Economics 240B: Econometrics
Recitation Notes

Jeffrey Greenbaum
University of California, Berkeley

This document contains my teaching notes for Graduate Econometrics: Econ 240B. The instructor for course was James Powell. Carolina Caetano also led some of the recitations, and greatly inspired and provided significant input for the content and pedagogy of my recitations.

Econ 240B is the second semester of the core graduate sequence in econometrics at Berkeley. Econ 240A concludes with deriving the Gauss-Markov Theorem, and 240B discusses the implications of relaxing each assumption. Topics include asymptotics, time series, generalized least squares, seemingly unrelated regressions, heteroskedasticity and serial correlation, panel data, and instrumental variables estimation. Additional themes not covered in my sections include maximum likelihood estimation and inferences for nonlinear statistical models as well as generalized method of moments estimation and inference. Specific topics include discrete dependent variables, censoring, and truncation.

The material draws upon Paul Ruud's *An Introduction to Classical Econometric Theory*, and is supplemented with Arthur Goldberger's *A Course in Econometrics* and William Greene's *Econometric Analysis*.

GLS and SUR

Jeffrey Greenbaum

February 16, 2007

Contents

1 Section Preamble

In the next few sections we relax the spherical covariance matrix assumption – $Var(\varepsilon|X) = \sigma^2 I$, or equivalently $Var(y|X) = \sigma^2 I$.

Recall from 240A that this assumption means that the errors are:
1. Homoskedastic – all of the errors have variance σ^2: $Var(\varepsilon_i|x_i) = \sigma^2 \ \forall i$. This property corresponds with equal values along the main diagonal of $Var(\varepsilon|X)$. It is implied when assuming that

the errors are identically distributed with finite second moments. We now allow for heteroskedastic errors whose variances usually vary with the observed regressors: $Var(\varepsilon_i|x_i) = \sigma^2(x_i)$.

2. Not Serially Correlated – none of the factors unobserved to the econometrician are correlated across individuals: $Cov(\varepsilon_i, \varepsilon_j|x_i, x_j) = 0 \; \forall i \neq j$. This property corresponds with the off-diagonal elements of the covariance matrix being zero. It is implied when assuming that the errors are independently distributed.

We now allow the covariance matrix to be of the general form: $Var(y|X) = \Sigma = \sigma^2\Omega$, and require that it retains its statistical properties of being nonsingular, positive definte, and symmetric. We continue to assume that we know all of the elements of Σ, in which we had previously assumed it to be the specific case of $\sigma^2 I$ with σ^2 known and unique. σ^2 is no longer unique but its value does not affect our results.

We retain all of the other classical regression assumptions of linear expectations, nonstochastic regressors, and full rank regressors, and call this model the generalized classical regression model. If the regressors are not nonstochastic then we can obtain equivalent calculations for most of what we do in this part of 240B by conditioning on them. In fact nonstochastic regressors are rare in economics because most empirical work is based on nonexperimental data rather than controlled experiments. For these reasons we will generally work in terms of the conditional.

As usual we ask the two questions related to relaxing an assumption:

1. Where did we use this assumption? What changes without it?

In 240A we used the error vector's covariance matrix to compute $Var(\hat{\beta}_{OLS}|X)$. In proving the Gauss-Markov Theorem, we showed that the spherical covariance matrix assumption makes $\hat{\beta}_{OLS}$ the most efficient estimator of β among the class of linear unbiased estimators. Without this assumption $Var(\hat{\beta}_{OLS}|X)$ can change and $\hat{\beta}_{OLS}$ is no longer always the most efficient linear unbiased estimator. Moreover it is no longer obvious how to consistently estimate $Var(\hat{\beta}_{OLS}|X)$, which is important for statistical inference. $\hat{\beta}_{OLS}$ remains consistent and unbiased however, because these two properties are affected only by the errors' first moment.

2. How can we remedy these problems?

i) OLS. Despite these two concerns we can still proceed with OLS because a series of advances in the 1980s introduced robust estimation procedures that correct the standard errors so that they are estimated consistently. There are different correction procedures based on whether we believe Ω suffers from just heteroskedasticity, or serial correlation as well. What is meant by robust is that these procedures result in consistent estimators without having to make any structurally parametric assumptions, such as the way in which the errors are heteroskedastic by specifying the form of $\sigma^2(x_i)$. We will devote more attention to these robust procedures next week.

Most of the empirical literature proceeds in this direction because we have a reasonable solution for inference, which is the only concrete problem that arises when transitioning to this generalized

framework. The loss of efficiency with OLS and the amount of error introduced by using robust standard errors is negligible in sufficiently large samples. In fact some econometric research has been devoted to adjusting these robust standard errors to improve the accuracy of small sample inference. We prefer to use OLS when we can do so because it is a straightforward estimator to interpret, and in this model $\hat{\beta}_{OLS}$ remains unbiased and consistent.

ii) GLS. The alternative to proceeding with OLS is to compute Aitken's Generalized Least Squares estimator because it is BLUE. Unfortunately we cannot compute $\hat{\beta}_{GLS}$ unless we know all of the elements of Ω because $\hat{\beta}_{GLS}$ is a function of Ω. That is a problem in practice because Ω is based on information about random variables that the econometrician does not observe unlike X or y. Yet if we can estimate Ω consistently then we can use $\hat{\Omega}$ to construct a feasible estimator that is asymptotically equivalent to $\hat{\beta}_{GLS}$. Estimating Ω consisently however, is not simple because it has more elements than data points. We can reduce this dimensionality concern by making assumptions about the structure of Ω, and we will devote the next few sections to this objective.

GLS appears much less frequently in the empirical literature than OLS because we rarely have reason to believe we know Ω. Similarly Feasible GLS (FGLS) is not widely used because the structural assumptions can be difficult to motivate. However when they can be, FGLS tends to be used as an interesting robustness check to OLS.

2 GLS

In this section we derive $\hat{\beta}_{GLS}$ and prove that it is BLUE in the generalized regression model. Recall that we assume to know all of the elements of Σ. We proceed with Ω in our notation to resemble the classical model, which is a special case of the generalized model where $\Omega = I$.

2.1 The GLS Estimator

We derive $\hat{\beta}_{GLS}$ by transforming the generalized classical regression model and computing its least squares estimate. If this transformed model satisifes the Gauss-Markov assumptions then we know that $\hat{\beta}_{GLS}$ is BLUE. Because Ω is positive definite, there exists a nonsingular $\Omega^{1/2}$ such that $\Omega = \Omega^{1/2}\Omega^{1/2'}$, and we can choose $\Omega^{1/2}$ such that $\Omega = \Omega^{1/2'}\Omega^{1/2}$.

In this subsection we transform the generalized regression model by multiplying $y = X\beta + \varepsilon$ through by $\Omega^{-1/2}$, which exists because Ω is nonsingular. We confirm that this model satisfies the classical linear regression assumptions so we can apply the Gauss-Markov Theorem. In the subsequent subsection we show that we make this specific transformation because no other linear unbiased estimator for β can be more efficient.

Accordingly the transformed model is:

$$\Omega^{-1/2}y = \Omega^{-1/2}X\beta + \Omega^{-1/2}\varepsilon$$

Full Rank Regressors -

We still assume that rank$(X) = K$. As Ruud proves on p.855, it follows that rank$(\Omega^{-1/2}X) = K$ because $\Omega^{-1/2}$ is nonsingular.

Nonstochastic Regressors -
We still assume that X is nonstochastic. $\Omega^{-1/2}X$ is nonstochastic because $\Omega^{-1/2}$ is assumed to be known. Note that if we were to relax the nonstochastic assumption that we could condition on either X or $\Omega^{-1/2}X$ because they contain the same information about the design matrix, X.

Linear Expectation -
We still assume that $E(\varepsilon|X) = 0$:

$$\begin{aligned} E(\Omega^{-1/2}\varepsilon|\Omega^{-1/2}X) &= E(\Omega^{-1/2}\varepsilon|X) \\ &= \Omega^{-1/2}E(\varepsilon|X) \\ &= \Omega^{-1/2}0 \\ &= 0 \end{aligned}$$

Spherical Covariance Matrix -
We now allow for a generalized covariance matrix: $Var(\varepsilon|X) = \sigma^2\Omega = \sigma^2\Omega^{1/2}\Omega^{1/2'}$:

$$\begin{aligned} Var(\Omega^{-1/2}\varepsilon|\Omega^{-1/2}X) &= Var(\Omega^{-1/2}\varepsilon|X) \\ &= \Omega^{-1/2}Var(\varepsilon|X)\Omega^{-1/2'} \\ &= \Omega^{-1/2}(\sigma^2\Omega^{1/2}\Omega^{1/2'})\Omega^{-1/2'} \\ &= \sigma^2 I \end{aligned}$$

Therefore the least squares estimate of this model is BLUE by the Gauss-Markov Theorem:

$$\begin{aligned} \hat{\beta}_{GLS} &= ((\Omega^{-1/2}X)'(\Omega^{-1/2}X))^{-1}(\Omega^{-1/2}X)'(\Omega^{-1/2}y) \\ &= (X'\Omega^{-1/2'}\Omega^{-1/2}X)^{-1}X'\Omega^{-1/2'}\Omega^{-1/2}y \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y \end{aligned}$$

Note that $\hat{\beta}_{GLS} = \hat{\beta}_{OLS}$ if $Var(y|X) = \sigma^2 I$ as expected from substitution of I into the model.

2.2 Relative Efficiency

We confirm that no other linear unbiased estimator of β is more efficient than $\hat{\beta}_{GLS}$ in the generalized model. This confirmation validates that the specific transformation we made by multiplying through by $\Omega^{-1/2}$ produces a least squares estimator that is BLUE for this model. The proof is very similar to the proof of the Gauss-Markov Theorem for $\hat{\beta}_{OLS}$.

$\hat{\beta}_{GLS}$ is BLUE for any non-singular Ω if it is relatively efficient to any other linear unbiased estimate of β, which we denote as $\tilde{\beta}$.

Recall that $\hat{\beta}_{GLS}$ is efficient relative to $\tilde{\beta}$ if and only if:

$$Var(\tilde{\beta}|X) - Var(\hat{\beta}_{GLS}|X) \text{ is positive semi-definite}$$

We first confirm that $\hat{\beta}_{GLS}$ is linear in y and is an unbiased estimator of β.

1. Let $A = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}$. $\hat{\beta}_{GLS}$ = Ay is linear in y because A is nonstochastic.
2. $\hat{\beta}_{GLS}$ is unbiased:

$$\begin{aligned} E(\hat{\beta}_{GLS}|X) &= E((X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y|X) \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}E(y|X) \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}X\beta = \beta \end{aligned}$$

$\tilde{\beta}$ is a linear in y and unbiased estimator of β if:

1. $\tilde{\beta} = Ay$ for some KXN nonstochastic matrix A that is not a function of y.
2. $E(\tilde{\beta}|X) = \beta$.

Combining these two statements:

$$\begin{aligned} E(\tilde{\beta}|X) = \beta &\iff E(Ay|X) = \beta \\ &\iff AE(y|X) = \beta \\ &\iff AX\beta = \beta \\ &\iff AX = I \text{ and } X'A' = I' = I \end{aligned}$$

We now take the conditional variance of both estimators to evaluate the relative efficiency claim:

$$\begin{aligned} Var(\hat{\beta}_{GLS}|X) &= Var((X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y|X) \\ &= ((X'\Omega^{-1}X)^{-1}X'\Omega^{-1})Var(y|X)(X'\Omega^{-1}X)^{-1}X'\Omega^{-1})' \\ &= ((X'\Omega^{-1}X)^{-1}X'\Omega^{-1})(\sigma^2\Omega)(\Omega^{-1}X(X'\Omega^{-1}X)^{-1}) \\ &= \sigma^2(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}X(X'\Omega^{-1}X)^{-1} \\ &= \sigma^2(X'\Omega^{-1}X)^{-1} \\ Var(\tilde{\beta}|X) &= Var(Ay|X) = AVar(y|X)A' = \sigma^2 A\Omega A' \end{aligned}$$

We thus want to show whether $\sigma^2(A\Omega A') - \sigma^2(X'\Omega^{-1}X)^{-1}$ is positive semi-definite. $\sigma^2 > 0$ so it is equivalent to factor it out and check whether $A\Omega A' - (X'\Omega^{-1}X)^{-1}$ is positive semi-definite.

We prove that this difference is positive semi-definite by making use of the property:

For any A and B that are invertible, $A - B$ is positive semi-definite if and only if $B^{-1} - A^{-1}$ is positive semi-definite (Amemiya, p. 461, Property 17).

We use this property and check whether $X'\Omega^{-1}X - (A\Omega A')^{-1}$ is positive semi-definite:

$$\begin{aligned} X'\Omega^{-1}X - (A\Omega A')^{-1} &= X'\Omega^{-1/2'}\Omega^{-1/2}X - (A\Omega^{1/2'}\Omega^{1/2}A')^{-1} \\ &= X'\Omega^{-1/2'}\Omega^{-1/2}X - X'A'(A\Omega^{1/2'}\Omega^{1/2}A')^{-1}AX \\ &= X'\Omega^{-1/2'}I\Omega^{-1/2}X - X'\Omega^{-1/2'}\Omega^{1/2}A'(A\Omega^{1/2'}\Omega^{1/2}A')^{-1}A\Omega^{1/2'}\Omega^{-1/2}X \\ &= X'\Omega^{-1/2'}(I - \Omega^{1/2}A'(A\Omega^{1/2'}\Omega^{1/2}A')^{-1}A\Omega^{1/2'})\Omega^{-1/2}X \\ &= Z'(I - W(W'W)^{-1}W')Z \\ &= Z'(I - P)Z \end{aligned}$$

where $Z = \Omega^{-1/2}X$, $W = \Omega^{1/2}A'$, and $I - P$ is the projection onto $Col(\Omega^{1/2}A')^{\perp}$. Recall that we previously derived that $X'A' = I = AX$ as used in the second equality.

Recall that projection matrices are idempotent and symmetric, and the identity minus a projection matrix is also a projection matrix:

$$\begin{aligned} Z'(I - P)Z &= Z'(I - P)(I - P)Z \\ &= Z'(I - P)'(I - P)Z \\ &= ((I - P)Z)'((I - P)Z) \\ &= \|(I - P)Z\| \end{aligned}$$

This norm must have a nonnegative length. Therefore $Z'(I - P)Z$ must be positive semi-definite.

2.3 Exercises

Professor Powell has used versions of questions from Goldberger in previous exams in the True/False section, especially those pertaining to the topics in GLS that we will cover this week and next. The first question in this section comes from Professor Powell's exam in 2004, which is in the spirit of Goldberger 27.1 The second reviews the derivation that $\hat{\beta}_{GLS}$ is BLUE in the generalized model and is meant to be instructive. It is a good example of how intuition can be used answer the question correctly and earn a lot of the credit before doing any of the math. In the third question we derive an asymptotic test statistic in the context of the generalized regression model and FGLS. This question comes from Professor Powell's 2004 exam, and it is not unusual that he asks a question that requires deriving an asymptotic test statistic in the free response part.

2.3.1 2004 Exam, Question 1A

Question: True/False/Explain. If the Generalized Regression models holds – that is, $E(y|X) = X\beta$, $Var(y|X) = \sigma^2\Omega$, and X full rank with probability one – then the covariance matrix between

Aitken's Generalized LS estimator of $\hat{\beta}_{GLS}$ (with known Ω matrix) and the classical LS estimator $\hat{\beta}_{LS}$ is equal to the variance matrix of the LS estimator.

Answer: False.

$$\begin{aligned}
Cov(\hat{\beta}_{GLS}, \hat{\beta}_{LS}|X) &= Cov((X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y, (X'X)^{-1}X'y|X) \\
&= ((X'\Omega^{-1}X)^{-1}X'\Omega^{-1})Cov(y, y|X)((X'X)^{-1}X')' \\
&= ((X'\Omega^{-1}X)^{-1}X'\Omega^{-1})(\sigma^2\Omega)X(X'X)^{-1} \\
&= \sigma^2(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\Omega X(X'X)^{-1} \\
&= \sigma^2(X'\Omega^{-1}X)^{-1}X'X(X'X)^{-1} \\
&= \sigma^2(X'\Omega^{-1}X)^{-1} \\
&= Var(\hat{\beta}_{GLS}|X)
\end{aligned}$$

The correct statement would be that the covariance of the GLS and the LS estimators is equal to the variance of the *GLS* estimator.

2.3.2 Relative Efficiency of GLS to OLS

Question: True/False/Explain. $\hat{\beta}_{GLS}$ is efficient relative to $\hat{\beta}_{OLS}$ in the generalized regression model.

Answer: True. We expect this statement to be true because both are linear unbiased estimators of β and the case in which $\hat{\beta}_{OLS}$ is the most efficient estimator is a special case of the generalized regression model. $\hat{\beta}_{OLS}$ is as efficient as $\hat{\beta}_{GLS}$ in this special case of $\Sigma = \sigma^2 I$ but is less efficient for all other nonsingular, positive definite, symmetric Σ.

As usual we prove this claim by showing that $Var(\hat{\beta}_{OLS}) - Var(\hat{\beta}_{GLS})$ is positive semi-definite.

$$\begin{aligned}
Var(\hat{\beta}_{OLS}|X) &= Var((X'X)^{-1}X'y|X) \\
&= ((X'X)^{-1}X')Var(y|X)((X'X)^{-1}X')' \\
&= \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}
\end{aligned}$$

This question reduces to showing that $\sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1} - \sigma^2(X'\Omega^{-1}X)^{-1}$ is positive semi-definite. σ^2 does not affect the positive semi-definiteness of this difference because it is postive. Accordingly, we use Amemiya (p. 461) and check the positive semi-definiteness of:

$$
\begin{aligned}
&(X'\Omega^{-1}X) - ((X'X)^{-1}(X'\Omega X)(X'X)^{-1})^{-1} \\
&= (X'\Omega^{-1}X) - (X'X)(X'\Omega X)^{-1}(X'X) \\
&= (X'\Omega^{-1/2'}\Omega^{-1/2}X) - (X'\Omega^{-1/2'}\Omega^{1/2}X)(X'\Omega^{1/2'}\Omega^{1/2}X)^{-1}(X'\Omega^{1/2'}\Omega^{-1/2}X) \\
&= X'\Omega^{-1/2'}(I - \Omega^{1/2}X(X'\Omega^{1/2'}\Omega^{1/2}X)^{-1}X'\Omega^{1/2'})\Omega^{-1/2}X \\
&= X'\Omega^{-1/2'}(I - P_{\Omega^{1/2}X})\Omega^{-1/2}X \\
&= \|(I - P_{\Omega^{1/2}X}))\Omega^{-1/2}X\|
\end{aligned}
$$

This expression is positive semi-definite since it is a norm that must have a nonnegative length.

2.3.3 2004 Exam, 2

Question: A feasible GLS fit of the generalized regression model with $K = 3$ regressors yields the estimates $\hat{\beta} = (2, -1, 2)$ where the GLS covariance matrix $V = \sigma^2[X'\Omega^{-1}X]^{-1}$ is estimated as

$$
\hat{V} = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}
$$

using consistent estimators of σ^2 and Ω. The sample size $N = 403$ is large enough so that it is reasonable to assume a normal approximation holds for the GLS estimator.

Use these results to test the null hypothesis $H_0 : \theta = 1$ against a two-sided alternative asymptotic 5% level, where

$$
\theta = g(\beta) = ||\beta|| = (\beta_1^2 + \beta_2^2 + \beta_3^2)^{\frac{1}{2}}
$$

Answer: We reject the null hypothesis by using the delta method to construct an approximate t-statistic.

Recall that $\sqrt{N}(\hat{\beta}_{GLS} - \beta) \longrightarrow_d N(0, V)$ where $V = \sigma^2(X'\Omega^{-1}X)^{-1}$. We are given a $\hat{V}$ such that $\hat{V} \longrightarrow_p V$.

We are interested in the limiting distribution of $\hat{\theta} = g(\hat{\beta})$, which we analyze by the Delta Method: $\sqrt{N}(\hat{\theta} - \theta) \longrightarrow_d N(0, GVG')$ where

$$
\begin{aligned}
G &= \frac{\partial g(\beta)}{\partial \beta'} \\
&= \frac{\partial (\beta_1^2 + \beta_2^2 + \beta_3^2)^{\frac{1}{2}}}{\partial \beta'} \\
&= \frac{1}{(\beta_1^2 + \beta_2^2 + \beta_3^2)^{\frac{1}{2}}} (\beta_1, \beta_2, \beta_3) \\
&= \frac{1}{g(\beta)} (\beta_1, \beta_2, \beta_3)
\end{aligned}
$$

Therefore an approximate test statistic is $\frac{\hat{\theta} - \theta}{\sqrt{GVG'}} \overset{A}{\sim} N(0,1)$.

We estimate G with $\hat{G}$ because $\hat{G} \longrightarrow_p G$ by the Continuous Mapping Theorem where

$$
\begin{aligned}
\hat{G} &= \frac{1}{g(\hat{\beta})} (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) \\
&= \frac{1}{(2^2 + (-1)^2 + (-2)^2)^{\frac{1}{2}}} (2, -1, 2) \\
&= \frac{1}{3} (2, -1, 2)
\end{aligned}
$$

By Slutsky's Theorem $\hat{G}\hat{V}\hat{G}' \longrightarrow_p GVG'$ where

$$
\begin{aligned}
\hat{G}\hat{V}\hat{G}' &= \frac{1}{3} \begin{pmatrix} 2, & -1, & -2 \end{pmatrix} * \begin{pmatrix} 2 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} * \frac{1}{3} \begin{pmatrix} 2 \\ -1 \\ -2 \end{pmatrix} \\
&= \frac{1}{9} \begin{pmatrix} 3, & 1, & -2 \end{pmatrix} \begin{pmatrix} 2 \\ -1 \\ -2 \end{pmatrix} \\
&= 1
\end{aligned}
$$

Thus to test $H_0 : \theta = 1$ against a two-sided alternative, the absolute value of the t-statistic is

$$
\frac{|\hat{\theta} - \theta_0|}{\sqrt{\hat{G}\hat{V}\hat{G}'}} = \frac{|3 - 1|}{1} = 2
$$

which exceeds 1.96, the upper 97.5% critical value of a standard normal. We thus (barely) reject H_0 at an asymptotic 5% level. As is often the case, the sample size $N = 403$ does not directly figure into the solution, though it is implicit in the estimate $\hat{V}$ of the approximate covariance matrix of $\hat{\beta}$.

An alternative solution entails deriving an approximate Wald statistic though it is simpler to compute a t-statstic since there is only one degree of freedom.

3 Robust OLS Estimation

Why don't we always use $\hat{\beta}_{GLS}$, considering that the generalized model is more realistic and that $\hat{\beta}_{GLS} = \hat{\beta}_{OLS}$ in the case that $Var(\varepsilon|X) = \sigma^2 I$? Calculating $\hat{\beta}_{GLS}$ hinges upon knowing all of the elements of Ω, which in practice we do know with certainty because we do not observe ε let alone anything about its second moment. We should still allow for $Var(\varepsilon|X)$ to be nonspherical because this framework is more realistic than the classical regression model, and we could try to compute a feasible GLS estimator by first consistently estimating the elements of Ω using our N data points. However it is difficult to easily obtain a consistent estimate for the $\frac{N(N+1)}{2}$ parameters of Ω because there are more parameters to estimate than data points.

The next few sections present various solutions to this problem depending on what assumptions we are willing to make about Ω. In this section we analyze the properties of $\hat{\beta}_{OLS}$ in this generalized model. Because $\hat{\beta}_{OLS}$ retains some of its properties from the classical regression model, one solution to GLS is to compute $\hat{\beta}_{OLS}$ and correct the aspects that no longer hold in the generalized context.

