Classical Least Squares Theory

CHUNG-MING KUAN

Department of Finance & CRETA

National Taiwan University

September 23, 2010

Lecture Outline

Lecture Outline (cont'd)

Simple Linear Regression

Given the variable of interest y, we are interested in finding a function of another variable x that can characterize the systematic behavior of y.

- y: Dependent variable or regressand
- x: Explanatory variable or regressor
- Specifying a linear function of x: $\alpha + \beta x$ with unknown parameters α and β
- The non-systematic part is the error: $y - (\alpha + \beta x)$

Together we write:

$$y = \underbrace{\alpha + \beta x}_{\text{linear model}} + \underbrace{e(\alpha, \beta)}_{\text{error}}.$$

The objective is to find the "best" fit of the data (y_t, x_t), $t = 1, \ldots, T$.

1. Minimizing a least-squares (LS) criterion function wrt α and β:

$$Q_T(\alpha, \beta) := \frac{1}{T}\sum_{t=1}^{T}(y_t - \alpha - \beta x_t)^2.$$

2. Minimizing a least-absolute-deviation (LAD) criterion wrt α and β:

$$\frac{1}{T}\sum_{t=1}^{T}|y_t - \alpha - \beta x_t|.$$

3. Minimizing asymmetrically weighted absolute deviations:

$$\frac{1}{T}\left(\theta \sum_{t:y_t>\alpha-\beta x_t} |y_t - \alpha - \beta x_t| + (1-\theta) \sum_{t:y_t<\alpha-\beta x_t} |y_t - \alpha - \beta x_t|\right),$$

with $0 < \theta < 1$.

The objective is to find the "best" fit of the data (y_t, x_t), $t = 1, \ldots, T$.

1. Minimizing a least-squares (LS) criterion function wrt α and β:

$$Q_T(\alpha, \beta) := \frac{1}{T}\sum_{t=1}^{T}(y_t - \alpha - \beta x_t)^2.$$

2. Minimizing a least-absolute-deviation (LAD) criterion wrt α and β:

$$\frac{1}{T}\sum_{t=1}^{T}|y_t - \alpha - \beta x_t|.$$

3. Minimizing asymmetrically weighted absolute deviations:

$$\frac{1}{T}\left(\theta \sum_{t:y_t>\alpha-\beta x_t} |y_t - \alpha - \beta x_t| + (1-\theta) \sum_{t:y_t<\alpha-\beta x_t} |y_t - \alpha - \beta x_t|\right),$$

with $0 < \theta < 1$.

The objective is to find the "best" fit of the data (y_t, x_t), $t = 1, \ldots, T$.

1. Minimizing a least-squares (LS) criterion function wrt α and β:

$$Q_T(\alpha, \beta) := \frac{1}{T} \sum_{t=1}^{T} (y_t - \alpha - \beta x_t)^2.$$

2. Minimizing a least-absolute-deviation (LAD) criterion wrt α and β:

$$\frac{1}{T} \sum_{t=1}^{T} |y_t - \alpha - \beta x_t|.$$

3. Minimizing asymmetrically weighted absolute deviations:

$$\frac{1}{T} \left(\theta \sum_{t: y_t > \alpha - \beta x_t} |y_t - \alpha - \beta x_t| + (1 - \theta) \sum_{t: y_t < \alpha - \beta x_t} |y_t - \alpha - \beta x_t| \right),$$

with $0 < \theta < 1$.

- The first order conditions (FOCs) of LS minimization are:

$$\frac{\partial Q(\alpha,\beta)}{\partial \alpha} = -\frac{2}{T}\sum_{t=1}^{T}(y_t - \alpha - \beta x_t) = 0,$$

$$\frac{\partial Q(\alpha,\beta)}{\partial \beta} = -\frac{2}{T}\sum_{t=1}^{T}(y_t - \alpha - \beta x_t)x_t = 0.$$

- The solutions are known as the ordinary least squares (OLS) estimators:

$$\hat{\beta}_T = \frac{\sum_{t=1}^{T}(y_t - \bar{y})(x_t - \bar{x})}{\sum_{t=1}^{T}(x_t - \bar{x})^2},$$

$$\hat{\alpha}_T = \bar{y} - \hat{\beta}_T \bar{x}.$$

Note: x_t can not be a constant.

- The estimated regression line is $\hat{y} = \hat{\alpha}_T + \hat{\beta}_T x$, with the t-th fitted value $\hat{y}_t = \hat{\alpha}_T + \hat{\beta}_T x_t$ and the t-th residual:

$$\hat{e}_t = e_t(\hat{\alpha}_T, \hat{\beta}_T) = y_t - \hat{y}_t.$$

- Substituting $\hat{\alpha}_T$ and $\hat{\beta}_T$ into the first order conditions:

$$\sum_{t=1}^{T}(y_t - \alpha - \beta x_t) = 0, \qquad \sum_{t=1}^{T}(y_t - \alpha - \beta x_t)x_t = 0,$$

  we have the following algebraic results:
  - $\sum_{t=1}^{T} \hat{e}_t = 0$.
  - $\sum_{t=1}^{T} \hat{e}_t x_t = 0$.
  - $\sum_{t=1}^{T} y_t = \sum_{t=1}^{T} \hat{y}_t$ so that $\bar{y} = \bar{\hat{y}}$.
  - $\bar{y} = \hat{\alpha}_T + \hat{\beta}_T \bar{x}$.

Multiple Linear Regression

- With k regressors $x_1, \ldots, x_k$ (x_1 is usually the constant one):

$$y = \beta_1 x_1 + \cdots + \beta_k x_k + e(\beta_1, \ldots, \beta_k).$$

- With data $(y_t, x_{t1}, \ldots, x_{tk})$, $t = 1, \ldots, T$, we can write

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}(\boldsymbol{\beta}), \qquad (1)$$

where $\boldsymbol{\beta} = (\beta_1 \ \beta_2 \ \cdots \ \beta_k)'$,

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{T1} & x_{T2} & \cdots & x_{Tk} \end{bmatrix}, \quad \mathbf{e}(\boldsymbol{\beta}) = \begin{bmatrix} e_1(\boldsymbol{\beta}) \\ e_2(\boldsymbol{\beta}) \\ \vdots \\ e_T(\boldsymbol{\beta}) \end{bmatrix}.$$

- Least-squares criterion function:

$$Q_T(\beta) := \frac{1}{T}\mathbf{e}(\beta)'\mathbf{e}(\beta) = \frac{1}{T}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta). \tag{2}$$

- Identification Requirement [ID-1]: $\mathbf{X}$ is of full column rank k.
  - Any column of $\mathbf{X}$ is not a linear combination of other columns.
  - $\mathbf{X}'\mathbf{X}$ is positive definite and hence invertible.
- FOCs: $-2\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta)/T = \mathbf{0}$, leading to the normal equations:

$$\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{y}.$$

- The unique solution to the normal equations is

$$\hat{\beta}_T = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{3}$$

- Second order condition: $\nabla^2_\beta Q_T(\beta) = 2(\mathbf{X}'\mathbf{X})/T$ is p.d. under [ID-1].

Theorem 3.1

Given specification (1), suppose $[\text{ID-1}]$ holds. Then, the OLS estimator $\hat{\boldsymbol{\beta}}_T = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ uniquely minimizes the criterion function (2).

- Theorem 3 holds regardless of the "true" relation between $\mathbf{y}$ and $\mathbf{X}$.
- When $\mathbf{X}$ is not of full column rank, we have exact multicollinearity. Then, $\mathbf{X}'\mathbf{X}$ is not invertible, and $\hat{\boldsymbol{\beta}}_T$ is not uniquely defined.
- The magnitude of $\hat{\boldsymbol{\beta}}_T$ is affected by the measurement units of the dependent and explanatory variables. Thus, a larger coefficient does not imply that the associated regressor is more important.
- OLS fitted values: $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}_T$; OLS residuals: $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{e}(\hat{\boldsymbol{\beta}}_T)$.
  - $\mathbf{X}'\hat{\mathbf{e}} = \mathbf{0}$; if $\mathbf{X}$ contains a vector of ones, $\sum_{t=1}^{T} \hat{e}_t = 0$.
  - $\hat{\mathbf{y}}'\hat{\mathbf{e}} = \hat{\boldsymbol{\beta}}_T'\mathbf{X}'\hat{\mathbf{e}} = 0$.

Geometric Interpretations

$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the orthogonal projection matrix that projects vectors onto span($\mathbf{X}$), and $\mathbf{I}_T - \mathbf{P}$ is the orthogonal projection matrix that projects vectors onto span$(\mathbf{X})^{\perp}$, the orthogonal complement of span($\mathbf{X}$). Thus, $\mathbf{PX} = \mathbf{X}$ and $(\mathbf{I}_T - \mathbf{P})\mathbf{X} = \mathbf{0}$.

- The vector of fitted values, $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}_T = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{Py}$, is the orthogonal projection of $\mathbf{y}$ onto span($\mathbf{X}$).
- The residual vector, $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}_T - \mathbf{P})\mathbf{y}$, is the orthogonal projection of $\mathbf{y}$ onto span$(\mathbf{X})^{\perp}$.
- $\hat{\mathbf{e}}$ is orthogonal to $\mathbf{X}$, i.e., $\mathbf{X}'\hat{\mathbf{e}} = \mathbf{0}$, and it is also orthogonal to $\hat{\mathbf{y}}$ because $\hat{\mathbf{y}}$ is in span($\mathbf{X}$), i.e., $\hat{\mathbf{y}}'\hat{\mathbf{e}} = 0$.

