Non-Spherical Errors

Krishna Pendakur

February 15, 2016

1 Efficient OLS

1. Consider the model

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

2. Consider the estimated OLS parameter vector

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

3. Its bias is

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

The variance of the estimated parameter vector is the expectation of the square of $\left(X'X\right)^{-1} X'\varepsilon$, which is the deviation between the estimate and its mean (which happily is its target):

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

4. How do we know it is the lowest variance linear unbiased estimator? [Gauss-Markov Theorem]

(a) Let

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

and note that $E[\varepsilon] = 0_N \Longrightarrow E\left[X'\varepsilon\right] = 0_K$ for fixed X.

(b) Let CY be another estimator, also linear, where $C = \left(X'X\right)^{-1} X' + D$ and D is a KxN matrix. This deviates from the OLS estimator by the matrix D.

(c) We have already shown the OLS estimator to be unbiased under the conditions above, so D will have to be pretty special.

$$E\left[CY\right] - \beta = \left(X'X\right)^{-1} X'X\beta + DX\beta + E[D\varepsilon] - \beta$$

$$E\left[CY\right] - \beta = \beta + DX\beta + 0 - \beta$$

$E[\varepsilon] = 0_N \Longrightarrow E\left[D'\varepsilon\right] = 0_K$. Unbiasedness thus requires $DX = 0$.

(d) As with the OLS estimator, since it is unbiased, $CY - \beta = C\varepsilon$. Its variance is the expectation of the square of this:

$$V[CY] = V[C\varepsilon] = \sigma^2 CC'$$

$$= \sigma^2[\left(X'X\right)^{-1} X'X\left(X'X\right)^{-1} + DX\left(X'X\right)^{-1} + \left(X'X\right)^{-1} X'D' + DD']$$

$$= \sigma^2[\left(X'X\right)^{-1} + 0 + 0 + DD']$$

(e) Since DD' is a square, it is positive semidefinite, and its minimum is zero when $D = 0$. Consequently, the lowest-variance unbiased estimator for the homoskedastic linear model is when $D = 0$, which is the OLS estimator.

5. Getting back to $V\left[\hat{\beta}\right] = \sigma^2\left(X'X\right)^{-1}$, a problem is that σ^2 is not observed. So, we don't yet have a useful object.

(a) 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?

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

e is a linear transformation of ε. However, although $\left[I - X\left(X'X\right)^{-1}X'\right]$ is an NxN matrix, it is not a full rank matrix: its columns are related. Indeed, this NxN 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 $\left[I - X\left(X'X\right)^{-1}X'\right]$ has rank $N - K$.

7. Matrices like $\left[I - X\left(X'X\right)^{-1}X'\right]$ and $X\left(X'X\right)^{-1}X'$ 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. Getting back to σ^2 and our estimate of it, compute $e'e$ in terms of ε:

   $$e = M_X \varepsilon$$

   so,

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

   because $\varepsilon' P_X \varepsilon = \varepsilon' P_X' P_X \varepsilon$, which is the sum of squares of $P_X\varepsilon$. P_X has rank K and projects ε on to X. Even though ε is noise, you can still get K perfect fits with K regressors, so its sum of squares picks up these K perfect fits. Consequently,

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

So, we can use an estimate

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

An *estimate* of the variance of the OLS estimator is thus given by

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

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

9. Time to crack open Kennedy for Ballentines (Kennedy, Peter E. 2002. More on Venn Diagrams for Regression, *Journal of Statistics Education* Volume 10, Number 1), linked on the course website. Go through ε vs $cov(X, Y)$ and the 2 regressor case.

   (a) size of circle is amount of variation.

   (b) overlap in circles is amount of covariation.

   (c) more overlap means more information to identify the parameter associated with that regressor.

   (d) with 2 regressors, there is overlap between all 3 variables. this covariation cannot be *uniquely* attributed to either regressor. this covariation can, however, be attributed to a linear combination of the two regressors.

