EC771: Econometrics, Spring 2004

Greene, Econometric Analysis (5th ed, 2003)

Chapters 10, 11, 12:
Generalized Least Squares, Heteroskedasticity, Serial Correlation

The generalized linear regression model

The generalized linear regression model may be stated as:

$$\begin{aligned} y &= X\beta + \epsilon \\ E[\epsilon|X] &= 0 \\ E[\epsilon\epsilon'|X] &= \sigma^2\Omega = \Sigma \end{aligned}$$

where Ω is a positive definite matrix. This allows us to consider data generating processes where the assumption that $\Omega = I$ does not hold. Two special cases are of interest: pure

heteroskedasticity, where Ω is a diagonal matrix, and some form of serial correlation, in which

$$\Omega = \begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \cdots & \rho_{n-2} \\ & & \vdots & \\ \rho_{n-1} & \rho_{n-2} & \cdots & 1 \end{pmatrix}$$

where the ρ parameters represent the correlations between successive elements of the error process. In an ordinary cross–sectional or time–series data set, we might expect to encounter one of these violations of the classical assumptions on $E[\epsilon\epsilon'|X]$. In a pooled cross–section time–series data set, or the special case of that data structure known as panel (longitudinal) data, we might expect to encounter both problems.

We consider first the damage done to the OLS estimator by this violation of classical assumptions, and an approach that could be used

to repair that damage. Since that approach will often be burdensome, we consider an alternative strategy: the robustification of least squares estimates to deal with a Σ of unknown form.

OLS and IV in the GLM context

In estimating the linear regression model under the full set of classical assumptions, we found that OLS estimates are best linear unbiased (BLUE), consistent and asymptotically normally distributed (CAN), and under the assumption of normally distributed errors, asymptotically efficient. Which of these desirable properties hold up if $\Omega \neq I$?

Least squares will retain some of its desirable properties in the generalized linear regression model: it will still be unbiased, consistent, and asymptotically normally distributed. However,

it will no longer be efficient, and the usual inference procedures are no longer appropriate, as the interval estimates are inconsistent.

The least squares estimator, given $X \perp \epsilon$, will be unbiased, with sampling variance (conditioned on X of:

$$\begin{aligned} Var[b|X] &= E[(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}] \\ &= \sigma^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1} \end{aligned}$$

The inconsistency of the least squares interval estimates arises here: this expression for the sampling variance of b does not equal that applicable for OLS (which is only the first term of this expression). Not only is the wrong matrix being used, but there is no guarantee that an estimate of σ^2 will be unbiased. Generally we cannot state any relation between the respective elements of the two covariance matrices; the OLS standard errors may be larger or smaller than those computed from the generalized linear regression model.

Robust estimation of asymptotic covariance matrices

If we know Ω (up to a scalar), then as we will see an estimator may be defined to make use of that information and circumvent the difficulties of OLS. In many cases, even though we must generate an estimate of $\hat{\Omega}$, use of that estimate will be preferable to ignoring the issue and using OLS. But in many cases we may not be well informed about the nature of Ω, and deriving an estimator for the asymptotic covariance matrix of b may be the best way to proceed.

If Ω was known, the appropriate estimator of that asymptotic covariance matrix would be

$$V[b] = \frac{1}{n}\left[\frac{1}{n}X'X\right]^{-1}\left[\frac{1}{n}X'(\sigma^2\Omega)X\right]\left[\frac{1}{n}X'X\right]^{-1}$$

in which the only unknown element is $\sigma^2\Omega = \Sigma$ (Ω is only known up to a scalar multiple).

It might seem that to estimate $\frac{1}{n}X'\Sigma X$, an object containing $n(n+1)/2$ unknown parameters, might be a hopeless task using a sample of size n. But what is needed is an estimator of

$$\text{plim } Q = \text{plim } \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{ij}x_i x_j'$$

where the matrix Q, which is a matrix of sums of squares and cross products, has $K(K+1)/2$ unknown elements. Thus, the approach to estimation of the asymptotic covariance matrix will be to work with X and e, the least squares residuals, which are consistent estimators of their population counterparts given the consistency of b, from which they are computed.