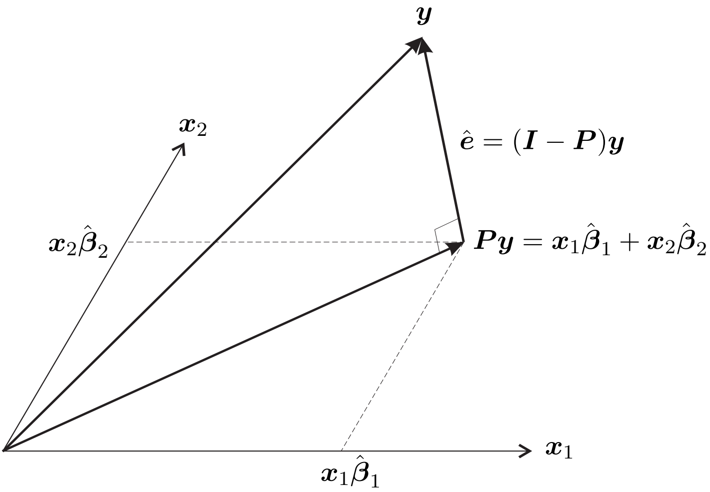

Figure: The orthogonal projection of $\mathbf{y}$ onto span($\mathbf{x}_1$,$\mathbf{x}_2$).

Theorem 3.3 (Frisch-Waugh-Lovell)

Given $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}$, the OLS estimators of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are

$$\hat{\boldsymbol{\beta}}_{1,T} = [\mathbf{X}_1'(\mathbf{I} - \mathbf{P}_2)\mathbf{X}_1]^{-1}\mathbf{X}_1'(\mathbf{I} - \mathbf{P}_2)\mathbf{y},$$

$$\hat{\boldsymbol{\beta}}_{2,T} = [\mathbf{X}_2'(\mathbf{I} - \mathbf{P}_1)\mathbf{X}_2]^{-1}\mathbf{X}_2'(\mathbf{I} - \mathbf{P}_1)\mathbf{y},$$

where $\mathbf{P}_1 = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$ and $\mathbf{P}_2 = \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'$.

- This result shows that $\hat{\boldsymbol{\beta}}_{1,T}$ can be computed from regressing $(\mathbf{I} - \mathbf{P}_2)\mathbf{y}$ on $(\mathbf{I} - \mathbf{P}_2)\mathbf{X}_1$, where $(\mathbf{I} - \mathbf{P}_2)\mathbf{y}$ and $(\mathbf{I} - \mathbf{P}_2)\mathbf{X}_1$ are the residual vectors of $\mathbf{y}$ on $\mathbf{X}_2$ and $\mathbf{X}_1$ on $\mathbf{X}_2$, respectively.
- Similarly, regressing $(\mathbf{I} - \mathbf{P}_1)\mathbf{y}$ on $(\mathbf{I} - \mathbf{P}_1)\mathbf{X}_2$ yields $\hat{\boldsymbol{\beta}}_{2,T}$.
- The OLS estimator of regressing $\mathbf{y}$ on $\mathbf{X}_1$ is not the same as $\hat{\boldsymbol{\beta}}_{1,T}$, unless $\mathbf{X}_1$ and $\mathbf{X}_2$ are orthogonal to each other.

Proof: Writing $\mathbf{y} = \mathbf{X}_1\hat{\beta}_{1,T} + \mathbf{X}_2\hat{\beta}_{2,T} + (\mathbf{I} - \mathbf{P})\mathbf{y}$, where $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ with $\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_2]$, we have

$$\begin{aligned}
&\mathbf{X}_1'(\mathbf{I} - \mathbf{P}_2)\mathbf{y} \\
&\quad = \mathbf{X}_1'(\mathbf{I} - \mathbf{P}_2)\mathbf{X}_1\hat{\beta}_{1,T} + \mathbf{X}_1'(\mathbf{I} - \mathbf{P}_2)\mathbf{X}_2\hat{\beta}_{2,T} + \mathbf{X}_1'(\mathbf{I} - \mathbf{P}_2)(\mathbf{I} - \mathbf{P})\mathbf{y} \\
&\quad = \mathbf{X}_1'(\mathbf{I} - \mathbf{P}_2)\mathbf{X}_1\hat{\beta}_{1,T} + \mathbf{X}_1'(\mathbf{I} - \mathbf{P}_2)(\mathbf{I} - \mathbf{P})\mathbf{y}.
\end{aligned}$$

We know $\text{span}(\mathbf{X}_2) \subseteq \text{span}(\mathbf{X})$, so that $\text{span}(\mathbf{X})^{\perp} \subseteq \text{span}(\mathbf{X}_2)^{\perp}$. Hence, $(\mathbf{I} - \mathbf{P}_2)(\mathbf{I} - \mathbf{P}) = \mathbf{I} - \mathbf{P}$, and

$$\begin{aligned}
\mathbf{X}_1'(\mathbf{I} - \mathbf{P}_2)\mathbf{y} &= \mathbf{X}_1'(\mathbf{I} - \mathbf{P}_2)\mathbf{X}_1\hat{\beta}_{1,T} + \mathbf{X}_1'(\mathbf{I} - \mathbf{P})\mathbf{y} \\
&= \mathbf{X}_1'(\mathbf{I} - \mathbf{P}_2)\mathbf{X}_1\hat{\beta}_{1,T},
\end{aligned}$$

from which we obtain the expression for $\hat{\beta}_{1,T}$.

Frisch-Waugh-Lovell Theorem

Observe that $(\mathbf{I}-\mathbf{P}_1)\mathbf{y} = (\mathbf{I}-\mathbf{P}_1)\mathbf{X}_2\hat{\beta}_{2,T} + (\mathbf{I}-\mathbf{P}_1)(\mathbf{I}-\mathbf{P})\mathbf{y}$.

- $(\mathbf{I}-\mathbf{P}_1)(\mathbf{I}-\mathbf{P}) = \mathbf{I}-\mathbf{P}$, so that the residual vector of regressing $(\mathbf{I}-\mathbf{P}_1)\mathbf{y}$ on $(\mathbf{I}-\mathbf{P}_1)\mathbf{X}_2$ is identical to the residual vector of regressing $\mathbf{y}$ on $\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2]$:

$$(\mathbf{I}-\mathbf{P}_1)\mathbf{y} = (\mathbf{I}-\mathbf{P}_1)\mathbf{X}_2\hat{\beta}_{2,T} + (\mathbf{I}-\mathbf{P})\mathbf{y}.$$

- $\mathbf{P}_1 = \mathbf{P}_1\mathbf{P}$, so that the orthogonal projection of $\mathbf{y}$ directly on $\text{span}(\mathbf{X}_1)$ (i.e., $\mathbf{P}_1\mathbf{y}$) is equivalent to iterated projections of $\mathbf{y}$ on $\text{span}(\mathbf{X})$ and then on $\text{span}(\mathbf{X}_1)$ (i.e., $\mathbf{P}_1\mathbf{P}\mathbf{y}$). Hence,

$$(\mathbf{I}-\mathbf{P}_1)\mathbf{X}_2\hat{\beta}_{2,T} = (\mathbf{I}-\mathbf{P}_1)\mathbf{P}\mathbf{y} = (\mathbf{P}-\mathbf{P}_1)\mathbf{y}.$$

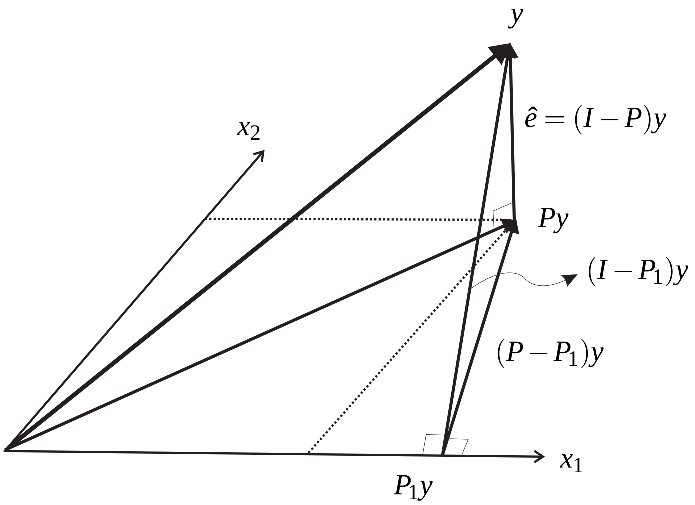

Figure: An illustration of the Frisch-Waugh-Lovell Theorem.

Measures of Goodness of Fit

- Given $\hat{\mathbf{y}}'\hat{\mathbf{e}} = 0$, we have $\mathbf{y}'\mathbf{y} = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \hat{\mathbf{e}}'\hat{\mathbf{e}}$, where $\mathbf{y}'\mathbf{y}$ is known as TSS (total sum of squares), $\hat{\mathbf{y}}'\hat{\mathbf{y}}$ is RSS (regression sum of squares), and $\hat{\mathbf{e}}'\hat{\mathbf{e}}$ is ESS (error sum of squares).
- The non-centered coefficient of determination (or non-centered R^2),

$$R^2 = \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\text{ESS}}{\text{TSS}}, \tag{4}$$

  measures the proportion of the total variation of y_t that can be explained by the model.
  - It is invariant wrt measurement units of the dependent variable but not invariant wrt constant addition.
  - It is a relative measure such that $0 \leq R^2 \leq 1$.
  - It is nondecreasing in the number of regressors. (Why?)

