Multiple linear regression

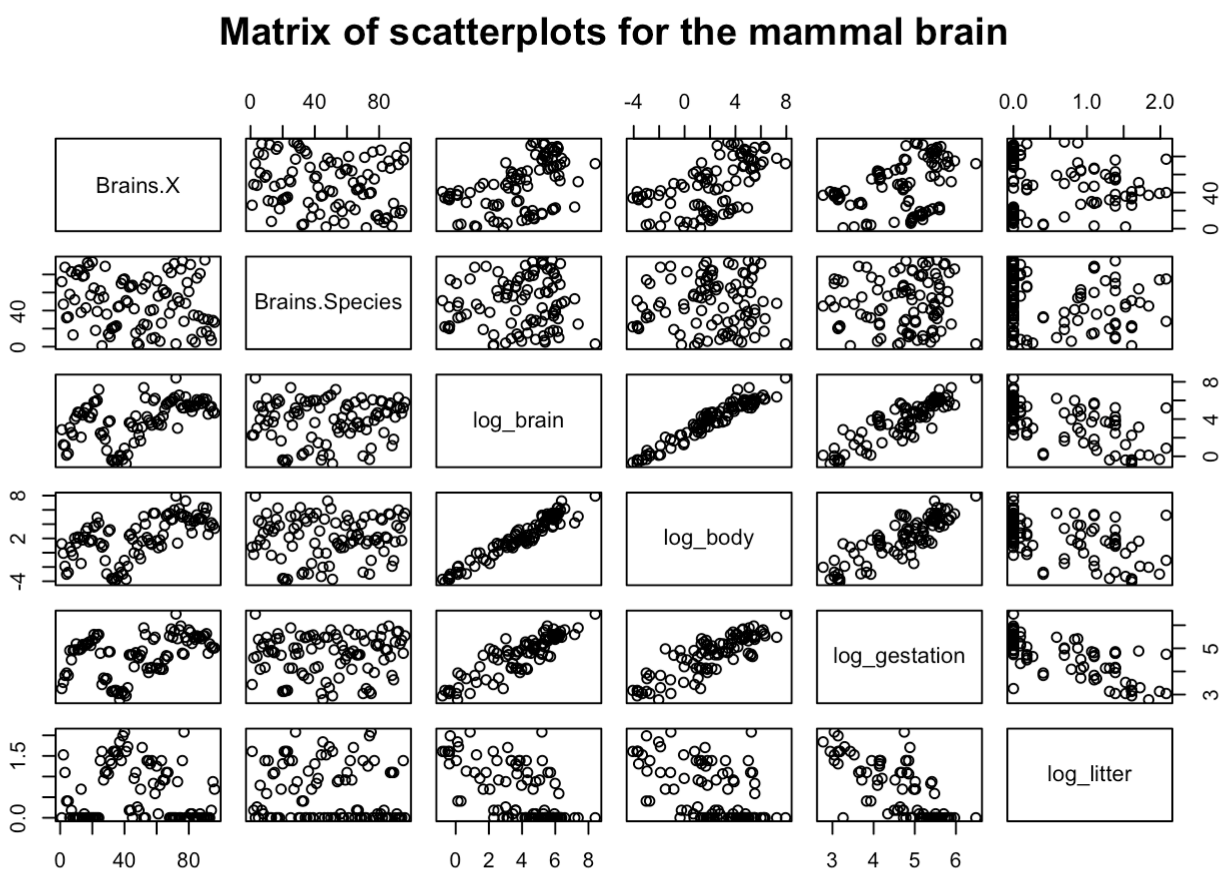
Oblasova Iuliia

9/14/2018

1. Mammal Brain Weights.

a). Draw a matrix of scatterplots for the mammal brain weight data with all variableds transformed to their logarifms.

```
Brains <- read.csv("~/Downloads/Ex0912.csv")
log_brain=log(Brains$Brain)
log_body=log(Brains$Body)
log_gestation=log(Brains$Gestation)
log_litter=log(Brains$Litter)
logBrains<-data.frame(Brains$X, Brains$Species, log_brain,log_body,log_gestation,log_lit
ter)
plot(logBrains, main="Matrix of scatterplots for the mammal brain")
```



We can observe strong linear association between log_brain~log_body variables and log_brain~log_gestation and suspect a negative assosiation between log_body and log_litter.

b. Fit the multiple linear regression of the log brain on the log body weight, log gestation and log litter size.

```
brain_size_all_logs = lm(log_brain ~ log_body + log_gestation + log_litter, data = logBr
ains)
summary(brain_size_all_logs)
```

```
##
## Call:
## lm(formula = log_brain ~ log_body + log_gestation + log_litter,
##     data = logBrains)
##
## 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
## log_body       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
```

Large R-squared indicates that model fits well.

Confidence interval for coefficients:

```
exp(confint(brain_size_all_logs))
```

```
##                   2.5 %    97.5 %
## (Intercept)   0.6317151 8.7491850
## log_body      1.6658725 1.8960897
## log_gestation 1.1483612 2.0088216
## log_litter    0.5825662 0.9232733
```

c. Does the relation between log brain weight and litter size appeared to be any better than the relationship between log brain weign and log litter size?

Slightly better pattern of a staight line is observed from the scatter plot of log brain and litter, therefore this relationship might be stronger than the relationship between log brain and log litter.

d. Fit the regression model with log(brain size) on log(body weight), log(gestation) and litter size on its natural scale. Report the output of the model (include a table with coefficients and SEs, their associated confidence intervals, and somewhere in the text or table the estimated regression standard deviation and R2).

```
brain_size = lm(log_brain ~ log_body + log_gestation + Brains$Litter, data = logBrains)
summary(brain_size)
```

```
##
## Call:
## lm(formula = log_brain ~ log_body + log_gestation + Brains$Litter,
##     data = logBrains)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93895 -0.27922 -0.00929  0.28646  1.59743
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.82338    0.66206   1.244  0.21678
## log_body       0.57455    0.03264  17.601  < 2e-16 ***
## log_gestation  0.43964    0.13698   3.210  0.00183 **
## Brains$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(brain_size)
```

```
##                    2.5 %      97.5 %
## (Intercept)   -0.4915254  2.13829063
## log_body       0.5097143  0.63937813
## log_gestation  0.1675856  0.71169994
## Brains$Litter -0.1943220 -0.02643223
```

e. Provide an interpretation of each of the coefficients from the regression in Part D on the natural scale of brain weight and each predictor. Include 95% confidence intervals in your interpretations.

We took the log of both respond and predictor variable. All calculations below based on the following calculations:

Avg log(y_cx) = B_o + B_1*log(cX) - Avg log(y_x) = B_o + B_1*log(X)

