Regression Diagnostics with R

The purpose of this exercise is to consider some of the aspects of regression diagnostics for examining the adequacy of multiple regression models. The data we will use come from a study by Duncan (1961), examining the relation between Prestige of occupational categories and the average Income and Education in those jobs. (A more complete analysis might include the type of occupation, but we ignore this here.) The first few cases are:

```
           type income education prestige
accountant prof     62        86       82
pilot      prof     72        76       83
architect  prof     75        92       90
author     prof     55        90       76
chemist    prof     64        86       90
minister   prof     21        84       87
```

The script used below is contained in **N:\psy6140\tutorials\duncan-reg.R**

1. The data is in the car package. Read it in with
   ```
   data(Duncan, package="car")
   ```

2. As always, it is useful to get an overview of the data by plotting. We should look for indications that relationships are seriously nonlinear or other oddities.
   ```
   library(car)
   scatterplotMatrix(~prestige + income + education,
           data=Duncan, id.n=2)
   ```

3. Proceed to fit a model using both income and education as predictors. What would you conclude from this?

   ```
   duncan.mod <- lm(prestige ~ income + education, data=Duncan)
   summary(duncan.mod)
   ```

4. Examine the four plots produced by `plot(model)`. What do they indicate about possible problems with the model?

   ```
   plot(duncan.mod)
   ```

5. Other plots of residuals are useful too. In well-behaved data, they should all look unstructured, with no systematic patterns.
   ```
   residualPlots(duncan.mod, id.n=2)
   ```

6. For today, we are interested mainly in whether there are any highly influential observations in the data – those that could change the regression coefficients depending on whether they were included or omitted. Probably the single most useful plot you can make is an "influence plot" that shows the residual and leverage, with bubbles proportional to a measure of total influence – Cook's D statistic.

   ```
   influencePlot(duncan.mod, id.n=2)
   ```

7. Finally, partial residual plots (aka added variable plots) are very useful, because they show the **unique** partial relationship of Y to each X, controlling (or adjusting) for all other predictors.

```
avPlots(duncan.mod, id.n=2)
# fancier version
avPlots(duncan.mod, id.n=2,
   ellipse=TRUE, ellipse.args=list(levels=0.68, fill=TRUE))
```

Several observations stand out in both the influence plot and in the partial residual plots, and were identified in the plots by their labels (`id.n=`). Try to think why these might be unusual in terms of the context of this data.

You can also view the relationships in 3D using scatter3d().

```
scatter3d(prestige ~ income + education, data=Duncan, id.n=2)
```