Chapter 6 outline, Econometrics
Heteroscedasticity and Serial Correlation

Recall that we made 5 assumptions in order to obtain the results of the Gauss-Markov theorem, and 1 assumption so that we can perform statistical testing easily. We will now see what happens when we violate two of those assumptions, numbers 4 and 5. Assumption number 4 was that the error terms have constant variance (are homoscedastic), or $Var(\varepsilon_i) = \sigma_\varepsilon^2$. Assumption number 5 was that the error terms were independent of one another (are not serially correlated), or $E[\varepsilon_i \varepsilon_j] = E[\varepsilon_i] E[\varepsilon_j] = 0$ for all $i \neq j$.

1 Heteroscedasticity

The term heteroscedasticity means that the variance of a variable is not constant. We know that if the error term in our regression model is heteroscedastic that the Gauss-Markov theorem does not hold and that we cannot be certain if our estimators are the best linear unbiased estimators (BLUE). Although there are formal tests for heteroscedasticity outlined below, there are a number of informal tests for heteroscedasticity (and serial correlation as well). These tests involve the "eyeball method". When looking at a plot of your residuals against the predicted value of the dependent variable they should be randomly dispersed around the x-axis. If they are not then this implies you may have a heteroscedasticity problem or a serial correlation problem. If the plot of the residuals is "funnel-shaped" this means that the variance of your error terms is heteroscedastic. By funnel-shaped I mean the residuals can either start out very tightly packed around the x-axis and grow more dispersed as the predicted value of the dependent variable increases OR the residuals may start out widely dispersed around the x-axis and grow more tightly packed as the predicted value of the dependent variable increases.

1.1 The problem

The problem that heteroscedasticity causes for our estimators is inefficiency. This arises because the variance of the slope estimator ($\hat{\beta}$) when the error term is heteroscedastic becomes:

$$Var(\hat{\beta}) = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2 \sigma_i^2}{\left(\sum_{i=1}^{N} (X_i - \bar{X})^2 \right)^2}$$

When the error term is homoscedastic (constant variance) the variance of $\hat{\beta}$ is $\frac{\sigma^2}{\sum_{i=1}^{N} (X_i - \bar{X})^2}$.

1.2 Corrections for heteroscedasticity

As I said earlier in the course, if the Gauss-Markov assumptions are not met we will try to transform the regression model so that they are met. Two methods for obtaining homoscedastic error terms follow below, based on the amount of knowledge one has about the form of the heteroscedasticity.

1.2.1 Known variances

The simplest case to deal with involves "known error variances". In practice this is rarely seen, but it provides a useful first attempt at correcting for heteroscedasticity. We will apply a technique called weighted least squares.

Suppose that all of the Gauss-Markov assumptions are met, except for assumption 4. This means we have a heteroscedastic error term, or $Var(\varepsilon_i) = \sigma_i^2$. Notice that this looks very similar to the condition necessary for the Gauss-Markov theorem to hold, save for one small difference. The variance of the error term is now observation specific, as can be seen by the subscript i. If the variance of the error term was not observation specific we would write $Var(\varepsilon_i) = \sigma^2$, dropping the subscript i and this would imply homoscedasticity. The question is, how can we transform our regression model to obtain homoscedastic error terms?

Recall that our regression model (in observation specific notation) is:

$Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i$

We assume that $\varepsilon_i \tilde{} N(0, \sigma_i^2)$

Our goal now is to rewrite the model in a form such that we have an error term that has constant variance. Suppose we estimate the following model:

$\frac{Y_i}{\sigma_i} = \frac{\beta_1 + \beta_2 X_{2i} + ... + \beta_k X_{ki}}{\sigma_i} + \nu_i$, where $\nu_i = \frac{\varepsilon_i}{\sigma_i}$

Now, what is the $E[\nu_i]$? What is $Var[\nu_i]$?

$E[\nu_i] = E\left[\frac{\varepsilon_i}{\sigma_i}\right] = \frac{1}{\sigma_i} E[\varepsilon_i] = 0$

The above follows because σ_i is a constant, and $E[\varepsilon_i] = 0$.

