An automatic report for the dataset : gpadata

(A very basic version of) The Automatic Statistician

Abstract

This is a report analysing the dataset gpadata. 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 sauGPA. There are 6 input columns and 255 rows of data. A summary of these variables is given in table 1.

Name	Minimum	Median	Maximum
sauGPA	0	2.8	3.8
hsGPA	2	3.4	4
hsRANK	6	66	1e+02
athlete	0	0	1
ACTscore	17	22	34
hoursSTUDY	0	8	50
female	0	1	1

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	0.527	47.7
LASSO	0.535	46.4
BIC stepwise	0.543	44.1

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.

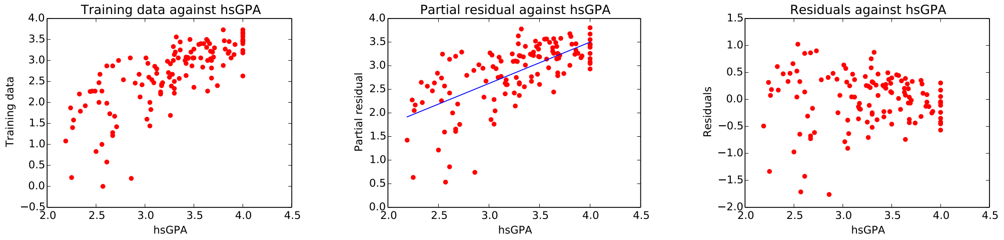
3.1 Summary

The output sauGPA:

- increases linearly with input hsGPA
- increases linearly with input ACTscore
- increases linearly with input female
- increases linearly with input hoursSTUDY
- increases linearly with input athlete
- increases linearly with input hsRANK

3.2 Detailed plots

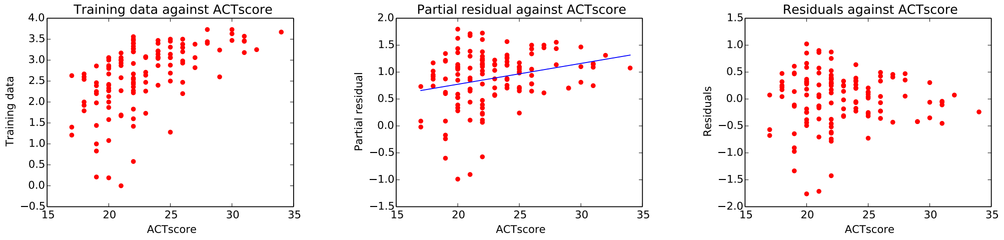
Increase with hsGPA The correlation between the data and the input hsGPA is 0.75 (see figure 1a). Accounting for the rest of the model, this changes slightly to a part correlation of 0.66 (see figure 1b).



a) b) c)

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

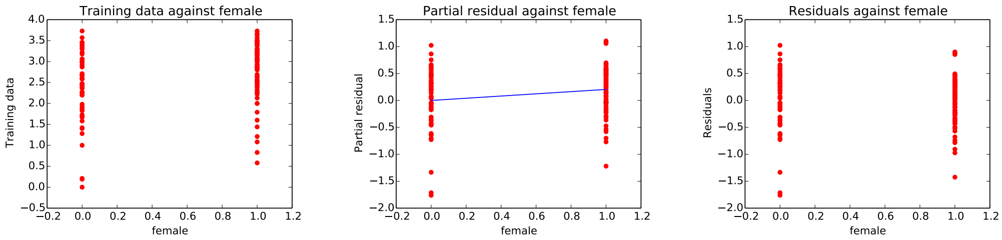
Increase with ACTscore The correlation between the data and the input ACTscore is 0.54 (see figure 2a). Accounting for the rest of the model, this changes moderately to a part correlation of 0.27 (see figure 2b).



a) b) c)

