Econometrics Lecture Notes (I)

Montserrat Farell*

Universitat Autònoma de Barcelona

12th April 2004

Contents

*This notes are slight modifications of part of the book *Lecture Notes in Internet* at http://pareto.uab.es/omega/Project_001 by Professor Michael Creel

1 Generalized least squares

One of the assumptions we've made up to now is that

$$\varepsilon_t \sim IID(0, \sigma^2),$$

or occasionally

$$\varepsilon_t \sim IIN(0, \sigma^2).$$

Now we'll investigate the consequences of nonidentically and/or dependently distributed errors. The model is

$$\begin{aligned} y &= X\beta + \varepsilon \\ \mathcal{E}(\varepsilon) &= 0 \\ V(\varepsilon) &= \Sigma \\ \mathcal{E}(X'\varepsilon) &= 0 \end{aligned}$$

where Σ is a general symmetric positive definite matrix (we'll write β in place of β_0 to simplify the typing of these notes).

- The case where Σ is a diagonal matrix gives uncorrelated, nonidentically distributed errors. This is known as *heteroscedasticity.*

- The case where Σ has the same number on the main diagonal but nonzero elements off the main diagonal gives identically (assuming higher moments are also the same) dependently distributed errors. This is known as *autocorrelation*.

- The general case combines heteroscedasticity and autocorrelation. This is known as "nonspherical" disturbances, though why this term is used, I have no idea.

Perhaps it's because under the classical assumptions, a joint confidence region for ε would be an $n-$ dimensional hypersphere.

1.1 Effects of nonspherical disturbances on the OLS estimator

The least square estimator is

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'y \\ &= \beta + (X'X)^{-1}X'\varepsilon\end{aligned}$$

- Conditional on X, or supposing that X is independent of ε, we have unbiasedness, as before.

- The variance of $\hat{\beta}$, supposing X is nonstochastic, is

$$\begin{aligned}\mathcal{E}\left[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'\right] &= \mathcal{E}\left[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\right] \\ &= (X'X)^{-1}X'\Sigma X(X'X)^{-1}\end{aligned}$$

  Due to this, any test statistic that is based upon $\widehat{\sigma^2}$ or the probability limit $\widehat{\sigma^2}$ of is invalid. In particular, the formulas for the t, F, χ^2 based tests given above do not lead to statistics with these distributions.

- $\hat{\beta}$ is still consistent, following exactly the same argument given before.

- If ε is normally distributed, then, conditional on X

$$\hat{\beta} \sim N\left(\beta, (X'X)^{-1}X'\Sigma X(X'X)^{-1}\right)$$

  The problem is that Σ is unknown in general, so this distribution won't be useful for testing hypotheses.

Summary: OLS with heteroscedasticity and/or autocorrelation is:

- unbiased in the same circumstances in which the estimator is unbiased with iid errors
- has a different variance than before, so the previous test statistics aren't valid
- is consistent
- is asymptotically normally distributed, but with a different limiting covariance matrix. Previous test statistics aren't valid in this case for this reason.
- is inefficient, as is shown below.

1.2 The GLS estimator

Suppose Σ were known. Then one could form the Cholesky decomposition

$$PP' = \Sigma^{-1}$$

We have

$$PP'\Sigma = I_n$$

so

$$P'\left(P\Sigma P'\right) = P',$$

which implies that

$$P'\Sigma P = I_n$$

Consider the model

$$P'y = P'X\beta + P'\varepsilon,$$

or, making the obvious definitions,

$$y^* = X^*\beta + \varepsilon^*.$$

This variance of $\varepsilon^* = P'\varepsilon$ is

$$\begin{aligned} \mathcal{E}\left(P'\varepsilon\varepsilon' P\right) &= P'\Sigma P \\ &= I_n \end{aligned}$$

Therefore, the model

$$\begin{aligned} y^* &= X^*\beta + \varepsilon^* \\ \mathcal{E}\left(\varepsilon^*\right) &= 0 \\ V(\varepsilon^*) &= I_n \\ \mathcal{E}\left(X^{*\prime}\varepsilon^*\right) &= 0 \end{aligned}$$

satisfies the classical assumptions (with modifications to allow stochastic regressors and nonnormality of ε). The GLS estimator is simply OLS applied to the transformed model:

$$\begin{aligned} \hat{\beta}_{GLS} &= (X^{*\prime}X^*)^{-1}X^{*\prime}y^* \\ &= (X'PP'X)^{-1}X'PP'y \\ &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y \end{aligned}$$

