

1 Efficient OLS

1. Consider the model

$$\begin{aligned} Y &= X\beta + \varepsilon \\ E[X'\varepsilon] &= 0_K \\ E[\varepsilon\varepsilon'] &= \Omega = \sigma^2 I_N. \end{aligned}$$

This is OLS happyland! OLS is BLUE here.

2. So, you get an estimated parameter vector

$$\hat{\beta}_{OLS} = (X'X)^{-1} X'Y.$$

3. You know that it is the lowest variance estimator, but what is its variance? Its bias is

$$\begin{aligned} E\left[\left(\hat{\beta}_{OLS} - \beta\right)\right] &= E\left[(X'X)^{-1} X'X\beta + (X'X)^{-1} X'\varepsilon - \beta\right] \\ &= E\left[(X'X)^{-1} X'\varepsilon\right] = (X'X)^{-1} 0_K = 0_K \end{aligned}$$

The variance of the estimated parameter vector is the expectation of the square of the quantity in square brackets:

$$\begin{aligned} V\left[\hat{\beta}_{OLS}\right] &= E\left[\left(\hat{\beta}_{OLS} - \beta\right)\left(\hat{\beta}_{OLS} - \beta\right)'\right] \\ &= E\left[(X'X)^{-1} X'\varepsilon\varepsilon'X(X'X)^{-1}\right] \\ &= (X'X)^{-1} X'E[\varepsilon\varepsilon']X(X'X)^{-1} \\ &= (X'X)^{-1} X'\sigma^2 I_N X(X'X)^{-1} \\ &= \sigma^2 (X'X)^{-1} X'X(X'X)^{-1} \\ &= \sigma^2 (X'X)^{-1} \end{aligned}$$

4. Precision is good. Low variance is precision. How do you get a precise estimate? One can think about $V\left[\hat{\beta}_{OLS}\right] = \sigma^2 (X'X)^{-1}$ in 3 pieces:

- (a) The variance of ε is the variance of Y conditional on X . Less variation of Y around the regression line yields greater precision.
- (b) N is the number of observations. It shows up, implicitly, inside $X'X$. This is easiest to see if X has just one column: in this case, $X'X = \sum_{i=1}^N (x_i)^2$, which for x_i drawn from some density $f(x)$ has an expectation that increases linearly with N . So, $V\left[\hat{\beta}_{OLS}\right]$ goes inversely proportionally with N .

(c) $X'X$ is related to the covariance matrix the vectors $x'_i, i = 1, \dots, N$. If each column of X is mean-zero, then $X'X$ is the covariance matrix of the K columns of X . For this reason, if X has a lot of variance, then $X'X$ is bigger, so $(X'X)^{-1}$ is smaller, so $V[\hat{\beta}_{OLS}]$ is smaller and the estimate $\hat{\beta}_{OLS}$ is more precise.

5. The only problem here is that σ^2 is not observed. However, we have a sample analog: the sample residual e :

$$e = Y - X\hat{\beta}_{OLS}.$$

6. So how exactly does e relate to ε ?

$$\begin{aligned} e &= Y - X(X'X)^{-1}X'Y \\ &= [I - X(X'X)^{-1}X']Y \\ &= [I - X(X'X)^{-1}X']X\beta + [I - X(X'X)^{-1}X']\varepsilon \\ &= X\beta - X\beta + [I - X(X'X)^{-1}X']\varepsilon \\ &= [I - X(X'X)^{-1}X']\varepsilon \end{aligned}$$

e is a linear transformation of ε . However, although $[I - X(X'X)^{-1}X']$ is an $N \times N$ matrix, it is not a full rank matrix: its columns are related. Indeed, this $N \times N$ weighting matrix is all driven by the identity matrix, which has rank N , and the matrix X , which only has K columns. The full matrix $[I - X(X'X)^{-1}X']$ has rank $N - K$.

7. Matrices like $[I - X(X'X)^{-1}X']$ and $X(X'X)^{-1}X'$ are called *projection* matrices, and they come up a lot.

(a) for any matrix Z , denote its projection matrix $P_Z = Z(Z'Z)^{-1}Z'$ and its error projection as $M_Z = I - Z(Z'Z)^{-1}Z'$

(b) These are convenient. We can write the OLS estimate of $X\beta$ as

$$X\hat{\beta}_{OLS} = P_X Y,$$

and the OLS residuals $Y - X\hat{\beta}_{OLS}$ as

$$e = M_X Y$$

and also,

$$e = M_X \varepsilon$$

(c) We say stuff like "The matrix P_X projects X onto Y ."

