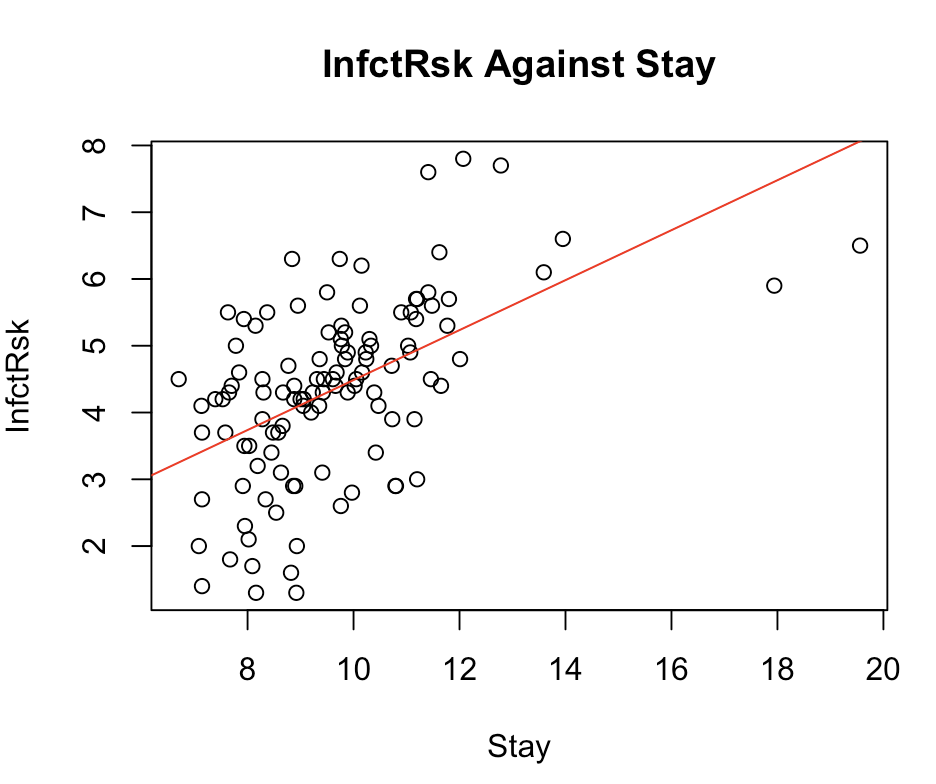
Yunlu Li (yl4df)

STAT 5120

Homework 3

1. (a) There are outliers: two patients stay longer than 18 days, whereas most patients stay less than 14 days. The linearity assumption and the constant variance assumption seems to be violated, due to the existence of outliers. If we can ignore the outliers, then the assumption for linear regression seems to hold.



