EXST SAS Lab
Lab #12: Regression Analysis

Objectives

1. Prepare a scatter plot of the dependent variable (Y_i) on the independent variable (X_i)
2. Do a regression analysis in **PROC REG** including:
   - Confidence intervals on the regression coefficients (β_0 and β_1)
   - Tests a specific value of the slope (and/or intercept)
   - Obtain confidence limits for (1) the regression line at a some value of the dependent variable (X_i) and (2) the individual observations
   - Output the residuals, the standardized residuals and the confidence limits on predicted values,
   - Make a scatterplot and residual plot from within the PROC REG procedure
3. Plot the standardized residuals (RSTUDENT)
4. Print the output data set
5. Use **PROC UNIVARIATE** to test the residuals for normality

File from Wikimedia Commons: Fight of the *Poursuivante* against the British ship Hercules, 6/28/1803. Louis-Philippe Crépin (1772-1851)

Regression analysis

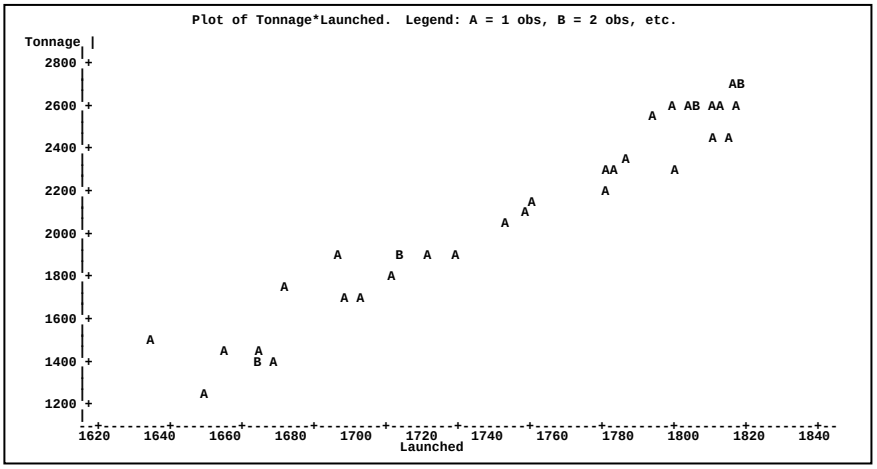
The example dataset for this regression is the tonnage (weight in tons) of sail powered battleships from the English navy over the years 1630 to 1833. The development of steam power and iron-clad vessels following this period altered the relationship, hence the restricted period. The name of the vessel is also included in the dataset, taking up to 21 characters. The input statements are:

```
TITLE1 'Simple linear Regression Example (SLR)';
DATA ONE; LENGTH name $ 21;
   INFILE "Battleship Weight.csv" missover DSD dlm="," firstobs=2;
   TITLE2 'Change in weight of sail powered battle ships from 1637 to 1833';
   INPUT Launched Tonnage Name $ ;
     LABEL Name     = 'Name of the ship'
           Launched = 'Year of launch'
           Tonnage  = 'Weight in tons';
datalines;
;
RUN;
```

In regression analysis it is advisable to first prepare a scatterplot of the dependent variable (Y_i) on the independent variable (X_i). This will help to determine if the potential relationship between the two variables is linear. It also aids in determining if there are some observations that do not appear to conform to the relationship between the variables (outliers) and to get some idea of the homogeneity of variance.

```
proc plot data=one; plot Tonnage*Launched;
   TITLE3 'Raw data scatterplot';
options ps=52 ls=111 ; run; options ps=512 ls=99;
```

Note that I altered the page and line sizes for the graphic. I like to use a large page size to keep output together without SAS splitting the output and adding page headers. However, a 10 page tall graphic is not useful. I shorten and widen the page to 52 lines tall by 111 wide before the run statement. After the **RUN** statement I return the size to my preferred height and width.



