GLS and Serial Correlation

Greene, CH. 20
R script `mod4s3a, mod4s3b`

Introduction

The phenomenon of serial correlation (often also referred to as "*autocorrelation*") arises when regression errors are correlated with one or more adjacent errors in a data set. This usually occurs in any type of time series data, where we have repeated observations over time for a given person, household, firm, or any other cross-sectional unit.

If serial correlation (SC) is present that OLS estimator exhibits the same flaws as under HSK: While still consistent, its variance is mis-specified, and we can't trust standard errors, t-values, or hypothesis tests. If the "correct" version of its asymptotic variance is used, the estimator still remains inefficient compared to GLS or a (correctly specified) FGLS model.

Before we look into the specifics of SC it will be useful to review some basic concepts of time series analysis, as presented in Greene's Ch. 20.

The Analysis of Time Series Data

A typical time series (TS) model relates time-specific observations of the dependent variable to contemporaneous (and possibly lagged) explanatory variables, optionally lagged observations of the dependent variable, and an error term, often referred to as "*disturbance*," "*shock*" or "*innovation*". Here is a simple example:

$$y_t = \beta_1 + \mathbf{x}_t'\boldsymbol{\beta}_{\mathbf{x}} + \beta_y y_{t-1} + \varepsilon_t, \tag{1}$$

where t indicates a specific time period (e.g. day, month, quarter, year, etc.) and $\boldsymbol{\beta}_{\mathbf{x}}$ contains as many elements as $\mathbf{x}_t$.

It should be immediately noted that the *inclusion of one or more lags of the dependent variable* raises a whole set of econometric concerns, not just autocorrelation. For the remainder of this chapter we will thus abstract from this case and focus instead on other reasons that may lead to correlated errors.

The conceptual relationship between a "sample" and an "underlying population" is a bit different for TS models than for cross-sectional (CS) data (i.e. all models we have considered so far). For CS data we have interpreted a "sample" as a set of random draws from an underlying, much larger, population. We have also assumed that at least in theory, we could repeat the sampling process many times, which justifies the notion of a "sampling distribution" of statistical estimators (such as the sample mean, or the OLS estimator **b**). We have further used the idea that one could in theory increase the sample size to "infinity", which led to the derivation of asymptotic properties of estimators.

For TS models the equivalent to an underlying "*population*" is the hypothetical "infinite" series, or "TS process":

$$\{y_t\}_{t=-\infty}^{t=\infty} \tag{2}$$

Each TS process is characterized by (1) the time ordering (i.e. the order in which the time periods are considered, usually chronological), and (2) the correlation patterns between observations in the sequence. The analogous concept to a "sample" in CS analysis is a "*time window*", i.e. a set of (usually consecutive) observations of y_t from $t = 1....T$, where "$t = 1$" is just an arbitrary labeling of the starting point. The analog of "repeated sampling" in CS analysis is the notion that we could "draw" an infinite number of different time windows from $\{y_t\}_{t=-\infty}^{t=\infty}$. Finally, the analog to "increasing sample size" in CS analysis is the notion of an ever increasing time window in the TS context.

Disturbance Processes

In the usual TS setting we assume that the error terms have mean zero and equal variance, but that they are correlated across observations, i.e.

$$E(\boldsymbol{\varepsilon}) = 0 \qquad E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \boldsymbol{\Omega} \qquad \text{where}$$

$$\boldsymbol{\Omega} = \begin{bmatrix} \sigma^2 & \text{cov}(\varepsilon_1, \varepsilon_2) & \cdots & \text{cov}(\varepsilon_1, \varepsilon_T) \\ \text{cov}(\varepsilon_2, \varepsilon_1) & \sigma^2 & \cdots & \text{cov}(\varepsilon_2, \varepsilon_T) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(\varepsilon_T, \varepsilon_1) & \text{cov}(\varepsilon_T, \varepsilon_2) & \cdots & \sigma^2 \end{bmatrix} \tag{3}$$

It is generally assumed that each error term is correlated with all other disturbances. However, an important constraint is assumed for these correlations: Specifically, for a "well-behaved" time series, we assume that for any *t, s,* $\text{cov}(\varepsilon_t, \varepsilon_{t+s})$ is a function only of $|t+s|$, i.e. the "distance" between the two time periods, and not of *t* or *s* in specific. This means that, for example, all disturbances have the same correlation with their immediate neighbors, the same correlation with disturbances 2 periods ahead or behind, and so on. This key property is called "*stationarity*". A TS with this property is referred to as "*covariance-stationary*".

