STAT 231 End of Term Report

Problem Part

This report is conducted to help the Canadian government to anticipate the Twitter activities of provincial health agencies during the period from 1st January 2022 to 31st March 2022. So we define the set of all tweets published by all provincial health agencies during the period from 1st January 2022to 31st March 2022 in Canada as the target population, and the unit is a tweet published by any provincial health agency during the period from 1st January 2022 to 31st March 2022 in Canada.

In this study, the data of tweets posted by those official accounts of provincial health agencies includes variates: 'urls.binary'(binary indicators of whether or not the tweet features at least one URL), 'username'(the name of the account that sent the tweet), 'time.of.day'(the time of day the tweet was published, expressed in seconds), 'likes'(number of users who have liked a tweet), 'retweets'(number of users who have retweeted a tweet), 'long.words'(the number of words 10 characters or longer in the tweet), 'hashtags.binary'(binary indicators of whether or not the tweet features at least one hashtag), 'media.binary'(binary indicators of whether or not the tweet features at least one media item), 'is.retweet'(whether the tweet is a retweet of another account). Among these variates, 'urls.binary', hashtags.binary', 'media.binary' and 'is.retweet' are all binary variates, which are taking value of 1 for yes and 0 for otherwise. We are also interested in some attribute of these variates, such as the proportion of tweets contained at least one URL (website link), hashtag or media item (likes image or videos) among all tweets published by provincial health agencies during the period from 1st January 2022 to 31st March 2022 in Canada, which is our target population, and the proportion of retweets from other accounts among all tweets posted by provincial health agencies during the period from 1st January 2022 to 31st March 2022 in Canada, which is the target population. The variates 'likes', 'retweets' and 'long.words' are discrete variates. In this study, we will focus on the mean number of users who have liked a tweet, the mean number of users who have retweeted a tweet and the mean number of words 10 characters or longer among the target population: all tweets published by provincial health agencies during the specific period separately. Finally, 'time.of.day' is a continuous variate and 'username' is a categorical variate. Meanwhile, the mean time of day that tweet was published among all tweets published by provincial health agencies during the period from 1st January 2022 to 31st March 2022 in Canada and the proportion of tweets sent by each username's account among the target population: all tweets posted by provincial health agencies during the specific period in Canada, will help us draw the conclusion of interest in this report.

Some motivating questions of different types will be used to guide a comprehensive analysis in this report as follows: (6 descriptive problems and 2 predictive problems)

One motivating question for analysis about the variate ‘urls.binary’ is:

What is the proportion of tweets contain at least one URL in the target population? This is an descriptive problem.

Two motivating questions for analysis about the variate ‘time.of.day.hour’ are:

Is the standard deviation of time of day in hours that the tweet was published by @ONThealth in the target population equals to the standard deviation of time of day in hours that the tweet was published by @GoAHealth in the target population? This is an descriptive problem.

Is there a difference between the provincial health agencies @ONThealth and @GoAHealth with respect to their mean time of day in hours that the tweet was published in the target population? This is an descriptive problem as well.

Two motivating questions for analysis about the variates ‘retweets.log’ and ‘likes.log’ are:

What is the change of mean transformed number of likes received in the target population for 1 unit increase in transformed number of retweets. This is an descriptive problem.

Suppose that 30 tweets are drawn at random from the target population, what is the total number of likes received for this 30 tweets? This is a predictive problem.

One motivating question for analysis about the variate ‘long.words’:

Suppose that one tweet is chosen randomly in the target population, what is the probability that this tweet contains at least 2 words of 10 characters or longer? This is a predictive problem.

One motivating question for analysis about the variates ‘hashtags.binary’ and ‘media.binary’:

Is the proportion of tweets contain both hashtags and medias higher than the proportion of tweets contain only medias in the target population? This is an descriptive problem.

One motivating question for analysis about the variates ‘username’ and ‘is.retweet’:

Are the probabilities that a randomly chosen tweet from the tweets in the target population for each various provincial health agency (for @GoAHealth, @HCS_GovNL, @Health_PEI, @HealthNS, @ONThealth, @PHSAofBC, @sante_qc, @SaskHealth separately) is a retweet, equal to each other? This is an descriptive problem.

This report will focus on these motivating questions and try to answer them through both numerical and graphical analysis. In addition, we should notice that these data may not be used to examine any causative problem. Because the experiment that we conducted in this report is observational, which means that we cannot control the value of explanatory variate, hence it is hard to make any inference about causation based on these data.

Plan Part

The study population for this study is a total of 21,883 tweets stored in a 'primary' dataset and posted by eight Canadian provincial health official accounts, which are Alberta (@GoAHealth), British Columbia (@PHSAofBC), Newfoundland and Labrador (@HCS-FovNL), Nova Scotia (@HealthNS), Ontario (@ONThealth), Prince Edward Island (@Health PEI), Quebec (@sante_qc) and Saskatchewan (@SaskHealth), on or after 20th April 2021, and before 20th October 2021. Notice that all tweets that were replies to other accounts were excluded and tweets that were retweets of other users were not excluded.

For consideration of time and costs , some restrictions are imposed on our study population, which may lead to study error. There exists a possible source of study error related to the variate 'time.of.day'. We want to predict the tweet activities during the period from 1st January 2022 to 31st March 2022, while all tweets in the study population are from 20th April 2021 to 20th October 2021. It is likely that the mean time of day that tweet was published is smaller (published earlier on average) in the study population than in the target population, since people usually get up relatively earlier during the warm season than freezing winter, so health official accounts might tend to post some daily information such as new COVID-19 cases report much earlier during the period from 20th April 2021 to 20th October 2021 than the period from 1st January 2022 to 31st March 2022.