Centered R^2

- When the specification contains a constant term,

$$\sum_{t=1}^{T}(y_t - \bar{y})^2 = \sum_{t=1}^{T}(\hat{y}_t - \bar{\hat{y}})^2 + \sum_{t=1}^{T}\hat{e}_t^2,$$

  i.e., centered TSS = centered RSS + ESS.

- The centered coefficient of determination (or centered R^2),

$$R^2 = \frac{\sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2} = \frac{\text{Centered RSS}}{\text{Centered TSS}} = 1 - \frac{\text{ESS}}{\text{Centered TSS}},$$

  measures the proportion of the total variation of y_t that can be explained by the model, excluding the effect of the constant term.

  - It is invariant wrt constant addition.
  - $0 \leq R^2 \leq 1$, and it is non-decreasing in the number of regressors.
  - It may be negative when the model does not contain a constant term.

Centered R^2: Alternative Interpretation

- When the specification contains a constant term,

$$\sum_{t=1}^{T}(y_t - \bar{y})(\hat{y}_t - \bar{y}) = \sum_{t=1}^{T}(\hat{y}_t - \bar{y} + \hat{e}_t)(\hat{y}_t - \bar{y}) = \sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2,$$

  because $\sum_{t=1}^{T} \hat{y}_t \hat{e}_t = \sum_{t=1}^{t} \hat{e}_t = 0$.

- Centered R^2 can also be expressed as

$$R^2 = \frac{\sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2} = \frac{[\sum_{t=1}^{T}(y_t - \bar{y})(\hat{y}_t - \bar{y})]^2}{[\sum_{t=1}^{T}(y_t - \bar{y})^2][\sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2]},$$

  which is the the squared sample correlation coefficient of y_t and $\hat{y}_t$, also known as the squared multiple correlation coefficient.

- Models for different dep. variables are not comparable in terms of R^2.

Adjusted R^2

- Adjusted R^2 is the centered R^2 adjusted for the degrees of freedom:

$$\bar{R}^2 = 1 - \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}/(T-k)}{(\mathbf{y}'\mathbf{y} - T\bar{y}^2)/(T-1)}.$$

- $\bar{R}^2$ adds a penalty term to R^2:

$$\bar{R}^2 = 1 - \frac{T-1}{T-k}(1-R^2) = R^2 - \frac{k-1}{T-k}(1-R^2),$$

where the penalty term depends on the trade-off between model complexity and model explanatory ability.

- $\bar{R}^2$ may be negative and need not be non-decreasing in k.

Classical Conditions

To derive the statistical properties of the OLS estimator, we assume:

[A1] $\mathbf{X}$ is non-stochastic.

[A2] $\mathbf{y}$ is a random vector such that

(i) $\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}_o$ for some $\boldsymbol{\beta}_o$;

(ii) $\text{var}(\mathbf{y}) = \sigma_o^2 \mathbf{I}_T$ for some $\sigma_o^2 > 0$.

[A3] $\mathbf{y}$ is a random vector s.t. $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}_o, \sigma_o^2 \mathbf{I}_T)$ for some $\boldsymbol{\beta}_o$ and $\sigma_o^2 > 0$.

- The specification (1) with [A1] and [A2] is the classical linear model; (1) with [A1] and [A3] is the classical normal linear model.
- The OLS estimator of σ_o^2 is

$$\hat{\sigma}_T^2 = \frac{1}{T-k} \sum_{t=1}^{T} \hat{e}_t^2.$$

Without Normality

Theorem 3.4

Consider the linear specification (1).

(a) Given [A1] and [A2](i), $\hat{\boldsymbol{\beta}}_T$ is unbiased for $\boldsymbol{\beta}_o$.

(b) Given [A1] and [A2], $\hat{\sigma}_T^2$ is unbiased for σ_o^2.

(c) Given [A1] and [A2], $\text{var}(\hat{\boldsymbol{\beta}}_T) = \sigma_o^2(\mathbf{X}'\mathbf{X})^{-1}$.

Proof: By [A1], $\mathbb{E}(\hat{\boldsymbol{\beta}}_T) = \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathbb{E}(\mathbf{y})$. [A2](i) gives $\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}_o$, so that

$$\mathbb{E}(\hat{\boldsymbol{\beta}}_T) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}_o = \boldsymbol{\beta}_o,$$

proving unbiasedness.

Proof (cont'd): Given $\hat{\mathbf{e}} = (\mathbf{I}_T - \mathbf{P})\mathbf{y} = (\mathbf{I}_T - \mathbf{P})(\mathbf{y} - \mathbf{X}\beta_o)$,

$$\begin{aligned}
\mathbb{E}(\hat{\mathbf{e}}'\hat{\mathbf{e}}) &= \mathbb{E}\big[\text{trace}\big((\mathbf{y} - \mathbf{X}\beta_o)'(\mathbf{I}_T - \mathbf{P})(\mathbf{y} - \mathbf{X}\beta_o)\big)\big] \\
&= \mathbb{E}\big[\text{trace}\big((\mathbf{y} - \mathbf{X}\beta_o)(\mathbf{y} - \mathbf{X}\beta_o)'(\mathbf{I}_T - \mathbf{P})\big)\big] \\
&= \text{trace}\big(\mathbb{E}[(\mathbf{y} - \mathbf{X}\beta_o)(\mathbf{y} - \mathbf{X}\beta_o)'](\mathbf{I}_T - \mathbf{P})\big) \\
&= \text{trace}\big(\sigma_o^2\mathbf{I}_T(\mathbf{I}_T - \mathbf{P})\big) \\
&= \sigma_o^2\,\text{trace}(\mathbf{I}_T - \mathbf{P}).
\end{aligned}$$

where the 4-th equality follows from [A2](ii) that $\text{var}(\mathbf{y}) = \sigma_o^2\mathbf{I}_T$. As $\text{trace}(\mathbf{I}_T - \mathbf{P}) = \text{rank}(\mathbf{I}_T - \mathbf{P}) = T - k$, we have $\mathbb{E}(\hat{\mathbf{e}}'\hat{\mathbf{e}}) = \sigma_o^2(T - k)$ and

$$\mathbb{E}(\hat{\sigma}_T^2) = \mathbb{E}(\hat{\mathbf{e}}'\hat{\mathbf{e}})/(T - k) = \sigma_o^2.$$

Proof (cont'd): By [A1] and [A2](ii),

$$\begin{aligned}
\operatorname{var}(\hat{\boldsymbol{\beta}}_T) &= \operatorname{var}\big((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\big) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\operatorname{var}(\mathbf{y})]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma_o^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{I}_T\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma_o^2(\mathbf{X}'\mathbf{X})^{-1},
\end{aligned}$$

proving (c).

- Theorem 3.4(a) suggests that the OLS fitted values $\mathbf{X}\hat{\boldsymbol{\beta}}_T$ are estimates of $\mathbb{E}(y)$.
- Intuitively, $\hat{\boldsymbol{\beta}}_T$ can be more precisely estimated (i.e., with a smaller variance) when $\mathbf{X}$ has larger variation.

Theorem 3.5 (Gauss-Markov)

Given linear specification (1), suppose that $[\mathrm{A1}]$ and $[\mathrm{A2}]$ hold. Then the OLS estimator $\hat{\boldsymbol{\beta}}_T$ is the best linear unbiased estimator (BLUE) for $\boldsymbol{\beta}_o$.

Proof: Consider an arbitrary linear estimator $\check{\boldsymbol{\beta}}_T = \mathbf{A}\mathbf{y}$, where $\mathbf{A}$ is non-stochastic. Writing $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{C}$, $\check{\boldsymbol{\beta}}_T = \hat{\boldsymbol{\beta}}_T + \mathbf{C}\mathbf{y}$. Then,

$$\operatorname{var}(\check{\boldsymbol{\beta}}_T) = \operatorname{var}(\hat{\boldsymbol{\beta}}_T) + \operatorname{var}(\mathbf{C}\mathbf{y}) + 2\operatorname{cov}(\hat{\boldsymbol{\beta}}_T, \mathbf{C}\mathbf{y}).$$

By [A1] and [A2](i), $\mathbb{E}(\check{\boldsymbol{\beta}}_T) = \boldsymbol{\beta}_o + \mathbf{C}\mathbf{X}\boldsymbol{\beta}_o$, which is unbiased iff $\mathbf{C}\mathbf{X} = \mathbf{0}$. This condition implies $\operatorname{cov}(\hat{\boldsymbol{\beta}}_T, \mathbf{C}\mathbf{y}) = \mathbf{0}$. Thus,

$$\operatorname{var}(\check{\boldsymbol{\beta}}_T) = \operatorname{var}(\hat{\boldsymbol{\beta}}_T) + \operatorname{var}(\mathbf{C}\mathbf{y}) = \operatorname{var}(\hat{\boldsymbol{\beta}}_T) + \sigma_o^2\mathbf{C}\mathbf{C}'.$$

This shows that $\operatorname{var}(\check{\boldsymbol{\beta}}_T) - \operatorname{var}(\hat{\boldsymbol{\beta}}_T)$ is p.s.d., so that $\hat{\boldsymbol{\beta}}_T$ is more efficient than any linear unbiased estimator $\check{\boldsymbol{\beta}}_T$.