It is common for TS analysis to work with correlations instead of covariances. Let $\text{cov}(\varepsilon_t, \varepsilon_{t-s}) = \Omega_{t,t-s} = \text{cov}(\varepsilon_{t+s}, \varepsilon_t) = \Omega_{t+s,t} = \gamma_s$ and $\text{cov}(\varepsilon_t, \varepsilon_t) = \Omega_{t,t} = V(\varepsilon_t) = \sigma^2 = \gamma_0$. We can then define the *autocorrelation* between two disturbances as

$$corr(\varepsilon_t, \varepsilon_{t-s}) = \frac{\text{cov}(\varepsilon_t, \varepsilon_{t-s})}{\sqrt{V(\varepsilon_t)V(\varepsilon_{t-s})}} = \frac{\gamma_s}{\gamma_0} = \rho_s \tag{4}$$

Thus, for the full TS we have

$$E(\varepsilon\varepsilon') = \boldsymbol{\Gamma} = \gamma_0 \mathbf{R} \qquad \text{with}$$

$$\boldsymbol{\Gamma} = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{T-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{T-1} & \gamma_{T-2} & \cdots & \gamma_0 \end{bmatrix} \qquad \mathbf{R} = \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{T-1} \\ \rho_1 & 1 & \cdots & \rho_{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{T-1} & \rho_{T-2} & \cdots & 1 \end{bmatrix}, \tag{5}$$

where $\boldsymbol{\Gamma}$ is the *auto-covariance matrix* and $\mathbf{R}$ is the *autocorrelation* matrix. Different types of TS processes imply different patterns in $\mathbf{R}$. Perhaps the most frequently analyzed process is a *first-order autoregression*, or AR(1), subject of the next topic.

AR(1) Disturbances

Similar to the CLRM for regression analysis, the AR(1) model is a time-proven "workhorse" for many applications. It is often a very reasonable approximation for much more complex TS processes that would be extremely difficult to analyze. The AR(1) model is characterized by the following properties for the disturbance terms:

$$\varepsilon_t = \rho\varepsilon_{t-1} + \mu_t \qquad E(\mu_t) = 0, \quad V(\mu_t) = \sigma_\mu^2 \tag{6}$$

The μ_t term is often referred to as "white noise". Think of it as the analog to the error term in the CLRM. We assume that the μ_t's are uncorrelated with each other and with any of the ε_t's. To assure stationarity we add the additional stipulation that $|\rho| < 1$. Under this assumption we obtain the following properties for ε_t (see Greene p. 911 for details):

$$E(\varepsilon_t) = 0, \quad V(\varepsilon_t) = \frac{\sigma_\mu^2}{1-\rho^2} = \gamma_0, \quad \text{cov}(\varepsilon_t, \varepsilon_{t-s}) = \frac{\rho^s \sigma_\mu^2}{1-\rho^2} = \gamma_0 \rho^s, \quad \text{corr}(\varepsilon_t, \varepsilon_{t-s}) = \rho^s \tag{7}$$

The autocorrelation matrix takes the following form:

$$\mathbf{R} = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & 1 \end{bmatrix}, \tag{8}$$

Thus, with the stationarity assumption the autocorrelations diminish ("fade") over time. If $\rho > 0$, the TS residuals will exhibit clusters of positive and then negative observations. If $\rho < 0$, the TS residuals will have regular oscillations (= changes) in sign. See Greene p. 904-906 for some example graphs.

Least Squares Estimation

As for HSK, when **R** (and thus $\mathbf{\Gamma}$, formerly labeled $\mathbf{\Omega}$) is fully known, $\mathbf{b_{GLS}}$ will be the most efficient estimator. The OLS estimator will be unbiased but inefficient (if the correct V(**b**) is used). As a rule of thumb, the stronger the correlation between adjacent errors (i.e. the larger $|\rho|$), the greater the efficiency gain from using GLS.

