GENERALIZED LEAST SQUARES REGRESSION WITH PARTIALLY CORRELATED DATA

Andrew J. R. GILLESPIE[1], Syracuse University Research Consulting Services

ABSTRACT

One assumption of the ordinary least squares (OLS) regression model is that the observations on the response variable Y are statistically uncorrelated. In data sets obtained from cluster sampling (usually a more cost-effective sampling method than simple random sampling), this assumption is violated; observations may be uncorrelated between clusters, but correlated within clusters. Application of OLS methods to such data can lead to serious underestimates of the precision of estimated regression coefficients.

Many methods have been proposed for dealing with such intracluster correlation, including generalized least squares (GLS), random coefficient regressions, and jackknife estimators of variance. This paper describes two algorithms written in the SAS Interactive Matrix Language (SAS/IML™) to fit GLS models based on work by Scott and Holt (1984) and Fuller and Battese (1973). The algorithms pass through the data twice. The first pass is used to obtain an estimate of intracluster correlation; the second pass transforms the data into a form suitable for analysis by the common OLS procedures.

INTRODUCTION

The general linear model, expressed in matrix form, is commonly written as

$$Y = \mathbf{X}'\beta + \varepsilon \qquad (1)$$

where Y is an observation on some response variable of interest, $\mathbf{X}$ is a p by 1 vector of observations on p predictor variables used to estimate Y, β is a p by 1 vector of unknown population parameters, ε is a random errors expressing the difference between $\mathbf{X}'\beta$ and Y, and $\mathbf{X}'$ denotes the transposed vector $\mathbf{X}$.

This model may be used to derive estimates of the response variable Y as a function of a set of predictor variables X_k and an constant parameter vector β. Under certain assumptions, the ordinary least squares (OLS) estimate $\mathbf{b}_0$ of β is unbiased and of minimum variance among all unbiased estimators. These assumptions are:

1. The predictor variables X_k ($k = 1$ to p) are fixed and known without error;
2. The expected value of ε given $\mathbf{X}$ is 0, i.e. $E(\varepsilon|\mathbf{X}) = 0$;
3. The conditional variance of ε given $\mathbf{X}$ is a constant parameter, i.e. $Var(\varepsilon|\mathbf{X}) = \sigma^2$;
4. The ε_i are statistically uncorrelated, i.e. $Cov(\varepsilon_i, \varepsilon_j: i \neq j) = 0$.

To estimate the variance of Y for hypothesis testing, it is necessary to assume that the ε_i are independent random variables, or are from a Normal (Gaussian) probability distribution. Assumptions 3 and 4 together imply a specific variance-covariance (or "dispersion") matrix structure for the vector ε of residuals, namely

$$\mathbf{D}(\varepsilon) = \sigma^2\mathbf{I} \qquad (2)$$

where $\mathbf{I}$ = an n by n identity matrix. However, the assumptions described above are seldom simultaneously satisfied in real life; rather, the assumptions are taken to be approximately true, and departures are assumed small enough to be unimportant.

One common violation of these assumptions occurs as a result of cluster sampling. In cluster sampling, a population is first subdivided into overlapping or non-overlapping clusters. For example, a forest may be conceptually divided into an infinite number of plots as determined by locations of points as plot centers. A sample of m clusters is selected, then some or all of the elements in each cluster are measured to obtain the sample observations. The result is that observations within clusters will tend to be more similar to each other than observations from different clusters; i.e., there exists a certain amount of <u>intracluster correlation</u> which violates assumption 4 above. Instead of the dispersion structure of equation (2), the covariance between observations from the same cluster is non-zero, so the dispersion matrix for a sample consisting of m clusters takes a

block diagonal form:

$$\mathbf{D(e)} = \begin{bmatrix} \mathbf{V}_1 & 0 & 0 & \cdots & 0 \\ 0 & \mathbf{V}_2 & 0 & \cdots & 0 \\ 0 & 0 & \mathbf{V}_3 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{V}_m \end{bmatrix}$$

where $\mathbf{V}_i$ is an n_i by n_i matrix which represents the correlation structure within cluster i, n_i is the number of sample elements in cluster i, and 0 is a matrix of 0's of convenient order which implies the absence of correlations between clusters. $\mathbf{V}_i$ has the form

$$V_i = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{bmatrix}$$

where ρ = a measure of the intracluster correlation among the observed residuals $e_{ij} = (Y_{ij} - \hat{Y}_{ij})$. Note that ρ measures correlation after accounting for the effects of X on Y.