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

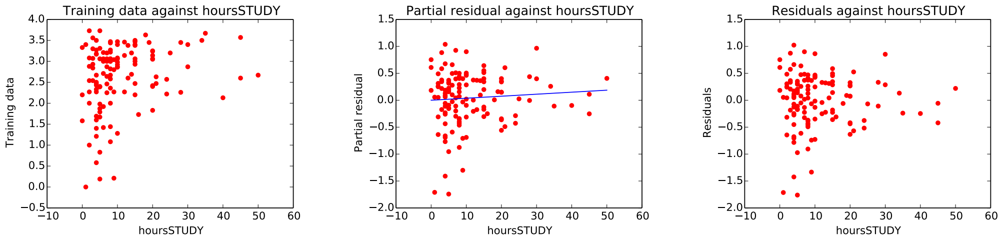
Increase with female The correlation between the data and the input female is 0.28 (see figure 3a). Accounting for the rest of the model, this changes slightly to a part correlation of 0.20 (see figure 3b).



a) b) c)

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

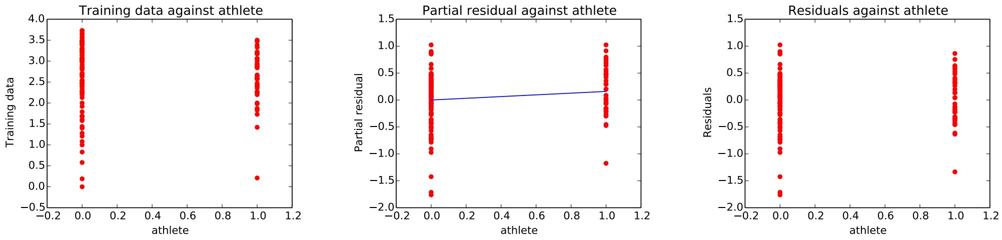
Increase with hoursSTUDY The correlation between the data and the input hoursSTUDY is 0.22 (see figure 4a). Accounting for the rest of the model, this changes moderately to a part correlation of 0.07 (see figure 4b).



a) b) c)

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

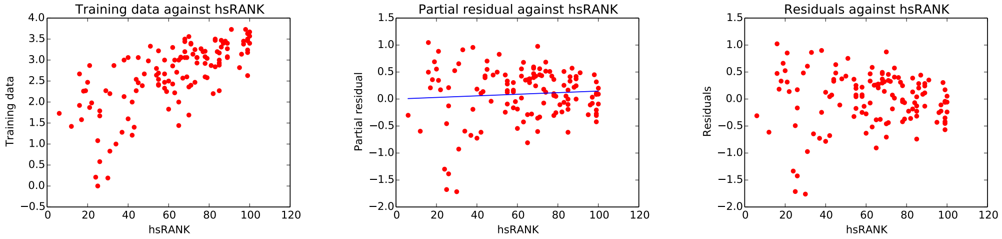
Increase with athlete The correlation between the data and the input athlete is -0.04 (see figure 5a). Accounting for the rest of the model, this changes moderately to a part correlation of 0.15 (see figure 5b).



a) b) c)

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

Increase with hsRANK The correlation between the data and the input hsRANK is 0.70 (see figure 6a). Accounting for the rest of the model, this changes substantially to a part correlation of 0.07 (see figure 6b).



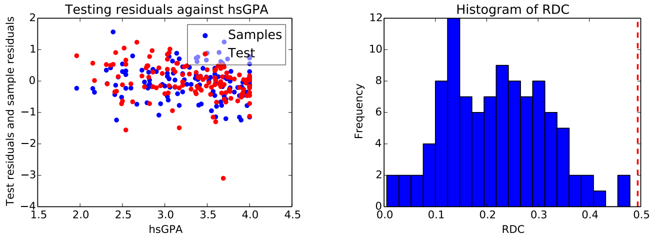
a) b) c)

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

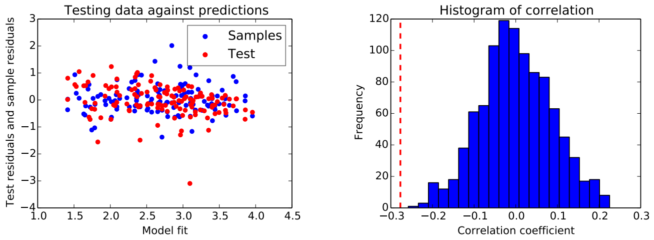
High dependence between residuals and hsGPA There is an unexpectedly high dependence between the residuals and input hsGPA (see figure 7a). The dependence as measured by the randomised dependency coefficient (RDC) has a substantially larger value of 0.49 compared to its median value under the proposed model of 0.22 (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).

