Statistical_Learning_Coursework final.pdf

by Tristan Peroy

Submission date: 26-Feb-2023 05:59PM (UTC+0000)
Submission ID: 197758622
File name: Statistical_Learning_Coursework_final.pdf (129.58K)
Word count: 992
Character count: 4737

Interesting: 2.5 Analysis: 2.5 Clarity: 1.5 Figures: 1.5

Statistical Learning Coursework

1 The Dataset

I have selected a dataset from Kaggle (by Diva Dugar) of house prices in King County, USA, with various descriptors of the house such as the number of bedrooms, the square footing of living room and the quality of the view. By regressing on this data, we can obtain coefficients to predict the price a house should be sold at, based on a description. This is of personal interest as someone who knows people moving to the US. I could recommend what to offer for a house they have identified so they avoid getting charged above market value. However, this data could also be useful selling houses, both in designing a house with high value (i.e. what can most easily be improved), as well as at the stage of correctly pricing it. We need interpretable models, hence I will start with multiple linear regression, then try to improve on it using lasso and ridge regression.

Interesting

Give some reasons for using ridge or lasso, e.g. to prevent the model from overfitting the data, etc.

2 Multiple Linear Regression

I first create a pairplot (Figure 1) to try to identify correlation between any 2 predictors, to check for redundency/ overfitting. We can see that descriptors such as waterfront and view often takes value 0 so must be handled with care, and that the number of bedrooms, bathrooms and square feet of living space are all weakly positively correlated as we would expect, as a big house is likely to have more of all. I first do simple linear regression between each predictor and the prices and find that the best R squared value is for square feet of living space (0.49), followed by bathrooms and view. These make sense as both space and utilities as well as nice views are likely to increase demand. The year built and waterfront variables have low predictive value alone, for the latter most likely because most houses do not have a waterfront at all. I then fit the model using multiple linear regression with stepwise regression using backwards deletion. In Figure 2, the coefficients are displayed when all features are included. I aim to iteratively remove any features from the model with a p value above 0.05, as a p value below this suggests there is some evidence against the null hypothesis, the null hypothesis being that there is no relationship at all between the feature and the price variable, however all variables are sufficiently significant to be kept (much below 0.05 in p value). In terms of F-statistic, I compare it to the 99th quantile of F, which I find to be 2.408, hence as the F statistic is 3568 (much larger), we can reject the null hypothesis that all feature coefficients should be

Nice

set to 0. In Figure 4, we see that the residuals of the model roughly follow a bell shaped curve centred at 0 as they should. The qqplot in Figure 3 however, suggests the errors are not exactly normally distributed, with heavier tails on both ends. Finding outliers from the residual vs fitted plot, and regressing again with these removed, the R squared is hardly improved (from 0.5979). I conclude that the residuals are not especially skewed one way or other but may follow a more heavy tailed distribution such as a t-distribution. 0.5979 is not a perfect fit but provides some useful predictive value (though with high variance). Perhaps the model can at least provide a confidence range of prices to buy/sell a house within. More features, i.e. more information on the house or transformation of variables may be needed to further improve the model.

3 Ridge Regression and Lasso Regression

I will now attempt to use ridge and lasso shrinkage methods to see if by adapting some of the coefficients, the model has improved R squared values. I split the data into testing and training, and training into 5 folds. By implementing cross-validation using R squared score as the metric, I obtain optimal penalty hyperparameters for lasso and ridge. When we retrain the model fixing the optimal penalty on the whole dataset, we find as shown in Figure 5 that some parameters have been shrunk, have reduced magnitude. For lasso, the coefficients for sqft living, bedrooms, condition and waterfront are increased and the rest shrunk. Similarly for ridge some parameters increase and some decrease. With much larger penalties, the coefficients will all be forced to shrink to zero, introducing more bias and but attempting to reduce variance. Of the 3 models, lasso has the lowest mean squared error and highest R squared (marginally) whilst ridge has the lowest R squared and worst performance on test data. For lasso compared to linear regression introducing this amount of bias is enough the reduce the cross-validation variance to overall minimise mean squared error on the set set too, i.e. the slight improvement also generalises to the test data. However, ridge and lasso still produce out of sample R squared below 0.6 so are no more useful than linear regression, and significant improvements to the model would be best made by having more useful features (or perhaps using nonlinear models). Lasso performs slightly better as the data is suited to a penalty that encourages sparsity (whereas ridge tends to reduce all parameters simulataneously and less aggressively).

Usually one standardises and centres the data before fitting lasso or ridge. The entering is useful so that the intercept can be removed from the data and added back later (so that it isn't shrunk)

Could you suggest other models that could be used instead in order to improve the fit?

