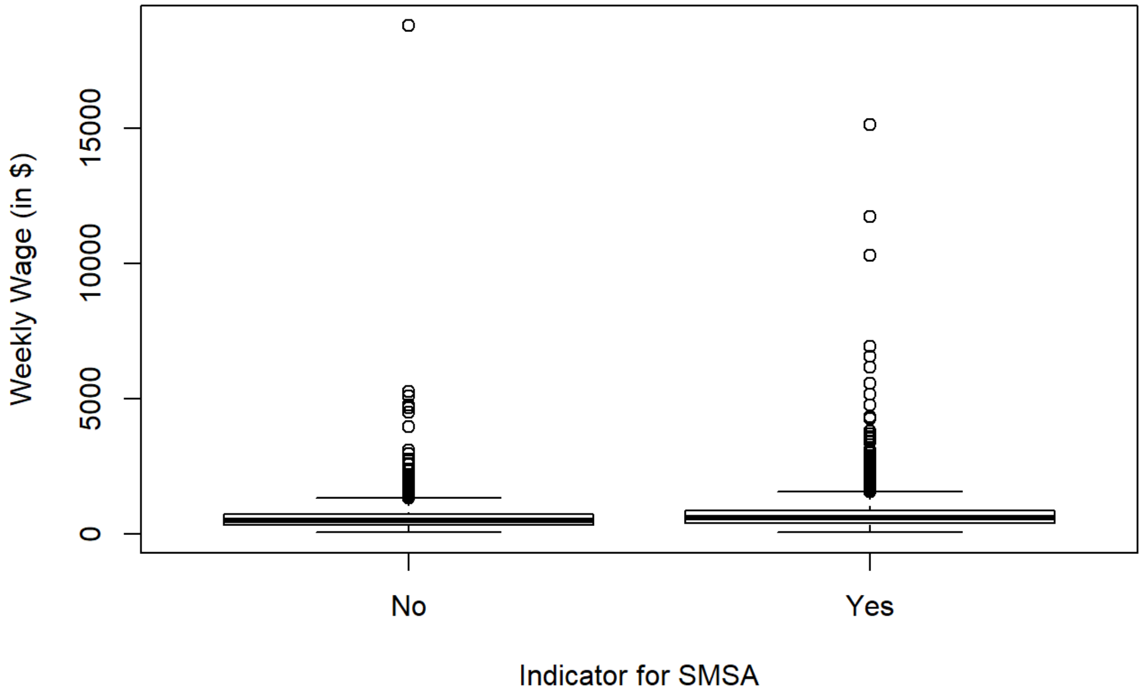
Exploratory Data Analysis

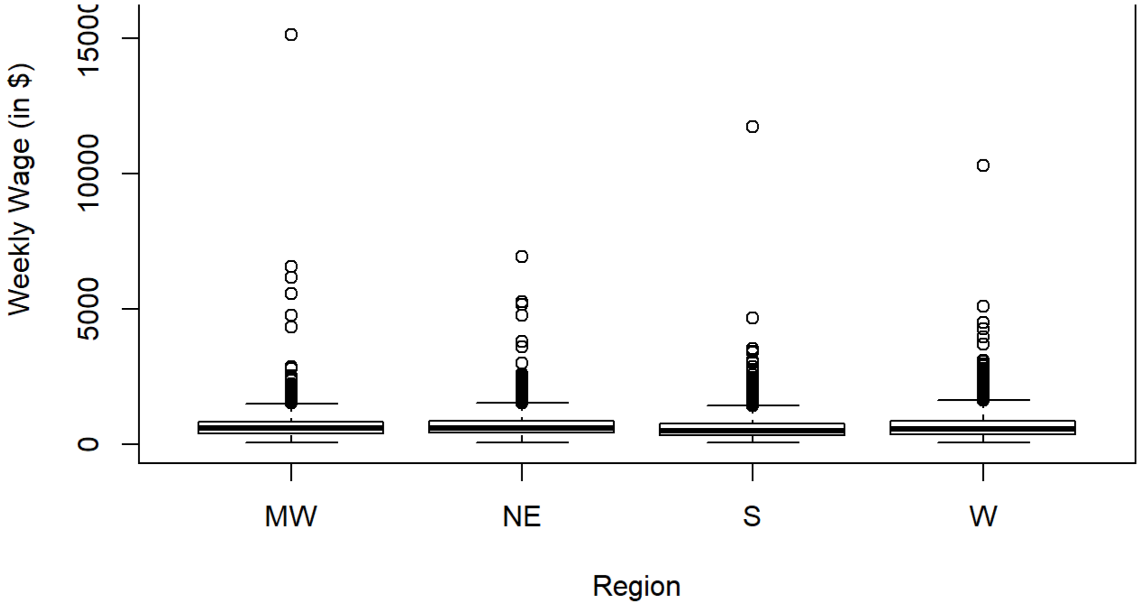
Plotting weekly wage against all predictors:


Weekly Wage (in $)
0
5000
10000
15000
0
5
10
15
Eduacation (in years)


Weekly Wage (in $)
0
5000
10000
15000
0
10
20
30
40
50
60
Experience (in years)

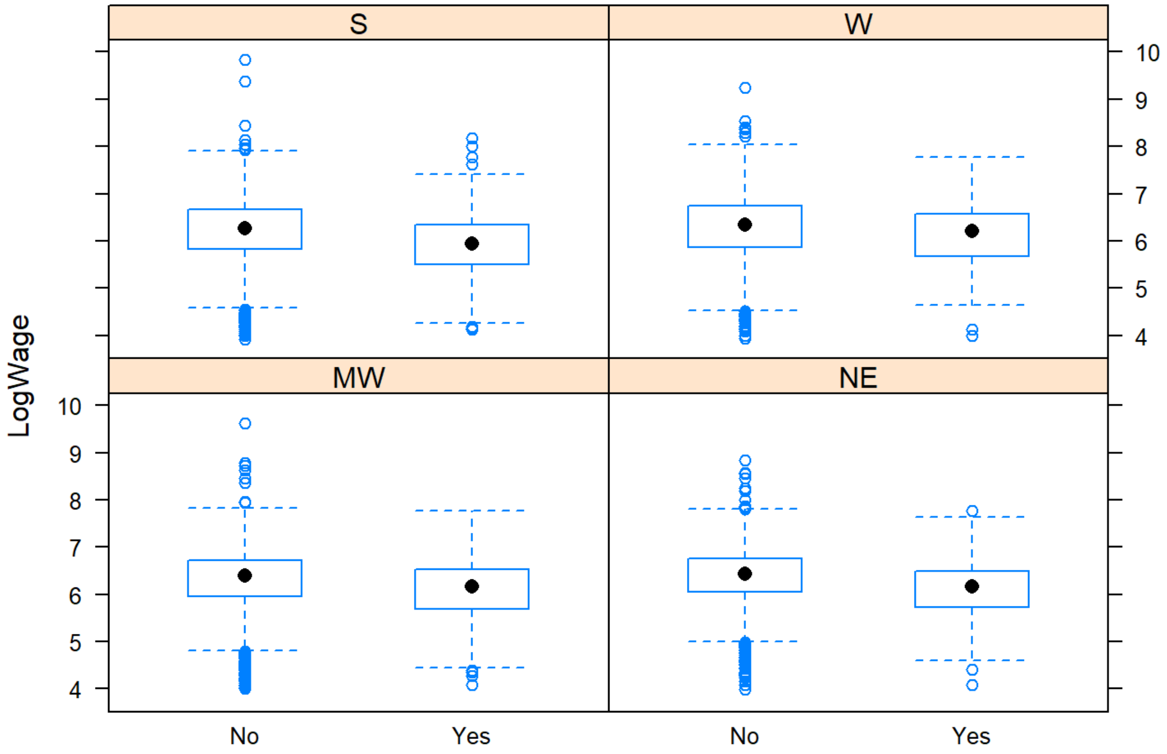

Weekly Wage (in $)
0
5000
10000
15000
No
Yes
Indicator for Black


Weekly Wage (in $)
0
5000
10000
15000
No
Yes
Indicator for SMSA



From the plot of weekly wages versus region, we can see that variability increases as the mean of the response variable increases (a fanning-out shape). So we could first try to transform Wage to its logarithm.

```
#make log of weekly wages
weeklywage$LogWage <- log(weeklywage$Wage)

#The problem also mentions to only consider the interaction effect between region and whether being black or not. Therefore, we make a trellis plot to explore.
bwplot(LogWage~Black | Region, data = weeklywage)
```

