BIOMETRICS INFORMATION

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

PAMPHLET NO. # 29 DATE: November 13, 1990

SUBJECT: Simple Regression: Confidence intervals for a predicted X-value.

This pamphlet discusses the use of a regression line in reverse. For example, suppose that there are two methods available for measuring some quantity, say water content in soil samples. One method is very accurate but expensive and complicated to operate, while the second one is cheaper and faster but less accurate. If a relationship between the results of the two methods could be established then the faster method could be used for future samples, and the relationship used to estimate the value the more expensive method would have given. This is commonly referred to as a calibration problem.

Suppose we have a regression line $Y = a + bX$ obtained from n pairs of observations (x_i,y_i). Suppose further that we have m new observations $(y_1^*, y_2^*, \ldots , y_m^*)$ known to have arisen from the same but unknown X value, say x^*. We wish to estimate x^* and construct confidence limits for x^*. From the regression line, x^* can be estimated by:

$$\hat{x}^* = \frac{\bar{y}^* - a}{b}$$

where a and b are the intercept and slope of the fitted regression line and $\bar{y}^*$ is the mean of the m new observations. The $(1-\alpha)100\%$ confidence interval for x^* is

$$\bar{x} + \frac{1}{1-c^2}\left\{(\hat{x}^*-\bar{x}) \pm t\,\frac{S_{y\cdot x}}{b}\sqrt{\left[\frac{1}{m} + \frac{1}{n}\right](1-c^2) + \frac{(\hat{x}^*-\bar{x})^2}{\sum(x_i-\bar{x})^2}}\right\}$$

where $$c^2 = \frac{t^2 S^2_{y\cdot x}}{b^2\sum(x_i-\bar{x})^2}$$ (see note 1) ,

$S^2_{y\cdot x}$ is the mean square error (MSE), $\bar{x}$ and $\bar{y}$ are the means of the n original observations and t is obtained from the t-table with $(n-2)$ degrees of freedom and probability level $\alpha/2$.

A SAS program has been written to compute $\hat{x}^*$ and the associated 95% confidence interval. It requires three input variables, x, y and group where group identifies those y-values which are known to have the same x^*. For the n pairs of original observations, the group variable has a missing value (use . in SAS). Similarly, the m new y observations have missing x values.

[1] Note that c^2 is often close to zero so that $1/(1-c^2)$ is close to 1.

BRITISH COLUMBIA
Ministry of Forests
Research Program

The input data file would have the form:

x	y	group
x_1	y_1	.
x_2	y_2	.
⋮	⋮	⋮
x_n	y_n	.
.	y_1*	1
.	y_2*	1
⋮	⋮	⋮
.	y_m*	1

Application of the program is illustrated using an example adapted from Bowker & Lieberman (1965, p.241), "Gravimetric Determination of Calcium in the Presence of Magnesium" where X is the amount of CaO (mg) found using the present method and Y is the amount found by a new method. The task is to estimate the X values that correspond to the three groups of new Y observations (29.6, 29.3, 29.4), (30), and (22.0, 22.3). The SAS program and output are shown on the next page.

Using the first 10 observations, the SAS program computed the regression line to be:

$$Y = -0.29278 + 1.00652 X.$$

This fitted line was used to estimate x* for the three groups. For group 1, $\bar{y}^* = 29.433$, $\hat{x}^* = 29.5335$ and the 95% confidence interval for x* is (28.2690, 30.7719).The corresponding values for groups 2 and 3 are also given in the output.

The program presently computes 95% confidence intervals via the assignment statement within the last data step: `t = tinv(0.975, n-2).` If a different confidence level is required, change the bracketed value 0.975 to the desired value. For a (1-α)100% confidence interval, the appropriate value is (1-α/2). For example, use 0.995 for a 99% confidence interval.

An alternative approach to the calibration problem, namely the 'inverse method', is based on the regression of x on y. See Chow & Shao for a recent disscussion of the comparison of the 2 methods.

References:

Bowker, A. H. & Leiberman, G. J. (1965). *Engineering Statistics.* Prentice-Hall Inc., Englewood Cliffs, N.J.

Chow & Shao. (1990). *Applied Statistics*, **39**: 219-228.

CONTACT: Vera Sit or Wendy Bergerud
356-0435 387-5676

NEW PROBLEM

It is generally thought that the precentage of wormy fruits is greater on apple trees bearing a small crop. A group of 12 trees are investigated and the results are shown below.

