

Day 1

Ordinary Least Squares and GLS

© A. Colin Cameron
Univ. of Calif.- Davis

Advanced Econometrics
Bavarian Graduate Program in Economics

*Based on A. Colin Cameron and Pravin K. Trivedi (2009,2010),
Microeconometrics using Stata (MUS), Stata Press.
and A. Colin Cameron and Pravin K. Trivedi (2005),
Microeconometrics: Methods and Applications (MMA), C.U.P.*

July 22-26, 2013

1. Introduction

- OLS for the linear model is the building block for other regression.
- Here we provide
 - ▶ model in matrix notation
 - ▶ statistical properties
 - ▶ hypothesis testing
 - ▶ simulations to show consistency and asymptotic normality.
- Additionally
 - ▶ More efficient FGLS with heteroskedastic data

Overview

- 1 Introduction
- 2 OLS: Data example
- 3 OLS: Matrix Notation
- 4 OLS: Properties
- 5 GLS: Generalized Least Squares
- 6 Tests of linear hypotheses (Wald tests)
- 7 Simulations: OLS Consistency and Asymptotic Normality
- 8 Stata commands
- 9 Appendix: OLS in matrix notation example

2. Data Example: OLS for doctor visits

- Cross-section data on individuals (from MUS chapter 10).
 - ▶ Dependent variable `docvis` is a count. Here do OLS (later Poisson).
 - ▶ Begin with data description and summary statistics.

```
. use mus10data.dta, clear
. quietly keep if year02==1
. describe docvis private chronic female income
```

variable name	storage type	display format	value label	variable label
<code>docvis</code>	int	%8.0g		number of doctor visits
<code>private</code>	byte	%8.0g		= 1 if private insurance
<code>chronic</code>	byte	%8.0g		= 1 if a chronic condition
<code>female</code>	byte	%8.0g		= 1 if female
<code>income</code>	float	%9.0g		Income in \$ / 1000

```
. summarize docvis private chronic female income
```

variable	Obs	Mean	Std. Dev.	Min	Max
<code>docvis</code>	4412	3.957389	7.947601	0	134
<code>private</code>	4412	.7853581	.4106202	0	1
<code>chronic</code>	4412	.3263826	.4689423	0	1
<code>female</code>	4412	.4718948	.4992661	0	1
<code>income</code>	4412	34.34018	29.03987	-49.999	280.777

- OLS regression with default standard errors: assumes i.i.d error.

```
. * OLS regression with default standard errors
. regress docvis private chronic female income
```

Source	SS	df	MS
Model	35771.7188	4	8942.92971
Residual	242846.27	4407	55.1046676
Total	278617.989	4411	63.1643594

Number of obs = 4412
 F(4, 4407) = 162.29
 Prob > F = 0.0000
 R-squared = 0.1284
 Adj R-squared = 0.1276
 Root MSE = 7.4233

docvis	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
private	1.916263	.2881911	6.65	0.000	1.351264	2.481263
chronic	4.826799	.2419767	19.95	0.000	4.352404	5.301195
female	1.889675	.2286615	8.26	0.000	1.441384	2.337967
income	.016018	.004071	3.93	0.000	.0080367	.0239993
_cons	-.5647368	.2746696	-2.06	0.040	-1.103227	-.0262465

- Overall fit poor as $R^2 = 0.13$. Often the case for cross-section data.
- Yet all regressors are stat. significant and have large impact.
 - ▶ For income: annual income \uparrow \$10,000 \Rightarrow income \uparrow 10 units
 \Rightarrow docvis \uparrow $10 \times 0.016 = 0.16$.

- OLS regression with robust standard errors for OLS estimator
 - ▶ preferred at this permits model error to be heteroskedastic

```
. * OLS regression with robust standard errors
. regress docvis private chronic female income, vce(robust)
```

Linear regression

Number of obs = 4412
 F(4, 4407) = 107.01
 Prob > F = 0.0000
 R-squared = 0.1284
 Root MSE = 7.4233

docvis	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
private	1.916263	.2347443	8.16	0.000	1.456047	2.37648
chronic	4.826799	.3001866	16.08	0.000	4.238283	5.415316
female	1.889675	.2154463	8.77	0.000	1.467292	2.312058
income	.016018	.005606	2.86	0.004	.0050275	.0270085
_cons	-.5647368	.2069188	-2.73	0.006	-.9704017	-.159072

- Same coefficient estimates. Different standard errors.

```

. * Comparison of standard errors
. quietly regress docvis private chronic female income

. estimates store DEFAULT

. quietly regress docvis private chronic female income, vce(robust)

. estimates store ROBUST

. estimates table DEFAULT ROBUST, b(%9.4f) se stats(N r2 F)

```

Variable	DEFAULT	ROBUST
private	1.9163	1.9163
	0.2882	0.2347
chronic	4.8268	4.8268
	0.2420	0.3002
female	1.8897	1.8897
	0.2287	0.2154
income	0.0160	0.0160
	0.0041	0.0056
_cons	-0.5647	-0.5647
	0.2747	0.2069
N	4412.0000	4412.0000
r2	0.1284	0.1284
F	162.2899	107.0104

Legend: b/se

- The preferred heteroskedastic-robust standard errors are within 25% of default, sometimes more and sometimes less.