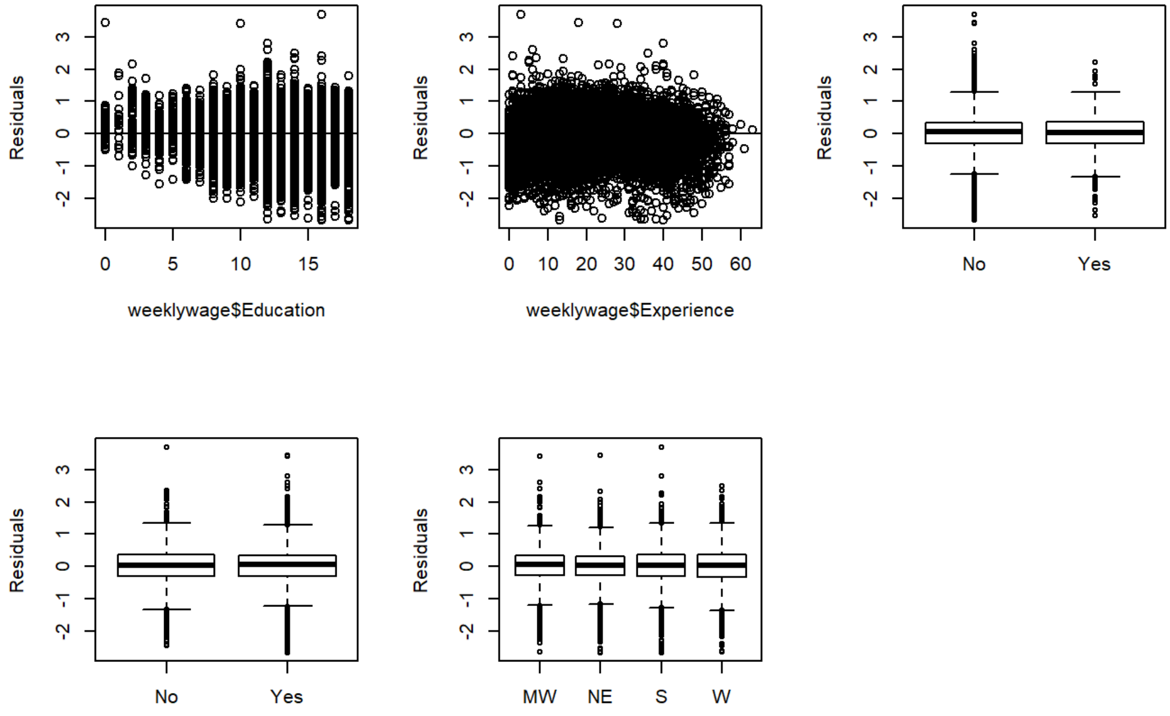


No strong pattern exists. But as the question asks for the extent of the interaction effect, we could later put the interaction effect in the model and use nested F test to test.

Model 1: log(Wage)

```
#Fitting the multiple regression
regweeklywage1 <- lm(LogWage~Education + Experience + as.factor(Black) + as.factor(SMSA) + as.factor(Regio
n), data = weeklywage)
```

To do check of residuals versus each predictor:



There are inconsistent variance problems in first two scatter plots. Thus the constant variance assumption is not satisfied.

