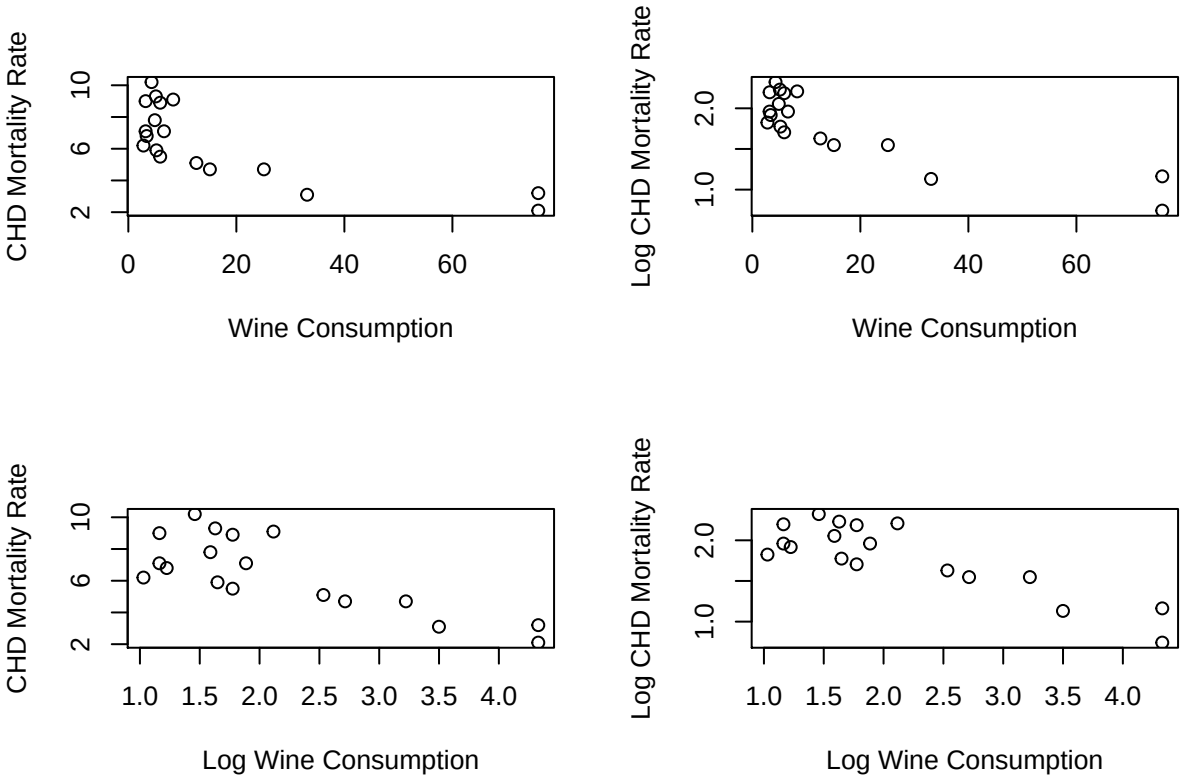
PHP 2514 - Homework 2

Jess Kaminsky

3/05/2018

QUESTION 1

When exploring the relationship between CHD mortality rate from wine consumption level, the plot of of the raw data - mortality rate vs. wine consumption - appears nonlinear. To further explore the trend between our outcome and predictor, we will consider log transformations on the outcome, the predictor, or both. Based on the scatterplots below, plot 4, the plot that shows log wine consumption vs log mortality appears the most linear.



To further assess which transformation will generate the best predictive model, we will generate a linear model for the raw data and each of the transformations to be considered.

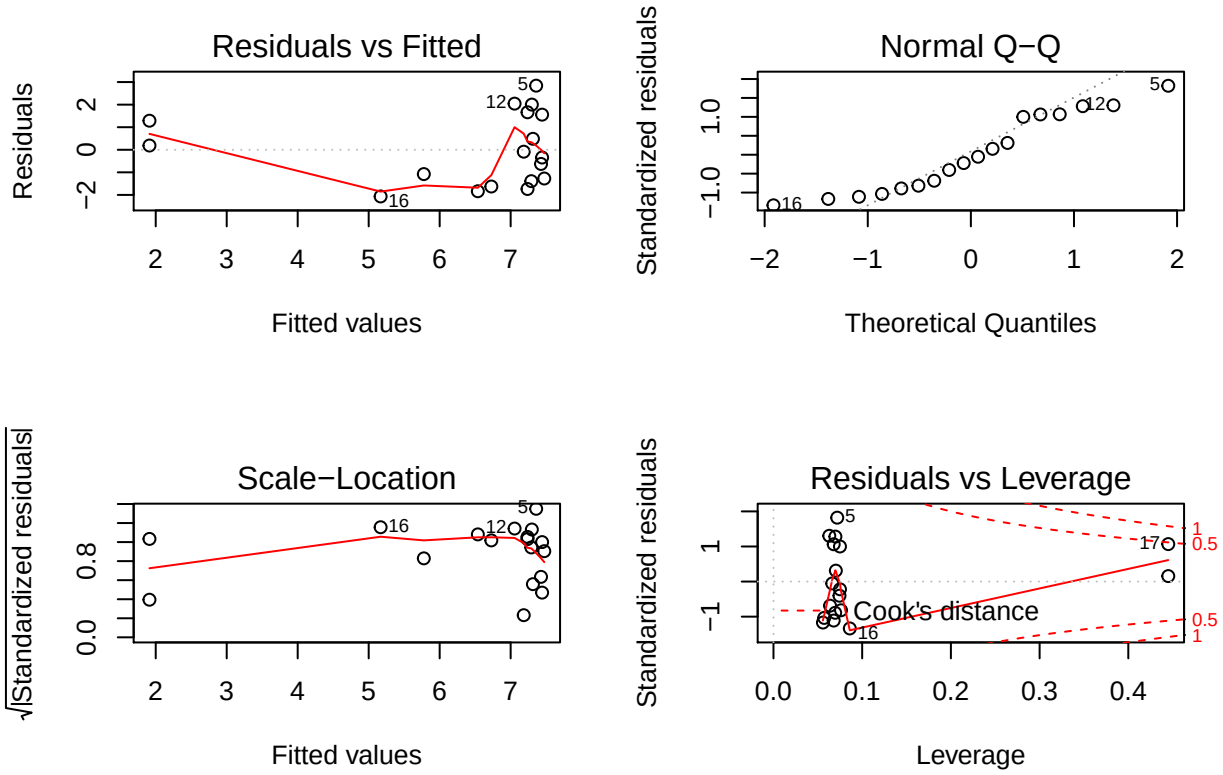
When assessing the four linear models that have been fit, model 4 appears to be the best model for predicting CHD mortality from wine consumption. Model 4 has the highest r-squared value - 72.21% of the variation in CHD mortality rate can be explained by wine consumption. Also, when comparing the normal Q-Q plots among all 4 models, we see that the points on the Q-Q plot for model 4 fall closest to the line overall - therefore satisfying the normality assumption of a linear model. The points on the plot of residuals vs fitted for model 4 appears to be the most randomly and evenly distributed around 0 when compared to the other 3 models. The residual standard error for this model is 0.2885, the smallest among all 4 models.

When fitting a linear model on the raw data, predicting mortality rate from wine consumption, the results indicate:

CHD Mortality Rate = 7.68655 - 0.07608*Wine Consumption

This model suggets that the mortality rate for a country with wine consumption equal to 0 is 7.68655. For every 10 unit increase in wine consumption, we expect CHD mortality rate to decrease by .7608.

```
##
## Call:
## lm(formula = wine$MORTALITY ~ wine$WINE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0683 -1.3616 -0.2138  1.4897  2.8406
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.68655    0.47332  16.240 2.31e-11 ***
## wine$WINE   -0.07608    0.01700  -4.475 0.000383 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.619 on 16 degrees of freedom
## Multiple R-squared:  0.5559, Adjusted R-squared:  0.5281
## F-statistic: 20.03 on 1 and 16 DF,  p-value: 0.0003828
```



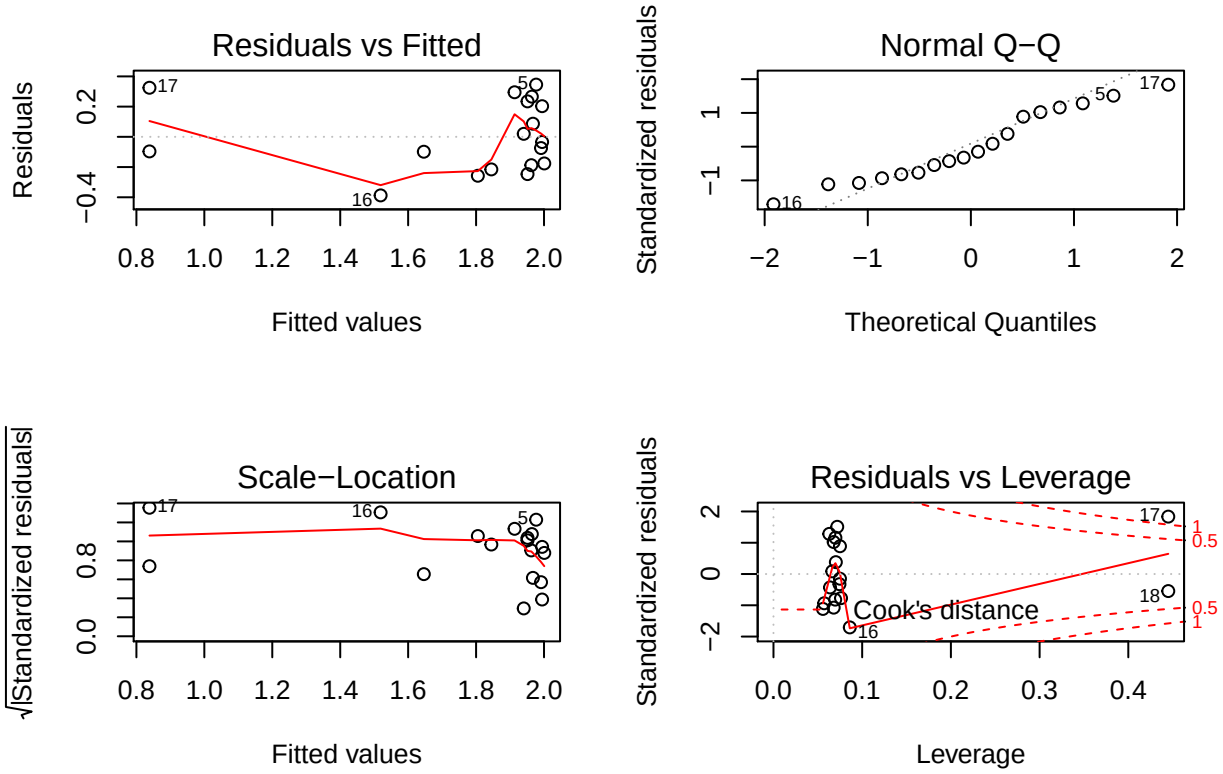
Model 2 fits a linear model predicting the log mortality rate from wine consumption.