Consider the case of pure heteroskedasticity, where we allow $E\epsilon_i\epsilon_i' = \sigma_i^2$. That assumption involves n unknown variances which cannot be estimated from samples of size 1. But in this

case the formula for Q, given that Σ is a diagonal matrix, simplifies to:

$$Q = \frac{1}{n}\sum_{i=1}^{n} \sigma_i x_i x_i'$$

White (1980) shows that under very general conditions the feasible estimator

$$S_0 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 x_i x_i'$$

has a plim equal to that of Q. Note that Q is a weighted sum of the outer products of the rows of X. We seek not to estimate Q, but to find a function of the sample data that will approximate Q arbitrarily well in the limit. If Q converges to a finite, positive definite matrix, we seek a function of the sample data that will converge to this same matrix. The matrix S_0 above has been shown to possess that property of convergence, and is thus the basis for White's heteroskedasticity–consistent estimator of the asymptotic covariance matrix:

$$\text{Est. Asy. Var.}[b] = n(X'X)^{-1}S_0(X'X)^{-1}$$

which gives us an interval estimate that is robust to unknown forms of heteroskedasticity. It is this estimator that is utilized in any computer program that generates "robust standard errors"; for instance, the `robust` option on a Stata estimation command generates the standard errors via White's formula.

However, the deviation of Σ from I may involve more than pure heteroskedasticity; Σ need not be a diagonal matrix. What if we also must take serial correlation of the errors into account? The natural counterpart to White's formula would be

$$\hat{Q} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} e_i e_j x_i x_j'$$

but as it happens this estimator has two problems. Since this quantity is $\frac{1}{n}$ times a sum of n^2 terms, it is difficult to establish that it will converge to anything, let alone Q. To obtain convergence, the terms involving products

of residuals—which are estimates of the autocorrelations between ϵ_i and ϵ_j—must decline as the distance between i and j grows. This unweighted sum will not meet that condition. If we weight the terms in this summation, and the weights decline sufficiently rapidly, then the sums of these n^2 terms can indeed converge to constant values as $n \to \infty$. There is still a practical problem, however, in that even a weighted sum may not yield a positive definite matrix, since the sample autocorrelations contain sampling error. The matrix autocovariogram must be positive definite, but estimates of it may not be. Thus, some sort of kernel estimator is needed to ensure that the resulting matrix will be positive definite.

The first solution to this issue was posed by Newey and West (1987), who proposed the estimator

$$\hat{Q} = S_0 + \frac{1}{n} \sum_{l=1}^{L} \sum_{t=l+1}^{n} w_l e_t e_{t-l} (x_t x'_{t-l} + x_{t+l} x'_t)$$

which takes a finite number L of the sample autocorrelations into account, employing the Bartlett kernel estimator

$$w_l = 1 - \frac{l}{L+1}$$

to generate the weights. Newey and West have shown that this estimator guarantees that $\hat{Q}$ will be positive definite. The estimator is said to be "HAC" : heteroskedasticity– and autocorrelation-consistent, allowing for any deviation of Σ from I up to L^{th} order autocorrelation. The user must specify her choice of L, which should be large enough to encompass any likely serial correlation in the error process. One rule of thumb that has been used is to choose $L = \sqrt[4]{n}$. This estimator is that available in the Stata command `newey`, which may be used as an alternative to `regress` for OLS estimation with HAC standard errors.

Two issues remain with the HAC estimator of the asymptotic covariance matrix: first, although the Newey–West estimator is widely

used, there is no particular justification for the use of the Bartlett kernel. There are a number of alternative kernel estimators that may be employed, and some may have better properties in specific instances. The only requirement is that the kernel deliver a positive definite covariance matrix.

Second, if there is no reason to question the assumption of homoskedasticity, it may be attractive to deal with serial correlation under that assumption. One may want the "AC" without the "H". The standard Newey–West procedure does not allow this.

The `ivreg2` routine can estimate robust, AC, and HAC standard errors for either OLS, IV, or IV-GMM models. It provides a choice of a number of alternative kernels.

Efficient estimation via generalized least squares

Efficient estimation of β requires knowledge of Ω. Consider the case where that matrix is known, positive definite and symmetric. It may be factored as $\Omega = C\Lambda C'$ where the columns of C are the eigenvectors of Ω and $\Lambda = diag(\lambda)$ where λ is the vector of eigenvalues of Ω. Let $T = C\Lambda^{1/2}$, such that $\Omega = TT'$. Also, let $P' = C\Lambda^{-1/2}$, such that $\Omega^{-1} = P'P$. Thus we can premultiply the regression model $y = X\beta + \epsilon$ by P, $Py = PX\beta + P\epsilon$, and $E[P\epsilon\epsilon'P'] = \sigma^2 I$. Since Ω is known, the observed data y, X may be transformed by P, and the resulting estimator is merely OLS on the transformed model. The efficient estimator of β, given the Gauss–Markov theorem, is thus the generalized least squares or "Aitken estimator":

$$\begin{aligned}\widehat{\beta} &= (X'P'PX)^{-1}(X'P'Py) \\ &= (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}y)\end{aligned}$$

