MATH 424: Homework Chapter 7: Residual Analysis

Due on Saturday, November 18, 2017

Kafai 11:10am

Jonathan Dombrowski

Contents

Q 4

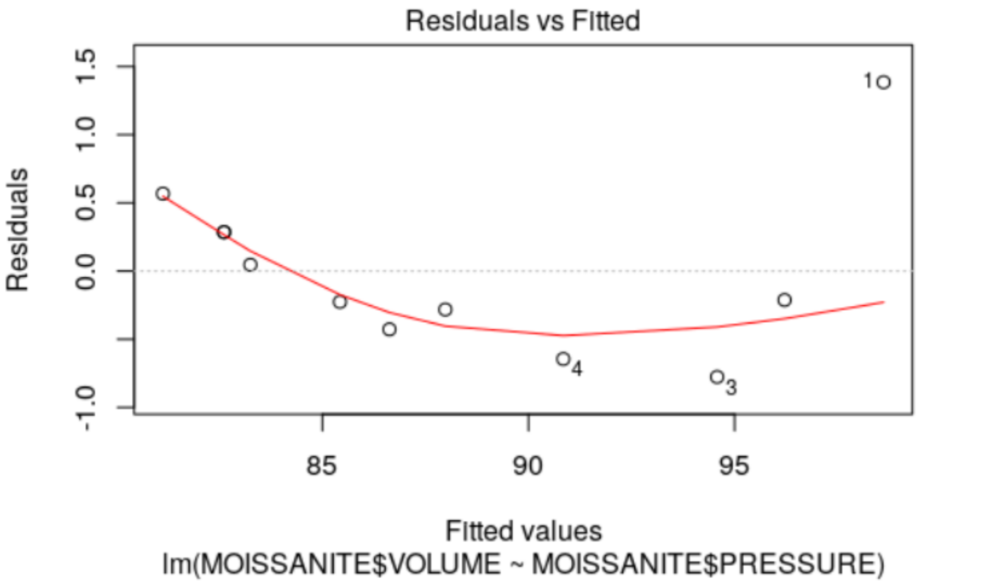
Elasticity of moissanite. Moissanite is a popular abrasive material because of its extreme hard- ness. Another important property of moissanite is elasticity. The elastic properties of the material were investigated in the Journal of Applied Physics (September 1993). A diamond anvil cell was used to compress a mixture of moissanite, sodium chloride, and gold in a ratio of 33:99:1 by volume. The com- pressed volume, y, of the mixture (relative to the zero-pressure volume) was measured at each of 11 different pressures (GPa). The results are displayed in the table (p. 397). A MINITAB printout for the straight-line regression model E(y) = 0 + 1 x and a MINITAB residual plot are displayed at left.
(a) Calculate the regression residuals.
(b) Plot the residuals against x. Do you detect a trend?
(c) Propose an alternative model based on the plot, part b.
(d) Fit and analyze the model you proposed in part c.

a

The residuals for the given model are as follows:

```
         1          2          3          4          5          6 
7 
-161.49037  -77.65049  -35.77246  176.28837 -117.29612   53.68403  141.98597 
         8          9         10         11         12         13         14 
 50.66572  116.91583   71.68961   17.19792  -73.17866   52.86881  -34.07440 
        15         16         17         18         19         20         21 
-117.08463  210.10916  100.71196 -117.45702  213.49551  176.15586 -153.66866 
        22         23         24         25         26         27         28 
 22.62320 -185.37244  -33.27300  176.25143   76.81172 -189.54715 -138.15989 
        29         30         31         32 
108.07673   16.11443 -206.48496 -141.13600 
```

