# EE239AS Special Topics in Signals and Systems

Project 3 Report:

**Popularity Prediction on Twitter**

Winter 2015

3/20/2015

## Question 1

This question requires calculating the statistics from the six given tweets data files. From the data, we can see that the start and end dates are different for each hashtag, so we need to find the beginning and ending date first from the crawling data. The dates for each hashtags are listed below:

|  |  |  |  |
| --- | --- | --- | --- |
| hashtag | Start date | End date | Hour expended |
| #gopatriots | 2015 Jan 09 20:30:45 | 2015 Feb 07 07:54:35 | 684 |
| #gohawks | 2014 Dec 28 22:14:35 | 2015 Feb 07 10:17:49 | 973 |
| #patriots | 2014 Dec 28 22:21:19 | 2015 Feb 07 18:55:00 | 981 |
| #nfl | 2014 Dec 31 04:21:23 | 2015 Feb 07 18:55:36 | 927 |
| #sb49 | 2015 Jan 14 12:31:15 | 2015 Feb 07 18:55:36 | 583 |
| #superbowl | 2014 Dec 31 04:21:23 | 2015 Feb 07 18:00:08 | 964 |

Table 1 Start and end date of each hashtag

In the data, the postdate, number of follower and number of retweet can be found in these categories. Note that the number of retweets should be found in the "citation", "total".

*result["tweet"]["created\_at"]*

*result["tweet"]["user"]["followers\_count"]*

*result["metrics"]["citations"]["total"]*

Then we can find the total number of hours, tweets, retweets and find the numbers in each hour. The results are listed below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Statistics  hashtag | Total number of tweets | Average number of tweets per hour | Average number of followers per user | Average number of retweets per hour |
| #gopatriots | 26232 | 38.3509 | 1602.0099 | 1.4001 |
| #gohawks | 188136 | 193.3566 | 2393.5823 | 2.0146 |
| #patriots | 489713 | 502.7495 | 3641.6884 | 1.7828 |
| #Sb49 | 826951 | 1418.4408 | 10230.0453 | 2.5111 |
| #superbowl | 1348767 | 1399.1359 | 9958.1157 | 2.3883 |
| #nfl | 259024 | 279.4218 | 4763.3265 | 1.5385 |

Table 2 The statistics of each hashtag

Then plot the "number of tweets in hour" over time for #superbowl and #nfl. From the figures we can see that the peak happens at round the 800th hours for both hashtags while #nfl has another local maximum at 450th hour.

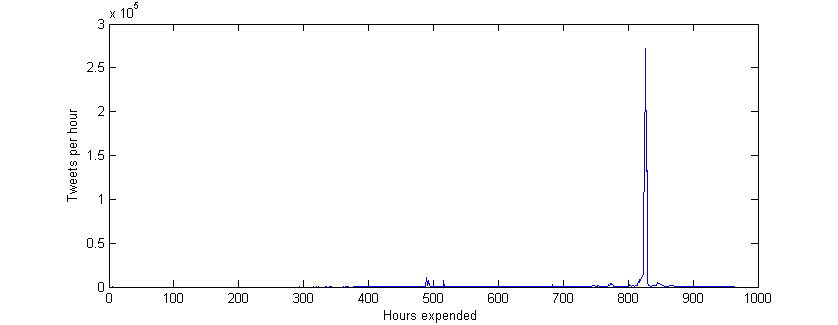


Figure 1 #superbowl tweets per hour over time

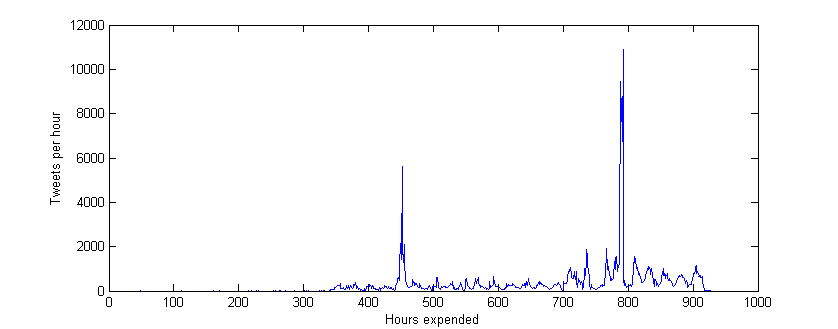


Figure 2 #nfl tweets per hour over time

## Question 2

The time window in this question is 1 hour, we calculate the

*Number of tweets,*

*Total number of retweets,*

*Sum of the number of followers posting the hashtag,*

*Maximum number of followers in users posting the hashtag,*

*Time of the day*

These five features are taken as the input value to our regression model and the output of the model is to predict the number of tweets in the next hour.

These five features are : true\_num, retweet\_each\_hour, follower\_each\_hour, max\_follower\_each\_hour, timeofday. each of them are calculated from the previous questions and saved as the format of *.csv*. e.g.

*csvWriter.writerow(retweet\_each\_hour)*

Input these features in to *statsmodel* and take the tweets per hour as the previous hour data to predict the next hour. Then we can plot the tweets predicted over each feature. The feature figures of #NFL are listed below.

