**EE219 Project 5**



Popularity Prediction on Twitter

Winter 2017

Mojtaba Sahraee 204517437

Kia Karbasi 804617014

Heng Zhao 304436191

**1)**

The hashtags that were provided in the training data were #gohawks, #gopatriots, #nfl, #patriots, #sb49, #superbowl. Here are the following statistics that we obtained for each of the hashtags:

|  |  |  |
| --- | --- | --- |
| #**gohawks**  Avg tweets/hour: 193  Avg retweets/tweet: 2.014  Avg num. of followers: 2203 | #**gopatriots**  Avg tweets/hour: 38  Avg retweets/tweet: 1.400  Avg num. of followers: 1401 | #**nfl**  Avg tweets/hour: 279  Avg retweets/tweet: 1.538  Avg num. of followers: 4653 |
| #**patriots**  Avg tweets/hour: 499  Avg retweets/tweet: 1.782  Avg num. of followers: 3309 | #**sb49**  Avg tweets/hour: 1419  Avg retweets/tweet: 2.511  Avg num. of followers: 10267 | #**superbowl**  Avg tweets/hour: 1401  Avg retweets/tweet: 2.388  Avg num. of followers: 8858 |

As one can clearly see, the hashtags #sb49 and #superbowl got the most number of tweets per hour, the most number of retweets per tweet and the most average number of followers. This might be because hashtags like #sb49 and #superbowl are common hashtags and all the people watching the Superbowl are likely to be tweet them and hence they have a high chance of becoming burst hashtags.

On the other hand, hashtags like #gopatriots and #gohawks are team specific and are not likely to get retweeted by the entire population watching the game but rather only specific sections and supporters. Hence the average tweets per hour and average number of followers is low compared to the other popular hashtags.

Our histograms of the number of tweets in one hour over time for #nfl and #superbowl are shown below

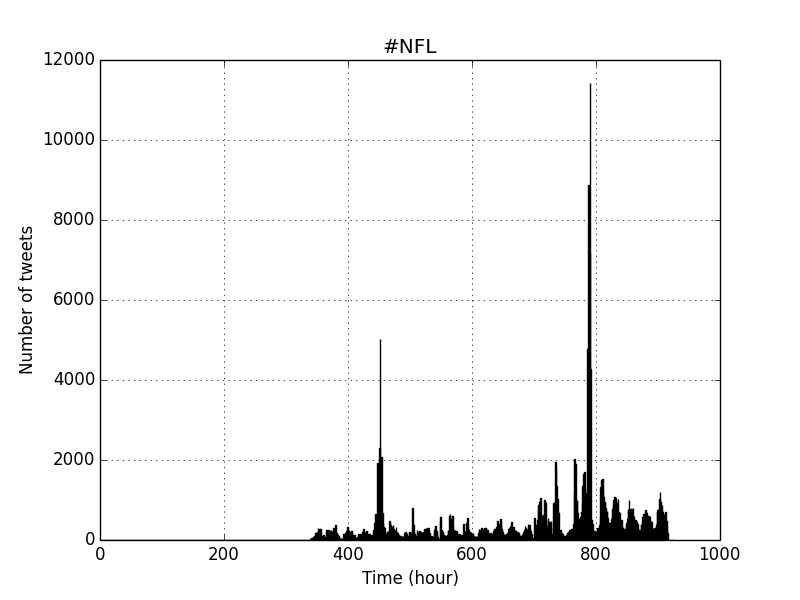


Figure #nfl over one hour bins



Figure #Superbowl over one hour bins

As is evidently seen from the graphs, we can see clear peaks during the 800th hours for both the #nfl and #superbowl. This is the burst time of the tweets and we hypothesize that this might correspond to the hours in which the superbowl match might actually have actually taking place and hence the very high number of tweets.

**2)**

A linear regression for predicting the number of tweets in the next hour was built with the features of the tweets from the previous hour as the features. The features we considered for our algorithm were the following: number of tweets, total number of retweets, sum of the number of followers of the users posting the hashtag, maximum number of followers of the users posting the hashtag, and time of the day.

So, we created a matrix in which the columns correspond to these features. Hence there are totally five columns in our feature matrix (six including the offset column). And our predictant vector will basically be the number of tweets for the next time step.

Intuitively, different types of tweets (segregated by their hashtags) should follow similar behavior. Here, we fit different linear models for each hashtag data separately. We used the OLSregression toolkit in Python to do OLS regression and we tabulate the result as follows. We use five features for our linear model. Please note that in all of our regressions models in this section,

* x1 is the number of tweets
* x2 is the total number of retweets
* x3 is the sum of the followers of the users posting the hashtag
* x4 is the maximum number of followers of the user posting the hashtag
* x5 is the time of day

