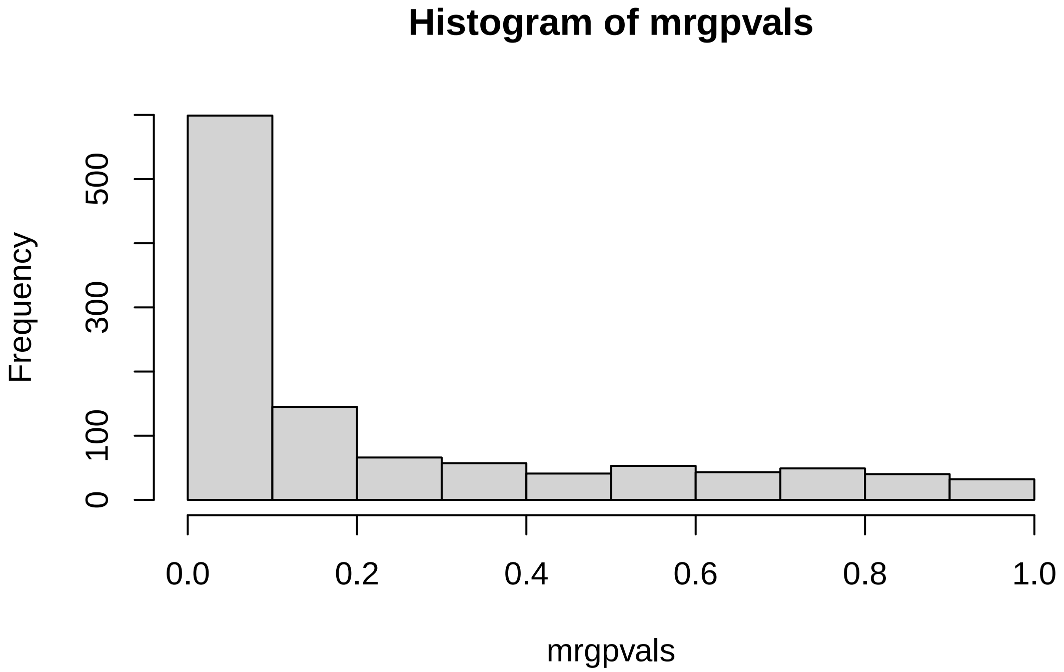
Homework 1 Assignment

Charlie Marcou, Carrie Mecca, Jessie Bustin and Jasmine Zhang

Homework Questions:

(1) Plot the p-values from the marginal screening and comment on their distribution. (10 point)

By plotting a histogram of the p-values from the marginal screening, we can see the distribution of the p-values is heavily right skewed.

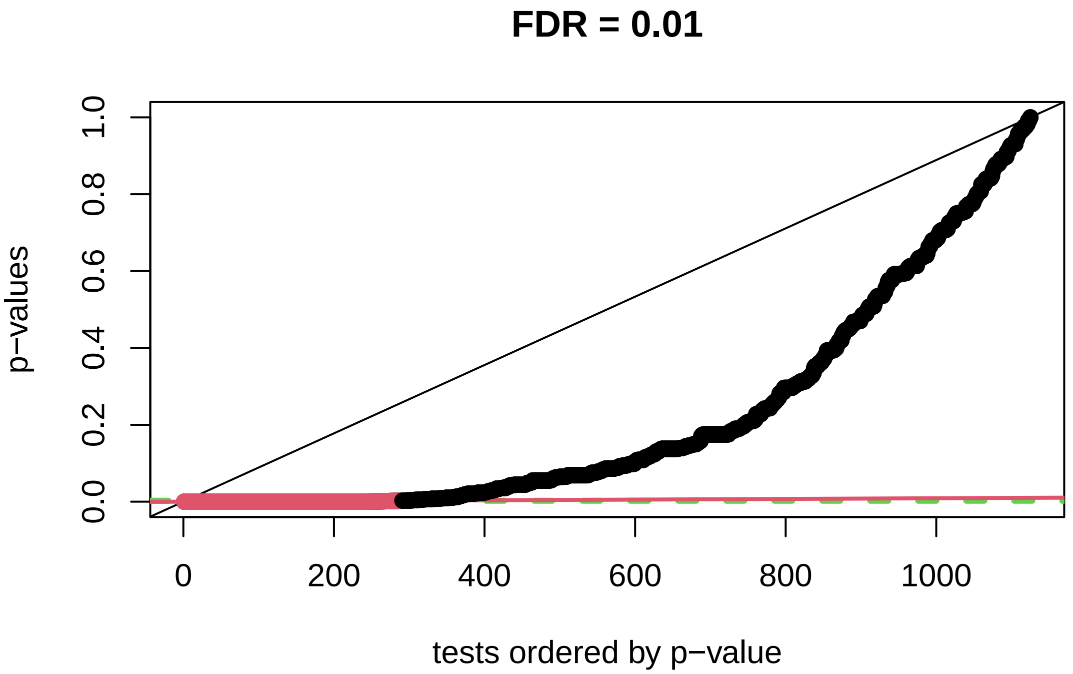


(2) Let's do standard statistical testing. How many tests are significant at the alpha level 0.05 and 0.01? (10 point)

The marginal p-values are already the p-values for t-tests of the predictor in each marginal regression. At alpha = .05, we have 461 significant tests. At alpha =.01 we have 348 significant tests.

(3) What is the p-value cutoff for 1% FDR? Plot and describe the rejection region. (10 point)

The p-value cutoff for 1% FDR is .002413249. The rejection region is visible in the plot and encompasses the points in red. This area is defined by a cutoff for the p-value that is found using the set FDR of 1% for conducting multiple tests. The false discovery rate, or number of tests that return a false positive is controlled and a region of p-values is identified where the null hypothesis is rejected.



(4) How many discoveries do you find at q=0.01 and how many do you expect to be false? (10 point)

At q=.01, there are 290 discoveries. We expect 2.9, or approximately 3 since the number of tests is an integer, of those discoveries to be false.

(5) What are the 10 most significant words? Do these results make sense to you? What are the advantages and disadvantages of our FDR anaysis? (10 point)

The 10 most significant words were: not, horrible, great, bad, nasty, disappointed, new, but,same, poor. Many of these words are direct expressions of a reviewer's sentiment about the product and so make sense as being significant such as "disappointed" or "great". However, there are a few significant words such as "not", "but" and "same" which are less obviously indicative of product rating. It is possible that these words frequently co-occur with other more meaningful words, which could explain the significance. This is a potential disadvantage of the FDR analysis, because we are not considering additive or interactive effect that words might have together. An advantage of the FDR analysis though is that it allows us to limit how many results that are considered significant are false positives.