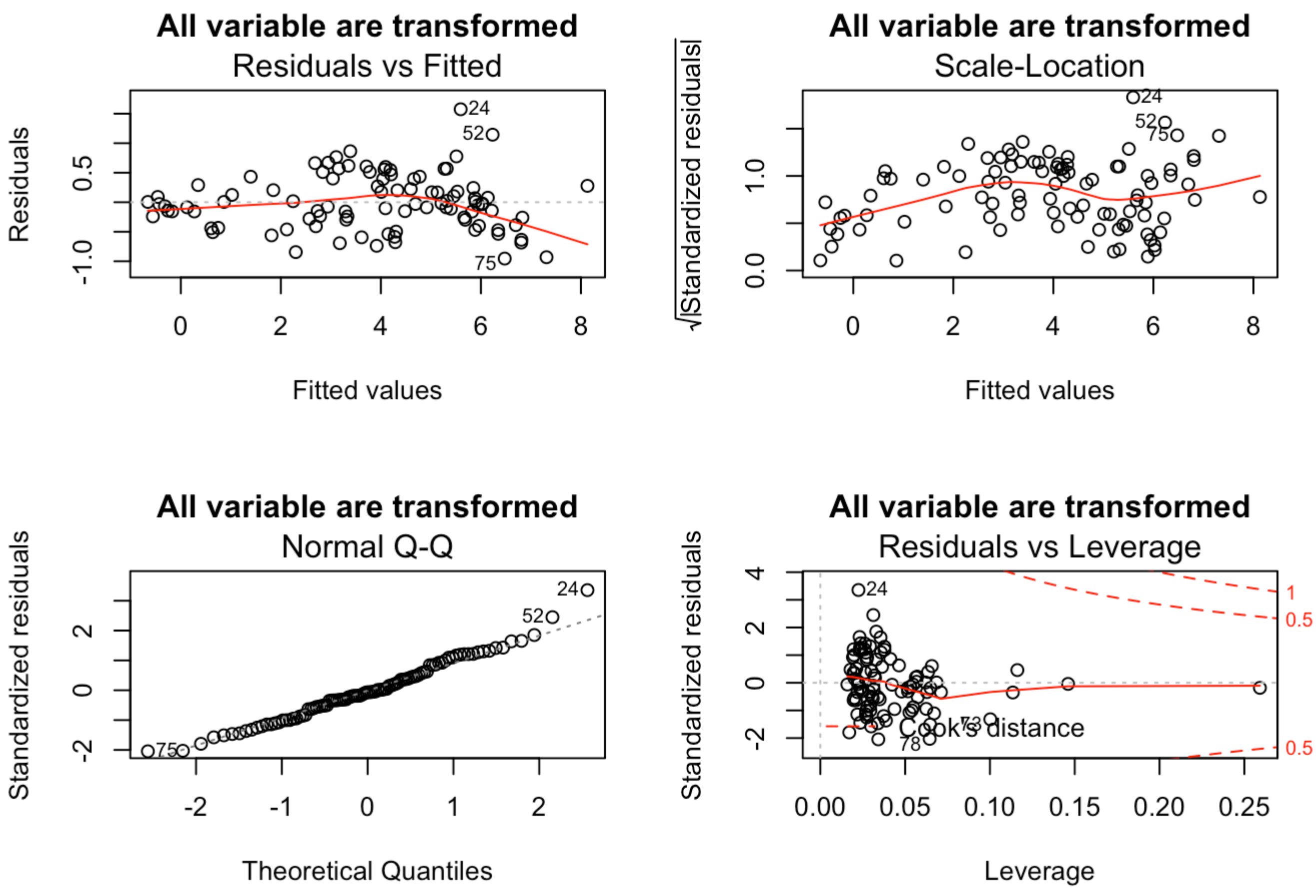
= B_1(log c)

To calculate it on the natural scale, expontatiate it: exp(B_1(log c)) = c^B_1. Therefore,

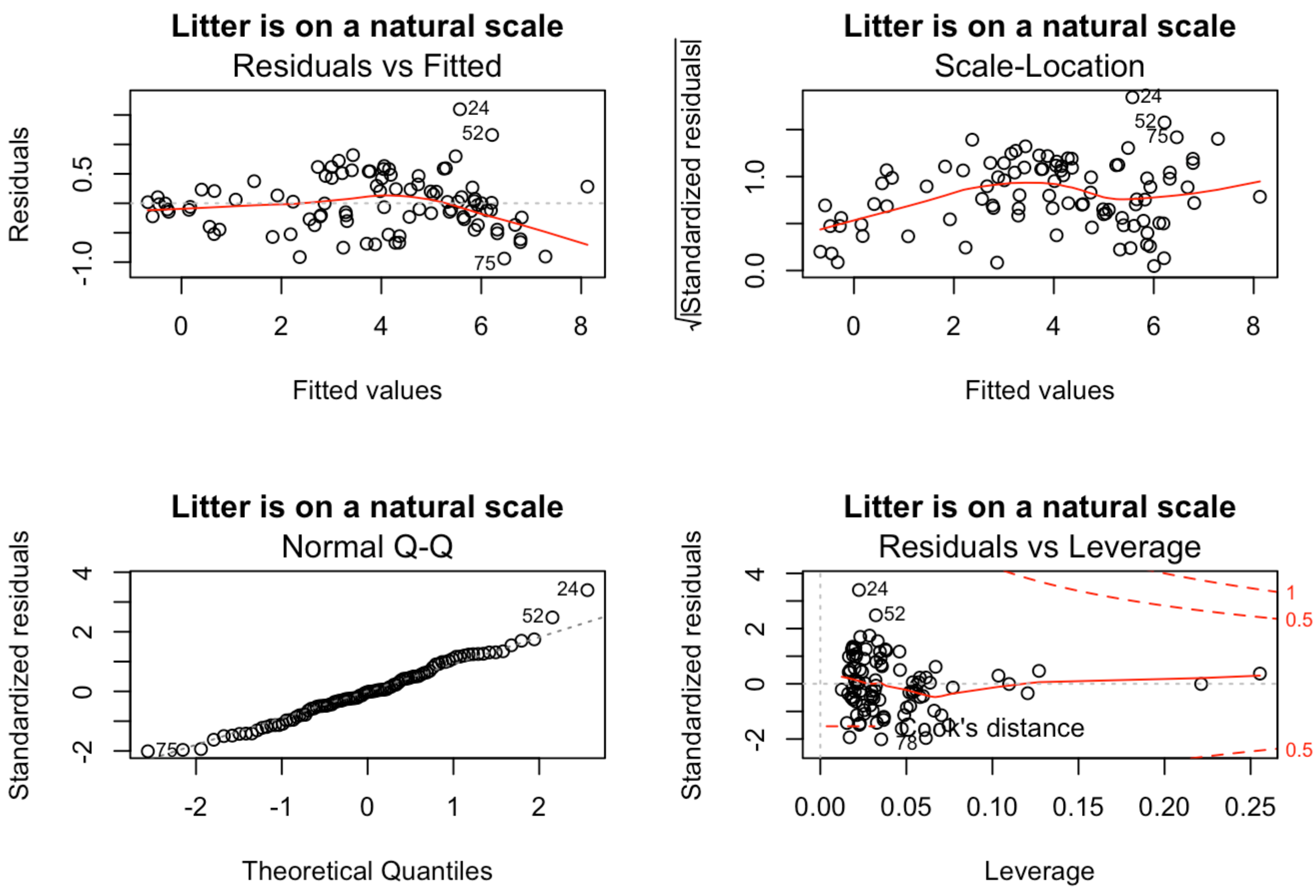
- For increase of body mass for 10%, the increase of median value of brain mass is expected to be 1.1^0.575 ≈ 5% with 95% confidence interval (4.98%, 6.28%).
- For increase of body mass for 10%, the increase of median value ofgestation is expected to be 1.1^0.4396 ≈ 4.3% with 95% confidence interval (1.6%, 7%)
- For increase of body mass for 10%, the decrease in meadian value of litter is expected to be 0.11 ≈ 1% with 95% confidence interval (1.94, 2.64).

f. Based on the quality of the residual plots and the value of R2, which model do you prefer: the one in Part B or the one in Part D?

```
par(mfcol=c(2,2))
plot(brain_size_all_logs, main = "All variable are transformed")
```



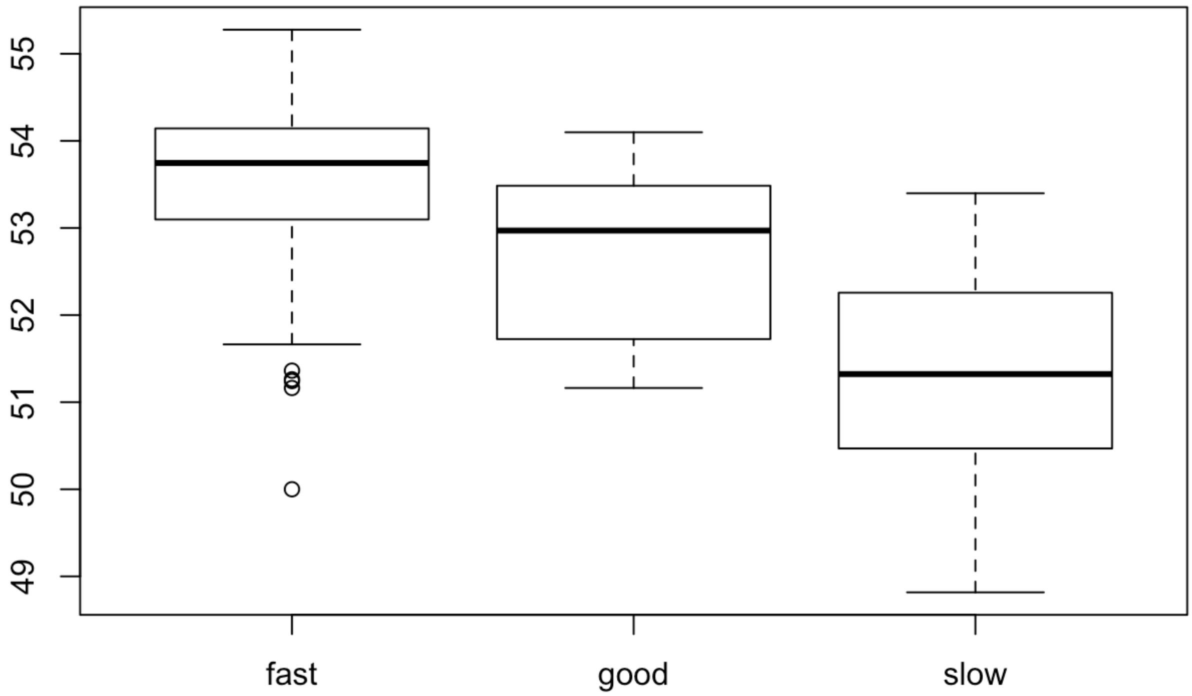

```
par(mfcol=c(2,2))
plot(brain_size, main = "Litter is on a natural scale")
```



