Section 6: Heteroskedasticity and Serial Correlation

Jeffrey Greenbaum, University of California, Berkeley
Section 6: Heteroskedasticity and Serial Correlation

Jeffrey Greenbaum

February 23, 2007

Contents

1 Section Preamble 2

2 Weighted Least Squares 3

3 Feasible WLS 3
   3.1 Multiplicative Heteroskedasticity Models 3
   3.2 Testing for Heteroskedasticity 4
   3.3 Feasible Estimator 6
   3.4 Exercises 6
      3.4.1 2002 Exam, 1B 6
      3.4.2 2004 Exam, 1D 7
      3.4.3 Grouped-Data Regression Model 7
      3.4.4 Multiplicative Model 8

4 Eicker-White Standard Errors 9

5 Structural Approach to Serial Correlation 9
   5.1 First-Order Serial Correlation 10
   5.2 Testing for Serial Correlation 11
   5.3 Feasible GLS 13
   5.4 Exercises 14
      5.4.1 2002 Exam, Question 1C 14
      5.4.2 2003 Exam, Question 1B 14
      5.4.3 2004 Exam, Question 1B 15

6 Nonstructural Approach to Serial Correlation 16
1 Section Preamble

This week we continue with the relaxed scalar variance-covariance assumption and GLS estimation. Specifically, we consider two more cases in which we can possibly construct a Feasible GLS estimator that has the same asymptotic properties as the GLS estimator. We will first analyze when the variance-covariance matrix demonstrates pure heteroskedasticity, and then serial correlation.

Recall that when we relax the scalar variance-covariance assumption that \( \hat{\beta}_{GLS} \) is the most efficient of the linear unbiased estimators. However, \( \hat{\beta}_{GLS} \) requires that we know the values of \( \Sigma \), which is not realistic because we do not know the values of \( \varepsilon \). Nevertheless, if we can consistently estimate \( \Sigma \), then we can construct a feasible estimator, \( \hat{\beta}_{FGLS} \), which has the same asymptotic properties as \( \hat{\beta}_{GLS} \). Finding a consistent estimator for \( \Sigma \), however, is often not possible because generally as the sample size increases, the number of elements in \( \Sigma \) to estimate increases by a greater amount. Moreover, we generally assume that each element effectively comes from a distinct distribution unless we pose structure on how \( \Sigma \) is formed. Posing the correct structure can enable us to consistently estimate \( \Sigma \).

The spirit of the solutions to heteroskedasticity and serial correlation are similar. As previously suggested, one approach is to assume a functional form for the structure of \( \Sigma \), estimate this structure, and use this estimate to construct \( \hat{\beta}_{FGLS} \). If the correct structure is chosen then this estimator has the same asymptotic properties as \( \hat{\beta}_{GLS} \), and so asymptotically, \( \hat{\beta}_{FGLS} \) is consistent and BLUE. Before proceeding with feasible estimation, we should conduct an hypothesis test where the null hypothesis is homoskedasticity or the lack of serial correlation, as appropriate to the case. If we believe that \( \Omega = I \), then we can use \( \hat{\beta}_{GLS} \), which has desirable finite sample properties. Nevertheless, hypothesis testing may spuriously lead to detecting heteroskedasticity or serial correlation. Moreover, we may either assume the wrong structure of \( \Sigma \), or more simply have no intuition about what its structure might be. Lastly, point estimates from [F]GLS do not lend themselves to the desirable ceteris paribus interpretation.

An alternative approach is use OLS, which remains unbiased and consistent, and to correct the standard errors so that they are consistently estimated. Although OLS is no longer BLUE if \( \text{Var}(y|X) = \Sigma \), this method is more frequently used because of these concerns about posing a structure for \( \Sigma \). Despite the loss of efficiency and some caution about the finite-sample properties of these corrected standard errors, if hypothesis testing produces highly statistically significant point estimates, then we can likely trust inferences with OLS. Moreover, OLS point estimates are appealing for policy applications because they lend to lend to a ceteris paribus interpretation.

Although the OLS and [F]GLS estimators are both unbiased, at least asymptotically, point estimates inevitably differ unless \( \text{Var}(y|X) = \sigma^2 I \). Nevertheless, it is not necessary to be concerned with such differences unless the difference is economically significant, such as a difference in sign, while inference on both are each highly statistically significant. In this case, another classical assumption is likely to be faulty, such as the linear expectations assumption, as we will begin to discuss next week.
2 Weighted Least Squares

GLS estimation with pure heteroskedasticity is known as weighted least squares. We first consider the case in which all of the elements along the diagonal of $\Sigma$, or equivalently $\Omega$, are known. Then, we consider a more realistic scenario where we can construct a feasible estimator by estimating the heteroskedasticity given an assumed functional form.

In the case of pure heteroskedasticity, the generalized classical regression assumption of $\text{Var}(y|X) = \Sigma$ reduces to $\text{Var}(y|X) = \text{Diag}[\sigma^2_i]$. That is, pure heteroskedasticity denotes that the diagonal elements of $\Sigma$ are heteroskedastic but the off diagonal elements are not serially correlated. Using the framework established for GLS, $\hat{\beta}_{WLS}$ is the ordinary least squares estimate applied to the linear model that is multiplied through by $\Sigma^{-1/2}$. Although the interpretation of WLS coefficients are not as desirable as OLS coefficients, $\hat{\beta}_{WLS}$ is BLUE. Moreover, if we assume that the errors are independent and distributed normally, then finite sample inference is most desirable with $\hat{\beta}_{WLS}$.