The GLS estimator is unbiased in the same circumstances under which the OLS estimator is unbiased. For example, assuming X is nonstochastic

$$\begin{aligned}
\mathcal{E}(\hat{\beta}_{GLS}) &= \mathcal{E}\left\{(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y\right\} \\
&= \mathcal{E}\left\{(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}(X\beta+\varepsilon\right\} \\
&= \beta.
\end{aligned}$$

The variance of the estimator, conditional on X can be calculated using

$$\begin{aligned}
\hat{\beta}_{GLS} &= (X^{*\prime}X^{*})^{-1}X^{*\prime}y^{*} \\
&= (X^{*\prime}X^{*})^{-1}X^{*\prime}\left(X^{*}\beta+\varepsilon^{*}\right) \\
&= \beta+(X^{*\prime}X^{*})^{-1}X^{*\prime}\varepsilon^{*}
\end{aligned}$$

so

$$\begin{aligned}
\mathcal{E}\left\{\left(\hat{\beta}_{GLS}-\beta\right)\left(\hat{\beta}_{GLS}-\beta\right)'\right\} &= \mathcal{E}\left\{(X^{*\prime}X^{*})^{-1}X^{*\prime}\varepsilon^{*}\varepsilon^{*\prime}X^{*}(X^{*\prime}X^{*})^{-1}\right\} \\
&= (X^{*\prime}X^{*})^{-1}X^{*\prime}X^{*}(X^{*\prime}X^{*})^{-1} \\
&= (X^{*\prime}X^{*})^{-1} \\
&= (X'\Sigma^{-1}X)^{-1}
\end{aligned}$$

Either of these last formulas can be used.

- All the previous results regarding the desirable properties of the least squares estimator hold, when dealing with the transformed model.
- Tests are valid, using the previous formulas, as long as we substitute X^{*} in place of X. Furthermore, any test that involves σ^2 can set it to 1. This is preferable to re-deriving the appropriate formulas.

- The GLS estimator is more efficient than the OLS estimator. This is a consequence of the Gauss-Markov theorem, since the GLS estimator is based on a model that satisfies the classical assumptions but the OLS estimator is not. To see this directly, not that

$$\begin{aligned} Var(\hat{\beta}) - Var(\hat{\beta}_{GLS}) &= (X'X)^{-1}X'\Sigma X(X'X)^{-1} - (X'\Sigma^{-1}X)^{-1} \\ &= \end{aligned}$$

- As one can verify by calculating fonc, the GLS estimator is the solution to the minimization problem

$$\hat{\beta}_{GLS} = \arg\min (y - X\beta)'\Sigma^{-1}(y - X\beta)$$

so the *metric* Σ^{-1} is used to weight the residuals.

1.3 Feasible GLS estimation

The problem is that Σ isn't known usually, so this estimator isn't available.

- Consider the dimension of Σ : it's an $n \times n$ matrix with $\left(n^2 - n\right)/2 + n = \left(n^2 + n\right)/2$ unique elements.

- The number of parameters to estimate is larger than n and increases faster than n. There's no way to devise an estimator that satisfies a LLN without adding restrictions.

- The *feasible GLS estimator* is based upon making sufficient assumptions regarding the form of Σ so that a consistent estimator can be devised.

Suppose that we *parameterize* Σ as a function of X and θ, where θ may include β as well as other parameters, so that

$$\Sigma = \Sigma(X, \theta)$$

where θ is of fixed dimension. If we can consistently estimate θ, we can consistently estimate Σ, as long as $\Sigma(X, \theta)$ is a continuous function of θ (by the Slutsky theorem). In this case,

$$\widehat{\Sigma} = \Sigma(X, \hat{\theta}) \xrightarrow{p} \Sigma(X, \theta)$$

If we replace Σ in the formulas for the GLS estimator with $\widehat{\Sigma}$, we obtain the FGLS estimator. **The FGLS estimator shares the same asymptotic properties as GLS. These are**

1. Consistency
2. Asymptotic normality
3. Asymptotic efficiency *if* the errors are normally distributed. (Cramer-Rao).
4. Test procedures are asymptotically valid.

In practice, the usual way to proceed is

1. Define a consistent estimator of θ. This is a case-by-case proposition, depending on the parameterization $\Sigma(\theta)$. We'll see examples below.
2. Form $\widehat{\Sigma} = \Sigma(X, \hat{\theta})$
3. Calculate the Cholesky factorization $\widehat{P} = Chol(\hat{\Sigma}^{-1})$.