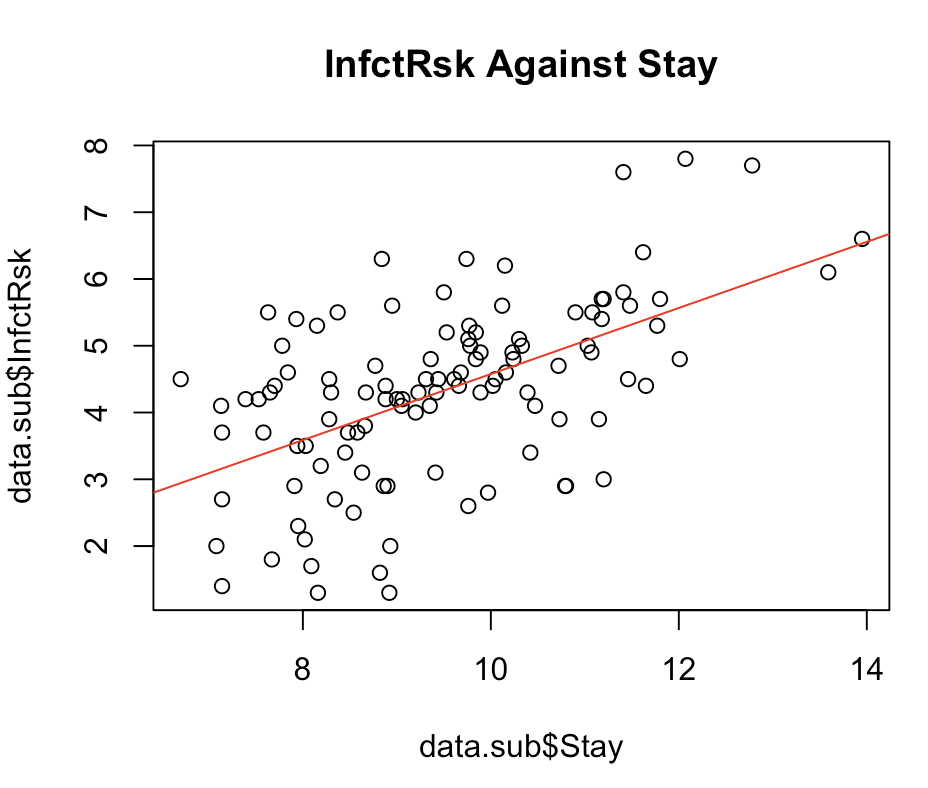
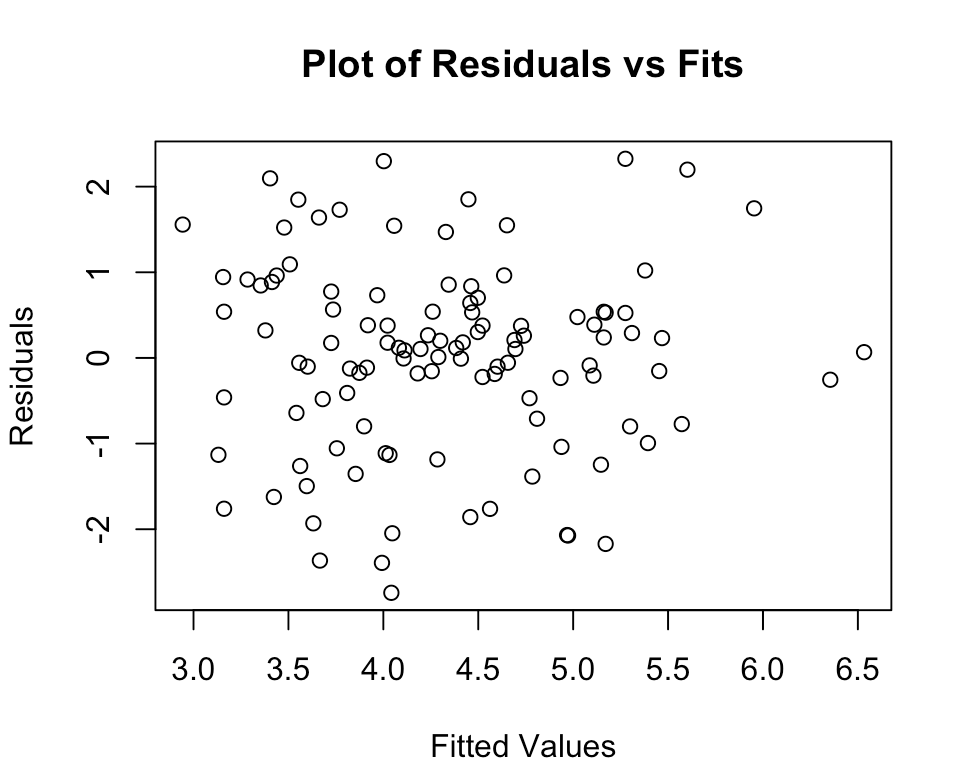
(b) The residuals, except two outliers, fall in a horizontal band around 0 with no apparent pattern and have similar variation for different fitted values. However, the two outliers make the assumption of regression violated.