Let $w_i = \frac{1}{\sigma_i^2}$. Because $\Sigma$ is diagonal, $\Sigma^{-1/2} = \text{Diag}[w_i^{1/2}]$. As a result,

$$\hat{\beta}_{WLS} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y = (X'(\text{Diag}[w_i])X)^{-1}X'(\text{Diag}[w_i])y$$

$$= \left(\sum_{i=1}^{n} \omega_i(x_i'x_i')^{-1}\right)^{-1}\sum_{i=1}^{n} \omega_i x_i y_i$$

Accordingly, this estimator is known as weighted least squares because it is equivalently derived by minimizing the weighted sum of the squared residuals. Specifically, each squared residual, $e_i^2$ is multiplied by $w_i$. The reason for this interpretation is because we are transforming our linear model by $\Sigma^{-1/2}$. As with all GLS estimation, this transformation is equivalent to finding the estimator that thus minimizes $(y - X\beta)'\Sigma^{-1}(y - X\beta)$. By construction, $\Sigma^{-1} = \text{Diag}[w_i]$ for GLS with pure heteroskedasticity, which when rewritten in summation notation, lends to the weighted least squares interpretation.

3 Feasible WLS

In practice, $\Sigma$ contains unknown parameters, aside from an arbitrary constant of proportionality. We can thus construct a feasible estimate for $\hat{\beta}_{WLS}$ by estimating a model for $\text{Var}(y_i) = \sigma^2_i$ and substituting $\hat{\Sigma}$ in place of $\Sigma$. This feasible estimator has the same asymptotic properites as the WLS estimator if $\hat{\Sigma} \xrightarrow{p} \Sigma$.

3.1 Multiplicative Heteroskedasticity Models

In lecture, Professor Powell presented the multiplicative heteroskedasticity model because of its wide use in Feasible WLS, which is the linear model, $y_i = x_i\beta + u_i$, with error terms of the form:
where \( \varepsilon_i \sim iid, E(\varepsilon_i) = 0 \), and \( Var(\varepsilon_i) = \sigma^2 \).

Furthermore, we assume that the function \( c_i^2 \) is a linear function of the observable variables in an underlying linear way:

\[
c_i^2 = h(z_i^\prime \theta)
\]

where the variables \( z_i \) are some observable functions of the regressors, \( x_i \), excluding the constant term, and \( \theta \) is a vector of coefficients to be estimated. We will elaborate on the estimation procedure for \( \theta \) shortly when discussing the feasible WLS procedure. Moreover, \( h(\cdot) \), which must always be positive so that \( Var(y_i|x_i) > 0 \ \forall i \), is normalized so that \( h(0) = 1 \) and \( h'(0) \neq 0 \). For example, we would expect to observe greater variation in the profits of large firms than in those of small firms, even after accounting for firm sizes.

Combining these two assumptions about the structure of the variance yields:

\[
Var(u_i) = Var(c_i\varepsilon_i) = c_i^2 Var(\varepsilon_i) = h(z_i^\prime \theta)\sigma^2
\]

This variance is thus homoskedastic if \( Var(u_i) \) is constant \( \forall i \), or equivalently, if \( h(z_i^\prime \theta) \) is constant \( \forall i \). By our normalization, we know that \( h(z_i^\prime \theta) \) is constant \( \forall i \) if \( z_i^\prime \theta = 0 \) because \( h(0) = 1 \). If \( \theta = 0 \) then we are guaranteed that \( z_i^\prime \theta = 0 \). On the other hand, it is not sensible to expect \( z_i = 0 \). Therefore, if \( \theta = 0 \) then \( Var(u_i) = 1 \times \sigma^2 = \sigma^2 \), and \( u_i \) is homoskedastic.

### 3.2 Testing for Heteroskedasticity

Accordingly, a test for heteroskedasticity reduces to testing the null hypothesis \( H_0 : \theta = 0 \). The alternative hypothesis is \( H_1 : \theta \neq 0 \). We now derive a linear regression that lends to this hypothesis test. Note that this test presumes that we have correctly assumed the functional form for \( h(\cdot) \).

Under the null hypothesis,

\[
\begin{align*}
Var(\varepsilon_i) &= E(\varepsilon_i^2) - E(\varepsilon_i)^2 \\
&= E(\varepsilon_i^2) - 0^2 = E(\varepsilon_i^2) \\
Var(u_i) &= E(u_i^2) - E(u_i)^2 \\
&= E(u_i^2) - E(c_i\varepsilon_i)^2 \\
&= E(u_i^2) - c_i^2 E(\varepsilon_i)^2 = E(u_i^2) - (1)(0)^2 = E(u_i^2) \\
Var(u_i) &= h(z_i^\prime \theta)\sigma^2 = \sigma^2 \Rightarrow E(u_i^2) = \sigma^2 \\
\Rightarrow E(\varepsilon_i^2) &= E(u_i^2) = \sigma^2 = h(z_i^\prime \theta)\sigma^2
\end{align*}
\]
A first order Taylor Series approximation for \( h(z_i' \theta) \) about \( \theta = 0 \) is
\[
h(z_i' \theta) = h(0) + h'(0)z_i' \theta + R(z_i' \theta).
\]
We assume that as \( z_i' \theta \to 0 \), \( R(z_i' \theta) \to 0 \) at rate that is at least quadratic. This assumption can be quite problematic depending on the functional form for \( h(.) \), although it is effectively a footnote to the many other assumptions we are making. Nevertheless, we proceed with this assumption so that we can write
\[
h(z_i' \theta) = h(0) + h'(0)z_i' \theta = 1 + h'(0)z_i' \theta,
\]
which mathematically is false:
\[
E(u_i^2) = \sigma^2 h(z_i' \theta) = \sigma^2 (1 + h'(0)z_i' \theta).
\]

