A GENERAL SAS MACRO FOR PERFORMING WEIGHTED LEAST SQUARES

Wanda H. Burton, Medical College of Virginia

The SAS procedure GLM provides an excellent means for performing least squares regression analysis when the usual model assumptions can be made. The model referred to is

$$Y = X\beta + \varepsilon,$$

where ε is normally distributed with mean 0 and variance $\sigma^2 I$.

The case to be considered here is that in which the observations remain independent, but their variances are not all equal. The form of the variance-covariance matrix is $\sigma^2 V$ where V is a diagonal matrix with unequal diagonal elements,

$$\sigma^2 V = \begin{bmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_n^2 \end{bmatrix}$$

A unique nonsingular symmetric matric P can be found such that

$$P^2 = V.$$

A transformation can then be made on our original model by premultiplying by P^{-1}, obtaining

$$P^{-1}Y = P^{-1}X\beta + P^{-1}\varepsilon$$

or

$$Z = Q\beta + f$$

with obvious substitutions. This model satisfies the necessary assumptions for carrying out the usual least squares regression analysis; that is, $f \sim N(0,\sigma^2 I)$.

The MACRO to be presented provides a thorough analysis for a simple linear regression model,

$$E(y_i) = \beta_0 + \beta_1 x_i.$$

Let us use the following notation for the variance of Y:

$$\mathrm{Var}(Y) = \sigma^2 V = \sigma^2 \begin{bmatrix} w_1 & & & 0 \\ & w_2 & & \\ & & \ddots & \\ 0 & & & w_n \end{bmatrix}$$

where the w's are known weights.

For this situation,

$$P^{-1} = \begin{bmatrix} 1/\sqrt{w_1} & & & 0 \\ & 1/\sqrt{w_2} & & \\ & & \ddots & \\ 0 & & & 1/\sqrt{w_n} \end{bmatrix}$$

and a simple transformation of the variables is appropriate. These calculations are carried out in the MACRO WT_REGR in statements 4-10, creating the variables Z,Q_0 and Q_1.

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad P^{-1}Y = Z = \begin{bmatrix} y_1/\sqrt{w_1} \\ y_2/\sqrt{w_2} \\ \vdots \\ y_n/\sqrt{w_n} \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix},$$

$$P^{-1}X = Q = (Q\_0, Q\_1) = \begin{bmatrix} 1/\sqrt{w_1} & X_1/\sqrt{w_1} \\ 1/\sqrt{w_2} & X_2/\sqrt{w_2} \\ \vdots & \vdots \\ 1/\sqrt{w_n} & X_n/\sqrt{w_n} \end{bmatrix}$$

These transformations are made after the regression variables, X and Y, and the weights, WT, are input separately and merged in statements 27-65. An identification variable, ID, is included in each of the input data sets and is necessary for the merger.

PROC GLM is used to compute the ordinary least squares regression of Z on Q_0 and Q_1. Notice that the MODEL statement is written with NOINT option because the vector of ones normally present in a simple linear regression model has been transformed to a non-constant vector, Q_0. The OUTPUT facilities are used for access to the predicted values and the residuals from the fitted line.

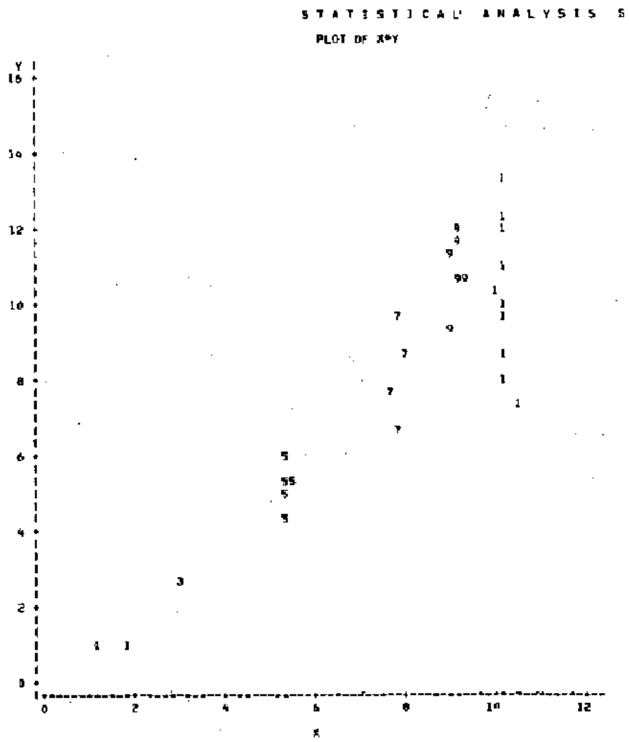
The MACRO WT_REGR includes a scatter of the original dependent variable (Y) against the independent variable (X). It provides plots which allow the user to examine the residuals from the analysis and thereby judge the effectiveness of applying the techniques of weighted regression to the data. A listing of the original variables with the scaled predicted values and residuals is included. This transformation back to the original scale of the data is carried out in statements 16-19.