To define the sampling protocol, we access a browser-based tweet downloader and entered the passwords and ID number, then the downloader generated a sample of tweets published on or after 20th April 2021, and before 20th October 2021 from each of the eight provincial health accounts from the 'primary' dataset randomly. Furthermore, there is a total of 982 tweets in our sample, including 179 tweets from @GoAHealth, 200 tweets from @HCS-FovNL, 45 tweets from @Health PEI, 117 tweets from @HealthNS, 149 tweets from @ONThealth, 124 tweets from @PHSAofBC, 40 tweets from @sante_qc and 128 tweets from @SaskHealth separately.

We should also notice the possible source of measurement error related to the variate 'long.words'. Since the length of a word includes punctuation, it is probably that some words of less than 10 characters leaded or followed by some punctuations will be identified as long words incorrectly, which causes measurement error. For example: a word with actually 9 characters and followed by a comma will be counted as a long word of 10 characters.

Data Part

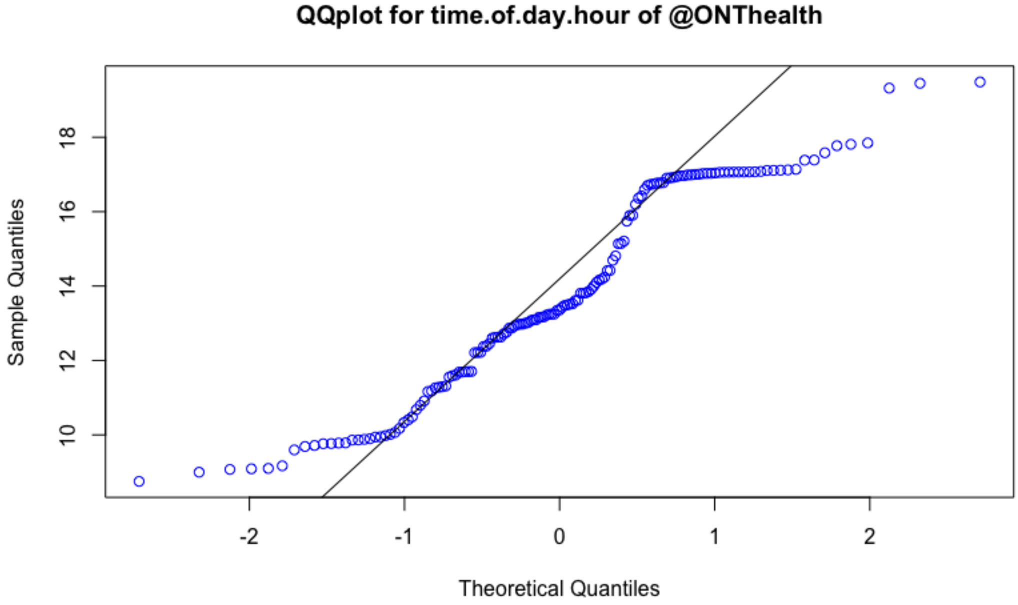
The data of tweets used in this study were downloaded by using R package ***retweet*** on October 25, 2021, and stored in a 'primary' dataset. Since mistakes can occur in entering data into data base, it is essential to put checks in place to avoid mistakes. Furthermore, a browser-based tweet downloader is used for downloading a sample of tweets from the primary dataset manually, some incorrect operation of the downloader in this step may lead to wrong data, so we should also double-check the sample generated by this downloader carefully.

