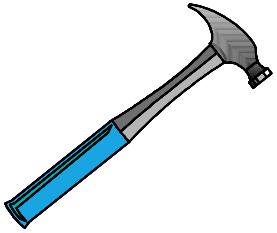
EXST7015

Numerical Example of a Simple Linear Regression



Simple Linear Regression

- **Our analyses will be done in SAS. Other, simpler options, such as EXCEL, work well for simple linear regression, but not other option will cover all of the analyses with all of the options that we want to cover this semester.**
- **If you are not familiar with SAS, see information available on my EXST7005 page, and talk to the TA about getting up to speed.**

Numerical Example

- **The example used is from your textbook. It is a data set taken from 47 trees. Each tree was measured for diameter, height, weight of harvestable wood and other values.**
- **Our objective will be to predict the weight of harvestable wood using just the diameter. The diameter variable is measured about 4 feet off the ground and is called Diameter at Breast Height (DBH).**

SAS Programming

- **The SAS program. I will presume you are familiar with the SAS data step. I will discuss it briefly only for this first example.**
- **SAS Statements - all SAS statements end in a semicolon;**
- **Comments - comments are statements that start with an asterisk. They do nothing in the program, they are included only for the purpose of documenting the program.**

SAS Programming *(continued)*

- **My Simple Linear Regression (SLR) example starts with the comments,**

```
*****************************************;
*** Data from Freund & Wilson (1993)      ***;
*** TABLE 8.24 : ESTIMATING TREE WEIGHTS  ***;
*****************************************;
```

- **This is for documentation purposes only. It does not affect the program.**

SAS Programming *(continued)*

- **Options - options can be specified to modify output appearance. The option statement I usually use is,**
  - **`options ps=61 ls=78 nocenter nodate nonumber;`**
  - **This option creates a page size (ps) of 61 lines (use 54 for the lab)**
  - **a line size of 78 character columns, and**
  - **suppresses the centering of output and printing of the date and page numbers.**

SAS Programming *(continued)*

- **The DATA step. All our programs will include a DATA section. In this section the data to be analyzed is entered into the SAS system and, if necessary, modified for analysis.**
  - `data one;`

SAS Programming *(continued)*

- A second statement informs SAS that the data is included in the program (CARDS)
- and that if there are missing values the system should NOT to the next line to get the data (MISSOVER).
  - `infile cards missover;`

SAS Programming *(continued)*

- The next statement in my program is a TITLE statement. Up to 9 titles can be active (TITLE1 through TITLE9) and once set are printed at the top of each page.
- Setting a new title, say TITLE3, would not affect lower numbered titles (TITLE1 and TITLE2) but would delete all higher numbered titles (TITLE4 ...).

SAS Programming *(continued)*

- **The TITLE statement ends in a semicolon as usual, and the text to be used a the title is enclosed in single quotes.**
- `TITLE1 'Estimating tree weights from other morphometric variables';`

SAS Programming *(continued)*

- The input statement. Along with the DATA statement, this is an important statement. It names the variables to be used, tells SAS what type of variables they are (numeric or alphanumeric) and gets the data into the SAS data set.
- `input ObsNo Dbh Height Age Grav Weight ObsID $;`

SAS Programming *(continued)*

- **Note that only one variable in the list is followed by a $. This will cause SAS to assume that all variables are numeric except the variable called OBSID.**
- **The variable OBSID is one I created by adding to each observation a different letter. The first line got an "a", the second a "b", etc. The 26th observation got a "z" and the 27th an "A", etc. This was done to have a way of distinguishing each observation.**

SAS Programming *(continued)*

- The LABEL statement provides a way of identifying each variable. It is optional, but if present will be used by SAS in a number of places to identify the variables.
  - `label ObsNo = 'Original observation number'`
  - `      Dbh = 'Diameter at breast height (inches)'`
  - etc. ... ;
- I have deactivated the labels by making them a comment statement.