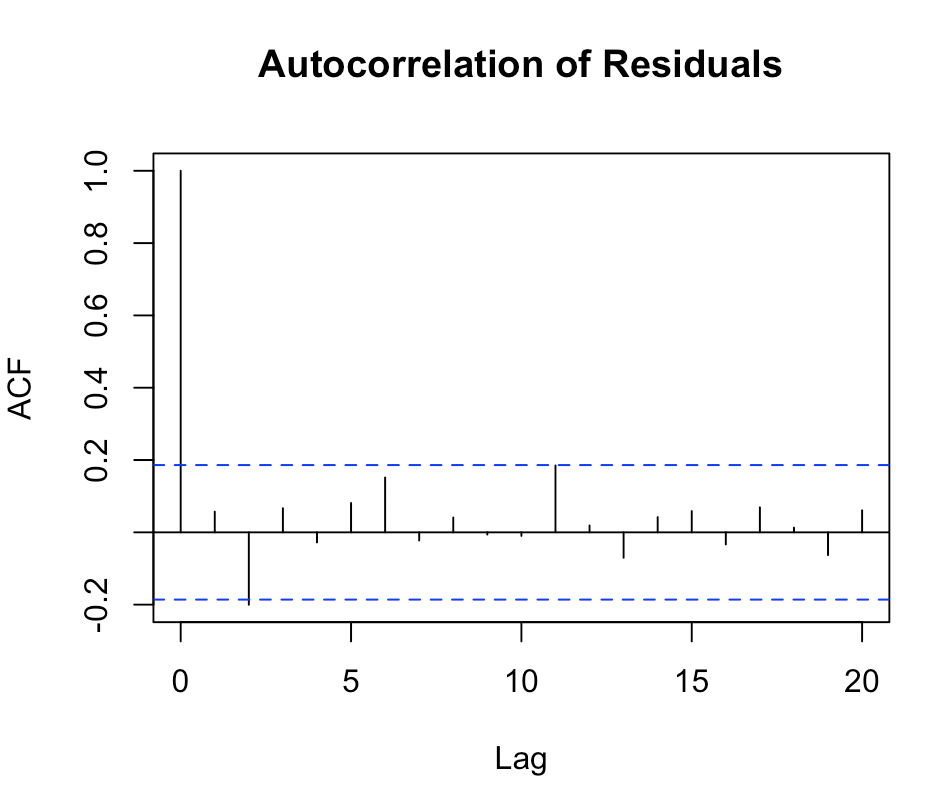
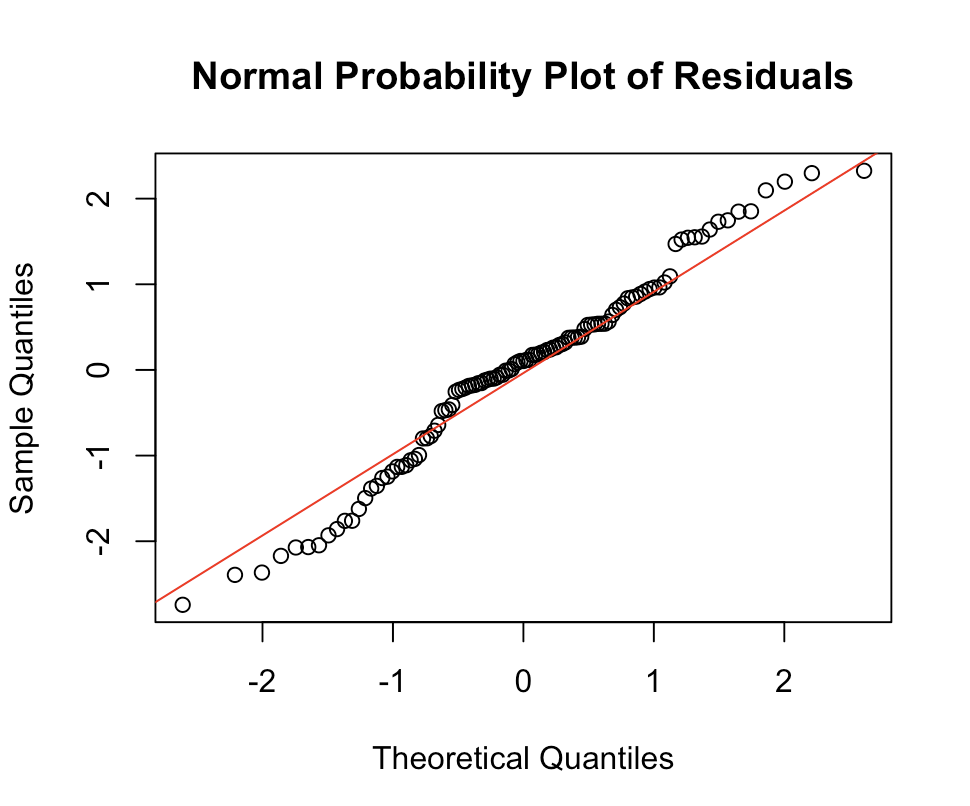


(c) The estimated slope is 0.37422.

(d) The is 0.2846.

(e) The following plots agree with the assumptions for a regression model. The relationship appears to be linear. The residuals fall in a horizontal band around 0 with no apparent pattern and maintain a constant variance. The ACF indicates no significant correlations between residuals. The normality plot indicates no significant deviation from normality assumption.