Based on plots above, there is almost no difference in two models and none of the models fits well enough. The R-squared for the model considering log litter is higher by 0.0002, which is not a significant value; however, in terms of interpretation and building the model it is simplier to work with data when all values have the same logarifmic transformation.

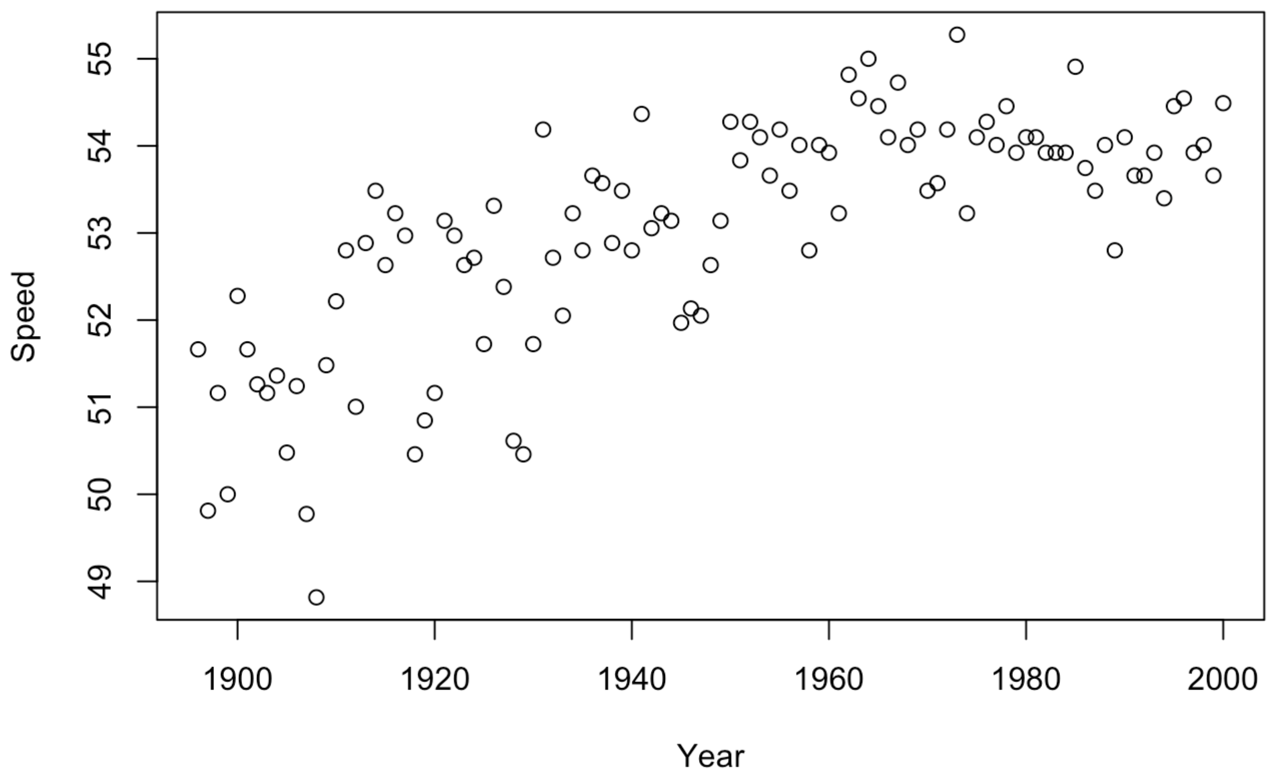
3. Kentucky Derby.

```
derby <- read.csv("~/Downloads/Ex0920.csv")
boxplot(Speed~Condition, data=derby)
```



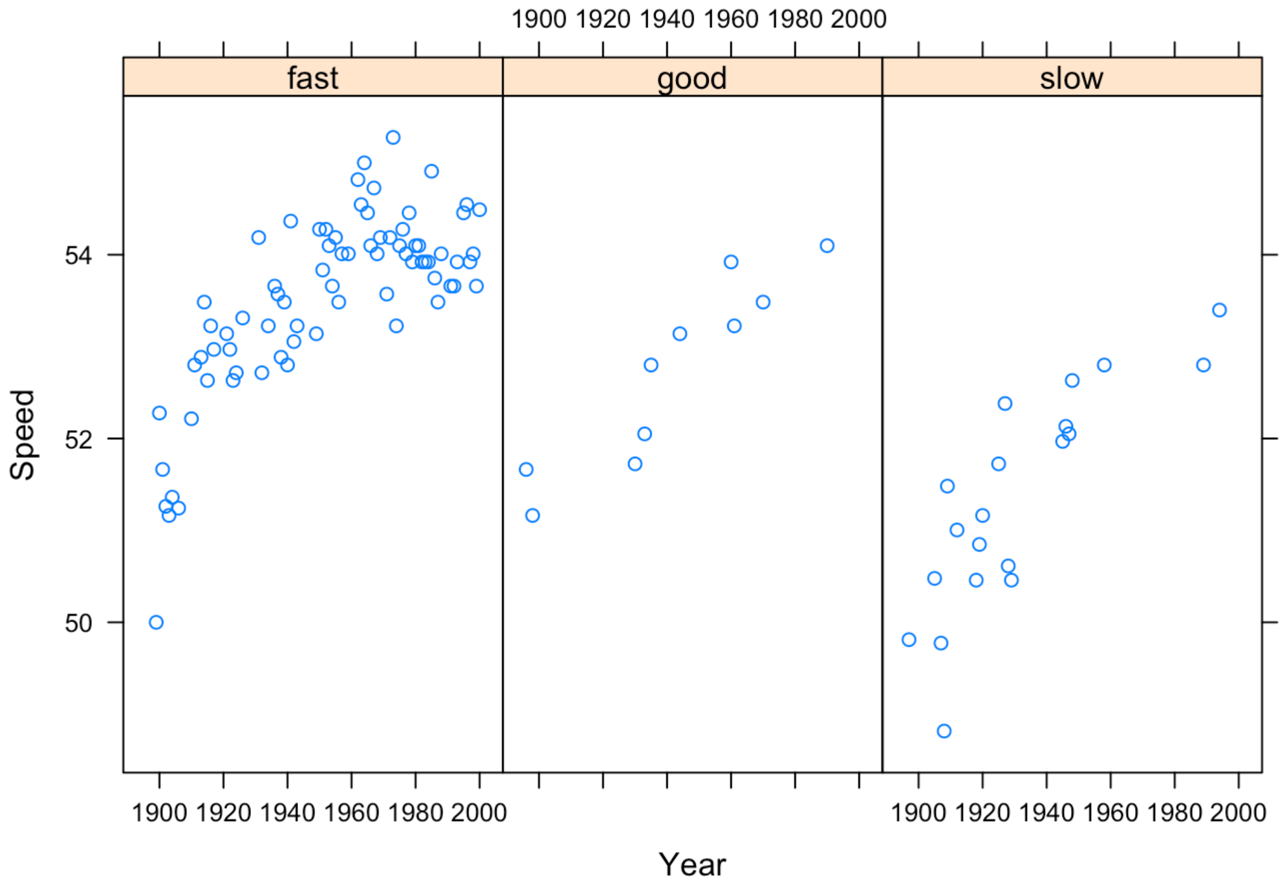
The boxplot above suggests differences in speed for the different levels of categorical variable (fast, good and slow) and a few possible outliers for the level = “fast”.

```
plot(derby$Speed~derby$Year, xlab = "Year", ylab="Speed")
```



