Exercise 1: Refer to problem 8.5 for a short explanation of the BigMac2003 data set. Then do the following,

1. Check the five diagnostics on model assumptions in the linear model that includes Bigmac as the response and all nin predictors. Based on these diagnostics, which model assumption do not appear to be valid.

   Solution:
   First we can check for non linearity in the model, plotting the residuals and performing Tukey's test we get the following,

Figure 1: Residual Plot for Each Predictor

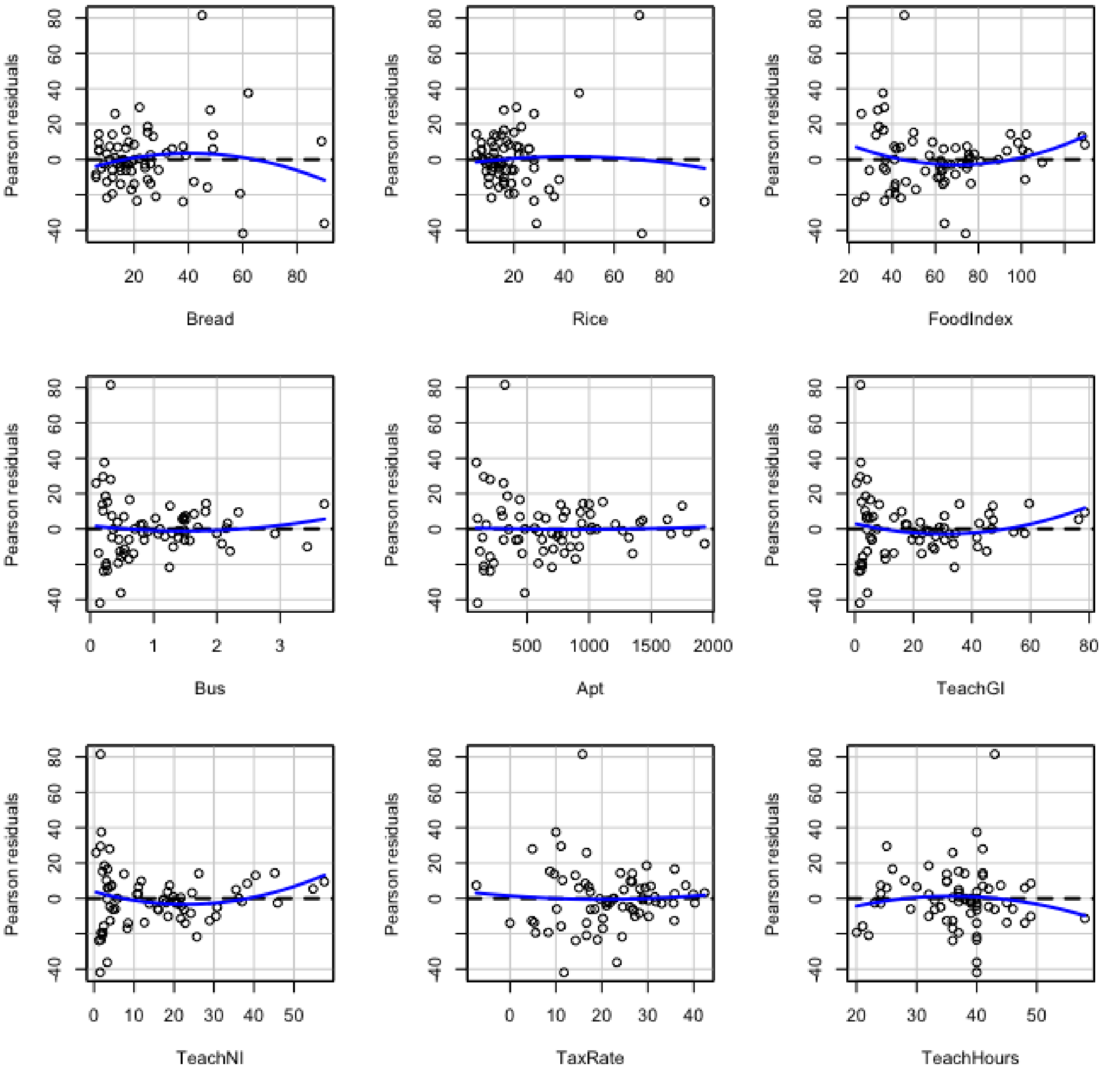
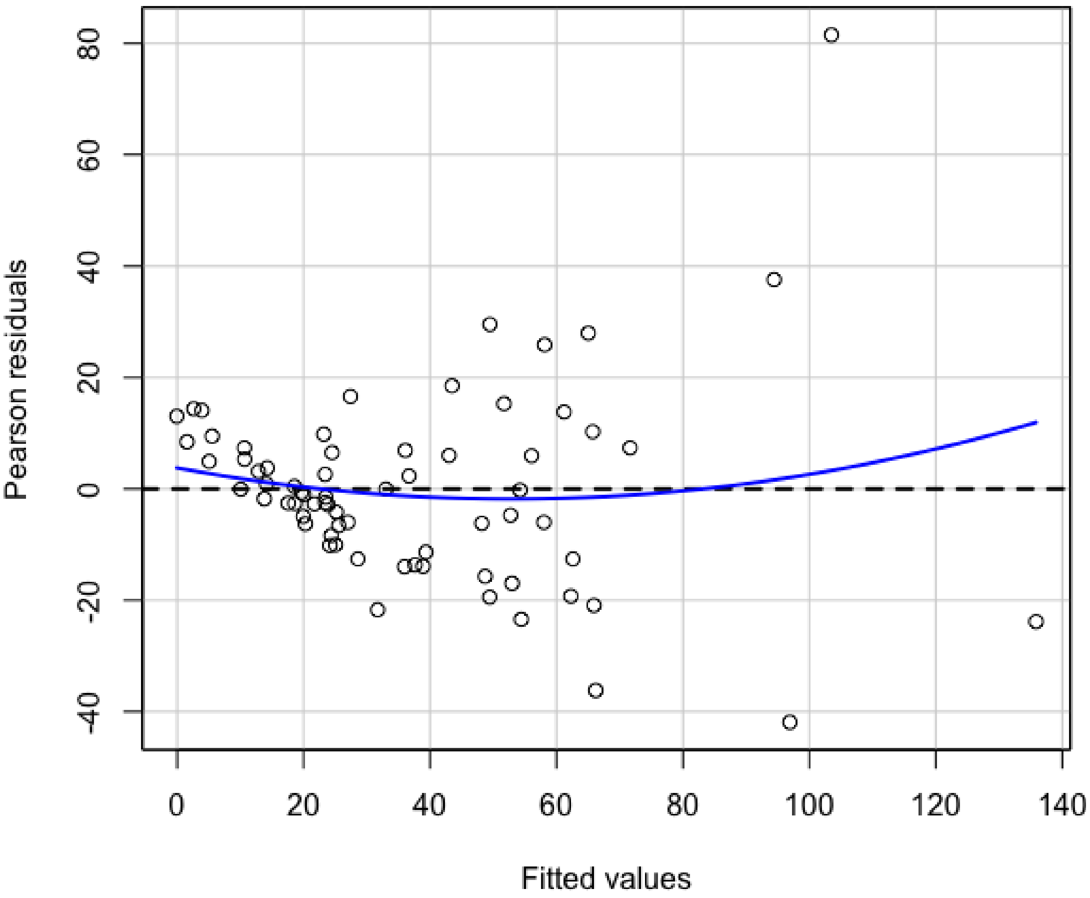


Figure 2: Residual Plot for Entire Model



Code:

```
> df <- BigMac2003
> View(df)
> model <- lm(BigMac ~. , data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-41.916 -10.053  -1.024   7.359  81.512

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 40.359743  16.884721   2.390   0.0200 *
```

```
Bread        0.387238   0.183319   2.112   0.0389 *
Rice         0.965387   0.182620   5.286  1.9e-06 ***
FoodIndex   -0.512792   0.194416  -2.638   0.0107 *
Bus         -0.229961   4.533740  -0.051   0.9597
Apt          0.003929   0.007795   0.504   0.6161
TeachGI      1.848863   1.363304   1.356   0.1802
TeachNI     -2.287929   1.830213  -1.250   0.2162
TaxRate     -0.775878   0.397161  -1.954   0.0555 .
TeachHours   0.295898   0.335860   0.881   0.3819
-------------------------------------------------------------
Residual standard error: 18.73 on 59 degrees of freedom
Multiple R-squared:  0.6918,    Adjusted R-squared:  0.6448
F-statistic: 14.71 on 9 and 59 DF,  p-value: 3.744e-12


> residualPlots(model)
           Test stat Pr(>|Test stat|)
Bread        -1.5537          0.12569
Rice         -0.5239          0.60232
FoodIndex     1.7418          0.08685 .
Bus           0.6758          0.50184
Apt           0.1695          0.86598
TeachGI       1.6793          0.09848 .
TeachNI       1.8961          0.06293 .
TaxRate       0.3386          0.73611
TeachHours   -0.8711          0.38727
Tukey test    1.5634          0.11797
-------------------------------------------------------------
```

Generally the residual plot show a low degree of curvature in the fitted second order model. As suggested by the results of Tukey's test I would say that this model achieves a high degree of linearity.
Since we already have the residual plots it would be appropriate to test for non-constant variance. Looking at the fitted value residuals we can see a wedge pattern in the residuals, this suggests a very strong degree of non constant variance. We can also see this in almost all of the predictors. Testing this with the Breusch-Pagan test we can see that, with a p-value on the order of 10^{16} we cannot assume constant variance in our error.
Code:

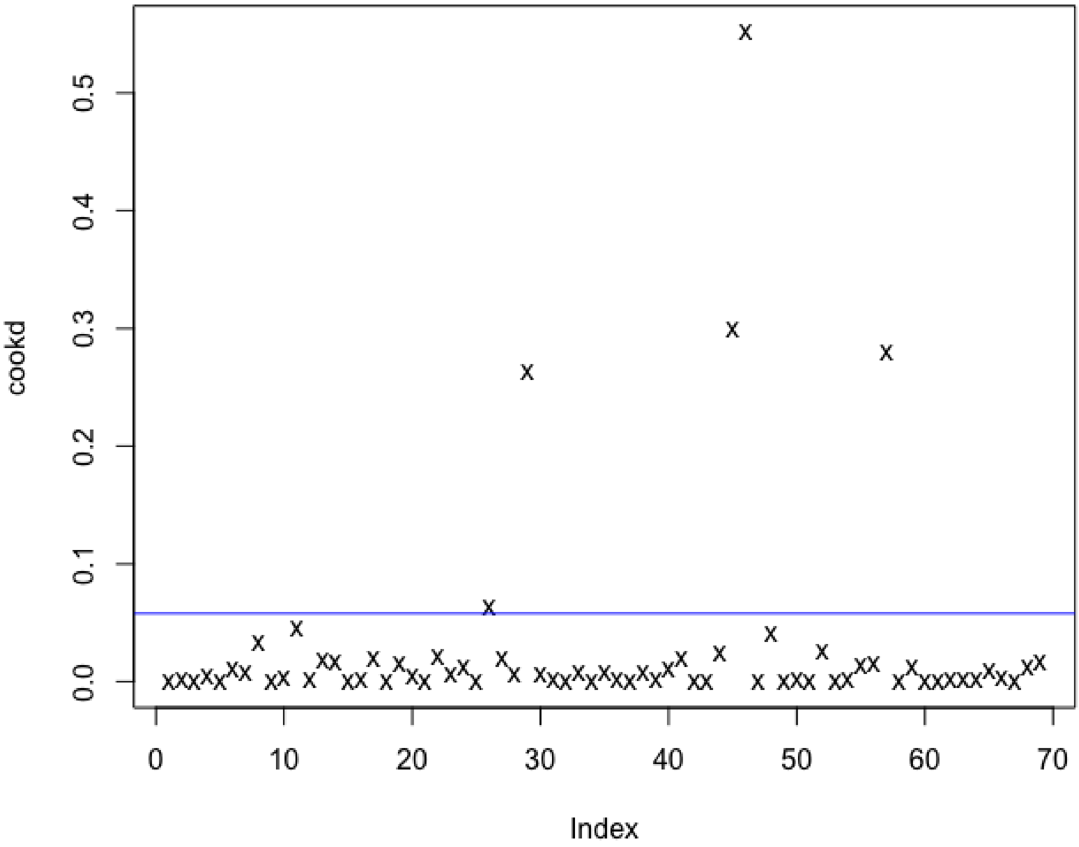
```
> ncvTest(model)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 74.03771, Df = 1, p = < 2.22e-16
```

Looking at the residual plot again it does seem possible there could be a couple of outliers, checking the residuals using the outlierTest() function we get, that there is at least one outlier, which would warrant further analysis.
Code:

```
> outlierTest(model)
        rstudent unadjusted p-value Bonferroni p
Nairobi  6.17679         6.9534e-08   4.7978e-06
```

In the same vein we can test for influence points using Cooks Distance. Using the $4/n$ criteria there are 5 influential places in the data set which we attained using cooks.distance() function,

Figure 3: Residual Plot Cooks Distance



Code:

```
> cookd <- cooks.distance(model)
> plot(cookd, pch = 'x')
> abline(h = 4/length(cookd), col = 'blue')
> # Influential observations with 4/n criteria
> Influential_index <- (names(cookd)[(cookd > (4/length(cookd)))])
> Influential_index
[1] "Karachi" "Lagos" "Mumbai" "Nairobi" "Shanghi"
```

Checking for autocorrelation in our residuals, we can use the Durbin-Watson test. Doing so we get a p-value of .98 which means on the $\alpha = .05$ level our residuals do not show autocorrelation.
Code:

```
> dwt(model)
 lag Autocorrelation D-W Statistic p-value
   1    -0.006492702      1.999359    0.98
 Alternative hypothesis: rho != 0
```

Finally using the Shapiro-Wilk test we can test for normality among our residuals. Doing so we get a p-value on the order of 10^{-5} so our residuals are not normally distributed. **Code:**

```
> shapiro.test(residuals(model))

        Shapiro-Wilk normality test

data:  residuals(model)
W = 0.90482, p-value = 6.78e-05
```

In general this model fails a majority of our diagnostics, specifically the model has several leverage points and outliers, demonstrates non-constant variance, and non-normally distributed residuals. I would not trust the results of this model.

b. Find the Box-Cox tranformation for the response variable and make theneares ladder-of-powers transformation to it. Recheck the five diagnostics with the transformed response. What is the transformation and which assumptions still appear to be invalid?

Solution:
We can find the nearest ladder-of-powers transformation using the powerTransformation() function. Doing so we get that the response should be raised to the power of $-.5$. Performing the diagnostics again we get that the model now has constant variance in the residuals, normally distributed residuals, and smaller outliers. There are still leverage points under the $4/n$ criteria, but the model is significantly more trustworthy.
Code:

```
---------------------------------------------------------------
## Finding Box-Cox Transformation
> summary(powerTransform(model))
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1   -0.3117         -0.5       -0.5657      -0.0577
---------------------------------------------------------------

> Transformed_model<-lm(BigMac^(-.5)~., data = df)
> summary(Transformed_model)
Residuals:
      Min        1Q     Median        3Q       Max
-0.056716 -0.021262 -0.000239  0.016476  0.095504

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.287e-01  2.675e-02   4.810 1.08e-05 ***
Bread       -4.722e-04  2.904e-04  -1.626  0.10927
Rice        -8.220e-04  2.893e-04  -2.841  0.00616 **
FoodIndex    5.846e-04  3.080e-04   1.898  0.06260 .
Bus          3.580e-03  7.183e-03   0.498  0.62007
Apt          7.107e-06  1.235e-05   0.575  0.56714
TeachGI     -2.274e-03  2.160e-03  -1.053  0.29680
TeachNI      4.554e-03  2.900e-03   1.571  0.12161
TaxRate      1.198e-03  6.292e-04   1.904  0.06179 .
TeachHours   4.111e-05  5.321e-04   0.077  0.93868

Residual standard error: 0.02967 on 59 degrees of freedom
Multiple R-squared:  0.8009,    Adjusted R-squared:  0.7705
F-statistic: 26.37 on 9 and 59 DF,  p-value: < 2.2e-16
```

```
-----------------------------------------------------------
## Testing Linearity
## Passes
> residualPlots(Transformed_model)
           Test stat Pr(>|Test stat|)
Bread         1.5873           0.1179
Rice          1.6303           0.1085
FoodIndex    -0.9765           0.3329
Bus          -1.5038           0.1381
Apt          -0.3876           0.6998
TeachGI      -4.8664        9.095e-06 ***
TeachNI      -4.7459        1.399e-05 ***
TaxRate      -0.5404           0.5910
TeachHours    0.0509           0.9596
Tukey test   -1.9126           0.0558 .
-----------------------------------------------------------
## Testing Constant Variance
## Passes
> ncvTest(Transformed_model)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 2.31749, Df = 1, p = 0.12793
-----------------------------------------------------------
## Checking For Outliers
## Doesn’t Pass but larger p-value this time
> outlierTest(Transformed_model)
      rstudent unadjusted p-value Bonferroni p
Miami  3.77982         0.00037274     0.025719
-----------------------------------------------------------
## Checking For Leverage Points
## Still has Leverage Points
> cookdTransformed <- cooks.distance(Transformed_model)
> plot(cookdTransformed, pch = ’x’)
> abline(h = 4/length(cookdTransformed), col = ’blue’)
> Influential_index <- (names(cookdTransformed)
                      [(cookdTransformed > (4/length(cookdTransforme
> Influential_index
[1] ”Basel”   ”Geneva”  ”Miami”   ”Mumbai”  ”Oslo”    ”Shanghi”
-----------------------------------------------------------
## Testing AutoCorrelation
## Passes, No AutoCorrelation
> dwt(Transformed_model)
 lag Autocorrelation D-W Statistic p-value
   1     -0.07018174      2.117556    0.63
 Alternative hypothesis: rho != 0
```

```
-----------------------------------------------------------------
## Testing Normality
## Passes, Residuals appear normal.
> shapiro.test(residuals(Transformed_model))

        Shapiro-Wilk normality test

data:  residuals(Transformed_model)
W = 0.97451, p-value = 0.1702
-----------------------------------------------------------------
```

