-First we construct a linear regression model using all the variables in the dataset, from the result above, we can see that the coefficients of review\_count.x, yelp.votes.cool, emark are positive and significant, which means these features have a positive impact on ratings. The coefficients of yelp.votes.funny, yelp.votes.useful, wcount, qmark are negative and significant, which means these features have a negative impact on ratings. And we get a RMSE of 1.117211 in this model.

The explanation of stepAIC

The **stepwise regression** (or stepwise selection) consists of iteratively adding and removing predictors, in the predictive model, in order to find the subset of variables in the data set resulting in the best performing model, that is a model that lowers prediction error.

1. **Stepwise selection** (or sequential replacement), which is a combination of forward and backward selections. You start with no predictors, then sequentially add the most contributive predictors (like forward selection). After adding each new variable, remove any variables that no longer provide an improvement in the model fit (like backward selection). The default one

-To improve the performance of linear model, we use stepAIC method to select some of the variables in the dataset as independent variables, then we got the final formula as shown above. We get a RMSE of 1.11369 in this model. According to the result of this model, most of the coefficients are significant here. We will focus on the variables that have larger absolute value. And we will explain these variables one by one.

-$yelp.votes.funny$, the coefficient is significant and negative, which means if a review is voted as funny, it always the case that the content of the review is to make jokes about the restaurant, so the overall sentiment should be negative, such as the food tastes like shit. -4.37729

-$yelp.votes.useful$,the coefficient is significant and negative, which means if a review is useful to other customers, the review should be a negative review that remind other customers not to come to this restaurant again -9.80270

-$yelp.votes.cool$,the coefficient is significant and positive, which means if a review is voted as cool, this review is more likely to be a positive review. 17.20662

-$emark$, the coefficient is significant and positive, which means the more exclamation mark in a review, the higher the rating tend to be. 5.45072

-$qmark$, the coefficient is significant and negative, which means the more question mark in a review, the more likely it is a negative review. -4.04347

-$wcount$, the coefficient is significant and negative, which means nowadays people tend to write more words in a negative review to criticize the detail of a restaurant. -1.41087

The RMSE we got from this model is 1.11369, has improved a little compare to the simple linear regression using all the variables.

One of the major aspects of training your machine learning model is avoiding overfitting. *The model will have a low accuracy if it is overfitting.* This happens because your model is trying too hard to capture the noise in your training dataset. *By noise we mean the data points that don’t really represent the true properties of your data, but random chance*. Learning such data points, makes your model more flexible, at the risk of overfitting.

-Then, we fit three glmnet models. The most important variables are all the same with the linear models above.The best models of lasso, ridge and elastic net give us the RMSE of 1.116176,1.118149 and 1.116241 respectively.

We try several baseline models in this part. First we construct a linear regression model using all the variables in the dataset, To improve the performance of linear model, we use stepAIC method to select some of the variables in the dataset as independent variables, then we use the final formula to run the linear regression again. Then we find that One of the major aspects of training machine learning model is avoiding overfitting. The model will have a low accuracy if it is overfitting. This happens because our model is trying too hard to capture the noise in training dataset. To handle this problem, we fit three glmnet models using different penalty term, we also tune for the optimal parameters. Last, we also fit a knn model. The RMSE for these models are all shown here. We can see that the stepAIC perform best over all these baseline models.

Then we will focus on the interpretation of the important variables, from the result of all the baseline models, we find that the coefficient of some variables are always significant and have a larger absolute value than other variables, so we will just show the result from the stepAIC part to explain these variables in detail. The important variables are marked out in this slide.

-$yelp.votes.funny$, the coefficient is significant and negative, which means if a review is voted as funny, it always the case that the content of the review is to make jokes about the restaurant, so the overall sentiment should be negative, such as the food tastes like shit. -4.37729

-$yelp.votes.useful$,the coefficient is significant and negative, which means if a review is useful to other customers, the review should be a negative review that remind other customers not to come to this restaurant again -9.80270

-$yelp.votes.cool$,the coefficient is significant and positive, which means if a review is voted as cool, this review is more likely to be a positive review. 17.20662

-$emark$, the coefficient is significant and positive, which means the more exclamation mark in a review, the higher the rating tend to be. 5.45072

-$qmark$, the coefficient is significant and negative, which means the more question mark in a review, the more likely it is a negative review. -4.04347

-$wcount$, the coefficient is significant and negative, which means nowadays people tend to write more words in a negative review to criticize the detail of a restaurant. -1.41087