Linear Statistical Models

SYS 6021

Project 2

Analysis of Spam and Ham Datasets

Debajyoti Datta
dd3ar@virginia.edu

Honor Pledge: On my honor, I pledge that I am the sole author of this paper and I have accurately cited all help and references used in its completion.

Summary

The spam dataset [3] which was part of the collection of emails from the Hewlett Packard Labs was analyzed using static and time series filters. Some features or variables were more prominent than the other as could be seen in equation 4. The MSE value of the stepwise regression was 0.06246 and that of the main effects model was 0.06236. Certain variables have more distinguishable features in spam and some have more distinguishable features for ham which helped in a reasonably distinction of the two models in the biplot. The base model to which this was compared was the mean(V58) (equation 2) and the model with only capital letters. (equation 3). The static filter was then developed with the help of a stepwise regression model. The confidence intervals for the model of the mean of spam dataset and the ham dataset had a Chi-Squared value of 2.2e-16 suggesting that there are definitely variables that affect the classification of spam from ham and the distinction is not arbitrary. The comparison of the main model with the model of capital letters also yielded a confidence level of 2.2e-16 suggesting that not only capital letters but other variables also affect the probability of an email being spam.

For the time series model the p-value for the spam trend model of 1.05e-05 was significant whereas the p-value for the ham trend model was not significant at 0.05339. The mean of the residuals of the spam and ham trend model were 1.79e-16 and 1.8e-16 respectively which was quite close to zero. The best spam model which had an AIC of 2493.081 had p,d, and q values of 1,0 and 1 respectively. The d value is zero and that is understandable since from the initial plots of the ts we found that the constantvariance has already been accounted for in the plot of the time series data without any difference. For ham however the best model has an AIC 2482.727 and that however has a p,d and q value of 3, 0 and 2 respectively. The final prediction done on the test set for the ham dataset did significantly well since ham had a strong seasonal component and the root mean square error was 6.446556. Plot of the confidence level shows that the prediction is well within the 95% confidence level.

Problem Description

1.1 Situations

Nowadays dealing with spam may not seem like a big deal since the spam filters are really good but the cost of spam is huge. In fact 94 billion spam messages are sent daily [5]. And not only that, the annual cost to the soceity is 20 billion per year due to spam. That eleven-figure number is derived from the cost of developing the software required to filter out spam emails and the few seconds it takes to delete every spam email that isn't successfully blocked. You're affected in more subtle ways as well: keep in mind that spam forces the engineers at Google, Yahoo or any other email provider to spend their time fighting spam, rather building new fun features. And because it's not just a few people footing the bill, but pretty much everyone who's ever used email, there's little political incentive for laws that really crack down on spammers.

The concept of spam is diverse. Spam emails also sometimes known as a junk email or unsolicited bulk email includes nearly identical emails sent to numerous recipients via email [6]. It can include, but not limited to advertisements for products/web sites, various promotional events, making money schemes or winning award schemes, or chain letters. The spam [3], is a collection of emails from the Hewlett Packard Labs, and they have been classfied into spam by their internal postmaster or individuals who filed spam. This project [1] uses the template [2] and the source code [4].

1.2 Goal

The goal of the project is to be able to be able to find different features for classifying emails into spam or ham based on the different features of email text like word frequency, number of capital letters, the distribution of capital letters and certain character counts.

The goal for the time series model is to be able to accurately figure out the trend, seasonal, cyclical components of the time series models for spam and ham.

1.3 Metrics

- For the static filter the tradeoff between the predicted model and the actual data in terms of the errors made in detection of spam or ham.
- For the time series filter the metric is to actually see the performance of the time series model of spam and ham on the test set to validate the accuracy of the model.

1.4 Hypothesis

H0 : That all the variables in the spam dataset are not good predictors to distinguish spam from ham.

H1 : That some of the variables in the spam dataset are good predictors to distinguish spam from ham.

H0 : In the time series model the ham does not have any seasonal component to it.

H1 : In the time series model the ham does have a seasonal componet to it.

H0 : In the time series model the spam does not have any seasonal component to it.

H1 : In the time series model the spam does have a seasonal componet to it.

Approach

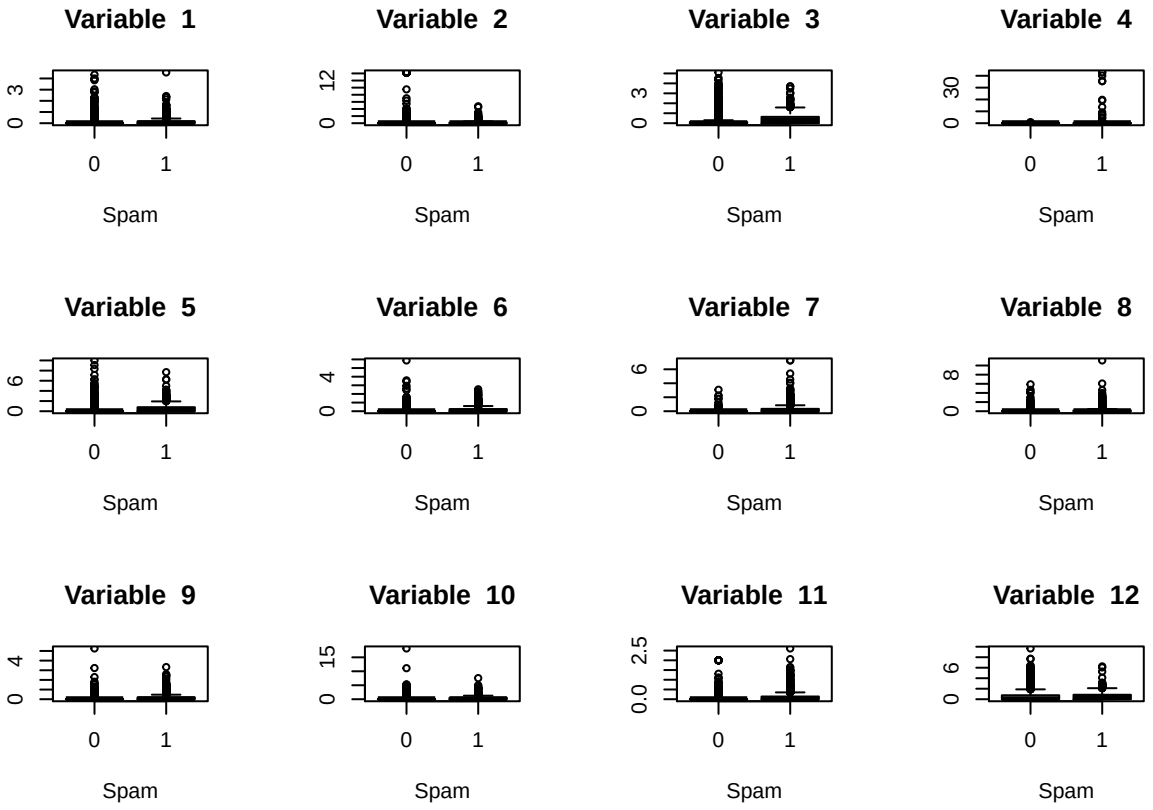
2.1 Data

The spam dataset[3] , is a collection of emails from the Hewlett Packard Labs, and they have been classfied into spam by their internal postmaster or individuals who filed spam.

This dataset with the variables mentioned above does not have any missing values. Also this dataset has no significant bias that has been observed.

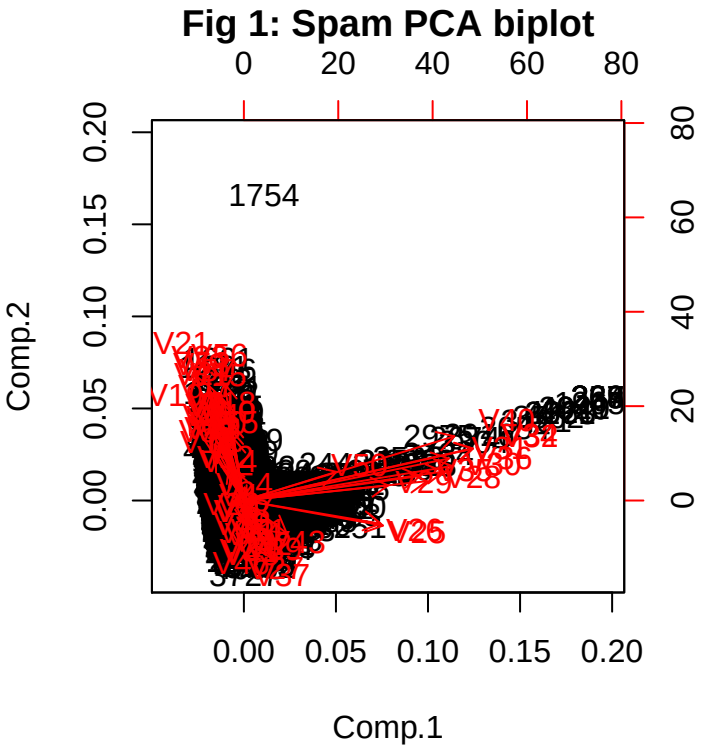
There are 48 continuous real [0, 100] attributes of word frequencies, i.e 100 * (number of times the WORD appears in the e-mail) / total number of words in e-mail. There are 6 continuous real attribute of type char frequency similarly. The last 3 attributes attributes (55-57) measure the length of sequences of consecutive capital letters. The final column of the dataset is a nominal {0,1} class attribute of type spam where 1 denotes it is a spam and 0 otherwise.

An observation of some of the first 12 predictor variables through boxplots in spam and ham are as follows:



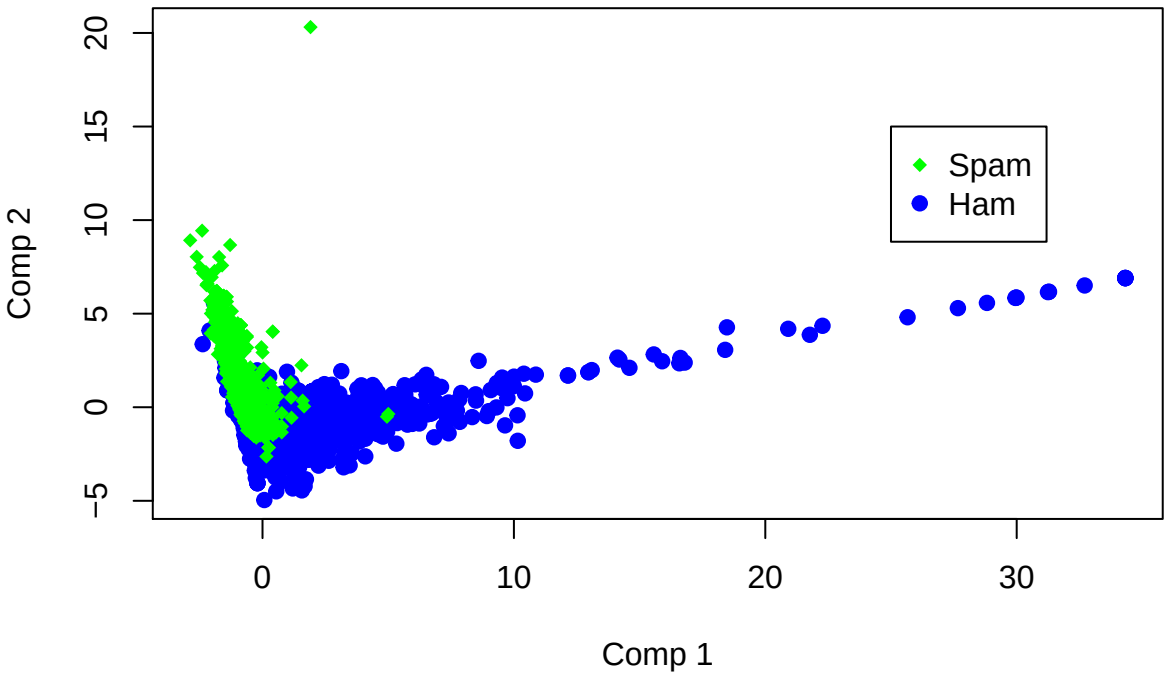
As can be seen that there are certain variables, namely variables 3,4,5,7,11 and 12 have distinguishable differences between spam (1) and ham (0).

Now creating a biplot to see the various effect of different variables in the principal component analysis we see a distinct pattern in the various variables, and that most of the variables account for the variance in the first and second principal components.



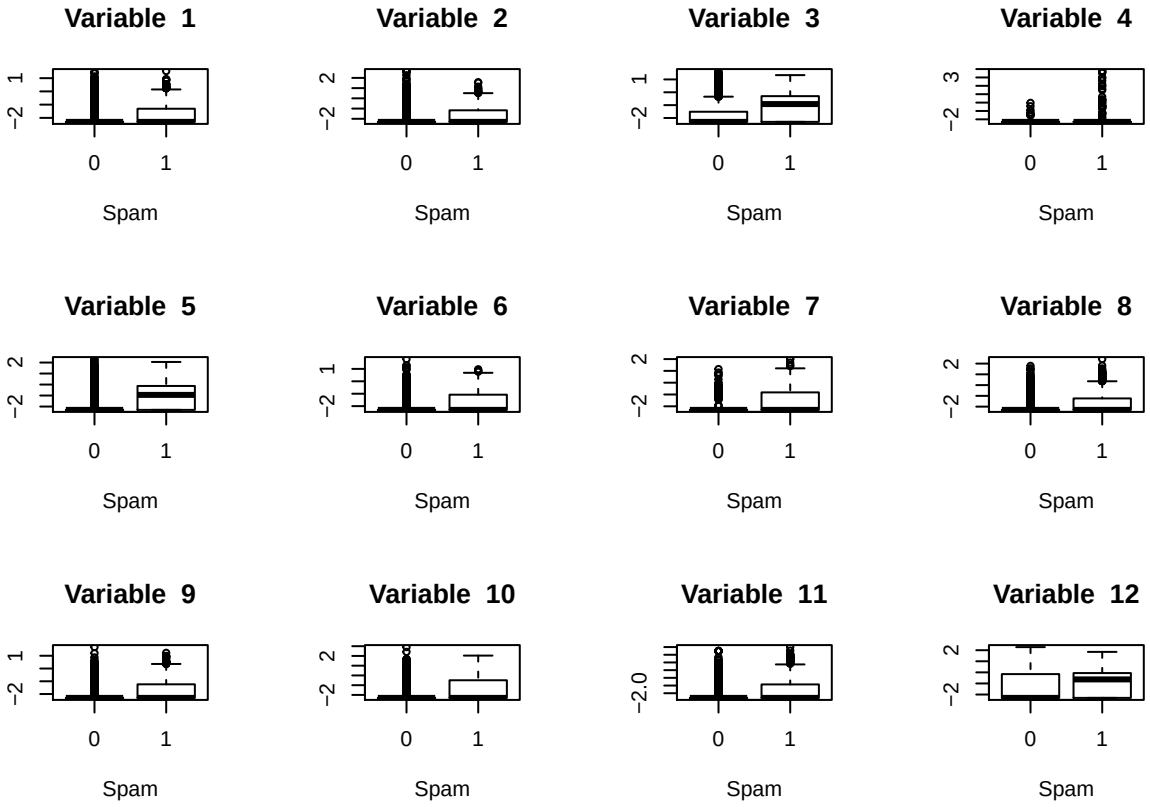
However the data point 1754 seems to be an outlier, as can be seen from the pca graph. The other points are not very different from each other and from the preliminary information not much can be said about the bias in the data.

