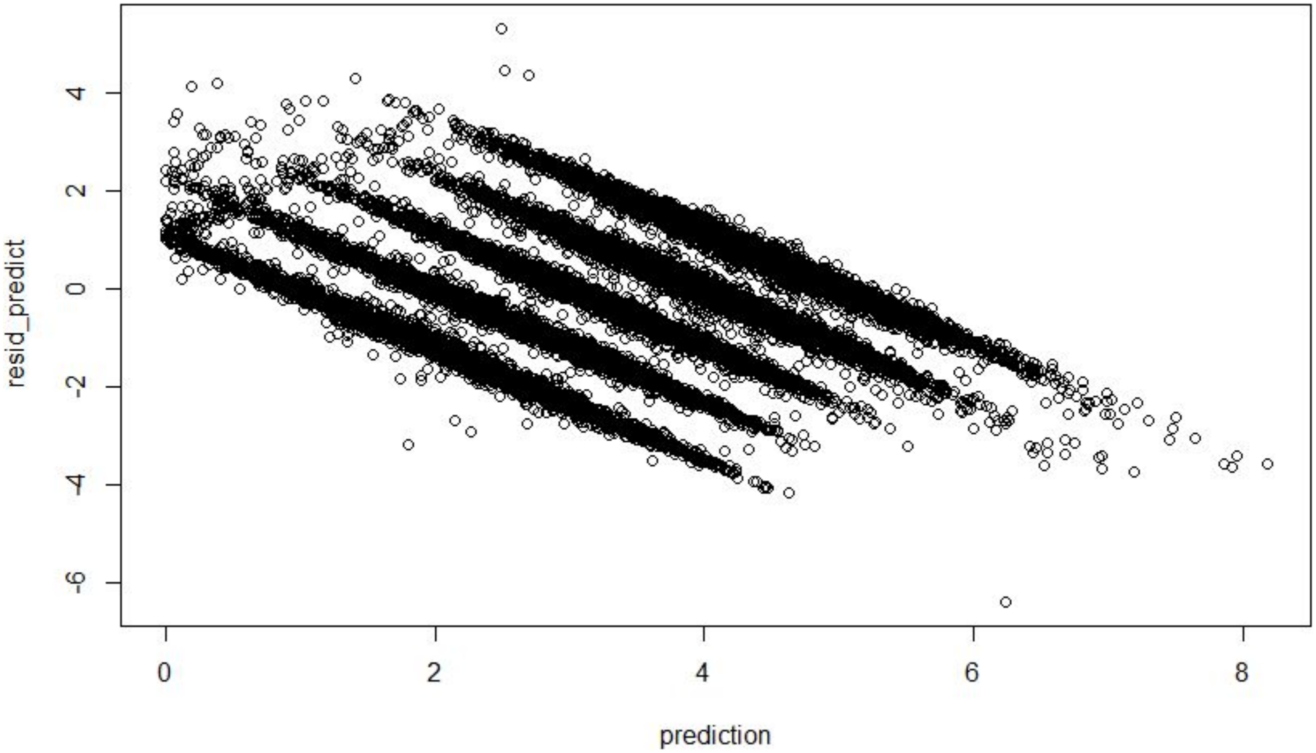
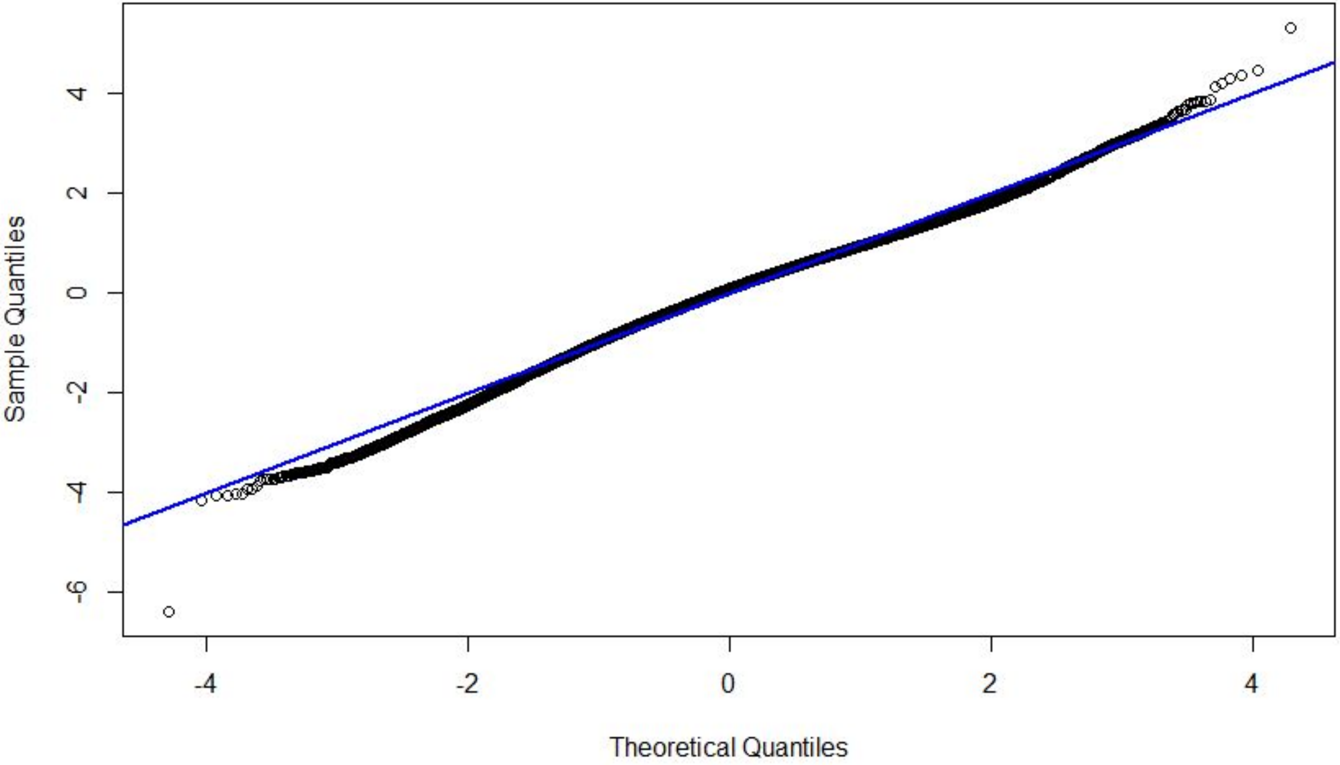
Group 17