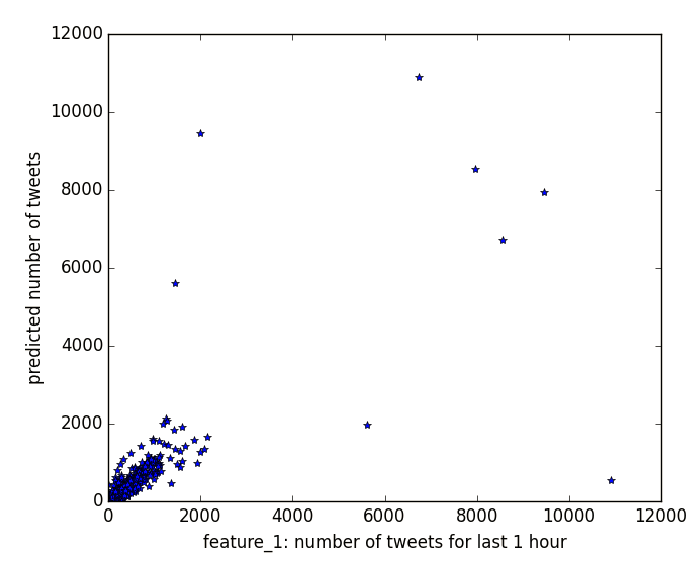


Figure 3 Predicated tweets over Number of tweets each hour

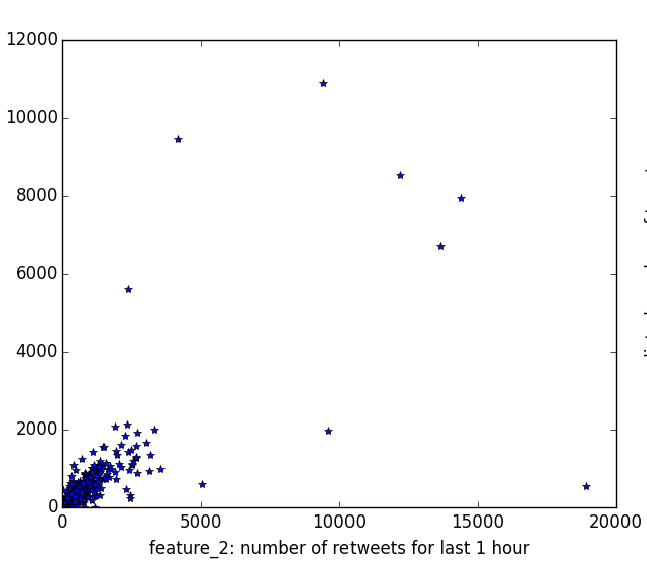


Figure 4 Predicated tweets over Number of retweets each hour

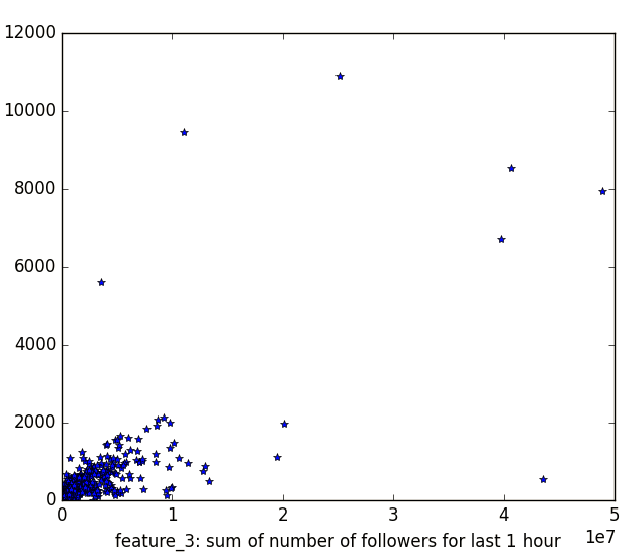


Figure 5 Predicated tweets over Number of followers each hour

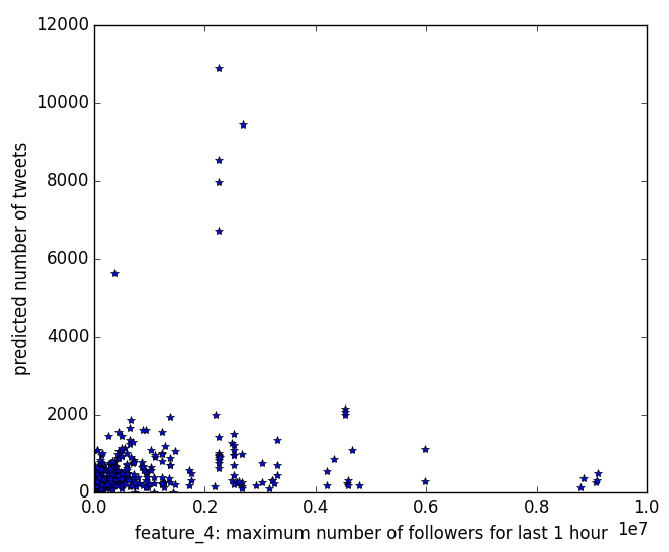


Figure 6 Predicated tweets over maximum number of followers each hour

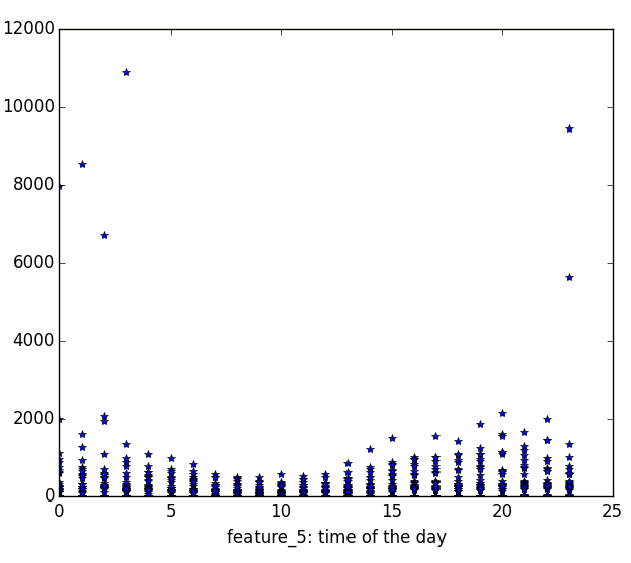


Figure 7 Predicated tweets over time of the day

Figure 8 shows the predicted number of tweets over the actual number of tweets in the next hour