3.1 OLS Properties

Although $\hat{\beta}_{OLS}$ is no longer efficient, it is still unbiased and consistent because these properties depend on the first moment of ε and the generalized classical regression model relaxes only the second moment assumption.

Accordingly recall the usual calculations from 240A and the asymptotics sections:

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

$\hat{\beta}_{OLS}$ is unbiased because

$$\begin{aligned}E(\hat{\beta}_{OLS}) - \beta &= E((X'X)^{-1}X'\varepsilon|X) \\ &= (X'X)^{-1}X'E(\varepsilon|X) \\ &= (X'X)^{-1}X'0 \\ &= 0\end{aligned}$$

$\hat{\beta}_{OLS}$ is consistent because $\hat{\beta}_{OLS} - \beta = \left(\frac{(X'X)^{-1}}{n}\right)\left(\frac{X'\varepsilon}{n}\right)$ where $\frac{(X'X)^{-1}}{n} \longrightarrow_p E(X'X)^{-1}$ and $\frac{X'\varepsilon}{n} \longrightarrow_p 0$ by the law of large numbers and $\hat{\beta} - \beta \longrightarrow_p 0$ by Slutsky's Theorem.

$Var(\hat{\beta}_{OLS})$ however is neither unbiased nor consistent because these properties depend on the second moment assumption. We now show how the limiting distribution for $\hat{\beta}_{OLS}$ depends on the second moment assumption:

$$\begin{aligned}\sqrt{n}(\hat{\beta}_{OLS} - \beta\ ) &= \left(\frac{X'X}{n}\right)^{-1}(\sqrt{n})\left(\frac{X'\varepsilon}{n}\right) \\ &\longrightarrow_d N(0, E(X'X)^{-1}Var(X'\varepsilon)E(X'X)^{-1})\end{aligned}$$

In the generalized model,

$$Var(X'\varepsilon) = plim_{n\to\infty}\frac{\sigma^2(X'\Omega X)}{n}$$

Rearranging the limiting distribution expression further yields:

$$\frac{\sqrt{n}(\hat{\beta}_{OLS} - \beta)}{\sqrt{\sigma^2\left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'\Omega X}{n}\right)\left(\frac{X'X}{n}\right)^{-1}}} \longrightarrow_d N(0,1)$$

Thus, a consistent estimator of $Var(\hat{\beta}_{OLS})$ is $\frac{1}{n}\left(\frac{X'X}{n}\right)^{-1}\left(\frac{\sigma^2 X'\Omega X}{n}\right)\left(\frac{X'X}{n}\right)^{-1}$.

$\frac{X'X}{n}^{-1}$ is straightforward to compute, but as previously mentioned we do not know the values of Ω and cannot estimate it consistently without further structural assumptions. Advances in the 1980s however now allow us to consistently estimate this middle term nonparametrically without estimating Ω consistently or making any structural assumptions about it. In these procedures we estimate β with $\hat{\beta}_{OLS}$ and replace our standard errors with a robust estimator. We will return to these procedures next week when we discuss heteroskedasticity and serial correlation in greater detail.

4 Feasible GLS Alternatives: SUR

An alternative to correcting the $\hat{\beta}_{OLS}$ standard errors is to use the unbiased, efficient GLS estimator and to make assumptions to consistently estimate Ω. This approach is possible by arguing that Ω has a specific structure. Often the least squares residuals are used to estimate $\hat{\Omega}$. We then substitute $\hat{\Omega}$ for Ω into $\hat{\beta}_{GLS}$ to compute a feasible estimator for GLS, $\hat{\beta}_{FGLS}$. Because $\hat{\Omega}$ is a consistent estimator of Ω, $\hat{\beta}_{GLS}$ and $\hat{\beta}_{FGLS}$ have the same asymptotic distribution under reasonable regularity conditions that we assume are true in the models we consider in 240B. With this consistent estimator for Ω we thus argue that in sufficiently large samples that $\hat{\beta}_{FGLS}$ has the same properties as $\hat{\beta}_{GLS}$. It is only asymptotically equivalent however if we posed the correct structure on Ω.

The first model that we consider that lends itself to Feasible GLS estimation is Arnold Zellner's Seemingly Unrelated Regressions (SUR) estimator, which he published in 1962.

4.1 Motivation and Examples

SUR is least squares estimation on a system of equations where each individual equation, j, is first stacked by each individual, i, and then by j. The system thus contains at least two distinct dependent variables, and each individual should be represented in each j. The important requirement is that the errors associated with each individual's equations across j are correlated. However, they are not correlated across individuals within equation j.

For example, suppose you would like to study factors associated with better GRE scores. It is conceivable that at least one factor that is unobserved to the econometrician and helps someone do well on the math section also helps for the verbal and writing sections. This factor can be something about test-taking ability. Then the errors in the equation for the math score, the equation for the verbal score, and the equation for the writing score are correlated for an individual because these unobserved factors affect all three equations in the same way for each individual. However after controlling for observable factors such as neighborhood and family income, it is conceivable that unobserved factors are not correlated across individuals for math scores. If there are observed regressors that are important for explaining verbal or writing but not math then this set-up would be an excellent case for SUR.

SUR has not appeared frequently in the empirical literature simply because there are not numerous models that lend itself to estimating j equations, each stacked first by i individuals. When such models arise, it is not always easy to demonstrate that the SUR assumptions are satisfied or that the SUR estimator is more efficient than OLS (which we discuss below). Accordingly SUR is often used as benchmark against OLS or to simply argue that we could proceed with OLS since it would be just as efficient as SUR.

For example, Justin McCrary (2002) responds to Steve Levitt (1997)'s paper about whether there are electoral cycles in police hiring and whether these cycles should instrument for the causal effect of police hiring on different types of crime. Levitt considers various crimes, such as murder, rape, and burglarly for a series of cities over time, and finds police reduce violent crime but have a smaller effect on property crime. McCrary cites Zellner (1962) to argue that SUR would be more appropriate than Levitt's two-step estimation procedure for improving efficiency, but OLS for each crime category equation separately is most appropriate because the model is a special case in which OLS for each category separately is as efficient as GLS to the stacked SUR model.

Orley Ashenfelter has used SUR in a series of papers in which he examines the returns to education in which he has data for multiple members of the same family. For example in his well-known paper with Alan Krueger in 1994 they analyze the returns to education for twins. They use OLS for the complete sample as a baseline estimate and then stack the equations and use SUR. For each twin pair they designate a 1st twin and a 2nd twin and they first stack each returns to education equation across families for each twin number and then by twin number. The assumption is that there are unobserved factors that affect income for both twins in a family but not across families within twin number. They then argue that SUR is more efficient than OLS.

4.2 SUR Model

The SUR model that we analyze is:

$$\begin{aligned} y_{ij} &= x'_{ij}\beta_j + \epsilon_{ij} \qquad i = 1, .., N \quad j = 1, ..., M \\ y_j &= X_j\beta_j + \epsilon_j \end{aligned}$$

where i tracks the individuals in the sample and j tracks the different categories of dependent variables.

y_j is the Nx1 vector obtained by stacking the y_{ij} for a fixed j.
X_j is the NxK_j matrix obtained by stacking the row vectors x'_{ij} for a fixed j and is indexed by K_j, which reflects that we do not need to constrain the model to having the same explanatory variables for each equation j.
It follows that β_j is a K_jx1 vector.

Each equation in terms of j satisfies the assumptions of the classical regression model, and we add one assumption about how the equations are related to each other.
The assumptions of the SUR model are thus:

1) $E(y_j|X_j) = X_j\beta$

2) $V(y_j|X_j) = \sigma_{jj}I_N$

2') $Cov(y_j, y_k|X_j, X_k) = \sigma_{jk}I_N$

3) X_j are nonstochastic and full rank with probability 1

Assumptions 1, 2, and 3 have the same interpretation as the classical regression model. Assumption 2 states that for each category j, the conditional variance of each error is σ_{jj}.
Assumption 2' is the addition. It says that the errors are correlated only within an individual across equations. Across equations the errors for different individuals are not correlated. For categories j and k where $j \neq k$, all individual's error terms have equal correlation of σ_{jk}.

Stacking once more over j yields the general representation of $y = X\beta + \epsilon$.
y is the NMx1 vector obtained by stacking over y_j. X is a NMx$\sum_{j=1}^{M} K_j$ block-diagonal matrix, with each block being a X_j matrix. This representation is necessary so that in the matrix multiplication of $X\beta$ we can back out each equation in terms of j.

$Var(y|X)$ requires use of the Kronecker product representation. Professor Powell provides some detail about the definition and properties of the Kronecker product in his notes.
By assumptions 2 and 2',

$$V(y|X) = \begin{pmatrix} \sigma_{11}I_N & \sigma_{12}I_N & ... & \sigma_{1M}I_N \\ . & . & . & . \\ . & . & . & . \\ \sigma_{M1}I_N & . & . & \sigma_{MM}I_N \end{pmatrix} = \Sigma \otimes I_N$$

Substituting this variance into β_{OLS} and β_{GLS} thus yields:

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

The conditional variances of each estimator are:

$$\begin{aligned}Var(\hat{\beta}_{OLS}|X) &= ((X'X)^{-1}X')Var(y|X)((X'X)^{-1}X')' \\ &= (X'X)^{-1}X'(\Sigma \otimes I_N)X(X'X)^{-1}\end{aligned}$$

$$\begin{aligned}Var(\hat{\beta}_{GLS}|X) &= [(X'(\Sigma \otimes I_N)^{-1}X)^{-1}X'(\Sigma \otimes I_N)^{-1}]Var(y|X)[(X'(\Sigma \otimes I_N)^{-1}X)^{-1}X'(\Sigma \otimes I_N)^{-1}]' \\ &= (X'(\Sigma \otimes I_N)^{-1}X)^{-1}X'(\Sigma \otimes I_N)^{-1}(\Sigma \otimes I_N)(\Sigma \otimes I_N)^{-1}X(X'(\Sigma \otimes I_N)^{-1}X)^{-1} \\ &= (X'(\Sigma \otimes I_N)^{-1}X)^{-1}\end{aligned}$$

Professor Powell derives in his lectures notes two distinct cases in which GLS in the SUR model is equivalent to estimating each dependent variable category separately with OLS:

a) The equations are unrelated (no seemingly): Σ is diagonal because $\sigma_{jk} = 0$ for $j \neq k$.

b) Each equation has the same explanatory variables: $X_j = X_0$ for each j.

Finally as usual we rarely know Ω, but now we can consistently estimate it. Professor Powell's notes discuss a feasible estimator based on residuals that is biased but consistent. Under reasonable regularity conditions, using these estimates yields an estimator that is asymptotically equivalent to $\hat{\beta}_{GLS}$, that with a sufficiently large sample is unbiased, consistent, and has a consistent covariance matrix. These results hinge upon the SUR assumptions being correct.

4.3 Exercises

A version of Goldberger 30.1 appeared in both the 2002 and 2005 exams. A version of Goldberger 30.2 appeared in 2003. This section thus presents solutions to 30.1, 30.2, and 30.3 in Goldberger.

4.3.1 Goldberger 30.1

Question: True or False? In the SUR model, if the explanatory variables in the two equations are identical, then the LS residuals from the two equations are uncorrelated with each other.

Answer: The statement is false unless $\sigma_{12} = 0$, thereby making the equations urelated.

$$\text{Let} \quad \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \quad \text{where} \quad Var(y|X) = \begin{pmatrix} \sigma_{11}I & \sigma_{12}I \\ \sigma_{21}I & \sigma_{22}I \end{pmatrix}$$

Suppose $X_1 = X_2 = X$.

Then using OLS, $\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'y_1 = (X'X)^{-1}X'y_1$ and $\hat{\beta}_2 = (X_2'X_2)^{-1}X_2'y_2 = (X'X)^{-1}X'y_2$.

The residual vector from the first equation is $e_1 = y_1 - X_1\hat{\beta}_1 = Iy_1 - X(X'X)^{-1}X'y_1 = (I - P_X)y_1$ where $P_X = X(X'X)^{-1}X'$ is a projection matrix so $(I - P_X)$ is a projection matrix.

Similarly for the second equation, $e_2 = y_2 - X_2\hat{\beta}_2 = Iy_2 - X(X'X)^{-1}X'y_2 = (I - P_X)y_2$.

$$\begin{aligned} Cov(e_1, e_2|X) &= Cov((I - P_X)y_1, (I - P_X)y_2)|X) \\ &= (I - P_X)Cov(y_1, y_2|X)(I - P_X)' \\ &= (I - P_X)\sigma_{12}I(I - P_X) \\ &= \sigma_{12}(I - P_X)(I - P_X) = \sigma_{12}(I - P_X) \neq 0 \end{aligned}$$

4.3.2 Goldberger 30.2

Question: True or False? 1. In the SUR Model, if the explanatory variables in the two equations are orthogonal to each other, then the LS coefficient estimates for the two equations are uncorrelated with each other. 2. The GLS estimate reduces to the LS estimate.

Answer: The first statement is true, the second statement is false.

1. Let $\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$ where $Var(y|X) = \begin{pmatrix} \sigma_{11}I & \sigma_{12}I \\ \sigma_{21}I & \sigma_{22}I \end{pmatrix}$

Using OLS, $\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'y_1$ and $\hat{\beta}_2 = (X_2'X_2)^{-1}X_2'y_2$.

If the explanatory variables in the two equations are orthogonal to each other, then $X_1'X_2 = 0$.

$$\begin{aligned} Cov(\hat{\beta}_1, \hat{\beta}_2|X) &= ((X_1'X_1)^{-1}X_1')Cov(y_1, y_2|X)((X_2'X_2)^{-1}X_2')' \\ &= (X_1'X_1)^{-1}X_1'\sigma_{12}I(X_2(X_2'X_2)^{-1}) \\ &= \sigma_{12}(X_1'X_1)^{-1}X_1'X_2(X_2'X_2)^{-1} \\ &= \sigma_{12}(X_1'X_1)^{-1}(0)(X_2'X_2)^{-1} = 0 \end{aligned}$$

Thus, it is true that the covariance of OLS estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ is zero.

2. (Note Professor Powell added this part to Goldberger 30.2 in the 2003 exam.)

$$
\begin{aligned}
\hat{\beta}_{GLS} &= \left(\begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix}' \begin{pmatrix} \sigma_{11}I & \sigma_{12}I \\ \sigma_{21}I & \sigma_{22}I \end{pmatrix} \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \right)^{-1} \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix}' \begin{pmatrix} \sigma_{11}I & \sigma_{12}I \\ \sigma_{21}I & \sigma_{22}I \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \\
&= \left(\begin{pmatrix} \sigma_{11}X_1' & \sigma_{12}X_1' \\ \sigma_{12}X_2' & \sigma_{22}X_2' \end{pmatrix} \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \right)^{-1} \begin{pmatrix} \sigma_{11}X_1' & \sigma_{12}X_1' \\ \sigma_{12}X_2' & \sigma_{22}X_2' \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \\
&= \begin{pmatrix} \sigma_{11}X_1'X_1 & \sigma_{12}X_1'X_2 \\ \sigma_{12}X_2'X_1 & \sigma_{22}X_2'X_2 \end{pmatrix}^{-1} \begin{pmatrix} \sigma_{11}X_1'y_1 + \sigma_{12}X_1'y_2 \\ \sigma_{12}X_2'y_1 + \sigma_{22}X_2'y_2 \end{pmatrix} \\
&= \begin{pmatrix} \sigma_{11}X_1'X_1 & 0 \\ 0 & \sigma_{22}X_2'X_2 \end{pmatrix}^{-1} \begin{pmatrix} \sigma_{11}X_1'y_1 + \sigma_{12}X_1'y_2 \\ \sigma_{12}X_2'y_1 + \sigma_{22}X_2'y_2 \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{\sigma_{11}}(X_1'X_1)^{-1} & 0 \\ 0 & \frac{1}{\sigma_{22}}(X_2'X_2)^{-1} \end{pmatrix} \begin{pmatrix} \sigma_{11}X_1'y_1 + \sigma_{12}X_1'y_2 \\ \sigma_{12}X_2'y_1 + \sigma_{22}X_2'y_2 \end{pmatrix} \\
&= \begin{pmatrix} (X_1'X_1)^{-1}X_1'y_1 + \frac{\sigma_{12}}{\sigma_{11}}(X_1'X_1)^{-1}X_1'y_2 \\ \frac{\sigma_{21}}{\sigma_{22}}(X_2'X_2)^{-1}X_2'y_1 + (X_2'X_2)^{-1}X_2'y_2 \end{pmatrix} \\
&\neq \begin{pmatrix} (X_1'X_1)^{-1}X_1'y_1 \\ (X_2'X_2)^{-1}X_2'y_2 \end{pmatrix} \\
&= \hat{\beta}_{OLS}
\end{aligned}
$$

Thus, $\hat{\beta}_{GLS}$ does not reduce to $\hat{\beta}_{OLS}$ in this case.

4.3.3 Goldberger 30.3

Question: Suppose that $E(y_1) = x_1\beta_1, E(y_2) = x_2\beta_2, V(y_1) = 4I, V(y_2) = 5I$, and $C(y_1, y_2) = 2I$. Here y_1, y_2, x_1, and x_2 are $n \times 1$, with $x_1'x_1 = 5, x_2'x_2 = 6, x_1'x_2 = 3$. Calculate the variances of the OLS and GLS estimators.

Answer:

$$
\text{Let} \quad \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \quad \text{where} \quad Var(y|X) = (\Sigma \otimes I_N) = \begin{pmatrix} 4I & 2I \\ 2I & 5I \end{pmatrix}
$$

OLS Variance -

Recall that $Var(\beta_{OLS}|X) = Var((X'X)^{-1}X'y|X) = (X'X)^{-1}X'(\Sigma \otimes I_N)X(X'X)^{-1}$:

$$
\begin{aligned}
(X'X)^{-1} &= \left(\begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix}' \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \right)^{-1} = \begin{pmatrix} X_1'X_1 & 0 \\ 0 & X_2'X_2 \end{pmatrix}^{-1} \\
&= \begin{pmatrix} 5 & 0 \\ 0 & 6 \end{pmatrix}^{-1} = \begin{pmatrix} 1/5 & 0 \\ 0 & 1/6 \end{pmatrix} \\
X'(\Sigma \otimes I_N)X &= \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix}' \begin{pmatrix} 4I & 2I \\ 2I & 5I \end{pmatrix} \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} = \begin{pmatrix} 4X_1' & 2X_1' \\ 2X_2' & 5X_2' \end{pmatrix} \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \\
&= \begin{pmatrix} 4X_1'X_1 & 2X_1'X_2 \\ 2X_2'X_1 & 5X_2'X_2 \end{pmatrix} = \begin{pmatrix} 20 & 6 \\ 6 & 30 \end{pmatrix} \\
(X'X)^{-1}X'\Sigma X(X'X)^{-1} &= \begin{pmatrix} 1/5 & 0 \\ 0 & 1/6 \end{pmatrix} \begin{pmatrix} 20 & 6 \\ 6 & 30 \end{pmatrix} \begin{pmatrix} 1/5 & 0 \\ 0 & 1/6 \end{pmatrix} \\
&= \begin{pmatrix} 4/5 & 1/5 \\ 1/5 & 5/6 \end{pmatrix}
\end{aligned}
$$

GLS Variance -

Recall that $Var(\hat{\beta}_{GLS}|X) = (X'(\Sigma \otimes I_N)^{-1}X)^{-1}$:

$$
\begin{aligned}
(\Sigma \otimes I_N)^{-1} &= \begin{pmatrix} 4I & 2I \\ 2I & 5I \end{pmatrix}^{-1} = \frac{1}{16} \begin{pmatrix} 5I & -2I \\ -2I & 4I \end{pmatrix} \\
(X'(\Sigma \otimes I_N)^{-1}X)^{-1} &= \left[\begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix}' \left(\frac{1}{16} \begin{pmatrix} 5I & -2I \\ -2I & 4I \end{pmatrix} \right) \left(\begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \right] ^{-1} \\
&= \left(\frac{1}{16} \begin{pmatrix} 5X_1'X_1 & -2X_1'X_2 \\ -2X_2'X_1 & 4X_2'X_2 \end{pmatrix} \right)^{-1} = \left(\frac{1}{16} \begin{pmatrix} 25 & -6 \\ -6 & 24 \end{pmatrix} \right)^{-1} = \begin{pmatrix} \frac{32}{47} & \frac{8}{47} \\ \frac{8}{47} & \frac{100}{141} \end{pmatrix}
\end{aligned}
$$

Note that the difference between the OLS and GLS variances is positive definite, which is what we expect in this case since GLS is more efficient.

Heteroskedasticity and Serial Correlation

Jeffrey Greenbaum

February 23, 2007

Contents

1 Section Preamble

This week we continue with the generalized regression model and two cases in which we can construct a feasible estimator that has the same asymptotic properties as $\hat{\beta}_{GLS}$. We also present two robust estimators for the standard errors of $\hat{\beta}_{OLS}$ as alternatives to imposing structure to estimate Ω. The first case is when $Var(\varepsilon|X)$ is purely heteroskedastic, and the second is serial correlation.

Recall the problem with the generalized regression model that the standard errors of $\hat{\beta}_{OLS}$ are no longer consistent. $\hat{\beta}_{GLS}$ is the most efficient linear unbiased estimator of β, but computing it requires knowing $Var(\varepsilon|X) = \Sigma$ though ε is unobserved. A consistent estimator of Σ can produce the feasible estimator, $\hat{\beta}_{FGLS}$, that is asymptotically equivalent to $\hat{\beta}_{GLS}$. However it is difficult to consistently estimate Σ because it has more parameters than data points. We can potentially reduce this dimensionality concern by posing structure on how the elements of Σ are formed such that there are no longer more parameters to estimate than data points.

We saw one such case of FGLS last week with SUR and this week we examine pure heteroskedasticity and serial correlation. The solution for the two are similar. Our approach is to assume a functional form for how the errors are heteroskedastic or serially correlated; estimate this structure using our data; and use this estimate to construct $\hat{\beta}_{FGLS}$. If the correct structure is chosen then this estimator has the same asymptotic properties as $\hat{\beta}_{GLS}$, wherein $\hat{\beta}_{FGLS}$ is asymptotically BLUE with consistently estimated standard errors.

FGLS may exacerbate the problem however if incorrectly applied. Hypothesis testing of our structure where the null is homoskedasticity or zero serial correlation as appropriate to the case could suggest that $\Omega = I$. If so we can use $\hat{\beta}_{GLS}$, which would be equivalent to $\hat{\beta}_{OLS}$. Yet hypothesis testing may spuriously lead to the wrong conclusion. Moreover we may either assume the wrong structure of Σ, or have no intuition about what its structure might be. In any of these situations $\hat{\Sigma}$ might contain more noise than information about Σ and FGLS will likely do worse than OLS.

An alternative approach is to use $\hat{\beta}_{OLS}$ – which remains unbiased and consistent – and to instead use consistently estimated standard errors. Although it is longer BLUE if $Var(y|X) \neq \sigma^2 I$, most empirical papers prefer this method because of these concerns about posing a structure for Σ. In fact many papers automatically compute robust standard erorrs without considering whether $\Omega \neq I$ because doing so does not change $\hat{\beta}_{OLS}$; we do not know ε so it is highly plausible that $\Omega \neq I$; and comparing them to $\hat{\sigma}^2(X'X)^{-1}$ reveals the extent to which $\Omega \neq I$. In large samples the loss of efficiency and amount of error introduced with these standard errors is negligible for hypothesis testing, and adjustments have been proposed for smaller samples. Moreover OLS point estimates are appealing for policy applications because they have a ceteris paribus interpretation.