James Thomason, Xiaosheng Liu, Zhelong LI

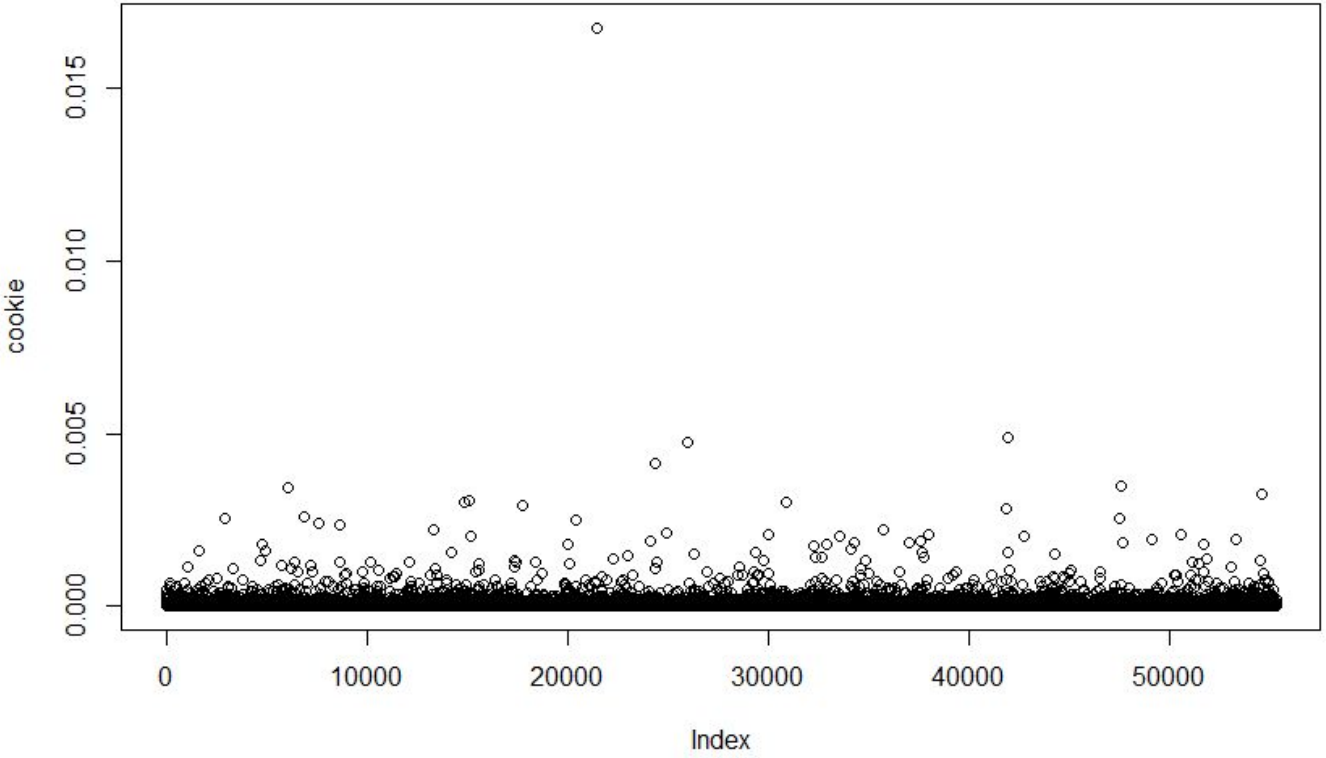
Residual plot



QQ Plot

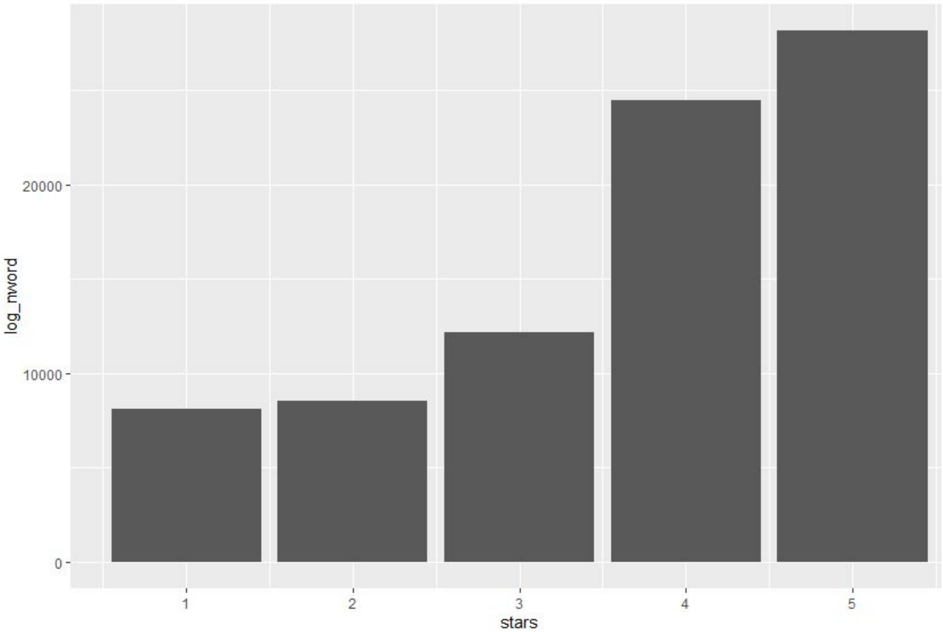