Example: $\mathbb{E}(\mathbf{y}) = \mathbf{X}_1\mathbf{b}_1$ and $\text{var}(\mathbf{y}) = \sigma_o^2\mathbf{I}_T$. Two specification:

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{e}.$$

with the OLS estimator $\hat{\mathbf{b}}_{1,T}$, and

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}.$$

with the OLS estimator $\hat{\boldsymbol{\beta}}_T = (\hat{\boldsymbol{\beta}}_{1,T}' \ \hat{\boldsymbol{\beta}}_{2,T}')'$. Clearly, $\hat{\mathbf{b}}_{1,T}$ is the BLUE of $\mathbf{b}_1$ with $\text{var}(\hat{\mathbf{b}}_{1,T}) = \sigma_o^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}$. By the Frisch-Waugh-Lovell Theorem,

$$\mathbb{E}(\hat{\boldsymbol{\beta}}_{1,T}) = \mathbb{E}\left([\mathbf{X}_1'(\mathbf{I}_T - \mathbf{P}_2)\mathbf{X}_1]^{-1}\mathbf{X}_1'(\mathbf{I}_T - \mathbf{P}_2)\mathbf{y}\right) = \mathbf{b}_1,$$
$$\mathbb{E}(\hat{\boldsymbol{\beta}}_{2,T}) = \mathbb{E}\left([\mathbf{X}_2'(\mathbf{I}_T - \mathbf{P}_1)\mathbf{X}_2]^{-1}\mathbf{X}_2'(\mathbf{I}_T - \mathbf{P}_1)\mathbf{y}\right) = \mathbf{0}.$$

That is, $\hat{\boldsymbol{\beta}}_T$ is unbiased for $(\mathbf{b}_1' \ \mathbf{0}')'$.

Example (cont'd):

$$\begin{aligned}\text{var}(\hat{\beta}_{1,T}) &= \text{var}([\mathbf{X}_1'(\mathbf{I}_T - \mathbf{P}_2)\mathbf{X}_1]^{-1}\mathbf{X}_1'(\mathbf{I}_T - \mathbf{P}_2)\mathbf{y}) \\ &= \sigma_o^2[\mathbf{X}_1'(\mathbf{I}_T - \mathbf{P}_2)\mathbf{X}_1]^{-1}.\end{aligned}$$

As $\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'(\mathbf{I}_T - \mathbf{P}_2)\mathbf{X}_1 = \mathbf{X}_1'\mathbf{P}_2\mathbf{X}_1$ is p.s.d.,

$$[\mathbf{X}_1'(\mathbf{I}_T - \mathbf{P}_2)\mathbf{X}_1]^{-1} - (\mathbf{X}_1'\mathbf{X}_1)^{-1}$$

is ps.d. Hence, $\hat{\mathbf{b}}_{1,T}$ is more efficient than $\hat{\beta}_{1,T}$, as it ought to be.

With Normality

- Under [A3] that $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}_o, \sigma_o^2 \mathbf{I}_T)$, the log-likelihood function of $\mathbf{y}$ is

$$\log L(\boldsymbol{\beta}, \sigma^2) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

- The score vector is

$$\mathbf{s}(\boldsymbol{\beta}, \sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2}\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{bmatrix},$$

- Solutions to $\mathbf{s}(\boldsymbol{\beta}, \sigma^2) = \mathbf{0}$ are the (quasi) maximum likelihood estimators (MLEs). Clearly, the MLE of $\boldsymbol{\beta}$ is the OLS estimator, and the MLE of σ^2 is

$$\tilde{\sigma}_T^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_T)'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_T)}{T} = \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{T} \neq \hat{\sigma}_T^2.$$

Theorem 3.7

Given the linear specification (1), suppose that $[\mathrm{A1}]$ and $[\mathrm{A3}]$ hold.

(a) $\hat{\boldsymbol{\beta}}_T \sim \mathcal{N}\big(\boldsymbol{\beta}_o, \sigma_o^2(\mathbf{X}'\mathbf{X})^{-1}\big)$.

(b) $(T-k)\hat{\sigma}_T^2/\sigma_o^2 \sim \chi^2(T-k)$.

(c) $\hat{\sigma}_T^2$ has mean σ_o^2 and variance $2\sigma_o^4/(T-k)$.

Proof: For (a), we note that $\hat{\boldsymbol{\beta}}_T$ is a linear transformation of $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}_o, \sigma_o^2\mathbf{I}_T)$ and hence also a normal random vector. As for (b), writing $\hat{\mathbf{e}} = (\mathbf{I}_T - \mathbf{P})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_o)$, we have

$$(T-k)\hat{\sigma}_T^2/\sigma_o^2 = \hat{\mathbf{e}}'\hat{\mathbf{e}}/\sigma_o^2 = \mathbf{y}^{*\prime}(\mathbf{I}_T - \mathbf{P})\mathbf{y}^*,$$

where $\mathbf{y}^* = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_o)/\sigma_o \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_T)$ by [A3].

Proof (cont'd): Let $\mathbf{C}$ orthogonally diagonalizes $\mathbf{I}_T - \mathbf{P}$ such that $\mathbf{C}'(\mathbf{I}_T - \mathbf{P})\mathbf{C} = \mathbf{\Lambda}$. Since $\text{rank}(\mathbf{I}_T - \mathbf{P}) = T - k$, $\mathbf{\Lambda}$ contains $T - k$ eigenvalues equal to one and k eigenvalues equal to zero. Then,

$$\mathbf{y}^{*\prime}(\mathbf{I}_T - \mathbf{P})\mathbf{y}^* = \mathbf{y}^{*\prime}\mathbf{C}[\mathbf{C}'(\mathbf{I}_T - \mathbf{P})\mathbf{C}]\mathbf{C}'\mathbf{y}^* = \boldsymbol{\eta}' \begin{bmatrix} \mathbf{I}_{T-k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \boldsymbol{\eta}.$$

where $\boldsymbol{\eta} = \mathbf{C}'\mathbf{y}^*$. As $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_T)$, η_i are independent, standard normal random variables. It follows that

$$\mathbf{y}^{*\prime}(\mathbf{I}_T - \mathbf{P})\mathbf{y}^* = \sum_{i=1}^{T-k} \eta_i^2 \sim \chi^2(T - k),$$

proving (b). (c) is a direct consequence of (b) and the facts that $\chi^2(T - k)$ has mean $T - k$ and variance $2(T - k)$.

Theorem 3.8

Given the linear specification (1), suppose that $[\mathrm{A1}]$ and $[\mathrm{A3}]$ hold. Then the OLS estimators $\hat{\boldsymbol{\beta}}_T$ and $\hat{\sigma}^2_T$ are the best unbiased estimators (BUE) for $\boldsymbol{\beta}_o$ and σ^2_o, respectively.

Proof: The Hessian matrix of the log-likelihood function is

$$\mathbf{H}(\beta, \sigma^2) = \begin{bmatrix} -\frac{1}{\sigma^2}\mathbf{X}'\mathbf{X} & -\frac{1}{\sigma^4}\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta) \\ -\frac{1}{\sigma^4}(\mathbf{y} - \mathbf{X}\beta)'\mathbf{X} & \frac{T}{2\sigma^4} - \frac{1}{\sigma^6}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \end{bmatrix}.$$

Under [A3], $\mathbb{E}[\mathbf{s}(\boldsymbol{\beta}_o, \sigma^2_o)] = \mathbf{0}$ and

$$\mathbb{E}[\mathbf{H}(\boldsymbol{\beta}_o, \sigma^2_o)] = \begin{bmatrix} -\frac{1}{\sigma^2_o}\mathbf{X}'\mathbf{X} & \mathbf{0} \\ \mathbf{0} & -\frac{T}{2\sigma^4_o} \end{bmatrix}.$$

Proof (cont'd):

By the information matrix equality, $-\mathbb{E}[\mathbf{H}(\boldsymbol{\beta}_o, \sigma_o^2)]$ is the information matrix. Then, its inverse,

$$-\mathbb{E}[\mathbf{H}(\boldsymbol{\beta}_o, \sigma_o^2)]^{-1} = \begin{bmatrix} \sigma_o^2(\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{2\sigma_o^4}{T} \end{bmatrix},$$

is the Cramér-Rao lower bound.

- $\text{var}(\hat{\boldsymbol{\beta}}_T)$ achieves this lower bound (the upper-left block) so that $\hat{\boldsymbol{\beta}}_T$ is the best unbiased estimator for $\boldsymbol{\beta}_o$.
- Although $\text{var}(\hat{\sigma}_T^2) = 2\sigma_o^4/(T-k)$ is greater than the lower bound (lower-right element), it can be shown that $\hat{\sigma}_T^2$ is still the best unbiased estimator for σ_o^2; see Rao (1973, p. 319) for a proof.

Tests for Linear Hypotheses

- Linear hypothesis: $\mathbf{R}\boldsymbol{\beta}_o = \mathbf{r}$, where $\mathbf{R}$ is $q \times k$ with full row rank q and $q < k$, $\mathbf{r}$ is a vector of hypothetical values.
- A natural way to construct a test statistic is to compare $\mathbf{R}\hat{\boldsymbol{\beta}}_T$ and r; we would reject the null if their difference is very "large."
- Given [A1] and [A3],

$$\mathbf{R}\hat{\boldsymbol{\beta}}_T \sim \mathcal{N}(\mathbf{R}\boldsymbol{\beta}_o, \sigma_o^2[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']).$$

Consider the case that $q = 1$. Under the null hypothesis,

$$\frac{\mathbf{R}\hat{\boldsymbol{\beta}}_T - \mathbf{r}}{\sigma_o[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{1/2}} = \frac{\mathbf{R}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o)}{\sigma_o[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{1/2}} \sim \mathcal{N}(0, 1).$$

An operational statistic is obtained by replacing σ_o with $\hat{\sigma}_T$:

$$\tau = \frac{\mathbf{R}\hat{\beta}_T - \mathbf{r}}{\hat{\sigma}_T[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{1/2}}.$$

Theorem 3.9

Given the linear specification (1), suppose that $[\mathrm{A1}]$ and $[\mathrm{A3}]$ hold. When $\mathbf{R}$ is $1 \times k$, $\tau \sim t(T - k)$ under the null hypothesis.

