Aitken's Generalized Least Squares

James L. Powell
Department of Economics
University of California, Berkeley

The Generalized Regression Model

Departures from the standard assumption of a scalar covariance matrix – that is, $\mathbf{V}(\mathbf{y}) = \sigma^2\mathbf{I}$ – yield a particular extension of the classical regression model known as the *generalized regression model*, or sometimes "generalized classical regression model." A concise statement of the assumptions on the N-dimensional vector $\mathbf{y}$ of dependent variables and the $N \times K$ matrix $\mathbf{X}$ of regressors is:

1. (Linear Expectation) $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, for some K-dimensional vector of unknown "regression coefficients" $\boldsymbol{\beta}$.

2. (Noncalar Covariance Matrix) $\mathbf{V}(\mathbf{y}) \equiv \mathbf{E}\left((\mathbf{y} - \mathbf{E}(\mathbf{y}))(\mathbf{y} - \mathbf{E}(\mathbf{y}))'\right) = \boldsymbol{\Sigma}$, for some positive definite $(N \times N)$ matrix $\boldsymbol{\Sigma}$.

3. (Nonstochastic Regressors) The $(N \times K)$ matrix $\mathbf{X}$ is nonrandom.

4. (Full Rank Regressors) The rank of the matrix $\mathbf{X}$ is K, or, equivalently, the $(K \times K)$ matrix $(\mathbf{X}'\mathbf{X})$ is invertible.

To this set of assumptions is sometimes appended the following assumption, which yields the *generalized normal regression model*:

5. (Multinormality) The vector $\mathbf{y}$ has a multivariate normal distribution.

Often matrix $\boldsymbol{\Sigma}$ will be written as

$$\boldsymbol{\Sigma} = \sigma^2\boldsymbol{\Omega},$$

where σ^2 is an unknown scaling parameter; as is true for the best linear unbiased estimator for the classical regression model (namely, classical LS), the BLU estimator for the generalized regression model does not depend upon the value of σ^2. Typically the matrix $\boldsymbol{\Omega}$ will also depend upon unknown parameters, but extension of Gauss-Markov arguments to this model will require it to be known exactly.

There are several varieties of linear models which yield a nonscalar covariance matrix, each with its own nomenclature; four leading examples, to be studied in greater detail later, are:

Seemingly Unrelated Regression Model: The matrix $\mathbf{\Omega}$ has a "Kronecker product form" (to be defined later). That is $\mathbf{\Omega} = \mathbf{\Sigma} \otimes \mathbf{I}_N$, where $\mathbf{\Sigma}$ is a $J \times J$ covariance matrix and $\mathbf{I}_N$ is an $N \times N$ identity matrix, where now $\dim(\mathbf{y}) = J \times N$. Such a model arises when N observations on J dependent variables y_{ij} are each assumed to be generated from separate linear models

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta}_j + \varepsilon_{ij},$$

where the errors ε_{ij} are assumed to satisfy the assumptions of the standard regression model for j fixed, but with $Cov(\varepsilon_{ij}, \varepsilon_{ik}) \equiv \sigma_{jk}$, which might be nonzero if equations j and k are "related" through common components in their error terms.

Models of Heteroskedasticity ("different scatter"): The matrix $\mathbf{\Omega}$ is diagonal, i.e., $\mathbf{\Omega} = diag[\omega_{ii}]$; this arises from the model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + c_i \cdot \varepsilon_i,$$

where $c_i = \sqrt{\omega_{ii}}$ and ε_i satisfies the assumptions of the standard regression model.

Models of Serial Correlation (or "autocorrelated errors"): The matrix $\mathbf{\Omega}$ is band-diagonal, i.e., $\mathbf{\Omega} \equiv [\omega_{ij}] = [c(|i-j|)]$ for some "autocorrelation function" $c(\cdot)$. A model which generates a special case is

$$\begin{aligned} y_t &= \mathbf{x}_t'\boldsymbol{\beta} + u_t, \\ u_t &= \rho u_{t-1} + \varepsilon_t, \end{aligned}$$

where ε_t satisfies the assumptions of the standard regression model with $Var(\varepsilon_t) \equiv \sigma^2(1-\rho^2)$. For this case, known as the *first-order serial correlation model*, the autocovariance function takes the form $c(s) = \rho^s$.