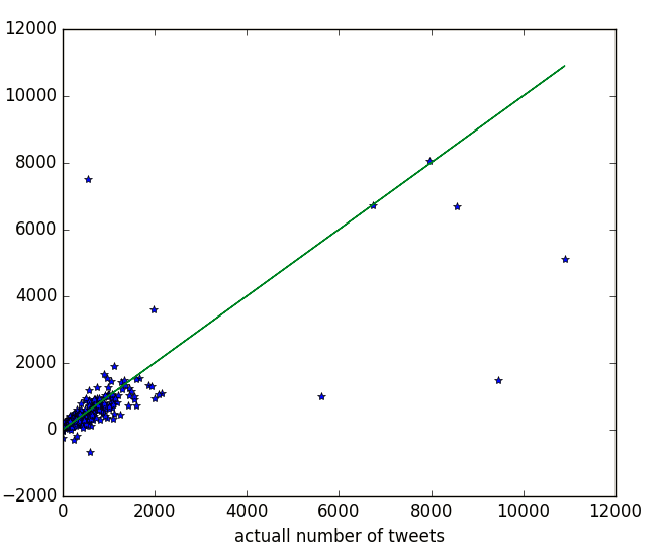
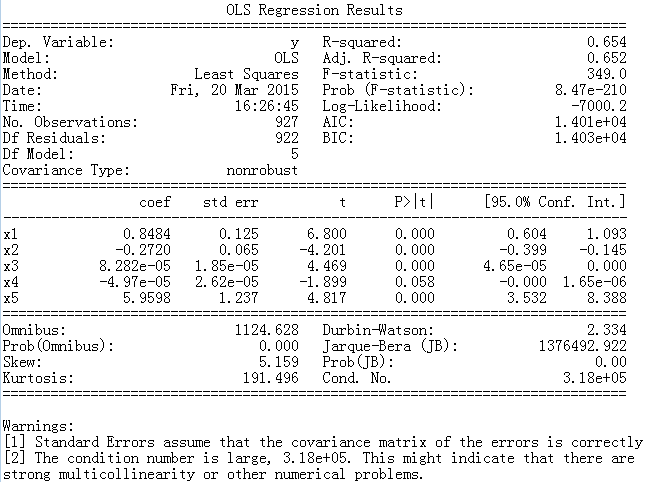


Figure 8 Predicated tweets over actual tweets

From the figures above, we can see that, the feature: number of tweets per hour and number of retweets per hour shows good ability to predict the next hour tweets while the maximum number of followers in each hour shows a poor relation to the number of tweets. Figure 7 shows the trend of the tweets in the day and because the data in the #NFL is not so big and there is no obvious spikes in it, so in this case it can somehow predict the tweets.

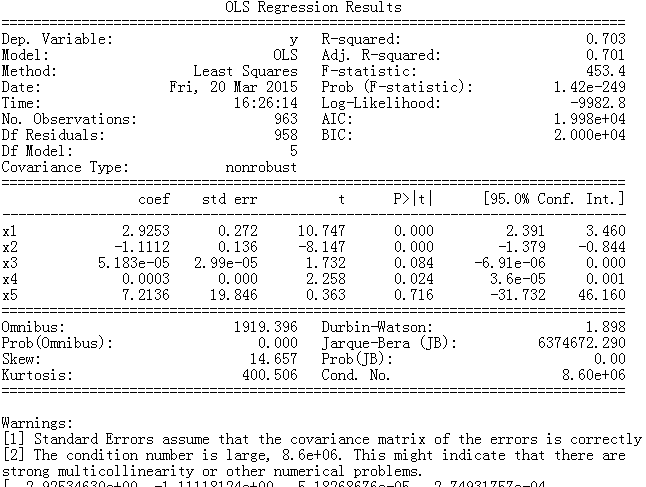


The summary provided by *statsmodel* can give an insight into the significance of each feature to the model.

The T-test and P value can be used to describe the statistical relationship between the features and models. The t statistic is the coefficient divided by its standard error, so the larger the t is the more significant the feature may be.

In this case, the **x1(Number of tweets), x2(Number of retweets), x3(Number of followers), x4(maximum number of followers) and x5(time of the day)** have very close t value so t cannot be taken as a criterion to decide which one is better.

The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis so a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. We have a P value for x4 which is larger than 0.05, this means **the time of the day** is not a good feature for the model. The result of #NFL is not that obvious to decide which feature has a better relationship to the model, so we check the summary of other hashtags and find that x3 and x5, the **number of followers** and the **maximum followers** are also not good features, for their P value is larger than 0.05. This can be supported by Figure 3.



The R-squared value and coefficients can both help us to decide which feature is better. More details will be discussed in Question 3 for how to pick good features. The figures over each feature of each hashtag and the summary are attached in the Appendix.

## Question 3

Design a regression model using any features from the paper or other new features you may find useful for this problem. Fit your model on the data and report fitting accuracy and significance of variables. For the top 3 features in your measurements, draw a scatter plot of predictant (number of tweets for next hour) versus feature value, using all the samples you have extracted.

For this question, we extracted all the possible feature that is possible to see their relation with the sum of tweets number in every hour. Some of the features can be easily gotten from the tweets data, (like Number of followers, Number of retweets), while others are extracted from data and text analysis. (like User emotion, Number of user mentioned in each tweet)

Here is a table of all the features and their meanings:

|  |  |  |
| --- | --- | --- |
| **No.** | **Features** | **Meanings of features** (specifically for the feature in 1-hour window ) |
| **1** | True Number | Number of tweets in 1-hour window |
| **2** | Retweet | Sum of the Numbers of retweets in 1-hour window |
| **3** | Follower | Number of followers of all authors in 1-hour window |
| **4** | Fax\_Follower | The Max follower of the 1-hour window |
| **5** | Time of the day | Time of the day, represented as whole hour like 18:00 |
| **6** | Acceleration | Sum of acceleration values for all tweets in 1-hour window  Acceleration: The attention concerning this topic, from -100 to 100 |
| **7** | Favorite | Sum of all the favorite numbers of tweets in 1-hour window |
| **8** | Statuses | Sum of all the history statuses numbers for users in the 1-hour window |
| **9** | Ranking | Sum of ranking values for all the tweets in 1-hour window |
| **10** | Influential | Sum of the influence value of all the tweets in 1-hour window |
| **11** | Matching | Sum of citation matching values for all the tweets in 1-hour window |
| **12** | Momentum | Sum of momentum values for all the tweets in 1-hour window |
| **13** | UserMention | Sum of No. of Users mentioned @ in each tweet in 1-hour window |
| **14** | Emotion | No. of “emotional” tweets (punctuation like !!! ???) in 1-hour window. |