Note: This t distribution result holds when the normality condition [A3] is true.

Proof: We write the statistic τ as

$$\tau = \frac{\mathbf{R}\hat{\boldsymbol{\beta}}_T - \mathbf{r}}{\sigma_o[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{1/2}} \Bigg/ \sqrt{\frac{(T-k)\hat{\sigma}_T^2/\sigma_o^2}{T-k}},$$

where the numerator is $\mathcal{N}(0,1)$ and $(T-k)\hat{\sigma}_T^2/\sigma_o^2$ is $\chi^2(T-k)$ by Theorem 3.7(b). The assertion follows when the numerator and denominator are independent. This is indeed the case, because $\hat{\boldsymbol{\beta}}_T$ and $\hat{\mathbf{e}}$ are jointly normally distributed with

$$\begin{aligned}
\operatorname{cov}(\hat{\mathbf{e}}, \hat{\boldsymbol{\beta}}_T) &= \mathbb{E}[(\mathbf{I}_T - \mathbf{P})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_o)\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\
&= (\mathbf{I}_T - \mathbf{P})\,\mathbb{E}[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_o)\mathbf{y}']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma_o^2(\mathbf{I}_T - \mathbf{P})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{0}.
\end{aligned}$$

Examples

To test $\beta_i = c$, let $\mathbf{R} = [0 \ \cdots \ 0 \ 1 \ 0 \ \cdots \ 0]$ and m^{ij} be the (i, j) th element of $\mathbf{M}^{-1} = (\mathbf{X}'\mathbf{X})^{-1}$. Then,

$$\tau = \frac{\hat{\beta}_{i,T} - c}{\hat{\sigma}_T \sqrt{m^{ii}}} \sim t(T-k),$$

where $m^{ii} = \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$. τ is a *t* statistic; for testing $\beta_i = 0$, τ is also referred to as the *t* ratio.

It is straightforward to verify that to test $a\beta_i + b\beta_j = c$, with a, b, c given constants, the corresponding test reads:

$$\tau = \frac{a\hat{\beta}_{i,T} + b\hat{\beta}_{j,T} - c}{\hat{\sigma}_T \sqrt{[a^2 m^{ii} + b^2 m^{jj} + 2ab m^{ij}]}} \sim t(T-k).$$

When $\mathbf{R}$ is a $q \times k$ matrix with full row rank, note that

$$(\mathbf{R}\hat{\beta}_T - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta}_T - \mathbf{r})/\sigma_o^2 \sim \chi^2(q).$$

An operational statistic is

$$\begin{aligned}\varphi &= \frac{(\mathbf{R}\hat{\beta}_T - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta}_T - \mathbf{r})/(\sigma_o^2 q)}{(T-k)\hat{\sigma}_T^2/[\sigma_o^2(T-k)]} \\ &= \frac{(\mathbf{R}\hat{\beta}_T - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta}_T - \mathbf{r})}{\hat{\sigma}_T^2 q}.\end{aligned}$$

When $q = 1$, $\varphi = \tau^2$.

Theorem 3.10

Given the linear specification (1), suppose that [A1] and [A3] hold. When $\mathbf{R}$ is $q \times k$ with full row rank, $\varphi \sim F(q, T-k)$ under the null hypothesis.

Example: $H_o: \beta_1 = b_1$ and $\beta_2 = b_2$. The F statistic,

$$\varphi = \frac{1}{2\hat{\sigma}_T^2} \begin{pmatrix} \hat{\beta}_{1,T} - b_1 \\ \hat{\beta}_{2,T} - b_2 \end{pmatrix}' \begin{bmatrix} m^{11} & m^{12} \\ m^{21} & m^{22} \end{bmatrix}^{-1} \begin{pmatrix} \hat{\beta}_{1,T} - b_1 \\ \hat{\beta}_{2,T} - b_2 \end{pmatrix},$$

is distributed as $F(2, T-k)$.

Example: $H_o: \beta_2 = 0$, and $\beta_3 = 0$, $\cdots$ and $\beta_k = 0$,

$$\varphi = \frac{1}{(k-1)\hat{\sigma}_T^2} \begin{pmatrix} \hat{\beta}_{2,T} \\ \hat{\beta}_{3,T} \\ \vdots \\ \hat{\beta}_{k,T} \end{pmatrix}' \begin{bmatrix} m^{22} & m^{23} & \cdots & m^{2k} \\ m^{32} & m^{33} & \cdots & m^{3k} \\ \vdots & & & \vdots \\ m^{k2} & m^{k3} & \cdots & m^{kk} \end{bmatrix}^{-1} \begin{pmatrix} \hat{\beta}_{2,T} \\ \hat{\beta}_{3,T} \\ \vdots \\ \hat{\beta}_{k,T} \end{pmatrix},$$

is distributed as $F(k-1, T-k)$ and known as regression F test.

Test Power

To examine the power of the F test, we evaluate the distribution of φ under the alternative hypothesis: $\mathbf{R}\beta_o = \mathbf{r} + \boldsymbol{\delta}$, with $\mathbf{R}$ is a $q \times k$ matrix with rank $q < k$ and $\delta \neq \mathbf{0}$.

Theorem 3.11

Given the linear specification (1), suppose that $[\mathrm{A1}]$ and $[\mathrm{A3}]$ hold. When $\mathbf{R}\beta_o = \mathbf{r} + \boldsymbol{\delta}$,

$$\varphi \sim F(q, T-k; \boldsymbol{\delta}'\mathbf{D}^{-1}\boldsymbol{\delta}, 0),$$

where $\mathbf{D} = \sigma_o^2[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']$, and $\boldsymbol{\delta}'\mathbf{D}^{-1}\boldsymbol{\delta}$ is the non-centrality parameter of the numerator of φ.

Proof: When $\mathbf{R}\boldsymbol{\beta}_o = \mathbf{r} + \boldsymbol{\delta}$,

$$[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1/2}(\mathbf{R}\hat{\boldsymbol{\beta}}_T - \mathbf{r})/\sigma_o = \mathbf{D}^{-1/2}[\mathbf{R}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o) + \boldsymbol{\delta}],$$

which is distributed as $\mathcal{N}(\mathbf{0}, \mathbf{I}_q) + \mathbf{D}^{-1/2}\boldsymbol{\delta}$. Then,

$$(\mathbf{R}\hat{\boldsymbol{\beta}}_T - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}_T - \mathbf{r})/\sigma_o^2 \sim \chi^2(q;\, \boldsymbol{\delta}'\mathbf{D}^{-1}\boldsymbol{\delta}),$$

a non-central χ^2 distribution with the non-centrality parameter $\boldsymbol{\delta}'\mathbf{D}^{-1}\boldsymbol{\delta}$. It is also readily seen that $(T-k)\hat{\sigma}_T^2/\sigma_o^2$ is still distributed as $\chi^2(T-k)$. Similar to the argument before, these two terms are independent, so that φ has a non-central F distribution.

- Test power is determined by the non-centrality parameter $\boldsymbol{\delta}'\mathbf{D}^{-1}\boldsymbol{\delta}$, where $\boldsymbol{\delta}$ signifies the deviation from the null. When $\mathbf{R}\boldsymbol{\beta}_o$ deviates farther from the hypothetical value $\mathbf{r}$ (i.e., $\boldsymbol{\delta}$ is "large"), the non-centrality parameter $\boldsymbol{\delta}'\mathbf{D}^{-1}\boldsymbol{\delta}$ increases, and so does the power.
- Example: The null distribution is $F(2, 20)$, and its critical value at 5% level is 3.49. Then for $F(2, 20; \nu_1, 0)$ with the non-centrality parameter $\nu_1 = 1, 3, 5$, the probabilities that φ exceeds 3.49 are approximately 12.1%, 28.2%, and 44.3%, respectively.
- Example: The null distribution is $F(5, 60)$, and its critical value at 5% level is 2.37. Then for $F(5, 60; \nu_1, 0)$ with $\nu_1 = 1, 3, 5$, the probabilities that φ exceeds 2.37 are approximately 9.4%, 20.5%, and 33.2%, respectively.

Alternative Interpretation

- Constrained OLS: Finding the saddle point of the Lagrangian:

$$\min_{\beta,\boldsymbol{\lambda}} \frac{1}{T}(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)+(\mathbf{R}\beta-\mathbf{r})'\boldsymbol{\lambda},$$

where $\boldsymbol{\lambda}$ is the $q \times 1$ vector of Lagrangian multipliers, we have

$$\ddot{\boldsymbol{\lambda}}_T = 2[\mathbf{R}(\mathbf{X}'\mathbf{X}/T)^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta}_T-\mathbf{r}),$$

$$\ddot{\beta}_T = \hat{\beta}_T - (\mathbf{X}'\mathbf{X}/T)^{-1}\mathbf{R}'\ddot{\boldsymbol{\lambda}}_T/2.$$

- The constrained OLS residuals are

$$\ddot{\mathbf{e}} = \mathbf{y}-\mathbf{X}\hat{\beta}_T+\mathbf{X}(\hat{\beta}_T-\ddot{\beta}_T) = \hat{\mathbf{e}}+\mathbf{X}(\hat{\beta}_T-\ddot{\beta}_T),$$

with $\hat{\beta}_T-\ddot{\beta}_T = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta}_T-\mathbf{r})$.