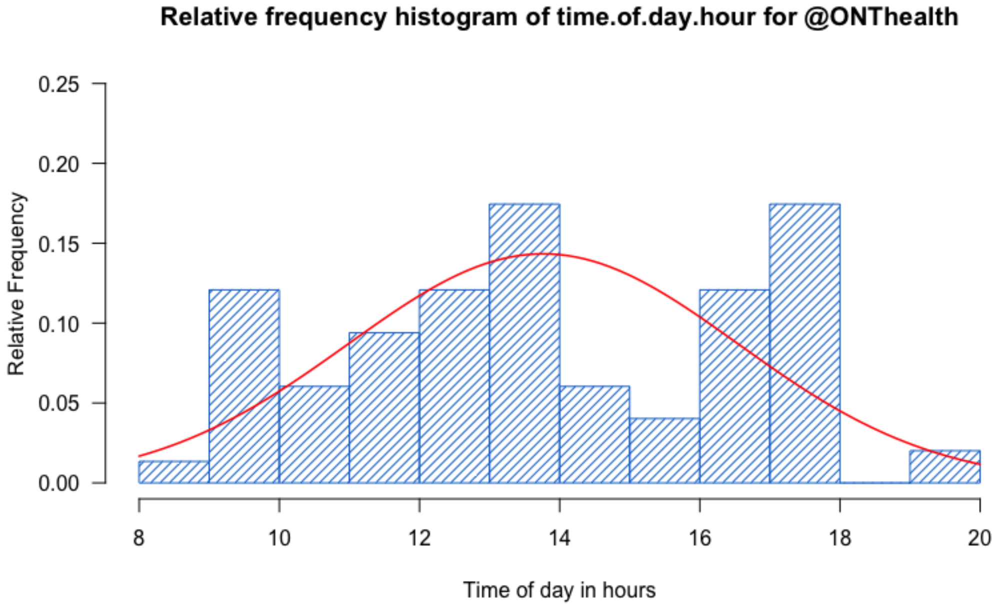
Analysis Part

Firstly, we need analysis the variate ‘urls.binary’. Since the variate ‘urls.binary’ has only 2 types of distinct outcomes: 1 for containing URL, 0 for not containing URL. Besides, the outcomes of each experiment are independent with each other and the probability of containing at least one URL in a tweet is fixed every time. Therefore, we believe that the Binomial model is an appropriate model for this variate. Assume Y=the number of tweets which contain at least on URL, and $Y \sim Binomial(n, \theta)$ where n is the sample size. Hence, the parameter θ corresponds to the probability that the tweet contains at least one URL among the total of 21,883 tweets stored in a ‘primary’ dataset and posted by eight Canadian provincial health official accounts on or after 20th April 2021, and before 20th October 2021, which is our study population. The maximum likelihood estimate for $\theta,\ \hat{\theta}$ can be obtained through the sample proportion: observed number of tweets containing URL(s) dividing by the sample size, which is $\frac{355}{982} \approx 0.361507$. And the 15% likelihood interval for θ is [0.3320578, 0.3917252]. Based on a recent study, we estimate that approximately 19% of tweets contains URL(s). So now we need test the null hypothesis: $H_0 : \theta = 0.19$ by using the likelihood ratio test statistic. And the observed value of the test statistic is $\lambda = 158.3566$, which is approximately in a Chi-squared distribution with degree freedom of 1. After using R, we got that the approximate p-value is extremely close to 0. And since $p - value\ \leq 0.001$, so we can conclude that there is very strong evidence against $H_0 : \theta = 0.19$ based on the observed data.

For analyzing the variate 'time.of.day', which represents the time of day that the tweet published in seconds, we convert it into 'time.of.day.hour' expressed as hours and focus on two accounts: @ONThealth and @GoAhealth. At first, we assume $G(\mu_O, \sigma_O)$ for the transformed variate 'time.of.day.hour' for @ONThealth and $G(\mu_A, \sigma_A)$ for 'time.of.day.hour' for @GoAhealth. The parameter μ_O corresponds to the mean time of day the tweet was published on twitter by @ONThealth, expressed as hours after midnight, among the total of 21,883 tweets stored in a 'primary' dataset and posted by eight Canadian provincial health official accounts during the specific period, which is the study population. The parameter μ_A corresponds to the mean time of day the tweet was published on twitter by @GoAhealth, expressed as hours after midnight, among the total of 21,883 tweets stored in a 'primary' dataset and posted by eight Canadian provincial health official accounts during the specific period, which is the study population. Furthermore, some numerical and graphical summaries are used to check how well the Gaussian model fits the variates for two accounts.

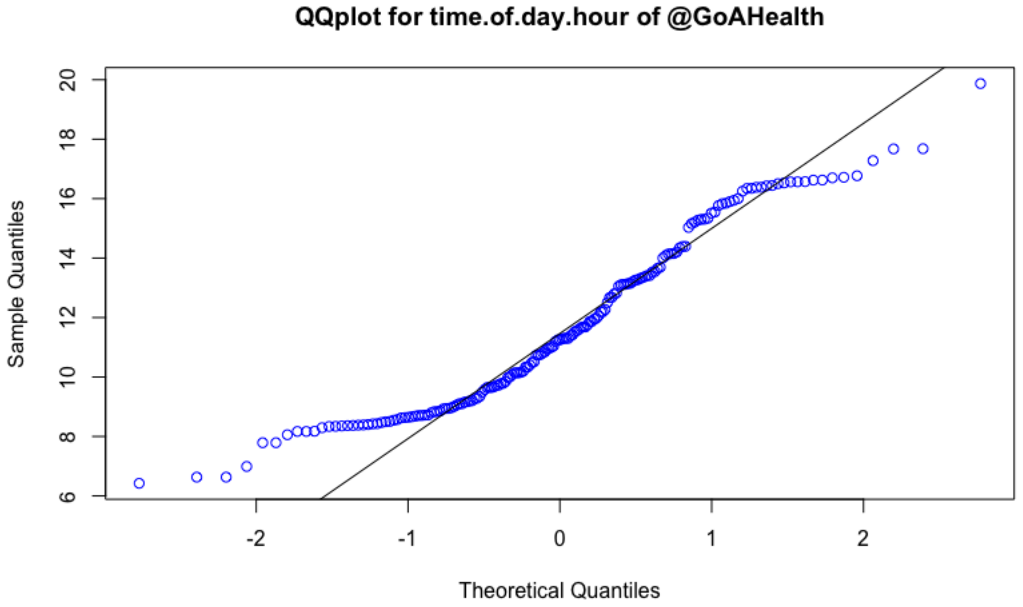
'time.of.day.hour' of @ONThealth			
Sample mean	Sample median	Sample skewness	Sample kurtosis
13.76514	13.35833	0.02883311	1.864081