The scatter plot does not appear to show any obvious outliers; anomalous points that do not appear to belong to the same population or fit the same pattern as the other points. The line also appears to be a straight line, with no obvious curvature, and the variability about the line appears to be similar at both ends of the line. All of these judgments can be made from a scatter plot, but are often more easily evaluated from a plot of the residual (i.e. the deviations of individual points from the regression line).

The programming statements to run the regression analysis are given below, followed by a line by line description of the effect of each line.

```
a  PROC REG DATA=ONE LINEPRINTER;
b     TITLE3 'Regression with confidence limits';
c     MODEL Tonnage = Launched / clb CLI CLM P R; ID Launched;
d     SlopeTest:TEST Launched = 6.2;
e     OUTPUT OUT=NEXT P=Predicted R=Resid STUDENT=student
f        rstudent=rstudent lcl=lcl lclm=lclm ucl=ucl uclm=uclm;
g  RUN; OPTIONS PS=45; TITLE4 'Plots of raw data & residuals';
h  PLOT PREDICTED.*Launched='P' Tonnage*Launched='O' / OVERLAY;
i     PLOT RESIDUAL.*Launched='E';
j  RUN;
```

The **PROC REG** statement (line a) has one new option, **LINEPRINTER**. SAS now puts all graphics in the Results Viewer window. The **LINEPRINTER** requests that graphics also be done with the “old fashion” line printer graphics and be placed in the output listing.

The model statement (line c) has the dependent variable (tonnage) on the left of the equal sign and the independent variable on the right (Launched). This is the basic model and the first results are in the form of an Analysis of Variance source table. For this simple linear regression, there only are two sources; the model and the error. The table has the sum of squares for the model (SSModel or SSRegression) and for the error (SSError or SSResiduals). The corrected total sum of squares (CSS) is the sum of the two sources, SSReg and SSError. The CSS value is adjusted, or “corrected”, for the mean (i.e. $CSS = \sum_i^n (Y_i - \bar{Y})^2$) which adjusts the level (intercept) and accounts for the loss of one degree of freedom. The single degree of freedom for the model is for fitting the slope. Therefore, the test of the model is essentially a test of the slope against zero (i.e. $H_0 : \beta_1 = 0$). The best estimate of variance is provided by the Mean Square Error. This is a measure of random variation about the regression line and it forms the basis for the estimates of standard errors and confidence intervals about the regression

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7204891	7204891	674.58	<.0001
Error	35	373822	10681		
Corrected Total	36	7578713			

coefficients (intercept and slope) and predicted values from the regression. As a result all confidence intervals and related t-tests have the error degrees of freedom.

Under the ANOVA source table there are some summary statistics. One statistic in this group that is frequently cited is the R-Square value (i.e. R^2). In this case $R^2 = 0.9507$, indicating that 95.07% of the corrected total was accounted for by the model. The remaining 4.93% is "error"; random variation not accounted for by the model.

Root MSE	103.34715	R-Square	0.9507
Dependent Mean	2103.62162	Adj R-Sq	0.9493
Coeff Var	4.91282		

The next section contains the parameter estimates, their standard errors and a t-test of the parameter estimate against zero $H_0 : \beta_i = 0$. The estimates are sample estimates (b_0 and b_1) of the population

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-10293	477.58286	-21.55	<.0001	-11262	-9323.04421
Launched	1	7.05815	0.27175	25.97	<.0001	6.50646	7.60984

regression coefficients (β_0 and β_1). Based on these estimate, the regression equation ($\hat{Y}_i = b_0 + b_1 X_i$) can be written with numeric coefficients as $\hat{Y}_i = -10293 + 7.058 X_i$. The units on the slope are "Y units per X unit", in this case tons per year. That means that ships increase by an average rate of just over 7 tons a year during the period being studied. The intercept is the value of the dependent variable *Y* when the independent variable *X* is zero and it is often a value of interest. However, in this particular case it would be the weight of a British ship-of-the-line in the year zero, so it is not a meaningful value.

