

1 Example: Weighting to Improve Precision

We have seen so far that if the data are heteroskedastic, then ordinary least squares will be unbiased and consistent, but its variance will have a more complicated form. So, we can still use OLS, but we have to modify the standard errors, test statistics, and confidence intervals to take this into account.

Since the Gauss-Markov theorem assumed homoskedasticity, it is not necessarily the case that OLS is minimum variance among linear, unbiased estimators. There may be other estimators that have lower variance. In this note, we explore how to modify OLS to improve its efficiency.

First, let's start with a very simple case with heteroskedasticity. Suppose that we have a sample of size 2 for y_i . Assume that

$$\begin{aligned}E[y_1] &= \mu \\E[y_2] &= \mu \\V[y_1] &= \sigma_1^2 \\V[y_2] &= \sigma_2^2 \\Cov(y_1, y_2) &= 0\end{aligned}$$

Letting $y = (y_1, y_2)'$ we can write this compactly as

$$E[y] = \begin{pmatrix} \mu \\ \mu \end{pmatrix}, \quad V[y] = \Omega = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}.$$

Assume that Ω is known. Usually, of course, we would not know Ω , but we start with this idealized case to get some intuition.

Let $\hat{\mu}$ be the sample average $\hat{\mu} = \bar{y}_n = (y_1 + y_2)/2$. As an estimator for μ , it is unbiased:

$$E[\hat{\mu}] = E[(y_1 + y_2)/2] = \mu.$$

Its variance is

$$V[\hat{\mu}] = V[(y_1 + y_2)/2] = \frac{1}{4}(\sigma_1^2 + \sigma_2^2).$$

Now consider an alternative estimator for μ , a weighted average:

$$\tilde{\mu}_w = w_1 \cdot y_1 + w_2 \cdot y_2.$$

Note that the sample average is a special case, with $w_1 = w_2 = 1/2$, and also note that we can think of this estimator as a linear function of y :

$$\tilde{\mu}_w = (w_1, w_2) \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}.$$

The mean and variance are:

$$E[\tilde{\mu}_w] = \mu \cdot (w_1 + w_2), \quad V[\tilde{\mu}_w] = w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2.$$

So, if we want the estimator to be unbiased, we clearly need $w_1 + w_2 = 1$, and if we want to then pick the weights that give the lowest variance, we would need to solve:

$$\min_{w_1, w_2} w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 \quad \text{s.t. } w_1 + w_2 = 1.$$

This is easy to solve with basic calculus. The solution turns out to be:

$$w_1 = \frac{\frac{1}{\sigma_1^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}, \quad w_2 = \frac{\frac{1}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}.$$

This is actually pretty intuitive: suppose that $\sigma_1 < \sigma_2$, so that observation 2 is “noisier” than observation 1. Then the optimal weighted average puts *less weight* on observation 2 and more on observation 1.

Note that the optimal weights give the *minimum variance linear unbiased estimator* of μ . All linear estimators (that is, estimators that are linear functions of the vector y) must have the weighted average form, and to be unbiased the weights also must add to 1.

There is a different way to motivate this estimator. Suppose we assume in addition that y is multivariate normal:

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N \left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \Omega \right).$$

Then the likelihood function (again, assuming Ω is known) is

$$f(y|\mu) = C \cdot \exp \left(-\frac{1}{2} \left(y - \begin{bmatrix} \mu \\ \mu \end{bmatrix} \right)' \Omega^{-1} \left(y - \begin{bmatrix} \mu \\ \mu \end{bmatrix} \right) \right),$$

and the log likelihood is

$$\log f(y|\mu) = C' - \frac{1}{2} \left(y - \begin{bmatrix} \mu \\ \mu \end{bmatrix} \right)' \Omega^{-1} \left(y - \begin{bmatrix} \mu \\ \mu \end{bmatrix} \right).$$

So the MLE for μ solves:

$$\max_{\mu} -\frac{1}{2} \left(y - \begin{bmatrix} \mu \\ \mu \end{bmatrix} \right)' \Omega^{-1} \left(y - \begin{bmatrix} \mu \\ \mu \end{bmatrix} \right).$$

Equivalently, we can solve:

$$\min_{\mu} \begin{pmatrix} y_1 - \mu \\ y_2 - \mu \end{pmatrix}' \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{pmatrix} \begin{pmatrix} y_1 - \mu \\ y_2 - \mu \end{pmatrix} = \sum_{i=1}^2 \frac{1}{\sigma_i^2} (y_i - \mu)^2. \quad (1)$$

Again, using calculus, we can solve this to get:

$$\hat{\mu}_{ML} = \frac{\frac{1}{\sigma_1^2} y_1 + \frac{1}{\sigma_2^2} y_2}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}.$$

So the MLE is identical to the optimal weighted average estimator we derived above.

2 Generalized Least Squares

We can generalize the weighting idea to models with regressors, and with more general covariance matrices.

Suppose that y is $n \times 1$ and X is $n \times k$, and assume that

$$E[y|X] = X\beta,$$

$$V[y|X] = \Omega,$$

where Ω is a known, positive definite symmetric matrix. In addition we will assume that X has full column rank k .

Consider the estimator $\hat{\beta}_{gls}$ which solves the following minimization problem:

$$\min_{\beta} (y - X\beta)' \Omega^{-1} (y - X\beta).$$

By the same logic as before, this estimator will be the conditional MLE if we were to assume that $y|X \sim N(X\beta, \Omega)$.

The solution to this generalized least-squares minimization problem is:

$$\hat{\beta}_{gls} = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} y.$$

We call this the **generalized least squares**, or GLS, estimator.

