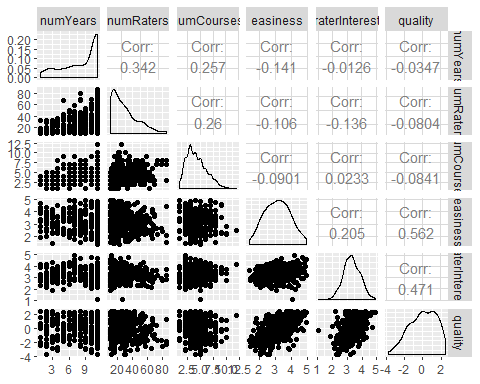
HW 6

Chris Trippe & Ethan Kemeny

1. In your own words, summarize the overarching problem and any specific questions that need to be answered using the rating dataset. Discuss how statistical modeling will be able to answer the posed questions.

In order to know what makes a good professor, we will make a regression model by examining nine different variables (number of years the professor has been teaching, number of courses taught by professor, is the professor "hot, the professor’s discipline, average easiness rating of the professor, etc.) from ratemyprofessor.com and determine which, if any, contribute to a professor’s overall quality.

1. Construct a scatterplot matrix of the data and assess any potential issues related to collinearity. Calculate variance inflation factors and discuss what variables, if any, are collinear. Comment on what affect collinearity can have on a regression analysis.



When referencing to the scatterplot matrix and associated correlation values, there doesn’t appear to be any collinearity or strong relationship between two explanitory variables. The largest correlation is 0.342 between number of years and number of raters, but we’ll consider than relationship insignificant enough. When calculating variance inflation factors (VIF), which works for qualitative and quantitative variables, values above ten are considered to be multicollinear. That means that variables that should be significant will not be considered significant because of their relationship with other explanitory variables. For this data set, disciplinePre, disciplineSocSci, disciplineSTEM, deptBiology, deptBusiness, and deptChemistry all have VIFs above ten.

gender numYears numRaters   
 1.196730 1.343714 1.506927   
 numCourses pepper easiness   
 1.483144 1.214964 1.379581   
 raterInterest disciplinePre.prof disciplineSocSci   
 1.462005 22.969870 12.224703   
 disciplineSTEM deptBiology deptBusiness   
 32.671820 11.181774 20.853459   
 deptChemistry deptCommunication deptCS   
 10.841703 6.800717 3.739853   
 deptEducation deptEnglish deptGeology   
 3.606844 4.718047 7.130056   
 deptHistory deptKins deptLanguages   
 3.011782 6.104904 3.287530   
 deptMath deptMusic deptPhilosophy   
 17.609677 1.715088 1.608090   
 deptPhysics deptPolySci deptPsychology   
 7.068442 3.539709 6.825622   
 deptReligion deptSocialScience deptSociology   
 1.672561 2.173193 4.077505

1. Use a variable selection technique to determine a MLR model that will answer the questions posed in #1 (do NOT consider any interactions). Justify your choice of your selected variable selection procedure (e.g., state why you chose to use backward instead of forward selection). Justify your choice of a model comparison criterion (e.g. state why you chose to base your variable selection procedure on AIC vs. BIC).

We will be using this model to make predictions which means AIC or Akaike Information Criterion, which is less restrictive on the variables allowed in “best” model, is better for making predictions than BIC or Bayesian Informaiton Criterion. The “exhaustive” selection method was also used because it finds the best model for each length one through the total number of variables based on their AIC values, and then of those models, the best is picked. That means every possible model is tested and the one with the best AIC value is chosen.

1. Write out (in mathematical form with greek letters) your selected MLR model from #3. Clearly state any assumptions you are using in your model. Provide an interpretation of each mathematical term (variable or parameter) included in your model. Using the mathematical form, discuss how your model, after fitting it to the data, will be able to answer the questions in this problem.

When using an exhaustive AIC variable selection test, the following model is built:

By using this model, we assume that the model is linear, the values of quality are independent, the residuals are normally distributed about the regression line, and the residuals have equal variance about the line.

: the intercept; when all categorical variables are “no” and all quantitative variables are 0, the value of quality is , on average.

: Holding all other explanitory variables constant, as the number of years increases by one, quality increases by , on average.

: Holding all other explanitory variables constant, as the number of courses increases by one, quality increases by , on average.

: Holding all other explanitory variables constant, as a review changes from “no” chili pepper to “yes” chili pepper, quality increases by , on average.

: Holding all other explanitory variables constant, as easiness increases by one, quality increases by , on average.

