HUDM 5123 - Linear Models and Experimental Design
Lab 02 - OLS Diagnostics

1 The Data

For lab today we will use the `state.x77` data that is built into R. Begin by accessing the help information on the data set by typing

```
help(state.x77)
```

Scroll down to the name `state.x77` to see that it is a matrix with 50 rows (one for each state) and 8 columns, each of which represents a state-level variable such as per capita murder rate, state area, percent high school graduates, etc. Note that `state.x77` is not a data frame, so we will make it a data frame and call it "dat".

```
dat <- data.frame(state.x77)
```

Use the functions `names()`, `head()`, `tail()`, `dim()`, and `str()` to examine the data. Also discuss the use of the $Y \sim .$ shortcut for regression formulas.

Task 1 *Run a multiple regression of murder rate, using all the other variables as predictors in the model, and assign it to the name `lm1`. Write out the model using βs, Y, Xs, and ϵ. Then report and interpret the R^2 value and the residual standard error and its degrees of freedom.*

2 Diagnostics

The package **car**, written by the author of our textbook, has most of the functions in it we will use for diagnostics in lab today. Install (if you haven't already) and load the package:

```
install.packages("car")
library(car)
```

2.1 Leverage, Discrepancy, and Influence

In line with the notes, we will begin by checking for points with high influence. Access the help file on the `influencePlot()` function with

```
help(influencePlot)
```

and read the description.

Task 2 *Describe the difference between a point with high leverage and a point that has high influence.*

Task 3 *Run the function to create an influence plot. Save a jpeg of the plot by going to "export" in the plot pane in Rstudio. Which points (a) have highest leverage, (b) most discrepancy, and (c) are most influential?*

Task 4 *Should influential points be thrown out here? Why or why not? Paste a copy of the plot into your lab write-up.*

2.2 Non-normality

With only 50 observations, it will be a challenge to assess normality using a histogram. Instead, we will use a QQ plot using the `qqPlot()` function from package **car**.

Task 5 *Create and interpret a QQ plot of studentized residuals for `lm1` with the function `qqPlot(lm1)`. Does the plot show evidence of non-normality or not? Save the QQ plot as a jpeg and copy and paste it into your lab document.*

2.3 Non-constant Error Variance

To check for constant error variance we will examine a plot of studentized residuals against the ordered fitted (i.e., predicted) values. To create this plot, use the function

```
residualPlot(lm1, type = "rstudent")
```

Task 6 *Is there evidence for non-constant error variance here? Why or why not? Save the residual plot as a jpeg and copy and paste it into your lab document.*

2.4 Nonlinearity

Component-plus-residual plots allow us to check on the linearity assumption for each predictor variable. To create the CR plot for the "Population variable", for example, use the following code:

```
crPlot(lm1, variable = "Population")
```

To create the CR plots for other variables, simply swap out the variable names.

Task 7 *After examining all seven of the CR plots, what do you conclude about the tenability of the linearity assumption? Be specific.*

2.5 Multicollinearity

It's not a bad idea to start by examining the pairwise correlations with `round(cor(dat), 1)`. If you want to see more digits, change 1 to 2 or 3, for example. Next, use the vif() function to get VIFs.

Task 8 *Which variable and/or variables have the largest variance inflation factors? Calculate the value of R_j^2 for those variables and describe how R_j^2 is calculated.*