Panel Data Models (or "pooled cross-section / time series models"): The matrix $\mathbf{\Omega}$ again has Kronecker product form, with $\mathbf{\Omega} = \sigma_\varepsilon^2 \mathbf{I}_{NT} + \sigma_u^2(\boldsymbol{\iota}_N \boldsymbol{\iota}_N' \sigma_\alpha^2 \otimes \mathbf{I}_T) + \sigma_v^2(\mathbf{I}_T \otimes \boldsymbol{\iota}_T \boldsymbol{\iota}_T')$, where σ_u^2, σ_v^2, and σ_ε^2 are positive constants and $\boldsymbol{\iota}_N$ denotes an N-dimensional column vector of ones, etc. This case arises for a doubly-indexed dependent variable $\mathbf{y} = vec([y_{it}]')$ with $\dim(\mathbf{y}) = N \times T$, satisfying the structural equation

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + u_i + v_t + \varepsilon_{it},$$

where the error terms $\{u_i\}, \{v_t\}$, and $\{\varepsilon_{it}\}$ are all mutually uncorrelated with variances σ_u^2, σ_v^2, and σ_ε^2, respectively.

There are many other variants, which typically combine two or more of these sources of differing variances and/or nonzero correlations.

Properties of Classical Least Squares

Recalling the algebraic form of the LS estimator

$$\hat{\boldsymbol{\beta}} \equiv \hat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

this estimator remains unbiased under the assumptions of the generalized regression model; as before, assumptions 1, 3, and 4 imply

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}_{LS}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{y}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}. \end{aligned}$$

But the form of the covariance matrix for $\hat{\boldsymbol{\beta}}_{LS}$ for the generalized regression model differs from that under the classical regression model:

$$\begin{aligned} \mathbf{V}(\hat{\boldsymbol{\beta}}_{LS}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}(\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}, \end{aligned}$$

which generally does not equal $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ except for special forms of $\boldsymbol{\Omega}$ (like $\boldsymbol{\Omega} = \mathbf{I}$). Also, in general

$$\begin{aligned} E(s^2) &\equiv E\left(\frac{1}{N-K}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{LS})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{LS})\right) \\ &\neq \sigma^2. \end{aligned}$$

Similar conclusions hold for the large-sample properties of $\hat{\boldsymbol{\beta}}_{LS}$. Assuming

$$\begin{aligned} \text{plim } \frac{1}{N}\mathbf{X}'\mathbf{X} &= \mathbf{D}, \\ \text{plim } \frac{1}{N}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X} &= \mathbf{C}, \end{aligned}$$

and assuming suitable limit theorems are applicable, the classical LS estimator will have an asymptotically normal distribution,

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}\right) \xrightarrow{d} N(0, \sigma^2 \mathbf{D}^{-1}\mathbf{C}\mathbf{D}^{-1}),$$

but in general $\mathbf{D}^{-1}\mathbf{C}\mathbf{D}^{-1} \neq \mathbf{D}^{-1}$ and plim $s^2 \neq \sigma^2$.

The bias and inconsistency of the usual estimator $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}_{LS}) = s^2(\mathbf{X}'\mathbf{X})^{-1}$ of the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_{LS}$ means that the standard normal-based inference will be incorrect. There are two types of "solution" to this problem: either construct a consistent estimator of the asymptotic covariance matrix of LS, or find an alternative estimation method to LS which does not suffer from this problem.

Finally, the classical LS estimator is no longer best linear unbiased in general; the BLU estimator $\hat{\boldsymbol{\beta}}_{GLS}$, the *generalized least squares estimator*, was derived by Aitken and is named after him.

Aitken's Generalized Least Squares

To derive the form of the best linear unbiased estimator for the generalized regression model, it is first useful to define the square root $\mathbf{H}$ of the matrix $\boldsymbol{\Omega}^{-1}$ as satisfying

$$\boldsymbol{\Omega}^{-1} = \mathbf{H}'\mathbf{H},$$

which implies

$$\mathbf{H}\boldsymbol{\Omega}\mathbf{H}' = \mathbf{I}_N.$$

In fact, several such matrices $\mathbf{H}$ exist, so that, for convenience, we can assume $\mathbf{H} = \mathbf{H}'$.