- The sum of squared, constrained OLS residuals are:

$$\begin{aligned}\ddot{\mathbf{e}}'\ddot{\mathbf{e}} &= \hat{\mathbf{e}}'\hat{\mathbf{e}} + (\hat{\boldsymbol{\beta}}_T - \ddot{\boldsymbol{\beta}}_T)'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_T - \ddot{\boldsymbol{\beta}}_T) \\ &= \hat{\mathbf{e}}'\hat{\mathbf{e}} + (\mathbf{R}\hat{\boldsymbol{\beta}}_T - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}_T - \mathbf{r}),\end{aligned}$$

  where the 2nd term on the RHS is the numerator of the F statistic.
- Letting $\mathrm{ESS_c} = \ddot{\mathbf{e}}'\ddot{\mathbf{e}}$ and $\mathrm{ESS_u} = \hat{\mathbf{e}}'\hat{\mathbf{e}}$ we have

$$\varphi = \frac{\ddot{\mathbf{e}}'\ddot{\mathbf{e}} - \hat{\mathbf{e}}'\hat{\mathbf{e}}}{q\hat{\sigma}_T^2} = \frac{(\mathrm{ESS_c} - \mathrm{ESS_u})/q}{\mathrm{ESS_u}/(T-k)},$$

  suggesting that F test in effect compares the constrained and unconstrained models based on their lack-of-fitness.
- The regression F test is thus $\varphi = \frac{(R_u^2 - R_c^2)/q}{(1-R_u^2)/(T-k)}$ which compares model fitness of the full model and the model with only a constant term.

- The sum of squared, constrained OLS residuals are:

$$\begin{aligned}\ddot{\mathbf{e}}'\ddot{\mathbf{e}} &= \hat{\mathbf{e}}'\hat{\mathbf{e}} + (\hat{\boldsymbol{\beta}}_T - \ddot{\boldsymbol{\beta}}_T)'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_T - \ddot{\boldsymbol{\beta}}_T) \\ &= \hat{\mathbf{e}}'\hat{\mathbf{e}} + (\mathbf{R}\hat{\boldsymbol{\beta}}_T - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}_T - \mathbf{r}),\end{aligned}$$

  where the 2nd term on the RHS is the numerator of the F statistic.

- Letting $\mathrm{ESS_c} = \ddot{\mathbf{e}}'\ddot{\mathbf{e}}$ and $\mathrm{ESS_u} = \hat{\mathbf{e}}'\hat{\mathbf{e}}$ we have

$$\varphi = \frac{\ddot{\mathbf{e}}'\ddot{\mathbf{e}} - \hat{\mathbf{e}}'\hat{\mathbf{e}}}{q\hat{\sigma}_T^2} = \frac{(\mathrm{ESS_c} - \mathrm{ESS_u})/q}{\mathrm{ESS_u}/(T-k)},$$

  suggesting that F test in effect compares the constrained and unconstrained models based on their lack-of-fitness.

- The regression F test is thus $\varphi = \frac{(R_\mathrm{u}^2 - R_\mathrm{c}^2)/q}{(1-R_\mathrm{u}^2)/(T-k)}$ which compares model fitness of the full model and the model with only a constant term.

Confidence Regions

- A confidence interval for $\beta_{i,o}$ is the interval $(\underline{g}_\alpha, \overline{g}_\alpha)$ such that

$$\mathbb{P}\{\, \underline{g}_\alpha \le \beta_{i,o} \le \overline{g}_\alpha \} = 1 - \alpha,$$

where $(1 - \alpha)$ is known as the confidence coefficient.

- Letting $c_{\alpha/2}$ be the critical value of $t(T - k)$ with tail prob. $\alpha/2$,

$$\mathbb{P}\left\{ \left| (\hat{\beta}_{i,T} - \beta_{i,o}) / (\hat{\sigma}_T \sqrt{m^{ii}}) \right| \le c_{\alpha/2} \right\}$$

$$\mathbb{P}\left\{ \hat{\beta}_{i,T} c_{\alpha/2} \hat{\sigma}_T \sqrt{m^{ii}} \le \beta_{i,o} \le \hat{\beta}_{i,T} + c_{\alpha/2} \hat{\sigma}_T \sqrt{m^{ii}} \right\}$$

$$= 1 - \alpha.$$

- The confidence region for a vector of parameters can be constructed by resorting to F statistic.
- For $(\beta_{1,o} = b_1, \beta_{2,o} = b_2)'$, suppose $T - k = 30$ and $\alpha = 0.05$. Then, $F_{0.05}(2, 30) = 3.32$, and

$$\mathbb{P}\left\{ \frac{1}{2\hat{\sigma}_T^2} \begin{pmatrix} \hat{\beta}_{1,T} - b_1 \\ \hat{\beta}_{2,T} - b_2 \end{pmatrix}' \begin{bmatrix} m^{11} & m^{12} \\ m^{21} & m^{22} \end{bmatrix}^{-1} \begin{pmatrix} \hat{\beta}_{1,T} - b_1 \\ \hat{\beta}_{2,T} - b_2 \end{pmatrix} \leq 3.32 \right\}$$

is $1 - \alpha$, which results in an ellipse with the center $(\hat{\beta}_{1,T}, \hat{\beta}_{2,T})$.

Note: It is possible that (β_1, β_2) is outside the confidence box formed by individual confidence intervals but inside the joint confidence ellipse. That is, while a t ratio may indicate statistic significance of a coefficient, the F test may suggest the opposite based on the confidence region.

Near Multicollinearity

It is more common to have near multicollinearity: $\mathbf{X}\mathbf{a} \approx \mathbf{0}$.

- Writing $\mathbf{X} = [\mathbf{x}_i \; \mathbf{X}_i]$, we have from the FWL Theorem that

$$\text{var}(\hat{\beta}_{i,T}) = \sigma_o^2[\mathbf{x}_i'(\mathbf{I} - \mathbf{P}_i)\mathbf{x}_i]^{-1} = \frac{\sigma_o^2}{\sum_{t=1}^{T}(x_{ti} - \bar{x}_i)^2(1 - R^2(i))},$$

  where $\mathbf{P}_i = \mathbf{X}_i(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'$, and $R^2(i)$ is the centered R^2 from regressing $\mathbf{x}_i$ on $\mathbf{X}_i$.

- Consequence of near multicollinearity:
  - $R^2(i)$ is high, so that $\text{var}(\hat{\beta}_{i,T})$ tend to be large and that $\hat{\beta}_{i,T}$ are sensitive to data changes.
  - Large $\text{var}(\hat{\beta}_{i,T})$ lead to small (insignificant) t ratios. Yet, regression F test may suggest that the model (as a whole) is useful.

How do we circumvent the problems from near multicollinearity?

- Try to break the approximate linear relation.
  - Adding more data if possible.
  - Dropping some regressors.
- Statistical approaches:
  - Ridge regression: For some $\lambda \neq 0$,

$$\hat{\mathbf{b}}_{\text{ridge}} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k)^{-1}\mathbf{X}'\mathbf{y}.$$

  - Principal component regression:
- Note: Multicollinearity vs. "micronumerosity" (Goldberger)

Digression: Regression with Dummy Variables

Example: Let y_t be wage and x_t be working experience (in years). The dummy variable $D_t = 1$ if t is a male ($D_t = 0$ otherwise). Then,

$$y_t = \alpha_0 + \alpha_1 D_t + \beta x_t + e_t;$$

regressions for female and male have the intercepts α_0 and $\alpha_0 + \alpha_1$.

Example: $D_{1,t} = 1$ if t is a high school graduate ($D_{1,t=0}$ otherwise), and $D_{2,t} = 1$ if t has college degree or higher ($D_{2,t=0}$ otherwise). We have:

$$y_t = \alpha_0 + \alpha_1 D_{1,t} + \alpha_2 D_{2,t} + \beta x_t + e_t,$$

with the intercepts for 3 regressions: α_0, $\alpha_0 + \alpha_1$, and $\alpha_0 + \alpha_2$.

Dummy variable trap: To avoid exact multicollinearity, the number of dummy variables in a model should be one less than the number of groups.

Limitation of the Classical Conditions

- [A1] $\mathbf{X}$ is non-stochastic: Economic variables can not be regarded as non-stochastic; also, lagged dependent variables may be used as regressors.
- [A2](i) $\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}_o$: $\mathbb{E}(\mathbf{y})$ may be a linear function with more regressors or a nonlinear function of regressors.
- [A2](ii) $\text{var}(\mathbf{y}) = \sigma_o^2 \mathbf{I}_T$: The elements of $\mathbf{y}$ may be correlated (serial correlation, spatial correlation) and/or may have unequal variances.
- [A3] Normality: $\mathbf{y}$ may have a non-normal distribution.
- The OLS estimator loses the properties derived before when some of the classical conditions fail to hold.