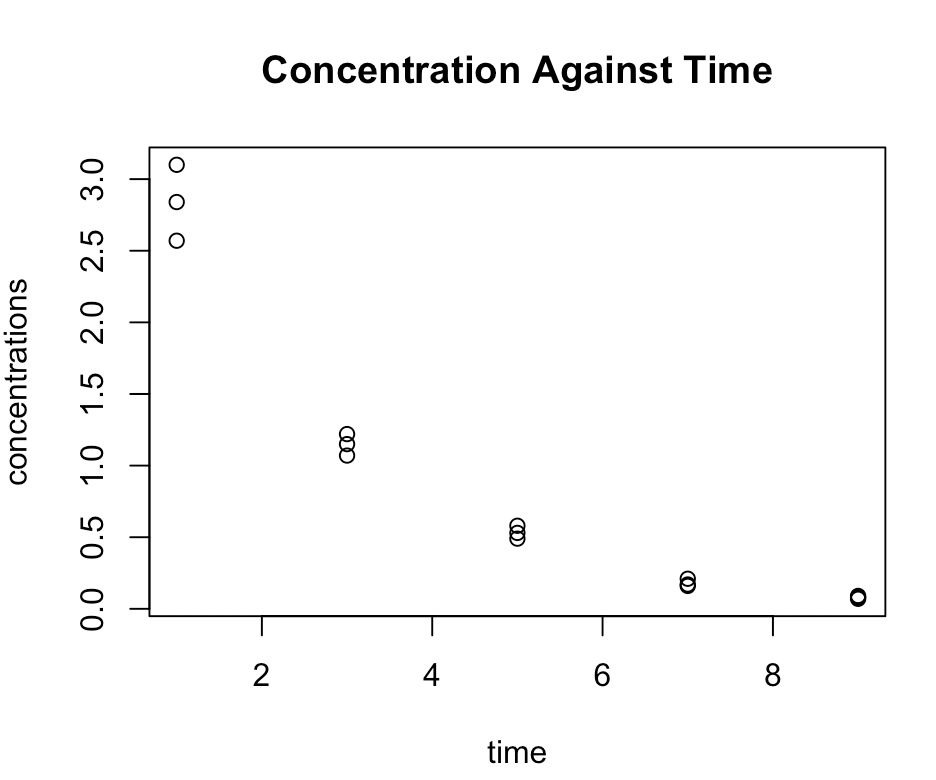
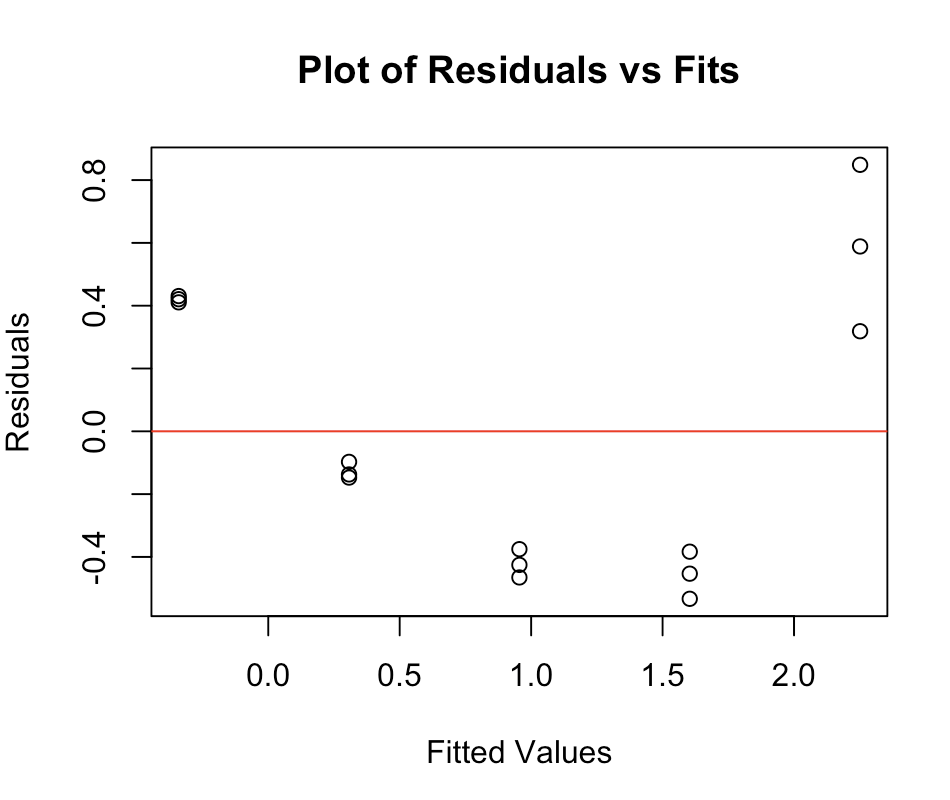
 

(f) The estimated slope is 0.49515, which is greater than the one from 1 (c).

(g) The is 0.3019, which is greater than the one from 1 (c).