Although $\hat{\beta}_{OLS}$ and $\hat{\beta}_{GLS}$ are both unbiased estimators of β, point estimates inevitably differ unless $Var(y|X) = \sigma^2 I$. It is not necessary to be concerned with such differences however unless the difference is economically significant, such as a difference in sign while inference on both are highly statistically significant. In this case another classical assumption is likely to be faulty such as the linear expectations assumption, which we will begin to relax next week.

2 Weighted Least Squares

GLS estimation with pure heteroskedasticity is known as weighted least squares. In pure heteroskedasticity we assume zero serial correlation wherein all of the off-diagonal elements of Σ, or equivalently Ω, are zero. If the diagonal elements are equal than $\Omega = I$, and the errors are homoskedastic. In this section we assume to know all of the elements along the main diagonal of Σ. In the next we analyze a more realistic setting in which we do not know the errors but can construct a feasible estimator by estimating a model of how the errors are heteroskedastic. We then return to OLS and consider how to correct the standard errors nonparametrically so they are consistent.

2.1 WLS Estimator

In the case of pure heteroskedasticity $Var(y|X) = \Sigma = Diag[\sigma_i^2]$. Following the derivation of $\hat{\beta}_{GLS}$, $\hat{\beta}_{WLS}$ is BLUE if we use OLS to estimate the generalized linear model that is multiplied through by $\Sigma^{-1/2}$. If we were to additionally assume that the errors are independent and distributed normally then finite sample inference should use $\hat{\beta}_{WLS}$.

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

$$\begin{aligned}
\hat{\beta}_{WLS} &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y \\
&= (X'(Diag[w_i])X)^{-1}X'(Diag[w_i])y \\
&= \left(\sum_{i=1}^{n}\omega_i(x_i x_i')\right)^{-1}\sum_{i=1}^{n}\omega_i x_i y_i
\end{aligned}$$

$\hat{\beta}_{WLS}$ is known as weighted least squares because it is equivalently derived by minimizing the weighted sum of the squared residuals. Specifically each squared residual is multiplied by the inverse of σ_i^2 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 minimizes $(y - X\beta)'\Sigma^{-1}(y - X\beta)$. The weighted least squares interpretation becomes clear when expressing this statement in summation notation since $\Sigma = Diag[w_i]$.

3 Feasible WLS

In practice Σ contains unknown parameters because we do not know ε_i let alone $Var(\varepsilon_i|x_i)$. Instead we construct a feasible weighted least squares estimator, $\hat{\beta}_{FWLS}$, by estimating $Var(y_i) = \sigma_i^2$ and estimating $\hat{\beta}_{WLS}$ with $\hat{\Sigma}$ in place of Σ. As with feasible GLS estimation we exploit that $\hat{\Sigma} \longrightarrow_p \Sigma$ enables $\hat{\beta}_{FWLS}$ to be asymptotically equivalent to $\hat{\beta}_{WLS}$ if the correct structure for the heteroskedasticity function is chosen.

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:

$$u_i = c_i\varepsilon_i$$

where $\varepsilon_i \sim iid(0, \sigma^2)$.
It thus follows that $E(\varepsilon_i^2) = Var(\varepsilon_i) + E(\varepsilon_i)^2 = \sigma^2$.

Furthermore we assume that the function c_i^2 has an underlying linear form:

$$c_i^2 = h(z_i'\theta)$$

where the variables z_i are some observable functions of the regressors, x_i, excluding the constant term. θ is a vector of coefficients to be estimated, whose estimation we will return to when discussing how to construct a feasible estimator. Moreover $h(.) > 0$ so that $Var(y_i|x_i) > 0 \quad \forall i$. It is normalized so that $h(0) = 1$ and $h'(0) \neq 0$. Professor Powell provides examples of such functions in his notes.

Combining these assumptions about the structure of the variance:

$$\begin{aligned} Var(u_i) &= Var(c_i\varepsilon_i) = c_i^2 Var(\varepsilon_i) = h(z_i'\theta)\sigma^2 \\ E(u_i) &= E(c_i\varepsilon_i) = c_i E(\varepsilon_i) = c_i * 0 = 0 \\ \Rightarrow Var(u_i) &= E(u_i^2) \end{aligned}$$

The error in this model, u, is homoskedastic if $Var(u_i)$ is constant $\forall i$, or equivalently if $h(z_i'\theta)$ is constant $\forall i$. By our normalization we know that $h(z_i'\theta)$ is constant $\forall i$ if $z_i'\theta = 0$ because $h(0) = 1$. It is not sensible to expect that $z_i = 0$ so if $\theta = 0$ then $z_i'\theta = 0$. Therefore, if $\theta = 0$ then $Var(u_i) = 1 * \sigma^2 = \sigma^2$ 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 assumed the functional form for h(.) correctly.

Under the null hypothesis where $c_i^2 = 1$, $Var(u_i) = h(z_i'\theta)\sigma^2 = \sigma^2$. In addition

$$\begin{aligned} E(u_i^2) &= Var(u_i) = \sigma^2 = h(z_i'\theta)\sigma^2 \\ E(\varepsilon_i^2) &= \sigma^2 = h(z_i'\theta)\sigma^2 \\ \Rightarrow E(u_i^2) &= E(\varepsilon_i^2) \end{aligned}$$

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 potentially limit functional forms of the heteroskedasticity, but we accept it as a reasonable regularity condition. We thus assume that in the neighborhood near $\theta = 0$, $h(z_i'\theta) = h(0) + h'(0)z_i'\theta = 1 + h'(0)z_i'\theta$.

We now derive a regression function to test our errors for heteroskedasticity:

$$\begin{aligned} E(\varepsilon_i^2) &= \sigma^2 h(z_i'\theta) \\ &= \sigma^2(1 + h'(0)z_i'\theta) \\ &= \sigma^2 + \sigma^2 h'(0)z_i'\theta \end{aligned}$$

Let $\delta = \sigma^2 h'(0)\theta$. Moreover if we include an error, r_i, and assume that $E(r_i|z_i) = 0$ and $Var(r_i|z_i) = \tau$, then this model satisfies the classical regression assumptions. Therefore, we can test the regression:

$$\varepsilon_i^2 = \sigma^2 + z_i'\delta + r_i$$

Since $\theta = 0 \Rightarrow \delta = 0$, we test the null hypothesis that $H_0 : \delta = 0$ in this model. Note that we could use our composite error u_i^2 in place of disturbance ε_i^2 because $E(\varepsilon_i^2) = E(u_i^2)$.

However we cannot estimate this model because we do not observe ϵ_i. We use the results of Breusch and Pagan (1979) to test this model, which is based on the least squares residuals in place of the errors. Although the justification for the method is beyond the scope of the class, Professor Powell expects that you know the steps of the test and that you could apply it to data.

Here is the 3-step procedure from Breusch and Pagan (1979) to test the null hypothesis of homoskedasticity:

1. Compute $\hat{\epsilon_i}^2 = (y_i - x_i'\hat{\beta}_{OLS})^2$ and use it as a proxy for ε_i^2 because the squared residuals are observable and are consistent estimators of the squared errors.

2. Regress $\hat{\epsilon_i}^2$ on 1 and z_i and obtain the usual constant-adjusted $R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2}$ from this squared residual regression.

3. Under the null hypothesis, Breusch and Pagan (1979) prove that the statistic

$$T = NR^2 \longrightarrow_d \chi_p^2$$

where $p = dim(\delta) = dim(z_i)$.

We reject H_0 if T exceeds the upper critical value of a chi-squared variable with p degrees of freedom.

Professor Powell discusses a few other test statistics depending on what assumptions we are willing to make about the data or errors. You are responsible for them insofar as Professor Powell presents them. Here is a summary:

Table 1: Summary of Tests for Heteroskedasticity

Name	Expression	Distribution	Comment
Breusch-Pagan	$T = NR^2$	χ^2_p	$p = dim(z_i)$
F	$\mathcal{F} = \frac{(N-K)R^2}{(1-R^2)p}$	$\mathcal{F}_{(p,N-K)}$	$\mathcal{F} \cong T/p$
Studentized LM	$T' = \frac{RSS}{\hat{\tau}}$	χ^2_p	if ε_i gaussian, $\tau = 2\sigma^4$
Goldfeld-Quandt	s_1^2/s_2^2	$\mathcal{F}_{([N/2]-k,N-[N/2]-k)}$	gaussian ε_i, one-sided

3.3 Feasible Estimator

If we reject the null hypothesis of homoskedasticity, then we must account for heteroskedasticity. To compute $\hat{\beta}_{FWLS}$ we must estimate $\hat{\Sigma} = \text{Diag}[E(\varepsilon_i^2)]$. Since $E(\varepsilon_i^2) = \sigma^2 h(z_i'\theta)$, we must estimate θ and σ^2:

1. Use $\hat{e_i^2} = (y_i - x_i'\hat{\beta}_{OLS})^2$ as a proxy for ε_i^2 because the least squares residuals are consistent estimators of the squared errors. Express the heteroskedasticity in terms of $E(\varepsilon_i^2)$ and estimate θ and σ^2 using least squares with $\hat{e_i^2}$ as the dependent variable. It is often possible to transform the heteroskedasticity function so that the function is linear. Professor Powell provides examples of this step in his notes.

2. Do least squares with $y_i^* = y_i * h(z_i'\hat{\theta})^{-1/2}$ and $x_i^* = x_i * h(z_i'\hat{\theta})^{-1/2}$. Doing so yields $\hat{\beta}_{FWLS}$ where $\hat{\Sigma} = \hat{\sigma^2} Diag[h(z_i'\hat{\theta})]$.

If the variance structure is correctly specified, then $\hat{\beta}_{FWLS}$ is asymptotically equivalent to $\hat{\beta}_{GLS}$. It would thus be asymptotically BLUE with the same asymptotic variance as $\hat{\beta}_{GLS}$. Moreover each estimated variance must be positive or $\hat{\beta}_{FWLS}$ is not well defined.

3.4 Exercises

The first two exercises are questions from previous exams. As with last week's GLS questions, Feasible WLS – specifically Breusch-Pagan – tends to appear in the True/False section. The third exercise is to demonstrate a very appropriate application of WLS that does not require feasible estimation. 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 regrssion will be (approximately) distributed as chi-square random variable under the null hypothesis of no heteroskedastidcity, with

degress of freedome equal to the number of no-constant functions of the regressors in the second-stage.

Answer: False. The statement would be correct if "R-squared" were replaced by the "sample size times R-squared." Under the null of homoskedasticity $R^2 \longrightarrow_p 0$, but as Breusch and Pagan (1979) show $N * R^2 \longrightarrow_d \chi^2_r$ 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 $NR^2 = 200 * 0.06 = 12$ 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. As a result the null hypothesis tests whether 5 parameters equal zero. Since the upper 5-percent critical value for a χ^2 random variable with 5 degrees of freedom is 11.07 is less than our test statistic of 12, we reject the null hypothesis of homoskedasticity.

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 the average for each group 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 to only group averages, 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 for example $\bar{\varepsilon_j} = m_j^{-1} \sum_{i=1}^{m_j} \varepsilon_{ij}$.

We multiply this model by $m_j^{1/2}$ and show it satisfies the Gauss-Markov assumptions:

$$
\begin{aligned}
E(m_j^{1/2}\bar{\varepsilon_j}) &= m_j^{1/2}E(\bar{\varepsilon_j}) \\
&= m_j^{1/2}E(m_j^{-1}\sum_{i=1}^{m_j}\varepsilon_{ij}) \\
&= m_j^{1/2} * m_j^{-1}\sum_{i=1}^{m_j}E(\varepsilon_{ij}) \\
&= m_j^{-1/2} * \sum_{i=1}^{m_j} 0 \\
&= m_j^{-1/2} * (m_j * 0) = 0
\end{aligned}
$$

$$
\begin{aligned}
Var(m_j^{1/2}\bar{\varepsilon_j}) &= m_j Var(\bar{\varepsilon_j}) \\
&= m_j Var(m_j^{-1}\sum_{i=1}^{m_j}\varepsilon_{ij}) \\
&= m_j * m_j^{-2}\sum_{i=1}^{m_j}Var(\varepsilon_{ij}) \\
&= m_j^{-1} * \sum_{i=1}^{m_j}\sigma^2 \\
&= m_j^{-1} * (m_j * \sigma^2) = \sigma^2
\end{aligned}
$$

As a result, this weighting causes $\hat{\beta}_{WLS}$ to be BLUE. Note that this model is applicable for any possible aggregator j, such as individuals in a company's firms, 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=125, and the random variables y_i are independent with $E(y_i) = \beta x_i$ and $V(y_i) = \sigma^2(1 + \beta x_i)^2$.

1) Is this a multiplicative model?

Yes. The model is: $y_i = \beta x_i + \varepsilon_i$ where $\varepsilon_i = u_i(1 + \beta x_i)$ for $u_i \sim iid(0, \sigma^2)$.

This error produces the correct form of heteroskedasticity since $Var(y_i) = Var(\varepsilon_i) = Var(u_i(1+\beta x_i)) = \sigma^2(1 + \beta x_i)^2$. Moreover $E(\varepsilon_i) = 0$.

Let $h(z_i'\theta) = (1 + \theta z_i)^2$ where $\theta = \beta$ and $z_i = x_i$. For this $h(.)$, $h(0) = 1$ and $h'(0) \neq 0$.

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

$E(\epsilon_i^2) = Var(\varepsilon_i)$ so we test the null $H_0 : \delta_1 = \delta_2 = 0$ in the model $\epsilon_i^2 = \sigma^2 + \delta_1 x_i + \delta_2 x_i^2 + r_i$. We assume r_i is homoskedastic and mean zero. We derive this model by expanding $h(.)$ and capturing each coefficient by one parameter. Homoskedasticity corresponds with the parameters of the nonconstant terms being equal to zero, which as expected would be equivalent to $\theta = 0$.

We proxy ϵ_i^2 with $e_i^2 = (y_i - \hat{\beta} x_i)^2$, the squared least squares residuals. We estimate

$$e_i^2 = \sigma^2 + \delta_1 x_i + \delta_2 x_i^2 + r_i$$

We compute the fitted values: $\hat{e_i}^2 = \hat{\sigma}^2 + \hat{\delta_1} x_i + \hat{\delta_2} x_i^2$.

We compute $R^2 = \frac{(\hat{e}-\bar{e})'(\hat{e}-\bar{e})}{(e-\bar{e})'(e-\bar{e})}$.

We reject H_0 if $125R^2 > q_{\chi_2^2=0.95}$ where $q_{\chi_2^2=0.95}$ is the 95th percentile of the χ_2^2 distribution.

3) Construct a GLS estimator of β.

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

where $\hat{\Sigma} = Diag[\hat{\sigma}^2(1 + \hat{\beta}_{OLS} x_i)^2]$ and $\hat{\sigma}^2$ is as previously estimated.

4 Eicker-White Robust Standard Errors

Alternatively we can use $\hat{\beta}_{OLS}$ – which is unbiased and consistent – and correct the standard errors nonparametrically so that they are consistent. 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 covariance estimator. Moreover the interpretation of OLS estimates is desirable for policy because of its ceteris paribus nature.

Specifically, the variance-covariance matrix for $\hat{\beta}_{OLS}$ is $Var(\hat{\beta}_{OLS}|X) = (X'X)^{-1}X'\Sigma X(X'X)^{-1}$. Recall that these standard errors cannot be consistently estimated because of the difficulty in consistently estimating Σ without imposing structure since there are more parameters to estimate than data points. Nevertheless, White (1980) generalizes Eicker (1967) to show that it is possible to consistently estimate $plim\left(\sigma^2\left(\frac{X'\Omega X}{n}\right)\right)$. With pure heteroskedasticity, Σ must be a diagonal matrix. Accordingly White proves that a consistent covariance estimator draws upon the ordinary least squares residuals:

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

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 Σ, but $\frac{X'Diag[(y_i - x_i'\hat{\beta}_{OLS})^2]X}{n}$ is a consistent estimator of $plim\frac{X'\Sigma X}{n}$.

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 and to construct the estimator in Matlab. Note that in Stata one would type ", robust" after the regression.

Although Professor Powell motivates Eicker-White standard errors as a correction to FGLS when the incorrect heteroskedasticity function is assumed, as he acknowledges most researchers go straight to the case of classical least squares estimation since we prefer the interpretation of $\hat{\beta}_{OLS}$ to $\hat{\beta}_{FGLS}$. In finite samples several adjustments based on degrees of freedom have been proposed to help make small sapmle 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 FGLS, there is likely a problem 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: $\Sigma = E(\varepsilon\varepsilon'|X)$ has non-zero elements off the diagonal. In this section we consider time series data because it is plausible to express the relationship between the errors mathematically. We usually assume the error terms are weakly stationary, wherein $Var(y_t) = \sigma_y^2 \quad \forall t$, thus returning to homoskedasticity and the diagonal elements of Σ being σ^2 so that we can factor them out and get a diagonal of ones.

As with pure heteroskedasticity we consider how to construct consistent standard errors if they are serially correlated. Our first approach is to assume a functional form for the serial correlation; estimate it; and test it for serial correlation. If we find evidence of serial correlation then we can use our estimated functional form to construct a feasible GLS estimator. Just as with pure heteroskedasticity, the standard errors will only be consistent if we have assumed the correct functional form of serial correlation. Alternatively we can proceed with OLS and use the nonparametric Newey-West estimator to correct the standard errors so they are consistent.

Although we only discuss serial correlation in time series data in this section and in 240B, cross-sectional data can also have correlated errors. At the least empiricists argue that unobservable factors are correlated within a geographic unit or within a household whenever possible. We account for this correlation by clustering our standard errors. For example, one might argue in Ashenfelter and Krueger (1994)'s returns to education experiment on twins that the unobservable characteristics are correlated within twin pair but not necessarily across twin pair. In an OLS regression that pools all of the twins data together should thus cluster standard errors by twin pair. In Stata, type ", cluster" after the regression; it embeds the robust command. A standard reference is Moulton

(1986, 1990), and one would discuss clustering in an applied econometrics or labor economics class or in public policy/public economics.

5.1 First-Order Serial Correlation

Consider the linear model:

$$y_t = x_t'\beta + \varepsilon_t, \qquad t = 1, ...T$$

where $Cov(\varepsilon_t, \varepsilon_s) \neq 0$. 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 ε_t is the same $\forall t$.

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

By recursion we can repress ε_t as

$$\begin{aligned} \varepsilon_t &= \rho\varepsilon_{t-1} + u_t = \rho(\rho\varepsilon_{t-2} + u_{t-1}) + u_t \\ &= \rho^2\varepsilon_{t-2} + \rho u_{t-1} + u_t = \rho^2(\rho\varepsilon_{t-3} + u_{t-2}) + \rho u_{t-1} + u_t \\ &= \rho^3\varepsilon_{t-3} + \rho^2 u_{t-2} + \rho u_{t-1} + u_t \\ ... &= \rho^s\varepsilon_{t-s} + \sum_{i=0}^{s-1} \rho^i u_{t-i} \end{aligned}$$

We use this result to compute the off-diagonal covariances in the variance-covariance matrix:

$$
\begin{aligned}
Cov(\varepsilon_t, \varepsilon_{t-s}) &= Cov(\rho^s \varepsilon_{t-s} + \sum_{i=0}^{s-1} \rho^i u_{t-i}, \varepsilon_{t-s}) \\
&= \rho^s Cov(\varepsilon_{t-s}, \varepsilon_{t-s}) + Cov(\sum_{i=0}^{s-1} \rho^i u_{t-i}, \varepsilon_{t-s}) \\
&= \rho^s Var(\varepsilon_{t-s}) + 0 \\
&= \rho^s \frac{\sigma^2}{1-\rho^2}
\end{aligned}
$$

Using these results

$$
Var(\varepsilon) = \sigma^2 \Omega = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & ..... & \rho^{T-1} \\ \rho & 1 & \rho & ..... & \rho^{T-2} \\ . & . & . & . & . \\ . & . & . & . & . \\ \rho^{T-1} & \rho^{T-2} & ..... & ... & 1 \end{pmatrix}_{TxT} \frac{1}{1-\rho^2}
$$

We can compute the matrix square root to derive $\hat{\beta}_{GLS}$. Specifically we compute Ω^{-1} and factor it into $\Omega^{-1} = H'H$ where

$$
H = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & 0 & 0 & ..... & 0 \\ -\rho & 1 & 0 & 0 & ..... & 0 \\ 0 & -\rho & 1 & 0 & . & 0 \\ . & . & . & 1 & . & 0 \\ . & . & . & & . & 0 \\ 0 & 0 & ..... & 0 & -\rho & 1 \end{pmatrix}
$$

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

$$
\begin{aligned}
&y_1^* = \sqrt{1-\rho^2} y_1, \quad x_1^* = \sqrt{1-\rho^2} x_1 \\
&y_t^* = y_t - \rho y_{t-1}, \quad x_t^* = x_t - \rho x_{t-1} \qquad \text{for t = 2,...T}
\end{aligned}
$$

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

5.2 Testing for Serial Correlation

If $\rho \neq 0$ in the AR(1) model, then there is serial correlation. If we fail to the null hypothesis: $H_0 : \rho = 0$, the model reduces to the classical regression model. We assume that ε_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 ρ is:

$$\tilde{\rho} = \frac{\sum_{t=1}^{T} \varepsilon_t \varepsilon_{t-1}}{\sum_{t=1}^{T} \varepsilon_{t-1}^2}$$

This estimator can be rewritten to compute its limiting distribution:

$$\sqrt{T}(\tilde{\rho} - \rho) = \frac{\sqrt{T}\frac{1}{T}\sum_{t=1}^{T} \varepsilon_{t-1} u_t}{\frac{1}{T}\sum_{t=1}^{T} \varepsilon_{t-1}^2}$$

Recall the limiting distributions for the numerator and denominator:

$$\sqrt{T}\frac{1}{T}\sum_{t=1}^{T} \varepsilon_{t-1} u_t \longrightarrow_d N(0, \frac{\sigma^4}{1-\rho^2})$$

$$\frac{1}{T}\sum_{t=1}^{T} \varepsilon_{t-1}^2 \longrightarrow_p \frac{\sigma^2}{1-\rho^2}$$

Thus by Slutsky's Theorem:

$$\sqrt{T}(\tilde{\rho} - \rho) = \frac{\sqrt{T}\frac{1}{T}\sum_{t=1}^{T} \varepsilon_{t-1} u_t}{\frac{1}{T}\sum_{t=1}^{T} \varepsilon_{t-1}^2} \longrightarrow_d N\left(0, \frac{\frac{\sigma^4}{1-\rho^2}}{\left(\frac{\sigma^2}{1-\rho^2}\right)^2}\right) = N(0, 1-\rho^2)$$

The problem with this estimator, however, is that we do not know ε_t so we cannot calculate $\tilde{\rho}$. 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 $e_{t,T} \underset{T\to\infty}{\longrightarrow_p} \varepsilon_t$. As a result, we can use probability theorems to show that $\frac{\sum_{t=1}^{T} e_t e_{t-1}}{\sum_{t=1}^{T} e_{t-1}^2} - \frac{\sum_{t=1}^{T} \varepsilon_{t-1}\varepsilon_t}{\sum_{t=1}^{T} \varepsilon_{t-1}^2} \longrightarrow_p 0$ as $T \to \infty$.
Accordingly, an asymptotically equivalent estimator based on the least squares residuals is:

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

$$\sqrt{T}(\hat{\rho} - \rho) \longrightarrow_d N(0, 1-\rho^2)$$

Under the null hypothesis,

$$\sqrt{T}\hat{\rho} \longrightarrow_d N(0,1)$$

Thus, this test statistic implies rejecting the null hypothesis if $\sqrt{T}\hat{\rho}$ exceeds the upper α critical value $z(\alpha)$ of a standard normal distribution.

