IDS 702 - Homework #1 (edited)

Ana Belen Barcenas J.

9/22/2018

1) Interpretation of the Age coefficient:

```
lm_resp_trans <- lm(log(Rate) ~ Age, data = RespRates)
summary(lm_resp_trans)
```

```
##
## Call:
## lm(formula = log(Rate) ~ Age, data = RespRates)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62571 -0.13201 -0.00402  0.13489  0.54771
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.8451185  0.0126277  304.50   <2e-16 ***
## Age         -0.0190090  0.0007357  -25.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1964 on 616 degrees of freedom
## Multiple R-squared:  0.5201, Adjusted R-squared:  0.5193
## F-statistic: 667.6 on 1 and 616 DF,  p-value: < 2.2e-16
```

Now, we are capturing a larger variance (R-square = 52%) than before log transforming the dependent variable (R-square = 47%).

The Age coefficient suggests that an increase of one month in the age of the child is associated with a multiplicative change in the median respiratory rate of exp(.0.019) = 0.98117. Thus, when the child gets older, the median respiratory rate decreases (negative relation).

2) Interpretation of the prediction intervals in natural scale (not log transformed):

Let's predict the 95% confidence intervals for the respiratory rate for 3 individual children: 1 month, 18 months, and 29 months:

```
months = c(1,18,29)
newdata = data.frame(Age = months)
exp(predict(lm_resp_trans, newdata, interval = "prediction"))
```

```
##        fit      lwr      upr
## 1 45.88368 31.17725 67.52721
## 2 33.21353 22.57614 48.86302
## 3 26.94664 18.30537 39.66714
```

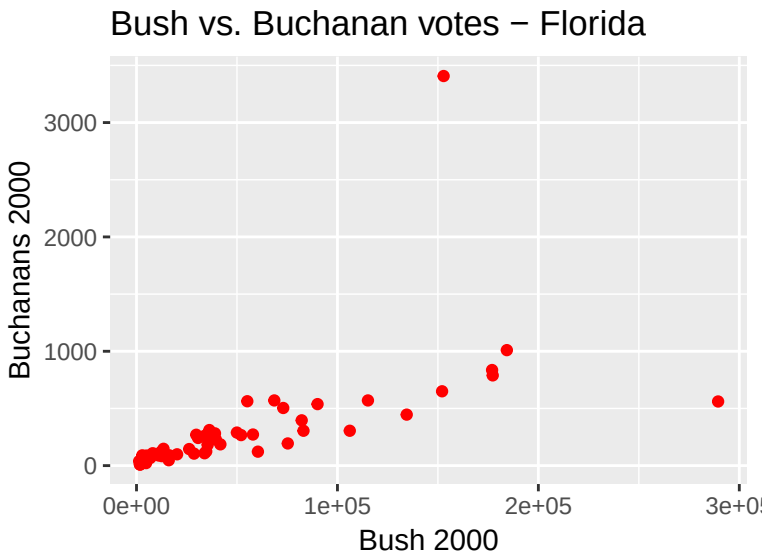
The prediction intervals suggests that:

- A 1 month old children will have a respiratory rate between 31.18 and 67.53 with 95% confidence
- A 18 months child will have a respiratory rate between 22.58 and 48.86 with 95% confidence
- A 29 months child will have a respiratory rate between 18.31 and 39.67 with 95% confidence.

3) Compare final model with outlier vs. without outlier:

Let's see a scatterplot of the votes in each Florida county:

```
ggplot(data = Elections) + geom_point(mapping = aes(x = Bush2000, y = Buchanan2000), colour = "red") + g
```



From the scatter plot is evident that the number of votes Buchanans received in Palm Beach is an atypical result given the regular relation between Bush and Buchanans number of votes.

After fitting a linear regression to the data without the outlier and without transforming the variables, the linearity and constant variance assumptions seems to be violated. Therefore, I'll try log transforming both the dependent and independent variables.