b

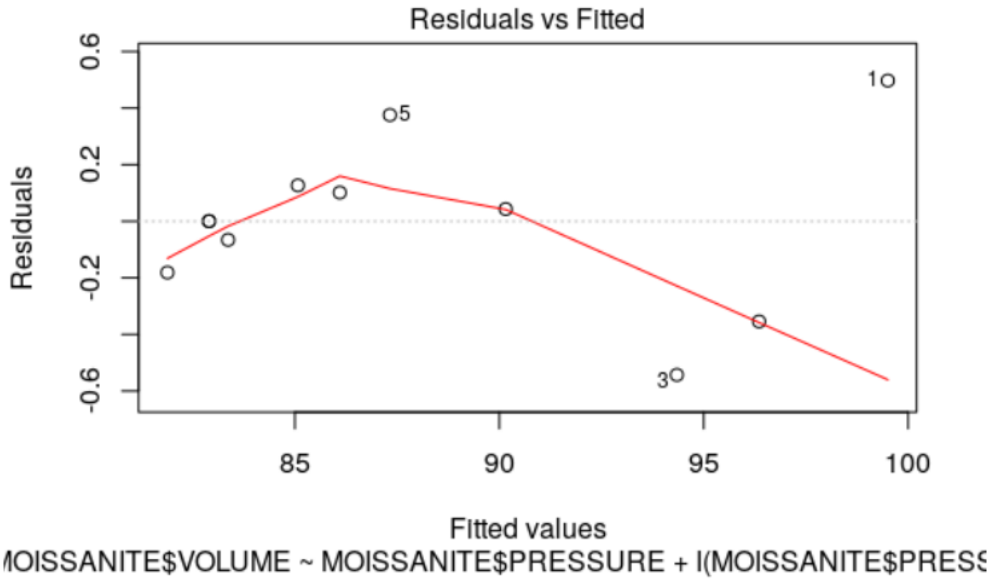


After plotting the residuals of the model, the plot of Fitted values vs Residuals, we can see that there is a trend in the residuals and instead of being linear, appears to be quadratic. By changing the model to a quadratic one, we can hopefully quell some of these errors.

c

The new proposed model is $E(y) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1^2$ A second degree model will hopefully deal with the likely quadratic residual plot, and normalize the residual.

d



Which we can see that the residuals are within 2s and appear to be far more randomnly distributed than the purely linear model. We can then conclude that the second order model fits the data more accurately.

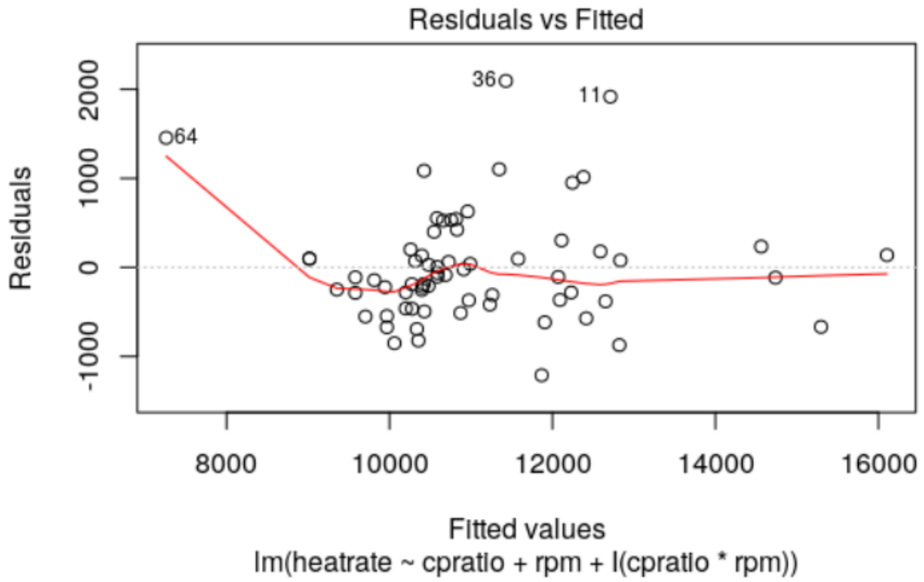
Q 11

E(y) = 0 + 1 x 1 + 2 x 2 + 3 x 1 x 2 .
Fit the model to the data saved in the GASTURBINE file, then plot the residuals against predicted heat rate. Is the assumption of a constant error variance reasonably satisfied? If not, suggest how to modify the model.