Let \( \delta = \sigma^2 h'(0) \theta \). Moreover, if we assume an error term, \( r_i \), such that \( E(r_i \mid z_i) = 0 \) and \( \text{Var}(r_i \mid z_i) = \tau \), then this model satisfies the classical linear assumptions. Therefore, we can test the regression:
\[
u_i^2 = \sigma^2 + z_i' \delta + r_i
\]
Since \( \theta = 0 \Rightarrow \delta = 0 \), we actually test the null hypothesis that \( H_0 : \delta = 0 \). Moreover, because \( E(z_i^2) = E(u_i^2) \), we could substitute \( z_i^2 \) in place of \( u_i^2 \). However, we cannot immediately proceed because we do not observe \( u_i \). Accordingly, the prevailing baseline test comes from Breusch and Pagan (1979). The justification for the test is beyond the scope of the course. Nevertheless, Professor Powell expects that you know the steps in Breusch and Pagan and that you could apply it to data.

Here is a summary of the tests Professor Powell presented. You are responsible for these tests insofar as Professor Powell presented them:

<table>
<thead>
<tr>
<th>Name</th>
<th>Expression</th>
<th>Distribution</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breusch-Pagan</td>
<td>( T = NR^2 )</td>
<td>( \chi_p^2 )</td>
<td>( p = \text{dim}(z_i) )</td>
</tr>
<tr>
<td>F</td>
<td>( F = (N-K)R^2 )</td>
<td>( \mathcal{F}_{p,N-K} )</td>
<td>( \mathcal{F} \approx T/p )</td>
</tr>
<tr>
<td>Studentized LM</td>
<td>( T' = \frac{\text{RSS}}{\hat{\varepsilon}_i^2} )</td>
<td>( \chi_p^2 )</td>
<td>if ( \varepsilon_i ) gaussian, ( \tau = 2\sigma^4 )</td>
</tr>
<tr>
<td>Goldfeld-Quandt</td>
<td>( s_1^2 / s_2^2 )</td>
<td>( \mathcal{F}_{[N/2]-k,N-[N/2]-k} )</td>
<td>gaussian ( \varepsilon_i ), one-sided</td>
</tr>
</tbody>
</table>

Here is the procedure for Breusch and Pagan (1979) to test the null hypothesis:

1. Use \( \hat{u}_i^2 = (y_i - x_i' \hat{\beta}_{OLS})^2 \) as a proxy for \( u_i^2 \) because the squared residuals are observable and are consistent estimators of the true squared errors. Note that \( \hat{u}_i \) actually makes more sense than \( \hat{\varepsilon}_i \) because that is technically our residual.

2. Regress the \( \hat{u}_i^2 \) on 1 and \( z_i \) and obtain the usual constant-adjusted \( R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \hat{\mu})^2}{\sum_{i=1}^{n}(y_i - \hat{\mu})^2} \).
3. Under the null, the statistic

\[ T = NR^2 \rightarrow_d \chi^2_p \]

where \( p = \text{dim}(\delta) = \text{dim}(z_i) \).

Reject \( H_0 \) if \( T \) exceeds the upper critical value of a chi-squared variable with \( p \) degrees of freedom.

### 3.3 Feasible Estimator

If the null hypothesis of homoskedasticity is rejected, then a correction for heteroskedasticity is needed. Using Feasible WLS, it is necessary to to estimate \( \hat{\Sigma} \), that is, \( \text{Diag}[E(\varepsilon_i^2)] \). Since \( E(\varepsilon_i^2) = \sigma^2 h(z'_i\hat{\theta}) \), we must estimate \( \theta \) and \( \sigma^2 \):

1. Use \( \varepsilon_i^2 = (y_i - x'_i\hat{\beta}_{OLS})^2 \) as a proxy for \( \varepsilon_i^2 \) because the squares residuals are consistent estimators of the true squared errors. Estimate \( \theta \) and \( \sigma^2 \) using least squares.

2. Replace \( y_i \) with \( y_i^* = y_i h(z'_i\hat{\theta})^{-1/2} \) and \( x_i \) with \( x_i^* = x_i h(z'_i\hat{\theta})^{-1/2} \). Do least squares using \( y_i^* \) and \( x_i^* \). In doing so, \( \hat{\Sigma} = \text{Diag}[h(z'_i\hat{\theta})] \).

If the variance structure is correctly specified, then \( \hat{\beta}_{FWLS} \) is asymptotically BLUE, and thus has the same asymptotic variance as \( \hat{\beta}_{GLS} \). This assumption requires the functional form of the heteroskedasticity to be correctly specified.

### 3.4 Exercises

The first two exercises are questions from previous exams. Feasible WLS, specifically Breusch-Pagan, has only appeared in True/False questions in the last 5 years, although part of a 2003 long question required listing the steps for constructing an appropriate Feasible WLS estimator. The third exercise is to demonstrate a very appropriate application of WLS and the fourth is to provide some practice with multiplicative models.

#### 3.4.1 2002 Exam, 1B

Note that a version of this question also appeared in the 2005 Exam as question 1B.

**Question:** True/False/Explain. To test for heteroskedastic errors in a linear model, it is useful to regress functions of the absolute values of least-squares residuals (eg. the squared residuals) on functions of the regressors. The R-squared from this second stage regression will be (approximately) distributed as chi-square random variable under the null hypothesis of no heteroskedasticity, with degrees of freedom equal to the number of no-constant functions of the regressors in the second-stage.

**Answer:** False. The statement would be correct if the words "R-squared" were replaced by the words "sample size times R-squared," since, under the null of homoskedasticity, \( R^2 \) tends to zero...
in probability, but $N \times R^2$, the ‘studentized Breusch-Pagan’ test statistic, has a limiting $\chi^2$ distribution under $H_0$, where $r$ is the number of non-constant regressors in the second stage regression.