Now, to derive the form of the BLU estimator of $\boldsymbol{\beta}$ for the generalized regression model under the assumption that $\boldsymbol{\Omega}$ is known, define

$$\begin{aligned} \mathbf{y}^* &\equiv \mathbf{H}\mathbf{y}, \\ \mathbf{X}^* &\equiv \mathbf{H}\mathbf{X}; \end{aligned}$$

by the usual mean-variance calculations,

$$\begin{aligned} E(\mathbf{y}^*) &= \mathbf{H}\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{X}^*\boldsymbol{\beta} \end{aligned}$$

and

$$\begin{aligned}\mathbf{V}(\mathbf{y}^*) &= \mathbf{H}\mathbf{V}(\mathbf{y})\mathbf{H}' \\ &= \mathbf{H}[\sigma^2\boldsymbol{\Omega}]\mathbf{H}' \\ &= \sigma^2\mathbf{I}_N.\end{aligned}$$

Since $\mathbf{X}^* = \mathbf{H}\mathbf{X}$ is clearly nonrandom with full column rank if $\mathbf{X}$ satisfies assumptions 3 and 4, the classical regression model applies to $\mathbf{y}^*$ and $\mathbf{X}^*$, so the Gauss-Markov theorem implies that the best linear (in terms of $\mathbf{y}^*$) unbiased estimator of $\boldsymbol{\beta}$ is

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{GLS} &\equiv (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{y}^* \\ &= (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}.\end{aligned}$$

But since this estimator is also linear in the original dependent variable $\mathbf{y}$, it follows that this "generalized least squares" (GLS) estimator is best linear unbiased using $\mathbf{y}$. Also, the usual estimator of the scalar variance parameter σ^2 will also be unbiased if $\mathbf{y}^*$ and $\mathbf{X}^*$ are used:

$$\begin{aligned}s^2_{GLS} &\equiv \frac{1}{N-K}(\mathbf{y}^*-\mathbf{X}^*\hat{\boldsymbol{\beta}}_{GLS})'(\mathbf{y}^*-\mathbf{X}^*\hat{\boldsymbol{\beta}}_{GLS}) \\ &= \frac{1}{N-K}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{GLS})'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{GLS})\end{aligned}$$

has $E[s^2_{GLS}] = \sigma^2$ by the usual arguments.

If $\mathbf{y}$ is assumed multinormal,

$$\mathbf{y} \sim N\left(\mathbf{X}\boldsymbol{\beta},\sigma^2\boldsymbol{\Omega}\right),$$

then the existing results for classical LS imply that $\hat{\boldsymbol{\beta}}_{GLS}$ is also multinormal,

$$\hat{\beta}_{GLS} \sim N\left(\boldsymbol{\beta},\sigma^2(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\right),$$

and is independent of s^2_{GLS}, with

$$\frac{(N-K)\,s^2_{GLS}}{\sigma^2} \sim \chi^2_{N-K}.$$

And the same arguments for consistency of the classical LS estimator under the classical regression model imply the corresponding large-sample results for the GLS estimator under the generalized regression model, assuming the usual limit theorems are applicable:

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}}_{GLS}-\boldsymbol{\beta}\right) \xrightarrow{d} N(\mathbf{0},\mathbf{V}),$$

with

$$\begin{aligned}\mathbf{V} &\equiv \boldsymbol{\sigma}^2 \text{ plim } \left(\frac{1}{N}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1} \\ &= \text{ plim } s^2_{GLS}\left(\frac{1}{N}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}.\end{aligned}$$

It is worth noting in passing that the definition of the squared multiple correlation coefficient R^2 generally must be adjusted for the GLS estimator. Even if one column $\boldsymbol{\iota}$ of the original matrix of regressors $\mathbf{X}$ has elements identically equal to one, that is not generally true for the transformed regressors $\mathbf{X}^* = \mathbf{HX}$. Thus, the correct restricted sum of squares in the denominator of the R^2 formula (imposing the restriction that all coefficients except the "intercept" are zero) is different from the usual $(\mathbf{y}-\bar{y}\boldsymbol{\iota})'(\mathbf{y}-\bar{y}\boldsymbol{\iota})$.