- Hypothesis tests can be implemented using Stata command `test`

$$H_0 : \beta_{\text{private}} = 0, \beta_{\text{chronic}} = 0$$

$$H_a : \text{at least one of } \beta_{\text{private}} \neq 0, \beta_{\text{chronic}} \neq 0.$$

- Stata post-estimation command `test` yields

```
. * wald test of restrictions
. quietly regress docvis private chronic female income, vce(robust) noheader
. test (private = 0) (chronic = 0)

( 1) private = 0
( 2) chronic = 0

      F( 2, 4407) = 165.11
      Prob > F = 0.0000
```

- Reject H_0 at level 0.05 since $p < 0.05$
or $165.11 > F_{.05}(2, 4407) = 3.00$ using `invFtail(2, 4407, .05)`.

3. OLS: Definition in matrix notation

- For the i^{th} observation

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + u_i$$

- ▶ Usually $x_{1i} = 1$ (an intercept).
- Introduce vector and matrix representation.
 - ▶ Regressor vector \mathbf{x}_i and parameter vector $\boldsymbol{\beta}$ are $K \times 1$ column vectors.

$$\begin{matrix} \mathbf{x}_i \\ (K \times 1) \end{matrix} = \begin{bmatrix} x_{1i} \\ \vdots \\ x_{Ki} \end{bmatrix} \quad \text{and} \quad \begin{matrix} \boldsymbol{\beta} \\ (K \times 1) \end{matrix} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}.$$

$$\mathbf{x}_i' \boldsymbol{\beta} = \begin{bmatrix} x_{1i} & \cdots & x_{Ki} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix} = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki}$$

- ▶ Note that all vectors are defined to be column vectors
- For the i^{th} observation

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + u_i.$$

- Now combine all N observations from sample $\{(y_i, \mathbf{x}_i), i = 1, \dots, N.\}$
- The linear regression model is

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} \mathbf{x}'_1 \boldsymbol{\beta} \\ \vdots \\ \mathbf{x}'_N \boldsymbol{\beta} \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ u_N \end{bmatrix}$$

- This is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

where

$$\underset{(N \times 1)}{\mathbf{y}} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \quad \underset{(N \times K)}{\mathbf{X}} = \begin{bmatrix} \mathbf{x}'_1 \\ \vdots \\ \mathbf{x}'_N \end{bmatrix} \quad \underset{(N \times 1)}{\mathbf{u}} = \begin{bmatrix} u_1 \\ \vdots \\ u_N \end{bmatrix} .$$

- The OLS estimator derived below is

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

OLS: matrix notation example

- Example: $N = 4$ with (x, y) equal to $(1, 1)$, $(2, 3)$, $(2, 4)$, and $(3, 4)$.
- Then \mathbf{y} is 4×1 and \mathbf{X} is 4×2 with

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 4 \\ 4 \end{bmatrix}; \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \mathbf{x}'_3 \\ \mathbf{x}'_4 \end{bmatrix} = \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ x_{13} & x_{23} \\ x_{14} & x_{24} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}.$$

- So (see appendix for detailed computation)

$$\hat{\beta}_{\text{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} 4 & 8 \\ 8 & 18 \end{bmatrix}^{-1} \begin{bmatrix} 12 \\ 27 \end{bmatrix} = \begin{bmatrix} 0 \\ 1.5 \end{bmatrix}$$

- Intercept $\hat{\beta}_1 = 0$ and slope coefficient $\hat{\beta}_2 = 1.5$.

Derivation of formula for OLS estimator

- The OLS estimator minimizes the sum of squared errors

$$Q(\beta) = \sum_{i=1}^N u_i^2 = \sum_{i=1}^N (y_i - \mathbf{x}_i' \beta)^2.$$

- The first-order conditions (f.o.c.) are

$$\frac{\partial Q(\beta)}{\partial \beta} = -2 \sum_{i=1}^N \mathbf{x}_i (y_i - \mathbf{x}_i' \beta) = -2 \mathbf{X}'(\mathbf{y} - \mathbf{X}\beta) = \mathbf{0}.$$

- Then

$$\begin{aligned} \mathbf{X}'(\mathbf{y} - \mathbf{X}\beta) &= \mathbf{0} && \text{from f.o.c.} \\ \Rightarrow \mathbf{X}'\mathbf{y} &= \mathbf{X}'\mathbf{X}\beta && K \text{ linear equations in } K \text{ unknowns } \beta \\ \Rightarrow \beta &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} && \text{if the inverse exists (i.e. } \text{rank}[\mathbf{X}] = K) \end{aligned}$$

- So

$$\hat{\beta}_{\text{OLS}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \left(\sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^N \mathbf{x}_i y_i.$$