(d) These matrices have a few useful properties:

i. they are symmetric.

ii. they are *idempotent*, which means they equal their own square:

$$P_Z P_Z = P_Z, M_Z M_Z = M_Z$$

8. Compute $e'e$ in terms of ε :

$$e = M_X \varepsilon$$

so,

$$\begin{aligned} E[e'e] &= E[\varepsilon' M_X M_X \varepsilon] \\ &= E[\varepsilon' M_X \varepsilon] \\ &= E[\varepsilon' \varepsilon] - E[\varepsilon' X (X'X)^{-1} X' \varepsilon] \\ &= N\sigma^2 - K\sigma^2 \end{aligned}$$

because P_X has rank K . Consequently,

$$\frac{E[e'e]}{N - K} = \sigma^2.$$

So, we can use an estimate

$$\hat{\sigma}^2 = \frac{e'e}{N - K}$$

9. The estimated variance of the OLS estimator is thus given by

$$\hat{V}[\hat{\beta}_{OLS}] = \hat{\sigma}^2 (X'X)^{-1}.$$

Now, we can compute the BLUE estimate, and say something about its bias (zero) and its sampling variability if we have spherical disturbances.

2 NonSpherical Disturbances

1. In a model

$$Y = g(X, \beta) + \varepsilon,$$

if disturbances satisfy

$$E[\varepsilon\varepsilon'] = \sigma^2 I_N,$$

we call them *spherical*. Independently normal disturbances are spherical, but the assumption of independent normality is much stronger than the assumption that disturbances are spherical, because normality restricts all products of all powers of all disturbances. In contrast, the restriction that disturbances are spherical restricts only the squares of disturbances and cross-products of disturbances:

(a) The first implication of spherical disturbances is

$$E [(\varepsilon_i)^2] = \sigma^2,$$

for all $i = 1, \dots, N$, which usually call *homoskedasticity*. Homoskedastic disturbances have the same variance for all observations.

(b) The second implication is that

$$E [\varepsilon_i \varepsilon_j] = 0,$$

for all $i \neq j$. This means that there are no correlations in disturbances across observations. This rules out over-time correlations in time-series data, and spatial correlations in cross-sectional data.

2. OLS is inefficient if disturbances are nonspherical. This is easy to see by example. Imagine that we have a linear model with a constant α and one regressor (the vector X):

$$\begin{aligned} Y &= \alpha + X\beta + \varepsilon, \\ E[\varepsilon] &= 0_N \\ E[\varepsilon\varepsilon'] &= \Omega \neq \sigma^2 I_N \end{aligned}$$

where

$$\Omega = \begin{bmatrix} 0 & 0 & 0 \\ 0 & I_{N-2} & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

That is, we have an environment where we know that the first and last observation have a disturbance term of zero, and all the rest are of the usual kind.

(a) Consider a regression line that connects the first and last data points, and ignores all the rest. This regression line is exactly right. Including other data in the estimate only adds wrongness. Thus, the best linear unbiased estimator in this case is the line connecting the first and last dots. Consequently, OLS is inefficient—it does not have the lowest variance.

(b) The point is that you want to pay close attention where the disturbances have low variance and not pay much attention where the disturbances have high variance.

(c) Alternatively, imagine that

$$\Omega = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 \iota_{N-1} \iota'_{N-1} & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}.$$

Here, the notation ι_K indicates a K -vector of ones. Thus, $\iota_{N-2}\iota'_{N-2}$ is an $(N-2) \times (N-2)$ matrix of ones, and $\sigma^2 \iota_{N-1}\iota'_{N-1}$ is a matrix filled with σ^2 . This covariance matrix would arise if observations 1 and N had independent disturbances with variance σ^2 , and observations 2, ..., $N-1$ had the same disturbance term. Not just disturbance terms drawn from the same distribution, but literally the same value of ε for each of those observations.

- (d) In this case, you'd want to treat observations 2, ..., $N-1$ as if they were just one observation: for example, they all had a big positive disturbance, you wouldn't want to pull the regression line up very much, because you'd know that what seemed like a lot of positive disturbances was really just one big outlier. Consequently, since OLS wouldn't do any grouping like this, OLS is not efficient.

