An automatic report for the dataset : automaticstatisticianM

(A very basic version of) The Automatic Statistician

Abstract

This is a report analysing the dataset automaticstatisticianM. Three simple strategies for building linear models have been compared using 5 fold cross validation on half of the data. The strategy with the lowest cross validated prediction error has then been used to train a model on the same half of data. This model is then described, displaying the most influential components first. Model criticism techniques have then been applied to attempt to find discrepancies between the model and data.

1 Brief description of data set

To confirm that I have interpreted the data correctly a short summary of the data set follows. The target of the regression analysis is the column y. There are 12 input columns and 74 rows of data. A summary of these variables is given in table 1.

Name	Minimum	Median	Maximum
y	-0.00046	3.8e-05	0.00049
Frequency	1	12	93
Anger	0	0	4
Negative	0	2	13
Positive	0	3	15
Skepticism	0	1	10
Trust	0	0	6
Total Frequency	24	1.2e+02	5.5e+02
Total Anger	0	7	1e+02
Total Negative	1	17	2.4e+02
Total Positive	7	34	2.1e+02
Total Skepticism	4	12	1.7e+02
Total Trust	0	6	22

Table 1: Summary statistics of data

2 Summary of model construction

I have compared a number of different model construction techniques by computing cross-validated root-mean-squared-errors (RMSE). I have also expressed these errors as a proportion of variance explained (negative values indicate performance that is worse than just predicting the mean value). These figures are summarised in table 2.

Method	Cross validated RMSE	Cross validated variance explained (%)
BIC stepwise	0.000214	-18.3
LASSO	0.000224	-16.3
Full linear model	0.000228	-39.7

Table 2: Summary of model construction methods and cross validated errors

The method, BIC stepwise, has the lowest cross validated error so I have used this method to train a model on half of the data. In the rest of this report I have described this model and have attempted to falsify it using held out test data.

3 Model description

In this section I have described the model I have constructed to explain the data. A quick summary is below, followed by quantification of the model with accompanying plots of model fit and residuals.

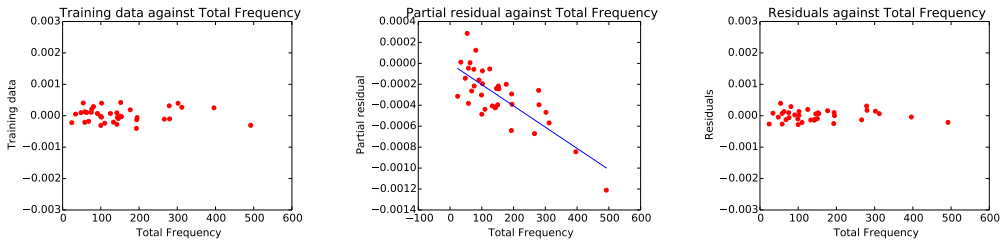
3.1 Summary

The output y:

- decreases linearly with input Total Frequency
- increases linearly with input Frequency
- increases linearly with input Total Anger
- decreases linearly with input Positive

3.2 Detailed plots

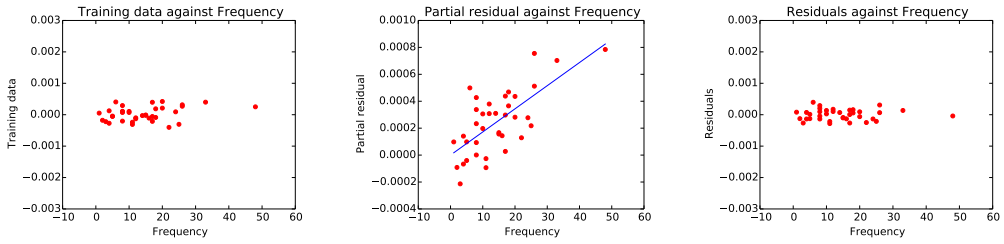
Decrease with Total Frequency The correlation between the data and the input Total Frequency is 0.01 (see figure 1a). Accounting for the rest of the model, this changes substantially to a part correlation of -0.78 (see figure 1b).



a) b) c)

Figure 1: a) Training data plotted against input Total Frequency. b) Partial residuals (data minus the rest of the model) and fit of this component. c) Residuals (data minus the full model).