4. OLS Properties: Summary

- $\hat{\beta}_{OLS}$ is always estimable, provided $\text{rank}[X] = K$.
- But properties of $\hat{\beta}_{OLS}$ depend on the true model
 - ▶ called the data generating process (d.g.p.)
- Essential result:
 - ▶ If the d.g.p. is correctly specified and the error u_i is uncorrelated with regressors \mathbf{x}_i
 - ▶ Then
 - (1) $\hat{\beta}$ is consistent for β
 - (2) $\hat{\beta}$ is normally distributed in large samples (“asymptotically”)
 - (3) Variance of $\hat{\beta}$ varies with assumptions on error u_i
 - ★ default: u_i are independent ($0, \sigma^2$)
 - ★ heteroskedastic: u_i are independent ($0, \sigma_i^2$)
 - ★ clustered: u_i are correlated within cluster, uncorrelated across cluster
 - ★ HAC: u_i are serially correlated (u_i are correlated with u_{i-1})

OLS Properties

- If the d.g.p. is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ then

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{OLS}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \\ &= \boldsymbol{\beta} + (\sum_i \mathbf{x}_i\mathbf{x}_i')^{-1} \sum_i \mathbf{x}_i u_i\end{aligned}$$

- So assumptions on \mathbf{x}_i and u_i are crucial.

OLS Finite Sample Properties

- If $\mathbf{u} \sim \mathcal{N}[\mathbf{0}, \Omega]$ and regressors \mathbf{X} are fixed (nonstochastic) then

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \\ &\sim \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \times \mathcal{N}[\mathbf{0}, \Omega] \\ &\sim \mathcal{N}[\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]\end{aligned}$$

- ▶ using linear transformation of the normal is normal
 $\mathbf{z} \sim \mathcal{N}[\boldsymbol{\mu}, \Omega] \implies \mathbf{A}\mathbf{z} + \mathbf{b} \sim \mathcal{N}[\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\Omega\mathbf{A}']$.

- We instead use asymptotic theory
 - ▶ this permits \mathbf{u} to be nonnormal distributed.
 - ▶ but does require a large sample so $N \rightarrow \infty$.

OLS Consistency

- Consistency

- Means that the probability limit (plim) of $\hat{\beta}$ equals β
- That is: $\lim_{N \rightarrow \infty} \Pr[|\hat{\beta} - \beta| < \varepsilon] = 1$ for any $\varepsilon > 0$.

- We have (using results below)

$$\begin{aligned}
 \text{plim } \hat{\beta} &= \text{plim} \{ \beta + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{u} \} \\
 &= \text{plim } \beta + \text{plim} \left\{ (\sum_i \mathbf{x}_i \mathbf{x}_i')^{-1} \sum_i \mathbf{x}_i u_i \right\} \\
 &= \text{plim } \beta + \text{plim} \left(\frac{1}{N} \sum_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \times \text{plim } \frac{1}{N} \sum_i \mathbf{x}_i u_i \\
 &= \beta + \left(\text{plim } \frac{1}{N} \sum_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \times \mathbf{0} \\
 &= \beta
 \end{aligned}$$

- $\text{plim} \{ \mathbf{A}_N \times \mathbf{b}_N \} = \text{plim } \mathbf{A}_N \times \text{plim } \mathbf{b}_N$ if the plim's are constants
- The plim's exist using laws of large numbers (as averages)
- For $\text{plim } \frac{1}{N} \sum_i \mathbf{x}_i u_i = \mathbf{0}$ the key assumption is $E[u_i | \mathbf{x}_i] = 0$.

OLS Limit Distribution

- $\hat{\beta}$ has limit distribution with all mass at β (since $\hat{\beta} \xrightarrow{p} \beta$).
 - ▶ To get a nondegenerate distribution inflate $\hat{\beta}$ by \sqrt{N} .
- Then limit normal distribution is

$$\begin{aligned}\sqrt{N}(\hat{\beta} - \beta) &= \left(\frac{1}{N} \sum_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \frac{1}{\sqrt{N}} \sum_i \mathbf{x}_i u_i \\ &\xrightarrow{d} \text{plim} \left(\frac{1}{N} \sum_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \times \mathcal{N}[\mathbf{0}, \mathbf{B}] \text{ for some } \mathbf{B} \\ &\xrightarrow{d} \mathcal{N}\left[\mathbf{0}, \text{plim} \left(\frac{1}{N} \sum_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \times \mathbf{B} \times \text{plim} \left(\frac{1}{N} \sum_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1}\right]\end{aligned}$$

- ▶ If $\mathbf{H}_N \xrightarrow{p} \mathbf{H}$ and $\mathbf{b}_N \xrightarrow{d} \mathcal{N}[\mu, \Omega]$ then $\mathbf{H}_N \mathbf{b}_N \xrightarrow{p} \mathcal{N}[\mathbf{H}\mu, \mathbf{H}\Omega\mathbf{H}']$
- ▶ $\frac{1}{\sqrt{N}} \sum_i \mathbf{x}_i u_i \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{B}]$ by a central limit theorem
- ▶ $\mathbf{B} = \text{plim} \left(\frac{1}{\sqrt{N}} \sum_i \mathbf{x}_i u_i\right) \left(\frac{1}{\sqrt{N}} \sum_i \mathbf{x}_i u_i\right)' = \text{plim} \frac{1}{N} \sum_i \sum_j u_i u_j \mathbf{x}_i \mathbf{x}_j'$