SAS Programming *(continued)*

- **If data must be modified, it is done in the data step after the INPUT statement. I have two statements that create logarithms. These are not used in the first analysis, but will be used later in the semester.**
  - `lweight = log(weight);`
  - `ldbh = log(DBH);`
- **These statements create two new variables (LWEIGHT and LDBH) that are the natural logs of the original variables.**

SAS Programming *(continued)*

- Two last statements before the data. The CARDS statement tells SAS that the data step is done and data follows. The RUN statement tells SAS to process all information that it has so far and output any messages about the analysis to the LOG.
- `cards; run;`
- Note that two statements can occur on the same line.

SAS Programming *(continued)*

- **The SAS DATA step is now complete. The data will be entered into the SAS system and processing will continue.**
- **The rest of the statements in this program are procedures (PROCs) and associated statements.**

SAS Programming *(continued)*

- **I will briefly discuss some of these statements. For most of the semester we will concentrate on the PROCs that actually do statistics, such as REG, GLM, LOGISTIC, ANOVA, and MIXED.**

SAS Programming *(continued)*

- **The first PROC is,**
  - `proc print data=one; TITLE2 'Raw data print'; run;`
- **This PROC causes the data to be printed with the second title line added as**
  - **"Raw data print".**

SAS Programming *(continued)*

- Data list from PROC PRINT,

```
EXST7015: Estimating tree weights from other morphometric variables
Raw data print
      Obs                                                        Obs
Obs    No     Dbh    Height    Age     Grav    Weight    ID     lweight      ldbh
  1     1     5.7      34       10    0.409      174      a     5.15906   1.74047
  2     2     8.1      68       17    0.501      745      b     6.61338   2.09186
  3     3     8.3      70       17    0.445      814      c     6.70196   2.11626
  4     4     7.0      54       17    0.442      408      d     6.01127   1.94591
  5     5     6.2      37       12    0.353      226      e     5.42053   1.82455
  6     6    11.4      79       27    0.429     1675      f     7.42357   2.43361
  7     7    11.6      70       26    0.497     1491      g     7.30720   2.45101
  8     8     4.5      37       12    0.380      121      h     4.79579   1.50408
...
 44    44     4.0      38       13    0.407       76      R     4.33073   1.38629
 45    45     8.0      61       13    0.508      614      S     6.41999   2.07944
 46    46     5.2      47       13    0.432      194      T     5.26786   1.64866
 47    47     3.7      33       13    0.389       66      U     4.18965   1.30833
```

SAS Programming *(continued)*

- Notice that this is a TITLE2, so any previous title1 is kept.
- Also notice I usually follow PROCs with a RUN statement. This causes the procedure to be executed and any comments regarding the statement are placed in the LOG prior to the next PROC.

SAS Programming *(continued)*

- **The next PROC is a PLOT.**
  - `options ls=111 ps=61; proc plot data=one; plot weight*Dbh=obsid;`
  - `TITLE1 'Scatter plot'; run;`
  - `options ps=256 ls=132;`
- **It is surrounded by option statements. Although I usually like a large page size of (256), I don't want the plot to cover 256 lines, so I put the page size to 61 for the plot, and then reset it to 256 for subsequent output.**

SAS Programming *(continued)*

- **The plot is for weight on DBH. Notice the "=ObsID" at the end of the plot statement. This will cause SAS to plot a single character (the ObsID I created) as a symbol representing each observation in the plot. I do this to be able to distinguish between the observations in the plot.**
- **Output from the PLOT statement follows.**

SAS Programming *(continued)*

```
Scatter plot
                of WEIGHT*DBH.  Symbol is value of OBSID.
H      |
a      |
r 2000 +
v      |
e      |
s      |
t      |                                                     f   q
       |
w 1500 +                                                      g
e      |
i      |
g      |
h      |
t      |
  1000 +
o      |
f      |                                    c z
       |                                 N b     t
t      |
h      |                                  S
e  500 +                               Dy    s
       |                             d
t      |                       CO uo
r      |                    Bm  J
e      |              G   AT E
e      |         iUKpHwr
     0 +
(      |
l      --+----------+----------+----------+----------+----------+----------+--
b        2          4          6          8          10         12         14
s                         Diameter at breast height (inches)
NOTE: 16 obs hidden.
```