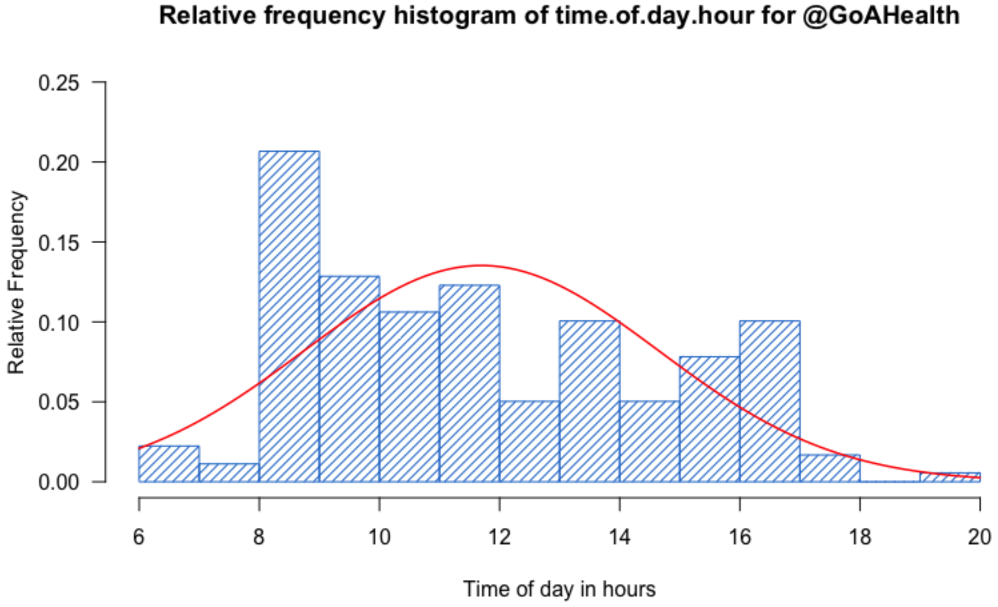




Based on summaries above, we can say that the Gaussian model does not fit the data for @ONThealth very well. The sample mean is 13.76514 and sample median is 13.35833, which are quite close to each other. Besides, the sample skewness is 0.02883311, which is very close to 0. However, the sample kurtosis is 1.864081, which is not close to 3. The set of points in qqplot seems more like a S-shape instead of lying along the straight line reasonably. Finally, it shows that there is large difference between the relative frequency histogram based on observed data and the theoretical probability density function assumed Gaussian distribution. Therefore, the Gaussian model is not reasonable for the data of @ONThealth.

'time.of.day.hour' of @GoAhealth			
Sample mean	Sample median	Sample skewness	Sample kurtosis
11.69688	11.24722	0.4451963	2.133412





Based on summaries above, we can say that the Gaussian model does not fit the data for @GoAHealth very well. The sample mean is 11.69688 and sample median is 11.24722, which are close to each other. However, the sample skewness is 0.4451963, which is not close enough to 0. And the sample kurtosis is 2.133412, which is not close to 3. The set of points in qqplot forms a S-shape instead of lying along the straight line reasonably. Finally, it shows that the relative frequency histogram based on observed data does not follow the theoretical probability density function assumed Gaussian distribution. Therefore, the Gaussian model is not reasonable for the variate ‘time.of.day.hour’ of @GoAHealth.

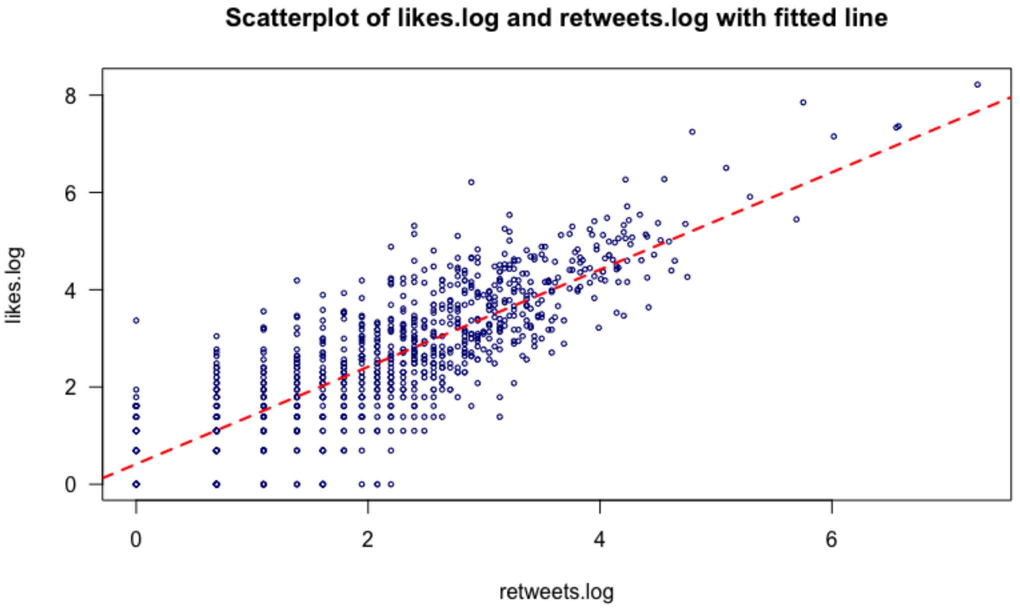
Then we construct 99% confidence intervals for σ_O and σ_A separately. The 99% confidence interval for σ_O is [2.417507, 3.265727], and the 99% confidence interval for σ_A is [2.592412, 3.409685]. Based on these intervals, since the lower bound and upper bound of 99% confidence intervals for σ_O are very close to the lower bound and upper bound for σ_A separately, which means that these two intervals have quite large overlap, so it is reasonable to assume that $\sigma_O = \sigma_A$.

Now we assume $\sigma_O = \sigma_A$ and test the null hypothesis $H_0: \mu_O = \mu_A$. The point estimate of the mean difference $\widehat{\mu_O} - \widehat{\mu_A}$ is 2.068262. The 95% confidence interval for $\mu_O - \mu_A$ is [1.441122, 2.695403]. And the p-value for the test of hypothesis $H_0: \mu_O = \mu_A$ is 3.232694e-10, which is extremely close to 0. Since the $p - value \leq 0.001$, we can conclude that there is very strong evidence against $H_0: \mu_O = \mu_A$ based on the observed data. Notice that in order to carry out these steps above, we assume two Gaussian models with same standard deviation (variance) for ‘time.of.day.hour’ for @ONThealth and ‘time.of.day.hour’ for @GoAHealth, which means $G(\mu_O, \sigma_O)$, $G(\mu_A, \sigma_A)$ for the two data groups with $\sigma_O = \sigma_A$. Besides, we also assume that both standard deviations σ_O and σ_A are unknown.