$Var[\nu_i] = Var\left[\frac{\varepsilon_i}{\sigma_i}\right] = \frac{1}{\sigma_i^2} Var[\varepsilon_i] = \frac{\sigma_i^2}{\sigma_i^2} = 1$

The above follows because σ_i is a constant, and $Var[\varepsilon_i] = \sigma_i^2$

So $\nu_i \tilde{} N(0, 1)$, and since 1 is a number and does not vary our new error term has constant variance. Thus, the assumptions of the Gauss-Markov theorem are met and we now know that our estimators of the β's are the best linear unbiased estimators.

Notice that this approach is very limiting since it assumes that you know the true error variance.

1.2.2 Error variances vary directly with an independent variable

Suppose you do not know the true error variance but you do have reason to believe that the error term varies directly with one of the independent variables. In this particular case it is not very difficult to construct a new regression model that meets the Gauss-Markov theorem assumptions.

Suppose we have the regression model $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + \varepsilon_i$

Also suppose that $\varepsilon_i \tilde{} N(0, C(X_{2i})^2)$, where C is some non-zero, positive constant.

Now, transform the original regression model by dividing through by X_{2i}.

$\frac{Y_i}{X_{2i}} = \frac{\beta_1}{X_{2i}} + \frac{\beta_2 X_{2i}}{X_{2i}} + \frac{\beta_3 X_{3i} + ... + \beta_k X_{ki}}{X_{2i}} + \frac{\varepsilon_i}{X_{2i}}$

Notice what happens. The term that was the intercept is now a variable and the term with X_{2i} in the original model is now the intercept. Rewriting, we have:

$\frac{Y_i}{X_{2i}} = \frac{1}{X_{2i}}\beta_1 + \beta_2 + \frac{\beta_3 X_{3i} + ... + \beta_k X_{ki}}{X_{2i}} + \nu_i$, where $\nu_i = \frac{\varepsilon_i}{X_{2i}}$

We can show that $\nu_i \tilde{} N(0, C)$ using the same process as we did in the section of the known variances. Since our model now meets the assumptions of the Gauss-Markov theorem we know that we have best linear unbiased estimators.

1.3 Tests for heteroscedasticity

We will now discuss two formal tests for heteroscedasticity. I will not go into the intricate details of these tests, but I will try to provide some intuition behind them. For those of you interested in the intricate details, see Breusch and Pagan (1979), "A Simple Test for Heteroscedasticity and Random Coefficient Variation," *Econometrica*, vol. 47, pp. 1287–1294 and White (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, vol. 48, pp. 817–838. I will warn you in advance that these are fairly difficult articles to read, and they require knowledge of matrix algebra (among other things).

1.3.1 Breusch-Pagan test

For the Breusch-Pagan test, begin with the usual regression form, $Y = \beta_1 + \beta_2 X_2 + \varepsilon$.

Now, assume that the error variance, σ^2, is a function of some variables, or $\sigma^2 = f(\gamma + \delta Z)$. We will assume that the error variance is a linear function of an intercept and one or more independent variables, the Z's. The Z could be the independent variable X_2, or the Z could be some other variable or group of variables. One problem with the Breusch-Pagan test is that we have to make a guess as to what the Z variable(s) is based on our knowledge of the problem. Another "problem" with the Breusch-Pagan test is that it relies on a normally distributed error term.

Procedure The procedure for performing the Breusch-Pagan test works as follows.

1. Run the regression $Y = \beta_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$
2. Obtain the residuals, which are the $\hat{\varepsilon}_i$'s, from this regression

3. Calculate the estimated error variance, $\hat{\sigma}_{\varepsilon}^2 = \frac{\sum (\hat{\varepsilon}_i)^2}{N}$

4. Run the regression $\frac{(\hat{\varepsilon}_i)^2}{\hat{\sigma}_{\varepsilon}^2} = \gamma + \delta Z_i + \nu_i$, where ν_i is the error term and γ is the intercept

5. Note the regression sum of squares (RSS) from $\frac{(\hat{\varepsilon}_i)^2}{\hat{\sigma}_{\varepsilon}^2} = \gamma + \delta Z_i + \nu_i$

Hypothesis testing The hypothesis test involves the use of one-half the regression sum of squares from above.