Table Features and meanings

For the No. 13th feature “UserMention”, we extract the tweet text from the original data, and count the number of @ in the text to find the No. of users mentioned in this tweet. Theoretically, when a user is mentioned by other people in some tweets, he would be more likely to engage in tweet communication activity, and more like to post a tweet afterwards. So we believe it reasonable to use this “UserMention” as a feature to predicted the tweet activity.

For the No.14th feature “Emotion”, we also extract the tweet text, and detect whether there is a strong emotion in this tweet. Such as the use of punctuation like multiple question mark “???” and multiple exclamation mark “!!!”. We guess that the more “strongly emotional” tweets in one hour, the more popular and active the hashtag will be.

So, we extract the 14 features for each of the 6 hashtags, and save the feature data to csv file. Before creating a fitting model, we draw the plot of tweets in current hour versus feature value, to see the correlation of the two values. Below are some plot of current No. of tweets V.S. features:

|  |  |  |
| --- | --- | --- |
| #nfl_acceleration_vs_tweet | #nfl_emotion_vs_tweet | #nfl_favorite_vs_tweet |
| #nfl_follower_vs_tweet | #nfl_influential_vs_tweet | #nfl_matching_vs_tweet |
| #nfl_max_follower_vs_tweet | #nfl_momentum_vs_tweet | #nfl_ranking_vs_tweet |
| #nfl_retweet_vs_tweet | #nfl_statuses_vs_tweet | #nfl_time_of_the_day_vs_tweet |
| #nfl_usermetion_vs_tweet |  |  |

Figure #nfl: No. of tweets V.S. features ----A: acceleration B: emotion C:favorite D:follower E:influential F:matching G:maxfollower H:momentum I:ranking J:retweet K:statuses L:timeoftheday M:usermention

From the plot above, we can see that some of the features have a strong correlation with tweets values, like influential value, ranking value, matching value, number of retweets, statuses value and UserMention value. While there are some features have nearly no relation with tweet, such as max number of followers and time of the day.

Then we also draw the figure of the “tweets for next hour” versus “features for this hour”, and this time, we want to see the relation between features and Numbers of tweets in next hour, to determine the performance of the features in prediction.

|  |  |
| --- | --- |
| #nfl_acceleration_vs_tweet | #nfl_ranking_vs_tweet |
| #nfl_retweet_vs_tweet | #nfl_statuses_vs_tweet |
| #nfl_tweet_vs_tweet | #nfl_usermetion_vs_tweet |

Figure 10 #NFL: “tweets for next hour” versus “features for this hour”

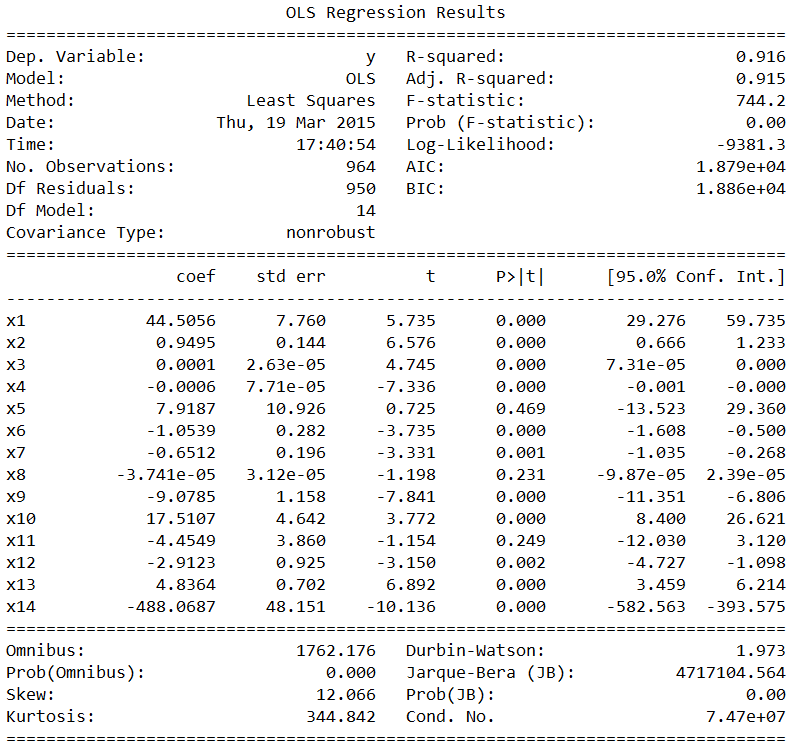
From these picture, we can see that the correlation is not as good as the figures before, since the feature is 1 hour before the time of tweet we want to predict. From this, we can also know that there might be some inaccuracy of the prediction in our model.

