An automatic report for the dataset : stovesmoke

(A very basic version of) The Automatic Statistician

Abstract

This is a report analysing the dataset stovesmoke. 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 Totaldust. There are 6 input columns and 117 rows of data. A summary of these variables is given in table 1.

Name	Minimum	Median	Maximum
Totaldust	0.4	1.6	51
fuelCV	1.4e+04	1.6e+04	3.1e+04
Inputfuel	0.7	1.1	2.5
CO	0.05	0.16	0.52
OutputkW	3.4	5.8	17
Efficiency	53	77	85
Dust	23	90	1e+03

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. These figures are summarised in table 2.

Method	Cross validated RMSE	Cross validated variance explained (%)
Full linear model	0.991	88.0
BIC stepwise	0.997	88.0
LASSO	1.04	87.5

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

The method, Full linear model, 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 Totaldust:

- increases linearly with input fuelCV
- increases linearly with input Dust
- decreases linearly with input Efficiency
- increases linearly with input Inputfuel
- increases linearly with input OutputkW
- increases linearly with input CO

3.2 Detailed plots

Increase with fuelCV The correlation between the data and the input fuelCV is 0.94 (see figure 1a). Accounting for the rest of the model, this changes slightly to a part correlation of 0.99 (see figure 1b).

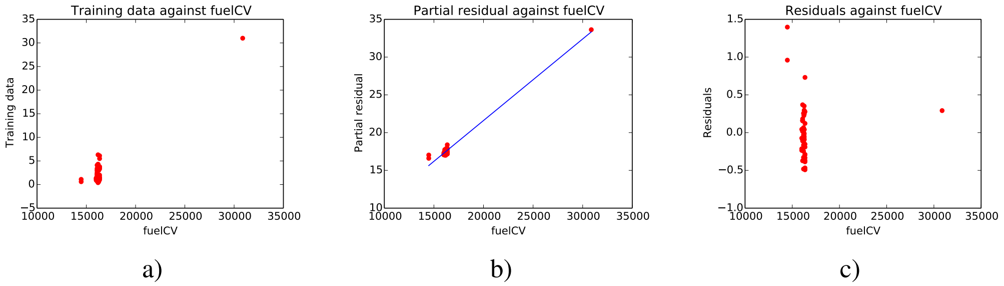


Figure 1: a) Training data plotted against input fuelCV. b) Partial residuals (data minus the rest of the model) and fit of this component. c) Residuals (data minus the full model).

Increase with Dust The correlation between the data and the input Dust is 0.94 (see figure 2a). Accounting for the rest of the model, this changes slightly to a part correlation of 0.98 (see figure 2b).

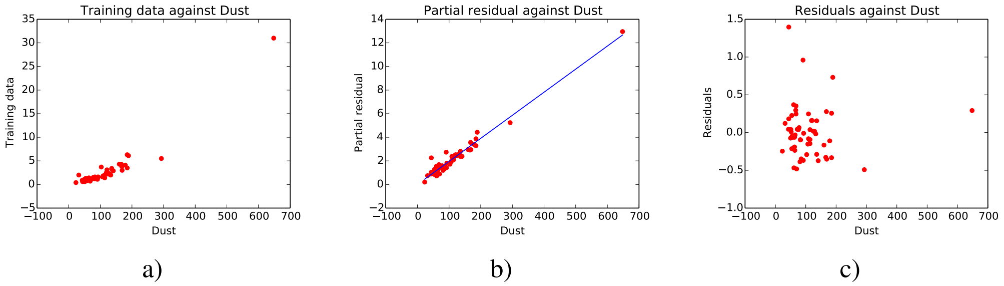
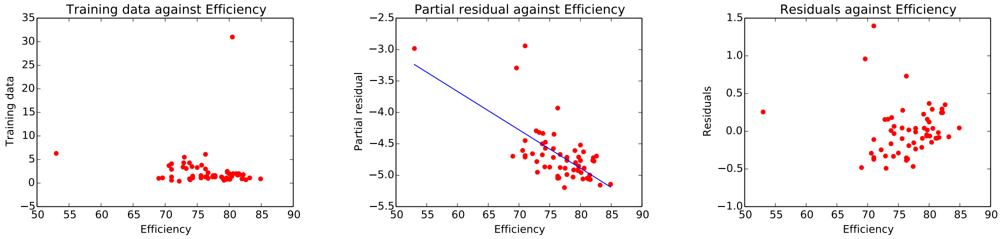


Figure 2: a) Training data plotted against input Dust . b) Partial residuals (data minus the rest of the model) and fit of this component. c) Residuals (data minus the full model).