3.4.2 2004 Exam, 1D

Question: True/False/Explain. In a linear model with an intercept and two nonrandom, nonconstant regressors, and with sample size $N = 200$, it is suspected that a ‘random coefficients’ model applies, i.e., that the intercept term and two slope coefficients are jointly random across individuals, independent of the regressors. If the squared values of the LS residuals from this model are themselves fit to a quadratic function of the regressors, and if the $R^2$ from this second-step regression equals 0.06, the null hypothesis of no heteroskedasticity should be rejected at an approximate 5-percent level.

Answer: True. The Breusch-Pagan test statistic for the null homoskedasticity is $NR^2 = 200 \times 0.06 = 12$ for these data; the second-step regresses the squared LS residuals on a constant term and five explanatory variables for the ‘random coefficients’ alternative, specifically, $x_1, x_2, x_1^2, x_2^2$, and $x_1x_2$, where $x_1$ and $x_2$ are the non-constant regressors in the original LS regression. Since the upper 5-percent critical value for a $\chi^2$ random variable with 5 degrees of freedom is 11.07, the null hypothesis of no homoskedasticity should be rejected.

3.4.3 Grouped-Data Regression Model

Question: True/False/Explain. Suppose we are interested in estimating a linear model, $y_{ij} = x_{ij}'\beta + \varepsilon_{ij}$, that satisfies the classical linear assumptions, including a scalar variance-covariance matrix. However, we only have access to data that is aggregated as averages for each $j$. Moreover, we know the amount of observations in the original model for each $j$. The WLS squares estimator that is weighted by square root of the number of observations in $j$, $\forall j$ is BLUE.

Answer: True. Suppose $E(\varepsilon_{ij}) = 0$ and $Var(\varepsilon_{ij}) = \sigma^2$. Given our limitation in data, we are analyze the model $\bar{y}_j = \bar{x}_j'\beta + \bar{\varepsilon}_j$. Let $m_j$ be the number of observations in the the original model for each unit $j$. Then $\bar{\varepsilon}_j = m_j^{-1} \sum_{i=1}^{m_j} \varepsilon_{ij}$.

We can show that this transformed model satisfies the Gauss-Markov assumptions:

\[
E(m_j^{1/2} \bar{\varepsilon}_j) = m_j^{1/2} E(\varepsilon_j) \\
= m_j^{1/2} E(m_j^{-1} \sum_{i=1}^{m_j} \varepsilon_{ij}) \\
= m_j^{1/2} \times m_j^{-1} \sum_{i=1}^{m_j} 0 \\
= m_j^{-1/2} \times (m_j \times 0) = 0
\]
\[ \text{Var}(m_j^{1/2} \varepsilon_j) = m_j \text{Var}(\varepsilon_j) \]

\[ = m_j \text{Var}(m_j^{-1} \sum_{i=1}^{m_j} \varepsilon_{ij}) \]

\[ = m_j \ast m_j^{-2} \sum_{i=1}^{m_j} \sigma^2 \]

\[ = m_j^{-1} \ast (m_j \ast \sigma^2) = \sigma^2 \]

As a result, this weighting causes \( \hat{\beta}_{WLS} \) to be BLUE. Note that this model is applicable for any possible aggregator \( j \). For example, \( j \) can be firms in a company, US states, or countries in a cross-country study. However, if the original linear model is not homoskedastic, then we would proceed with Eicker-White standard errors.

### 3.4.4 Multiplicative Model

Question: Suppose that the sample has size \( N=100 \), and the random variables \( y_i \) are independent with \( E(y_i) = \beta x_i \) and \( V(y_i) = \sigma^2 x_i^2 \).

1) Is this a multiplicative model?

Yes. The model is: \( y_i = \beta x_i + \varepsilon_i \). Let \( u_i \sim iid(0, \sigma^2) \). Then \( \text{Var}(y_i) = \text{Var}(\varepsilon_i) = \sigma^2 x_i^2 \) implies that \( \varepsilon_i = x_i u_i \) because \( \text{Var}(x_i u_i) = \sigma^2 x_i^2 \).

2) How could you test for heteroskedasticity in this model?

\[ z_i = x_i^2 \]

\[ h(z_i' \theta) = x_i^2 \theta \]

We use OLS to estimate:

\[ e_i^2 = \sigma^2 + \delta x_i^2 + r_i \]

\[ \Rightarrow e_i^2 = \sigma^2 + \hat{\delta} x_i^2 \Rightarrow R^2 = \frac{(\hat{y}_i - \bar{y}_i)'(\hat{y}_i - \bar{y}_i)}{(y_i - \bar{y}_i)'(y_i - \bar{y}_i)} \]

We reject \( H_0 : \delta = 0 \) if:

\[ 100R^2 > q_{\chi_1^2=0.95} \], where \( q_{\chi_1^2=0.95} \) is the 95th percentile of the \( \chi_1^2 \) distribution.

3) Construct a GLS estimator of \( \beta \).

\[ \hat{\beta}_{FWLS} = (X' \hat{\Sigma}^{-1} X)^{-1} X' \hat{\Sigma}^{-1} y \]

where \( \hat{\Sigma} = \text{Diag}[\hat{\delta} x_i^2] \) and \( \hat{\delta} = (\sum_{i=1}^{n} x_i^4)^{-1} \sum_{i=1}^{n} x_i^2 e_i^2 \) by least squares estimation.
4 Eicker-White Standard Errors

Alternatively, we can use OLS and correct the standard errors. The benefit of this approach is that it does not require any structure on the nature of the heteroskedasticity. In addition, the structure of the heteroskedasticity may not be correctly specified, and a diagnostic test may falsely reject the hypothesis that the errors are homoskedastic. An incorrectly specified structure would cause $\hat{\beta}_{FGLS}$ to not be asymptotically BLUE nor have a consistent variance-covariance estimator. Further, the interpretation of OLS estimates is desirable for policy because of its ceteris paribus nature.