a

Fitting the model to the data:

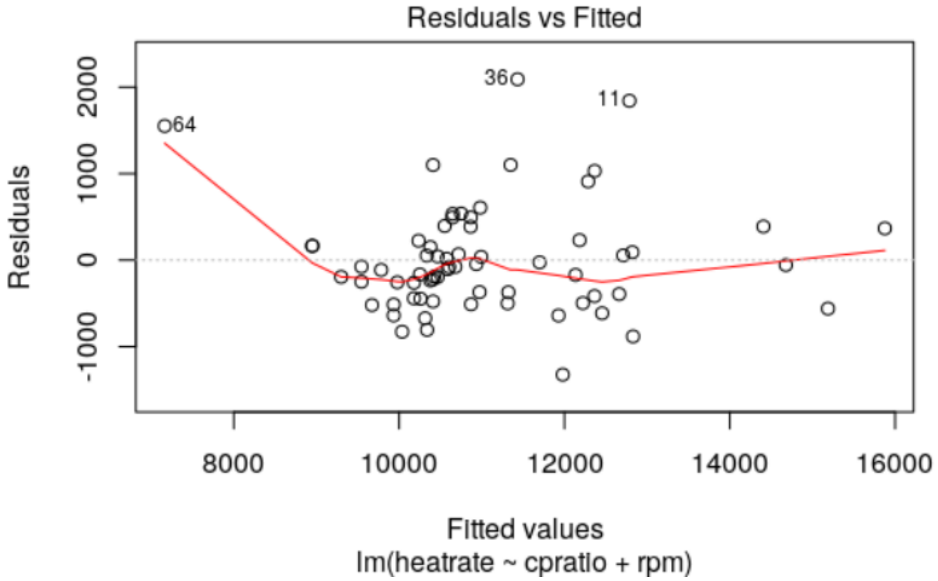
```
lm(formula = heatrate ~ cpratio + rpm + I(cpratio * rpm))

Residuals:
    Min       1Q   Median       3Q      Max
-1211.7   -375.6   -107.2    189.7   2095.0

Coefficients:
                     Estimate  Std. Error  t value  Pr(>t)
(Intercept)         1.207e+04   4.185e+02   28.828   < 2e-16 ***
cpratio            -1.461e+02   2.666e+01   -5.479  7.98e-07 ***
rpm                 1.697e-01   3.467e-02    4.895  7.16e-06 ***
I(cpratio * rpm)   -2.425e-03   3.120e-03   -0.777      0.44
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 633.8 on 63 degrees of freedom
Multiple R-squared:  0.8492,    Adjusted R-squared:  0.8421
F-statistic: 118.3 on 3 and 63 DF,  p-value: < 2.2e-16
```



The residuals seem to be randomly distributed, possibly binomially distrubuted if anything. The residuals are not the issue with this model, instead we propose that since the t-score of the interaction term is only -0.777, with a p-value of 0.44, that it is not statistically significant. Changing to a new first degree model is what I suggest. Testing that suggestion:
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$

```
lm(formula = heatrate ˜ cpratio + rpm)

Residuals:
     Min        1Q    Median        3Q       Max
-1323.67   -428.36    -78.54    227.83   2090.48

Coefficients:
              Estimate Std. Error t value Pr(>t)
(Intercept)  1.220e+04  3.809e+02  32.026  < 2e-16 ***
cpratio     -1.587e+02  2.103e+01  -7.548 2.02e-10 ***
rpm          1.446e-01  1.271e-02  11.383  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1          1

Residual standard error: 631.9 on 64 degrees of freedom
Multiple R-squared:  0.8478,    Adjusted R-squared:  0.843
F-statistic: 178.3 on 2 and 64 DF,  p-value: < 2.2e-16
```

We see that all terms are statistically significant and that the residuals have roughly the same distribution.