Liner Model for #gohawks

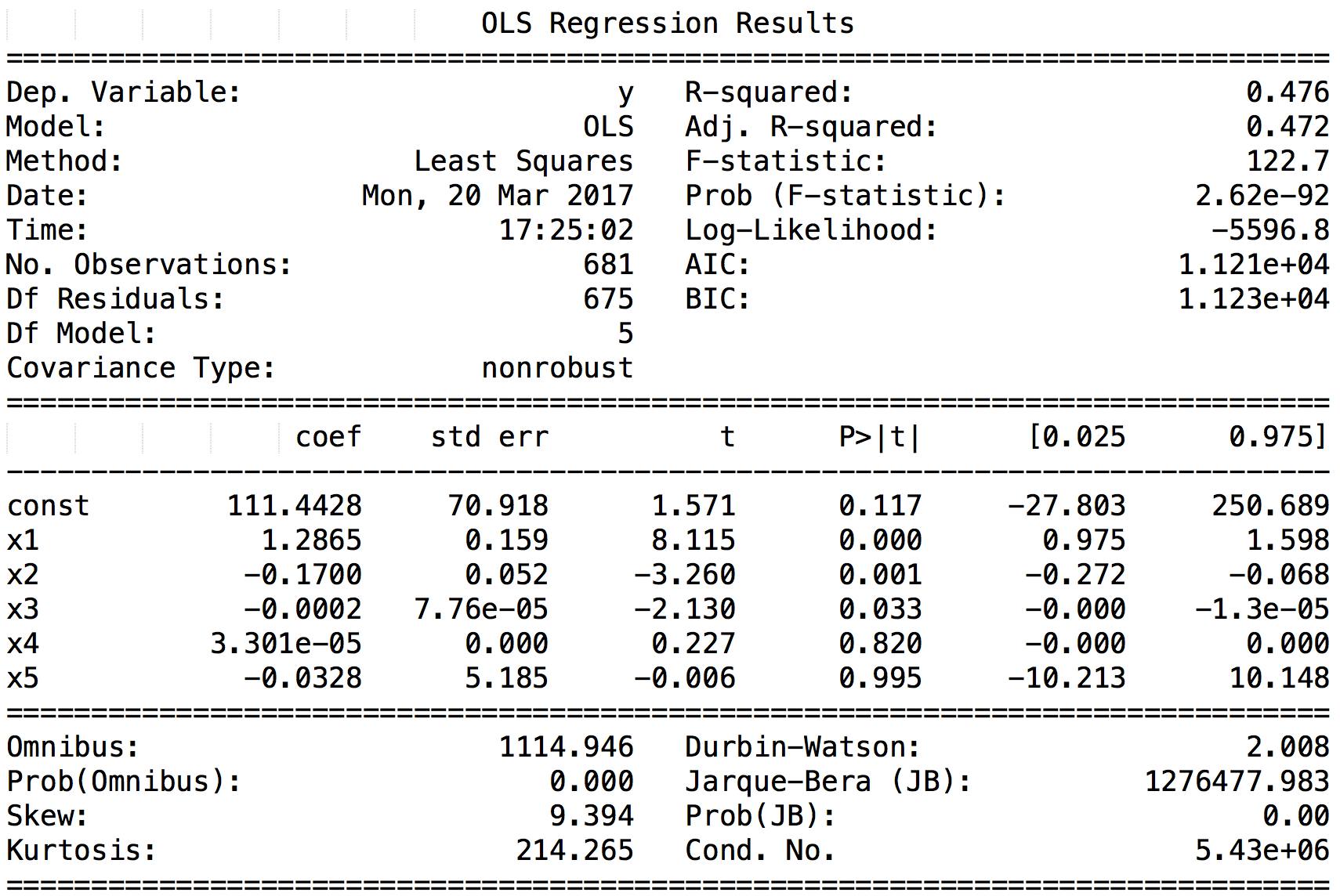


Figure 3 Regression Statistics for #gohawks

Looking at the t-value and p-value (we consider features with higher t-value magnitude and lower p-value to be significant), we can say number of tweets, total number of retweets and sum of followers of the users posting the hashtag to be significant features. For our regression model, R-squared value is 0.476, which isn't great. It's could be that we don't have large number of data-points to train or that a linear model can't explain the data well.

Liner Model for #gopatriots

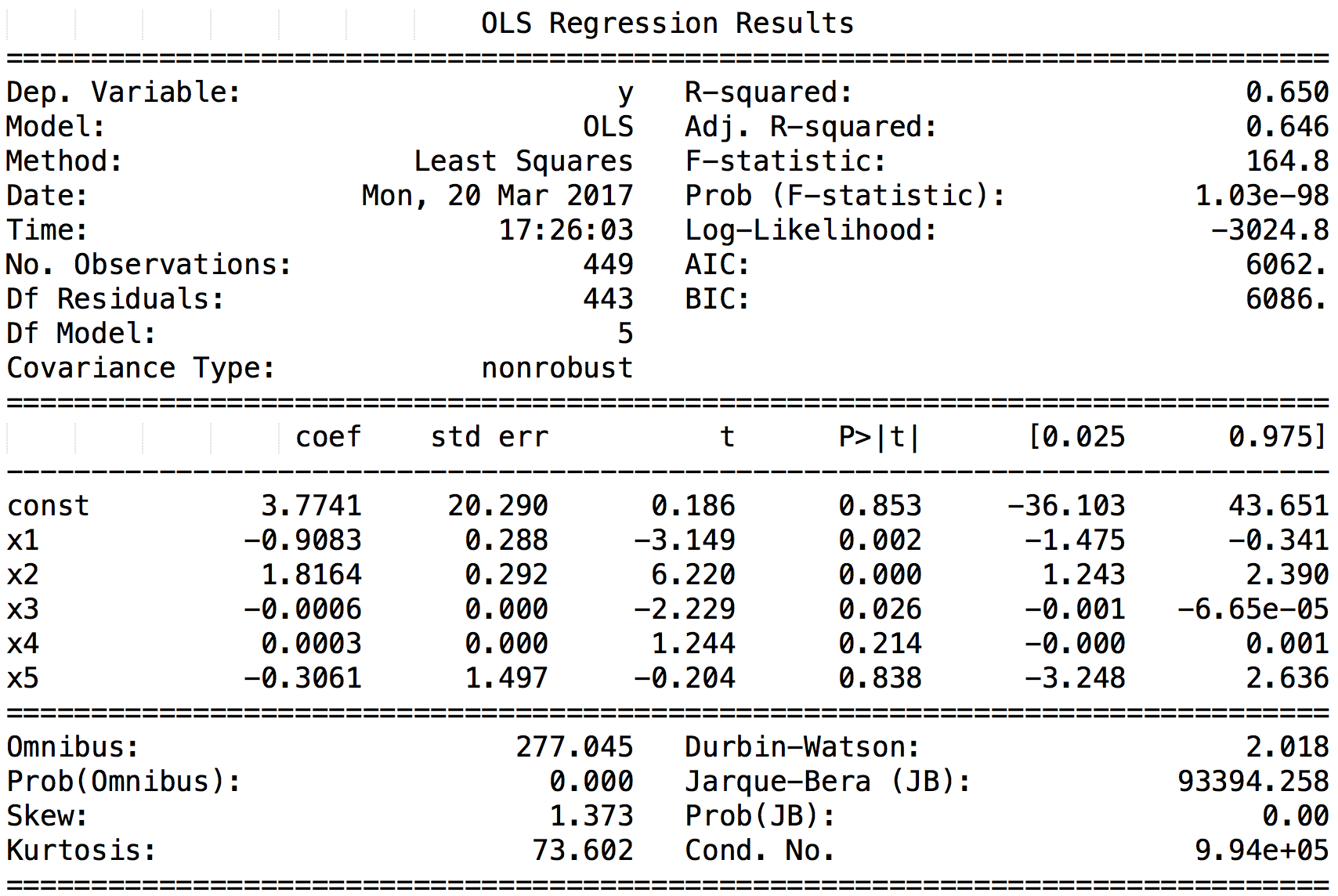


Figure 4 Regression Statistics for #gopatriots

From the table, we conclude number of tweets, total number of retweets and sum of followers of the users posting the hashtag to be significant features. For our regression model, R-squared value is 0.650 which means that our model is reasonable.