OLS Asymptotic Distribution

- All we need for theory is the previous result.
 - ▶ but rescale from $\sqrt{N}(\hat{\beta} - \beta)$ to $\hat{\beta}$ for “friendlier” looking results
 - ▶ drop plims and replace \mathbf{B} by a consistent estimate $\hat{\mathbf{B}}$
- The so-called “asymptotic distribution” is

$$\hat{\beta} \overset{a}{\sim} \mathcal{N} \left[\beta, \left(\sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \times N \hat{\mathbf{B}} \times \left(\sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \right]$$

- ▶ Usually $\mathbf{B} = \text{Var} \left[\frac{1}{\sqrt{N}} \mathbf{X}' \mathbf{u} \right] = \text{Var} \left[\frac{1}{\sqrt{N}} \sum_i \mathbf{x}_i u_i \right]$
- ▶ For independent heteroskedastic errors $\hat{\mathbf{B}} = \frac{1}{N} \sum_i \hat{u}_i^2 \mathbf{x}_i \mathbf{x}_i'$.

White Estimate of VCE

- Most often used: requires data to be independent over i .
- Then $\mathbf{B} = \text{plim } \frac{1}{N} \sum_i \sum_j u_i u_j \mathbf{x}_i \mathbf{x}_j' = \text{plim } \frac{1}{N} \sum_i u_i^2 \mathbf{x}_i \mathbf{x}_i'$.
- White (1980) showed that can use $\hat{\mathbf{B}} = \frac{1}{N} \sum_i \hat{u}_i^2 \mathbf{x}_i \mathbf{x}_i'$.
- Yields the heteroskedastic-consistent estimate of the variance-covariance matrix of the OLS estimator (VCE)

$$\hat{\mathbf{V}}_{\text{robust}}[\hat{\boldsymbol{\beta}}] = \left(\sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^N \hat{u}_i^2 \mathbf{x}_i \mathbf{x}_i' \left(\sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i' \right)^{-1}$$

- ▶ $\hat{u}_i = y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}$
- ▶ Leads to “heteroskedastic robust” or “robust” standard errors.
- ▶ In Stata this is option `vce(robust)` for cross-section commands

Other Estimates of VCE

- **Default:** Independent homoskedastic errors: $V[u_i | \mathbf{x}_i] = \sigma^2$

$$\widehat{V}[\widehat{\beta}] = s^2 \left(\sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i' \right)^{-1}; \quad s^2 = \frac{1}{N-K} \sum_i \widehat{u}_i^2$$

- ▶ Simplification as then $\mathbf{B} = \text{plim } \frac{1}{N} \sum_i u_i^2 \mathbf{x}_i \mathbf{x}_i' = \sigma^2 \text{plim } \sum_i \mathbf{x}_i \mathbf{x}_i'$

- **Cluster robust:** Errors correlated within cluster but independent across cluster.

$$\widehat{V}[\widehat{\beta}] = \left(\sum_{g=1}^G \mathbf{X}_g \mathbf{X}_g' \right)^{-1} \sum_{g=1}^G \mathbf{X}_g \widehat{\mathbf{u}}_g \widehat{\mathbf{u}}_g' \mathbf{X}_g' \left(\sum_{g=1}^G \mathbf{X}_g \mathbf{X}_g' \right)^{-1}.$$

- ▶ Here observations are stacked in cluster g as $\mathbf{y}_g = \mathbf{X}_g \beta + \mathbf{u}_g$.
 - ▶ In Stata this is option `vce(cluster id)` for cross-section commands
 - ▶ and is option `vce(robust)` for most xt panel commands.

- **Heteroskedasticity and autocorrelation (HAC) robust:** time series
 - ▶ Not covered here but extends White to an MA(q) error.

5. Generalized least squares (GLS) Overview

- OLS is efficient (best linear unbiased estimator) if errors are i.i.d. so that $V[\mathbf{u}|\mathbf{X}] = \sigma^2\mathbf{I}$.
 - ▶ In practice errors are rarely i.i.d.
- So we usually do OLS and obtain robust VCE that permits $V[\mathbf{u}|\mathbf{X}] \neq \sigma^2\mathbf{I}$
 - ▶ could be heteroskedastic robust, cluster-robust, HAC,
- More efficient feasible GLS (FGLS) assumes a model for $V[\mathbf{u}|\mathbf{X}]$
 - ▶ yields more precise estimates (smaller standard errors and bigger t-statistics)
 - ▶ but then obtain robust VCE that allows for misspecified model for $V[\mathbf{u}|\mathbf{X}]$.
 - ▶ called weighted LS or working matrix LS.