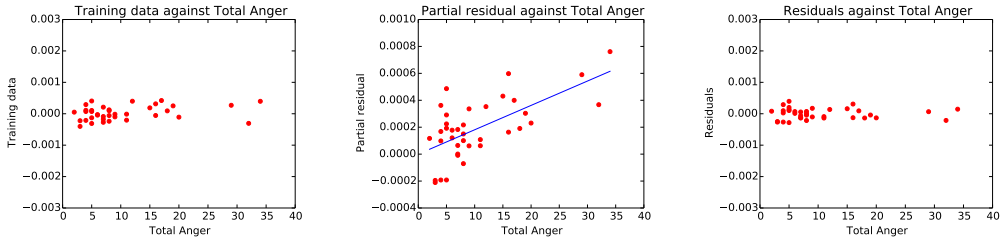
Increase with Frequency The correlation between the data and the input Frequency is 0.31 (see figure 2a). Accounting for the rest of the model, this changes substantially to a part correlation of 0.70 (see figure 2b).



a) b) c)

Figure 2: a) Training data plotted against input Frequency. b) Partial residuals (data minus the rest of the model) and fit of this component. c) Residuals (data minus the full model).

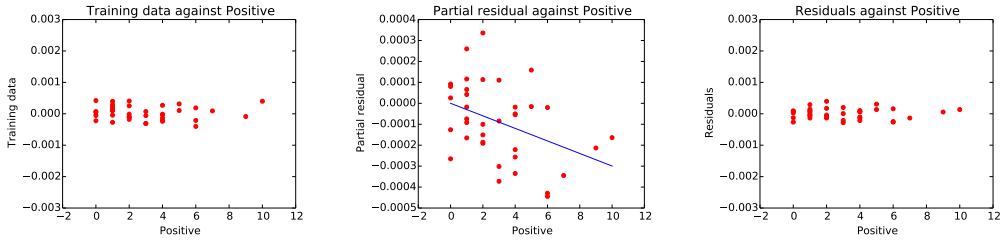
Increase with Total Anger The correlation between the data and the input Total Anger is 0.29 (see figure 3a). Accounting for the rest of the model, this changes substantially to a part correlation of 0.65 (see figure 3b).



a) b) c)

Figure 3: a) Training data plotted against input Total Anger. b) Partial residuals (data minus the rest of the model) and fit of this component. c) Residuals (data minus the full model).

Decrease with Positive The correlation between the data and the input Positive is -0.05 (see figure 4a). Accounting for the rest of the model, this changes substantially to a part correlation of -0.40 (see figure 4b).



a) b) c)

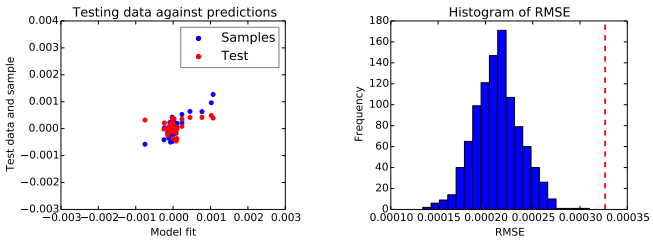
Figure 4: a) Training data plotted against input Positive. b) Partial residuals (data minus the rest of the model) and fit of this component. c) Residuals (data minus the full model).

4 Model criticism

In this section I have attempted to falsify the model that I have presented above to understand what aspects of the data it is not capturing well. This has been achieved by comparing the model with data I held out from the model fitting stage. In particular, I have searched for correlations and dependencies within the data that are unexpectedly large or small. I have also compared the distribution of the

residuals with that assumed by the model (a normal distribution). There are other tests I could perform but I will hopefully notice any particularly obvious failings of the model. Below are a list of the discrepancies that I have found with the most surprising first. Note however that some discrepancies may be due to chance; on average 10% of the listed discrepancies will be due to chance.

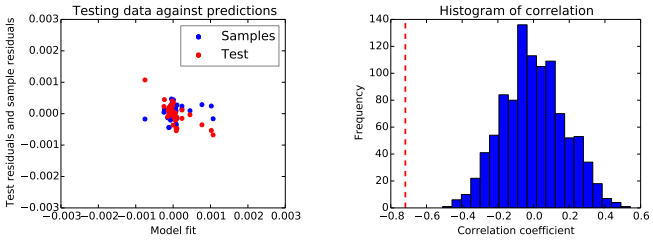
High test set error There is an unexpectedly high RMSE on the test data (see figure 5a). The RMSE has a slightly larger value of 0.00033 compared to its median value under the proposed model of 0.00021 (see figure 5b).



a) b)

Figure 5: a) Test set and model samples. b) Histogram of RMSE evaluated on random samples from the model (values that would be expected if the data was generated by the model) and value on test data (dashed line).