: Holding all other explanitory variables constant, as rater interest increases by one, quality increases by , on average.

: Holding all other explanitory variables constant, as a review changes from “no” discipline STEM to “yes” discipline STEM, quality increases by , on average.

: Holding all other explanitory variables constant, as a review changes from “no” department Business to “yes” department Business, quality increases by , on average.

: Holding all other explanitory variables constant, as a review changes from “no” department English to “yes” department English, quality increases by , on average.

: Holding all other explanitory variables constant, as a review changes from “no” department Geology to “yes” department Geology, quality increases by , on average.

: Holding all other explanitory variables constant, as a review changes from “no” department Longuage to “yes” department Language, quality increases by , on average.

: Holding all other explanitory variables constant, as a review changes from “no” department Physics to “yes” department Physics, quality increases by , on average.

: Holding all other explanitory variables constant, as a review changes from “no” department PolySci to “yes” department PolySci, quality increases by , on average.

: Holding all other explanitory variables constant, as a review changes from “no” department Psychology to “yes” department Psychology, quality increases by , on average.

In order to use this model, the estimated quality value is equal to any of the quantivative values multiplied by their beta or slope value plus the categorical values beta values when that categarical value is “yes”.

1. Fit your model in #3 to the professor rating data and summarize the results by displaying the estimated coefficients in a table with 95% confidence intervals for each parameter. Provide an example of how to interpret 1 interval for a quantitative explanatory variable and 1 interval for a categorical explanatory variable correctly in the context of the problem.

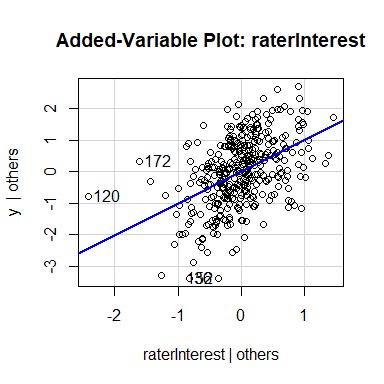
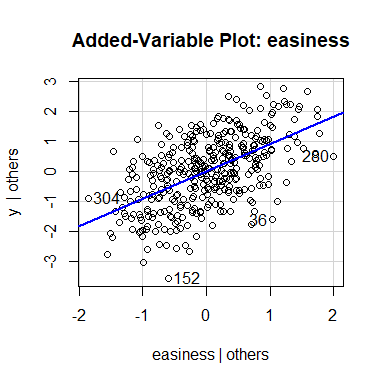
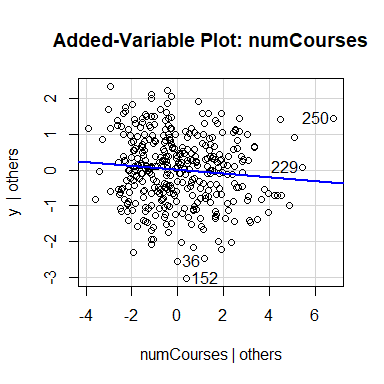
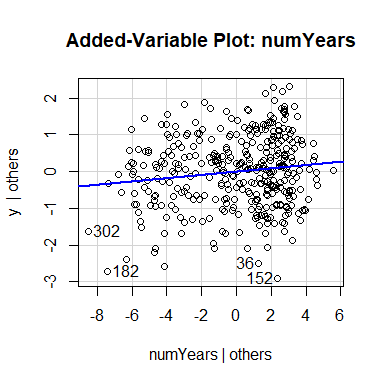
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Symbol | Estimate | Lower Bound | Upper Bound |
| (Intercept) |  | -6.8927151 | -7.7525968 | -6.0328334 |
| numYears |  | 0.0436752 | 0.009359 | 0.0779913 |
| numCourses |  | -0.0527503 | -0.1094181 | 0.0039175 |
| pepperyes |  | 0.9454868 | 0.617937 | 1.2730366 |
| easiness |  | 0.9089336 | 0.7685679 | 1.0492993 |
| raterInterest |  | 1.0074449 | 0.8063025 | 1.2085873 |
| disciplineSTEMYes |  | 0.6530968 | 0.3621027 | 0.9440908 |
| deptBusinessYes |  | 0.6989131 | 0.3444793 | 1.0533469 |
| deptEnglishYes |  | 0.5135504 | 0.1689766 | 0.8581242 |
| deptGeologyYes |  | 0.4527604 | -0.0814628 | 0.9869836 |
| deptLanguagesYes |  | 0.3381718 | -0.0760327 | 0.7523763 |
| deptPhysicsYes |  | 0.9328262 | 0.3349629 | 1.5306894 |
| deptPolySciYes |  | 0.5948989 | -0.1260427 | 1.3158405 |
| deptPsychologyYes |  | 0.4757963 | 0.0018108 | 0.9497818 |

: the intercept; when all categorical variables are “no” and all quantitative variables are 0, we are 95% confident that the value of quality is between -7.7525968 and -6.0328334, on average.

