Adops & Data Scientist Sample Data - Q2 Regression

Jae Hyun Ha

The data contains 300 rows and 3 columns (from the left, A, B, and C). Please build a good regression model which explains column C by a function of A and B.

Note: Please do not use any ML libraries or packages. You can simply attach plot of data points and your regression model that fits the data points and of course, code point (preferably Github).

import data

```
data <- read.csv("D:/Job Search/Adops & Data Scientist Sample Data - Q2 Regression.csv")
d <- data.frame(data)
colnames(d) <- c('A','B','C')
A <- d$A
B <- d$B
C <- d$C
```

Finding the best Regression Model

We try a simple multiple linear regression fit on the model to check the its goodness of fit

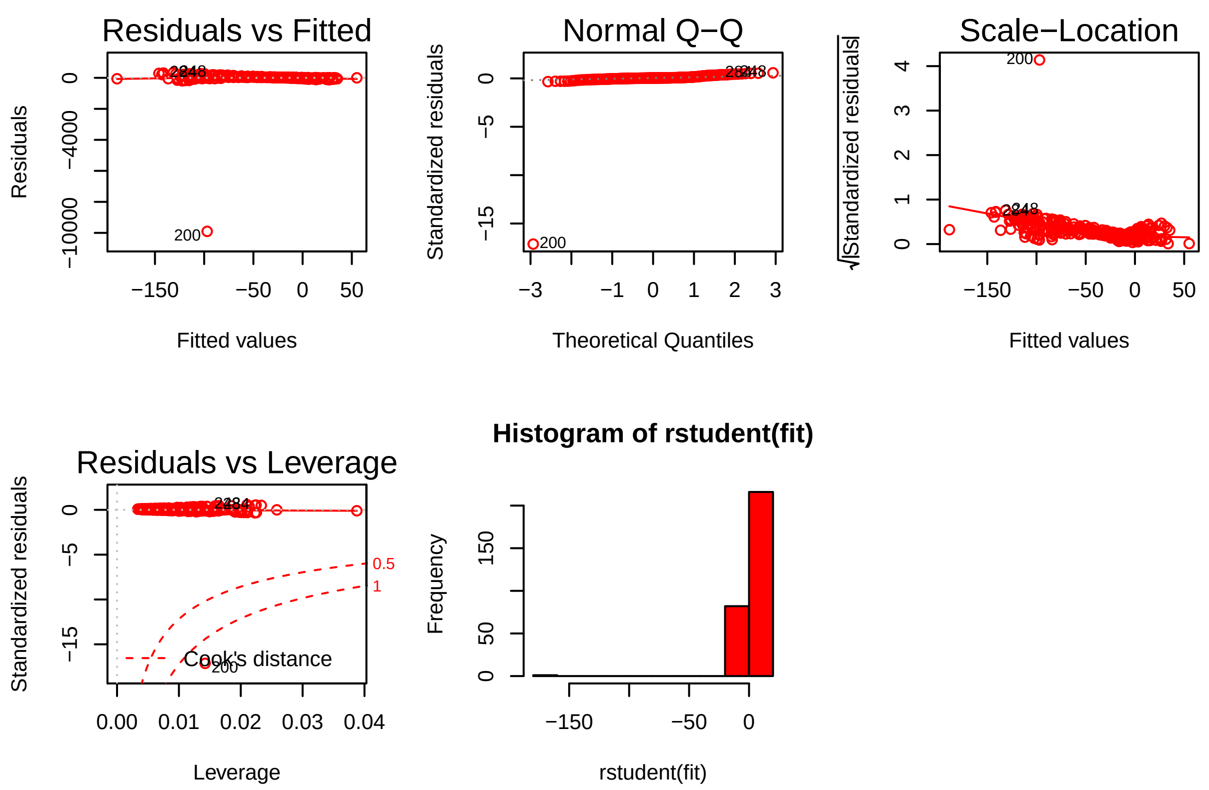
```
fit <- lm(C ~ A+B)
summary(fit)
```

```
##
## Call:
## lm(formula = C ~ A + B)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9903.0    -5.2    23.0    54.1   333.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -19.496     44.632  -0.437    0.663
## A             -8.057      6.029  -1.336    0.182
## B             -1.746     11.459  -0.152    0.879
##
## Residual standard error: 582.4 on 296 degrees of freedom
## Multiple R-squared:  0.006037,   Adjusted R-squared:  -0.0006787
## F-statistic: 0.8989 on 2 and 296 DF,  p-value: 0.4081
```

```
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: C
##            Df    Sum Sq Mean Sq F value Pr(>F)
## A           1    601980  601980  1.7747 0.1838
## B           1      7878    7878  0.0232 0.8790
## Residuals 296 100405289  339207
```

```r
par(mfrow=c(2,3))
plot(fit, col="red")
hist(rstudent(fit), col="red")
```