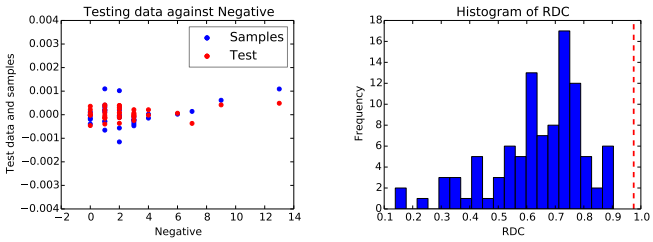
Low correlation between residuals and model fit There is an unexpectedly low correlation between the residuals and model fit (see figure 6a). The correlation has a substantially smaller value of -0.72 compared to its median value under the proposed model of -0.00 (see figure 6b).



a) b)

Figure 6: a) Test set and model sample residuals. b) Histogram of correlation evaluated on random samples from the model (values that would be expected if the data was generated by the model) and value on test data (dashed line).

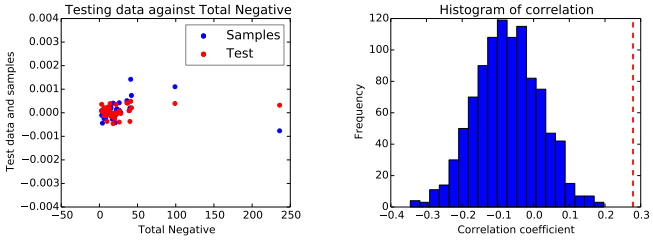
High dependence between data and Negative There is an unexpectedly high dependence between the data and input Negative (see figure 7a). The dependence as measured by the randomised dependency coefficient (RDC) has a substantially larger value of 0.97 compared to its median value under the proposed model of 0.67 (see figure 7b).



a) b)

Figure 7: a) Test set and model samples. b) Histogram of RDC evaluated on random samples from the model (values that would be expected if the data was generated by the model) and value on test data (dashed line).

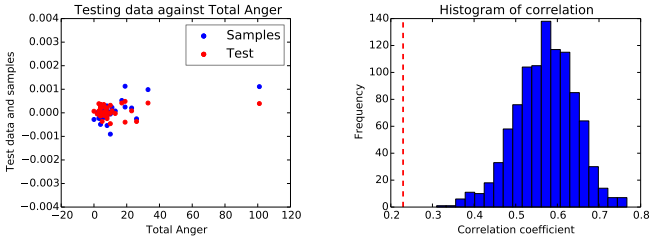
High correlation between data and Total Negative There is an unexpectedly high correlation between the data and input Total Negative (see figure 8a). The correlation has a substantially larger value of 0.28 compared to its median value under the proposed model of -0.07 (see figure 8b).



a) b)

Figure 8: a) Test set and model samples. b) Histogram of correlation coefficient evaluated on random samples from the model (values that would be expected if the data was generated by the model) and value on test data (dashed line).

Low correlation between data and Total Anger There is an unexpectedly low correlation between the data and input Total Anger (see figure 9a). The correlation has a substantially smaller value of 0.23 compared to its median value under the proposed model of 0.57 (see figure 9b).

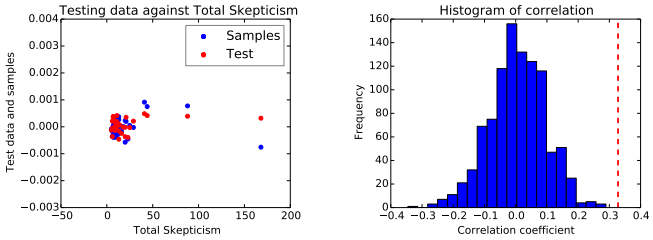


a) b)

Figure 9: a) Test set and model samples. b) Histogram of correlation evaluated on random samples from the model (values that would be expected if the data was generated by the model) and value on test data (dashed line).