Observing how spam and ham varies according to these features in the biplot we can see that there are quite a few features that are present more in spam and there are some features that account for most of the variance in ham.

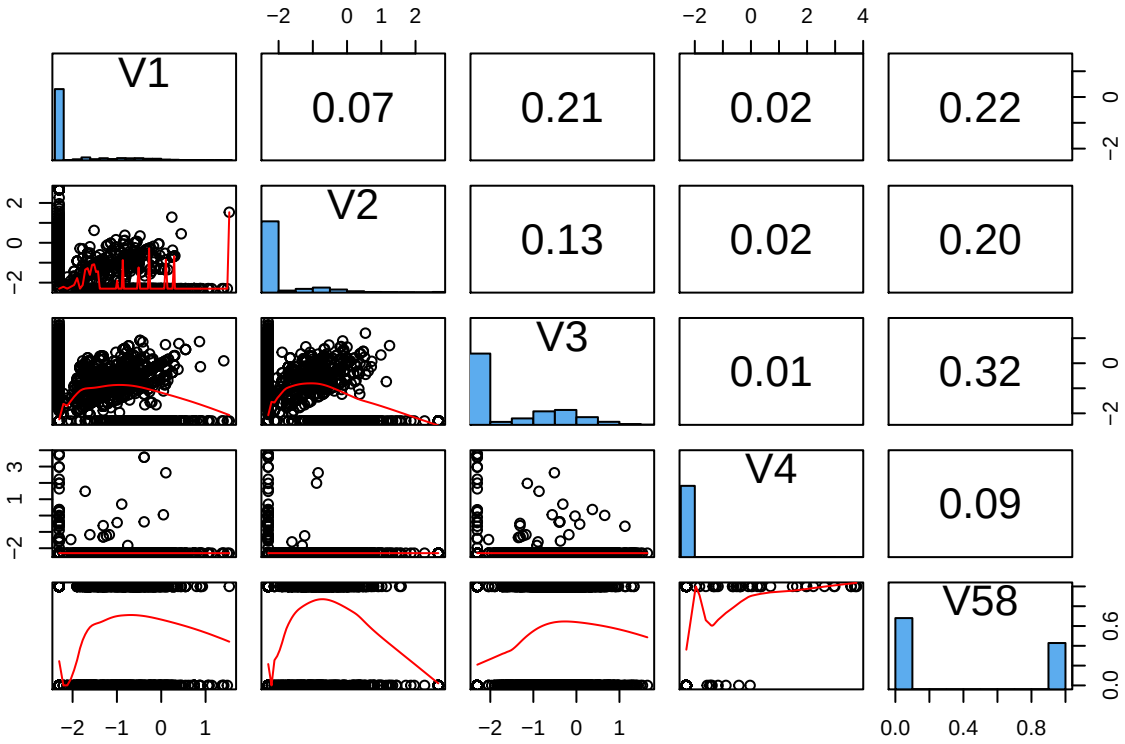


Taking a log of all the predictor variables, (and adding 0.1, because some of the variables are zero) and plotting a biplot we see that the distinction between spam and ham is more pronounced for each of the

variables.



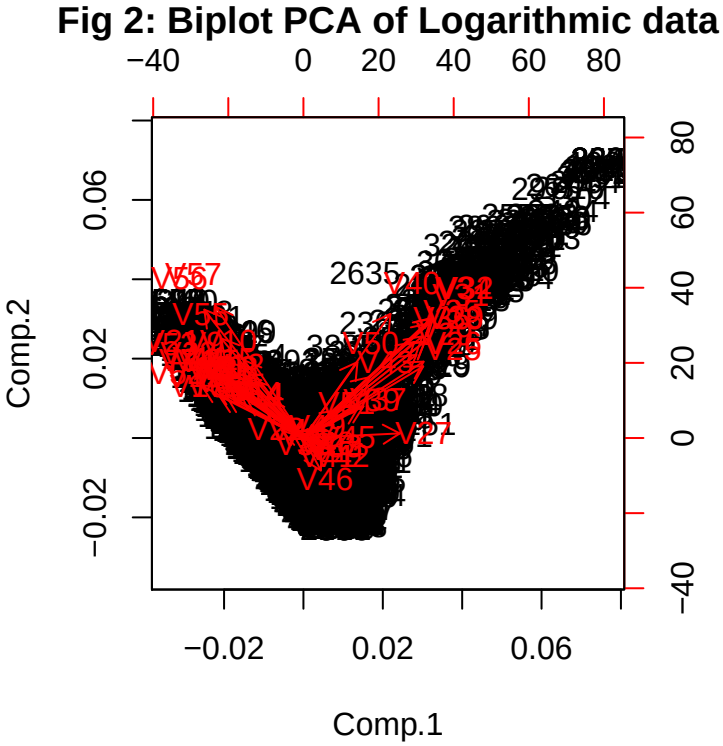
The scatter plot matrix of the first 10 variables after taking a log transform.



2.2 Analysis

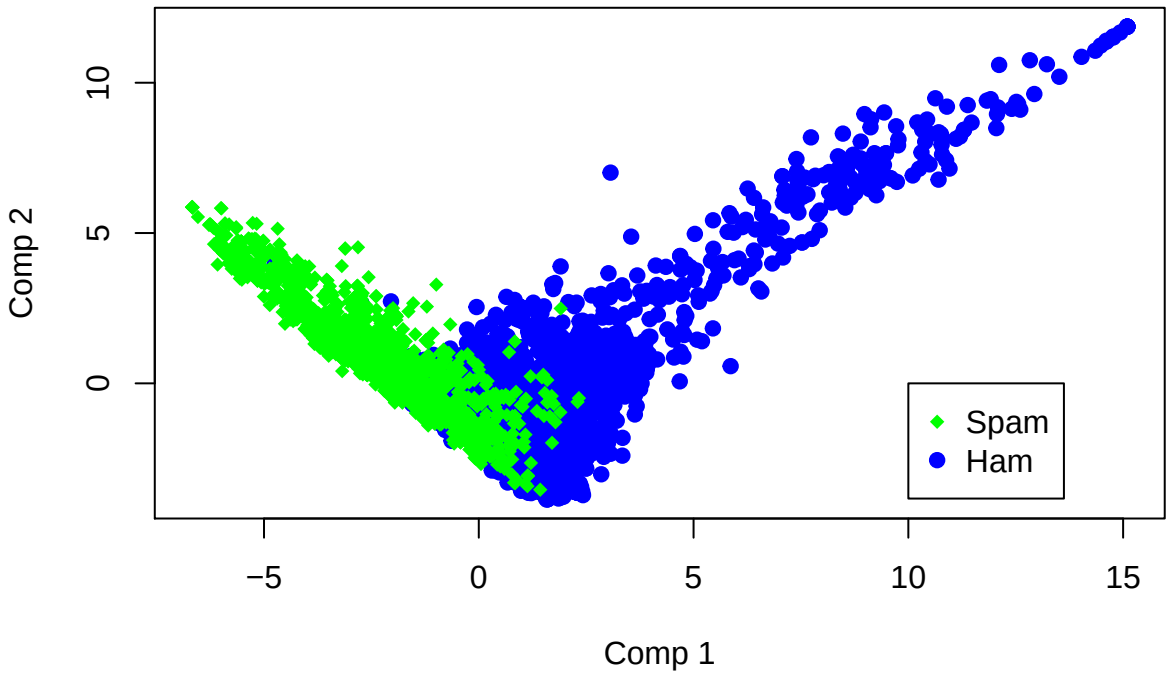
2.2.1 Static Analysis for Spam Filter Design

In the data section in the box plots we had seen that the difference between some of the predcitor variables was not very pronounced. Thus taking a log transform of the entire dataset of the predictor variable we explore the biplot below.



From the biplot we see that lots of variables are correlated. the two branches of the L like thing show how their are two sets of the correlation that exists.

Now the biplot above did not tell us which was spam or ham so a biplot factor graph below is drawn that shows the different variations.



As can be seen some of the variables have the most amount of significance in a spam, especially in principal component 2 and some variables have more significance and impact on principal component 1 in ham.

This clear distinction between the variables for spam and ham in the factor biplot above will help in coming up with a better predictor model for spam and ham.

Now for the model we use logistic regression to predict for variables V58, against all the predictor variables.

The model is as follows:

$$\begin{aligned} V58 = & \beta_0 \\ & + \beta_1 * V1 \\ & + \beta_2 * V2 \\ & + ... \\ & + \beta_{56} * V56 \\ & + \beta_{57} * V57 \end{aligned} \tag{1}$$

The base model against which it is compared is:

$$avg(V58) \tag{2}$$

Another model of capital letters only is also used to compare it with the model with all variables.

$$\begin{aligned} V58 = & \beta_0 \\ & + \beta_1 * V55 \\ & + \beta_2 * V56 \\ & + \beta_{57} * V57 \end{aligned} \tag{3}$$

2.2.2 Time Series Analysis for Spam Filter Design

The time series analysis provides a lot of key insights in data related to variations in time or periodicity and so on. The plot of the time series data of the count of ham and spam is below.

Fig 3: Ham Trend Series plot

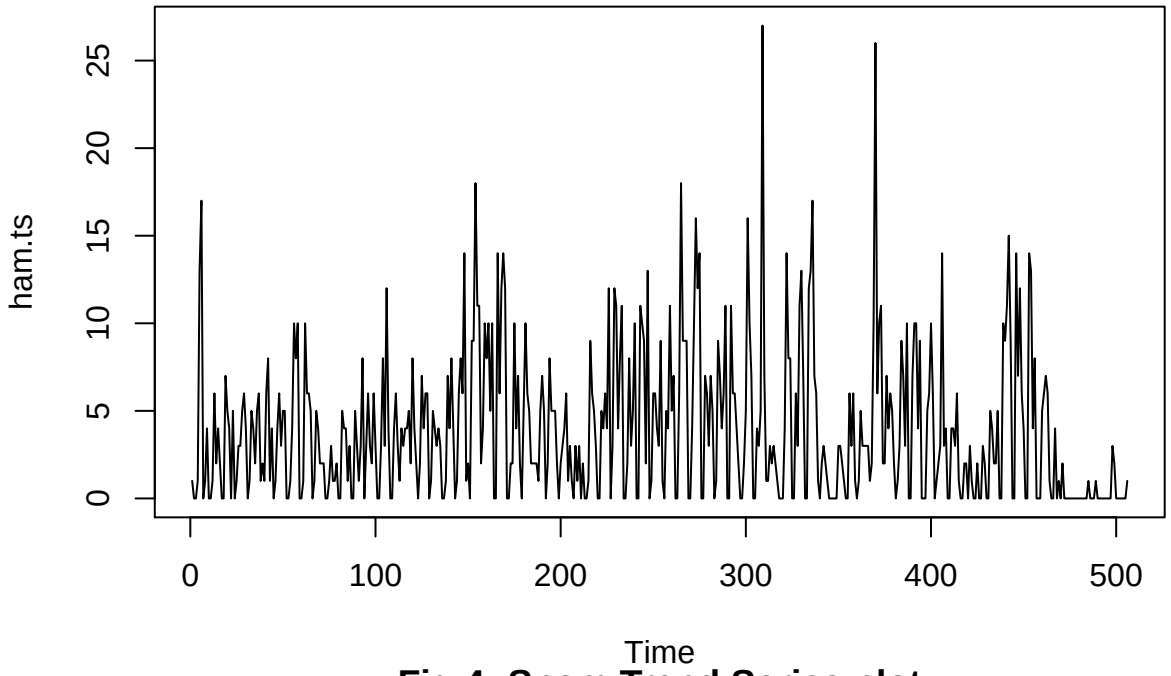
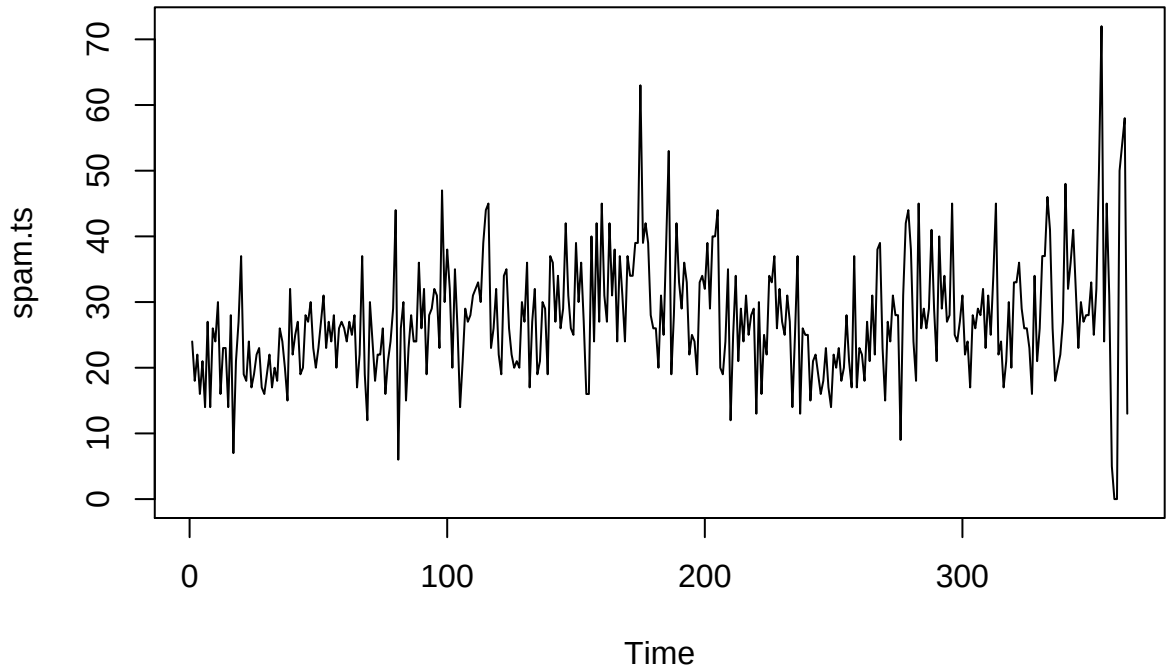


Fig 4: Spam Trend Series plot



Now we are checking if the insample forecast shows non-zero auto correlationsfor spam and ham.