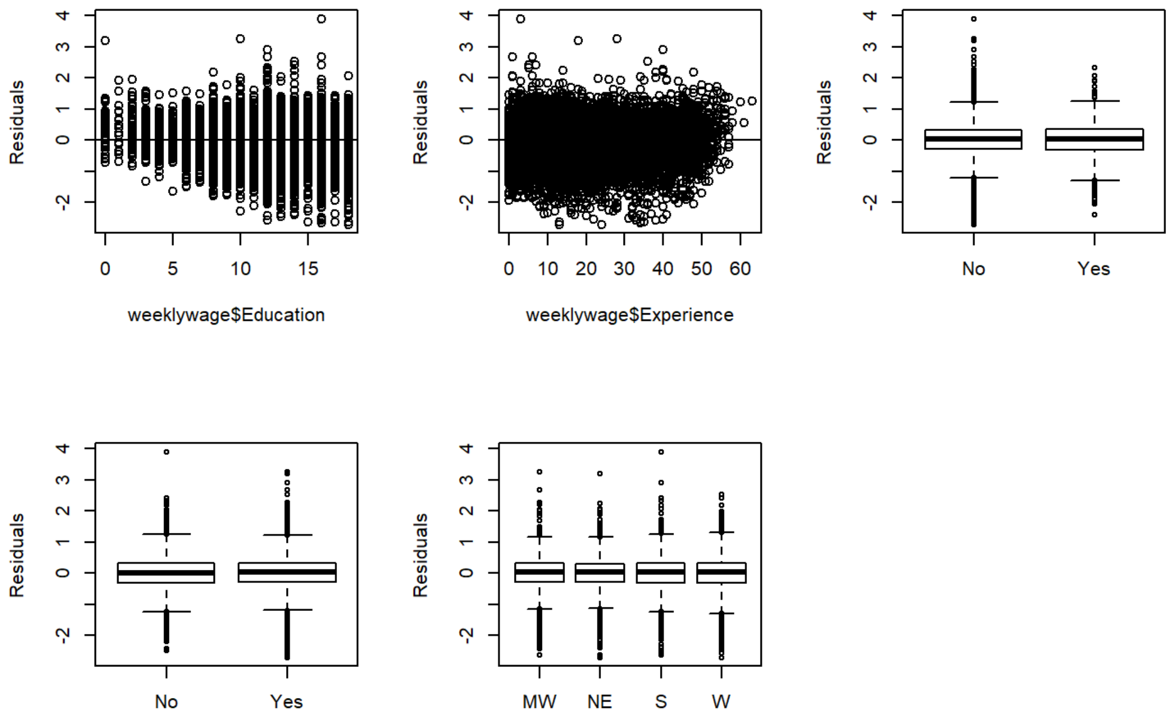
```
##
## Call:
## lm(formula = LogWage ~ Education + Experience + as.factor(Black) +
##     as.factor(SMSA) + as.factor(Region), data = weeklywage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6881 -0.2997  0.0432  0.3425  3.7050
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4.5878455  0.0196372 233.631  < 2e-16 ***
## Education            0.0971174  0.0011958  81.213  < 2e-16 ***
## Experience           0.0178191  0.0002798  63.677  < 2e-16 ***
## as.factor(Black)Yes -0.2331872  0.0126232 -18.473  < 2e-16 ***
## as.factor(SMSA)Yes   0.1592142  0.0076933  20.695  < 2e-16 ***
## as.factor(Region)NE  0.0370482  0.0097032   3.818 0.000135 ***
## as.factor(Region)S  -0.0597360  0.0090537  -6.598 4.25e-11 ***
## as.factor(Region)W  -0.0026049  0.0098397  -0.265 0.791215
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5286 on 25429 degrees of freedom
## Multiple R-squared:  0.2827, Adjusted R-squared:  0.2825
## F-statistic:  1432 on 7 and 25429 DF,  p-value: < 2.2e-16
```

R-square of this model is 0.2827, which is relatively low. We could try to improve the fit of the model by adding quadratic term for experience (since there seems to be a quadratic relationship between wage and experience in the EDA stage.)

Model 2: log(Wage) and Experience squared

```
#make experience squared--it makes sense to directly square experience, because experience can be 0 year.
weeklywage$Experience2 <- weeklywage$Experience^2

#fit multiple regression model
regweeklywage2 <- lm(LogWage~Education + Experience + Experience2 + as.factor(Black) + as.factor(SMSA) + a
s.factor(Region), data = weeklywage)
```

To do check of residuals versus each predictor:



Similar with residual plots for the first model, there are inconsistent variance problems in first two scatter plots. Thus the constant variance assumption is not satisfied.

```
summary(regweeklywage2)
```

```
##
## Call:
## lm(formula = LogWage ~ Education + Experience + Experience2 +
##     as.factor(Black) + as.factor(SMSA) + as.factor(Region), data = weeklywage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7136 -0.2850  0.0349  0.3254  3.9057
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4.421e+00  1.937e-02 228.249  < 2e-16 ***
## Education            8.862e-02  1.172e-03  75.597  < 2e-16 ***
## Experience           5.496e-02  9.112e-04  60.315  < 2e-16 ***
## Experience2         -8.356e-04  1.958e-05 -42.681  < 2e-16 ***
## as.factor(Black)Yes -2.352e-01  1.219e-02 -19.288  < 2e-16 ***
## as.factor(SMSA)Yes   1.648e-01  7.433e-03  22.167  < 2e-16 ***
## as.factor(Region)NE  4.297e-02  9.374e-03   4.584 4.58e-06 ***
## as.factor(Region)S  -6.147e-02  8.746e-03  -7.029 2.14e-12 ***
## as.factor(Region)W  -1.136e-02  9.507e-03  -1.195    0.232
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5106 on 25428 degrees of freedom
## Multiple R-squared:  0.3307, Adjusted R-squared:  0.3305
## F-statistic:  1570 on 8 and 25428 DF,  p-value: < 2.2e-16
```

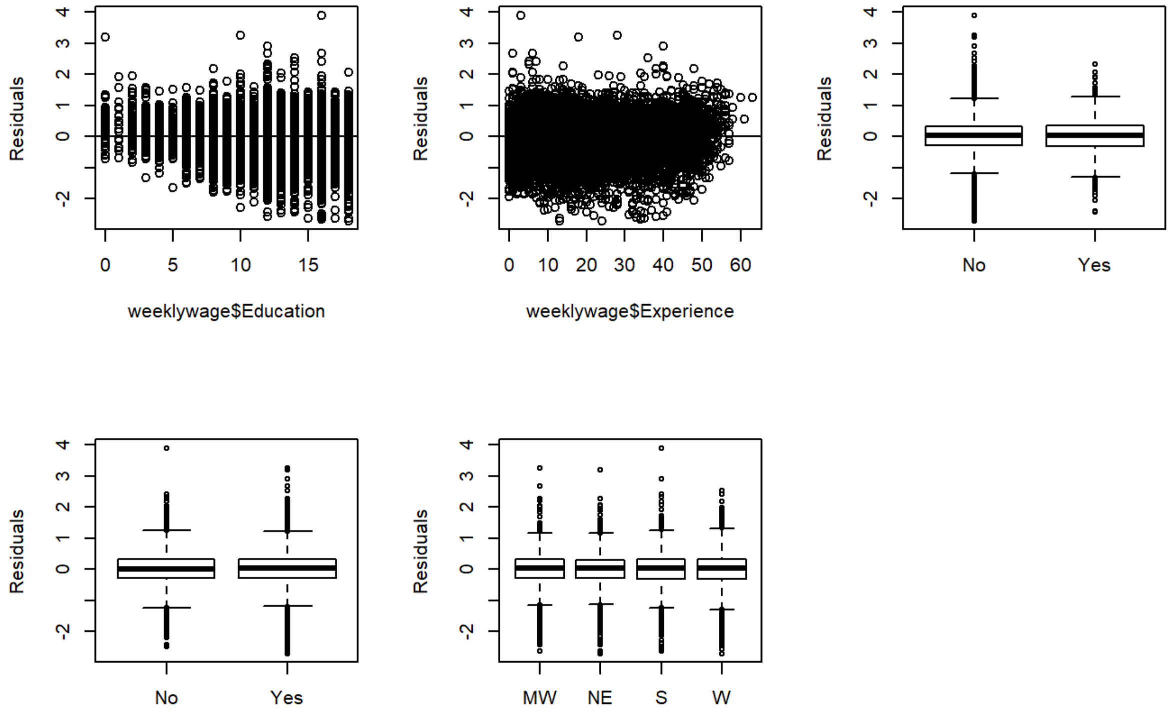
```
#confidence interval for coefficients
exp(confint(regweeklywage2))
```

```
##                          2.5 %     97.5 %
## (Intercept)         80.0737261 86.3903224
## Education            1.0901550  1.0951762
## Experience           1.0546146  1.0583886
## Experience2          0.9991264  0.9992030
## as.factor(Black)Yes  0.7717392  0.8095265
## as.factor(SMSA)Yes   1.1620638  1.1964225
## as.factor(Region)NE  1.0249053  1.0632702
## as.factor(Region)S   0.9243947  0.9566378
## as.factor(Region)W   0.9704471  1.0072985
```

R-square increases to 0.3307, which is better than the first model. P values for all predictors are significant except for W(region). In the final model, let's add the interaction effect in the model and decide to the extent to which blacks were paid differently than nonblacks may depend on region.