Following the slash on the model statement (line c) there are a series of options. We are most interested in **CLB** which produces confidence intervals for the regression coefficients (β_0 and β_1). As mentioned, the value of the intercept is not meaningful, so the confidence interval is irrelevant. However, the slope was interpretable and the confidence interval suggests that the true mean of the rate of increase in tonnage of these ships was between 6.5 and 7.6 tons per year. One of the best way to express this confidence interval is as a probability statement; $P\left(6.506 \le \beta_1 \le 7.610\right) = 0.95$.

A **TEST** statement can also be used to test a slope against a specific, hypothesized value. I have included a test of the observed slope against a hypothesized value of 6.2 (line d). This output can be assigned a name to be able to distinguish among several such tests. The name precedes the TEST statement and is separated by a colon (not a semi-colon). I used the name SlopeTest in line d. The results show that the null hypothesis $H_0 : \beta_1 = 6.2$ would also be rejected (P=0.0033).

Test SlopeTest Results for Dependent Variable Tonnage

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	106506	9.97	0.0033
Denominator	35	10681		

Diagnostics and Statistics for Individual Observations

Other options following the slash on the model statement produce diagnostics for each observation; **CLI** (upper and lower confidence limits for individual observations around the regression line), **CLM** (produces upper and lower confidence limits for the mean or predicted values ($\hat{Y}_i$), the regression line itself), **P**

Output Statistics

Obs	Launched	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict	
1	1637	1522	1262	36.6018	1187	1336	1039	1484
2	1655	1258	1389	32.3489	1323	1454	1169	1608
3	1670	1403	1495	28.9594	1436	1553	1277	1712
4	1671	1416	1502	28.7397	1443	1560	1284	1719
5	1673	1443	1516	28.3032	1458	1573	1298	1733

...

(predicted values) and **R** (residuals). I usually prefer to get these in an "output" statement because it gives more control over the output presentation. Both types of output are included in this example for comparison and partial results of first few lines are shown here. Results would be complete in the example output.

Also included on line "c" is an **ID** statement. By default, the independent variable (Launched) would not be included in the **Output Statistics** mentioned in the preceding paragraph. Variables specified in an ID statement are included. I could also have put the vessel name in the statement.

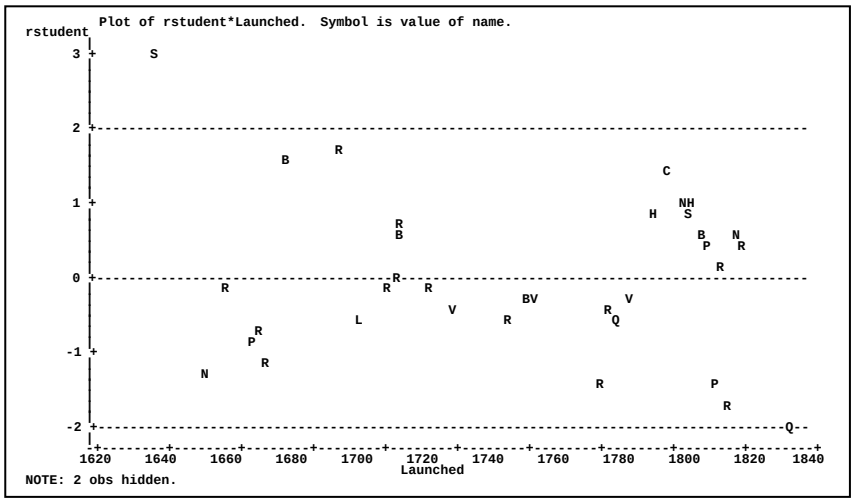
A large number of regression diagnostics can be output into a new SAS dataset using an **OUTPUT** statement (lines e & f). The only diagnostics we are interested in at present are the predicted and residual values,

Obs	Launched	Tonnage	Predicted	Resid	student	rstudent	lcl	ucl	lclm	uclm
1	1637	1522	1261.60	260.397	2.69426	2.98276	1039.03	1484.18	1187.30	1335.91
2	1655	1258	1388.65	-130.650	-1.33107	-1.34644	1168.81	1608.49	1322.98	1454.32
3	1670	1403	1494.52	-91.522	-0.92254	-0.92053	1276.64	1712.41	1435.73	1553.31
4	1671	1416	1501.58	-85.580	-0.86209	-0.85886	1283.81	1719.35	1443.24	1559.93
5	1673	1443	1515.70	-72.697	-0.73139	-0.72643	1298.17	1733.23	1458.24	1573.16
...										