Fig 5: ACF plot of spam time series

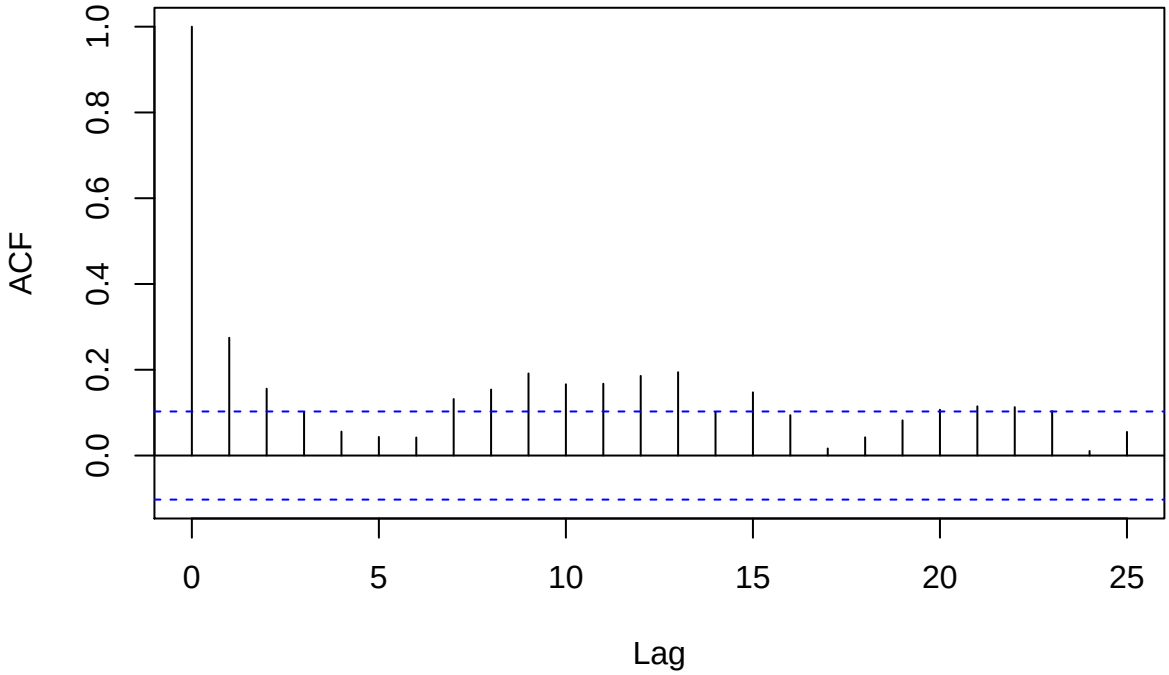
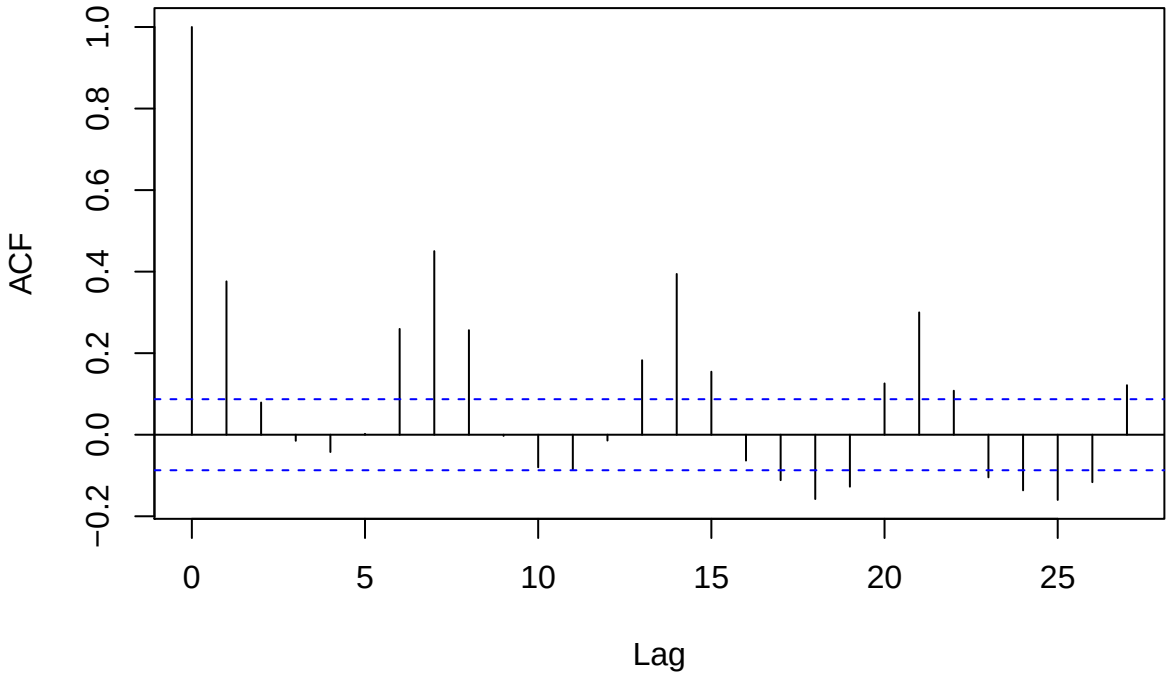


Fig 6: PACF plot of spam time series



As we can see that for certain lag values in both spam and ham it exceeds the significance level. Now carrying out a Ljung-Box test to see if any lags upto lag 27 (this value is obtained from the acf diagrams above) is significant.

As we can see from the Ljung-Box test that the p-value in case of both spam and ham happen to be less than

2.2e-16 showing that there is sufficient evidence of non-zero auto correlations in the data.

Thus it is useful to check the periodogram of the spam and the ham data.

Fig 7: Periodogram for spam trend series

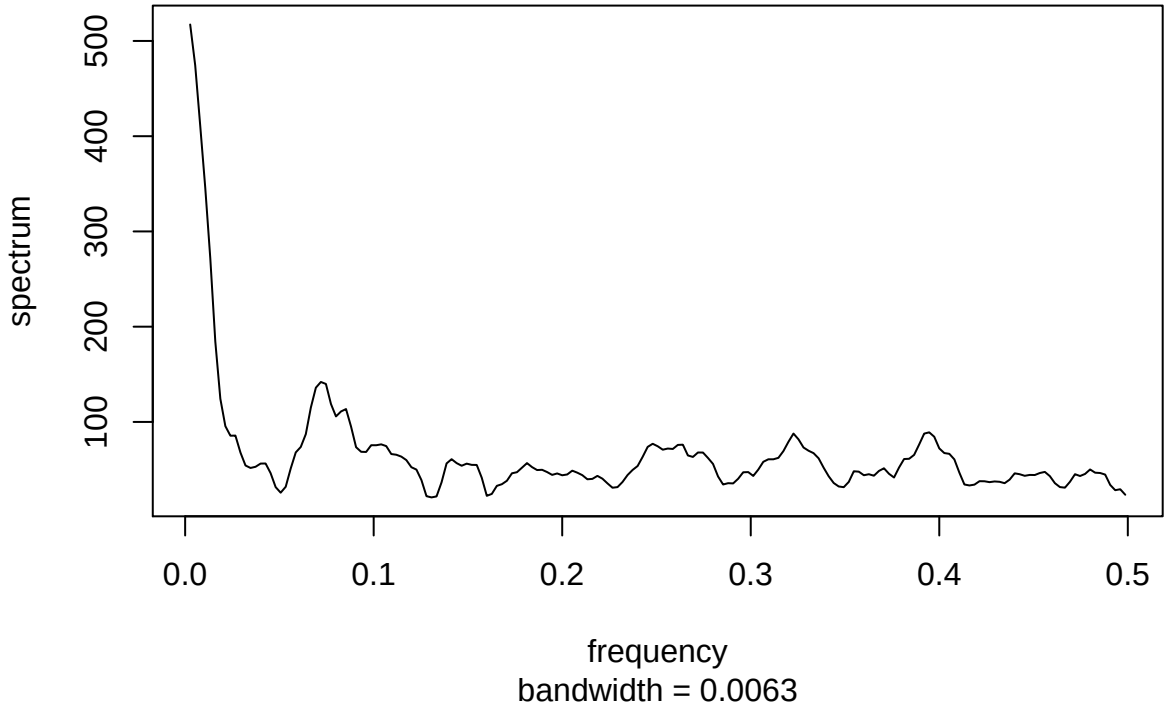
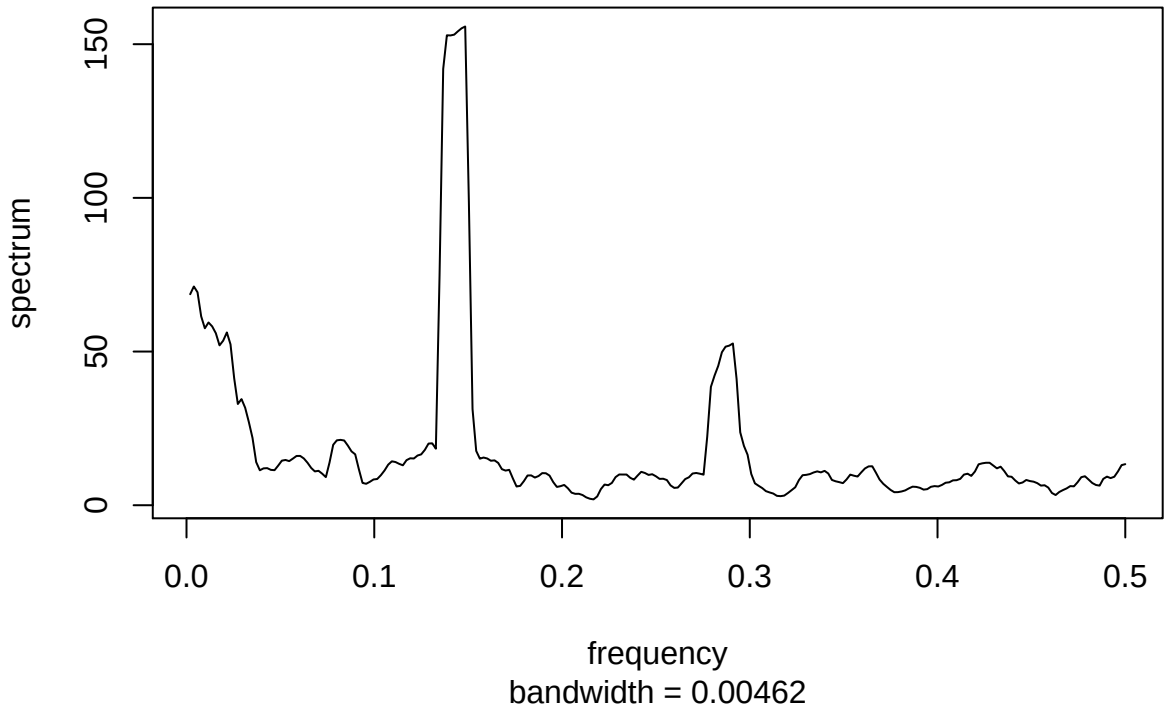


Fig 8: Periodogram for ham trend series

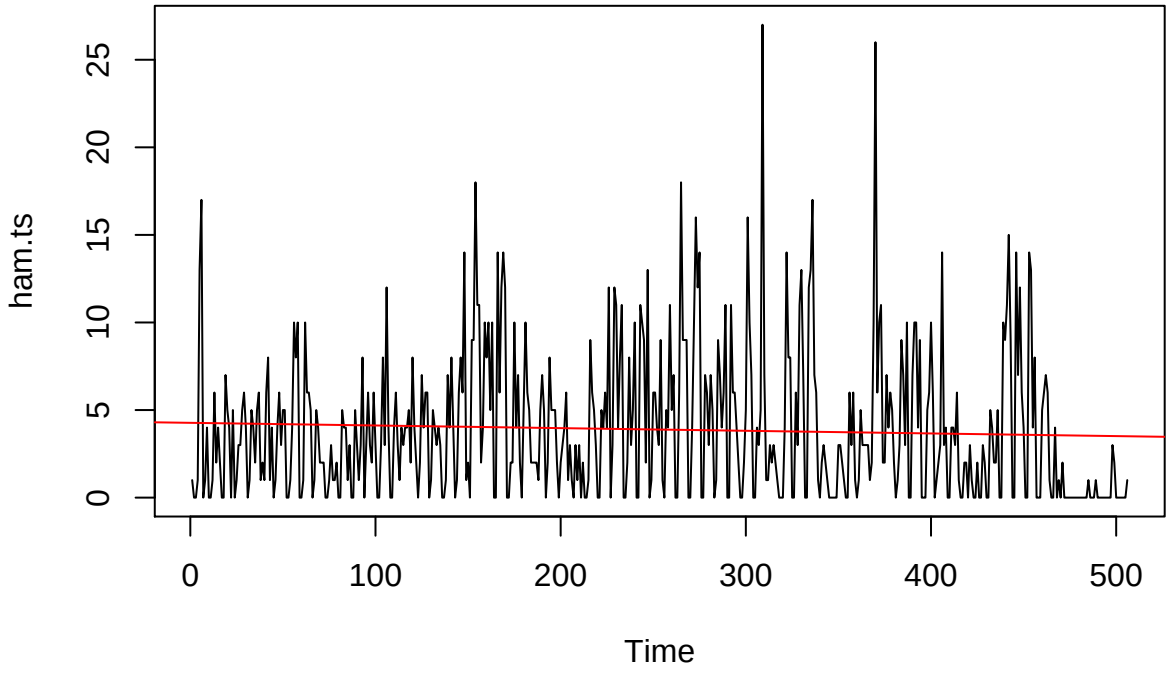


The period of maximum variation for the spam and ham dataset is 375 for spam and about 6.7 for ham. Now since we are exploring just 2 years of data the value 375 does not make much sense. So the approximate variation of ham seems to have a period of 7.

Now based on all these observations a time series analysis of a linear model is being done here.