Cook’s distance

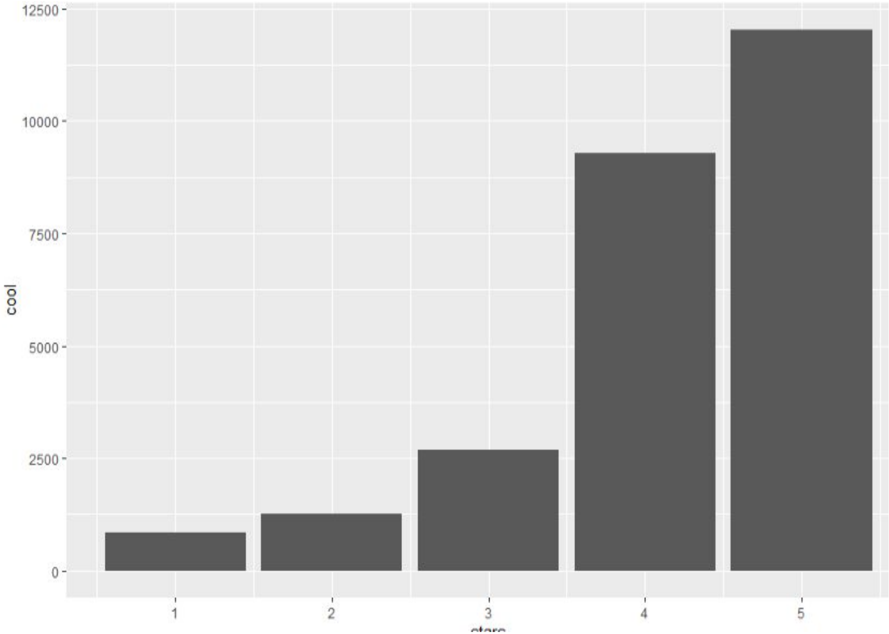


Graphs of stars vs variables

Log of nword



Cool



Why Lasso?