Log CHD Mortality Rate = 2.045 - 0.0159*Wine Consumption

When wine consumption is equal to 0 the expected log mortality rate is 2.045. To transform this expectation back from the log scale, we should compute eˆ2.045 = 7.729. That is - if a country does not consume wine, its CHD Mortality Rate is expected to be 7.729. For every 10 unit increase in wine consumption, we expect CHD mortality rate to decrease by 79.6% - CHD mortality is 20.39% of its original value determined by calculating eˆ(-0.0159*10).

```
##
## Call:
## lm(formula = wine$log.mortality ~ wine$WINE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38757 -0.18476 -0.05428  0.22667  0.34550
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.045256   0.069438  29.454 2.29e-15 ***
## wine$WINE   -0.015900   0.002494  -6.375 9.21e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2375 on 16 degrees of freedom
```

```
## Multiple R-squared:  0.7175, Adjusted R-squared:  0.6999
## F-statistic: 40.64 on 1 and 16 DF,  p-value: 9.208e-06
```



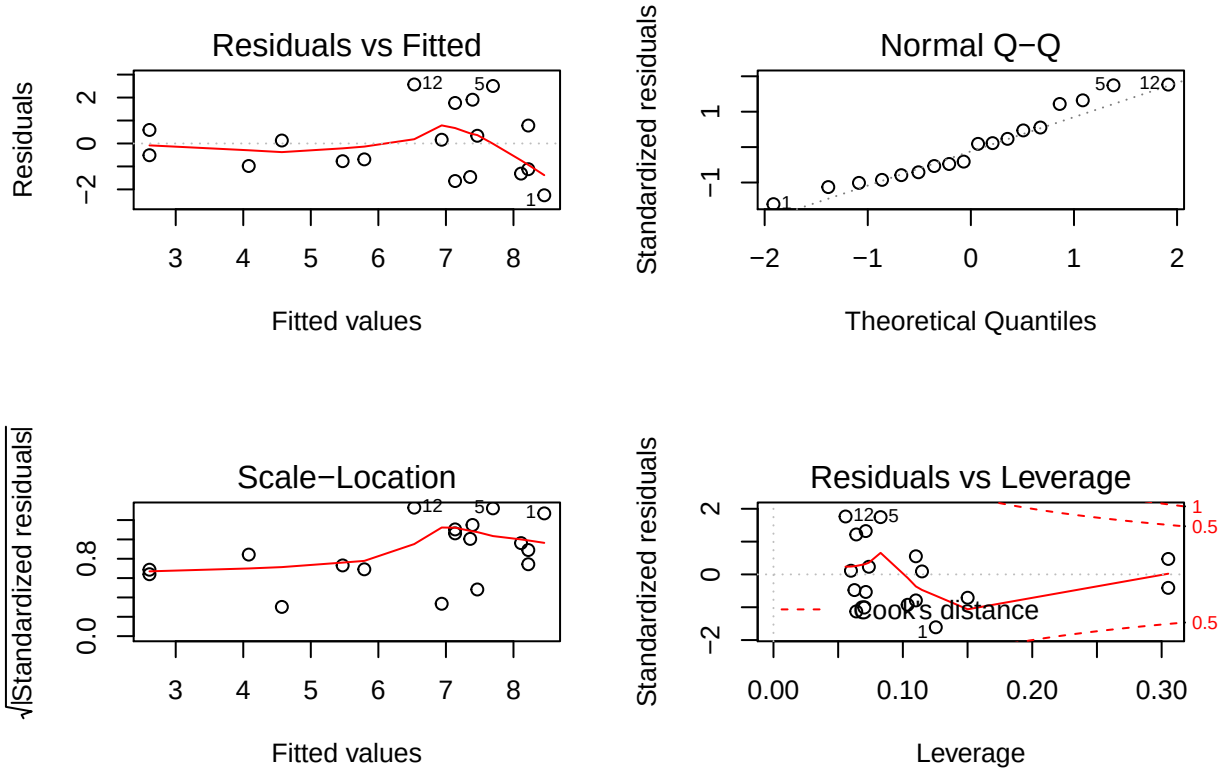
Model 3 fits a linear model predicting CHD mortality rate from log wine consumption.

CHD Mortality Rate = 10.2795 - 1.771*Log Wine Consumption**

When wine consumption is equal to 1 - log wine consumption is equal to 0 - we expect CHD mortality rate to be 10.2795. Doubling the wine consumption rate decreases the CHD mortality rate by log(2)*1.771 = 1.227.

```
##
## Call:
## lm(formula = wine$MORTALITY ~ wine$log.wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2559 -1.0849 -0.1914  0.7326  2.5687
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.2795     0.8316  12.360 1.34e-09 ***
## wine$log.wine  -1.7712     0.3468  -5.108 0.000105 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.498 on 16 degrees of freedom
## Multiple R-squared:  0.6199, Adjusted R-squared:  0.5961
## F-statistic: 26.09 on 1 and 16 DF,  p-value: 0.0001054
```



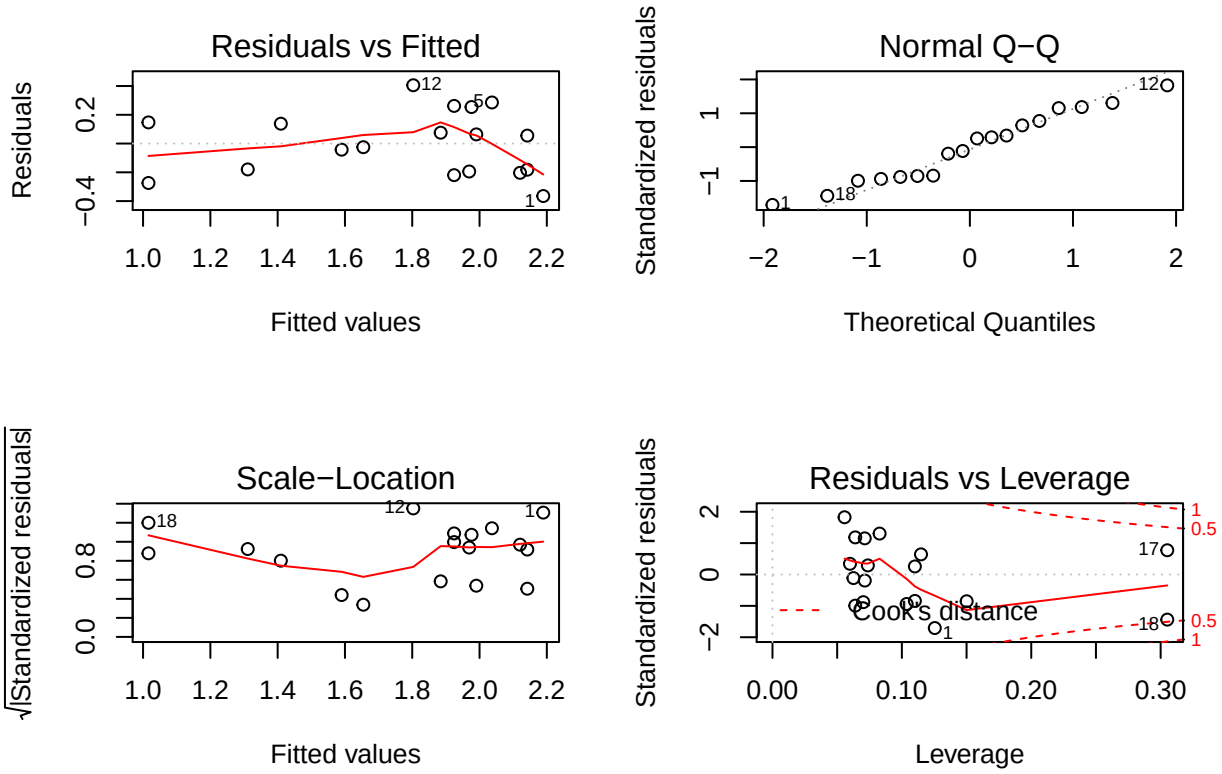
Model 4 fits a linear model predicting log CHD mortality rate from log wine consumption.

Log CHD Mortality Rate = 2.55 - 0.3556*Log Wine Consumption**

When wine consumption is equal to 1 - log wine consumption is equal to 0 - we expect mortality rate to be e^2.55 = 12.81. If wine consumption doubles then CHD mortality decreases by 21.84% - CHD mortality is 78.15% of its original value determined by calculating e^(-0.355*log(2)) = 0.7815.

```
##
## Call:
## lm(formula = wine$log.mortality ~ wine$log.wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36487 -0.19122  0.01497  0.14485  0.40525
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.55555    0.12690  20.139 8.60e-13 ***
## wine$log.wine -0.35560    0.05291  -6.721 4.91e-06 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2285 on 16 degrees of freedom
## Multiple R-squared:  0.7384, Adjusted R-squared:  0.7221
## F-statistic: 45.17 on 1 and 16 DF,  p-value: 4.914e-06
```



QUESTION 2

We can see in Display 8.25 that Buchanan received more votes than expected in Palm Beach County because the point falls much higher on the y-axis than every other data point - the rest of the points in the scatterplot seem to follow a somewhat linear trend, but the data point for Palm Beach is an extreme outlier.

