Statistics 512: Final Project

Instructions

- Each group will find a dataset containing at least four predictors, and check with me to see if dataset is suitable for final project by Friday, March 31, 2017.
- All statistical analyses must be performed in SAS.
- Every member is expected to contribute to the analysis of the dataset, written report and oral assessment.
- Presentations and oral assessment are scheduled during the week of April 24-28, 2017. Every member will be given at most 3 minutes to present their work plus one minute for a question. I will stop you if you go beyond 3 minutes. Make ONE PowerPoint presentation per group, you should organize the slides so that the transition is smooth within each group. To optimize time and have smooth transition between presentations, please have your PowerPoint files ready, I will load all presentations on the desktop at the beginning of class.
- The deadline to submit the written report is Friday, April 28, 2017 at 3 pm; please submit one final report per group. No late submission will be accepted under any circumstances. Hard copies only.
- The deadline to submit the peer evaluation form (which has been posted on the course website) is also Friday, April 28, 2017 at 3 pm. Students who don't turn in the peer evaluation form won't receive any points (0%) even if their peers give them points.
- Please address the following questions in the final report and presentation.

Questions

1. Using techniques you learned in class, determine whether the response variable and any of the predictors need to be transformed. Indicate the reasoning for your decision. If a variable needs to be transformed, transform it and keep it in the full model for the rest of the questions.
2. Use the C_p criterion to select the best subset of variables for your data (i.e. use the options "`/ selection = cp b;`"). Summarize the results and explain your choice of the best model.
3. Use the stepwise option to report the best subset of variables for your data (i.e. use the options "`/ selection = stepwise;`"). Summarize the results and explain your choice of the best model.
4. Use either one of the best models you chose in questions 2 and 3 as a final "best" model. Check the assumptions of this final "best" model using all the usual plots (you know what they are by now). Explain in detail whether or not each assumption appears to be substantially violated.

5. Use the "best" model to predict the response variable. Examine other diagnostics such as (but not necessarily exclusively) studentized and studentized-deleted residuals, Cook's D, hat matrix diagonals, tolerance or vif, and partial residual plots. Explain any problems such as outliers, highly influential observations or multicollinearity that these diagnostics point out. (Do not include in your output any tables of values for all observations. Use plots and verbal summaries instead. You may include values for a few selected individuals if you wish.)

6. For the "best" model report the following:

   (a) Equation of the regression model.

   (b) 90% confidence interval for the mean of the response variable.

   (c) 90% prediction interval for individual observations.

   (d) 90% confidence intervals for the regression coefficients.

Grading

Evaluation of the final project is based on the following components:

- Written Report (8%), I will grade the reasoning and accuracy of addressing the 6 questions.
- Oral Assessment and Presentations (9%). 3% for presentation and 6% for answering the question.
- Peer Evaluation (8%); each member of the group will have to evaluate all other members.