3 Weighted Least Squares

It's useful to consider the special case where Ω is diagonal:

$$\Omega = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & & \\ \vdots & & \ddots & \\ 0 & \cdots & & \sigma_n^2 \end{pmatrix}.$$

Then

$$\Omega^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & & \\ \vdots & & \ddots & \\ 0 & \cdots & & \frac{1}{\sigma_n^2} \end{pmatrix}.$$

Multiplying out the expression for $\hat{\beta}_{gls}$, we can express the estimator as:

$$\hat{\beta}_{gls} = \left(\sum_{i=1}^n \frac{1}{\sigma_i^2} \cdot x_i x_i' \right)^{-1} \sum_{i=1}^n \frac{1}{\sigma_i^2} \cdot x_i y_i,$$

which is a weighted version of the usual least squares estimator.

If we form

$$\tilde{y}_i = \frac{1}{\sigma_i} y_i, \quad \tilde{x}_i = \frac{1}{\sigma_i} x_i,$$

(notice we are using the standard deviations!) then we can write

$$\hat{\beta}_{gls} = \left(\sum_{i=1}^n \tilde{x}_i \tilde{x}_i' \right)^{-1} \sum_{i=1}^n \tilde{x}_i \tilde{y}_i.$$

So one way to implement this estimator is to construct the modified variables, then run ordinary least squares.

4 Properties of GLS

Return to the general version of $\hat{\beta}_{gls}$. What properties does this estimator have? It is unbiased:

$$\begin{aligned} E[\hat{\beta}_{gls}|X] &= E[(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y|X] \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}E[y|X] \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}X\beta \\ &= \beta \end{aligned}$$

Its conditional variance is:

$$\begin{aligned} V[\hat{\beta}_{gls}|X] &= V[(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y|X] \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}V[y|X]\Omega^{-1}X(X'\Omega^{-1}X)^{-1} \\ &= (X'\Omega^{-1}X)^{-1}. \end{aligned}$$

Finally, let us consider the issue of efficiency. Our intuition is that the GLS estimator should be a “better” estimator than OLS in general. It turns out that it is in fact efficient:

Theorem 1 *The GLS estimator $\hat{\beta}_{gls}$ is efficient relative to all other linear unbiased estimators for β .*

The proof of this result is worth understanding. We first use the Cholesky factorization of the variance matrix Ω :

$$\Omega = CC',$$

where C is a lower triangular matrix. This is a version of a matrix “square root” of Ω – see Ruud, 7.6.1. Consider the linear transformation $C^{-1}y$:

$$\begin{aligned} E[C^{-1}y|X] &= C^{-1}X\beta, \\ V[C^{-1}y|X] &= C^{-1}\Omega C^{-1'} = I_n. \end{aligned}$$

So, if we set $\tilde{y} = C^{-1}y$, and $\tilde{X} = C^{-1}X$, then

$$E[\tilde{y}|\tilde{X}] = \tilde{X}\beta,$$

$$V[\tilde{y}|\tilde{X}] = I_n.$$

Since we have a homoskedastic model, the conditions for the Gauss-Markov theorem hold, and the estimator

$$(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y}$$

is minimum variance among linear estimators. But this is equal to

$$[(C^{-1}X)'(C^{-1}X)]^{-1}(C^{-1}X)'(C^{-1}y) = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y = \hat{\beta}_{gls}.$$

5 Feasible Generalized Least Squares

In practice, we usually do not know Ω , so the GLS estimator cannot be used. But in some cases, we might be able to estimate Ω , if we can make some assumptions about its structure.

Suppose for example that we assume that (y_i, x_i) are IID, with

$$E[y_i|x_i] = x_i'\beta,$$

$$V[y_i|x_i] = z_i'\gamma,$$

where z_i could be equal to x_i or could be a subvector of x_i . So

$$\Omega = \begin{pmatrix} z_1'\gamma & 0 & \cdots & 0 \\ 0 & z_2'\gamma & & 0 \\ \vdots & & \ddots & \\ 0 & \cdots & & z_n'\gamma \end{pmatrix}.$$

We could estimate β and γ jointly by conditional MLE (if we assume conditional normality of y_i). However, there isn't a simple formula for the MLE in this case, so it will require a numeric solution method.

A simple alternative is a two-step estimator. Consider the following procedure:

1. First, estimate β by OLS, to get a preliminary estimate $\hat{\beta}$. Form $e_i = y_i - x_i'\hat{\beta}$, and regress e_i^2 on z_i to get $\hat{\gamma}$.
2. Form $\hat{\Omega}$ by using the estimate $\hat{\gamma}$. Run GLS of y on X using $\hat{\Omega}$ in place of Ω .

The intuition is that if we observed $\epsilon_i = y_i - x_i'\beta$, then since $E[\epsilon_i^2|X] = V[\epsilon_i|X] = V[y_i|X] = z_i'\gamma$, we could regress ϵ_i^2 on z_i to get a consistent estimator of γ . We do not know β , so we cannot form ϵ_i . Instead, we first estimate β using an inefficient, but consistent estimator. Then we use e_i in place of ϵ_i to get an estimate of γ . Then we plug in this estimate to get an estimated version of Ω to use for GLS.

This two-step estimator is called the *feasible generalized least squares* estimator, or FGLS. Provided that $\hat{\Omega} \xrightarrow{P} \Omega$, it turns out that this estimator has the same asymptotic distribution as the infeasible GLS estimator that uses the true variance matrix Ω . However, in finite samples, it may have different behavior from the GLS estimator.