1. H_0 : Homoscedasticity vs. H_A : Heteroscedasticity

2. $\frac{RSS}{2} \sim \chi_p^2$, where p is the number of Z variables included in the regression in step 4 above

3. Reject the null if the test statistic is greater than the critical value

The intuition behind this test is that the larger the regression sum of squares from the regression in step 4 is, the more highly correlated Z is with the error variance, and the less likely the null hypothesis will hold. You should note that if you reject the null hypothesis that you are ONLY rejecting heteroscedasticity for the particular Z variable(s) you have used in step 4. Heteroscedasticity may still be present in your model, you just haven't found the cause of it yet. A suggestion for correcting for heteroscedasticity would be to include the Z variable in the original regression (if it is not already included) or to use weighted least squares, as described in the section above.

1.3.2 The White Test

The White test is very similar to the Breusch-Pagan test, only it does not depend on the assumption of a normally distributed error term.

Procedure

1. Run the regression $Y = \beta_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$

2. Obtain the residuals, which are the $\hat{\varepsilon}_i$'s, from this regression

3. Run the regression $(\hat{\varepsilon}_i)^2 = \gamma + \delta Z_i + \nu_i$, where ν_i is the error term and γ is the intercept

4. Note the R^2 and the number of observations (N) from the regression $(\hat{\varepsilon}_i)^2 = \gamma + \delta Z_i + \nu_i$

As you can see this is very similar to the Breusch-Pagan test. However, with the White test we do NOT use $\frac{(\hat{\varepsilon}_i)^2}{\hat{\sigma}_{\varepsilon}^2}$ as the dependent variable in our second regression, but $(\hat{\varepsilon}_i)^2$.

Hypothesis testing

1. H_0 : Homoscedasticity vs. H_A : Heteroscedasticity
2. $N * R^2 \tilde{} \chi^2_p$, where p is the number of Z variables included in the regression in step 3 above
3. Reject the null if the test statistic is greater than the critical value

Once again, failing to reject the null hypothesis does NOT mean that there is no heteroscedasticity in your regression model. It does mean that any heteroscedasticity that does occur is NOT likely to come from the Z variable(s).

Since both the White and Breusch-Pagan test are so similar, either can usually be applied with the same amount of effort. Therefore, what is most important is the choice of the Z variable(s). White suggests using combinations of Z and Z^2 if Z is suspected to cause the heteroscedasticity or X, Z, and XZ if X and Z are suspected to cause heteroscedasticity jointly.

2 Serial Correlation

Serial correlation occurs when the occurrence of a particular value of a random variable has some effect on the next random variable drawn from the probability distribution. In regression analysis, we are concerned that our error terms may be serially correlated. This would violate assumption 5 of the Gauss-Markov theorem. Recall that we assume that error terms are independent, meaning one error term does not affect the other error terms.

Serial correlation poses a problem for us because it affects the efficiency of our parameter results. This occurs because the standard errors of our estimators are smaller than they should be, which leads us to rejecting the null hypothesis more often than we should (recall that the test statistic for checking $\beta = 0$ is $\frac{\hat{\beta}}{s_{\hat{\beta}}}$, and if $s_{\hat{\beta}}$ is smaller than it really is then we will have a larger test statistic, leading us to reject the null hypothesis more often).

The problem of serial correlation is shown by the following model. Suppose we have our regression equation (I have used t subscripts instead of is because serial correlation often occurs with time series data):

$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + ... + \beta_k X_{kt} + \varepsilon_t$

Now suppose $\varepsilon_t = \rho \varepsilon_{t-1} + \nu_t$, where $0 \leq |\rho| \leq 1$, $\nu_t \tilde{} N(0, \sigma^2_\nu)$ with independent errors and $\varepsilon_t \tilde{} N(0, \sigma^2_\varepsilon)$. Also, assume that ν_t and ε_t are independent, for all t.

The process described above, $\varepsilon_t = \rho \varepsilon_{t-1} + \nu_t$, is known as a first-order autoregressive process. It is abbreviated as an AR(1). It is a first-order autoregressive process because ε_t depends on "only" its lagged value (ε_{t-1}) and a random error term with zero mean (ν_t).

There are three "useful" results we can show with AR(1) processes:

1. The effect of an error term in any given time period is felt in all future periods, with a magnitude that diminishes over time.