Liner Model for #nfl

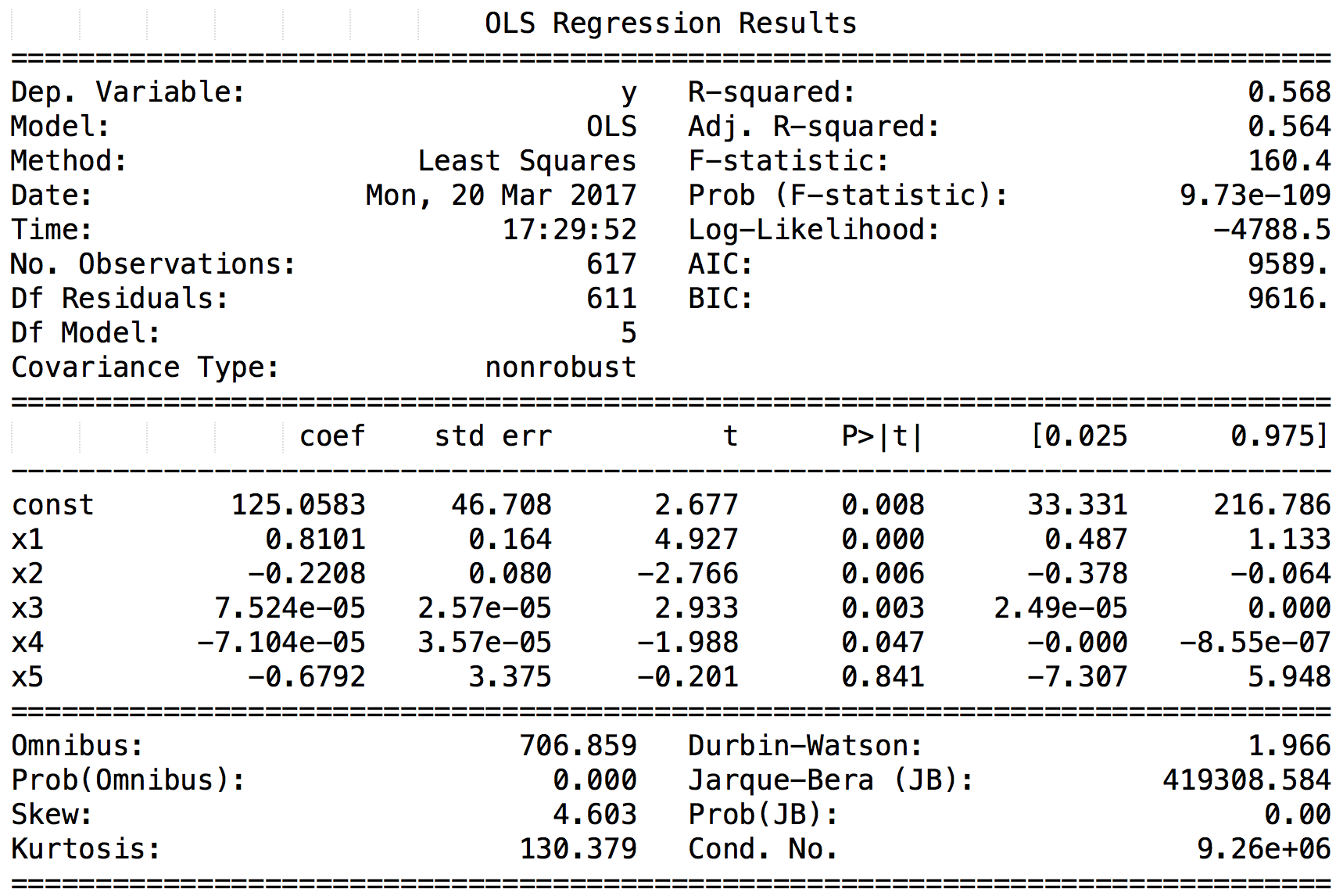


Figure 5 Regression Statistics for #nfl

Figure 6: OLS Regression Statistics for #nfl

Number of tweets, total number of retweets and sum of followers of the users posting the hashtag are the three most significant features. For our regression model, R-squared value is 0.568. This could imply that linear model can't explain data well.

Liner Model for #patriots

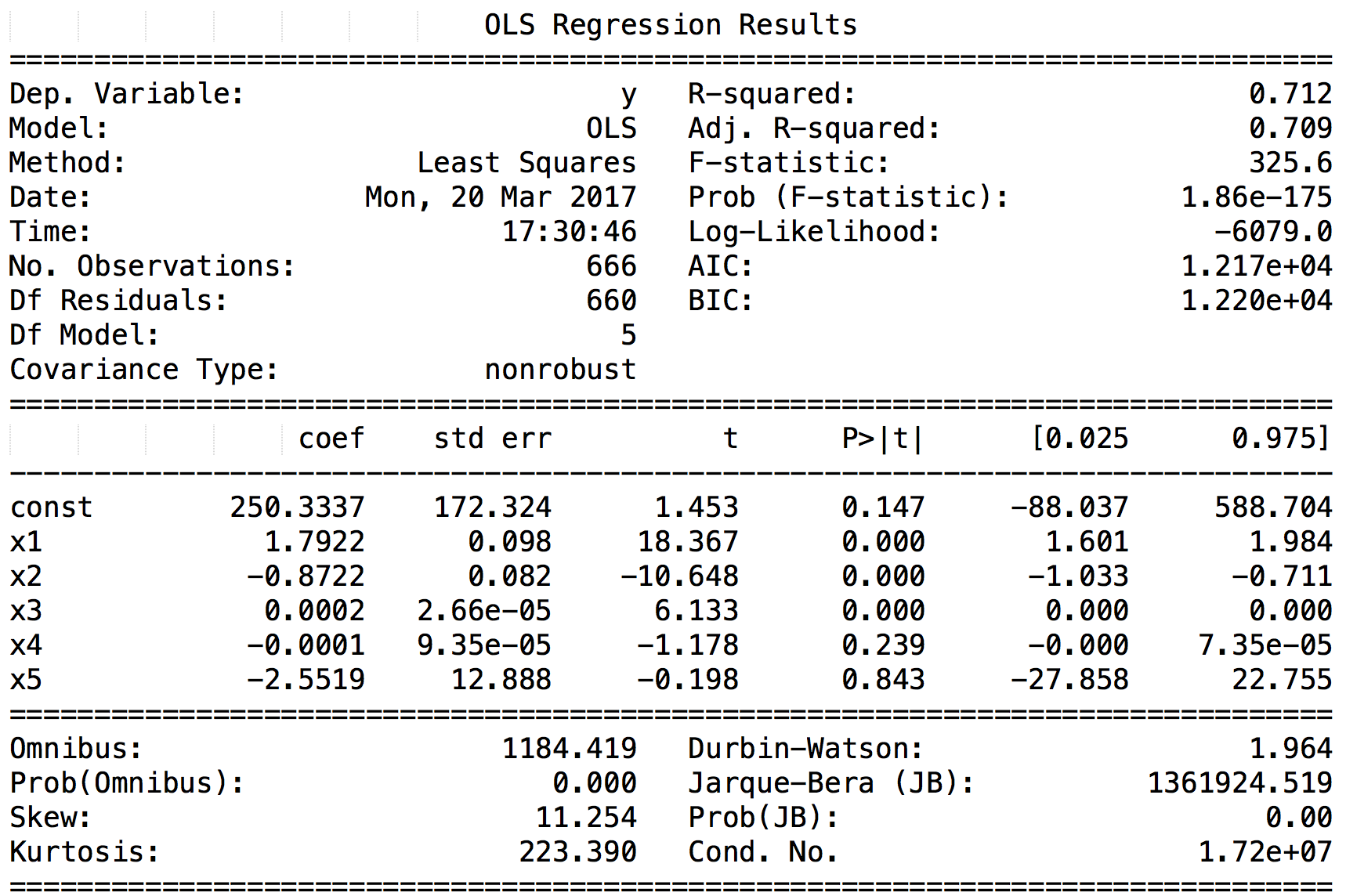


Figure 6 Regression Statistics for #patriots

Looking at the t-value and p-value (we consider features with higher t-value magnitude and lower p-value to be significant), we can say number of tweets, total number of retweets and sum of followers of the users posting the hashtag to be significant features. For our regression model, R-squared value is 0.712, which is good.