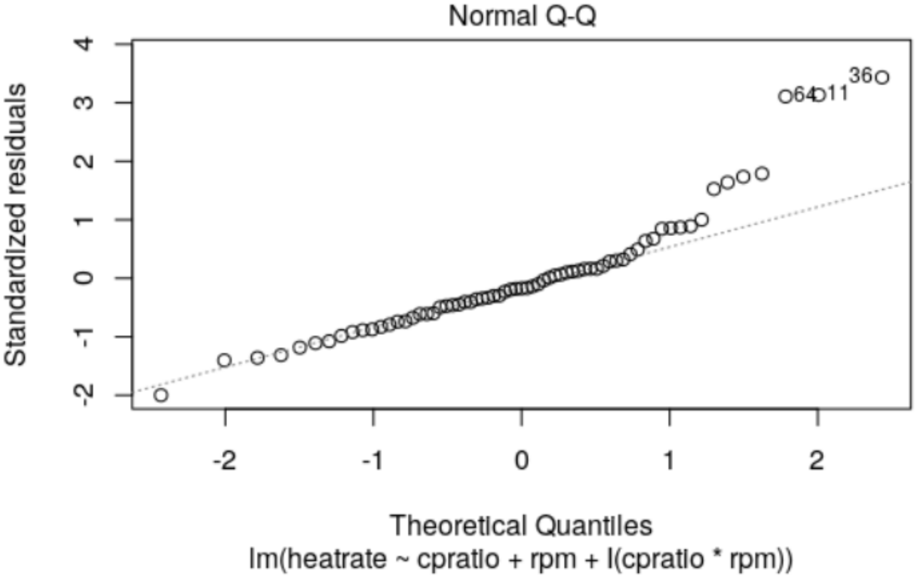
Therefore the only suggested change to the model is to remove the interaction term with the final model being :
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$

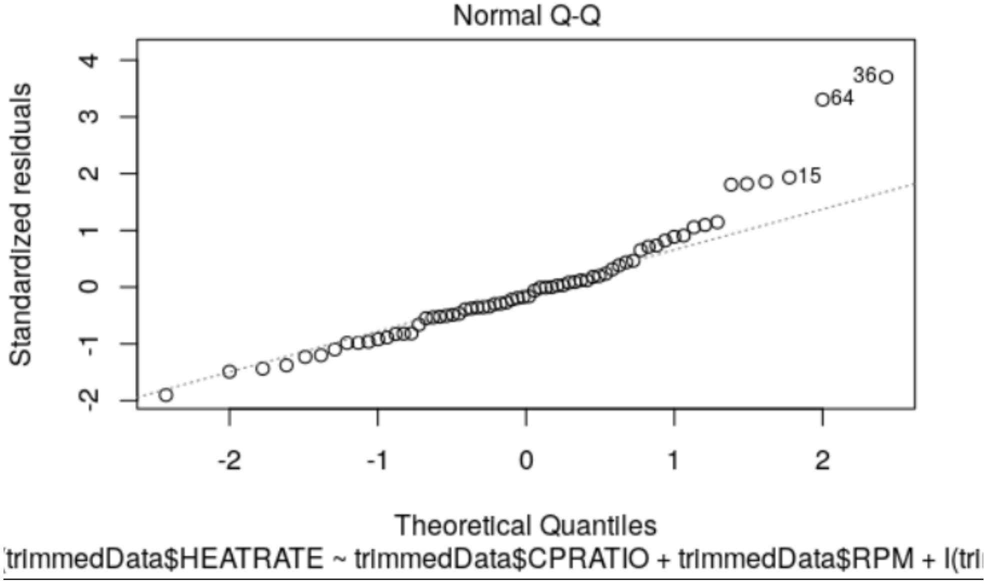
Q 20

Refer to the Journal of Engineering for Gas Turbines and Power (January 2005) study of a high-pressure inlet fogging method for a gas turbine engine, Exercise 8.11 (p. 407). Use a residual graph to check the assumption of normal errors for the interaction model for heat rate (y). Is the normal- ity assumption reasonably satisfied? If not, suggest how to modify the model.