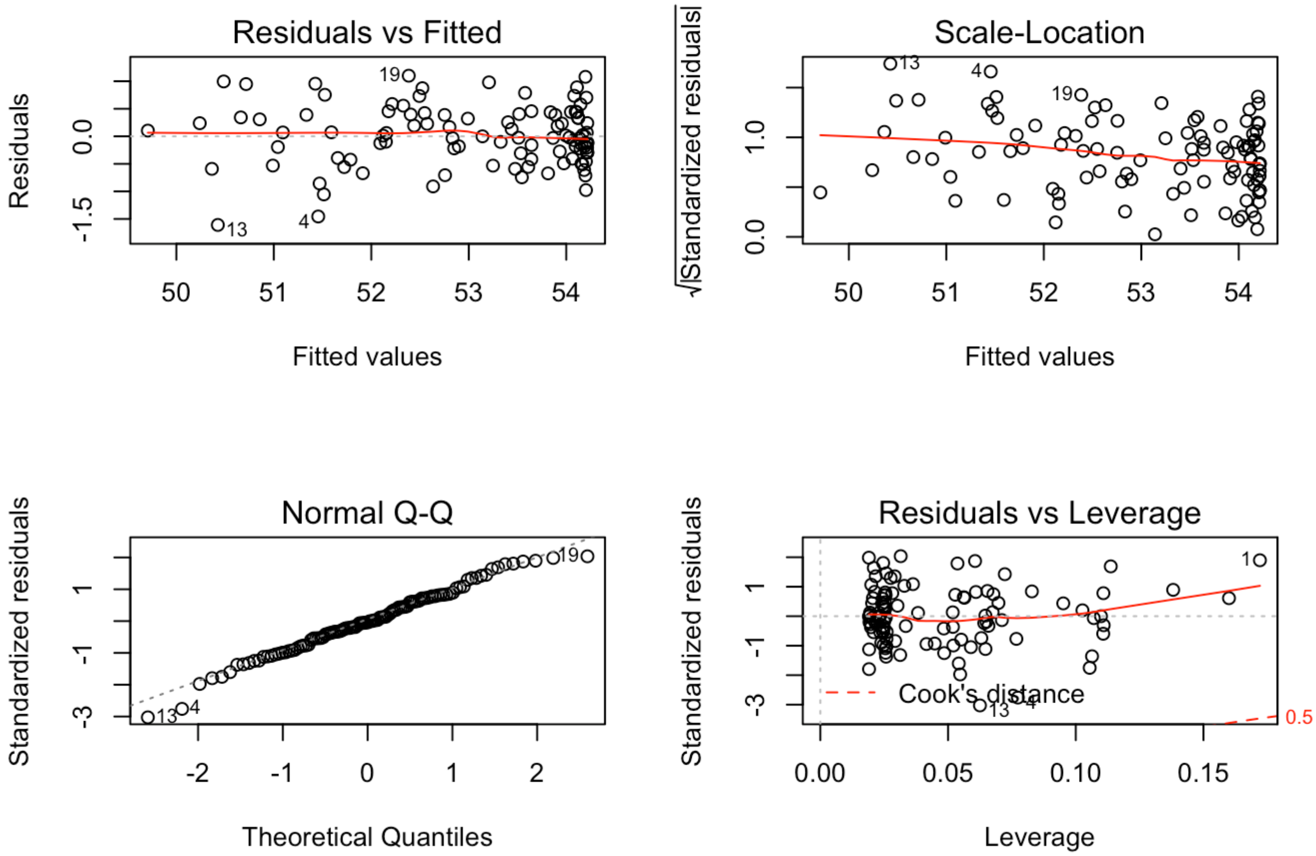
The light quadratic pattern can be observed, so we apply sqr transformation.

```
#Check for interactions between Year and Condition.
library(lattice)
xyplot(Speed~Year | Condition, data = derby)
```



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

```
year_sqr = derby$Year^2
model_derby=lm(Speed ~ Year+ year_sqr + as.factor(Condition), data=derby)
par(mfcol=c(2,2))
plot(model_derby)
```



Residuals plot do not show any patters, so we assume that this model fits well.

```
summary(model_derby)
```

```
##
## Call:
## lm(formula = Speed ~ Year + year_sqr + as.factor(Condition),
##     data = derby)
##
## 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)               -1.598e+03  2.476e+02  -6.452 3.97e-09 ***
## Year                       1.669e+00  2.543e-01   6.563 2.37e-09 ***
## year_sqr                  -4.214e-04  6.526e-05  -6.457 3.89e-09 ***
## as.factor(Condition)good -5.319e-01  1.862e-01  -2.857   0.0052 **
## as.factor(Condition)slow -1.610e+00  1.439e-01 -11.189  < 2e-16 ***
## ---
## 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
```

```
confint(model_derby)
```

```
##                                   2.5 %        97.5 %
## (Intercept)               -2.088874e+03 -1.106401e+03
## Year                       1.164145e+00  2.173011e+00
## year_sqr                  -5.508393e-04 -2.918965e-04
## as.factor(Condition)good -9.012659e-01 -1.625939e-01
## as.factor(Condition)slow -1.895303e+00 -1.324404e+00
```

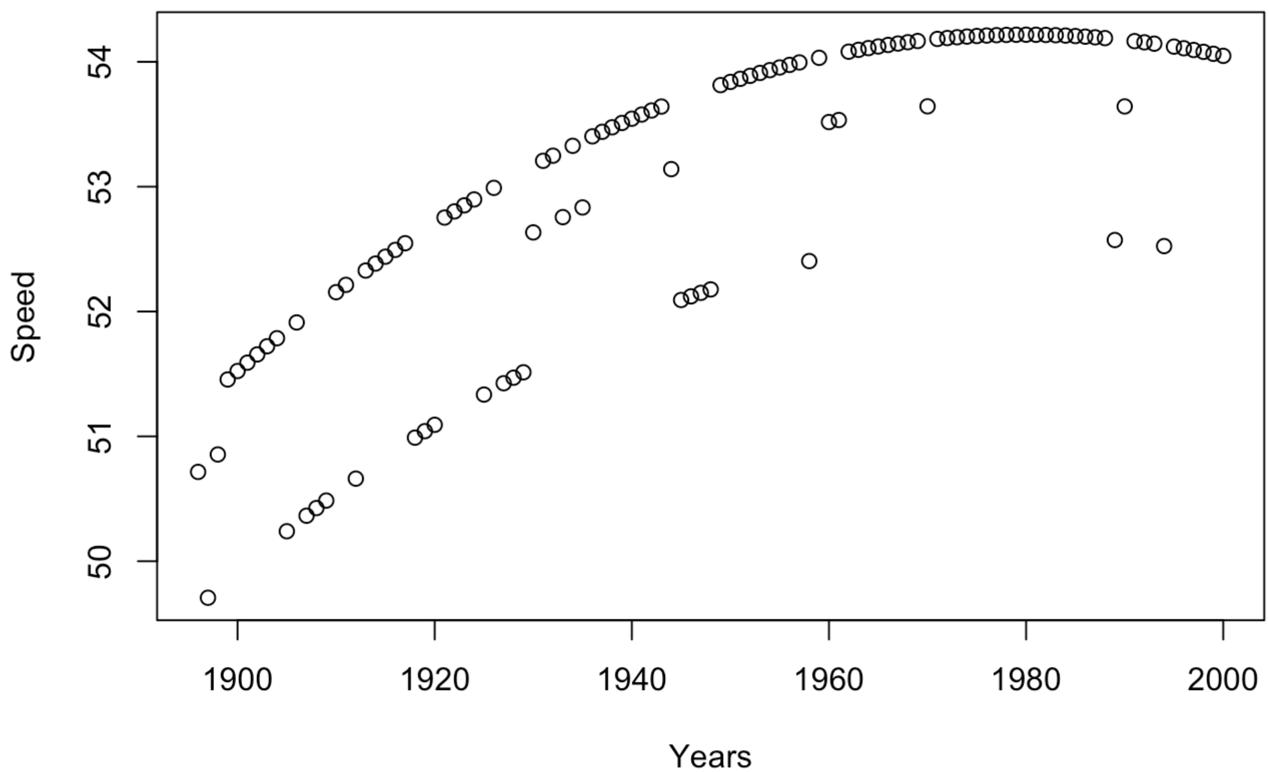
R-squared = 0.8365, which suggests that 83.65% of the data is explained by the model. Relatively large p-value for the Conditiongood = 0.0052 may be caused by small number of observations.