: Holding all other explanitory variables constant, as the number of years increases by one, we are 95% confident that quality increases by between 0.009359 and 0.0779913, on average.

: Holding all other explanitory variables constant, as a review changes from “no” chili pepper to “yes” chili pepper, we are 95% confident that quality increases by between 0.617937 and 1.2730366, on average.

1. Assess the fit of your model and justify your model assumptions using appropriate graphics or summary statistics. Provide discussion of your assessment of the model fit and assumptions on the level of your target audience (e.g. interpret your model R2).



One-sample Kolmogorov-Smirnov test  
  
data: stdres(slr\_best)  
D = 0.031783, p-value = 0.8535  
alternative hypothesis: two-sided

studentized Breusch-Pagan test  
  
data: slr\_best  
BP = 17.766, df = 13, p-value = 0.1666

By looking at the AV Plots, we can see that the model is approximately linear. A line seems to be a reasonable model for the data.

As for normality, the Kolmogorov\_Smirnov test, which tests for normality, has a p-value of 0.8535, which means we fail to reject the null hypothesis and conclude that the residuals are normally distributed.

When testing for equal variance, we used the Breusch-Pagan test and obtain a value of 0.1666 which means fail to reject the null hypothesis and conculde that the residuals have equal variance.

Lastly, we must consider independence of the values of quality each professor receives. One could argue that some students might rate more leniently than others, and if those students take a certain class form a professor in one department, they potentially will not take a the classes from a professor in the same department that teaches similar classes. This, however, and other arguments seem like a stretch and we can assume the effect of one professor’s quality rating to be negligable on another professor’s quality rating. Therefore, we can assume independence.

Using a cross-validation study of 500 samples we are able to get a sense of how accurate our model is at proffesor ratings. The bias is -0.0082881. This means our model typically predicts life expectancy -0.0082881 above the actual value on average. That value is relatively small so we feel our model isn’t very biased. The root prediction mean squared error value of 0.990053 tells us our predictions miss the mark by an average of 0.990053. Even with the response variables being smaller values, a value of 0.990053 is about 16% of the overal range of the response variable. This is a little on the high side for margin of error, but we feel that our model is accurate enough in its predictions.

We also wanted to look at our prediction intervals or the low and high end of our estimated values. The study returned a coverage of 0.9532222 which is the percentage of prediction intervals that contain the actual value. We have a width of 3.9469844, we can conclude that our prediction intervals are about half the size of the response range. This means our model will get us close to the proffesor’s quality but the width is fairly large given the scale of the data.

The for this model is 0.5562445 this means our model of years, number of courses, pepper, easiness, rater interest, stem discipline, and the selected departments can explain 55.6244522% of the variation in proffesor quality. This value isn’t great but it is considerable. We feel that our model does an okay job at fitting the data with only about 45% of the variation in rating left unexplained

1. Provide a discussion to an incoming BYU professor about important qualities of a good professor (according to rate my professor anway). As an example, use your model to predict (and show a corresponding 95% prediction interval) Dr. Heaton’s quality rating score given: numYears= 5, numRaters= 54, numCourses= 2, pepper=no (§ - at least my wife thinks I’m cute) , discipline=STEM, dept=Math, easiness= 2.5, raterInterest= 4.2.

What makes a good professor according to our data and model is their experience, if they are attractive, how easy they are, what department they are in and how interested the students who rate the proffesors are. The number of courses seems to hurt how good a proffessor is according to their students. THis makes sense as the teacher might be constrained time wise and in other ways due to their busyness with other classes and the students might feel a lack of availability from the proffesor.

Using our model we predict that Dr. Heaton would recieve a rating of 0.3768594. This is positive so on average we would expect students to think he is a decent proffesor. However, the prediction interval shows that his true rating could be as low as -1.5873576 or as high as 2.3410765. This interval tells us that his rating could be slightly negative or it could be close to the highest rating we have data of. With this in mind, we would say that Dr. Heaton would be considered a “good” proffesor on average by his students.