IDS 702 - Homework #2

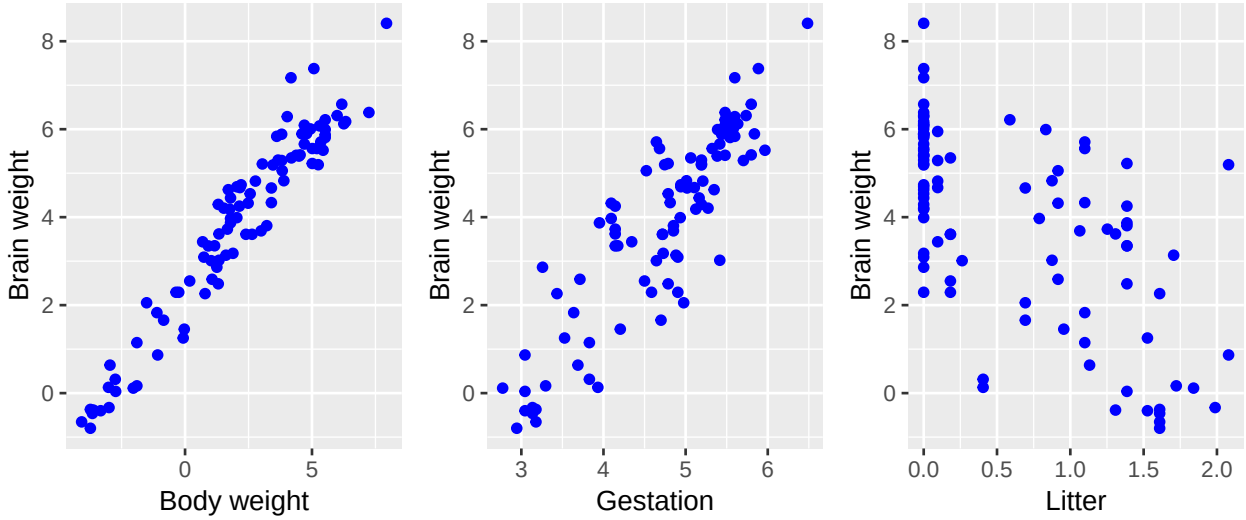
Ana Belen Barcenas J.

9/17/2018

1. Mammal Brain Weights

a) Mammal brain weight vs. covariates (log transformed):

```
grid.arrange(body, gestation, litter, ncol=3)
```



b) Fitting a multiple linear regression.

Before fitting a regression model, I will substract the mean from log(body weight) to obtain a more accurate interpretation of the intercept and I'll compute the log(brain weight):

```
brain_weight$log_BodyCent = log(brain_weight$Body) - mean(log(brain_weight$Body))

brain_weight$log_Brain = log(brain_weight$Brain)
```

Now I'll fit the regression model:

```
reg_bw <- lm(log_Brain ~ log_BodyCent + log(Gestation) + log(Litter), data=brain_weight)

summary(reg_bw)
```

```
##
## Call:
## lm(formula = log_Brain ~ log_BodyCent + log(Gestation) + log(Litter),
##     data = brain_weight)
##
## 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)     2.08001    0.71305   2.917  0.00444 **
```

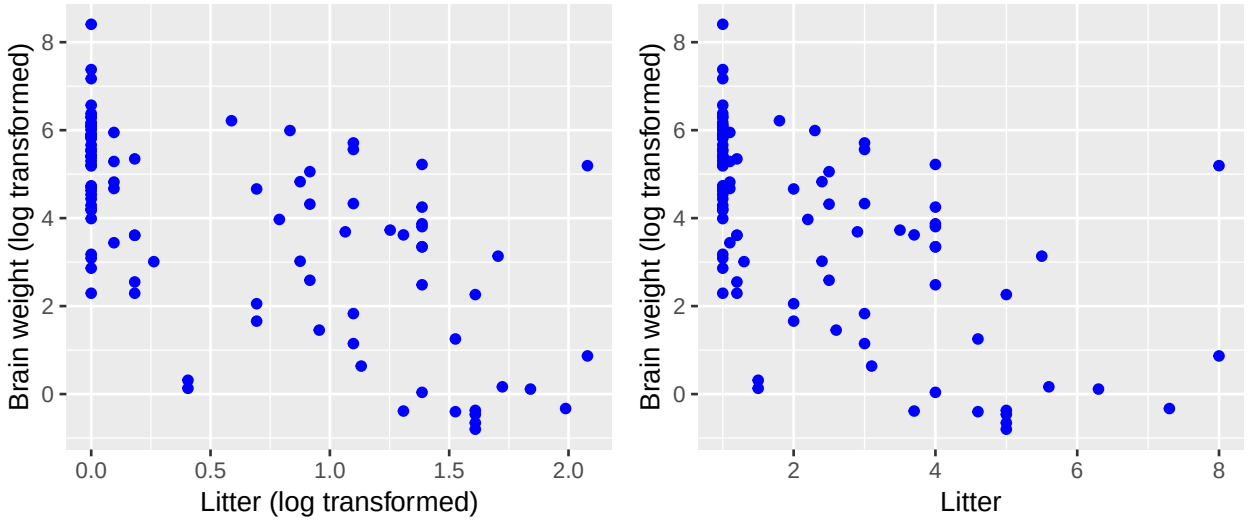
```
## log_BodyCent    0.57507    0.03259  17.647  < 2e-16 ***
## log(Gestation)  0.41794    0.14078   2.969  0.00381 **
## log(Litter)    -0.31007    0.11593  -2.675  0.00885 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4748 on 92 degrees 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
```

```
confint(reg_bw, level=0.95)
```

```
##                     2.5 %      97.5 %
## (Intercept)     0.6638274  3.49619182
## log_BodyCent    0.5103490  0.63979373
## log(Gestation)  0.1383359  0.69754827
## log(Litter)    -0.5403124 -0.07982996
```

c) Comparing log transformed litter vs. non-log transformed litter:

```
grid.arrange(litter_log, litter_nonlog, ncol=2)
```



The relationship seems to be better when litter is on its natural scale. On the second plot, the points are less spreaded than in the first (log transformed) plot, what suggests a clearer negative relationship. However, the results obtanied after fitting a model on each case are not being taken into account for this conclusion.