The fit is very bad as you can see by the R-quared value of 0.006037. You can also see from the histogram that the distribution is not normal and is very skwed. There are multiple approaches to fix these issues by modifying the model. Based on the plots we will look at how we can fix some of the issues that can be seen in the data.

We will try a log fit model. This will fix any non-linearity, outliers, and bunching in the x axis.

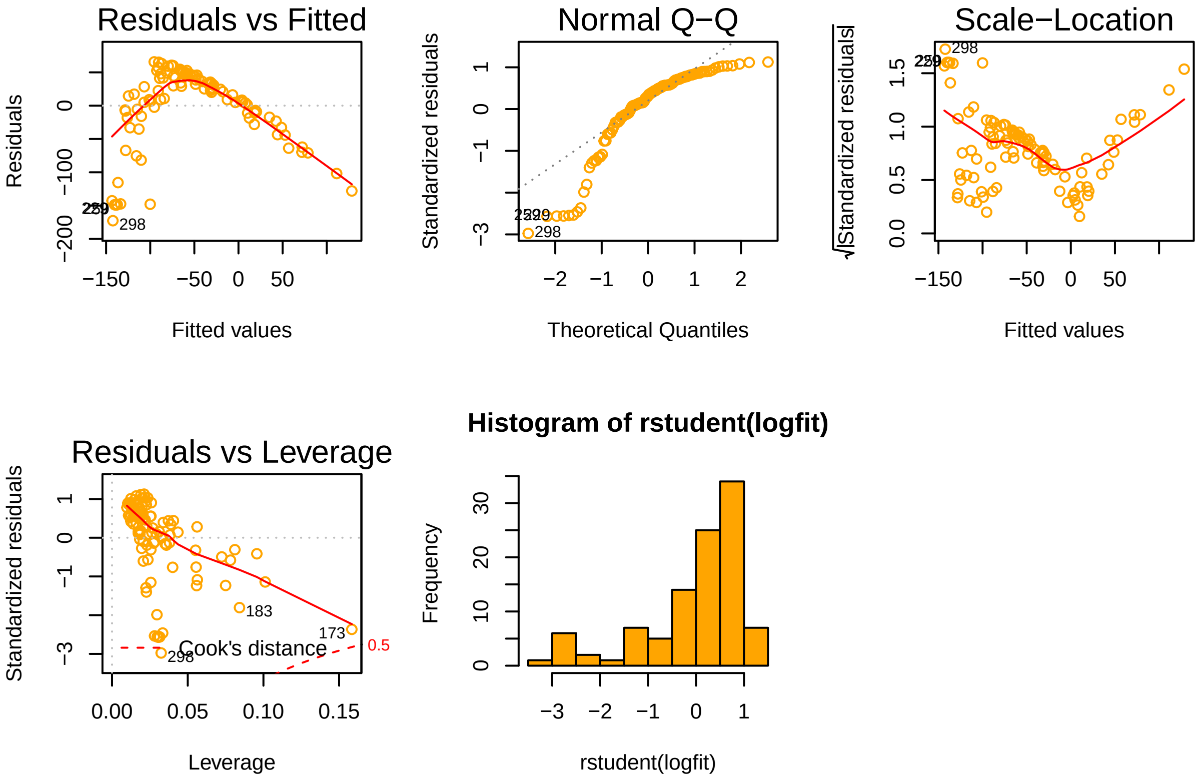
```r
logfit <- lm(C ~ log(A)+log(B))
summary(logfit)
```

```
##
## Call:
## lm(formula = C ~ log(A) + log(B))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -172.67  -17.88   18.62   41.88   65.69
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    43.708     11.044   3.958 0.000142 ***
## log(A)        -38.264      4.617  -8.288 5.78e-13 ***
## log(B)       -50.526      7.011  -7.206 1.15e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.98 on 99 degrees of freedom
##   (197 observations deleted due to missingness)
## Multiple R-squared:  0.5173, Adjusted R-squared:  0.5075
## F-statistic: 53.04 on 2 and 99 DF,  p-value: < 2.2e-16
```

```
anova(logfit)
```

```
## Analysis of Variance Table
##
## Response: C
##           Df Sum Sq Mean Sq F value    Pr(>F)
## log(A)     1 188378  188378  54.159 5.502e-11 ***
## log(B)     1 180622  180622  51.930 1.152e-10 ***
## Residuals 99 344343    3478
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(2,3))
plot(logfit, col="orange")
hist(rstudent(logfit), col="orange")
```