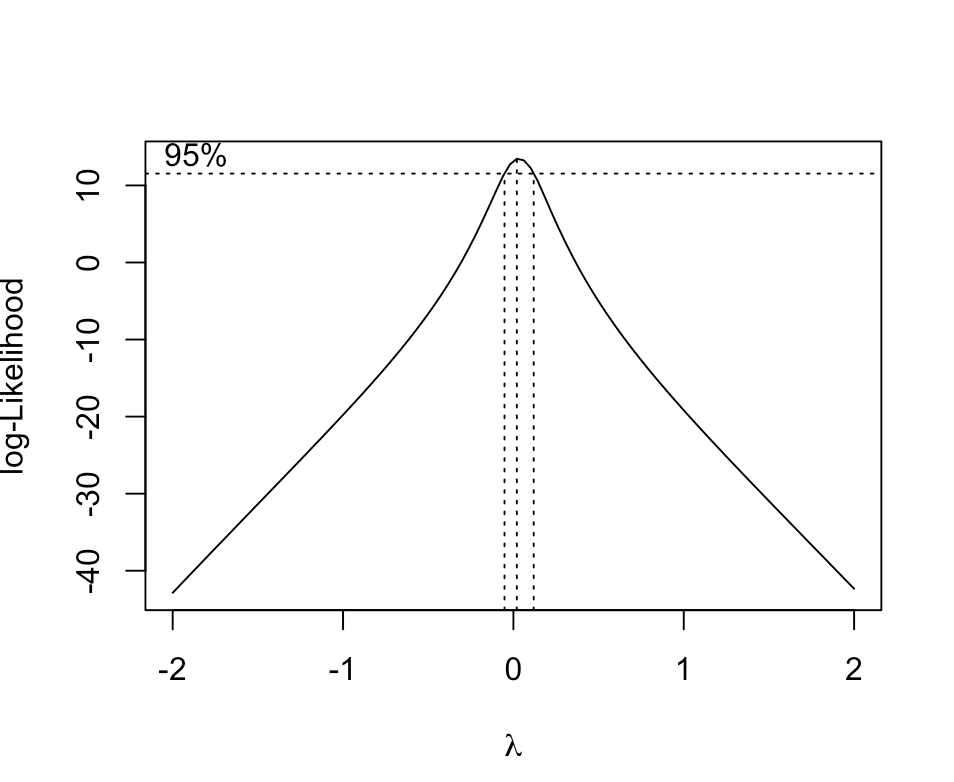
(h) Though the coefficient of determinant does not increase greatly, the slope change from 0.374 to 0.495. This implies that the relationship between InfctRsk and Stay should be considered separately when Stay is short (< 15 days) or long (>15 days). When focusing on the short length of stay, overall it is a good step to remove outliers.

1. (a) The relationship between solution and time is not linear and the variance of residuals is not constant.