2. $Cov(\varepsilon_t, \varepsilon_{t-k}) = \rho^k \sigma_\varepsilon^2$, for all $k \geq 0$

3. $\rho^k = Correlation(\varepsilon_t, \varepsilon_{t-k})$

$Cov(\varepsilon_t, \varepsilon_{t-1}) = E\left[\left(\varepsilon_t - E\left[\varepsilon_t\right]\right)\left(\varepsilon_{t-1} - E\left[\varepsilon_{t-1}\right]\right)\right] = E\left[\varepsilon_t \varepsilon_{t-1}\right]$ since $E\left[\varepsilon_t\right] = E\left[\varepsilon_{t-1}\right] = 0$

$E\left[\varepsilon_t \varepsilon_{t-1}\right] = E\left[\left(\rho\varepsilon_{t-1} + \nu_t\right)\varepsilon_{t-1}\right] = E\left[\rho\left(\varepsilon_{t-1}\right)^2 + \varepsilon_{t-1}\nu_t\right] = E\left[\rho\left(\varepsilon_{t-1}\right)^2\right] + E\left[\varepsilon_{t-1}\nu_t\right]$

Now, the last term, $E\left[\varepsilon_{t-1}\nu_t\right]$, equals zero because ε_{t-1} and ν_t are independent. This leaves us with:

$Cov(\varepsilon_t, \varepsilon_{t-1}) = E\left[\rho\left(\varepsilon_{t-1}\right)^2\right] = \rho E\left[\left(\varepsilon_{t-1}\right)^2\right]$, since ρ is a constant

But we know $E\left[\left(\varepsilon_{t-1}\right)^2\right] = \sigma_\varepsilon^2$, so:

$Cov(\varepsilon_t, \varepsilon_{t-1}) = \rho\sigma_\varepsilon^2$

We can also show:

$Cov(\varepsilon_t, \varepsilon_{t-2}) = \rho^2\sigma_\varepsilon^2$

$Cov(\varepsilon_t, \varepsilon_{t-3}) = \rho^3\sigma_\varepsilon^2$

We can show the first by plugging in for ε_t (we would then have an equation with ε_{t-1} and ε_{t-2}). Then plug in for ε_{t-1} so that we only have an equation with ε_{t-2}. For the second we would need to plug in for ε_{t-2} so that we would only have an equation with ε_{t-3}.

Now, we can show that $Corr(\varepsilon_t, \varepsilon_{t-k}) = \rho^k$.

Recall that $Corr(a,b) = \frac{Cov(a,b)}{\sqrt{Var(a)Var(b)}}$

For any ε_t and ε_{t-k}, the denominator will be σ_ε^2. This is because $Var(\varepsilon_t) = \sigma_\varepsilon^2$ and $Var(\varepsilon_{t-k}) = \sigma_\varepsilon^2$. If we multiply the two together and take the square root, we get $\sqrt{\left(\sigma_\varepsilon^2\right)^2} = \sigma_\varepsilon^2$. Now, the numerator will be $\rho^k\sigma_\varepsilon^2$ based on our formula above for the covariance. This means the correlation will be $\frac{\rho^k\sigma_\varepsilon^2}{\sigma_\varepsilon^2} = \rho^k$.

What does this mean for us? It means that whatever happened at time $t-k$ will have a ρ^k effect on what happens at time t. Notice that this effect is SMALLER than the effect that $t-k$ would have at time $t-1$ (the effect would be ρ^{k-1}) because as we raise ρ to higher powers the number becomes SMALLER (because $0 \leq |\rho| \leq 1$). This shows our first point and the others have been shown throughout.

2.1 Corrections for serial correlation

Since the Gauss-Markov assumptions are not met we need to know how to transform the regression model so that the assumptions are met. The basic process is called generalized differencing, and is described below. All of the processes below involve generalized differencing, and the purpose of the Cochrane-Orcutt and Hildreth-Lu procedures is to show how to estimate ρ when it is unknown.

2.1.1 Generalized differencing

Once again we start by assuming that we know what ρ is, just like we assumed that we knew the form of the heteroscedasticity in the section above. Here is the model we will use:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \varepsilon_t$$

This looks like every other model, only now: $\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t$. What we need to realize is that the model holds FOR ALL time periods. So we have the "same" model for time $t-2, t-1, t, t+1, t+2$ etc. Thus,

$$Y_{t-1} = \beta_1 + \beta_2 X_{2(t-1)} + \beta_3 X_{3(t-1)} + \varepsilon_{t-1}$$

where $\varepsilon_{t-1} = \rho\varepsilon_{t-2} + \nu_{t-1}$.

