Unit 12 HW: The Classical Linear Model

1.3 - Linear conditional expectation

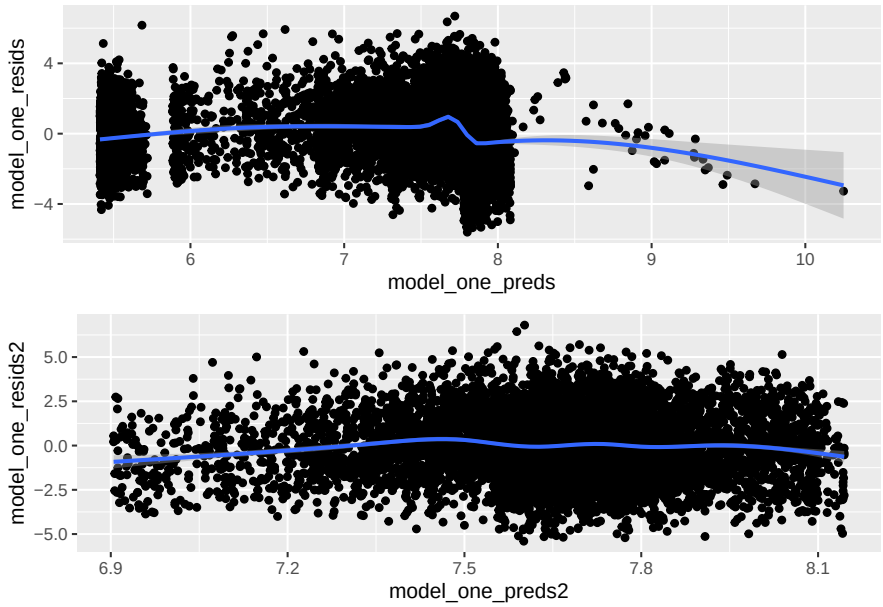

Figure 1: Predicted vs Residuals - Pre and Post Cleaning

By the above *Residual vs. Estimators* graphs, it's possible to note that a non-linear relationship to the data. That said, linear conditional expectation assumption is violated. Specially toward the right end of the data (around 7.5 and on), our model seem to shift pattern and start to produce different patterned residuals. In order to go around this, one could try to breakdown the model in two sets, as they appear to behave distinctively, or maybe a different variable transformation, hoping to end up with a more linear pattern. For the sake of this analysis, by excluding what appears to be outliers on the far right, and also removing videos without a rate (rate == 0), it is possible to achieve a much more reasonable curvature as shown on the lower graph, although not really linear. That said, it is important to say that some rows were excluded, it is still a big sample with 8068 observations.

1.4 Homoskedastic Error

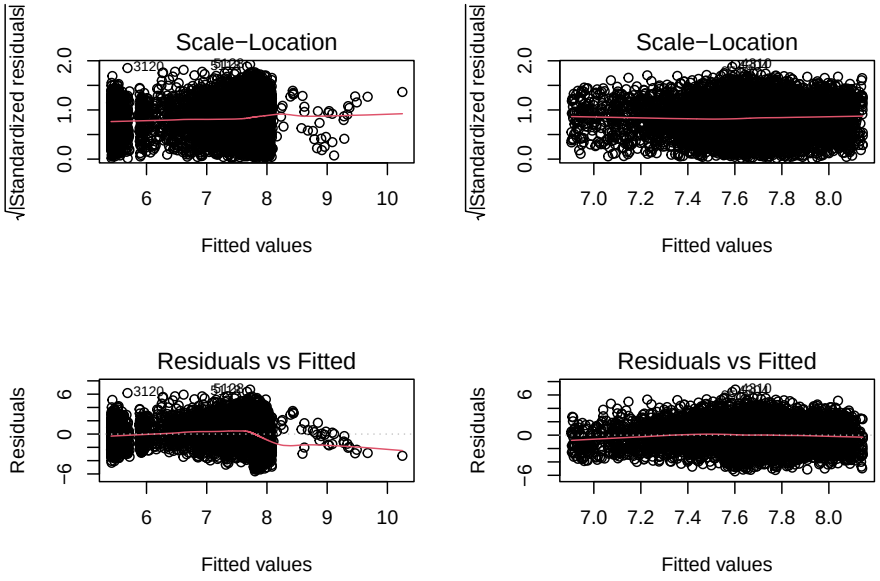

Figure 2: Homoskedastic Errors on Residuals vs Fitted

To assess whether the distribution of the errors is homoskedastic, we can examine the residuals versus fitted plot again and also run a Breusch-Pagan Test. With a strong p-value significance (p-value $<$ 2.2e-16 when using the whole dataset, and a little more modest p-value = 0.003997 when removing data that appears to be outliers and zero - as described in 1.3), the null hypothesis is rejected, leading us to believe that Heteroskedastic errors are present. By the plot above, it does look there is some in balance in the variance of the residuals specially after the before mention tipping point. Meaning that not only the linearity is compromised, but also pointing out to the presence of different variance patterns. Again, as mentioned in 1.3, the data on the far right of the set, seems to compose a group of outliers, or just differ from the rest of the set. Those are long videos (over 11 minutes), and when excluded, although still display a larger variance on errors on the mid and right portion of the graphs, the heteroskedasticity appears more mild. All that said, the assumption was not met. In order to correct this, there's the possibility to use a robust covariance matrix which in mild conditions could help with this issue.