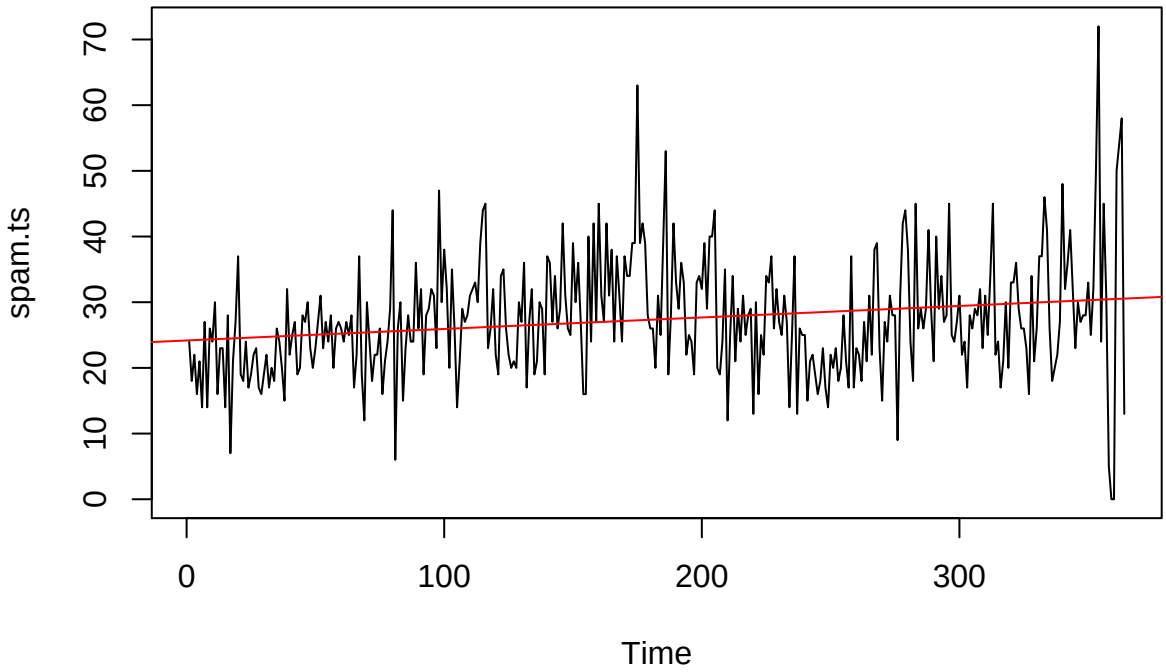
Now we want to model the trend of ham. In order to do so we create a time variable that runs from time unit 1 to the entire length of ham. A linear regression model is built on this time variable and it's effect on ham to see if the count of ham has increased over time. As can be seen from the graph that the trend of ham does not show any significant increase or decrease over time.

Fig 9: Ham Trend



Now we want to model the trend of spam. In order to do so we create a time variable that runs from time unit 1 to the entire length of spam. A linear regression model is built on this time variable and it's effect on spam to see if the count of spam has increased over time. As can be seen from the graph that the trend of spam, unlike ham shows a significant increase in the count.

Fig 10: Spam Trend



Now we will account for seasonality of the data set. In the analysis before we found that ham showed a time period of approximately 7 days from the periodogram. Thus in this case seasonality is modelled using dummy variables, using day of the week as the interval. The level values were coded as 'F,Sa,S,M,T,W,Th' for understanding. Here Friday is the base case. We observe from this analysis that ham.count is significant on Friday, Saturday and Sundays with respect to p-values of 2e-16, 6.20e-13 and 3.97e-14 suggesting that these are significance of these days on the count of ham.

2.2.3 Integrated Filter Design

The time series filter helps take into account the trend component and the time component along with the features of the static filters. This makes the overall model not only more robust but in this case the prediction of spam and ham can take into account for the properties of the feature variables and their effect on spam and ham. Bayes theorem can help us combine the static and the time series filters. In the equation below, ST represent static filter and TS represents time series filter.

Thus using Bayes rule can combine static and time series filters to get probability that email will be spam can be given by the following equation.

$$\Pr(E = S|ST, TS) = \frac{\Pr(ST|E = S)\Pr(TS|E = S)\Pr(E = S)}{\Pr(ST|E = S)\Pr(TS|E = S)\Pr(E = S) + \Pr(ST|E = \neg S)\Pr(TS|E = \neg S)\Pr(E = \neg S)} \tag{4}$$

Now if we consider that the probabilities of static filter and the time series filters are independent then we can calculate the final probability of an email being spam would be:

$$\Pr(email = spam) = \frac{spam}{totalEmail} \tag{5}$$

3 Evidence

3.1 Static Filter Design

For the evidence, in this case we are using CHI squared tests to compare between the two models.

In the above model as we can see that the variables do have a significance in the classification of spam and ham.

It turns out that the main effects model is significant and at least some of the variables are non-zero. The probability of this not being the case is 2.2e-16. Since this value is really small one can reject the null hypothesis with a very high level of confidence.

In fact the main effects model is also compared with the model of all capital letters and it turns out that the other variables other than the capital letters are also significant.

It turns out that the main effects model is significant and at least some of the variables are non-zero. The probability of this not being the case is 2.2e-16 when compared against the capital letters. Since this value is really small one can say with confidence that not only the capital letters but other variables are also signifcant predictors in the distinction of spam and ham.

The problems with some of the models are that some variables are more signifcant than the others and some add noise and do not lead to significant development of the model. Thus variables which have low significance were dropped using a step wise regression to adjust for the inconsistencies.

The model diagnostics for the final model is as follows:

Fig 11: Residual vs. Fitted

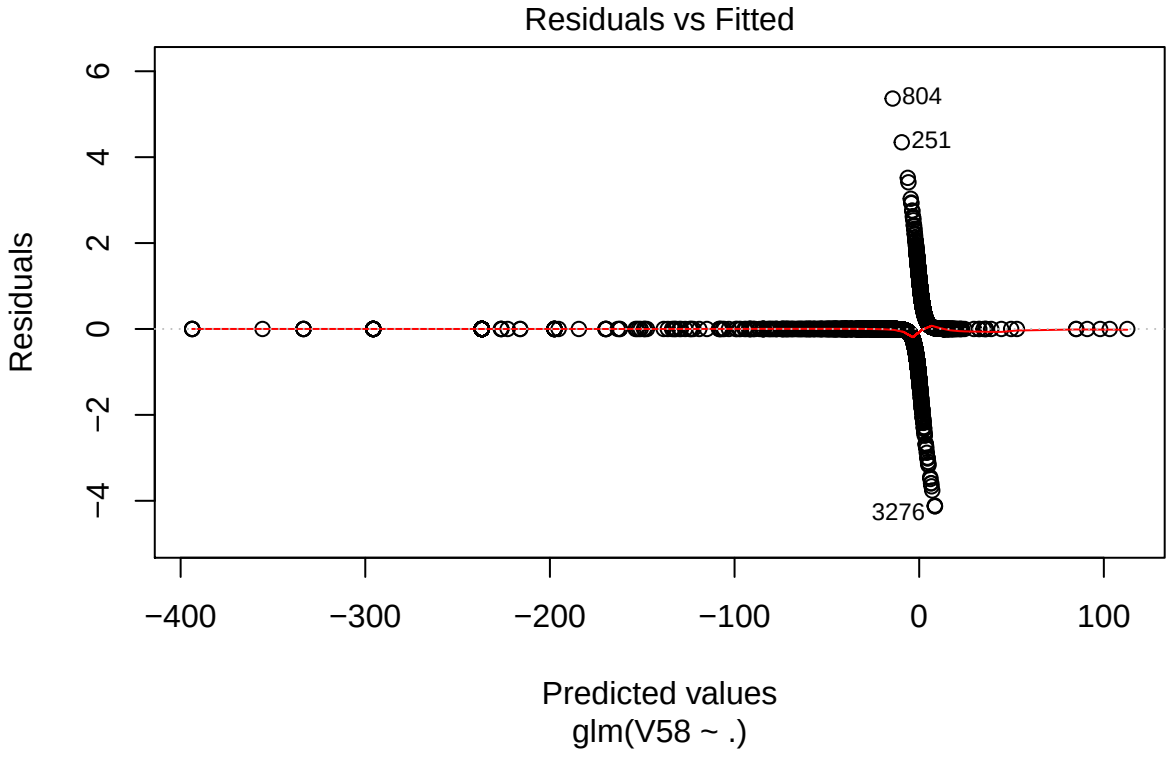


Fig 12: QQ Plot

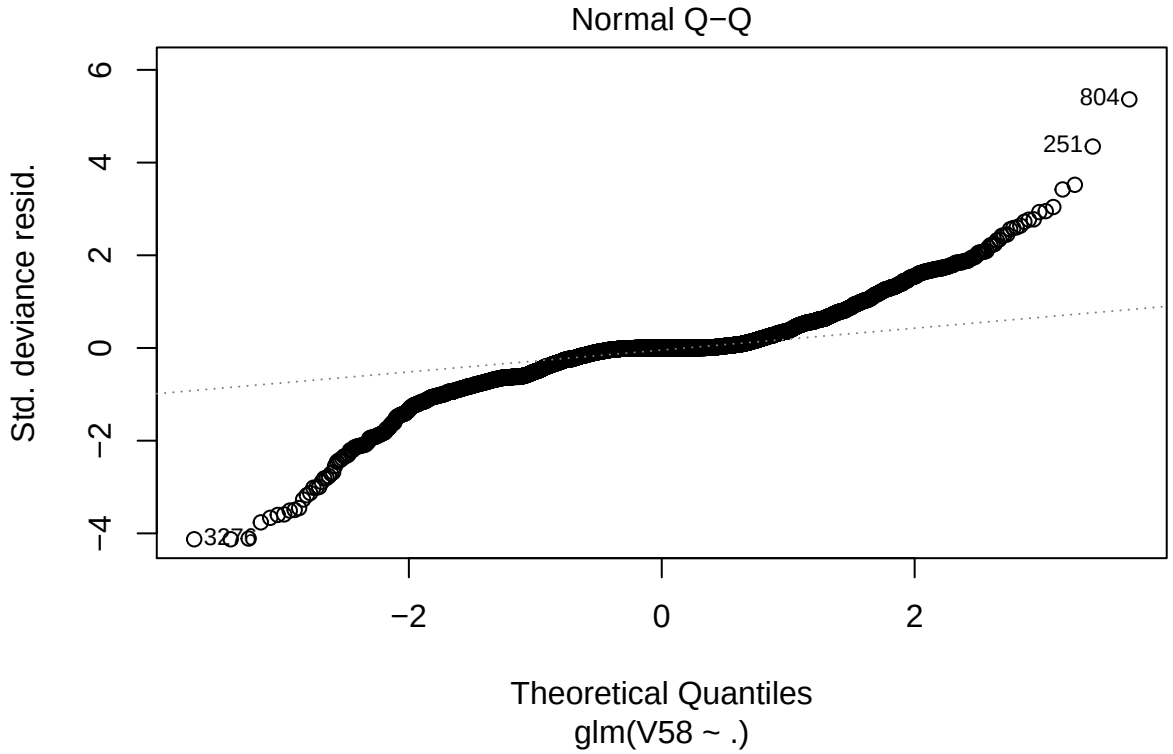


Fig 13: Scale Location Plot

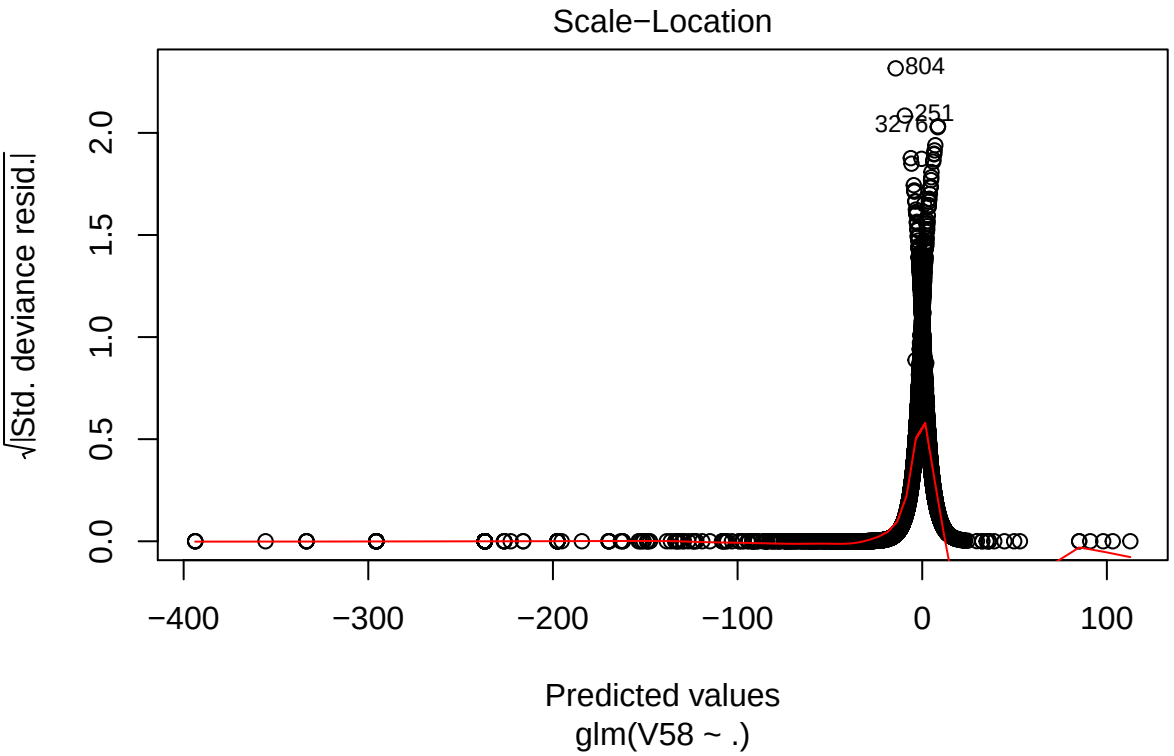


Fig 14: Cook's distance plot

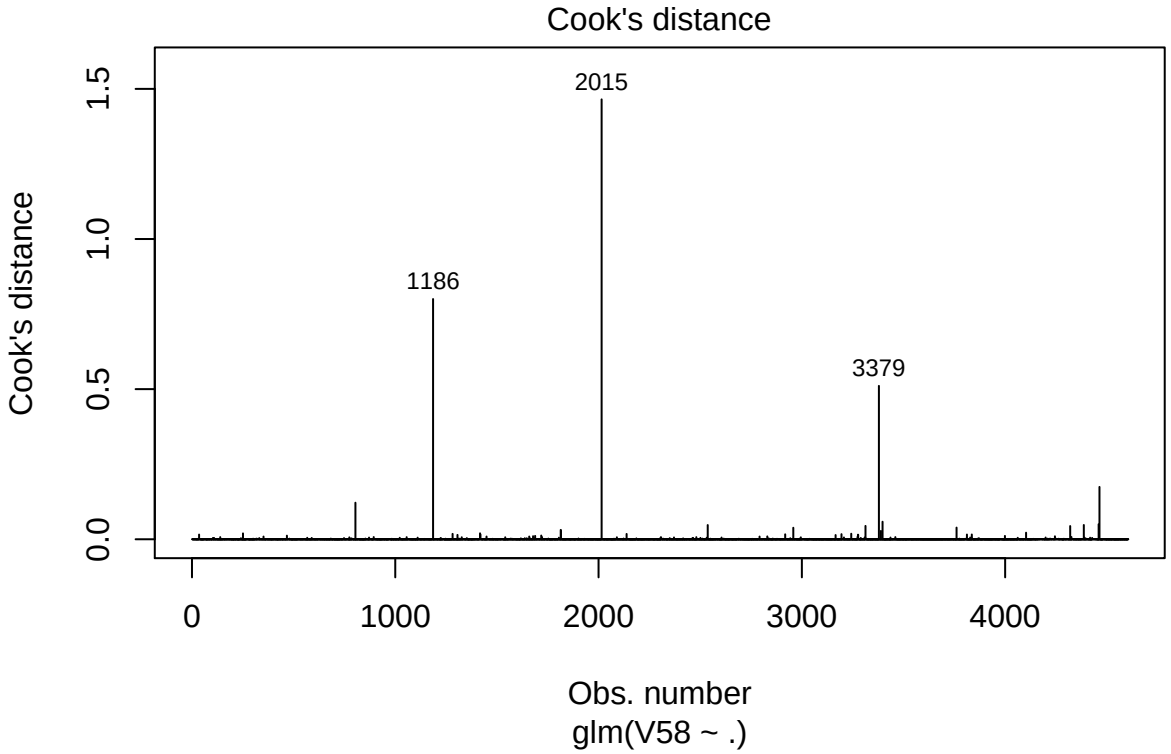


Fig 15: Residuals vs Leverage

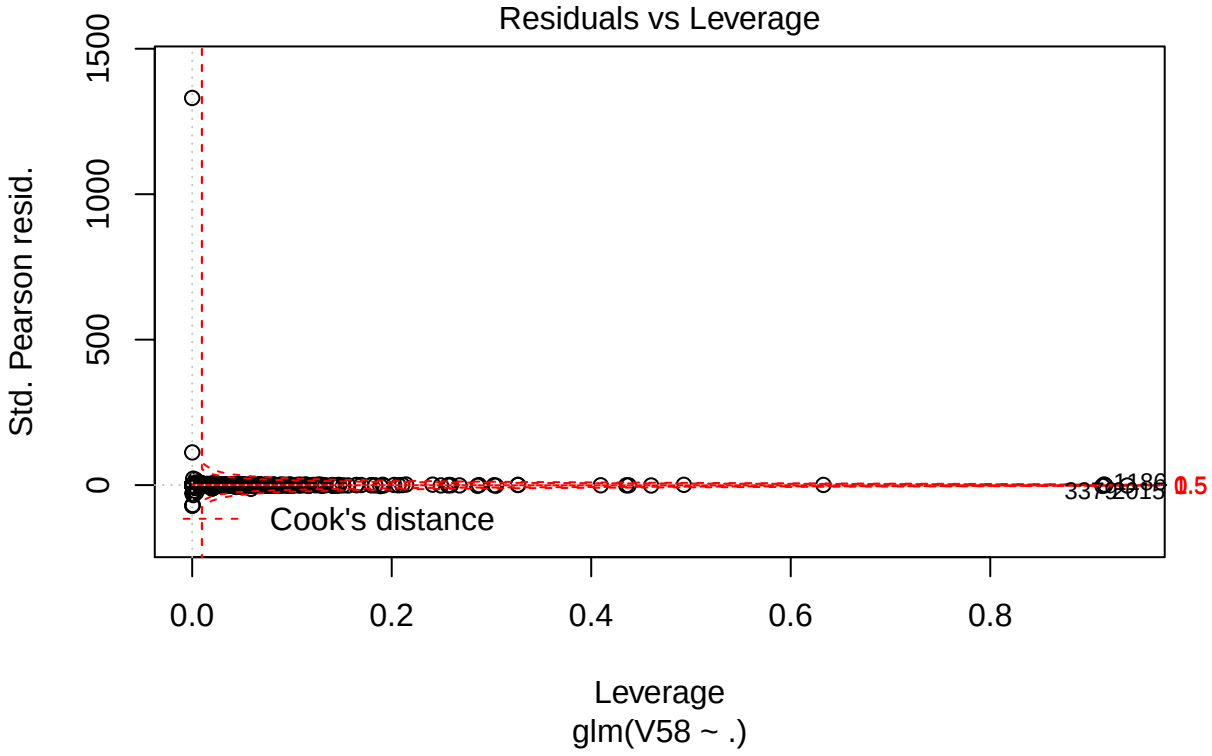
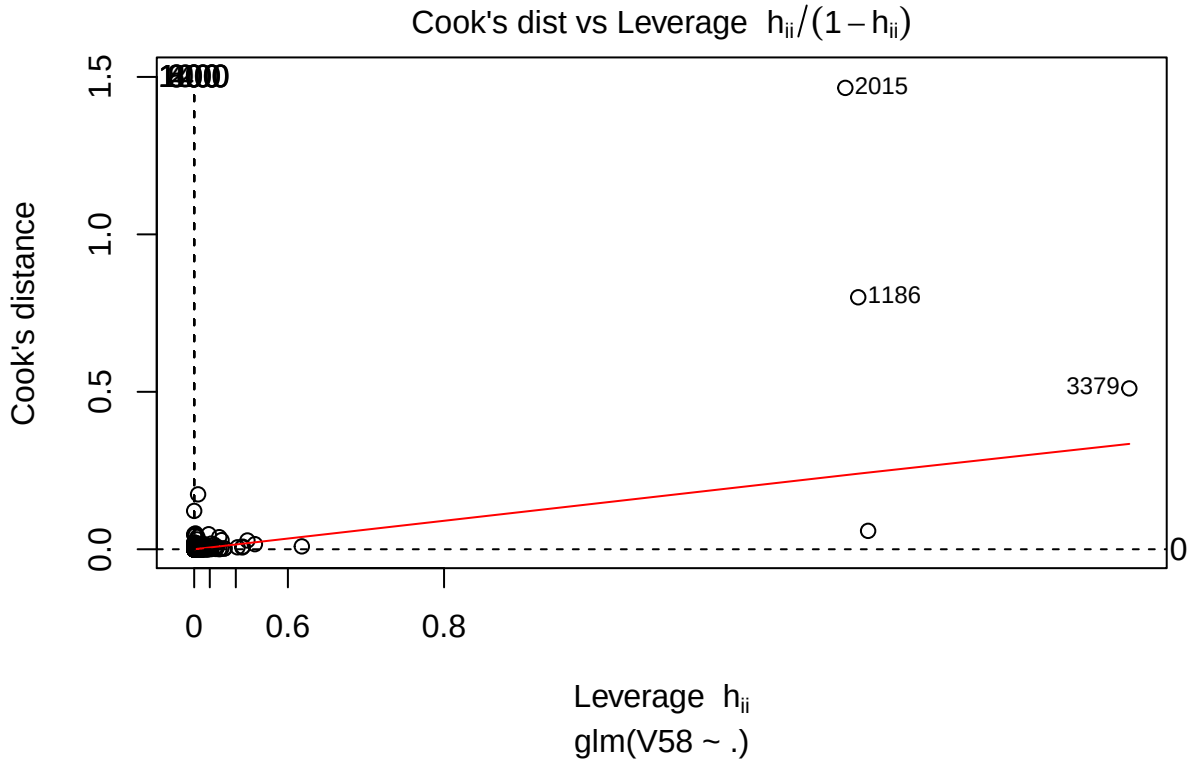


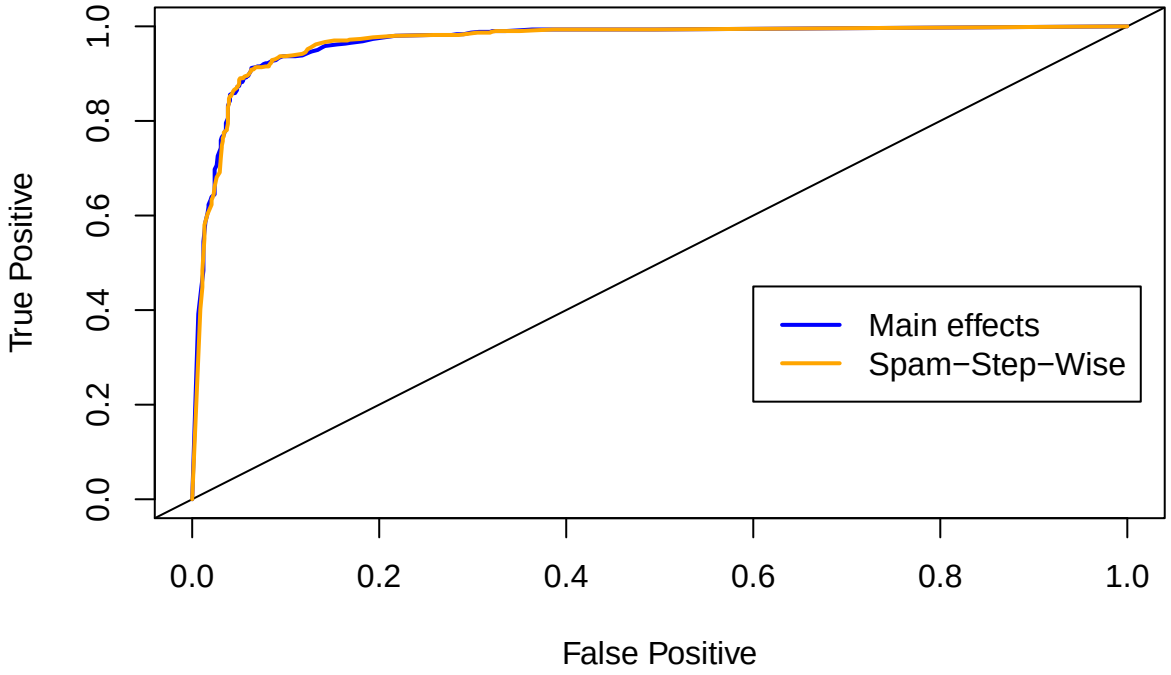
Fig 16: Cook's dist vs. Leverage



So the final model is as follows:

$$
\begin{aligned}
V58 = {} & \beta_0 + \beta_1 * V1 + \beta_2 * V2 + \beta_3 * V4 + \beta_4 * V5 + \beta_5 * V6 + \beta_6 * V7 + \beta_7 * V8 + \beta_8 * V9 + \beta_9 * V10 \\
& + \beta_{10} * V12 + \beta_{11} * V15 + \beta_{12} * V16 + \beta_{13} * V17 + \beta_{14} * V19 + \beta_{15} * V20 + \beta_{16} * V21 + \beta_{17} * V22 \\
& + \beta_{18} * V23 + \beta_{19} * V24 + \beta_{20} * V25 + \beta_{21} * V26 + \beta_{22} * V27 + \beta_{23} * V28 + \beta_{24} * V29 + \beta_{25} * V33 \\
& + \beta_{26} * V35 + \beta_{27} * V36 + \beta_{28} * V38 + \beta_{29} * V39 + \beta_{30} * V41 + \beta_{31} * V42 + \beta_{32} * V43 + \beta_{33} * V44 \\
& + \beta_{34} * V45 + \beta_{35} * V46 + \beta_{36} * V47 + \beta_{37} * V48 + \beta_{38} * V49 + \beta_{39} * V52 + \beta_{40} * V53 + \beta_{41} * V54 \\
& + \beta_{42} * V56 + \beta_{43} * V57
\end{aligned} \tag{6}
$$

Fig 17: ROC Curve − SPAM Filter



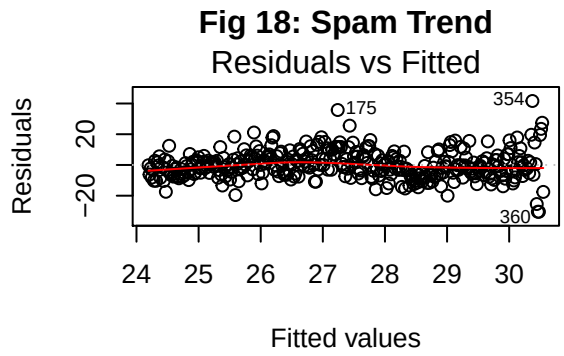
We see that the from the model utility test, the chi squared value between the entire model and this model is not very significant on the test set and thus the final model is the one shown with lesser number of terms. This is also validated by the ROC curves above that were computed on the test set. The PMSE value on the two models are also not very significant. The PMSE on the stepwise is 0.06246 and the PMSE on the other model is 0.06236.

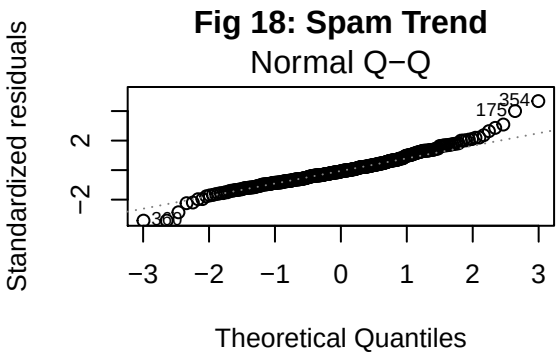
Thus we can see that certain variables are better at predicting if emails are spams or hams and in this case the variables that contributed to the distinguishability of spam and ham are mentioned in model 1 (as we saw in the last equation above), which was the goal of this static analysis.

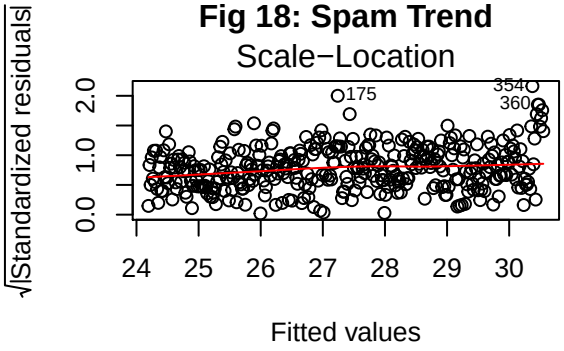
Also the hypothesis that all the variables are not good predictors to distinguish spam from ham is not true and that we could see both in the scatterplots of the predictor variables before and also from the analysis and that certain variables play a more significant role in determining spam from ham. (as we saw in the last equation above.)