When initially exploring the data, we see that a scatter plot of the votes for Bush vs. the votes for Buchanan - after remove the Palm Beach data - appear to be somewhat linearly correlated. The correlation between the 2 variables is 0.867.

Display 8.25 Votes for George W. Bush and Pat Buchanan in all Florida counties

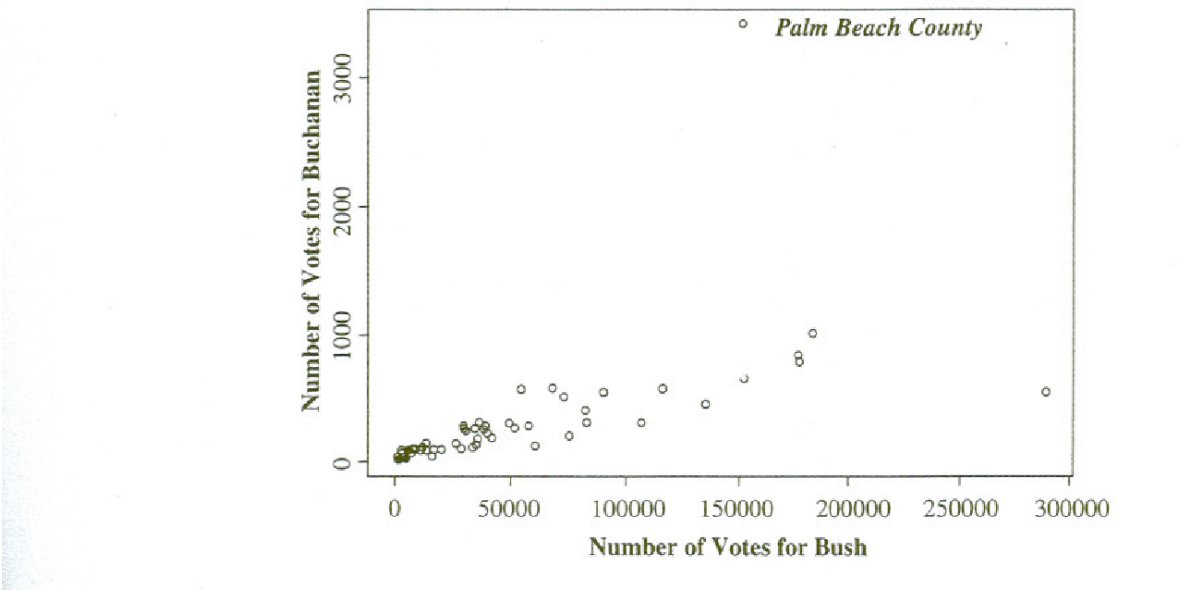
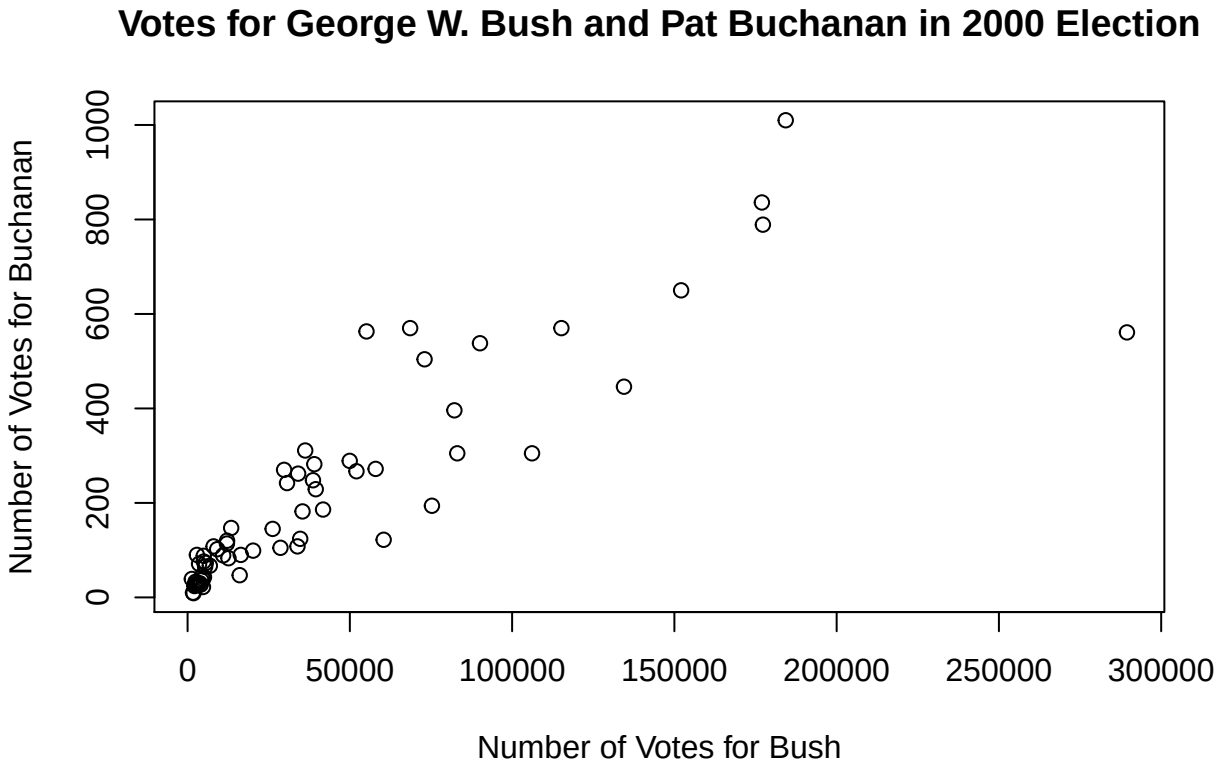


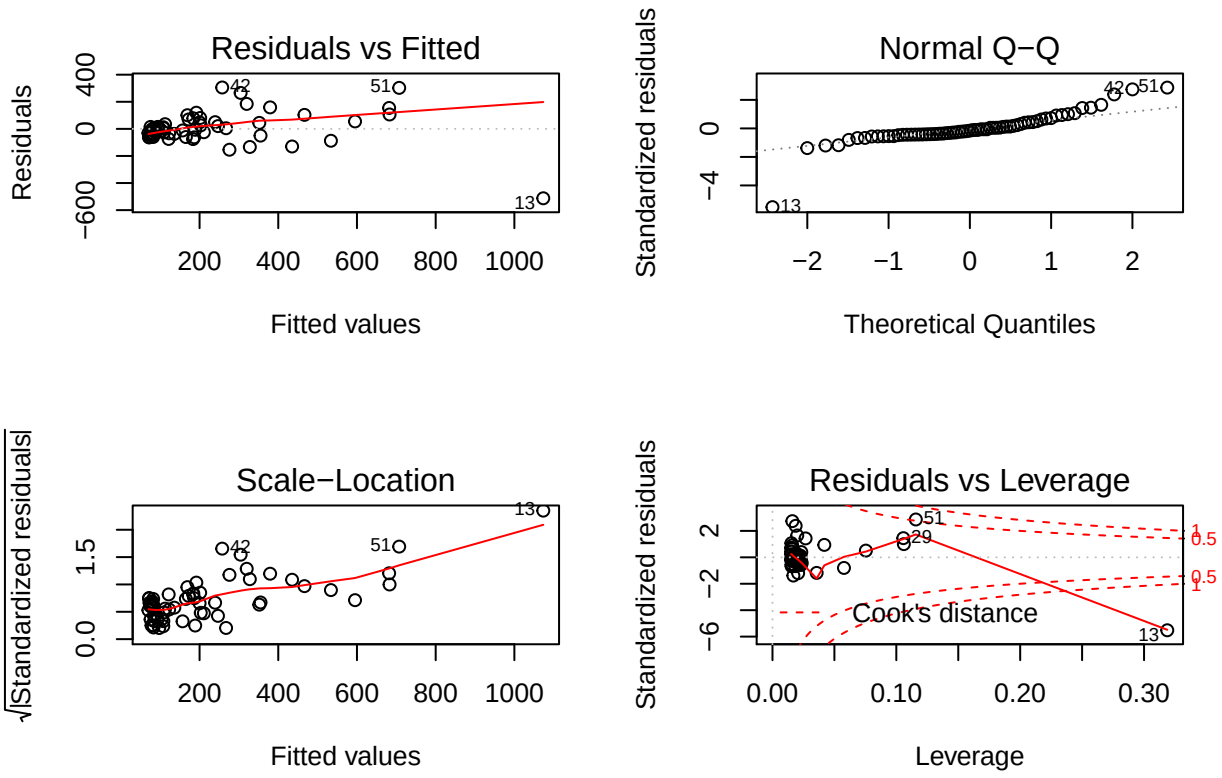
Figure 1:



Although, after looking at Q-Q plot from a simple model predicting Buchanan's votes from