   (e) show correlated missing regressor case. show uncorrelated missing regressor case.

   (f) go through Frisch-Waugh-Lovell with the circles.

10. Precision is good. Low variance is precision. How do you get a precise estimate? One can think about $V\left[\hat{\beta}_{OLS}\right] = \sigma^2 \left(X'X\right)^{-1}$ 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 variance matrix of the vectors x_i', $i = 1, ..., N$. Indeed, for $x_i \sim iid$, $X'X$ is an estimate of $E[x_i'x_i]$. By definition, $E[x_i'x_i] = E[x_i]E[x_i]' + V[x_i]$. For x_i with a given mean $E[x_i]$, increasing $V[x_i]$ with a mean-preserving spread implies increasing $E[x_i'x_i]$, which in turn is associated with larger $X'X$. If $X'X$ is bigger, then $(X'X)^{-1}$ is smaller, so $V\left[\hat{\beta}_{OLS}\right]$ is smaller and the estimate $\hat{\beta}_{OLS}$ is more precise.

    (d) If the columns of X covary a lot, then the off-diagonals of $X'X$ are larger. If the off-diagonals of $X'X$ are larger, then the diagonals of $(X'X)^{-1}$, which give the variance of each estimated coefficient, are bigger, which means we have less precision.

i. Consider a model with 2 columns in X and no constant, and let the columns of X be positively correlated.

ii. Then, $X'X$ has elements $X_j'X_k$ for $j,k = 1,2$. Its diagonals are the sum of squares of each column (call them a and b), and its off-diagonal is the cross-product (call that c). So, $X'X$ is given by $\begin{bmatrix} a & c \\ c & b \end{bmatrix}$. Then, $(X'X)^{-1}$ is given by $\frac{1}{ab-c^2}\begin{bmatrix} b & -c \\ -c & a \end{bmatrix}$, whose diagonal elements are $\frac{b}{ab-c^2}$ and $\frac{a}{ab-c^2}$. These elements increase as c, the cross-product of the columns, increases. So, more correlation of the X's means higher variance of the estimates.

(e) If the columns of X covary a lot, then although we have less precision on any one regressor's coefficient, we may have a lot of precision on particular linear combinations of coefficients.

i. Suppose the two variables mentioned above strongly positively covary, so that c is positive and big. The covariance of the two coefficients is $-c/(ab-c^2)$. When c goes up, the numerator goes up and the denominator goes down, so the overall ratio goes up a lot. Since a, b are both sums of squares, they are positive. The cross-product c cannot exceed a or b, so the denominator is positive. Thus, these positively covarying regressors would result in estimated coefficients that each have high variance, but are strongly negatively correlated.

ii. Consider the variance of the sum of the two coefficients: $V(\beta_1 + \beta_2) = \frac{1}{ab-c^2}(b + b - 2c)$. The larger is c, the smaller is this variance.

11. The Frisch-Waugh-Lovell Theorem can be expressed in a simple way using projection matrices. Let

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

and consider premultiplying the whole thing by the error-maker matrix for X_2, M_{X_2}. This gives

$$M_{X_2}Y = M_{X_2}X_1\beta_1 + M_{X_2}X_2\beta_2 + M_{X_2}\varepsilon.$$

The projection of X_2 onto itself is perfect, so it has no error, so $M_{X_2}X_2 = 0$. Thus, we have

$$M_{X_2}Y = M_{X_2}X_1\beta_1 + M_{X_2}\varepsilon.$$

Writing $\hat{Y} = M_{X_2}Y$ and $\hat{X}_1 = M_{X_2}X_1$, we have

$$\hat{Y} = \hat{X}\beta_1 + M_{X_2}\varepsilon.$$

Since X_2 is uncorrelated with ε by assumption, $M_{X_2}\varepsilon = \varepsilon - X_2(X_2'X_2)^{-1}X_2'\varepsilon$ is also uncorrelated with $\hat{X}$:

$$E[\hat{X}_1'M_{X_2}\varepsilon] = E[X_1'M_{X_2}'M_{X_2}\varepsilon] = E[X_1'M_{X_2}\varepsilon]$$

$$= E[X_1'\varepsilon] - E[X_1'X_2(X_2'X_2)^{-1}X_2'\varepsilon] = 0 - X_1'X_2(X_2'X_2)^{-1}E[X_2'\varepsilon] = 0$$

so this error term is exogenous with respect to $\hat{X}$.