When $\text{var}(\mathbf{y}) \neq \sigma_o^2 \mathbf{I}_T$

Given the linear specification $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$, suppose, in addition to [A1] and [A2](i), $\text{var}(\mathbf{y}) = \mathbf{\Sigma}_o \neq \sigma_o^2 \mathbf{I}_T$, where $\mathbf{\Sigma}_o$ is p.d. That is, the elements of $\mathbf{y}$ may be correlated and have unequal variances.

- The OLS estimator $\hat{\boldsymbol{\beta}}_T$ remains unbiased with

$$\text{var}(\hat{\boldsymbol{\beta}}_T) = \text{var}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}_o\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

- $\hat{\boldsymbol{\beta}}_T$ is not the BLUE for $\boldsymbol{\beta}_o$, and it is not the BUE for $\boldsymbol{\beta}_o$ under normality.
- The estimator $\widehat{\text{var}}(\hat{\boldsymbol{\beta}}_T) = \hat{\sigma}_T^2(\mathbf{X}'\mathbf{X})^{-1}$ is a biased estimator for $\text{var}(\hat{\boldsymbol{\beta}}_T)$. Consequently, the t and F tests do not have t and F distributions, even when $\mathbf{y}$ is normally distributed.

The GLS Estimator

Consider the specification: $\mathbf{Gy} = \mathbf{GX}\beta + \mathbf{Ge}$, where $\mathbf{G}$ is nonsingular and non-stochastic.

- $\mathbb{E}(\mathbf{Gy}) = \mathbf{GX}\beta_o$ and $\text{var}(\mathbf{Gy}) = \mathbf{G}\mathbf{\Sigma}_o\mathbf{G}'$.
- $\mathbf{GX}$ has full column rank so that the OLS estimator can be computed:

$$\mathbf{b}(\mathbf{G}) = (\mathbf{X}'\mathbf{G}'\mathbf{GX})^{-1}\mathbf{X}'\mathbf{G}'\mathbf{Gy},$$

  which is still linear and unbiased. It would be the BLUE provided that $\mathbf{G}$ is chosen such that $\mathbf{G}\mathbf{\Sigma}_o\mathbf{G}' = \sigma_o^2\mathbf{I}_T$.
- Setting $\mathbf{G} = \mathbf{\Sigma}_o^{-1/2}$, where $\mathbf{\Sigma}_o^{-1/2} = \mathbf{C}\mathbf{\Lambda}^{-1/2}\mathbf{C}'$ and $\mathbf{C}$ orthogonally diagonalizes $\mathbf{\Sigma}_o$: $\mathbf{C}'\mathbf{\Sigma}_o\mathbf{C} = \mathbf{\Lambda}$, we have $\mathbf{\Sigma}_o^{-1/2}\mathbf{\Sigma}_o\mathbf{\Sigma}_o^{-1/2\prime} = \mathbf{I}_T$.

- With $\mathbf{y}^* = \mathbf{\Sigma}_o^{-1/2}\mathbf{y}$ and $\mathbf{X}^* = \mathbf{\Sigma}_o^{-1/2}\mathbf{X}$, we have the GLS estimator:

$$\hat{\beta}_{\mathrm{GLS}} = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{y}^* = (\mathbf{X}'\mathbf{\Sigma}_o^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{\Sigma}_o^{-1}\mathbf{y}). \tag{5}$$

- The $\hat{\beta}_{\mathrm{GLS}}$ is a minimizer of weighted sum of squared errors:

$$Q(\beta; \mathbf{\Sigma}_o) = \frac{1}{T}(\mathbf{y}^* - \mathbf{X}^*\beta)'(\mathbf{y}^* - \mathbf{X}^*\beta) = \frac{1}{T}(\mathbf{y} - \mathbf{X}\beta)'\mathbf{\Sigma}_o^{-1}(\mathbf{y} - \mathbf{X}\beta).$$

- The vector of GLS fitted values, $\hat{\mathbf{y}}_{\mathrm{GLS}} = \mathbf{X}(\mathbf{X}'\mathbf{\Sigma}_o^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{\Sigma}_o^{-1}\mathbf{y})$, is an oblique projection of $\mathbf{y}$ onto span($\mathbf{X}$), because $\mathbf{X}(\mathbf{X}'\mathbf{\Sigma}_o^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}_o^{-1}$ is idempotent but not asymmetric. The GLS residual vector is $\hat{\mathbf{e}}_{\mathrm{GLS}} = \mathbf{y} - \hat{\mathbf{y}}_{\mathrm{GLS}}$.
- The sum of squared OLS residuals is less than the sum of squared GLS residuals. (Why?)

Stochastic Properties of the GLS Estimator

Theorem 4.1 (Aitken)

Given linear specification (1), suppose that [A1] and [A2](i) hold and that $\text{var}(\mathbf{y}) = \boldsymbol{\Sigma}_o$ is positive definite. Then, $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ is the BLUE for $\boldsymbol{\beta}_o$.

- Given [A3′] $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}_o, \boldsymbol{\Sigma}_o)$,

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} \sim \mathcal{N}\big(\boldsymbol{\beta}_o, (\mathbf{X}'\boldsymbol{\Sigma}_o^{-1}\mathbf{X})^{-1}\big).$$

- Under [A3′], the log likelihood function is

$$\log L(\boldsymbol{\beta}; \boldsymbol{\Sigma}_o) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\log(\det(\boldsymbol{\Sigma}_o)) - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}_o^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

with the FOC: $\mathbf{X}'\boldsymbol{\Sigma}_o^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$. Thus, the GLS estimator is also the MLE under normality.

- Under normality, the information matrix is

$$\mathbb{E}[\mathbf{X}'\boldsymbol{\Sigma}_o^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}_o^{-1}\mathbf{X}]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_o} = \mathbf{X}'\boldsymbol{\Sigma}_o^{-1}\mathbf{X}.$$

  Thus, the GLS estimator is the BUE for $\boldsymbol{\beta}_o$, because its covariance matrix reaches the Crámer-Rao lower bound.

- Under the null hypothesis $\mathbf{R}\boldsymbol{\beta}_o = \mathbf{r}$, we have

$$(\mathbf{R}\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\boldsymbol{\Sigma}_o^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} - \mathbf{r}) \sim \chi^2(q).$$

- A major difficulty: How should the GLS estimator be computed when $\boldsymbol{\Sigma}_o$ is unknown?

- Under normality, the information matrix is

$$\mathbb{E}[\mathbf{X}'\boldsymbol{\Sigma}_o^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}_o^{-1}\mathbf{X}]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_o} = \mathbf{X}'\boldsymbol{\Sigma}_o^{-1}\mathbf{X}.$$

Thus, the GLS estimator is the BUE for $\boldsymbol{\beta}_o$, because its covariance matrix reaches the Crámer-Rao lower bound.

- Under the null hypothesis $\mathbf{R}\boldsymbol{\beta}_o = \mathbf{r}$, we have

$$(\mathbf{R}\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\boldsymbol{\Sigma}_o^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} - \mathbf{r}) \sim \chi^2(q).$$

- A major difficulty: How should the GLS estimator be computed when $\boldsymbol{\Sigma}_o$ is unknown?

The Feasible GLS Estimator

- The Feasible GLS (FGLS) estimator is

$$\hat{\beta}_{\text{FGLS}} = (\mathbf{X}'\widehat{\boldsymbol{\Sigma}}_T^{-1}\mathbf{X})^{-1}\mathbf{X}'\widehat{\boldsymbol{\Sigma}}_T^{-1}\mathbf{y},$$

  where $\widehat{\boldsymbol{\Sigma}}_T$ is an estimator of $\boldsymbol{\Sigma}_o$.
- Further difficulties in FGLS estimation:
  - The number of parameters in $\boldsymbol{\Sigma}_o$ is $T(T+1)/2$. Estimating $\boldsymbol{\Sigma}_o$ without some prior restrictions on $\boldsymbol{\Sigma}_o$ is practically infeasible.
  - Even when an estimator $\widehat{\boldsymbol{\Sigma}}_T$ is available under certain assumptions, the finite-sample properties of the FGLS estimator are still difficult to derive.

The Feasible GLS Estimator

- The Feasible GLS (FGLS) estimator is

$$\hat{\beta}_{\mathrm{FGLS}} = (\mathbf{X}'\widehat{\mathbf{\Sigma}}_T^{-1}\mathbf{X})^{-1}\mathbf{X}'\widehat{\mathbf{\Sigma}}_T^{-1}\mathbf{y},$$

  where $\widehat{\mathbf{\Sigma}}_T$ is an estimator of $\mathbf{\Sigma}_o$.

- Further difficulties in FGLS estimation:
  - The number of parameters in $\mathbf{\Sigma}_o$ is $T(T+1)/2$. Estimating $\mathbf{\Sigma}_o$ without some prior restrictions on $\mathbf{\Sigma}_o$ is practically infeasible.
  - Even when an estimator $\widehat{\mathbf{\Sigma}}_T$ is available under certain assumptions, the finite-sample properties of the FGLS estimator are still difficult to derive.

Tests for Heteroskedasticity

A simple form of $\boldsymbol{\Sigma}_o$ is

$$\boldsymbol{\Sigma}_o = \begin{bmatrix} \sigma_1^2 \mathbf{I}_{T_1} & \mathbf{0} \\ \mathbf{0} & \sigma_2^2 \mathbf{I}_{T_2} \end{bmatrix},$$

with $T = T_1 + T_2$; this is known as groupwise heteroskedasticity.