Several simulation studies (Briggs, 1981; Kotimaki 1981; Gillespie, 1985) have shown that application of ordinary least squares procedures to samples containing such intracluster correlation may result in an underestimate of σ_{bb}, the variance-covariance matrix of the estimated least squares parameter vector $\mathbf{b}_o$. An alternative estimator of β is the generalized least squares estimator

$$\beta_G = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}) \qquad (3)$$

where **X**, **Y**, and **V** are as defined above. An operational estimator of β_G requires an estimate of ρ, the intracluster correlation coefficient. This paper discusses two different approaches for dealing with intracluster correlation.

The first approach, discussed by Scott and Holt (1983), works with an estimate of the matrix **V** in equation (3) to calculate an estimate $\mathbf{b}_G$ of β_G. A modified form of an equation given in Cochran (1977) is used to estimate the intracluster correlation coefficient ρ from a sample with unequal cluster size. The SAS/IML routine first fits the OLS regression of Y on **X** to obtain the observed residuals e_{ij}. Next, one must pass through the data one cluster at a time to complete the calculation of $\hat{\rho}$ = the estimate of ρ:

$$\hat{\rho} = \frac{2n \sum_{i=1}^{m} \sum_{j<u} (e_{ij})(e_{iu})}{\sum_{i=1}^{m} n_i (n_i-1)(n-p) S_{yy|x}} \qquad (4)$$

Theoretically, once ρ is estimated, one could construct **Y**, **X**, and **V** and proceed by the usual matrix manipulations to derive the parameter estimates. However, **V** is a square matrix of dimension n = sample size. For an interesting problem where n is large (say over 300), **V** might be too large for operations such as inversion or multiplication. One can take advantage of the block diagonal structure of **V** and proceed iteratively, cluster by cluster, since it can be shown that

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} = \Sigma(\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i) \qquad (5)$$

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} = \Sigma(\mathbf{X}_i\mathbf{V}^{-1}{}_i\mathbf{Y}_i) \qquad (6)$$

where $\mathbf{X}_i$, $\mathbf{V}_i$, and $\mathbf{Y}_i$ correspond to the subsets of **X**,**V**, and **Y** contributed by cluster i (i = 1 to m). Thus the second pass through the data consists of constructing the submatrices $\mathbf{X}_i$, $\mathbf{V}_i$, and $\mathbf{Y}_i$, then summing the appropriate products over all m clusters. The resulting products are then used to obtain the following parameter estimates:

$$\mathbf{b}_G = [\Sigma(\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i)]^{-1}\,[\Sigma\mathbf{X}_i\mathbf{V}_i^{-1}\mathbf{Y}_i)] \qquad (7)$$

an estimate of the GLS regression coefficient vector;

$$S_{yy|x} = [\{\Sigma(\mathbf{Y}_i'\mathbf{V}_i^{-1}\mathbf{Y}_i)\} - \{\mathbf{b}_G'(\mathbf{X}_i\mathbf{V}_i^{-1}\mathbf{Y}_i))\}]/(n-p) \qquad (8)$$

an estimate of the conditional variance of Y given X; and

$$\mathbf{S}_{bb} = S_{yy|x}\,[(\Sigma(\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i))^{-1}] \qquad (9)$$

an estimate of the covariance matrix of $\mathbf{b}_G$

The second approach, proposed by Fuller and Battese (1983), involves the construction of a transformational matrix **T**. This matrix is used to premultiply **Y** and **X**

to produce transformed matrices, say **U** and **V**. The standard OLS calculations, applied to the transformed matrices **U** and **V** , yield parameter estimates which account for the intracluster correlation effect. Like the **V** matrix in model 1, **T** is block diagonal, where $\mathbf{T}_i$ is constructed for each cluster i. $\mathbf{T}_i$ is defined as

$$\mathbf{T}_i = \mathbf{I}_{ni} - (\{1-[\sigma^2_e/(\sigma^2_e+n_i\sigma^2_u)]^{1/2}\}[1/n_i(J_{ni})]) \quad (10)$$

where σ^2_e is a variance component due to the variation of the sampled individuals, σ^2_u is a variance component due to the correlation among individuals within a cluster, and J_{ni} is an n_i by n_i matrix of 1's.