Generalized least squares (GLS)

- Suppose $V[\mathbf{u}|\mathbf{X}] = \Omega$ where Ω is known
 - ▶ and $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, $E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$ as before.
- The generalized least squares estimator is efficient:

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}\mathbf{y}.$$

- Derivation:

- ▶ Premultiply $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ by $\Omega^{-1/2}$ so

$$\Omega^{-1/2}\mathbf{y} = \Omega^{-1/2}\mathbf{X}\boldsymbol{\beta} + \Omega^{-1/2}\mathbf{u}.$$

- ▶ This model has i.i.d. errors since

$$V[\Omega^{-1/2}\mathbf{u}|\mathbf{X}] = E[(\Omega^{-1/2}\mathbf{u})(\Omega^{-1/2}\mathbf{u})'|\mathbf{X}] = \Omega^{-1/2}\Omega\Omega^{-1/2} = \mathbf{I}_N.$$
- ▶ Then GLS is OLS in this transformed model:

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

Feasible generalized least squares (FGLS)

- To implement GLS we need a consistent estimate of Ω . Assume a model for $\Omega = \Omega(\gamma)$, estimate $\hat{\gamma} \xrightarrow{P} \gamma$, and form $\hat{\Omega} = \Omega(\hat{\gamma}) \xrightarrow{P} \Omega$.
- The feasible GLS estimator (FGLS) is

$$\hat{\beta}_{\text{GLS}} = (\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\Omega}^{-1}\mathbf{y},$$

and then

$$\hat{\beta}_{\text{GLS}} \stackrel{a}{\sim} \mathcal{N} \left[\beta, (\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X})^{-1} \right].$$

- Examples:
 - Heteroskedasticity: $V[u_i^2 | \mathbf{x}_i] = \exp(\mathbf{z}_i' \gamma)$
 - Seemingly unrelated equations: $y_{ig} = \mathbf{x}_{ig}' \beta_g + u_{ig}$, $g = 1, \dots, G$.
 u_{ig} independent over i and homoskedastic with $\text{Cov}[u_{ig}, u_{ih}] = \sigma_{gh}$.
 - Systems of equations: SUR with $\beta_g = \beta$.
 - Panel data: random effects estimator.

Weighted least squares (WLS)

- Now do FGLS but allow for possibility that model for $V[\mathbf{u}|\mathbf{X}]$ is incorrectly specified
 - ▶ So then obtain robust VCE for FGLS.
- Distinguish between
 - ▶ the assumed (working) error variance matrix, denoted $\Sigma = \Sigma(\gamma)$ with estimate $\hat{\Sigma} = \Sigma(\hat{\gamma})$.
 - ▶ the true (unknown) error variance matrix Ω
- The weighted least squares (WLS) estimator is

$$\hat{\beta}_{\text{WLS}} = (\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{y}.$$

- Asymptotically $\hat{\beta}_{\text{WLS}} \overset{a}{\sim} \mathcal{N}[\beta, V[\hat{\beta}]]$ where robust VCE is

$$\hat{V}[\hat{\beta}] = (\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\hat{\Sigma}^{-1}\hat{\Omega}\hat{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{X})^{-1},$$

- ▶ for cross-section data $\hat{\Omega} = \text{Diag}[(y_i - \mathbf{x}_i'\hat{\beta}_{\text{WLS}})^2]$.

Hypothesis test of single restriction

- Consider test of a single restriction, for notational simplicity β

$$H_0 : \beta = \beta^*$$

$$H_a : \beta \neq \beta^*.$$

- A Wald test rejects H_0 if $\hat{\beta}$ differs greatly from β^* .
- Define $\sigma_{\hat{\beta}}$ to be the asymptotic standard deviation of $\hat{\beta}$. Then

$$\begin{aligned} \hat{\beta}_j &\overset{a}{\sim} \mathcal{N}[\beta, \sigma_{\hat{\beta}}^2] && \text{for unknown } \beta \\ \Rightarrow \frac{\hat{\beta} - \beta}{\sigma_{\hat{\beta}}} &\overset{a}{\sim} \mathcal{N}[0, 1] && \text{standardizing} \\ \Rightarrow z_j = \frac{\hat{\beta} - \beta^*}{\sigma_{\hat{\beta}}} &\overset{a}{\sim} \mathcal{N}[0, 1] && \text{under } H_0 : \beta = \beta^* \end{aligned}$$

- To implement this, replace $\sigma_{\hat{\beta}}$ by $s_{\hat{\beta}}$, the standard error of $\hat{\beta}$.
 - This makes no difference asymptotically (so still $\mathcal{N}[0, 1]$).

- The Wald z-statistic is

$$z_j = \frac{\hat{\beta} - \beta^*}{s_{\hat{\beta}}} \stackrel{a}{\sim} \mathcal{N}[0, 1] \quad \text{under } H_0 : \beta = \beta^*$$

- Implementation by two equivalent methods

- ▶ Test using p-values: reject H_0 at level 0.05 if

$$p = \Pr[|Z| > |z_j|] < 0.05, \quad \text{where } Z \sim \mathcal{N}[0, 1].$$

- ▶ Test using critical values: reject H_0 at level 0.05 if

$$|z_j| > z_{.025} = 1.96.$$

- Many packages such as Stata use $T(N - k)$ rather than $\mathcal{N}[0, 1]$
 - ▶ More conservative (less likely to reject H_0)
 - ▶ Exact in unlikely special case that $u_i \sim \mathcal{N}[0, \sigma^2]$.