The the results are better but the fit is only moderately good. The standardized residuals are still not normally distributed. To account for any bunching in the y axis and any outliers with high leverage we will try a log-log fit.

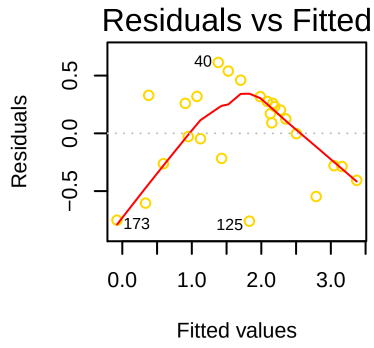
```
loglogfit <- lm(log(C) ~ log(A)+log(B))
summary(loglogfit)
```

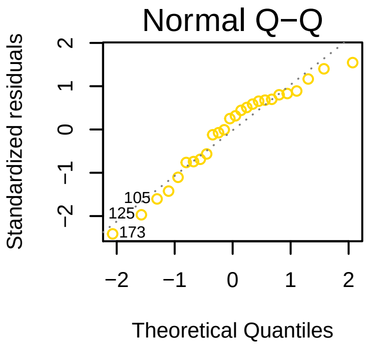
```
##
## Call:
## lm(formula = log(C) ~ log(A) + log(B))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7604 -0.2754  0.1076  0.2719  0.6136
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.65528    0.08861  18.680 2.13e-15 ***
## log(A)       0.18643    0.07252   2.571   0.0171 *
## log(B)      -0.88760    0.12716  -6.980 4.10e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4065 on 23 degrees of freedom
##   (273 observations deleted due to missingness)
## Multiple R-squared:  0.8435, Adjusted R-squared:  0.8299
## F-statistic:    62 on 2 and 23 DF,  p-value: 5.44e-10
```

