Day 1A
Ordinary Least Squares and GLS

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

Frontiers in 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.

March 21-25, 2011
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.

```stata
use mus10data.dta, clear
quietly keep if year02==1
use mus10data.dta, clear
```

```
income 4412 34.34018 29.03987 -49.999 280.777
female 4412 .4718948 .4992661 0 1
chronic 4412 .3263826 .4689423 0 1
private 4412 .7853581 .4106202 0 1
```

```
describe docvis private chronic female income
```

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

```
describe docvis private chronic female income
```

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>docvis</td>
<td>4412</td>
<td>3.95739</td>
<td>7.947601</td>
<td>0</td>
<td>134</td>
</tr>
<tr>
<td>private</td>
<td>4412</td>
<td>.785358</td>
<td>.4106202</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>chronic</td>
<td>4412</td>
<td>.326382</td>
<td>.4689423</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>female</td>
<td>4412</td>
<td>.471894</td>
<td>.4992661</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>income</td>
<td>4412</td>
<td>34.34018</td>
<td>29.03987</td>
<td>-49.999</td>
<td>280.777</td>
</tr>
</tbody>
</table>
OLS regression with default standard errors: assumes i.i.d error.

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

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>35771.7188</td>
<td>4</td>
<td>8942.92971</td>
</tr>
<tr>
<td>Residual</td>
<td>242846.27</td>
<td>4407</td>
<td>55.1046676</td>
</tr>
<tr>
<td>Total</td>
<td>278617.989</td>
<td>4411</td>
<td>63.1643594</td>
</tr>
</tbody>
</table>

| 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     | 0.016018 | .004071  | 3.93 | 0.000  | 0.008037 - 0.0239993|
| _cons      | -0.5647368 | .2746696 | -2.06| 0.040  | -1.103227 - 0.0262465|