3 Generalised Least Squares

1. *Generalised Least Squares* (GLS) is used when we face a model like

$$\begin{aligned} Y &= \alpha + X\beta + \varepsilon, \\ E[\varepsilon] &= 0_N \\ E[\varepsilon\varepsilon'] &= \Omega \end{aligned}$$

Here, if $\Omega \neq \sigma^2 I_N$, you have some form of nonspherical disturbances: either heteroskedasticity, or correlations across observations.

2. We know that OLS is the efficient estimator given homoskedastic disturbances, but what about the above case?
3. The trick is to convert this problem back to a homoskedastic problem. Consider premultiplying Y and X by $\Omega^{-1/2}$

$$\Omega^{-1/2}Y = \Omega^{-1/2}X\beta + \Omega^{-1/2}\varepsilon$$

Here is a model with the disturbance term premultiplied by this weird inverse-matrix-square-root thing.

4. What is the mean and variance of this new transformed disturbance term?

$$\begin{aligned} E\left[\Omega^{-1/2}\varepsilon\right] &= \Omega^{-1/2}E[\varepsilon] = 0 \\ E\left[\Omega^{-1/2}\varepsilon\varepsilon'\Omega^{-1/2}\right] &= \Omega^{-1/2}E[\varepsilon\varepsilon']\Omega^{-1/2} \\ &= \Omega^{-1/2}\Omega\Omega^{-1/2} \\ &= \Omega^{-1/2}\Omega^{1/2}\Omega^{1/2}\Omega^{-1/2} = I_N I_N = I_N \end{aligned}$$

(see Kennedy's appendix "All About Variance" for more rules on variance computations).

5. So the premultiplied model is homoskedastic with unit variance disturbances.
6. Given that the coefficients in the transformed model are the same as those in the untransformed model, we can estimate them by using OLS on the transformed model.
7. Transforming data by a known variance matrix and then applying OLS is called *Generalised Least Squares*.
8. We refer to the matrix

$$T = \Omega^{-1/2}$$

as the *Transformation Matrix*.

9. One such known variance matrix is that associated with dependent variable data whose elements are group means: eg, average income in a country. In this case, the averages have known relative variances: the variance of the mean of something goes with the square root of the sample size used to compute it. If every country has the same variance in each observation it uses to calculate its average income, the averages will have variances inversely proportional to the sample sizes used to compute them. So, in the model where i indexes countries, and each country computes its mean off of a sample with size S_i , and the disturbances are not correlated across countries, the covariance matrix must be

$$\Omega = \sigma^2 \begin{bmatrix} \frac{1}{S_1} & 0 & 0 \\ 0 & \frac{1}{S_i} & 0 \\ 0 & 0 & \frac{1}{S_N} \end{bmatrix}$$

and, therefore,

$$T = \frac{1}{\sigma} \begin{bmatrix} \sqrt{S_1} & 0 & 0 \\ 0 & \sqrt{S_i} & 0 \\ 0 & 0 & \sqrt{S_N} \end{bmatrix}$$

10. The transformation matrix which amounts to multiplying each Y and each X by the square root of the sample size used in each country.
11. This strategy, in which you premultiply each observation separately, rather than premultiplying a whole vector of Y and a whole matrix of X , is appropriate when the covariance matrix is diagonal as it is in the grouped mean data case. This strategy is referred to as *Weighted Least Squares* (WLS).
12. GLS is all great if you know the covariance matrix of the disturbances, but usually, you don't. A similar strategy, called *Feasible Generalised Least Squares* (FGLS) covers the case where you don't know this covariance matrix, but you can estimate it.

13. FGLS uses two steps:

- (a) Get a consistent estimate $\hat{\Omega}$ of Ω .
 - i. A *consistent* estimate is one which is asymptotically unbiased and whose variance declines as the sample size increases.
 - ii. Not all things can be estimated consistently. Examples will come somewhat later.
- (b) Compute $\hat{T} = \hat{\Omega}^{-1/2}$, and run GLS.

14. The Random Effects Model uses FGLS

- (a) Assume that

$$\begin{aligned} Y_{it} &= X_{it}\beta + \theta_i + \varepsilon_{it} \\ E[\theta_i | X_{it}] &= E[\varepsilon_{it} | X_{it}] = E[\theta_i \varepsilon_{js} | X_{it}] = 0, \\ E[(\theta_i)^2 | X_{it}] &= \sigma_\theta^2 \quad E[(\varepsilon_{it})^2 | X_{it}] = \sigma_\varepsilon^2 \end{aligned}$$

(Actually, this is a bit stronger than what is needed: you just need θ_i orthogonal to X_{it} , but the differing subscripts makes that assumption notationally cumbersome.) The fact that θ_i are mean zero no matter what value X takes is strong. For example, if X includes education and θ_i is meant to capture smartness, we would expect correlation between them. We also need the variance of θ_i to be independent of X . For example, if half of all people are lazy and lazy people never go to college, then the variance of θ_i would covary positively with X observed post-secondary schooling.