4. Transform the model using

$$\hat{P}'y = \hat{P}'X\beta + \hat{P}'\varepsilon$$

5. Estimate using OLS on the transformed model.

1.4 Heteroscedasticity

Heteroscedasticity is the case where

$$\mathcal{E}(\varepsilon\varepsilon') = \Sigma$$

is a diagonal matrix, so that the errors are uncorrelated, but have different variances. Heteroscedasticity is usually thought of as associated with cross sectional data, though there is absolutely no reason why time series data cannot also be heteroscedastic (topic for a more advanced course).

Consider a supply function

$$q_i = \beta_1 + \beta_p P_i + \beta_s S_i + \varepsilon_i$$

where P_i is price and S_i is some measure of size of the i^{th} firm. One might suppose that unobservable factors (e.g., talent of managers, degree of coordination between production units, *etc.*) account for the error term ε_i. If there is more variability in these factors for large firms than for small firms, then ε_i may have a higher variance when S_i is high than when it is low.

Another example, individual demand.

$$q_i = \beta_1 + \beta_p P_i + \beta_m M_i + \varepsilon_i$$

where P is price and M is income. In this case, ε_i can reflect variations in preferences. There are more possibilities for expression of preferences when one is rich, so it is possible that the variance of ε_i could be higher when M is high.

Add example of group means.

1.4.1 Detection

There exist many tests for the presence of heteroscedasticity. We'll discuss three methods.

Goldfeld-Quandt The sample is divided in to three parts, with n_1, n_2 and n_3 observations, where $n_1 + n_2 + n_3 = n$. The model is estimated using the first and third parts of the sample, separately, so that $\hat{\beta}^1$ and $\hat{\beta}^3$ will be independent. Then we have

$$\frac{\hat{\varepsilon}^{1\prime}\hat{\varepsilon}^1}{\sigma^2} = \frac{\varepsilon^{1'}M^1\varepsilon^1}{\sigma^2} \xrightarrow{d} \chi^2(n_1 - K)$$

and

$$\frac{\hat{\varepsilon}^{3\prime}\hat{\varepsilon}^3}{\sigma^2} = \frac{\varepsilon^{3'}M^3\varepsilon^3}{\sigma^2} \xrightarrow{d} \chi^2(n_3 - K)$$

so

$$\frac{\hat{\varepsilon}^{1\prime}\hat{\varepsilon}^1/(n_1 - K)}{\hat{\varepsilon}^{3\prime}\hat{\varepsilon}^3/(n_3 - K)} \xrightarrow{d} F(n_1 - K, n_3 - K).$$

The distributional result is exact if the errors are normally distributed. This test is a two-tailed test. Alternatively, and probably more conventionally, if one has prior ideas about the possible magnitudes of the variances of the observations, one could order the observations accordingly, from largest to smallest. In this case, one would use a conventional one-tailed F-test. *Draw picture.*

- Ordering the observations is an important step if the test is to have any power.
- The motive for dropping the middle observations is to increase the difference between the average variance in the subsamples, supposing that there exists heteroscedasticity. This can increase the power of the test. On the other hand, dropping too many observations will substantially increase the variance of the statistics $\hat{\varepsilon}^{1\prime}\hat{\varepsilon}^{1}$ and $\hat{\varepsilon}^{3\prime}\hat{\varepsilon}^{3}$. A rule of thumb, based on Monte Carlo experiments is to drop around 25% of the observations.
- If one doesn't have any ideas about the form of the het. the test will probably have low power since a sensible data ordering isn't available.

White's test When one has little idea if there exists heteroscedasticity, and no idea of its potential form, the White test is a possibility. The idea is that if there is homoscedasticity, then

$$\mathcal{E}\left(\varepsilon_t^2\right) = \sigma^2, \forall t$$

so that x_t or functions of x_t shouldn't help to explain $\mathcal{E}\left(\varepsilon_t^2\right)$. The test works as follows:

1. Since ε_t isn't available, use the consistent estimator $\hat{\varepsilon}_t$ instead.
2. Regress

$$\hat{\varepsilon}_t^2 = z'\gamma + v_t$$

   where z_t is a P -vector. z_t may include some or all of the variables in x_t, as well as other variables. White's original suggestion was the set of all unique squares and cross products of variables in x_t.
3. Test the hypothesis that $\gamma = 0$. Note that this is the R^2 or the artificial regression used to test for heteroscedasticity, not the R^2 of the original model.