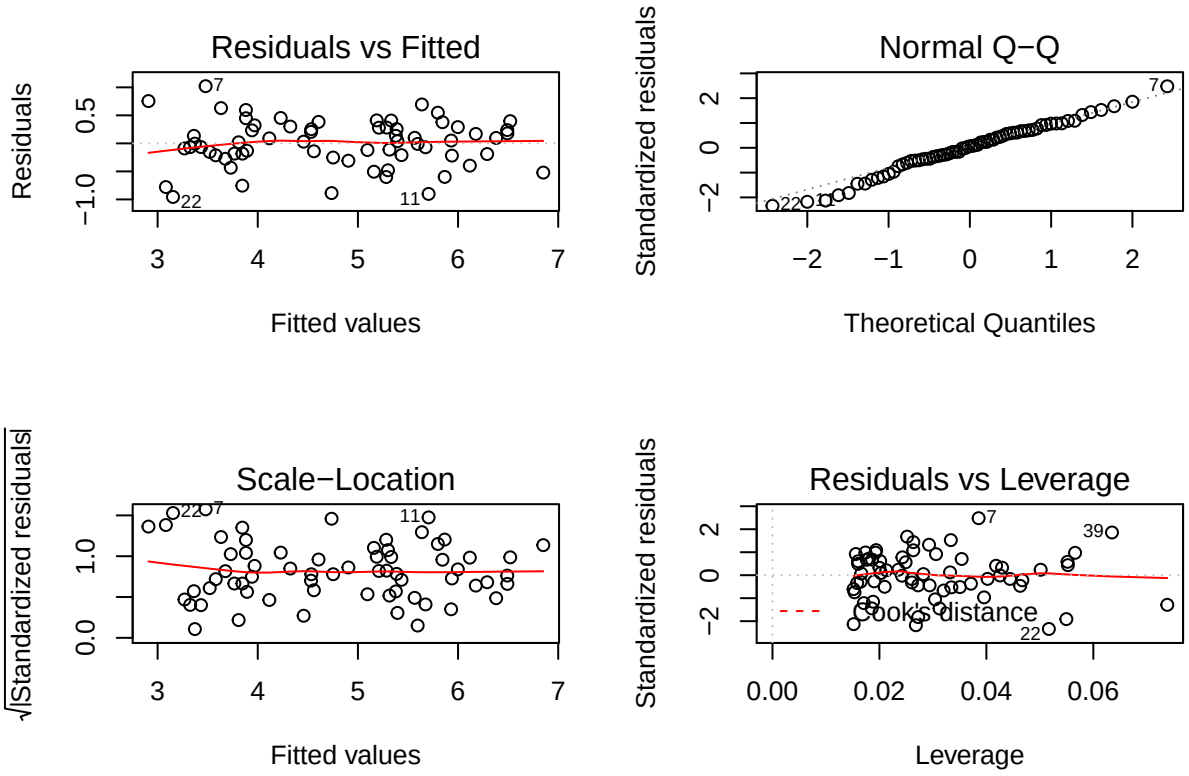
Bush's votes in 2000, we see that the data does not appear normally distributed.

Plots of the Linear Model Prediciting Buchanan Votes from Bush Votes on the raw data



When considering a transformation of the data, applying a log transformation to both the outcome and predictor variables appears to provide the best model fit.

Plots of the Linear Model Prediciting Log of Buchanan Votes from Log of Bush Votes



The predictive linear model generated on the log transformed data is:

Log(Buchanan Votes in 2000) = -2.341 + 0.731*Log(Bush Votes in 2000)

If Bush received 1 vote, log of his votes would be equal to 0 - we would then expect Buchanan to receive e^(-2.341) = 0.096 votes; however this is extrapolation as Bush did not receive only 1 vote in any county in Florida. When interpeting the predictive model, we can see that when the vote count for Bush doubles, Buchanan's votes will increase by 65.9%

```
##
## Call:
## lm(formula = buchanan2000 ~ bush2000, data = log.remove_pb)
##
## 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 ***
## 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
```

Based on the linear model presented above and the Palm Beach data from 2000, we expect that Buchanan should have received approximately 592 votes. We are 95% confident that the true number of votes for Buchanan should have been between 250 and 1399 votes - based on a 95% prediction interval.

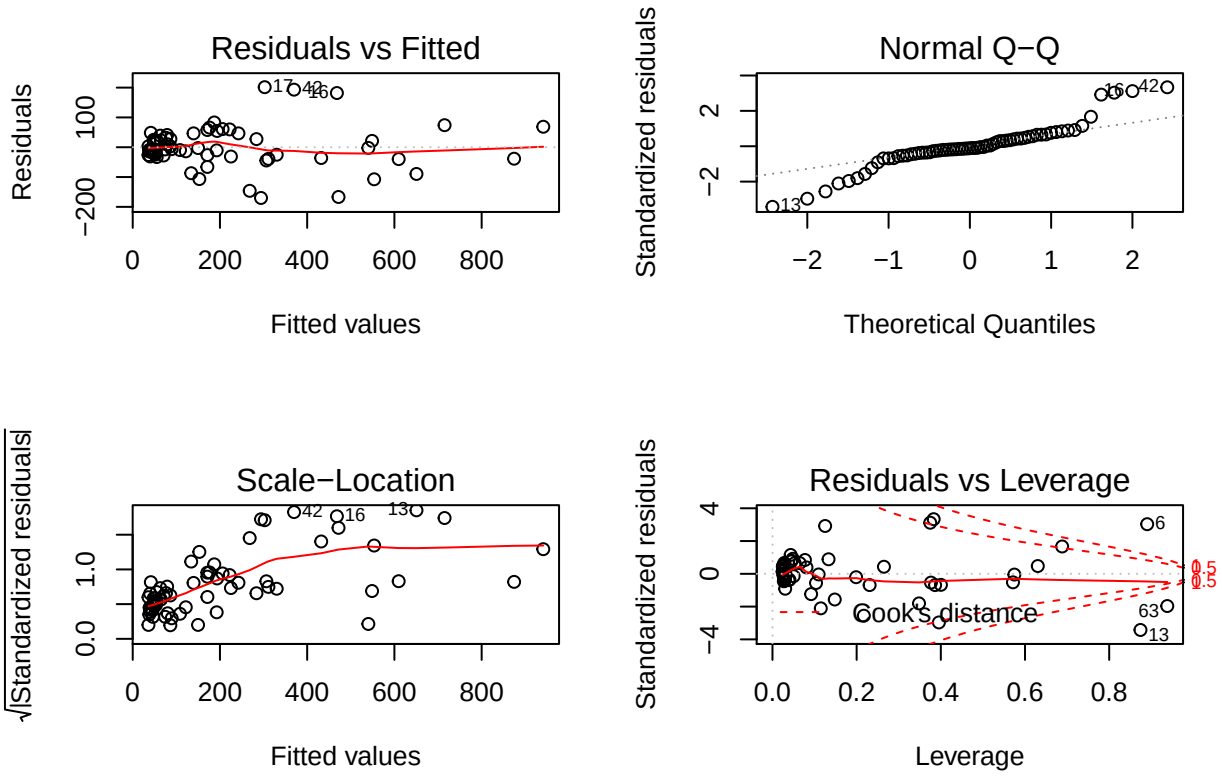
fit	lwr	upr
592.3769	250.8001	1399.164

In Palm Beach County in 2000, Buchanan received 3407 votes. Using the prediction interval and assuming that this vote count actually contains a number of votes intended for Gore it is likely that between 2008 and 3157 of the votes Buchanan received were actually intended for Gore. These numbers were calculated by subtracting the upper and lower bounds of the prediction interval of Buchanan votes in Palm Beach from the actualy count of Buchanana votes in Palm Beach.

We will now continue to explore and analyze our data by trying to find the best linear model for predicting the number of votes for Buchanan in 2000 using all the variables provided in our dataset.

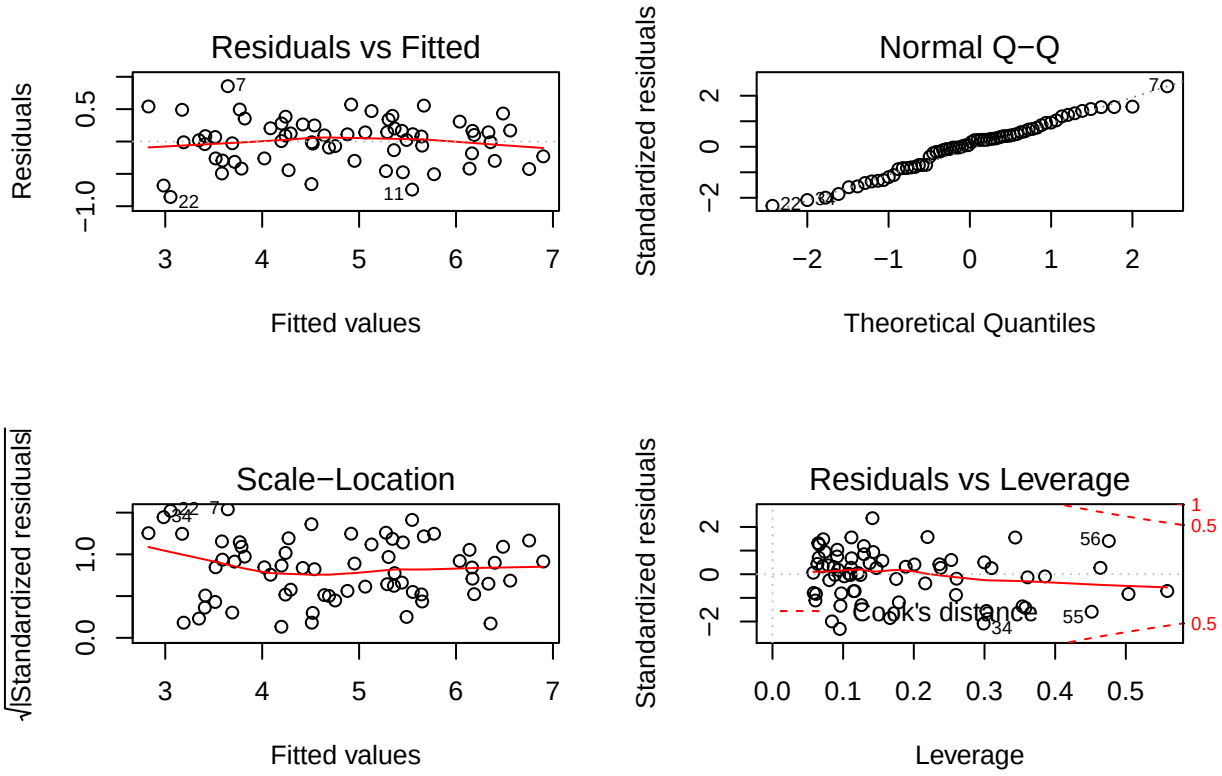
The first step is to explore the linear model prediciting Buchanan's votes from all the variables in our dataset without any transformation. From the Q-Q plot below, we can see that many of the points deviate from the line, indicating that a transformation of our data is likely appropriate in order to ensure normality of our data.

