20: GENERALIZED LEAST SQUARES

So far, we have assumed that the u_i are *iid*, so that

$$Cov(u) = \sigma_u^2 I \quad . \tag{1}$$

According to the **Gauss-Markov Theorem,** if (1) holds, then the least squares estimator b is **best linear unbiased.** This means that b is a linear combination of y with $E[b] = \beta$, and that b is at least as good as any other linear unbiased estimator c in the sense that

$$Var(\alpha' b) \leq Var(\alpha' c)$$

for any nonrandom $(p+1)\times 1$ vector α.

In practice, (1) may not be a reasonable assumption. For example, we may have *heteroscedasticity*,

i.e., $Var(u_i)$ may not be constant. Furthermore, if the subscript in y_i represents time, there may be *autocorrelation*, i.e., a correlation between y_i and y_j for $i \neq j$. In general, we will have

$$Cov(u) = \sigma_u^2 \Omega \ , \qquad (2)$$

where Ω is an $n \times n$ symmetric matrix whose diagonal entries may not be the same, and whose off-diagonal entries may not be zero. For the moment, we assume that Ω is known. Using the theory of linear algebra, we can construct an invertible $n \times n$ matrix P such that $P'P = \Omega^{-1}$.

If $Cov(u) = \sigma_u^2 \Omega$ with $\Omega \neq I$, the estimator b is still linear and unbiased, but it is no longer *best* linear unbiased. Define the **generalized least squares estimator** by

$$\tilde{b} = (X^{*\prime} X^{*})^{-1} X^{*\prime} y^{*} \quad ,$$

where $X^{*} = PX$, and $y^{*} = Py$.

The model $y = X\beta + u$ can be re-expressed as

$$y^{*} = X^{*}\beta + u^{*} \quad , \qquad (3)$$

where $u^{*} = Pu$.

The error term u^{*} in (3) satisfies

$$Cov(u^{*}) = E[u^{*} u^{*\prime}] = E[P\ uu'P']$$

$$= \sigma_u^2 P \Omega P' = \sigma_u^2 P P^{-1} (P')^{-1} P' = \sigma_u^2 I \quad .$$

Since $\tilde{b}$ is the least squares estimator of β in the model (3) and since the error term u^* in (3) satisfies $Cov(u^*) = \sigma_u^2 I$, it follows that $\tilde{b}$ is best linear unbiased.

The covariance matrix for $\tilde{b}$ is given by

$$Cov(\tilde{b}) = \sigma_u^2 (X^{*\prime} X^*)^{-1} \ ,$$

which can be estimated without bias by $s^2(X^{*\prime} X^*)^{-1}$, where $s^2 = \frac{1}{n-p-1} ||y^* - X^* \tilde{b}||^2$ is an unbiased estimator of σ_u^2. If the u_i are normal then $s^2 \sim \frac{1}{n-p-1} \sigma_u^2 \chi^2_{n-p-1}$, independently of $\tilde{b}$.

If we use the ordinary least squares estimator b when $\Omega \neq I$, then our inferences about β will be incorrect, since the covariance matrix for b will *not* be equal to the presumed $\sigma_u^2(X'X)^{-1}$. In other words, tests and confidence intervals based on b which assume that $Cov(b) = \sigma_u^2(X'X)^{-1}$ will not have the nominal (i.e., intended) level. This drawback, together with the sub-optimality of b when $\Omega \neq I$, provides strong motivation for using the generalized least squares estimator $\tilde{b}$ instead of b whenever possible. Unfortunately, Ω will usually be unknown, and an estimator of β based on an *estimate* of Ω may not be optimal.

Autocorrelation

Suppose our observations are a *time series*, y_t, $t = 1, \ldots, T$, collected at equally spaced intervals of time, e.g., annually, quarterly, monthly, weekly, or daily. We can write the linear model as

$$y_t = \beta_0 + \sum_{j=1}^{p} x_{tj} \beta_j + u_t \quad ,$$

where the errors u_t may be autocorrelated. We should consider the possibility of autocorrelated errors whenever the observations are a time series. Since the exact autocorrelation structure, i.e., Ω, will not be known, and since the specification of $Corr(u_i, u_j)$ for $i < j$ would require $T(T-1)/2$ coefficients, it is wise to postulate a model for the

error autocorrelations which has a small number of parameters, and then try to estimate these parameters from the data. This yields an estimate of Ω, which can then be used to obtain the ("feasible" version of the) generalized least squares estimate of β.

A useful model for error autocorrelation is the *first order autoregression* ($AR(1)$) model,

$$u_t = \rho\, u_{t-1} + \varepsilon_t \quad , \tag{4}$$

where the parameter ρ is assumed to be between -1 and 1, $E[\varepsilon_t] = 0$, $Var[\varepsilon_t] = \sigma_\varepsilon^2$, and $E[\varepsilon_t \varepsilon_s] = 0$, $t \neq s$. The term "autoregression" is used because, from (4), u_t satisfies a regression model in terms of

an "explanatory variable", u_{t-1}, which is just a time-lagged version of u_t itself. Under this model, it can be shown that $E[u_t] = 0$, $Var[u_t] = \sigma_\varepsilon^2/(1-\rho^2)$ for all t, and $Corr(u_t, u_{t+k}) = \rho^k$ for any integer $k > 0$. Thus, if $\rho > 0$, then y_t and y_{t+k} will be positively correlated. This means that if the observation y_t happens to be above average (i.e., larger than its mean value, as given by the model), then y_{t+1}, y_{t+2}, etc., will tend to be above average as well. The "memory" that y_t was above average will decline at the exponential rate, ρ^k, which can be quite slow if ρ is close to 1.