The authors suggested estimating the variance components σ^2_e and σ^2_u by the "fitting of constants" method outlined by Searle (1971). The first step is to regress the Y deviations from cluster means $(Y_{ij} - Y_{i.})$ on the analogous X deviations $(X_{ijk} - X_{i.k})$, where $Y_{i.}$ and $X_{i.k}$ denote the means in cluster i of Y and X_k respectively. This regression includes only those X variables for which the deviations are not 0 (i.e., ignoring the column of ones included for an intercept, since 1-1=0 in each case). Then σ^2_e is unbiasedly estimated by

$$\widehat{\sigma^2}_e = \mathbf{e'e}/(n - m - p + \delta) \quad (11)$$

where **e** = the vector of observed residuals from the above regression, and δ = the number of predictor variables for which all deviations are 0 (ie, the number of predictor variables ignored in the regression of residuals). This computation requires one pass through the data, to calculate the cluster residuals. The regression can then be done with normal matrix procedures.

The second variance component, σ^2_u, is unbiasedly estimated by

$$\widehat{\sigma^2}_u = \frac{\mathbf{u'u} - (n - p)\widehat{\sigma^2}_e}{n - \text{trace}[(\mathbf{X'X})^{-1} \sum (n_i{}^2\mathbf{x}_i.'\mathbf{x}_i.)]} \quad (12)$$

where **u** is the residual vector from the least squares regression of Y on X, and $\mathbf{x}_i.$ is the mean vector of the observations on X_{ijk} in cluster i.

At this point, one has enough information to estimate $\mathbf{T}_i$ for each cluster. The original cluster matrices $\mathbf{Y}_i$ and $\mathbf{X}_i$ are premultiplied by $\mathbf{T}_i$ and concatenated, yielding transformed vectors **U** = **TY** and **V** = **TX**. This requires a second pass through the data. Finally, the generalized least squares model is fit by applying the OLS procedures to the transformed matrices **U** and **V**.

SUMMARY

Two algorithms have been developed using the SAS Interactive Matrix Language (SAS/IML) to fit Generalized Least Squares (GLS) models for data collected by a cluster sample. Such models are better suited to data from complex samples than the common Ordinary Least Squares (OLS) model, because the GLS model accounts for the effect of intracluster correlation among observations from the same cluster.

Copies of the SAS/IML code used to estimate the GLS models are available on request from Syracuse University Research Computing Services, 732 Ostrom Avenue, Syracuse, NY 13244. Bitnet address: RESCON@SUVM.

LITERATURE CITED

Briggs, E.F. 1980. "Biomass table construction by regression modified for cluster sampling." M.S. Thesis, SUNY College of Environmental Science and Forestry, Syracuse, NY. 326 pp.

Cochran, W.G. 1977. Sampling Techniques. 3rd. edition. New York: John Wiley & Sons. 428 pp.

Gillespie, A.G. 1985. "Estimation of biomass tables by cluster sampling: Results of a simulation study." M.S. Thesis, SUNY College of Environmental Science and Forestry, Syracuse, NY. 314 pp.

Fuller, W.A., and G.E. Battese. 1973. "Transformations for estimation of linear models with nested-error structure". Journal of the American Statistical Association 68:626-632.

Scott, A.J., and D. Holt. 1982. "The effect of two stage sampling on ordinary least squares methods". Journal of the American Statistical Association 77:848-854.

Searle, S.R. 1971. "Topics in variance component estimation". Biometrics 27:1-76.

[1]Current address: Institute of Tropical Forestry, Box 25000, Rio Piedras, Puerto Rico 00928-2500.