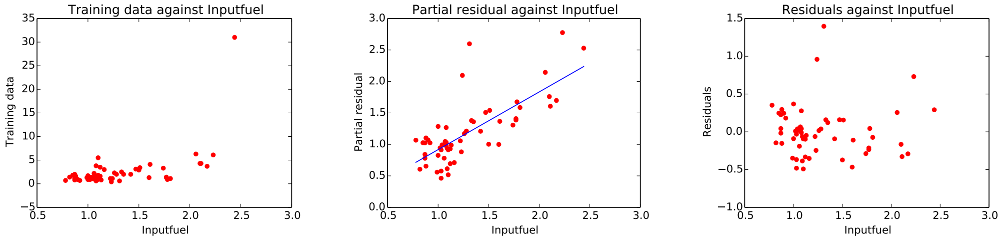
Decrease with Efficiency The correlation between the data and the input Efficiency is -0.07 (see figure 3a). Accounting for the rest of the model, this changes substantially to a part correlation of -0.67 (see figure 3b).



a) b) c)

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

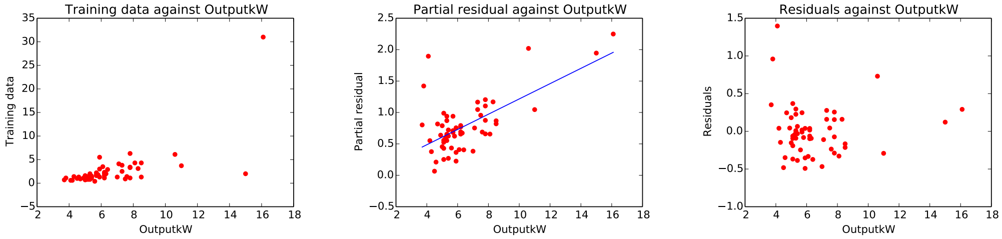
Increase with Inputfuel The correlation between the data and the input Inputfuel is 0.55 (see figure 4a). Accounting for the rest of the model, this changes moderately to a part correlation of 0.74 (see figure 4b).



a) b) c)

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

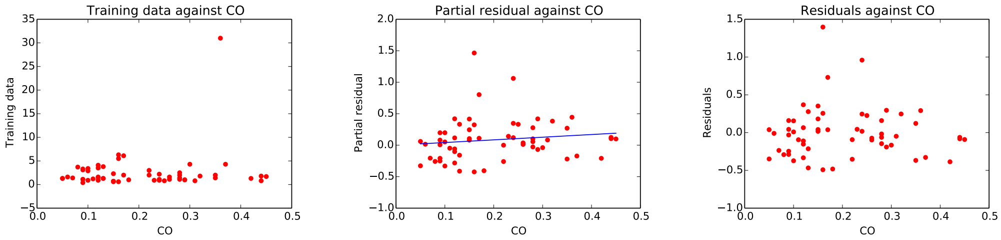
Increase with OutputkW The correlation between the data and the input OutputkW is 0.68 (see figure 5a). Accounting for the rest of the model, this changes slightly to a part correlation of 0.64 (see figure 5b).



a) b) c)

Figure 5: a) Training data plotted against input OutputkW. b) Partial residuals (data minus the rest of the model) and fit of this component. c) Residuals (data minus the full model).

Increase with CO The correlation between the data and the input CO is 0.15 (see figure 6a). Accounting for the rest of the model, this changes slightly to a part correlation of 0.14 (see figure 6b).



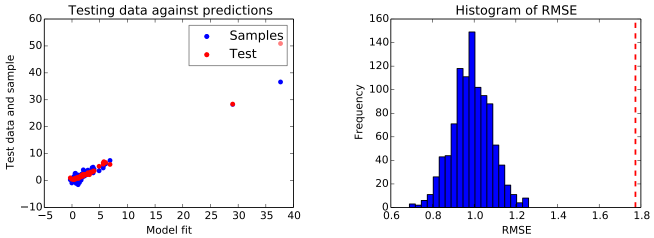
a) b) c)

Figure 6: a) Training data plotted against input CO. 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.

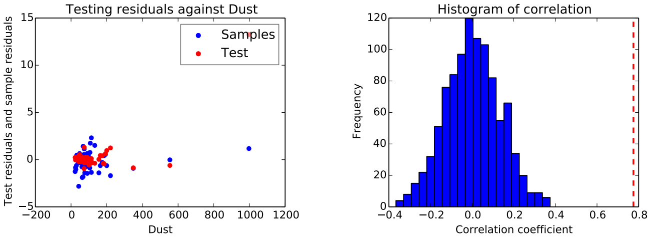
High test set error There is an unexpectedly high RMSE on the test data (see figure 7a). The RMSE has a substantially larger value of 1.8 compared to its median value under the proposed model of 0.99 (see figure 7b).



a) b)

Figure 7: 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).