- (b) Given the assumption on θ_i , we get

$$Y_{it} = X_{it}\beta + u_{it}$$

where

$$u_{it} = \theta_i + \varepsilon_{it}$$

is a composite error term which satisfies exogeneity, but does not satisfy the spherical error term requirement for efficiency of OLS.

- (c) One could use OLS of Y on X and get unbiased consistent estimates of β . The reason is that the nonspherical error term only hurts the efficiency of the OLS estimator; it is still unbiased.
- (d) However, this approach leaves out important information that could improve the precision of our estimate. In particular, we have assumed that the composite errors have a chunk which is the same for every t for a given i . There is a GLS approach to take advantage of this

assumption. If we knew the variance of the θ_i terms, σ_θ^2 , and knew the variance of the true disturbances, σ_ε^2 , we could take advantage of this fact.

- (e) Under the model, we can compute the covariance of errors of any two observations:

$$\Omega = E[u_{it}u_{js}] = E[(\theta_i + \varepsilon_{it})(\theta_j + \varepsilon_{js})] = I[i = j]\sigma_\theta^2 + I[s = t]\sigma_\varepsilon^2$$

where $I[\cdot]$ is the indicator function. This covariance matrix is block diagonal, where each block consists of the sum of the two variances σ_θ^2 and σ_ε^2 on the diagonal, and just σ_θ^2 off the diagonal. These blocks lie on the diagonal of the big matrix, and the off-diagonal blocks are all zero. (see Green around p 295 for further exposition). So, Ω has diagonal elements equal to $\sigma_\theta^2 + \sigma_\varepsilon^2$ and within-person off-diagonal elements equal to σ_θ^2 and across-person off-diagonal elements equal to 0.

- (f) The GLS transformation matrix is computed as $T = \Omega^{-1/2}$, which is the matrix square-root of this composite error covariance matrix. Then, FGLS regresses transformed Y on transformed X :

$$\begin{aligned} TY &= TX\beta + Tu \\ E[XT'Tu] &= 0 \\ E[Tuu'T] &= 1_N \end{aligned}$$

Note that $E[XT'Tu] = 0$ because we imposed the strong conditional mean-independence condition above: $E[\theta_i | X_{it}] = E[\varepsilon_{it} | X_{it}] = 0$ implies both $E[X'u] = 0$ and $E[XT'Tu] = 0$ (check for yourself!). In contrast, the weaker orthogonality condition $E[X'u] = 0$ does not imply $E[XT'Tu] = 0$.

- (g) This GLS approach is only easy to implement with a balanced panel in which each and every observation is observed for the same number of periods (so that T is well-defined). But, even with an unbalanced panel, you can still create the block diagonal matrix and invert it.
- (h) FGLS requires a consistent estimate of the two variances. A fixed effects model can be run in advance to get estimates of these variances. Or, one could run OLS and construct an estimate of the error covariance matrix directly. Either yields a consistent estimate.
15. The trick with FGLS is that the covariance matrix Ω has $N(N - 1)/2$ elements (it is symmetric, so it doesn't have $N \times N$ elements). Thus, it always has more elements than you have observations. So, you cannot estimate the covariance matrix of the disturbances without putting some structure on it. We'll do this over and over later on.

4 Inefficient OLS

1. What if disturbances are not spherical? OLS is inefficient, but so what? Quit your bellyachin'—it still minimizes prediction error, it still forces orthogonality of disturbances to regressors, it is still easy to do, easy to explain, just plain easy.
2. But, with non-spherical disturbances, the OLS estimated coefficient variance is different from when disturbances are spherical. Consider the model

$$\begin{aligned} Y &= X\beta + \varepsilon \\ E[X'\varepsilon] &= 0_K \\ E[\varepsilon\varepsilon'] &= \Omega \neq \sigma^2 I_N. \end{aligned}$$

Recall that

$$\begin{aligned} E\left[\left(\hat{\beta}_{OLS} - \beta\right)\right] &= E\left[\left(X'X\right)^{-1}X'X\beta + \left(X'X\right)^{-1}X'\varepsilon - \beta\right] \\ &= E\left[\left(X'X\right)^{-1}X'\varepsilon\right] = \left(X'X\right)^{-1}0_K = 0_K \end{aligned}$$