An asymptotically equivalent statistic, under the null of no heteroscedasticity (so that R^2 should tend to zero), is

$$nR^2 \overset{a}{\sim} \chi^2(P-1).$$

This doesn't require normality of the errors, though it does assume that the fourth moment of ε_t is constant, under the null. **Question**: why is this necessary?

- The White test has the disadvantage that it may not be very powerful unless the z_t vector is chosen well, and this is hard to do without knowledge of the form of heteroscedasticity.

- It also has the problem that specification errors other than heteroscedasticity may lead to rejection.

- Note: the null hypothesis of this test may be interpreted as $\theta = 0$ for the variance model $V(\varepsilon_t^2) = h(\alpha + z_t'\theta)$, where $h(\cdot)$ is an arbitrary function of unknown form. The test is more general than is may appear from the regression that is used.

Plotting the residuals A very simple method is to simply plot the residuals (or their squares). *Draw pictures here*. Like the Goldfeld-Quandt test, this will be more informative if the observations are ordered according to the suspected form of the heteroscedasticity.

1.4.2 Correction

Correcting for heteroscedasticity requires that a parametric form for $\Sigma(\theta)$ be supplied, and that a means for estimating θ consistently be determined. The estimation method will be specific to the for supplied for $\Sigma(\theta)$. We'll consider two examples. Before this, let's consider the general nature of GLS when there is heteroscedasticity.

Multiplicative heteroscedasticity Suppose the model is

$$\begin{aligned} y_t &= x_t'\beta + \varepsilon_t \\ \sigma_t^2 &= \mathcal{E}\left(\varepsilon_t^2\right) = \left(z_t'\gamma\right)^{\delta} \end{aligned}$$

but the other classical assumptions hold. In this case

$$\varepsilon_t^2 = \left(z_t'\gamma\right)^{\delta} + v_t$$

and v_t has mean zero. Nonlinear least squares could be used to estimate γ and δ consistently, were ε_t observable. The solution is to substitute the squared OLS residuals $\hat{\varepsilon}_t^2$ in place of ε_t^2, since it is consistent by the Slutsky theorem. Once we have $\hat{\gamma}$ and $\hat{\delta}$, we can estimate σ_t^2 consistently using

$$\hat{\sigma}_t^2 = \left(z_t'\hat{\gamma}\right)^{\hat{\delta}} \overset{p}{\to} \sigma_t^2 \,.$$

In the second step, we transform the model by dividing by the standard deviation:

$$\frac{y_t}{\hat{\sigma}_t} = \frac{x_t'\beta}{\hat{\sigma}_t} + \frac{\varepsilon_t}{\hat{\sigma}_t}$$

or

$$y_t^* = x_t^{*\prime}\beta + \varepsilon_t^*.$$

Asymptotically, this model satisfies the classical assumptions.

- This model is a bit complex in that NLS is required to estimate the model of the variance. A simpler version would be

$$\begin{aligned} y_t &= x_t'\beta + \varepsilon_t \\ \sigma_t^2 &= \mathcal{E}(\varepsilon_t^2) = \sigma^2 z_t^\delta \end{aligned}$$

where z_t is a single variable. There are still two parameters to be estimated, and the model of the variance is still nonlinear in the parameters. However, the *search method* can be used in this case to reduce the estimation problem to repeated applications of OLS.

- First, we define an interval of reasonable values for δ, e.g., $\delta \in [0, 3]$.
- Partition this interval into M equally spaced values, e.g., $\{0, .1, .2, ..., 2.9, 3\}$.
- For each of these values, calculate the variable $z_t^{\delta_m}$.
- The regression

  $$\hat{\varepsilon}_t^2 = \sigma^2 z_t^{\delta_m} + v_t$$

  is linear in the parameters, conditional on δ_m, so one can estimate σ^2 by OLS.
- Save the pairs (σ_m^2, δ_m), and the corresponding ESS_m. Choose the pair with the minimum ESS_m as the estimate.
- Next, divide the model by the estimated standard deviations.
- Can refine. *Draw picture.*
- Works well when the parameter to be searched over is low dimensional, as in this case.

Groupwise heteroscedasticity A common case is where we have repeated observations on each of a number of economic agents: e.g., 10 years of macroeconomic data on each of a set of countries or regions, or daily observations of transactions of 200 banks. This sort of data is a *pooled cross-section time-series model.* It may be reasonable to presume that the variance is constant over time within the cross-sectional units, but that it differs across them (e.g., firms or countries of different sizes...). The model is

$$\begin{aligned} y_{it} &= x_{it}'\beta + \varepsilon_{it} \\ \mathcal{E}\left(\varepsilon_{it}^2\right) &= \sigma_i^2, \forall t \end{aligned}$$

where $i = 1, 2, ..., G$ are the agents, and $t = 1, 2, ..., n$ are the observations on each agent.

