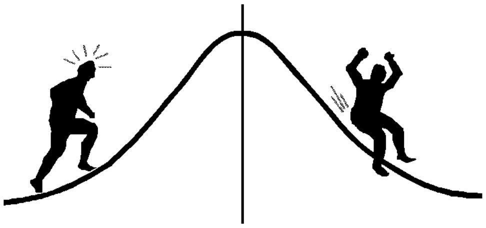

BIOMETRICS INFORMATION

(You're 95% likely to need this information)

PAMPHLET NO. # 28 DATE: July 12, 1990

SUBJECT: Simple regression with replication: testing for lack of fit

Textbooks usually describe simple regression without replication, i.e., for each X-value there is only one observed Y-value. Frequently though, researchers have two or more points for some or all of the X values. In this case, how does the analysis change?

Suppose that there are k different X_i values, each with n_i observed Y_{ij} values[1]. There are three likely models which could be fit:

1) Simple regression on the individual data points (X_i,Y_{ij}).
2) Simple regression on the means $(X_i,\overline{Y}_i)$.
3) Weighted regression on the means $(X_i,\overline{Y}_i)$ using the sample size n_i as the weight.

When the sample sizes are equal all three models will have the same parameter estimates. The weighted and individual regressions will always have the same parameter estimates, but when the sample sizes are unequal these will be different from those produced by the unweighted means regression. Whether the weighted or unweighted model should be used in the unbalanced case depends on the situation and this question will not be addressed in this pamphlet.

A big advantage of having replication with regression is the opportunity to test for lack of fit of the straight line model. The lack-of-fit test is conducted by separating the residual sums of squares (SSR) into two parts: a lack-of-fit sums of squares (SSF), and a within group sums of squares (SSE) also called the pure error.

The SSE can be determined from an ANOVA on Y_{ij} using X_i as the class variable. The error or residual sums of squares from this ANOVA is the SSE. To compare the three models above we can use the following terminology for the various sums of squares (SS).

Model:	Individual Data		Unweighted Means		Weighted Means	
Source of Variation	df	SS	df	SS	df	SS
Regression	1	SSLI	1	SSLU	1	SSLW
Residual	$\sum n_i$-2	SSRI	k-2	SSRU	k-2	SSRW
i) Lack of fit	k-2	SSF				
ii) Within Groups	$\sum(n_i$-1)	SSE				

[1] The index i identifies particular X-values and has values i = 1, 2, ... k, while j identifies a specific Y-value and has values j = 1, 2, ... n_i.

BRITISH COLUMBIA
Ministry of Forests
Research Program

These SS's have the following relationships:

a) SSF = SSRI - SSE

b) SSF = SSRW, regardless of sample size

hence SSE = SSRI - SSRW.

Also c) SSF = n SSRU , but only if the sample sizes are equal.

The sums of squares required for a lack-of-fit test can be calculated using two of the following three analyses:

1) Regression on the individual data
2) ANOVA using the X_i values as the class variable
3) Weighted regression on the means.

The F-value for the lack-of-fit test is calculated by:

$$F = \frac{\sum(n_i - 1)}{k\text{-}2} \left\{ \frac{SSF}{SSE} \text{ or } \frac{SSRI - SSE}{SSE} \text{ or } \frac{SSRW}{SSRI - SSRW} \text{ or } \frac{SSRW}{SSE} \right\}$$

with df = k-2, $\sum(n_i\text{-}1)$.

All three analyses can be accomplished with SAS using the following program (which includes some example data).

```
TITLE 'Different ways to analyse a replicated regression';
    DATA EXAMPLE;
      INPUT X Y @@;
    CARDS;
      10 70 10 78
      15 85
      20 90 20 97
      25 97 25 86 25 91
      30 95
      35 105;
TITLE3 'Regression analysis on individual data ignoring the replication';
    PROC REG DATA=EXAMPLE;
      MODEL Y = X; RUN;                              /* get SSRI */

TITLE3 'Obtaining the Within Sums of Squares (also known as Pure Error)';
    PROC ANOVA DATA=EXAMPLE;
      CLASS X; MODEL Y = X; RUN;                      /* get SSE */

TITLE3 'Weighted regression on the means';
    PROC MEANS N MEAN NWAY;
      CLASS X; VAR Y;
      OUTPUT OUT=MEANS N=NUM MEAN=MY;
    PROC REG DATA=MEANS;
      WEIGHT NUM;
      MODEL MY = X; RUN;                        /* get SSRW = SSF */

TITLE3 'Unweighted regression on the means';
    PROC REG DATA=MEANS;
      MODEL MY = X; RUN;
```

Now let us compare the output of the different procedures:

Model:	Individual Data		Unweighted Means		Weighted Means	
Source of Variation	df	SS	df	SS	df	SS
Total	9	930.40	5	542.48	5	813.23
Regression	1	699.09	1	477.54	1	699.90
Residual	8	231.31	4	65.04	4	114.14
i) Lack of fit	4	114.14				
ii) Within Groups	4	117.17				
Fitted equation:	Y = 66.24 + 1.077X		Y = 67.13 + 1.045X		Y = 66.24 + 1.077X	
R-square	0.7514		0.8801		0.8596	
Adjusted R-square	0.7203		0.8502		0.8246	

Note the following:

1) The best fit line is the same for the individual regression and the weighted regression, but since replication was unequal for each X_i-value, the estimated line for the unweighted regression is different.

2) The R-square values are markedly improved for the regressions on means. This occurs because the total variability in the data has been reduced by using means (note that the total df is reduced for the regressions on means). It is misleading to quote the R-square for the regression on the means instead of that from the individual data (especially if it is not stated that means were used). This is particularly true if the regression is to be used for prediction. A regression on the means will always suggest greater predictive ability than a regression on individual data because it predicts mean values, not individual values. And mean values are always less variable than are individual values.

3) The relationships between the various sums of squares are confirmed:

   i) SSF = 114.14 = SSRI - SSE = 231.31 - 117.17 = 114.14.

   ii) SSF = 114.14 = SSRW = 114.14.

   iii) SSE = 117.17 = SSRI - SSRW = 231.31 - 114.14 = 117.17.

The F-value to test for lack of fit of the regression is:

$$F = \frac{4}{4}\left\{\frac{114.14}{117.17} \text{ or } \frac{231.31\text{-}117.17}{117.17} \text{ or } \frac{114.14}{231.31\text{-}114.14} \text{ or } \frac{114.14}{117.17}\right\}$$

$= 0.974$ with df = 4, 4 which is not significant.

CONTACT: Wendy Bergerud
387-5676

NEW PROBLEM

Test the following two sets of data for lack of fit from a straight line regression model.

SET 1:		SET 2:	
X = 1	Y = 9, 4, 14	X = 1	Y = 9, 4, 14
X = 2	Y = 14	X = 2	Y = 15, 18, 9
X = 3	Y = 16, 18	X = 3	Y = 16, 12, 23
X = 4	Y = 27, 28, 20	X = 4	Y = 27, 28, 20
X = 5	Y = 19, 23	X = 5	Y = 19, 17, 27
X = 6	Y = 27	X = 6	Y = 24, 24, 33
X = 7	Y = 30, 31, 20	X = 7	Y = 30, 31, 20

What do you notice about the regressions for these two sets of data?