Liner Model for #sb49

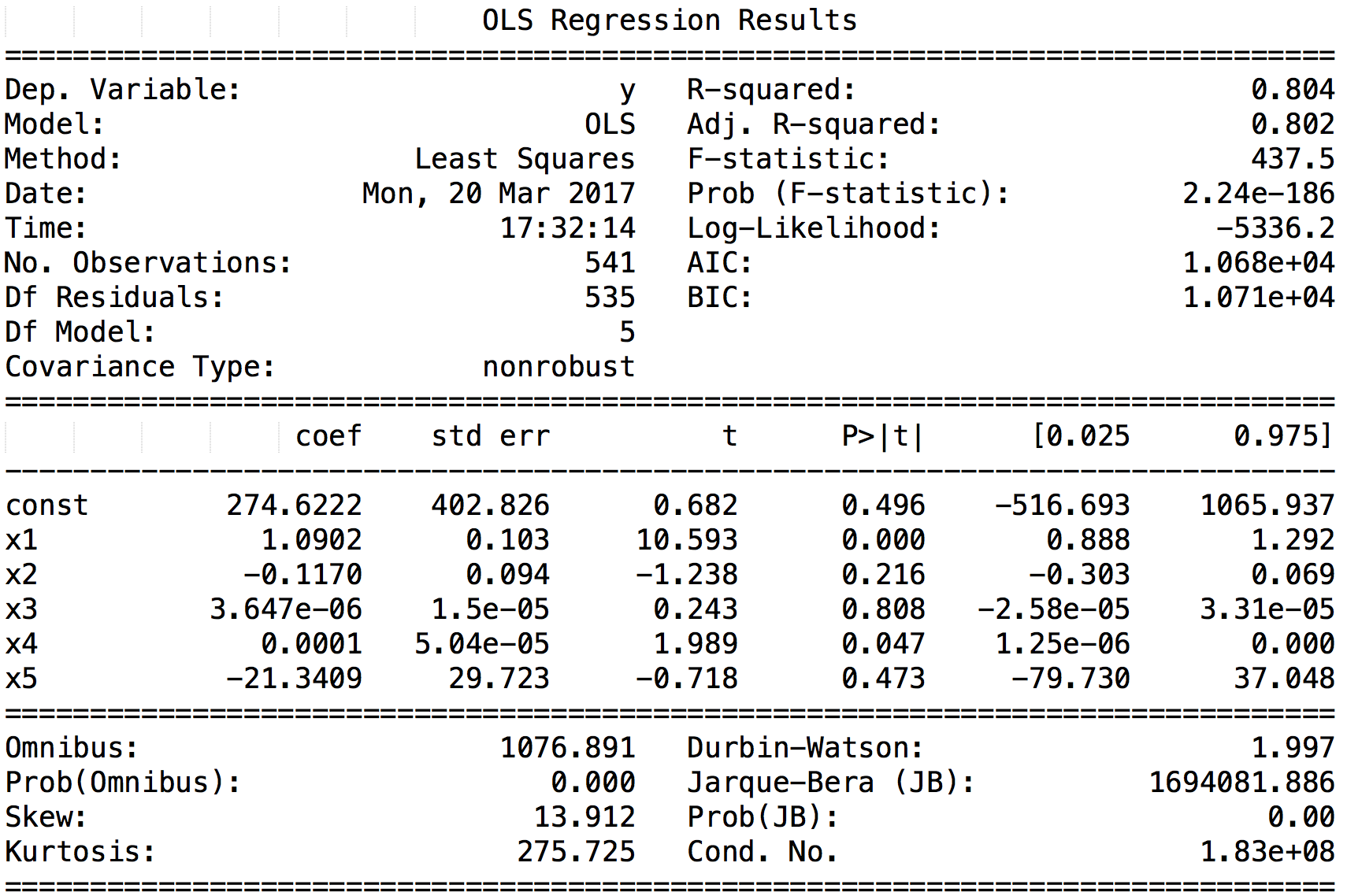


Figure 7 Regression Statistics for #sb49

From the table, we can say number of tweets, total number of retweets and maximum number of followers of the user posting the hashtag to be significant features. We have a good training accuracy as our R-squared value is 0.804.