Since “fast” condition is taken as a base line, “good” condition will decrease the speed for 0.5319 units on average and “slow” for 1.61 units on average.

Possible limitations for the model: the limited range of years and conditions. Some other factors could effect the speed (weather or experience of a experience of a jockey).

For visualization of the model, we will build a plot: three lines represent three levels of categorical variable Condition: slow, good, fast respectively from bottom to up.

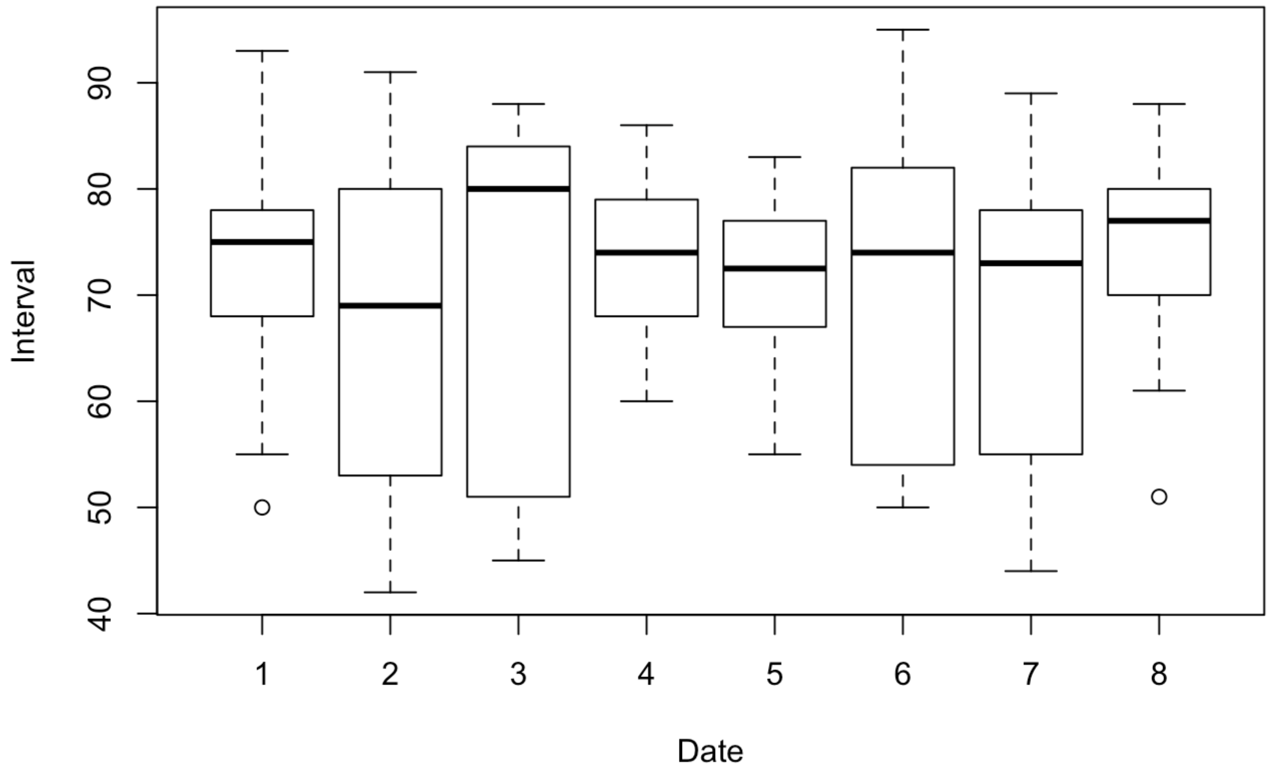
```
plot(y = model_derby$fitted.values, x=derby$Year, xlab = "Years", ylab = "Speed")
```



4. Old Faithful. Report the value of the F statistic, the p-value, and your conclusion.

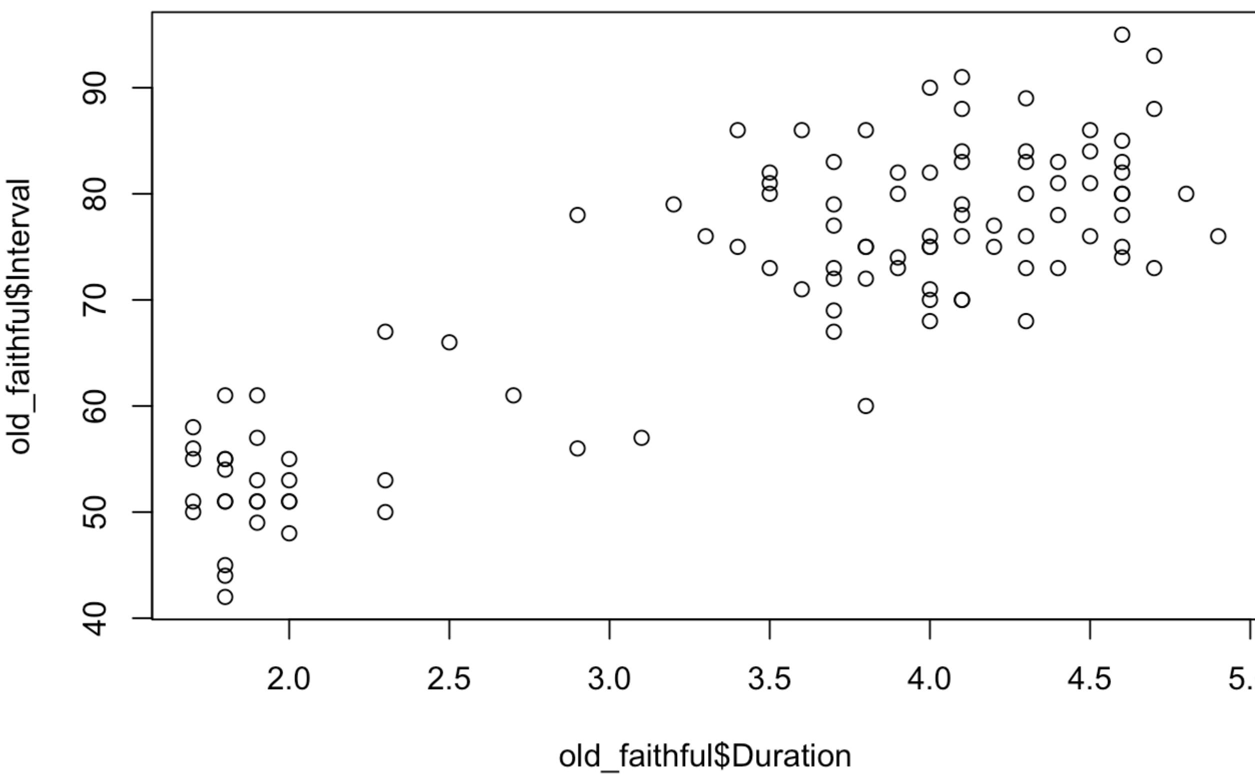
Data includes 107 observations of 4 variables; no missing values.

```
old_faithful <- read.csv("~/Downloads/Ex1015.csv")
boxplot(Interval~Date, data=old_faithful, xlab="Date", ylab="Interval")
```