some standardized residuals and some confidence limits. There are two types of confidence limits. Confidence limits for the regression line itself are referred to as "confidence limits for the mean" and denoted **CLM**. The lower and upper limits are designated **LCLM** and **UCLM**, respectively. Confidence limits for individual observations, which have more scatter and are wider intervals, are designated **LCL** and **UCL** for the lower and upper limits respectively. I included a **PROC PRINT** to list the variables that had been output. Of particular interest are the upper and lower limits of both the regression line (mean) and the individual points.

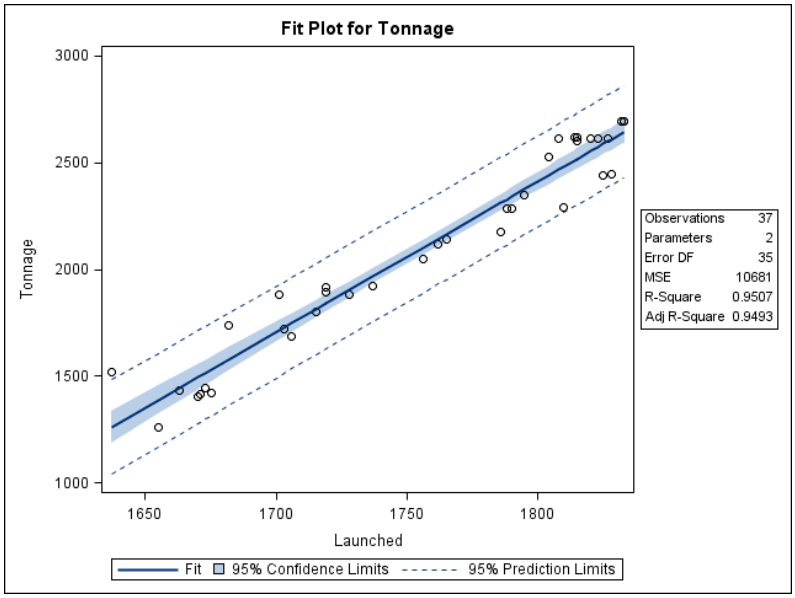
```
proc print data=next;
   TITLE4 'Listing of output from PROC REG';
   var Launched Tonnage Predicted Resid student rstudent
       lcl ucl lclm uclm; run;

proc plot data=next; plot rstudent * Launched = name
                          / vref = -2.030 0 +2.030;
   TITLE4 'Deleted standardized residuals with 95% interval';
options ps=52 ls=111 ; run; options ps=512 ls=99;
```

The standardized residuals (**STUDENT** and **RSTUDENT**) are also of interest. Raw residuals values are scaled to the units of the dependent variable. For the tonnages in this example the largest values are in the hundreds. In other analysis the value can be much larger or much smaller. Therefore, there is no useful information to be gleaned from the size of the raw residual values. However, since we have a variance and standard deviation of the residuals, these values can be scaled, like a Z score or t score. Once scaled, they will have values in the range of ±2 or ±3 and follow a t-distribution if the assumptions of normality and homogeneity are met. In SAS they are called **STUDENT** values.



Since an outlier would tend to pull the regression line towards itself it is often advantageous to see what the deviation for each point would be if the regression line was calculated without the point participating in the calculations. This deviation is standardized and called the **RSTUDENT** value. These are called *deleted standardized residuals*. Since this example had 35 degrees of freedom in the error term, we expect 95% of the

observations to fall within the t-value limits ±2.030, and 99% within the limits ±2.724.This is usually the best option for detecting outliers. To this end, a residual plot of the standardized residuals was done and the VREF= option used to locate critical levels of the standardized residual. Usually a horizontal reference line is also drawn at the zero value because we expect to see a random scatter of points about zero.