Confidence interval

- A $100(1 - \alpha)\%$ confidence interval for β is

$$\hat{\beta} \pm z_{\alpha/2} \times s_{\hat{\beta}}.$$

- ▶ in particular a 95% confidence interval is $\hat{\beta} \pm 1.96s_{\hat{\beta}}$.
- ▶ can replace $z_{\alpha/2}$ by $T_{N-k;\alpha/2}$ for better finite sample performance

Hypothesis test of multiple linear restrictions

- Now consider test of several restrictions
 - e.g. Test $H_0 : \beta_2 = 0, \beta_3 = 0$ against H_a : at least one $\neq 0$.
- In matrix algebra we test

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

against $H_a : \mathbf{R}\boldsymbol{\beta} \neq \mathbf{r}.$

- Example: Test $H_0 : \beta_2 = 0, \beta_3 = 0$ against H_a : at least one $\neq 0$

$$\begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\text{or } \underset{(2 \times K)}{\mathbf{R}} \times \underset{(K \times 1)}{\boldsymbol{\beta}} = \underset{(2 \times 1)}{\mathbf{r}}$$

- A Wald test rejects $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ if $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ differs greatly from $\mathbf{0}$.
- Now $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ is normal as linear combination of normals is normal.

$$\begin{aligned}
 \hat{\boldsymbol{\beta}} &\stackrel{a}{\sim} \mathcal{N}[\boldsymbol{\beta}, \mathbf{V}[\hat{\boldsymbol{\beta}}]] \\
 \Rightarrow \quad \mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r} &\stackrel{a}{\sim} \mathcal{N}[\mathbf{R}\boldsymbol{\beta} - \mathbf{r}, \mathbf{RV}[\hat{\boldsymbol{\beta}}]\mathbf{R}'] \\
 \Rightarrow \quad \mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r} &\stackrel{a}{\sim} \mathcal{N}[\mathbf{0}, \mathbf{RV}[\hat{\boldsymbol{\beta}}]\mathbf{R}'] \quad \text{under } H_0 \\
 \Rightarrow \quad (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\mathbf{RV}[\hat{\boldsymbol{\beta}}]\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) &\sim \chi^2(h) \text{ under } H_0
 \end{aligned}$$

- ▶ The last step converts to chi-square using the result

$$\mathbf{z} \sim \mathcal{N}[\mathbf{0}, \Omega] \quad \Rightarrow \quad \mathbf{z}'\Omega^{-1}\mathbf{z} \sim \chi^2(\dim[\Omega]).$$

- To implement this test, replace $\mathbf{V}[\hat{\boldsymbol{\beta}}]$ by $\hat{\mathbf{V}}[\hat{\boldsymbol{\beta}}]$.
 - ▶ This makes no difference asymptotically.

- The Wald chi-squared statistic is

$$W = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\mathbf{R}\hat{\mathbf{V}}[\hat{\boldsymbol{\beta}}]\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) \stackrel{a}{\sim} \chi^2(h) \text{ under } H_0$$

- Implementation by two equivalent methods

- ▶ Test using p-values: reject H_0 at level 0.05 if

$$p = \Pr[\chi^2(h) > W] < 0.05.$$

- ▶ Test using critical-values: reject H_0 at level 0.05 if

$$W > \chi_{.05}^2(h).$$

- The alternative Wald F-test statistic is

$$F = \frac{W}{h} \sim F(h, N - k) \text{ under } H_0$$

- ▶ Makes no difference asymptotically as $F(h, N) \rightarrow \chi^2(h)/h$ as $N \rightarrow \infty$.
- ▶ More conservative (less likely to reject H_0)
- ▶ Exact in unlikely special case that $u_i \sim \mathcal{N}[0, \sigma^2]$.

Further test details

- Wald test is the commonly-used method to test H_0 against H_a .
 - ▶ Estimate β without imposing H_0 .
 - ▶ Then ask does $\hat{\beta}$ approximately satisfy H_0 ?
- The other two test methods used at times are
 - ▶ Likelihood ratio test: Estimate under both H_0 & H_a and compare $\ln L$.
 - ▶ Lagrange multiplier or score test: Estimate under H_a only.
 - ▶ Asymptotically equivalent to Wald under H_0 and local alternatives
 - ▶ Choice is mainly one of convenience, though Wald does have the weakness of lack of invariance to reparameterization.
- Also as already noted for Wald test
 - ▶ asymptotic theory: use Z and $\chi^2(q)$
 - ▶ better finite sample approximation: use $T(N - k)$ and $F(q, N - k)$
 - ▶ even better still: bootstrap with asymptotic refinement.