Note that in either case we can no longer clearly interpret (and thus use) the t- and F- statistics, since their distribution no longer follows *t* or *F*, even when μ_t follows a normal distribution. Thus, we need to resort to asymptotic theory to implement any specification tests. Asymptotically, the OLS estimator will be consistent for $\boldsymbol{\beta}$ and normally distributed, as long as the model does not include any lagged dependent variable (which continues to be our ongoing assumption).

In analogy to the "White-corrected" estimator for HSK, there exists a **robust** (i.e. asymptotically consistent) estimator for $\hat{V}_a(\mathbf{b})$, based on Newey and West (1987). As for HSK, this is useful when $\mathbf{\Omega}$ is unknown and there is little guidance as to its general structure. The NW estimator for autocorrelated disturbances with unspecified structure is given as

$$\hat{\hat{V}}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{S_1}(\mathbf{X}'\mathbf{X})^{-1} \qquad \text{where}$$

$$\mathbf{S_1} = (\mathbf{X}'\mathbf{E}\mathbf{X}) + \sum_{j=1}^{L}\sum_{t=j+1}^{T}\left(\left(1-\frac{j}{L+1}\right)e_t e_{t-j}\left(\mathbf{x_t'x_{t-j}} + \mathbf{x_{t-j}'x_t}\right)\right) \quad \text{with} \tag{9}$$

$$\mathbf{E} = \begin{bmatrix} e_1^2 & & & \\ & e_2^2 & & \\ & & \ddots & \\ & & & e_T^2 \end{bmatrix}$$

As before, the *e*-terms are OLS residuals. *L* denotes a "lag" of a specific number of time periods. This value has to be chosen arbitrarily. A larger *L* would reflect the assumption of a slowly fading correlation. In absence of any guidance on this *L* is often set to $T^{1/4}$ (rounded to the nearest integer) .

See script `mod4s3a` for an example.

Testing for Serial Correlation

Basic residual regression

A simple first check if autocorrelation may be an issue is to regress the OLS residuals on their lagged values, i.e using the auxiliary "CLRM" model

$$e_t = \rho e_{t-1} + v_t \tag{10}$$

The estimated slope of this regression will be an estimator of $\rho = corr(\varepsilon_t, \varepsilon_{t-1})$. A standard t- or F-test can be used to assess significance. A flat slope would imply no autocorrelation ($\rho = 0$).

A scatterplot of residuals vs. lagged residuals may also be illustrative to examine if serial correlation may be present.

Breusch-Godfrey LM test

A refined version of this elementary test is the *Breusch-Godfrey LM test*. The test can be used for a set of alternative hypotheses, each of which describes a different AR(P) process, P = 1,2, etc. The null hypothesis is always "$\rho = 0$", i.e. "no autocorrelation.

As a preparatory step to derive the BG test statistic the original matrix of regressors, **X**, needs to be augmented with P columns of lagged residuals from the OLS model. The first added column contains lag-1 residuals, with a "0" added at the beginning to preserve the row dimension of **X**. The second column has two leading 0's followed by lag-2 residuals, and so on. For the AR(1) version of this test only the lag-1 residuals need to be added to **X**. Calling the augmented **X** matrix "$\mathbf{X_0}$", the PG test statistic is computed as

$$BG = T\left(\frac{\mathbf{e}'\mathbf{X_0}\left(\mathbf{X_0}'\mathbf{X_0}\right)^{-1}\mathbf{X_0}'\mathbf{e}}{\mathbf{e}'\mathbf{e}}\right) \sim \chi^2_{(P)} \tag{11}$$

This statistic is equivalent to T times the R^2 from a regression of OLS residuals **e** on $\mathbf{X_0}$. Since $\mathbf{e}'\mathbf{X} = 0$ by the properties of the OLS model, any model "fit" (i.e. a non-zero R^2) will be due to correlation between current and lagged residuals. In contrast to the simple residual regression shown above, this test procedure controls for possible correlation between **X** and lagged error terms.

A word of caution: The alternative hypothesis for this test specifies autocorrelation (AR) of order *P OR* Moving Average (MA) of order *P*. Correlation plots will usually provide guidance to assess which time series process applies. (We will not consider MA- type processes in this course).