Plots of the Linear Model Prediciting Buchanan Votes from all variables



After exploring different combinations of transformations on the dependent variables and predictors, the transformation that appears to satisfy the normality assumption best is a log transformation of all variables. This can be seen in the Q-Q plot below.

Plots of the Linear Model Prediciting Log of Buchanan Votes from Log of all variables



Results of the log transformed model are presented below. The intercept term, Browne2000, and Perot96 are the only significant coefficients in our model. The model has a fairly strong correlation coefficient of 0.8829. When assessing potential collinearity among the predictors in the model, the collinearity matrix below has a fairly high condition index for the reform.reg and total.reg predictors. This makes sense because reform.reg is the registration count in Buchanan's reform party while total.reg is the total political party registration. Reform.reg plus the registration of all other parties is equal to total.reg.

```
##
## Call:
## lm(formula = buchanan2000 ~ bush2000 + gore2000 + nader2000 +
##     browne2000 + total2000 + clinton96 + dole96 + perot96 + buchanan96p +
##     reform.reg + total.reg, data = log.remove_pb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.85705 -0.26158  0.04603  0.20383  0.85470
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.00620    1.93622  -2.586  0.01245 *
## bush2000    -0.51637    1.20167  -0.430  0.66911
## gore2000    -0.57245    0.91255  -0.627  0.53310
```

```
## nader2000   -0.37026    0.21043  -1.760  0.08415 .
## browne2000   0.26943    0.11448   2.354  0.02226 *
## total2000    1.80476    2.08594   0.865  0.39075
## clinton96   -0.47357    0.38239  -1.238  0.22091
## dole96       0.16585    0.42034   0.395  0.69471
## perot96      0.72757    0.25090   2.900  0.00539 **
## buchanan96p -0.27231    0.13808  -1.972  0.05373 .
## reform.reg   0.02002    0.01954   1.024  0.31025
## total.reg    0.12542    0.07794   1.609  0.11341
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3892 on 54 degrees of freedom
## Multiple R-squared:  0.9027, Adjusted R-squared:  0.8829
## F-statistic: 45.55 on 11 and 54 DF,  p-value: < 2.2e-16
```

```
## Condition
## Index    Variance Decomposition Proportions
##              intercept bush2000 gore2000 nader2000 browne2000 total2000
## 1      1.000 0.000     0.000    0.000    0.000     0.000      0.000
## 2      4.308 0.000     0.000    0.000    0.000     0.000      0.000
## 3     14.462 0.003     0.000    0.000    0.002     0.058      0.000
## 4     44.076 0.013     0.000    0.000    0.012     0.544      0.000
## 5     57.346 0.022     0.000    0.000    0.062     0.014      0.000
## 6     68.140 0.001     0.000    0.002    0.051     0.189      0.000
## 7     84.014 0.095     0.000    0.000    0.301     0.022      0.000
## 8    103.157 0.021     0.004    0.001    0.000     0.043      0.000
## 9    118.495 0.019     0.002    0.009    0.041     0.007      0.000
## 10   215.706 0.084     0.004    0.020    0.470     0.037      0.001
## 11   467.085 0.007     0.080    0.117    0.029     0.045      0.001
## 12  1820.042 0.735     0.909    0.852    0.032     0.042      0.997
##    clinton96 dole96 perot96 buchanan96p reform.reg total.reg
## 1  0.000     0.000  0.000   0.000       0.001      0.000
## 2  0.000     0.000  0.000   0.000       0.427      0.000
## 3  0.000     0.000  0.000   0.004       0.188      0.004
## 4  0.000     0.000  0.001   0.161       0.221      0.002
## 5  0.000     0.002  0.001   0.147       0.015      0.375
## 6  0.008     0.000  0.005   0.141       0.084      0.528
## 7  0.025     0.009  0.004   0.017       0.005      0.026
## 8  0.012     0.010  0.213   0.005       0.005      0.008
## 9  0.021     0.024  0.166   0.365       0.048      0.006
## 10 0.123     0.137  0.418   0.121       0.005      0.000
## 11 0.751     0.812  0.074   0.038       0.001      0.051
## 12 0.061     0.005  0.119   0.001       0.000      0.001
```

We will use stepwise model selection by AIC to find the best predictive model that increases R-squared, gives us more significant predictors, and minimizes collinearity among predictors. The stepwise process is shown below.

```
## Start:  AIC=-113.81
## buchanan2000 ~ bush2000 + gore2000 + nader2000 + browne2000 +
##     total2000 + clinton96 + dole96 + perot96 + buchanan96p +
##     reform.reg + total.reg
##
##               Df Sum of Sq    RSS     AIC
```

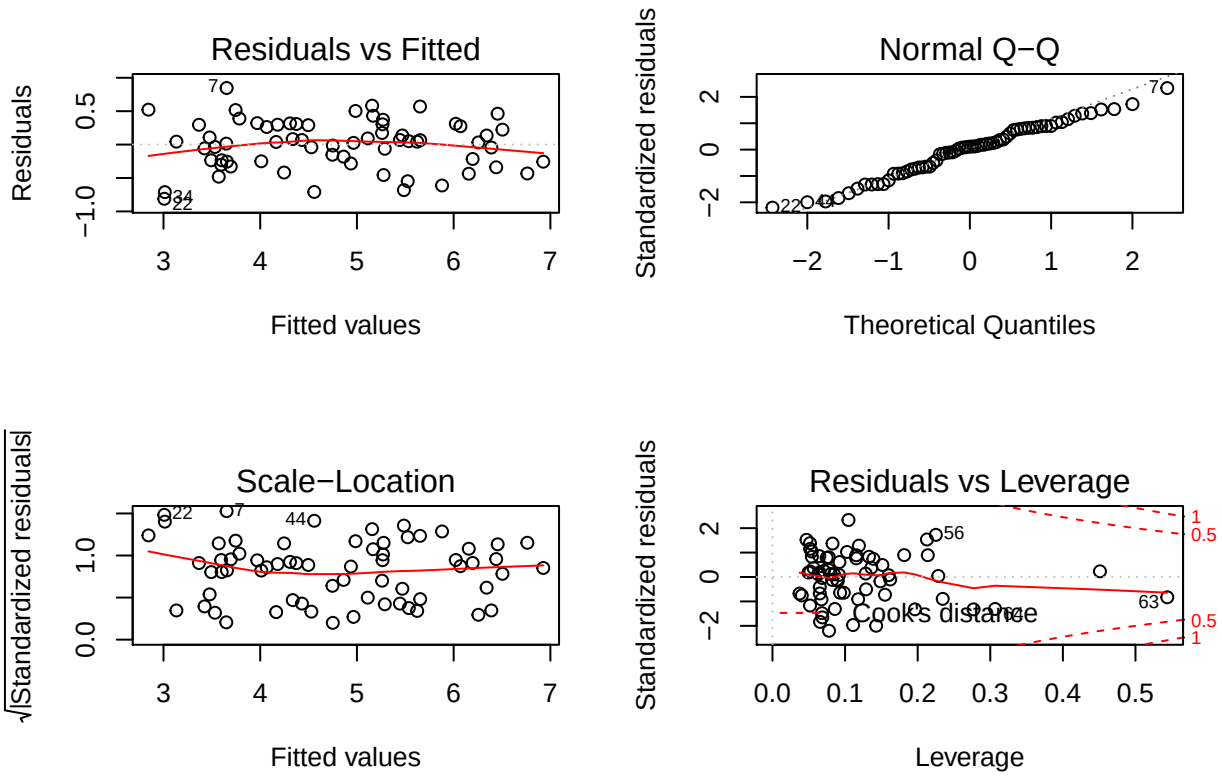
```
## - dole96       1   0.02358 8.2028 -115.62
## - bush2000     1   0.02797 8.2072 -115.59
## - gore2000     1   0.05961 8.2388 -115.33
## - total2000    1   0.11338 8.2926 -114.90
## - reform.reg   1   0.15892 8.3381 -114.54
## - clinton96    1   0.23231 8.4115 -113.96
## <none>                     8.1792 -113.81
## - total.reg    1   0.39220 8.5714 -112.72
## - nader2000    1   0.46893 8.6481 -112.13
## - buchanan96p  1   0.58907 8.7683 -111.22
## - browne2000   1   0.83907 9.0183 -109.37
## - perot96      1   1.27370 9.4529 -106.26
##
## Step:  AIC=-115.62
## buchanan2000 ~ bush2000 + gore2000 + nader2000 + browne2000 +
##     total2000 + clinton96 + perot96 + buchanan96p + reform.reg +
##     total.reg
##
##               Df Sum of Sq    RSS     AIC
## - bush2000     1   0.01952 8.2223 -117.47
## - gore2000     1   0.09428 8.2971 -116.87
## - total2000    1   0.12669 8.3295 -116.61
## - reform.reg   1   0.15089 8.3537 -116.42
## <none>                     8.2028 -115.62
## - clinton96    1   0.25378 8.4566 -115.61
## - total.reg    1   0.37086 8.5736 -114.70
## + dole96       1   0.02358 8.1792 -113.81
## - nader2000    1   0.49034 8.6931 -113.79
## - buchanan96p  1   0.59251 8.7953 -113.02
## - browne2000   1   0.82114 9.0239 -111.33
## - perot96      1   1.27181 9.4746 -108.11
##
## Step:  AIC=-117.47
## buchanan2000 ~ gore2000 + nader2000 + browne2000 + total2000 +
##     clinton96 + perot96 + buchanan96p + reform.reg + total.reg
##
##               Df Sum of Sq     RSS     AIC
## - reform.reg   1   0.15314  8.3754 -118.25
## - gore2000     1   0.19318  8.4155 -117.93
## - clinton96    1   0.24109  8.4634 -117.56
## <none>                      8.2223 -117.47
## - total.reg    1   0.36052  8.5828 -116.63
## - nader2000    1   0.47082  8.6931 -115.79
## + bush2000     1   0.01952  8.2028 -115.62
## + dole96       1   0.01513  8.2072 -115.59
## - buchanan96p  1   0.61147  8.8338 -114.73
## - browne2000   1   0.88729  9.1096 -112.70
## - perot96      1   1.34812  9.5704 -109.44
## - total2000    1   1.78673 10.0090 -106.49
##
## Step:  AIC=-118.25
## buchanan2000 ~ gore2000 + nader2000 + browne2000 + total2000 +
##     clinton96 + perot96 + buchanan96p + total.reg
##
```

```
##               Df Sum of Sq     RSS     AIC
## - gore2000     1   0.16100  8.5364 -118.99
## - clinton96    1   0.25686  8.6323 -118.25
## <none>                      8.3754 -118.25
## + reform.reg   1   0.15314  8.2223 -117.47
## - buchanan96p  1   0.46193  8.8374 -116.70
## + bush2000     1   0.02177  8.3537 -116.42
## + dole96       1   0.00865  8.3668 -116.31
## - nader2000    1   0.51957  8.8950 -116.28
## - total.reg    1   0.58084  8.9563 -115.82
## - browne2000   1   0.80962  9.1851 -114.16
## - perot96      1   1.27102  9.6465 -110.92
## - total2000    1   1.71053 10.0860 -107.98
##
## Step:  AIC=-118.99
## buchanan2000 ~ nader2000 + browne2000 + total2000 + clinton96 +
##     perot96 + buchanan96p + total.reg
##
##               Df Sum of Sq     RSS     AIC
## <none>                      8.5364 -118.99
## - buchanan96p  1   0.31244  8.8489 -118.62
## + gore2000     1   0.16100  8.3754 -118.25
## + dole96       1   0.13722  8.3992 -118.06
## + reform.reg   1   0.12096  8.4155 -117.93
## + bush2000     1   0.09403  8.4424 -117.72
## - total.reg    1   0.64363  9.1801 -116.19
## - browne2000   1   0.82930  9.3657 -114.87
## - clinton96    1   1.08738  9.6238 -113.08
## - nader2000    1   1.20657  9.7430 -112.27
## - total2000    1   1.96845 10.5049 -107.30
## - perot96      1   2.02449 10.5609 -106.94
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## buchanan2000 ~ bush2000 + gore2000 + nader2000 + browne2000 +
##     total2000 + clinton96 + dole96 + perot96 + buchanan96p +
##     reform.reg + total.reg
##
## Final Model:
## buchanan2000 ~ nader2000 + browne2000 + total2000 + clinton96 +
##     perot96 + buchanan96p + total.reg
##
##
##           Step Df   Deviance Resid. Df Resid. Dev       AIC
## 1                                   54   8.179193 -113.8120
## 2     - dole96  1 0.02358159        55   8.202774 -115.6220
## 3   - bush2000  1 0.01951568        56   8.222290 -117.4652
## 4 - reform.reg  1 0.15314425        57   8.375434 -118.2472
## 5   - gore2000  1 0.16099743        58   8.536432 -118.9906
```