Size of crop X (hundreds of fruit)	8	6	11	22	14	17	18	24	19	23	26	40
Percentage wormy, Y	59	58	56	53	50	45	43	42	39	38	30	27

Based on the above data, estimate the size of crop and the 90% confidence limits for a tree with 40% wormy fruit.

```
/*  SAS program to determine 95 % confidence intervals for predicted
    X-values given one or more Y-values  */

DATA ORIGDATA (keep = x y xy)
     NEWDATA (keep = y1 group);
      input x y group;
      if x = . then do;  y1 = y;  output NEWDATA; end;
        else do;       xy = x*y;  output ORIGDATA; end;
    CARDS;
    20.0  19.8    .
    22.5  22.8    .
    25.0  24.5    .
    28.5  27.3    .
    31.0  31.0    .
    33.5  35.0    .
    35.5  35.1    .
    37.0  37.1    .
    38.0  38.5    .
    40.0  39.0    .
      .   29.6    1
      .   29.3    1
      .   29.4    1
      .   30.0    2
      .   22.0    3
      .   22.3    3
;
PROC MEANS data=ORIGDATA noprint;
    output out = STATS (keep = n mx my mxy sxx syy)
             n = n mean = mx my mxy css = sxx syy;
PROC MEANS data=NEWDATA noprint nway;
    class group;
    output out = YSTATS (keep = m ynew group ) n = m mean = ynew;
PROC REG data=ORIGDATA noprint outest=EST ;
    model y = x;
DATA EST ;
    merge EST STATS ;
    sxy  = n*mxy - n*mx*my;             ahat = intercep;
    bhat = x;                           sdyx = _rmse_;
    keep ahat bhat sdyx n mx my mxy sxx syy sxy ;
```

```
DATA RESULT;
    if _n_=1 then set EST;
    set YSTATS;
    t    = tinv(0.975,n-2);              xnew = (ynew-ahat)/bhat;
    c2   = (t*sdyx/bhat)**2 /sxx;        c1   = 1 - c2;
    dx   = xnew - mx;
    temp = ((1/m + 1/n)*c1 + (dx ** 2)/sxx)**0.5;
    v    = t * sdyx * temp / bhat;       x1   = (dx - v)/c1 + mx;
    x2   = (dx + v)/c1 + mx;
    keep m ynew xnew x1 x2 group ;
PROC PRINT data=EST noobs;
    title1 'Point Estimate and 95% Confidence Interval of X on a
Given Y';
    title4 'Model: Y = ahat + bhat * X';
    var ahat bhat;
PROC PRINT data=RESULT noobs split='_';
    by group;                            title;
    label x1   ='Lower_Limit'            x2   ='Upper_Limit'
          xnew ='Estimated_X'            ynew ='Mean of_Given Y'
          m    ='No. of_Y-values_used';
run;
```

SAS Output for example:

```
    Point Estimate and 95% Confidence Interval of X on a Given Y

                     Model: Y = ahat + bhat * X

                        AHAT              BHAT

                      -0.29278          1.00652

----------------------------- GROUP=1 ------------------------------

          No. of
         Y-values   Mean of    Estimated    Lower       Upper
           used     Given Y        X        Limit       Limit

             3      29.4333     29.5335    28.2690     30.7719

----------------------------- GROUP=2 ------------------------------

          No. of
         Y-values   Mean of   Estimated     Lower       Upper
           used     Given Y        X        Limit       Limit

             1         30       30.0965    28.1053     32.0710

----------------------------- GROUP=3 ------------------------------

          No. of
         Y-values   Mean of    Estimated    Lower       Upper
           used     Given Y        X        Limit       Limit

             2        22.15     22.2974    20.5534     23.8947
```

PROBLEM FROM BI# 28

For Data Set # 1:

Model:	Individual Data		Unweighted Means		Weighted Means	
Source of Variation	df	SS	df	SS	df	SS
Total	14	862.00	6	290.00	6	690.00
Regression	1	582.06	1	38.00	1	582.06
Residual	13	279.94	4	252.00	5	107.94
i) Lack of fit	5	107.94				
ii) Within Groups	8	172.00				
Fitted equation:	Y = 8.12 + 2.97X		Y = 8.00 + 3.00X		Y = 8.12 + 2.97X	
R-square	0.6752		0.8690		0.8436	
Adjusted R-square	0.6503		0.8428		0.8123	

For Data Set # 2:

Model:	Individual Data		Unweighted Means		Weighted Means	
Source of Variation	df	SS	df	SS	df	SS
Total	20	1246.00	6	290.00	6	870.00
Regression	1	756.00	1	38.00	1	756.00
Residual	19	490.00	4	252.00	5	114.00
i) Lack of fit	5	114.00				
ii) Within Groups	14	376.00				
Fitted equation:	Y = 8.00 + 3.00X		Y = 8.00 + 3.00X		Y = 8.00 + 3.00X	
R-square	0.6067		0.8690		0.8690	
Adjusted R-square	0.5860		0.8428		0.8428	

What you should notice is that the mean for each X-value is the same in both datasets and that the unweighted regression on the means provided exactly the same ANOVA table and regression coefficients. The individual regression and weighted regressions agreed on sums of squares and on regression coefficients. When the dataset is balanced, as it is for the second dataset, the three methods provide identical estimates for the regression coefficients.