Liner Model for #superbowl

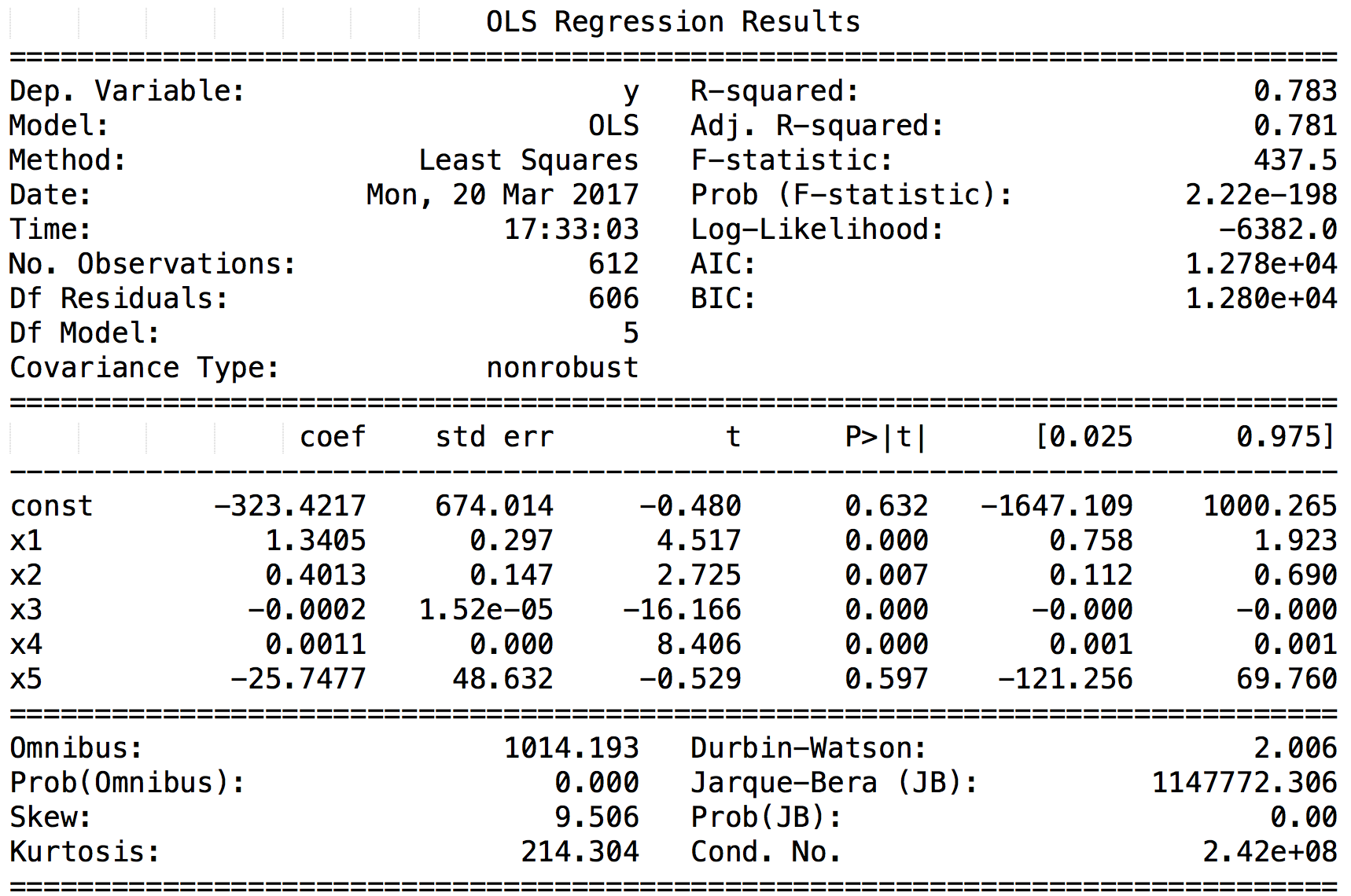


Figure 8 OLS Regression Statistics for #sb49

Looking at the t-value and p-value (we consider features with higher t-value magnitude and lower p-value to be significant), we can say number of tweets, total number of retweets and sum of the followers of the users posting the hashtag to be significant features. The training model is about 80% accurate, as the R-squared value is 0.783

**3)**

For the third section of the problem, we removed the features from section 2 that we deemed to be insignificant and in-place of those features, we added three new user-dependent-features:

**Sum of friends count**, **sum of statuses count** and **sum of ranking score**.

**Sum of friends count**: In Twitter, you can become a friend to a person if you follow him/her and he/she follows you back. This was one of our crucial ideas behind selecting this as a feature. For example, if you are friends with one person on Twitter, then he/she is more likely to see your status and retweet it and vice versa. Since we believe that this aspect is crucial to making a hashtag popular, this was one of our features.

**Sum of statuses count**: Another feature we considered we was the sum of the statuses count of the users. A user with more statuses count is more likely to be an active user of Twitter and one with a lower status count is not likely to be very active. Hence, using this feature will help us predict the number of tweets in the next hour more accurately.

**Sum of Ranking score**: Ranking score indicates the importance of the user who is generating the tweet. Higher the ranking score, the more influential a user is and hence there is a greater probability of his tweets reaching a wider audience and making the hashtag popular. Hence this was naturally our third and final feature for our prediction.

We fitted a linear regression model with the most significant features from section 2 and these three new features of our own adding to a total of 7 features. We label them x1; : : : ; x7 where

* x1 is sum of friends count
* x2 is sum of statuses count
* x3 is sum of Ranking score
* x4 is the number of tweets
* x5 is the total number of retweets
* x6 is the sum of the followers of the users posting the hashtag
* x7 is the maximum number of followers of the user posting the hashtag

**Liner Model for #gohawks**

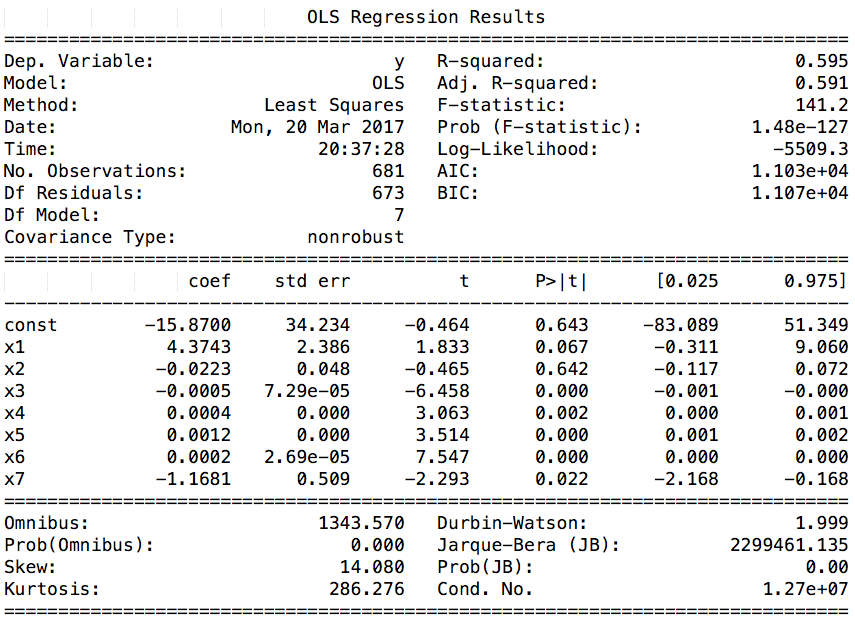


Figure 9 Regression Statistics for #gohawks

Looking at the t-value and p-value (we consider features with higher t-value magnitude and lower p-value to be significant), we can say sum of the followers of the users posting the hashtag, sum of Ranking score and the maximum number of followers of the user posting the hashtag are the 3 most significant features. For our regression model, after adding features, the accuracy has improved. This can be seen from increase in R-squared value to 0.591.

|  |  |  |
| --- | --- | --- |
|  |  |  |

Figure 10 Relation between number of tweets in the next hour and 3 major features for #gohawks

**Liner Model for #gopatriots**

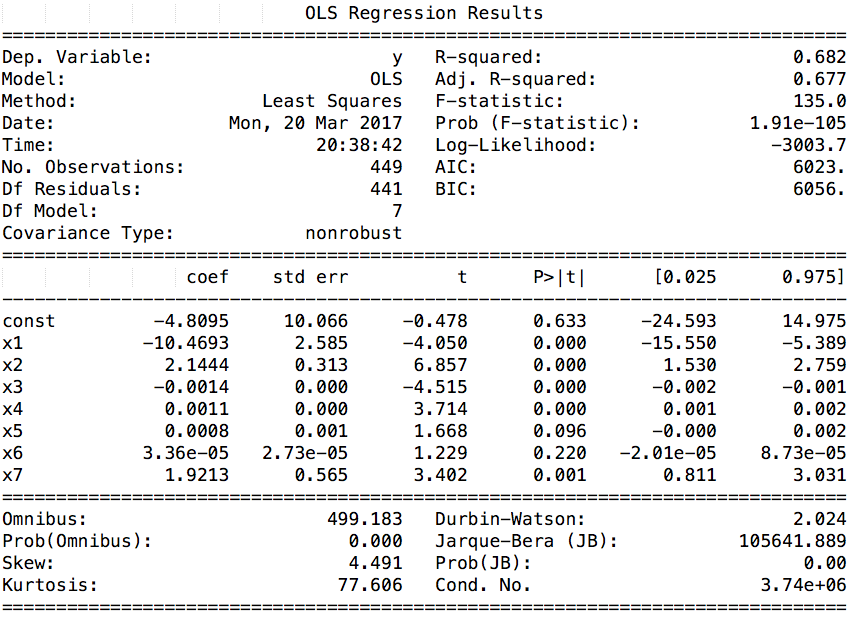
****

Figure 11 OLS Regression Statistics for #gopatriots

For this hashtag, sum of statuses count, sum of Ranking score and sum of friends count are the 3 most significant features. We again see an improvement in Training accuracy as compared to section 2. R-squared value is 0.682.

