Regression Project

Analysis 1

My sample contains 1168 tweets, of which 464 contain at least one hashtag. The maximum likelihood estimate of θ is 0.3972603.

The observed value of the test statistic for Test A is 25.3385, and the resulting *p*-value is 0. The observed value of the test statistic for Test B is 467.3876, and the resulting *p*-value is 0.

For tests A and B, we conclude that there is strong evidence against the null hypothesis that Theta = 0.14.

I was not surprised by how similar my test results were, because the sample size is 1168. As the sample size is this large the central limit theorem and the likelihood ratio statistic converge to a similar p-value.

Analysis 2

I will be analyzing the Square Root transformed variate.

The value of μ_0 is 3.

To test H_0: $\mu = \mu_0$, we calculate the observed value of the test statistic using |Y - μ_0|/(s/sqrt(n)), where, n is the sample size of the distribution; s is the sample standard deviation of the distribution, Y is the sample mean of the distribution, sqrt(n) is the square root of n, |Y - μ_0| is the absolute value of the difference of Y and μ_0. The value of the test statistic for my sample is 15.66575. To calculate the *p*-value we 2*(1-P(T <= |Y - μ_0|/(s/sqrt(n)))) for T belonging to the t-distribution with n-1 degrees of freedom, and the resulting *p*-value is 0.

Based on the results of Analysis 2c, I conclude that there is very little probability of observing data greater than the test statistics with the test hypothesis of μ = μ_0. As the p-value is 0, hence there is very strong evidence against μ_0.

Analysis 3

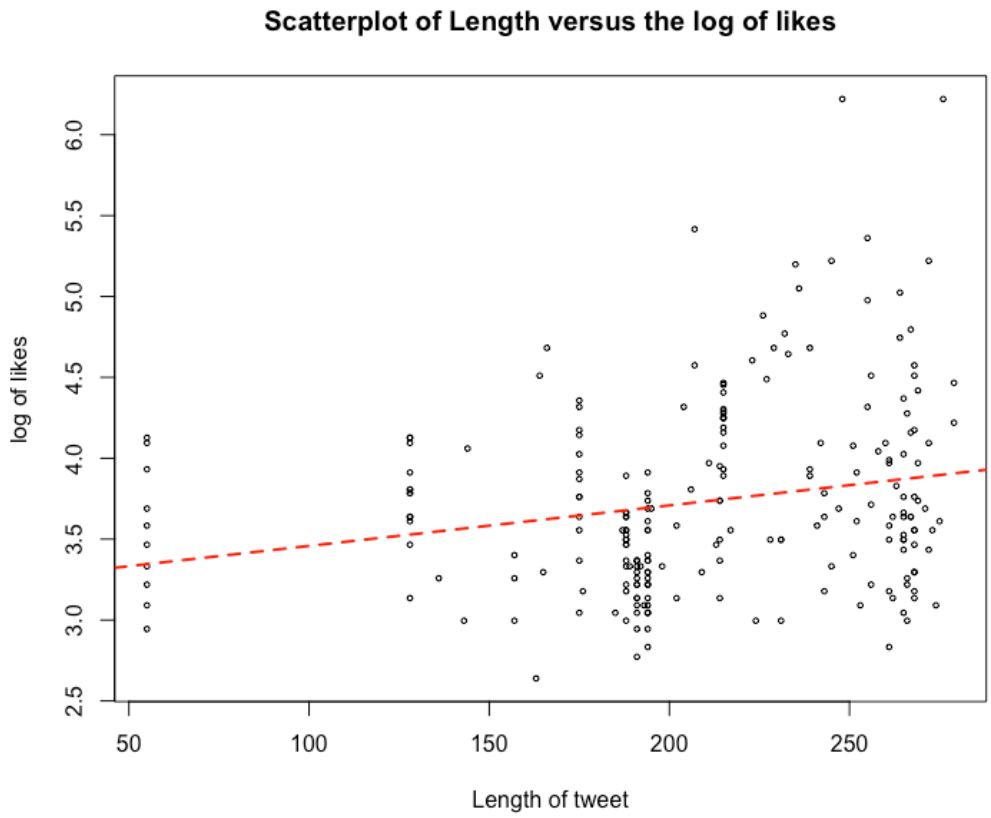
I will be analyzing CPHO_Canada's tweets.

The least squares estimate of α is 3.2076781, with a 95% confidence interval [2.899731417, 3.515624795]. The least squares estimate of β is 0.0025046, with a 95% confidence interval [0.001067099, 0.003942161].

The estimate of σ is 0.5773665.

