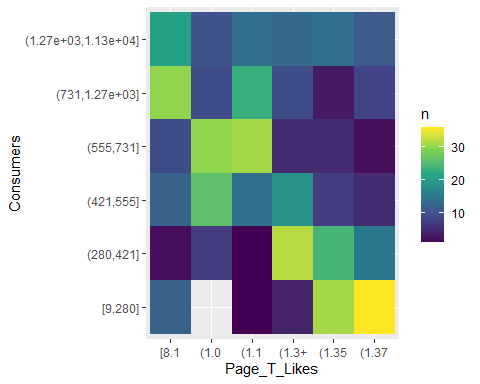
final\_code

575 C1 Team 3

2020/11/11

## Model choice and Heatmap

In previous labs, the results from OLS and QLS show that lower order least squares does not produce optimal regression fits. Considering that multiple variables could potentially affect Lifetime Post Consumers, multiple least square would be a more appropriate approach. The 7 input features are determined by the process of data collection and the nature of the Facebook dataset. Thus, Page Total Likes, Type, Category, Month, Weekday, Hour, and Paid are used as covariates. The only continuous variable among the covariates is Page Total Likes. The heatmap below does confirm the existence of a correlation with between Page Total Likes and Lifetime Post Consumers.



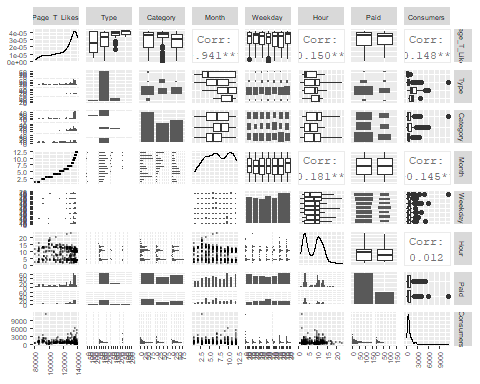
## Outliers

Before proceeding with analysis, the issue of outliers need to be addressed. For the continuous variable Page Total Likes, the calculated percentage of datapoints that lie outside of 3 standard deviations from the mean is 0, meaning that all datapoints are non-outliers.

## # A tibble: 1 x 2  
## Variable Percentage\_Outside\_3\_SD  
## <chr> <dbl>  
## 1 Page Total Likes 0

## Scatter Plots and Covariance Matrices

Notice that variables Month and Hour can only take on discrete values, and should therefore be treated as factors. However, treating them as factors triggers a cardinality threshold error in R as Hour’s factor level exceed 15. This error can be fixed by specifying the threshold, but drawing 12 or 24 boxplots are not visually significant at all due to extreme clustering and very slow processing time. Thus, it will be treated as a continuous value for both the matrix plots and the regression analysis. They will be interpreted as factors when making a conclusion.

 Page Total Like is indeed a significant predictor, noted by the 3 asterisks top the right of -0.148. Categorical variables do not have a correlation index with the response variable because the response variable is being distributed into different bins. To compare those values for significance, ANOVA should be used. Based on the scatterplot matrix and multiple experiments on various combinations of predictors, the best MLS outcome that has the most significant p values and the highest R-Square value seems to be the model with Page Total Like, Type, and Paid as predictors. The summary of the model is illustrated below.

## Performing MLS

##   
## Call:  
## lm(formula = Consumers ~ Page\_T\_Likes + Type + Paid - 1, data = fb.clean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1998.6 -287.0 -137.3 125.3 10126.7   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## Page\_T\_Likes -1.398e-02 2.173e-03 -6.433 2.99e-10 \*\*\*  
## TypeLink 1.863e+03 3.013e+02 6.184 1.32e-09 \*\*\*  
## TypePhoto 2.347e+03 2.691e+02 8.722 < 2e-16 \*\*\*  
## TypeStatus 3.758e+03 3.102e+02 12.113 < 2e-16 \*\*\*  
## TypeVideo 3.354e+03 4.135e+02 8.111 4.07e-15 \*\*\*  
## Paid1 2.067e+02 7.663e+01 2.698 0.00722 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 763.3 on 489 degrees of freedom  
## Multiple R-squared: 0.5971, Adjusted R-squared: 0.5922   
## F-statistic: 120.8 on 6 and 489 DF, p-value: < 2.2e-16