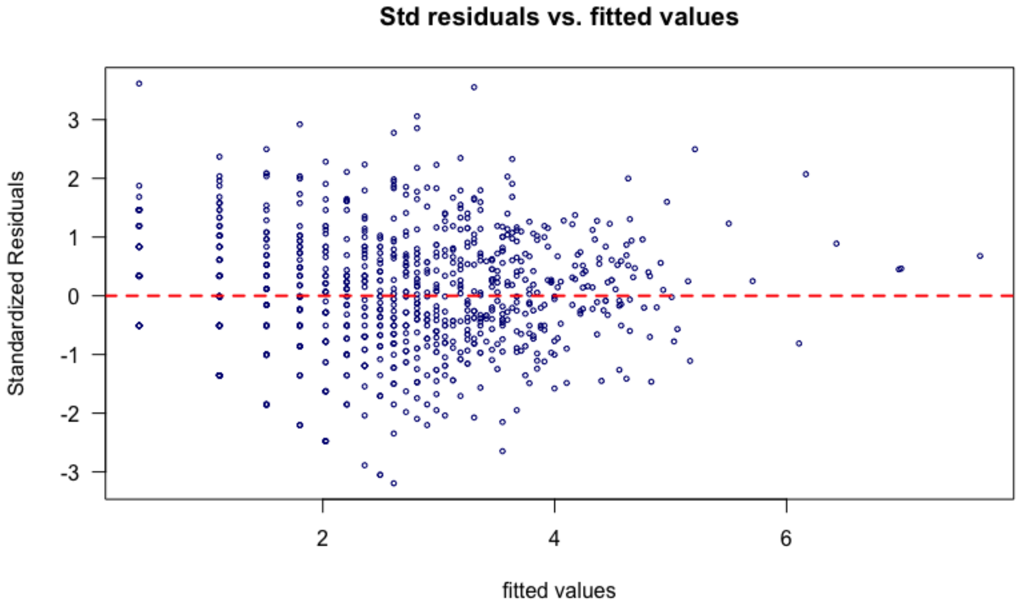
In conclusion, we test the hypothesis that no matter for health official accounts @ONThealth or @GoAHealth, their mean time of day the tweet was published in hours are the same. And the extremely small p-value means that the probability, calculated assuming the hypothesis is true, of observing a value greater than or equal to the observed value, is extremely small. But we indeed observed such a value, which is contradictory to the p-value, so we find a very strong evidence to against the hypothesis based on this observed value.

For convenience, we transform the variates ‘retweets’ and ‘likes’ into ‘retweets.log’ and ‘likes.log’ by plus 1 and taking log. To analyze the relationship between ‘retweets.log’ and ‘likes.log’, we fit a simple linear regression model $y = \alpha + \beta x$ to the data where ‘retweets.log’ is the explanatory variate and ‘likes.log’ is the response variate. By using Least Squares Estimation, we find $\hat{\alpha} = 0.41588, \hat{\beta} = 0.99960$.

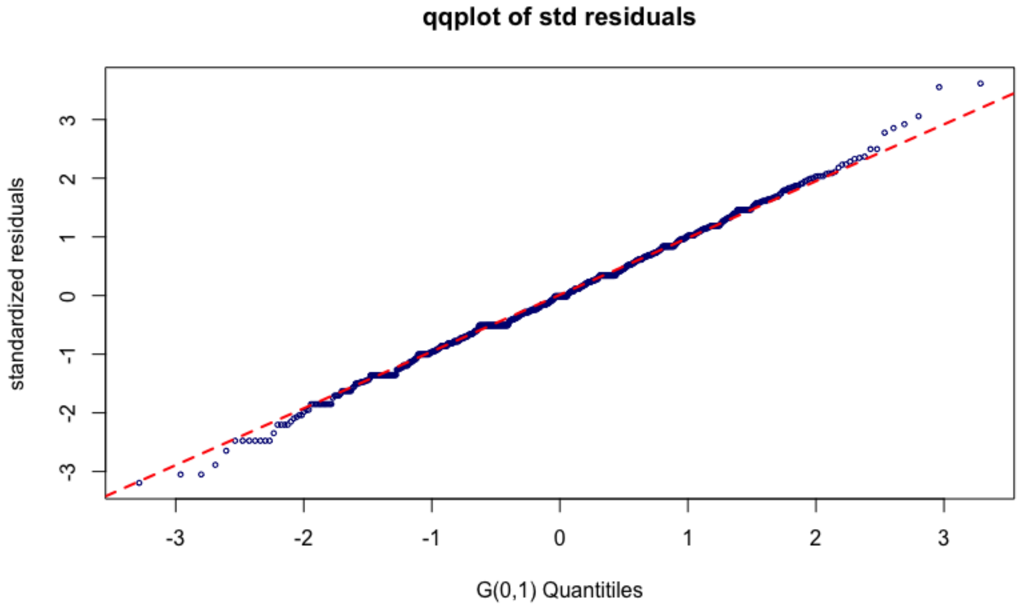
In the following, we will use a scatterplot of the data including the fitted line, a plot of the standardized residuals versus the fitted values and a qqplot of the standardized residuals to check the adequacy of the simple linear regression model.



Based on the scatterplot including the fitted line, we observe that points lie more or less along the fitted line reasonably, even though a few points are relatively far from the fitted line. Hence, the scatterplot suggests that the assumption: the mean of response variate is a linear combination of observed explanatory variate, is valid.



Based on the plot of the standardized residuals versus the fitted values, we observe that points lie more or less horizontally around the line $\hat{r}^* = 0$ with most points inside the range from -3 to 3 and approximately half the points lie on either side of the line. But the variability of points seem to be relatively larger in the middle, which means the spread of points about the line has a increase followed by a decrease as x increases and makes the 'band' looks wider in the middle and narrower on both sides. The plot verifies the assumption that the mean of response variate is a linear combination of observed explanatory variate. However, it also shows that the standard deviation is not constant over the range of values of the explanatory variate.



