ST 512 - Lab 5 - Correlation and Introduction to SLR

1. Open the `Soil Water Data.sas` file. Here you will find a data set with measurements on soil depth (cm) and soil water content (mL/cm^3). Note: We could treat the depth variable as a factor and run a One-Way ANOVA model here. However, since this variable actually has a continuous underlying scale, we can investigate other questions as well!

   Some questions we may want to answer from this type of data set:

   (i) Does there appear to be an association between the two variables?

   (ii) If so, does that association appear to be linear?

   (iii) Can we conduct a statistical test to determine this linear relationship is statistically significant?

   (iv) Can we fit a linear regression line to this data?

   (v) How can we use that line to predict for a future observation?

2. Use PROC CORR to answer questions (i) and (ii) above.

   (a) Start by using the following code.

   ```
   ods graphics on;
   proc corr data=soilwater plots=matrix;
       var depth swc;
       run;
   ods graphics off;
   ```

   Inspect the resulting plot(s). How would you describe the relationship? (Form, Strength, Direction)

   (b) To answer part(iii) above, we can do a test for the correlation: $H_0 : \rho = 0 \quad vs \quad H_1 : \rho \neq 0$
   To get the Fisher's Z test and the t-test for correlation from class, we can run the following code:

   ```
   ods graphics on;
   proc corr data=soilwater plots=matrix fisher(biasadj=NO);
       var depth swc;
   run;
   ods graphics off;
   ```

   Note: The t-test is the default test, we have to ask for the Fisher's Z test.

   - What is the sample correlation, r_{XY}?
   - According to the output from the Fisher's Z test, is there a significant correlation?
   - What is a 95% CI for ρ? How would you interpret this interval?
   - According to the output for the t-test, is there a significant correlation?

3. Rather than inspect the linear relationship through correlation, we could instead fit a line to the data.

   To answer (iv) and 'fit' this line (i.e. get estimates for the parameters β_0, β_1, and σ^2), we can use PROC GLM. Note: We could also use the PROC REG (shown in lecture) or PROC MIXED (discussed later in the course).

```
ods graphics on;
proc glm data=soilwater plots = all;
    model swc=depth;
run;
ods graphics off;
```

   Use the provided output to do the following:

   (a) Write down the equation for the estimated (or 'fitted') line based on your output.

   (b) Explain what hypothesis is tested by each p-value.

   (c) Determine if the estimated line appears to fit the data on the scatter plot well.

   (d) Determine if any assumptions have been violated.

4. Why might fitting a line be more useful than treating depth as a factor with only the 4 levels? To answer (v) from above, let's run the following code:

```
proc glm data=soilwater plots = none;
    model swc = depth / clparm clm;
run;
```

   (a) What is the predicted value when depth is 40? How is SAS coming up with this value?

   (b) Get CI for the mean response when the depth is 12.5 cm or 49 cm. Use the following code to help out!

```
data newdepths;
    input depth;
    cards;
    12.5
    49
    ;
run;

data alldata;
    set soilwater newdepths;
run;
```

   Be sure to check your log first to make sure this worked!

   (c) Now, adjust your code to get the PI for an individual response is 12.5 cm or 49 cm.

If you have time and would like more practice, there is another data set available named `mother.xls`.

Weight gain of the mother during pregnancy is known to be a critical factor in determining the birthweight of the infant. Some data collected in a study of the relationship between average weight gain and mother's age are given in this data set.

For practice, you may want to try and answer the following questions about the mother data set:

1. What is the sample correlation between weight gain and age?
2. Is the sample correlation significantly different from zero?
3. Which variable would we consider the response and which the independent variable? Why?
4. Fit a regression line to the data, is the slope significantly different from 0?
5. Do the data appear to satisfy the assumption of constant variance?
6. Predict the value of weight gain for someone who is 20 years old and someone who is 35 years old.