Table 2: Summary of Tests for Serial Correlation

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

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 β and ρ 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, $E(y_t) = x_t'\beta, Var(y_t) = \frac{\sigma^2}{1-\rho^2}$, and $Cov(y_t, y_{t-1}) = \frac{\rho\sigma^2}{1-\rho^2}$. 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 β 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 autocvoariance of y_t is the same as that for the errors $\varepsilon_t = y_t - x_t'\beta$ because the regressors are assumed nonrandom, the sample autocovaraince 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 $y_t = x_t'\beta + \varepsilon_t$, if the conditional covariances of the errors terms, ε_t have the mixed heteroskedastic/autocorrelated form

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

(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 ρ, 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 θ coefficients.

Answer: False. Assuming x_t is stationary and $E[\varepsilon_t|X] = 0$, the probability limit of the LS regression of e_t on e_{t-1} will be

$$
\begin{aligned}
\rho^* &= \frac{Cov(\varepsilon_t, \varepsilon_{t-1})}{Var(\varepsilon_{t-1})} \\
&= \frac{E[Cov(\varepsilon_t, \varepsilon_{t-1})] + Cov[E(\varepsilon_t|X), E(\varepsilon_{t-1}|X)]}{E[Var(\varepsilon_{t-1})] + Var[E(\varepsilon_t|X)]} \\
&= \frac{E[Cov(\varepsilon_t, \varepsilon_{t-1})]}{E[Var(\varepsilon_{t-1})]} \\
&= \frac{E[\rho\sqrt{(}x_t'\theta)\sqrt{(}x_s'\theta)]}{E[(x_t'\theta)]} \\
&\neq \rho
\end{aligned}
$$

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

$$Cov(\varepsilon_t, \varepsilon_t|X) = \rho^{|t-t|}\sqrt{x_t'\theta}\sqrt{x_t'\theta} = x_t'\theta$$

we could regress ε_t^2 on x_t to estimate θ or, since ε_t is unobserved, regress e_t^2 on x_t ($\grave{a}$ la Breusch-Pagan). Given $\hat{\theta}$, we can reweight the residuals to form $\hat{u}_t = e_t/\sqrt{x_t'\theta}$. Since $Cov(u_t, u_{t-1}|X) = \rho$, a least squares regression of $\hat{u}_t$ on $\hat{u}_{t-1}$ will consistently estimate ρ (as long as the least squares residuals e_t are consistent for the true errors ε_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., $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$, where u_t is an i.i.d. sequence with zero mean, variance σ^2, and is independent of x_s for all t and s. For this model, the classical LS estimators will be inconsistent for β and γ, but Aitken's GLS estimator (for a known Ω matrix) will consistently estimate these parameters.

Answer: True. While the classical LS estimators of β and γ are indeed inconsistent because of the covariance between y_{t-1} and ε_t, the GLS estimator, with the correct value of ρ, 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:

$$
\begin{aligned}
y_t^* &= y_t - \rho y_{t-1} \\
&= (x_t - \rho x_{t-1})'\beta + \gamma(y_{t-1} - \rho y_{t-2}) + (\varepsilon_t - \rho\varepsilon_{t-1}) \\
&= x_t^{*\prime}\beta + \gamma y_{t-1}^* + u_t
\end{aligned}
$$

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 β and γ with a consistent estimator of ρ, but rather it is the difficulty of getting a consistent estimator of ρ, since the usual least squares residuals invovle inconsistent estimators of the regression coefficients

6 Nonstructural Approach to Serial Correlation

A handful of robust 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. The key advantage is that we can use $\hat{\beta}_{OLS}$ and do not need to assume a form for the variance-covariance matrix. However the estimator does not perform very well in small samples, and some macroeconomists prefer to use FGLS in small samples if they have good reason to argue a structural for the standard errors (eg. C. Hsieh and C. Romer, 2006).

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

$$\sqrt{T}(\hat{\beta}_{LS} - \beta) \longrightarrow_d \mathcal{N}(0, D^{-1}VD^{-1})$$

where

$$D = plim\frac{1}{T}X'X, \quad \text{and} \quad V = plim\frac{1}{T}X'\Sigma X$$

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 $\hat{\Gamma} = T^{-1}\sum_{t=j+1}^{T}\hat{e}_t\hat{e}_{t-j}x_t x_{t-j}'$ and M is the *bandwidth parameter*. This parameter is important because we weigh down autocovariances near this threshold and we have a positive semidefinite 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.

Panel Data & Endogenous Regressors

Jeffrey Greenbaum

March 2, 2007

Contents

1 Section Preamble

In this section we complete our discussion of the generalized regression model and GLS estimation for a class of panel data models. We will then relax our last assumption of linear expectations. We first introduce the panel data model, which is when we observe a cross-section in multiple time periods; this cross-section can be individuals, geographic units, or firms. Many empirical microeconomics papers estimate panel data models, and it is an active topic of econometric research. We also study panel data because for random effects models, a class of panel data models, we can construct a feasible GLS estimator that can be asymptotically equivalent to $\hat{\beta}_{GLS}$. The model thus fits well with the theme of relaxing the spherical covariance assumption.

We will then return to the classical regression model and discuss endogenous regressors for the rest of Professor Powell's part of 240B. The final assumption to relax is the linear expectations assumption that $E(y) = X\beta \Rightarrow E(\varepsilon) = 0 \Rightarrow E(\varepsilon|X) = 0$.

This assumption implies that $E(X'\varepsilon) = 0$ by the law of iterated expectations:

$$E(X'\varepsilon) = E(E(X'\varepsilon|X)) = E(X'E(\varepsilon|X) = E(X'0) = 0$$

As a result $E(X'\varepsilon) \neq 0 \Rightarrow E(\varepsilon|X) \neq 0$.

Per usual, we ask the two questions associated with relaxing an assumption:

1. What happens to the classical model if we relax $E(X'\varepsilon) = 0$?

As we will show β is no longer identified because it cannot be written as a function of population moments with sample moment counterparts. Not surprisingly $\hat{\beta}_{OLS}$ is no longer unbiased nor consistent. As with $\hat{\beta}_{OLS}$ in the generalized regression model , an inconsistent estimator is incredibly problematic because we want to get closer to the true parameter if we collect more data. Clive Granger, a Nobel Laureate econometrician, once remarked, "If you can't get it right as n goes to infinity, you should not be in this business."

2. How can we solve this problem?

We need to find an instrumental variable for the regressors that are preventing it from being zero. With a valid instrument then we can identify β and construct an estimator that is unbiased, consistent, and asymptotically normal. We conclude that we have a good instrument, Z, if it is [highly] correlated with the variable it is instrumenting for, X, and is uncorrelated with all remaining unobservable characteristics that affect Y, which are captured by ε. For identification we require that Z contains at least as many variables as we seeks to instrument in X. Moreover our instrumental variable matrix must contain at least as many variables as parameters in our original model so we usually include all of the other exogenous variables from our original model. In some models we can deduce a valid instrument from our data. However in most applications, it is necessary to collect more data about a new variable to argue for the validity of an instrument. As is seen in the empirical literature, an economist must often motivate intuitively that $Cov(Z, \varepsilon) = 0$ by showing that the instrument is not correlated with any of the hypothetical components of the error term. Just like with the nature of hypothesis testing, it may not be possible to prove that an instrument is valid but it is possible to reject the validity of an instrument by arguing that an unobserved variable is correlated with the instrument.

2 Panel Data Models

Panel data models are those in which we have data about a cross-section over a set of time periods. The panel is balanced if there is data for the same cross-section in each time period of the sample.

Although this set-up resembles a SUR model for multiple time periods, we will show that the stacking occurs differently for panel data models.

The general framework for the panel data model is:

$$y_{it} = x'_{it}\beta + \alpha_i + \epsilon_{it}, \qquad i = 1, ..., N; \qquad t = 1, ..., T$$

where we assume $E(\epsilon_{it}|X) = E(\epsilon_{it}) = 0$, $Var(\epsilon_{it}) = \sigma^2_\epsilon$ and $Cov(\epsilon_{it}, \epsilon_{js}) = 0$ if $i \neq j$ and $t \neq s$. i tracks the cross-sectional units, and t tracks time periods.

Stacking observations for each individual over time and then across individuals yields:

$$y = X\beta + D\alpha + \epsilon$$

where y is a NTx1 vector, X is a NTxK matrix, D is a NTxN matrix with T NxN vertically stacked identity matrices. As Professor Powell proved in lecture, X does not include an intercept because if it did, $[X, D]$ would not be full column rank.
α_i is our vector of individual-level fixed effects that capture all time-invariant characteristics for individual i. These characteristics are all characteristics that do not vary over time – both observed and unobserved to the econometrician. By unobserved, we mean that they are unobserved to the econometrician, or in other words, we do not have reliable data to measure these relevant variables. Accordingly we would no longer explicitly control for the observed time-invariant characteristics.

For example Hausman and Taylor (1981) analyze the returns to education with the PSID panel data. We would want to include regressors like schooling and unemployment rate, which are included in the data. We would also like to account for characteristics like charisma, motivation, and IQ, but we do not have measures for such in our data set and are arguably difficult to measure reliably. Assuming that they are time-invariant, then if we include them in our model as individual fixed effects then we should also not include observable time-invariant variables like gender that would be multicolinear with the fixed effects matrix.

Accordingly our error term, ε_{it}, includes all individual-year shocks, in addition to individual-invariant shocks for each year in the absence of time fixed effects. Note that we could include time fixed effects if we believed these were more appropriate for our model; we could also include both individual fixed effects and time fixed effects.

If we were to generalize to a larger panel that say indexes individuals various geographic regions over multiplie time periods we could have 6 different types of fixed effects. The only requirement is that we must leave some shocks in the error term, so including both individual and year fixed effects leaves the individual-time shocks in our model. We choose not to account for these shocks because it is more sensible to motivate the individual or year fixed effects.

2.1 Fixed Effects Model

We allow for an arbitrary relationship between α_i and x_i where $\alpha_i = z_i^{*'}\delta$. z_i^* are the collection of time-invariant variables. We do not necessarily care about δ or in fact know all of the variables that belong in z, but we want our estimator to account for these characteristics; otherwise we would not satisfy the linear expectations assumption. This model is effectively an OLS regression with our controls, x_i and N binary variables – one for each unit of observation that equals 1 if it is the variable for individual i and 0 otherwise.

The *fixed effects* (FE) or *within* (W) or *least squares dummy variable* (DV) estimator for β can be obtained by partitioned regression. We do so because are not directly interested in the effects of the remaining variables but must control for them in our model. In our application, the second set of variables are the fixed effects that are relevant for properly specifying the model but not are directly meaningful because we do not observe any of them.

Accordingly applying the expression of the Frish-Waugh Theorem:

$$\hat{\beta}_{FE} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y}$$

where $\tilde{X} = (I_{NT} - D(D'D)^{-1}D')X$ and $\tilde{y} = (I_{NT} - D(D'D)^{-1}D')y$ which are the residuals of the regression of X on D and Y on D respectively.

Note that $\tilde{X}'$ is,

$$\begin{pmatrix} X_1 - l_T(T^{-1}\sum_{t=1}^{T} x_{1t}) \\ . \\ . \\ X_N - l_T(T^{-1}\sum_{t=1}^{T} x_{Nt}) \end{pmatrix} = \begin{pmatrix} X_1 - l_T x_{1.} \\ . \\ . \\ X_N - l_T x_{N.} \end{pmatrix}$$

Writing these expressions in summation notation yields:

$$\hat{\beta}_{DV} = \hat{\beta}_{FE} = \hat{\beta}_W = [\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - x_{i.})(x_{it} - x_{i.})']^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - x_{i.})(y_{it} - y_{i.})$$

As Professor Powell presented in lecture, these two estimators come from reexpressing our model such that the individual fixed effects drop from the regression. Such estimation is the spirit of our partitioned regression estimator.

Note that the difference-in-differences framework can be viewed as a special case of the fixed effects model. In the baseline case, we have two groups, control and treatment, and two time periods of data, pre-treatment and post-treatment. We allow for there to be individual and time fixed effects. We take first-differences and then run the regression. In doing so, individual fixed effects drop because they are constant for all individuals in both periods. Also with only one control – the presence of being in the treatment group – this variable reduces to 0 for the control and 1 for treatment. The least squares estimator that comes from this framework is the difference between treatment and control of the difference in y over each time period for both groups.

Finally we estimate σ^2 with our usual degrees of freedom adjusted estimate s^2. In doing so we have NT observations and must account for $K + N$ degrees of freedom to represent our K regressors and our fixed effects variables for N units. This estimator is both unbiased and consistent.

2.2 Random Effects Model

The fixed effects model fails to identify any components of β that correspond to regressors that constant over time for a given individual. Moreover Professor Powell presented in class that $\alpha\hat{O}_{LS}$ is not consistent in the panel data model. For this model to yield a consistent estimator, α_i must be uncorrelated with x_{it}. Accordingly we treat the α's as random variables and assume the following in a random effects model:

- $y_{it} = x'_{it}\beta + \alpha_i + \epsilon_i$
- α_i is independent of ϵ_{it}
- α_i is independent of x_{it} and
- $E(\alpha_i) = \alpha$, $Var(\alpha_i) = \sigma^2_\alpha$, $Cov(\alpha_i, \alpha_j) = 0$ if $i \neq j$.

We can then rewrite the model as:

$$\begin{aligned} y_{it} &= x'_{it}\beta + \alpha_i + \epsilon_i \\ &= x'_{it}\beta + \alpha + u_{it} \end{aligned}$$

where $u_{it} = \epsilon_{it} + (\alpha_i - \alpha)$ and $E(u_{it}) = 0$, $Var(u_{it}) = \sigma^2_\epsilon + \sigma^2_\alpha$, $Cov(u_{it}, u_{js}) = 0$ if $i \neq j$, and $Cov(u_{it}, u_{is}) = \sigma^2_\alpha$.

Stacking the model we have,

$$y = X\beta + \alpha l_{NT} + u$$

which produces a non-spherical variance-covariance matrix for each individual:

$$Var(u_i) = \begin{pmatrix} \sigma^2_\epsilon + \sigma^2_\alpha & \sigma^2_\alpha & ... & \sigma^2_\alpha \\ \sigma^2_\alpha & \sigma^2_\epsilon + \sigma^2_\alpha & ... & \sigma^2_\alpha \\ . & . & . & . \\ \sigma^2_\alpha & . & .. & \sigma^2_\epsilon + \sigma^2_\alpha \end{pmatrix}_{T\text{x}T}$$

and

$$Var(u) = \sigma^2_\epsilon I_{NT} + \sigma^2_\alpha (I_N \otimes l_T l'_T)$$

The least squares estimate of the RE model can be found using Frisch-Waugh theorem again:

$$\hat{\beta}_{LS} = (X^{*'}X^*)^{-1}X^{*'}y^*$$

where $X^* = (I_{NT} - l_{NT}(l'_{NT}l_{NT})^{-1}l'_{NT})X$ and $y^* = (I_{NT} - l_{NT}(l'_{NT}l_{NT})^{-1}l'_{NT})y$ which are the residuals of the regression of X on l_{NT} and Y on l_{NT} respectively.

Expanding this estimator gives the the following representation in summation:

$$\hat{\beta}_{LS} = [\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - x_{..})(x_{it} - x_{..})']^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - x_{..})(y_{it} - y_{..})$$

where $x_{..}$ is the *grand mean*, i.e. the average of x_{it} over i and t. This estimator is unbiased and consistent but inefficient though.

We know that GLS is efficient relative OLS. We call it the GLS Random Effects Estimator, which is given by:

$$(\hat{\beta}_{GLS}, \hat{\alpha}_{GLS})' = (Z'\Omega^{-1}(\theta)Z)^{-1}Z'\Omega^{-1}(\theta)y$$

where $X = [l_{NT}X]$, $\Omega(\theta) = I_{NT} + \theta(I_N \otimes l_T l_T')$ and $\theta = \sigma_\alpha^2/\sigma_\epsilon^2$

It can be shown that the GLS or RE estimator is a matrix-weighted average between the *within* and the *between* groups estimators:

$$\hat{\beta}_{RE} = A(w_0)\hat{\beta}_{FE} + [I_K - A(w_0)]\hat{\beta}_B$$

where $\hat{\beta}_B$ is the between estimator that captures variation only between groups since there is none within groups:

$$\hat{\beta}_B = [\sum_{i=1}^{N}(x_{i.} - x_{..})(x_{i.} - x_{..})']^{-1}\sum_{i=1}^{N}(x_{i.} - x_{..})(y_{i.} - y_{..})$$

As $T \longrightarrow \infty$ and N is fixed, it can be proved that $A(w_0) \longrightarrow I_K$, hence FE and RE are asymptotically equivalent. See section 24.9 for more detail.

It should be clear that we have the usual problems with hypothesis testing since in practice we do not observe our error terms, let alone anything about their variances. Fixed effects models can be relaxed so that they are written with variance-covariance matrices that are purely heteroskedastic. In that case, we would want to use heteroskedastic-robust consistent standard errors based on Eicker-White. Similarly if we do not know the elements of the variance-covariance matrix for random effects, then we must construct a feasible estimator; Professor Powell presented a feasible estimator in his lecture.

One final note is that not all models lend themselves to random effects estimation. For example in the Hausman and Taylor returns to education example, education attainment is likely correlated with some of the factors in the fixed effect, such as ability. In that case we fail to satisfy the assumption that α_i is independent of x_{it}.

2.3 2004 Exam, 1C

Professor Powell acknowledges that "this is a tricky problem" and that he initially had an incorrect answer in mind when making up the question.

Question: For a balanced panel data regression model with random individual effects, $y_{it} = x'_{it}\beta + \alpha_i + \varepsilon_{it}$ (where the α_i are independent of ε_{it}, and all error terms have mean zero, constant variance, and are serially independent across i and t), suppose that only the number of time periods T tends to infinity, wile the number of individuals N stays fixed. The the "fixed effect" estimator for β will be consistent as $T \longrightarrow \infty$, but the "random effects" GLS estimator is infeasible, since the joint covariance matrix of the error terms is not consistently estimable.

Answer: False. It is true that the joint covariance matrix of the error terms is not consistently estimable - specifically σ^2_α isn't estimable because there are only N realizations of α_i available in the sample, and N is fixed - but this does not mean GLS is either "infeasible" or inconsistent. It is also true that the "fixed effect" estimator β_{FE} is consistent; as in Ruud's text, the FGLS estimator can be written as a matrix-weighted average of the fixed-effect and "between" estimators, where the latter is inconsistent (being based upon only N time averages). However, inspection of the weight matrices for the FGLS estimator reveals that the weight on the "between" estimator goes to zero, and the corresponding weight on "fixed effects" goes to the identity matrix, as $T \longrightarrow \infty$. Moreover, it can be shown that FGLS and "fixed effects" are asymptotically equivalent,

$$\sqrt{T}(\hat{\beta}_{FGLS} - \hat{\beta}_{RE}) \longrightarrow_p 0$$

under the usual conditions on the regressors, etc. So, at least as $T \longrightarrow \infty$, FGLS behaves just like the fixed effect estimator for β, and is consistent.

2.4 2006 Exam, 1B

Question: Suppose that, for the population of firms in the U.S., the relationship over time between dividends and some observable regressors (which include firm size) follows the assumptions of the Classical Normal Linear Regression model, conditional on the realized values of the regressors. Rather than a random sample of firms over time, though, suppose only that a sample of T average values of dividends and the regressors are available for the Fortune 500 largest firms. Using this sample, the Classical Least Squares estimators of the regression coefficients will be efficient, and F-tests using the usual normal-theory for the linear regression model will have correct size.

Answer: True. Denote our linear model of firms as $y_{it} = x'_{it}\beta + \varepsilon_{it}$ where $\varepsilon_{it} \sim N(0, \sigma^2)$ are i.i.d and independent of x_{it}

However instead of observations for each firm i in each year t we have averages of the 500 firms for each year. We thus estimate $y_{.t} = x'_{.t}\beta + \varepsilon_{.t}$ where $y_{.t} = \frac{1}{500}\sum_{i=1}^{500} y_{it}$ for $t = 1, 2, ..., T$

If this model still satisfies the Gauss-Markov assumptions then the classical least squares estimators of the regression coefficients will be efficient. Moreover if the errors are normally distributed then the F-tests will have the correct size.

$\varepsilon_{.t} = \frac{1}{500}\sum_{i=1}^{500}\varepsilon_{it} \sim \frac{1}{500}N(0,\sigma^2) = N(0,\frac{\sigma^2}{500^2})$ for $t = 1, 2, ..., T$. $\varepsilon_{.t}$ are still independent over t and independent of $x_{.t}$.

As a result this linear model still satisfies the Gauss-Markov assumptions as well as the normality assumption. Therefore, the results of the classical normal regression model are applicable and the statement is true.

3 OLS problems with endogeneity

In the remainder of Professor Powell's part of 240B we analyze endogenous regressors and relax the linear expectations assumption and allow $E(X'\varepsilon) \neq 0$. In the endogenous regressor linear model β is no longer identified. Moreover $\hat{\beta}_{OLS}$ is no longer unbiased nor consistent. We now show these properties:

- Identification:

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

  $E(X'\varepsilon)$ is not a population moment that has a sample counterpart because ε is unobserved. Thus β is no longer identified because it cannot be written as population moments that have sample moment counterparts.

- Bias:

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

  The model implies that $E(\varepsilon|X) \neq 0 \Rightarrow E(\hat{\beta}_{OLS}|X) \neq \beta$. $\hat{\beta}_{OLS}$ is thus biased in this model.

- Inconsistency:

$$\hat{\beta} = \beta + (X'X)^{-1}X'\varepsilon \longrightarrow_p \beta + \left(plim\frac{X'X}{n}\right)^{-1} plim\frac{X'\varepsilon}{n}$$

  Recall that in the classical regression model that $\hat{\beta}_{OLS} \longrightarrow_p \beta$ because $plim\frac{X'\varepsilon}{n} = E(X'\varepsilon) = 0$. Because it is no longer 0, the OLS estimator does not converge to β.

3.1 Motivation and Examples

In social science research one should often be suspicious of whether this assumption is satisfied. We can only be completely certain that we have avoided this problem in a laboratory experiment where the scientist can randomly assign the treatment to a representative sample and then isolate the experiment from all external influences. If the two groups are identical before the treatment then by isolating the experiment we can attribute any difference in the treatment group after the experiment to the treatment.

For instance, a biologist can put two groups of bacteria in the same environment, and change the amount of oxygen in one group. If more bacteria grow in the group with more oxygen, then we can conclude that this additional amount of oxygen causes the differential bacteria growth. If we repeat this experiment many times then we can see if we consistently obtain the same results.

A physician can get really close to experimental results in a double blind experiment. Suppose the doctor randomly assigns pills to some patients and a placebo to others and neither the patients nor the physician know who is taking what. We can thus expect the groups to be the same on average and that the only difference in the medicine intake since there are no external factors affecting either group once the experiment begins. On the other hand one can argue that this study will likely not have a large sample size to assure that randomization created two nearly identical groups. Such experiments can be expensive, and since individuals must choose to participate it is also questionable the extent to which the sample is representative of the target population.

If we were to estimate the treatment effect with OLS in the model:

$$y_i = x_i\beta + \varepsilon_i$$

where x is a binary variable for having taken the medicine instead of the placebo, and y is the result of a certain health exam, then β is the effect of the treatment in the health exam if the experiment is ideally run. Of course if we think that some observable characteristics such as age also affect the outcome, then we should explicitly control for such characteristics and compute our average treatment effect through partitioned regression.