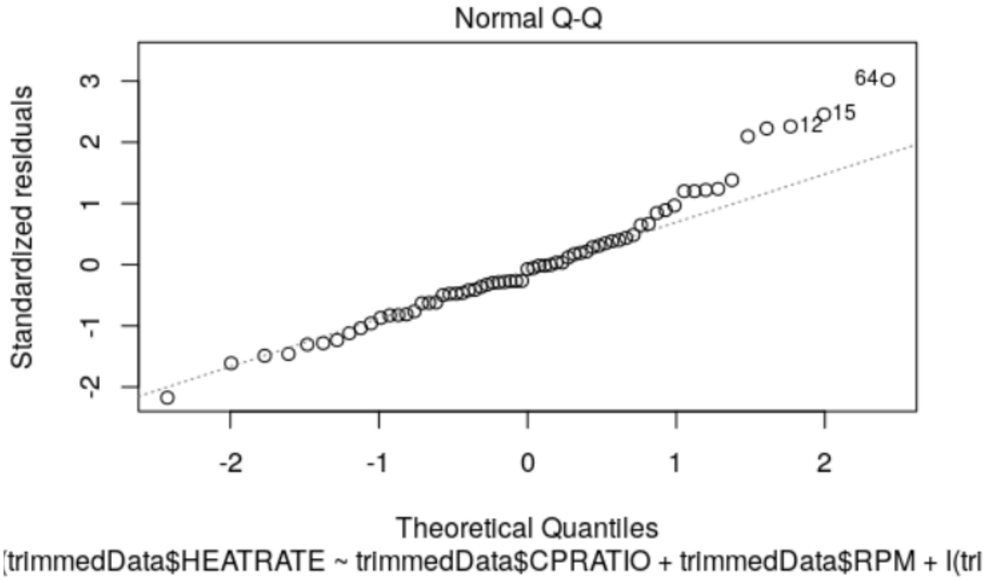
a

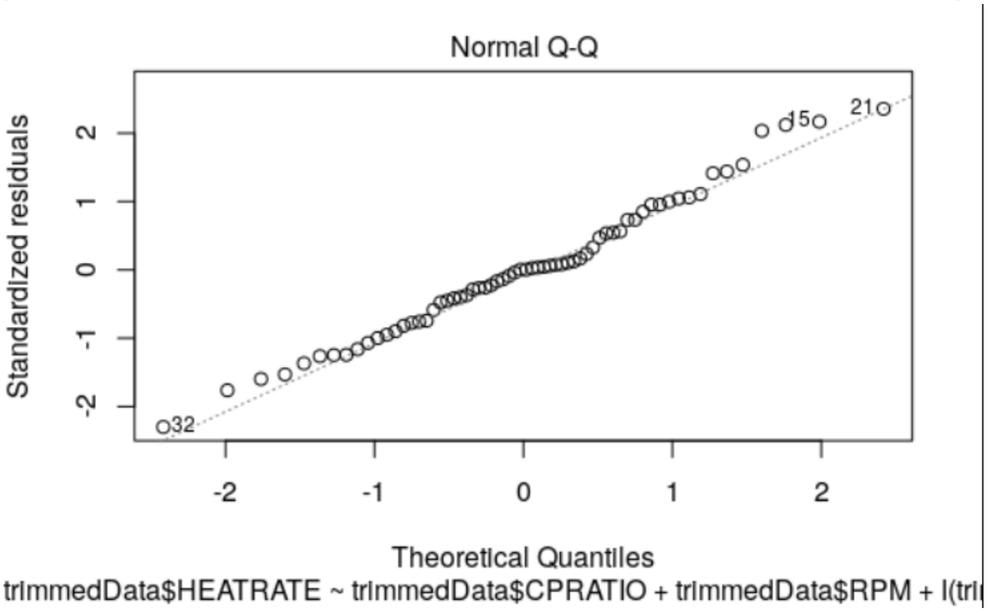
Here, we graph the residuals versus the expected residuals under the assumption of normality. If we assume normality, then we expect to see a near straight line from plotting th two against eachother. Failing to see this would be akin to a proof by contradiction. We assume the plot will be a straight line and see if the shows us otherwise. In this case, excluding the 3 extreme outliers, 11,36,64, this can be viewed as sufficiently normal.





