Assignment 4 LaTeXTemplate

Analysis 1

1a: My ID number is 20892405.

1b: My sample contains 1133 tweets, of which 264 contain at least one hashtag. The maximum likelihood estimate of θ is 0.233.

1c: The observed value of the test statistic for Test A is 9.022, and the resulting p-value is 0. The observed value of the test statistic for Test B is 70.055, and the resulting p-value is 0.

1d: For Test A and B we conclude that since our $p \leq 0,001$, then we have strong evidence against our H_o for Test B we conclude that since $p \leq 0,001$ then ther is strong evidence against our H_o .

1e: I was not surprised by how similar my test results were, because since n is large then we would expect to have similar p-values for 2 different test statistics .

Analysis 2 2a: My ID number is 20892405. I will be analyzing the Square Root transformed variate.

2b: The value of μ_0 is 3.

2c: To test $H_0 : \mu = 3$ we calculate the observed value of the test statistic using $\frac{|\bar{Y}-\mu_0|}{S/\sqrt{n}}$, where $\bar{Y}$ is our sample mean (2.316283), S is our standard deviation (0.8353614), n is our sample size (1132) and μ_o is our null hypothesis (3). We are using this test statistic since the mean and standard deviation is unknown. The value of the test statistic for my sample is 13.986. To calculate the p-value we find solve $2P(T \geq 13.986)$ where T $t(1132)$ a t distribution, and the resulting p-value is 0.

2d: Based on the results of Analysis 2d, I conclude that there is strong evidence against the null hypothesis since we have a $p - value \leq 0.001$. This makes sense as we examine our 95% confidence interval for μ [2.220069, 2.412498], we can see that our null hypothesis is not within our interval and if it it not within our interval then $p - value$ must be ≤ 0.05 which it is.

Analysis 3

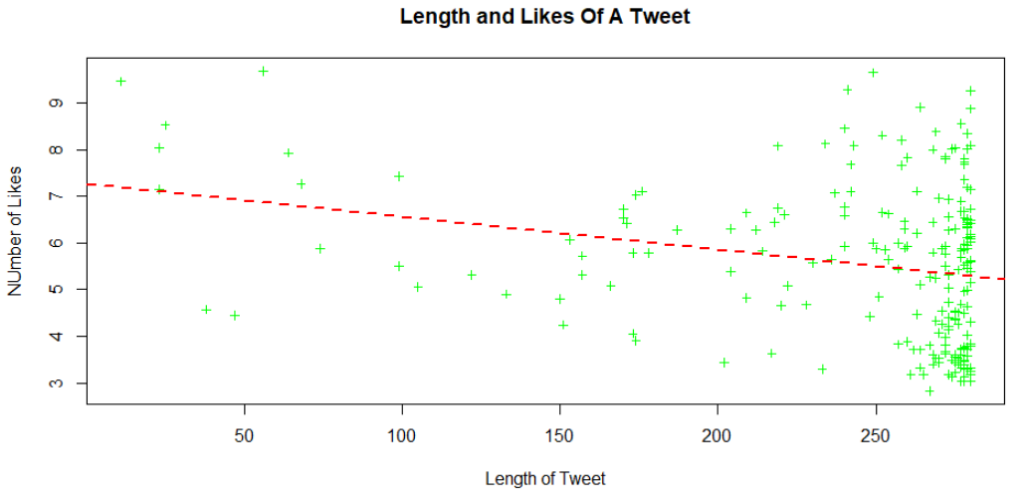
3a: My ID number is 20892405. I will be analyzing [@theJagmeetSingh]'s tweets.

3b: The least squares estimate of α is 7.263972530, with 95% confidence interval [6.3395720, 8.188373014]. The least squares estimate of β is [-0.007017032, with 95% confidence interval [-0.0106826, -0.003351467].

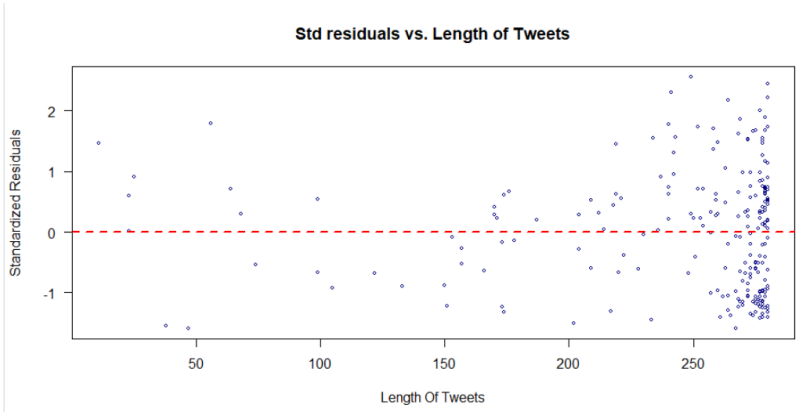
3c: The estimate of σ is 1.614346.