7. Simulations: OLS consistency and asymptotic normality

- D.g.p.: $y_i = \beta_1 + \beta_2 x_i + u_i$ where $x_i \sim \chi^2(1)$ and $\beta_1 = 1, \beta_2 = 2$.
Error: $u_i \sim \chi^2(1) - 1$ is skewed with mean 0 and variance 2.

```
. * Small sample: parameters differ from dgp values
. clear all

. quietly set obs 30

. set seed 10101

. quietly generate double x = rchi2(1)

. quietly generate y = 1 + 2*x + rchi2(1)-1      // demeaned chi^2 error

. regress y x, noheader
```

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	2.713073	.5743189	4.72	0.000	1.536634	3.889512
_cons	1.150439	.6148461	1.87	0.072	-.1090161	2.409894

- For $N = 30$: $\hat{\beta}_2 = 2.713$ differs appreciably from $\beta_2 = 2.000$.
 - This is due to sampling error as $se[\hat{\beta}_2] = 0.574$.

- How to verify consistency: set N very large.

```
. * Consistency: Large sample: parameters are very close to dgp values
. clear all

. quietly set obs 100000

. set seed 10101

. quietly generate double x = rchi2(1)

. quietly generate y = 1 + 2*x + rchi2(1)-1    // demeaned chi^2 error

. regress y x, noheader
```

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	1.998675	.0031725	630.00	0.000	1.992457	2.004893
_cons	1.005819	.0054945	183.06	0.000	.9950495	1.016588

- For $N = 100,000$: $\hat{\beta}_2 = 1.999$ is very close to $\beta_2 = 2.000$.

- How to check asymptotic results: compute $\hat{\beta}$ many times.

```
. * Central limit theorem
. * Write program to obtain betas for one sample of size numobs (= 150)
. program chi2data, rclass
1.     version 10.1
2.     drop _all
3.     set obs $numobs
4.     generate double x = rchi2(1)
5.     generate y = 1 + 2*x + rchi2(1)-1           // demeaned chi^2 error
6.     regress y x
7.     return scalar b2 = _b[x]
8.     return scalar se2 = _se[x]
9.     return scalar t2 = (_b[x]-2)/_se[x]
10.    return scalar r2 = abs(return(t2))>invttail($numobs-2,.025)
11.    return scalar p2 = 2*ttail($numobs-2,abs(return(t2)))
12. end

. * Run this program 1,000 times to get 1,000 betas etcetera
. * Results differ from MUS (2008) as MUS did not reset the seed to 10101
. * First define global macro numobs for sample size
. global numobs 150

. set seed 10101

. quietly simulate b2f=r(b2) se2f=r(se2) t2f=r(t2) reject2f=r(r2) p2f=r(p2), ///
>     reps(1000) saving(chi2datares, replace) nolegend nodots: chi2data
```

- Then look at the distribution of these $\hat{\beta}'$ s and test statistics.

```
. * Summarize the 1,000 sample means
. summarize b2f se2f t2 reject2f p2f
```

Variable	Obs	Mean	Std. Dev.	Min	Max
b2f	1000	2.000506	.08427	1.719513	2.40565
se2f	1000	.0839776	.0172588	.0415919	.145264
t2f	1000	.0028714	.9932668	-2.824061	4.556576
reject2f	1000	.046	.2095899	0	1
p2f	1000	.5175818	.2890325	.0000108	.9997772

```
. mean b2f se2f t2 reject2f p2f
```

```
Mean estimation              Number of obs   =    1000
```

	Mean	Std. Err.	[95% Conf. Interval]	
b2f	2.000506	.0026649	1.995277	2.005735
se2f	.0839776	.0005458	.0829066	.0850486
t2f	.0028714	.0314099	-.0587655	.0645082
reject2f	.046	.0066278	.032994	.059006
p2f	.5175818	.00914	.499646	.5355177

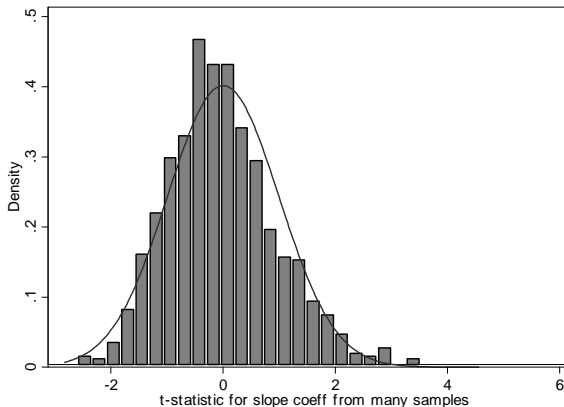
- For $S = 1,000$ simulations each with sample size $N = 150$.

- $\hat{\beta}_2^{(1)}, \hat{\beta}_2^{(2)}, \dots, \hat{\beta}_2^{(1000)}$ has distn. with mean 2.001 close to $\beta_2 = 2.000$
 - and standard deviation 0.089 close to $\sqrt{1/150} = 0.082$

★ using $V[\hat{\beta}_2] \simeq (\sigma_u^2 / V[x_i]) / N = (2/2) / 150 = 1/150$.

- Test $\beta_2 = 2$ using $z = (\hat{\beta}_2 - \beta_2) / \text{se}[\hat{\beta}_2] = (\hat{\beta}_2 - 2.0) / \text{se}[\hat{\beta}_2]$ to test $H_0 : \beta_2 = 2$.