The best final model predicts Buchanan from Nader2000, Browne2000, Total2000, Clinton96, Perot96, Buchanan96p, and total.reg. In this final model, all predictors included are significant except for Buchanan96p. Our correlation coefficient is slightly higher than the full model at

0.8862. There is much less evidence of collinearity among the predictors.




```
##
## Call:
## lm(formula = buchanan2000 ~ nader2000 + browne2000 + total2000 +
##     clinton96 + perot96 + buchanan96p + total.reg, data = log.remove_pb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.81111 -0.25387  0.04132  0.29480  0.84700
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.62874    1.03880  -4.456 3.87e-05 ***
## nader2000   -0.49058    0.17134  -2.863 0.005827 **
## browne2000   0.25389    0.10696   2.374 0.020942 *
## total2000    0.89874    0.24575   3.657 0.000551 ***
## clinton96   -0.52193    0.19202  -2.718 0.008644 **
## perot96      0.76847    0.20720   3.709 0.000468 ***
## buchanan96p -0.14635    0.10045  -1.457 0.150507
## total.reg    0.14925    0.07137   2.091 0.040901 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3836 on 58 degrees of freedom
## Multiple R-squared:  0.8985, Adjusted R-squared:  0.8862
## F-statistic: 73.31 on 7 and 58 DF,  p-value: < 2.2e-16
```

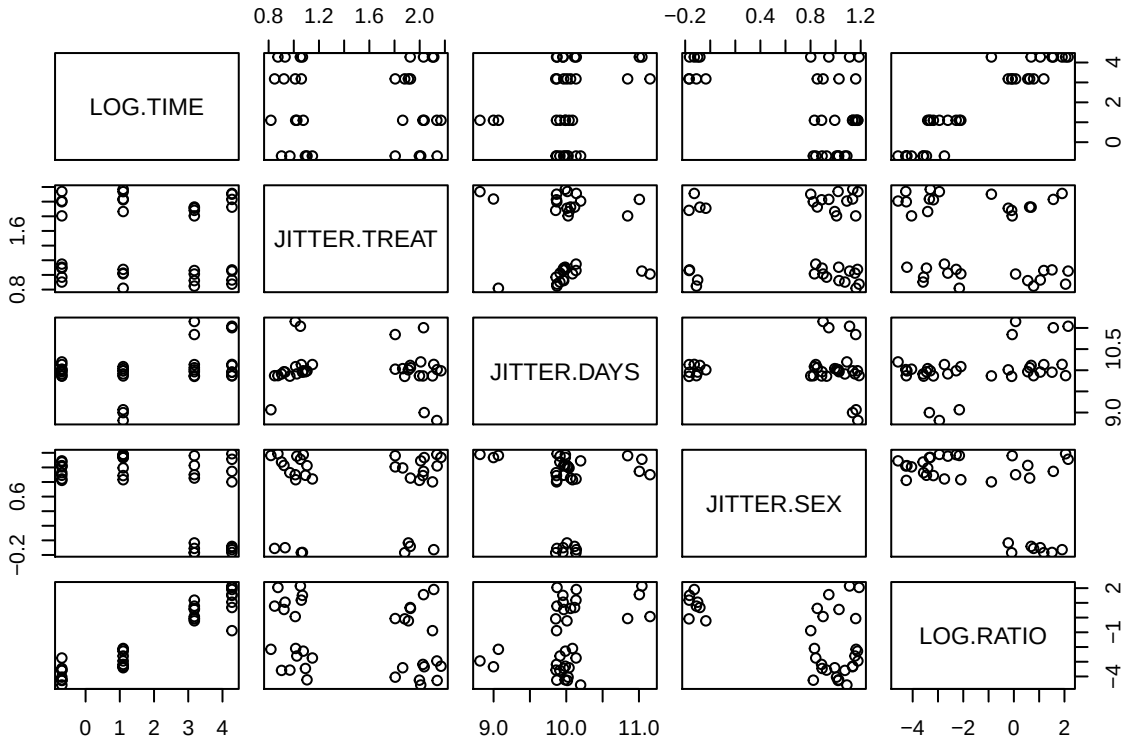
```
## Condition
## Index    Variance Decomposition Proportions
##              intercept nader2000 browne2000 total2000 clinton96 perot96
## 1    1.000 0.000     0.000     0.000       0.000     0.000     0.000
## 2   10.311 0.011     0.002     0.035       0.000     0.000     0.000
## 3   31.675 0.026     0.006     0.504       0.000     0.000     0.000
## 4   49.274 0.075     0.033     0.013       0.000     0.000     0.005
## 5   58.497 0.016     0.283     0.331       0.002     0.026     0.025
## 6   74.054 0.247     0.417     0.016       0.000     0.186     0.053
## 7  108.991 0.002     0.001     0.017       0.114     0.152     0.713
## 8  184.145 0.622     0.258     0.085       0.883     0.636     0.204
##   buchanan96p total.reg
## 1 0.000       0.000
## 2 0.007       0.003
## 3 0.199       0.030
## 4 0.351       0.623
## 5 0.372       0.290
## 6 0.017       0.003
## 7 0.017       0.039
## 8 0.037       0.013
```

Below is the predicted value and prediction interval for Buchanan votes in 2000 based on the Palm Beach data and the model generated from stepwise regression. Our prediction interval is narrower than the the previous prediction interval generated from the simple model and the predicted value is slightly lower. The mean number of votes Buchanan received in all counties - excluding Palm Beach is 210.7576. The predicted number of Buchanan votes is closer to that value and likely closer to the true number of votes Buchanan was expected to receive.

fit	lwr	upr
431.9028	164.1014	1136.736

QUESTION 3

In order to obtain a clearer picture of the spread of our data - particularly discrete and categorical data - we will create jittered versions of the treatment, sex, and days after inoculation variables. Below, we see a matrix of scatter plots among log of sacrifice time, jittered treatment group, jittered days after inoculation, jittered sex, and log of the brain tumor to liver antibody ratio.



