An automatic report for the dataset : affairs

(A very basic version of) The Automatic Statistician

Abstract

This is a report analysing the dataset affairs. 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 affairs. There are 6 input columns and 601 rows of data. A summary of these variables is given in table 1.

Name	Minimum	Median	Maximum
affairs	0	0	12
age	18	32	57
yearsmarried	0.12	7	15
religiousness	1	3	5
education	9	16	20
occupation	1	5	7
rating	1	4	5

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	3.22	7.0
BIC stepwise	3.31	2.1
LASSO	3.33	1.2

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.

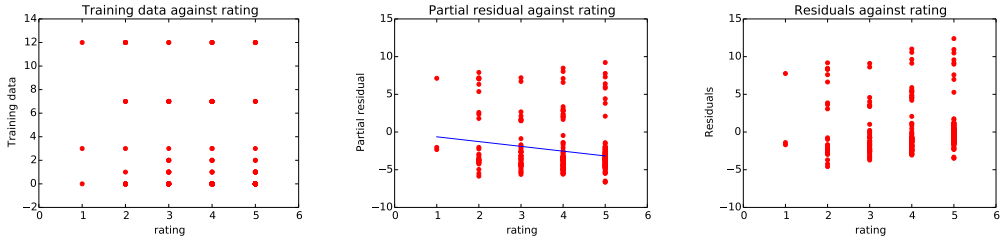
3.1 Summary

The output affairs:

- decreases linearly with input rating
- decreases linearly with input religiousness
- increases linearly with input yearsmarried
- decreases linearly with input age
- increases linearly with input occupation
- decreases linearly with input education

3.2 Detailed plots

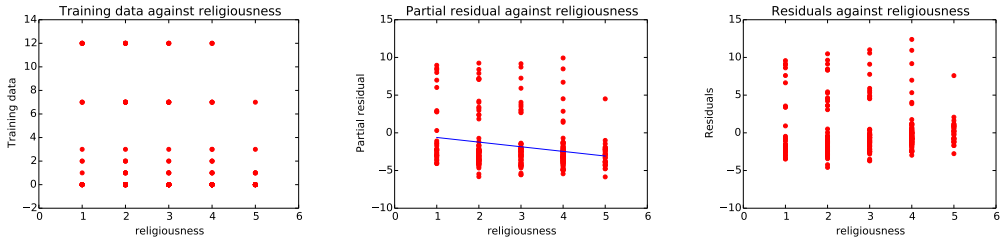
Decrease with rating The correlation between the data and the input rating is -0.25 (see figure 1a). Accounting for the rest of the model, this changes slightly to a part correlation of -0.20 (see figure 1b).



a) b) c)

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

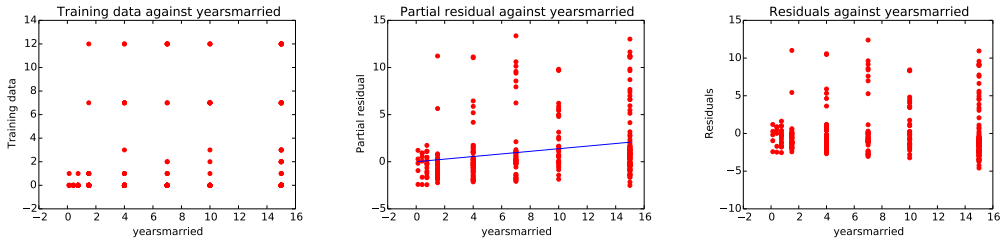
Decrease with religiousness The correlation between the data and the input religiousness is -0.22 (see figure 2a). This correlation does not change when accounting for the rest of the model (see figure 2b).



a) b) c)

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

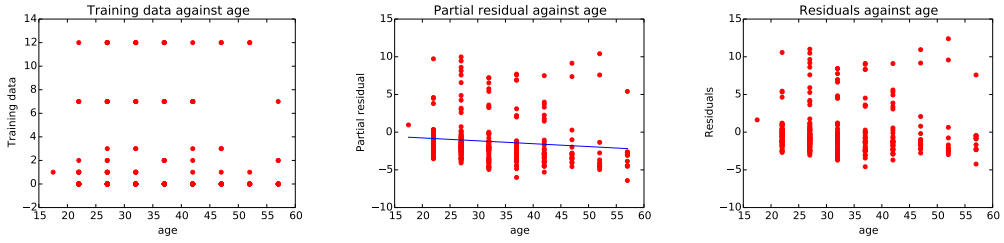
Increase with yearsmarried The correlation between the data and the input yearsmarried is 0.16 (see figure 3a). Accounting for the rest of the model, this changes slightly to a part correlation of 0.24 (see figure 3b).