Based on the qqplot, we observe that the points lie along the straight line reasonably with little more variability in the tails as we expected. The plot suggests that the response variate is modeled by Gaussian distribution perfectly.

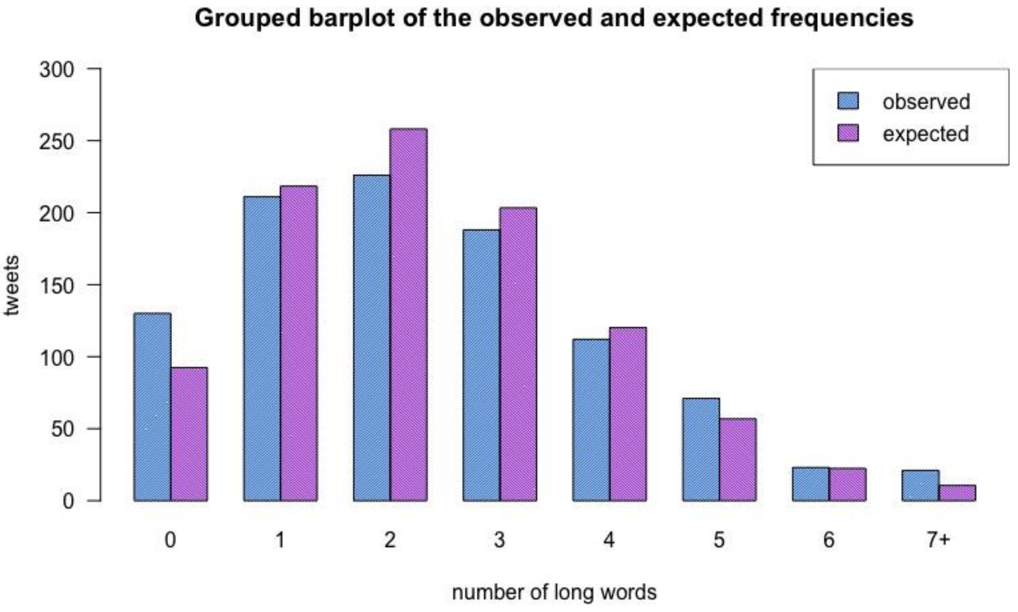
According to the plots above, we can conclude that the linear regression is not a perfect model to the data, since the response variate's standard deviation is not constant over the range of values of explanatory variate.

The parameter β corresponds to the change in mean of transformed number of users who have liked the tweet (transformed by plus 1 and taking log) in the total of 21,883 tweets stored in a 'primary' dataset and posted by eight Canadian provincial health official accounts during a specific period (which is our study population) for 1 unit increase in transformed number of users who have retweeted a tweet (transformed by plus 1 and taking log). And the 95% confidence interval for β is [0.9545545, 1.0446468]. Since $\beta = 1$ is inside the 95% confidence interval for β, so we can say that the p-value for testing the hypothesis $H_0: \beta = 1$ is greater than or equal to (1-95%)=0.05. The reason why we are interested in the hypothesis $H_0: \beta = 1$ is that we are hypothesizing that, in the study population, for every 1 unit increases in the transformed number of retweets there is 1 unit increase in the mean of transformed number of likes received. Now we assume that 30 tweets are chosen randomly from the total of 21,883 tweets stored in a 'primary' dataset and posted by eight Canadian provincial health official accounts during a specific period (which is our study population), the point estimate and a 90% prediction interval for the number of likes received with these 30 tweets is 45.92266 and [11.17848, 179.789] separately.

For analyzing the variate 'long.word', we assume a $Poisson(\theta)$ model for this variate. The point estimate of θ by using the maximum likelihood estimate is $\hat{\theta} = 2.363544$, and the approximately 95% confidence interval for θ based on the asymptotic Gaussian pivotal quantity is [2.267388, 2.459699]. Then, we apply the likelihood ratio goodness of fit test to test the hypothesis that a Poisson model is reasonable for these data, which is equivalent to $H_0{:}\,\theta_j = \frac{\theta^j e^{-\theta}}{j!}$. Following is the table of observed and expected values.

No. of long.words	0	1	2	3	4	5	6	7+	Total
Observed frequency: y_i	130	211	226	188	112	71	23	21	982
Expected frequency:e_i	92.39	218.37	258.07	203.32	120.14	56.79	22.37	10.55	982

The observed value of the likelihood ratio statistic is $\lambda = 2\sum_{j=1}^{8} y_j \log\left(\frac{y_j}{e_j}\right) = 31.06118$, which follows the approximate Chi-squared distribution with degree freedom of $8-1-1=6$. After using R, the p-value is 2.467649e-05, which is extremely close to 0. Since $p - value \leq 0.001$, we can conclude that there is very strong evidence against the hypothesis that a Poisson model is reasonable for these data. We also attach a grouped barplot of the observed and expected frequencies that we used for testing the goodness of fit.



If we assume that the Poisson model fits these data appropriately, and then a tweet is chosen randomly from the total of 21,883 tweets stored in a ‘primary’ dataset and posted by eight Canadian provincial health official accounts during a specific period (which is our study population), the point estimate of the probability that this tweet contains at least 2 long words is 0.6835369.

The two binary variates ‘hashtags.binary’ and ‘media.binary’ will be examined as following. We use the likelihood ratio test for the independence between the two variates. We create a two-way table of observed and expected frequencies. Notice that the values inside brackets are the corresponding expected counts.