Now, to perform generalized differencing, we multiply the equation with Y_{t-1} as the dependent variable by ρ. This gives us:

$$\rho Y_{t-1} = \rho\beta_1 + \rho\beta_2 X_{2(t-1)} + \rho\beta_3 X_{3(t-1)} + \rho\varepsilon_{t-1}$$

Now, subtract the equation with ρY_{t-1} as the dependent variable from the equation with Y_t as the dependent variable. This gives us:

$$Y_t - \rho Y_{t-1} = (\beta_1 - \rho\beta_1) + (\beta_2 X_{2t} - \rho\beta_2 X_{2(t-1)}) + (\beta_3 X_{3t} - \rho\beta_3 X_{3(t-1)}) + (\varepsilon_t - \rho\varepsilon_{t-1})$$

We now have a "new" model,

$$Y_t^* = \beta_1(1-\rho) + \beta_2 X_{2t}^* + \beta_3 X_{3t}^* + \mu_t$$

where:

- $Y_t^* = Y_t - \rho Y_{t-1}$
- $X_{2t}^* = X_{2t} - \rho X_{2(t-1)}$
- $X_{3t}^* = X_{3t} - \rho X_{3(t-1)}$
- $\mu_t = \varepsilon_t - \rho\varepsilon_{t-1}$

Our "new" model will satisfy all of the Gauss-Markov assumptions. To see this we need to look at the $Cov(\mu_t, \mu_{t-1})$. If $Cov(\mu_t, \mu_{t-1}) = 0$ then we will not have serial correlation in the model. By definition, $Cov(\mu_t, \mu_{t-1}) = E[(\mu_t - E[\mu_t])(\mu_{t-1} - E[\mu_{t-1}])]$. We need one piece of information first, $E[\mu_t] = 0$ for all t. This is because $\mu_t = \varepsilon_t - \rho\varepsilon_{t-1}$. So $E[\mu_t] = E[\varepsilon_t - \rho\varepsilon_{t-1}] = E[\varepsilon_t] - E[\rho\varepsilon_{t-1}] = 0 - \rho 0 = 0$. Now,

$Cov(\mu_t, \mu_{t-1}) = E[(\mu_t - E[\mu_t])(\mu_{t-1} - E[\mu_{t-1}])]$

Reason: By definition

$E[(\mu_t - E[\mu_t])(\mu_{t-1} - E[\mu_{t-1}])] = E[\mu_t \mu_{t-1}]$

Reason: The result shown above where $E[\mu_t] = 0$.

$E[\mu_t \mu_{t-1}] = E[(\varepsilon_t - \rho\varepsilon_{t-1})(\varepsilon_{t-1} - \rho\varepsilon_{t-2})]$

Reason: By substitution

$E[(\varepsilon_t - \rho\varepsilon_{t-1})(\varepsilon_{t-1} - \rho\varepsilon_{t-2})] = E[\varepsilon_t\varepsilon_{t-1} - \rho(\varepsilon_{t-1})^2 - \rho\varepsilon_t\varepsilon_{t-2} + \rho^2\varepsilon_{t-1}\varepsilon_{t-2}]$

Reason: Expansion of terms

$E[\varepsilon_t\varepsilon_{t-1} - \rho(\varepsilon_{t-1})^2 - \rho\varepsilon_t\varepsilon_{t-2} + \rho^2\varepsilon_{t-1}\varepsilon_{t-2}] = E[\varepsilon_t\varepsilon_{t-1}] - E[\rho(\varepsilon_{t-1})^2] - E[\rho\varepsilon_t\varepsilon_{t-2}] + E[\rho^2\varepsilon_{t-1}\varepsilon_{t-2}]$

Reason: Properties of the expectations operator

$E[\varepsilon_t\varepsilon_{t-1}] - E[\rho(\varepsilon_{t-1})^2] - E[\rho\varepsilon_t\varepsilon_{t-2}] + E[\rho^2\varepsilon_{t-1}\varepsilon_{t-2}] =$

$Cov(\varepsilon_t\varepsilon_{t-1}) - \rho Var(\varepsilon_{t-1}) - \rho Cov(\varepsilon_t\varepsilon_{t-2}) + \rho^2 Cov(\varepsilon_{t-1}\varepsilon_{t-2})$

Reason: Definitions of variance and covariance when random variables have a zero mean.