Figure 4: Residual Plot for Transformed Model

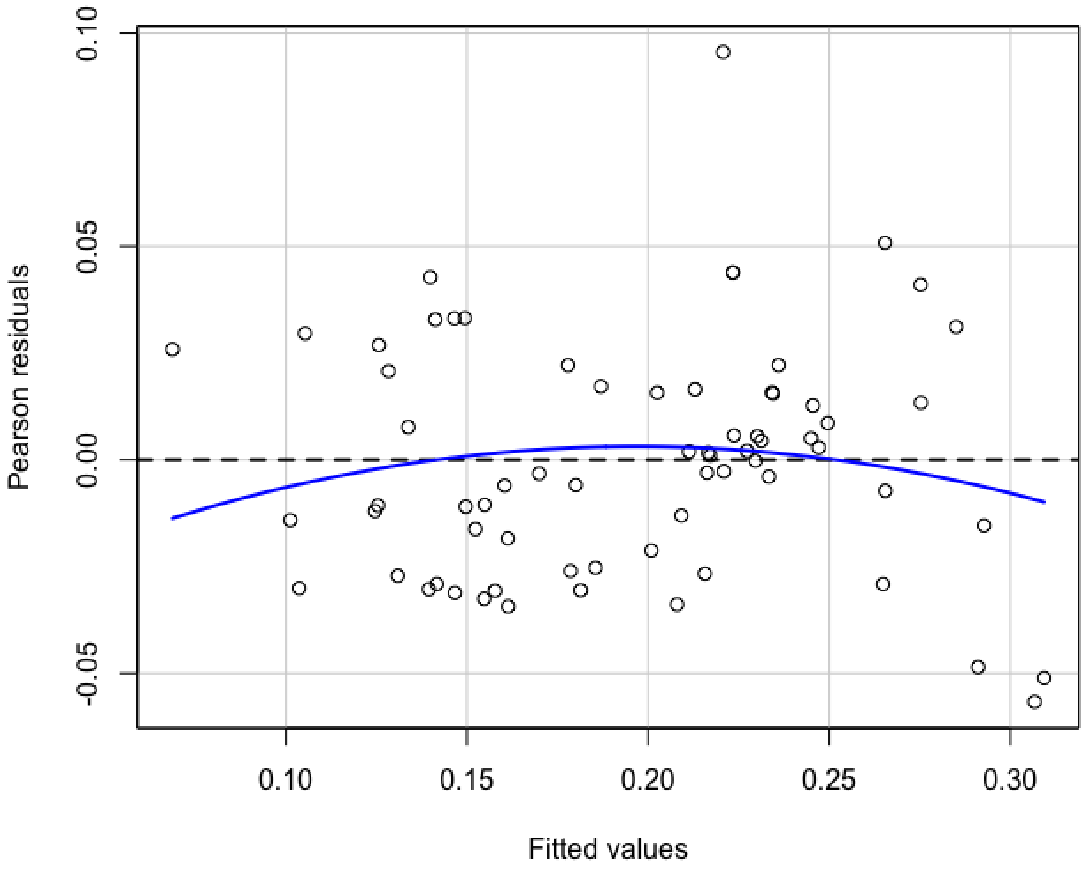
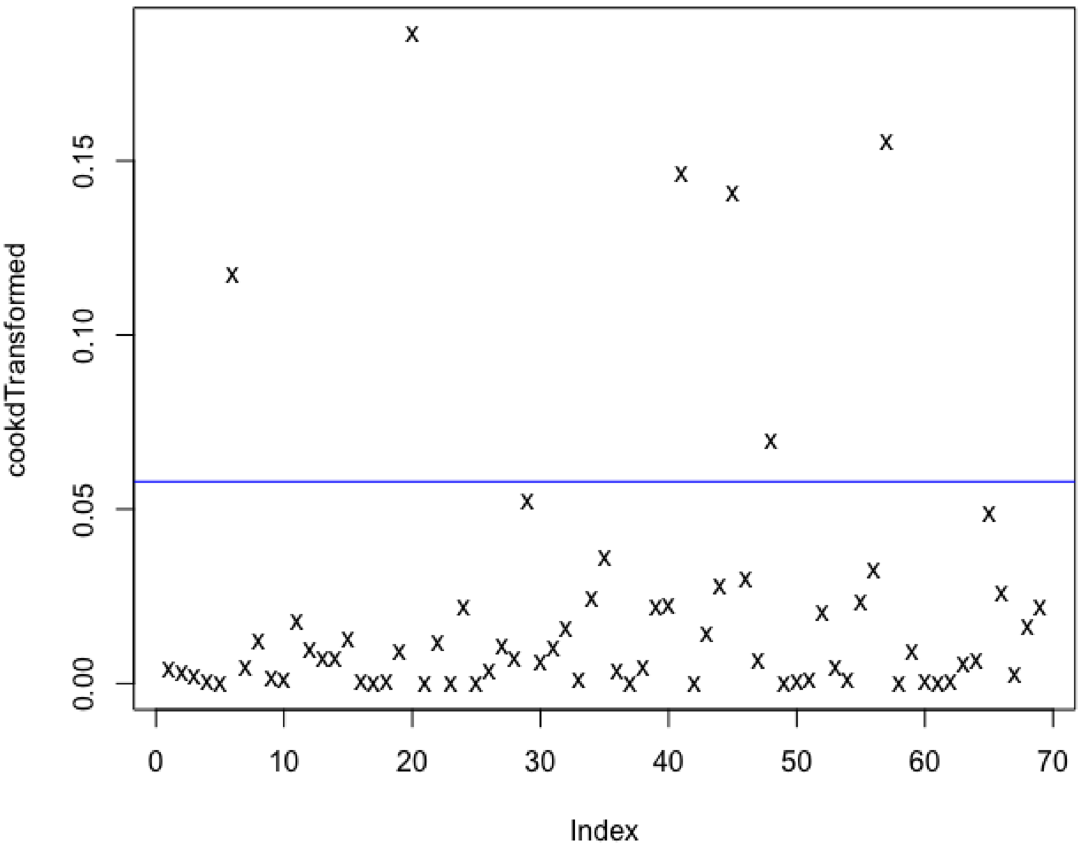


Figure 5: Residual Plot Cooks Distance



c. Use the generalized Box-Cox tranformaiton for the predictors and find transformation from the ladder of powers to make the transformed predictors as close to linearly related as possible. Recheck the five diagnostics. what is the transformation for each of the nine predictors and which assumptions still appear to be invalid.

Solution:
First we need to do as the hint suggests and smear the data that this is less than or equal to zero to use the powerTransform() function. Doing so we get that we should log transform Bread, Rice, FoodIndex, Bus, TeachGI and TeachNI. We should also square root Apt and raise TaxRate to the 1.1 power. Running diagnostics on the fitted model, we see that it performs significantly worse than when we just transformed the response. The model show significant non-linear, non-normal residuals with non-constant variance. It also shows multiple leverage points and even greater outliers. The model does appear to show no autocorrelation. This model is not trustworthy.
Code:

```
-------------------------------------------------------
## Changing data to use powerTransform()
> df<- BigMac2003
> for(i in 1:ncol(df)){
+     df[,i] = ifelse(df[,i]<=0, .01,df[,i])
+ }
-------------------------------------------------------
## Finding Box-Cox Transformation For Predictors
> summary(powerTransform(cbind(Bread,Rice, FoodIndex, Bus,
+                              Apt, TeachGI, TeachNI,TaxRate,
+                              TeachHours) ~ 1, data = df ))
bcPower Transformations to Multinormality
           Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Bread        -0.1078         0.0      -0.3897       0.1740
Rice         -0.2062         0.0      -0.4649       0.0525
FoodIndex    -0.0023         0.0      -0.4257       0.4211
Bus           0.1180         0.0      -0.0829       0.3189
Apt           0.3668         0.5       0.1234       0.6102
TeachGI      -0.0023         0.0      -0.0244       0.0198
TeachNI      -0.0024         0.0      -0.0263       0.0214
TaxRate       1.1025         1.1       1.0481       1.1569
TeachHours    1.4190         1.0       0.5312       2.3069
-------------------------------------------------------
## Settiing up Transformed Model
> model_Pred_Transform <- lm(BigMac ~ log(Bread) +
                                  log(Rice) + log(FoodIndex) +
                                  log(Bus) + sqrt(Apt) +
                                  log(TeachGI) + log(TeachNI) +
                            I(TaxRate^(1.1)) + TeachHours,
                            data = df)
> summary(model_Pred_Transform)
Call:
Residuals:
    Min      1Q  Median      3Q     Max
-35.169  -9.168  -3.041   5.440  98.832

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       48.3088    41.8111   1.155    0.253
log(Bread)         8.8089     5.8101   1.516    0.135
log(Rice)         15.2424     6.6757   2.283    0.026 *
log(FoodIndex)   -19.8378    12.3561  -1.606    0.114
log(Bus)           1.2699     5.6740   0.224    0.824
sqrt(Apt)          0.2240     0.4695   0.477    0.635
log(TeachGI)     -40.3002   278.9727  -0.144    0.886
log(TeachNI)      32.3130   278.6002   0.116    0.908
I(TaxRate^(1.1))   0.2254     2.5051   0.090    0.929
TeachHours         0.4367     0.3609   1.210    0.231

Residual standard error: 20.31 on 59 degrees of freedom
Multiple R-squared:  0.6374,    Adjusted R-squared:  0.5821
F-statistic: 11.52 on 9 and 59 DF,  p-value: 3.449e-10

------------------------------------------------------
## Testing Linearity
## Fails
> residualPlots(model_Pred_Transform)
                 Test stat Pr(>|Test stat|)
log(Bread)          1.7908        0.0785428 .
log(Rice)           3.7712        0.0003833 ***
log(FoodIndex)      0.4511        0.6536311
log(Bus)            1.5751        0.1206813
sqrt(Apt)           0.3206        0.7496302
log(TeachGI)        2.8610        0.0058638 **
log(TeachNI)        2.5545        0.0132827 *
I(TaxRate^(1.1))    0.8455        0.4013336
TeachHours         -1.6091        0.1130152
Tukey test          5.1710        2.329e-07 ***
------------------------------------------------------
## Testing Constant Variance
## Fails
> ncvTest(model_Pred_Transform)
Non-constant Variance Score Test
```

```
Variance formula: ~ fitted.values
Chisquare = 44.31609, Df = 1, p = 2.7941e-11
--------------------------------------------------
## Checking For Outliers
## Doesn't Pass and has smaller p-value this time
> outlierTest(model_Pred_Transform)
        rstudent unadjusted p-value Bonferroni p
Nairobi 7.126503         1.7969e-09   1.2399e-07
--------------------------------------------------
## Checking For Leverage Points
## Still has Leverage Points
> cookdPredTransformed <- cooks.distance(model_Pred_Transform)
> plot(cookdPredTransformed, pch = 'x')
> abline(h = 4/length(cookdPredTransformed), col = 'blue')
> Influential_index <- (names(cookdPredTransformed)
+                         [(cookdPredTransformed > (4/length(cookdPredT
> Influential_index
[1] "Karachi" "Lagos"   "Lima"    "Nairobi" "Shanghi"
--------------------------------------------------
## Testing AutoCorrelation
## Passes, No AutoCorrelation
> dwt(model_Pred_Transform)
 lag Autocorrelation D-W Statistic p-value
   1       0.1115053      1.775927   0.354
 Alternative hypothesis: rho != 0

--------------------------------------------------
## Testing Normality
## Fails, Residuals do not appear normal.
> shapiro.test(residuals(model_Pred_Transform))

        Shapiro-Wilk normality test

data:  residuals(model_Pred_Transform)
W = 0.83207, p-value = 2.213e-07
--------------------------------------------------
```

