Lab 4: Modeling predictors of lung health

Create a short reproducible report answering the questions below. The report should be less than 3 pages, including all figures. You do not need to show your code in the PDF report.

This lab is due at 5pm on Friday, October 12th. You should submit your assignment, in the form of both a knitted RMarkdown PDF as well as the .Rmd file that created the PDF, by uploading them to your personal Google Drive folder that is shared with the TA and the instructor. While you may collaborate with other students on this assignment, you must write up your own code and answers to the questions. Absolutely no cutting and pasting of any portion of the answers. This assignment, like the others, will be worth 50 points.

Data exploration

This lab will cover model selection and checking for multiple linear regression using the FEV dataset (found in the Vanderbilt Datasets repository). This dataset contains 654 observations on 6 variables, including a continuous outcome variable, forced expiratory volume (FEV) which is a measure of lung health and strength, age, height, sex, and smoking status. To begin, load (install if necessary) the Hmisc package to load the dataset from the Vanderbilt website. The brief FEV dataset codebook can be found on this webpage.

```
library("Hmisc")
getHdata(FEV)
```

Exercise 1 (5 pts) Before looking at the data, what types of relationships do you expect to see between each of these variables and FEV? Justify your answers briefly.

Exercise 2 (5 pts) Think about the different measures variables that you have at your disposal. Hypothesize at least two possible interactions, and come up with a plausible justification about why such an ineraction might exist.

Exercise 3 (10 pts) Generate a few simple plots of the data to evaluate the possible bivariate relationships that exist. Based on these plots, do you think that linear relationships will be sufficient to explain the data?

Recall (from lecture 6) that when you have a large sample size relative to the number of possible covariates (which we do), one justifiable way to build a model is to

- Include key covariates of interest.
- Include covariates needed because they might be confounders.
- Include covariates that your colleagues/reviewers/collaborators will demand be included for face validity.
- Test a reasonable/limited number of interaction terms if it seems appropriate.

Exercise 4 (15 pts) Before fitting any models, identify somewhere between 2 and 6 multiple linear regression models that you think follow the rules above, i.e. write down what each model is, and which covariates and/or interactions it includes. Fit each model. Compare the coefficients from each of the models as well as the adjusted R^2 values. For the 2-3 models with the best adjusted R^2 values, look at residual plots. Do you see any worrisome patterns that suggest the model assumptions aren't met? If so, do something to try to improve it.

Exercise 5 (5 pts) Based on your work above, pick one model that represents what you think is the "best" model for this data. Justify your choice, and interpret what you think are the most important

or interesting relationships with the covariates and FEV. Illustrate one of these relationships with a figure.

Exercise 6 (5 pts) Based on your thinking about this dataset and correlates of lung health more generally, if you were to collect similar data like this in a new study, what additional variables would you be interested in collecting and why?

Exercise 7 (5 pts) Ensure that your entire report is reproducible. That means that the instructors should be able to run your .Rmd file and have it generate the same PDF that you handed in. Note that you can assume that we have the Hmisc package installed on our machine but you should not rely on us having any particular filepath or directory structure when running the file. (In other words, don't try to load the data from a local file whose path is specific to your computer.)