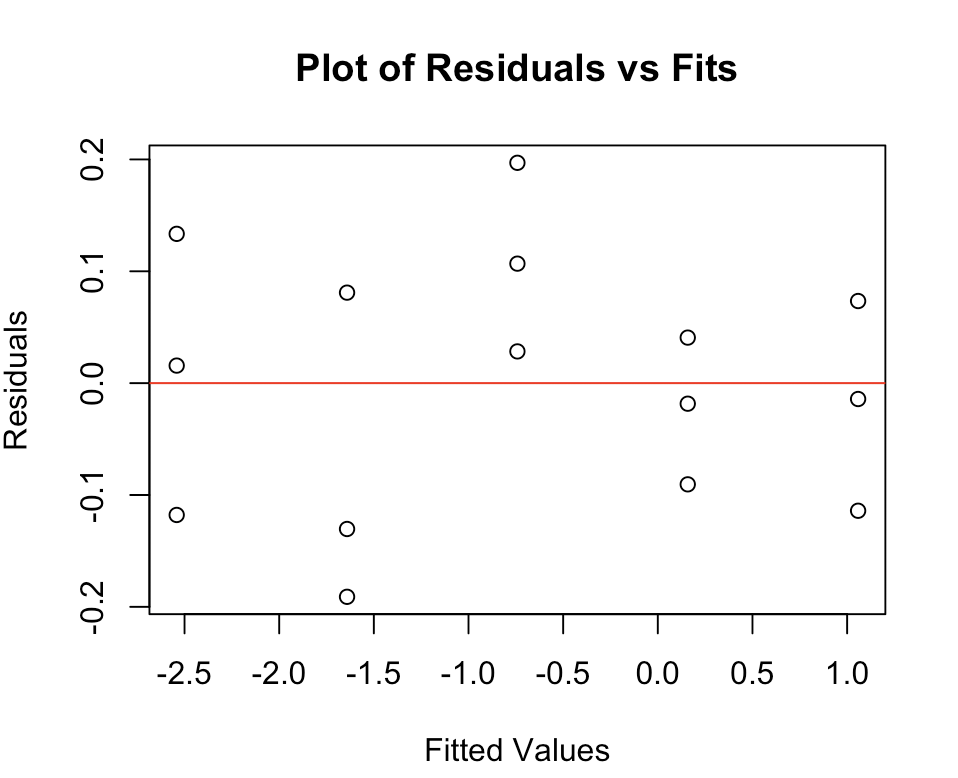
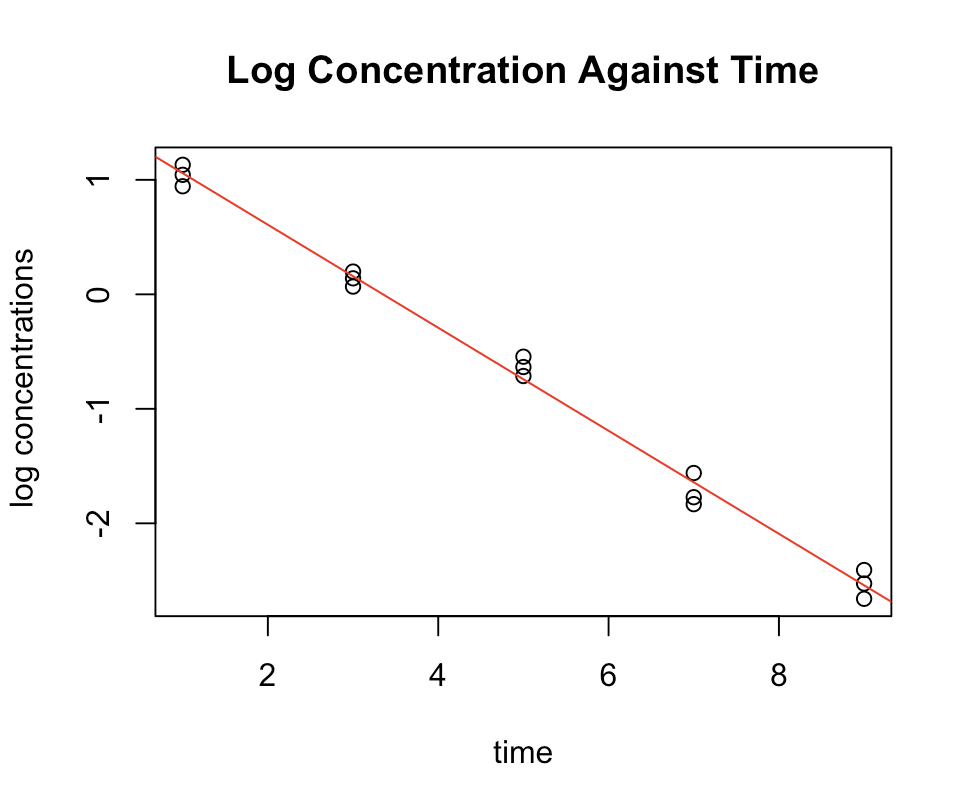
(b) I sugggest trasnforming “concentration” first, because both non-linearity and non-constant variance occur. This way can help to stablize variance and then observe the linearity to determine whether we should trasnform “time” as well.

(c) The boxcox() tells us that λ is 0 so transformation is log(y).