SAS Programming *(continued)*

- **The means statement is often used to examine variables and determine the number of observations of each variable, its minimum and maximum.**

```
proc means data=one n mean max min var std stderr;
     TITLE1 'Raw data means';
     var Dbh Height Age Grav Weight; run;
```

SAS Programming *(continued)*

- **This has limited utility for regression analysis.**

```
Raw data means

Variable  Label                                  N          Mean       Maximum
------------------------------------------------------------------------------
DBH       Diameter at breast height (inches)   47     6.1531915    12.1000000
HEIGHT    Height of the tree (feet)            47    49.5957447    79.0000000
AGE       Age of the tree (years)              47    16.9574468    27.0000000
GRAV      Specific gravity of the wood         47     0.4452979     0.5080000
WEIGHT    Harvest weight of the tree (lbs)     47   369.3404255       1692.00
------------------------------------------------------------------------------
```

- **You might use it to look for outliers, or to get the range of values for a plot.**

SAS Programming *(continued)*

- **The SAS UNIVARIATE procedure is very useful in regression analysis. However, the application to the RAW variables is not very useful.**

```
proc univariate data=one normal plot;
TITLE1 'Raw data Univariate analysis';
          var Weight Dbh; run;
```

SAS Programming *(continued)*

- **We will be interested in using this PROC to evaluate normality. We will be ESPECIALLY interested in the tests,**
  - `Shapiro-Wilk   W  0.710878    Pr < W   <0.0001`
  - `Shapiro-Wilk   W  0.89407     Pr < W    0.0005`
- **We will also be interested in other tools to evaluate normality (STEM & LEAF, BOX PLOT, NORMAL PROBABILITY PLOT), but NOT FOR THE RAW DATA for either variable (X or Y).**

SAS Programming *(continued)*

- **Note that these tests of normality are not useful.**
- **We will be assuming normality and testing for normality, but <u>not on the original variables</u>.**
- **We will later be testing the Deviations or Residuals!!! These are the appropriate tests of normality, not the tests of the original variables!!!**

Regression analysis

- As far a regression is concerned, the preceding material is ancillary, used to prepare or enhance our analysis.
- The important information for regression will be provided by PROC REG or PROC GLM.

```
proc reg data=one LINEPRINTER; ID ObsID DBH;
     TITLE2 'Simple linear regression';
     model Weight = Dbh / p xpx i influence clb
                          alpha=0.01; *** CLI CLM;
          Slope:Test DBH = 200;
          Joint:TEST intercept = 0, DBH = 200;
      run;
NOTE: 47 observations read.
NOTE: 47 observations used in computations.
```

Regression analysis *(continued)*

- **Additional useful statements that can be added to PROC REG include.**

```
!           options ls=78 ps=45;
        plot residual.*predicted.=obsid; run;
        OUTPUT OUT=NEXT1 P=YHat R=E STUDENT=student
        rstudent=rstudent lcl=lcl lclm=lclm ucl=ucl uclm=uclm;
  run;
  options ps=61 ls=132;
```

Regression analysis *(continued)*

- PROC REG Output
- The model is fitted by the statements,

```
84          proc reg data=one LINEPRINTER; ID ObsID DBH;
86                 model Weight = Dbh / p xpx i influence clb
                                   alpha=0.01;  *** CLI CLM;
```

```
Analysis of Variance
                                    Sum of           Mean
Source                   DF        Squares         Square    F Value    Pr > F
Model                     1        6455980        6455980     433.49    <.0001
Error                    45         670191          14893
Corrected Total          46        7126171

Root MSE            122.03740    R-Square     0.9060
Dependent Mean      369.34043    Adj R-Sq     0.9039
Coeff Var            33.04198

                                       Parameter Estimates
                        Parameter       Standard
Variable     DF          Estimate          Error    t Value    Pr > |t|       99% Confidence Limits
Intercept     1        -729.39630       55.69366     -13.10      <.0001     -879.18914     -579.60346
Dbh           1         178.56371        8.57640      20.82      <.0001      155.49675      201.63067
```