If **ODS GRAPHICS** are "**ON**" some high resolution graphics are produced. One of the graphics given with **PROC REG** provides some summary information and shows a scatter plot of the data and the fitted regression line. The shaded area around the regression line shows the **CLM** area, the calculated 95% confidence of the mean, or the regression line itself. The dotted lines show the 95% confidence bands for the individual points, estimated as **LCL** and **UCL** in the output discussed previously.

There are some graphics that can be produced within the **PROC REG**. These were done (lines h and i), though we have the necessary output and could do a **PROC PLOT** to accomplish the same type of plots. I put a run (line g) statement in to execute the **PROC REG** to that point and allow me to change the page size and to add a **TITLE** line.

The two plots produced by **PROC REG** are included in the example program (lines h and i). The first is a scatter plot of the observed values of tonnage on the year launched variable. The "**='O'**" 'will cause these to be plotted as the character **O**. The predicted values, plotted as **P** were superimposed on the observed points with an overlay option. The second plot is a simple residual plot on the independent variable. Note that to facilitate these plots SAS has some built in variables, **RESIDUAL.** and **PREDICTED.** that are followed by a period as a part of the variable name. These graphics are not included here.

Finally, I included a test of the residuals for normality using **PROC UNIVARIATE**.

```
PROC UNIVARIATE DATA=NEXT NORMAL PLOT; VAR Resid;
  TITLE4 'Residual analysis with PROC UNIVARIATE';
  ods exclude BasicMeasures ExtremeObs ExtremeValues Modes
              MissingValues Quantiles TestsForLocation;
RUN;
```

The interpretation of this test of the **assumption of normality** is the same as previous analyses that we have done. Three other regression assumptions are a little more difficult to check. The residual plot can be examined for the **assumption of homogeneity of variance**, but there is not a ready test of homogeneity provided in the **REG** procedure in SAS as there was for ANOVA with **PROC MIXED**. Failure to meet the **assumption of independence** among the observations is sometimes evidenced by strings of positive and negative residuals when the data is examined in the order that the data was acquired. The randomness of the strings can be tested with a "runs test". Finally, the **assumption that the independent variable is measured without error** is not easily checked, but random sampling helps in meeting this assumption. In ordinary least squares the variance is measured only in the vertical, Y-variable direction, and there are few software packages that allow for any alternative. Fortunately, the analysis is "robust" to this assumption, which in statistics means that good results can be expected even when the assumption is violated.

Assignment 12

Assignment 12 is to do a regression analysis of a dataset from you textbook. The dataset description is as follows: In an effort to determine the cost of air conditioning, a resident in College Station, TX, for the period from September 19 through November 4. He recorded the daily values of two variables, the mean temperature (Tavg) and electricity consumption (Kwh). Month and day are included in the dataset, but will not be needed in our analysis.

Following the usual lead statements and input section complete the following tasks.

http://en.wikipedia.org/wiki/File:Reddy_Kilowatt_iconic_pose.jpg

Task 1: Produce a scatterplot of the dependent variable (electricity consumption) on the independent variable (temperature).	1 point
Task 2: Produce a regression analysis and answer the following questions	
2a) Find the value and confidence interval for the intercept	1 point
2b) Find the value and confidence interval for the slope	1 point
2c) Test the slope against an hypothesized value of 2 Kwt per degree of temperature	1 point
2d) Estimate a predicted value and confidence interval for a single day temperature of 80 degrees	1 point
2e) Estimate a predicted value and confidence interval for a mean day temperature of 60 degrees	1 point
Task 3: Print the dataset output from the **PROC REG**.	1 point
Task 4: Prepare a residual plot with the **RSTUDENT** values	
Question 1: Do you see any evidence of an outlier in the residuals?	1 point
Question 2: Do you see any evidence of curvature in the residuals?	1 point
Task 5: Prepare a **PROC UNIVARIATE** with a test of normality of the residuals	
Question 3: Is the hypothesis of “normality” rejected for either of the two samples?	1 point