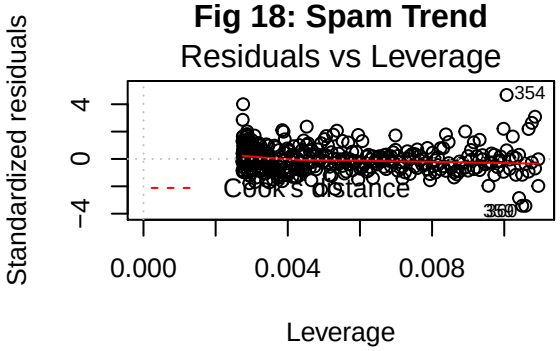
3.2 Time Series Filter Design

The diagnostic plots for the spam trend model is as follows:

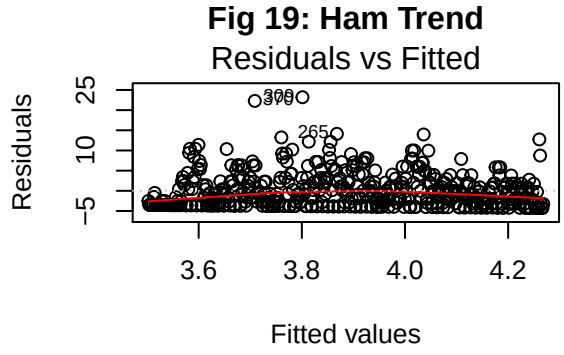


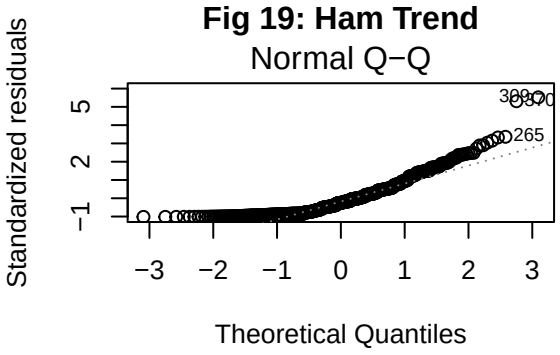


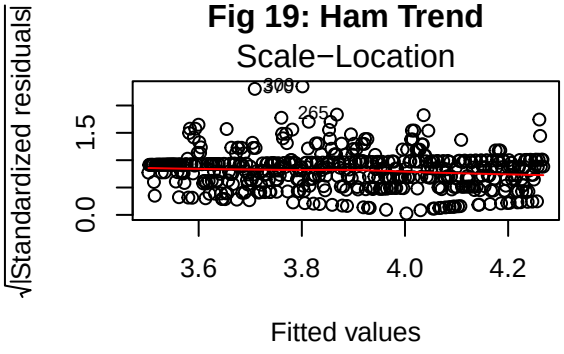


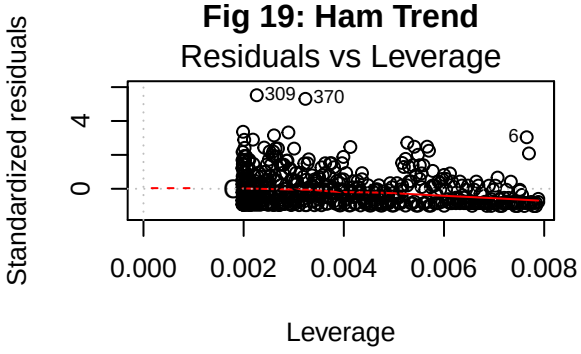


From the diagnostic plots we can see the residuals vs fitted does not show any pattern and the residuals seem to have a mean of zero. The QQ plot shows that the data does not quite follow a normal distribution towards the tails. The residuals vs leverage plots give us some insight into points like 175,354 and 353 which seem to have high standardized residuals.









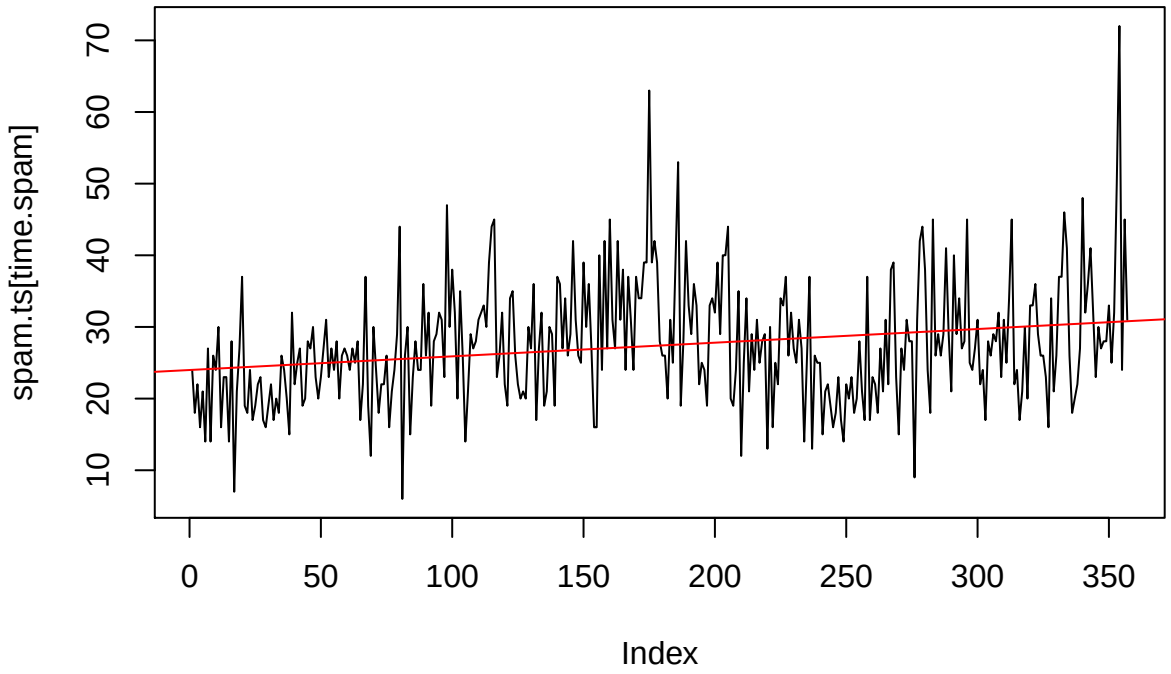
From the diagnostic plots we can see the residuals vs fitted does not show any pattern and the residuals seem to have a mean of zero. The QQ plot shows that the data does not quite follow a normal distribution towards the tails. The residuals vs leverage plots give us some insight into points like 309,370 and 6 which seem to have high standardized residuals.

Now the p-value for the spam-trend model is 1.05e-05 and it thus seems to be very significant. For the ham-trend model the p-value is 0.05339 and it is not very significant.

However, a good way to model this would be to see how this performs on the test set. In order to do that the last 7 days will be used for prediction. A linear model of the spam time series and of time index is then created to check the relationship of time (i.e index of the day) with the count of spam.

From the summary of the model with a p-value of 1.05e-05, this model seems significant. However when we plot the model and the trend of spam we see that there are still some correlations that have not been modelled.

Fig 20: Spam Time Series



From the auto correlations plot and the partial auto correlations plot we saw there was a lag for ham of 7. The lag for spam of 375 does not make much sense in this context since we only have two years worth of data. But weekly count of ham periodicity makes sense in this context.

Now the final model will be a sum of the main model that account for trends, seasonality and cyclical components and the residuals.

$$F(t) = (E(Y_t) + \epsilon_t) \tag{7}$$

Here ϵ_t takes into account if the residuals show a correlation. ϵ_t is defined as follows:

$$\epsilon_t = \left(\sum_{j=1}^{k} \theta_j \epsilon_{t-j} + w_t \right) \tag{8}$$

Fig 21: Spam trend model

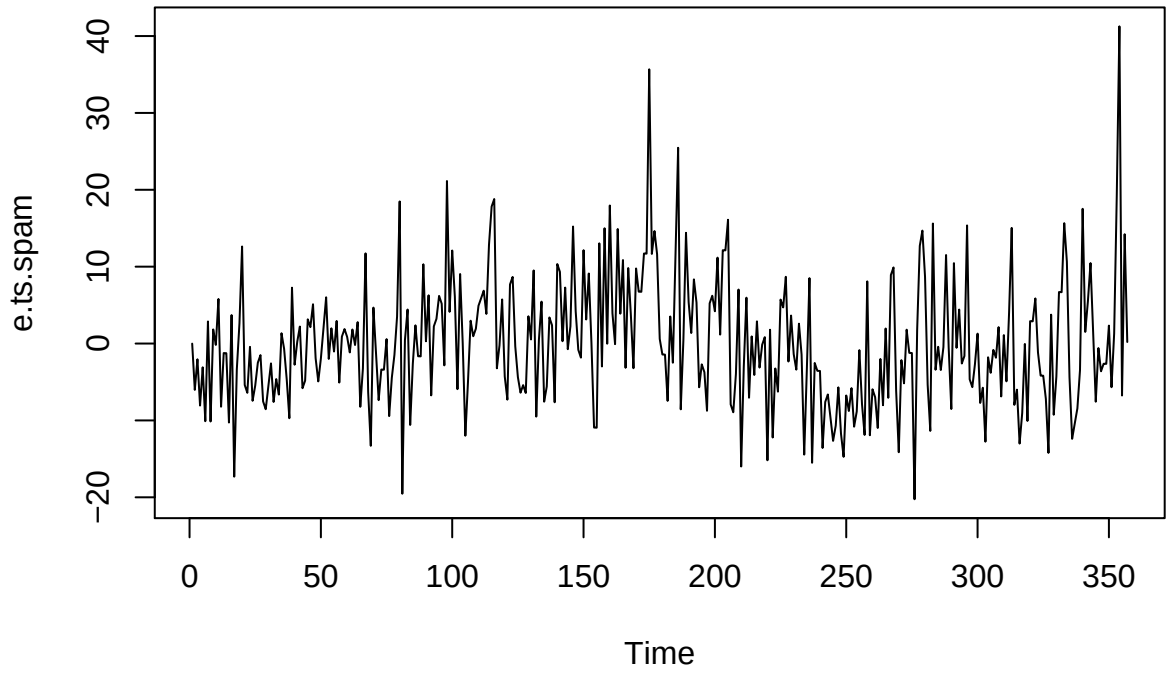
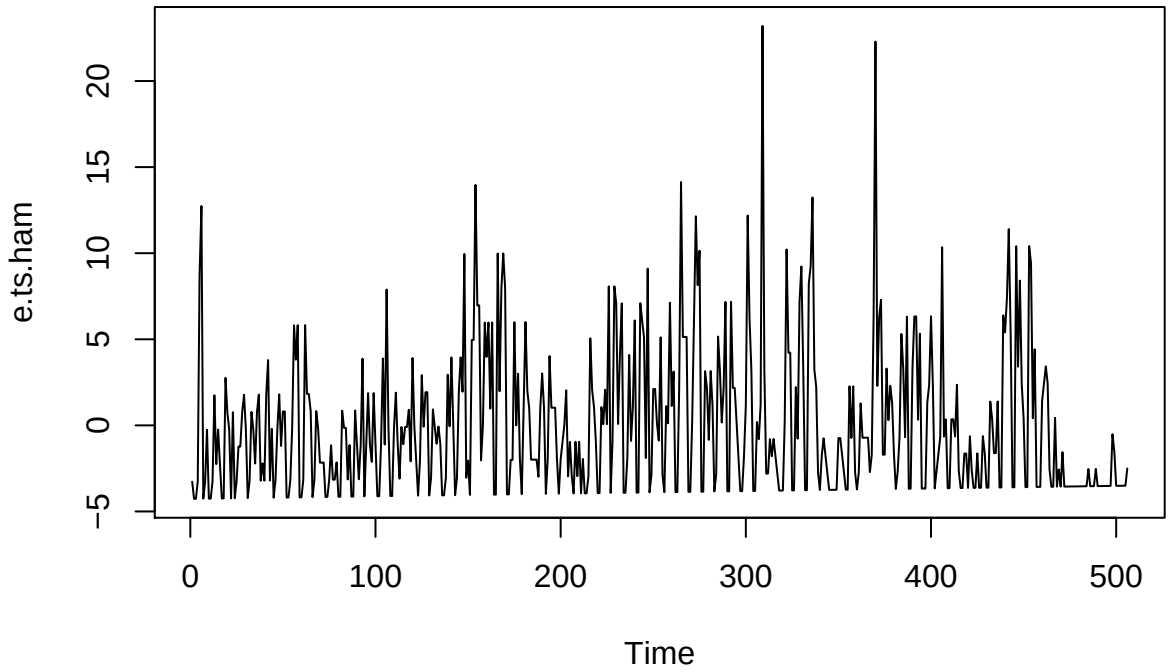


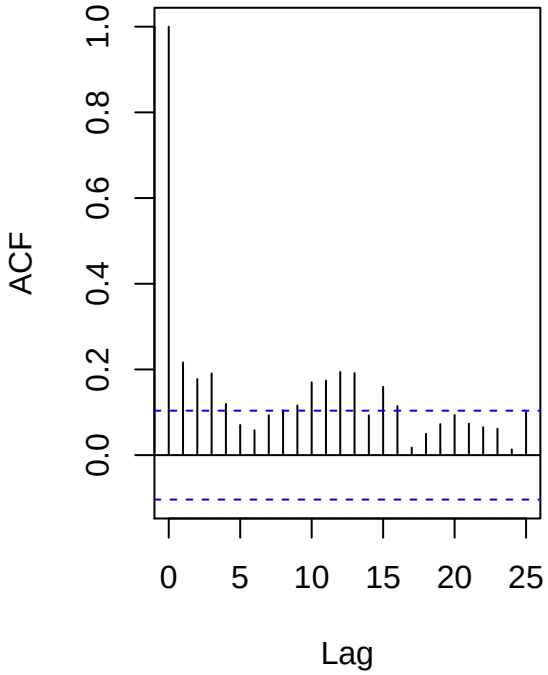
Fig 22: Ham trend model



We can also see that the mean of the residuals is approximately 0, (for spam the mean is 1.79e-16 and for ham it is 1.8e-16), and the variance is also constant showing no other significant pattern in them.