However in the social sciences field experiments are rarely done because they are costly and raise ethical concerns when randomly assigning a resource to human beings. Moreover it is difficult to assure a controlled environment that is free of external influence. For starters, the researcher cannot involve humans in an experiment without them knowing they are receiving a benefit. Known as the Hawthoren Effect, a subject may temporarily modify his behavior in response to a change in the environmental conditions of the experiment. More generally it may be difficult to control for external influences and spillover effects since we are not isolating people as in a science experiment; for example the treatment may improve the treatment group, which might have a positive spillover effect on the control group.

Our experimental model will suffer from endogeneity if after the experiment begins, the control group is affected by anything related to the treatment or the treatment group is affected by anything

other than the experiment. We are not concerned if the outcome for either group is affected by pre-experimental characteristics; we can control for these characteristics and attribute the remainder of the outcome to the treatment in the ideal experiment. A good way to check that our treatment effect is not absorbing the results of a pre-experimental characteristic is to make sure that on the aggregate level pre-experimental characteristics are nearly identical between the two groups. If there are any striking diffrences between the groups then we have not done a good job randomly assigning the treatment since it would be a valid argument that the treatment effect is partially picking up such differences between the two groups.

In the absence of field experiments we look for naturally occuring instances that effectively create randomly assigned treatments. This setting is known as a natural experiment, and the researcher must argue that the treatment is not confounded by unobserved factors. Often such arguments are made in the social sciences in response to policy changes. A classic example is David Card and Alan Krueger's (1994) study of how minimum wage affects employment in the fast-food industry. They compare employment in the industry between New Jersey and Pennsylvania and exploit a minimum wage increase in NJ in 1992 as a natural experiment. They control for a host of observable characteristics, such as the unemployment rate, and argue that the difference between employment between the two states after the legislation can be attributed to the legislation. Meyer (1995) outlines the application of natural experiments in the empirical literature.

Most observational studies however do not lend themselves to a natural experiment that can be analyzed by OLS. For example a classic question in labor economics about the returns to education is difficult to assess because of omitted variables bias and measurement error (Card, 2001). If we regress earnings on education then we can argue that we have an endogenous regressor because ability, an omitted variable because it is unobserved to the econometrician, affects both education and earnings. Moreover it is not unusual that survey respondants misreport their education. Likewise the classic question in macroeconomics of why some countries grow faster than others is plagued by omitted variables bias, measurement error, and simultaneity (Acemoglu, Johnson, and Robinson, 2005). They argue for example that institutions are an important determinant of economic growth, but a growing economy likely develops better institutions.

Professor Powell will discuss some classical situations such as these that present endogeneity. Although some of them have a standard solution, others require a creative solution. The four most problematic cases are:

1. Lagged dependent variable and Serial correlation
2. Omitted Variables
3. Measurement error
4. Simultaneous Equations

All of the solutions require finding an instrumental variable, and performing an Instrumental Variable Regression (IV), a Two Stage Least Squares Regression (2SLS), or a Generalized Method of Moments regression (GMM) as appropriate.

4 Instrumental Variables

The solution to endogenous regressors is to change transform the model so β identifiable. We do so with an *"instrument"*, a set of variables which we will refer to as Z. The amount of instruments that we have affects our method of estimation.

We must satisfy two conditions about Z to consider it a valid instrument:

1. Z must be uncorrelated with ε: $E(Z'\varepsilon) = 0$.
2. Z must be correlated with X, and preferably, this correlation is as high as possible: $E(Z'X) \neq 0$. If were to regress x – the variable we must instrument – on z – our instrumental variables for x – and any other control variable:

$$x_i = \alpha_0 + \alpha_1 w_{1i} + \cdots + \alpha_p w_{pi} + z_i'\gamma + v_i$$

   γ must be statiatically different from 0. In this regression, we assume the classical linear assumptions wherein $E(v_i) = 0$, $Var(v_i) = \sigma_v^2$, and full row rank regressors.

If Z satisfies these conditions then it will solve the identification problem because we can β as a function of population moments that have corresponding sample counterparts:

$$\begin{aligned} Y &= X\beta + \varepsilon \\ Z'Y &= Z'X\beta + Z'\varepsilon \\ E(Z'Y) &= E(Z'X)\beta + E(Z'\varepsilon) \\ E(Z'Y) &= E(Z'X)\beta \\ &\Rightarrow E(Z'(Y - X'\beta)) = 0 \\ &\Rightarrow E(Z'\varepsilon) = 0 \end{aligned}$$

The caveat is that $dim(Z) \geq dim(X)$ to identify the K parameters of β. If $dim(Z) = K$ then we have K equations with K unknowns and it is clear that we could identify β if the instrument is valid. We cover this case in this section.

4.1 Motivation and Examples

Finding an instrument is not usually easy because it is often not something that can be derived mathematically since we do not observe ε. Instead with omitted variables, for instance, the researcher must be creative and argue that $E(Z'\varepsilon) = 0$. Often in the empirical literature, the contribution is the instrumental variable strategy since convincing instruments are difficult to come by. Accordingly a significant part of the paper is devoted to motivating the instrument, and then defending that it is not correlated with possible components of the error term. The latter is known as the robustness checks; if someone has any doubt about the lack of correlation then the instrument

is no longer valid. Moreover we want an instrument that is not weakly correlated with regressors though this condition is straightforward to from a least squares regression of X on Z because both are data.

The classic instrumental variables example in an observational study is Josh Angrist's (1990) research on the effect of having served in the Vietnam War on one's lifetime income. In this case the treatment is going to war while the control is not going; there is no doubt that the latent decision is not random. If we find that people that went to war are poorer in the future, it can be because of war but also because the people who went to war were less prepared for the labor market in the first place. Why might they have enlisted? Some of them could have done so because they could not find another job, or because the payment in the army is better than what they would get otherwise. Also, there is considerable government support for veterans, which could have influenced the decision to enlist. Accordingly it is not clear a prior whether the loss of labor market experience outweighs the post-war benefits.

Consider the regression:

$$y_i = x_i\beta + \varepsilon_i$$

where y_i is income, x_i is a binary variable of having gone to war.

Is x_i uncorrelated with ε_i? Hardly, because there are plenty of things that are correlated with earnings that are also correlated with the decision of enlisting. The error is everything that affects earnings (y_i) other than having served in the war (x_i), and a person's earnings are of course not only determined by having been in a war. They are determined by other things such as education, ability, experience, personal networks, appearance, race, and so forth. We then ask whether these factors are correlated with x_i, having served in war? Quite likely. For instance, a less educated person will earn less in average than a more educated one, and also a less educated person is more likely to enlist, because the army is a good job that pays better than working in say pizza delivery and earns more respect. In this case, $\hat{\beta}_{OLS}$ will not only reflect the exact effect of going to war in earnings, but also the effect of education in earnings **through** going to war as seen by the identifiability math.

However we could control for education, health, parents' education, work experience, and all other relevant observable characteristics that are relevant to both income and serving in the war. The regression would be:

$$y_i = \alpha_0 + \alpha_1 w_{1i} + \cdots + \alpha_p w_{pi} + \beta x_i + \varepsilon_i$$

where the each w_i' is our vector of control variables. If we do OLS by partitioned regression, $\hat{\beta}_{OLS}$ would be the effect of going to war on income, controlling for the effect of each w_i'. However some variable could still be missing from the data set, or worse unobservable. That is, there could still be something in individuals that enlist that makes them earn systematically less in the future. It could be a psychological characteristic, such as ability or motivation. If so we would still have endogeneity, and we would not be able to identify β through OLS.

Fortunately for Angrist's research there was a moment during the Vietnam era when the government needed people to enlist and voluntary enlistment was not covering the demand for soldiers. Accordingly the government instituted compulsory enlistment, which at that time meant going to war provided one passed certain physical and mental requirements. All of the men in a certain age range were randomly assigned a lottery number. Then the army would call starting from the lowest number and going up until it satisfied its demand. Because the number was absolutely random, a person being called for war did not depend on any social, psychological or economic characteristics, in such a way that those who stayed were on average similar to the ones who ended up going to war [1]

Angrist thus exploits the person's number in the lottery draft as an instrument. [2]
Let's see if the instrument satisfies the two requirements:

- $E(z_i\varepsilon_i) = 0$? The lottery number is randomly assigned so the number assigned must be independent of any unobservable characteristic that could influence earnings. At least, we will assume this to be the case for all practical purposes; as previously noted any doubt in this argument must be motivated conceptually and would present the instrument from being valid.

- $E(z_i x_i') \neq 0$? Although the correlation is not perfect, there is a huge increase in the probability of going to war if one's lottery number is low; x and z are definitely correlated.

Instrumental variables has provoked very interesting discussions amongst empirical researchers, at least in labor economics and economics history/economic growth, as well as in econometric research. If the instrumental variable is not valid – usually, $E(Z'\varepsilon) \neq 0$ – then the researcher has not solved the problems that arise with OLS without the linear expectations hypothesis, that is the lack of identification, consistency, and unbiasedness. In other words, the researcher does not have an identification strategy for estimating β. Angrist's paper is often cited as one of the most convincing applications of instrumental variables because there is no better reason to argue in favor of randomization in an observational study than one based on a lottery. Nevertheless some are not convinced that $E(z_i x_i') = 0$. Without going into too many of the details, some economists are more open to the use of instruments that require creativity because they do not require any distributional assumptions about the error terms. However instrumental variables papers can be controversial if researchers are not convinced that $E(Z'\varepsilon) = 0$. And even if the instrument is valid, inference may not be as convincing if the instrument is considered to be weak, which has been an

[1] I exaggerate to make a point. There were some problems with the lottery and people's decisions to enlist because those who had low numbers tended to enlist voluntarily before being called. However, even if the lottery number did not absolutely determine one's going or not to war, a lower number certainly meant a higher probability of going to war on average. Angrist discusses all of this in the paper, but here we just want to understand the nature of an instrument, so I'm taking some permissions.

[2] In the applied econometrics liteerature would argue that Angrist is identifying a local average treatment effect by analyzing the effect of people who are induced to serve in the war only because of the lottery versus those who would have served in the absence of the lotery. For this course we are not concerned about this interpretation though it is something to keep in mind for empirical research.

area of research amongst econometricians. Although this discussion may sound cynical, it is too emphasize the difficulty in asserting causality in empirical economic research. At least in labor, some alternative research designs such as the regression discontinuity design, have been developed and recently somewhat widely applied, to assert causality at least locally in a population.

Although this class is not about the empirical literature, hopefully it is clear how prevalent endogeneity is, how serious it is, and how difficult it is to convincingly solve. For these reasons it is a highly important topic, and as we will shortly see there is quite a bit of math that we can do that is within the scope of the course to analyze endogenous regressors. It should thus not be surprising that instrumental variables has appeared on every recent exam, and as comprised as many as half of the exam's points.

5 Just-Identified IV Estimation

We say that are in the just-identifed case if the dimension of z is the same as the dimension of x so that we have as many instrumental variables as we have endogenous regressors. We assure that $dim(Z) = dim(X)$ by making the remaining instrumental variables the remaining controls in X. Because we assume that these controls are exogenous then they satsify the validity conditions of being independent of the unobservables as a result of exogeneity and correlated with the regressors by being regressors themselves. As a result $rank(E(Z'X)) = K$. Moreover we assume that Z is nonstochastic. Our identification equation translates to

$$\beta = E(Z'X)^{-1}E(Z'Y)$$

and we estimate it by using sample moments:

$$\hat{\beta}_{IV} = \left[\frac{1}{n}\sum_{i=1}^{n} z_i x_i'\right]^{-1} \frac{1}{n}\sum_{i=1}^{n} z_i y_i = (Z'X)^{-1}Z'Y$$

We now show that $\hat{\beta}_{IV}$ is an unbiased estimator for β:

$$\begin{aligned}
\hat{\beta}_{IV} &= (Z'X)^{-1}Z'(X\beta + \varepsilon) \\
&= \beta + (Z'X)^{-1}Z'\varepsilon \\
\Rightarrow \hat{\beta}_{IV} - \beta &= (Z'X)^{-1}Z'\varepsilon \\
\Rightarrow E(\hat{\beta}_{IV} - \beta) &= E[(Z'X)^{-1}Z'\varepsilon] \\
&= Z'X^{-1}E(Z'\varepsilon) = 0
\end{aligned}$$

5.1 Asymptotics for the IV estimator

We now return to our asymptotic calculations and modify them for our new estimator. We want to demonstrate that our estimator is consistent and that we conduct hypothesis testing – at least in large samples without assuming that our errors are normally distributed These two reasons are the main motivation for why asymptotics is so important to this course. We are interested in these two

asymptotic properties for each estimator we have discussed in Econ 240B, although for some the math is beyond the scope of the course.
We $\hat{\beta}_{IV} - \beta$ as

$$\hat{\beta}_{IV} - \beta = (Z'X)^{-1}Z'\varepsilon = (\frac{1}{n}\sum_{i=1}^{n} z_i x_i)^{-1}\frac{1}{n}\sum_{i=1}^{n} z_i\varepsilon_i$$

As usual we treat each part separately so we can ultimately apply Slutsky's Theorem:

$\underline{\frac{1}{n}\sum_{i=1}^{n} z_i x_i'}$:

Since we still have random sampling, the x_i' and z_i are iid and thus so is $z_i x_i'$. Also assuming that the second moments are finite, the weak law of large number states that:

$$\frac{1}{n}\sum_{i=1}^{n} z_i x_i' \longrightarrow_p E(zx')$$

$\underline{\frac{1}{n}\sum_{i=1}^{n} z_i\varepsilon_i}$:

Since the ε_i are iid[3], so are the $z_i\varepsilon_i$. Also assuming that $E(z_i^2\varepsilon_i^2)$ is finite[4], we can apply the weak law of large numbers and the central limit theorem to obtain:

$$\frac{1}{n}\sum_{i=1}^{n} z_i\varepsilon_i \longrightarrow_p E(z\varepsilon) = 0$$

and

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} z_i\varepsilon_i \longrightarrow_d N(0, E(z^2\varepsilon^2))$$

Convergence in probability implies convergence in distribution. Apply that to $\frac{1}{n}\sum_{i=1}^{n} z_i\varepsilon_i \longrightarrow_p E(z\varepsilon) = 0$. As a result of the Continuous Mapping Theorem and then Slutsky's Theorem,

$$(\frac{1}{n}\sum_{i=1}^{n} z_i x_i)^{-1}\frac{1}{n}\sum_{i=1}^{n} z_i\varepsilon_i \longrightarrow_p 0 \Rightarrow \hat{\beta}_{IV} \longrightarrow_p \beta$$

and

$$\sqrt{n}(\hat{\beta}_{IV} - \beta) \longrightarrow_d N(0, E(zx')^{-1}E(z^2\varepsilon^2)E(xz')^{-1})$$

Thus a pivotal statistic for hypothesis testing is

[3]Observe that we don't have heteroskedasticity here since we're only relaxing the linear expectations assumption.

[4]We only assumed z_i and ε_i are uncorrelated. If we want to go further and assume independence, then $E(z_i^2\varepsilon_i^2) = E(z_i^2)E(\varepsilon_i^2) = \sigma^2 E(z_i^2)$

$$\frac{\sqrt{n}(\hat{\beta}-\beta)}{\left(\left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i'\right)^{-1}\hat{V}\left(\frac{1}{n}\sum_{i=1}^{n} x_i z_i'\right)^{-1}\right)^{-1/2}} \longrightarrow_d N(0, I_K)$$

where $\hat{V}$ is a consistent estimator of $plim\left(\frac{Z'\Omega Z}{n}\right)$ and $\Omega = E(\varepsilon\varepsilon')$. The middle term is: $s^2 I_n$ in the homoskedasticity case. If we would like to relax the scalar covariance assumption however then we could use Eicker-White, $Diag(y_i - x_i\hat{\beta}_{IV})$, for pure heteroskedasticity, and Newey-West if there is also serial correlation.

Next week we will discuss the over-identified case and GMM estimation.

Endogeneity and Exam Practice

Jeffrey Greenbaum

March 9, 2007

Contents

1 Section Preamble

In this section we complete our discussion of endogeneous regressors by analyzing the over-identified case, which lends to 2SLS and GMM estimation. Recall last week's section preamble for our motivation of endogenous regressors. Over-identifcation is the case in which we have more instrumental variables than regressors, and thus $dim(Z) > dim(X)$. We will then discuss a slew of old exam questions on endogeneity.

2 Overidentified case: 2SLS

When there are more instruments than regressors $dim(Z) = L > K$ and $(Z'X)$ is no longer invertible since it is not a square matrix. Estimating β in $E(Z'(Y - X\beta)) = 0$ with sample counterparts would reduce to more equations than unknowns, which can be problematic to solve. We could throw away instruments so that we have only K instruments but it would be senseless to do so since we can only improve our estimate with more instruments. Instead we solve the dimensionality problem by premultiplying Z by $\hat{\Pi}$ so $rank(\hat{\Pi}'Z'X) = K$ and there are K equations. We thus generalize $\hat{\beta}_{IV}$ to $\hat{\beta}_{GIV}$ which solves the sample counterpart of $E(\Pi'Z'(Y - X\beta)) = 0$:

$$\begin{aligned}
&\hat{\Pi}'Z'X\hat{\beta}_{GIV} = \hat{\Pi}'Z'Y \\
&\Rightarrow \hat{\beta}_{GIV} = (\hat{\Pi}'Z'X)^{-1}\hat{\Pi}'Z'Y
\end{aligned}$$

The traditional choice for Π comes from two stage least squares esimtation. We first regress X on Z, the first stage, and then we regress Y on the predicted values of X given Z, the reduced form equation. The first stage removes from X all of its correlation with ε, and the reduced form analyzes how X affectes Y through Z.

More formally we compute $\hat{\beta}_{2SLS}$ first by estimating Π in $X = Z\Pi + v_i$ through OLS where we assume that Z is exogenous. As a result $\hat{\Pi} = (Z'Z)^{-1}Z'X$, and we obtain $\hat{X} = Z\hat{\Pi} = Z(Z'Z)^{-1}Z'X = P_Z X$. Then we run the regression $Y = \hat{X}\beta + \varepsilon$ through OLS and obtain $\hat{\beta}$, which is our two stage least squares estimator:

$$\begin{aligned}
\hat{\beta}_{2SLS} &= (\hat{X}'\hat{X})^{-1}\hat{X}'y \\
&= (X'P_Z'P_Z X)^{-1}X'P_Z y \\
&= (X'P_Z X)^{-1}X'P_Z y \\
&= (X'Z(Z'Z)^{-1}Z'X)^{-1}(X'Z(Z'Z)^{-1}Z'y)
\end{aligned}$$

In the case that $dim(Z) = K$ we can show that $\hat{\beta}_{2SLS} = \hat{\beta}_{IV}$.

2.1 Properties of 2SLS Estimator

$\hat{\beta}_{2SLS}$ is unbiased, consistent, and asymptotically normal because $\hat{\Pi} \longrightarrow_p \Pi$ and is nonstochastic. The derivations follow that of $\hat{\beta}_{IV}$.
We start by computing $\hat{\beta}_{2SLS} - \beta$:

$$\begin{aligned}
\hat{\beta}_{2SLS} - \beta &= (X'P_Z X)^{-1}X'P_Z(X\beta + \varepsilon) - \beta \\
&= (X'P_Z X)^{-1}X'P_Z\varepsilon \\
&= (\frac{1}{n}\sum_{i=1}^{n} x_i z_i'(z_i z_i')^{-1}z_i x_i')^{-1}(\frac{1}{n}\sum_{i=1}^{n} x_i z_i'(z_i z_i')^{-1}z_i\varepsilon_i)
\end{aligned}$$

$\hat{\beta}_{2SLS}$ is unbiased and consistent because $E(Z'\varepsilon) = 0$, as was used for our proof in $\hat{\beta}_{IV}$ As a result $\frac{1}{n}\sum_{i=1}^{n} z_i\varepsilon_i \longrightarrow_p 0$, and the remaining terms converge in probability to finite nonzero moments. Thus by Slutsky's Theorem the entire expression converges in probability to zero.

Moreover by the Central Limit Theorem, $\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} z_i\varepsilon_i \longrightarrow_d N(0, V_0)$.
Again we repeat the work of $\hat{\beta}_{IV}$ to show that

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta) \longrightarrow_d N\left(0, plim\left(\left(\frac{X'P_ZX}{n}\right)^{-1}\frac{X'P_Z\hat{V}P_ZX}{n}\left(\frac{X'P_ZX}{n}\right)^{-1}\right)\right)$$

We estimate $\hat{V}$ based on our assumptions of the covariance matrix.

3 GMM

In the case that $V_0 \neq \sigma^2 M_{zz}$ then $\hat{\beta}_{2SLS}$ will not have the smallest asymptotic covariance matrix. The asymptotic variance of β_{GIV} is $[\Pi' M_{zx}]^{-1}\Pi' V_0 \Pi [M_{xz}\Pi]^{-1}$ where $\frac{1}{T}Z'X \longrightarrow_p M_{zx}$. In this section we extend our two stage estimation to other possible matrices for $\hat{\Pi}$.

We obtain an efficient estimator by choosing $\hat{\Pi}$ that minimizes our estimate of this asymptotic variance matrix. This problem is analogous to that of the generalized regression model where we transformed the linear model to find the most efficient estimator.

As such we multiply the linear model by $\frac{1}{\sqrt{T}}Z'$ so we are estimating $\frac{1}{\sqrt{T}}Z'y = \frac{1}{\sqrt{T}}Z'X + \frac{1}{\sqrt{T}}Z'\varepsilon$. We continue to assume that $E[z_i\varepsilon_i] = 0 \Rightarrow E[\frac{1}{\sqrt{T}}Z'\varepsilon] = 0$.

As a result we can apply the Central Limit Theorem to show that $\frac{1}{\sqrt{T}}Z'\varepsilon \longrightarrow_d N(0, V_0)$. Moreover, we can show that asymptotically this model satisfies the generalized regression assumptions, particularly that asymptotically the covariance of the new design matrix with the new error vector is zero.

Thus we use the GLS framework to define $\hat{\beta}_{GMM}$:

$$\begin{aligned}\hat{\beta}_{GLS} &= ((\frac{1}{\sqrt{T}}Z'X)'V_0^{-1}\frac{1}{\sqrt{T}}Z'X)^{-1}(\frac{1}{\sqrt{T}}Z'X)'V_0^{-1}\frac{1}{\sqrt{T}}Z'y \\ &= [X'ZV_0^{-1}Z'X]^{-1}X'ZV_0^{-1}Z'y\end{aligned}$$

Accordingly $\hat{\pi} = V_0^{-1}(\frac{1}{T}Z'X)$ where we need a consistent estimator of V_0 through methods such as Eicker-White or Newey-West.

Moreover we can use our GLS and 2SLS frameworks to derive $\sqrt{T}(\hat{\beta}_{GMM} - \beta) \longrightarrow_d N(0, [M_{zx}V_0^{-1}M_{xz}]^{-1})$.

4 Endogeneity Exercises

Endogeneity has appeared on every exam since 2002 in both True/False and free-response questions! In fact recently it has tended to appear in multiple questions.

4.1 2002 Exam, 2

Question: Suppose the coefficients $\beta = (\beta_1, \beta_2)'$ in the lienar model $y = X\beta + \varepsilon$ are estimated by two-stage least squares, where it is assumed that the errors ε are independent of the matrix Z of instruments with scalar covariance matrix $Var(\varepsilon) = Var(\varepsilon|Z) = \sigma^2 I$. An analysis of $N = 163$ observations yields

$$\hat{\beta}_{2SLS} = \begin{pmatrix} 2 \\ 5 \end{pmatrix} \quad \hat{\sigma}^2_{2SLS} = 4, \quad \hat{X}'\hat{X} = (X'Z)(Z'Z)^{-1}(Z'X) = \begin{pmatrix} 5 & 1 \\ 1 & 1 \end{pmatrix}$$

Construct an approximate 95% confidence interval for $\gamma = \beta_1 * \beta_2$, under the (possibly heroic) assumption that the sample size is large enough for the usual limit theorems and linear approximations to be applicable. Is $\gamma_0 = 0$ in this interval?