Model 3: log(Wage), Experience squared and interaction between Black and Region

```
regweeklywage3 <- lm(LogWage~Education + Experience + Experience2 + as.factor(Black)* as.factor(Region)+ a
s.factor(SMSA), data = weeklywage)
```



Still, there are inconsistent variance problems in first two scatter plots. Thus the constant variance assumption is not satisfied.

```
## 
## Call:
## lm(formula = LogWage ~ Education + Experience + Experience2 + 
##     as.factor(Black) * as.factor(Region) + as.factor(SMSA), data = weeklywage)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -2.7128 -0.2848  0.0346  0.3252  3.9043 
## 
## Coefficients:
##                                           Estimate Std. Error t value
## (Intercept)                              4.422e+00  1.941e-02 227.812
## Education                                8.859e-02  1.173e-03  75.540
## Experience                               5.497e-02  9.113e-04  60.319
## Experience2                             -8.358e-04  1.958e-05 -42.688
## as.factor(Black)Yes                     -2.430e-01  2.941e-02  -8.263
## as.factor(Region)NE                      4.199e-02  9.641e-03   4.356
## as.factor(Region)S                      -6.056e-02  9.120e-03  -6.641
## as.factor(Region)W                      -1.326e-02  9.709e-03  -1.366
## as.factor(SMSA)Yes                       1.644e-01  7.451e-03  22.068
## as.factor(Black)Yes:as.factor(Region)NE  1.922e-02  4.129e-02   0.465
## as.factor(Black)Yes:as.factor(Region)S  -1.250e-03  3.358e-02  -0.037
## as.factor(Black)Yes:as.factor(Region)W   5.449e-02  4.908e-02   1.110
##                                         Pr(>|t|)    
## (Intercept)                              < 2e-16 ***
## Education                                < 2e-16 ***
## Experience                               < 2e-16 ***
## Experience2                              < 2e-16 ***
## as.factor(Black)Yes                      < 2e-16 ***
## as.factor(Region)NE                     1.33e-05 ***
## as.factor(Region)S                      3.19e-11 ***
## as.factor(Region)W                         0.172    
## as.factor(SMSA)Yes                       < 2e-16 ***
## as.factor(Black)Yes:as.factor(Region)NE    0.642    
## as.factor(Black)Yes:as.factor(Region)S     0.970    
## as.factor(Black)Yes:as.factor(Region)W     0.267    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.5106 on 25425 degrees of freedom
## Multiple R-squared:  0.3307, Adjusted R-squared:  0.3305 
## F-statistic:  1142 on 11 and 25425 DF,  p-value: < 2.2e-16
```

R-square is 0.3307, which is exactly the same as r-square in the second model. P-values for the interaction terms are bigger than 0.5, indicating that we could actually rule out this effect. Let’s run a nested f test to verify my observation.

```
#nested f test
anova(regweeklywage3,regweeklywage2)
```

```
## Analysis of Variance Table
## 
## Model 1: LogWage ~ Education + Experience + Experience2 + as.factor(Black) * 
##     as.factor(Region) + as.factor(SMSA)
## Model 2: LogWage ~ Education + Experience + Experience2 + as.factor(Black) + 
##     as.factor(SMSA) + as.factor(Region)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1  25425 6629.4                           
## 2  25428 6629.9 -3  -0.50605 0.6469 0.5848
```

