# GLS (Generalized Least Squares)

Paul Johnson

October 24, 2005

## 1 Ordinary Least Squares

These are givens:
$y_i$ is a column (N-vector) of observations
$X$ is a matrix of observations with $N$ rows.
$e$ is a column (N-vector) of errors
$b$ is a column (p-vector) of parameters

$$y = Xb + e \tag{1}$$

Choose estimates $\hat{b}$ so as to minimize the sum of squares

$$S(\hat{b}) = \sum_{i=1}^{N}(y_i - X_i b)^2 = (y - X\hat{b})'(y - X\hat{b}) \tag{2}$$

The OLS solution assumes $E(e) = 0$ and that $Cov(e_i, e_j) = 0$

$$\hat{b} = (X'X)^{-1}X'y \tag{3}$$

$$Var(\hat{b}) = \sigma^2(X'X)^{-1} \tag{4}$$

## 2 Weighted Least Squares

In OLS, the "variance-covariance" matrix of the error terms is a very simple, clean thing:

$$Var(e) = \begin{bmatrix} \sigma_e^2 & 0 & & & 0 & 0 \\ 0 & \sigma_e^2 & & & 0 & 0 \\ & & \ddots & & & 0 \\ & & & \ddots & & \\ 0 & & & & \sigma_e^2 & 0 \\ 0 & 0 & & & 0 & \sigma_e^2 \end{bmatrix} = \sigma_e^2 \begin{bmatrix} 1 & 0 & & & 0 & 0 \\ 0 & 1 & & & 0 & 0 \\ & & \ddots & & & 0 \\ & & & \ddots & & \\ 0 & & & & 1 & 0 \\ 0 & 0 & & & 0 & 1 \end{bmatrix} \tag{5}$$

If you have "heteroskedasticity", then the off-diagonal elements are the same–all 0. But on the diagonal, the elements differ:

$$Var(e) = \begin{bmatrix} \sigma^2_{e_1} & 0 & & & 0 & 0 \\ 0 & \sigma^2_{e_2} & & & 0 & 0 \\ & & \ddots & & & 0 \\ & & & \ddots & & \\ 0 & & & & \sigma^2_{e_{N-1}} & 0 \\ 0 & 0 & 0 & & 0 & \sigma^2_{e_N} \end{bmatrix} \tag{6}$$

In weighted least squares, we use estimates of $\sigma^2_{e_i}$ as weights.

$$S(\hat{b}) = \sum_{i=0}^{N} \frac{1}{\widehat{\sigma^2_{e_i}}}(y_i - \hat{y}_i)^2 = \sum_{i=0}^{N} w_i(y_i - \hat{y}_i)^2 \tag{7}$$

Consider the WLS problem. If the heteroskedastic case occurs

$$S(\hat{b}) = \sum_{i=1}^{N} w_i(y_i - \hat{y}_i)^2 = \sum_{i=1}^{N}(\sqrt{w_i}y_i - \sqrt{w_i}\hat{y}_i)^2 \tag{8}$$

$$= w_1(y_1 - \hat{y}_1)^2 + w_2(y_2 - \hat{y}_2)^2 + \cdots + w_N(y_N - \hat{y}_N)^2 \tag{9}$$

# 3 Generalized Least Squares (GLS)

Suppose, generally, the Variance/Covariance matrix of residuals is

$$V = \begin{bmatrix} \sigma^2_1 & Cov(e_1, e_2) & Cov(e_1, e_2) & \cdots & Cov(e_1, e_N) \\ Cov(e_1, e_2) & \sigma^2_2 & & & \\ & & \sigma^2_3 & & \\ \vdots & & & \ddots & \\ & & & \sigma^2_{N-1} & Cov(e_1, e_{N-1}) \\ Cov(e_1, e_N) & & & Cov(e_1, e_{N-1} & \sigma^2_N \end{bmatrix} \tag{10}$$

You can factor out a constant $\sigma^2$ if you care to (MM&V do, p. 51).

If all the off diagonal elements of $V$ are set to $0$, then this degenerates to a problem of heteroskedasticity, know as Weighted Least Squares (WLS).

If the off diagonal elements of $V$ are not 0, then there can be correlation across cases. In a time series problem, that amounts to "autocorrelation." In a cross-sectional problem, it means that various observations are interrelated.