Figure 6: Residual Plot for Each Transformed Predictor

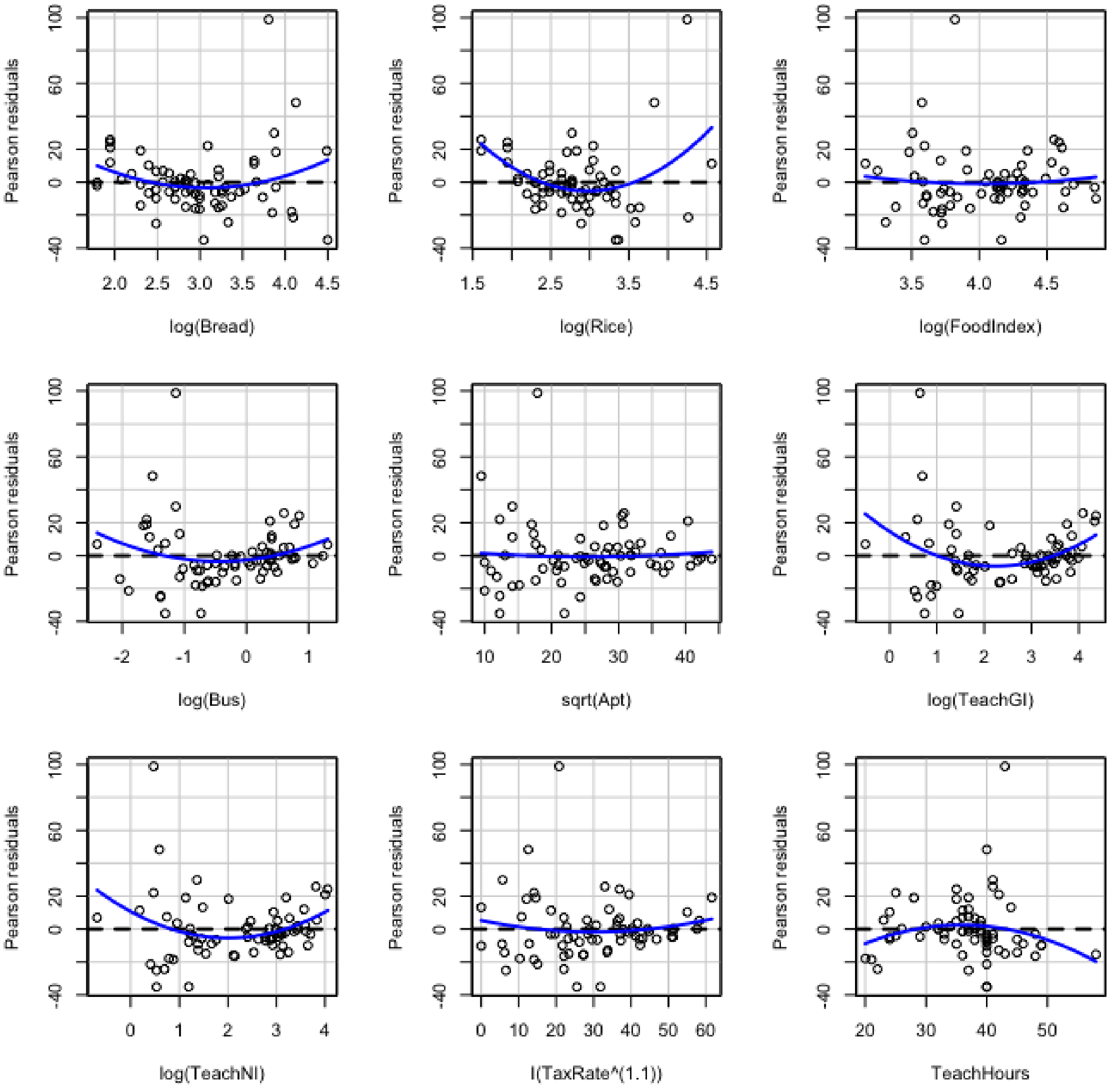


Figure 7: Residual Plot for Transformed Predictor Model

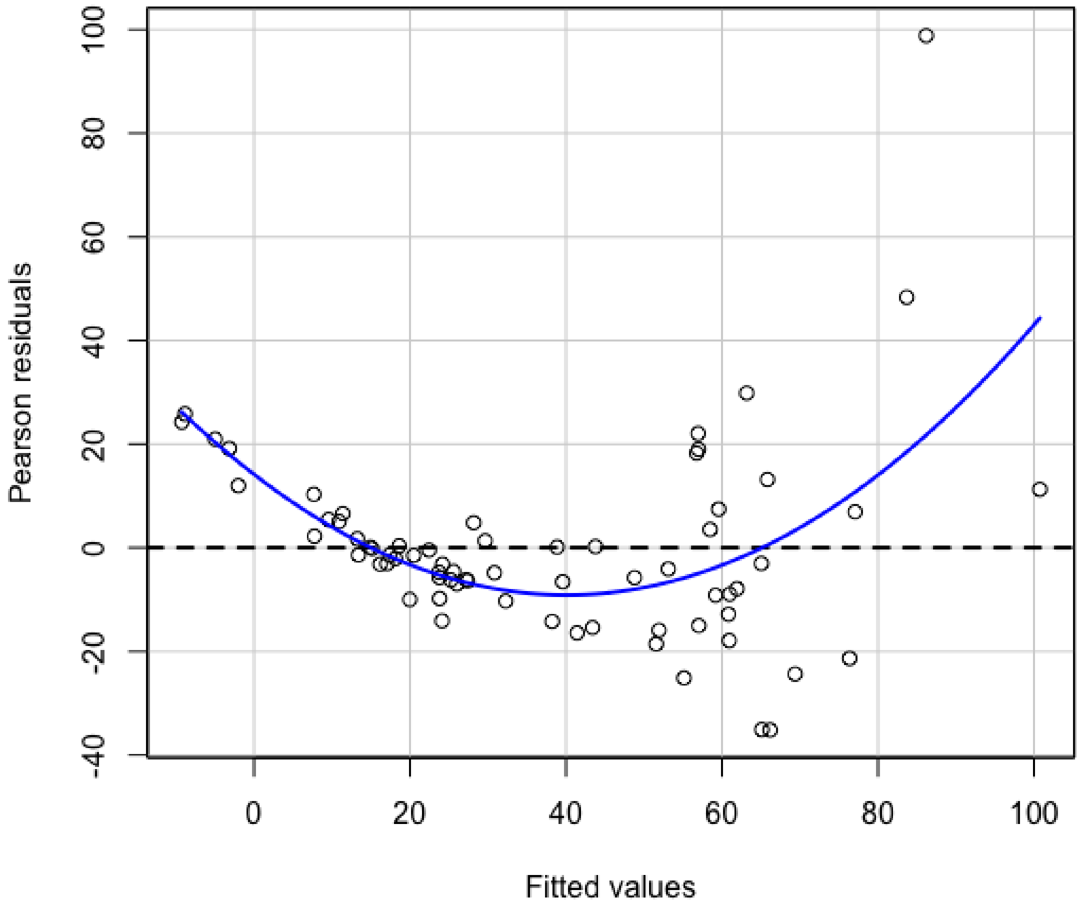
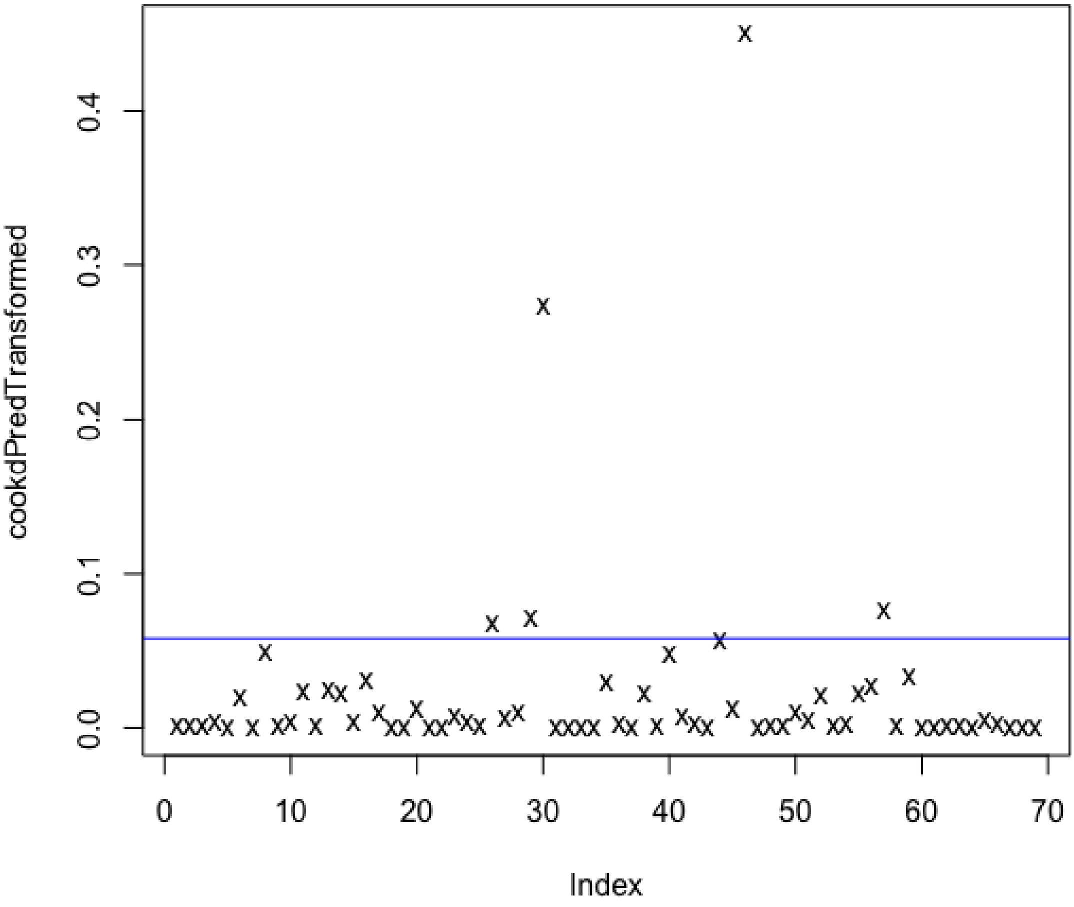


Figure 8: Residual Plot Cooks Distance



d. Given the predictor transformation made in the last part, use the Box-Cox method to find a transformation of BigMac. Recheck the five diagnostics. What is the transformation ans which assumptions still appear to be invalid.