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 ↑ $10,000 \Rightarrow income \uparrow 10 \text{ units} \Rightarrow docvis \uparrow 10 \times 0.016 = 0.16.$
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)
```

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

Same coefficient estimates. Different standard errors.
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 \( 165.11 > F_{0.05}(2, 4407) = 3.00 \) 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 $x_i$ and parameter vector $\beta$ are $K \times 1$ column vectors.

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

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

  - Note that all vectors are defined to be column vectors

- For the $i^{th}$ observation

$$y_i = x_i' \beta + u_i.$$
Now combine all $N$ observations from sample $\{(y_i, x_i), i = 1, \ldots, N\}$.

The linear regression model is

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

This is

$$
y = X\beta + u
$$

where

$$
\begin{align*}
y &= \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \\
X &= \begin{bmatrix} x'_1 \\ \vdots \\ x'_N \end{bmatrix} \\
u &= \begin{bmatrix} u_1 \\ \vdots \\ u_N \end{bmatrix}
\end{align*}
$$

The OLS estimator derived below is

$$
\hat{\beta}_{OLS} = (X'X)^{-1}X'y.
$$
Example: \( N = 4 \) with \((x, y)\) equal to \((1, 1)\), \((2, 3)\), \((2, 4)\), and \((3, 4)\).

Then \( y \) is \(4 \times 1\) and \(X\) is \(4 \times 2\) with

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

So (see appendix for detailed computation)

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

\( \hat{\beta}_1 = 0 \) and slope coefficient \( \hat{\beta}_2 = 1.5 \).
Derivation of formula for OLS estimator

- The OLS estimator minimizes the sum of squared errors

\[ Q(\beta) = \sum_{i=1}^{N} u_i^2 = \sum_{i=1}^{N} (y_i - x_i' \beta)^2. \]

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

\[ \frac{\partial Q(\beta)}{\partial \beta} = -2 \sum_{i=1}^{N} x_i (y_i - x_i' \beta) = -2 X'(y - X\beta) = 0. \]

- Then

\[ X'(y - X\beta) = 0 \quad \text{from f.o.c.} \]

\[ \Rightarrow \quad X'y = X'X\beta \quad K \text{ linear equations in } K \text{ unknowns } \beta \]

\[ \Rightarrow \quad \beta = (X'X)^{-1}X'y \quad \text{if the inverse exists (i.e. rank}[X] = K) \]

- So

\[ \hat{\beta}_{OLS} = (X'X)^{-1}X'y = \left(\sum_{i=1}^{N} x_ix_i'\right)^{-1} \sum_{i=1}^{N} x_iy_i. \]
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 $x_i$;
  - Then
    1. $\hat{\beta}$ is consistent for $\beta$
    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 \( y = X\beta + u \) then

\[
\hat{\beta}_{OLS} = (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + u) = (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u
\]

\[
= \beta + (X'X)^{-1}X'u
\]

\[
= \beta + (\sum_i x_ix'_i)^{-1} \sum_i x_iu_i
\]

- So assumptions on \( x_i \) and \( u_i \) are crucial.
OLS Finite Sample Properties

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

\[
\hat{\beta} = \beta + (X'X)^{-1}X'u \\
\sim \beta + (X'X)^{-1}X' \times \mathcal{N}[0, \Omega] \\
\sim \mathcal{N}[\beta, (X'X)^{-1}X'\Omega X(X'X)^{-1}]
\]

- using linear transformation of the normal is normal
\( z \sim \mathcal{N}[\mu, \Omega] \implies Az + b \sim \mathcal{N}[A\mu + b, A\Omega A'] \).

- We instead use asymptotic theory
  - this permits \( u \) to be non-normal distributed.
  - but does require a large sample so \( N \to \infty \).
OLS Consistency

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

- We have (using results below)

\[
\begin{align*}
\text{plim} \hat{\beta} &= \text{plim}\{\beta + (X'X)^{-1}X'u}\} \\
&= \text{plim} \beta + \text{plim} \left\{ (\sum_i x_ix_i')^{-1} \sum_i x_iu_i \right\} \\
&= \text{plim} \beta + \text{plim} \left( \frac{1}{N} \sum_i x_ix_i' \right)^{-1} \times \text{plim} \frac{1}{N} \sum_i x_iu_i \\
&= \beta + (\text{plim} \frac{1}{N} \sum_i x_ix_i')^{-1} \times 0 \\
&= \beta
\end{align*}
\]

- \( \text{plim}\{A_N \times b_N\} = \text{plim} A_N \times \text{plim} b_N \) if the plim’s are constants
- The plim’s exist using laws of large numbers (as averages)
- For \( \text{plim} \frac{1}{N} \sum_i x_iu_i = 0 \) the key assumption is \( E[u_i|x_i] = 0 \).
OLS Limit Distribution

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

- Then limit normal distribution is
  \[
  \sqrt{N}(\hat{\beta} - \beta) = \left(\frac{1}{N} \sum_i x'_i x_i\right)^{-1} \frac{1}{\sqrt{N}} \sum_i x_i u_i
  \]
  \[
  \xrightarrow{d} \text{plim} \left(\frac{1}{N} \sum_i x'_i x_i\right)^{-1} \times \mathcal{N}[0, B] \text{ for some } B
  \]
  \[
  \xrightarrow{d} \mathcal{N} \left[0, \text{plim} \left(\frac{1}{N} \sum_i x'_i x_i\right)^{-1} \times B \times \text{plim} \left(\frac{1}{N} \sum_i x'_i x_i\right)^{-1}\right]
  \]

- If $H_N \xrightarrow{p} H$ and $b_N \xrightarrow{d} \mathcal{N} [\mu, \Omega]$ then $H_N b_N \xrightarrow{p} \mathcal{N} [H\mu, H\Omega H']$
- $\frac{1}{\sqrt{N}} \sum_i x_i u_i \xrightarrow{d} \mathcal{N}[0, B]$ by a central limit theorem
- $B = \text{plim} \left(\frac{1}{\sqrt{N}} \sum_i x_i u_i\right) \left(\frac{1}{\sqrt{N}} \sum_i x_i u_i\right)' = \text{plim} \frac{1}{N} \sum_i \sum_j u_i u_j x_i x'_j$
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 $B$ by a consistent estimate $\hat{B}$

- The so-called “asymptotic distribution” is
  \[
  \hat{\beta} \sim \mathcal{N} \left[ \beta, \left( \sum_{i=1}^{N} x_i x'_i \right)^{-1} \times N \hat{B} \times \left( \sum_{i=1}^{N} x_i x'_i \right)^{-1} \right]
  \]

- Usually $B = \text{Var} \left[ \frac{1}{\sqrt{N}} X'u \right] = \text{Var} \left[ \frac{1}{\sqrt{N}} \sum_i x_i u_i \right]$
- For independent heteroskedastic errors $\hat{B} = \frac{1}{N} \sum_i u_i^2 x_i x'_i$. 
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 $\widehat{\mathbf{B}} = \frac{1}{N} \sum_i \widehat{u}_i^2 \mathbf{x}_i \mathbf{x}_i'$.
- Yields the heteroskedastic-consistent estimate of the variance-covariance matrix of the OLS estimator (VCE)

