Collinearity in Regression

In this exercise we will examine some of the ways of detecting and dealing with problems of collinearity – high correlations among the predictors in regression models. (The concepts involved here will be explained in the lecture.)

We will use a data set, cars, containing 406 observations on the following 8 variables: MPG (miles per gallon), # cylinders, engine displacement (cu. inches), horsepower, vehicle weight (lbs.), time to accelerate from 0 to 60 mph (sec.), model year (modulo 100), and origin of car (American, European, Japanese). The first few observations are shown below.

```
      make      model mpg cylinder engine horse weight accel year origin
1     chev   chevelle  18        8    307   130   3504  12.0   70      A
2    buick    skylark  15        8    350   165   3693  11.5   70      A
3 plymouth satellite   18        8    318   150   3436  11.0   70      A
4      amc      rebel  16        8    304   150   3433  12.0   70      A
5     ford     torino  17        8    302   140   3449  10.5   70      A
6     ford    galaxie  15        8    429   198   4341  10.0   70      A
```

Here, we are interested in predicting gas mileage, MPG from the other quantitative variables (excluding ORIGIN, which is categorical). The steps in this exercise are a subset of those included in the script, `http://friendly.apps01.yorku.ca/psy6140/tutorials/cars-collin.R`. You can open this in a browser or notepad window if you prefer.

1. The data is stored on my server. Read it into R using

```
cars <-
read.csv("http://friendly.apps01.yorku.ca/psy6140/tutorials/cars.csv",
             header=TRUE)
str(cars)
```

2. Need I say this again? As always, it is useful to get an overview of the data by plotting. One simple way is a scatterplot matrix. We will also take a look at the correlation matrix, and visualize it using a ‘corrgram’.

```
library(car)
scatterplotMatrix(~ mpg + weight + year + engine + horse + accel +
      cylinder,      data=cars, smooth=FALSE)

cor(cars[,3:9], use="complete")
# visualize correlations
library(corrgram)
corrgram(cars[,3:9], upper.panel=panel.cor, diag.panel=panel.density)
```

Examine the scatterplot matrix and correlation matrix. The focus here should be on predictors that seem to be highly correlated.

3. Start with a regression model using all quantitative predictors.. What do you observe about the *t* statistics for the individual predictors?

```
# fit model with all numeric predictors
cars.mod <- lm(mpg ~ weight + year + engine + horse + accel + cylinder,
                   data=cars)
summary(cars.mod)
```

4. Let’s ask for variance inflation factors (`car::vif()`) and collinearity diagnostics (`perturb::colldiag()`). You may need to install the `perturb` package first. As a rough rule, VIF values > 10 signal a predictor that is highly correlated with the other predictors.

```
library(car)       # for vif
# maybe need to install this first
if(!require("perturb")) {install.packages("perturb"); library(perturb)}
# colinearity diagnostics
vif(cars.mod)
colldiag(cars.mod, center=TRUE)
```

5. One way to deal with the problem of highly correlated predictors is to use ideas of model selection to remove possibly redundant variables. This should be done based in part on substantive grounds, so here we will remove the variables that are necessarily related to ENGINE size.

```
# Remove variables strongly related to engine size
cars.mod2 <- lm(mpg ~ weight + year + engine, data=cars)
vif(cars.mod2)
```

6. Finally, another idea is to transform the predictors to principal components (which are always uncorrelated) and use those as predictors of `mpg`. Here, `prcomp` produces an output data set containing new variables PC1 – PC6.

```
# PCA of predictors
OK <- complete.cases(cars)
cars.pca <- prcomp(cars[OK, 4:9])
cars.pca

# PCA regression: use uncorrelated components as predictors
Cars.merged <- cbind(cars[OK,], cars.pca$x)
cars.pcareg <- lm( mpg ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6,
                   data=cars.merged )
summary(cars.pcareg)
vif(cars.pcareg)
```

Compare the R^2 for this model with your first model that uses all the predictors. (They should be the same. Why?) Compare the VIF values for the model using the principal component scores with those for the original predictors.