a) b) c)

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

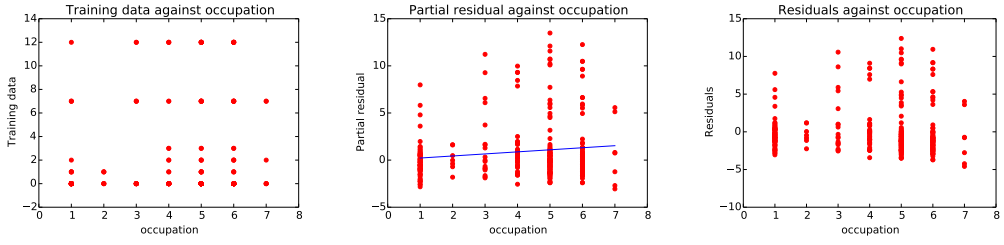
Decrease with age The correlation between the data and the input age is 0.10 (see figure 4a). Accounting for the rest of the model, this changes moderately to a part correlation of -0.12 (see figure 4b).



a) b) c)

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

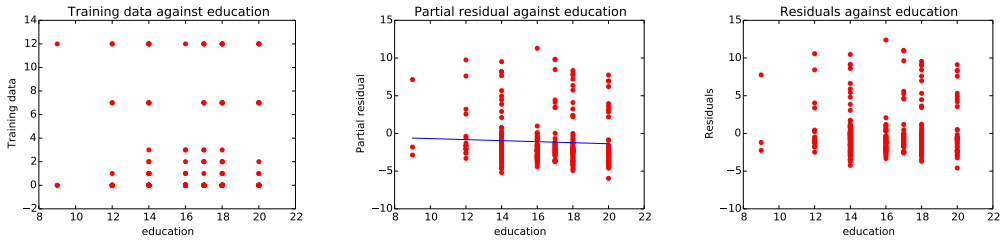
Increase with occupation The correlation between the data and the input occupation is 0.12 (see figure 5a). This correlation does not change when accounting for the rest of the model (see figure 5b).



a) b) c)

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

Decrease with education The correlation between the data and the input education is 0.04 (see figure 6a). Accounting for the rest of the model, this changes slightly to a part correlation of -0.05 (see figure 6b).



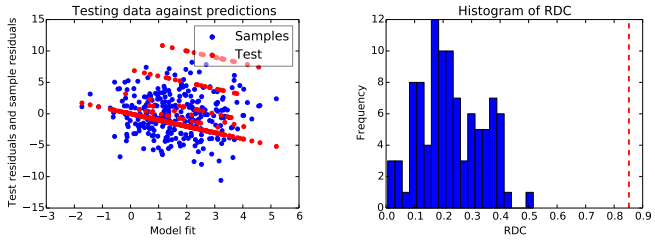
a) b) c)

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

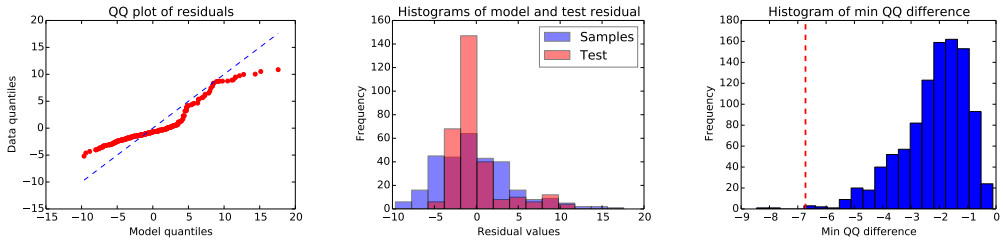
High dependence between residuals and model fit There is an unexpectedly high dependence between the residuals and model fit (see figure 7a). The dependence as measured by the randomised dependency coefficient (RDC) has a substantially larger value of 0.85 compared to its median value under the proposed model of 0.21 (see figure 7b).



a) b)

Figure 7: 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).

Low negative deviation between quantiles of test residuals and model There is an unexpectedly low negative deviation from equality between the quantiles of the residuals and the assumed noise model (see figure 8a). The minimum of this deviation occurs at the 99th percentile indicating that the test residuals have unexpectedly light positive tails. The minimum value of this deviation is -6.7 which is moderately lower than its median value under the proposed model of -1.9 (see figure 8c). To demonstrate the discrepancy in simpler terms I have plotted histograms of test set residuals and those expected under the model in figure 8b.



a) b) c)

Figure 8: a) Test set residuals vs model sample quantile quantile plot. b) Histograms of test set residuals and model residuals. c) Histogram of min QQ difference 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).