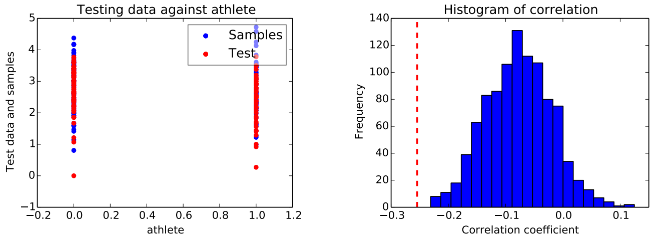
Low correlation between residuals and model fit There is an unexpectedly low correlation between the residuals and model fit (see figure 8a). The correlation has a substantially smaller value of -0.28 compared to its median value under the proposed model of -0.00 (see figure 8b).



a) b)

Figure 8: a) Test set and model sample residuals. b) Histogram of correlation 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).

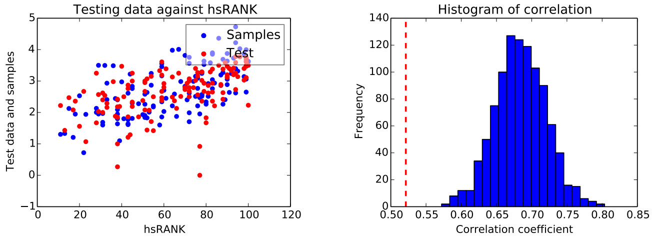
Low correlation between data and athlete There is an unexpectedly low correlation between the data and input athlete (see figure 9a). The correlation has a slightly smaller value of -0.25 compared to its median value under the proposed model of -0.08 (see figure 9b).



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).

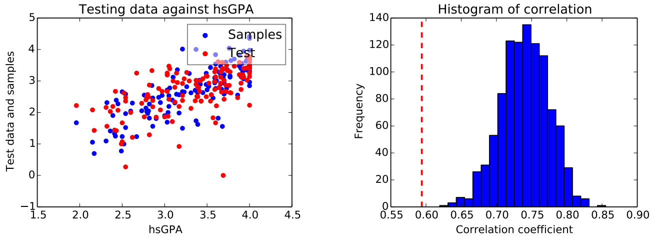
Low correlation between data and hsRANK There is an unexpectedly low correlation between the data and input hsRANK (see figure 10a). The correlation has a slightly smaller value of 0.52 compared to its median value under the proposed model of 0.68 (see figure 10b).



a) b)

Figure 10: 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).

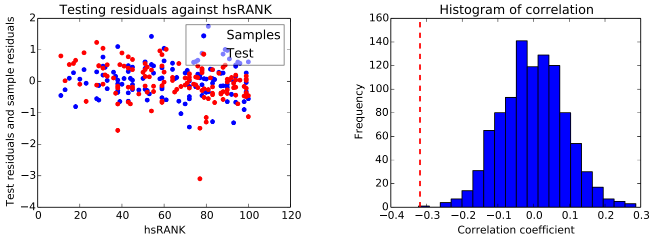
Low correlation between data and hsGPA There is an unexpectedly low correlation between the data and input hsGPA (see figure 11a). The correlation has a slightly smaller value of 0.59 compared to its median value under the proposed model of 0.74 (see figure 11b).



a) b)

Figure 11: 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).

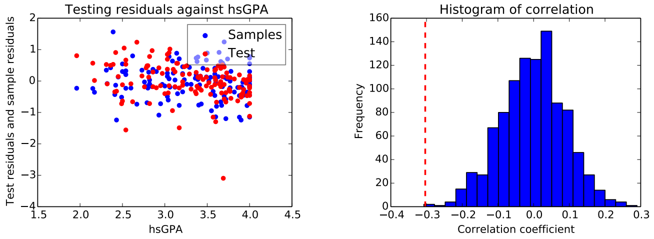
Low correlation between residuals and hsRANK There is an unexpectedly low correlation between the residuals and input hsRANK (see figure 12a). The correlation has a substantially smaller value of -0.32 compared to its median value under the proposed model of -0.00 (see figure 12b).



a) b)

Figure 12: a) Test set and model sample residuals. b) Histogram of correlation 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).