Regression analysis *(continued)*

- The ANOVA table

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6455980	6455980	433.49	<.0001
Error	45	670191	14893		
Corrected Total	46	7126171			

- Supplemental information

Root MSE	122.03740	R-Square	0.9060
Dependent Mean	369.34043	Adj R-Sq	0.9039
Coeff Var	33.04198		

- Parameter estimates and tests

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	99% Confidence Limits	
Intercept	1	-729.39630	55.69366	-13.10	<.0001	-879.18914	-579.60346
Dbh	1	178.56371	8.57640	20.82	<.0001	155.49675	201.63067

Parameter estimates and tests *(continued)*

- The parameter estimates are,
  - Intercept = -729.396300
  - Slope = 178.563714
  - Equation: $Y_i = -729.4 + 178.6*X_i$
  - Interpretation : The weight starts at -729 when the diameter is zero and increases by 179 pounds for each additional inch in diameter.

Parameter estimates and tests *(continued)*

- **For a t-test of either parameter against an hypothesized value or a confidence interval on either parameter we would use the standard errors provided by SAS.**
- **S_{b0} = 55.69366336**
- **S_{b1} = 8.57640103**

Parameter estimates and tests *(continued)*

- **A 95% confidence interval is calculated as, Parameter ± tvalue*standard error**
  - **The t-value has n-2=45 d.f. and can be found in a t-table. For a two tailed interval and a value of α = 0.05, the t-value is 2.014**
  - **For the slope the estimate is 178.6**
  - **The standard error is 8.576**
  - **The confidence interval is**
  - **178.6 ± 2.014*8.576**
  - **$P(161.328 \leq \beta_1 \leq 195.872) = 0.95$**

Parameter estimates and tests *(continued)*

- **A 99% confidence interval is calculated by SAS because the option CLB was requested on the model statement and a value of alpha-0.01 was specified.**
-

```
proc reg data=one LINEPRINTER; ID ObsID DBH;
    TITLE2 'Simple linear regression';
    model Weight = Dbh / p xpx i influence clb
                  alpha=0.01;
         Slope:Test DBH = 200;
         Joint:TEST intercept = 0, DBH = 200;
```

Parameter estimates and tests *(continued)*

- **Output with 99% confidence interval.**

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	99% Confidence Limits	
Intercept	1	-729.39630	55.69366	-13.10	<.0001	-879.18914	-579.60346
Dbh	1	178.56371	8.57640	20.82	<.0001	155.49675	201.63067

Parameter estimates and tests *(continued)*

- **A t-test of an hypothesized value for the slope would be calculated as**

$$t = \frac{b_1 - \beta_{1|H_0}}{S_{b_1}}$$

Parameter estimates and tests *(continued)*

- **SAS automatically provides a t-test of each parameter against an hypothesized value of zero, the most common test.**

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	99% Confidence Limits	
Intercept	1	-729.39630	55.69366	-13.10	<.0001	-879.18914	-579.60346
Dbh	1	178.56371	8.57640	20.82	<.0001	155.49675	201.63067

- **T values and P values are**
  - **Intercept : t=-13.097, P value < 0.0001**
  - **Slope : t=20.820, P value < 0.0001**

Parameter estimates and tests *(continued)*

- **Interpretation: The slope and intercept differ from zero. Therefore, the line does not pass through the origin, and the line is not "flat", basically the regression line is an improvement over the original flat line fitted by the correction factor.**
- **Other values may be of interest besides zero. These can be tested by hand, or with at "TEST" statement in SAS.**

Parameter estimates and tests *(continued)*

- I added an additional, optional, test. I decided to test two specific hypotheses about the regression coefficients.

```
87         Slope:Test DBH = 200;
88         Joint:TEST intercept = 0, DBH = 200;
```

- SAS provides a mechanism to do this. The statement "TEST DBH = 200;" is added to the program after the model statement. The test outputs the test result (in this program the output follows the list of observation diagnostics).