|  |  |  |
| --- | --- | --- |
|  |  |  |

Figure 12 Relation between number of tweets in the next hour and 3 major features for #gopatriots

**Liner Model for #nfl**

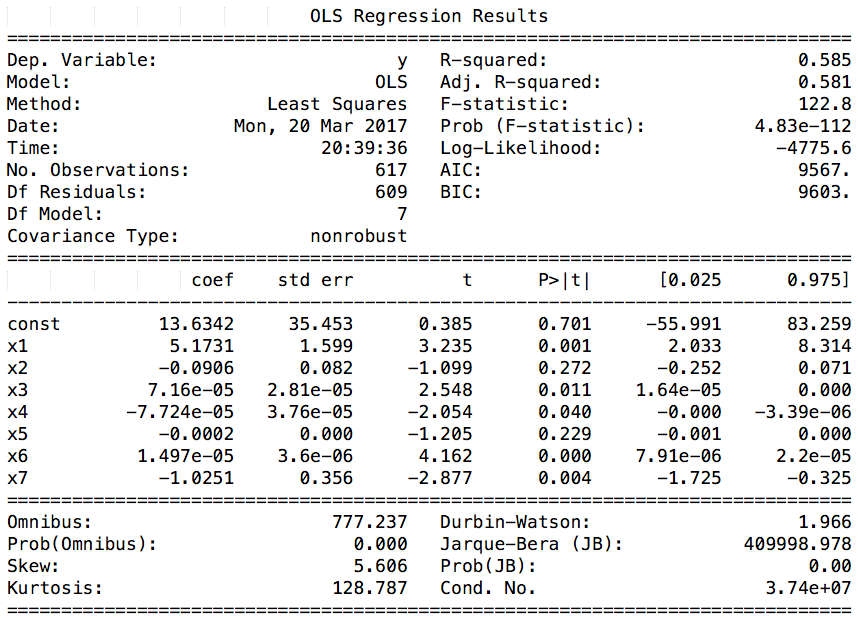
****

Figure 13 OLS Regression Statistics for #nfl

From the table, the sum of the followers of the users posting the hashtag, sum of friends count and sum of Ranking score are the 3 most significant features. For our regression model, R-squared value is 0.581 which is better than the model from section 2.

|  |  |  |
| --- | --- | --- |
|  |  |  |

Figure 14 Relation between number of tweets in the next hour and 3 major features for #nfl

**Liner Model for #patriots**

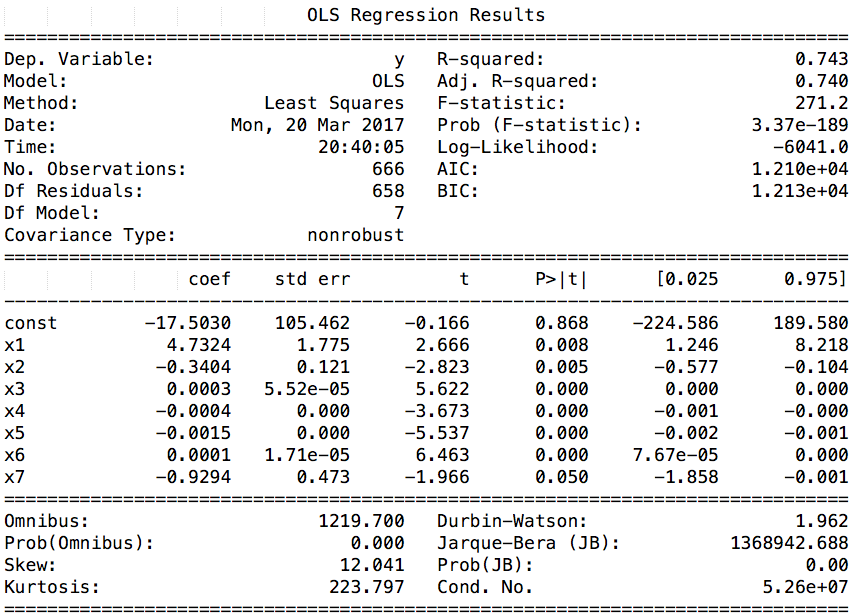
****

Figure 15 OLS Regression Statistics for #patriots

Here sum of Ranking score and the maximum number of followers of the user posting the hashtag are the 3 most significant features. For our regression model, R-squared value is 0.591, which is lower than section 2. Overfitting could be responsible for brining the training accuracy lower.

|  |  |  |
| --- | --- | --- |
|  |  |  |

Figure 16 Relation between number of tweets in the next hour and 3 major features for #patriots

**Liner Model for #sb49**

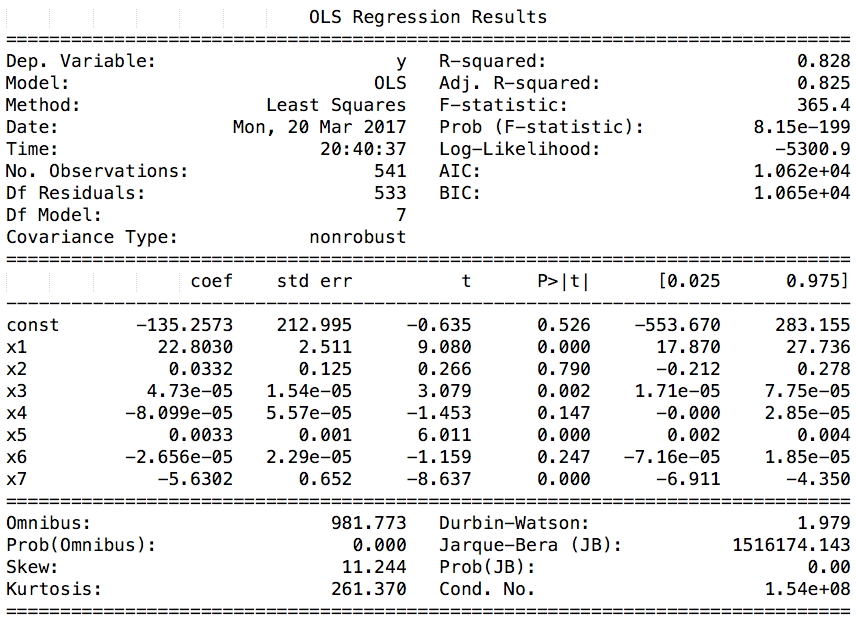
****

Figure 17 OLS Regression Statistics for #sb49

Looking at the t-value and p-value (we consider features with higher t-value magnitude and lower p-value to be significant), we can say sum of friends count, the sum of the followers of the users posting the hashtag and number of re-tweets are the 3 most significant features. For our regression model, R-squared value is 0.825 which means that the model is very good fit.

|  |  |  |
| --- | --- | --- |
|  |  |  |

Figure 18 Relation between number of tweets in the next hour and 3 major features for #sb49