2. Mammal Brain Weights (additional problems)

d) Fitting a regression model without log transforming litter:

```
reg_bw2 <- lm(log_Brain ~ log_BodyCent + log(Gestation) + Litter, data=brain_weight)

summary(reg_bw2)
```

```
##
## Call:
## lm(formula = log_Brain ~ log_BodyCent + log(Gestation) + Litter,
##     data = brain_weight)
##
```

```
## 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)     2.04745    0.71355   2.869  0.00510 **
## log_BodyCent    0.57455    0.03264  17.601  < 2e-16 ***
## log(Gestation)  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
```

```
confint(reg_bw2, level=0.95)
```

```
##                    2.5 %      97.5 %
## (Intercept)     0.6302741  3.46462873
## log_BodyCent    0.5097143  0.63937813
## log(Gestation)  0.1675856  0.71169994
## Litter         -0.1943220 -0.02643223
```

e) Coefficients and CI interpretations:

– Intercept: The average brain weight is exp(2.0475) = 7.7481 grams when the mammal body weight is the average of the sample and the months of gestation are zero as well as the litter. From the CI we can say that we are 95% confident that the average brain weight falls between 1.88 and 31.94.

– log_BodyCent: For a 10% increase in the body weight (assuming that litter and gestation are zero), we expect about (1.10 ^ 0.5756) - 1 = 5.64% increase in the median mammal brain weight. From the CI we can say that we are 95% confident that the increase in median brain weight falls between 4.98% and 6.28% when there is a 10% increase in body weight.

– log(Gestation): For a 10% increase in gestation (assuming litter = 0 and the average body weight of the sample), we expect about (1.10 ^ 0.4396) - 1 = 4.28% increase in the median mammal brain weight. From the CI we can say that we are 95% confident that the increase in median brain weight falls between 1.61% and 7.02% when there is a 10% increase in gestation.

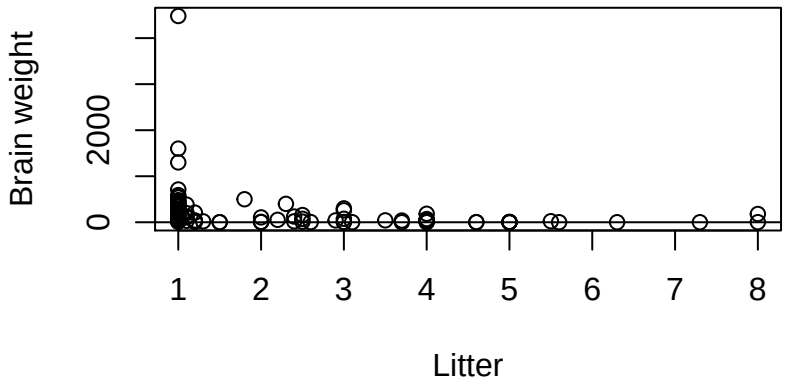
– Litter: An increase of one unit of litter (asumming that gestation = 0 and the body weight is the average of the sample) produces a decrease of exp(0.1104) - 1 = -11.67% in the median mammal brain weight. From the CI we can say that we are 95% confident that the decrease in median brain weight falls between -17.65% and -2.61% when litter increase by one unit.

– R-Squared: This model seems to be pretty good! We are capturing 95.35% of the variance.

Reference: https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faqhow-do-i-interpret-a-regression-model-when-some-varial

f) Let's check the variance of Litter:

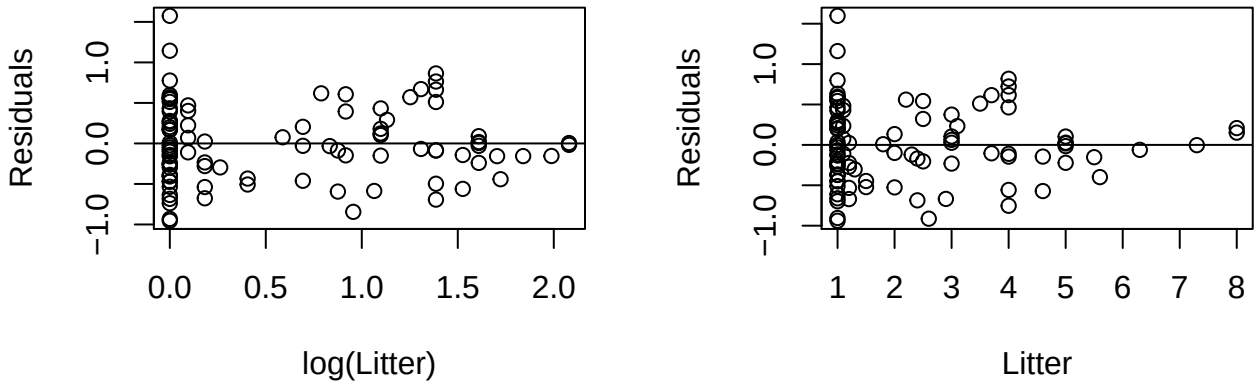
```
plot(y = brain_weight$Brain, x = brain_weight$Litter, xlab = "Litter", ylab = "Brain weight")
abline(0,0)
```



It seems that the constant variance assumption is violated because of the cases where Litter is small. At the greater values of Litter, the variance seems to be constant. Let's see the residual plots to check if those graphs give us a better clue.

