Print

Practice: Using PROC CORR to Describe the Relationship between Continuous Variables

The percentage of body fat, age, weight, height, and 10 body circumference measurements (for example, abdomen) were recorded for 252 men by Dr. Roger W. Johnson of Calvin College in Minnesota. The data are in the **stat1.bodyfat2** data set. Body fat, one measure of health, has been accurately estimated by a water displacement measurement technique.

1. Generate scatter plots and correlations for the VAR variables **Age**, **Weight**, and **Height**, and the circumference measures **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist** versus the WITH variable, **PctBodyFat2**.

   IMPORTANT: For PROC CORR, ODS Graphics will display a maximum of 10 VAR variable plots at a time. This practice analyzes thirteen variables, so it requires two PROC CORR steps to generate all thriteen plots. This limitation only applies to the ODS graphics. The correlation table displays all variables in the VAR statement by default.

   Analyze the relationships:

   - Write a PROC CORR step to analyze all thirteen variables (**Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist**). This will generate a correlation table for all of the variables, but it will display plots for only the first ten.
   - Write an ODS statement to limit the graphic output to scatter plots.
   - Write another PROC CORR step, to look at only the last three variables, **Biceps**, **Forearem**, and **Wrist**.

   Submit the code. The output should include a correlation table for all thirteen variables followed by a plots for the first ten, and then plots for the last three.

```
/*st102s03.sas*/  /*Part A*/
%let interval=Age Weight Height Neck Chest Abdomen Hip
              Thigh Knee Ankle Biceps Forearm Wrist;

ods graphics / reset=all imagemap;
proc corr data=STAT1.BodyFat2
          plots(only)=scatter(nvar=all ellipse=none);
   var &interval;
   with PctBodyFat2;
   id Case;
   title "Correlations and Scatter Plots";
run;

%let interval=Biceps Forearm Wrist;

ods graphics / reset=all imagemap;
ods select scatterplot;
proc corr data=STAT1.BodyFat2
          plots(only)=scatter(nvar=all ellipse=none);
   var &interval;
   with PctBodyFat2;
   id Case;
   title "Correlations and Scatter Plots";
run;
```

   Here are the results.

2. Examine the plots. Can straight lines adequately describe the relationships?

   Yes. **Height** seems to be the only variable that shows no real linear relationship. **Age** and **Ankle** show little linear trend.

3. Are there any outliers that you should investigate?

   One person has outlying values for several measurements. In addition, there are one or two values that seem to be outliers for **Ankle**.

4. Which variable has the highest correlation with **PctBodyFat2**?

   Abdomen, with r=0.81343, is the variable with the highest correlation with **PctBodyFat2**.

5. What is the *p*-value for the coefficient? Is it statistically significant at the 0.05 level?

   The *p*-value is <.0001.

6. Generate correlations among all the variables previously mentioned (**Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist**) minus **PctBodyFat2**. Use the OUT= option in the PROC CORR statement to output the correlation table into a data set named **pearson**. Use the BEST= option to select only the highest five per variable.

   Submit the code and review the results.

```
/*st102s03.sas*/  /*Part B*/
ods graphics off;
%let interval=Age Weight Height Neck Chest Abdomen Hip Thigh
              Knee Ankle Biceps Forearm Wrist;

proc corr data=STAT1.BodyFat2
          nosimple
          best=5
          out=pearson;
   var &interval;
   title "Correlations of Predictors";
run;
```

   Here are the results.

7. Are there any notable relationships?

   Several relationships appear to have high correlations (such as those among **Hip**, **Thigh**, and **Knee**). **Weight** seems to correlate highly with all circumference variables.

8. **Challenge:** Use the **pearson** data set to print only the correlations whose absolute values are 0.70 and above, or note them with an asterisk in the full correlation table.

   Submit the code and review the results.

   Potential solution to printing the correlation matrix with asterisks in the full correlation table:

```
/*st102s03.sas*/  /*Part B*/
%let big=0.7;
proc format;
    picture correlations &big -< 1 = '009.99' (prefix="*")
                         -1 <- -&big = '009.99' (prefix="*")
                         -&big <-< &big = '009.99';
run;

proc print data=pearson;
    var _NAME_ &interval;
    where _type_="CORR";
    format &interval correlations.;
run;
```

Here are the results.

Potential solution to printing only the correlations whose absolute values are 0.7 and above:

```
/*st102s03.sas*/  /*Part B*/
%let big=0.7;
data bigcorr;
    set pearson;
    array vars{*} &interval;
    do i=1 to dim(vars);
        if abs(vars{i})<&big then vars{i}=.;
    end;
    if _type_="CORR";
    drop i _type_;
run;

proc print data=bigcorr;
    format &interval 5.2;
run;

title;
```

Here are the results.

Hide Solution