This may be viewed as a case of weighted least squares, where OLS uses the improper weighting matrix I, rather than the appropriate weights of Ω^{-1}. The GLS estimator is the minimum variance linear unbiased estimator of the generalized least squares model, of which OLS is a special case. The residuals from this model are based on the transformed data, so that the GLS estimate of σ^2 is

$$\widehat{\sigma^2} = \frac{(y - X\hat{\beta})'\Omega^{-1}(y - X\hat{\beta})}{n - K}$$

There is no precise counterpart to R^2 in the GLS context. For instance, we could consider the R^2 of the OLS model estimated above, but that model need not have a constant term. In any case, that model reflects how well the parameters fit the transformed data, not the original data of interest. We might rather consider the GLS parameters applied to the original data, which can be used to generate $\hat{y}$ and a residual series in that metric. However, one

must note that the objective of GLS is to minimize the sum of squares of the transformed residuals (that is, those based on the transformed data), and that does not necessarily imply that the sum of squared residuals based on the original data will be minimized in the process.

Feasible generalized least squares

If we relax the assumption that Ω is known, we confront the issue of how to estimate it. Since it contains $n(n+1)/2$ distinct elements, it cannot be estimated from n observations. We must impose constraints to reduce the number of unknown parameters, as $\theta = \theta(\Omega)$, where the number of elements in θ is much less than n. In the time series context, for instance, a specification of AR(1) will reduce the number of unknown parameters to one: ρ in $\epsilon_t =$

$\rho\epsilon_{t-1} + v_t$, which causes all off–diagonal elements of Ω to be powers of ρ. Likewise, one can specify a model of pure heteroskedasticity which only contains one additional parameter, such as $\sigma_i^2 = \sigma^2 z_i^\gamma$ where z_i is some observable magnitude (such as the size of a firm, or the income of a household) and γ is to be estimated. In either case, we may consider that we have a consistent estimator of θ (ρ in the former case, γ in the latter). Then feasible GLS estimation will involve $\hat{\Omega} = \Omega(\hat{\theta})$ rather than the true Ω. What will be the consequences of this replacement?

If the plim of the elements of $\hat{\theta}$ equal the respective elements of θ, then using $\hat{\Omega}$ is asymptotically equivalent to using Ω itself. The feasible GLS estimator is then

$$\hat{\hat{\beta}} = (X'\hat{\Omega}^{-1}X)^{-1}(X'\hat{\Omega}^{-1}y)$$

and we need not have an efficient estimator of θ to ensure that this feasible estimator of

β is asymptotically efficient; we only need a consistent estimator of θ. Except for the simplest textbook cases, the finite–sample properties and exact distributions of feasible GLS (FGLS) estimators are unknown. With normally distributed disturbances, the FGLS estimator is also the maximum likelihood estimator of β. An important result due to Oberhofer and Kmenta (1974) is that if β and θ have no parameters in common, a "back–and–forth" approach which estimates first one, then the other of those vectors will yield the MLE of estimating them jointly; and that there is in that case no gain in asymptotic efficiency in knowing Ω over consistently estimating its contents.

Heteroskedasticity

Heteroskedasticity appears in many guises in economic and financial data, in both cross–section and time–series contexts. In the former, we often find that disturbance variances

are related to some measure of size: total assets or total sales of the firm, income of the household, etc. Alternatively, we may have a dataset in which we may reasonably assume that the disturbances are homoskedastic within groups of observations, but potentially heteroskedastic between groups. As a third potential cause for heteroskedasticity, consider the use of grouped data, in which each observation is the average of microdata (e.g., state–level data for the US, where the states have widely differing populations). Since means computed from larger samples are more accurate, the disturbance variance for each observation is known up to a factor of proportionality.

We often find heteroskedasticity in time–series data: particularly a phenomenon known as volatility clustering, which appears in high–frequency data from financial markets. We will not discuss this context of (conditional) heteroskedasticity at length, but you should be aware that

the widespread use of ARCH and GARCH models for high–frequency time–series data is based on the notion that the errors in these contexts are conditionally heteroskedastic.

What happens if we use OLS in a heteroskedastic context? The OLS estimator b is still unbiased, consistent, and asymptotically normal, but its covariance matrix is based on the wrong formula. Thus, the interval estimators are biased (although we can show that the plim of s^2 is σ^2 as long as we use a consistent estimator of b). The greater is the dispersion of ω_i (the diagonal element of Ω for the i^{th} observation) the greater the degree of inefficiency of the OLS estimator, and the greater the gain to using GLS (if we have the opportunity to do so). If the ω_i are correlated with any of the variables in the model, the difference between the OLS and GLS covariance matrices will be sizable, since the difference Δ depends on

$$\frac{1}{n}\sum_{i=1}^{N}(1-\omega_i)x_i x_i'$$

where x_i is the i^{th} row of the X matrix.