Solution:
Finally transforming both the predictor and response using the powerTransform() function we get that BigMac should once again be raised to the power of $-.5$. Running the diagnostics to the fitted model we get passes all assumptions except that its residuals are not normally distributed, and it still has multiple leverage points.
Code:

```
-------------------------------------------------------
## Finding Box-Cox Transformation
> summary(powerTransform(model_Pred_Transform))
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1   -0.2834        -0.5      -0.5009      -0.0658
-------------------------------------------------------
## Settiing up Transformed Model
> model_Final_Transform <- lm(I(BigMac^(-.5)) ~ log(Bread) +
                             log(Rice) + log(FoodIndex) +
                             log(Bus) + sqrt(Apt) +
                             log(TeachGI) + log(TeachNI) +
                             I(TaxRate^(1.1)) + TeachHours,
                             data = df)

> summary(model_Final_Transform)
Residuals:
      Min          1Q      Median          3Q         Max
-0.048095  -0.015159  -0.001807   0.014231   0.087000

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       9.055e-02  5.863e-02   1.545  0.12780
log(Bread)       -2.208e-02  8.147e-03  -2.710  0.00879 **
log(Rice)        -1.196e-02  9.360e-03  -1.278  0.20625
log(FoodIndex)    4.054e-02  1.733e-02   2.340  0.02271 *
log(Bus)         -3.244e-04  7.956e-03  -0.041  0.96762
sqrt(Apt)        -8.399e-05  6.584e-04  -0.128  0.89892
log(TeachGI)      8.517e-02  3.912e-01   0.218  0.82839
log(TeachNI)     -6.074e-02  3.906e-01  -0.155  0.87697
I(TaxRate^(1.1)) -6.812e-04  3.513e-03  -0.194  0.84689
TeachHours       -3.291e-04  5.061e-04  -0.650  0.51812
```

```
Residual standard error: 0.02848 on 59 degrees of freedom
Multiple R−squared:  0.8165,    Adjusted R−squared:  0.7885
F−statistic: 29.17 on 9 and 59 DF,  p−value: < 2.2e−16
——————————————————————————————————————————————————
## Testing Linearity
## Passes
> residualPlots(model_Final_Transform)
                  Test stat Pr(>|Test stat|)
log(Bread)          −0.3724          0.71095
log(Rice)           −0.4575          0.64902
log(FoodIndex)       1.3154          0.19355
log(Bus)            −0.5345          0.59503
sqrt(Apt)            0.6856          0.49570
log(TeachGI)         1.3456          0.18366
log(TeachNI)         1.4999          0.13907
I(TaxRate^(1.1))    −1.6770          0.09893 .
TeachHours           0.3949          0.69435
Tukey test          −0.0653          0.94793
——————————————————————————————————————————————————
## Testing Constant Variance
## Passes
> ncvTest(model_Final_Transform)
Non−constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.568242, Df = 1, p = 0.21046
——————————————————————————————————————————————————
## Has no outliers with less conservative Bonferroni
## value
> outlierTest(model_Final_Transform)
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
         rstudent unadjusted p−value Bonferroni p
Miami 3.545876....         0.00078178     0.053943
——————————————————————————————————————————————————
## Still has several influence\leverage points
> cookdFinalTransformed <− cooks.distance(model_Final_Transform)
> plot(cookdFinalTransformed, pch = ’x’)
> abline(h = 4/length(cookdFinalTransformed), col = ’blue’)
> Influential_index <− (names(cookdFinalTransformed)
                        [(cookdFinalTransformed > (4/length(cookdFina
> Influential_index
[1] ”Lima”        ”Mexico_City” ”Miami”       ”Shanghi”
”Tokyo”
——————————————————————————————————————————————————
```

```
## Testing AutoCorrelation
## Passes, No AutoCorrelation
> dwt(model_Final_Transform)
 lag Autocorrelation D-W Statistic p-value
   1      -0.1296902      2.256308   0.254
 Alternative hypothesis: rho != 0
--------------------------------------------------
## Testing Normality
## Fails, Residuals do not appear normal.
> shapiro.test(residuals(model_Final_Transform))

        Shapiro-Wilk normality test

data:  residuals(model_Final_Transform)
W = 0.95508, p-value = 0.01446
--------------------------------------------------
```

Figure 9: Predictor Residual Plots for Full Transformed Model

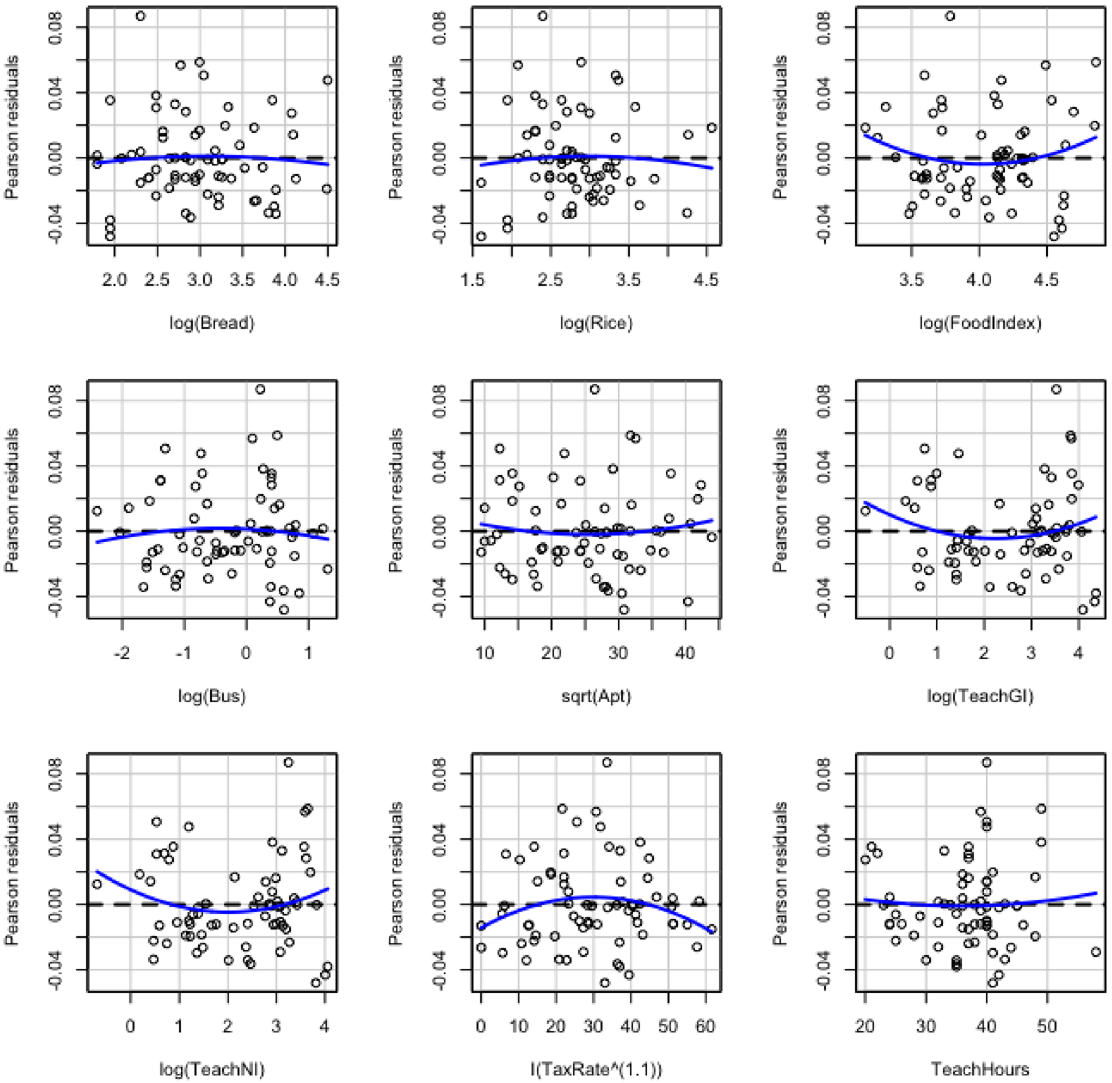


Figure 10: Residual Plot for Full Transformed Model

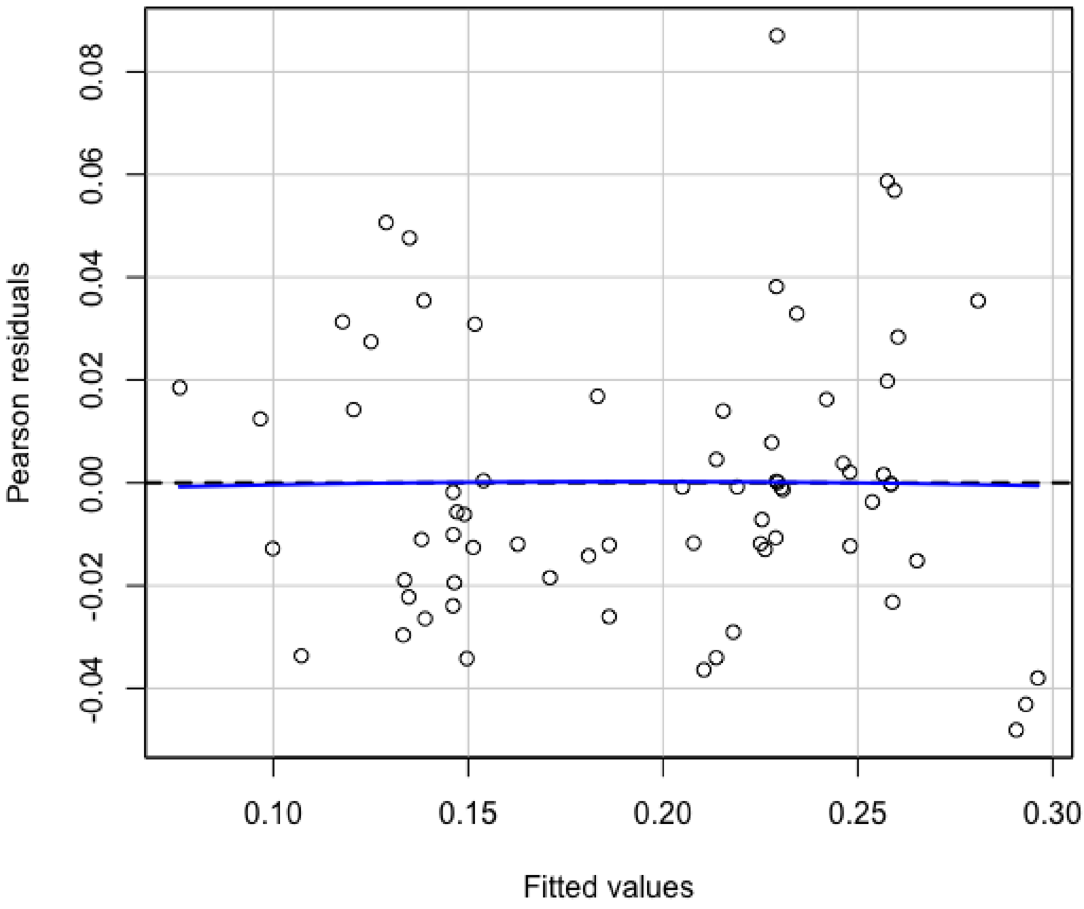
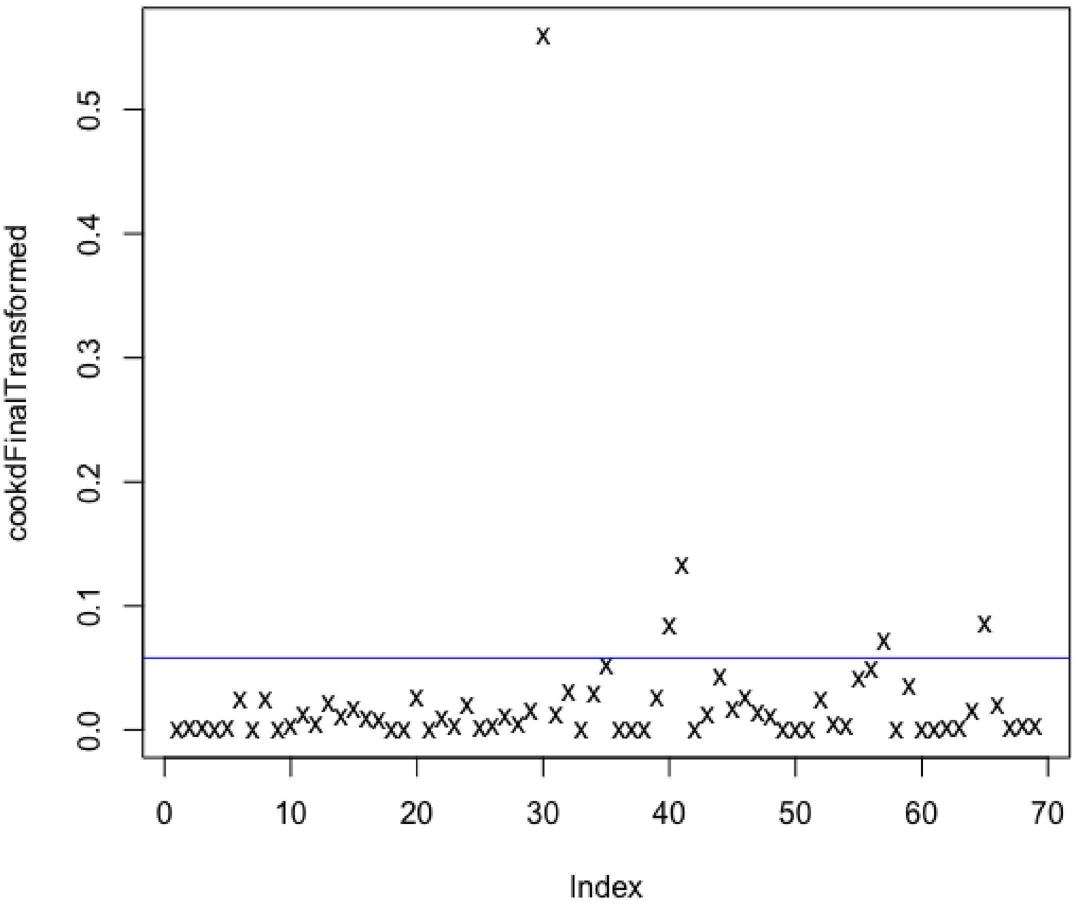