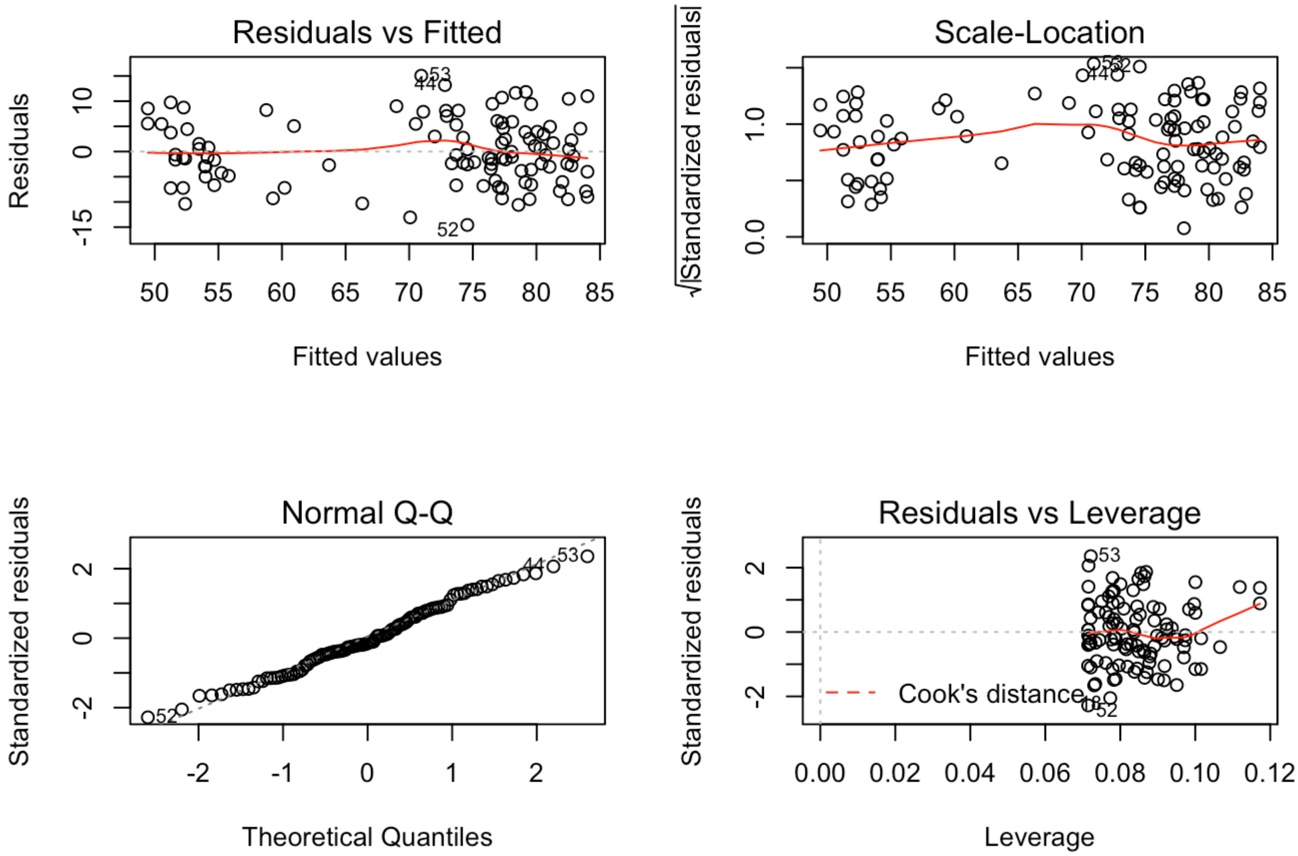
The plot suggests that there is no interaction between Interval and Date (medians of data are almost on the same level).

```
plot(old_faithful$Interval~old_faithful$Duration)
```



R-squared for original untransformed data is 0.7408 with residual standard error 6.866. Light “funnel” pattern suggests non-costant variance, so we apply log transformation.

```
faith_model=lm(Interval~log(Duration)+as.factor(Date), data=old_faithful)
par(mfcol=c(2,2))
plot(faith_model)
```




```
summary(faith_model)
```

```
##
## Call:
## lm(formula = Interval ~ log(Duration) + as.factor(Date), data = old_faithful)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.5672  -4.1247  -0.9158   4.7741  15.0518
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      32.18940    2.97160  10.832   <2e-16 ***
## log(Duration)    32.53717    1.88943  17.221   <2e-16 ***
## as.factor(Date)2  1.07286    2.62175   0.409    0.683
## as.factor(Date)3  0.91387    2.60726   0.351    0.727
## as.factor(Date)4 -1.05940    2.55871  -0.414    0.680
## as.factor(Date)5 -0.52076    2.55422  -0.204    0.839
## as.factor(Date)6  2.16646    2.56763   0.844    0.401
## as.factor(Date)7 -0.06155    2.60976  -0.024    0.981
## as.factor(Date)8 -0.80711    2.60355  -0.310    0.757
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.631 on 98 degrees of freedom
## Multiple R-squared:  0.7582, Adjusted R-squared:  0.7385
## F-statistic: 38.42 on 8 and 98 DF,  p-value: < 2.2e-16
```

Multiple R-squared: 0.7582 > 0.7408 and Residual standard error: 6.631 < 6.866, so second model it more precise. F-statistic: 38.42 on 8 and 98 DF, p-value: < 2.2e-16.

```
faith_model_nodate = lm(Interval~log(Duration), data=old_faithful)
anova(faith_model_nodate, faith_model)
```

```
## Analysis of Variance Table
##
## Model 1: Interval ~ log(Duration)
## Model 2: Interval ~ log(Duration) + as.factor(Date)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    105 4418.4
## 2     98 4309.0  7    109.41 0.3555 0.9256
```

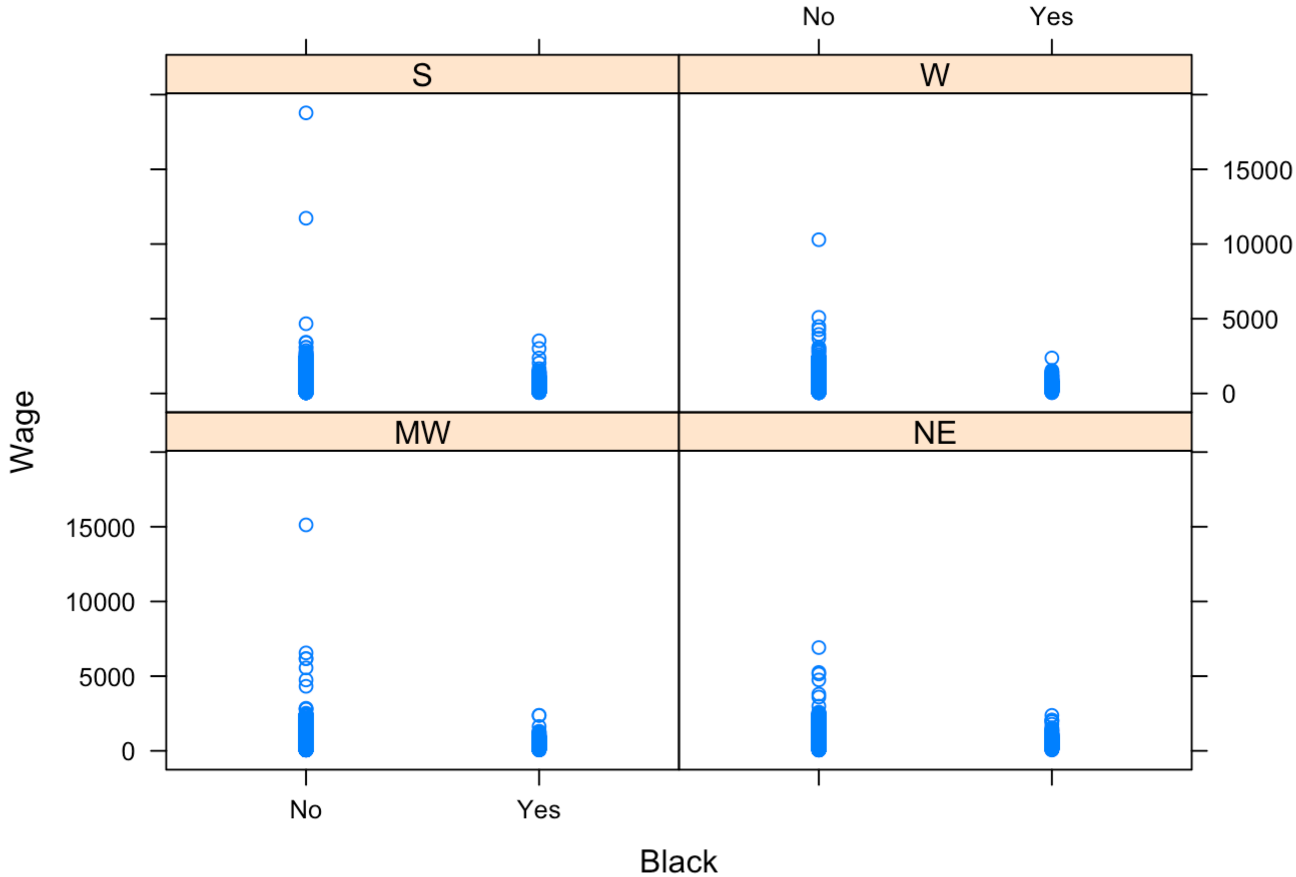
Large p-value for the levels of categorical variable suggest that Date parameter has very small effect on the respond variable and may be excluded from the model.