Parameter estimates and tests *(continued)*

Test Slope Results for Dependent Variable Weight

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	93041	6.25	0.0162
Denominator	45	14893		

- **This tests the hypothesis H_0: $\beta_{DBH} = 200$, and you can see that it is rejected here. SAS used an F test to test this (more flexible), we would probably use a t-test (computationally and conceptually easier).**

Parameter estimates and tests *(continued)*

Test Joint Results for Dependent Variable Weight

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	17479620	1173.67	<.0001
Denominator	45	14893		

- This tests the second hypothesis, a joint test of the two hypotheses
- H_0: $\beta_0 = 0$ and H_0: $\beta_{DBH} = 200$.
- This is a two degree of freedom test.

Other useful information

- **Other useful output from PROC REG includes observation diagnostics, residual plots and the ability to output residuals for testing.**
- **AS YOU KNOW, WE TEST NORMALITY OF THE RESIDUALS, NOT THE RAW DATA!**
-

Observation diagnostics

- There are a few diagnostics calculated from individual observations that are of interest.
- First the residuals are of interest only for their sign. Long strings of residuals with the same sign can indicate either curvature or a lack of independence.
- Since we don't know what constitutes an overly large residual, these are not very useful for detecting outliers.

Observation diagnostics *(continued)*

- Another value of interest is the standardized residuals, in SAS the values "STUDENT" and "RSTUDENT".
  - These are standardized residuals, and should have a mean of zero and a variance of one. They should follow a t distribution, so that for our example with 45 observations we expect that 99% would be between $\pm$2.690.

Observation diagnostics *(continued)*

- The HAT diag values.
  - Hat diag is a relative measure of how far an X value is from the center of the X space. A high value indicates an unusual value of X. This is not necessarily bad, but unusual values should be examined for correctness.

Observation diagnostics *(continued)*

- The HAT diag values (continued).
  - The hat diag values will sum to "p", where p is the number of parameters estimated in the model (2 for SLR).
  - The mean of the hat diag values will be p/n, and any values more than twice this value are considered "large". Again, this is not necessarily a problem.

Observation diagnostics *(continued)*

- **Influence diagnostics examine how the regression would change if an observation were removed from the analysis. If an observation is removed an the regression does not change, the observation is not influential. If the regression changes a lot, the observation is very influential.**
-

Observation diagnostics *(continued)*

- Influence diagnostics (continued)
  - DFFITS measures the change in terms of the "fit", as judged by the predicted (Yhat) value. If a point is removed and Yhat changes a lot, the point is influential.
  - for small to medium size databases, DFFITS should not exceed 1, while for large databases it should not exceed 2*sqrt(p/n)

Observation diagnostics *(continued)*

- **Influence diagnostics (continued)**
  - **DFBETAS measures the change in terms of the "fit", as judged by changes in b_0 and b_1. If a point is removed and b_0 or b_1 change a lot, the point is influential.**
  - **for small to medium size databases, DFBETAS should not exceed 1, while for large databases it should not exceed 2/sqrt(n)**

Observation diagnostics

- **Look for RSTUDENT values over 2.7**
- **Look for Hat diag values over 0.04**
- **Look for DFFITS & DFBETAS over 1.**

