ECONOMICS 7344, Spring 2017
Bent E. Sørensen
April 12, 2017

1 Introduction to Generalized Least Squares

Consider the model

$$Y = X\beta + \epsilon \ ,$$

where the $N \times K$ matrix of regressors X is fixed, independent of the error term, and of full rank, but the error variance is no longer $\sigma^2 I$ but $var(\epsilon) = \sigma^2\Omega$. We will start with the general theory and work on the common examples next. Also notice that sometimes people will write $var(\epsilon) = \Omega$, but it is often convenient to have a factor of proportionality "outside." In most applications, Ω is not known, but first analyze the case where it is known. (Sometimes one knows from the way the data is constructed.)

Because Ω is a variance matrix it is symmetric and positive definite, so we can take the square root of both Ω and Ω^{-1}. Let us assume for simplicity that we take a symmetric square root (although I will later make another choice—it does not matter for the following). What we want to use is that $\Omega^{-1/2}\Omega\Omega^{-1/2} = I$. Consider then the transformed equation

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

If we define

$$\tilde{Y} = \Omega^{-1/2}Y \ ,$$

and

$$\tilde{X} = \Omega^{-1/2}X \ ,$$

and

$$U = \Omega^{-1/2}\epsilon \ ,$$

we have

$$(*) \quad \tilde{Y} = \tilde{X}\beta + U \ .$$

This is a linear equation with $var(U) = \sigma^2 I$, so it satisfies all the OLS assumptions. (Also, if ϵ is normally distributed, so is U.) Of course, we know to estimate equation (*) efficiently, namely by the OLS estimator

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

So now let us substitute in the definition of these variables and get

$$\hat{\beta} = [(\Omega^{-1/2}X)'\Omega^{-1/2}X]^{-1}(\Omega^{-1/2}X)'\Omega^{-1/2}Y \ ,.$$

or, because $\Omega^{-1/2}\Omega^{-1/2} = \Omega^{-1}$,

$$\hat{\beta} = [X'\Omega^{-1}X]^{-1}X'\Omega^{-1}Y \ ,$$

which is the GLS-estimator. (Sometimes, I will label it $\hat{\beta}^{gls}$ or something like that if we need to discuss both OLS and GLS estimators.) Fortunately, it is easy implement because we do not actually need to take the square roots of the matrices...although, for modern computers and algorithms, it doesn't matter much.

The feasible GLS estimator. In many cases, the variances and covariances of the residuals are not known, so we need to estimate them from the data. Later, we will talk about Maximum Likelihood estimation, but commonly people use a 2-step estimator.

Step 1 (OLS), estimate $\hat{\beta}^{ols} = (X'X)^{-1}X'Y$. (Show that OLS is unbiased, later we will show that it is consistent.). Because OLS is unbiased and consistent, the error terms $e = Y - X\hat{\beta}^{ols}$ are unbiased estimates of the true errors. We can therefore try and estimate

$$\hat{\sigma^2\Omega} = \frac{1}{N-K} e\, e' \ .$$

This of course does not work because there are even more elements in Ω than we have data points. So, we need to make assumptions on Ω in order to limit be able to estimate the variance matrix. We will discuss the main cases next, but first: assuming we have a valid estimate of $\hat{\Omega}$, we do

Step 2 (Feasible GLS)

$$\hat{\beta}^{fgls} = [X'\hat{\Omega}^{-1}X]^{-1}X'\hat{\Omega}^{-1}Y \ .$$

This estimator will be consistent if $\hat{\Omega}$ is, although it will not be unbiased, because $\hat{\Omega}$ is a random variable and it is not easy to find the expectation of the inverse of a random matrix (further, because it is estimated from the data, $\hat{\Omega}$ is no independent of the error terms).

Main example 1: Heteroskedasticity

Sometimes we suspect, or can reasonably assume, that the error terms are independent but not identically distributed. In, other words

$$E\epsilon_i^2 = \sigma_i^2 ; \quad E\epsilon_i\epsilon_j = 0 \,\text{if}\ i \neq j \ .$$

If you know σ_i for all i, you can divide by σ_i and obtain

$$\tilde{y}_i = \tilde{x}_i\beta + u_i$$

where u_i satisfies the OLS conditions. u_i is ϵ_i/σ_i although we of course divide only y and all the columns of x. (Because $u_i = \tilde{y}_i - \tilde{x}_i\beta$ this is equivalent to having also divided ϵ_i by its standard

deviation.) So dividing by σ_i is the same as multiplying the vector with $\Omega^{-1/2}$ in this case. As you will see more clearly for the case of autocorrelation, if you know the linear functions (of y and x) that creates independent homoskedastic error terms, you have found a version of $\Omega^{-1/2}$.