$Cov(\varepsilon_t\varepsilon_{t-1}) - \rho Var(\varepsilon_{t-1}) - \rho Cov(\varepsilon_t\varepsilon_{t-2}) + \rho^2 Cov(\varepsilon_{t-1}\varepsilon_{t-2}) = \rho\sigma^2 - \rho\sigma^2 - \rho^3\sigma^2 + \rho^3\sigma^2$

Reason: Use the results derived in the section above.

$\rho\sigma^2 - \rho\sigma^2 - \rho^3\sigma^2 + \rho^3\sigma^2 = 0$

Reason: Addition

So our new model has error terms that are NOT serially correlated, and we can now obtain parameter estimates for the β coefficients that are efficient. You should note that if the original model has T observations that the "new" model only has $T-1$ observations. This is because we lose the first observation from the differencing process (since we have nothing to subtract from it). Also, we have assumed that we know ρ. If we do not know ρ we must estimate it. Two procedures for estimating ρ are listed after the small section on first differencing.

First differencing First differencing is a special case of generalized differencing. First differencing sets $\rho = 1$ and then performs the process of generalized differencing. One important note is that a constant term is NOT included in a model that is first differenced. To see this note that the constant term in our generalized differencing model is $\beta_1(1-\rho)$. If $\rho = 1$, then the constant term drops out of the equation. We can calculate the constant term from its formula, $\beta_1 = \bar{Y} - \beta_2\overline{X}_2 - \beta_3\overline{X}_3$. I mention first differencing for two reasons. One is because it is a very common "quick and dirty method" for removing serial correlation. The second is because the case of $\rho = 1$ is the source of many debates in the economics and statistics literatures (it is usually discussed as a "unit root"). The reason is because if $|\rho| < 1$ then we have a well-defined problem. If $|\rho| > 1$ then we have a very badly behaved problem. But if $|\rho| = 1$ then we have a problem that is between well-defined and very badly behaved.

2.1.2 Cochrane-Orcutt procedure

One procedure that we can use to estimate ρ is the Cochrane-Orcutt procedure. The Cochrane-Orcutt procedure involves an iterative method of estimating ρ. The first thing we do is run the original regression model, $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \varepsilon_t$. Notice that this equation has not been transformed by the generalized differencing process. Once we have estimated the regression

model, we then take the residuals of that model (the $\hat{\varepsilon}_t$'s) and estimate the following model:

$\hat{\varepsilon}_t = \rho\widehat{\varepsilon}_{t-1} + \nu_t$

Since the $\hat{\varepsilon}_t$'s are now "data" we can obtain an estimate for ρ using the least squares procedure. Once we obtain an estimate for ρ (I will call it $\hat{\rho}_1$ for the "first estimate of ρ"), we take this estimate of ρ and use it as the value of ρ in our generalized differencing equation.

Now, run the generalized differenced equation. Obtain the residuals from that regression. Run the regression of $\hat{\varepsilon}_t = \rho\widehat{\varepsilon}_{t-1} + \nu_t$ using the residuals from the generalized differencing equation. We will now get a new estimate of ρ, call it $\hat{\rho}_2$. Now take $\hat{\rho}_2$ and use it as the estimate of ρ. Run a second generalized differenced regression using $\hat{\rho}_2$ as the value of ρ. Obtain the residuals from that equation and reestimate ρ. You can continue this process until two successive estimates of ρ "change very little". Generally we call this the tolerance level. If you set your tolerance level to 0.001 this means that your iterative procedure will stop when two successive estimates of ρ are less than 0.001 apart. You then use either estimate of ρ as your "true value" and run the generalized differenced regression using that value.

There is one big problem using the Cochrane-Orcutt method. The value of ρ obtained may NOT be the value of ρ that minimizes the sum of squares. This could occur if your sum of squares function was very badly behaved. In that case, you could actually iterate to a local rather than a global minimum.