In the case of unknown heteroskedasticity, we will probably employ the White (Huber, sandwich) estimator of the covariance matrix that is implied by the "robust" option of Stata. If we have knowledge of Ω, we should of course use that information. If $\sigma_i^2 = \sigma^2 \omega_i$, then we should use the weighted least squares (WLS) estimator in which each observation is weighted by $\frac{1}{\omega_i}$: that is, the P matrix is diag($\frac{1}{\omega_i}$). Observations with smaller variances receive a larger weight in the computation of the sums, and therefore have greater weight in computing the weighted least squares estimates.

Consider the case where the firm–specific error variance is assumed to be proportional to firm size, so that $\omega_i = x_{ik}^2$, where x_k is the variable measuring size. Then the transformed regression model involves dividing through the equation by x_{ik}. This can be achieved in Stata by

creating a variable that is proportional to the observation's error variance, e.g., `gen size2 = size*size`, and then specifying to Stata that this variable is to be used in the expression for the analytical weight (`aw`) in the regression: e.g. `regress q l k [aw=1/size2]`, in which the analytical weight is assumed to be (proportional to) the inverse of the observation variance. Note that the way in which you specify WLS differs from package to package; in some programs, you would give $size2$ itself as the weight!

What about the case in which we have grouped data, representing differing numbers of microdata records? Then we have a known Ω, depending on the n underlying each observation. Each observation in our data stands for an integer number of records in the population. Say that we have, for each U.S. state, the population, recorded in variable `pop`. Then we might

say `regress saving income [fw=pop]`, in which we specify what Stata calls a frequency weight: the number of observations in the population corresponding to each observation in the sample. This will cause the total N of the regression to be reported as the sum of the `pop` variable.

What if we do not have knowledge of Ω, and must make some assumptions on its contents? Then we face the issue: how good is our information (or ability to estimate from OLS), and would we be better off using an estimated Ω, or using a robust estimation technique which makes no assumptions on its contents? There is the clear possibility that using faulty information on Ω, although it may dominate OLS, may be worse than using a robust covariance matrix. And the most egregious (but not uncommon) error, weighting "upside down", will exacerbate the heteroskedasticity rather than removing it!

Estimation of an unknown Ω

The estimation of Ω proceeds from the use of OLS to obtain estimates of σ_i^2 from the least squares residuals. The OLS residuals, being functions of the point estimates, are consistent, even though OLS is not efficient in this context. They may be used to estimate the variances associated with groups of observations (in which some sort of groupwise heteroskedasticity is to be modeled) or the variance of individual observations as a function of a set of auxiliary variables z via a regression of the squared residuals on those variables. In this latter case, one may want to consider reformulating the model by using the information in z: for instance, if it appears that the residuals' variances are related to (some power of) size, the regression model might be scaled by the size variable. The common use of per capita measures, logarithmic functional

forms, and ratios of level variables may be considered as specifications designed to mitigate problems of heteroskedasticity that would appear in models containing level variables.

Testing for heteroskedasticity

The Breusch–Pagan/Godfrey/Cook–Weisberg and White/Koenker statistics are standard tests of the presence of heteroskedasticity in an OLS regression. The principle is to test for a relationship between the residuals of the regression and p indicator variables that are hypothesized to be related to the heteroskedasticity. Breusch and Pagan (1979), Godfrey (1978) and Cook and Weisberg (1983) separately derived the same test statistic. This statistic is distributed as χ^2 with p degrees of freedom under the null of no heteroskedasticity, and under the maintained hypothesis that the error of the regression is normally distributed.

Koenker (1981) noted that the power of this test is very sensitive to the normality assumption, and presented a version of the test that relaxed this assumption. Koenker's test statistic, also distributed as χ^2_p under the null, is easily obtained as nR^2_c, where R^2_c is the centered R^2 from an auxiliary regression of the squared residuals from the original regression on the indicator variables. When the indicator variables are the regressors of the original equation, their squares and their cross-products, Koenker's test is identical to White's (1980) nR^2_c general test for heteroskedasticity. These tests are available in Stata, following estimation with `regress`, using `ivhettest` as well as via `hettest` and `whitetst`.

As Pagan and Hall (1983) point out, the above tests will be valid tests for heteroskedasticity in an IV regression only if heteroskedasticity is present in that equation and *nowhere else in*

the system. The other structural equations in the system (corresponding to the endogenous regressors) must also be homoskedastic, even though they are not being explicitly estimated. Pagan and Hall derive a test which relaxes this requirement. Under the null of homoskedasticity in the IV regression, the Pagan–Hall statistic is distributed as χ^2_p, irrespective of the presence of heteroskedasticity elsewhere in the system. A more general form of this test was separately proposed by (White (1982). Our implementation is of the simpler Pagan–Hall statistic, available with the command `ivhettest` after estimation by `ivreg` or `ivreg2`.

