EXST 7015

Fall 2014

Lab02: More Simple Linear Regression Analysis

This example will help reinforce some of the concepts of simple linear regression. We will look at some of the same questions, but with a different example and include a few new options.

Objectives

1. Prepare a scatter plot of the dependent variable (Y_i) on the independent variable (X_i)
2. Print the data set
3. Do a regression analysis in **PROC REG** including:
   - 99% Confidence intervals on the regression coefficients (b_0 and b_1)
   - Test the hypothesis; ***H_0: $\beta_1 = 0.5$***
   - Use the PLOT statement in PROC REG to get a residual plot
   - Output of the predicted values, residuals and standardized residuals
4. Create a residual plot using the standardized residuals (RSTUDENT)
5. Use **PROC UNIVARIATE** to test the residuals for normality

Housekeeping Statements

Continue to include the "housekeeping " statements.

```
dm 'log;clear;output;clear';
ods html close;
ods html style=minimal body='EXST7015 Assign02.html';
ods graphics off;
options nodate nocenter pageno=1 ls=111 ps=53;
options FORMCHAR="|----|+|---+=|-/\<>*";
ODS listing;
```

The include comment statements that identify log printouts as your own,

```
************************************************;
*** James P Geaghan                          ***;
*** EXST 7015 Regression Example             ***;
*** Assignment 02                            ***;
************************************************;
```

and TITLE# statements to be printed at the top of each page that also serve to identify the owner of the results.

```
title1 "James P Geaghan";
title2 "Assignment 2 - Regression example";
```

Dataset

The data is from your textbook, chapter 7, problem 2. The values are:

0 7.9, 1 12, 3 9.5, 4 11.3, 5 11.8, 6 11.3, 7 4.2, 8 0.4

You can use the values below in your program. The first value is "days after picking" and the second is a measure of sugar content.

0	7.9
1	12.0
3	9.5
4	11.3
5	11.8
6	11.3
7	4.2
8	0.4

The following program performs a SLR using SUGAR as the dependent variable and DAYS as the independent variable. The code provided is minimal. Feel free to embellish the program and make it your own.

```
data fw07p02;
   input Days Sugar;
cards;
0    7.9
1   12.0
3    9.5
4   11.3
5   11.8
6   11.3
7    4.2
8    0.4
;
Proc Print data = fw07p02;
   title3 'Sugar content on days after picking';
Run;
```

Creating a Scatter Plot

When performing a regression analysis, it is always advisable to look at scatter plots of the data in order to get an idea of the type of relationship that exists between the response variable and the explanatory variables.

```
Proc plot data=fw07p02;
   Title3 'Scatter plot of Sugar content on days';
   plot Sugar * Days;
run;
```

The PROC PLOT statement above will create a scatter plot. To create more professional graphics, you will want to use procedures in SAS GRAPHICS package. Try changing the "PROC PLOT" to "PROC **G**PLOT" to see the results. The resulting graphic will be placed in a new graphics tab. You may also wish to explore the SAS ODS option "ODS graphics on;" option in the housekeeping statements.

Fitting the Least-Squares Regression line Using SAS

Based on the scatter plot produced above, we will assume that an appropriate regression model relating SUGAR to DAYS is the linear model given by

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where Y is the SUGAR , X is the DAYS, and ε is a random error term that is normally distributed with mean 0 and unknown variances σ^2. β_0 is the estimate of the Y-intercept, and β_1 is the estimate of the slope.

SAS has several procedures that would fit this model to the data. In this lab we will use PROC REG. Include the "alpha=0.01" option in addition to CLB. This will give a 99% confidence interval instead of the 0.05 default value of alpha. We will also test the hypothesis that
H_0: β_1 = 0.5 with a "TEST" statement.

Include a plot statement to obtain a residual plot from PROC REG. Note that the variable "residual." is an internal SAS variable. Make sure your page size is not excessively large for the plot statement.

```
Proc reg data=fw07p02 lineprinter;
  title3 'SLR between SUGAR and DAYS';
  Model SUGAR  = DAYS / CLB alpha=0.01;
  TEST days = 0.5;
  plot residual.*days;
  output out=next01 r=resid p=predicted rstudent=rstudent;
run;
```

The usual residual plot from PROC REG is not scaled to standard deviation units. Use the "RSTUDENT" variable in the NEXT01 dataset to plot scaled residuals. In addition to lines at ±3 and ±2, include a reference line at zero. This plot should also be done on a page not much over 50 lines.

```
Proc plot data=next01;
  Title3 'RESIDUAL plot of SUGAR and DAYS';
  plot rstudent * DAYS / vref = -3 -2 0 2 3;
run;
```

Finally, use proc univariate to test for normality and to obtain plots to help examine the assumptions.

```
PROC UNIVARIATE data=next01 normal plots;
 TITLE3 'Test of residuals from REG';
 VAR resid;
 ods exclude BasicMeasures ExtremeObs ExtremeValues Modes
      MissingValues Quantiles TestsForLocation;
RUN;

ods html close;
```

Previously stated objectives

1. Prepare a scatter plot of the dependent variable (Y_i) on the independent variable (X_i)
2. Print the data set
3. Do a regression analysis in **PROC REG** including:
    - 99% Confidence intervals on the regression coefficients (b_0 and b_1)
    - Test the hypothesis; H_0: $\beta_1 = 0.5$
    - Use the PLOT statement in PROC REG to get a residual plot
    - Output of the predicted values, residuals and standardized residuals
4. Create a residual plot using the standardized residuals (RSTUDENT)
5. Use **PROC UNIVARIATE** to test the residuals for normality

LAB ASSIGNMENT

1. Run the program and print the data set. **(1 point)**

2. Make a scatter plot to show the relationship between SUGAR and DAYS. How does it look? Does there appear to be a good linear association? **(1 point)**

3. Use PROC REG to fit the linear model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. Explain briefly your findings. In particular; what are the parameter estimates, do the estimated differ significantly from zero (give a P value) and do the confidence intervals include zero? **(2 points)**

4. Write the estimated regression function. **Is the sugar content increasing or decreasing over time? (1 point)**

5. Previous studies have shown that the sugar content will increase by about 0.5 units daily for about a week. Test this value against the current data set. Are your results consistent with previous studies? **(1 point)**

6. What proportion of the variability in the dependent variable RANGE is accounted for by DAYS through the regression line? **(1 point)**

7. Examine the residual plot from PROC REG and the scaled plot done in PROC PLOT. Except for the different scale, do the plots appear to be the same? **(1 point)**

8. Does the residual plot appear to represent just a random scattering of points? If there are outliers do they appear to be explicable? **(1 point)**

9. Does the residual analysis indicate that the assumption of normality has been met? **(1 point)**

Please turn in the **SAS log** and as much of the output as you need to answer the questions.

Thank you.