5. Wages and Race

```
Wage <- read.csv("~/Downloads/Ex1029.csv")
```

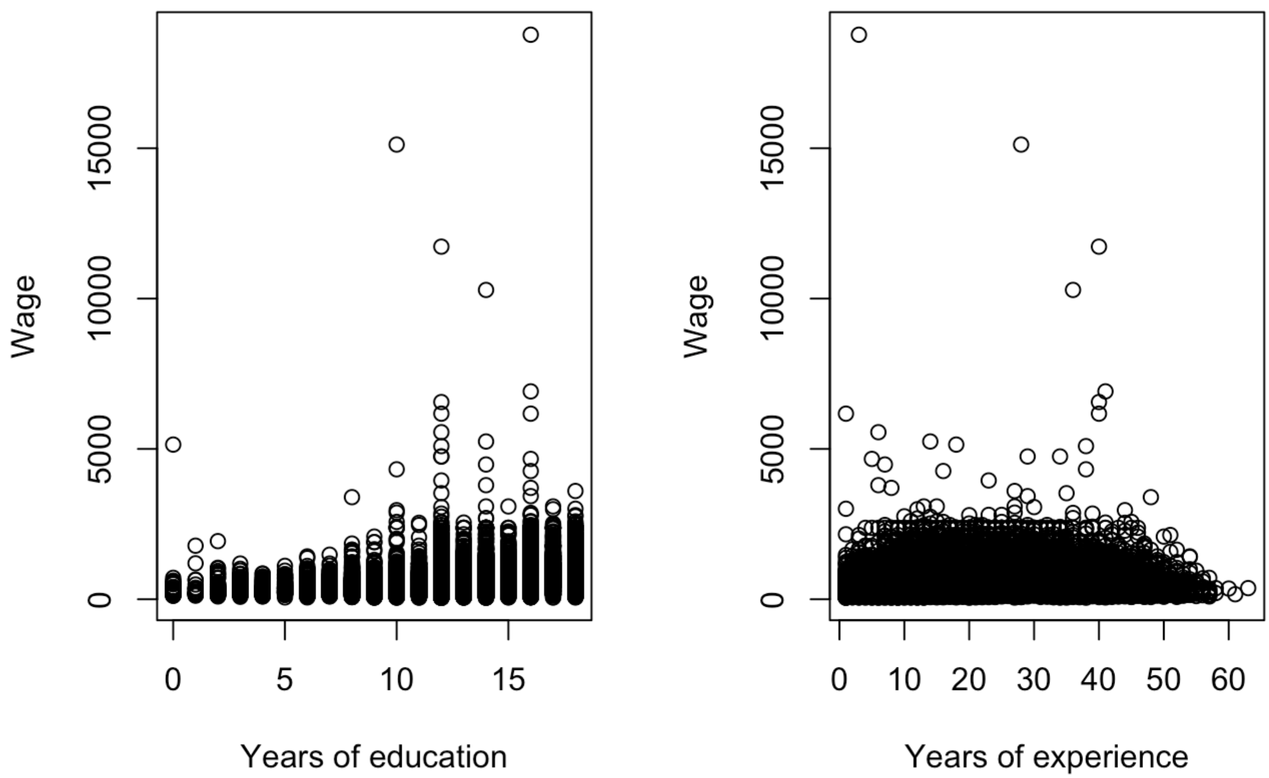
We remove the rows with negative number of years of experience assuming incorrect data enter.

```
Wage <- Wage[Wage$Experience > 0,]
library(lattice)
xyplot(Wage ~ Black | Region, data = Wage)
```



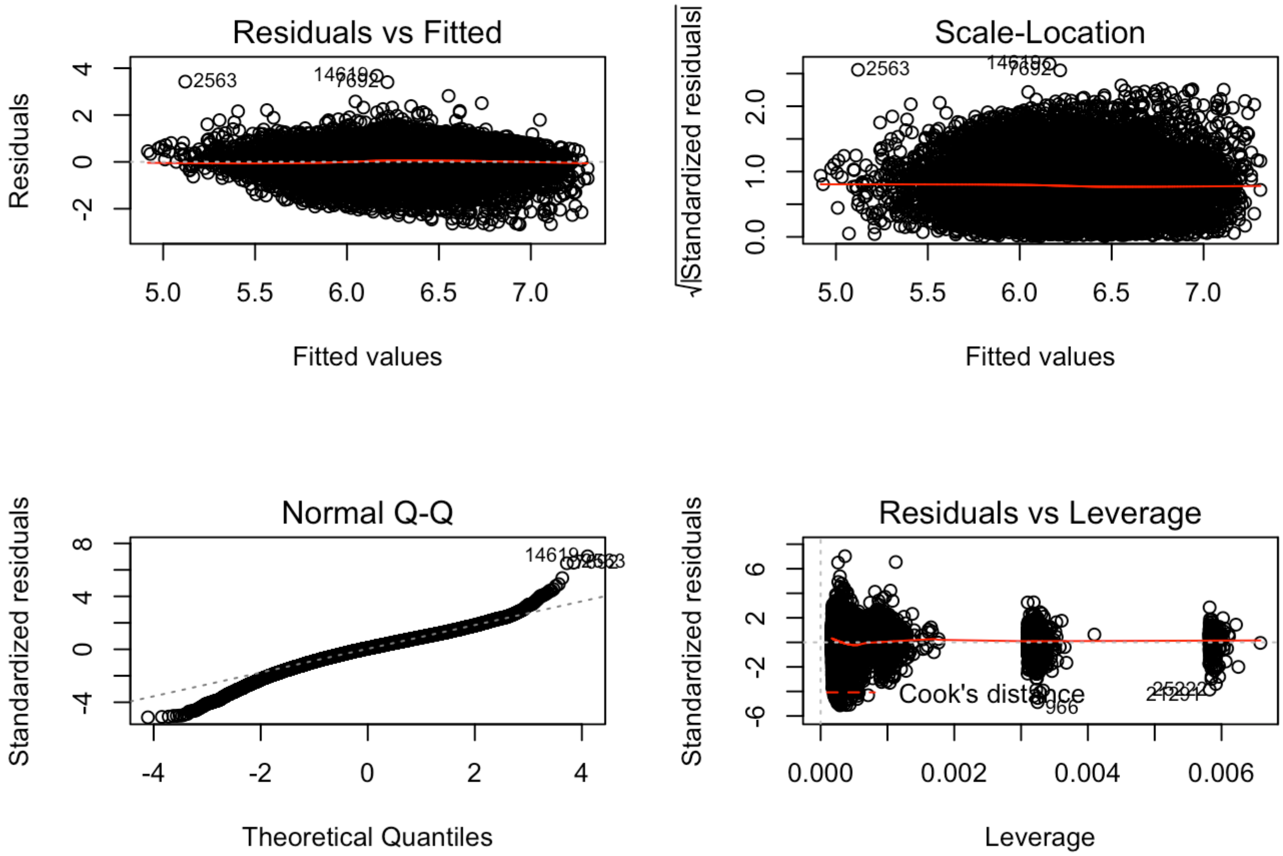
Plots look differemnt, so we suspect the interaction between Black ~ Region, which we test later by using anova() test.

```
par(mfcol=c(1,2))
plot(Wage$Wage~Wage$Education, xlab = "Years of education", ylab = "Wage")
plot(Wage$Wage~Wage$Experience, xlab="Years of experience", ylab = "Wage")
```



