Print

Practice: Using the Linear Regression Task to Generate Potential Outliers

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

1. Use the Linear Regression task 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.

    1. In the Navigation pane, select **Tasks and Utilities**.
    2. Expand **Tasks**.
    3. Expand **Statistics** and open the **Linear Regression** task.
    4. Select the **stat1.bodyfat2** table.
    5. Assign **PctBodyFat2** to the Dependent variable role.
    6. Assign **Abdomen**, **Weight**, **Wrist**, and **Forearm** to the Continuous variables role.
    7. On the MODEL tab, click the **Edit this model** icon, select all variables, and click **Add**. Then click **OK**.
    8. On the OPTIONS tab, expand **Diagnostic and Residual Plots** and clear the check boxes for **Diagnostic plots** and **Residuals for each explanatory variable**.
    9. Expand **More Diagnostics Plots** and select all four check boxes. This will display diagnostic plots with labels for influential observations.
    10. Expand **Scatter Plots** and clear the check box for **Observed values by predicted values**.
    11. Modify the code to add the Cook's D influence statistics and to export the RSTUDENT, DFFITS, DFBETAS, and Cook's D statistics.
        a. On the CODE tab, click the **Edit SAS code** icon.
            1. In the PROC REG step, enter **cooksd** within the parentheses where the plots are listed.
            2. Add an ODS OUTPUT statement to save the data from the plots.
                - Save RSTUDENTBYPREDICTED in a data set named **Rstud**.
                - Save COOKSDPLOT in a data set named **Cook**.
                - Save DFFITSDPLOT in a data set named **Dffits**.
                - Save DFBETAPANEL in a data set named **Dfbs**.
    12. Click **Run**.

Modified Code

```
ods noproctitle;
ods graphics / imagemap=on;
ods output RSTUDENTBYPREDICTED=Rstud
           COOKSDPLOT=Cook
           DFFITSPLOT=Dffits
           DFBETASPANEL=Dfbs;

proc reg data=STAT1.BODYFAT2 alpha=0.05 plots(only label)=(rstudentbypredicted cooksd dffits dfbetas);
   model PctBodyFat2=Weight Abdomen Forearm Wrist /;
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 a DATA step to merge the four output data sets based on the common variable, **observation**. Add code to subset the data to select only the observations that are potentially influential outliers. Name the output data set **influential.** Submit the code and use the PRINT procedure or the Table Viewer in SAS Sudio to display the **influential** data set.

```
/*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