Residuals plots log transforming Litter and without log transfomring litter:

```
par(mfrow=c(1,2))
plot(reg_bw$resid~log(brain_weight$Litter), xlab = "log(Litter)", ylab = "Residuals")
abline(0,0)
plot(reg_bw2$resid~brain_weight$Litter, xlab = "Litter", ylab = "Residuals")
abline(0,0)
```



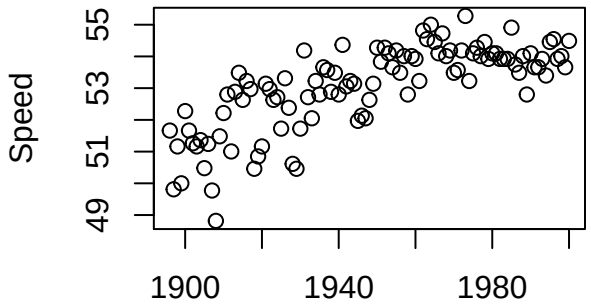
Based on the previous two residual plots I prefer the model where Litter is log transformed. Applying the transformation seems to slightly correct the non-constant variance problem. Moreover, the R squared is a little bit bigger in the regression where Litter is log transformed: 0.9537 vs. 0.9535.

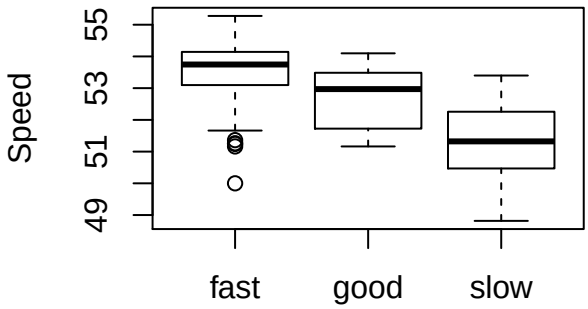
3. Winning speeds at the Kentucky Derby

Exploratory analysis:

Let's see the relation of the speed with the year and the track conditions as well as the relation between years and track condition:

```
par(mfrow=c(1,2))
plot(y = kentucky$Speed, x = kentucky$Year, xlab = "Year", ylab = "Speed")
boxplot(Speed~Condition, data = kentucky, ylab = "Speed", xlab = "Condition")
```





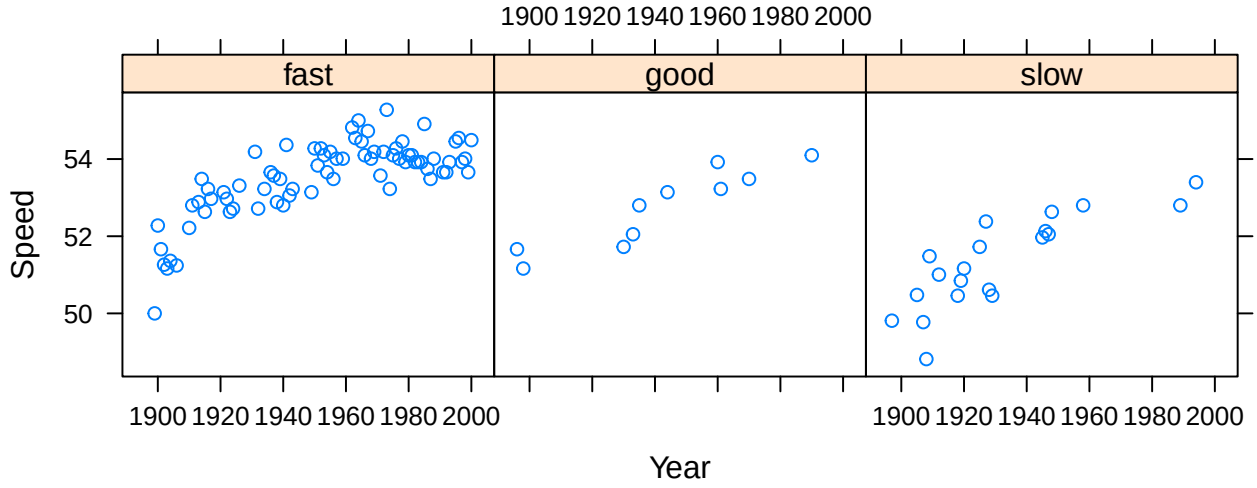
– Both graphs suggest that the constant variance assumption is not remarkably violated. Therefore, I will not log transform the speed to keep the interpretation as simple as possible.

– On the other hand, the first graph shows that there is a positive effect between speed and year but the relations seems to be non-linear. I will try log transformation as well as quadratic transformation of the year variable to adjust that curve better.

– From the second plot we can identify a negative relation between speed and track condition, what suggests that the condition variable will be important to predict accurately the speed.

Now, let's look for interaction effects between years and track condition:

```
library(lattice)
xyplot(Speed~Year | Condition, data = kentucky)
```



– The slope on each graph seems to be pretty similar. Which is to say that there is not an interaction between the track condition and the year variables. Therefore, is not necessary use interaction terms in the regression model.

Now that we have an good idea about the relation between the variables, let's create the year squared and a series of dummy variables for each speed. Also, I will transform the year variable assigning a zero to the first year which is 1896 to obtain an interpretation of the intercept that makes sense:

```
n = nrow(kentucky)
kentucky$fast = rep(0, n)
kentucky$fast[kentucky$Condition == "fast"] = 1
kentucky$good = rep(0, n)
kentucky$good[kentucky$Condition == "good"] = 1
kentucky$slow = rep(0, n)
```

```r
kentucky$slow[kentucky$Condition == "slow"] = 1

min_year <- min(kentucky$Year)
kentucky$Year_t <- kentucky$Year - min_year

kentucky$Year_t2 = kentucky$Year_t^2
```