Let Ψ be the $n \times p$ matrix of indicator variables hypothesized to be related to the heteroskedasticity in the equation, with typical row Ψ_i. These indicator variables must be exogenous, typically either instruments or functions of the instruments. Common choices would be:

1. The levels only of the instruments Z (excluding the constant). This is available in `ivhettest` by specifying the `ivlev` option, and is the default option.

2. The levels and squares of the instruments Z, available as the `ivsq` option.

3. The levels, squares, and cross-products of the instruments Z (excluding the constant), as in the White (1980) test. This is available as the `ivcp` option.

4. The "fitted value" of the dependent variable. This is *not* the usual fitted value of the dependent variable, $X\widehat{\beta}$. It is, rather, $\widehat{X}\widehat{\beta}$, i.e., the prediction based on the IV estimator $\widehat{\beta}$, the exogenous regressors Z_2,

and the fitted values of the endogenous regressors $\widehat{X}_1$. This is available in `ivhettest` by specifying the `fitlev` option.

5. The "fitted value" of the dependent variable and its square (`fitsq` option).

6. A user–defined set of indicator variables may also be provided for `ivhettest`.

The trade-off in the choice of indicator variables is that a smaller set of indicator variables will conserve degrees of freedom, at the cost of being unable to detect heteroskedasticity in certain directions.

Let

$$\begin{aligned}
&\overline{\Psi} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \Psi_i && \text{dimension} = n \times p \\
&\hat{D} \equiv \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \Psi_i'(\hat{u}_i^2 - \hat{\sigma}^2) && \text{dimension} = n \times 1 \\
&\hat{\Gamma} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} (\Psi_i - \hat{\Psi})' X_i \hat{u}_i && \text{dimension} = p \times K \\
&\hat{\mu}_3 = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \hat{u}_i^3 \\
&\hat{\mu}_4 = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \hat{u}_i^4 \\
&\hat{X} = P_z X
\end{aligned} \tag{1}$$

If u_i is homoskedastic and independent of Z_i, then Pagan and Hall (1983) (Theorem 8) show that under the null of no heteroskedasticity,

$$n\hat{D}'\hat{B}^{-1}\hat{D} \overset{A}{\sim} \chi_p^2 \tag{2}$$

where

$$\hat{B} = B_1 + B_2 + B_3 + B_4$$

$$B_1 = (\hat{\mu}_4 - \hat{\sigma}^4)\tfrac{1}{n}(\Psi_i - \overline{\Psi})'(\Psi_i - \overline{\Psi})$$

$$B_2 = -2\hat{\mu}^3\tfrac{1}{n}\Psi'\hat{X}(\tfrac{1}{n}\hat{X}'\hat{X})^{-1}\hat{\Gamma}' \qquad (3)$$

$$B_3 = B_2'$$

$$B_4 = 4\hat{\sigma}^2\tfrac{1}{n}\hat{\Gamma}'(\tfrac{1}{n}\hat{X}'\hat{X})^{-1}\hat{\Gamma}$$

This is the default statistic produced by `ivhettest`. Several special cases are worth noting:

- If the error term is assumed to be normally distributed, then $B_2 = B_3 = 0$ and $B_1 = 2\hat{\sigma}^4\frac{1}{n}(\Psi_i - \overline{\Psi})'(\Psi_i - \overline{\Psi})$. This is available from `ivhettest` with the `phnorm` option.

- If the rest of the system is assumed to be homoskedastic, then $B_2 = B_3 = B_4 = 0$

and the statistic in (2) becomes the White / Koenker nR_c^2 statistic. This is available from `ivhettest` with the `nr2` option.

- If the rest of the system is assumed to be homoskedastic and the error term is assumed to be normally distributed, then $B_2 = B_3 = B_4 = 0$, $B_1 = 2\hat{\sigma}^4 \frac{1}{n}(\Psi_i - \overline{\Psi})'(\Psi_i - \overline{\Psi})$, and the statistic in (2) becomes the Breusch–Pagan/Godfrey/Cook–Weisberg statistic. This is available from `ivhettest` with the `bpg` option.

All of the above statistics will be reported with the `all` option. `ivhettest` can also be employed after estimation via OLS or HOLS using `regress` or `ivreg2`. In this case the default test statistic is the White/Koenker nR_c^2 test.

The Pagan–Hall statistic has not been widely used in practice, perhaps because it is not a standard feature of most regression packages.

The Breusch–Pagan (/Cook–Weisberg) statistic can also be computed via Stata user–written command `bpagan` or built–in command `hettest` in the context of a model estimated with `regress`. Likewise, White's general test (and the variant using fitted values) may be computed via Stata user–written command `whitetst` after `regress`.