Acknowledgment

The author would like to thank Maria Sauer for the use of labortary data from the Pulmonary Division of the Department of Medicine, Medical College of Virginia. This work was supported in part by grant 1 R01 01899-02 from the National Center for Health Services Research, HRA.

Reference

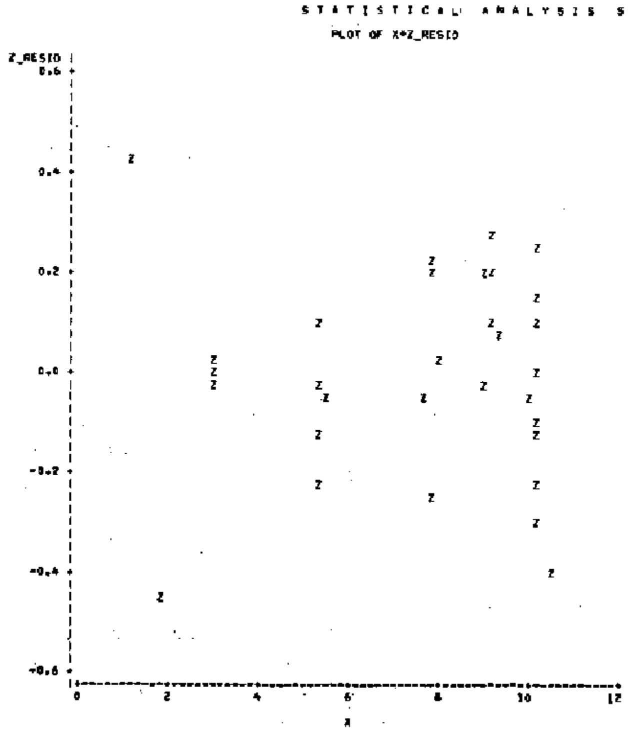
Draper, N. R. and Smith, H. Applied Regression Analysis, John Wiley & Sons, Inc., 1966, pp. 77-81.

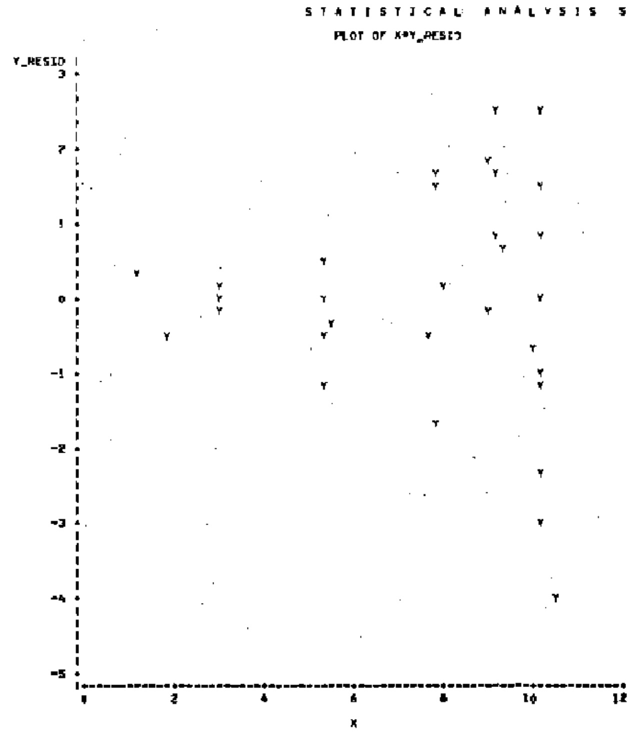



```
1                                    S T A T I S T I C A L   A N A L Y S I S   S
NOTE: THE JOB SORWTS HAS BEEN RUN UNDER RELEASE 76.2 OF SAS AT VIRGINIA COMMONWEALT

1          MACRO WT_REGR
2          PROC SORT DATA=ORIG; BY ID;
3          PROC SORT DATA=WEIGHTS; BY ID;
4          DATA TRANS;
5          MERGE ORIG WEIGHTS;
6          BY ID;
7          W=SQRT(WT);
8          Z=Y/W;
9          Q_0=1/W;
10         Q_1=X/W;
11         PROC SCATTER;
12         PLOT X*Y=ID / HPOS=75;
13         PROC GLM;
14         MODEL Z = Q_0 Q_1 / NOINT SS1 P;
15         OUTPUT OUT=PRED PREDICTED=Z_HAT RESIDUAL=Z_RESID;
16         DATA BACK;
17         SET PRED;
18         Y_HAT=W*Z_HAT;
19         Y_RESID=Y-Y_HAT;
20         PROC PRINT;
21         VAR ID WT X Y Y_HAT Y_RESID;
22         PROC SCATTER;
23         PLOT X*Z_RESID='Z' / HPOS=75;
24         PROC SCATTER;
25         PLOT X*Y_RESID='Y' / HPOS=75;
26         %
27         DATA ORIG;   INPUT X 1-5 Y 6-10 ID 11-12;
28         CARDS;

NOTE: DATA SET WORK.ORIG HAS 35 OBSERVATIONS AND 3 VARIABLES.
NOTE: THE DATA STATEMENT USED 0.41 SECONDS AND 96K.

64         DATA WEIGHTS;   INPUT ID WT;
65         CARDS;

NOTE: DATA SET WORK.WEIGHTS HAS 6 OBSERVATIONS AND 2 VARIABLES.
NOTE: THE DATA STATEMENT USED 0.14 SECONDS AND 96K.

72         ;
73         WT_REGR

NOTE: DATA SET WORK.ORIG HAS 35 OBSERVATIONS AND 3 VARIABLES.
NOTE: THE PROCEDURE SORT USED 1.34 SECONDS AND 110K.

NOTE: DATA SET WORK.WEIGHTS HAS 6 OBSERVATIONS AND 2 VARIABLES.
NOTE: THE PROCEDURE SORT USED 1.38 SECONDS AND 110K.

NOTE: DATA SET WORK.TRANS HAS 35 OBSERVATIONS AND 8 VARIABLES.
NOTE: THE DATA STATEMENT USED 0.27 SECONDS AND 104K.
```