And we also take the statistical value in to consideration:

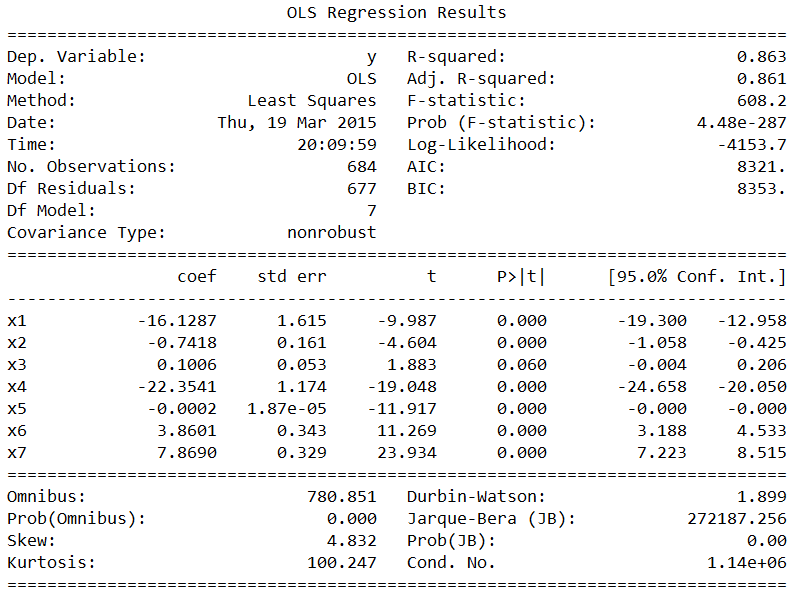
From the Ordinary Least Square Regression Result summary below, we can see not only the coefficient of each features in our linear fitting model, but also their t-test and P-value results. A simple way to evaluate the performance of fitting model is R-squared value. The closer R-Squared value to 1, the better the model is.

We also know that when the P-value of the feature is high (larger than 0.05), it means the feature is not really efficient, and it might have no relation with our tweets number in prediction.

Below is a result of model with 14 features (#superbowl)



So, after integrating the factor of feature V.S. tweet plot, feature significance (coefficient), feature P-value, t-test value and R-squared value, we come up with a New Regression Model. This new model use the feature of tweets, retweet, acceleration, favorite, statuses, ranking, and UserMention. The regression result is shown below (#gopatriots):



From this result, we can see that the P-value is all very low, and the coefficient is homogeneous. The 1st feature No. of tweet and 4th feature acceleration takes a significance role in the model. And also the feature of UserMention is really important. This result is quite reasonable, because the No. of tweets in previous hour is certainly the best feature to predict the No. of tweets in next hour, and the acceleration value is a coefficient that represent the tweet’s popularity. And the UserMention is also a latent factor that influence the future tweet number.

We can see that the R-squared value is 0.86, it is smaller than 0.916, which is the R-squared value in the model with 14 features. However it doesn't means this model is inferior than that one. Because we know that the R-squared value is only a rough way to judge the performance of a prediction model, it not really accurate. And also too many features can lead to the problem of overfitting.

Then, we draw the scatter plot of predictant versus feature value for the top three features:

From the plots, we can easily see that most of the data point are scattered on a straight line. Some of the data points are extremely large, which is hard to predict because of lack of sufficient data to learn. So, these features show a nice performance overall.