2.1.3 Hildreth-Lu procedure

The Hildreth-Lu procedure involves a method of grid searching for the true value of ρ. To perform a grid search the first thing you need to do is set up your grid values. Since ρ has to be between -1 and 1, we know that we can only choose values within that range. Suppose we thought that positive serial correlation is likely. Then we might set up a grid as 0, .1, .2, .3, .4, .5, .6, .7, .8, and .9. This means we will run 10 regressions, each one with a different value of ρ, where the values of ρ are given by our grid choice. We then find the value of ρ that minimizes the sum of squared errors (just use the regression model with the lowest ESS). Suppose it is .3. We can now do another grid search around .3. We could choose .26, .27, .28, .29, .3, .31, .32, .33, and .34 as our new grid. We could then estimate 9 more models using our new grid as our values for ρ. Now choose the model with the lowest ESS from the new models, and use that estimate of ρ as your true ρ. We can make the grid as fine as we like. One problem with this method is that we may again find a local rather than a global minimum. For instance, suppose the true value of ρ was .62. If the model with .3 had a lower ESS than the one with .6 then we may start expanding the grid around .3 rather than .6. Thus we may find that of all the models we have run that .34 has the lowest ESS, but that is because we started expanding the grid around .3 and not around .6.

2.2 Tests for serial correlation

So far we have discussed methods of correcting for serial correlation and estimating ρ. Now we will discuss a statistical test for detecting serial correlation.

2.2.1 Durbin-Watson Test

The most popular test for serial correlation is the Durbin-Watson test. Formally, the null hypothesis is no serial correlation or $H_0 : \rho = 0$. The Durbin-Watson test statistic is provided by the following formula:

$$DW = \frac{\sum_{t=2}^{T} \left(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1}\right)^2}{\sum_{t=1}^{T} \left(\hat{\varepsilon}_t\right)^2}$$

Note that in the numerator we use only $T-1$ terms since we have no residual to subtract the first residual from. Also, the Durbin-Watson statistic will fall between 0 and 4. We can see this by making some approximations in the formula. The approximation we can use is $DW = 2(1 - \hat{\rho})$. Plugging in a value of 1 (the highest value ρ can take) for $\hat{\rho}$ gives us 0. Plugging in a value of -1 (the lowest value ρ can take) for $\hat{\rho}$ gives us 4. This also gives us some insight as to what values of the Durbin-Watson statistic are going to lead to positive serial correlation, no serial correlation, and negative serial correlation. A Durbin-Watson test statistic that is close to 0 tells us that positive serial correlation is present. A Durbin-Watson test statistic that is close to 4 tells us that negative serial correlation is present. A Durbin-Watson test statistic close to 2 tells us that there is no serial correlation.

The Durbin-Watson test is slightly different than the classical statistical tests we have been using so far in the class. The normal, chi-square, t-distribution, and F-distribution are all well-defined statistical distributions. The "Durbin-Watson distribution" (for lack of a better term) involves approximating the critical values. You can find the table in the back of the book for the Durbin-Watson statistic. To use the table you need to note 2 things. You need to know the number of regressors (excluding the constant term, so $k-1$) as well as the degrees of freedom. Also, there are critical values for upper and lower portions of the test. Since the Durbin-Watson test is used to test for serial correlation (and not specifically only positive or negative serial correlation) we need to have critical values that will let us determine if positive or negative serial correlation is present. So, how do we perform the Durbin-Watson test?

The null hypothesis is that there is no serial correlation ($H_0 : \rho = 0$). The alternative hypothesis is $H_A : \rho \neq 0$. We can use the following table to determine if we reject the null hypothesis in favor of positive serial correlation or negative serial correlation.

Value of DW	**Result**
4 - dl $<$ DW $<$ 4	Reject null hypothesis; negative serial correlation
4 - du $<$ DW $<$ 4 - dl	Result indeterminate
du $<$ DW $<$ 4 - du	Fail to reject null hypothesis
dl $<$ DW $<$ du	Result indeterminate
0 $<$ DW $<$ dl	Reject null hypothesis; positive serial correlation

In the table dl stands for the lower value of the Durbin-Watson distribution for the given number of regressors and the given degrees of freedom. The term du stands for the upper value of the Durbin-Watson distribution for the given number of regressors and the given degrees of freedom. Once we find these two numbers from the Durbin-Watson distribution, plug them into the table and you will have ranges which will allow you to accept or reject the null hypothesis.

Two caveats need to be made when using this particular test. One is that the regression model MUST include a constant term. The other is that the regression model CANNOT include lagged dependent variables as independent variables (well, it can, you just won't be very likely to find serial correlation present even if it is present). This means that if your regression model is in the form of $Y_t = \beta_1 + \beta_2 Y_{t-1} + \beta_3 X_{3t} + \varepsilon$ you are very unlikely to find serial correlation present if you calculate the Durbin-Watson statistic.

DO NOT worry about memorizing the table for the exam!!! I will provide the table for you if necessary.