**Liner Model for #superbowl**

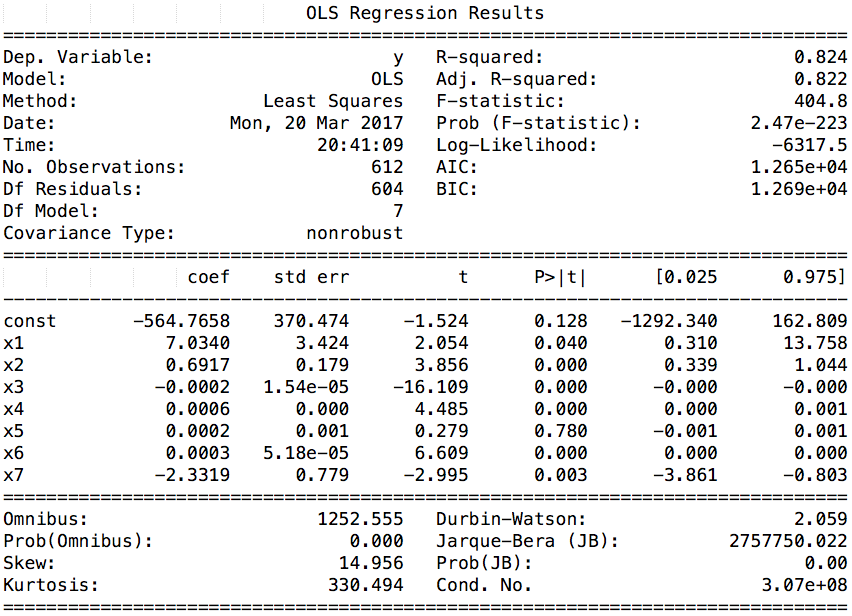
****

Figure 19 OLS Regression Statistics for #superbowl

From the table, we conclude sum of Ranking score, sum of the followers of the users posting the hashtag and number of tweets are the 3 most significant features. By introducing new features, we have increased our training accuracy. For our regression model, R-squared value is 0.822 which is higher than that of the linear model in section 2.

|  |  |  |
| --- | --- | --- |
|  |  |  |

Figure Relation between number of tweets in the next hour and 3 major features for #superbowl

**4)**

In this section, instead of running a general regression model over the entire period, we run a regression model for three specific time periods:

* Before Feb. 1, 8:00 a.m
* Between Feb. 1, 8:00 a.m. and 8:00 p.m
* After Feb. 1, 8:00 p.m

Feb. 1 is the day on which the Superbowl took place and hence our hypothesis is that this time period should contain the maximum burst of the tweets. We trained different linear models for each hashtag separately.

Here, we do cross-validation training and testing for each of the six linear models separately. Like the previous approach, we adopt a 10-fold cross validation for each linear model. The results of our cross-validation training and test are as follows:

Hashtag: #superbowl File Loaded

The average prediction error for #superbowl is 2792.566726

The average prediction error for #superbowl in period 1 is 247.894642

The average prediction error for #superbowl in period 2 is 137604.112462

The average prediction error for #superbowl in period 3 is 224.387692

Hashtag: #gohawks File Loaded

The average prediction error for #gohawks is 227.449555

The average prediction error for #gohawks in period 1 is 302.696504

The average prediction error for #gohawks in period 2 is 5606.527057

The average prediction error for #gohawks in period 3 is 23.302044

Hashtag: #sb49 File Loaded

The average prediction error for #sb49 is 1161.742542

The average prediction error for #sb49 in period 1 is 47.998492

The average prediction error for #sb49 in period 2 is 85209.898420

The average prediction error for #sb49 in period 3 is 163.329313

Hashtag: #nfl File Loaded

The average prediction error for #nfl is 188.692228

The average prediction error for #nfl in period 1 is 123.608549

The average prediction error for #nfl in period 2 is 12782.782564

The average prediction error for #nfl in period 3 is 117.572069

Hashtag: #gopatriots File Loaded

The average prediction error for #gopatriots is 103.550765

The average prediction error for #gopatriots in period 1 is 18.641701

The average prediction error for #gopatriots in period 2 is 4350.164763

The average prediction error for #gopatriots in period 3 is 7.326925

Hashtag: #patriots File Loaded

The average prediction error for #patriots is 590.008351

The average prediction error for #patriots in period 1 is 206.766123

The average prediction error for #patriots in period 2 is 25835.863690

The average prediction error for #patriots in period 3 is 91.166772

We also trained the linear model using tweets from all hashtags instead of isolating them and we arrived at the following results:

For our experiments, we first split the data set into three specific time periods (before Feb.1,8:00 a.m, between Feb.1,8:00 a.m. and 8:00 p.m, after Feb.1,8:00 p.m) based on the time stamp (one hour bin) of the tweet. Hence, we now have three sets of training data each for a specific period. We now train a regression model for each one of these periods. We split each data into train-test split and perform a 10-fold cross-validation. We divide the training data into ten subsets and for each of the ten folds, we fit a model on 9 parts of the data and test on the final part. We then average our errors over the ten folds. We arrived at the following results.

The average prediction error overall in period 1 is 573.005528

The average prediction error in period 2 is 449075.994620

The average prediction error overall in period 3 is 519.353667

The average prediction error overall is 3267.789455

**Analysis**:

From the results above we can see that for the time period before and after the Superbowl the errors are very less. This is because there are relatively lesser number of tweets over a longer time period for Before Feb.1, 8:00 a.m and After Feb.1, 8:00 p.m. However for the time period during the Superbowl: Between Feb.1, 8:00 a.m. and 8:00 p.m., there is a huge burst of tweets. Also, the number of training examples that we have for this period is very less compared to the other two periods as we are taking

One hour bins and the time period is just 12 hours. Hence, our model is prone to high errors during this time. If we had indeed had a smaller time-period, say 10 minute bins instead of 1 hour bins, our model would have performed better since we would have a larger number of examples which might have led to a better accuracy. For the sake of completeness, we are also posting the regression statistics we got for each of the above periods.

**5)**

For the final problem, we are given test data which contain 10 windows of 6 hour time periods: **sample1-period1.txt, sample2-period2.txt, sample3-period3.txt, sample4-period1.txt, sample5-period1.txt, sample6-period2.txt, sample7-period3.txt, sample8-period1.txt, sample9-period2.txt, sample10-period3.txt.**

We are now supposed to make predictions for the number of tweets in the next hour for each of these six-hour windows. We can leverage that fact that we have built three different regression models for each interval:

* Before Feb.1, 8:00 am (period 1),
* Between Feb.1, 8:00 a.m. and 8:00 pm (period 2),
* After Feb.1, 8:00 pm (period 3).

Also, we know from the test data as to which period each sample belongs to. Hence, we segregate the test data into three sets each belonging to period 1, period 2 and period 3.

For example, if we want to test for sample1-period1.txt, we will send it to the model trained on tweets during period 1 i.e Before Feb.1, 8:00 a.m. If we want to test for sample6-period2.txt, then we would be using the model trained on tweets during period 2 i.e Between Feb.1, 8:00 a.m. and 8:00 p.m.