Obs	OBSID	Dep Var WEIGHT	Predict Value	Residual	Rstudent	Hat Diag H	Cov Ratio	Dffits	INTERCEP Dfbetas	DBH Dfbetas
1	i	58.0000	-104.4	162.4	1.3837	0.0560	1.0176	0.3372	0.3180	-0.2656
2	U	66.0000	-68.7106	134.7	1.1368	0.0510	1.0402	0.2635	0.2450	-0.2012
3	I	70.0000	-68.7106	138.7	1.1716	0.0510	1.0365	0.2716	0.2525	-0.2073
4	K	99.0000	-32.9978	132.0	1.1105	0.0464	1.0378	0.2448	0.2236	-0.1801
5	p	60.0000	-15.1414	75.1414	0.6255	0.0442	1.0751	0.1345	0.1216	-0.0968
6	R	76.0000	-15.1414	91.1414	0.7603	0.0442	1.0661	0.1634	0.1478	-0.1177
7	n	84.0000	-15.1414	99.1414	0.8280	0.0442	1.0610	0.1780	0.1609	-0.1282
8	Q	89.0000	-15.1414	104.1	0.8705	0.0442	1.0576	0.1871	0.1692	-0.1347
9	H	84.0000	20.5713	63.4287	0.5262	0.0401	1.0761	0.1076	0.0949	-0.0737
10	w	100.0	38.4277	61.5723	0.5102	0.0382	1.0748	0.1017	0.0885	-0.0678
11	G	125.0	38.4277	86.5723	0.7195	0.0382	1.0624	0.1435	0.1247	-0.0955
12	r	74.0000	74.1404	-0.1404	-0.0012	0.0348	1.0837	-0.0002	-0.0002	0.0001
13	h	121.0	74.1404	46.8596	0.3871	0.0348	1.0763	0.0735	0.0617	-0.0458
14	x	122.0	74.1404	47.8596	0.3954	0.0348	1.0760	0.0751	0.0631	-0.0468
15	A	194.0	163.4	30.5777	0.2515	0.0278	1.0728	0.0426	0.0315	-0.0207
16	T	194.0	199.1	-5.1350	-0.0422	0.0258	1.0735	-0.0069	-0.0047	0.0029
17	L	200.0	199.1	0.8650	0.0071	0.0258	1.0736	0.0012	0.0008	-0.0005
18	B	280.0	234.8	45.1522	0.3709	0.0241	1.0651	0.0583	0.0363	-0.0199
19	F	229.0	252.7	-23.7041	-0.1944	0.0234	1.0692	-0.0301	-0.0177	0.0090
20	E	200.0	270.6	-70.5605	-0.5806	0.0228	1.0542	-0.0887	-0.0490	0.0228
21	m	209.0	270.6	-61.5605	-0.5061	0.0228	1.0580	-0.0773	-0.0427	0.0199
22	v	210.0	270.6	-60.5605	-0.4978	0.0228	1.0584	-0.0760	-0.0420	0.0196
23	M	214.0	270.6	-56.5605	-0.4647	0.0228	1.0599	-0.0710	-0.0392	0.0183
24	a	174.0	288.4	-114.4	-0.9471	0.0223	1.0275	-0.1430	-0.0736	0.0305
25	k	220.0	288.4	-68.4169	-0.5627	0.0223	1.0546	-0.0850	-0.0437	0.0181
26	C	296.0	342.0	-45.9860	-0.3773	0.0214	1.0620	-0.0558	-0.0217	0.0041