Durbin-Watson Test

This test is specifically designed to examine if the error terms may follow an AR(1) process. The test statistic is given by

$$d = \frac{\left(\mathbf{e_t} - \mathbf{e_{t\text{-}1}}\right)'\left(\mathbf{e_t} - \mathbf{e_{t\text{-}1}}\right)}{\mathbf{e_t}'\mathbf{e_t}} \approx 2\left(1 - \hat{\rho}\right) \tag{12}$$

where $\mathbf{e_t}$ is the vector of original OLS residuals, $\mathbf{e_{t\text{-}1}}$ is a vector of lag-1 residuals, and $\hat{\rho}$ is the estimated autocorrelation coefficient from a regression of residuals on lagged residuals (our "basic residual regression" from above). For the denominator, the full T observations of $\mathbf{e_t}$ are used. For the numerator, the first element of $\mathbf{e_t}$ is dropped to assure compatibility of dimension with the lagged residual vector.

The *d*-statistic follows a distribution that is different from any we have encountered so far. For each combination of sample size T and k (columns in the original **X**) there are a lower and upper critical value, denoted as d_L and d_U, respectively. The null hypothesis is always: $H_0 : \rho = 0$, i.e. absence of serial correlation. The researcher has to choose one of two "one-sided" alternative hypotheses: $H_1 : \rho > 0$ ("positive autocorrelation"), or $H_1 : \rho < 0$ ("negative autocorrelation"). The use of the critical thresholds for test decisions depends on which H_1 is specified.

For $H_1 : \rho > 0$, the null is rejected if $d<d_L$ and not rejected if $d>d_U$. The test is inconclusive if $d_L < d < d_U$.

For $H_1 : \rho < 0$ the null is rejected if $d > 4-d_L$ and not rejected if $d < 4-d_U$. The test is inconclusive if $4 - d_U < d < 4 - d_L$.

Feasible GLS

If tests reveal that the regression errors follow an AR(1) process, we immediately know the structure of $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \boldsymbol{\Omega} = \frac{\sigma_\mu^2}{1-\rho^2}\mathbf{R} = \sigma_\varepsilon^2\mathbf{R}$, where **R** takes the form of (8). Note that we revert to using σ_ε^2 instead of γ_0 to denote the variance of ε_t for consistency with the notation in section 19.9 of the textbook.

Step 1: Find a consistent estimator of the autocorrelation coefficient ρ. We can use again the estimated slope coefficient of our residual regression above, i.e. $\hat{\rho} = \left(\mathbf{e'_{t-1}e_{t-1}}\right)^{-1}\mathbf{e'_{t-1}e_t}$, where as before the first element of $\mathbf{e_t}$ is dropped to assure compatibility of dimension with the lagged residual vector. This allows for the computation of the estimated correlation matrix $\hat{\mathbf{R}}$

Step 2: Derive the FGLS estimator. Here it is important to note that this can be done without knowing σ_ε^2, since $\mathbf{b_{FGLS}} = \left(\mathbf{X'\hat{\Omega}^{-1}X}\right)^{-1}\mathbf{X'\hat{\Omega}^{-1}y} = \left(\mathbf{X'}\left(\boldsymbol{\sigma}_\varepsilon^2\hat{\mathbf{R}}\right)^{-1}\mathbf{X}\right)^{-1}\mathbf{X'}\left(\boldsymbol{\sigma}_\varepsilon^2\hat{\mathbf{R}}\right)^{-1}\mathbf{y} = \left(\mathbf{X'\hat{R}^{-1}X}\right)^{-1}\mathbf{X'\hat{R}^{-1}y}$.

Step 3: Derive an estimate of σ_ε^2 using

$$\hat{\sigma}_\varepsilon^2 = \frac{\left(\mathbf{y} - \mathbf{Xb_{FGLS}}\right)'\hat{\mathbf{R}}^{-1}\left(\mathbf{y} - \mathbf{Xb_{FGLS}}\right)}{T} \tag{13}$$

Step 4: Derive the variance of the FGLS estimator using

$$\hat{V}_a\left(\mathbf{b_{FGLS}}\right) = \hat{\sigma}_\varepsilon^2\left(\mathbf{X'\hat{R}^{-1}X}\right)^{-1} \tag{14}$$

See script `mod4s3a` for an example.

Additional ***R*** example:

Script `mod4s3b` covers the same steps as `mod4s3a` for a 96-month time series of average monthly water consumption by hotels in the Reno area (data set `hotel`).