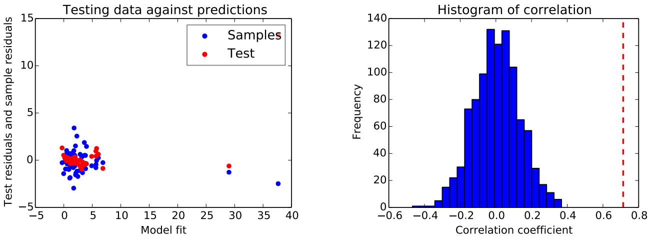
High correlation between residuals and Dust There is an unexpectedly high correlation between the residuals and input Dust (see figure 8a). The correlation has a substantially larger value of 0.78 compared to its median value under the proposed model of 0.00 (see figure 8b).



a) b)

Figure 8: a) Test set and model sample residuals. 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).

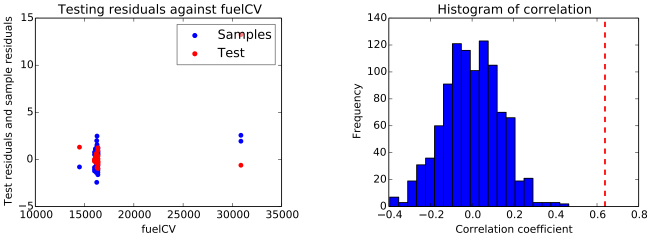
High correlation between residuals and model fit There is an unexpectedly high correlation between the residuals and model fit (see figure 9a). The correlation has a substantially larger value of 0.71 compared to its median value under the proposed model of 0.01 (see figure 9b).



a) b)

Figure 9: a) Test set and model sample residuals. 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).

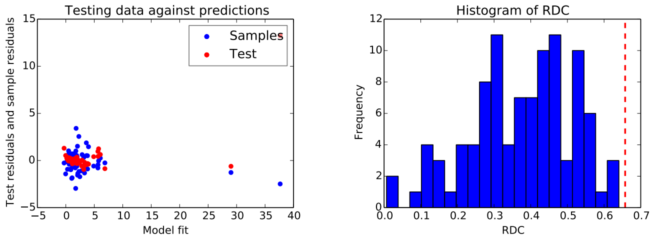
High correlation between residuals and fuelCV There is an unexpectedly high correlation between the residuals and input fuelCV (see figure 10a). The correlation has a substantially larger value of 0.64 compared to its median value under the proposed model of -0.00 (see figure 10b).



a) b)

Figure 10: a) Test set and model sample residuals. 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).

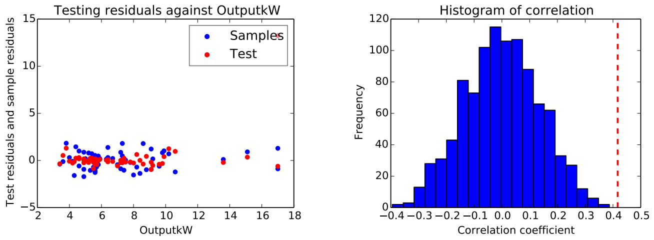
High dependence between residuals and model fit There is an unexpectedly high dependence between the residuals and model fit (see figure 11a). The dependence as measured by the randomised dependency coefficient (RDC) has a substantially larger value of 0.66 compared to its median value under the proposed model of 0.40 (see figure 11b).



a) b)

Figure 11: 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).

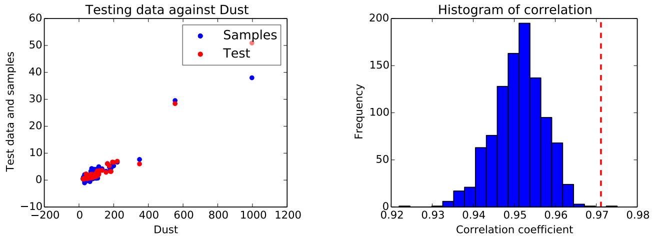
High correlation between residuals and OutputkW There is an unexpectedly high correlation between the residuals and input OutputkW (see figure 12a). The correlation has a substantially larger value of 0.42 compared to its median value under the proposed model of -0.00 (see figure 12b).



a) b)

Figure 12: a) Test set and model sample residuals. 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).

High correlation between data and Dust There is an unexpectedly high correlation between the data and input Dust (see figure 13a). The correlation has a slightly larger value of 0.97 compared to its median value under the proposed model of 0.95 (see figure 13b).



a) b)

Figure 13: 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).