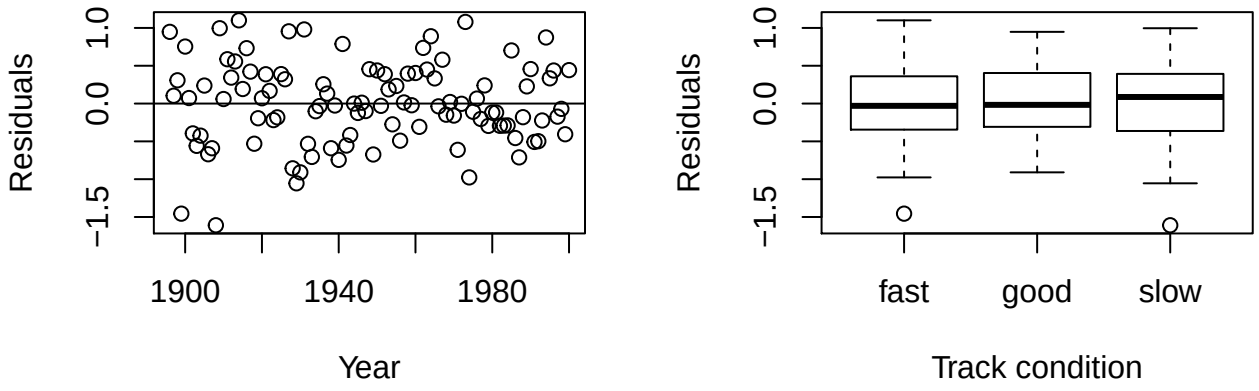
Let's fit the multiple regresion with the transformations proposed and see the results. The baseline will be the track condition "slow":

```r
reg_ken = lm(Speed~Year_t + Year_t2 + fast + good, data = kentucky)
summary(reg_ken)
```

```
##
## Call:
## lm(formula = Speed ~ Year_t + Year_t2 + fast + good, data = kentucky)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.60905 -0.30796 -0.02224  0.38851  1.10047
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.964e+01  1.808e-01 274.520  < 2e-16 ***
## Year_t       7.075e-02  7.033e-03  10.060  < 2e-16 ***
## Year_t2     -4.214e-04  6.526e-05  -6.457 3.89e-09 ***
## fast         1.610e+00  1.439e-01  11.189  < 2e-16 ***
## good         1.078e+00  2.136e-01   5.047 2.02e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5492 on 100 degrees of freedom
## Multiple R-squared:  0.8365, Adjusted R-squared:  0.8299
## F-statistic: 127.9 on 4 and 100 DF,  p-value: < 2.2e-16
```

The R-squared seems to be great, we are explaining the 83.65% of the variance. Let's check the residual plots to detemrine if the model meets the required assumptions:

```r
par(mfrow=c(1,2))
plot(reg_ken$resid, x=kentucky$Year, ylab = "Residuals", xlab = "Year")
abline(0,0)
boxplot(reg_ken$resid~kentucky$Condition, ylab = "Residuals", xlab = "Track condition")
```



No patterns appear in these graphs. Moreover, both the scatterplot and the box plots seems to be equally distributed along the x-axis. In other words, the linearity, constant variance, and normality assumptions seems to be met. It's time to interpret the coefficients and its confidence intervals.

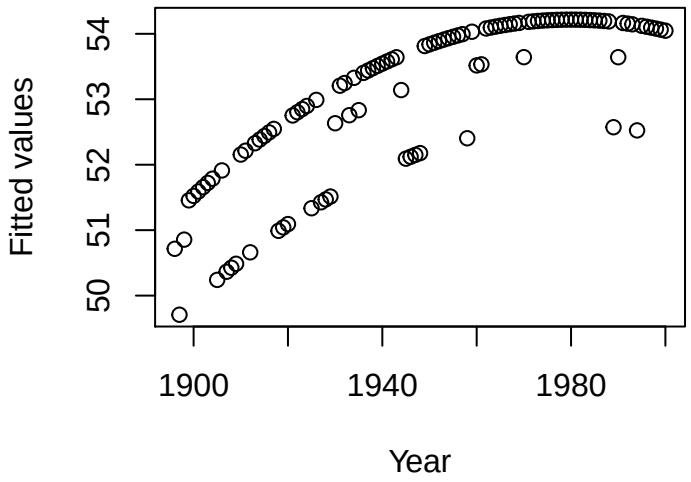
```
confint(reg_ken)
```

```
##                      2.5 %        97.5 %
## (Intercept) 49.2785542956 49.9960163523
## Year_t       0.0567983709  0.0847042477
## Year_t2     -0.0005508393 -0.0002918965
## fast         1.3244040523  1.8953027706
## good         0.6541672128  1.5016797590
```

Interpretation:

To interpret the quadratic term year, I will refer to a plot that shows how does the relation between speed and years looks like:

```
kentucky$fitted.values <- reg_ken$fitted.values
plot(y=kentucky$fitted.value, x = kentucky$Year, xlab = "Year", ylab = "Fitted values")
```



– Year and Year^2: From the previous graph is clear that there is a positive relation between the average speed and the years: every year the speed seems to increase. Nevertheless, the increment is not linear. Each year the speed increment is less than the previous year.

– Intercept: Assuming we are situated in the first year when the data was collected (1896), the average speed in a slow track condition (the baseline) is 49.64 feet per second. We are 95% confident that the average speed lies between 49.28 and 49.99.

– Fast: If the track condition switch from slow to fast, the speed will increase 1.61 feet per second. We are 95% confident that the average speed will increase somewhere between 1.32 and 1.90.

– Good: If the track condition switch from slow to good, the speed will increase 1.08 feet per second. We are 95% confident that the average speed will increase somewhere between 0.65 and 1.50.

– R-squared: Fitting this model we are capturing 83.65% of the variance, which is great!

Conclusions and limitations

The model seems to predict the speed of the horses pretty well. I would not say that the model has limitations since we are capturing the 83.65% of the variance. On the other hand, the model is simple which is helpful to interpret the results to non-technical forums and it is also a parsimonious model.

4. Old Faithful

Let's compute a nested F test to check if the whole set of date dummy variables are significant:

```
lm_oldf <- lm(Interval ~ Duration, data = old_faithful)
lm_oldf_fact <- lm(Interval ~ Duration + as.factor(Date), data = old_faithful)

anova(lm_oldf_fact, lm_oldf)
```

```
## 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
```

The p-value in the nested F test is too large. This suggests that there is not enough information to accept the null hypothesis. In other words, the Date variable is not significant to predict the Duration of the eruption. Given the previous evidence and for parsimonious reasons, I would select the model that excludes the Date variable to predict the Interval between each eruption.