Answer: The sample size is sufficiently large that we can use the limiting distribution of $\hat{\beta}_{2SLS}$ to approximate our test statistic. Recall that $\hat{\beta}_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'y$ where

$$\begin{aligned} Var(\hat{\beta}_{2SLS}) &= (\hat{X}'\hat{X})^{-1}\hat{X}'Var(y)\hat{X}(\hat{X}'\hat{X})^{-1} \\ &= (\hat{X}'\hat{X})^{-1}\hat{X}'(\sigma^2 I)\hat{X}(\hat{X}'\hat{X})^{-1} \\ &= \sigma^2(\hat{X}'\hat{X})^{-1}\hat{X}'\hat{X}(\hat{X}'\hat{X})^{-1} \\ &= \sigma^2(\hat{X}'\hat{X})^{-1} \end{aligned}$$

Based on the usual least squares asymptotics, $\sqrt{N}(\hat{\beta}_{2SLS} - \beta) \longrightarrow_d N(0, \sigma^2 \frac{(\hat{X}'\hat{X})^{-1}}{N})$.
However, we are interested in the asymptotic distribution of $\gamma = \beta_1 * \beta_2 = g(\beta_1, \beta_2)$.
We thus use the Delta Method to show that $\sqrt{N}(\hat{\gamma} - \gamma) \longrightarrow_d N(0, \sigma^2 G \frac{(\hat{X}'\hat{X})^{-1}}{N} G')$ where $G = \frac{\partial g(\beta_1,\beta_2)}{\partial(\beta_1,\beta_2)'} = (\beta_2, \beta_1)$.

We estimate G with $\hat{G} = (\hat{\beta}_2, \hat{\beta}_1)$ because $\hat{G} \longrightarrow_p G$.
Moreover we know that $\hat{\sigma}^2 \longrightarrow_p \sigma^2$ by law of large numbers.
We can thus apply Slutksy's Theorem twice to show that $\frac{\sqrt{N}(\hat{\gamma}-\gamma)}{\sqrt{\frac{1}{N}\hat{G}\hat{\sigma}^2(\hat{X}'\hat{X})^{-1}\hat{G}'}} \longrightarrow_d N(0,1)$.
Equivalently $\frac{(\hat{\gamma}-\gamma)}{\sqrt{(\hat{\beta}_2,\hat{\beta}_1)\hat{\sigma}^2(\hat{X}'\hat{X})^{-1}(\hat{\beta}_2,\hat{\beta}_1)'}} \longrightarrow_d N(0,1)$ where $\hat{V} = \hat{\sigma}^2(\hat{X}'\hat{X})^{-1}$.

For these data, the estimate of γ is $\hat{\gamma} = \hat{\beta}_1 * \hat{\beta}_2 = 2 * 5 = 10$.
We now calculate the asymptotic standard error, $\sqrt{\hat{G}\hat{V}(\hat{\beta}_{2SLS})\hat{G}'}$:

$$\hat{V}(\hat{\beta}_{2SLS}) = \hat{\sigma}^2(\hat{X}'\hat{X})^{-1}$$
$$4 * \left(\frac{1}{4}\right)\begin{pmatrix} 1 & -1 \\ -1 & 5 \end{pmatrix}$$
$$= \begin{pmatrix} 1 & -1 \\ -1 & 5 \end{pmatrix}$$

$$\hat{G}\hat{V}\hat{G}' = \begin{pmatrix} 5 & 2 \end{pmatrix} * \begin{pmatrix} 1 & -1 \\ -1 & 5 \end{pmatrix} * \begin{pmatrix} 5 \\ 2 \end{pmatrix}$$
$$= \begin{pmatrix} 3 & 5 \end{pmatrix}\begin{pmatrix} 5 \\ 2 \end{pmatrix}$$
$$= 25$$

As a result, $SE(\hat{\gamma}) = \sqrt{25} = 5$.

Therefore, an approximate 95% confidence interval for γ is:
$CI = (\hat{\gamma} \pm 1.96 * SE(\hat{\gamma})) = (10 \pm 1.96 * 5) = (10 \pm 9.8) = (0.2, 19.8)$.
0 is not in this confidence interval.

4.2 2003 Exam, 1A

Question: True/False/Explain. The Two-Stage Least Squares estimator $\hat{\beta}_{2SLS}$ is unchanged if the original NXL matrix of instrumental variables Z is replaced by a new matrix Z* of instruments if Z* = ZH, where H is an invertible L X L matrix.

Answer: True. The first stage projection matrix, $P_{Z^*} = Z^*((Z^*)'Z^*)^{-1}(Z^*)'$ for the transformed instruments Z^* is identical to the corresponding projection matrix $P_Z = Z(Z'Z)^{-1}Z'$ for the original instruments,

$$\begin{aligned} P_{Z^*} &= Z^*((Z^*)'Z^*)^{-1}(Z^*) \\ &= ZH((ZH)'(ZH))^{-1}(ZH)' \\ &= ZH(H'Z'ZH)^{-1}H'Z' \\ &= ZHH^{-1}(Z'Z)^{-1}(H')^{-1}H'Z' \\ &= Z(Z'Z)^{-1}Z' = P_Z \end{aligned}$$

Since the two-stage least squares estimator is defined as $\hat{\beta}_{2SLS} = (X'P_ZX)^{-1}X'P_Zy$, it does not change if P_{Z^*} replaces P_Z.
Note that Z^{-1} does not exist because Z is not a square matrix.

4.3 2004 Exam, 3

Question: Consider the estimation of two scalar coefficients, β_1 and β_2, in the linear equation

$$y = x_1\beta_1 + x_2\beta_2 + \varepsilon$$

where y, x_1, and x_2 are observable N-dimensional random vectors. In addition, two N-dimensional vectors of instrumental variables, z_1 and z_2, are available. In a sample size of $N = 227$, the following matrix of cross-products of the variables is observed:

$$\begin{bmatrix} y'y & y'x_1 & y'x_2 & y'z_1 & y'z_2 \\ x_1'y & x_1'x_1 & x_1'x_2 & x_1'z_1 & x_1'z_2 \\ x_2'y & x_2'x_1 & x_2'x_2 & x_2'z_1 & x_2'z_2 \\ z_1'y & z_1'x_1 & z_1'x_2 & z_1'z_1 & z_1'z_2 \\ z_2'y & z_2'x_1 & z_2'x_2 & z_2'z_1 & z_2'z_2 \end{bmatrix} = \begin{bmatrix} 22 & -11 & 10 & 8 & 8 \\ -11 & 21 & 10 & -8 & -8 \\ 10 & 10 & 20 & -2 & 0 \\ 8 & -8 & -2 & 6 & 4 \\ 8 & -8 & 0 & 4 & 6 \end{bmatrix}$$

A. For these data, calculate the classical LS estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ of the unknown regression coefficients, and compute the instrumental variables estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ using z_1 and z_2 as instruments for x_1 and x_2.

B. Suppose the error terms ε are independent of z_1 and z_2, so that $Var[\varepsilon|z_1, z_2] = \sigma^2 I$, ie., ε has a scalar covariance matrix. If you had to conduct a test of $H_0 : \beta_2 = 1$ versus $H_A : \beta_2 \neq 1$ at an asymptotic 5% level using the IV estimator, and were given a consistent estimator $\tilde{\sigma}^2$ of the unknown variance parameter σ^2, how small would $\tilde{\sigma}^2$ have to be to reject H_0? That is, find the largest value of $\tilde{\sigma}^2$ for which you could (barely) reject the null hypothesis.

Answer: We first calculate the least squares and instrumental variable estimates using the usual formulae.

$$\begin{aligned} \hat{\beta}_{OLS} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} &= (X'X)^{-1}X'y = \left(\begin{pmatrix} x_1 & x_2 \end{pmatrix}' \begin{pmatrix} x_1 & x_2 \end{pmatrix} \right)^{-1} \begin{pmatrix} x_1 & x_2 \end{pmatrix}' \begin{pmatrix} y \end{pmatrix} \\ &= \left(\begin{pmatrix} x_1' \\ x_2' \end{pmatrix} \begin{pmatrix} x_1 & x_2 \end{pmatrix} \right)^{-1} \begin{pmatrix} x_1' \\ x_2' \end{pmatrix} \begin{pmatrix} y \end{pmatrix} \\ &= \begin{pmatrix} x_1'x_1 & x_1'x_2 \\ x_2'x_1 & x_2'x_2 \end{pmatrix}^{-1} \begin{pmatrix} x_1'y \\ x_2'y \end{pmatrix} \\ &= \begin{bmatrix} 21 & 10 \\ 10 & 20 \end{bmatrix}^{-1} \begin{pmatrix} -11 \\ 10 \end{pmatrix} \\ &= \frac{1}{320} \begin{bmatrix} 20 & -10 \\ -10 & 21 \end{bmatrix} \begin{pmatrix} -11 \\ 10 \end{pmatrix} \\ &= \frac{1}{320} \begin{pmatrix} -320 \\ 320 \end{pmatrix} \\ &= \begin{pmatrix} -1 \\ 1 \end{pmatrix} \end{aligned}$$

$$
\begin{aligned}
\hat{\beta}_{IV} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} &= (Z'X)^{-1}Z'y = \left(\begin{pmatrix} z_1 & z_2 \end{pmatrix}' \begin{pmatrix} x_1 & x_2 \end{pmatrix} \right)^{-1} \begin{pmatrix} z_1 & z_2 \end{pmatrix}' \begin{pmatrix} y \end{pmatrix} \\
&= \left(\begin{pmatrix} z_1' \\ z_2' \end{pmatrix} \begin{pmatrix} x_1 & x_2 \end{pmatrix} \right)^{-1} \begin{pmatrix} z_1' \\ z_2' \end{pmatrix} \begin{pmatrix} y \end{pmatrix} \\
&= \begin{pmatrix} z_1'x_1 & z_1'x_2 \\ z_2'x_1 & z_2'x_2 \end{pmatrix}^{-1} \begin{pmatrix} z_1'y \\ z_2'y \end{pmatrix} \\
&= \begin{bmatrix} -8 & -2 \\ -8 & 0 \end{bmatrix}^{-1} \begin{pmatrix} 8 \\ 8 \end{pmatrix} \\
&= \frac{-1}{16} \begin{bmatrix} 0 & 2 \\ 8 & -8 \end{bmatrix} \begin{pmatrix} 8 \\ 8 \end{pmatrix} \\
&= \frac{1}{16} \begin{pmatrix} -16 \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} -1 \\ 0 \end{pmatrix}
\end{aligned}
$$

Recall that the asymptotic variance matrix of the IV estimator is given by

$$AVar(\hat{\beta}_{IV}) = (Z'X)^{-1}Z'Var(y|X,Z)Z(X'Z)^{-1}$$

It is given that $Var(\varepsilon|Z) = Var(y|Z) = \sigma^2 I$. In addition for these data, $Z'X = X'Z$. It thus follows that:

$$
\begin{aligned}
AVar(\hat{\beta}_{IV}) &= (Z'X)^{-1}Z'\sigma^2 I Z(X'Z)^{-1} \\
&= \sigma^2 (Z'X)^{-1}Z'IZ(Z'X)^{-1} \\
&= \sigma^2 (Z'X)^{-1}Z'Z(Z'X)^{-1} \\
&= \sigma^2 \left(\frac{1}{16}\right) \begin{bmatrix} 0 & -2 \\ -8 & 8 \end{bmatrix} \begin{bmatrix} z_1'z_1 & z_1'z_2 \\ z_2'z_1 & z_2'z_2 \end{bmatrix} \left(\frac{1}{16}\right) \begin{bmatrix} 0 & -2 \\ -8 & 8 \end{bmatrix} \\
&= \sigma^2 \left(\frac{1}{16^2}\right) \begin{bmatrix} 0 & -2 \\ -8 & 8 \end{bmatrix} \begin{bmatrix} 6 & 4 \\ 4 & 6 \end{bmatrix} \begin{bmatrix} 0 & -2 \\ -8 & 8 \end{bmatrix} \\
&= \sigma^2 \left(\frac{1}{16^2}\right) \begin{bmatrix} 96 & 80 \\ -120 & 160 \end{bmatrix}
\end{aligned}
$$

Replacing the unknown value of σ^2 by an estimator $\tilde{\sigma}^2$ would give $\hat{AVar}(\hat{\beta}_2) = \tilde{\sigma}^2 \left(\frac{5}{8}\right)$. Thus, the asymptotic t-test for $H_0 : \beta_2 = 1$ would reject the null hypothesis if

$$T = \frac{|\beta_2 - 1|}{\sqrt{Var(\beta_2)}} = \frac{|0-1|}{\sqrt{(\tilde{\sigma}^2)\frac{5}{8}}} > 1.96 \Rightarrow \tilde{\sigma}^2 < 1.6 * 1.96^{-2} \approx 0.4.$$

4.4 2005 Exam, 3

Question: Suppose that, for the sample linear model with no intercept term,

$$y_i = \beta x_i + \varepsilon_i$$

that both $z_i = 1$ and $z_2 = x_i$ are valid instrumental variables for x_i, that is

$$E(z_{i1}\varepsilon_i) = E(\varepsilon_i) = 0$$
$$E(z_{i2}\varepsilon_i) = E(x_i\varepsilon_i) = 0$$

and

$$E(z_{i1}x_i) = E(x_i) = \mu \neq 0$$
$$E(z_{i2}x_i) = E(x_i^2) = \tau^2 \neq 0$$

A. Under the assumption that ε_i and x_i are jointly i.i.d. and ε_i is independent of x_i with $E(\varepsilon_i^2) = \sigma^2$, derive the asymptotic distribution of the IV estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ which use $z_{i1} = 1$ or $z_{i2} = x_i$, respectively, as an instrument for x_i, and compare the asymptotic variances of these two estimators.

B. Under the same assumptions as in part A, explicitly derive the asymptotic variance for the GMM estimator $\hat{\beta}_{GMM}$ which optimally uses both $z_{i1} = 1$ and $z_{i2} = x_i$ as instrumental variables, and show that this variance reduces to the asymptotic variance of one of the estimators in part A. [Hint: the relevant matrices M_{XZ} and V_0 can be written in term of the parameters given above.]

Answer: Recall the asymptotic distribution for the IV estimator:

$$\sqrt{n}(\hat{\beta}_{IV} - \beta) \longrightarrow_d N(0, E(z_i x_i')^{-1} E(z_i^2 \varepsilon_i^2) E(x_i z_i')^{-1})$$

For $z_{i1} = 1$, $E(z_i^2\varepsilon_i^2) = E(\varepsilon_i^2) = \sigma^2$.
$E(z_i x_i') = E(x_i') = E(x_i) = \mu$
As a result, the asymptotic variance is $E(x_i)^{-1}\sigma^2 E(x_i)^{-1} = \sigma^2 E(x_i)^{-2} = \sigma^2\mu^{-2}$.
Accordingly, $\sqrt{n}(\hat{\beta}_1 - \beta) \longrightarrow_d N(0, \sigma^2\mu^{-2})$

For $z_{i2} = 1$, $E(z_i^2\varepsilon_i^2) = E(x_i^2\varepsilon_i^2) = E(x_i^2)E(\varepsilon_i^2) = \sigma^2 E(x_i^2)$.
$E(z_i x_i') = E(x_i * x_i') = E(x_i^2) = \tau^2$
As a result, the asymptotic variance is $E(x_i^2)^{-1}\sigma^2 E(x_i^2) E(x_i^2)^{-1} = \sigma^2 E(x_i^2)^{-1} = \sigma^2\tau^{-2}$.
Accordingly, $\sqrt{n}(\hat{\beta}_2 - \beta) \longrightarrow_d N(0, \sigma^2\tau^{-2})$.

We know that $0 < Var(x_i) = E(x_i^2) - E(x_i)^2 \Rightarrow E(x_i^2) > E(x_i)^2 \Rightarrow E(x_i^2)^{-1} < E(x_i)^{-2}$.
As a result, $AVar(\hat{\beta}_1) < AVar(\hat{\beta}_2)$ when using instrumental variables estimation.

Recall that the asymptotic variance for the $\hat{\beta}_{GMM}$ is $(\frac{X'Z}{n}(\frac{Z'\hat{V}Z}{n})^{-1}\frac{Z'X}{n})^{-1}$.
In this set-up, $z_i = (1, x_i)'$ and $\hat{V} = \sigma^2 I$.

$\frac{X'Z}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i z_i' = \frac{1}{n}\sum_{i=1}^{n} x_i(1, x_i) = (\frac{1}{n}\sum_{i=1}^{n} x_i, \frac{1}{n}\sum_{i=1}^{n} x_i^2) \longrightarrow_p (\mu, \tau^2)$.

Accordingly, $\frac{XZ'}{n} \longrightarrow_p (\mu, \tau^2)'$.

Looking at the middle term,
$\frac{Z'\hat{V}Z}{n} = \frac{\sigma^2 Z'Z}{n} = \sigma^2(\frac{1}{n}\sum_{i=1}^{n} z_i z_i') \longrightarrow_p \sigma^2 E(z_i z_i')$
By Continuous Mapping Theorem, if $E(z_i z_i') \neq 0$ then $\frac{Z'\hat{V}Z}{n}^{-1} \longrightarrow_p \frac{1}{\sigma^2}E(z_i z_i')^{-1}$.

$$\frac{1}{\sigma^2}E(z_i z_i')^{-1} = \frac{1}{\sigma^2}E[(1, x_i)'(1, x_i)] = \sigma^2 \begin{pmatrix} 1 & \mu \\ \mu & \tau^2 \end{pmatrix}^{-1} = \frac{1}{\sigma^2(\tau^2 - \mu^2)} \begin{pmatrix} \tau^2 & -\mu \\ -\mu & 1 \end{pmatrix}$$

Therefore, Asymptotic Var$(\hat{\beta}_{GMM}) = (\frac{X'Z}{n}(\frac{Z'\hat{V}Z}{n})^{-1}\frac{Z'X}{n})^{-1} =$

$$\left(\frac{1}{\sigma^2(\tau^2 - \mu^2)} \begin{pmatrix} \mu & \tau^2 \end{pmatrix} \begin{pmatrix} \tau^2 & -\mu \\ -\mu & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau^2 \end{pmatrix}\right)^{-1} = \sigma^2\tau^{-2} = \sigma^2(E(x_i^2))^{-1}.$$

Note that $\hat{\beta}_{GMM}$ is equivalent to the second IV estimator that uses the instrument x_i.

4.5 2006 Exam, 1A

Question: True/False/Explain. The Two-Stage Least Squares estimator $\hat{\beta}_{2SLS}$ is unchanged if the original NXL matrix of instrumental variables Z is replaced by a new matrix Z^* of instruments if $Z^* = HZ$, where H is an invertible NXN matrix.

Answer: False. $\hat{\beta}_{2SLS}$ is unchanged by using Z^* in place of Z if $P_Z = P_{Z^*}$ because the projection matrix is the only place in the estimator in which Z is used.

$$\begin{aligned} P_{Z^*} &= Z^*((Z^*)'Z^*)^{-1}(Z^*)' \\ &= HZ((HZ)'(HZ))^{-1}(HZ)' \\ &= HZ(Z'H'HZ)^{-1}Z'H' \end{aligned}$$

Because we assume that $N > L$, neither Z nor ZH is invertible. As a result, this expression cannot be simplied further and does not equal $P_Z = Z(Z'Z)^{-1}Z$ in general for all invertible NXN matrices H. For a specific Z there are many possible matrices for H that can satisfy $P_Z = P_{Z^*}$, one of which is always the trivial case of the identity matrix, just as there are also many possible matrices for H for which they would not be equal.

However, if $Z^* = ZH$ and H is LXL so that the dimensions are valid for multiplicaiton then the statement would be true:

$$\begin{aligned} P_{Z^*} &= Z^*((Z^*)'Z^*)^{-1}(Z^*) \\ &= ZH((ZH)'(ZH))^{-1}(ZH)' \\ &= ZH(H'Z'ZH)^{-1}H'Z' \\ &= ZHH^{-1}(Z'Z)^{-1}(H')^{-1}H'Z' \\ &= Z(Z'Z)^{-1}Z' = P_Z \end{aligned}$$

4.6 2006 Exam, 3

Question: Consider the two equation system

$$\begin{aligned} y &= x\beta + \varepsilon \\ x &= Z\pi + \eta \end{aligned}$$

where y and x are observable T-dimensional vectors, ε and η are T-vectors of errors assumed jointly independent across rows, β is a scalar unknown parameter, Z is an observable TXL matrix of instrumental variables, and π is a L-dimensional vector of unknown coefficients. The error terms ε and η are jointly independent of the instruments Z, and assumed to have $E[\varepsilon] = E[\eta] = 0, E[\varepsilon\varepsilon'] = \sigma^2 I, E[\eta\eta'] = \tau^2 I$, and $E[\varepsilon\eta'] = \gamma I$, where γ may be nonzero (so x is endogenous in the equation for y).

Suppose you are given an estimator $\hat{\pi}$ from a separate sample (so it is statistically independent of ε and η) that satisfies

$$\sqrt{T}(\hat{\pi} - \pi) \longrightarrow_d N(0, V)$$

Defining $\hat{x} = Z\hat{\pi}$, consider the following two estimators of the scalar parameter β: the "instrumental variables" estimator

$$\hat{\beta}_{IV} = (\hat{x}'x)^{-1}\hat{x}'y,$$

and the "two-stage plug-in" estimator

$$\hat{\beta}_{2S} = (\hat{x}'\hat{x})^{-1}(\hat{x}'y)$$

Assuming $plim T^{-1}Z'Z = M_{ZZ} = E[z_t z_t']$ has $\pi' M_Z Z\pi \neq 0$, derive the limiting distributions of $\sqrt{T}(\hat{\beta}_{IV} - \beta)$ and $\sqrt{T}(\hat{\beta}_{2S} - \beta)$, assuming the relevant Law of Large Numbers and Central Limit Theorems apply. Are these asymptotic distributions the same? If not, is one more efficient than the other in general?

Hint: Substitute the equation for y into the expression for the estimators, and, where necessary, substitute $x = \hat{x} + (x - \hat{x}) = \hat{x} + Z(\pi - \hat{\pi}) + \eta$ as well. Note there is a typo in the hint on the actual exam.

Answer: The asymptotic distributions are not the same. We first analyze $\hat{\beta}_{IV}$ and use many of the same calculations to analyze $\hat{\beta}_{2S}$.

$$
\begin{aligned}
\hat{\beta}_{IV} &= (\hat{x}'x)^{-1}\hat{x}'y \\
&= (\hat{x}'x)^{-1}\hat{x}'(x\beta+\varepsilon) \\
&= (\hat{x}'x)^{-1}\hat{x}'x\beta + (\hat{x}'x)^{-1}\hat{x}'\varepsilon \\
&= \beta + (\hat{x}'x)^{-1}\hat{x}'\varepsilon \\
\Rightarrow \hat{\beta}_{IV} - \beta &= (\hat{x}'x)^{-1}\hat{x}'\varepsilon \\
&= ((Z\hat{\pi})'(Z\pi+\eta))^{-1}((Z\hat{\pi})'\varepsilon) \\
&= (\hat{\pi}'Z'(Z\pi+\eta))^{-1}(\hat{\pi}'Z'\varepsilon) \\
&= (\hat{\pi}'Z'Z\pi + \hat{\pi}'Z'\eta)^{-1}(\hat{\pi}'Z'\varepsilon) \\
&= (T^{-1}\hat{\pi}'Z'Z\pi + T^{-1}\hat{\pi}'Z'\eta)^{-1}(T^{-1}\hat{\pi}'Z'\varepsilon) \\
\Rightarrow \sqrt{T}(\hat{\beta}_{IV} - \beta) &= (T^{-1}\hat{\pi}'Z'Z\pi + T^{-1}\hat{\pi}'Z'\eta)^{-1}\sqrt{T}(T^{-1}\hat{\pi}'Z'\varepsilon)
\end{aligned}
$$

We proceed by analyzing each of the three terms separately. We will then apply Slutsky's Theorem to the sum and then again to the product to determine the overall distribution. In doing so we must also apply Slutsky's Theorem to each term along the way. We will proceed from left to right.