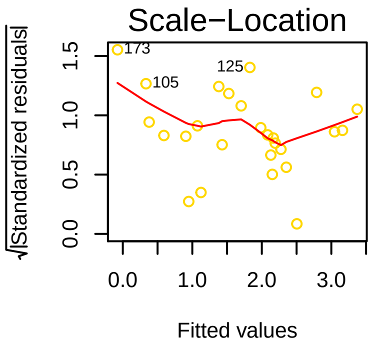
```
anova(loglogfit)
```

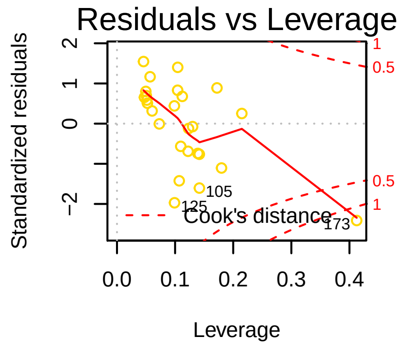
```
## Analysis of Variance Table
##
## Response: log(C)
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## log(A)     1 12.4395 12.4395  75.282 1.036e-08 ***
## log(B)     1  8.0505  8.0505  48.721 4.097e-07 ***
## Residuals 23  3.8005  0.1652
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

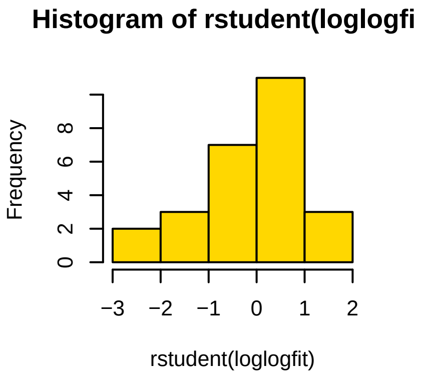
```
par(mfrow=c(2,3))
plot(loglogfit, col="gold")
hist(rstudent(loglogfit), col="gold")
```











This is definietely a better fit than the previous fit. You can see that the R-quared value has risen from 0.5173 to 0.8435. However, we can try to increase the R-quared by adding a variable.

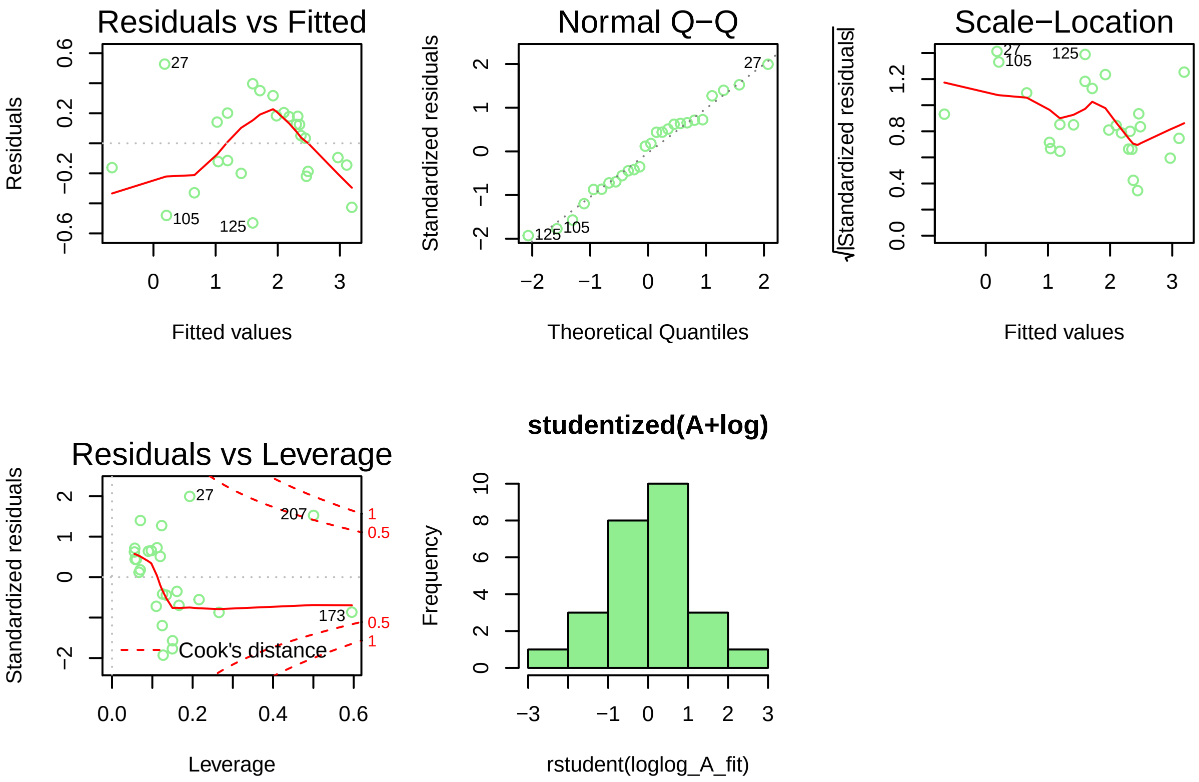
```
loglog_A_fit <- lm(log(C) ~ A+log(A) + log(B))
summary(loglog_A_fit)
```