Observation diagnostics *(continued)*

Obs	OBSID	Dep Var WEIGHT	Predict Value	Residual	Rstudent	Hat Diag H	Cov Ratio	Dffits	INTERCEP Dfbetas	DBH Dfbetas
27	l	342.0	342.0	0.0140	0.0001	0.0214	1.0688	0.0000	0.0000	-0.0000
28	J	224.0	359.8	-135.8	-1.1286	0.0213	1.0094	-0.1665	-0.0572	0.0043
29	P	238.0	359.8	-121.8	-1.0094	0.0213	1.0209	-0.1489	-0.0512	0.0038
30	O	297.0	359.8	-62.8424	-0.5163	0.0213	1.0559	-0.0761	-0.0262	0.0020
31	e	226.0	377.7	-151.7	-1.2648	0.0213	0.9950	-0.1865	-0.0556	-0.0042
32	j	278.0	377.7	-99.6987	-0.8228	0.0213	1.0366	-0.1213	-0.0362	-0.0027
33	u	345.0	431.3	-86.2678	-0.7108	0.0219	1.0452	-0.1063	-0.0169	-0.0175
34	o	313.0	467.0	-154.0	-1.2856	0.0228	0.9942	-0.1962	-0.0133	-0.0500
35	d	408.0	520.5	-112.5	-0.9326	0.0248	1.0314	-0.1488	0.0092	-0.0562
36	D	462.0	592.0	-130.0	-1.0829	0.0290	1.0220	-0.1870	0.0400	-0.0963
37	y	539.0	645.5	-106.5	-0.8857	0.0331	1.0442	-0.1639	0.0508	-0.0979
38	N	712.0	663.4	48.5993	0.4015	0.0347	1.0756	0.0761	-0.0258	0.0473
39	S	614.0	699.1	-85.1134	-0.7072	0.0381	1.0631	-0.1408	0.0551	-0.0936
40	b	745.0	717.0	28.0302	0.2319	0.0400	1.0869	0.0473	-0.0197	0.0324
41	c	814.0	752.7	61.3175	0.5096	0.0440	1.0814	0.1094	-0.0502	0.0786
42	s	515.0	806.3	-291.3	-2.6020	0.0508	0.8277	-0.6022	0.3106	-0.4592
43	z	815.0	842.0	-26.9644	-0.2250	0.0559	1.1053	-0.0547	0.0300	-0.0431
44	t	766.0	931.2	-165.2	-1.4200	0.0702	1.0285	-0.3901	0.2399	-0.3257
45	f	1675.0	1306.2	368.8	3.7355	0.1572	0.7154	1.6135	-1.2320	1.5004
46	g	1491.0	1341.9	149.1	1.3511	0.1678	1.1587	0.6067	-0.4681	0.5669
47	q	1692.0	1431.2	260.8	2.5208	0.1959	0.9932	1.2444	-0.9822	1.1749

- **Large Hat diag values on both ends of the regression**
- **Large DFFITS and DFBETAS for observation 45 & 47.**
- **Large RSTUDENT for observation 45**

Residual plot

- Residual plots are a useful tool for detecting various problems
  - Outliers
  - Curvature
  - Non-homogeneous variance
  - and more
- This was produced by the statements,

```
89                    options ls=78 ps=45;
90          plot residual.*predicted.=obsid; run;
```

Residual plot

```
           -+------+------+------+------+------+------+------+------+------+------+--
       400 +                                                                 +
           |                                                                 |
           |                                                     f           |
           |                                                                 |
           |                                                                 |
       300 +                                                                 +
           |                                                                 |
           |                                                         q       |
           |                                                                 |
           |                                                                 |
       200 +                                                                 +
           |                                                                 |
  RESIDUAL |   i                                                             |
           |     ?K                                               g          |
R          |                                                                 |
e      100 +      ?                                                          +
s          |      p G                                                        |
i          |        ?                        c                               |
d          |          ?  A B              N                                  |
u          |                                b                                |
a        0 +          r   ?    l                                             +
l          |                F                   z                            |
           |                   C                                             |
           |                ?k  O                                            |
           |                E     u        S                                 |
      -100 +                    j         y                                  +
           |                 a  P    d  D                                    |
           |                    J                                            |
           |                    e  o                t                        |
           |                                                                 |
      -200 +                                                                 +
           |                                                                 |
           |                                                                 |
           |                                                                 |
           |                                                                 |
      -300 +                                   s                             +
           -+------+------+------+------+------+------+------+------+------+------+--
          -200     0     200    400    600    800   1000   1200   1400   1600
                         Predicted Value of WEIGHT     PRED
```

Univariate tests & graphics

- Basic PROC UNIVARIATE information.