Whenever the data are a time series, it is wise to examine a plot of the least squares residuals versus time. If $\rho > 0$, this time series plot will tend to be smoother than in the uncorrelated errors case, $\rho = 0$. The plot may also show a cyclical pattern, perhaps indicating that we need to include a seasonal term in the model.

In practice, even if it is assumed that the errors are $AR(1)$, the parameter ρ will not be known. We therefore start by computing the ordinary least squares estimator, b. We then estimate ρ, using the residuals e_t from the least squares fit, by

$$\hat{\rho} = \frac{\sum_{t=2}^{T} e_t e_{t-1}}{\sum_{t=1}^{T} e_t^2} \quad .$$

This is (essentially) the Pearson correlation coefficient between the vectors $(e_2, \ldots, e_T)'$ and $(e_1, \ldots, e_{T-1})'$, both of which are assumed to have zero expectation. It can be shown that $\hat{\rho}$ is a consistent estimator of ρ, although $\hat{\rho}$ may be biased if T is small. We then construct the matrix $\hat{\Omega}$, given by $\hat{\Omega}_{ij} = \hat{\rho}^{|i-j|}$, and then find $\hat{P}$ such that $\hat{P}'\hat{P} = \hat{\Omega}^{-1}$. An explicit formula for $\hat{P}$ in the $AR(1)$ case is given in Jobson, p. 381. Finally, we obtain the approximate generalized least squares estimator $[(\hat{P}X)'(\hat{P}X)]^{-1}(\hat{P}X)'(\hat{P}y)$, which is called the *Prais–Winsten* estimator.

The Durbin-Watson Test

A popular method of testing for error autocorrelation, using the residuals from ordinary least squares, is to use the *Durbin-Watson Statistic,*

$$d = \frac{\sum_{t=2}^{T} (e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} \quad ,$$

which is approximately equal to $2(1-\hat{\rho})$. A value of d close to 2 indicates that there is no first-order autocorrelation.

Suppose we want to test $H_0 : \rho = 0$ versus $H_1 : \rho > 0$. We will be inclined to reject H_0 in favor of H_1 if d is sufficiently *small*. The standard

implementation of the test is somewhat complicated, due to computational problems. *Assuming that the* u_t *are normal*, we can, in principle, use numerical methods to determine a critical value d^* such that $Pr\{d < d^* \mid H_0\} = \alpha$.

- The ("pure" form of the) Durbin-Watson test rejects H_0 if $d < d^*$, and does not reject H_0 if $d \geq d^*$. Unfortunately, the distribution of d under H_0 depends on X, so it is not feasible to make a table of d^* values which would be universally applicable.

To avoid the computational burden of calculating d^* for each new data set, Durbin and Watson

developed a modified version of the test. Specifically, they found quantities d_1 and d_2 whose distributions under H_0 do not depend on X, and such that $d_1 < d < d_2$. If d_L and d_U are defined by

$$Pr\{d_1 < d_L \mid H_0\} = \alpha \ , \ Pr\{d_2 < d_U \mid H_0\} = \alpha \ ,$$

then it follows that $d_L < d^* < d_U$. Tables of d_L and d_U for different values of T and p are given in many textbooks.

To perform the (modified) DW test, we first calculate d, and then proceed as follows:

- If $d < d_L$, reject H_0.
- If $d > d_U$, do not reject H_0.
- If $d_L \leq d \leq d_U$, the test is said to be *inconclusive*.

To justify this test, note that if $d < d_L$, we can be sure that $d < d^*$, even without computing d^*. Therefore, we should reject H_0 in this case. If $d > d_U$, we can be sure that $d > d^*$, so we should not reject H_0. If $d_L \le d \le d_U$, then we cannot be sure (without calculating d^*) whether $d < d^*$ or $d \ge d^*$, and therefore we regard the results as inconclusive. It is clear that the resulting test has a level which is *less than or equal to* α, although we will not always reach a conclusion.

Here is a way around the problem: perform the test as before, but if $d_L \le d \le d_U$, go to the trouble of computing d^*; Reject H_0 if $d < d^*$, do not reject

if $d \geq d^*$. This amounts to performing the original ("pure") DW test, but often saves us the trouble of computing d^*. Another advantage of doing the test this way is that it will actually have the nominal level, α.

The main drawback of the DW test is that it can have low power if the actual time series model describing the errors is something other than $AR(1)$. Suppose, for example, that $Corr(u_t, u_{t-1}) = 0$ but $Corr(u_t, u_{t-2})$ is positive. Then the errors are not uncorrelated, but the DW test will have virtually no ability to detect this.