5. Wages and Race

Exploratory analysis:

Since we have around 25 thousands observations, it's easy to have an idea of how does the data behaves by means of a summary:

```
summary(wage_race)
```

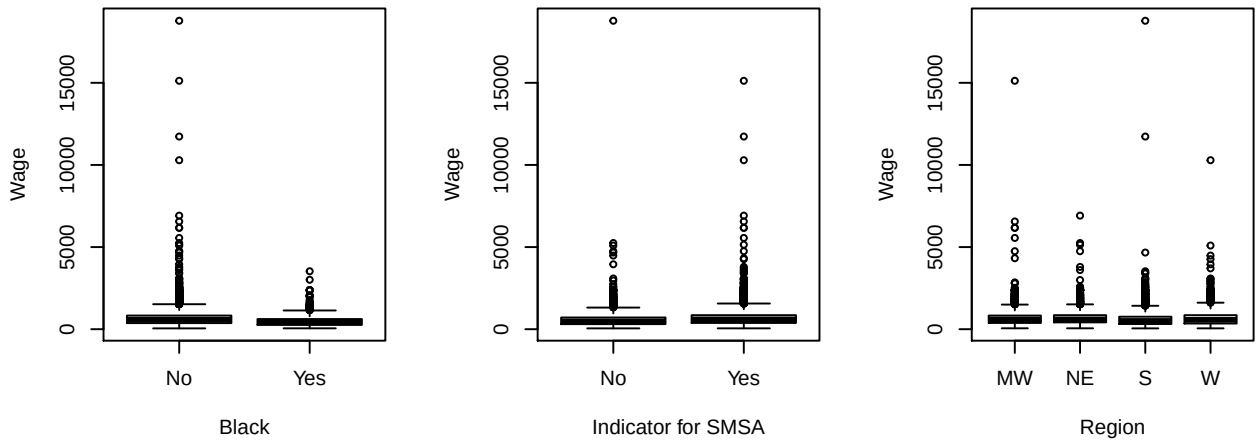
```
##        X              Wage            Education       Experience
##  Min.   :    1   Min.   :   50.39   Min.   : 0.00   Min.   :-4.00
##  1st Qu.: 6408   1st Qu.:  356.13   1st Qu.:12.00   1st Qu.: 9.00
##  Median :12816   Median :  567.23   Median :12.00   Median :16.00
##  Mean   :12816   Mean   :  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
##  No :23643   No : 6591   MW:6226
##  Yes: 1988   Yes:19040   NE:5949
##                          S :7991
##                          W :5465
##
##
```

From this summary we can observe that there are negative values for experience, which makes no sense. The best way to address this problem is to remove those rows since we do not have the opportunity to talk with the people whom collected the data and understand the nature of those values.

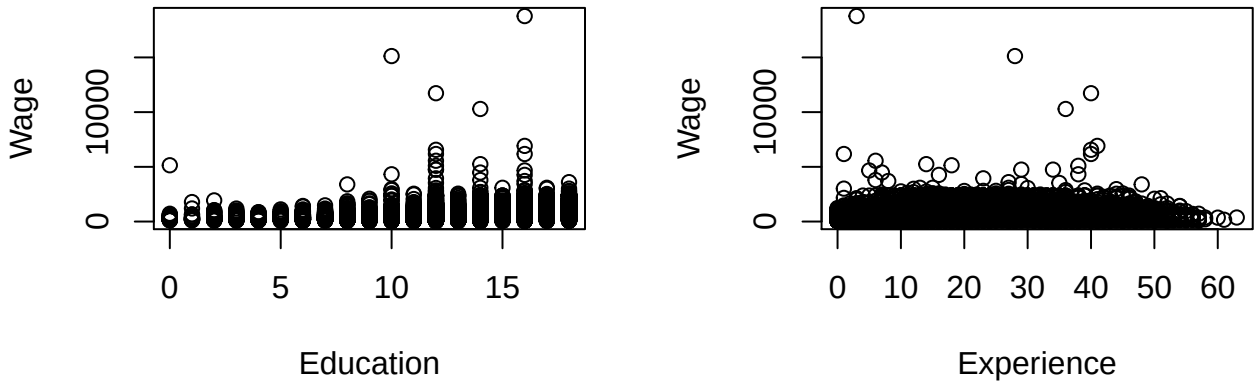
```
wage_race2 <- wage_race[wage_race$Experience >= 0,]
```

Let's see the relation between the speed and the covariates (black, indicator of SMSA, region, education, and experience) to have an idea of any required transformation and/or if the assumptions are met.

```
par(mfrow=c(1,3))
boxplot(Wage~Black, data = wage_race2, ylab = "Wage", xlab = "Black")
boxplot(Wage~SMSA, data = wage_race2, ylab = "Wage", xlab = "Indicator for SMSA")
boxplot(Wage~Region, data = wage_race2, ylab = "Wage", xlab = "Region")
```




```
par(mfrow=c(1,2))
plot(y = wage_race2$Wage, x = wage_race2$Education, xlab = "Education", ylab = "Wage")
plot(y = wage_race2$Wage, x = wage_race2$Experience, xlab = "Experience", ylab = "Wage")
```



– The relation in which we have special interest is between black males and the wages. The first graph shows that relation. From more than one graph in the chart above, we can assert that the constant variance assumption is violated (especially in the 1st, 4th and 5th plots). To solve this problem, the most common approach is to log transform the dependent variables (wage) before fitting the regression.