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

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

- **Default**: Independent homoskedastic errors: $V[u_i | x_i] = \sigma^2$

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

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

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

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

  - Here observations are stacked in cluster $g$ as $y_g = X_g \beta + u_g$.
  - 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 $V[u|X] = \sigma^2 I$.
  - In practice errors are rarely i.i.d.

- So we usually do OLS and obtain robust VCE that permits $V[u|X] \neq \sigma^2 I$
  - could be heteroskedastic robust, cluster-robust, HAC, ....

- More efficient feasible GLS (FGLS) assumes a model for $V[u|X]$
  - yields more precise estimates (smaller standard errors and bigger t-statistics)
  - but then obtain robust VCE that allows for misspecified model for $V[u|X]$.
  - called weighted LS or working matrix LS.
Generalized least squares (GLS)

- Suppose $V[u|X] = \Omega$ where $\Omega$ is known
  - and $y = X\beta + u$, $E[u|X] = 0$ as before.
- The generalized least squares estimator is efficient:
  \[
  \hat{\beta}_{\text{GLS}} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y.
  \]
- Derivation:
  - Premultiply $y = X\beta + u$ by $\Omega^{-1/2}$ so
    \[
    \Omega^{-1/2}y = \Omega^{-1/2}X\beta + \Omega^{-1/2}u.
    \]
  - This model has i.i.d. errors since
    \[
    V[\Omega^{-1/2}u|X] = E[(\Omega^{-1/2}u)(\Omega^{-1/2}u)'|X] = \Omega^{-1/2}\Omega\Omega^{-1/2} = I_N.
    \]
  - Then GLS is OLS in this transformed model:
    \[
    \hat{\beta}_{\text{GLS}} = [(\Omega^{-1/2}X)'(\Omega^{-1/2}X)](\Omega^{-1/2}X)'(\Omega^{-1/2}y)
    = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y.
    \]
Feasible generalized least squares (FGLS)

To implement GLS we need a consistent estimate of $\Omega$. Assume a model for $\Omega = \Omega(\gamma)$, estimate $\hat{\gamma} \overset{p}{\rightarrow} \gamma$, and form $\hat{\Omega} = \Omega(\hat{\gamma}) \overset{p}{\rightarrow} \Omega$.

The feasible GLS estimator (FGLS) is

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

and then

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

Examples:

- Heteroskedasticity: $V[u_i^2|x_i] = \exp(z_i'\gamma)$
- Seemingly unrelated equations: $y_{ig} = x_{ig}'\beta_g + u_{ig}$, $g = 1, \ldots, G$.
  $u_{ig}$ 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.
5. Generalized least squares

Weighted least squares (WLS)

- Now do FGLS but allow for possibility that model for \( V[u|X] \) 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 \( \Omega \)

- The weighted least squares (WLS) estimator is
  \[
  \hat{\beta}_{WLS} = (X'\hat{\Sigma}^{-1}X)^{-1}X'\hat{\Sigma}^{-1}y.
  \]

- Asymptotically \( \hat{\beta}_{WLS} \sim \mathcal{N}[\beta, \hat{V}[\beta]] \) where robust VCE is
  \[
  \hat{\Sigma} = (X'\hat{\Sigma}^{-1}X)^{-1}(X'\hat{\Sigma}^{-1}\Omega\hat{\Sigma}^{-1}X)^{-1}(X'\hat{\Sigma}^{-1}X)^{-1},
  \]
  - for cross-section data \( \Omega = \text{Diag}[(y_i - x_i'\hat{\beta}_{WLS})^2] \).
Hypothesis test of single restriction

- Consider test of a single restriction, for notational simplicity $\beta$

\[
\begin{align*}
H_0 & : \beta = \beta^* \\
H_a & : \beta \neq \beta^*.
\end{align*}
\]

- A Wald test rejects $H_0$ if $\hat{\beta}$ differs greatly from $\beta^*$.

- Define $\sigma_{\hat{\beta}}$ to be the asymptotic standard deviation of $\hat{\beta}$. Then

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

- 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

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

Implementation by two equivalent methods

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

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

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

  \[ |z_j| > z_{0.025} = 1.96. \]

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 $u_i \sim \mathcal{N}[0, \sigma^2]$. 

A. Colin Cameron Univ. of Calif.- Davis (Frontiers in 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.
Confidence interval

- A $100(1 - \alpha)\%$ confidence interval for $\beta$ 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
  
  $H_0 : R\beta = r$
  
  against
  
  $H_a : R\beta \neq r$.
- Example: Test $H_0 : \beta_2 = 0, \beta_3 = 0$ against $H_a$: at least one $\neq 0$

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

or

\[
R \times \beta = r
\]
A Wald test rejects $H_0 : R\beta = r$ if $R\hat{\beta} - r$ differs greatly from 0.

Now $R\hat{\beta} - r$ is normal as linear combination of normals is normal.

\[
\hat{\beta} \overset{a}{\sim} \mathcal{N}[\beta, V[\hat{\beta}]] \\
\Rightarrow \\
R\hat{\beta} - r \overset{a}{\sim} \mathcal{N}[R\beta - r, RV[\hat{\beta}]R'] \\
\Rightarrow \\
R\hat{\beta} - r \overset{a}{\sim} \mathcal{N}[0, RV[\hat{\beta}]R'] \quad \text{under } H_0 \\
\Rightarrow (R\hat{\beta} - r)'[RV[\hat{\beta}]R']^{-1}(R\hat{\beta} - r) \sim \chi^2(h) \quad \text{under } H_0
\]

The last step converts to chi-square using the result

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

To implement this test, replace $V[\hat{\beta}]$ by $\hat{V}[\hat{\beta}]$.

This makes no difference asymptotically.
6. Wald tests and Confidence intervals

Multiple linear restrictions

The Wald chi-squared statistic is

\[ W = (R\hat{\beta} - r)'[R\hat{\Sigma}[\hat{\beta}]R]\cdot^{-1}(R\hat{\beta} - r) \sim \chi^2(h) \text{ under } H_0 \]

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
  \[ W > \chi^2_{0.05}(h). \]

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 \( F(h, N) \rightarrow \chi^2(h)/h \text{ as } N \rightarrow \infty. \)
- 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 $\beta$ 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
```

