Print

## Practice: Using PROC REG to Generate Potential Outliers

Generate statistics for potential outliers in the **stat1.bodyfat2** data set. Write this data to an output data set, and print your results.

1. Use PROC REG to run a regression model of **PctBodyFat2** on **Abdomen**, **Weight**, **Wrist**, and **Forearm**. Create plots to identify potential influential observations that are based on the suggested cutoff values. Submit the code and view the results.

```
/*st105s02.sas*/  /*Part A*/
ods graphics on;
ods output RSTUDENTBYPREDICTED=Rstud
           COOKSDPLOT=Cook
           DFFITSPLOT=Dffits
           DFBETASPANEL=Dfbs;
proc reg data=STAT1.BodyFat2
         plots(only label)=
              (RSTUDENTBYPREDICTED
               COOKSD
               DFFITS
               DFBETAS);
   FORWARD: model PctBodyFat2
                 = Abdomen Weight Wrist Forearm;
   id Case;
   title 'FORWARD Model - Plots of Diagnostic Statistics';
run;
quit;
```

   Here are the results.

   - In the RStudent by Predicted for PctBodyFat2 scatter plot, only a modest number of observations are further than two standard error units from the mean of 0.
   - In the Cook's D for PctBodyFat2 plot, there are 10 labeled outliers, but observation 39 is clearly the most extreme.
   - In the Influence Diagnostics for PctBodyFat2 plot, the same observations are shown to be influential by the DFFITS statistic.
   - In the panel plot, DFBETAS are particularly high for observation 39 on the parameters for **Weight** and **Forearm** circumference.

2. Write the residuals output to a data set named **influential**, subset the data to select only observations that are potentially influential outliers, and print your results. Submit the code and view the results.

```
/*st105s02.sas*/  /*Part B*/
data influential;
/*  Merge datasets from above.*/
    merge Rstud
          Cook
          Dffits
                  Dfbs;
    by observation;

/*  Flag observations that have exceeded at least one cutpoint;*/
    if (ABS(Rstudent)>3) or (Cooksdlabel ne ' ') or Dffitsout then flag=1;
    array dfbetas{*} _dfbetasout: ;
    do i=2 to dim(dfbetas);
        if dfbetas{i} then flag=1;
    end;
```

```
/*  Set to missing values of influence statistics for those*/
/*  who have not exceeded cutpoints;*/
    if ABS(Rstudent)<=3 then RStudent=.;
    if Cooksdlabel eq ' ' then CooksD=.;

/*  Subset only observations that have been flagged.*/
    if flag=1;
    drop i flag;
run;

proc print data=influential;
    id observation ID1;
    var Rstudent CooksD Dffitsout _dfbetasout:;
run;
```

Here are the results.

The same observations appear in the PROC PRINT report as in the plots.

Examine the values of observation 39 to see what is causing problems. You might find it interesting.

Hide Solution