Now the Autocorrelation and the Partial Autocorrelation plots below show the details in much greater lenth.

ig 23: ACF of Residuals from spam.g 24: PACF of Residuals from spam



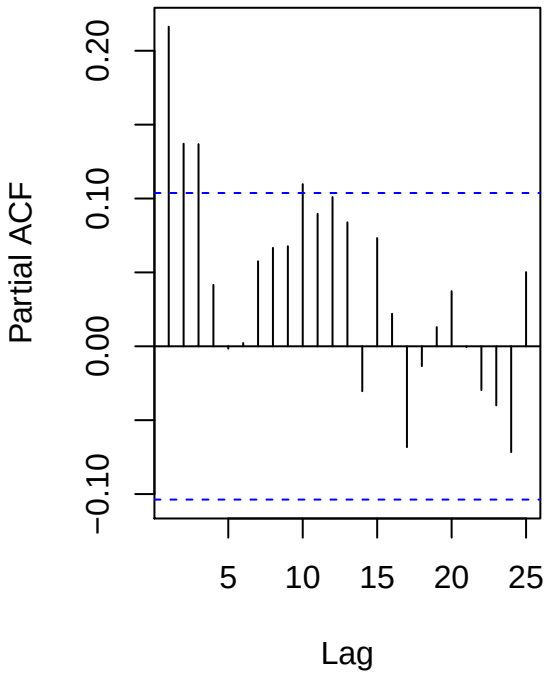
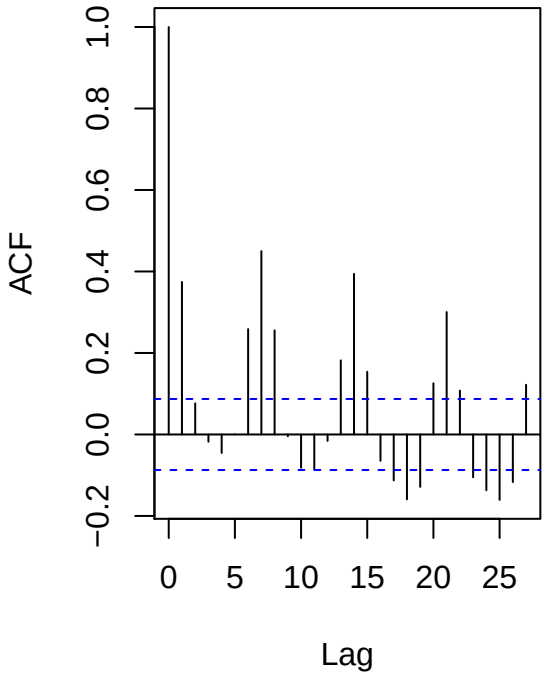
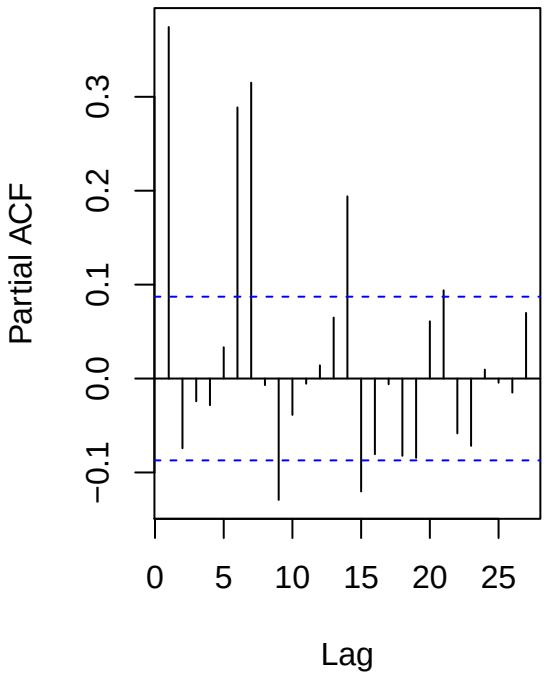


Fig 25: ACF of Residuals from ham.ig 26: PACF of Residuals from ham.





However in order to determine the optimal value for p, q and d for the ARIMA model we do the auto.arima for a stepwise model based on AIC.

The best model for spam has an AIC of 2493.081 and the p,d, and q values are 1,0 and 1 respectively. The d value is zero and that is understandable since from the initial plots of the ts we found that the constant

variance has already been accounted for in the plot of the time series data without any difference.

For ham however the best model has an AIC 2482.727 and that however has a p,d and q value of 3, 0 and 2 respectively.

We now plot the residuals in the autocorrelation and the partial autocorrelation plots below of the final models.

Fig 27: ACF of Residuals from spamig 28: PACF of Residuals from span

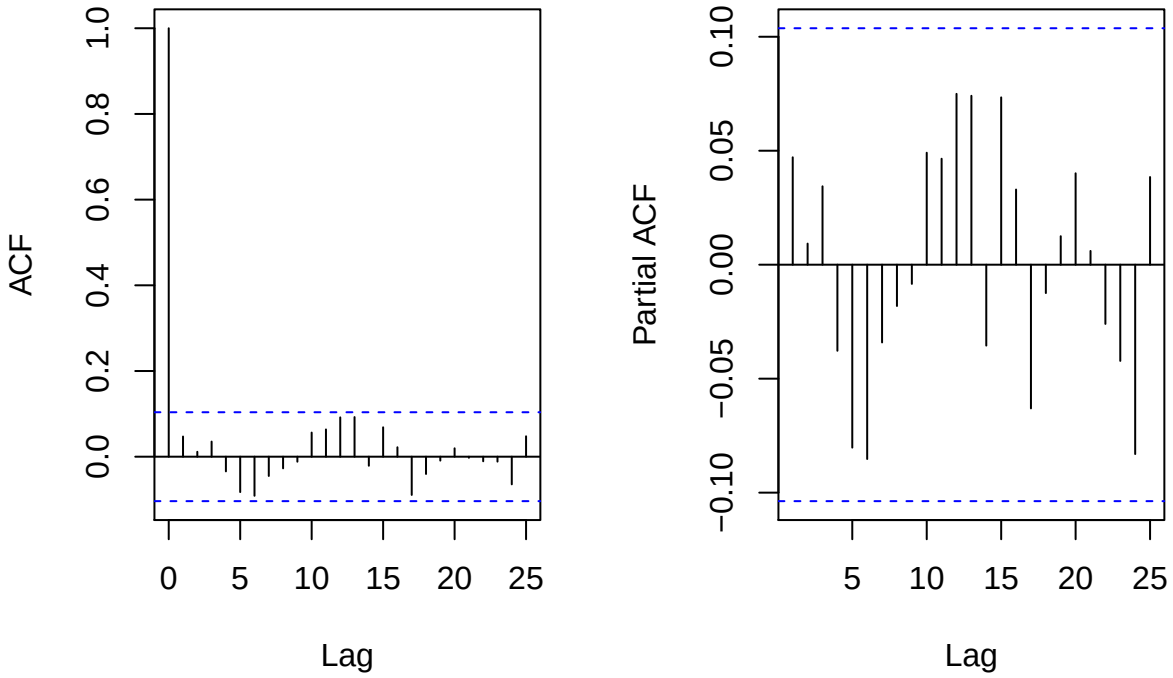
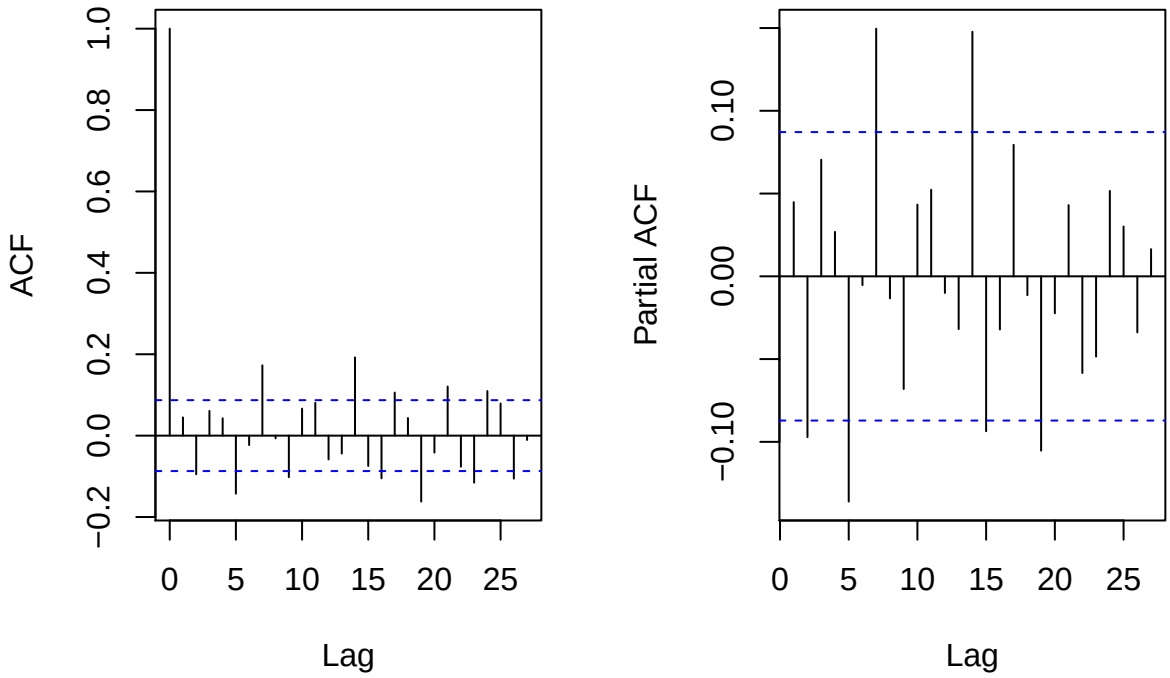
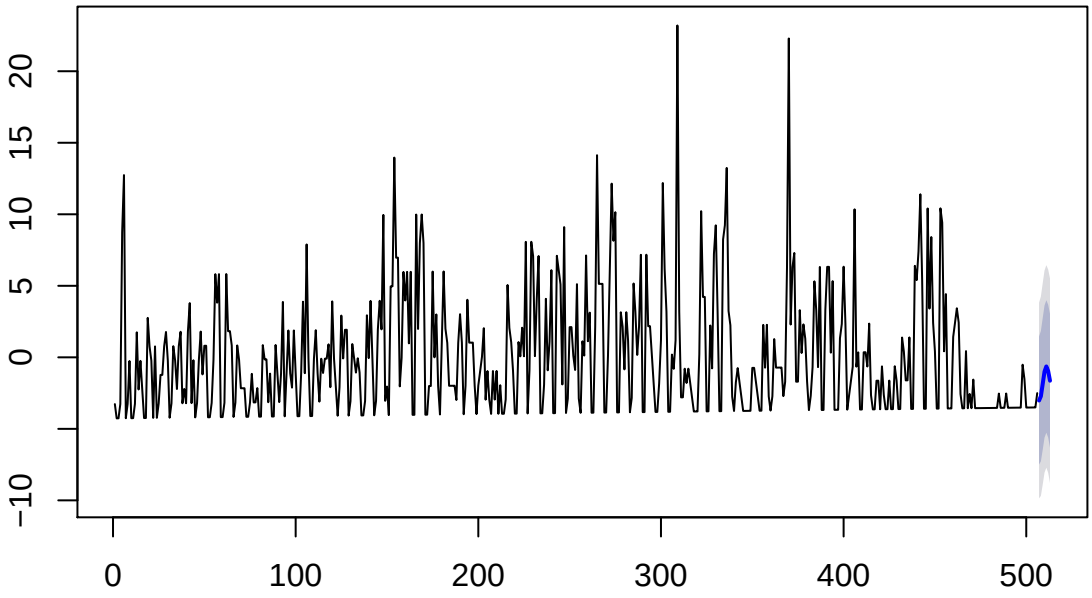


Fig 29: ACF of Residuals from ham.Fig 30: PACF of Residuals from ham



Now that the trends have been modelled we will use the forecast option in R to see how we predict for future time stamps.

Fig 31: Ham Auto Forecast Model



In order to evaluate the prediction performance a test set from ham data set with last 7 days is being created. A data frame with the next 7 days was created from the day on which the test data set ended. A prediction was then done on the next 7 days on this dataset. The root mean square error was found to be 6.446556.

The predicted values are then plotted for further demonstration of and the confidence values lines are also

added. This shows that ham has a very good temporal component.

