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 \( \varepsilon \) 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,

\[
\begin{bmatrix}
\sigma_1^2 & 0 \\
0 & \sigma_2^2 \\
\vdots & \ddots \\
0 & \cdots & \sigma_n^2
\end{bmatrix}
\]

A unique nonsingular symmetric matrix \( P \) 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,

\[ E(Y) = \beta_0 + \beta_1X. \]

Let us use the following notation for the variance of \( Y \):

\[
\begin{bmatrix}
w_1 \\
w_2 \\
\vdots \\
w_n
\end{bmatrix}
\]

\[ \text{Var}(Y) = \sigma^2 V = \sigma^2 
\begin{bmatrix}
w_1 & 0 \\
w_2 & 0 \\
\vdots & \ddots \\
0 & \cdots & 0 & w_n
\end{bmatrix}
\]

where the \( w_i \)'s are known weights.

For this situation,

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

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

\[
\begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{bmatrix}
\]

\[
\begin{bmatrix}
1 & X_1 \\
1 & X_2 \\
\vdots & \ddots \\
1 & X_n
\end{bmatrix}
\]

\[
\begin{bmatrix}
y_1/\sqrt{w_1} \\
y_2/\sqrt{w_2} \\
\vdots \\
y_n/\sqrt{w_n}
\end{bmatrix}
\]

\[
\begin{bmatrix}
1/\sqrt{w_1} & X_1/\sqrt{w_1} \\
1/\sqrt{w_2} & X_2/\sqrt{w_2} \\
\vdots & \ddots \\
1/\sqrt{w_n} & X_n/\sqrt{w_n}
\end{bmatrix}
\]

These transformations are made after the regression variables, \( X \) and \( Y \), and the weights, \( Wt \), are input separately and merged in statements 27-55. 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 Saucer for the use of laboratory data from the Pulmonary Division of the Department of Medicine, Medical College of Virginia. This work was supported in part by grant 1 R01 01893-02 from the National Center for Health Services Research, HRA.

Reference

STATISTICAL ANALYSIS B
PLLOT OF IN(RESID)

STATISTICAL ANALYSIS B
PLLOT OF IN(RESID)