The "sum of squared errors" approach uses this variance matrix to adjust the data so that the residuals are homoskedastic or that the cross-unit correlations are taken into account.

Let

$$W = V^{-1} \tag{11}$$

The idea behind WLS/GLS is that there is a way to transform $y$ and $X$ so that the residuals are homogeneous, some kind of weight is applied.

In matrix form, the representation of WLS/GLS is

$$S(\hat{b}) = (y - \hat{y})' W (y - \hat{y}) \tag{12}$$

and, if you write that out, the sum of squares in a GLS framework is a rather more complicated thing. All of those off-diagonal $w_{ij}$'s make the number of terms multiply.

$$\begin{bmatrix} y_1 - \hat{y}_1 & y_2 - \hat{y}_2 & \cdots & y_N - \hat{y}_N \end{bmatrix} \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1N} \\ w_{12} & w_{22} & & w_{2N} \\ \vdots & & \ddots & \vdots \\ w_{N1} & w_{N2} & \cdots & w_{NN} \end{bmatrix} \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_N - \hat{y}_N \end{bmatrix}$$

$$= \begin{bmatrix} y_1 - \hat{y}_1 & y_2 - \hat{y}_2 & \cdots & y_N - \hat{y}_N \end{bmatrix} \begin{bmatrix} w_{11}(y_1 - \hat{y}_1) & +w_{12}(y_2 - \hat{y}_2) & \cdots & +w_{1N}(y_N - \hat{y}_N) \\ w_{12}(y_1 - \hat{y}_1) & +w_{22}(y_2 - \hat{y}_2) & & +w_{2N}(y_N - \hat{y}_N) \\ \vdots & & \ddots & \vdots \\ w_{N1}(y_1 - \hat{y}_1) & +w_{N2}(y_2 - \hat{y}_2) & \cdots & +w_{NN}(y_N - \hat{y}_N) \end{bmatrix}$$

$$= \begin{bmatrix} w_{11}(y_1 - \hat{y}_1)^2 & +w_{12}(y_2 - \hat{y}_2)(y_1 - \hat{y}_1) & \cdots & +w_{1N}(y_N - \hat{y}_N)(y_1 - \hat{y}_1) \\ w_{12}(y_1 - \hat{y}_1)(y_2 - \hat{y}_2) & +w_{22}(y_2 - \hat{y}_2)^2 & & +w_{2N}(y_N - \hat{y}_N)(y_2 - \hat{y}_2) \\ \vdots & & \ddots & \vdots \\ w_{N1}(y_1 - \hat{y}_1)(y_N - \hat{y}_N) & +w_{N2}(y_2 - \hat{y}_2)(y_N - \hat{y}_N) & \cdots & +w_{NN}(y_N - \hat{y}_N)^2 \end{bmatrix}$$

$$\begin{aligned} &w_{11}(y_1 - \hat{y}_1)^2 + w_{12}(y_2 - \hat{y}_2)(y_1 - \hat{y}_1) + \cdots + w_{1N}(y_N - \hat{y}_N)(y_1 - \hat{y}_1) \\ = \quad &+w_{12}(y_1 - \hat{y}_1)(y_2 - \hat{y}_2) + w_{22}(y_2 - \hat{y}_2)^2 + \cdots + w_{2N}(y_N - \hat{y}_N)(y_2 - \hat{y}_2) \\ &+w_{N1}(y_1 - \hat{y}_1)(y_N - \hat{y}_N) + w_{N2}(y_2 - \hat{y}_2)(y_N - \hat{y}_N) + \cdots + w_{NN}(y_N - \hat{y}_N)^2 \end{aligned}$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(y_i - \hat{y}_i)(y_j - \hat{y}_j) \tag{13}$$

Supposing a linear model:

$$\hat{y} = X\hat{b} \tag{14}$$

The normal equations (the name for the first order conditions) are found by setting the derivatives of the Sum of Squares with respect to the parameters equal to 0.

$$for\, each\, parameter\, \hat{b}_j : \frac{\partial}{\partial \hat{b}_j} \left[ (y - \hat{y})' W (y - \hat{y}) \right] = 0$$

If you insert $\hat{y} = Xb$ in there and do the math (or look it up in a book :) )

$$(X'WX)\hat{b} - X'Wy = 0$$

The WLS/GLS estimator is the solution:

$$\hat{b} = (X'WX)^{-1} X'Wy \tag{15}$$

and, supposing you divide out a constant communal variance term $\sigma^2$ and the 'leftovers' remain in $W$, the Var/Cov matrix is

$$Var(\hat{b}) = \sigma^2(X'WX)^{-1} \tag{16}$$

The estimate of $\hat{b}$ is consistent and has lower variance than other linear estimators.