Moments			
N	47	Sum Weights	47
Mean	0	Sum Observations	0
Std Deviation	120.703619	Variance	14569.3637
Skewness	0.47869472	Kurtosis	1.04153074
Uncorrected SS	670190.732	Corrected SS	670190.732
Coeff Variation	.	Std Error Mean	17.6064324

Basic Statistical Measures

Location		Variability	
Mean	0.00000	Std Deviation	120.70362
Median	-0.14041	Variance	14569
Mode	.	Range	660.02160
		Interquartile Range	161.40929

Univariate tests & graphics

- PROC UNIVARIATE test of normality.

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.973389	Pr < W	0.3544
Kolmogorov-Smirnov	D	0.084574	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.044081	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.354877	Pr > A-Sq	>0.2500

Univariate tests & graphics *(continued)*

Univariate Procedure (*continued*)

```
   Stem Leaf                          #  Boxplot
      3 7                             1     0
      3
      2 6                             1     |
      2                                     |
      1 56                            2     |
      1 00334                         5     |
      0 5555666899                   10  +-----+
      0 0033                          4  |  +  |
     -0 3210                          4  *-----*
     -0 997766665                     9  +-----+
     -1 4321110                       7     |
     -1 755                           3     |
     -2                                     |
     -2 9                             1     |
        ----+----+----+----+
    Multiply Stem.Leaf by 10**+2
```

Univariate tests & graphics *(continued)*

Univariate Procedure (*continued*)

```
                       Normal Probability Plot
     375+                                               *
        |                                                  +
        |                                          *  ++++
        |                                         ++++
        |                                     +++*
        |                                 +*****
        |                             *****
        |                        +****
        |                    ++***
        |                ******
        |           ******
        |      * *+*+
        |    ++++
    -275+ ++*+
         +----+----+----+----+----+----+----+----+----+----+
             -2        -1         0        +1        +2
```

Summary

- For this relationship a significant correlation exists between the diameter of the tree and the weight of the wood harvested from the tree. In fact, we get 178.6 pounds of wood for each additional inch of diameter
  - $P(161.328 \le \beta_1 \le 195.872) = 0.95$
- The equation to predict wood harvest from diameter is $Y_i = -729.4 + 178.6*X_i$

Summary *(continued)*

- We might expect that a tree with a diameter of zero to have a weight of zero, but our model says that the weight for such a tree would actually be -729.4. The first question is whether this is STATISTICALLY SIGNIFICANT in differing from the hypothesized value of zero. It is ($P<0.0001$).
- This is impossible, so either there is something about tree growth we don't understand, or we do not have a good model.

Summary *(continued)*

- So we try to evaluate our model.
- Are the observations correct and reasonable?
- Examine the RSTUDENT values.
  - Potential problem for Obs #45
- Examine the residual plot. This plot appears to show that the line is actually curved and possibly has non-homogeneous variance!!!

Summary *(continued)*

- The Hat diag values indicated that the values on the end of the regression were possibly "unusual".
- This is not uncommon for simple linear regression, which is kind of one dimensional for X. This statistics will be more useful for multiple regression.

Summary *(continued)*

- **The influence diagnostics indicated that a number of observations were "influential".**
  - **If the observations are correct, and not outliers, this is not a problem.**
  - **Also if an observation IS an outlier, but it is not influential, we don't have much of a problem.**
  - **Problems occur when an observation is BOTH and outlier and influential.**
- **Like observation #45!!!**

Summary *(continued)*

- Examine the PROC UNIVARIATE output for tests and graphics of normality and for outliers.
  - The Shapiro-Wilk test indicates the residuals do not depart from normality.
  - The graphics do not show a great departure from normality, but there is a possible outlier (observation "f", its #45).
  - The normal probability plot shows only one departure, and it appears to be the outlier on the upper end (#45 again).

Summary *(continued)*

- So this regression appears to fit "well". Everything is significant and the R^2 is pretty high, but there are a lot of problems.
- The basic problem is that we do not have the right model. The model should really have some curvature (we will cover this later). Then, observations that are outliers on the ends might fit right on the line.