```
##
## Call:
## lm(formula = log(C) ~ A + log(A) + log(B))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5303 -0.1810  0.0424  0.1822  0.5273
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.11519    0.11729  18.034 1.14e-14 ***
## A           -0.18929    0.04042  -4.683 0.000114 ***
## log(A)       0.41432    0.07157   5.789 7.98e-06 ***
## log(B)      -1.05093    0.09840 -10.680 3.60e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2941 on 22 degrees of freedom
##   (273 observations deleted due to missingness)
## Multiple R-squared:  0.9216, Adjusted R-squared:  0.911
## F-statistic: 86.26 on 3 and 22 DF,  p-value: 2.548e-12
```

```
anova(loglog_A_fit)
```

```
## Analysis of Variance Table
##
## Response: log(C)
##           Df Sum Sq Mean Sq F value    Pr(>F)
## A          1 7.2095  7.2095  83.333 6.172e-09 ***
## log(A)     1 5.3097  5.3097  61.373 8.344e-08 ***
## log(B)     1 9.8679  9.8679 114.061 3.600e-10 ***
## Residuals 22 1.9033  0.0865
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(2,3))
plot(loglog_A_fit, col="lightgreen")
hist(rstudent(loglog_A_fit), col="lightgreen",main="studentized(A+log)")
```



This is a very good model, the R-squared value is 0.9216 and the standaradized residuals have a normal distribution with a residual standard error of only 0.2941. We will try to explore a couple more variations to see if we can maximize the goodness of fit.

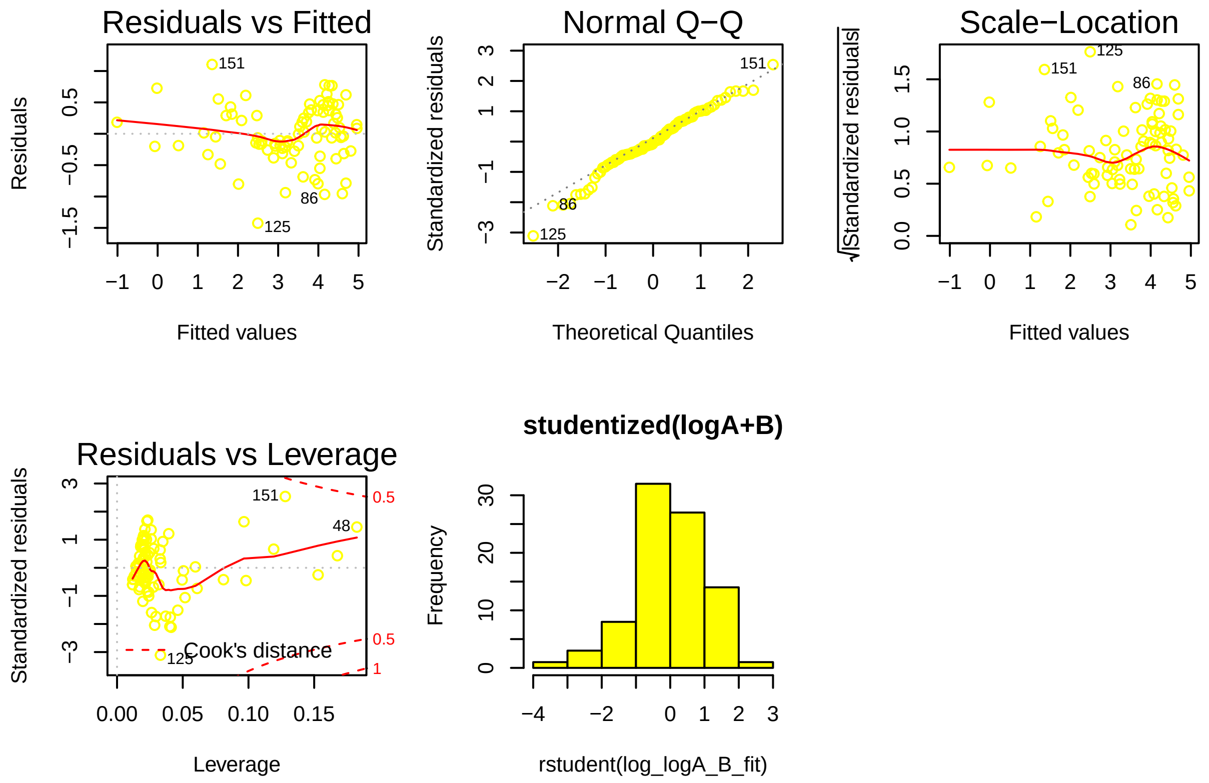
```
log_logA_B_fit <- lm(log(C) ~ log(A) + B)
summary(log_logA_B_fit)
```