- The other classical assumptions are presumed to hold.
- In this case, the variance σ_i^2 is specific to each agent, but constant over the n observations for that agent.
- In this model, we assume that $\mathcal{E}\left(\varepsilon_{it}\varepsilon_{is}\right) = 0$. This is a strong assumption that we'll relax later.

To correct for heteroscedasticity, just estimate each σ_i^2 using the natural estimator:

$$\hat{\sigma}_i^2 = \frac{1}{n}\sum_{t=1}^{n}\hat{\varepsilon}_{it}^2$$

- Note that we use $1/n$ here since it's possible that there are more than n regressors, so $n - K$ could be negative. Asymptotically the difference is unimportant.

- With each of these, transform the model as usual:

$$\frac{y_{it}}{\hat{\sigma}_i} = \frac{x'_{it}\beta}{\hat{\sigma}_i} + \frac{\varepsilon_{it}}{\hat{\sigma}_i}$$

  Do this for each cross-sectional group. This transformed model satisfies the classical assumptions, asymptotically.

1.5 Autocorrelation

Autocorrelation, which is the serial correlation of the error term, is a problem that is usually associated with time series data, but also can affect cross-sectional data. For example, a shock to oil prices will simultaneously affect all countries, so one could expect contemporaneous correlation of macroeconomic variables across countries.

1.5.1 Causes

Autocorrelation is the existence of correlation across the error term:

$$\mathcal{E}(\varepsilon_t\varepsilon_s) \neq 0, t \neq s.$$

Why might this occur? Plausible explanations include

1. Lags in adjustment to shocks. In a model such as

$$y_t = x'_t\beta + \varepsilon_t,$$

   one could interpret $x'_t\beta$ as the equilibrium value. Suppose x_t is constant over a number of observations. One can interpret ε_t as a shock that moves the system away from equilibrium. If the time needed to return to equilibrium is long with respect to the observation frequency, one could expect ε_{t+1} to be positive, conditional on ε_t positive, which induces a correlation.

2. Unobserved factors that are correlated over time. The error term is often assumed to correspond to unobservable factors. If these factors are correlated, there will be autocorrelation.

3. Misspecification of the model. Suppose that the DGP is

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + \varepsilon_t$$

but we estimate

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

Draw a picture here.

1.5.2 AR(1)

There are many types of autocorrelation. We'll consider two examples. The first is the most commonly encountered case: autoregressive order 1 (AR(1) errors. The model is

$$\begin{aligned} y_t &= x_t'\beta + \varepsilon_t \\ \varepsilon_t &= \rho\varepsilon_{t-1} + u_t \\ u_t &\sim iid(0, \sigma_u^2) \\ \mathcal{E}(\varepsilon_t u_s) &= 0, t < s \end{aligned}$$

We assume that the model satisfies the other classical assumptions.

- We need a stationarity assumption: $|\rho| < 1$. Otherwise the variance of ε_t explodes as t increases, so standard asymptotics will not apply.

- By recursive substitution we obtain

$$\begin{aligned}
\varepsilon_t &= \rho\varepsilon_{t-1} + u_t \\
&= \rho\left(\rho\varepsilon_{t-2} + u_{t-1}\right) + u_t \\
&= \rho^2\varepsilon_{t-2} + \rho u_{t-1} + u_t \\
&= \rho^2\left(\rho\varepsilon_{t-3} + u_{t-2}\right) + \rho u_{t-1} + u_t
\end{aligned}$$

In the limit the lagged ε drops out, since $\rho^m \to 0$ as $m \to \infty$, so we obtain

$$\varepsilon_t = \sum_{m=0}^{\infty} \rho^m u_{t-m}$$

With this, the variance of ε_t is found as

$$\begin{aligned}
\mathcal{E}\left(\varepsilon_t^2\right) &= \sigma_u^2 \textstyle\sum_{m=0}^{\infty} \rho^{2m} \\
&= \frac{\sigma_u^2}{1-\rho^2}
\end{aligned}$$