|     | Coef. | Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|-----|-------|-----------|------|------|-----------------------|
| y   |       |           |      |      |                       |
| 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 \( se[\hat{\beta}_2] = 0.574 \).
7. Simulations

OLS consistency and asymptotic normality

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 \( x = \text{rchi2}(1) \)
.quietly generate \( y = 1 + 2x + \text{rchi2}(1)-1 \) // demeaned \( \chi^2 \) error
.regress \( y \) \( x \), noheader

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

* 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.

Then look at the distribution of these $\hat{\beta}'s$ and test statistics.
For $S = 1,000$ simulations each with sample size $N = 150$.

- $\hat{\beta}_2^{(1)}, \hat{\beta}_2^{(2)}, \ldots, \hat{\beta}_2^{(1000)}$ 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 $V[\hat{\beta}_2] \approx (\sigma_u^2 / V[x_i]) / N = (2/2) / 150 = 1/150$. 

For $S = 1,000$ simulations each with sample size $N = 150$. 
- $\hat{\beta}_2^{(1)}, \hat{\beta}_2^{(2)}, \ldots, \hat{\beta}_2^{(1000)}$ 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 $V[\hat{\beta}_2] \approx (\sigma_u^2 / V[x_i]) / N = (2/2) / 150 = 1/150$. 