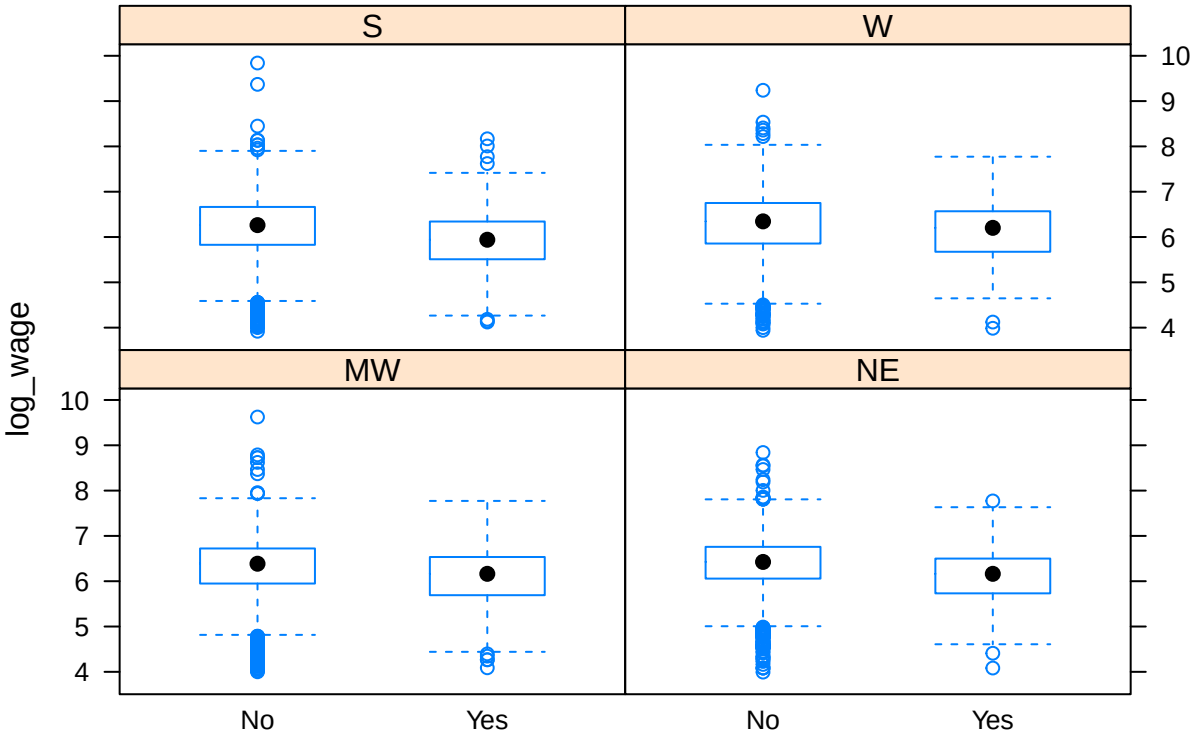
– On the other hand, the last two graphs suggest that there is a non-linear relation between wage and education as well as wage and experience. I will explore different regression models transforming education and experience using log and quadratic transformations.

– Lastly, each graph have a different relation with wages. Suggesting that each one explain something about the wage. For this reason, the five variables will be included in the regression model.

To assess if there is an interactive effect between skin color and region we can examine a box plot of each

category:

```
wage_race2$log_wage = log(wage_race2$Wage)
bwplot(log_wage~Black | Region, data = wage_race2)
```



– From the previous chart I would say that the skin color boxplots have the same order for each region: the average log(wage) is clearly greater for non-black people in every region. Therefore, there is no evidence of interactions between skin color and region. For the previous evidence, I will exclude such interaction in the model.

The suggested model based on the exploratory data analysis suggests the following:

1) Use the log transformed wage.
2) Refrain from interactive effects between skin color and region.
3) Try fitting quadratic transformations for education and experience. Analyze R-square and residual plots to choose the best model. Log transformations are not possible since education=0 and experience=0 are plausible and log(0) is undefined.

Let's compute the dummy variables and the quadratic transformations:

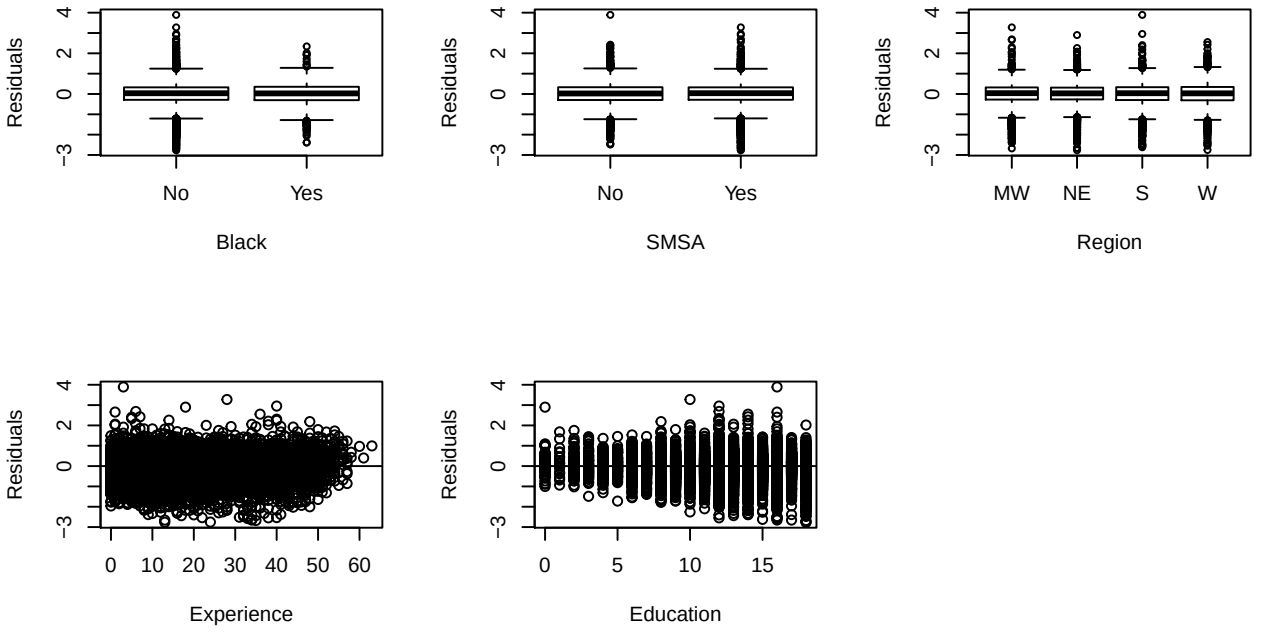
```
n = nrow(wage_race2)
wage_race2$NE = rep(0, n)
wage_race2$NE[wage_race2$Region == "NE"] = 1
wage_race2$MW = rep(0, n)
wage_race2$MW[wage_race2$Region == "MW"] = 1
wage_race2$S = rep(0, n)
wage_race2$S[wage_race2$Region == "S"] = 1
wage_race2$W = rep(0, n)
wage_race2$W[wage_race2$Region == "W"] = 1

wage_race2$education2 = wage_race2$Education^2
wage_race2$experience2 = wage_race2$Experience^2
```

