

Homework 4

Due April 12, 2011

Directions Please turn-in a hard copy of your R code along with a brief write-up of the solutions (do not submit raw output). Also submit via e-mail (njc23@pitt.edu) a copy of your R code.

The `attitude` dataset in R contains the results of a survey of clerical employees for a large financial organization. Questionnaires were completed by employees from 30 randomly selected departments. The numbers in the dataset give the percent of favorable responses to seven questions for each department. The seven questions dealt with,

Variable Name	Description
<code>rating</code>	Overall rating (response)
<code>complaints</code>	Handling of employee complaints
<code>privileges</code>	Does not allow special privileges
<code>learning</code>	Opportunity to learn
<code>raises</code>	Raises based on performance
<code>critical</code>	Too critical
<code>advance</code>	Advancement

The objective of the study is to determine the relationship between the overall rating and the six remaining variables. Analyze the `attitude` dataset by answering the following questions.

1. Make the `complaints` variable a discrete variable with three levels, "bad", "okay" and "good." Where 0-60 is "bad", "61-80" is "okay" and "81-100" is good. For the remaining questions use this categorical complaints variable, not the continuous version.
2. Calculate the average overall rating for each level of complaints.
3. Create at least one descriptive plot of the data that includes a lowess smoother.
4. Determine the "best" linear regression model using AIC, that is find the linear model that gives the smallest value for AIC. The response is the overall rating (`rating`) and the remaining six variables are all predictors. Consider all six main effects and all possible two-way interactions when selecting the "best" model.
5. For the best model in question 4, give the estimated regression coefficients with 95% confidence intervals and test the overall effect of `complaints`.
6. Verify the linear model assumptions by generating six different diagnostic plots. Are there any major violations of the assumptions?

7. Using `influence.measures()` find the points that R identifies as being influential. For these influential points print the values of the influential measures and the original data from `attitude`.
8. Refit the best model from question 4 without the influential points identified in question 7. Is there any major change in the model?
9. What is the predicted overall rating and corresponding prediction interval for the following two departments (use the model from question 4)?

	Complaints	Privileges	Learning	Raises	Critical	advance
Department 1	71	76	98	81	82	99
Department 2	92	96	61	96	78	78