Feasible GLS

If the matrix $\boldsymbol{\Omega}$ involves unknown quantities, there are (at least) three possible strategies for inference:

1. Parametrize the matrix $\boldsymbol{\Omega}$ in terms of a finite p-dimensional vector $\boldsymbol{\theta}$ of unknown parameters

   $$\boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\theta}),$$

   which is constructed so that

   $$\boldsymbol{\Omega}(\mathbf{0}) = \mathbf{I},$$

   and conduct a *diagonostic test* of the null hypothesis $H_0 : \boldsymbol{\Omega} = \mathbf{I} \Longleftrightarrow \boldsymbol{\Theta} = \mathbf{0}$. (The form of this test will depend upon the particular parametric structure of $\boldsymbol{\Omega}(\boldsymbol{\theta})$.) If that test fails to reject, common practice is to conclude that the classical regression model is appropriate, and use the usual LS methods for inference.

2. Again parametrize the matrix $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\theta})$ in terms of a finite-dimensional parameter vector $\boldsymbol{\theta}$, and use the classical LS residuals $\mathbf{e} \equiv \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{LS}\right)$ to obtain consistent estimators $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Omega}} = \boldsymbol{\Omega}(\hat{\boldsymbol{\theta}})$ of $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$. (The details of this step depend upon the particular model, e.g., heteroskedasticity, serial correlation, etc.) Then replace the unknown $\boldsymbol{\Omega}$ with the estimated $\hat{\boldsymbol{\Omega}}$ in the formula for GLS, yielding the *feasible GLS* estimator

   $$\hat{\boldsymbol{\beta}}_{FGLS} = (\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{y},$$

with a corresponding estimator s^2_{FGLS} of σ^2. Because $\hat{\boldsymbol{\theta}}$ is typically a function of $\mathbf{y}$, this estimator will no longer be linear in $\mathbf{y}$, and the finite-sample results for GLS will no longer be applicable; nevertheless, it is often possible to find conditions under which FGLS is asymptotically equivalent to GLS,

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}}_{FGLS} - \hat{\boldsymbol{\beta}}_{GLS}\right) \xrightarrow{d} \mathbf{0},$$

so that

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}}_{FGLS} - \boldsymbol{\beta}\right) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}),$$

where

$$\mathbf{V} = \text{plim } s^2_{FGLS}\left(\frac{1}{N}\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}\right)^{-1}.$$

3. If the form of the parametric form $\boldsymbol{\Omega}(\boldsymbol{\theta})$ for the covariance matrix is incorrect – for example, if it is mistakenly assumed $\boldsymbol{\Omega}(\boldsymbol{\theta}) = \mathbf{I}$ – then application of the usual calculations will yield a particular form for the asymptotically normal distribution of $\hat{\boldsymbol{\beta}}_{FGLS}$:

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}}_{FGLS} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{D}^{-1}\mathbf{C}\mathbf{D}^{-1}\right),$$

with

$$\mathbf{D} \equiv \text{plim } \left(\frac{1}{N}\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}\right) \equiv \text{plim } \hat{\mathbf{D}}$$

and

$$\mathbf{C} \equiv \text{plim } \frac{1}{N}\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Sigma}\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X},$$

for $\boldsymbol{\Sigma} = \mathbf{V}(\mathbf{y})$ as above. Consistent estimation of $\mathbf{D}$ is immediate; the trick is to find a consistent estimator of the matrix $\mathbf{C}$ without imposing a parametric structure on the matrix $\boldsymbol{\Sigma}$. The feasibility of consistent estimation of the asymptotic covariance matrix $\mathbf{D}^{-1}\mathbf{C}\mathbf{D}^{-1}$, termed *robust covariance matrix estimation*, will vary with the particular restrictions on distributional heterogeneity and dependence imposed (e.g., i.i.d. sampling, stationary data, etc.).