- If we had directly assumed that ε_t were covariance stationary, we could obtain this using

$$\begin{aligned}
V(\varepsilon_t) &= \rho^2\mathcal{E}\left(\varepsilon_{t-1}^2\right) + 2\rho\mathcal{E}\left(\varepsilon_{t-1}u_t\right) + \mathcal{E}\left(u_t^2\right) \\
&= \rho^2 V(\varepsilon_t) + \sigma_u^2,
\end{aligned}$$

so

$$V(\varepsilon_t) = \frac{\sigma_u^2}{1-\rho^2}$$

- The variance is the 0^{th} order autocovariance: $\gamma_0 = V(\varepsilon_t)$

- Note that the variance does not depend on t

Likewise, the first order autocovariance γ_1 is

$$\begin{aligned} Cov(\varepsilon_t, \varepsilon_{t-1}) &= \gamma_s = \mathcal{E}\left((\rho\varepsilon_{t-1} + u_t)\,\varepsilon_{t-1}\right) \\ &= \rho V(\varepsilon_t) \\ &= \tfrac{\rho\sigma_u^2}{1-\rho^2} \end{aligned}$$

- Using the same method, we find that for $s < t$

$$Cov(\varepsilon_t, \varepsilon_{t-s}) = \gamma_s = \frac{\rho^s \sigma_u^2}{1-\rho^2}$$

- The autocovariances don't depend on t: the process $\{\varepsilon_t\}$ is *covariance stationary*

The *correlation (*in general, for r.v.'s x and y) is defined as

$$\mathrm{corr}(x,y) = \frac{\mathrm{cov}(x,y)}{\mathrm{se}(x)\mathrm{se}(y)}$$

but in this case, the two standard errors are the same, so the s-order autocorrelation ρ_s is

$$\rho_s = \rho^s$$

- All this means that the overall matrix Σ has the form

$$\Sigma = \underbrace{\frac{\sigma_u^2}{1-\rho^2}}_{\text{this is the variance}} \underbrace{\begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \vdots & & \ddots & & \vdots \\ & & & \ddots & \rho \\ \rho^{n-1} & \cdots & & & 1 \end{bmatrix}}_{\text{this is the correlation matrix}}$$

So we have homoscedasticity, but elements off the main diagonal are not zero. All of this depends only on two parameters, ρ and σ_u^2. If we can estimate these consistently, we can apply FGLS.

It turns out that it's easy to estimate these consistently. The steps are

1. Estimate the model $y_t = x_t'\beta + \varepsilon_t$ by OLS. This is consistent as long as $\frac{1}{n}X'\Sigma X$ converges to a finite limiting matrix. It turns out that this requires that the regressors X satisfy the previous stationarity conditions and that $|\rho| < 1$, which we have assumed.

2. Take the residuals, and estimate the model

$$\hat{\varepsilon}_t = \rho\hat{\varepsilon}_{t-1} + u_t^*$$

Since $\hat{\varepsilon}_t \xrightarrow{p} \varepsilon_t$, this regression is asymptotically equivalent to the regression

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

which satisfies the classical assumptions. Therefore, $\hat{\rho}$ obtained by applying OLS to $\hat{\varepsilon}_t = \rho\hat{\varepsilon}_{t-1} + u_t^*$ is consistent. Also, since $u_t^* \xrightarrow{p} u_t$, the estimator

$$\hat{\sigma}_u^2 = \frac{1}{n}\sum_{t=2}^{n}(\hat{u}_t^*)^2 \xrightarrow{p} \sigma_u^2$$

3. With the consistent estimators $\hat{\sigma}_u^2$ and $\hat{\rho}$, form $\hat{\Sigma} = \Sigma(\hat{\sigma}_u^2, \hat{\rho})$ using the previous structure of Σ, and estimate by FGLS. Actually, one can omit the factor $\hat{\sigma}_u^2/(1-\rho^2)$, since it cancels out in the formula

$$\hat{\beta}_{FGLS} = \left(X'\hat{\Sigma}^{-1}X\right)^{-1}(X'\hat{\Sigma}^{-1}y).$$

- One can iterate the process, by taking the first FGLS estimator of β, re-estimating ρ and σ_u^2, etc. If one iterates to convergences it's equivalent to MLE (supposing normal errors).

- An asymptotically equivalent approach is to simply estimate the transformed model

$$y_t - \hat{\rho} y_{t-1} = (x_t - \hat{\rho} x_{t-1})' \beta + u_t^*$$

  using $n-1$ observations (since y_0 and x_0 aren't available). This is the method of Cochrane and Orcutt. Dropping the first observation is asymptotically irrelevant, but *it can be very important in small samples.* One can recuperate the first observation by putting

$$\begin{aligned} y_1^* &= \sqrt{1-\hat{\rho}^2} y_1 \\ x_1^* &= \sqrt{1-\hat{\rho}^2} x_1 \end{aligned}$$

  This somewhat odd result is related to the Cholesky factorization of Σ^{-1}. See Davidson and MacKinnon, pg. 348-49 for more discussion. Note that the variance of y_1^* is σ_u^2, asymptotically, so we see that the transformed model will be homoscedastic (and nonautocorrelated, since the $u's$ are uncorrelated with the $y's$, in different time periods.