Please note that the formula 16 ends up with such a small, simple formula because of the simplifying results that are invoked along the way. Observe (don't forget that $Var(e) = \sigma^2 V = \sigma^2 W^{-1}$, $W = W'$, and $Var(aW) = aVar(W)a'$.) All the rest is easy.

$$Var(\hat{b}) = Var[(X'WX)^{-1}X'Wy] \tag{17}$$

$$= Var[(X'WX)^{-1}X'W(Xb + e)] = Var[(X'WX)^{-1}X'We]$$

$$= (X'WX)^{-1}X'W[Var(e)]W'X(X'WX)^{-1}$$

$$= (X'WX)^{-1}X'W(\sigma^2 W^{-1})W'X(X'WX)^{-1}$$

$$= \sigma^2(X'WX)^{-1}X'WX(X'WX)^{-1}$$

$$= \sigma^2(X'WX)^{-1}$$

# 4 Robust estimates of $Var(\hat{b})$

The Huber-White "sandwich" estimator of $Var(\hat{b})$ is designed to deal with the problem that the model $W$ may not be entirely correct.

Ordinarily, what people do is to calculate an OLS model, and then use the "sandwich" estimator to make the estimates of the standard errors more robust. In an OLS context, we use the observed residuals:

$$\hat{e} = y - X\hat{b} \tag{18}$$

This is an "empirical" estimator, in the sense that "rough" estimates of the correlations among observations are used in place of the hypothesized values. Huber and White developed this estimator, which is more likely to be accurate when the assumption about $W$ is wrong. 16

When I get tired of gazing at statistics books and really want to know how things get calculated, I download the source code for R and some of its libraries and go read their code. Assuming they get it right, which they usually do, its quite easy to figure how things are actually done.

In the code for the "car" package by John Fox, for example, there's an R function hccm that calculates 5 flavors of heteroskedasticity consistent covariance matrices. The oldest, the original, the Huber-White formula (version hc 0) works on the hypothesis that one can estimate $Var(e_i)$ with $(\hat{e}_i^2)$. So, in the OLS model, the $Var(\hat{b})$starts out like so:

$$Var(\hat{b}) = (X'X)^{-1}X'Var[e]X(X'X)^{-1}$$

$$= (X'X)^{-1}X'(\sigma_e^2 I)X(X'X)^{-1}$$

and we replace the middle part with the empirical estimate of the variance matrix.

In the heteroskedasticity case, that matrix looks like this

$$\begin{bmatrix} \hat{e}_1^2 & 0 & 0 & 0 & 0 \\ 0 & \hat{e}_2^2 & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & \hat{e}_{N-1}^2 & 0 \\ 0 & 0 & 0 & 0 & \hat{e}_N^2 \end{bmatrix} = diag[e \cdot e']$$

The heteroskedasticity consistent estimator is thus:

$$Var_{hc0}(b) = (X'X)^{-1}X'\{diag[e \cdot e']\}X(X'X)^{-1} \tag{19}$$

It's called a "sandwich estimator" because somebody thought it was cute to think of the matching outer elements as pieces of bread. Its also called the "information sandwich estimator" because those things on the outside look an awful lot like information matrices.

TODO: In Dobson there are some special formulae for correction of standard errors from panel studies. Write in all those details.

Seems like every time I turn around somebody suggests a new improvement on robust standard errors, especially for panel studies. I guess when these get published, all the textbooks and stat packs will have to be redone.

Edward W. Frees and Chunfang Jin. 2003. "Empirical standard errors for longitudinal data mixed linear models" University of Wisconsin. for information contact jfrees@bus.wisc.edu