One practical issue. Consider the typical model

$$y_i = \alpha_0 + \alpha_1 x_i + \epsilon_i \ .$$

(There would typically be more than one regressor, but the logic would be the same for the following point.) If you divide by the standard error σ_i you get the relation

$$(*) \quad y_i/\sigma_i = \alpha_0(1/\sigma_i) + \alpha_1(x_i/\sigma_i) + u_i \ .$$

So you have to run a regression with two-regressors and no constant. This is the answer I expect if I ask, unless I explicitly suggest something else. In practice, people sometimes do not construct the regressor where the vector of ones are divided by the standard deviations) and sometimes they do and also include a constant. This may or may not make sense. If you are running a non-structural regression, you should think about which specification to use, although (*) is the default. If the original equation is derived from a model (a structural relation), you have to explicitly use the (*) regression. In the structural case, you now have the variance of the error term in the transformed regression having variance unity and you should use that in your tests.
In many situations, we do not know σ_i but suspect it varies with, say, x_i, which you may verify by a test or a more informally with a plot. If you have decided that $\sigma_i = \sigma x_i$ you can estimate by GLS (not just *feasible* GLS), because you divide the variable by the observable x_i and you the variance of u_i equal to the unknown σ^2, but that is the standard OLS situation. (This is why text-books often writes $\sigma^2\Omega$ for the variance matrix. If Ω somehow is know (or maybe estimated), we are back in the OLS case with the transformed variables if σ is unknown. (If it is known, you still do $(X'X)^{-1}X'Y$ to find the coefficients, but you use the known constant when calculating $t-$stats etc.)

In the case where for example $\sigma_i = \gamma_1 X_i^1 + \gamma_2 X_i^2$, you would usually not know the values of the γs so you would estimated them and use feasible GLS, dividing the variables by $\hat{\gamma}_1 X_i^1 + \hat{\gamma}_2 X_i^2$. Notice that because the γs are estimated, they are not equal to the true values, so we have introduced measurement error. More generally, this is known as the "*generated regressor*" issue. If your sample is large, $\hat{\gamma}_i \approx \gamma_i$ and you can ignore the issue. If not, you will have to correct the standard errors somehow (which we will talk about later, maybe in Econometrics II).

Main example 2: Autocorrelated residuals

A collection of stochastic variables $x_1, .., x_t, .., x_T$ indexed by an integer value t. The interpretation is that the series represent a vector of stochastic variables observed at equal-spaced time intervals. The series is also some times called a stochastic process.

Consider a time series of error terms ϵ_t where t is time. (In rare occasion, something else, like physical distance.) We assume the time series is stationary with auto-covariance $Ee_te_{t-k} = \gamma(k)$. The $\gamma(k)$'s for $k \neq 0$ are called **autocovariances** and if we divide by the variance we obtain the **autocorrelations** $\rho(k) = \gamma(k)/\gamma(0)$. These are the correlation of x_t with it own lagged values.

Note that if Ω_T is the matrix of variances and covariance of $e_1, ..., e_T$ then

$$\Omega_T \;=\; \begin{pmatrix} \gamma(0) & \gamma(1) & \gamma(2) & \dots & \gamma(T-1) \\ \gamma(1) & \gamma(0) & \gamma(1) & \dots & \gamma(T-2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma(T-2) & \dots & \dots & \dots & \gamma(1) \\ \gamma(T-1) & \gamma(T-2) & \dots & \gamma(1) & \gamma(0) \end{pmatrix} .$$

So if we let Ψ_T be the matrix of autocorrelations; i.e., $\Omega_T = \gamma(0)\Psi_T$ we will have

$$\Psi_T \;=\; \begin{pmatrix} 1 & \rho(1) & \rho(2) & \dots & \rho(T-1) \\ \rho(1) & 1 & \rho(1) & \dots & \rho(T-2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho(T-1) & \rho(T-2) & \dots & \rho(1) & 1 \end{pmatrix} .$$

Time series models are simple models for the (auto-) correlation of the e_t's that makes the auto-covariance matric a function of a small number of parameters that we can estimate. (This is the statisticians perspective, the models are use extensively as building blocks in modern macroeconomics.)