Recall that the problem with $\hat{\beta}_{OLS}$ if $\text{Var}(y|X) = \Sigma = \sigma^2 \Omega$ is that although the estimator is unbiased and consistent, it is not the most efficient. Moreover, its standard errors cannot be consistently estimated because of the difficulty in consistently estimating $\Sigma$.

Specifically, the variance-covariance matrix for $\hat{\beta}_{OLS}$ is $\text{Var}(\hat{\beta}_{OLS}|X) = (X'X)^{-1}X'\Sigma X(X'X)^{-1}$. Nevertheless, White (1980) draws upon the work of Eicker (1967) to show that it is possible to consistently estimate $\text{plim} \left( \sigma^2 \left( \frac{X'\Omega X}{n} \right) \right)$. With pure heteroskedasticity, $\Sigma$ must be a diagonal matrix. Accordingly, White proves that a consistent variance-covariance estimator draws upon the ordinary least squares residuals:

$$\text{Var}(\hat{\beta}_{OLS}|X) = (X'X)^{-1}X'Diag[(y_i - x_i'\hat{\beta}_{OLS})^2]X(X'X)^{-1}$$

That is, White proves that $\hat{\Sigma} = Diag[(y_i - x_i'\hat{\beta}_{OLS})^2]$, a diagonal matrix of the OLS residuals, is not a consistent estimator of $\Sigma$, but $X'Diag[(y_i - x_i'\hat{\beta}_{OLS})^2]X/n$ is a consistent estimator of $\text{plim} \frac{X'\Sigma X}{n}$.

This estimator is known as the heteroskedasticity-consistent covariance matrix estimator, and often includes combinations of the authors’ names. Note that Professor Powell does not prove this result because it is beyond the scope of the course. However, you should understand its purpose, be aware of its advantages and disadvantages, and know how to implement it.

In finite samples, several adjustments based on degrees of freedom have been proposed to help make small sample inference more accurate. Relative to an asymptotically correct $\hat{\beta}_{FGLS}$, hypothesis testing based on the corrected standard errors is likely over stated. If OLS yields highly statistically significant results, however, then we can likely trust inferences based on OLS. If OLS yields results that are economically different from [F]GLS, the probably is likely with another assumption.

5 Structural Approach to Serial Correlation

Serial Correlation means that in the linear model, $y_t = x_t'\beta + \varepsilon_t$, the variance of the errors: $\Omega = E(\varepsilon\varepsilon'|X)$ has non-zero elements off the diagonal. Equivalently, $E(\varepsilon_i\varepsilon_j|X) \neq 0 \ \forall i \neq j$. We consider time series data because we can write a plausible functional form for the relationship between the errors. We usually assume the error terms are weakly stationary, whereby $\text{Var}(y_t) = \sigma_y^2 \ \forall t$, thus returning to homoskedasticity.
As with pure heteroskedasticity, we consider how to construct consistent standard errors if we hypothesize there is serial correlation. We can test for serial correlation and construct a feasible GLS estimator based on the estimation associated with the assumed structural form of the serial correlation. Alternatively, the Newey-West estimator uses OLS and corrects its standard errors so that they are consistent. The former is known as the structural approach because of the assumed structure of the serial correlation, whereas the latter is known as the nonstructural approach because of its applicability for any possible structure of the serial correlation that need not be known.

5.1 First-Order Serial Correlation

Consider the linear model:

\[ y_t = x_t' \beta + \varepsilon_t, \quad t = 1, \ldots, T \]

where \( \text{Cov}(\varepsilon_t, \varepsilon_s) \neq 0 \). Specifically, we consider that the errors follow a weakly stationary AR(1) process:

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

where the \( u_t \) are i.i.d., \( E(u_t) = 0, Var(u_t) = \sigma^2 \), and \( u_t \) are uncorrelated with \( x_t \).

This last assumption eliminates the possibility of having a lagged \( y \) among the regressors.

By stationarity, the variance of each \( \varepsilon_t \) is the same \( \forall t \).

\[
\begin{align*}
Var(\varepsilon_t) &= Var(\rho \varepsilon_{t-1} + u_t) \\
&= \rho^2 Var(\varepsilon_{t-1}) + Var(u_t) + 2 Cov(\varepsilon_{t-1}, u_t) \\
&= \rho^2 Var(\varepsilon_{t}) + \sigma^2 + 0 \\
\Rightarrow Var(\varepsilon_t)(1 - \rho^2) &= \sigma^2 \\
\Rightarrow Var(\varepsilon_t) &= \frac{\sigma^2}{1 - \rho^2}
\end{align*}
\]

Also, note that by recursion, we can repress \( \varepsilon_t \) as

\[
\begin{align*}
\varepsilon_t &= \rho \varepsilon_{t-1} + u_t = \rho(\rho \varepsilon_{t-2} + u_{t-1}) + u_t \\
&= \rho^2 \varepsilon_{t-2} + \rho u_{t-1} + u_t = \rho^2(\rho \varepsilon_{t-3} + u_{t-2}) + \rho u_{t-1} + u_t \\
&= \rho^3 \varepsilon_{t-3} + \rho^2 u_{t-2} + \rho u_{t-1} + u_t \\
& \quad \vdots \\
&= \rho^s \varepsilon_{t-s} + \sum_{i=0}^{s-1} \rho^i u_{t-i}
\end{align*}
\]