Hence, in this problem, for each test case, we send it to one of the fitted models, and obtain the prediction of number of tweets for the next hour for each of them.

The results for number of tweets predicted in the next hour (using the features of the last hour of each window) for each sample window is reported below.

**Period 1:**

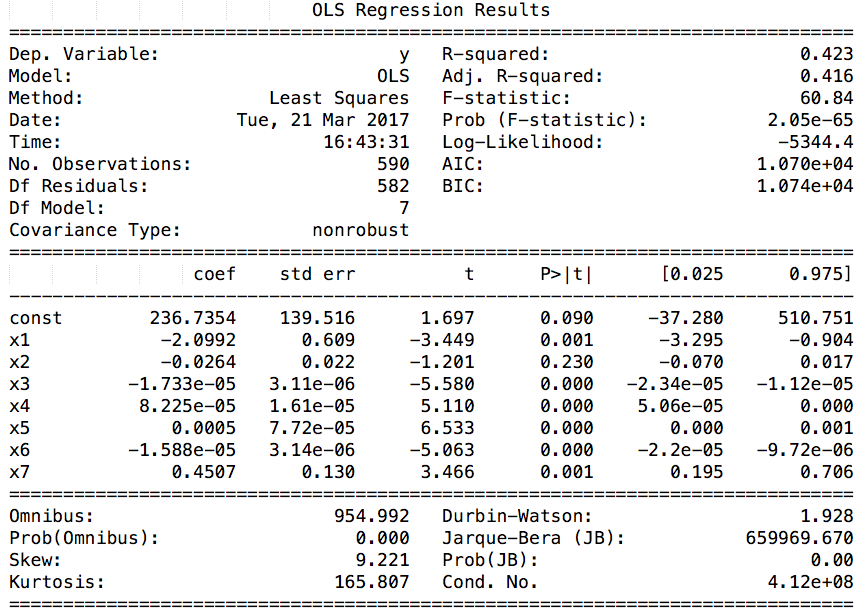
****

Figure Regression Statistics for period 1

The predicted number of tweets for the next hour of test file: sample1\_period1.txt is 651.543762

The predicted number of tweets for the next hour of test file: sample4\_period1.txt is 760.366689

The predicted number of tweets for the next hour of test file: sample5\_period1.txt is 738.350018

The predicted number of tweets for the next hour of test file: sample8\_period1.txt is 314.218468

**Period 2:**

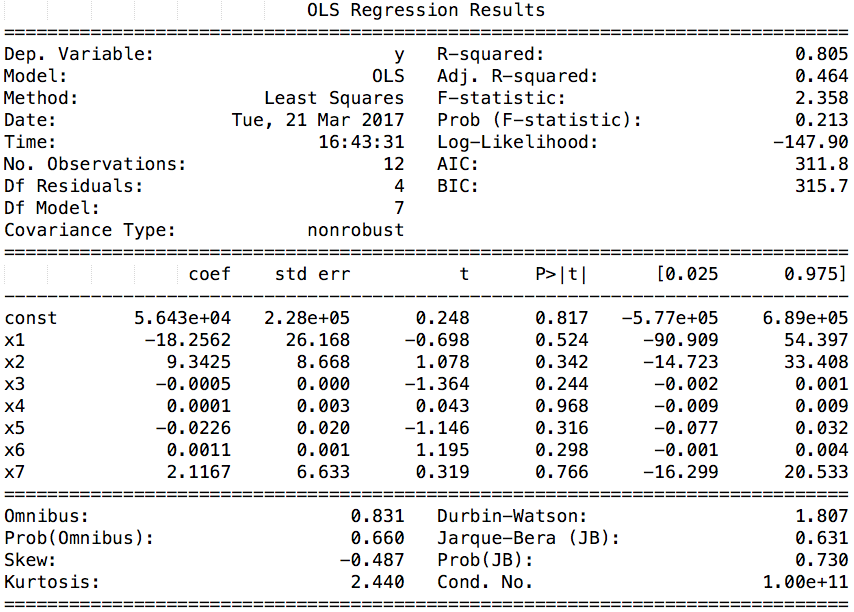
****

Figure Regression Statistics for period 2

The predicted number of tweets for the next hour of test file: sample2\_period2.txt is -440800.347151

The predicted number of tweets for the next hour of test file: sample6\_period2.txt is -563236.694304

The predicted number of tweets for the next hour of test file: sample9\_period2.txt is 10898.192715

**Period 3:**

****

Figure Regression Statistics for period 3

The predicted number of tweets for the next hour of test file: sample3\_period3.txt is 3767.510989

The predicted number of tweets for the next hour of test file: sample7\_period3.txt is 3950.676630

The predicted number of tweets for the next hour of test file: sample10\_period3.txt is 3851.193499

Few notes:

1. The negative values that we arrived at mean predict 0 as the optimum value for number of tweets.
2. Number of training data is small (12) for period 2 so it results in suboptimal predictions since the linear regression is not “well” trained to give us better results.
3. During period 1 (before the superbowl), the predicted number of tweets is very less.
4. But exactly during the period of the Superbowl (period 2), the predicted number of tweets shoots up to the thousands. This is expected as the number of tweets generated will be naturally higher during this period as it is the Superbowl and the popularity is high.
5. During period 3 (after the superbowl) we can see the predicted number of tweets to be lower. This goes along well with the fact that during this time the excitement of the Superbowl must have gone down and so is the number of tweets.

**6)**

Similar to project 2, we trained a classifier using the textual contents of the tweets to be able to predict the location of the author. Using the techniques and the algorithms we learned in project 2 we calculated the confusion matrix, the accuracy measures, and the ROC curves. Here are the results that we arrived at:

|  |  |
| --- | --- |
| SVM  Precision: 0.70477348738  Accuracy: 0.72971082089  Recall: 0.924980047885075 | C:\Users\Kia\Google Drive\Big Data Mining\Project 5\Figures\Part 6\ROC_SVM.png |
| Naive Bayes  Precision: 0. 71374045801  Accuracy: 0. 679570895522  Recall: 0. 7497995188452 | C:\Users\Kia\Google Drive\Big Data Mining\Project 5\Figures\Part 6\ROC_NaiveBayes.png |
| Logistic Regression  Precision: 0. 71068020929  Accuracy: 0. 73763992537  Recall: 0. 925821972734 |  |
| Logistic Regression with L1 regularizer  Precision: 0. 7100737100737101  Accuracy: 0. 7374067164179104  Recall: 0. 9270248596631917 | C:\Users\Kia\Google Drive\Big Data Mining\Project 5\Figures\Part 6\ROC_LogisticRegressionL1Regularization.png |
| Logistic Regression with L2 regularizer  Precision: 0. 7085504137296966  Accuracy: 0. 7357742537313433  Recall: 0. 9270248596631917 | C:\Users\Kia\Google Drive\Big Data Mining\Project 5\Figures\Part 6\ROC_LogisticRegressionL2Regularization.png |

We can see that the results are not satisfactory since the feature vectors are not optimized for classification and we have not achieved sharp ROC curves.