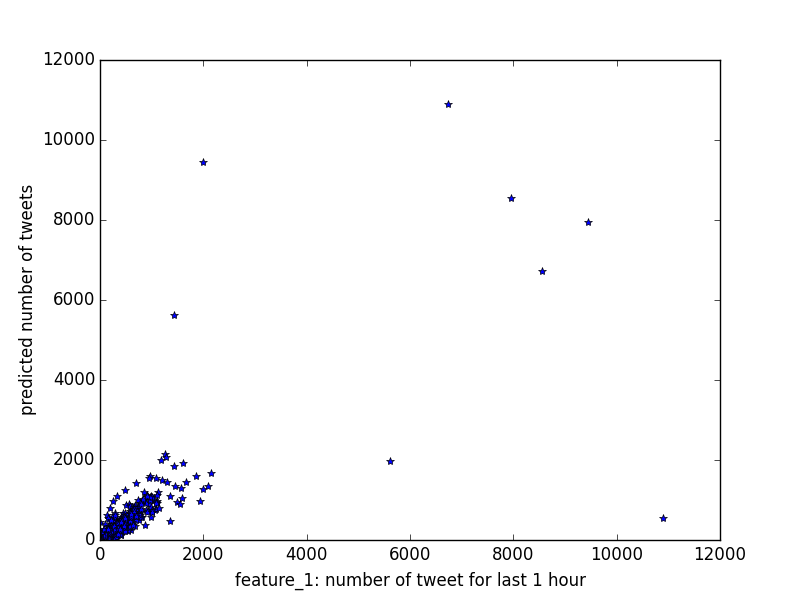


Figure Number of tweet for last 1 hour

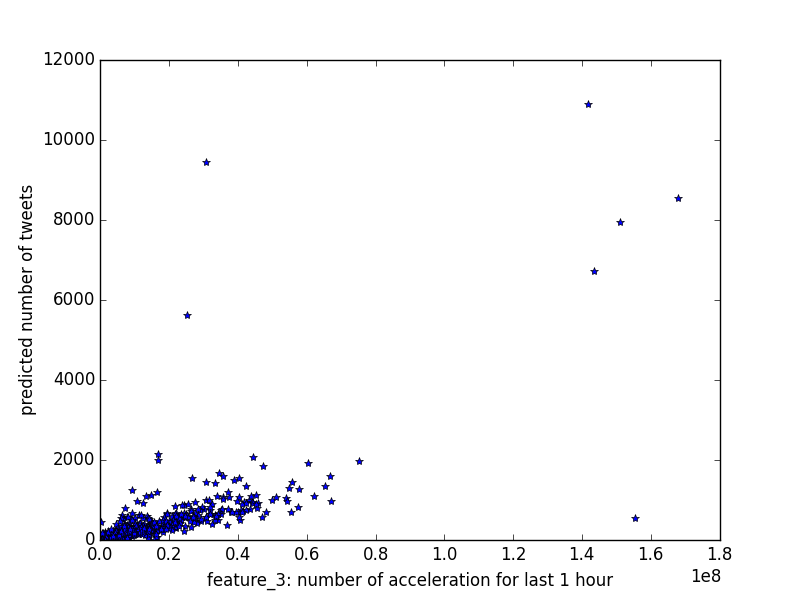


Figure Number of acceleration for last 1 hour

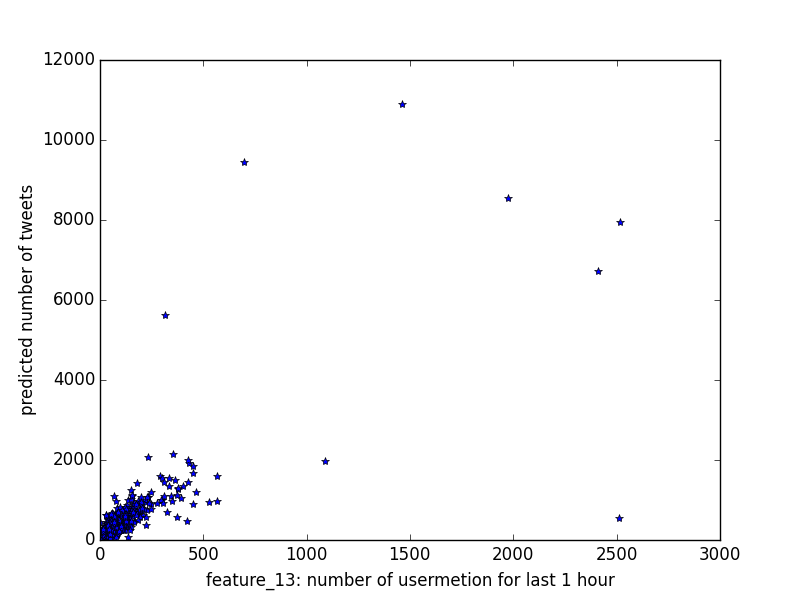


Figure Number of usermention for last 1 hour

At last, we draw the plot of predicted tweets number V.S. actually tweets number. Below are some of the figures:

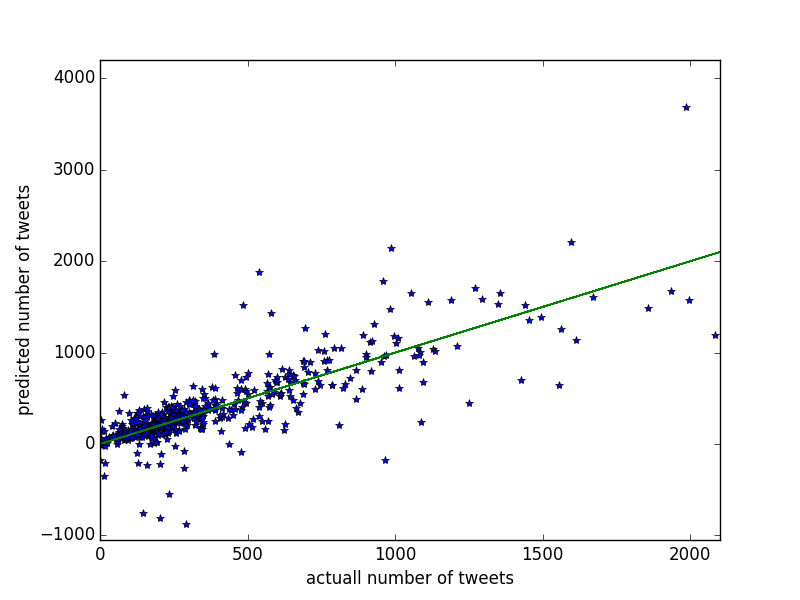


Figure #nfl: predicted tweets number V.S. actually tweets

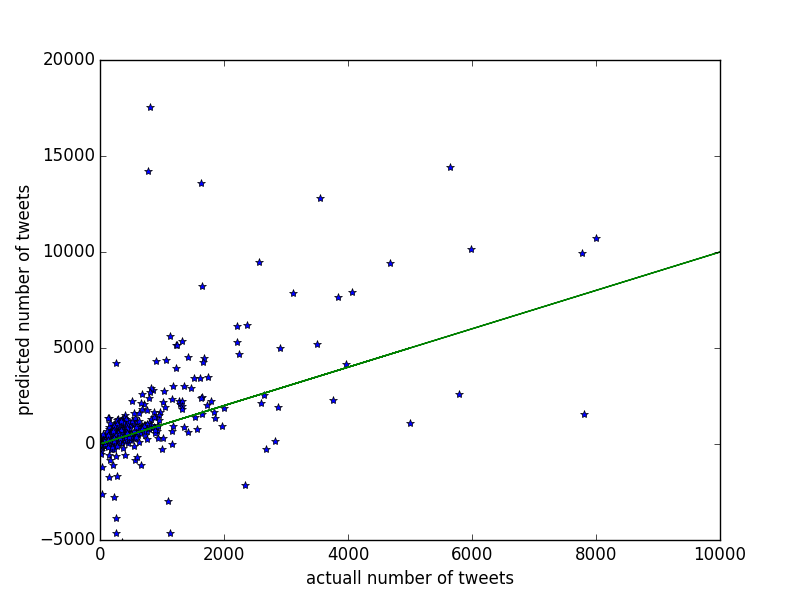


Figure #Superbowl: predicted tweets number V.S. actually tweets

## Question 4

This question requires us to split the data in each time frame into 10 parts and this can be done by dividing the hours in each timeframe by 10. If the number of hours in the window has residual then take the integer result and add the residual to the last part. From our calculation, the residual won't affect the result a lot and this will make the prediction more accurate.

Each time fit the 9 parts of data into the model we design and make prediction to the one training part, and compare the predicted data and the actual data. The errors of predicted results are shown below:

1st test date: Before Feb. 1, 8:00 a.m.

2nd test date: Between Feb. 1, 8:00 a.m. and 8:00 p.m.