Figure 11: Residual Plot Cooks Distance



e . Are you satisfied wiht the final model? Why or why not?

Solution:
I am not satisfied with the final model. The final model does not produce normally distributed residuals, it would likely be better to just use the first transformed model where we simply transformed the response variable.

Exercise 10.4: Use whatever transformations yuo prefer to satisfy your assesment of the diagnostics.
For the boys in Berkeley Guidance Study in Problem, 3.3. find a model for HT18 as a function of other variable for ages 9 and earlier. Perform a complete analysis included selection of transformation and diagnostic analysis, and summarize your results.

Solution:
Checking the Box-Cox, nearest ladder of powers transformation for the response results in no change to the response data. Checking the same transformation for the predictors results in a substantial transformation which doesn't seem to make a huge difference on the diagnostics. The following compared the diagnostics to the transformed and non-transformed models.
The standard model, no transformations and all predictors.
Code:

```
----------------------------------------------------

model <- lm(HT18~ WT2 + HT2 + WT9 + HT9 + LG9 + ST9, data = df)
> summary(model)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9146 -1.7280  0.1164  1.9112  7.0939

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 44.39974   16.29469   2.725  0.00845 **
WT2          0.53027    0.32276   1.643  0.10572
HT2         -0.30299    0.17635  -1.718  0.09102 .
WT9         -0.05333    0.19963  -0.267  0.79031
HT9          1.25099    0.11454  10.922 8.53e-16 ***
LG9         -0.61487    0.47459  -1.296  0.20017
ST9          0.03964    0.03375   1.174  0.24495

Residual standard error: 3.013 on 59 degrees of freedom
Multiple R-squared:  0.806,     Adjusted R-squared:  0.7862
F-statistic: 40.85 on 6 and 59 DF,  p-value: < 2.2e-16

----------------------------------------------------
## Testing Linearity
## Passes
> residualPlots(model)
           Test stat Pr(>|Test stat|)
WT2           0.8139           0.4190
HT2          -0.0444           0.9647
WT9           0.3643           0.7169
```

```
HT9          −0.4084         0.6844
LG9           0.1289         0.8979
ST9          −0.1674         0.8676
Tukey test    0.1365         0.8914
−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−
## Testing Constant Variance
## Passes
> ncvTest(model)
Non−constant Variance Score Test
 Variance formula: ~ fitted.values
 Chisquare = 0.6801585, Df = 1, p = 0.40953
−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−
## Has no outliers with less conservative Bonferroni
## value
> outlierTest(model)
No Studentized residuals with Bonferroni p < 0.05
 Largest |rstudent|:
   rstudent unadjusted p−value Bonferroni p
57 2.545374           0.013597      0.89742
−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−
## Checking For Leverage Points
## Still has Leverage Points
> cookd <− cooks.distance(model)
> plot(cookd, pch = ’x’)
> abline(h = 4/length(cookd), col = ’blue’)
> Influential_index <− (names(cookd)[(cookd > (4/length(cookd)))])
> Influential_index
[1] ”2”  ”16” ”44” ”60”
−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−
## Testing AutoCorrelation
## Passes, No AutoCorrelation
> dwt(model)
 lag Autocorrelation D−W Statistic p−value
   1      0.04818875      1.901781   0.702
 Alternative hypothesis: rho != 0
−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−
## Testing Normality
## Passes, Residuals appear normal.
> shapiro.test(residuals(model))
        Shapiro−Wilk normality test
data:  residuals(model)
W = 0.97876, p−value = 0.3164
−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−
```