Using this result, we can more easily compute the off-diagonal covariances in the variance-covariance matrix:
\[ \text{Cov}(\varepsilon_t, \varepsilon_{t-s}) = \text{Cov}(\rho^s \varepsilon_t, \varepsilon_{t-s}) + \sum_{i=0}^{s-1} \rho^i u_{t-i}, \varepsilon_{t-s}) \]

\[ = \rho^s \text{Cov}(\varepsilon_t, \varepsilon_{t-s}) + \text{Cov}(\sum_{i=0}^{s-1} \rho^i u_{t-i}, \varepsilon_{t-s}) \]

\[ = \rho^s \text{Var}(\varepsilon_{t-s}) + 0 = \rho^s \frac{\sigma^2}{1 - \rho^2} \]

Using these results,

\[ \text{Var}(\varepsilon) = \sigma^2 \Omega = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \cdots & 1 \end{pmatrix}_{T \times T} \]

\[ \frac{1}{1 - \rho^2} \]

We can use the Cholesky decomposition to derive \( \hat{\beta}_{GLS} \). Computing \( \Omega^{-1} \) and using the Cholesky Decomposition where \( \Omega^{-1} = H' H \) yields

\[ H = \begin{pmatrix} \sqrt{1 - \rho^2} & 0 & 0 & 0 & \cdots & 0 \\ -\rho & 1 & 0 & 0 & \cdots & 0 \\ 0 & -\rho & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & -\rho & 1 \end{pmatrix} \]

The transformed model thus uses \( y_t^* = Hy_t \) and \( x_t^* = Hx_t \), which expanded out is:

\[ y_1^* = \sqrt{1 - \rho^2} y_1, \quad x_1^* = \sqrt{1 - \rho^2} x_1 \]

\[ y_t^* = y_t - \rho y_{t-1}, \quad x_t^* = x_t - \rho x_{t-1} \quad \text{for} \ t = 2, \ldots, T \]

Accordingly, except for the first observation, this regression is known as 'generalized difference.'

### 5.2 Testing for Serial Correlation

If \( \rho \neq 0 \) in the AR(1) model, then there is serial correlation. If the null hypothesis: \( H_0 : \rho = 0 \) is rejected, the model reduces to the classical linear regression model. We assume that \( \varepsilon_0 \) equals zero so the sums start in \( t=1 \). This assumption is not necessary, but it helps some of the calculations.

Recall from the time series exercise done in section that an ordinary least squares estimate of \( \rho \) is:
\[ \hat{\rho} = \frac{\sum_{t=1}^{T} \varepsilon_t \varepsilon_{t-1}}{\sum_{t=1}^{T} \varepsilon_{t-1}^2} \]

This estimator can be rewritten to compute its limiting distribution:

\[ \sqrt{T}(\hat{\rho} - \rho) = \sqrt{T} \frac{\frac{1}{T} \sum_{t=1}^{T} \varepsilon_{t-1} u_t}{\frac{1}{T} \sum_{t=1}^{T} \varepsilon_{t-1}^2} \]

Recall the limiting distributions for the numerator and denominator:

\[ \sqrt{T} \frac{1}{T} \sum_{t=1}^{T} \varepsilon_{t-1} u_t \xrightarrow{d} N(0, \frac{\sigma^4}{1 - \rho^2}) \]
\[ \frac{1}{T} \sum_{t=1}^{T} \varepsilon_{t-1}^2 \xrightarrow{p} \frac{\sigma^2}{1 - \rho^2} \]

Thus by Slutsky’s Theorem:

\[ \sqrt{T}(\hat{\rho} - \rho) = \frac{\sqrt{T} \frac{1}{T} \sum_{t=1}^{T} \varepsilon_{t-1} u_t}{\frac{1}{T} \sum_{t=1}^{T} \varepsilon_{t-1}^2} \xrightarrow{d} N \left( 0, \frac{\sigma^4}{\left(\frac{\sigma^2}{1 - \rho^2}\right)^2} \right) = N(0, 1 - \rho^2) \]

The problem with this estimator, however, is that we do not know \( \varepsilon_t \) so we cannot calculate \( \hat{\rho} \). However, we can express the least squares residual, \( e_t \) as:

\[ e_t = \varepsilon_t + x_t' (\beta - \hat{\beta}) \]

Because \( \hat{\beta} \) depends on \( T \), we can rewrite \( e_t \) as \( e_{t,T} \), where \( e_{t,T} \xrightarrow{T \to \infty} p e_t \). As a result, we can use probability theorems to show that \( \frac{\sum_{t=1}^{T} \varepsilon_t \varepsilon_{t+1}}{\sum_{t=1}^{T} \varepsilon_{t-1}^2} - \frac{\sum_{t=1}^{T} \varepsilon_{t-1} \varepsilon_t}{\sum_{t=1}^{T} \varepsilon_{t-1}^2} \xrightarrow{p} 0 \) as \( T \to \infty \).

