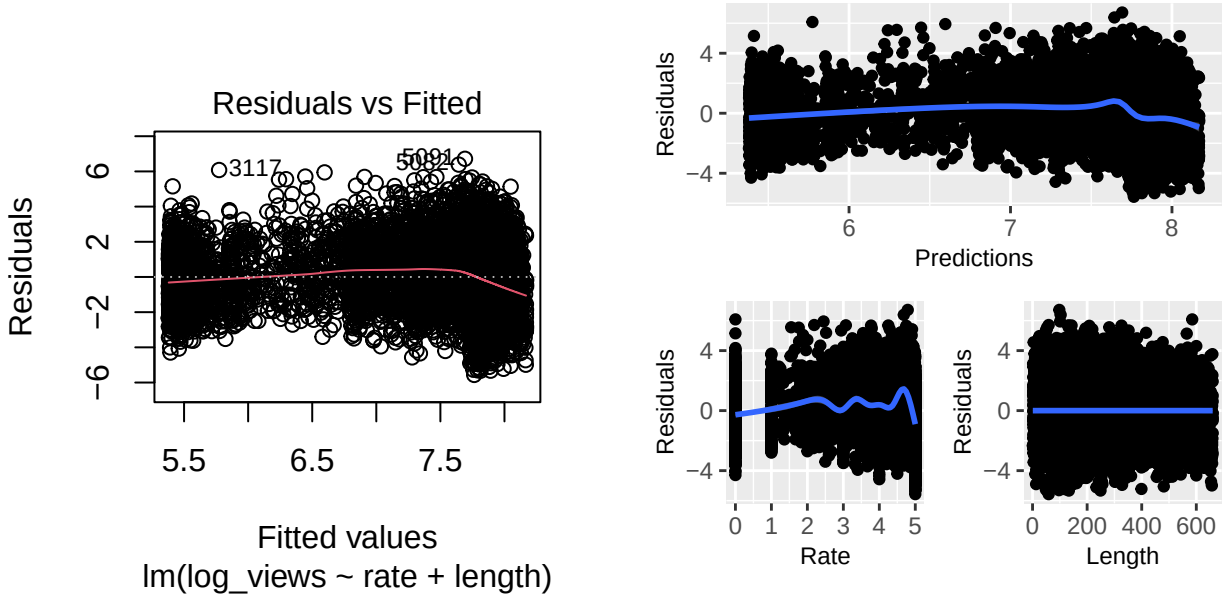
Q1.3 Linear Conditional Expectation

Assumption: The assumption of linearity means that the conditional expectation function of the response variable given the values of the predictor variables can be represented by a linear combination of the predictor variables. In other words, the relationship between the response variable and the predictor variables is linear.

Evaluating the Assumption:

To assess the linear conditional expectation of the given higher dimensional youtube video dataset, we can use the plot between predictors vs. residuals of the model. Here, a linear model is created with the predictor (rate & length) variables and target (ln(views)) variable. The residuals are computed for the predictors of the model.



Inference:

From the Residual vs Fitted and Predictions vs Residual plots, the expected mean of the residuals is around zero for most of the data points but there is squiggle at the end of the plot where most of the data points are concentrated. Visual inspection of these plots could be subjective. But if we take the squiggle at the end of the plot into consideration where most data points reside, this could be a violation in the assumption of the Linearity Conditional Expectation.

Note: The columns having null values are dropped in the dataset. To make the visualization better, the length column in seconds is transformed to minutes. The median views for videos greater than 11 min and videos smaller than 11 mins are very different. This is evident from histogram distribution of the video length variable which is right skewed, hence, the video length greater than 11 min are not significant and are dropped from dataset.