Extra Practice for Exam1

1. Return the Body Fat data from HW2. Recall that the response variable is BodyFat.
   A. Start by fitting the full model, including all 3 predictors (Triceps, Thigh, Midarm). Use the `vif()` function from `car` to find the VIF values. Verify the VIF value for Triceps using a hand calculation.
   B. Now refit the model, including just Triceps and Midarm. (This is the model chosen by backwards elimination.) Now recheck the VIF values.
   C. Using the model including just Triceps and Midarm, identify the observation with the largest Cook's distance. Considering DFBETAS, DFFITS and Cook's D (and using the various rules of thumb), is this observation influential?

2. The data Mortality.csv reports mortality in 60 U.S. cities, along with various environmental and background variables, and three pollution variables (HC, NOX, SO2). The <u>three pollution variables</u> should be <u>log transformed for all analyses</u>.

   MORTALITY - mortality rate per 100,000
   PRECIP - average annual precipitation in inches
   HUMIDITY - average annual humidity
   JANTEMP - average January temperature
   JULYTEMP - average July temperatures
   OVER65- percent of population over age 65
   HOUSE - average population per household
   EDUC - median educational attainment in years
   SOUND - percentage of housing that was judged to be sound
   DENSITY - population density per square mile
   NONWHITE - percent non-white
   WHITECOL – Percent employed in white-collar occupations
   POOR - percent below the poverty line
   HC - relative pollution potential of hydrocarbons
   NOX - relative pollution potential of nitrogen oxide
   SO2 - relative pollution potential of sulphur dioxides

The primary interest is in the effect of the pollution variables (HC, NOX, and SO2) on mortality, adjusting for the climate and demographic variables. The strategy that the researchers followed is to first fit a model that predicts mortality as a function of the climate and demographic variables, and then consider adding the three pollution variables to the previously selected model.

   A. Look at the correlation between the pollution variables (on the **log transformed scale**) and mortality.
   B. Select a multiple regression model for the mortality as a function of the background and climate variables (PRECIP, HUMIDITY, JANTEMP, JULYTEMP, OVER65, HOUSE, EDUC, SOUND, DENSITY, NONWHITE, WHITECOL, POOR). Use `dredge()` from `MuMin` to choose a model based on AIC best subsets selection.
   C. Add the pollutions variables, HC, NOX, and SO2 (**transformed to the log scale**) to the model selected above. Use best subsets selection to select a final model

with the restriction that all models considered should include the predictors selected in (B). This can be done with the “fixed” option.

D. Find the VIF values corresponding to the “final” model.
E. Interpret the parameters associated the pollution variables in the “final” model.