3rd test date: After Feb. 1, 8:00 p.m.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Test1 | Test2 | Test3 | Test4 | Test5 | Test6 | Test7 | Test8 | Test9 | Test10 | Average |
| 1st test date | 297.306 | 55.372 | 71.049 | 62.875 | 199.475 | 38.648 | 1.348 | 0.501 | 0.040 | 0.050 | 72.666 |
| 2nd test date | 8163.413 | 6596.606 | 5543.256 | 8353.510 | 5102.332 | 793.236 | **1472.302** | 404.538 | 1681.548 | 7203.943 | 4531.468 |
| 3rd test date | 61.396 | 137.809 | 73.832 | 72.732 | 64.600 | 147.859 | 166.690 | 93.693 | 219.461 | 101.179 | 113.925 |

**#GoPatriots Prediction Errors**

**#NFL Prediction Errors**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Test1 | Test2 | Test3 | Test4 | Test5 | Test6 | Test7 | Test8 | Test9 | Test10 | Average |
| 1st test date | 7.618 | 8.506 | 14.288 | 9.652 | 9.374 | 47.676 | 13.932 | 4.654 | 0.3968, | 0.393 | 11.649 |
| 2nd test date | 548.369 | 3455.977 | 1871.218 | 1891.251 | 400.363 | 705.802 | 424.140 | 1419.262 | 1583.471 | 19.105 | 1231.895 |
| 3rd test date | 0.000 | 0.728 | 0.156 | 0.000 | 1.009 | 0.923 | 0.629 | 0.969 | 46.191 | 18.256 | 6.886 |

**#GoHawks Prediction Errors**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Test1 | Test2 | Test3 | Test4 | Test5 | Test6 | Test7 | Test8 | Test9 | Test10 | Average |
| 1st test date | 183.412 | 104.782 | 21.673 | 20.645 | 10.884 | 9.033 | 7.741 | 22.081 | 1.783 | 9.520 | 39.155 |
| 2nd test date | 1857.236 | 4352.874 | 675.973 | 3899.488 | 2492.045 | 1634.968 | 2273.988 | 8863.653 | 6941.768 | 6766.916 | 3975.891 |
| 3rd test date | 14.417 | 52.133 | 27.495 | 26.202 | 30.557 | 45.252 | 62.734 | 55.190 | 55.190 | 781.697 | 147.860 |

**#Sb49 Prediction Errors**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Test1 | Test2 | Test3 | Test4 | Test5 | Test6 | Test7 | Test8 | Test9 | Test10 | Average |
| 1st test date | 115.106 | 337.803 | 167.248 | 624.824 | 357.829 | 45.262 | 39.019 | 1.794 | 2.156 | 1.112 | 169.215 |
| 2nd test date | 702.633 | 8378.557 | 2974.994 | 5120.371 | 1985.151 | 2628.239 | 1327.232 | 1084.528 | 2977.237 | 331.014 | 2750.995 |
| 3rd test date | 4.299 | 2.066 | 4.305 | 3.467 | 9.627 | 1.595 | 5.573 | 4.554 | 95.163 | 4552.016 | 456.967 |

**#SuperBowl Prediction Errors**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Test1 | Test2 | Test3 | Test4 | Test5 | Test6 | Test7 | Test8 | Test9 | Test10 | Average |
| 1st test date | 883.105 | 201.626 | 64.838 | 525.658 | 32.212 | 7.648 | 0.104 | 0.061 | 0.084 | 0.060 | 171.539 |
| 2nd test date | 2101.1246 | 7590.3455 | 1390.4964 | 1510.5366 | 6949.0058 | 4980.1010 | 26293.9222 | 2128.0402 | 8454.0573 | 5591.2663 | 6698.8896 |
| 3rd test date | 54.159 | 114.035 | 47.928 | 86.447 | 51.261 | 189.649 | 188.460 | 352.930 | 882.061 | 800.366 | 276.730 |

**#Patriots Prediction Errors**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Test1 | Test2 | Test3 | Test4 | Test5 | Test6 | Test7 | Test8 | Test9 | Test10 | Average |
| 1st test date | 142.280 | 179.730 | 413.547 | 559.150 | 62.751 | 19.517 | 4.844 | 0.290 | 0.220 | 0.398 | 138.273 |
| 2nd test date | 4524.449218 | 4497.218817 | 1838.295156 | 1806.203918 | 1000.971143 | 1352.983372 | 8387.973921 | 4158.853548 | 8068.959703 | 3371.669892 | 3900.757869 |
| 3rd test date | 18.542 | 27.007 | 26.703 | 22.384 | 22.488 | 28.580 | 35.128 | 111.251 | 117.174 | 355.701 | 76.496 |

From these six tables we can conclude that our model works well on the first and third time period in which we only have less than 200, however in the second time period, the errors are little biit large, which is attribute to two reasons. First, the data is not sufficient for the model. Second, the linear regression model has its limitation to predict the trend of tweets which actually is non-linear.

## Question 5

In this question, we download the test data and run our model to make predictions for the next hour in each case.

For this problem, we are given the original “txt” data without the “hashtag” and tweet “title”! So, we cannot use the features that derived from tweet text analysis. Besides, we cannot use the model created in question 4 directly. However, the 10 samples in this question belong to specific period. So we can use the data samples that belong to the same period together, creating a fitting model to predict the tweets in next hour.

After doing this, the data for training each model is about 10-20 observations, which is far better than the 6 hours period before. Therefore we can make the prediction more accurate and robust.