By assumption $\sqrt{T}(\hat{\pi} - \pi) \longrightarrow_d N(0, V)$.
This distribution implies that $\hat{\pi} \longrightarrow_p \pi \Rightarrow \hat{\pi}' \longrightarrow_p \pi' \Rightarrow \hat{\pi}' \longrightarrow_d \pi'$
It is given that $plim T^{-1}Z'Z = M_{ZZ} = E[z_t z_t']$.
As a result of Slutstky's Theorem, $T^{-1}\hat{\pi}'Z'Z\pi \longrightarrow_p \pi' M_{ZZ}\pi$

Now we analyze the second term and exploit the assumption that η is independent of Z.
$(\frac{1}{T}\sum_{t=1}^{T} z_t\eta_t) \longrightarrow_p E(z_t\eta_t) = E(z_t)E(\eta_t) = 0$ by law of large numbers because $z_t\eta_t$ is iid and $Var(z_t\eta_t) = E(z_t\eta_t\eta_t' z_t) - E(z_t\eta_t)^2 = \tau^2 E(z_t z_t')$ is finite.
As a result of Slutsky's Theorem, $T^{-1}\hat{\pi}'Z'\eta \longrightarrow_p \pi' * 0 = 0$.

As a result of Slutsky's Theorem, $(T^{-1}\hat{\pi}'Z'Z\pi + T^{-1}\pi'Z'\eta) \longrightarrow_p \pi' M_{ZZ}\pi + 0 = \pi' M_{ZZ}\pi$.
It is assumed that $\pi' M_{ZZ}\pi \neq 0$ and thus its inverse exists.
As a result of the Continuous Mapping Theorem, $(T^{-1}\hat{\pi}'Z'Z\pi + T^{-1}\pi'Z'\eta)^{-1} \longrightarrow_p (\pi' M_{ZZ}\pi)^{-1}$.

Now we analyze the asymptotic distribution of the third term.
$E(z_t\varepsilon_t) = E(z_t)E(\varepsilon_t) = 0$ by independence and $z_t\varepsilon_t$ is iid.
$Var(z_t\varepsilon_t) = E(z_t\varepsilon_t\varepsilon_t' z_t') - 0 = \sigma^2 E(z_t z_t') = \sigma^2 M_{zz}$.
As a result of the Central Limit Theorem, $\sqrt{T}(T^{-1}Z'\varepsilon) \longrightarrow_d N(0, Var(z_t\varepsilon_t)) = N(0, \sigma^2 M_{zz})$.
By Slutsky's Theorem, $\sqrt{T}(T^{-1}\hat{\pi}'Z'\varepsilon) \longrightarrow_d \pi' N(0, \sigma^2 M_{zz}) \sim N(0, \sigma^2 \pi' M_{zz}\pi)$.

Lastly we apply Slutsky's Theorem to the entire expression:
$\sqrt{T}(\hat{\beta}_{IV} - \beta) \longrightarrow_d (\pi' M_{zz}\pi)^{-1} N(0, \sigma^2 \pi' M_{zz}\pi) \sim N(0, (\pi' M_{zz}\pi)^{-1}\sigma^2\pi' M_{zz}\pi(\pi' M_{zz}\pi)^{-1}) \sim N(0, \sigma^2(\pi' M_{zz}\pi)^{-1})$.

Now we analyze the asymptotic distribution of $\hat{\beta}_{2S}$.

$$
\begin{aligned}
\hat{\beta}_{2S} &= (\hat{x}'\hat{x})^{-1}\hat{x}'y \\
&= (\hat{x}'\hat{x})^{-1}\hat{x}'(x\beta+\varepsilon) \\
&= (\hat{x}'\hat{x})^{-1}\hat{x}'((\hat{x}+Z(\pi-\hat{\pi})+\eta)\beta+\varepsilon) \\
&= (\hat{x}'\hat{x})^{-1}\hat{x}'\hat{x}\beta+(\hat{x}'\hat{x})^{-1}(\hat{x}'Z(\pi-\hat{\pi})\beta+\hat{x}'\eta\beta+\hat{x}'\varepsilon) \\
\Rightarrow \hat{\beta}_{2S}-\beta &= (\hat{x}'\hat{x})^{-1}(\hat{x}'Z(\pi-\hat{\pi})\beta+\hat{x}'\eta\beta+\hat{x}'\varepsilon) \\
&= (T^{-1}\hat{x}'\hat{x})^{-1}(T^{-1}(\hat{x}'Z(\pi-\hat{\pi})\beta+\hat{x}'\eta\beta+\hat{x}'\varepsilon) \\
\Rightarrow \sqrt{T}(\hat{\beta}_{2S}-\beta) &= (T^{-1}\hat{x}'\hat{x})^{-1}\sqrt{T}(T^{-1}\hat{x}'Z(\pi-\hat{\pi})\beta+T^{-1}\hat{x}'\eta\beta+T^{-1}\hat{x}'\varepsilon) \\
&= (T^{-1}(Z\hat{\pi})'(Z\hat{\pi}))^{-1}\sqrt{T}(T^{-1}(Z\hat{\pi})'Z(\pi-\hat{\pi})\beta+T^{-1}(Z\hat{\pi})'\eta\beta+T^{-1}(Z\hat{\pi})'\varepsilon) \\
&= (T^{-1}\hat{\pi}'Z'Z\hat{\pi})^{-1}[(T^{-1}\hat{\pi}'Z'Z)\sqrt{T}(\pi-\hat{\pi})\beta+\sqrt{T}(T^{-1}(\hat{\pi}'Z'\eta\beta+\hat{\pi}'Z'\varepsilon))] \\
&= (T^{-1}\hat{\pi}'Z'Z\hat{\pi})^{-1}[(T^{-1}\hat{\pi}'Z'Z)\sqrt{T}(\pi-\hat{\pi})\beta]+(T^{-1}\hat{\pi}'Z'Z\hat{\pi})^{-1}[\hat{\pi}'\sqrt{T}(T^{-1}Z'\eta\beta+T^{-1}Z'\varepsilon)] \\
&= T_1+T_2
\end{aligned}
$$

T_1 and T_2 are asymptotically independent because the limiting distribution of T_1 depends on $\hat{\pi}$, which is statistically independent of η and ε, the drivers of the limiting distribution of T_2. We analyze the distribution of each separately.

Using the previous calculations, we can use Slutsky's Theorem twice to show that $(T^{-1}\hat{\pi}'Z'Z\hat{\pi}) \longrightarrow_p \pi'M_{zz}\pi$, and as a result of the Continuous Mapping Theorem, $(T^{-1}\hat{\pi}'Z'Z\hat{\pi})^{-1} \longrightarrow_p (\pi'M_{ZZ}\pi)^{-1}$.

It is given that $\sqrt{T}(\hat{\pi}-\pi) \longrightarrow_d N(0,V)$.
This distribution implies that $\sqrt{T}(\pi-\hat{\pi}) = -\sqrt{T}(\hat{\pi}-\pi) \longrightarrow_d N(0,(-1)V(-1)) \sim N(0,V)$.

We now apply Slutsky's Theorem to T_1 to show that

$$
\begin{aligned}
&(T^{-1}\hat{\pi}'Z'Z\hat{\pi})^{-1}[(T^{-1}\hat{\pi}'Z'Z)\sqrt{T}(\pi-\hat{\pi})\beta] \\
&\qquad \longrightarrow_d (\pi'M_{ZZ}\pi)^{-1}\pi'M_{ZZ}\beta N(0,V) \\
&\qquad \sim N(0,(\pi'M_{ZZ}\pi)^{-1}\pi'M_{ZZ}\beta V\beta M_{ZZ}\pi(\pi M_{ZZ}\pi)^{-1})
\end{aligned}
$$

We now analyze the limiting distribution of T_2 and first apply the Central Limit Theorem to $\sqrt{T}(T^{-1}Z'\eta\beta+T^{-1}Z'\varepsilon)$.
η and ε are iid and jointly independent so $(Z'\eta\beta+Z'\varepsilon)$ is iid.
$E(T^{-1}Z'\eta\beta+T^{-1}Z'\varepsilon)=0$ because Z is orthogonal to η for the first-stage to the well specified and orthogonal to ε for it to be a valid instrument.
We now calculate the variance and confirm that it is finite.

$$
\begin{aligned}
Var(T^{-1}Z'\eta\beta+T^{-1}Z'\varepsilon) &= Var(Z'\epsilon)+Var(Z'\eta\beta)+2Cov(Z'\varepsilon,Z'\eta\beta) \\
&= E(Z'Z)Var(\varepsilon)+\beta^2Var(\eta)E(Z'Z)+2\beta E(Z'\varepsilon\eta'Z) \\
&= \sigma^2M_{ZZ}+\tau^2M_{ZZ}+2M_{ZZ}\beta\gamma
\end{aligned}
$$

We thus apply the Central Limit Theorem to show that $\sqrt{T}(T^{-1}Z'\eta\beta + T^{-1}Z'\varepsilon) \longrightarrow_d N(0, (\sigma^2 + \beta^2\tau^2 + 2\gamma\beta)M_{ZZ})$.

We now apply Slutsky's Theorem to T_2 to show that

$$\begin{aligned}(T^{-1}\hat{\pi}'Z'Z\hat{\pi})^{-1}[\hat{\pi}'\sqrt{T}(T^{-1}Z'\eta\beta + T^{-1}Z'\varepsilon)] \\ &\longrightarrow_d (\pi' M_{ZZ}\pi)^{-1}\pi' N(0, (\sigma^2 + \beta^2\tau^2 + 2\gamma\beta)M_{ZZ}) \\ &\sim N(0, (\sigma^2 + \beta^2\tau^2 + 2\gamma\beta)[(\pi' M_{ZZ}\pi)^{-1}\pi']M_{ZZ}[\pi(\pi' M_{ZZ}\pi)^{-1}]) \\ &\sim N(0, (\sigma^2 + \beta^2\tau^2 + 2\gamma\beta)[(\pi' M_{ZZ}\pi)^{-1}])\end{aligned}$$

By asymptotic independence of T_1 and T_2,

$$\sqrt{T}(\hat{\beta}_{2S} - \beta) \longrightarrow_d N(0, A^{-1}BA^{-1} + A^{-1}(\sigma^2 + \beta^2\tau^2 + 2\gamma\beta))$$

where $A = \pi' M_{ZZ}\pi$ and $B = \pi' M_{ZZ}V M_{ZZ}\pi$.

Therefore, the limiting distribution of $\sqrt{T}(\hat{\beta}_{2S} - \beta)$ differs from $\sqrt{T}(\hat{\beta}_{IV} - \beta)$.

5 Additional Exercises

This section includes solutions to previous 240B exam questions that have not already been distributed. Excluded questions are those about maximum likelihood estimation from the 2002 and 2003 exam as well as recycled questions as seen in the 2005 exam and question 1D in 2006.

5.1 2003 Exam, 1D

Question: By the Continuous Mapping theorem, if $\hat{\theta}$ is root-n consistent and asymptotically normal for the scalar parameter θ_0, then its squared value, when multiplied by an appropriate function of n, should have a limiting chi-square distribution.

Answer: False. As we will show though the statement is true only for $\theta_0 = 0$.
If $\theta_0 \neq 0$ then the delta method implies that the asymptotic distribution of θ^2 is nomral, not chi-squared. Letting $g(\theta) = \theta^2$ with derivative $g'(\theta) = 2\theta$, and assuming

$$\sqrt{n}(\hat{\theta} - \theta_0) \longrightarrow_d N(0, V_0)$$

the delta method implies that

$$\sqrt{n}(\hat{\theta}^2 - \theta_0^2) \longrightarrow_d N(0, [g'(\theta)]^2 V_0) = N(0, 4\theta_0^2 V_0)$$

Thus, the squared value $\hat{\theta}^2$ is equal to θ_0^2 plus an asymptotically normal variable, and cannot be rescaled to be asymptotically chi-squared.

However, if $\theta_0 = 0$, then this statement is true if we scale $\hat{\theta}^2$ by n wherein $\sqrt{n}\hat{\theta} \longrightarrow_d N(0, V_0)$.

Recall that if $Z_N = \sqrt{N}(\theta_N - \theta_0) \longrightarrow_d N(0, I_p)$ then $T = Z'_N Z_N = N(\hat{\theta} - \theta_0)'(\hat{\theta} - \theta_0) \longrightarrow_d \chi^2_p$ where $p = dim(\hat{\theta_N})$

*Note that this statement comes directly from Professor Powell's lecture notes though he has a very important typo. Instead he has Σ listed as the variance-covariance matrix rather than I_p.

Therefore, if $\theta_0 = 0$, $\left(\frac{n}{V_0}\right)\hat{\theta}^2 \longrightarrow_d \chi^2_1$ because $\left(\sqrt{\frac{N}{V_0}}\right)\hat{\theta} \longrightarrow_d N(0,1)$.

5.2 2006 Exam, 1C

Question: If y_t is a stationary $AR(2)$ process with no intercept term - specifically, $y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \varepsilon_t$, where ε_t is an i.i.d process with zero mean and all moments finite - then an asymptotically valid test of the null hypothesis $H_0 : \beta_1 = 0$ can be based upon the first-order (non-constant adjusted) sample autocovariance of y_t and y_{t-1},

$$\hat{\gamma} = T^{-1} \sum_{t=1}^{T} y_t y_{t-1}$$

Assuming the relevant Central Limit Theorem applies to this average, the estimator $\hat{\gamma}$ will have an approximate normal distribution which is centered at zero under H_0, so a t-statistic can be constructed using $\hat{\gamma}$ to obtain a large sample test of this null hypothesis.

Answer: True. Immediately, we impose the null hypothesis that $\beta_1 = 0$ because of our interest in the asymptotic distribution of $\hat{\gamma}$ under the null. Doing so simplifies the analysis immensly to $y_t = \beta_2 y_{t-2} + \varepsilon_t$.

Our goal is to argue that $\hat{\gamma}$ will have an approximate normal distribution and that it will be centered at zero under H_0. We argue that it will an approximate normal distribution because we are granted the assumption that the relevant Central Limit Theorem applies to this average. As a result, $\sqrt{T}\,(\hat{\gamma} - \mu) \longrightarrow_d N\,(0, V_0)$ where $\mu = E(\hat{\gamma})$ and V_0 is the asymptotic variance.

Next we demonstrate that that approximate normal distribution is centered at zero because $\mu = 0$. In doing so we exploit the stationarity of y_t wherein $E(y_t) = E(y_{t-2})$ and $Cov(y_t, y_{t-1}) = Cov(y_{t-1}, y_{t-2}) = \gamma_1$.

We first exploit stationarity to show that evaluating μ reduces to evaluating $E(y_t y_{t-1})$:

$$
\begin{aligned}
\mu &= E(\hat{\gamma}) \\
&= E(T^{-1}\sum_{t=1}^{T} y_t y_{t-1}) \\
&= T^{-1}E(\sum_{t=1}^{T} y_t y_{t-1}) \\
&= T^{-1} * T * E(y_t y_{t-1}) \\
&= E(y_t y_{t-1})
\end{aligned}
$$

We now show that $E(y_t) = 0$ and $\beta_2 \neq 1$:

$$
\begin{aligned}
E(y_t) &= E(\beta_2 y_{t-2}) + E(\varepsilon_t) \\
&= \beta_2 E(y_{t_2}) + 0 \\
&= \beta_2 E(y_t) \\
&\Rightarrow (1-\beta_2)E(y_t) = 0 \\
&\Rightarrow E(y_t) = \frac{0}{1-\beta_2} = 0 \\
&\Rightarrow \beta_2 \neq 1
\end{aligned}
$$

Finally we show that $E(y_t y_{t-1}) = 0$:

$$
\begin{aligned}
E(y_t y_{t-1}) &= Cov(y_t, y_{t-1}) + E(y_t)E(y_{t-1}) \\
&= \gamma_1 + 0 \\
&= Cov(\beta_2 y_{t-2} + \varepsilon_t, y_{t-1}) \\
&= \beta_2 Cov(y_{t-2}, y_{t-1}) + Cov(y_{t-1}, \varepsilon_t) \\
&= \beta_2 \gamma_1 + 0 \\
&\Rightarrow (1-\beta_2)\gamma_1 = 0 \\
&\Rightarrow \gamma_1 = 0
\end{aligned}
$$

As a result, $\sqrt{T}\hat{\gamma}$ is a normalized average of mean-zero stationary random variables, which will have a limiting normal distribution that is centered at zero.

Given an estimator of the asymptotic variance of $\sqrt{T}\hat{\gamma}$, which is V_0 in the expression above, a t-statistic can be constructed using $\hat{\gamma}$ to obtain a large sample test of this null hypothesis because asymptotically it will be a standard normal random variable under H_0.

5.3 2006 Exam, 2

Question: Suppose $\hat{\theta}$ is an asymptotically normal estimator of a 3-dimensional parameter $\theta = (\theta_1, \theta_2, \theta_3)'$, which has the asymptotic distribution

$$\sqrt{N}(\hat{\theta} - \theta) \longrightarrow_d N(0, V)$$

Suppose that $\hat{\theta} = (1, -1, -2)'$ is the realized value of this estimator, and that a consistent estimator $\hat{V}$ of V has the realized value

$$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 50 & 0 \\ 0 & 0 & 100 \end{pmatrix}$$

where it is assumed that the sample $N = 597$ is large enough so that the normal approximation is accurate for this problem.
Use these results to test the joint null hypothesis $H_0 : \theta_1^2 + \theta_2^2 = 1$ and $\theta_3^2 = 1$, against the alternative that one or both of these restrictions fail, at an asymptotic 5% level.

Answer: We approximate a Wald statistic, which converges in distribution to a chi-squared distribution, and fail to reject H_0 based on this test statistic.

The general form of the Wald statistic for $H_0 : \hat{\theta} = \theta_0$ against $H_1 : \hat{\theta} \neq \theta_0$ is $W_n = n(\hat{\theta} - \theta_0)'(\hat{V}_0)^{-1}(\hat{\theta} - \theta_0)$ where $\hat{V}_0$ is the asymptotic covariance matrix estimate of $\hat{\theta}$.

We first write the hypotheses so that $\theta_0 = 0$ for each one. We then stack them into the matrix γ:

$$\gamma = g(\theta) = \begin{pmatrix} \theta_1^2 + \theta_2^2 - 1 \\ \theta_3^2 - 1 \end{pmatrix}$$

Using our estimate for $\hat{\theta} = (1, -1, -2)'$, $g(\hat{\theta}) = (1, 3)'$ and by construction $g(\theta) = (0, 0)'$ so $g(\hat{\theta}) - g(\theta) = (1, 3)'$.

By the Delta Method, $\sqrt{N}(\hat{\gamma} - \gamma_0) \longrightarrow_d N(0, V_0)$ where $V_0 = GVG'$. V is as previously defined and it is given that $\hat{V} \longrightarrow_p V$. $G = \frac{\partial g}{\partial \theta'}$, and by the continuous mapping theorem $\hat{G} \longrightarrow_p G$ where

$$\hat{G} = \frac{\partial g}{\partial (\theta_1, \theta_2, \theta_3)'} = \begin{pmatrix} 2\hat{\theta_1} & 2\hat{\theta_2} & 0 \\ 0 & 0 & 2\hat{\theta_3} \end{pmatrix} = \begin{pmatrix} 2 & -2 & 0 \\ 0 & 0 & -4 \end{pmatrix}$$

We thus compute the Wald statistic, $W_n = n(\hat{\gamma} - \gamma_0)'(\hat{V}_0)^{-1}(\hat{\gamma} - \gamma_0)$ where $\hat{G}\hat{V}\hat{G}' \longrightarrow_p GVG'$.

Because G has full rank 2, GVG' is positive semi-definite. As a result we can use the Chloesky Decomposition to reexpress this statistic as $W_n = \sqrt{n}(\hat{\gamma} - \gamma_0)'\hat{V_0}^{-\frac{1}{2}\prime}\hat{V_0}^{-\frac{1}{2}}\sqrt{n}(\hat{\gamma} - \gamma_0) = Z'Z$ where $Z = \hat{V_0}^{-\frac{1}{2}}\sqrt{n}(\hat{\gamma} - \gamma_0)$. Using the previously established result of the Delta Method, $Z \longrightarrow_d N(0, V_0^{-\frac{1}{2}} V_0 V_0^{-\frac{1}{2}\prime}) = N(0, I_2)$. As a result, $W_n = Z'Z \longrightarrow_d \chi_2^2$.

$$
\begin{aligned}
(\hat{G}\hat{V}\hat{G}')^{-1} &= \left(\begin{pmatrix} 2 & -2 & 0 \\ 0 & 0 & -4 \end{pmatrix} \begin{pmatrix} 50 & 0 & 0 \\ 0 & 50 & 0 \\ 0 & 0 & 100 \end{pmatrix} \begin{pmatrix} 2 & 0 \\ -2 & 0 \\ 0 & -4 \end{pmatrix} \right)^{-1} \\
&= \left(\begin{pmatrix} 100 & -100 & 0 \\ 0 & 0 & -400 \end{pmatrix} \begin{pmatrix} 2 & 0 \\ -2 & 0 \\ 0 & -4 \end{pmatrix} \right)^{-1} \\
&= \begin{pmatrix} 400 & 0 \\ 0 & 1600 \end{pmatrix}^{-1} = \left(400 \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix} \right)^{-1} \\
&= \frac{1}{400} \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{4} \end{pmatrix} \\
\gamma'(\hat{G}\hat{V}\hat{G}')^{-1}\gamma &= \frac{1}{400} \begin{pmatrix} 1 & 3 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{4} \end{pmatrix} \begin{pmatrix} 1 \\ 3 \end{pmatrix} \\
&= \frac{1}{400}(1 + \frac{9}{4}) = \frac{9}{1600} \\
W_n = n\gamma'(\hat{G}\hat{V}\hat{G}')^{-1}\gamma &= \frac{(9)(597)}{1600} \\
&< \frac{(10)(600)}{1600} < \frac{60}{16} \\
&< 4
\end{aligned}
$$

The 5% critical value for χ^2_2 is 5.99, which exceeds our approximated Wald statistic by a non-negligble amount.

Therefore we fail to reject H_0.

2007 Exam Solutions

Jeffrey Greenbaum

March 2007

Contents

1 Question 1A

Question: Suppose that an IV regression of y_i on a scalar endogenous regressor x_{i1} and a vector x_{i2} of exogenous regressors, using an instrument vector z_i that includes the x_{i2} components, yields a coefficient on x_{i1} of 2.2. If, instead, x_{i1} is taken to be the dependent variable, and an IV fit of x_{i1} on y_i and x_{i2} is calculated using the same instruments z_i, then the IV estimate of the coefficient on y_i will be positive.

