36-617: Applied Linear Models
Fall 2019
HW06 – Due Mon Oct 12, 11:59pm

- Please turn the homework in online as a pdf to gradescope using the link in the hw05 page in our course webspace at canvas.cmu.edu.
- Please finish reading Chapter 7 of Scheather this week. For next week, read Chapter 8, on logistic regression.

Exercises

1. For this exercise we will consider some of the data analyses needed for Project 01. That way, you can use your work here as part of the technical appendix for the paper for Project 01.

   Consider the data file `maskdata.csv` in the `Project 01` folder in our files area on Canvas. The variables are defined in Table 1; a more complete description of the data can be found in the `project-01.pdf` assignment sheet.

Variable Name	Values	Description
study	"cold" or "flu"	The "flu" study ran from 2013 to 2014; the "cold" study ran from 2014 to 2016.
age	integer	Age of subjects, ranging from 12 to 84
female	0 or 1	1 = female
allo.wm	0 or 1	random allocation of subject to "mask" or "nomask" group; 1 = "mask"
mask.class	"ctrl", "mask", or "both"	"ctrl" = no mask, "mask" = wore mask; "both" = was included in both "ctrl" and "mask" samples
breath.num	integer	number of exhaled breaths collected (to measure viral load)
cough.num	integer	number of coughs during exhaled breath collection
virus	"hcov", "flu", or "cold"	type of virus measured: hcov = coronavirus, flu = influenza virus, cold = rhinovirus
nasal	real number	viral load (copies per ml) in nasal swab before experiment
throat	real number	viral load (copies per ml) in throat swab before experiment
aerosols	real number	viral load (copies per ml) in exhaled aerosols
droplets	real number	viral load (copies per ml) in exhaled droplets

Table 1: Variables in the file `maskdata.csv`. A value of 2 was used for the viral load when the amount of virus was too close to zero to be measured. NA's (missing data) were recorded whenever a measurement was not made for some reason, or the measurement was lost.

   (a) Data description.

   - Make a table or tables showing appropriate summary statistics for each variable in the data set. Note that summary statistics for continuous variables will be different from the summary statistics for categorical variables.

- Indicate where (in which variables) there is missing data (NA's), how much there is (in each variable) and why it might be there.
- Make some appropriate descriptive EDA plots to illustrate any important features of the variables.

(b) Use linear regression to answer the following two questions: *In this study, does wearing a mask make a difference? Is the answer different for aerosols, than it is for droplets?* Show the fitted models and explain your answer to these questions in terms of the results of fitting those models.

(c) Use linear regression to answer the following questions: *Is there a difference between coronavirus, influenza virus, and rhinovirus, in the efficacy of masks? Is the answer different for aerosols, than it is for droplets?* Show the fitted models and explain your answer to these questions in terms of the results of fitting those models.

(d) Find the multiple regression model that makes the best tradeoff between the following criteria:

- Best fit, as measured by *adjusted* R^2, AIC, or BIC.
- Best satisfies the assumptions of the linear model.
- Best for interpreting and explaining to a medical collaborator or client.

Feel free to use transformations, interactions, etc., to make the model as good a tradeoff as you can between these three criteria. If it seems to work better, also feel free to make separate models for each of the three virus types.

No matter what you do, you are likely to be unhappy with some or all of these criteria; the better you make one criterion, the worse another is likely to get. So you will have to find a compromise or tradeoff between the three criteria. Explain how you decided to make the tradeoff(s) you made.

2. Return again to the `beauty` data that you have worked on for several assignments. For this exercise, use the transformed variables that you found for HW05.

(a) Use the "all subsets" method to choose the best model for `coursevaluation` (or a transformation of it, if you found a good transformation for HW05), considering all the other variables (with whatever transformations you found for HW05) except for —tt profevaluation, `profnumber.multipleclass` and the 30 class variables (`class1` through `class30`.

(b) Repeat part (2a) using Stepwise Regression instead.

(c) Repeat part (2a) using the lasso instead.

(d) *Briefly* compare the models in parts (2a), (2b) and (2c), with the model you obtained for HW05, part (2c): Make a table showing which variables remain in the final model for each of the four methods, and then write a sentence or two saying which model or models makes the most sense to you, based on this table.