The result is as follow:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | period | 2nd hour | 3rd hour | 4th hour | 5th hour | 6th hour | Next hour |
| Sample1 | 1 | 82 | 68 | 94 | 171 | 178 | 290.0 |
| Sample2 | 2 | 9361 | 10377 | 20078 | 81973 | 82890 | 60527.4 |
| Sample3 | 3 | 550 | 610 | 888 | 616 | 523 | 155.3 |
| Sample4 | 1 | 257 | 236 | 266 | 267 | 201 | 187.0 |
| Sample5 | 1 | 508 | 353 | 362 | 281 | 213 | 1.6 |
| Sample6 | 2 | 12943 | 60627 | 52695 | 41016 | 37293 | 182416.5 |
| Sample7 | 3 | 102 | 66 | 60 | 55 | 120 | 0 |
| Sample8 | 1 |  | 72 | 56 | 41 | 11 | 4.0 |
| Sample9 | 2 | 1729 | 1619 | 1582 | 1857 | 2789 | 1821.2 |
| Sample10 | 3 | 54 | 68 | 62 | 58 | 61 | 34.8 |

Table Prediction of each sample

Besides we also plot the development trend of the tweet for the 10 sample.

|  |  |
| --- | --- |
| sample1_period1_tweets trend | sample2_period2_tweets trend |
| sample3_period3_tweets trend | sample4_period1_tweets trend |
| sample5_period1_tweets trend | sample6_period2_tweets trend |
| sample7_period3_tweets trend2 | sample8_period1_tweets trend |
| sample9_period2_tweets trend | sample10_period3_tweets trend |

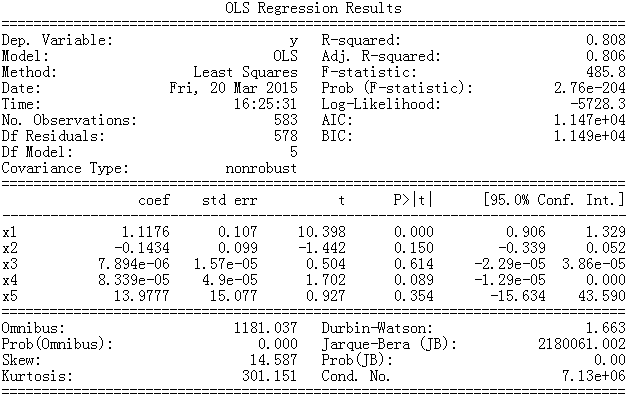
Figure development trend of the tweet for the 10 sample

From these figure above, we can see that most of the prediction value is consistence with the previous data. However, in sample7 and sample9, the prediction seems a little odd. This is mainly because of the lack of sufficient data. In summary, our model shows a good performance in predicting the number of tweets and popularity of a hashtag.

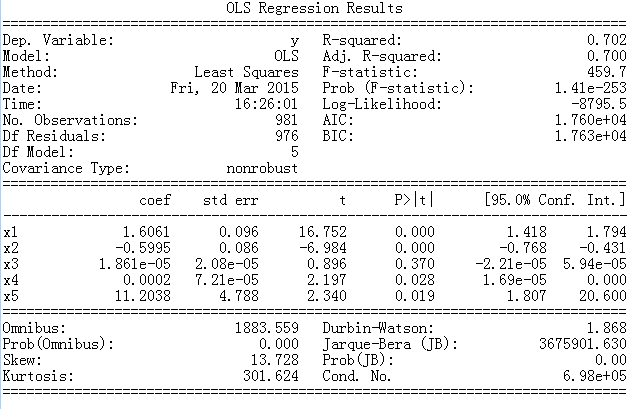
## Appendix

##### Summary

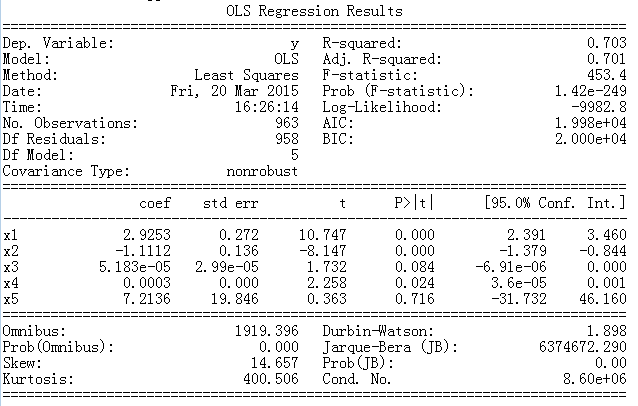
#Sb49



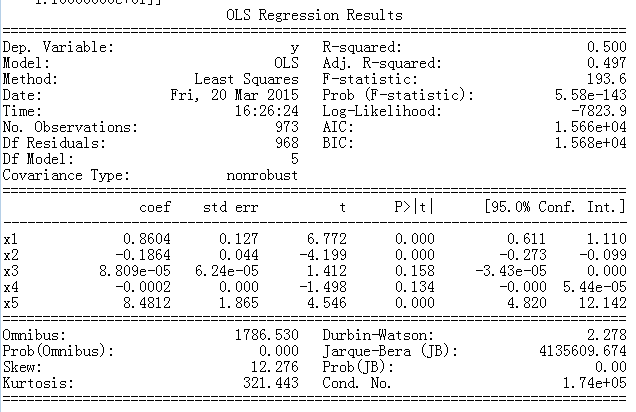
#patriots



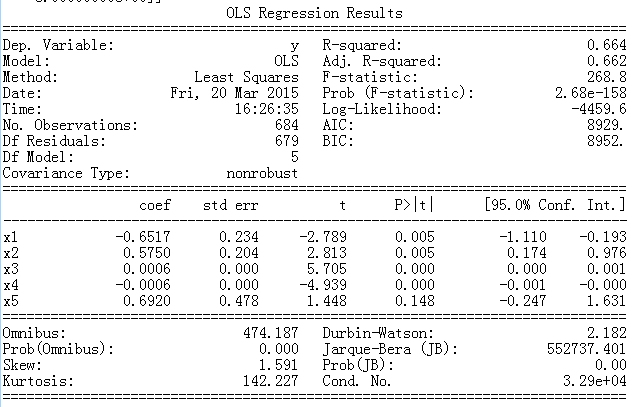
#superbowl



#GoHawks



#Patriots



##### Features