Q 20 [a] continued on next page. . .



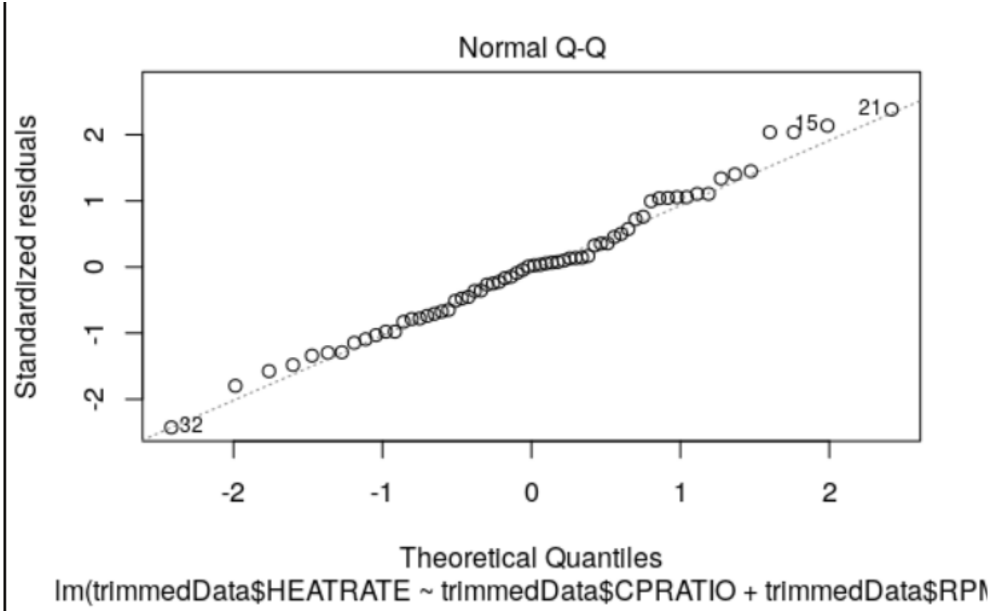


After excluding the 3 outliers in respect to residuals, the line appears resonably straight, and no values exceed the 3s limit, therefore we can assume that the constant variation of error is normal. The same model, with trimmed values is the suggested modified model. Doing analysis of the suggested model:

```
lm(formula = trimmedData$HEATRATE ~ trimmedData$CPRATIO + trimmedData$RPM,
    data = trimmedData)

Residuals:
     Min       1Q   Median       3Q      Max
-1150.79  -341.74     7.81   291.68  1138.87

Coefficients:
                      Estimate Std. Error t value Pr(>t)
(Intercept)          1.271e+04  3.789e+02  33.544  < 2e-16 ***
trimmedData$CPRATIO -1.939e+02  2.180e+01  -8.894 1.29e-12 ***
```

```
trimmedData$RPM      1.340e-01  1.058e-02  12.664  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 484.2 on 61 degrees of freedom
Multiple R-squared:  0.9003,    Adjusted R-squared:  0.897
F-statistic: 275.3 on 2 and 61 DF,  p-value: < 2.2e-16
```

From the previous problem we can confirm that all terms of the model are statistically significant. Doing residual analysis:



We can confirm that the distribution of the trimmed model is still normal and therefore we can suggest it as an alternative to the initial model.

