Team 23

Bob Dai, Jialuo Li

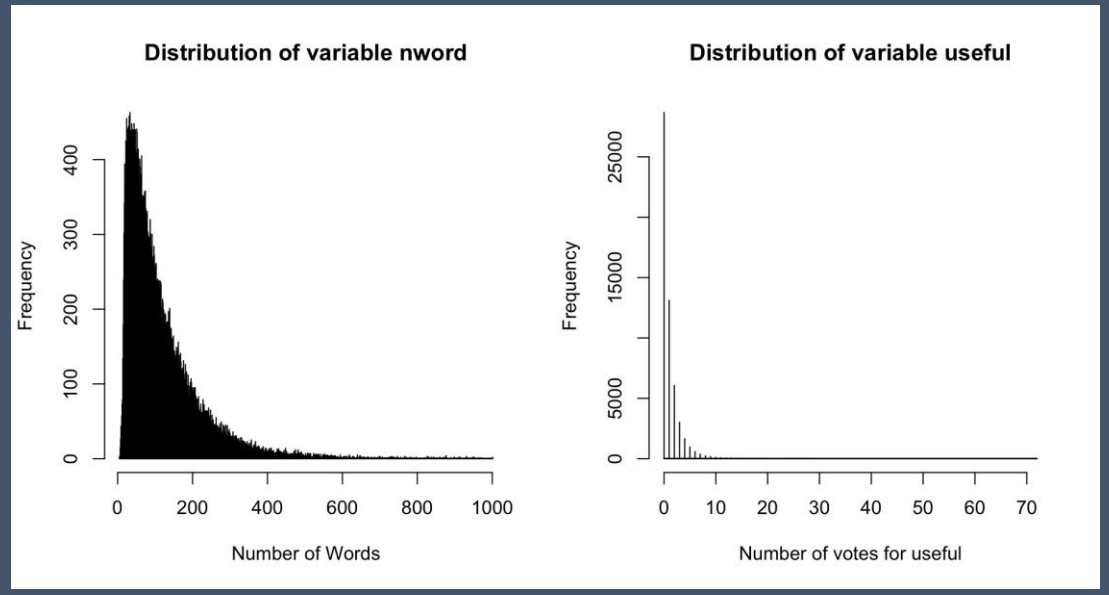
Predictors Selection

Criteria

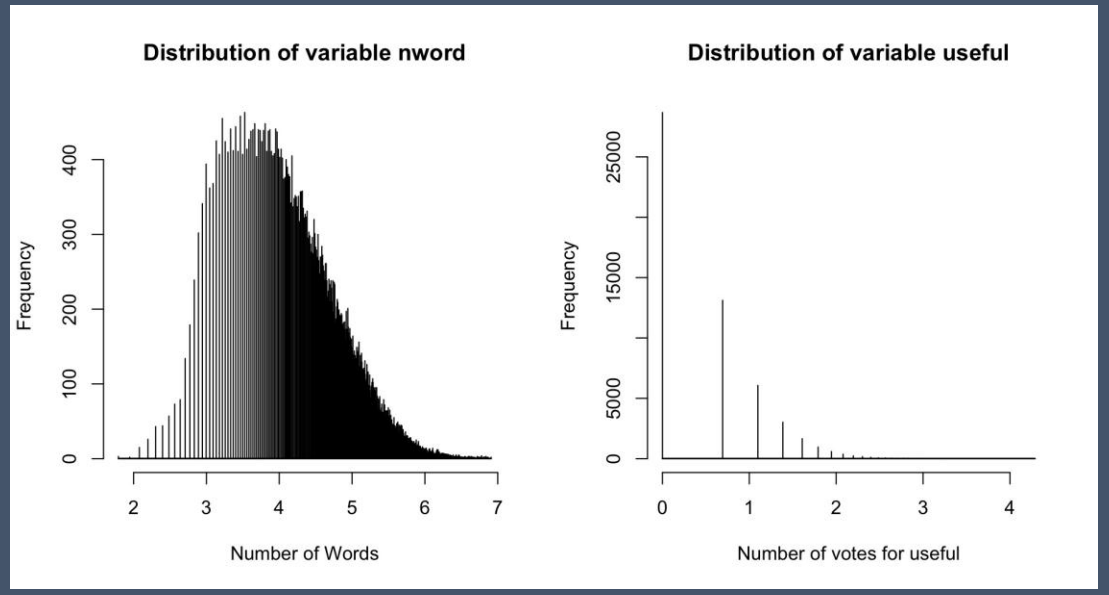
- Used frequency of words in the review as criteria to pre-select predictors
- Delete numbers and special symbols
- Lasso regression selects 2250 predictors in total

Data Preparation

- Raw data



- Log Transformation



- For each variables and nword (+1 before log transformation)

Modeling

- Result

R^2	0.6702
RSS	31024.95
TSS	94067.93
ESS	63042.98
RMSE	0.78323
R^2 adjusted	0.6562
Standard error	0.7644