(a) so you can regress $\hat{Y}$ on $\hat{X}$ to get an estimate of β_1.

(b) In terms of Kennedy's Ballentines, $M_{X_2}Y$ is the part of Y that has had X_2 cleaned out of it, and $M_{X_2}X_1$ is the part of X_1 that has had X_2 cleaned out of it.

2 NonSpherical errors

1. In a model

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

if errors satisfy

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

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

(a) The first implication of spherical errors is

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

for all $i = 1, ..., N$, which we usually call *homoskedasticity*. Homoskedastic errors 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 errors across observations. This rules out over-time correlations in time-series data, and spatial correlations in cross-sectional data.

2. OLS is inefficient if errors are nonspherical. This is easy to see by example.

(a) 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

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

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

i. 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.

ii. The point is that you want to pay close attention where the errors have low variance and not pay much attention where the errors have high variance.

(b) Alternatively, imagine that

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

Here, the notation ι_K indicates a $K-$vector of ones. Thus, $\iota_{N-2}\iota'_{N-2}$ is an $(N-2)x(N-2)$ 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 errors with variance σ^2, and observations $2, ..., N-1$ had the same error term. Not just error terms drawn from the same distribution, but literally the same value of ε for each of those observations.

i. 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 error, you wouldn't want to pull the regression line up very much, because you'd know that what seemed like a lot of positive errors 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 known form of nonspherical errors: either heteroskesticity, or correlations across observations. Note that $E[\varepsilon] = 0 \Rightarrow E[X'\varepsilon] = 0$ for fixed X.

2. We know that OLS is the efficient estimator given homoskedastic errors, 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 error term premultiplied by this weird inverse-matrix-square-root thing.

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

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

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

5. So the premultiplied model is homoskedastic with unit variance errors.

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. Tranforming data by a known variance matrix and then applying OLS is called *Generalised Least Squares (GLS).*

8. We refer to the matrix

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

   as the *Transformation Matrix.*

9. GLS in Stata is

10. `reg TY TX`

3.1 Group Means Regressands

1. 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 σ^2 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 errors are not correlated across countries, and the covariance matrix is

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

   and, therefore,

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

2. The transformation matrix T amounts to multiplying each Y and each X by the square root of the sample size used in each country.

3. One need not include the scalar σ in T, because leaving it out just loads it on to the second stage which would have variance $\sigma^2 I_N$ instead of I_N.

4. This strategy, in which you premultiply each observation separately, rather than pre-multiplying 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).

   (a) in Stata,

   (b) `reg Y X [aweight=S]`

3.2 Feasible Generalized Least Squares

1. GLS is all great if you know the covariance matrix of the errors, 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.

2. 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.

3. The Random Effects Model uses FGLS

   (a) Assume that

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

   (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 errors, 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:

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

where $I[.]$ 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) 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.

i. `reg Y X i.person`
ii. compute the variance of the person dummies for σ_θ^2 and use the estimate of the variance of the fixed effects error term for σ_ε^2.
iii. `reg Y X`
iv. take the average squared error as an estimate of σ_ε^2 and the average cross-product of errors for a given person as an estimate of $\sigma_\theta^2 + \sigma_\varepsilon^2$.

(g) Now, form Ω and T and run GLS.

(h) The FGLS estimator uses a consistent pre-estimate of Ω, but this estimate is only exactly right asymptotically. Thus, the FGLS estimator is only efficient asymptotically. In the small-sample, it could be kind of crappy because the pre-estimate of Ω might be kind of crappy.