	No medias	Medias	Total
No hashtags	436 (376.7475)	155 (214.2525)	591
Hashtags	190 (249.2525)	201 (141.7475)	391
Total	626	356	982

The observed value of the likelihood ratio test statistic is $\lambda = 2\sum_{i=1,j=1}^{i=2,j=2} y_{ij}\log\left(\frac{y_{ij}}{e_{ij}}\right) = 64.26768$, which follows the approximate Chi-squared distribution with degree freedom of (2-1)(2-1)=1. After using R, the p-value is 1.110223e-15, which is extremely close to 0. Since $p - value \leq 0.001$, so we can conclude that there is very strong evidence against the hypothesis of independence between the variates ‘hashtags.binary’ and ‘media.binary’. The hypothesis of independence between the two variates in this study means that whether a tweet contains hashtags or not has not effect on whether the tweet contains medias or not. Furthermore, we use the proportion of various types to demonstrate the relationship between the two variates. We find that the proportion of tweets with medias but without hashtags is $\frac{155}{982} \approx 15.8\%$, the proportion of tweets with hashtags but without medias is $\frac{190}{982} \approx 19.3\%$, while the proportion of tweets with both medias and hashtags is $\frac{201}{982} \approx 20.1\%$. The higher proportion of tweets with both medias and hashtags demonstrates that there perhaps exists positive relationship between whether a tweet contains hashtags or not and whether a tweet contains medias or not, in other words, a tweet is more likely to contain both hashtags and medias at the same time.

In the last part of our analysis, we will research the variates 'username' and 'is.retweet'. Let $Y_i\ with\ i = 1,2 \ldots,8$ represents the number of tweets which are retweets from other accounts posted by @GoAHealth, @HCS_GovNL, @Health_PEI, @HealthNS, @ONThealth, @PHSAofBC, @sante_qc, @SaskHealth separately. Assume that Y_i follows $Binomial(n_i, \theta_i)$ where n_i is the corresponding number of tweets for various provincial agencies and θ_i is the corresponding probability that the tweet from the study population is a retweet for various provincial agencies. Now, we raise the hypothesis $H_0: \theta_1 = \theta_2 = \cdots = \theta_8$, which means that for the eight official accounts of health agencies (@GoAHealth, @HCS_GovNL, @Health_PEI, @HealthNS, @ONThealth, @PHSAofBC, @sante_qc, @SaskHealth), the probabilities that a randomly chosen tweet from the tweets in the study population for each corresponding provincial health agency is a retweet, are all the same. And the maximum likelihood estimates of θ_i can be obtained by calculating the sample proportion of retweets for each health agencies, so $\widehat{\theta_1} = \frac{164}{179} = 0.9162011, \widehat{\theta_2} = \frac{155}{200} = 0.775, \widehat{\theta_3} = \frac{1}{45} = 0.02222222, \widehat{\theta_4} = \frac{31}{117} = 0.2649573, \widehat{\theta_5} = \frac{122}{149} = 0.8187919, \widehat{\theta_6} = \frac{57}{124} = 0.4596774$, $\widehat{\theta_7} = \frac{1}{40} = 0.025, \widehat{\theta_8} = \frac{31}{128} = 0.2421875.$ Meanwhile, under the hypothesis $H_0: \theta_1 = \theta_2 = \cdots = \theta_8$, the maximum likelihood estimate for θ, $\hat{\theta}$ can be obtained by calculating the sample proportion of retweets among the total tweets posted by the eight health agencies, which is $\hat{\theta} = \frac{562}{982} = 0.5723014$. Finally, the Pearson goodness of fit test is used to test the hypothesis $H_0: \theta_1 = \theta_2 = \cdots = \theta_8$. Following is the table of observed and expected values.

	@GoAHealth	@HCS_GovNL	@Health_PEI	@HealthNS	@ONThealth	@PHSAofBC	@sante_qc	@SaskHealth	Total
Observed frequency: y_i	164	155	1	31	122	57	1	31	562
Expected frequency: e_i	102.44	114.46	25.75	66.96	85.27	70.97	22.89	73.26	562

The observed value of the Pearson goodness of fit statistic can be obtained by $d = \sum_{i=1}^{8} \frac{(y_i - e_i)^2}{e_i} = 158.3283$, which follows the approximate Chi-squared distribution with degree freedom of 8-1-1=6. After using R, we get the result that $p - value$ is extremely close to 0. Since $p - value \leq 0.001$, so we can conclude that there is very strong evidence against the hypothesis $H_0: \theta_1 = \theta_2 = \cdots = \theta_8$.

Conclusion Part

Some conclusions targeted our motivating question at the beginning of the report can be drawn according to our data analysis.

What is the proportion of tweets contain at least one URL in the target population?

The proportion of tweets contain at least one URL in the total of 21,883 tweets stored in a 'primary' dataset and posted by eight Canadian provincial health official accounts (the study population) is estimated to be 0.361507, which is the point estimate obtained by our sample proportion. The uncertainty for this estimate is quite small, since the interval estimate for the proportion is [0.3320578, 0.3917252], which is narrow. Besides, the sample size for this study is 982, which is large enough. Furthermore, based on some study conducted by other researchers, we raise the hypothesis that 19% tweets contain at least one URL among the study population, but the p-value for testing it is extremely small, which suggests very strong evidence to against our hypothesis.

Is the standard deviation of time of day in hours that the tweet was published by @ONThealth in the target population equals to the standard deviation of time of day in hours that the tweet was published by @GoAHealth in the target population?

Is there a difference between the provincial health agencies @ONThealth and @GoAHealth with respect to their mean time of day in hours that the tweet was published in the target population?

