**Lab 2: Advanced Diagnostics**

The following PDF explains the use of most functions in this lab:

[www.sagepub.com/upm-data/38503\_Chapter6.pdf](http://www.sagepub.com/upm-data/38503_Chapter6.pdf)

**To submit your work, insert screenshots of your code and outputs (both numeric outputs and graphs) under respective problem prompts. Many steps also require a written answer, and you should insert your written or typed answer below the prompt.**

1. If you haven’t installed the **car** package before, run the first line below to install it. After you have installed a package, run the second line below to load the package each time you start a new R session.   
     
   install.packages("car")  
   library(car)
2. The dataset “Duncan” is provded by the **car** package. After you load the package, you can run **data(Duncan)** to load the dataset into your workspace. Then use **head(Duncan)** to take a look at the first a few rows of the dataset, and use **str(Duncan)** to inspect the structure of the dataset.
3. **Inspect each variable.** Before running regression analysis, always first understand and inspect each individual variable. For each variable in the dataset:  
   1. Read the dataset manual, which is obtained by **help(Duncan)**, to understand the nature of the variable.
   2. Identify whether it is a numeric or a categorical variable.
   3. For a categorical variable, use the **table( )** function to inspect the frequency table of the variable.
   4. For a numeric variable, create a histogram of the variable using **hist( )**.
   5. For each variable, discussion whether the data distribution look reasonable. Are there any signs of possible data errors?
4. **Inspect bivariate relationships.** Run **pairs(Duncan)** to inspect the bivariate relationships.
   1. In this data, *prestige* is the response variable (a.k.a., dependent variable) to be predicted. Inspect the row/column for *prestige* and describe your impression of its relationship with each predictor variable.
   2. The rest of the scatterplot matrix shows the bivariate relationships among the predictors themselves. Describe your impression of those scatterplots and how they imply multicollinearity.
5. **Fit the model.** Run to code below to fit a multiple regression model with *prestige* as the response variable, and *education*, *income*, and *type* as the predictors. Then generate the model summary. Write out the fitted model.  
     
   m <- lm(prestige ~ education + income + type, data=Duncan)  
   summary(m)
6. **Check multicollinearity.** Use **vif(m)** to generate VIF values for each predictor. Read the results in the *last column*. Report and interpret the results.   
     
   (Note: for regular continuous predictors, the last column is simply the square root of their VIFs. But for a categorical variable that generates multiple dummy variables, this quantity combines all dummy variables and produces a single VIF for the original categorical predictor. When the rule-of-thumb cutoff for VIF is 10, the cutoff for the last column is sqrt (10) = 3.16.)
7. **Residual plots.** Run the code below to generate the residual plot. (1) Does the residual plot suggest any nonlinear relationship? (2) Is the homoscedasticity (constant variance of the error term) assumption roughly met?

residualPlots(m, ~1, type="rstudent",  
 id=list(labels=row.names(Duncan)))

1. **Normality** **assumptions of residuals.** Run the code belowto generate the histogram and the Q-Q plot of the residuals of the fitted model. Based on the plot, is the normal distribution of residual assumption roughly met?  
     
   hist(residuals(m))  
   qqPlot(m)
2. **Detecting influential points: Cook’s D.** 
   1. Run the code below to generate the influence plot. Describe your impression of the plot.  
        
      influencePlot(m, id=list(labels=row.names(Duncan)))
   2. Run the code below to calculate the Cook’s D value of each case and find their corresponding percentile scores on the reference distribution *F(p, n – p)*. Does any case exceed the 50th percentile?

Cooks.d <- cooks.distance(m)

p <- 5

n <- nrow(Duncan)

percentile <- 100 \* pf(q=Cooks.d, df1=p, df2=n-p)

data.frame(Duncan, Cooks.d=round(Cooks.d, 3), percentile=round(percentile, 1))

1. **Detecting influenital points: dfbeta.**
   1. Generate the dfbeta matrix for the model by **dfbetas(m)**.
   2. For each predictor, find out the case that influences its coefficient to the largest extent. What occupation is it? Can you explain why this case increases or decreases the slope of this predictor?