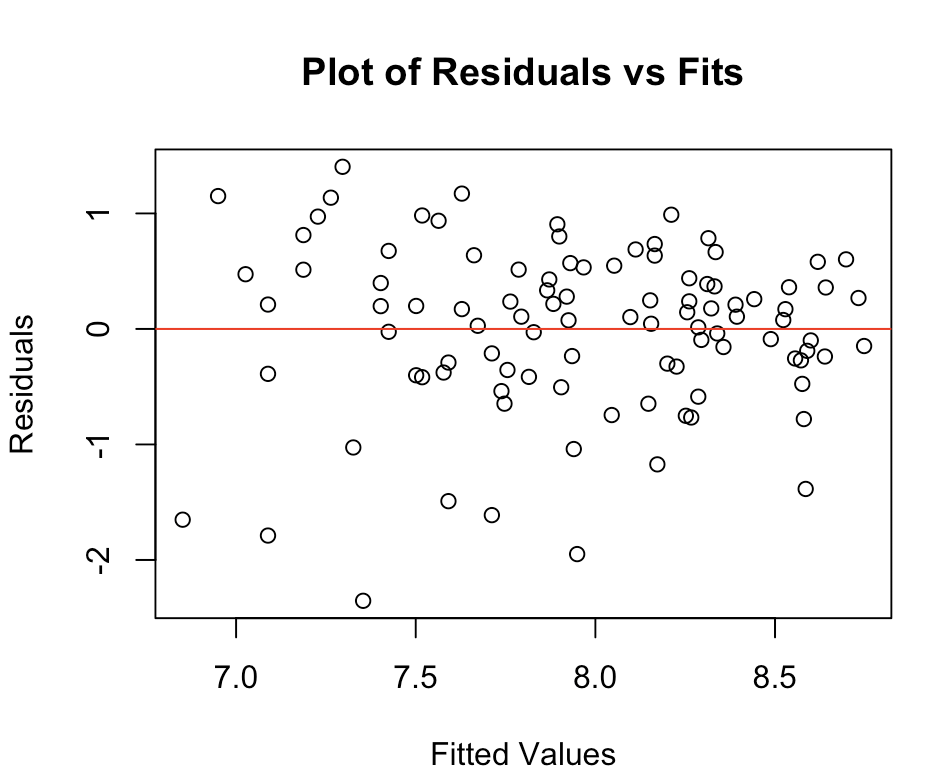
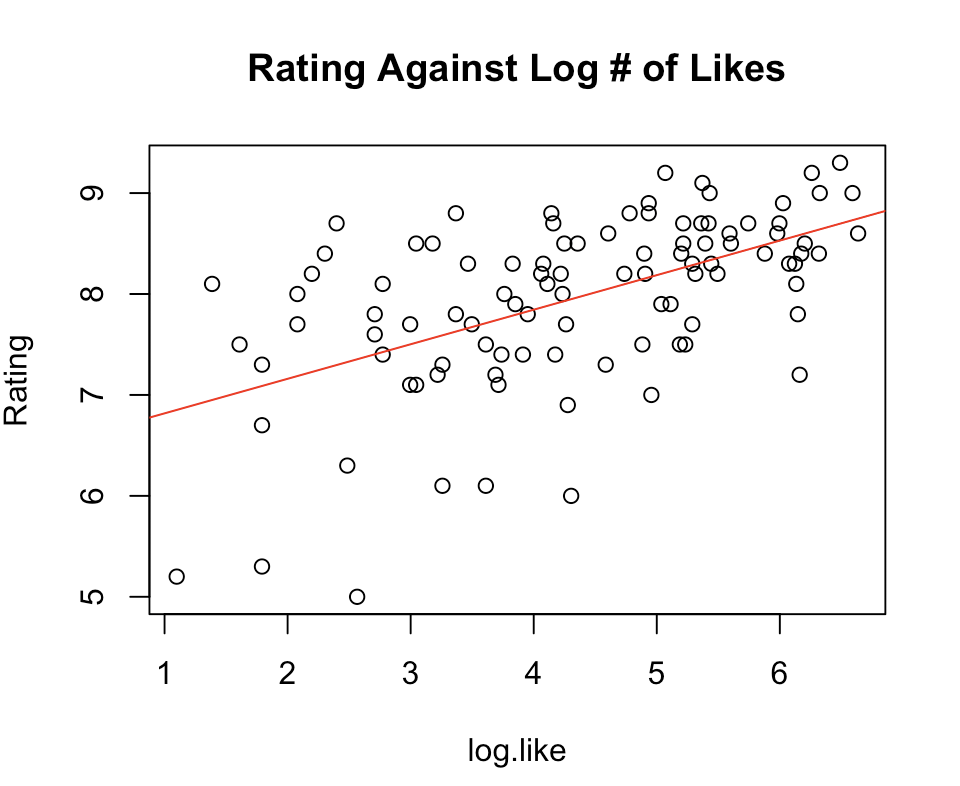
(d) The estimated slope is -0.44993. In context, a unit change in the time is associated with a multiplicative change of e^(-0.44993)= 0.6377 (about 36% decrease) in the median value of the concentration.

(e) The relationship is linear and the variance of error is constant.