4 Conclusion

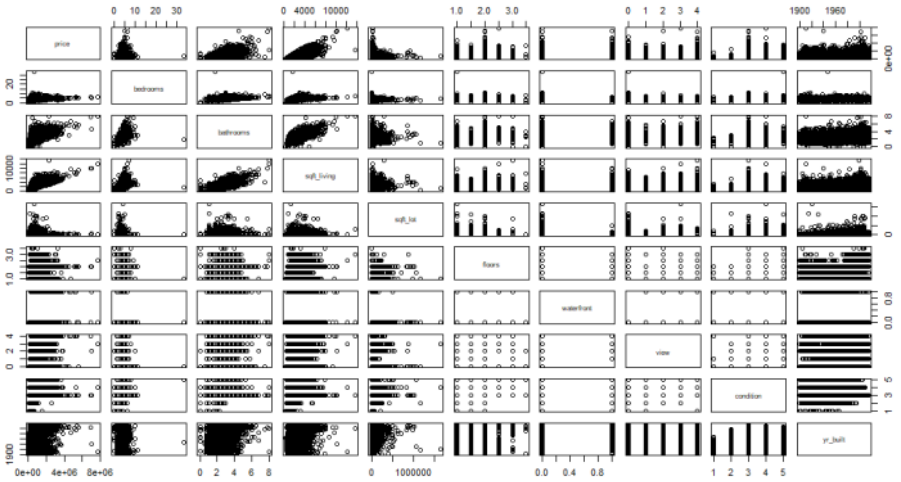
We can try to make conclusions about the relative importance of features of a house based on the relative R squared scores for simple linear regression (square footage then bathrooms are the best predictors). Also, as the coefficient is negative, we see interestingly that more recently built houses tend to have slightly lower value. Ultimately, our models can predict for us the price a house should have based on a description, though with high prediction variance.

Which coefficient are you referring to here?

Overall the report seems good. The conclusion could be expanded upon by including more insight into what predictors are best and why. Some of the figures seem a little redundant (e.g Figure 2 could have been included as part of Figure 5)

5 Figures

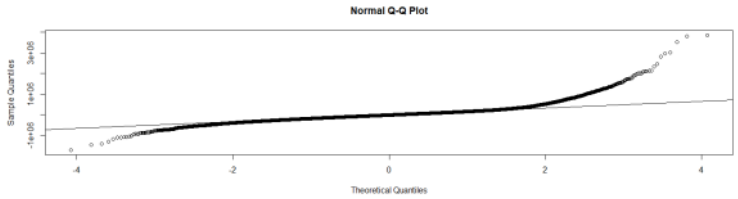
5.1 Figure 1



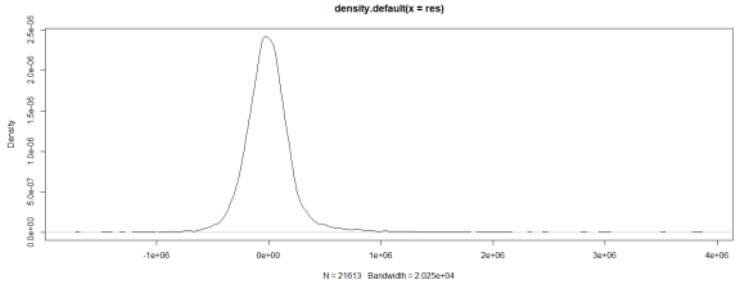
5.2 Figure 2

```
Residuals:
     Min       1Q   Median       3Q      Max
-1704390  -123246   -12663    98649  3852508

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.454e+06  1.386e+05  39.344  < 2e-16 ***
bedrooms    -5.671e+04  2.157e+03 -26.289  < 2e-16 ***
bathrooms    5.862e+04  3.663e+03  16.004  < 2e-16 ***
sqft_living  2.771e+02  2.921e+00  94.841  < 2e-16 ***
sqft_lot    -3.154e-01  3.916e-02  -8.055 8.34e-16 ***
floors       5.917e+04  3.627e+03  16.313  < 2e-16 ***
waterfront   5.445e+05  2.003e+04  27.183  < 2e-16 ***
view         5.928e+04  2.383e+03  24.872  < 2e-16 ***
condition    1.666e+04  2.649e+03   6.287 3.29e-10 ***
yr_built    -2.832e+03  7.032e+01 -40.269  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 233000 on 21603 degrees of freedom
Multiple R-squared:  0.5979,    Adjusted R-squared:  0.5977
F-statistic:  3568 on 9 and 21603 DF,  p-value: < 2.2e-16
```

5.3 Figure 3



5.4 Figure 4



5.5 Figure 5

coefficients	Multiple Linear Regression	Lasso Regression	Ridge Regression
intercept	545000	5399976	4754861
bedrooms	-5671	-6031	-4216
bathrooms	5862	5548.5	7031
sqft_living	277	285	243
sqft_lot	0.3154	-0.27	-0.176
floors	5917	5702	5527
waterfront	54450	54981	53191
view	5928	5325	6000
condition	1666	1765	1947
yr_built	-283.2	-281	-248
R squared	0.5978765	0.5979024	0.58758

Statistical_Learning_Coursework final.pdf

GRADEMARK REPORT

FINAL GRADE

80/100

GENERAL COMMENTS

PAGE 1

Text Comment. Interesting: 2.5 Analysis: 2.5 Clarity: 1.5 Figures: 1.5

Text Comment. Interesting

Text Comment. Give some reasons for using ridge or lasso, e.g. to prevent the model from overfitting the data, etc.

Text Comment. Nice

PAGE 2

Text Comment. Usually one standardises and centres the data before fitting lasso or ridge. The entering is useful so that the intercept can be removed from the data and added back later (so that it isn't shrunk)

Text Comment. Could you suggest other models that could be used instead in order to improve the fit?

Text Comment. Which coefficient are you referring to here?

Text Comment. Overall the report seems good. The conclusion could be expanded upon by including more insight into what predictors are best and why. Some of the figures seem a little redundant (e.g Figure 2 could have been included as part of Figure 5)

PAGE 3

PAGE 4