Based on the output of MLS, the model predicts the data moderately well with Adjusted R-Squared of 0.5922. It is not a high value, but there are high significance on all coefficients and the trend is clear. In this model, all p values are significant at a confidence level that’s greater than 99.9%, except for Paid1 which is significant at 99.9% confidence level. The estimated value for Page Total Likes is -1.398e-02, which means that for each unit of likes a page gets, the Lifetime Post Consumers variable decreases by 1.398e-02. Rejecting null for Page\_T\_Likes mean that Page Total Like is indeed a significant predictor for Lifetime Post Consumers. The null hypothesis on all four levels of the “Type” variable can be rejected, meaning that all types have a significant impact on the intercept of the prediction for Lifetime Post Consumers. Regression with each factor level in “Type” as baseline were also performed. All of them show that the mean of each level is different than that of all other levels. The paid posts also differs significantly with the unpaid posts, where the paid posts have a higher intercept at 2.067e+02. The F statistics tests for whether all the regression coefficients are equal to zero. In this model, the p value for the F statistics is extremely small at p < 2.2e-16. It can be concluded confidently that not all coefficients are equal to zero.

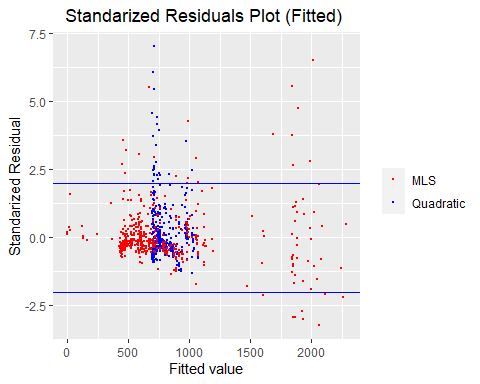
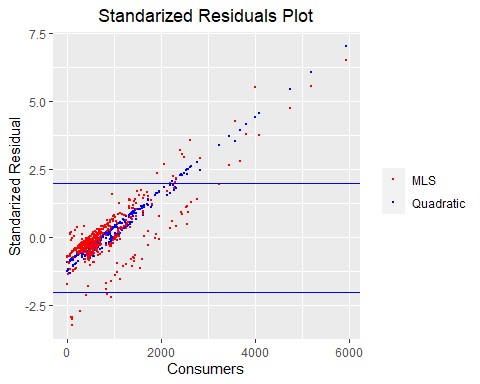
Notice that Category, Month, Weekday, and Hour were not included as part of the regression covariates due to insignificance of coefficients and low R-Square values that emerged when we were experimenting with different combinations of predictors. This outcome is indeed consistent with the correlation matrices. However, Month, Weekday, and Hour are time data that can be analyzed separately. The analysis will be carried out later in the report.

## 3D Graph for MLS output

In order to create visualization for 3 predictors, a 3D graph is illustrated below from two different angles.