1.5.3 MA(1)

The linear regression model with moving average order 1 errors is

$$\begin{aligned} y_t &= x_t'\beta + \varepsilon_t \\ \varepsilon_t &= u_t + \phi u_{t-1} \\ u_t &\sim iid(0, \sigma_u^2) \\ \mathcal{E}(\varepsilon_t u_s) &= 0, t < s \end{aligned}$$

In this case,

$$\begin{aligned} V(\varepsilon_t) &= \gamma_0 = \mathcal{E}\left[(u_t + \phi u_{t-1})^2\right] \\ &= \sigma_u^2 + \phi^2 \sigma_u^2 \\ &= \sigma_u^2 (1 + \phi^2) \end{aligned}$$

Similarly

$$\begin{aligned} \gamma_1 &= \mathcal{E}\left[(u_t + \phi u_{t-1})(u_{t-1} + \phi u_{t-2})\right] \\ &= \phi \sigma_u^2 \end{aligned}$$

and

$$\begin{aligned} \gamma_2 &= \left[(u_t + \phi u_{t-1})(u_{t-2} + \phi u_{t-3})\right] \\ &= 0 \end{aligned}$$

so in this case

$$\Sigma = \sigma_u^2 \begin{bmatrix} 1+\phi^2 & \phi & 0 & \cdots & 0 \\ \phi & 1+\phi^2 & \phi & & \\ 0 & \phi & \ddots & & \vdots \\ \vdots & & & \ddots & \phi \\ 0 & \cdots & & \phi & 1+\phi^2 \end{bmatrix}$$

Note that the first order autocorrelation is

$$\begin{aligned} \rho_1 &= \tfrac{\phi\sigma_u^2}{\sigma_u^2(1+\phi^2)} = \frac{\gamma_1}{\gamma_0} \\ &= \tfrac{\phi}{(1+\phi^2)} \end{aligned}$$

- This achieves a maximum at $\phi = 1$ and a minimum at $\phi = -1$, and the maximal and minimal autocorrelations are 1/2 and -1/2. Therefore, series that are more strongly autocorrelated can't be MA(1) processes.

Again the covariance matrix has a simple structure that depends on only two parameters. The problem in this case is that one can't estimate ϕ using OLS on

$$\hat{\varepsilon}_t = u_t + \phi u_{t-1}$$

because the u_t are unobservable and they can't be estimated consistently. However, there is a simple way to estimate the parameters.

- Since the model is homoscedastic, we can estimate

$$V(\varepsilon_t) = \sigma_\varepsilon^2 = \sigma_u^2(1+\phi^2)$$

using the typical estimator:

$$\widehat{\sigma_{\varepsilon}^2} = \sigma_u^2(\widehat{1+\phi^2}) = \frac{1}{n}\sum_{t=1}^{n}\hat{\varepsilon}_t^2$$

- By the Slutsky theorem, we can interpret this as defining an (unidentified) estimator of both σ_u^2 and ϕ, e.g., use this as

$$\widehat{\sigma_u^2}(1+\widehat{\phi}^2) = \frac{1}{n}\sum_{t=1}^{n}\hat{\varepsilon}_t^2$$

  However, this isn't sufficient to define consistent estimators of the parameters, since it's unidentified.

- To solve this problem, estimate the covariance of ε_t and ε_{t-1} using

$$\widehat{Cov}(\varepsilon_t, \varepsilon_{t-1}) = \widehat{\phi\sigma_u^2} = \frac{1}{n}\sum_{t=2}^{n}\hat{\varepsilon}_t\hat{\varepsilon}_{t-1}$$

  This is a consistent estimator, following a LLN (and given that the epsilon hats are consistent for the epsilons). As above, this can be interpreted as defining an unidentified estimator:

$$\hat{\phi}\widehat{\sigma_u^2} = \frac{1}{n}\sum_{t=2}^{n}\hat{\varepsilon}_t\hat{\varepsilon}_{t-1}$$

- Now solve these two equations to obtain identified (and therefore consistent) estimators of both ϕ and σ_u^2. Define the consistent estimator

$$\hat{\Sigma} = \Sigma(\hat{\phi}, \widehat{\sigma_u^2})$$

  following the form we've seen above, and transform the model using the Cholesky decomposition. The transformed model satisfies the classical assumptions asymptotically.