```
##
## Call:
## lm(formula = log(C) ~ log(A) + B)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.42494 -0.22582 -0.02173  0.33106  1.10451
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.43726    0.06613   36.86   <2e-16 ***
## log(A)       0.42811    0.03543   12.08   <2e-16 ***
## B           -0.57577    0.03459  -16.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4658 on 83 degrees of freedom
##   (213 observations deleted due to missingness)
## Multiple R-squared:  0.8812, Adjusted R-squared:  0.8784
## F-statistic: 307.9 on 2 and 83 DF,  p-value: < 2.2e-16
```

```r
anova(log_logA_B_fit)
```

```
## Analysis of Variance Table
##
## Response: log(C)
##           Df Sum Sq Mean Sq F value    Pr(>F)
## log(A)     1 73.518  73.518  338.81 < 2.2e-16 ***
## B          1 60.126  60.126  277.08 < 2.2e-16 ***
## Residuals 83 18.010   0.217
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
par(mfrow=c(2,3))
plot(log_logA_B_fit, col="yellow")
hist(rstudent(log_logA_B_fit), col="yellow",main="studentized(logA+B)")
```



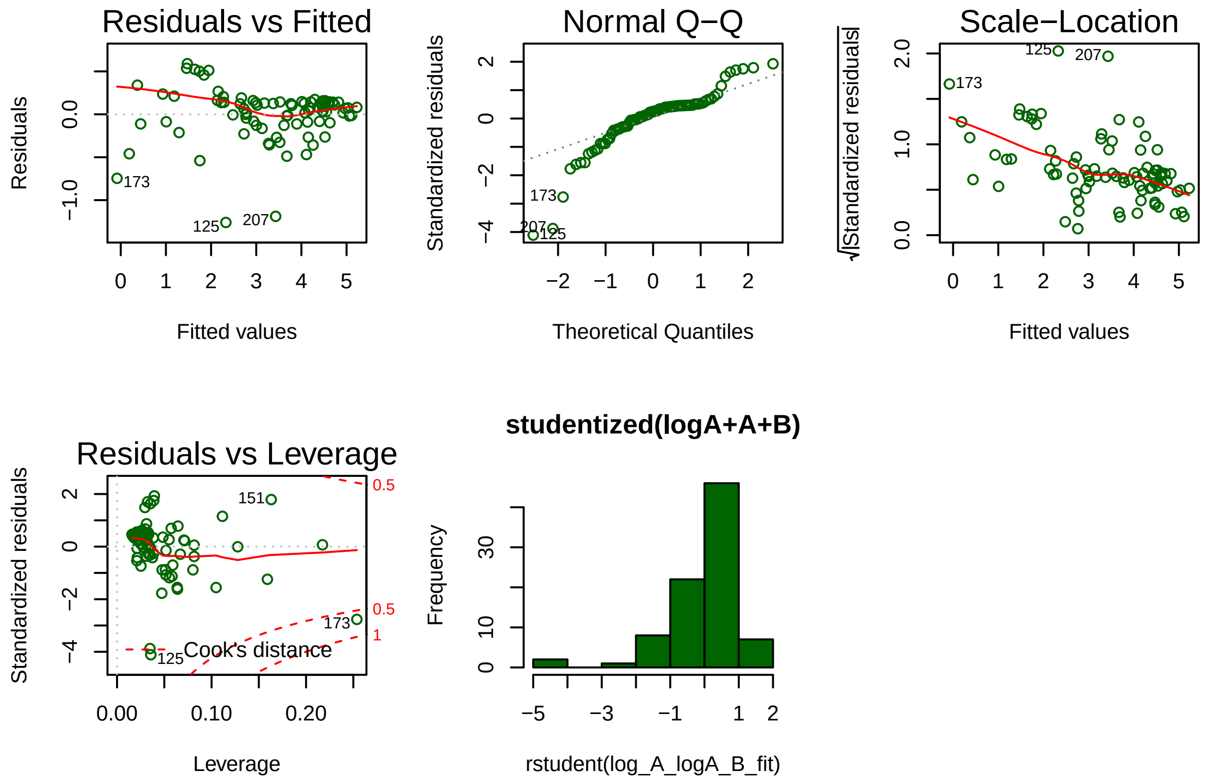
```r
log_A_logA_B_fit <- lm(log(C) ~ A+log(A) + B)
summary(log_A_logA_B_fit)
```