Below is a matrix of correlation coefficients and p-values testing the nulll hypothesis that the true correlation coefficient is equal to 0 against the alternative that it is not equal to 0 for all combinations of variables presented in the above scatterplot matrix.

```
##           LOG.TIME TREAT  DAYS   SEX LOG.RATIO
## LOG.TIME      1.00  0.03  0.33 -0.54      0.94
## TREAT         0.03  1.00 -0.06  0.00     -0.16
## DAYS          0.33 -0.06  1.00  0.04      0.40
## SEX          -0.54  0.00  0.04  1.00     -0.56
## LOG.RATIO     0.94 -0.16  0.40 -0.56      1.00
##
## n= 34
##
##
## P
##           LOG.TIME TREAT  DAYS   SEX    LOG.RATIO
## LOG.TIME           0.8771 0.0558 0.0010 0.0000
## TREAT     0.8771          0.7151 1.0000 0.3516
## DAYS      0.0558   0.7151        0.8397 0.0201
## SEX       0.0010   1.0000 0.8397        0.0006
## LOG.RATIO 0.0000   0.3516 0.0201 0.0006
```

When observing scatterplots and the correlations among log sacrifice time, treatment groups, days after inoculation, sex, and log of brain tumor to antibody ratio there are some trends to be observed. Based on the correlation coefficients, the strongest correlations is between log time and log ratio with a correlation coefficient = 0.94. Based on the p-values, we can be 95% confident that the true correlation coefficient between the following pairs variables

are not equal to 0 because their p-values are less than 0.05: sex and log time, log time and log ratio, days and log ratio, sex and log ratio. The significance of the correlation coefficients between the log transformed variables and many of the other variables indicates that this transformation is appropriate and helps to normalize our data in order to obtain the best linear model.

We will now fit a regression model predicting the log of the ratio on treatment group and sacrifice time, both as factor variables. The results of the model are presented below. The model appears to be quite successful at predicting the log ratio and has a strong linear fit. All predictors in this model are significant and the r-squared value is 0.9438.

```
##
## Call:
## lm(formula = LOG.RATIO ~ TREAT + FACTOR.TIME, data = bloodbrain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.74019 -0.17548 -0.01782  0.24772  1.05512
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -3.5049     0.1954 -17.937  < 2e-16 ***
## TREAT2         -0.7968     0.1834  -4.346 0.000155 ***
## FACTOR.TIME3    1.1341     0.2520   4.501 0.000101 ***
## FACTOR.TIME24   4.2573     0.2591  16.431 3.13e-16 ***
## FACTOR.TIME72   5.1539     0.2591  19.892  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5328 on 29 degrees of freedom
## Multiple R-squared:  0.9506, Adjusted R-squared:  0.9438
## F-statistic: 139.6 on 4 and 29 DF,  p-value: < 2.2e-16
```

The predicted values of the log brain tumor to liver antibody ratio and the transformed ratio for every combination of treatment and time are presented in the table below.

TREAT	FACTOR.TIME	LOG.ESTIMATED.MEAN	ESTIMATED.MEAN
1	0.5	-3.5048856	0.0300502
1	3	-2.3707373	0.0934118
1	24	0.7523685	2.1220201
1	72	1.6489790	5.2016662
2	0.5	-4.3016813	0.0135458
2	3	-3.1675330	0.0421073
2	24	-0.0444271	0.9565453
2	72	0.8521833	2.3447607

We will now fit a linear model predicting the log of the ratio on treatment as a factor variable, x = log of sacrifice time, x squared, and x-cubed. The results of the model are presented below. All predictors are significant and the r-squared is 0.9438 - the same as the model above.

```
##
## Call:
## lm(formula = LOG.RATIO ~ TREAT + LOG.TIME + LOG.TIME2 + LOG.TIME3,
##     data = bloodbrain)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.74019 -0.17548 -0.01782  0.24772  1.05512
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.45144    0.20120 -17.154  < 2e-16 ***
## TREAT2      -0.79680    0.18335  -4.346 0.000155 ***
## LOG.TIME     0.49528    0.17090   2.898 0.007081 **
## LOG.TIME2    0.54189    0.15413   3.516 0.001463 **
## LOG.TIME3   -0.08858    0.02871  -3.085 0.004442 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5328 on 29 degrees of freedom
## Multiple R-squared:  0.9506, Adjusted R-squared:  0.9438
## F-statistic: 139.6 on 4 and 29 DF,  p-value: < 2.2e-16
```

Below are the predicted values of log brain tumor to liver antibody ratio and the transformed ratio for every combination of treatment and log time. These predicted values are the same as the one generated from the model above. Both models produe the same results because each model includes three terms for the for level factor variable time which provides enough information to determine which combination of treatment and time an observation falls into. In this model log time and the 2 tranformations of log time act the same way as the factor variable in the previous model, but on a different scale.

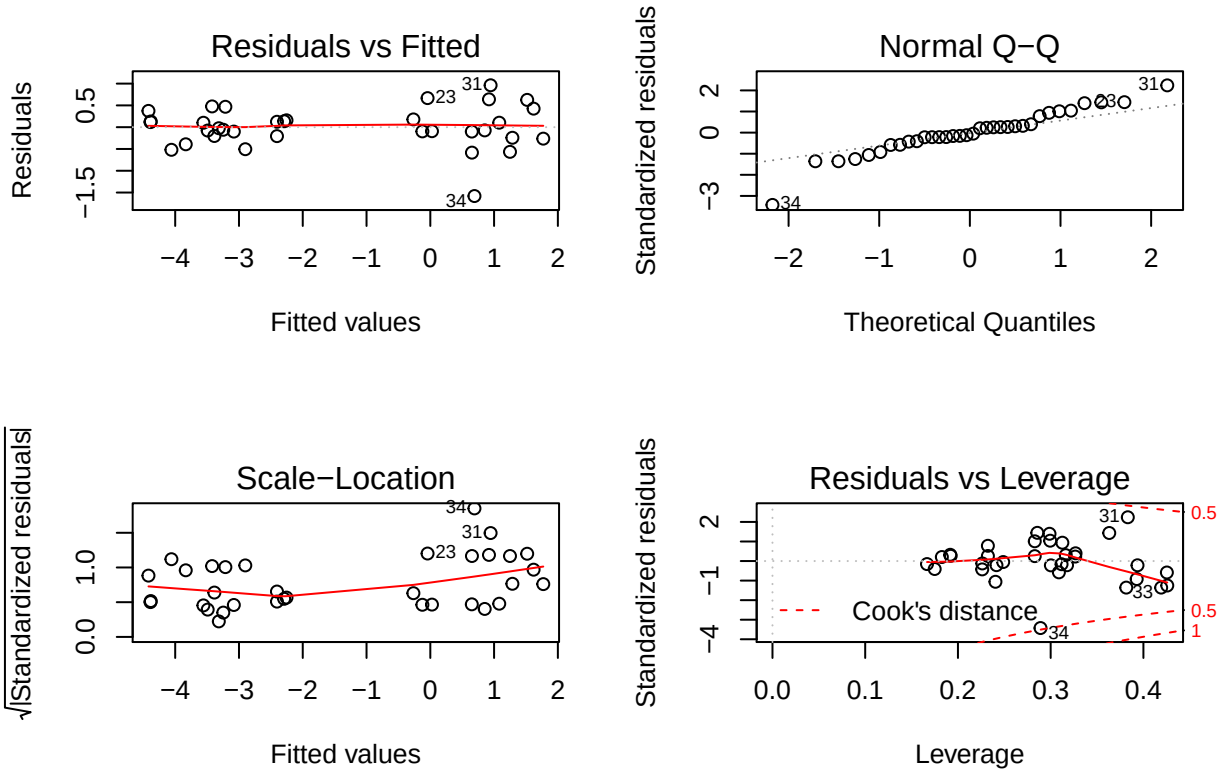
TREAT	LOG.TIME	LOG.TIME2	LOG.TIME3	LOG.ESTIMATED.MEAN	ESTIMATED.MEAN
1	-0.6931472	0.480453	-0.3330247	-3.5048856	0.0300502
1	1.0986123	1.206949	1.3259690	-2.3707373	0.0934118
1	3.1780538	10.100026	32.0984268	0.7523685	2.1220201
1	4.2766661	18.289873	78.2196806	1.6489790	5.2016662
2	-0.6931472	0.480453	-0.3330247	-4.3016813	0.0135458
2	1.0986123	1.206949	1.3259690	-3.1675330	0.0421073
2	3.1780538	10.100026	32.0984268	-0.0444271	0.9565453
2	4.2766661	18.289873	78.2196806	0.8521833	2.3447607

Below is a model predicting the log brain tumor to liver antibody ratio on all covariates. The r-squared is 0.9408 indicating a strong linear fit; however, only the treatment and time predictors are significant in this model.

```
##
## Call:
## lm(formula = LOG.RATIO ~ DAYS + SEX + WEIGHT + LOSS + TUMOR +
##     TREAT + FACTOR.TIME, data = bloodbrain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.58034 -0.20482 -0.04134  0.17296  0.96182
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -4.099278   3.110002  -1.318 0.199915
## DAYS           0.019350   0.282007   0.069 0.945864
```

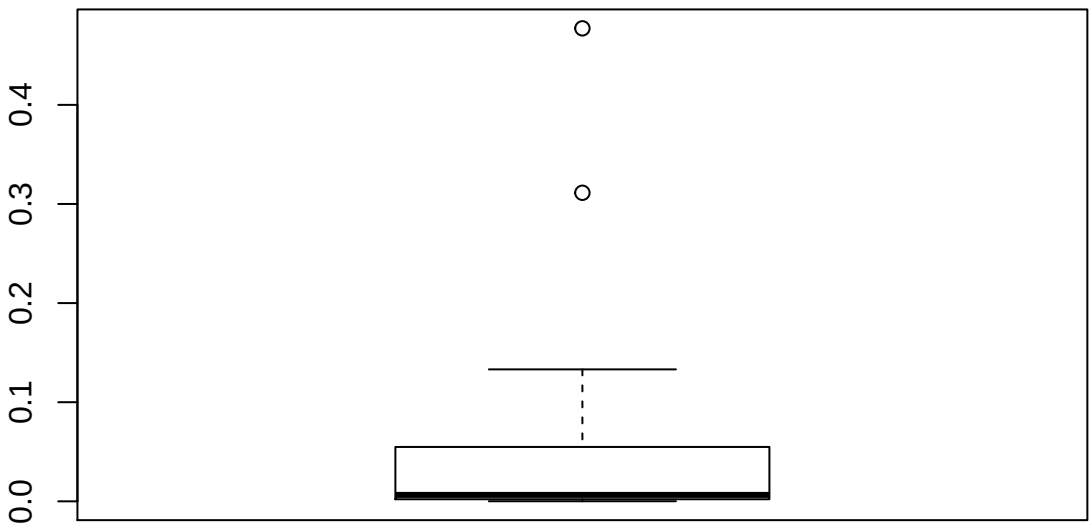
```
## SEX            0.035751   0.357884   0.100 0.921257
## WEIGHT         0.001502   0.004740   0.317 0.754079
## LOSS          -0.048216   0.027653  -1.744 0.094032 .
## TUMOR          0.001379   0.001160   1.189 0.246065
## TREAT2        -0.830930   0.197544  -4.206 0.000312 ***
## FACTOR.TIME3   1.089380   0.294004   3.705 0.001106 **
## FACTOR.TIME24  4.113695   0.337234  12.198 8.90e-12 ***
## FACTOR.TIME72  5.136627   0.340967  15.065 9.88e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5471 on 24 degrees of freedom
## Multiple R-squared:  0.9569, Adjusted R-squared:  0.9408
## F-statistic: 59.26 on 9 and 24 DF,  p-value: 3.287e-14
```