- Interpretation

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.125e+00  6.600e-02  62.501  < 2e-16 ***
useful        -9.101e-02  6.845e-03 -13.295  < 2e-16 ***
funny         -1.762e-01  1.036e-02 -17.015  < 2e-16 ***
cool           3.044e-01  9.234e-03  32.962  < 2e-16 ***
nword         -4.056e-01  3.215e-02 -12.616  < 2e-16 ***
sentiment      3.221e-01  4.529e-03  71.117  < 2e-16 ***
gem            1.654e-01  4.330e-02   3.821 0.000133 ***
incredible     3.269e-01  4.124e-02   7.927 2.29e-15 ***
perfection     6.626e-02  5.236e-02   1.265 0.205753
heaven         2.216e-01  5.096e-02   4.348 1.38e-05 ***
phenomenal     1.575e-01  9.199e-02   1.712 0.086815 .
```

Take “cool” as an example, its p-value is less than 2e-16, which means that it have significant effect on the rating of customer. If a customer use “cool” one time, the 3.044e-01 means that if he uses “cool” one more time while holding all other variables unchanged, the prediction on rating tends to increase 0.3044*(log(2+1)-log(1+1)) = 0.1234236

Inference

- **Hypothesis**:

  H_0: Coefficients of all the predictors equal to 0 at a = 0.05

  H_a: At least one Coefficient is not equal to 0 at a = 0.05

- **Test Statistic**:

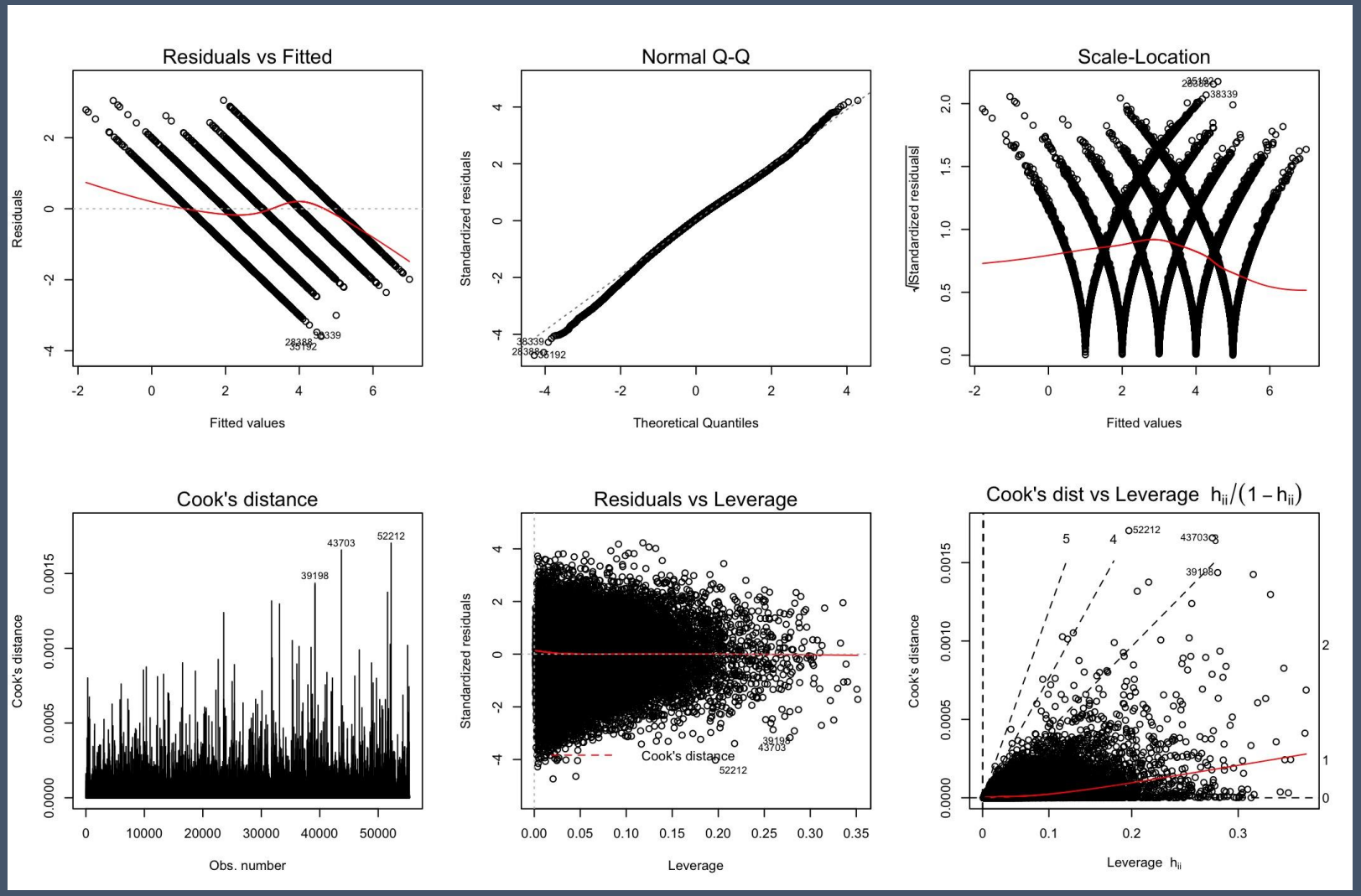
  F-value = 47.95

  P-value < 2.2e-16

- **Conclusion:**

  Reject H_0, at least one predictor should be useful in predicting the stars

Model Diagnosis



Reflection

- Strengths

1. Use frequency as criteria select predictors makes model more efficient
2. Log transformation greatly lower the error term of our model
3. Lasso regression remove predictors that have no use in model

- Weakness

1. Linearity assumption may not hold