```
##
## Call:
## lm(formula = log(C) ~ A + log(A) + B)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.25988 -0.09667  0.07465  0.14087  0.58973
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.02662    0.06000  33.779  < 2e-16 ***
## A            0.13167    0.01297  10.148 3.81e-16 ***
## log(A)       0.11182    0.03918   2.854  0.00546 **
## B           -0.52883    0.02363 -22.383  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.312 on 82 degrees of freedom
##   (213 observations deleted due to missingness)
## Multiple R-squared:  0.9474, Adjusted R-squared:  0.9454
## F-statistic: 491.9 on 3 and 82 DF,  p-value: < 2.2e-16
```

```
anova(log_A_logA_B_fit)
```

```
## Analysis of Variance Table
##
## Response: log(C)
##           Df Sum Sq Mean Sq F value    Pr(>F)
## A          1 93.602  93.602 961.367 < 2.2e-16 ***
## log(A)     1  1.292   1.292  13.269 0.0004715 ***
## B          1 48.777  48.777 500.977 < 2.2e-16 ***
## Residuals 82  7.984   0.097
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(2,3))
plot(log_A_logA_B_fit, col="darkgreen")
hist(rstudent(log_A_logA_B_fit), col="darkgreen",main="studentized(logA+A+B)")
```



This last results shows the best fit. The R-quared value for the model is 0.9474 which is greater than our previous best fit model. However, the residual standard error is 0.312 whihc is lower than our previous best fit model. Looking at the plots you can see that the distribution of the standardaized residuals is more skewed and the lines don't fit the model as well as the previous model. You could say that we have overfitted the model.

Conclusion:
Both of the green models are considered to have very good fit, but the best fitted model is not always the

best model. We can conclude that the last model has the best fit but the best regression model to explain the relationship between C explained by A, B is the light green model:

```
summary(loglog_A_fit)
```

```
##
## Call:
## lm(formula = log(C) ~ A + log(A) + log(B))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5303 -0.1810  0.0424  0.1822  0.5273
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.11519    0.11729  18.034 1.14e-14 ***
## A           -0.18929    0.04042  -4.683 0.000114 ***
## log(A)       0.41432    0.07157   5.789 7.98e-06 ***
## log(B)      -1.05093    0.09840 -10.680 3.60e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2941 on 22 degrees of freedom
##   (273 observations deleted due to missingness)
## Multiple R-squared:  0.9216, Adjusted R-squared:  0.911
## F-statistic: 86.26 on 3 and 22 DF,  p-value: 2.548e-12
```