The interval estimates for the standard deviation of time of day in hours that the tweet was published by @ONThealth and @GoAHealth in the study population are [2.417507, 3.265727] and [2.592412, 3.409685] separately, since the two intervals have quite large overlap, it is reasonable for us to assume that the standard deviations of the two official accounts in the study population are equal to each other. The uncertainty in the estimate is small, since the sample sizes for @ONThealth, @GoAHealth are 149 and 179 separately, which are large enough to convince us.

The difference between the mean time of day in hours that the tweet was posted by @ONThealth and @GoAHealth in the study population is estimated to be 2.068262, which is the gap between the two accounts' sample mean. The uncertainty of this point estimate is reasonably small, since the interval estimate for the mean difference constructed through 95% confidence interval is [1.441122, 2.695403], which is quite narrow. Besides, positive point estimate and interval estimate both suggest that, in the study population, the mean time of day in hours that the tweet was published by @ONThealth is larger than the mean of @GoAHealth. Furthermore, we also test the hypothesis that the two mean time of day for @ONThealth and @GoAHealth are equivalent under the reasonable assumption of same standard deviation. And the extremely small p-value makes us conclude that there is very strong evidence to against this hypothesis.

What is the change of mean transformed number of likes received in the target population for 1 unit increase in transformed number of retweets. This is an descriptive problem.

Suppose that 30 tweets are drawn at random from the target population. What is the total number of likes received for these 30 tweets? This is a predictive problem.

Among the total of 21,883 tweets posted by eight Canadian provincial health official accounts, which is our study population, the mean number of likes received is expected to increase around $e^{0.99960} - 1 \approx 1.717194733$ as response to the 1 unit increase in the transformed number of retweets. Due to the narrow interval estimate [0.9545545, 1.0446468] for this transformed parameter, the uncertainty of this estimate is reasonably small.

Suppose 30 tweets are drawn randomly from the total of 21,883 tweets, which is our study population, the total number of likes received for these 30 tweets is estimated to be 45.92266. we also notice that the interval estimate for the total number of likes received is [11.17848, 179.789]. The wide gap between the lower bound and upper bound of this interval shows that the uncertainty for the estimate is relatively large.

Suppose that one tweet is chosen randomly in the target population, what is the probability that this tweet contains at least 2 words of 10 characters or longer?

The point estimate of the mean number of words 10 characters or longer in the tweet among the study population (estimate of θ) is 2.363544, which is the sample mean, and the interval estimate is [2.267388, 2.459699]. We assume a Poisson model based on the estimate of the mean number of long word, suppose one tweet is chosen randomly in the study population, then the probability that the selected tweet contains at least 2 long words is estimated to be 0.6835369. However, as we mentioned before, it is probably that some words of less than 10 characters leaded or followed by some punctuations will be identified as long words incorrectly, so the existence of measurement error leads to relatively large uncertainty about the estimates of mean number of long words in the tweet and the probability that selected tweet contains at least 2 long words among the study population.

Is the proportion of tweets contain both hashtags and medias higher than the proportion of tweets contain only medias in the target population?

We use likelihood ratio test for the independence between whether a tweet contains hashtags and whether a tweet contains medias. The resulted p-value is extremely close to 0, which suggest that there is very strong evidence to against the hypothesis, in other words, it is likely that there exists relationship between the two variates. Besides, the point estimates for the proportion of tweets contain both hashtags and medias, the proportion of tweets contain only medias among the study population is around 20.1% and 15.8% separately, which are sample proportions. The uncertainty of these estimates is reasonably small, since the sample size for this study is quite large, which is 982. Therefore, based on the resulted p-value and the point estimates, we believe that the proportion of tweets contain both hashtags and medias is expected to be higher than the proportion of tweets contain only medias in the study population.

Are the probabilities that a randomly chosen tweet from the tweets in the target population for each various provincial health agency (for @GoAHealth, @HCS_GovNL, @Health_PEI, @HealthNS, @ONThealth, @PHSAofBC, @sante_qc, @SaskHealth separately) is a retweet, equal to each other?

For answering this motivating question, we test the hypothesis: the probabilities that a randomly chosen tweet from the tweets in the study population for each corresponding provincial health agency is a retweet, are all the same. The point estimate for such a same probabilities is 0.5723014, which is the sample proportion of retweets among the total tweets. Since the sample size is large, which is 982, the uncertainty of this estimate is reasonably small. Finally, the p-value for the hypothesis test is extremely small and suggests that there is very strong evidence against this hypothesis. Therefore, we can say that the probabilities that a randomly chosen tweet from the tweets in the study population for each various provincial health agency is a retweet, are expected to not equal to each other.

Besides, extreme values also bring us uncertainty about our result. There exists one to two tweets with extremely large number of likes and retweets, which need further study.

There are also several limitations to our conclusions. Because our samples are all selected within the period from April 20,2021 to October 20,2021, the conclusion may not apply to the tweets which will be posted from January 1,2022 to March 31,2022. It is possible that the time of day that tweet was published during warm season will be earlier than during freezing winter. Besides, there will be more concerns on public health issues, such as flu and pandemic, during freezing winter, which might lead to more likes and retweets received by tweets posted within the period from January 1,2022 to March 31,2022. In this study, our analysis is only limited to tweets which are not replies to other Twitter user, so the conclusion might be not suitable for tweets which are replies to other user among all tweets published by provincial health agencies. In most cases, the tweets which are replies to other user gain less likes and retweets than usual tweets. Finally, there are eight provincial health agencies are included in our study, so we also worry that the tweets posted by other provincial health agencies, such as Nunavut, Yukon and Northwest Territories are largely different with the eight provinces in our study due to the much smaller population and less information about public health need to be posted in these regions. Thus it might be not appropriate to apply our conclusions to all provincial health agencies.