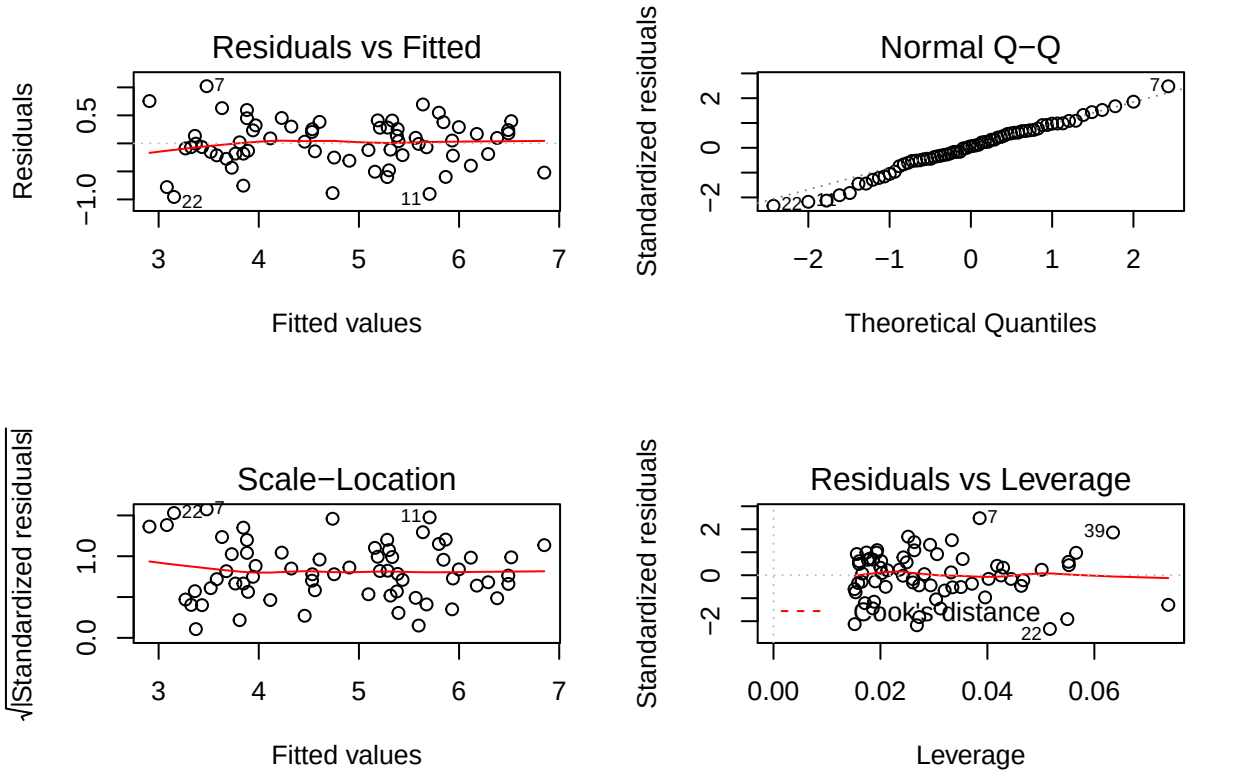
```
Elections_2 <- subset(Elections, County!='Palm Beach')

lm_elect <- lm(log(Buchanan2000) ~ log(Bush2000), data = Elections_2)
summary(lm_elect)
```

```
##
## Call:
## lm(formula = log(Buchanan2000) ~ log(Bush2000), data = Elections_2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95631 -0.21236  0.02503  0.28102  1.02056
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.34149    0.35442  -6.607 9.07e-09 ***
## log(Bush2000)  0.73096    0.03597  20.323  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4198 on 64 degrees of freedom
## Multiple R-squared:  0.8658, Adjusted R-squared:  0.8637
## F-statistic:   413 on 1 and 64 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(lm_elect)
```




```
par(mfrow=c(1,1))
```

Now both assumptions are met.