Accordingly, an asymptotically equivalent estimator based on the least squares residuals is:

\[ \hat{\rho} = \frac{\sum_{t=1}^{T} e_t e_{t-1}}{\sum_{t=1}^{T} e_{t-1}^2} \]

\[ \sqrt{T}(\hat{\rho} - \rho) \xrightarrow{d} N(0, 1 - \rho^2) \]

Under the null hypothesis,

\[ \sqrt{T}\hat{\rho} \xrightarrow{d} N(0, 1) \]

Thus, this test statistic implies rejecting the null hypothesis if \( \sqrt{T}\hat{\rho} \) exceeds the upper \( \alpha \) critical value \( z(\alpha) \) of a standard normal distribution.
### Table 2: Summary of Tests for Serial Correlation

<table>
<thead>
<tr>
<th>Name</th>
<th>Expression</th>
<th>Distribution under the null</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breusch-Godfrey</td>
<td>$T = NR^2$</td>
<td>$\chi^2_p$</td>
<td>Higher serial corr. and lagged dep var</td>
</tr>
<tr>
<td>usual test</td>
<td>$\sqrt{T}\hat{\rho}$</td>
<td>$\mathcal{N}(0, 1)$</td>
<td>also chi-square $T\hat{\rho}^2$</td>
</tr>
<tr>
<td>Durbin-Watson</td>
<td>$DW = \frac{\sum_{t=2}^{T} (\hat{e}<em>t - \hat{e}</em>{t-1})^2}{\sum_{t=1}^{T} \hat{e}_t^2}$</td>
<td>$DW$</td>
<td>normal approximation</td>
</tr>
<tr>
<td>Durbin’s h</td>
<td>$\frac{\sqrt{T}\hat{\rho}}{\sqrt{1-T\cdot [SE(\hat{\beta}_1)]^2}}$</td>
<td>$\mathcal{N}(0, 1)$</td>
<td>Lagged dep. variable</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$T \cdot [SE(\hat{\beta}_1)]^2 &lt; 1$</td>
</tr>
</tbody>
</table>

Other tests exist, and they have specific characteristics that you should study in Professor Powell’s notes. Here is a table that summarizes these tests.

In Table 2 the tests are ranked in decreasing order of generality. For instance, Breusch-Godfrey is general in the sense that we can test serial correlation of order $p$, and the test can be used with lagged dependent variable. The usual test and Durbin Watson allow us to test first order serial correlation, but recall that Durbin Watson has an inconclusive region. The usual test statistic is straight forward, and it can also be used against a two-sided alternative hypothesis whereas DW has exact critical values that depend on $X$. Durbin’s h is useful for testing in the presence of lagged dependent variable. With lagged dependent variables, $\sqrt{T}\hat{\rho}$ has a distribution that is more tightly distributed around zero than a standard normal, thus making it more difficult to reject the null.

### 5.3 Feasible GLS

After determining that there is indeed serial correlation, we can construct a feasible GLS estimator. Professor Powell presented 5 methods of constructing such an estimator that you should known insofar as he they were discussed in lecture:

i) Prais-Winsten

ii) Cochrane-Orcutt

iii) Durbin’s method

iv) Hildreth-Liu

v) MLE
Professor Powell also briefly discussed how to generalize FGLS construction to the case of AR(p) serially correlated errors.

As with heteroskedasticity, if the form of serial correlation is correctly specified, then these approaches give us estimators of $\beta$ and $\rho$ with the same asymptotic properties as $\hat{\beta}_{GLS}$.

5.4 Exercises

As with heteroskedasticity, serial correlation has appeared regularly on exams. However, it has only appeared in the True and False section.

5.4.1 2002 Exam, Question 1C

Note that a nearly identical question appeared in the 2005 Exam.

Question: In the regression model with first-order serially correlated errors and fixed (nonrandom) regressors, $E(y_t) = x_t'\beta$, $\text{Var}(y_t) = \frac{x_t'^2}{1-\rho^2}$, and $\text{Cov}(y_t, y_{t-1}) = \frac{\rho x_t'^2}{1-\rho^2}$. So if the sample correlation of the dependent variable $y_t$ with its lagged value $y_{t-1}$ exceeds $\frac{1.96}{\sqrt{T}}$ in magnitude, we should reject the null hypothesis of no serial correlation, and should either estimate $\beta$ and its asymptotic covariance matrix by FGLS or some other efficient method or replace the usual estimator of the LS covariance matrix by the Newey-West estimator (or some variant of it).

Answer: False. The statement would be correct if the phrase, "...sample correlation of the dependent variable $y_t$ with its lagged value $y_{t-1}$" were replaced with "...sample correlation of the least squares residual $e_t = y_t - x_t'\hat{\beta}_{LS}$ with its lagged value $e_{t-1}$...". While the population autocovariance of $y_t$ is the same as that for the errors $\epsilon_t = y_t - x_t'\beta$ because the regressors are assumed nonrandom, the sample autocovariance of $y_t$ will involve both the sample autocovariance of the residuals $e_t$ and the sample autocovariance of the fitted values $\hat{y} = x_t'\hat{\beta}_{LS}$, which will generally be nonzero, depending upon the particular values of the regressors.

5.4.2 2003 Exam, Question 1B

Question: In the linear model $y_t = x_t'\beta + \epsilon_t$, if the conditional covariances of the errors terms, $\epsilon_t$ have the mixed heteroskedastic/autocorrelated form

$$\text{Cov}(\epsilon_t, \epsilon_s | X) = \rho^{t-s}\sqrt{x_t'\theta} \sqrt{x_s'\theta}$$

(where it is assumed $x_t'\theta > 0$ with probability one), the parameters of the covariance matrix can be estimated in a multi-step procedure, first regressing least-squares residuals $e_t = y_t - x_t'\hat{\beta}_{LS}$ on their lagged values $e_{t-1}$ to estimate $\rho$, then regressing the squared generalized differenced residuals $\hat{u}_t^2$ (where $\hat{u}_t = e_t - \hat{\rho}e_{t-1}$) on $x_t$ to estimate the $\theta$ coefficients.