1.5.4 Testing for autocorrelation

Durbin-Watson test The Durbin-Watson test statistic is

$$\begin{aligned} DW &= \frac{\sum_{t=2}^{n}(\hat{\varepsilon}_t-\hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{n}\hat{\varepsilon}_t^2} \\ &= \frac{\sum_{t=2}^{n}\left(\hat{\varepsilon}_t^2-2\hat{\varepsilon}_t\hat{\varepsilon}_{t-1}+\hat{\varepsilon}_{t-1}^2\right)}{\sum_{t=1}^{n}\hat{\varepsilon}_t^2} \end{aligned}$$

- The null hypothesis is that the first order autocorrelation of the errors is zero: $H_0 : \rho_1 = 0$. The alternative is of course $H_A : \rho_1 \neq 0$. Note that the alternative is not that the errors are AR(1), since many general patterns of autocorrelation will have the first order autocorrelation different than zero. For this reason the test is useful for detecting autocorrelation in general. For the same reason, one shouldn't just assume that an AR(1) model is appropriate when the DW test rejects the null.

- Under the null, the middle term tends to zero, and the other two tend to one, so $DW \xrightarrow{p} 2$.

- .Supposing that we had an AR(1) error process with $\rho = 1$. In this case the middle term tends to -2, so $DW \xrightarrow{p} 0$

- Supposing that we had an AR(1) error process with $\rho = -1$. In this case the middle term tends to 2, so $DW \xrightarrow{p} 4$

- These are the extremes: DW always lies between 0 and 4.

- The distribution depends on the matrix of regressors, X, so tables can't give exact critical values. The give upper and lower bounds, which correspond to the extremes that are possible. *Picture here.* There are means of determining exact critical values conditional on X.

- Note that DW can be used to test for nonlinearity (add discussion).

Breusch-Godfrey test This test uses an auxiliary regression, as does the White test for heteroscedasticity. The regression is

$$\hat{\varepsilon}_t = x_t'\delta + \gamma_1\hat{\varepsilon}_{t-1} + \gamma_2\hat{\varepsilon}_{t-2} + \cdots + \gamma_P\hat{\varepsilon}_{t-P} + v_t$$

and the test statistic is the nR^2 statistic, just as in the White test. There are P restrictions, so the test statistic is asymptotically distributed as a $\chi^2(P)$.

- The intuition is that the lagged errors shouldn't contribute to explaining the current error if there is no autocorrelation.
- x_t is included as a regressor to account for the fact that the $\hat{\varepsilon}_t$ are not independent even if the ε_t are. This is a technicality that we won't go into here.
- The alternative is not that the model is an AR(P), following the argument above. The alternative is simply that some or all of the first P autocorrelations are different from zero. This is compatible with many specific forms of autocorrelation.

1.5.5 Lagged dependent variables and autocorrelation

We've seen that the OLS estimator is consistent under autocorrelation, as long as $plim\frac{X'\varepsilon}{n} = 0$. This will be the case when $\mathcal{E}(X'\varepsilon) = 0$, following a LLN. An important exception is the case where X contains lagged $y's$ and the errors are autocorrelated. A simple example is the case of a single lag of the dependent variable with AR(1) errors.

The model is

$$
\begin{aligned}
y_t &= x_t'\beta + y_{t-1}\gamma + \varepsilon_t \\
\varepsilon_t &= \rho\varepsilon_{t-1} + u_t
\end{aligned}
$$

Now we can write

$$
\begin{aligned}
\mathcal{E}\left(y_{t-1}\varepsilon_t\right) &= \mathcal{E}\left\{\left(x_{t-1}'\beta + y_{t-2}\gamma + \varepsilon_{t-1}\right)\left(\rho\varepsilon_{t-1} + u_t\right)\right\} \\
&\neq 0
\end{aligned}
$$

since one of the terms is $\mathcal{E}\left(\rho\varepsilon_{t-1}^2\right)$ which is clearly nonzero. In this case $\mathcal{E}\left(X'\varepsilon\right) \neq 0$, and therefore $plim\frac{X'\varepsilon}{n} \neq 0$. Since

$$
plim\hat{\beta} = \beta + plim\frac{X'\varepsilon}{n}
$$

the OLS estimator is inconsistent in this case. One needs to estimate by instrumental variables (IV), which we'll get to later