Answer: True. We estimate both models with $\hat{\beta}_{2SLS}$ because it is equivalent to $\hat{\beta}_{IV}$ in the just-identified case and more straightforward to analyze in partitioned regression. Moreover doing so encompasses the over-identified case where $\dim(z_i) > \dim(x_{1i})$.

We first analyze $y_i = x_{1i}\beta_1 + x'_{2i}\beta_2 + \epsilon_i$ that we instrument with z_i and x_{2i}.

In the just-identified case $X'Z$ and $Z'X$ are square. Let $P_Z = Z(Z'Z)^{-1}Z'$. As a result

$$\begin{aligned}\hat{\beta}_{2SLS} &= (X'P_ZX)^{-1}X'P_Zy \\ &= (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y \\ &= (Z'X)^{-1}(Z'Z)(X'Z)^{-1}X'Z(Z'Z)^{-1}Z'y \\ &= (Z'X)^{-1}Z'y \\ &= \hat{\beta}_{IV}\end{aligned}$$

Thus in the just-identified case an equivalent instrumental variables estimtor comes from the regression of $\hat{y} = P_Zy$ on $\hat{X} = P_ZX$.

We estimate the scalar $\hat{\beta}_1$ by partitioned regression to isolate the relationship between x_{1i} on y_i. Let $P_2 = I - \hat{X}_2(\hat{X}_2'\hat{X}_2)^{-1}\hat{X}_2'$. Then

$$\hat{\beta}_1 = (\hat{X}_1'P_2\hat{X}_1)^{-1}(\hat{X}_1'P_2\hat{y})$$

P_2 is a projection matrix so we can express the denominator as $\hat{X}_1'P_2\hat{X}_1 = \hat{X}_1'P_2'P_2\hat{X}_1 = \|P_2\hat{X}_1\|$, which must be positive since it a norm that is invertible. Thus the numerator must be positive because it must have the same sign as $\hat{\beta}_1$.

We now analyze the reverse regression: $x_{1i} = y_i\delta_1 + x_{2i}'\delta_2 + \eta_i$ with the same instruments so we proceed as before with 2SLS. In the partitioned regression P_2 remains the same since we are again partialing out the effect of x_{2i}' on both y_i and x_{1i}. Then

$$\hat{\delta_1} = (\hat{y}'P_2\hat{y})^{-1}(\hat{y}'P_2\hat{X}_1)$$

Like $\hat{\beta}_1$ the denominator can be reexpressed as $\|P_2\hat{y}\|$, which is a norm that must be positive. Moreover the numerator is the same as that of $\hat{\beta}_1$ because it is the transpose of a scalar; thus it is positive.

Therefore $\hat{\delta_1}$ is positive: $sgn(\hat{\beta}) = sgn(\hat{\delta}) = sgn(\hat{X_1}'M_2\hat{y}) = sgn(\hat{y}'M_2\hat{X}_1)$.

2 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 are $MA(1)$, i.e., $\varepsilon_t = u_t + \theta u_{t-1}$, where u_t is an i.i.d. sequence with zero mean, variance σ^2, and is independent of x_s for all t and s. For this model, the classical LS estimator will be inconsistent for β and γ when $|\gamma| < 1$, but an IV estimator using x_t and y_{t-2} as instrumental variables will consistently esimate these parameters.

Answer: True. First we show that the OLS estimator will be inconsistent and then that this IV estimator will be consistent. Note that throughout this question that x_t is an exogenous regressor because u_t is independent of x_s for all t and s.

The OLS estimator is inconsistent because y_{t-1} is an endogenous regressor. $\hat{\gamma}_{OLS}$ will thus be inconsistent because y_{t-1} is not uncorrelated with ε_t.

$$\begin{aligned} Cov(y_{t-1}, \varepsilon_t) &= Cov(y_{t-1}, u_t + \theta u_{t-1}) \\ &= Cov(y_{t-1}, u_t) + \theta Cov(y_{t-1}, u_{t-1}) \end{aligned}$$

Lagging the model one period, $y_{t-1} = x'_{t-1}\beta + \gamma y_{t-2} + \varepsilon_{t-1} = x'_{t-1}\beta + \gamma y_{t-2} + u_{t-1} + \theta u_{t-2}$. This representation makes it clear that a current or future shock does not affect y_{t-1} since they would be uncorrelated with any of the regressors or error terms. However in general, $\theta Cov(y_{t-1}, u_{t-1}) = \theta\sigma^2 \neq 0$ unless $\theta = 0$ since u_{t-1} is only correlated with u_{t-1} in the expression for y_{t-1}.

Through partitioned regression, $\hat{\beta}_{OLS}$ is inconsistent because its estimation is related to the endogenous regressor.

Thus, the OLS estimator for β and γ will not be consistent.

An IV estimator though can consistently estimate the parameters if we have a valid instrument for the endogenous regressor, y_{t-1}. In this question, we consider whether y_{t-2} as such a valid instrument.

For it to be a valid instrument, it must satisfy two conditions:

1. $Cov(y_{t-2}, y_{t-1}) \neq 0$
2. $Cov(y_{t-2}, \varepsilon_t) = 0$

1. $$\begin{aligned} Cov(y_{t-2}, y_{t-1}) &= Cov(y_{t-2}, x'_{t-1}\beta + \gamma y_{t-2} + \varepsilon_{t-1}) \\ &= Cov(y_{t-2}, x_{t-1})'\beta + \gamma Cov(y_{t-2}, y_{t-2}) + Cov(y_{t-2}, \varepsilon_{t-1}) \end{aligned}$$

$\gamma Cov(y_{t-2}, y_{t-2}) = \gamma Var(y_{t-2}) \neq 0$ in general unless $\gamma = 0$.
Moreover $Cov(y_{t-2}, \varepsilon_{t-1}) = Cov(y_{t-1}, \varepsilon_t) \neq 0$ in general unless $\theta = 0$ as previously discussed.
We thus expect that in general, $Cov(y_{t-2}, y_{t-1}) \neq 0$, unless say $\gamma = \theta = 0$.

2. $$\begin{aligned} Cov(y_{t-2}, \varepsilon_t) &= Cov(y_{t-2}, u_t + \theta u_{t-1}) \\ &= Cov(y_{t-2}, u_t) + Cov(y_{t-2}, \theta u_{t-1}) \end{aligned}$$

As we previously discussed, both of these terms are zero because future shocks are not related to the output of later periods.

Therefore absent some multicollinearity problem between y_{t-2} and x, y_{t-2} is a valid instrument for y_{t-1}, in contrast to the case with $AR(1)$ errors. As a result, an IV estimator using x_t and y_{t-2} as instruments for x_t and y_{t-1} will consistently estimate the parameters, β and γ.

3 Question 1C

Question: For a balanced panel data regression model with individual fixed effects, $y_{it} = x'_{it}\beta + \alpha_i + \varepsilon_{it}$ - where the α_i are not assumed to be uncorrelated with x_{it}, but the error terms ε_{it} are i.i.d. and independent of α_i and x_{it}, with $E(\varepsilon_{it}) = 0$ and $Var(\varepsilon_{it}) = \sigma^2$ - suppose that only the number of time periods T tends to infinity, while the number of individuals N stays fixed. Then the "fixed effect" estimator for β will be consistent as $T \to \infty$ provided the regressors and individual indicator variables are not asymptotically multicollinear. Furthermore, if $\hat{\sigma^2} = (NT)^{-1}(y_{it} - \hat{\alpha}_i - x'_{it}\hat{\beta}_{LS})^2$ is the (biased) LS estimator of σ^2, then the usual LS formulae for the standard errors of $\hat{\beta}_{LS}$ (replacing the unknown σ^2 by $\hat{\sigma}^2$) will be asymptotically valid.

Answer: True. We proceed by showing separately that both parts of the statement are true.

The Fixed Effects estimator is the classical least squares regression of y_{it} on x_{it} and N binary variables. It is given that the assumptions of the classical regression model are satisfied. As a result, for a fixed N, $\hat{\beta}_{OLS}$ and $\hat{\alpha}_i$ for $i = 1, ..., N$ are unbiased, and assuming that there is no asymptotic multicollinearity, then they are also consistent and asymptotically normal with the usual form for the asymptotic covariance matrix.

Finally the usual biased MLE estimator $\hat{\sigma}^2$ of σ^2 is consistent since, as $T \to \infty$ when N is fixed, the ratio of $\hat{\sigma}^2$ approaches the unbiased and consistent s^2. We can see this by analyzing the ratio in the limit.

$$\frac{\hat{\sigma}^2}{s^2} = \frac{NT - (N + K)}{NT} \to 1.$$

4 Question 1D

Question: By the so-called "Delta Method", if $\hat{\theta}$ is root-n consistent and asymptotically normal for a vector parameter θ_0, then the difference bewteen the squared length of $\hat{\theta}$ and the squared length of θ_0, when multiplied by the square root of the sample size, will generally have a limiting normal distribution.

Answer: True. The statement is generally true and is false in the case of θ_0 = 0.

There exists a V_0 such that $\sqrt{N}(\hat{\theta} - \theta_0) \longrightarrow_d N(0, V_0)$.

We are interested in the asymptotic properties of $\hat{\theta}'\hat{\theta}$. Let $\hat{\delta} = g(\hat{\theta}) = \hat{\theta}'\hat{\theta}$.

Since g is a differentiable function of θ, then by the Delta Method, $\sqrt{N}(\hat{\delta} - \delta_0) \longrightarrow_d N(0, G_0 V_0 G'_0)$ where $\hat{\delta} = g(\hat{\theta}) = \hat{\theta}'\hat{\theta}$, $\delta_0 = g(\theta_0) = \theta'_0\theta_0$, and $G_0 = \frac{\partial g(\theta_0)}{\partial \theta} = 2\theta'_0$.

Accordingly, $\sqrt{N}\hat{\delta} \longrightarrow_d N(\delta_0, 2\theta'_0 V_0 2\theta_0)$.

Therefore the statement is generally true that the distribution will be normal because $2\theta_0' V_0 2\theta_0 = 4\theta_0' V_0 \theta_0 \neq 0$ in general.

However in the special case that $\theta_0 = 0$, that is a vector of zeros, then the estimator converges to a singular normal limiting distribution with zero variance.

We could normalize $\hat{\theta}$ so its asymptotic distribution is standard normal: $V_0^{-\frac{1}{2}}\sqrt{N}(\hat{\theta} - \theta_0) \longrightarrow_d N(0, I_k)$ where $k = dim(\hat{\theta})$. If $\theta_0 = 0$ then $V_0^{-\frac{1}{2}}\sqrt{N}\hat{\theta} \longrightarrow_d N(0, I_k)$.

Then $N\hat{\theta}' V_0^{-1}\hat{\theta} \longrightarrow_d \chi_k^2$.

Therefore in the case of $\theta_0 = 0$, a better asymptotic approxmation would be based on the difference between the squared length of $\hat{\theta}$ and the squared length of θ_0, when multiplied by the sample size. That is, we would not want to use the square root of the sample size. This limiting distribution would be chi-squared rather than gaussian.

5 Question 2

Question: Suppose a dependent variable y_i and two (scalar) regressors x_i and z_i satisfy a random coefficients model

$$y_i = \alpha_i + \beta_i x_i + \gamma_i z_i, \quad i = 1, ..., N,$$

where the coefficients $(\alpha_i, \beta_i, \gamma_i)$ are assumed to be i.i.d. and independent of x_i and z_i. In this framework, under the null hypothesis $H_0 : Var(\beta_i) = 0 = Var(\gamma_i)$, the mean values $\beta = E(\beta_i)$ and $\gamma = E(\gamma_i)$ can be estimated by a least-squares regression of y_i; in turn, this null hypothesis can be tested using R^2 from a least-squares regression of LS residuals $\hat{e}_i = (y_i - \hat{\alpha} - \hat{\beta} x_i - \hat{\gamma} z_i)^2$ on functions of the regressors.

Given a sample of size $N = 500$, derive the algebraic form of all of the regressors in this "squared residual regression", and give a numerical value for the critical value C for an (asymptotic) 5 $\%$ test of homoskedasticity using the second-stage R^2. i.e., the value for which H_0 will be rejected if $R^2 > C$ with asymptotic size 5 $\%$.

Answer: As we will derive the squared residual regression has a constant and 5 regressors: x_i, z_i, $x_i z_i$, x_i^2, and z_i^2. The critical value is $C = 0.02214$.

We seek to reexpress our model so that we can collect all disturbances in one term and that there is a null hypothesis such that it satisfies the assumptions of the classical regression model, namely homoskedasticity.

First, we reexpress $y_i = \alpha_i + \beta_i x_i + \gamma_i z_i$ as $y_i = \alpha + \beta x_i + \gamma z_i + \varepsilon_i$.

By setting these two expressions equal to each other, $\varepsilon_i = (\alpha_i - \alpha) + (\beta_i - \beta) x_i + (\gamma_i - \gamma) z_i$.

Note that $E(\varepsilon_i) = E[(\alpha_i - \alpha) + (\beta_i - \beta)x_i + (\gamma_i - \gamma)z_i] = 0$ by the assumptions that $E(\beta_i) = \beta$, $E(\gamma_i) = \gamma$, and the additional assumption that $E(\alpha_i) = \alpha$.

As a result $E(\varepsilon_i^2) = Var(\varepsilon_i)$, which equals

$$\begin{aligned}
Var(\varepsilon_i) &= Var[(\alpha_i - \alpha) + (\beta_i - \beta)x_i + (\gamma_i - \gamma)z_i] \\
&= Var(\alpha_i - \alpha) + Var((\beta_i - \beta)x_i)) + Var((\gamma_i - \gamma)z_i) + \\
&2Cov((\alpha_i - \alpha), (\beta_i - \beta)x_i) + 2Cov((\alpha_i - \alpha), (\gamma_i - \gamma)z_i) + 2Cov((\beta_i - \beta)x_i, (\gamma_i - \gamma)z_i) \\
&= \sigma_\alpha^2 + \sigma_\beta^2 x_i^2 + \sigma_\gamma^2 z_i^2 + 2\sigma_{\alpha\beta}x_i + 2\sigma_{\alpha\gamma}z_i + 2\sigma_{\beta\gamma}x_i z_i \\
&= \sigma_\alpha^2(1 + \frac{2\sigma_{\alpha\beta}}{\sigma_\alpha^2}x_i + \frac{2\sigma_{\alpha\gamma}}{\sigma_\alpha^2}z_i + \frac{2\sigma_{\beta\gamma}}{\sigma_\alpha^2}x_i z_i + \frac{\sigma_\beta^2}{\sigma_\alpha^2}x_i^2 + \frac{\sigma_\gamma^2}{\sigma_\alpha^2}z_i^2)
\end{aligned}$$

Accordingly we analyze $\varepsilon_i^2 = \sigma_\alpha^2 + \delta_2 x_i + \delta_3 z_i + \delta_4 x_i z_i + \delta_5 x_i^2 + \delta_6 z_i^2 + r_i$, where we assume that r_i is a random disturbance term and each δ corresponds with the previously derived coefficient. Moreover we we assume this model satisfies the classical regression assumptions, namely that the disturbance term is expectation zero and homoskedastic.

We test the null hypothesis that $H_0 : \delta_2 = \delta_3 = \delta_4 = \delta_5 = \delta_6 = 0$ because under the null ϵ is homoskedastic since $Var(\varepsilon_i) = \sigma_\alpha^2 \quad \forall i$.

By Breusch-Pagan (1979) we can test this hypothesis by the following three steps:

First we collect the least squared residuals, $\hat{e_i}$, and use ordinary least squares to estimate the parameters of $\hat{e_i} = \sigma_\alpha^2 + \delta_2 x_i + \delta_3 z_i + \delta_4 x_i z_i + \delta_5 x_i^2 + \delta_6 z_i^2 + r_i$.

Second we compute R^2.

Third we reject the null hypothesis at the asymptotic 5 $\%$ level if $NR^2 > \chi_5^2(5\%)$ where N is the sample size. That is we reject the hypothesis of homoskedasticity if $R^2 > \frac{\chi_5^2(5\%)}{N} = \frac{11.07}{500} = \frac{22.14}{1000} = 0.02214$. Note that we have 5 degrees of freedom because our null hypothesis tests that 5 parameters all equal zero.

6 Question 3

Question: A feasible GLS fit of the generalized regression model with $K = 3$ regressors yields the estimates $\hat{\beta} = (2, -2, -1)$ where the GLS covariance matrix $V = \sigma^2[X'\Omega^{-1}X]^{-1}$ is estimated as

$$\hat{V} = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

using consistent estimators of σ^2 and Ω. The sample size $N = 403$ is large enough so that it is reasonable to assume a normal approximation holds for the GLS estimator.

Use these results to test the null hypothesis $H_0 : \beta_1^2 + \beta_2^2 + \beta_3^2 = 1$ at an asymptotic 5 $\%$ level.

Answer: We fail to reject the null hypothesis by using the delta method to construct an approximate t-statistic.

Recall that $\sqrt{N}(\hat{\beta}_{GLS} - \beta) \longrightarrow_d N(0, V)$ where $V = \sigma^2(X'\Omega^{-1}X)^{-1}$. We are given a $\hat{V}$ such that $\hat{V} \longrightarrow_p V$.

We are interested in the limiting distribution of $\hat{\theta} = g(\hat{\beta}) = \|\hat{\beta}\|^2 = \hat{\beta}'\hat{\beta}$, which we analyze by the Delta Method: $\sqrt{N}(\hat{\theta} - \theta) \longrightarrow_d N(0, GVG')$ where

$$\begin{aligned} G &= \frac{\partial g(\beta)}{\partial \beta'} \\ &= \frac{\partial(\beta'\beta)}{\partial \beta'} \\ &= 2\beta' = 2(\beta_1, \beta_2, \beta_3)' \end{aligned}$$

Therefore an approximate test statistic is $\frac{\hat{\theta}-\theta}{\sqrt{GVG'}} \overset{A}{\sim} N(0, 1)$.

We estimate G with $\hat{G}$ because $\hat{G} \longrightarrow_p G$ by the Continuous Mapping Theorem where

$$\begin{aligned} \hat{G} &= 2(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) \\ &= 2(2, -2, -1) \\ &= (4, -4, -2) \end{aligned}$$

By Slutsky's Theorem $\hat{G}\hat{V}\hat{G}' \longrightarrow_p GVG'$ where

$$\begin{aligned} \hat{G}\hat{V}\hat{G}' &= \begin{pmatrix} 4, & -4, & -2 \end{pmatrix} * \begin{pmatrix} 2 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} * \begin{pmatrix} 4 \\ -4 \\ -2 \end{pmatrix} \\ &= \begin{pmatrix} 6, & -4, & 2 \end{pmatrix} \begin{pmatrix} 4 \\ -4 \\ -2 \end{pmatrix} \\ &= 36 \end{aligned}$$

Thus to test $H_0 : \theta = 1$ against a two-sided alternative, the t-statistic is

$$\frac{\hat{\theta} - \theta_0}{\sqrt{\hat{G}\hat{V}\hat{G}'}} = \frac{[2^2 + (-2)^2 + (-1)^2] - 1}{\sqrt{36}} = \frac{8}{6} < 1.34$$

which is less than 1.96, the upper 97.5% critical value of a standard normal. We thus fail to reject H_0 at an asymptotic 5% level. As is often the case, the sample size $N = 403$ does not directly

figure into the solution, though it is implicit in the estimate $\hat{V}$ of the approximate covariance matrix of $\hat{\beta}$.

An alternative solution entails deriving an approximate Wald statistic though it is simpler to compute a t-statstic since there is only one degree of freedom.

7 Question 4

Question: If y_t is an $MA(1)$ process with zero mean, i.e., if

$$y_t = \varepsilon_t + \theta\varepsilon_{t-1}, \quad \varepsilon_t WN(\sigma^2)$$

and if $\gamma(s) = Cov(\varepsilon_t, \varepsilon_{t-s})$ is the autocovariance function and $\rho(s) = \frac{\gamma(s)}{\gamma(0)}$ is the autocorrelation function of y_t, show that

$$-1 < c^L \leq \rho(1) \leq c^U < 1,$$

i.e., the first autocorrelation is strictly bounded away from -1 and 1, by calculating the maximum and minimum values c^U and c^L of $\rho(1)$ over all possible θ.

Answer: First we calculate $\rho(1) = \frac{\gamma(1)}{\gamma(0)}$.

$$\begin{aligned}
\gamma(0) &= Var(y_t) \\
&= Var(\varepsilon_t + \theta\varepsilon_{t-1}) \\
&= Var(\varepsilon_t) + Var(\theta\varepsilon_{t-1}) + 2Cov(\varepsilon_t, \theta\varepsilon_{t-1}) \\
&= \sigma^2 + \theta^2 Var(\varepsilon_{t-1}) + 0 \\
&= \sigma^2 + \theta^2\sigma^2 \\
&= \sigma^2(1+\theta^2) \\
\gamma(1) &= Cov(y_t, y_{t-1}) \\
&= Cov(\varepsilon_t + \theta\varepsilon_{t-1}, \varepsilon_{t-1} + \theta\varepsilon_{t-2}) \\
&= Cov(\varepsilon_t, \varepsilon_{t-1}) + Cov(\varepsilon_t, \theta\varepsilon_{t-2}) + Cov(\theta\varepsilon_{t-1}, \varepsilon_{t-1}) + Cov(\theta\varepsilon_{t-1}, \theta\varepsilon_{t-2}) \\
&= \theta Cov(\varepsilon_{t-1}, \varepsilon_{t-1}) \\
&= \theta Var(\varepsilon_{t-1}) \\
&= \theta\sigma^2
\end{aligned}$$

Note that $Cov(\varepsilon_t, \varepsilon_{t-1}) = 0$ because the shocks of different periods are assumed to be independent of one another in a white noise process. Moreover $Var(\varepsilon_s) = \sigma^2 \quad \forall s$.

Accordingly, $\rho(1) = \frac{\theta\sigma^2}{\sigma^2(1+\theta^2)} = \frac{\theta}{(1+\theta^2)}$.

Next we seek to show that this function is strictly bounded away from -1 and 1 by calculating its maximum and minimum value over all possible θ. We proceed by analyzing the first derivative of $\rho(1)$.

$$\begin{aligned}\frac{d\rho(1)}{d\theta} &= \frac{(1+\theta^2)(1)-(\theta)(2\theta)}{(1+\theta^2)^2} \\ &= \frac{1-\theta^2}{(1+\theta^2)^2} \\ &= \frac{(1+\theta)(1-\theta)}{(1+\theta^2)^2}\end{aligned}$$

This expression equals zero if $\theta = 1$ or $\theta = -1$.

At $\theta = 1$, $\rho(1) = \frac{1}{1+1^2} = \frac{1}{1+1} = \frac{1}{2}$.
At $\theta = -1$, $\rho(1) = \frac{-1}{1+(-1)^2} = \frac{-1}{1+1} = -\frac{1}{2}$.
At the extreme value on the right, as $\theta \to \infty$, $\rho(1) \to 0$.
At the extreme value on the left, as $\theta \to -\infty$, $\rho(1) \to 0$.

We confirm that the maximum value occurs at $\theta = 1$ and the minimum value occurs at $\theta = -1$ by the second derivative.

$$\begin{aligned}\frac{d^2\rho(1)}{d\theta^2} &= \frac{(1+\theta^2)^2(-2\theta)-(1-\theta^2)(2)(1+\theta^2)(2\theta)}{(1+\theta^2)^4} \\ &= \frac{-2\theta(1+\theta^2)^2-4\theta}{(1+\theta^2)^4} \\ &= \frac{(1+\theta)(1-\theta)}{(1+\theta^2)^2}\end{aligned}$$

Therefore, $|\rho(1)|$ is strictly bounded by 1. Specifically, $-1 < -\frac{1}{2} \leq \rho(1) \leq \frac{1}{2} < 1$.