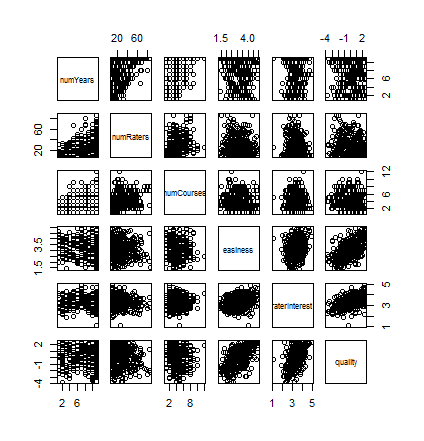
1. We want to analyze what which factors have the greatest impact on a professor's rating. We have a dataset that contains components of reviews of professors across many universities from ratemyprofessors.com. Using statistical modeling, we will be able to assess how much impact each of these variables has on the professor's rating. Also, we will be able to predict a professor's rating given new data.

2. To the right is a scatter-plot matrix of the numerical variables in this dataset. We want to check to see if there are any issues with collinearity that we should be aware of. Collinearity is an issue where two or more predictor variables are highly correlated with each other. This leads to inflated standard errors. When collinearity is present it is harder to detect significance, and predictions are thrown off.

It looks like there may be some collinearity going on with the number of years on the number of raters, so we will probably just include one of those in our model. Another way to detect collinearity is to use variation inflation factors. With these data, we get high variation inflation factors with the "discipline" variables, and a few of the "department" variables. This is probably because a teacher that is in the STEM discipline will also fall under one of the other department variables. So there is high correlation between those variables. This will lead us to not include overlapping departments and disciplines in our model.

3. We used a best subset selection technique to determine a multiple linear regression model that is appropriate to answer the questions at hand. We did best subset selection because we felt like it was the most exhaustive approach since there are not too many variables to make this a slow computation. We wanted to consider every possible model in our selection. For model comparison criterion, we used BIC, which will lend to a simpler model. We recognize that we were giving up about 2% of variation explained when using BIC over AIC, but we felt that the simplicity of the model when using BIC was more valuable to us.

4.

y­­i =0 + ­1I(Pepper=Yes)+ 2(Easiness) + 3­­(RaterInterest)+ β4­I(DisciplineSTEM=YES) + β5­I(DeptBusiness=Yes) + β6­I(DeptPhysics=Yes) + i

where εi ~ N(0,σ2)

yi = the quality rating for the ith professor

β0 = the average quality rating for a professor that has an easiness rating of 0 and has no pepper, is not in the STEM discipline, is not in the Business department, and is not in the physics department

β1 = the average change in quality rating as a professor goes from having no pepper to having a pepper

2 = the average change in quality rating as the easiness rating increases by 1

3 = the average change in quality rating as the rater's interest increases by 1

4 = the average change in quality rating as a professor goes from having not being in the STEM discipline to being in the STEM discipline

5 = the average change in quality rating as a professor goes from having not being in the Business department to being in the Business department

6 = the average change in quality rating as a professor goes from having not being in the Physics department to being in the Physics department

This model assumes linearity, independence, normality, and equal variance.

After fitting our model to the data, we will be able to assign a co-efficient to each of the factors in this model which will tell us how much of an effect each factor is having on the quality rating. This will help us understand what a professor must focus on if he or she wants to improve their quality rating. We will also be able to plug in amounts for each factor and predict the quality rating.

5.