Histogram and kernel density estimate for $z_1, z_2, \dots, z_{1000}$.



- Not quite standard normal: $N = 150$ is still not large enough for CLT.

- How to verify that standard errors are correctly estimated.

- ▶ The average of the computed standard errors of $\hat{\beta}_2$ is 0.0839 (see mean of se2f)
- ▶ This is close to the simulation estimate of $\text{se}[\hat{\beta}_2]$ of 0.0842 (see Std.Dev. of b2f)
- ▶ Aside: Actually for this dgp expect $\sqrt{1/150} \simeq 0.082$ using $V[\hat{\beta}_2] \simeq (\sigma_u^2/V[x_i])/N = (2/2)/150 = 1/150)$

- How to verify that test has correct size.

- ▶ The Wald test of $H_0 : \beta_2 = 2$ at level 0.05 has actual size 0.046 (see mean of reject2f)
- ▶ This is close enough as a 95% simulation interval when $S = 1000$ is

$$0.05 \pm 1.96 \times \sqrt{0.05 \times 0.95/1000} = 0.05 \pm 1.96 \times 0.007 = (0.046, 0.064).$$

8. Stata commands

- Command `regress` does OLS
 - ▶ option `vce(robust)` for heteroskedastic-robust standard errors
 - ▶ option `vce(cluster clid)` for cluster-robust standard errors (with cluster on `clid`)
- For Feasible GLS
 - ▶ command `regress [aweight=]` for known or estimated heteroskedasticity
 - ▶ command `sureg` for systems of linear equations
 - ▶ command `nlsur` for systems of nonlinear equations
 - ▶ command `xtreg, re` for panel random effects.
- For hypothesis tests
 - ▶ command `test` (and `nltest` for nonlinear hypotheses)

9. Appendix: OLS matrix notation example

- Example: $N = 4$ with (x, y) equal to $(1, 1)$, $(2, 3)$, $(2, 4)$, and $(3, 4)$.
- Vector \mathbf{y} and matrix \mathbf{X} are

$$\underset{(4 \times 1)}{\mathbf{y}} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 4 \\ 4 \end{bmatrix}$$

and

$$\underset{(4 \times 2)}{\mathbf{X}} = \begin{bmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \mathbf{x}'_3 \\ \mathbf{x}'_4 \end{bmatrix} = \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ x_{13} & x_{23} \\ x_{14} & x_{24} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}.$$

- Compute $\hat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 3 \end{bmatrix} \times \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 4 & 8 \\ 8 & 18 \end{bmatrix}.$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 4 & 8 \\ 8 & 18 \end{bmatrix}^{-1} = \frac{1}{72 - 64} \begin{bmatrix} 18 & -8 \\ -8 & 4 \end{bmatrix} = \begin{bmatrix} 9/4 & -1 \\ -1 & 1/2 \end{bmatrix}.$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 3 \end{bmatrix} \times \begin{bmatrix} 1 \\ 3 \\ 4 \\ 4 \end{bmatrix} = \begin{bmatrix} 12 \\ 27 \end{bmatrix}.$$

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} 9/4 & -1 \\ -1 & 1/2 \end{bmatrix} \begin{bmatrix} 12 \\ 27 \end{bmatrix} = \begin{bmatrix} 108/4 - 27 \\ -12 + 54/4 \end{bmatrix} = \begin{bmatrix} 0 \\ 1.5 \end{bmatrix}.$$

- OLS estimates:

▶ intercept $\hat{\beta}_1 = 0$ and slope coefficient $\hat{\beta}_2 = 1.5$.

- OLS on intercept and single regressor: $y_i = \beta_1 + \beta_2 x_i + u_i$.

$$\blacktriangleright \mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_N \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} = \begin{bmatrix} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix}$$

$$\blacktriangleright (\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{N \sum_i x_i^2 - (\sum_i x_i)^2} \begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & N \end{bmatrix}$$

$$= \frac{1}{\sum_i x_i^2 - N\bar{x}^2} \begin{bmatrix} N^{-1} \sum_i x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$$

$$\blacktriangleright \mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_N \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{bmatrix} = \begin{bmatrix} N\bar{y} \\ \sum_i x_i y_i \end{bmatrix}$$

$$\blacktriangleright (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \frac{1}{\sum_i x_i^2 - N\bar{x}^2} \begin{bmatrix} \bar{y} \sum_i x_i^2 - \bar{x} \sum_i x_i y_i \\ -\bar{x} N\bar{y} + \sum_i x_i y_i \end{bmatrix}$$

$$= \frac{1}{\sum_i (x_i - \bar{x})^2} \begin{bmatrix} \bar{y} \sum_i x_i^2 - \bar{x} \sum_i x_i y_i \\ \sum_i (x_i - \bar{x})(y_i - \bar{y}) \end{bmatrix} = \begin{bmatrix} \bar{y} - \hat{\beta}_2 \bar{x} \\ \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \end{bmatrix}$$

- So $\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$ and $\hat{\beta}_2 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$ as in introductory course.