An automatic report for the dataset : automaticstatisticianKD

(A very basic version of) The Automatic Statistician

Abstract

This is a report analysing the dataset automaticstatisticianKD. 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	-0.00012	0.00011	0.00044
Frequency	0	2.5	1.2e+02
Anger	0	0	6
Negative	0	0	8
Positive	0	1	6
Skepticism	0	0	6
Trust	0	0	3
Total Frequency	26	1.3e+02	5.8e+02
Total Anger	0	7.5	99
Total Negative	1	18	2.4e+02
Total Positive	5	37	2.2e+02
Total Skepticism	5	13	1.7e+02
Total Trust	1	7	23

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	0.000146	1.3
Full linear model	0.000228	-304.6
BIC stepwise	0.000277	-594.1

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.

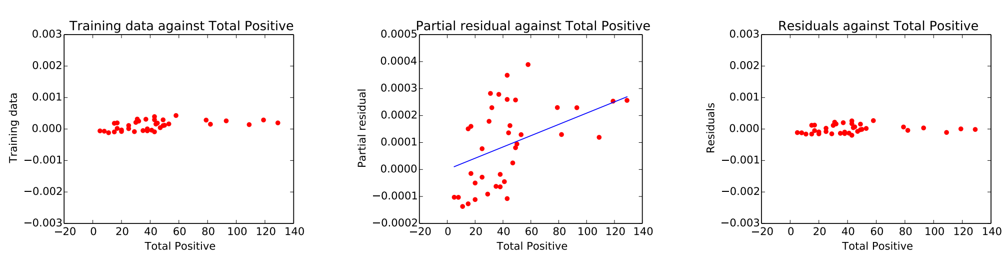
3.1 Summary

The output y:

- increases linearly with input Total Positive
- decreases linearly with input Negative
- increases linearly with input Frequency
- increases linearly with input Positive
- decreases linearly with input Total Skepticism

3.2 Detailed plots

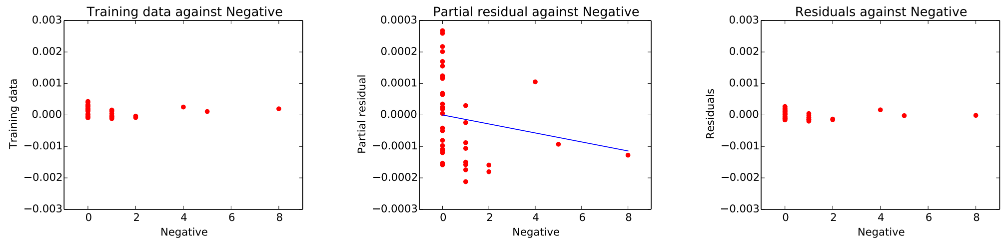
Increase with Total Positive The correlation between the data and the input Total Positive is 0.45 (see figure 1a). Accounting for the rest of the model, this changes slightly to a part correlation of 0.51 (see figure 1b).



a) b) c)

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

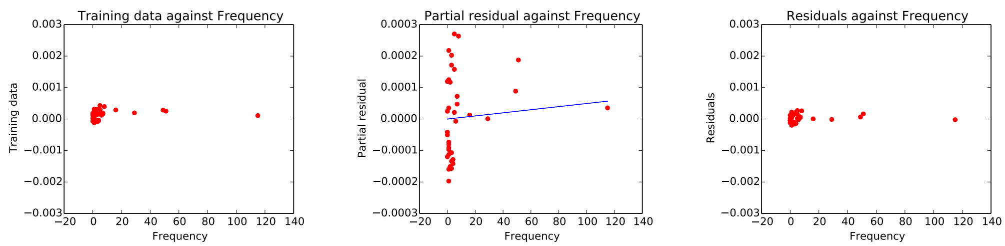
Decrease with Negative The correlation between the data and the input Negative is -0.06 (see figure 2a). Accounting for the rest of the model, this changes moderately to a part correlation of -0.28 (see figure 2b).



a) b) c)

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

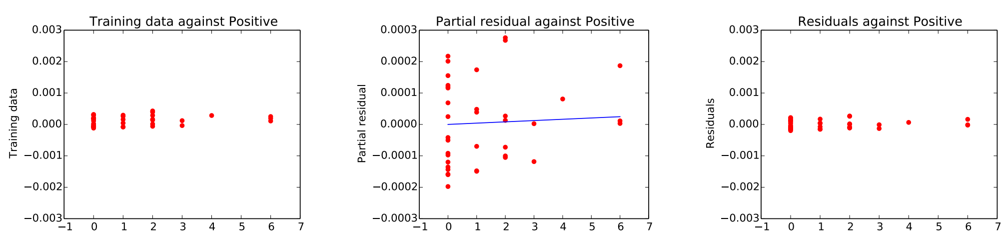
Increase with Frequency The correlation between the data and the input Frequency is 0.18 (see figure 3a). This correlation does not change when accounting for the rest of the model (see figure 3b).



a) b) c)

Figure 3: 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).

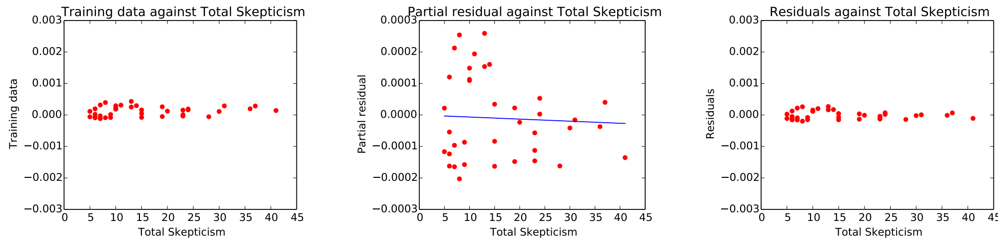
Increase with Positive The correlation between the data and the input Positive is 0.24 (see figure 4a). Accounting for the rest of the model, this changes slightly to a part correlation of 0.17 (see figure 4b).



a) b) c)

Figure 4: a) Training data plotted against input Positive. 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 Skepticism The correlation between the data and the input Total Skepticism is 0.18 (see figure 5a). Accounting for the rest of the model, this changes substantially to a part correlation of -0.16 (see figure 5b).



a) b) c)

Figure 5: a) Training data plotted against input Total Skepticism. 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.

No discrepancies were found with a false discovery rate of 10%.