# McMaster University

## Department of Mathematics and Statistics

## STATISTICS 3A03: Applied Regression Analysis with SAS

## Fall 2017

## SAS Lab 2. Week of September 18-22, 2017

**Topics Covered in this Lab**

1. PROC PLOT
2. PROC GPLOT
3. PROC CORR
4. PROC REG

**1. PROC PLOT** This procedure creates scatter plots. The basic form of the PROC is

```
PROC PLOT Data=S3A3.Heights;
    Plot Dheight*Mheight;
run;
```

In the `Plot` statement the first variable specified goes on the vertical (Y) axis and the second goes on the horizontal (X) axis. The SAS default is to use the letters `A`–`Z` as plotting symbols. An `A` is plotted when there is only one point at (or very close) to the plotting position. When two points need to be plotted on top of each other, a `B` is printed etc. This rarely makes for very nice plots! I suggest changing the plotting character used. Often people like to use the characters `*` or `+`. You can do this in the plot statement as follows

```
PROC PLOT Data=S3A3.Heights;
    Plot Dheight*Mheight="*";
run;
```

It is often helpful to give a title to your plot which you can do by adding a statement like

```
Title "Mothers and Daughters Heights"
```

The SAS default is to print out the variable names as the axis labels. This is rarely useful. The `Label` statement can be used to override this default.

```
PROC PLOT Data=S3A3.Heights;
    Plot Dheight*Mheight="*";
    Title "Mother and Daughter Heights";
    Label Dheight="Daughter's Height";
    Label Mheight="Mother's Height";
run;
```

**2. PROC GPLOT** The `PLOT` procedure is the basic plotting method in SAS. The output from `PLOT` is part of the regular SAS output so it does not often look great. A better approach is using the `GPLOT` procedure which produces a separate graphics plot. This plot can be saved as a number of different types of image files for later use. Plots produced from `PROC GPLOT` are usually much nicer than those produced by `PROC PLOT`. Here is an example

```
PROC GPLOT Data=S3A3.Heights;
    Plot Dheight*Mheight;
    Symbol Value=STAR Color=BLUE;
    Title "Mother and Daughter Heights";
    Label Dheight="Daughter's Height";
    Label Mheight="Mother's Height";
run;
```

**3. PROC CORR** This is the procedure to find sample correlation coefficients and to test the null hypothesis that the population correlation is 0.

```
PROC CORR Data=S3A3.Heights;
    VAR Dheight Mheight;
run;
```

The output from this includes basic summary statistics on each variable in the `VAR` statement. The correlations are displayed in a matrix. The diagonal elements are always 1. The off-diagonal elements give the observed sample correlation coefficient and the $p$-value for the hypothesis test as described in class.
An alternative way that is sometimes useful and gives a more compact output is:

```
PROC CORR Data=S3A3.Heights;
    VAR Dheight;
    WITH Mheight;
run;
```

This will correlate each of the variables in the `Var` statement with each variable in the `WITH` statement.
**4. PROC REG** This is the basic regression procedure in SAS.

```
PROC REG Data=S3A3.Heights Plots=none;
    Model Dheight=Mheight;
run;
```

The displayed output includes the Analysis of Variance Table, some important quantities such as the Root MSE ($\hat{\sigma}$), $R^2$ and the adjusted $R^2$, and the table of parameter estimates, their standard errors, the $t$-statistic for testing if the population parameter is 0 and the $p$-value of the two sided test.
You can also add a `PLOT` Statement in `PROC REG` and a scatterplot is produced with the fitted least squares regression line on the plot and some summary statistics.
We can also use `PROC REG` to construct a dataset that includes such things as fitted values and the corresponding residuals for each point. This is often very useful for subsequent analysis of the fitted model. The `OUTPUT` statement will do this as follows:

```
PROC REG data=S3A3.heights Plots=none;
     Model Dheight=Mheight;
     PLOT DHeight*MHeight;
Output Out=heights_out
     Predicted=Fitted
          Residual=Residuals;
run;
quit;
```

The `Out=heights.out` gives the name of the output dataset. Note that I did not give a library prefix so this dataset will be created in the Work library and will disappear when you end your SAS session. To make it permanent use something like `S3A3.heights_out` instead. The `Predicted=Fitted` tells SAS to store the predicted values in a column called `Fitted` in the output dataset and `Residual=Residuals` tells it to save the residuals in a column called `Residuals` in the output dataset. The names after the `=` can be any valid name but those before `must` be exactly as given above. Later we will see other things that can be added to this dataset too. Note that the original data columns are automatically included in the output dataset also.
We can save the estimates and also get more confidence intervals for the intercept and slope in a new data set using `Outest=` and `Tableout` options in the `PROC REG` statement as in

```
PROC REG data=S3A3.heights Plots=none Outest=S3A3.heights_est Tableout;
     Model Dheight=Mheight;
run;
quit;
```

We can then examine this dataset as any other dataset to extract the estimates, standard error, p-values, t-statistics and 95% confidence intervals for the intercept and each $X$ variable.

**Exercise:**

(a) Import the dataset `Heights2.txt` which contains the heights of married couples (in centimeters) as shown in Table 2.11 of your textbook.

(b) Draw a scatterplot of the data.

(c) Find the correlation coefficient and test the hypothesis that the heights of husbands and their wives are linearly related. What is your conclusion?

(d) Do a regression to see how well husband's height can be used to predict the height of his wife and find the estimated intercept and slope as well as the estimate of the error standard deviation $\sigma$.

(e) What is the $p$-value for the test of $H_0 : \beta_1 = 0$ V $H_1 : \beta_1 \neq 0$?

(f) Give a confidence interval for the slope.

(g) What would a slope of 1 imply about the relationship between the heights of married couples. What is the value of the test statistic to test the hypothesis that the slope of the line is 1?

**Advanced Example (optional):** This example is a bit more advanced than I expect you to be able to do but is included for those who are interested in learning more about using SAS. In class we saw a plot of the Height data with both the fitted regression line and the diagonal line for comparison. The following code can be used to create such a plot.

```
PROC REG data=S3A3.heights Plots=none;
     Model Dheight=Mheight;
Output Out=heights_out
     Predicted=Fitted;
run;
quit;

PROC SORT data=heights_out;
      BY=MHeight;
run;

symbol1 interpol=none value=dot color=black;

symbol2 interpol=join value=none color=red;

symbol3 interpol=join value=none color=blue;

PROC GPLOT data=heights_out;
     Plot Dheight*Mheight Mheight*Mheight Fitted*Mheight / overlay;
     Title "Mother and Daughter Heights";
     Label Dheight="Daughter's Height";
     Label Mheight="Mother's Height";
run;
quit;
```

This is what each portion of the code does

- The `PROC REG` fits the regression and saves the fitted values in an output dataset.
- The `PROC SORT` sorts the data using the column specified in the `BY` statement. Since I did not give an output dataset it will overwrite the input dataset with the sorted dataset.
- The `symbol1` statement gives the plotting symbol to be used for the first plot in any subsequent `PROC GPLOT`. In this case we specify to use a black dot and not to join the points so we get a regular scatter plot.
- Similary the `symbol2` and `symbol3` statements give the plotting symbol to be used for the second and third plot in any subsequent `PROC GPLOT` if there are multiple plots. `interpol=join` says to join the points together so we will get lines instead of multiple points.
- Finally `PROC GPLOT` is asked to produce three plots, the first is of the daughters heights against their mothers heights, the second is the mothers heights against themselves (so the diagonal line) and the third is the fitted values against the mothers heights (so the fitted line). The `/overlay` option tells SAS to overlay all three plots on the same graph.