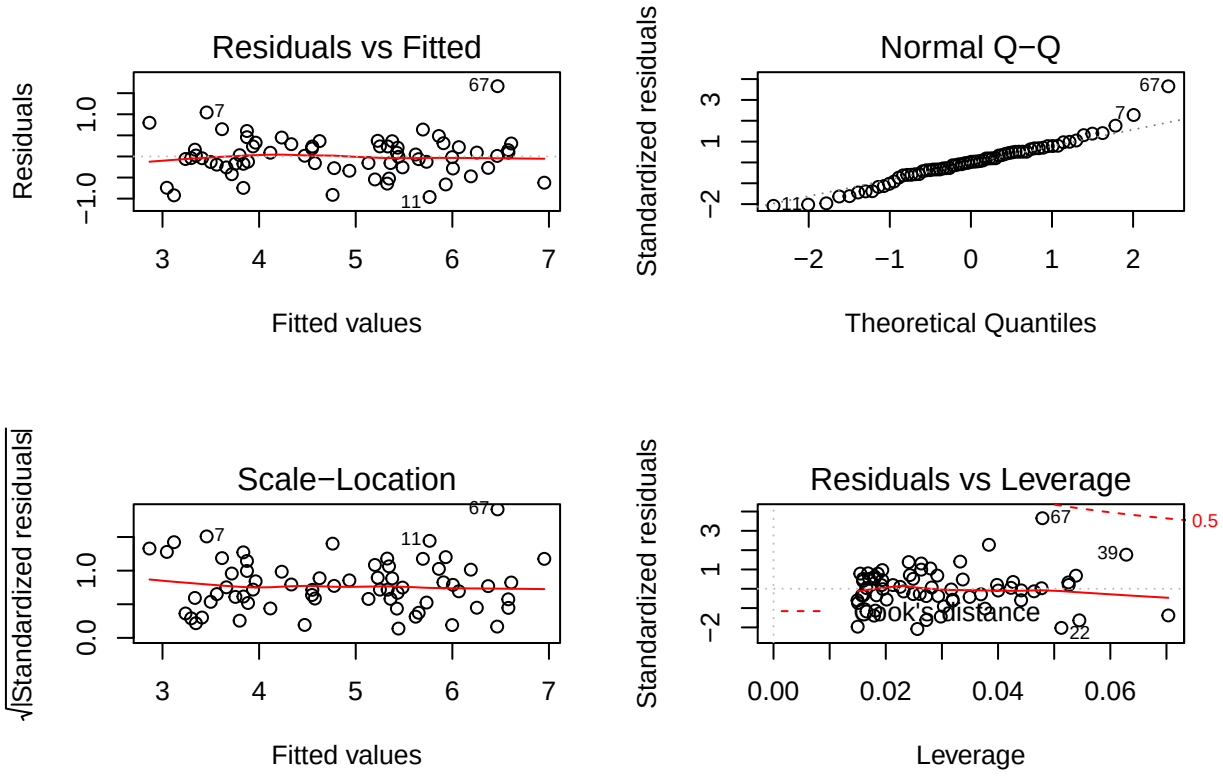
Let's fit the final model again but including the outlier I deleted before to compare the results:

```
lm_elect <- lm(log(Buchanan2000) ~ log(Bush2000), data = Elections)
summary(lm_elect)
```

```
##
## Call:
## lm(formula = log(Buchanan2000) ~ log(Bush2000), data = Elections)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.96075 -0.25949  0.01282  0.23826  1.66564
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.57712    0.38919  -6.622 8.04e-09 ***
## log(Bush2000)  0.75772    0.03936  19.251  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4673 on 65 degrees of freedom
## Multiple R-squared:  0.8508, Adjusted R-squared:  0.8485
## F-statistic: 370.6 on 1 and 65 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(lm_elect)
```




```
par(mfrow=c(1,1))
```

It seems that the normality assumption is violated if we include the outlier for Palm Beach. Moreover, the R-squared is smaller in the regression where the outlier is present. Thus, there is evidence suggesting that the number of votes Palm Beach received on the 2000 presidential election is atypical and represents an outlier (y-axis) given the results in other counties.