High correlation between data and Total Skepticism There is an unexpectedly high correlation between the data and input Total Skepticism (see figure 10a). The correlation has a substantially

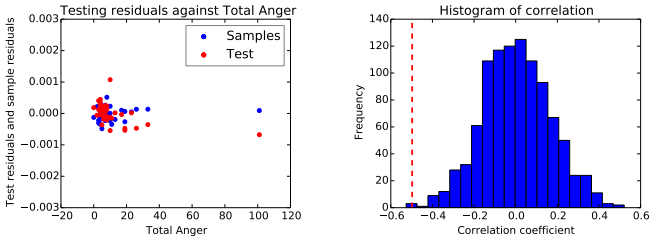
larger value of 0.33 compared to its median value under the proposed model of 0.00 (see figure 10b).



a) b)

Figure 10: a) Test set and model samples. b) Histogram of correlation coefficient evaluated on random samples from the model (values that would be expected if the data was generated by the model) and value on test data (dashed line).

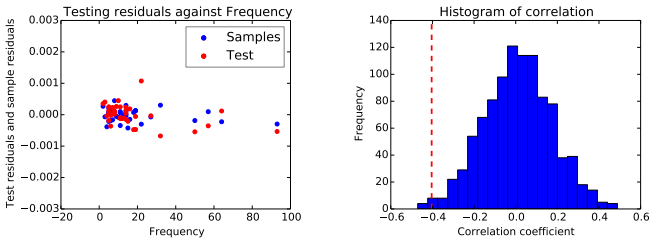
Low correlation between residuals and Total Anger There is an unexpectedly low correlation between the residuals and input Total Anger (see figure 11a). The correlation has a substantially smaller value of -0.50 compared to its median value under the proposed model of 0.00 (see figure 11b).



a) b)

Figure 11: a) Test set and model sample residuals. b) Histogram of correlation evaluated on random samples from the model (values that would be expected if the data was generated by the model) and value on test data (dashed line).

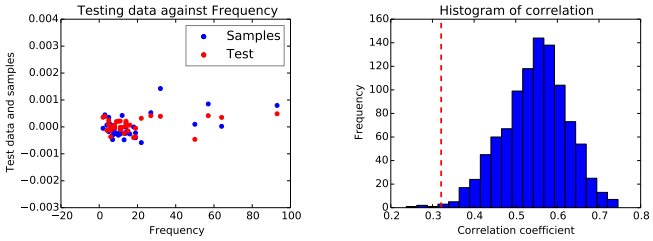
Low correlation between residuals and Frequency There is an unexpectedly low correlation between the residuals and input Frequency (see figure 12a). The correlation has a substantially smaller value of -0.40 compared to its median value under the proposed model of 0.01 (see figure 12b).



a) b)

Figure 12: a) Test set and model sample residuals. b) Histogram of correlation evaluated on random samples from the model (values that would be expected if the data was generated by the model) and value on test data (dashed line).

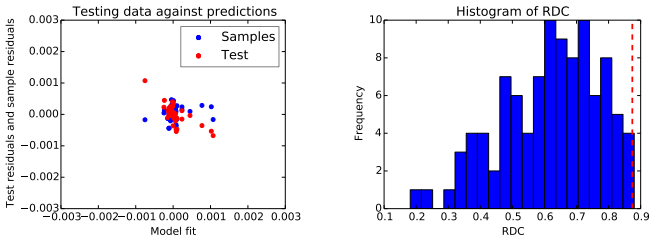
Low correlation between data and Frequency There is an unexpectedly low correlation between the data and input Frequency (see figure 13a). The correlation has a substantially smaller value of 0.32 compared to its median value under the proposed model of 0.55 (see figure 13b).



a) b)

Figure 13: a) Test set and model samples. b) Histogram of correlation evaluated on random samples from the model (values that would be expected if the data was generated by the model) and value on test data (dashed line).

High dependence between residuals and model fit There is an unexpectedly high dependence between the residuals and model fit (see figure 14a). The dependence as measured by the randomised dependency coefficient (RDC) has a substantially larger value of 0.87 compared to its median value under the proposed model of 0.64 (see figure 14b).



a) b)

Figure 14: a) Test set and model sample residuals. b) Histogram of RDC evaluated on random samples from the model (values that would be expected if the data was generated by the model) and value on test data (dashed line).