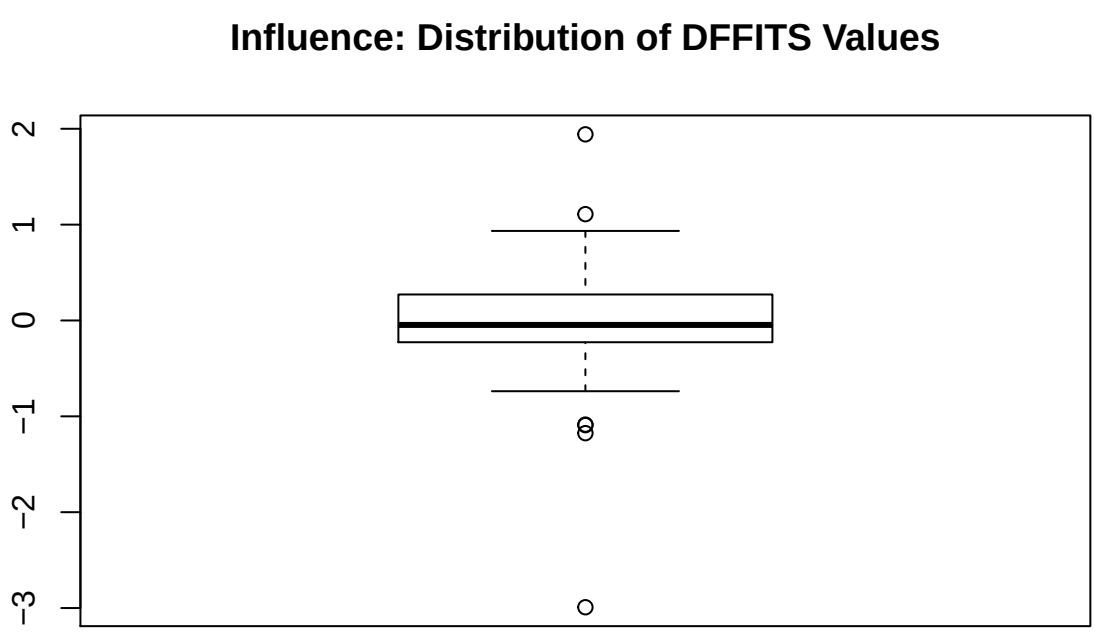
Observing the Q-Q plot suggests that this model does not satisfy the normality assumption.



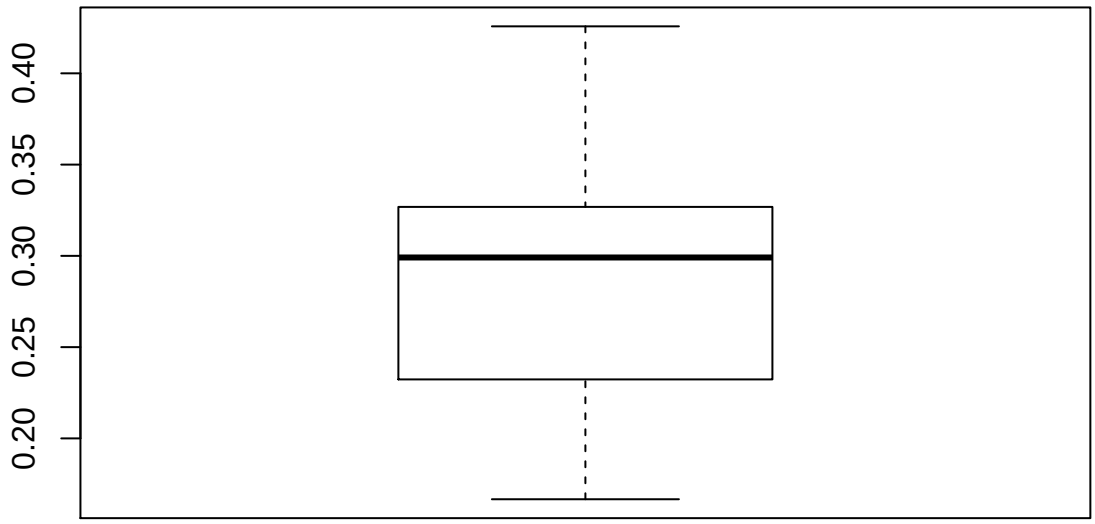
We will now assess measures of leverage and influence in this model.

We look at the distribution of Cook's distance values to assess the influence of points in the data set. Most of the cook's values are quite small and close in relation to each other. Looking at the boxplot, we can see that there are two points that may be influential as they are larger than the rest of the data. These two points with cook's equal to 0.4777 and 0.311 may be considered influential. To further assess influence, we look at the difference in fitted values standardized. As a general rule, large dffits values are those larger than 2 x squareroot((p+1)/n) - based on our model this value is 1.49. Our DFFITS values range from -2.99 to 1.94. The largest and smallest DFFITS values correspond to the two points that have large cook's d values. This is enough to consider these points - observations 31 and 34 - influential.

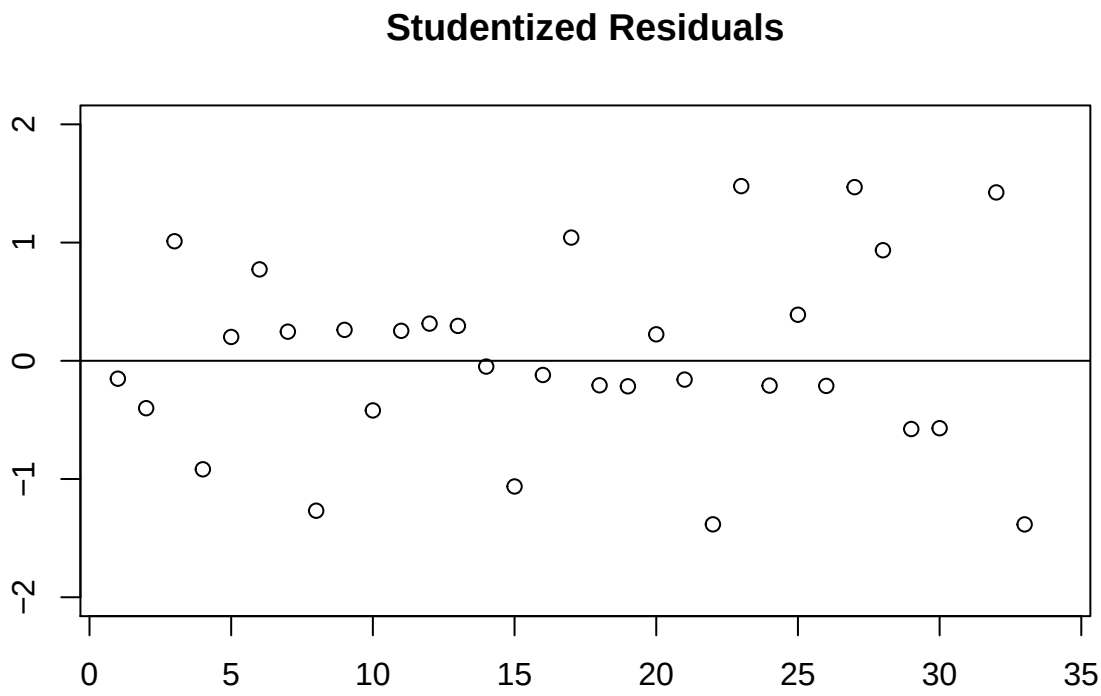

Influence: Distribution of Cook's D Values
0.0
0.1
0.2
0.3
0.4



To assess leverage in our model, we will look at the diagonal elements of the hat matrix. As a general rule, large hat values are those greater than (2 x number of predictors) / sample size. In our model, 2p/n = .529. None of the hat values are larger than this. Also, when looking at the boxplot of hat values, we can see that none of the values are extremely different from the others. There is not evidence of leverage in this model.



Finally we will look at the studentized residuals. The residuals appear to be evenly and randomly distributed around 0, indicating that none of our points are outliers.



QUESTION 5

```r
#y in an rXm matrix of Y values
#x is a vector of X values

library(stats4)
#library(dplyr)

set.seed(100)

lmq <-function(x,y){
  LL <- function(beta0,beta1,sigma,q){
    beta= t(c(beta0,beta1))
    resid = y - beta0 - beta1*x
    R = dnorm(resid, mean = 0, sd=sqrt(x^(q*2)*sigma^2), log=TRUE)
    -sum(R)
  }

  mle(LL, list(beta0=2, beta1=1, sigma=1, q=1), method = "L-BFGS-B")
}

x = rep(seq(1,3,by=0.5),2000)
y = 1 + 2*x + rnorm(length(x),0,1*x^1)
```

```r
lmq(x,y)
```

```
##
## Call:
## mle(minuslogl = LL, start = list(beta0 = 2, beta1 = 1, sigma = 1,
##     q = 1), method = "L-BFGS-B")
##
## Coefficients:
##     beta0     beta1     sigma         q
## 0.9788321 2.0161733 0.9945119 1.0002124
```