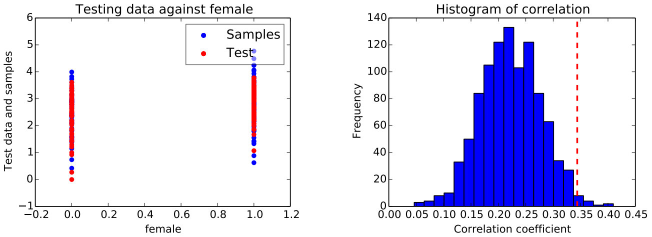
Low correlation between residuals and hsGPA There is an unexpectedly low correlation between the residuals and input hsGPA (see figure 13a). The correlation has a substantially smaller value of -0.30 compared to its median value under the proposed model of 0.00 (see figure 13b).



a) b)

Figure 13: 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).

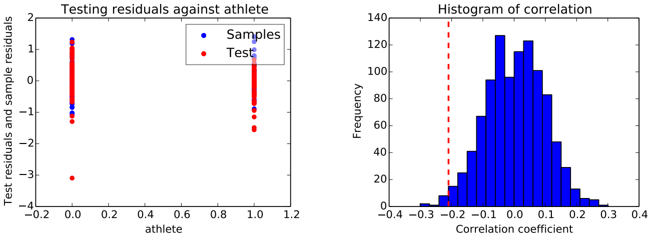
High correlation between data and female There is an unexpectedly high correlation between the data and input female (see figure 14a). The correlation has a slightly larger value of 0.34 compared to its median value under the proposed model of 0.22 (see figure 14b).



a) b)

Figure 14: 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).

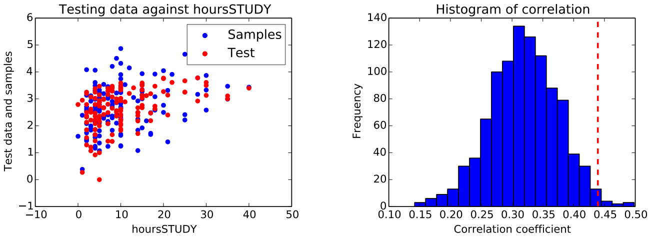
Low correlation between residuals and athlete There is an unexpectedly low correlation between the residuals and input athlete (see figure 15a). The correlation has a substantially smaller value of -0.21 compared to its median value under the proposed model of 0.01 (see figure 15b).



a) b)

Figure 15: 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).

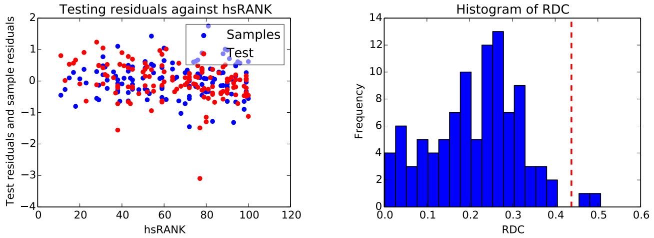
High correlation between data and hoursSTUDY There is an unexpectedly high correlation between the data and input hoursSTUDY (see figure 16a). The correlation has a slightly larger value of 0.44 compared to its median value under the proposed model of 0.32 (see figure 16b).



a) b)

Figure 16: 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).

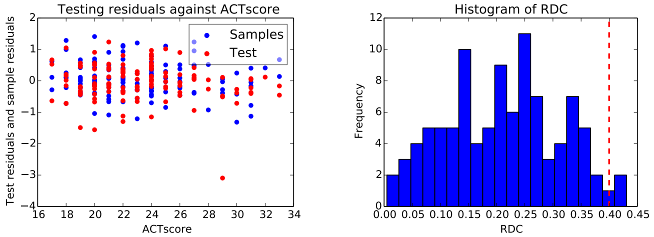
High dependence between residuals and hsRANK There is an unexpectedly high dependence between the residuals and input hsRANK (see figure 17a). The dependence as measured by the randomised dependency coefficient (RDC) has a substantially larger value of 0.44 compared to its median value under the proposed model of 0.23 (see figure 17b).



a) b)

Figure 17: 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 dependence between residuals and ACTscore There is an unexpectedly high dependence between the residuals and input ACTscore (see figure 18a). The dependence as measured by the randomised dependency coefficient (RDC) has a slightly larger value of 0.40 compared to its median value under the proposed model of 0.21 (see figure 18b).



a) b)

Figure 18: 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).