Q 26

Prices of antique clocks. Refer to the grandfather clock example, Example 4.1 (p. 183). The least squares model used to predict auction price, y,from age of the clock, x 1 , and number of bidders, x 2 , was determined to be = 1, 339 + 12.74x 1 + 85.95x 2
(a) Use this equation to calculate the residuals of each of the prices given in Table 4.1 (p. 171).
(b) Calculate the mean and the variance of the residuals. The mean should equal 0, and the variance should be close to the value of MSE given in the SAS printout shown in Figure 4.3 (p. 172).
(c) Find the proportion of the residuals that fall outside 2 estimated standard deviations (2s) of 0 and outside 3s.
(d) Rerun the analysis and request influence diag- nostics. Interpret the measures of influence given on the printout.

a

We can calculate the residuals plugging the appropriate values into the regression formula, then subtracting them from the actual values to get the difference. By fitting the same model in the question, R can do this very easily. The result is below.

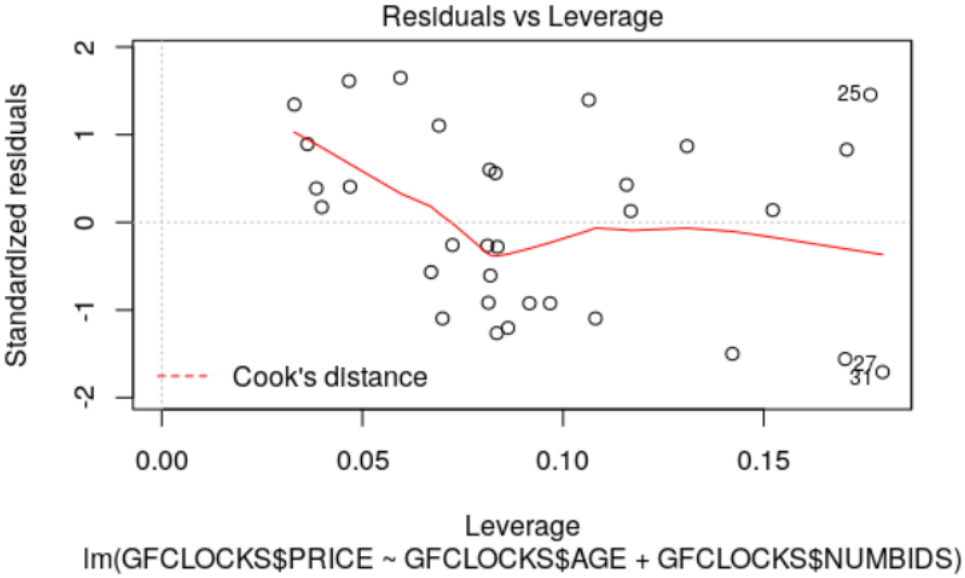
Q 26 [a] continued on next page. . .

```
 -161.49037 -77.65049   -35.77246   176.28837 -117.29612   53.68403  141.98597
50.66572    116.91583  71.68961   17.19792   -73.17866   52.86881   -34.07440
-117.08463  210.10916  100.71196  -117.45702 213.49551   176.15586  -153.66866
22.62320   -185.37244  -33.27300  176.25143   76.81172   -189.54715 -138.15989
108.07673   16.11443  206.48496  -141.13600
```

b

$\mu_{residuals}$ = sum(residuals(model))/n =-2.442491e-15≈ 0
Variance of model residuals = var(residuals(model)) = 16668.6
Variance of actual Y's = var(GFCLOCKS$PRICE)) = 154831.9
Which is only a 10% difference between the two variances.

c



We see that graphically, no residuals fall outside 2s or 3s. To numerically find this proportion, we compare this to 2s and 3s.
We find that s = 133.48, 2s = 266.97, and 3s = 400.45. By checking the maximum of the residuals, we can see there are any residuals that fall outside 2s or 3s. We find that the maximum value for a residual from this model is 213.50 and we can then conclude that there is not a proportion of residuals that fall outside 2s or 3s. The absolute maximum of the residuals of the model is 213.50, which is not outside 2s or 3s, which means that the proportion of residuals outside of 2s or 3s is zero.