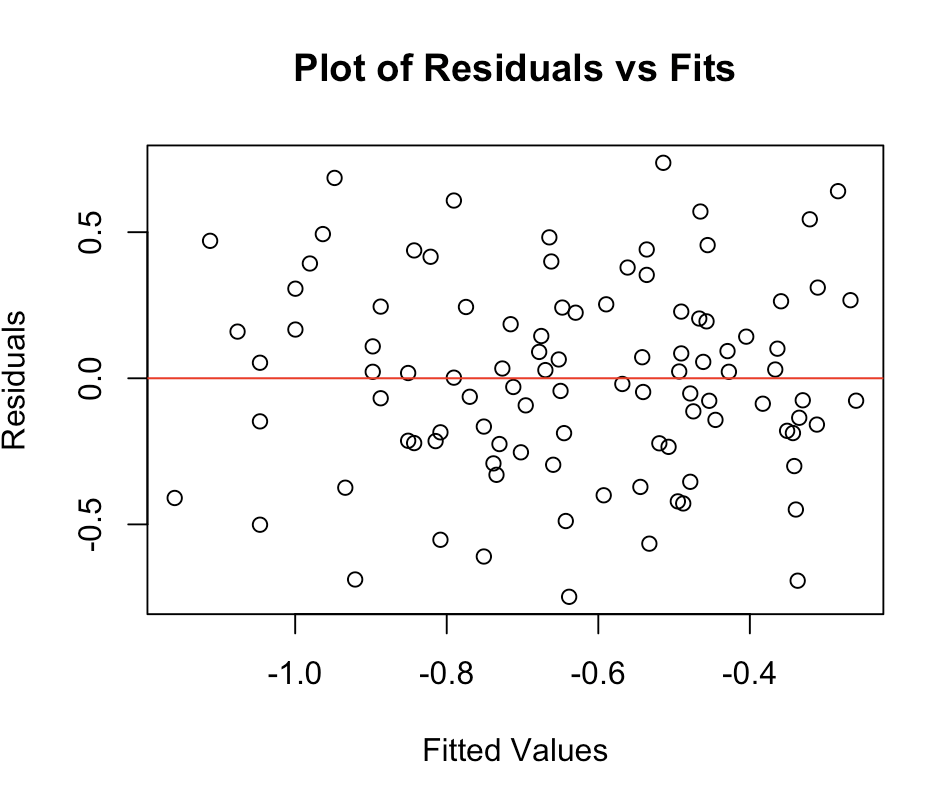
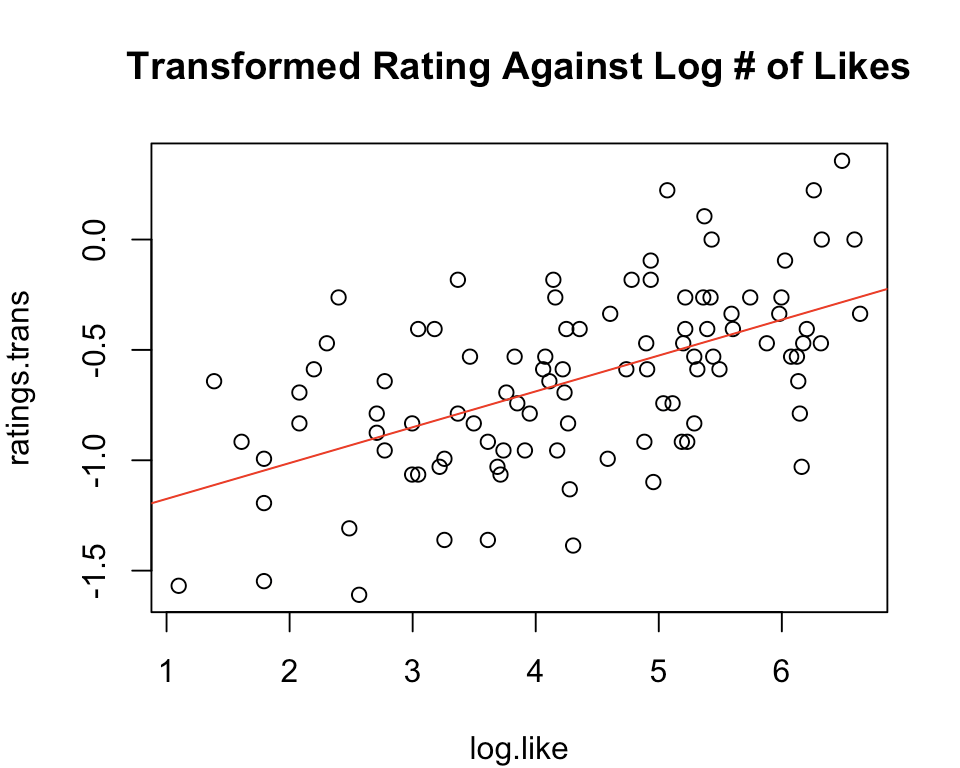
(f) The p-value given by the LOF test is 0.2217, so we fail to reject the null hypothesis. The regression equation fits the data well.

1. (a) The relationship seems to be linear but the variance of error terms is not constant.



(b) The interval is (6.61481, 9.489979)

(c) The relationship is linear and the variance of error terms is constant.



(d) The transformed interval is (-1.250347, 0.07119358). Then we transform the interval back and finally we get the predication interval (6.508447, 9.068718).

(f) As we can see from the original data, the Rating is in the range of 0 and 10, and there is low level of variability among the large values and high variability among the small values. By taking such transformation, we could make level of variability among the large values higher and level of variability among the small values lower, effectively balancing the level of variability in ratings.