We will not discuss the Goldfeld–Quandt (1965) test here, since its usefulness is limited relative to the other tests described above.

Serial correlation

Serial correlation in the errors may arise due to omitted factors in the regression model, in

which case its diagnosis represents misspecification. But there are cases where errors will be, by construction, serially correlated rather than independent across observations. Theoretical schemes such as partial–adjustment mechanisms and adaptive expectations can give rise to errors which cannot be serially independent. Thus, we also must consider this sort of deviation of Ω from I: one which is generally more challenging to deal with than is pure heteroskedasticity.

As with the latter case, OLS is inefficient, with unbiased and consistent point estimates, but an inappropriate covariance matrix of the estimated parameters, rendering hypothesis tests and confidence intervals invalid.

First, some notation: in applying OLS or GLS to time–series data, we are usually working with series that have been determined to be

stationary in some sense. If a time–series process is covariance stationary or weakly stationary, it has a constant mean and variance and an autocorrelation function whose elements only depend on the temporal displacement. It is obvious that many series would fail to meet these conditions, if only for their having a non–constant mean over time. Nevertheless, if the variation in the mean can be characterized as a deterministic trend, that trend can be removed. There may also be a concern of a time–varying variance. Since we test regression models for regime shifts, involving changes in the model's parameters over time in response to certain events, might we not also be concerned about a changing variance? Naturally, that is a possibility, and in this sense we might consider this a form of groupwise heteroskedasticity, where the groups are defined by various time periods.

One way in which a time–series process might fail to exhibit (covariance) stationarity would be that its variance might not be finite. Consider the process

$$y_t = \beta_1 + \beta_2 y_{t-1} + \epsilon_t$$

Assume that $\epsilon \sim N(0, \sigma_\epsilon^2)$ is a stationary process. (Since we assume that it is normally distributed, it will be strongly stationary, since its entire distribution is described by these two moments). Then the mean of this process is a function of the lag coefficient:

$$E[y] = \frac{\beta_1}{1 - \beta_2}$$

and the variance of the y process is merely an amplification of the variance of the ϵ process:

$$\gamma_0 = \frac{\sigma_\epsilon^2}{1 - \beta_2^2}$$

where γ_0 is the first element of the autocovariance function of y. The variance of y will

be finite and positive as long as β_2 is inside the unit circle. We may note that this expression, in which we have assumed covariance stationarity for y in order to state that $Var(y_t) = Var(y_{t-1})$, may also be derived from back–substitution of the DGP for y_t, showing that the current value y_t may be written in terms of an infinite sum of the ϵ process, with each element $\epsilon_{t-\tau}$ weighted by β_2^τ.

Now let us consider the autocovariances of the y process. Although the elements of ϵ are serially independent, the elements of y clearly will not be. In fact, the

$$cov[y_t, y_{t-1}] = cov[y_t, y_{t+1}] = \gamma_1 = \frac{\beta_2 \sigma_\epsilon^2}{1 - \beta_2^2}.$$

while the covariance between elements of y two periods apart will be

$$cov[y_t, y_{t-2}] = \gamma_2 = \frac{\beta_2^2 \sigma_\epsilon^2}{1 - \beta_2^2}$$

and so on. We may define the autocorrelation function as

$$Corr[y_t, y_{t-\tau}] = \frac{\gamma_\tau}{\gamma_0}.$$

For the y process, the autocorrelations are β_2, $\beta_2^2, \ldots$. Since y is a so–called $AR(1)$ process: an autoregressive model of order one, its autocorrelations will be geometric, defined by powers of β_2. We may write such a model using the lag operator L as

$$(1 - \beta_2 L)y_t = \beta_1 + \epsilon_t$$

and we may consider the root of the autoregressive polynomial $1 - \beta_2 L = 0$ as defining the behavior of the series. That root is β_2^{-1}, which must lie outside the unit circle if this first–order difference equation in y is to be stationary. What if the root lies on the unit circle? Then we have a so–called unit root process, which will possess an infinite variance. Such a process is said to be nonstationary, or integrated of order one ($I(1)$), since differencing

the process once will render it stationary (or $I(0)$). That is, if we consider the random walk

$$y_t = y_{t-1} + \epsilon_t$$
$$(1 - L)y_t = \epsilon_t$$

The first difference of this process will be white noise:

$$\Delta y_t = \epsilon_t$$

A nonstationary or integrated process should be used in a regression equation—either as the dependent variable or as a regressor—only with great care, since in general regressions containing such variables are said to be spurious, indicating the existence of correlations that do not exist in the data generating process. That is, an OLS regression of two independent random walks will not yield an unbiased and consistent estimate of the population slope parameter, which equals zero. There are circumstances where we may use regression on