Fitting the possible models:

```
reg_wage = lm(log_wage~Black + Experience + experience2 + Education + education2 + SMSA + NE + MW + S, d

reg_wage2 = lm(log_wage~Black + Experience + experience2 + Education + SMSA + NE + MW + S, data = wage_

reg_wage3 = lm(log_wage~Black + Experience + Education + SMSA + NE + MW + S, data = wage_race2)

reg_wage4 = lm(log_wage~Black + Experience  + Education + education2 + SMSA + NE + MW + S, data = wage_
```

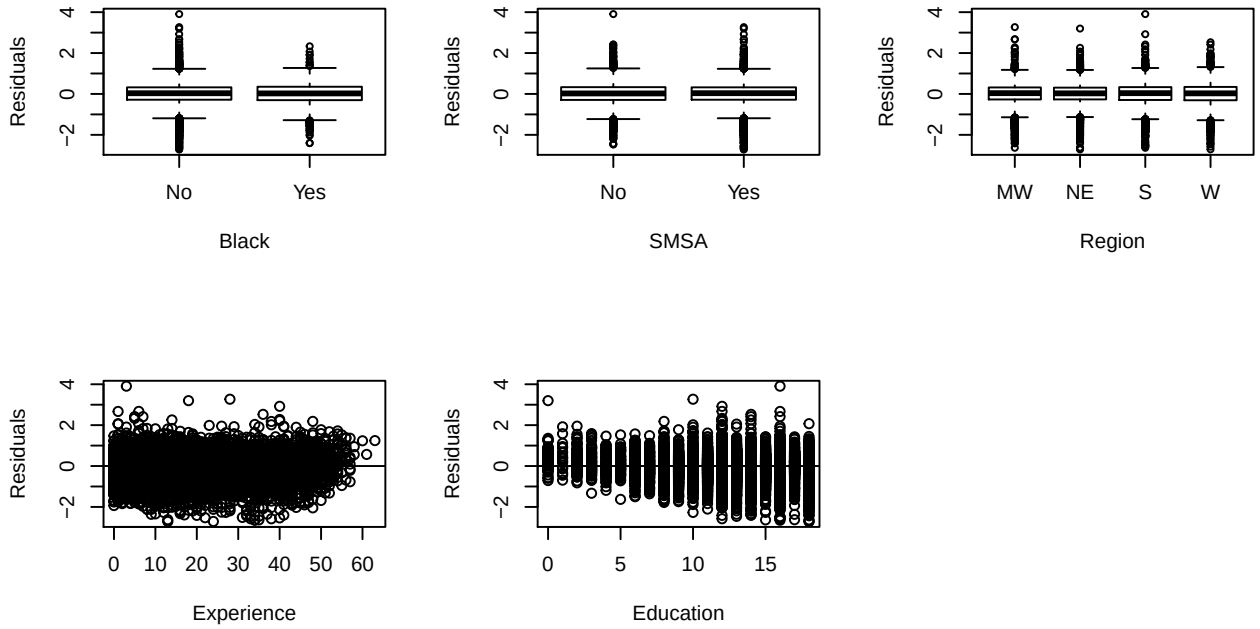
The first and the second one seems to explain more variance (bigger R-squared) than the last two options. Both models has in common the quadratic term of experience. Actually, within the labor economics field is well known that the experience has a positive but decreasing effect on wages. Thus, besides the evidence that the R-squared provides, there is an intuitive justification for including the quadratic term of experience in the final model. To choose between the first two, I will view the residual plots of each model.

```
par(mfrow=c(2,3))

boxplot(reg_wage$resid~wage_race2$Black, ylab = "Residuals", xlab = "Black")
boxplot(reg_wage$resid~wage_race2$SMSA, ylab = "Residuals", xlab = "SMSA")
boxplot(reg_wage$resid~wage_race2$Region, ylab = "Residuals", xlab = "Region")

plot(reg_wage$resid, x=wage_race2$Experience, ylab = "Residuals", xlab = "Experience")
abline(0,0)
plot(reg_wage$resid, x=wage_race2$Education, ylab = "Residuals", xlab = "Education")
abline(0,0)
```




```
par(mfrow=c(2,3))

boxplot(reg_wage2$resid~wage_race2$Black, ylab = "Residuals", xlab = "Black")
boxplot(reg_wage2$resid~wage_race2$SMSA, ylab = "Residuals", xlab = "SMSA")
boxplot(reg_wage2$resid~wage_race2$Region, ylab = "Residuals", xlab = "Region")

plot(reg_wage2$resid, x=wage_race2$Experience, ylab = "Residuals", xlab = "Experience")
abline(0,0)
```

```r
plot(reg_wage2$resid, x=wage_race2$Education, ylab = "Residuals", xlab = "Education")
abline(0,0)
```



The assumptions seems to be met in both models. The constant variance assumption could be weak in the case of education, but including education squared does not solve this problem. Thus, for parsimonious reasons and an easier interpretation, I will not include the quadratic education variable in the final model.

