Stat 6021 R Tutorial: Detecting Outliers

In this tutorial, we will learn how to use R to detect (influential) outliers in our data. The data for this tutorial is "bp.txt". The data set contains information on the weight and systolic blood pressure of 26 randomly selected males in the age group 25-30.

1. Residuals, studentized residuals, and externally studentized residuals. Create a plot of these residuals against the fitted values. Assume that the variable `result` stores the model, `result$residuals` produces the residuals, `rstandard(result)` produces the studentized residuals, and `rstudent(result)` produces the externally studentized residuals. Comment on similarities / differences in these plots.

2. Detecting outliers in the response variable. We use the externally studentized residuals, t_i, to detect outliers in the response. If the regression model is appropriate, each t_i will follow the t distribution with $n-p-1$ degrees of freedom. Using the Bonferroni procedure, we consider an observation i to be outlying in the response if $|t_i| > t_{1-\alpha/2n;n-p-1}$. Here are a few ways to detect outliers in the response:

   (a) Sorting the t_i's. Use `sort()` to sort the externally studentized residuals, and compare with the value of $t_{1-\alpha/2n;n-p-1}$.

   (b) Overlay the critical values $\pm t_{1-\alpha/2n;n-p-1}$ on the plot of the externally studentized residuals. For example,

   ```
   plot(ext.student.res, ylim=c(-4,4))
   abline(h=qt(1-0.05/(2*n), n-p-1), col="red")
   abline(h=-qt(1-0.05/(2*n), n-p-1), col="red")
   ```

   The optional argument `ylim` was specified in `plot()` to ensure the lines representing the critical values are visible.

   (c) Typing `ext.student.res[abs(ext.student.res)>qt(1-0.05/(2*n), n-p-1)]` will produce a list of observations that have $|t_i| > t_{1-\alpha/2n;n-p-1}$.

3. Detecting outliers in the predictors. We use leverages, h_{ii}, to detect outliers in the predictors. Typing `lm.influence(result)$hat` produces the leverages for our model. An observation i is considered outlying in the predictors if $h_{ii} > \frac{2p}{n}$. Just like detecting outliers in the responses, here are a few ways to identify outliers in the predictors:

(a) Use `sort()` to sort the leverages, and compare with $\frac{2p}{n}$.

(b) Overlay the lines corresponding to $\frac{2p}{n}$ on a plot of leverages.

```
plot(lev, main="Leverages", ylim=c(0,0.4))
abline(h=2*p/n, col="red")
```

Notice from the plot we have observations with high leverages. To identify which observations these are, you may type `identify(lev)`. Notice that if you move your cursor over the graphic window containing the plot of the leverages, the cursor is a cross-hair. To identify the high leverage observations move the cursor close to a plot with high leverage, and click. A number that represents the observation number appears near the plot. Continue moving your cursor and clicking to identify all the high leverage observations. When you are done, right-click on the graphic window, and select stop. The index numbers also appear on the main R console.

(c) Typing `lev[lev>2*p/n]` will output the observations that have leverages greater than $\frac{2p}{n}$.

4. Influential outliers. An outlier is influential if its presence in the data set changes the estimated regression coefficients (and as a result the predicted values as well) substantially. We use $DFFITS_i$, $DFBETAS_{j,i}$, and Cook's distance to check for influential observations.

(a) $DFFITS_i$. In R, we use the `dffits()` function, for example, `DFFITS<-dffits(result)`. What is the criterion used for $DFFITS_i$?

(b) $DFBETAS_{j,i}$. In R, we use the `dfbetas()` function.

(c) Cook's Distance. In R, we use the `cooks.distance()` function, for example, `COOKS<-cooks.distance(result)`. What is the criterion used for Cook's distance?