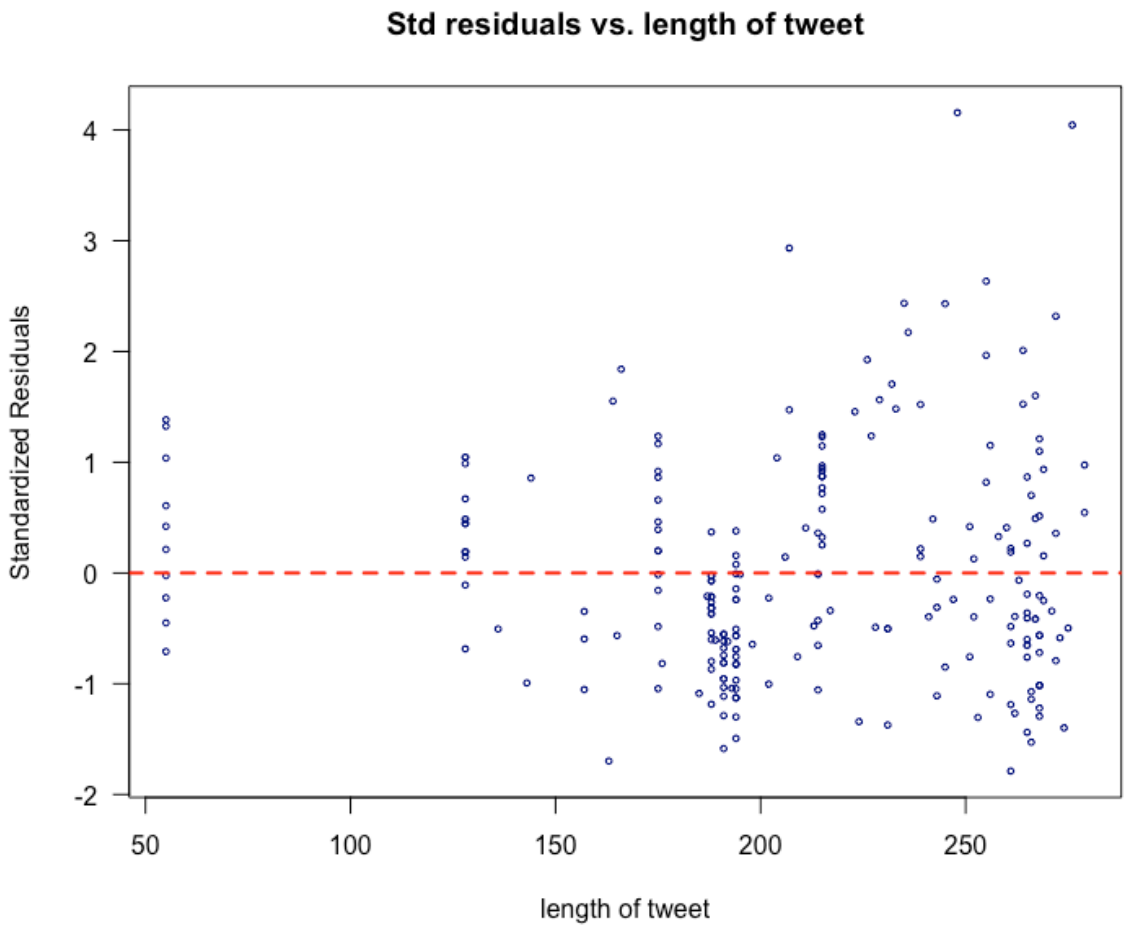
In the context of this study, α is uninterpretable. This is because α is the numerical value of the log of likes when the length of the tweet is zero. But for our sample, the length of tweets is in the range of [55,279]. Hence, we can never predict the value of α in the study.