The variance of the estimated parameter vector is the expectation of the square of this quantity:

$$\begin{aligned} V\left[\hat{\beta}_{OLS}\right] &= E\left[\left(\hat{\beta}_{OLS} - \beta\right)\left(\hat{\beta}_{OLS} - \beta\right)'\right] \\ &= E\left[\left(X'X\right)^{-1}X'\varepsilon\varepsilon X\left(X'X\right)^{-1}\right] \\ &= \left(X'X\right)^{-1}X'E\left[\varepsilon\varepsilon'\right]X\left(X'X\right)^{-1} \\ &= \left(X'X\right)^{-1}X'\Omega X\left(X'X\right)^{-1}. \end{aligned}$$

If $\Omega = \sigma^2 I_N$, a pair of $X'X$'s cancel leaving $\sigma^2 (X'X)^{-1}$. If not, then not.

3. It seems like you could do something like with the spherical case to get rid of the bit with Ω : After all $E[\varepsilon\varepsilon'] = \Omega$, so perhaps we could just substitute in some errors. For example, we could compute

$$\left(X'X\right)^{-1}X'ee'X\left(X'X\right)^{-1}.$$

Unfortunately, since OLS satisfies the moment condition $X'e = 0$, this would result in

$$\left(X'X\right)^{-1}0_K 0_K' X\left(X'X\right)^{-1} = 0_K 0_K'.$$

So, that's not gonna work.

4. The problem for estimating Ω is the same as with FGLS: Ω has too many parameters to consistently estimate without structure. You might think that a model like that used for WLS might be restrictive enough: you reduce Ω to just N variance parameters and no off-diagonal terms. Unfortunately, with N observations, you cannot estimate N parameters consistently.
5. The trick here is to come up with an estimate of $X'\Omega X$. There are many strategies, and they are typically referred to as 'robust' variance estimates (because they are robust to nonspherical disturbances) or as 'sandwich' variance estimates, because you sandwich an estimate $\widehat{X'\Omega X}$ inside a pair of $(X'X)^{-1}$'s. For the same reason as above, you cannot substitute ee' for Ω , because you'd get $X'\Omega X = X'ee'X = 0$.
- (a) General Heteroskedastic disturbances. Imagine that disturbances are not correlated with each other, but they don't have identical variances. We use the *Eicker-White Heterorobust variance estimator*: restrict Ω to be a diagonal matrix, construct

$$\widehat{X'\Omega X} = X'DX$$

where D is a diagonal matrix with $(e_i)^2$ on the main diagonal.

- (b) You cannot get a consistent estimate of D , because D has N elements: adding observations will not increase the precision of the estimate of any element of D .
- (c) However, $X'DX$ is only $K \times K$, which does not grow in size with N . Recall that *asymptotic variance* is equal to the variance divided by N , and it is used because the variance goes to 0 as the sample size goes to infinity. To talk about variance as the sample size grows, you have to reflate it by something, in this case N . (The choice of what to reflate it by underlies much of nonparametric econometric theory—in some models, you have to reflate by N raised to a power less than 1). So,

$$\text{asy.V} \left[\hat{\beta}_{OLS} \right] = \frac{1}{N} (X'X)^{-1} X'\Omega X (X'X)^{-1},$$

and

$$\begin{aligned} \text{asy.}\hat{V} \left[\hat{\beta}_{OLS} \right] &= \frac{1}{N} (X'X)^{-1} \widehat{X'\Omega X} (X'X)^{-1} = \\ &= (X'X)^{-1} \frac{\widehat{X'\Omega X}}{N} (X'X)^{-1}. \end{aligned}$$

Consider a model where $X = 1$, a column of ones. Then,

$$\frac{\widehat{X'\Omega X}}{N} = \frac{X'DX}{N} = \frac{\sum_{i=1}^N (e_i)^2}{N}.$$

As N grows, this thing gets closer and closer to σ^2 .

- (d) Spatially correlated disturbances. Imagine that within groups of observations, disturbances are correlated, but across groups, they are not. We use the *clustered variance estimator*: restrict Ω to be a block-diagonal matrix, construct

$$\widehat{X'\Omega X} = X'CX$$

where C is block diagonal, with elements equal to $e_i e_j$ (or their average) in the blocks and zero elsewhere.