Figure 12: Predictor Residual Plots for Standard Model

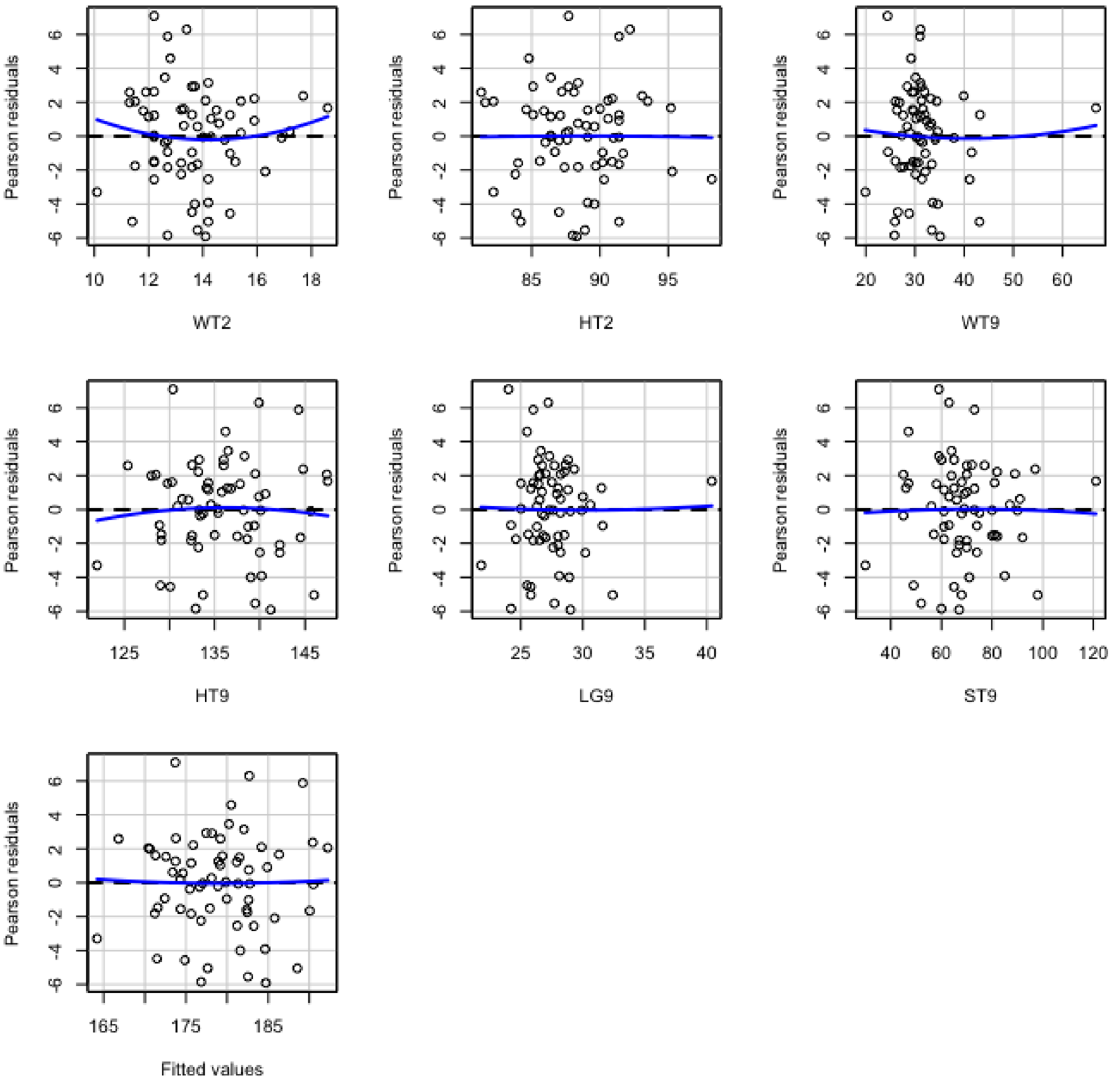


Figure 13: Residual Plot for Standard Model

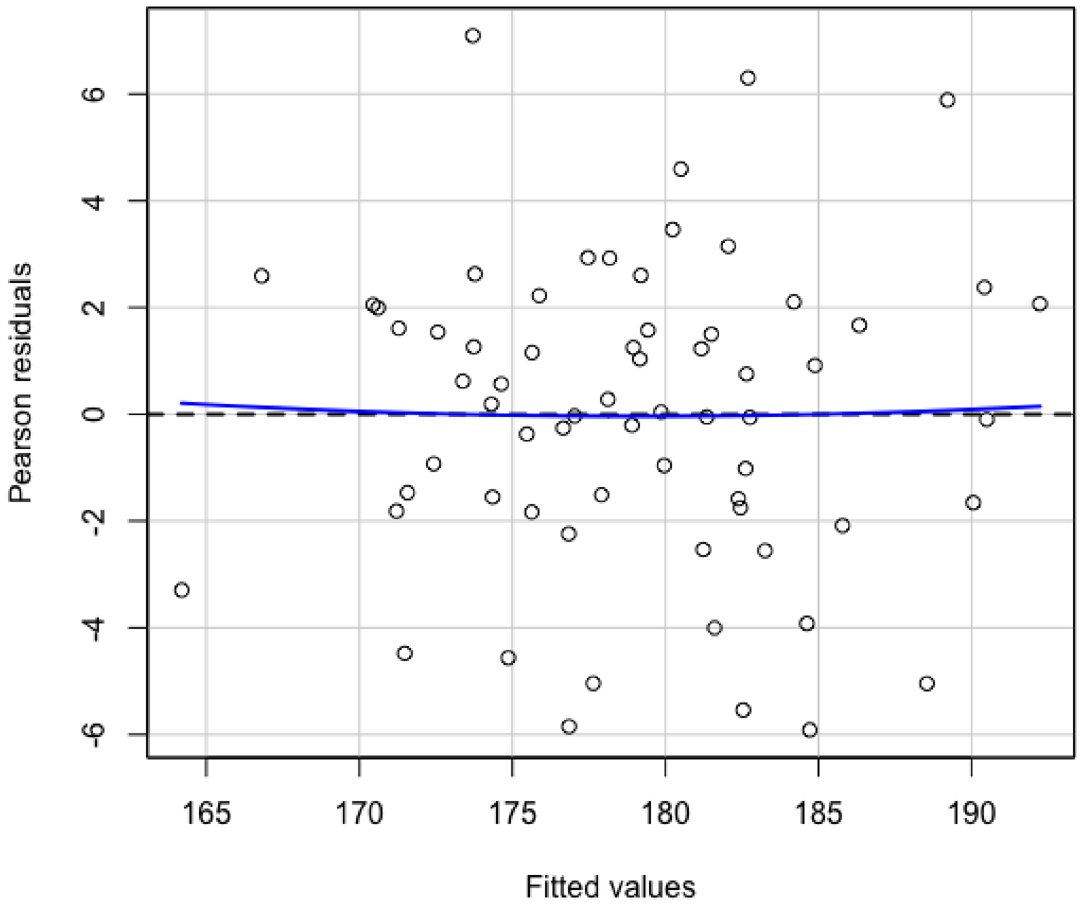
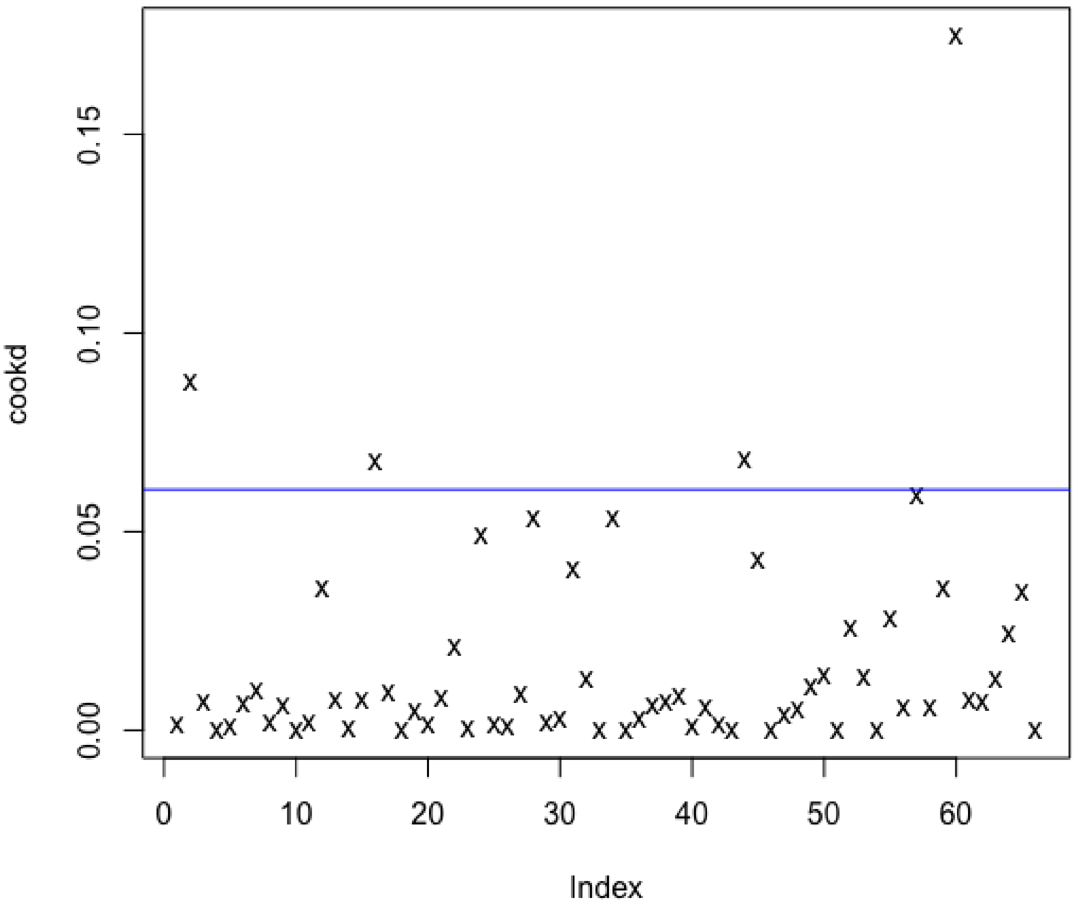


Figure 14: Residual Plot Cooks Distance



Predictor transformed model.
Code:

```
—————————————————————————————————————————————————————
## Finding Box−Cox Transformation For Predictors

> summary(powerTransform(cbind(WT2,HT2,WT9,HT9,LG9,ST9)~1, data = df))
bcPower Transformations to Multinormality
     Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
WT2    −1.2909           0     −2.8558       0.2740
HT2    −1.9862           1     −6.3381       2.3656
WT9    −1.3138          −1     −1.9487      −0.6789
HT9    −1.1724           1     −5.5173       3.1726
```

```
LG9    -2.1851             -1         -3.6035         -0.7667
ST9     0.5611              1         -0.1592          1.2815

------------------------------------------------------------
## Fitting the model
model_pred_Transformed <- lm(HT18~ log(WT2) +
                              HT2 + I(-1*WT9) +
                              HT9 + I(-1*LG9) +
                              ST9, data = df)
> summary(model_pred_Transformed)
Residuals:
    Min      1Q  Median      3Q     Max
-5.9521 -1.7210  0.1415  1.9947  7.0962

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.58455   16.34695   2.054   0.0444 *
log(WT2)     7.03091    4.49100   1.566   0.1228
HT2         -0.30056    0.17746  -1.694   0.0956 .
I(-1 * WT9)  0.04124    0.19979   0.206   0.8372
HT9          1.24669    0.11469  10.870 1.03e-15 ***
I(-1 * LG9)  0.62661    0.47696   1.314   0.1940
ST9          0.04013    0.03381   1.187   0.2400

Residual standard error: 3.019 on 59 degrees of freedom
Multiple R-squared:  0.8052,    Adjusted R-squared:  0.7854
F-statistic: 40.64 on 6 and 59 DF,  p-value: < 2.2e-16
------------------------------------------------------------
## Testing Linearity
## Passes
> residualPlots(model_pred_Transformed)
            Test stat Pr(>|Test stat|)
log(WT2)       0.6505           0.5179
HT2            0.0239           0.9810
I(-1 * WT9)    0.4474           0.6562
HT9           -0.3129           0.7555
I(-1 * LG9)    0.2108           0.8338
ST9           -0.0761           0.9396
Tukey test     0.1795           0.8576
------------------------------------------------------------
## Testing Constant Variance
## Passes
> ncvTest(model_pred_Transformed)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
```

```
 Chisquare = 0.8398594, Df = 1, p = 0.35944
---------------------------------------------------------
## Has no outliers with less conservative Bonferroni
## value
> outlierTest(model_pred_Transformed)
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferroni p
57  2.54096           0.013752      0.90766
---------------------------------------------------------
## Checking For Leverage Points
## Still has Leverage Points
> cookd <- cooks.distance(model_pred_Transformed)
> plot(cookd, pch = ’x’)
> abline(h = 4/length(cookd), col = ’blue’)
> Influential_index <- (names(cookd)[(cookd > (4/length(cookd)))])
> Influential_index
[1] ”2”  ”16” ”44” ”60”
---------------------------------------------------------
## Testing AutoCorrelation
## Passes, No AutoCorrelation
> dwt(model_pred_Transformed)
 lag Autocorrelation D-W Statistic p-value
   1      0.04450205      1.908687   0.704
 Alternative hypothesis: rho != 0
 ---------------------------------------------------------
## Testing Normality
## Passes, Residuals appear normal.
 > shapiro.test(residuals(model_pred_Transformed))
        Shapiro-Wilk normality test

 data:  residuals(model_pred_Transformed)
W = 0.97909, p-value = 0.3287
```

