Multiple Regressions

We have already calculated simple linear regressions (i.e., one X to predict Y) and AICs. Today we extend to multiple *quantitative* predictors, and evaluate alternative models with AIC. We set aside ANCOVAs that also use categorical predictors today for simplicity and to emphasize *collinearity* and *scaling* among quantitative variables.

Get the data set:

1. We use again the `cars93` data set in the `MASS` package. If `MASS` is already installed, then simply load it.
2. Also load `car` (Companion to Applied Regression – not to be confused with the Cars93 data set).
3. Because it comes with a package, we load Cars93 differently than if when we import a txt file:

```
data(Cars93)
attach(Cars93)
View(Cars93)
```

Your mission today: make the most plausible model you can to predict gas mileage (MPG.city). Below is an example – run through this first, and then use it as a template to proceed with your models.

Model 1: I think MPG.city is predicted by Price and RPM.

```
model1 <- lm(MPG.city ~ Price + EngineSize)
summary(model1)
```

Model 2: I think MPG.city is predicted by Weight and Passengers.

```
model2 <- lm(MPG.city ~ Weight + Passengers)
summary(model2)
```

BUT: These predictor variables have very different units and ranges. Because coefficients of a model are multiplied by the units (e.g., the a in $Y = aX + b$), it is then hard to compare coefficients for importance if they are on different scales. So we adjust model variables by computing a Z-score for each variable, so that they are now all in units of standard deviations. This nicely makes all terms comparable but retains relationships. So instead do this:

```
model1s <- lm(MPG.city ~ scale(Price) + scale(EngineSize))
summary(model1s)

model2s <- lm(MPG.city ~ scale(Weight) + scale(Passengers))
summary(model2s)
```

Notice that t- and p-values do not change, but *now coefficients are different* and can now be fairly compared. You can now say that Engine Size has three-fold the effect of Price.

ANOTHER BUT!: The assumption of independence among data is violated if variables are too correlated (a problem called *multicollinearity)*. Closely correlated variables are redundant and artificially ***inflate*** the explained variance of each other. We measure this with a Variable Inflation Factor (VIF), where the rule of thumb = ***values > 10 are too correlated*** and you should omit one from subsequent models. You must choose which one to omit. So run this command (in the `car` package):

```
vif(model1s)
vif(model2s)
```

Notice that you must first have a model to compute VIF scores!

A THIRD BUT!: How do you first choose potential variables to make models? You can see a correlation matrix of quantitative variables, and/or a grid of scatter plots (remember those?):

```
require(dplyr) # for the select_if command
quantCars93 <- select_if(Cars93, is.numeric) # keeps numeric columns
cor(quantCars93) # a correlation matrix of variables in quantCars93
```

Got the basic approach?
1. Make models that represent hypotheses, using scaled predictors
2. Evaluate collinearity & trim out redundant predictors
3. Compare alternative models with AICc
4. Evaluate residuals of your most plausible model(s).
5. Finally, examine the ***Adjusted*** R^2 and coefficients of the most plausible model.

Your mission: develop the most plausible and predictive model for MPG.city, using the template above.