An automatic report for the dataset : automaticstatisticianV

(A very basic version of) The Automatic Statistician

Abstract

This is a report analysing the dataset automaticstatisticianV. 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	-0.00024	-2.8e-05	0.0002
Frequency	1	17	1.1e+02
Anger	0	1	43
Negative	0	2	10
Positive	0	5	35
Skepticism	0	1	7
Trust	0	0	6
Total Frequency	24	1.2e+02	5.3e+02
Total Anger	0	6	60
Total Negative	0	16	2.4e+02
Total Positive	7	32	2.2e+02
Total Skepticism	3	11	1.7e+02
Total Trust	1	6	20

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 (%)
LASSO	9.89e-05	-74.5
BIC stepwise	0.000113	-211.1
Full linear model	0.000136	-518.4

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

The method, LASSO, 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:

- increases linearly with input Frequency
- decreases linearly with input Total Negative
- decreases linearly with input Trust
- increases linearly with input Total Frequency

3.2 Detailed plots

Increase with Frequency The correlation between the data and the input Frequency is 0.27 (see figure 1a). Accounting for the rest of the model, this changes slightly to a part correlation of 0.36 (see figure 1b).

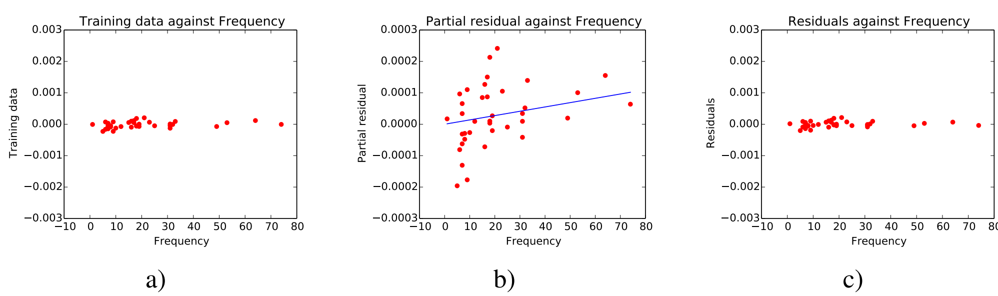
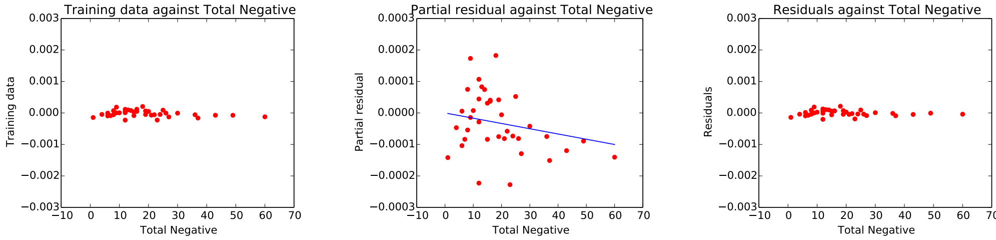


Figure 1: a) Training data plotted against input Frequency. b) Partial residuals (data minus the rest of the model) and fit of this component. c) Residuals (data minus the full model).

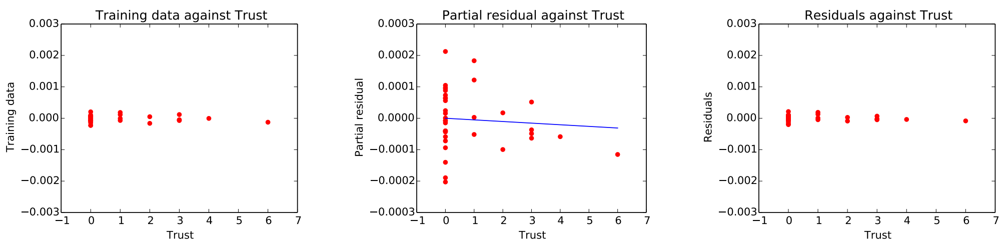
Decrease with Total Negative The correlation between the data and the input Total Negative is -0.26 (see figure 2a). Accounting for the rest of the model, this changes slightly to a part correlation of -0.34 (see figure 2b).



a) b) c)

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

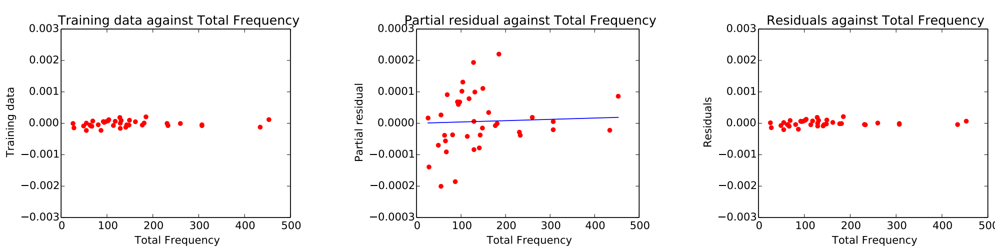
Decrease with Trust The correlation between the data and the input Trust is -0.09 (see figure 3a). Accounting for the rest of the model, this changes moderately to a part correlation of -0.21 (see figure 3b).



a) b) c)

Figure 3: a) Training data plotted against input Trust. 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 Frequency The correlation between the data and the input Total Frequency is 0.11 (see figure 4a). Accounting for the rest of the model, this changes slightly to a part correlation of 0.17 (see figure 4b).



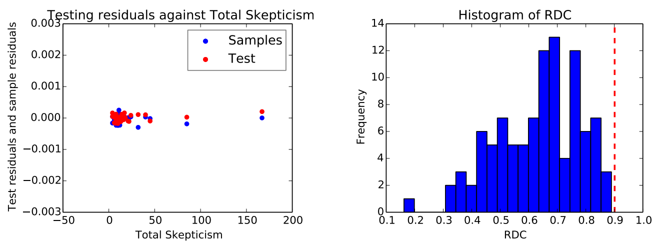
a) b) c)

Figure 4: 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).

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 dependence between residuals and Total Skepticism There is an unexpectedly high dependence between the residuals and input Total Skepticism (see figure 5a). The dependence as measured by the randomised dependency coefficient (RDC) has a substantially larger value of 0.90 compared to its median value under the proposed model of 0.65 (see figure 5b).



a) b)

Figure 5: 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).