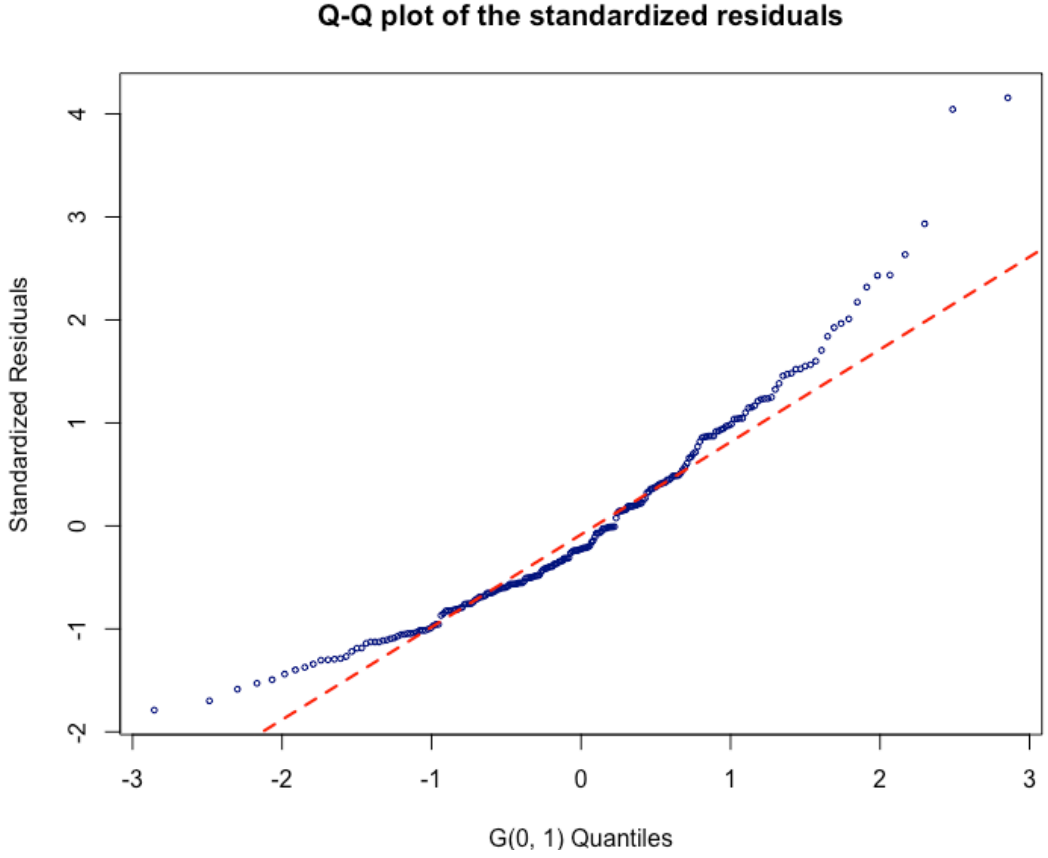
Scatterplot with the fitted line:



Scatterplot of likes.log vs length

Standardized residual plots:





3f 4

The linear model assumes Linearity, Independence, Normality and Equal variance. If these hold, we would expect to see a QQ plot that is completely linear and a standard residual vs length graph that is very close to the regression approximation with equal variance with respect to the length of the tweet. For my sample, we observe from the QQ plot that we have a right skew and the standard residuals increase with an increase in the length of tweets hence we cannot. Overall, the linear model does not seem suitable for my sample.

An estimate of the value of likes.log for a future tweet that is 200 characters long is 3.708604, with a 95% prediction interval [2.568496, 4.848712].

The p-value of a test of H_0: $\beta = 0$ is 0.0007082109. This was calculated using the t distribution with 230 degrees of freedom.

Based on the results of Analysis 3i, I conclude that there is very strong evidence against H_0: $\beta = 0$ being true. In the context of the study, there is very strong evidence against there not being a linear relationship between length and log value of likes.

Analysis 4

I will be comparing tweets without media vs. tweets with media from @CPHO_Canada's tweets.

We use an unpaired test to test H0: μ0 = μ1 because the number of tweets without media is 57 and the number of tweets with media is 175 for @CPHO_Canada's tweets. Hence pairing data is not possible as the number of observations is not equal.

The observed value of the test statistic is calculated by |y0 - y1|/(Sp*(sqrt(1/n1 + 1/n2))), where y0 is the sample mean value of the log of likes for the user = @CPHO_Canada when there is no use of media, y1 is the sample mean value of the log of likes for the user = @CPHO_Canada when there is the use of media, Sp is the pooled standard deviation calculated using sqrt((n0-1)s0^2 + (n1-1)s1^2)/(n0 + n1 - 2)), where n0 is the number of tweets when no media is used and n1 is the number of tweets when media is used, s0 is the sample standard deviation of tweets when no media is used and s1 is the sample standard deviation of tweets when media is used. We are using the square root and absolute value functions as well.

The value of the test statistic for my sample is 6.757547.

To calculate the *p*-value we calculate the 2*(1-pt(t, n1+n0 - 2)), where t is the test statistic calculated above, n0 is the number of tweets when no media is used and n1 is the number of tweets when media is used, we are finding the value of the test statistic on a t-distribution with 230 degrees of freedom, and the resulting *p*-value is 1.140124e-10.

The results in Analysis 4b rely on the following assumptions: We have an equal variance for both the log value of likes when the user is using media and when they are not. We are also assuming that the log value of likes is independent of each other.

Based on the results of Analysis 4b, I conclude the probability of obtaining data more extreme than our sample is rare. Hence there is strong evidence against the null Hypothesis. That is there is strong evidence against the sample mean value of the log of likes for the tweets with media is not equal to the sample mean value of the log of likes for the tweets without media. I would advise the Twitter user to use fewer or no media. This advice is revenant as the Chief Health Officer of Canada is liked because of their ability to explain complex concepts and give advice rather than post pictures.