d

Running the analysis with influence diagnostics:

```
        Influence measures of
         lm(formula = GFCLOCKS$PRICE ~ GFCLOCKS$AGE + GFCLOCKS$NUMBIDS) :

     dfb.1_ dfb.GFCLOCKS.A dfb.GFCLOCKS.N   dffit cov.r   cook.d    hat inf
1   0.02869        0.08543       -0.26148 -0.3855 1.023 0.048488 0.0835
2  -0.05862        0.10147       -0.06889 -0.1793 1.165 0.010954 0.0819
```

```
3  -0.06946          0.04707         0.05651 -0.0832 1.203 0.002384 0.0836
4   0.02387          0.03534        -0.03607  0.2519 0.948 0.020537 0.0330
5  -0.10784         -0.01563         0.20211 -0.2722 1.107 0.024830 0.0814
6  -0.11675          0.12356         0.07164  0.1526 1.233 0.007988 0.1158
7  -0.17188          0.13299         0.20699  0.3015 1.050 0.030061 0.0691
8   0.03122         -0.03108         0.00327  0.0763 1.138 0.002001 0.0385
9   0.08549         -0.05670        -0.04382  0.1725 1.061 0.009995 0.0363
10  0.13709         -0.13037        -0.05189  0.1669 1.173 0.009508 0.0831
11 -0.02517          0.00552         0.05138  0.0583 1.308 0.001172 0.1523
12 -0.07287          0.09580        -0.02787 -0.1504 1.152 0.007724 0.0671
13  0.05660         -0.03243        -0.04659  0.0887 1.146 0.002700 0.0469
14 -0.03577          0.00108         0.05932 -0.0779 1.200 0.002088 0.0812
15 -0.03668          0.12765        -0.17315 -0.3013 1.125 0.030415 0.0968
16  0.19705         -0.20502        -0.00326  0.3673 0.879 0.042407 0.0467
17 -0.29005          0.22077         0.30336  0.3738 1.247 0.047090 0.1707
18  0.12457         -0.21815         0.03515 -0.2924 1.118 0.028655 0.0916
19 -0.24865          0.24578         0.22009  0.4281 0.879 0.057341 0.0595
20 -0.34324          0.40949         0.14699  0.4901 1.009 0.077356 0.1064
21 -0.24163          0.09110         0.29752 -0.3731 1.042 0.045653 0.0863
22 -0.00628          0.01501        -0.00183  0.0346 1.154 0.000414 0.0399
23 -0.07864          0.29185        -0.37875 -0.6245 1.017 0.124219 0.1422
24  0.02310         -0.04715         0.01280 -0.0712 1.190 0.001744 0.0724
25  0.65137         -0.50561        -0.48185  0.6877 1.075 0.151355 0.1766
26 -0.09600          0.13758         0.01477  0.1771 1.165 0.010692 0.0817
27  0.02485          0.24471        -0.52572 -0.7250 1.030 0.166265 0.1703
28  0.18473         -0.30244         0.03236 -0.3829 1.097 0.048507 0.1081
29  0.31201         -0.25231        -0.20766  0.3356 1.181 0.037861 0.1309
30  0.04201         -0.03291        -0.02916  0.0459 1.256 0.000728 0.1169
31  0.13427         -0.50277         0.41400 -0.8280 0.985 0.212872 0.1796
32  0.00229         -0.13043         0.14359 -0.3018 1.052 0.030135 0.0699
```

Abiding by $H > \frac{2(k+1)}{n}$ we find the threshold for an H score to be overly influential to the model to be 0.1875; which none of the H values on the right side of the graph exceed, which confirms the previous conclusion that there are no residual values that exceed the influence limit.