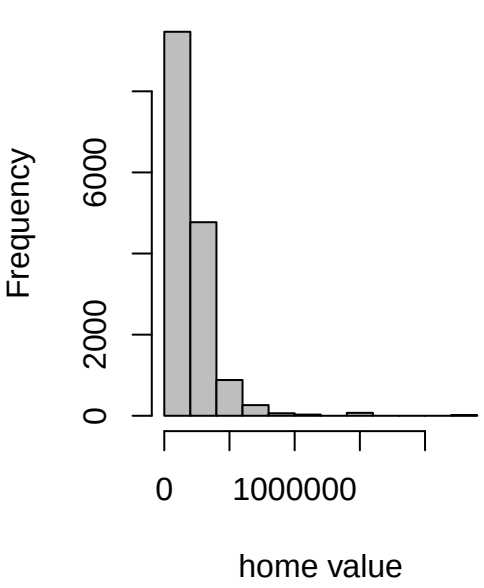
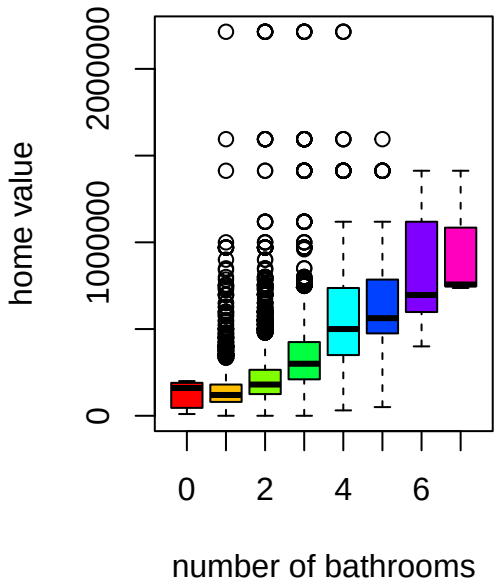
Homework 2 Assignment

Carrie Mecca, Charlie Marcou, Jessie Bustin and Jasmine Zhang

This assignment contains TWO data analyses. Each question is 10 points.

Predicting House Prices





Question 1

Regress log price onto all variables but mortgage. What is the R2? How many coefficients are used in this model and how many are significant at 10% FDR? Re-run regression with only the significant covariates, and compare R2 to the full model.

The R^2 is .447301. There are 40 coefficients. Of these 40 coefficients, we can see that 5 are not significant. The five that are no significant are: "ETRANSY" "STATECO" "STATECT" "BEDRMS" and "NUNITS"

We drop BEDRMS and NUNITS, but do not drop the insignificant categorial levels. After this, the R^2 of the reduced model is 0.4471981. In comparison to the full model, there is a slight decrease in the R^2. The shrinkage between the R^2 of the two models is .0001029.

Question 2

Fit a regression for whether the buyer had more than 20 percent down (onto everything but AMMORT and LPRICE). Interpret effects for Pennsylvania state, 1st home buyers and the number of bathrooms. Add and describe an interaction between 1st home-buyers and the number of baths.

For a buyer in pennsylvania we expect the log odds of the downpayment being greater than 20% to increase by .6011 For a one unit increase in bedrooms, we expect the log odds of the downpayment being greater than

20% to change by -.00286 For a first time home buyer, we expect the log odds of the the downpayment being greater than 20% to change by -.37

After expanding the model, we see a negative interaction between 1st time home buyers and number of bathrooms. This indicates that the log odds of the downpayment greater than 20% for each additional bathroom, will be lower for a first time home buyer than a buyer who has bough a home previously.

Question 3

Focus only on a subset of homes worth > 100k. Train the full model from Question 1 on this subset. Predict the left-out homes using this model. What is the out-of-sample fit (i.e. R^2)? Explain why you get this value.

For the training data, the subset of homes worth > 100k, the R^2 is 0.3919407. The out-of-sample data, however, has a lower R^2 at only 0.1834929. The reason the test set R^2 is lower is because the model is overfitting on the subset of higher-priced homes and, subsequently, is unable to perform as well on the lower-priced subset. Using the model to extrapolate to homes with smaller prices would not be reasonable.

Amazon Reviews

We will use the same datasets (review_subset.csv, word_freq.csv and words.csv) as in Assignment 1.

Question 4

We want to build a predictor of customer ratings from product reviews and product attributes. For these questions, you will fit a LASSO path of logistic regression using a binary outcome:

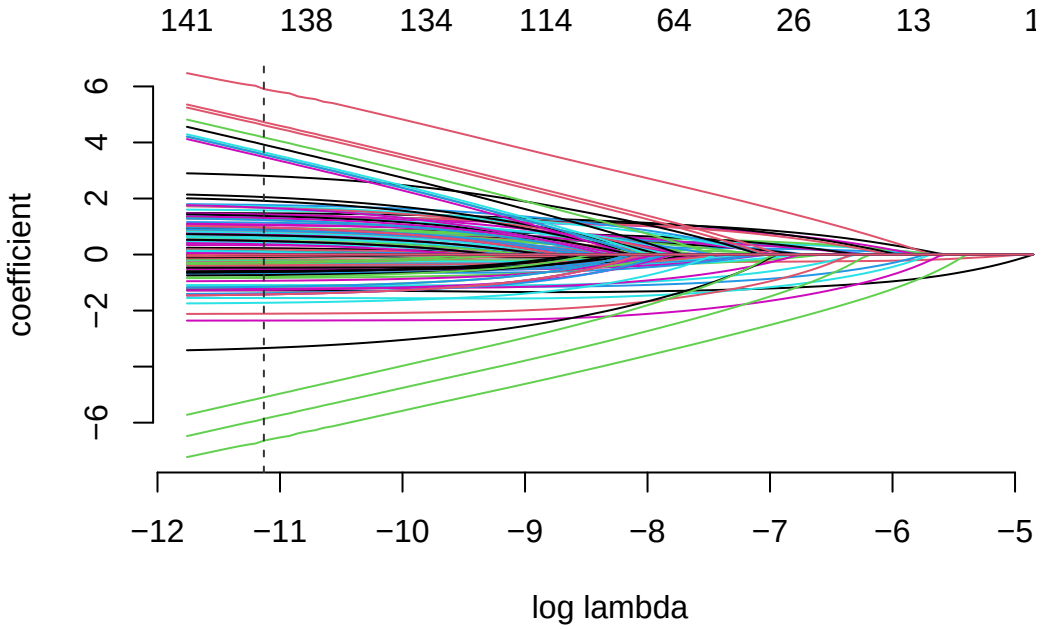
$$Y = 1 \quad \text{for 5 stars} \tag{1}$$
$$Y = 0 \quad \text{for less than 5 stars.} \tag{2}$$

Fit a LASSO model with only product categories. The start code prepares a sparse design matrix of 142 product categories. What is the in-sample R2 for the AICc slice of the LASSO path? Why did we use standardize FALSE?

We can use standardize false because the x's are on the same scale already because they are all factor levels. not sure what is meant by r2 for aicc slice of lasso path. The in-sample R2 for the best lambda is 0.1048737.

```
## Loading required package: Matrix
```

[1] "factor"



binomial gamlr with 142 inputs and 100 segments.

Question 5

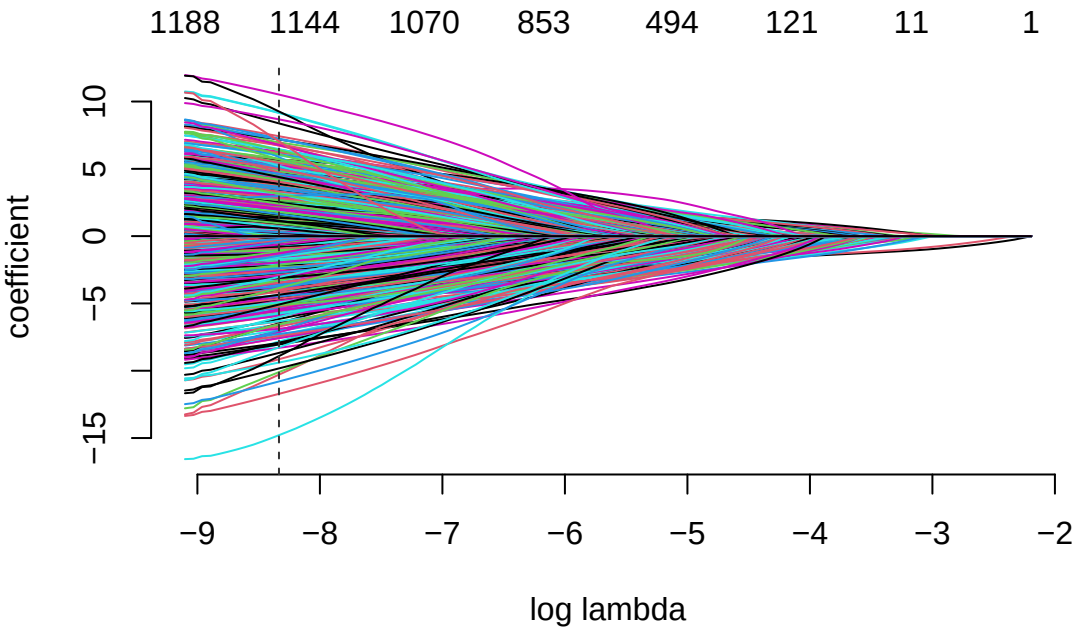
Fit a LASSO model with both product categories and the review content (i.e. the frequency of occurrence of words). Use AICc to select lambda. How many words were selected as predictive of a 5 star review? Which 10 words have the most positive effect on odds of a 5 star review? What is the interpretation of the coefficient for the word 'discount'?

At a log-lambda of -8.334091, we select 1154 words and categories

The top 10 words with positive effect on odds of a 5 star review are: worried, Breads, plus, Almond Leaveners & Yeasts, excellently, find, grains, Computers, Features, hound.

The coefficient of discount is: 6.961539. This means that we expect log odds to increase by 6.961539 when the word discount is included in a review.

[1] 13319 1125



seg89

-8.334091

Question 6

Continue with the model from Question 5. Run cross-validation to obtain the best lambda value that minimizes OOS deviance. How many coefficients are nonzero then? How many are nonzero under the 1se rule?

The lambda value that minimizes OOS deviance is 0.00137444. With that lambda, there are 974 non-zero coefficients. Using the 1se rule instead, there are only 831 non-zero coefficients (with a random seed of 250).

fold 1,2,3,4,5,done. [1] 0.00137444 [1] 974 [1] 831