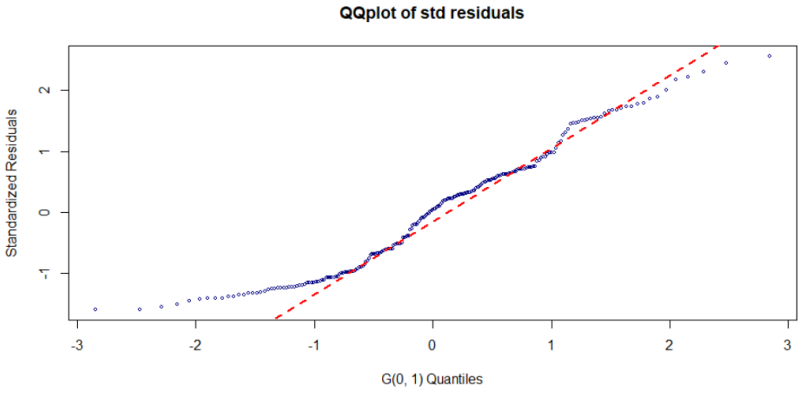
3d: In the context of this study, α represents the mean likes.log when our length is equal to 0.

3e: Scatterplot with fitted line:



3f: Standardized residual plots:





3g: The linear model assumes that our response variate can be modeled by a random variable where its mean is a linear function of the explanatory variable and has a constant standard deviation over the range of the explanatory variate. If these hold, we would expect to see our observed points along our fitted line. For my sample, we observe that the observe values were not on the line and had a large variability in distance from the line for each observed value. Overall, the linear model does not seem suitable for my sample.

3h: An estimate of the value of `likes.log` for a future tweet that is 200 characters long is 5.860566, with 95% prediction interval [2.667827, 9.053305].

3i: The p-value of a test of $H_0 : \beta = 0$ is 0.0002071595. This was calculated using t-distribution with degree of freedom of 223.

3j: Based on the results of Analysis 3i, I conclude there is strong evidence against our null hypothesis.

Analysis 4

4a: My ID number is 20892405. I will be comparing short tweets vs. long tweets from [@theJagmeetSingh]'s tweets.

4b: To test $H_0 : \mu_0 = \mu_1$ we use a paired test because the length of the short tweet can be dependent of wether a long tweet was tweeted before and there is less motivation to make another long tweet so a short tweet made and may get less/more likes. The observed value of the test statistic is calculated by $\frac{|\bar{y_1}-\bar{y_2}-(\mu_1-\mu_2)|}{\sqrt{\frac{{s_1}^2}{n_1}+\frac{{s_2}^2}{n_2}}}$, where $\bar{y_1} = 5.867862, \bar{y_2} = 5.235002$ are the sample means of short tweet likes and long tweets likes. μ_1 is the mean of short tweets likes and μ_2 is the mean of the long tweets likes. $s_1 = 1.641893, s_2 = 1.627474$ are the standard deviations of short and long tweets. Lastly $n_1 = 109, n_2 = 116$ are the sample size for short and long tweets. The value of the test statistic for my sample is 2.9018. To calculate the p-value we find the $2P(Z \geq \frac{|\bar{y_1}-\bar{y_2}-(\mu_1-\mu_2)|}{\sqrt{\frac{{s_1}^2}{n_1}+\frac{{s_2}^2}{n_2}}}) = 2P(Z \geq 2.9018)$, and the resulting p-value is 0.004085.

4c: The results in Analysis 4b rely on the following assumptions that we assumed both groups are Gaussian and we must check if that is if that is valid with numerical/graphical summaries.

4d: Based on the results of Analysis 4b, I conclude there is strong evidence against the null hypothesis. I would advise twitter users to tweet shorts tweets since we we can see they have a higher sample mean than long tweet likes. Also from our p-value there is strong evidence that our mean will be the same as the mean for long tweet. This may show that short tweets receive more likes but no guaranteed.