Lab 4: Model checking and selection in multiple linear regression

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).

Exercise 1 Load (install if necessary) the Hmisc package to load the dataset from the Vanderbilt website.

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

The brief FEV dataset codebook can be found on this webpage.

Exercise 2 Generate a few simple plots of the data to evaluate the possible bivariate relationships that exist. Based on these plots, do you think any data transformations or curvilinear relationships (i.e. polynomial regression) should be explored?

Exercise 3 Define a class of models that you will explore. What predictor variables will you use? What combinations will you evaluate? What, if any, transformations or "curvilinearities" will you explore? What metric will you use to pick the best model out of this subset of models?

Exercise 4 Fit each model in the class defined above, report the decision metric for each model in a table and identify the "best" model. Do NOT use a forward or backward selection procedure.

Exercise 5 Identifying the "best" model is just the first step. Now, evaluatethe residual vs. fitted value plot and the qqplot of standardized residuals. Are any adjustments needed, or does the fitted model seem to meet the assumptions of multiple linear regression? If any assumptions appear to be violated, try to diagnose the source of the problem, fix it, and generate a new fitted model.

Exercise 6 Summarize your model selection procedure and the results in a few sentences. Interpret the coefficients from the final model.

Exercise 7 For the chosen model, calculate Cook's distance without using the `cooks.distance()` function. Plot the Cook's distance values. Do any points look worrisomely influential?

Exercise 8 Add a new, synthetic observation to the dataset that has influence on the fit. Try to balance making a datapoint that isn't *too* unrealistic (i.e. isn't too much of an oulier in either y or x directions) given all of the other observations, but that has a Cook's Distance value at least twice as much as the highest value of the original datapoints. Experiment a little bit with making new data points. How much of an outlier would a datapoint have to be to have a Cook's Distance value greater than 1?

For this lab, you should submit a single PDF file that summarizes your findings and answers to these questions.