### Mean estimation

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>b2f</td>
<td>1000</td>
<td>2.000506</td>
<td>0.08427</td>
<td>1.719513</td>
<td>2.40565</td>
</tr>
<tr>
<td>se2f</td>
<td>1000</td>
<td>0.0839776</td>
<td>0.0172588</td>
<td>0.0415919</td>
<td>0.145264</td>
</tr>
<tr>
<td>t2f</td>
<td>1000</td>
<td>0.0028714</td>
<td>0.00295899</td>
<td>-2.824061</td>
<td>4.556576</td>
</tr>
<tr>
<td>reject2f</td>
<td>1000</td>
<td>0.046</td>
<td>0.2095899</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>p2f</td>
<td>1000</td>
<td>0.5175818</td>
<td>0.2890325</td>
<td>0.000108</td>
<td>0.999772</td>
</tr>
</tbody>
</table>

### Mean estimation

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>b2f</td>
<td>1000</td>
<td>2.000506</td>
<td>0.0026649</td>
<td>1.995277</td>
</tr>
<tr>
<td>se2f</td>
<td>1000</td>
<td>0.0839776</td>
<td>0.005458</td>
<td>0.0829066</td>
</tr>
<tr>
<td>t2f</td>
<td>1000</td>
<td>0.0028714</td>
<td>0.0314099</td>
<td>-0.587655</td>
</tr>
<tr>
<td>reject2f</td>
<td>1000</td>
<td>0.046</td>
<td>0.0066278</td>
<td>0.032994</td>
</tr>
<tr>
<td>p2f</td>
<td>1000</td>
<td>0.5175818</td>
<td>0.00914</td>
<td>0.499646</td>
</tr>
</tbody>
</table>
• 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 $z_1, z_2, \ldots, z_{1000}$.

• 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 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)
Example: \(N = 4\) with \((x, y)\) equal to \((1, 1)\), \((2, 3)\), \((2, 4)\), and \((3, 4)\).

Vector \(y\) and matrix \(X\) are

\[
\begin{bmatrix}
y_1 \\
y_2 \\
y_3 \\
y_4 
\end{bmatrix}_{(4 \times 1)} = \begin{bmatrix} 1 \\ 3 \\ 4 \\ 4 \end{bmatrix}
\]

and

\[
\begin{bmatrix}
x'_1 \\
x'_2 \\
x'_3 \\
x'_4 
\end{bmatrix}_{(4 \times 2)} = \begin{bmatrix}
x_{11} & x_{21} \\
x_{12} & x_{22} \\
x_{13} & x_{23} \\
x_{14} & x_{24} 
\end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}.
\]
Compute $\hat{\beta}_{OLS} = (X'X)^{-1}X'y$:

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

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

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

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

**OLS estimates:**  
> intercept $\hat{\beta}_1 = 0$ and slope coefficient $\hat{\beta}_2 = 1.5.$
OLS on intercept and single regressor: \( y_i = \beta_1 + \beta_2 x_i + u_i \).

\[
\begin{align*}
\mathbf{x}'\mathbf{x} &= \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_N \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} = \begin{bmatrix} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix} \\
(X'X)^{-1} &= \frac{1}{N \sum_i x_i^2 - (\sum_i x_i)^2} \begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & N \end{bmatrix} \\
X'y &= \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_N \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{bmatrix} = \begin{bmatrix} N\bar{y} \\ \sum_i x_i y_i \end{bmatrix} \\
(X'X)^{-1}X'y &= \frac{1}{\sum_i x_i^2 - N\bar{x}^2} \begin{bmatrix} \bar{y} \sum_i x_i^2 - \bar{x} \sum_i x_i y_i \\ -\bar{x} N\bar{y} + \sum_i x_i y_i \end{bmatrix} \\
&= \frac{1}{\sum_i (x_i-\bar{x})^2} \begin{bmatrix} \bar{y} \sum_i x_i^2 - \bar{x} \sum_i x_i y_i \\ \sum_i (x_i-\bar{x})(y_i-\bar{y}) \end{bmatrix} = \begin{bmatrix} \bar{y} - \hat{\beta}_2 \bar{x} \\ \frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sum_i (x_i-\bar{x})^2} \end{bmatrix} \\
\end{align*}
\]

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