nonstationary variables, but for such a model to make sense, we must demonstrate that the nonstationary variables are cointegrated in the sense of Granger (1986). Such variables are said to contain stochastic trends (since a constant in the equation above will imply the so–called random walk with drift model) and a major challenge for time series modelling is to distinguish between a deterministic trend, which may be extracted via detrending procedures, and a stochastic trend, which must be removed by differencing. Neither remedy is appropriate for the other. To establish whether a stochastic trend is present in a series, we must utilize a unit root test of the sort proposed by Dickey and Fuller (or the improved version of Elliott, Rothenberg, Stock known as DF–GLS, available in Stata as `dfgls`), Phillips and Perron (Stata routine `pperron`) or Kwiatkowski et al. (Stata user–contributed routine `kpss`).

The distinction between stationary and integrated processes also implies that if we have an autocorrelated error process, we would not generally want to apply a first difference operator to the series, since that would imply that we assumed that $\rho = 1$. If we believe that the error process follows an AR(1) model, with $|\rho| < 1$, then the first difference operator will not be appropriate; we should apply quasi–differencing using a consistent estimate of ρ.

The conclusion that in the presence of serial correlation one may apply OLS to generate unbiased and consistent estimates of the parameters b has one important exception. If the regression contains a lagged dependent variable as well as autocorrelated errors, the regressor and the disturbance are correlated by construction, and neither OLS nor GLS will generally

be consistent. The problem can be cast in terms of omitted variables: e.g. if

$$y_t = \beta y_{t-1} + \epsilon_t$$
$$\epsilon_t = \rho \epsilon_{t-1} + u_t$$

In which both β and ρ lie within the unit circle. If we subtract ρy_{t-1} from y_t, we arrive at

$$y_t = (\beta + \rho) y_{t-1} - \beta \rho y_{t-2} + u_t$$

which is a proper OLS regression, since u_t is not correlated with either lag of y. However, this regression does not yield estimates of the original parameters, but only of combinations of those parameters. The inconsistency of OLS in this context may be shown in terms of the plim of the OLS estimator b:

$$\frac{\beta + \rho}{1 + \beta \rho}$$

This quantity will only equal β if $\rho = 0$; otherwise it will lie between β and unity. An approach that would take account of the problem is an instrumental variables estimator: if

the error process is AR(1), then y_{t-2} is an appropriate instrument for y_{t-1} in the original regression.

What happens, in the absence of lagged dependent variables, if we use OLS rather than GLS to estimate the covariance matrix of the parameter estimates? In the presence of positive first–order serial correlation, we can show that the t–statistics are biased upward (that is, the variances are biased downward) by our ignoring the serial correlation in the error process. We may either apply the appropriate GLS estimator, or use a HAC estimator of the covariance matrix such as that proposed by Newey and West.

Testing for autocorrelation

How might we test for the presence of autocorrelated errors? Like the case of pure heteroskedasticity, we may base tests of serial correlation on the estimated moments of the OLS

residuals. If we estimate the regression of e_t on e_{t-1}, the slope estimate will be a consistent estimator of the first–order autocorrelation coefficient ρ_1 of the ϵ process. A generalization of this procedure is the Lagrange Multiplier (LM) test of Breusch and Godfrey, in which the OLS residuals are regressed on the original X matrix augmented with p lagged residual series. The null hypothesis is that the errors are serially independent up to order p, and proceeds by considering the partial correlations of the OLS residual process (with the X variables partialled off). Of course the residuals at time t are orthogonal to the columns of X at time t, but that need not be so for the lagged residuals. This is perhaps the most useful test, in that it allows the researcher to examine more than first–order serial independence of the errors in a single test, and is available in Stata as `bgodfrey`.

A variation on the BP test is the Q test of Box and Pierce (1970), as refined by Ljung and Box (1979), which examines the first p autocorrelations of the residual series:

$$Q = T(T+2) \sum_{j=1}^{p} \frac{r_j^2}{T-j}$$

where r_j^2 is the j^{th} empirical autocorrelation of the residual series. This test, unlike the BP test, does not condition on X: it is based on the simple correlations of the residuals rather than their partial correlations. It is less powerful than the BP test when the null hypothesis (of no serial correlation in ϵ up to order p) is false, but it is not model–dependent. Under the null hypothesis, $Q \sim \chi^2(p)$. The Q test is available in Stata as `wntestq`: labelled such to indicate that it may be used as a general test for white noise.