Final model results:

```r
summary(reg_wage2)
```

```
##
## Call:
## lm(formula = log_wage ~ Black + Experience + experience2 + Education +
##     SMSA + NE + MW + S, data = wage_race2)
##
## 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)  4.410e+00  1.937e-02 227.638  < 2e-16 ***
## BlackYes    -2.352e-01  1.219e-02 -19.288  < 2e-16 ***
## Experience   5.496e-02  9.112e-04  60.315  < 2e-16 ***
## experience2 -8.356e-04  1.958e-05 -42.681  < 2e-16 ***
## Education    8.862e-02  1.172e-03  75.597  < 2e-16 ***
## SMSAYes      1.648e-01  7.433e-03  22.167  < 2e-16 ***
## NE           5.434e-02  9.667e-03   5.621 1.92e-08 ***
## MW           1.136e-02  9.507e-03   1.195    0.232
## S           -5.011e-02  9.102e-03  -5.505 3.72e-08 ***
## ---
## 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
```

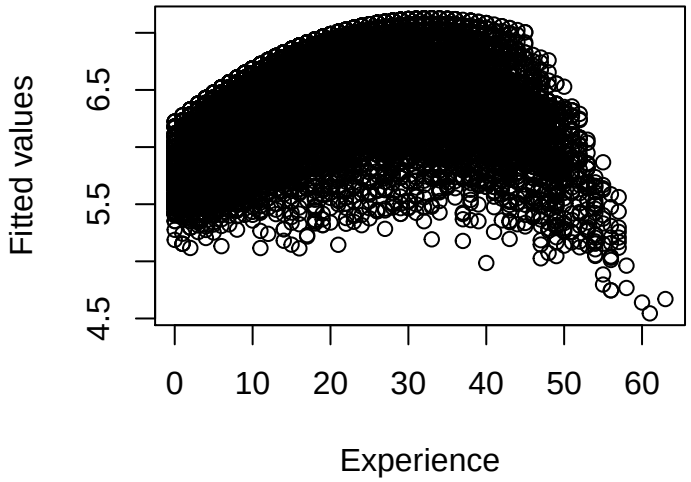
```
confint(reg_wage2)
```

```
##                     2.5 %      97.5 %
## (Intercept)  4.3715805260  4.44751645
## BlackYes    -0.2591086390 -0.21130579
## Experience   0.0531753840  0.05674756
## experience2 -0.0008740206 -0.00079727
## Education    0.0863199092  0.09091523
## SMSAYes      0.1501975514  0.17933585
## NE           0.0353904532  0.07328549
## MW          -0.0072719458  0.02999842
## S           -0.0679509882 -0.03226913
```

Interpretation:

To interpret the quadratic term experience, I will refer to a plot that shows how does the relation between wage and experience looks like:

```
wage_race2$fitted.values <- reg_wage2$fitted.values
plot(y=wage_race2$fitted.value, x = wage_race2$Experience, xlab = "Experience", ylab = "Fitted values")
```



– Experience and experienceˆ2: The above graph shows how years of experience determine higher wages until the person reaches an amount of approximately 30 years. This decreasing relation after 30 years is also associated with the age of the person. After this threshold, the wages decrease. This could be associated with how productive is the employee when is getting older.

– Intercept: The average weekly wage of white males with zero years of education and experience that has worked in a standard metropolitan statistical area (SMSA) located in the west (baseline), is exp(4.41) = 82.27. We are 95% confident that the average weekly wage when the person has the characteristics mentioned above falls between 79.16 and 85.41.

– Skin color (BlackYes): Being black (with zero years of education and experience that has worked in a standard metropolitan statistical area (SMSA) located in the west (baseline)) has a median weekly wage exp(-0.2352) - 1 = -20.96% smaller than a white men. We are 95% confident that the decrease in median weekly wage falls between -22.83% and -19.05% when the person is black.

– Education: An additional year of education (of a white male with zero years of experience that has worked in a standard metropolitan statistical area (SMSA) located in the west (baseline)) yields to an increase of exp(0.0886) -1 = 9.26% in the median weekly wage. We are 95% confident that the increment in median weekly wage when the person has the characteristics mentioned above falls between 9.01% and 9.52%.

– SMSA: Assuming a white male with zero years of education and experience that has worked in the west (baseline) receives a median weekly wage exp(0.1648) - 1 = 17.91% higher than a person that has not worked in a standard metropolitan statistical area. We are 95% confident that the increment in median weekly wage falls between 16.20% and 19.64% when the person is black.

– Region: Assuming a white male with zero years of education and experience that has worked in a standard metropolitan statistical area (SMSA) receives a median weekly wage exp(0.0543) - 1 = 5.58% higher if its work is located in the northeast instead of in the west (CI: 3.59% - 7.59%); exp(0.0113) - 1 = 1.14% higher if its work is located in the midwest instead of in the west (CI: -0.71% - 3.03%); and exp(-0.0501) - 1 = -4.89% smaller if its work is located in the south instead of in the west (CI: -6.5% - -3.17%).

– R-squares: With this model we are being able to explain 33.07% of the variance, which is not bad for a human related analysis.

Conclusions:

According to the final model, black males are being paid less than white males even in the same geographical areas and with the same education and work experience. The magnitude is around 21% higher wages- The data suggests that the labor market analyzed discriminate between white and black employees.

Limitations of the model:

Despite the good fit of the model, there are more variables that can be really useful to obtain more accurate results. For example, keeping the IQ of each person fixed, we can determine if the difference in wages is explained by the innate ability of the person instead or if there are discriminatory reasons. There are other examples of variables that could help us to fit a better model such as:

- Innate ability.
- Quality of education.
- Non-schooling investments.
- Family/friends networks (social mobility barriers).

Reference: https://economics.mit.edu/files/4689

Finally, the model does not address perfectly the constant variance assumption for the education variable. Including more observations could be helpful.