4. The trick with FGLS is that the covariance matrix Ω has $N(N-1)/2$ elements (it is symmetric, so it doesn't have NxN elements). Thus, it always has more elements than you have observations. So, you cannot estimate the covariance matrix of the errors without putting some structure on it. We'll do this over and over later on.

4 Inefficient OLS

1. What if errors are not spherical? OLS is inefficient, but so what? Quit your bellyachin'—it still minimizes prediction error, it still forces orthogonality of errors to regressors, it is still easy to do, easy to explain, just plain easy.

2. But, with non-spherical errors, the variance of the OLS estimated coefficient is different from when errors are spherical. Consider the model

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

   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 $\sigma^2 \left(X'X\right)^{-1}$. 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 $E\left[\varepsilon\varepsilon'\right] = \Omega$, 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. Robust Standard Errors.

   (a) The trick here is to come up with an estimate of $X'\Omega X$ (which is a KxK matrix). There are many strategies, and they are typically referred to as 'robust' variance estimates (because they are robust to nonspherical errors) 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 $\widehat{X'\Omega X} = X'ee'X = 0$.

   (b) General Heteroskedastic errors. Imagine that errors are not correlated with each other, but they don't have identical variances. We use the *Eicker-White Hetero-robust variance estimator*.

      i. First, restrict Ω to be a diagonal matrix with diagonal elements σ_i^2 and off-diagonal elements equal to 0. This is the structure you have imposed on the model: diagonal Ω.

      ii. Then, construct an estimate of $X'\Omega X$ that satisfies this structure:

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

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

   (c) 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.

   (d) However, $X'DX$ is only KxK, 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,

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

   and

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

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

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

As N grows, this thing gets closer and closer to the average σ_i^2.

(e) In Stata, you can get the hetero-robust standard errors as follows:

(f) `reg Y X, robust`

6. Clustered Standard Errors.

(a) Imagine that within groups of observations stratified by a grouping variable g, errors are correlated, but across groups, they are not. Let $g = 1, .., G$ and let each group have n_g observations (and note that $N = \sum_g n_g$).

i. For example, suppose your data are drawn from a multistage sample wherein first we sample city blocks and then we sample individuals within those city blocks. The individuals on the same block might actually know each other, and thus have correlations across each other.

ii. Or, suppose that you are interested in networks and your underlying data are people, but the data you run regressions on compare distances between pairs of people in different groups and the interactions among pairs of people in those groups. Then, you would almost certainly have correlations between people within groups.

iii. Or, suppose that you have data on patients in hospitals, many patients in each of many hospitals. Certainly, diseases can travel from person to person in a hospital, but less so across them. So, you'd expect correlations across patients in hospitals but not across them.

(b) In this environment, Ω has a blockish structure. Sort the data by groups. The across group blocks of Ω are 0, but the within-group blocks of Ω are non-zero. The upper-left block is an $n_1 x n_1$ symmetric matrix with unknown elements. To its right, we find an $n_1 x (N - n_1)$ matrix of zeroes, and below it, we find an $(N - n_1) x n_1$ matrix of zeroes. The next diagonal block is an $n_2 x n_2$ symmetric matrix with unknown elements. And so it goes.

(c) This is very like the random-effect model, except that we have not imposed the further structure that within-group correlations are all the same and all the groups are the same.

i. So, it is like a random-effects structure, but much more general. Unfortunately, it is so much more general that we cannot use FGLS as we would with random-effects. The reason is that FGLS requires a consistent estimate of Ω. This model has $\sum_g n_g (n_g + 1)/2$ parameters, which increases faster than N. So, we cannot construct a consistent estimate of Ω.

(d) So, analogous to the hetero-robust standard errors, we use the *clustered variance estimator*. We construct an estimate of $X'\Omega X$ that is consistent with the structure we've imposed. In particular, 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. Then,

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

(e) In Stata,

(f) `reg Y X, cluster(g)`

(g) A question for you: why can't we go maximally general, and have just 1 big cluster?