Non-constant variance is observed on both plots, so we take the log of the response variable.

```
model_wage = lm((log(Wage)) ~ Education + Experience + as.factor(SMSA) + as.factor(Regio
n)*as.factor(Black), data = Wage)
par(mfcol=c(2,2))
plot(model_wage)
```




```
summary(model_wage)
```

```
##
## Call:
## lm(formula = (log(Wage)) ~ Education + Experience + as.factor(SMSA) +
##     as.factor(Region) * as.factor(Black), data = Wage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6916 -0.2963  0.0404  0.3372  3.6786
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                             4.6270451  0.0196513 235.457
## Education                               0.0965802  0.0011893  81.207
## Experience                              0.0166857  0.0002823  59.098
## as.factor(SMSA)Yes                      0.1597390  0.0077036  20.736
## as.factor(Region)NE                     0.0329428  0.0099710   3.304
## as.factor(Region)S                     -0.0605791  0.0094327  -6.422
## as.factor(Region)W                     -0.0066413  0.0100380  -0.662
## as.factor(Black)Yes                    -0.2460614  0.0303670  -8.103
## as.factor(Region)NE:as.factor(Black)Yes  0.0234638  0.0426296   0.550
## as.factor(Region)S:as.factor(Black)Yes   0.0017202  0.0346618   0.050
## as.factor(Region)W:as.factor(Black)Yes   0.0473136  0.0506530   0.934
##                                        Pr(>|t|)
## (Intercept)                             < 2e-16 ***
## Education                               < 2e-16 ***
## Experience                              < 2e-16 ***
## as.factor(SMSA)Yes                      < 2e-16 ***
## as.factor(Region)NE                    0.000955 ***
## as.factor(Region)S                     1.37e-10 ***
## as.factor(Region)W                     0.508226
## as.factor(Black)Yes                    5.61e-16 ***
## as.factor(Region)NE:as.factor(Black)Yes 0.582042
## as.factor(Region)S:as.factor(Black)Yes  0.960420
## as.factor(Region)W:as.factor(Black)Yes  0.350275
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5238 on 25041 degrees of freedom
## Multiple R-squared:  0.2789, Adjusted R-squared:  0.2786
## F-statistic: 968.5 on 10 and 25041 DF,  p-value: < 2.2e-16
```

```
confint(model_wage)
```

```
##                                               2.5 %      97.5 %
## (Intercept)                              4.58852736  4.66556288
## Education                                0.09424906  0.09891130
## Experience                               0.01613226  0.01723906
## as.factor(SMSA)Yes                       0.14463952  0.17483850
## as.factor(Region)NE                      0.01339898  0.05248660
## as.factor(Region)S                      -0.07906768 -0.04209056
## as.factor(Region)W                      -0.02631640  0.01303384
## as.factor(Black)Yes                     -0.30558251 -0.18654034
## as.factor(Region)NE:as.factor(Black)Yes -0.06009270  0.10702030
## as.factor(Region)S:as.factor(Black)Yes  -0.06621906  0.06965939
## as.factor(Region)W:as.factor(Black)Yes  -0.05196929  0.14659652
```

The model fits the data poorly and explains only 29% of the data; however, after many trials done this was my best result achieved. This model should not be interpreted as a reasonable one or be used to predict values.

For educational purposes, compare the median Wages for black / non-black population.

```
model_wage_noblack = lm((log(Wage)) ~ Education + Experience + as.factor(SMSA) + as.fact
or(Region), data = Wage)
anova(model_wage_noblack, model_wage)
```

```
## Analysis of Variance Table
##
## Model 1: (log(Wage)) ~ Education + Experience + as.factor(SMSA) + as.factor(Region)
## Model 2: (log(Wage)) ~ Education + Experience + as.factor(SMSA) + as.factor(Region) *

##     as.factor(Black)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1  25045 6968.0
## 2  25041 6870.6  4    97.368 88.718 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Small p-value suggests that there is a difference in two models; therefore, Black variable effects the response variable Wage.

Check if the interaction statistically significant:

```
model_wage_interactions = lm((log(Wage)) ~ Education + Experience+ as.factor(SMSA) + as.
factor(Region) + as.factor(Black), data = Wage)
anova(model_wage_interactions, model_wage)
```

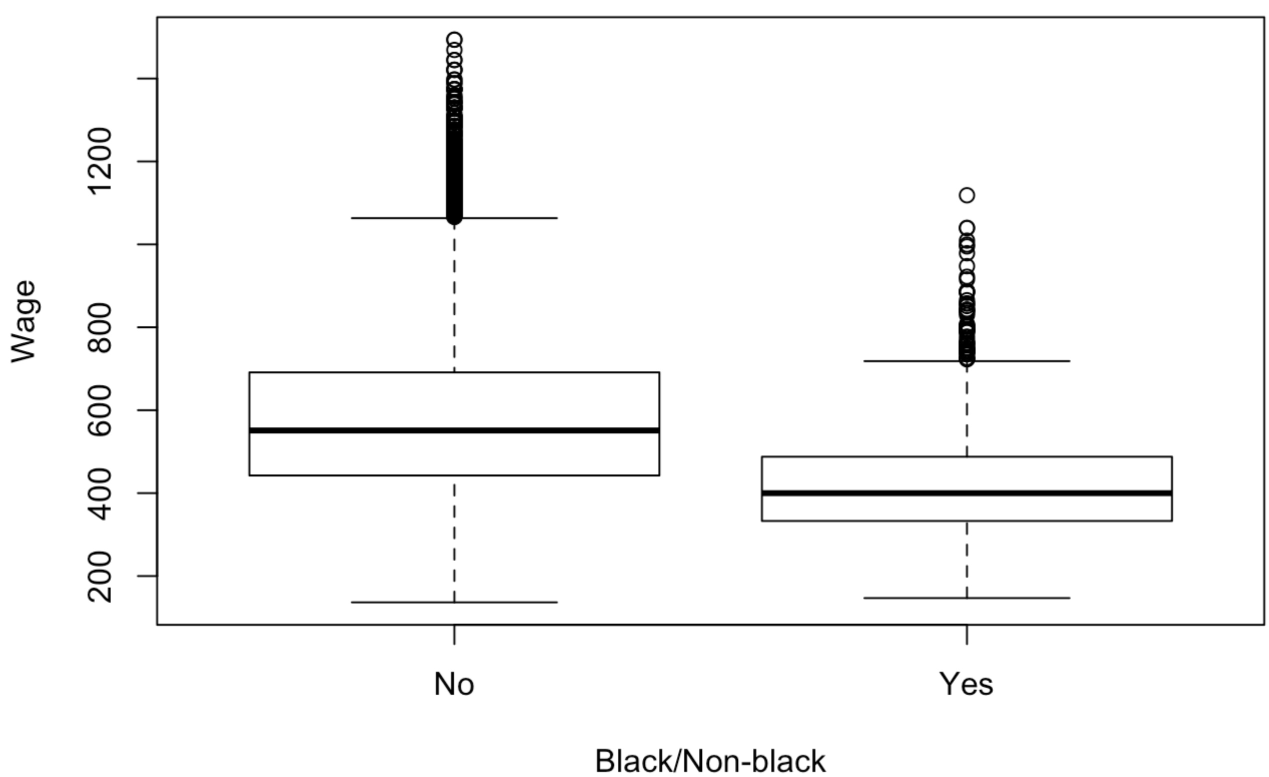
```
## Analysis of Variance Table
##
## Model 1: (log(Wage)) ~ Education + Experience + as.factor(SMSA) + as.factor(Region) +

##     as.factor(Black)
## Model 2: (log(Wage)) ~ Education + Experience + as.factor(SMSA) + as.factor(Region) *

##     as.factor(Black)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1  25044 6871.0
## 2  25041 6870.6  3   0.38333 0.4657 0.7062
```

The result is not statistically significant, and does not change R^2, so we may exclude it from the model.

```
plot(y = exp(model_wage$fitted.values), x=Wage$Black, xlab = "Black/Non-black", ylab =
"Wage")
```



Holding all variables constant, we may assume that the difference in means between Black/Non-Black is about $200. However, the model is not precise and should not be used as an evidence to support the hypothesis that black males are paid less.