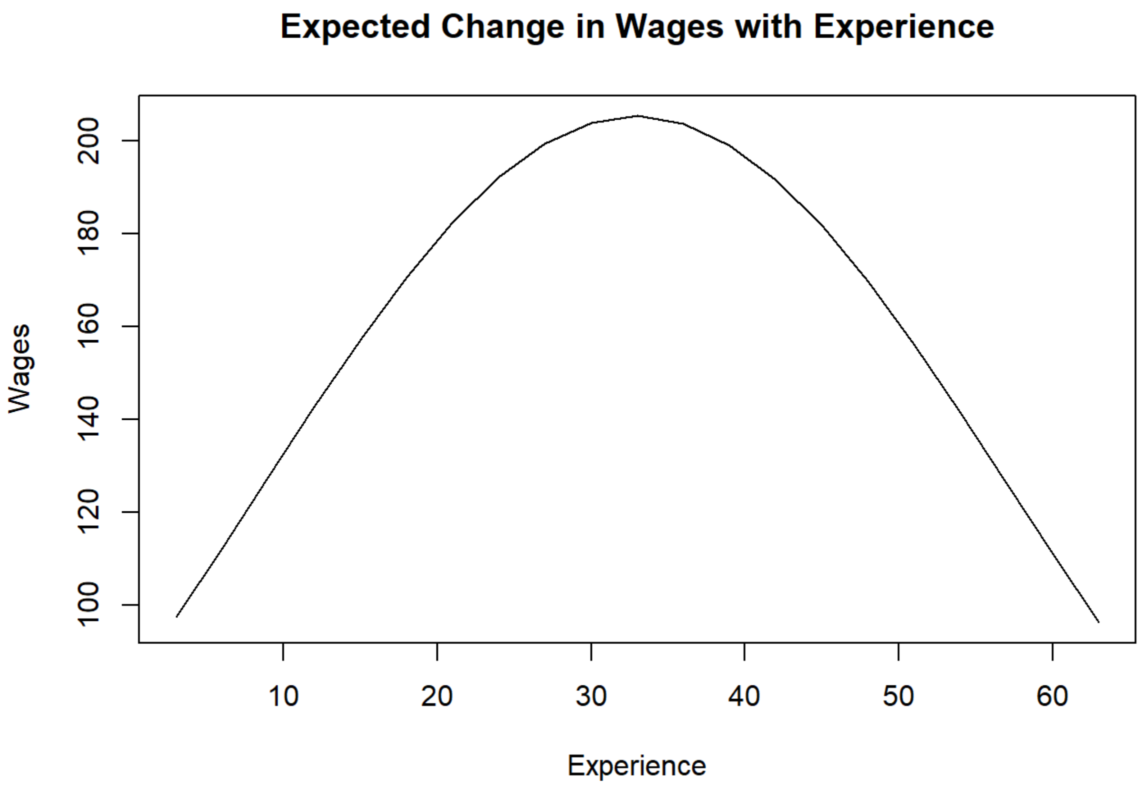
P-value for the nested f test is 0.5951 which is a lot bigger than 0.05. Thus we can say that the relationship between log(Wage) and being black is not affected by region. Therefore, we rule out the interaction effect.

Therefore, I choose the second model of all three models presented.

Interpretation

Education Slope estimation for Education is 8.862e-02 (95% CI: 0.0860298756 to 0.0906278757) in log scale. An increase in education of 1 year is associated with a multiplicative change of $e^{0.0862} \approx 1.09$ (CI:1.0898389, 1.094862) in median of wage.

Experience It's hard to interpret quadratic effect in words, thus I include a plot showing the relationship between Wage and Experience assuming nonblack males, having 0 year of education, living in non-city area of the middle-west of the country (all baseline conditions in the model).



Black Coefficient interpretation: Holding other variables constant (SMSA="No", Region="MW"), the median wage is expected to decrease $e^{-0.2352} \approx 0.7904$ if we change the "Black" indicator from nonblack to black. In other words, black males are actually paid less than nonblack males in the same region and with the same level of education and experience. Small p-value also verifies this.

Brief Conclusion:

The model I chose for Wage & Race data set is a multiple regression model with log(Wage) on Experience squared. I chose the log transformation of Wage and quadratic effects on Year in the innitial exploratory data analysis process. It turns out that interaction effect between "Black" and "region" could be negligible. R-square of the final model is 0.3307 which is not that high, but it's understandable that other important predictors such as occuptation are not included in the model. Limitation is that there are more relevant predictors existed, which could further boost r-square of the model. But for the pupose of comparing black and nonblack's wages, with the current data collected, we can still yield a conclusion. Also, since non-constant variance assumption is violated, we need to be cautious of using this model. In addition, quadratic effect for experience might imply that wages eventually will decrease and even be negative as experience further increases, which makes little sense.