The oldest test (but still widely employed and

reported, despite its shortcomings) is the Durbin–Watson d statistic:

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

The D–W test proceeds from the principle that the numerator of the statistic, when expanded, contains twice the variance of the process minus twice the (first) autocovariance of the series. If $\rho = 0$, that autocovariance will be near zero, and the D–W will equal 2.0. As ρ increases, D–W $\rightarrow 0$, while as $\rho \rightarrow -1$, D–W $\rightarrow 4$. However, the exact distribution of the statistic depends on the regressor matrix (which must contain a constant term, and must not contain a lagged dependent variable), so that rather than having a set of critical values, the D–W test has two, labelled d_L and d_U. If the statistic falls below d_L, a rejection is indicated; above d_U, one does not reject; and in between, the statistic is inconclusive. (For negative autocorrelation, one tests

$4 - d$ against the same tabulated critical values). The test is available in Stata as `dwstat`, and is automatically provided in the `prais` GLS estimation command.

In the presence of a lagged dependent variable, the D–W statistic is biased toward 2, and Durbin's alternative (or "h") test must be used. That test is a Lagrange multiplier test in which one regresses residuals on their own lags and the original X matrix. The test is asymptotically equivalent to the BP test for $p = 1$, and is available in Stata as command `durbina`.

GLS estimation with serial correlation

If the Ω matrix is known: in the case of AR(1) errors, if ρ_1 is known—then we may apply GLS to the data by constructing quasi–differences,

$y_t - \rho y_{t-1}$, $X_{j,t} - \rho X_{j,t-1}$, etc. for observations 2–T. The first observation is multiplied by $\sqrt{1-\rho^2}$. One may also apply an algebraic transformation in the case of AR(2) errors.

But what if we must estimate ρ_1 (or ρ_1 and ρ_2)? Then any consistent estimator of those parameters will suffice to define the feasible Aitken estimator. The Prais–Winsten estimator uses an estimate of ρ_1 based on the OLS residuals to create $\widehat{\Omega}$; the Cochrane–Orcutt variation on that estimator differs only in its treatment of the first observation. Either of these estimators may be iterated to convergence: essentially they operate by "ping-ponging" back and forth between estimates of β and θ (equal in this case to the single parameter ρ). Iteration refines the estimate of ρ: not asymptotically necessary, but recommended in small samples. Both estimators are available in Stata via the `prais` command.

Other approaches include that of maximum likelihood, which estimates a single parameter vector $[\beta \ \theta]'$, and the grid search approach of Hildreth and Lu. Although one might argue for the superiority of MLE in this context, Monte Carlo studies suggest that the Prais–Winsten estimator is nearly as efficient in practice.

In summary, although we may employ GLS to deal with detected problems of autocorrelation, we should always be open to the possibility that this diagnosis reflects misspecification of the model's dynamics, or omission of one or more key factors from the model. We may mechanically correct for first–order (or higher–order) serial correlation in a model, but we are then attributing this persistence to some sort of "clockwork" in the error process rather than explaining its existence.

Forecasting with serially correlated errors

It is easy to show that in the presence of AR(p) errors that the standard OLS forecast will no longer be the BLUP. Consider a one–step–ahead forecast beyond the model's horizon. We no longer have $E[\epsilon_{t+1}|\Psi_t] = 0$, where Ψ_t is the information set at time t. Since in the last period of the sample, the least squares residual was not zero, we would not expect the next error to be zero either; our conditional expectation of that quantity, based on the model, is $\widehat{\rho}\ e_T$. Thus, this quantity should be added to the least squares prediction Xb to generate that one–step–ahead forecast. Likewise, a two–step–ahead forecast would include a term $\widehat{\rho}^2\ e_T$, and so on. The interval estimates will likewise be modified for the presence of serial correlation. This logic will be applied in the case of more complex serial correlation structures (such as AR(p) or moving average, MA(q), error processes) as well.

The cluster estimator of the covariance matrix

One particular deviation from independence of the errors, which may apply in a cross–sectional context as well as a time–series context, is Stata's implementation of the "cluster" estimator. Application of this option does not affect the point estimates: like the "robust" option, it only affects the covariance matrix of the estimated parameters. But although the robust option may only deal with pure heteroskedasticity, the cluster option allows for both non–independent errors within "clusters" and heteroskedastic errors across clusters. Clusters are defined by some additional variable that indicates group membership. For instance, the 74 automobiles in Stata's `auto.dta` may be clustered by manufacturer, so that the manufacturer is the "cluster ID". We then assume that the errors may not be independent among a manufacturer's models, but are independent

(possibly with differing error variances) across manufacturers. The estimator may be invoked in most of Stata's estimation commands by using the `cluster(clid)` option, which implies `robust` as well. One could imagine applying cluster in a time–series context: for instance, the observations on macro aggregates corresponding to a particular presidential administration.