Fig 32: Next Week Time Series

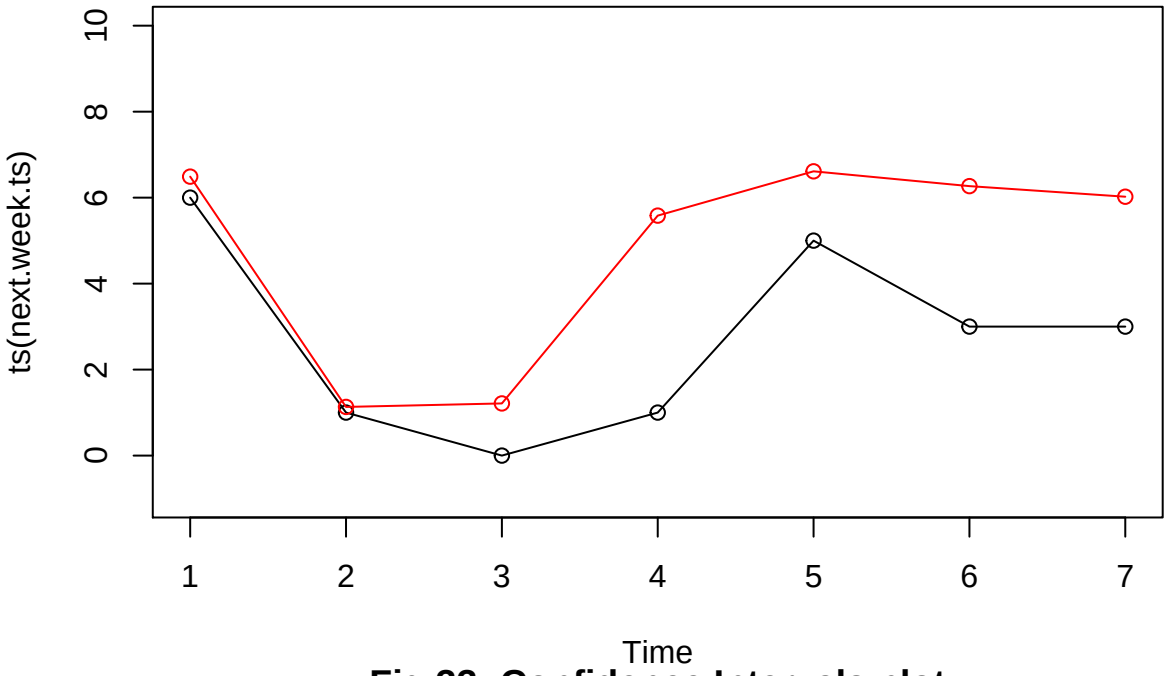
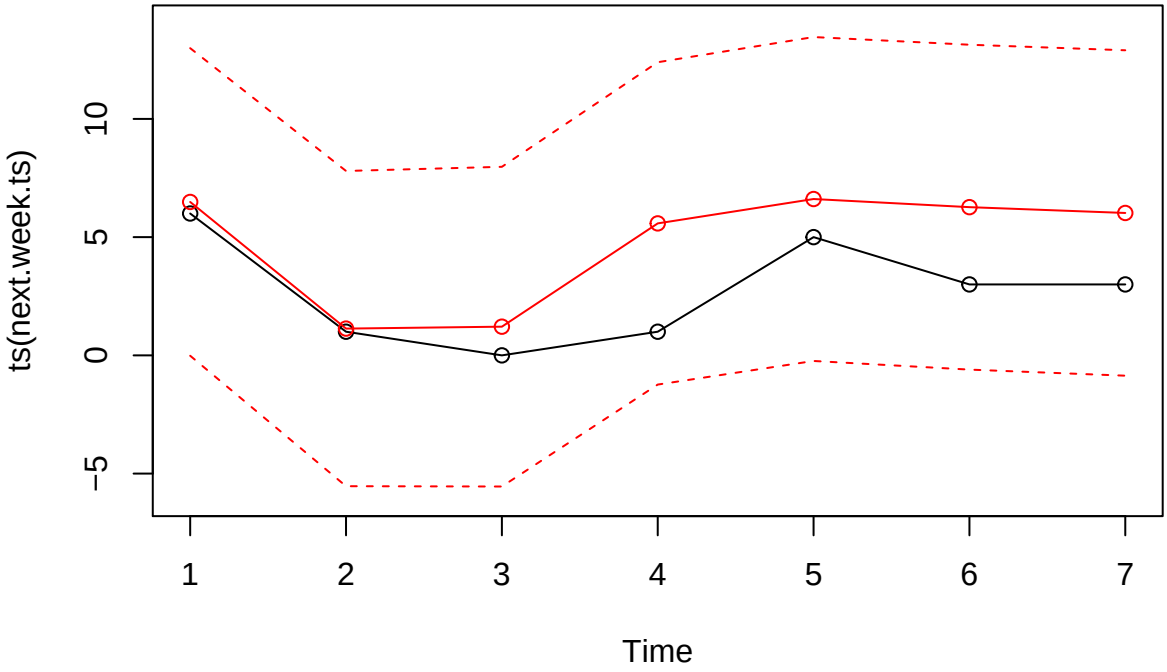
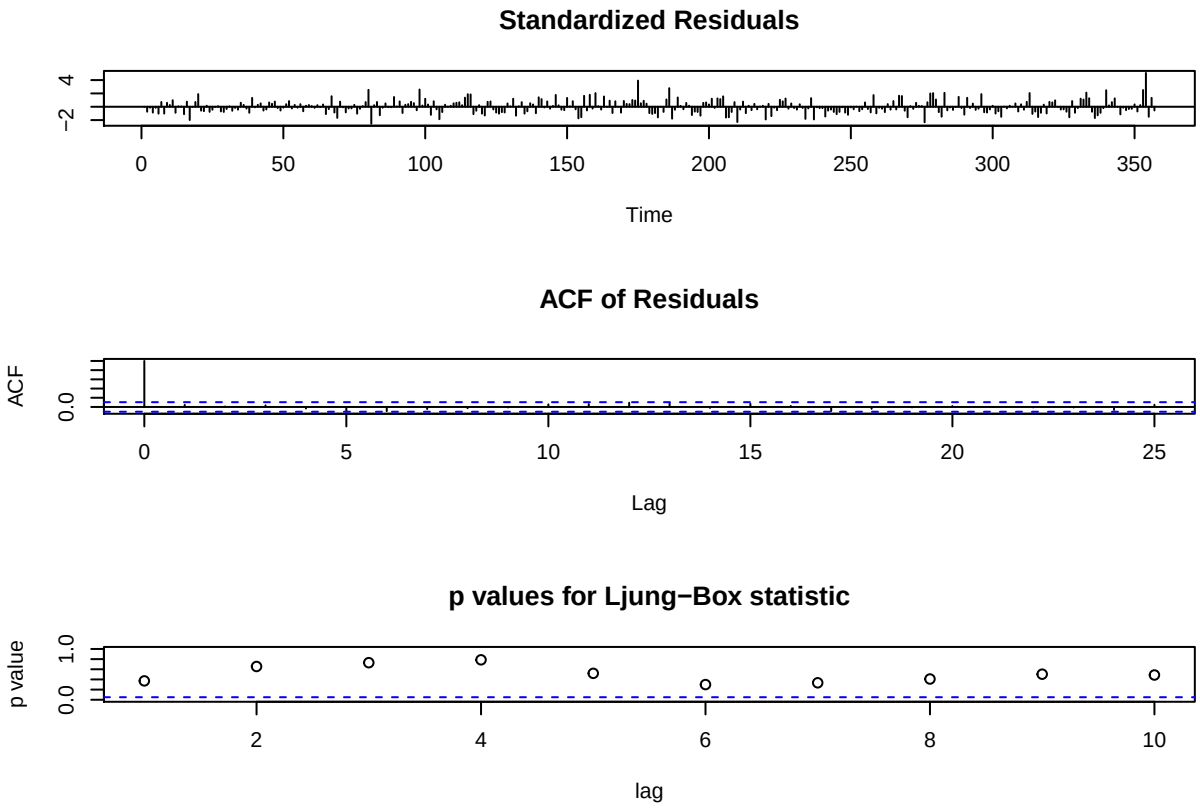


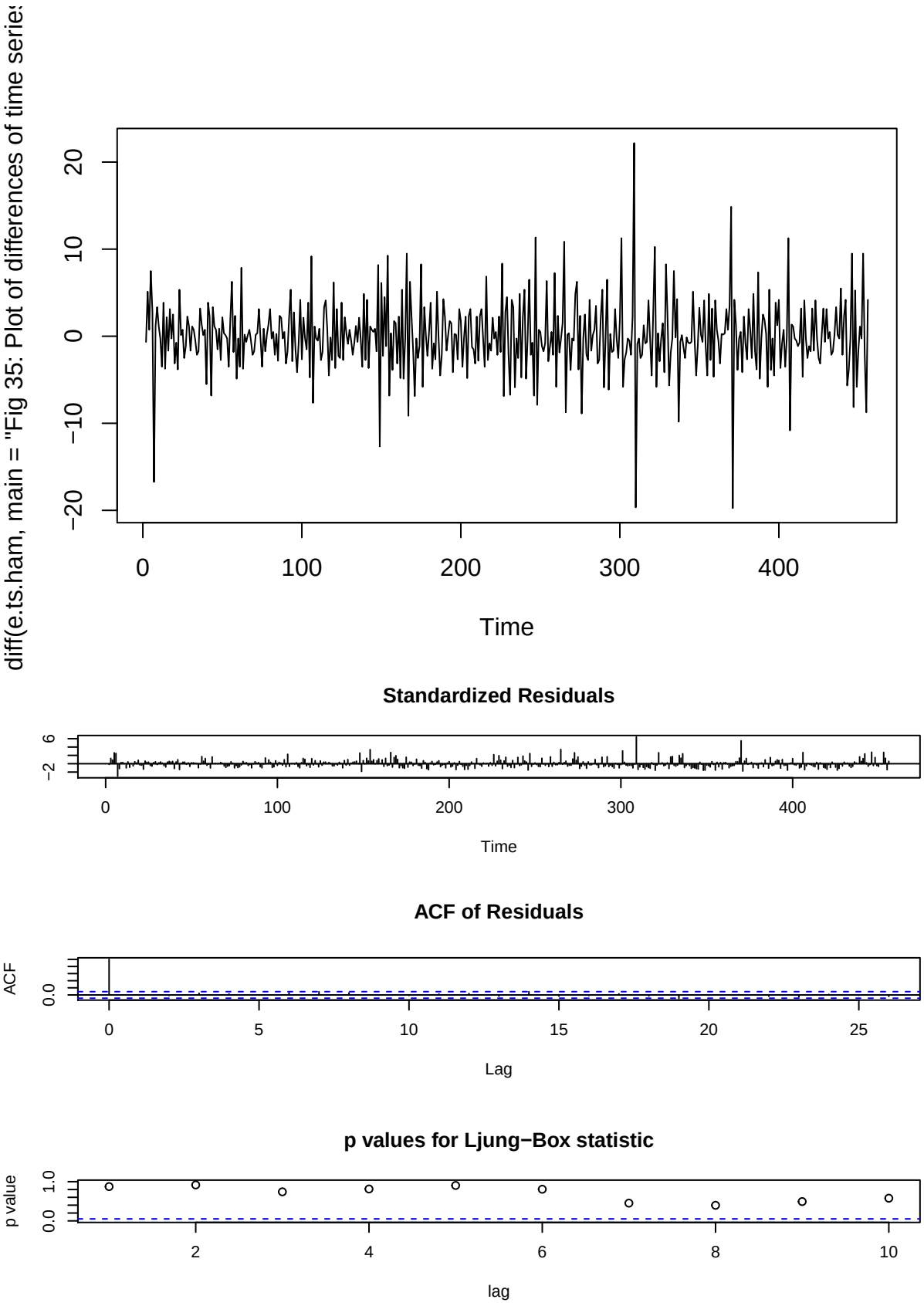
Fig 33: Confidence Intervals plot



Now that we have the model from auto arima, the diagnostic plots will give us more information about the performance of the model. Below is the diagnostic information of the final model for spam and ham.



As we can see from the standardized residuals plot, there are no significant patterns. The ACF of the residuals also do not show any noticeable lag. And none of the values are below the significance threshold in the Ljung-Box plot. This model is thus appropriate to be used as our final model for the spam senario.



For the ham we take the successive differences of each value since the model in the analysis section failed to

produce a good optimum using the auto arima. A successive calculation of the auto arima on this difference model performs much better and that can be seen in the three plots above. The standardized residual does not show any significant pattern, the acf of the residuals do not show any major lag and the Ljung-Box statistic does not have a significant p-value for any lags between 0 and 10.

Here we explored the time series data for ham and spam and found that the ham data has a significant trend component and that the null hypothesis can be rejected. The spam dataset does not have any such component and the null hypothesis cannot be rejected. The confidence levels have been plotted and we can see that the forecast on the next seven days with the ham trend model is extremely good and the confidence interval is within the 95%, to the actual prediction.

Recommendation

Here quite a few features were observed that helped in the design of static filters. Some features or variables were more prominent than the other as could be seen in equation 4. The MSE value of the stepwise regression was 0.06246 and that of the main effects model was 0.06236. Certain variables have more distinguishable features in spam and some have more distinguishable features for ham which helped in a reasonably distinction of the two models in the biplot. The base model to which this was compared was the mean(V58) (equation 2) and the model with only capital letters. (equation 3). The static filter was then developed with the help of a stepwise regression model. The confidence intervals for the model of the mean of spam dataset and the ham dataset had a Chi-Squared value of 2.2e-16 suggesting that there are definitely variables that affect the classification of spam from ham and the distinction is not arbitrary. The comparison of the main model with the model of capital letters also yielded a confidence level of 2.2e-16 suggesting that not only capital letters but other variables also affect the probability of an email being spam.

For the time series model the p-value for the spam trend model of 1.05e-05 was significant whereas the p-value for the ham trend model was not significant at 0.05339. The mean of the residuals of the spam and ham trend model were 1.79e-16 and 1.8e-16 respectively which was quite close to zero. The best spam model which had an AIC of 2493.081 had p,d, and q values of 1,0 and 1 respectively. The d value is zero and that is understandable since from the initial plots of the ts we found that the constantvariance has already been accounted for in the plot of the time series data without any difference. For ham however the best model has an AIC 2482.727 and that however has a p,d and q value of 3, 0 and 2 respectively. The final prediction done on the test set for the ham dataset did significantly well since ham had a strong seasonal component and the root mean square error was 6.446556. Plot of the confidence level shows that the prediction is well within the 95% confidence level.

References

[1] L. E. Barnes and D. E. Brown, *Project 2: Spam Classfication ,Class project in SYS 6021, 2014.*

[2] Project 2 template, Class template in SYS 4021, 2014.

[3] Mark Hopkins, Erik Reeber, George Forman, Jaap Suermondt Hewlett-Packard Labs, 1501 Page Mill Rd., Palo Alto, CA 94304

[4] Code sample provided in Class, SYS 6021, L. E. Barnes and D. E. Brown, 2014, UVa

[5] Spam Cost: http://www.huffingtonpost.com/ 2012/ 08/ 08/ cost-of-spam_n_1757726.html

[6] *Email Spam: http://en.wikipedia.org/wiki/Spamming*

Optional Appendices

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.57	0.14	-11.04	0.00
V1	-0.39	0.23	-1.68	0.09
V2	-0.15	0.07	-2.10	0.04
V3	0.11	0.11	1.03	0.30
V4	2.25	1.51	1.49	0.14
V5	0.56	0.10	5.52	0.00
V6	0.88	0.25	3.53	0.00
V7	2.28	0.33	6.85	0.00
V8	0.57	0.17	3.39	0.00
V9	0.73	0.28	2.58	0.01
V10	0.13	0.07	1.76	0.08
V11	-0.26	0.30	-0.86	0.39
V12	-0.14	0.07	-1.87	0.06
V13	-0.08	0.23	-0.35	0.73
V14	0.14	0.14	1.06	0.29
V15	1.24	0.73	1.70	0.09
V16	1.04	0.15	7.13	0.00
V17	0.96	0.23	4.26	0.00
V18	0.12	0.12	1.03	0.30
V19	0.08	0.04	2.32	0.02
V20	1.05	0.54	1.95	0.05
V21	0.24	0.05	4.61	0.00
V22	0.20	0.16	1.24	0.22
V23	2.25	0.47	4.76	0.00
V24	0.43	0.16	2.63	0.01
V25	-1.92	0.31	-6.14	0.00
V26	-1.04	0.44	-2.37	0.02
V27	-11.77	2.11	-5.57	0.00
V28	0.45	0.20	2.24	0.03
V29	-2.49	1.50	-1.66	0.10
V30	-0.33	0.31	-1.05	0.29
V31	-0.17	0.48	-0.35	0.72
V32	2.55	3.28	0.78	0.44
V33	-0.74	0.31	-2.37	0.02
V34	0.67	1.60	0.42	0.68
V35	-2.06	0.79	-2.61	0.01
V36	0.92	0.31	2.99	0.00
V37	0.05	0.18	0.27	0.79
V38	-0.60	0.42	-1.41	0.16
V39	-0.87	0.38	-2.26	0.02
V40	-0.30	0.36	-0.84	0.40
V41	-45.05	26.60	-1.69	0.09
V42	-2.69	0.84	-3.21	0.00
V43	-1.25	0.81	-1.55	0.12
V44	-1.57	0.53	-2.97	0.00
V45	-0.79	0.16	-5.09	0.00
V46	-1.46	0.27	-5.43	0.00
V47	-2.33	1.66	-1.40	0.16
V48	-4.02	1.61	-2.49	0.01
V49	-1.29	0.44	-2.92	0.00
V50	-0.19	0.25	-0.75	0.45
V51	-0.66	0.84	-0.78	0.43
V52	0.35	0.09	3.89	0.00
V53	5.34	0.71	7.55	0.00
V54	2.40	1.11	2.16	0.03
V55	0.01	0.02	0.64	0.52
V56	0.01	0.00	3.62	0.00
V57	0.00	0.00	3.75	0.00

Table 1: Table of coefficents and statistical confidence