The most commonly used type of time series models are the auto regressive (AR) models. We will focus on the AR(1) model, that is the most used by far. We have

$$e_t \;=\; ae_{t-1} \;+\; u_t \;,$$

where the innovation u_t is white noise with constant variance σ^2. Here k is a positive integer called the order of the AR-process. An AR(1) model is a way of writing the conditional expectation in a simple manner. We have

$$u_t = e_t - ae_{t-1} \;,$$

and because the u_t's are i.i.d., subtraction the ae_{t-1} from e_t is equivalent to multiplying the e vector by $\Omega^{-1/2}$. But wait: we cannot do that for the first innovation e_1. We know it is independent of u_2, u_3, etc. so we just need to normalize by it standard deviation. The variance of e_t (assuming stationarity) is $\sigma_u^2/(1-a^2)$, so $u_1 = e_1 * (1-a)$ has variance σ_u^2.
So if Ω is the variance matrix for $(e_1, ..., e_T)'$, what does $\Omega^{-1/2}$ look like. Well it is the transforma-

tion, such that $u = \Omega^{-1/2} e$ has variance I. So, based on what we have found ,

$$\Omega^{-1/2} \;=\; \frac{1}{\sigma_u} \begin{pmatrix} \sqrt{1-a^2} & 0 & \dots & 0 & 0 \\ -a & 1 & & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & -a & 1 \end{pmatrix} .$$

(Because everything is proportianal to the variance σ_u^2, I sometimes forget it.) Ω can easily be calculated as

$$\Omega \;=\; \frac{\sigma_u^2}{1-a^2} \begin{pmatrix} 1 & a & a^2 & \dots & a^{T-1} \\ a & 1 & a & a^2 & a^{T-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a^{T-1} & a^{T-2} & \dots & a & 1 \end{pmatrix} .$$

You should verify that $\Omega^{-1/2}\,\Omega\,\Omega^{-1/2'} = I$ for a low dimension—see homework (remember the transposition, this particular choice of $\Omega^{-1/2}$ is not symmetric.

The transformation is called the Prais-Winsten transformation. In large samples, people often just drop the first observation (meaning that they use it in the first stage, and then calculate $u_t = \hat{e}_t - \hat{a}\hat{e}_{t-1}$ for $t = 2, ...T$. This is called the Cochrane-Orcutt transformation.

You can do something similar for an AR(2) model, but this takes some thinking in order to figure out how to make the first two errors i.i.d. So usually people do Cochrane-Orcutt. Or, if your sample is small, you can have the computer solve the problem without calculating the exact formula, but you have to rely on numerical calculations. We will get to that later, maybe not until Econometrics II.

1.1 MA models:

The simplest time series models are the moving average (MA) models:

$$e_t \;=\; \mu \;+\; u_t \;+\; b_1 u_{t-1} \;+\; ... \;+\; b_l u_{t-l} \;=\; \mu \;+\; b(L) u_t,$$

where the innovation u_t is white noise and the lag-polynomial is defined by the equation. The positive integer l is called the **order** of the MA-process. MA processes are quite easy to *analyze* because they are given as a sum of independent (or uncorrelated) variables. However, they are not so easy to *estimate* econometrically: since (in almost all applications of this) is only the e_t's that are observed, the u_t's are unobserved, i.e., latent variables, that one cannot regress on. For the purpose of our class, where we use the models as modeling tools, this is a parenthetic remark.]

Consider the simple scalar MA(1)-model (I leave out the mean for simplicity)

$$(*)\ e_t \ = \ u_t \ + \ bu_{t-1} \ .$$

If u_t is an independent series of $N(0, \sigma_u^2)$ variables, then this model really only states that x_t has mean zero and autocovariances: $\gamma(0) = (1 + b^2)\sigma_u^2;\ \gamma(1) = \gamma(-1) = b\sigma_u^2;\ \gamma(k) = 0;\ k \neq -1, 0, 1.$ (Notice that I here figured out what the model says about the distribution of the observed x's. In some economic models, the u_t terms may have an economic interpretation, but in many applications of time series the the MA- (or AR-) model simply serves as a very convenient way of modeling the autocovariances of the $x-$ variables.)

The Ω matrix for the MA(1) is

$$\Omega \ = \ \sigma_u^2 \begin{pmatrix} 1+b^2 & b & 0 & \dots & 0 \\ b & 1+b^2 & b & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & b & 1+b^2 \end{pmatrix} .$$

This looks fairly simple, but I cannot deduce how to find the square root. If the sample is large, I can approximate by assuming $u_0 = 0$; if I do that then $u_1 = e_1$ and $u_2 = e_2 - b\,u_1$ and in general $u_t = e_t - b\,u_{t-1}$. The problem, as you see, is that the u's, while convenient building blocks in models are unobserved in the econometrics application. So if your sample is small, you may want to numerically invert Ω (again, we will cover that later).
Consider equation (*) again. In lag-operator notation it reads

$$e_t \ = \ (1 \ + \ bL)u_t \ ,$$

which can be inverted to

$$u_t \ = \ (1 \ + \ bL)^{-1}e_t \ = \ e_t \ - \ be_{t-1} \ + \ b^2e_{t-2} \ + \ ...$$

It is quite obvious that this expression is not meaningful if $|b| \geq 1$ since the power term blows up. In the case where $|b| < 1$ the right hand side converges to a well defined random variable. So, conceptually, we do not observe the u's because the are residuals in particular infinite AR-process.