Answer: False. Assuming $x_t$ is stationary and $E[\epsilon_t | X] = 0$, the probability limit of the LS regression of $e_t$ on $e_{t-1}$ will be
\[ \rho^* = \frac{\text{Cov}(\varepsilon_t, \varepsilon_{t-1})}{\text{Var}(\varepsilon_{t-1})} \]
\[ = \frac{E[\text{Cov}(\varepsilon_t, \varepsilon_{t-1})] + \text{Cov}[E(\varepsilon_t|X), E(\varepsilon_{t-1}|X)]}{E[\text{Var}(\varepsilon_{t-1})] + \text{Var}[E(\varepsilon_t|X)]} \]
\[ = \frac{E[\text{Cov}(\varepsilon_t, \varepsilon_{t-1})]}{E[\text{Var}(\varepsilon_{t-1})]} \]
\[ = \frac{E[\rho \sqrt{x_t'\theta} \sqrt{x_{t-1}'\theta}]}{E[(x_t'\theta)]} \]
\[ \neq \rho \]

in general. Note that the second line uses the conditional variance identity (See Casella and Berger, p. 167). The remaining substitutions use stationary and the expression given in the question about the conditional covariance of the errors.

To make this statement correct, we must reverse the order of autocorrelation and heteroskedasticity corrections. First, since
\[ \text{Cov}(\varepsilon_t, \varepsilon_{t-1}|X) = \rho^* \sqrt{x_t'\theta} \sqrt{x_{t-1}'\theta} = x_t'\theta \]
we could regress \( \varepsilon_t^2 \) on \( x_t \) to estimate \( \theta \) or, since \( \varepsilon_t \) is unobserved, regress \( \varepsilon_t^2 \) on \( x_t \) (à la Breusch-Pagan). Given \( \hat{\theta} \), we can reweight the residuals to form \( \hat{u}_t = \varepsilon_t/\sqrt{x_t'\hat{\theta}} \). Since \( \text{Cov}(u_t, u_{t-1}|X) = \rho \), a least squares regression of \( \hat{u}_t \) on \( \hat{u}_{t-1} \) will consistently estimate \( \rho \) (as long as the least squares residuals \( \varepsilon_t \) are consistent for the true errors \( \varepsilon_t \)).

5.4.3 2004 Exam, Question 1B

Question: In the linear model with a lagged dependent variable, \( y_t = x_t'\beta + \gamma y_{t-1} + \varepsilon_t \), suppose the error terms have first-order serial correlation, i.e., \( \varepsilon_t = \rho_{t-1} + u_t \), where \( u_t \) is an i.i.d. sequence with zero mean, variance \( \sigma^2 \), and is independent of \( x_s \) for all \( t \) and \( s \). For this model, the classical LS estimators will be inconsistent for \( \beta \) and \( \gamma \), but Aitken’s GLS estimator (for a known \( \Omega \) matrix) will consistently estimate these parameters.

Answer: True. While the classical LS estimators of \( \beta \) and \( \gamma \) are indeed inconsistent because of the covariance between \( y_{t-1} \) and \( \varepsilon_t \), the GLS estimator, with the correct value of \( \rho \), will be consistent. Apart from the first observation (which would not make a difference in large samples), the GLS estimator is LS applied to the ’generalized differenced’ regression:

\[ y_t^* = y_t - \rho y_{t-1} \]
\[ = (x_t - \rho x_{t-1})'\beta + \gamma(y_{t-1} - \rho y_{t-2}) + (\varepsilon_t - \rho\varepsilon_{t-1}) \]
\[ = x_t^*\beta + \gamma y_{t-1}^* + u_t \]
But because \( u_t = \varepsilon_t - \rho \varepsilon_{t-1} \) is i.i.d., it will be independent of \( x_t^* \) and \( y_{t-1}^* = y_{t-1} - \rho y_{t-2} \), so 
\[ E[u_t|x_t^*, y_{t-1}^*] = 0, \] as needed for consistency. So the problem with feasible GLS with lagged dependent variables isn’t consistency of the estimators of \( \beta \) and \( \gamma \) with a consistent estimator of \( \rho \), but rather it is the difficulty of getting a consistent estimator of \( \rho \), since the usual least squares residuals involve inconsistent estimators of the regression coefficients.

6 Nonstructural Approach to Serial Correlation

A handful of estimators have been proposed in the style of Eicker White to account for serial correlation. That is, we can use \( \hat{\beta}_{OLS} = (X'X)^{-1}X'y \) and adjust the standard errors to obtain a consistent estimator that accounts for possible serial correlation. Such methods do not require the structure of the serial correlation to be known, and have similar advantages and disadvantages to Eicker-White.

Recall that \( \hat{\beta}_{OLS} \) is inefficient if there is serial correlation, but still consistent and approximately normally distributed with

\[
\sqrt{T}(\hat{\beta}_{LS} - \beta) \overset{d}{\longrightarrow} N(0, D^{-1}VD^{-1})
\]

where

\[
D = \text{plim} \frac{1}{T}X'X, \quad \text{and} \quad V = \text{plim} \frac{1}{T}X'SX
\]

and \( \Sigma = E[\epsilon \epsilon' | X] \). Since we have a consistent estimator of \( D \), say \( \hat{D} = X'X/T \), we just need to get a consistent estimator for \( V \). One popular nonparametric choice is the Newey-West estimator which is consistent:

\[
\hat{V} = \hat{\Gamma}_0 + \sum_{j=1}^{M} (1 - \frac{j}{M})(\hat{\Gamma}_j + \hat{\Gamma}_j')
\]

where \( \hat{\Gamma} = T^{-1} \sum_{t=j+1}^{T} \hat{\epsilon}_t \hat{\epsilon}_{t-j} x_t x_{t-j}' \) and \( M \) is the bandwidth parameter. This parameter is important because we weight down autocovariances near this threshold and we have a p.s.d. matrix \( V \). Some technical requirements are that \( M = M(T) \to \infty \), \( M/T^{1/3} \to 0 \) as \( T \to \infty \). The proof for Newey-West is beyond the scope of the course, and you should be familiar with its existence, purpose, and vaguely its construction.