Figure 15: Predictor Residual Plots for Standard Model

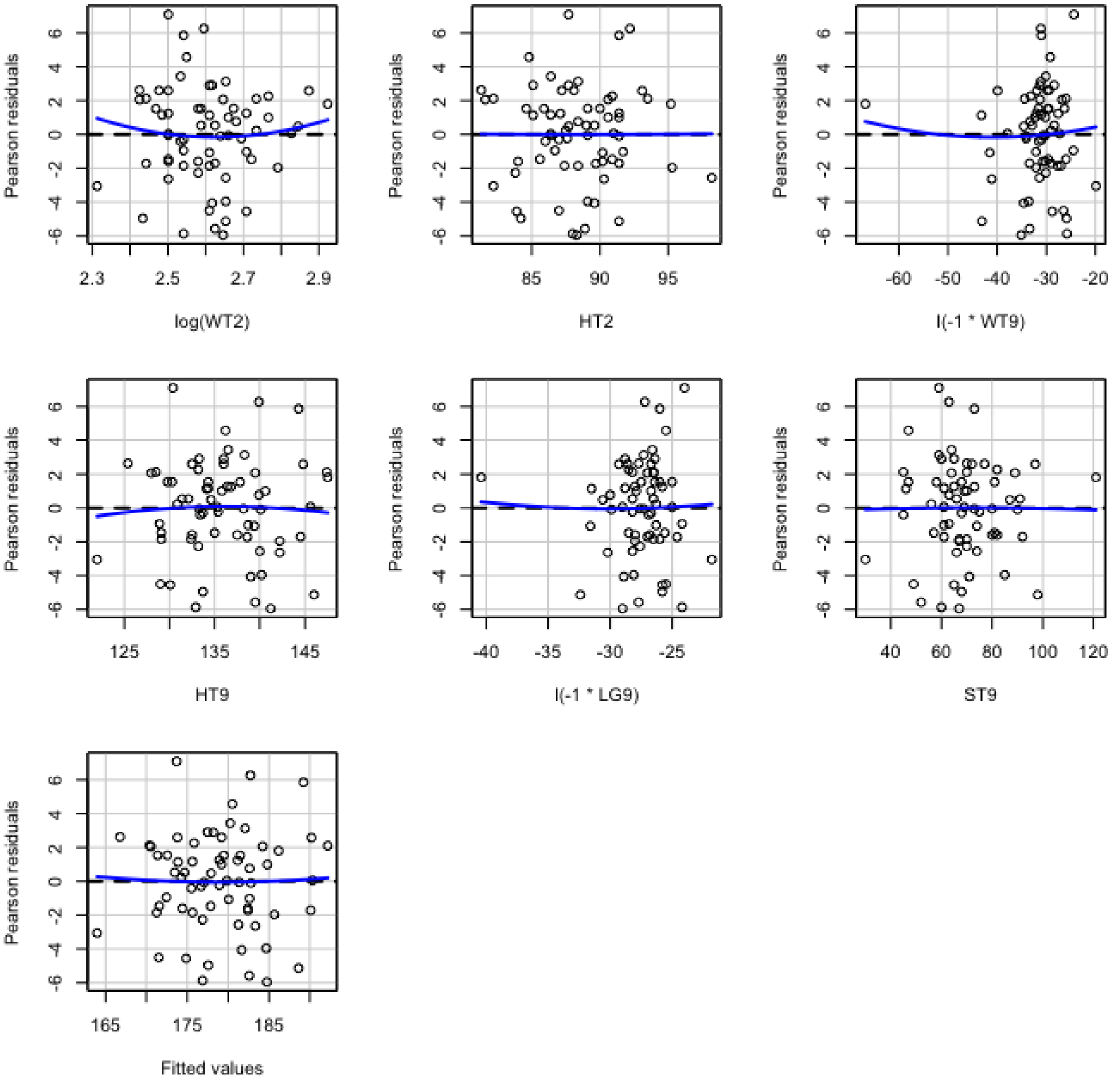


Figure 16: Residual Plot for Standard Model

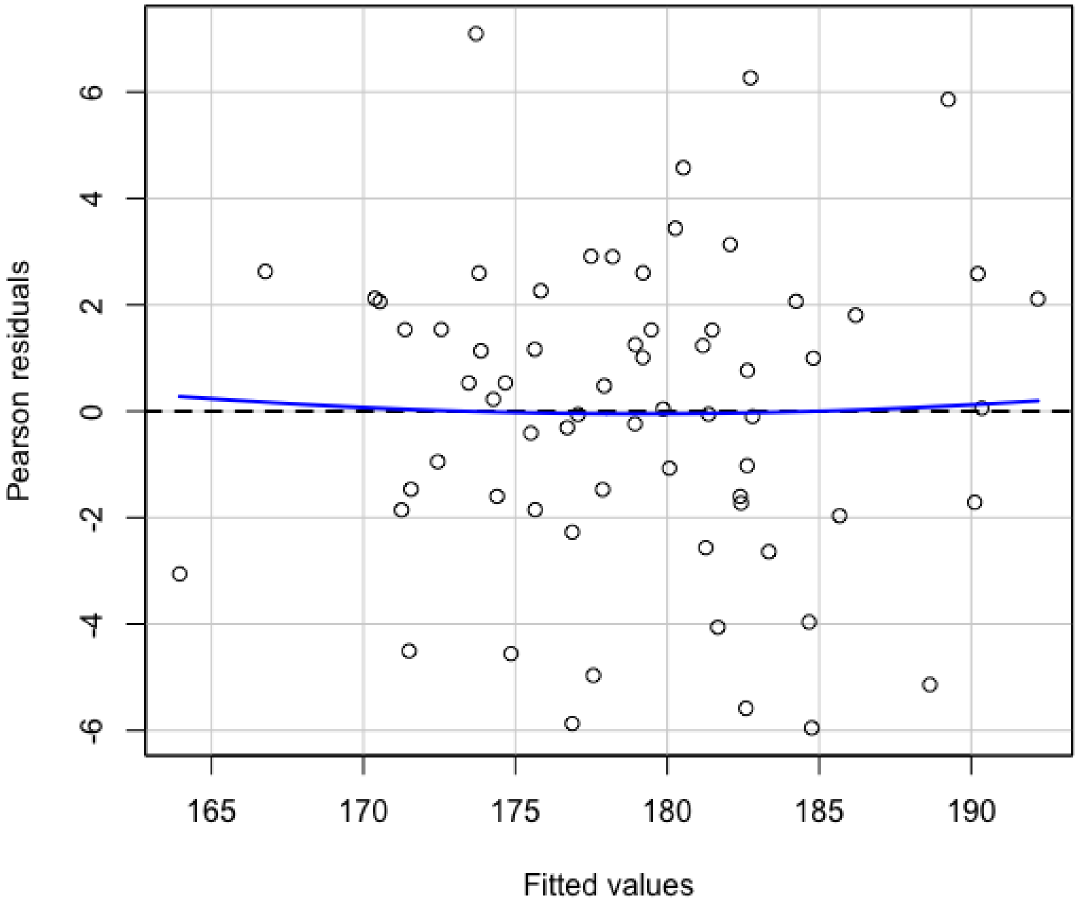
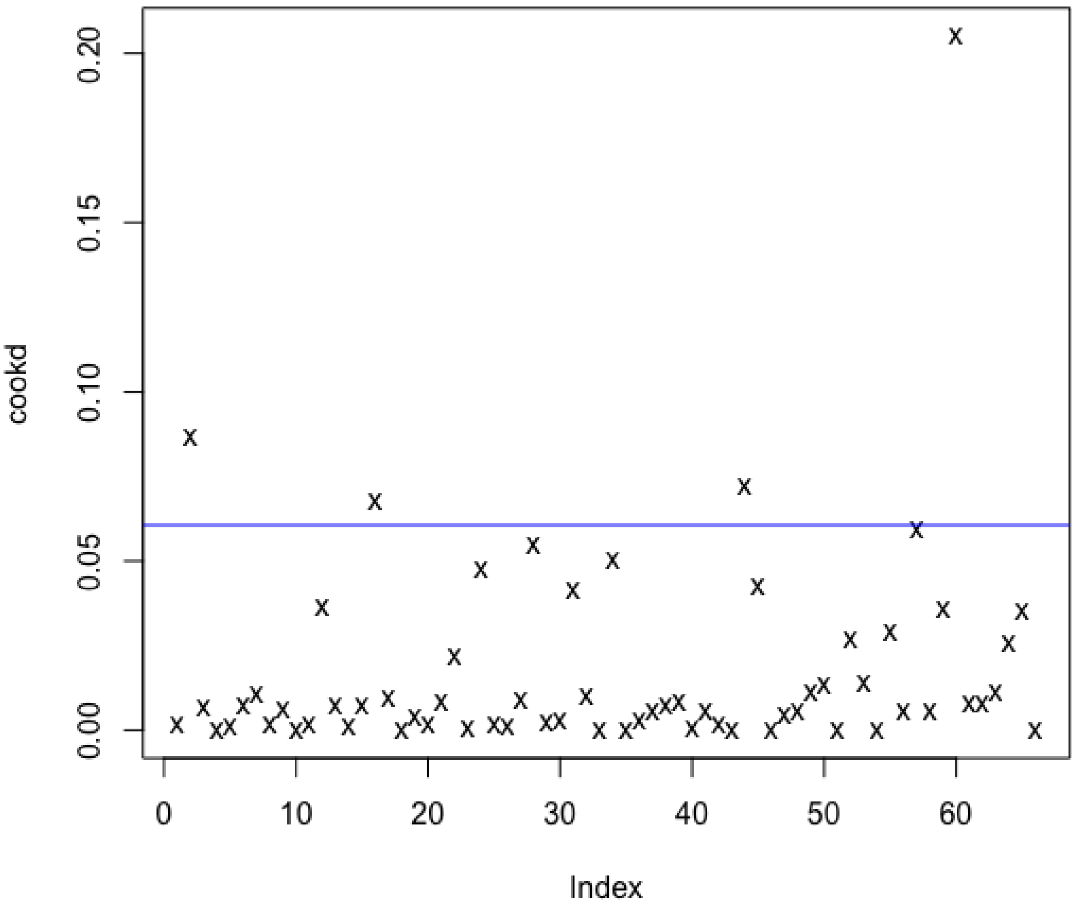


Figure 17: Residual Plot Cooks Distance



As we can see the diagnostic scored almost exactly the same, and comparing model summary reports they are also very similar. It seems as though there is no significant gain in transforming this data. I would likely use the plain data.

Exercise 4.6: In the simple linear regression of log(fertility) on pctUrban using UN11 data the fitted model is,

$$log(fertility) = 1.501 - 0.01 pctUrban$$

Provide an interpretation of the estimated coefficient for pct Urban.

Solution:
For every one unit increase of pctUrban we can expect the mean response of fertility to decrease by a factor of $e^{.01} \approx 1.01$ or approximately 1 percent.