STATISTICAL ANALYSIS SYSTEM

OBS	ID	WT	X	Y	Y_HAT	Y_RESID
1	1	1	1.15	0.99	0.5738	0.4162
2	1	1	1.90	0.98	1.4354	-0.4554
3	3	9	3.00	2.60	2.6990	-0.0990
4	3	9	3.00	2.67	2.6990	-0.0290
5	3	9	3.00	2.66	2.6990	-0.0390
6	3	9	3.00	2.78	2.6990	0.0810
7	3	9	3.00	2.80	2.6990	0.1010
8	5	25	5.34	5.92	5.3872	0.5328
9	5	25	5.38	5.35	5.4331	-0.0831
10	5	25	5.40	4.33	5.4561	-1.1261
11	5	25	5.40	4.89	5.4561	-0.5661
12	5	25	5.45	5.21	5.5135	-0.3035
13	7	49	7.70	7.68	8.0983	-0.4183
14	7	49	7.80	9.81	8.2132	1.5968
15	7	49	7.81	6.52	8.2246	-1.7046
16	7	49	7.85	9.71	8.2706	1.4394
17	7	49	7.87	9.82	8.2936	1.5264
18	7	49	7.91	9.81	8.3395	1.4705
19	7	49	7.94	8.50	8.3740	0.1260
20	9	81	9.03	9.47	9.6261	-0.1561
21	9	81	9.07	11.45	9.6721	1.7779
22	9	81	9.11	12.14	9.7180	2.4220
23	9	81	9.14	11.50	9.7525	1.7475
24	9	81	9.16	10.65	9.7755	0.8745
25	9	81	9.37	10.64	10.0167	0.6233
26	10	100	10.17	9.78	10.9357	-1.1557
27	10	100	10.18	12.39	10.9472	1.4428
28	10	100	10.22	11.03	10.9932	0.0368
29	10	100	10.22	8.00	10.9932	-2.9932
30	10	100	10.22	11.90	10.9932	0.9068
31	10	100	10.18	8.68	10.9472	-2.2672
32	10	100	10.50	7.25	11.3148	-4.0648
33	10	100	10.23	13.46	11.0047	2.4553
34	10	100	10.03	10.19	10.7749	-0.5849
35	10	100	10.23	9.93	11.0047	-1.0747


STATISTICAL ANALYSIS S
PLOT OF X*Z_RESID
Z_RESID
X


STATISTICAL ANALYSIS S
PLOT OF X*Y_RESID
Y_RESID
X