## Diagnostic Analysis

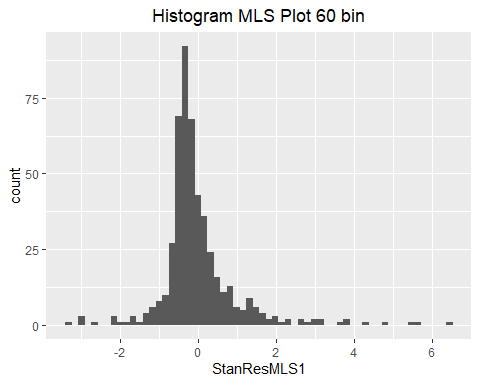
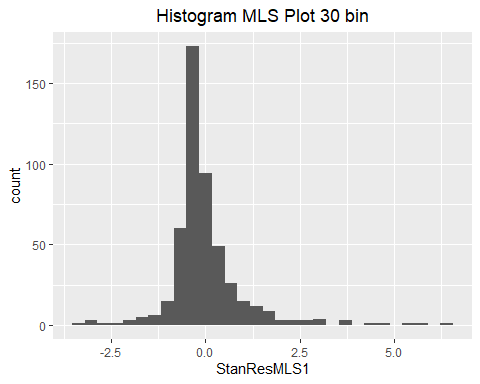
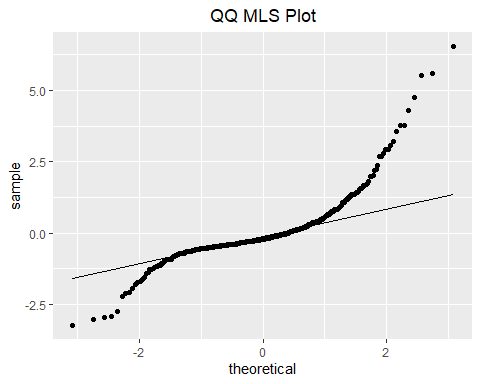
The prior analysis does not suggest for transformations of variables, therefore it would be useful to analyze the validity of the MLS. Construct residual plot and qq plot. There exist one datapoint with residual > 11000, whereas all other datapoints have a residual < 6000. The gap on the residual plot makes visual interpretation extremely difficult due to clustering. Thus, remove that outlier for now when inferring about trends in the residual plot.



On the standardized residual plot, there is a positive linear trend in both the MLS and the quadratic model, indicating that the model tend to consistently overestimate actual values as Consumers increases beyond the standard residual = 0 line. There is no good predictions for larger values of Consumers, partly because there are a lot less datapoints to generate the model from. The majority of the residual lie roughly within the horizontal band, but it clearly exhibit a non-horizontal trend. Even though residuals in both models exhibit linearity, MLS model is actually better than the quadratic model. This is because the linear trend is more scattered compared to the very concentrated line in the quadratic form. MLS residuals distributes itself almost evenly on both sides of the quadratic residuals.

The fitted model has a very nice-looking residual plot for MLS when compared to QLS. The MLS model does not exhibit any linear trend, and it is mostly distributed within the horizontal bands. There is a potential for a unequal variance, or conical trend, because variances for larger fitted values seem to be a lot more spread apart than the concentrated residuals around fitted value = 500. Based on the residual plots, the use of a constant noise model is not justified because the residual potentially violates some assumptions about the model.

To further analyze the model, construct the qq plot and histogram to check for distribution.



The qq plot does not align quite well with the line representing the standard normal distribution. There appears to be skewness on both tails of the plot, with the upper tail having a more extreme skew. The histogram indeed confirmed that the distribution is rightward skewed. The heavy tails on the qq plot also indicate that the data have more extreme values than would be expected if they truly came from a normal distribution. Under the 30 bin histogram, it appears that a significant amount of residuals fall into the residual = 0 bin. However, the rightward skewness becomes more and more obvious as bin numbers increase. Comparing the two histograms above, it seems that the mode occurs at a residual value (call it k) that’s smaller than 0, and high frequency residuals seems to cluster around this negative k value.

## Training and Validation

Previously, the MLS was generated based on all datapoints. In order to perform training and validation, randomly select half of the dataset

do ANOVA on the categorical predictors, or just include it as part of the next section? Search R ANOVA

Since Consumers includes like, comment, and shares, some metrics such as “LP\_Engage\_With\_Post” does not provide useful predictions for our response variable because they are composed of element from Consumers. Other variables such as “Comment” is simply a component of Consumers. Thus, we don’t consider these either. In other words, these variables are simply a function of Consumers.

aside from the seven predictors all variables other than the seven input features are dependent, but some might show relationships

analyze only the 3 time related variables, use ANOVA to find the maximum, posts made in what month, what day, what hour tend to be more popular? find holiday data and compare whether you should make posts on holidays. Does not exclude the fact that