$$\text{minimize } \frac{1}{2n}\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1X_{i1}-\cdots-\beta_pX_{ip})^2+\lambda(|\beta_1|+|\beta_2|+\cdots+|\beta_p|)$$

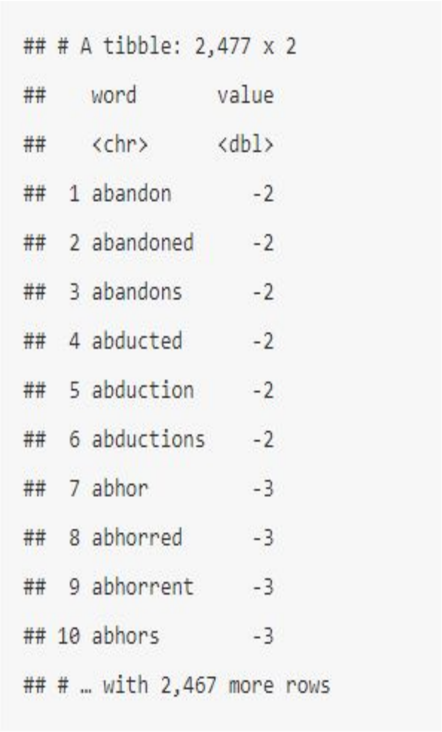
when λ is a “good” value in between 0 and ∞, MSE for estimation of β_J S will be smallest and prediction of Y will have the smallest variance. When λ=0 (i.e. least square), we need n<p to have unique solution for βs. By having λ≠0, even when n<p, there is still a unique solution for β. For lasso, if we move λ from 0 to ∞, the β_js will become zero one after another (i.e. predictors being removed one after another), with less important predictors being removed first. For ridge regression, βs will not be zero unless λ=∞. So Lasso can give you sparse estimation of β and can be used for model selection, which makes it very popular

Mention dictionaries

Most helpful strategy:

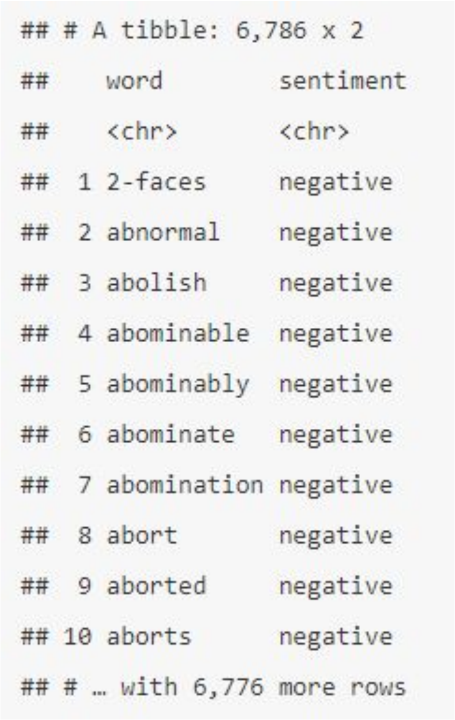
- Finding dictionary
  - AFINN
  - Bing
- Trial and error
  - P-value of each word.
  - Words that appeared frequently in reviews.
- Narrowing down the variables
  - Removing funny and cool
  - Absolute value of sentiment scores
  - Adding “non-words” like emojis and money signs($).

AFINN:



```
## # A tibble: 2,477 x 2
##    word       value
##    <chr>      <dbl>
##  1 abandon       -2
##  2 abandoned     -2
##  3 abandons      -2
##  4 abducted      -2
##  5 abduction     -2
##  6 abductions    -2
##  7 abhor         -3
##  8 abhorred      -3
##  9 abhorrent     -3
## 10 abhors        -3
## # … with 2,467 more rows
```

Bing:



```
## # A tibble: 6,786 x 2
##    word        sentiment
##    <chr>       <chr>
##  1 2-faces     negative
##  2 abnormal    negative
##  3 abolish     negative
##  4 abominable  negative
##  5 abominably  negative
##  6 abominate   negative
##  7 abomination negative
##  8 abort       negative
##  9 aborted     negative
## 10 aborts      negative
## # … with 6,776 more rows
```

Conclusion

Scoring a RMSE of 0.88133 means that the standard deviation of the prediction errors were relatively low and so our model had a standard deviation of unexplained variance less than one.