Casella Berger Oscars

Jackson Curtis

March 2019

1 Introduction

In this project we want to predict the number of stars given on an Amazon review of the textbook Statistical Inference by George Casella and Roger Berger. To make our predictions we are given the following information: the username of the author, the date of the review, whether its hardcover or paperback, whether its a verified user, how many people marked it as helpful, and the review itself. We have 68 of the approximately 100 reviews to train our model and we will predict on the unseen third, with the goal of minimizing MSE on the prediction of number of stars.

2 Model

To answer this question we will fit a ordinal regression model to the data. Because we have so little data (and the data we do have has weak relationships to the response) we will create a very simple model to avoid overfitting, and we will use the proportional odds assumption to limit the number of parameters we estimate. Two variables will be created from the text of the data using the sentimentr package in R. The first is a count of the words, and the second is a sentiment score which attempts to assign a positive or negative number that describes how positive the connotation of the words being used is. Our full model can be written as:

$$Y_i \sim Multinomial(\pi_{i1}, ..., \pi_{i5}) \tag{1}$$

$$log(\frac{\pi_{i1} + ... + \pi_{ij}}{\pi_{i(j+1)} + ... + \pi_{i5}}) = \beta_{0j} + \beta_1 * VerifiedPurchase + \beta_2 * log(wordCount) + \beta_3 * sentiment$$

This model models the odds of being in the lower groups vs. the higher groups, so a negative value of a β coefficient suggests a higher probability of receiving a higher star rating in the presence of that covariate. Because our model is so simple, it is also very interpretable. Our $\hat{\beta_1} = 1.21$ means that the odds ratio between a verified user vs. a non-verified user for the more negative category is 3.35. For a one unit increase in log(wordcount), the odds ratio of being in the low group is 2.57 ($\hat{\beta_2} = .945$). The strongest effect, $\hat{\beta_3} = -5.63$, means that for a one unit increase in sentiment (which ranges from -0.6 to 1.8) the odds ratio that the review with the higher sentiment has only 0.003 the odds of being in the lower rated group than the lower sentiment review.

The last two coefficients agree with our prior knowledge about reviews. Reviews that use mostly negatively connoted words are much more likely to be negative in nature. The longer a review is the more likely it is to be negative, which agrees with the general internet trend of negativity making people more vocal than positivity. The direction of the verified-review coefficient was a little surprising because if you look at the average rating before controlling for the other covariates, verified reviews are higher in general, but after controlling for the other factors they tend to be lower. This might make intuitive sense in that unverified reviews can sometimes be used to inflate product ratings, but the effect isn't large and is questionably significant.

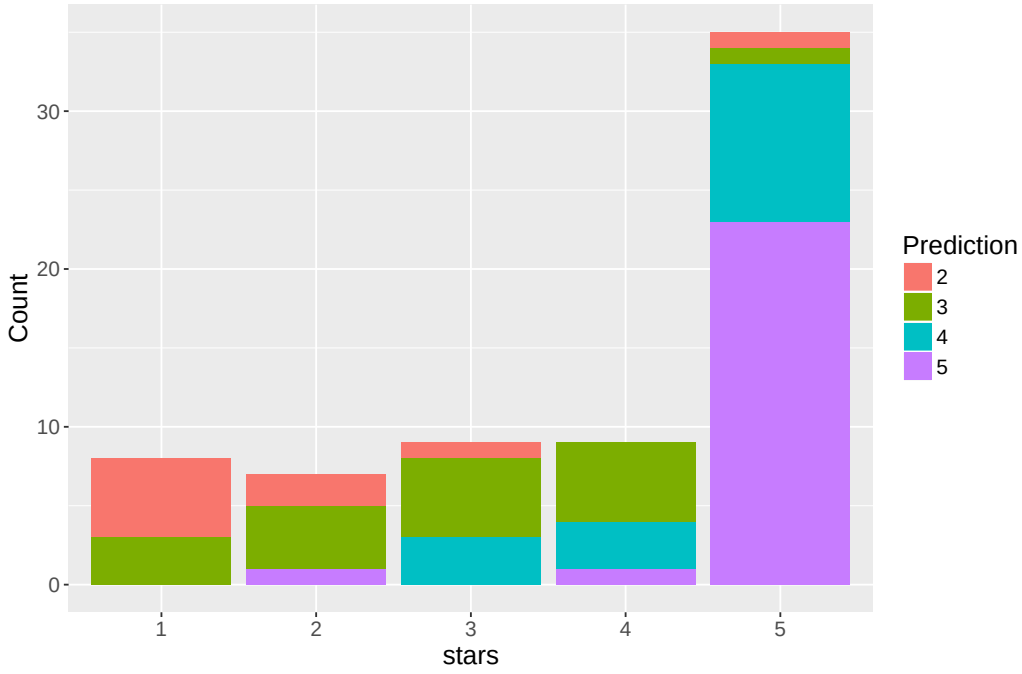

Figure 1: The number of stars predicted plotted against the number of stars actually received (1 star reviews were never predicted)

3 Model Diagnostics

To calculate a prediction for the model, we can calculate the probabilities of each category given the covariates. We then take a weighted average of the five probabilities and round to the nearest integer to obtain our prediction for the number of stars.

Our best guess at our out of sample MSE can be calculated using leave-one-out cross validation. We will hold out one observation each time we build the model, and predict the number of stars for the held out point. The mean squared error using this method is 0.9265, which corresponds to an average distance off of just under 1 star (0.96). Figure 1 shows what category we predicted for each of the actual categories.