- The null hypothesis of homoskedasticity: $\sigma_1^2 = \sigma_2^2 = \sigma_o^2$.
- Perform separate OLS regressions using the data in each group and obtain the variance estimates $\hat{\sigma}^2_{T_1}$ and $\hat{\sigma}^2_{T_2}$.
- Under [A1] and [A3′], the F test is:

$$\varphi := \frac{\hat{\sigma}^2_{T_1}}{\hat{\sigma}^2_{T_2}} = \frac{(T_1 - k)\hat{\sigma}^2_{T_1}}{\sigma_o^2 (T_1 - k)} \Bigg/ \frac{(T_2 - k)\hat{\sigma}^2_{T_2}}{\sigma_o^2 (T_2 - k)} \sim F(T_1 - k, T_2 - k).$$

- More generally, for some constants $c_0, c_1 > 0$, $\sigma_t^2 = c_0 + c_1 x_{tj}^2$.
- The Goldfeld-Quandt test:
  (1) Rearrange obs. according to the values of x_j in a descending order.
  (2) Divide the rearranged data set into three groups with T_1, T_m, and T_2 observations, respectively.
  (3) Drop the T_m observations in the middle group and perform separate OLS regressions using the data in the first and third groups.
  (4) The statistic is the ratio of the variance estimates:

$$\hat{\sigma}^2_{T_1} / \hat{\sigma}^2_{T_2} \sim F(T_1 - k, T_2 - k).$$

- Some questions:
  - Can we estimate the model with all observations and then compute $\hat{\sigma}^2_{T_1}$ and $\hat{\sigma}^2_{T_2}$ based on T_1 and T_2 residuals?
  - If Σ_o is not diagonal, does the F test above still work?

- More generally, for some constants $c_0, c_1 > 0$, $\sigma_t^2 = c_0 + c_1 x_{tj}^2$.
- The Goldfeld-Quandt test:
  (1) Rearrange obs. according to the values of x_j in a descending order.
  (2) Divide the rearranged data set into three groups with T_1, T_m, and T_2 observations, respectively.
  (3) Drop the T_m observations in the middle group and perform separate OLS regressions using the data in the first and third groups.
  (4) The statistic is the ratio of the variance estimates:

$$\hat{\sigma}^2_{T_1} / \hat{\sigma}^2_{T_2} \sim F(T_1 - k, T_2 - k).$$

- Some questions:
  - Can we estimate the model with all observations and then compute $\hat{\sigma}^2_{T_1}$ and $\hat{\sigma}^2_{T_2}$ based on T_1 and T_2 residuals?
  - If $\boldsymbol{\Sigma}_o$ is not diagonal, does the F test above still work?

- More generally, for some constants $c_0, c_1 > 0$, $\sigma_t^2 = c_0 + c_1 x_{tj}^2$.
- The Goldfeld-Quandt test:
  (1) Rearrange obs. according to the values of x_j in a descending order.
  (2) Divide the rearranged data set into three groups with T_1, T_m, and T_2 observations, respectively.
  (3) Drop the T_m observations in the middle group and perform separate OLS regressions using the data in the first and third groups.
  (4) The statistic is the ratio of the variance estimates:

$$\hat{\sigma}_{T_1}^2 / \hat{\sigma}_{T_2}^2 \sim F(T_1 - k, T_2 - k).$$

- Some questions:
  - Can we estimate the model with all observations and then compute $\hat{\sigma}_{T_1}^2$ and $\hat{\sigma}_{T_2}^2$ based on T_1 and T_2 residuals?
  - If $\boldsymbol{\Sigma}_o$ is not diagonal, does the F test above still work?

GLS and FGLS Estimation

Under groupwise heteroskedasticity,

$$\boldsymbol{\Sigma}_o^{-1/2} = \begin{bmatrix} \sigma_1^{-1}\mathbf{I}_{T_1} & \mathbf{0} \\ \mathbf{0} & \sigma_2^{-1}\mathbf{I}_{T_2} \end{bmatrix},$$

so that the transformed specification is

$$\begin{bmatrix} \mathbf{y}_1/\sigma_1 \\ \mathbf{y}_2/\sigma_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1/\sigma_1 \\ \mathbf{X}_2/\sigma_2 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{e}_1/\sigma_1 \\ \mathbf{e}_2/\sigma_2 \end{bmatrix}.$$

Clearly, $\text{var}(\boldsymbol{\Sigma}_o^{-1/2}\mathbf{y}) = \mathbf{I}_T$. The GLS estimator is:

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = \left[\frac{\mathbf{X}_1'\mathbf{X}_1}{\sigma_1^2} + \frac{\mathbf{X}_2'\mathbf{X}_2}{\sigma_2^2}\right]^{-1} \left[\frac{\mathbf{X}_1'\mathbf{y}_1}{\sigma_1^2} + \frac{\mathbf{X}_2'\mathbf{y}_2}{\sigma_2^2}\right].$$

With $\hat{\sigma}^2_{T_1}$ and $\hat{\sigma}^2_{T_2}$ from separate regressions, an estimator of $\mathbf{\Sigma}_o$ is

$$\widehat{\mathbf{\Sigma}} = \begin{bmatrix} \hat{\sigma}^2_{T_1}\mathbf{I}_{T_1} & \mathbf{0} \\ \mathbf{0} & \hat{\sigma}^2_{T_2}\mathbf{I}_{T_2} \end{bmatrix}.$$

The FGLS estimator is:

$$\hat{\boldsymbol{\beta}}_{\mathrm{FGLS}} = \left[\frac{\mathbf{X}_1'\mathbf{X}_1}{\hat{\sigma}_1^2} + \frac{\mathbf{X}_2'\mathbf{X}_2}{\hat{\sigma}_2^2}\right]^{-1} \left[\frac{\mathbf{X}_1'\mathbf{y}_1}{\hat{\sigma}_1^2} + \frac{\mathbf{X}_2'\mathbf{y}_2}{\hat{\sigma}_2^2}\right].$$

Note: If $\sigma_t^2 = c\, x_{tj}^2$, a transformed specification is

$$\frac{y_t}{x_{tj}} = \beta_j + \beta_1 \frac{1}{x_{tj}} + \cdots + \beta_{j-1}\frac{x_{t,j-1}}{x_{tj}} + \beta_{j+1}\frac{x_{t,j+1}}{x_{tj}} + \cdots + \beta_k \frac{x_{tk}}{x_{tj}} + \frac{e_t}{x_{tj}},$$

where $\mathrm{var}(y_t/x_{tj}) = c := \sigma_o^2$. Here, the GLS estimator is readily computed as the OLS estimator for the transformed specification.

Discussion and Remarks

- How do we determine the "groups" for groupwise heteroskedasticity?
- What if the diagonal elements of $\boldsymbol{\Sigma}_o$ take multiple values (so that there are more than 2 groups)?
- A general form of heteroskedasticity: $\sigma_t^2 = h(\alpha_0 + \mathbf{z}_t' \boldsymbol{\alpha}_1)$, with h unknown, $\mathbf{z}_t$ a $p \times 1$ vector and p a fixed number less than T.
- When the F test rejects the null of homoskedasticity, groupwise heteroskedasticity need not be a correct description of $\boldsymbol{\Sigma}_o$.
- When the form of heteroskedasticity is incorrectly specified, the resulting FGLS estimator may be less efficient than the OLS estimator.
- The finite-sample properties of FGLS estimators and hence the exact tests are typically unknown.

Serial Correlation

- When time series data y_t are correlated over time, they are said to exhibit serial correlation. For cross-section data, the correlations of y_t are known as spatial correlation.
- A general form of $\mathbf{\Sigma}_o$ is that its diagonal elements (variances of y_t) are a constant σ_o^2, and the off-diagonal elements ($\text{cov}(y_t, y_{t-i})$) are non-zero.
- In the time series context, $\text{cov}(y_t, y_{t-i})$ are known as the autocovariances of y_t, and the autocorrelations of y_t are

$$\text{corr}(y_t, y_{t-i}) = \frac{\text{cov}(y_t, y_{t-i})}{\sqrt{\text{var}(y_t)}\sqrt{\text{var}(y_{t-i})}} = \frac{\text{cov}(y_t, y_{t-i})}{\sigma_o^2}.$$

Simple Model: AR(1) Disturbances

- A time series y_t is said to be weakly (covariance) stationary if its mean, variance, and autocovariances are all independent of t.
  - i.i.d. random variables
  - White noise: A time series with zero mean, a constant variance, and zero autocovariances.
- Disturbance: $\boldsymbol{\epsilon} := \mathbf{y} - \mathbf{X}\boldsymbol{\beta}_o$ so that $\text{var}(\mathbf{y}) = \text{var}(\boldsymbol{\epsilon}) = \mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}')$. Suppose that ϵ_t follows a weakly stationary AR(1) (autoregressive of order 1) process:

$$\epsilon_t = \psi_1 \epsilon_{t-1} + u_t, \quad |\psi_1| < 1,$$

where $\{u_t\}$ is a white noise with $\mathbb{E}(u_t) = 0$, $\mathbb{E}(u_t^2) = \sigma_u^2$, and $\mathbb{E}(u_t u_\tau) = 0$ for $t \neq \tau$.

By recursive substitution,

$$\epsilon_t = \sum_{i=0}^{\infty} \psi_1^i u_{t-i},$$

a weighted sum of current and previous "innovations" (shocks). This is a stationary process because:

- $\mathbb{E}(\epsilon_t) = 0$, $\operatorname{var}(\epsilon_t) = \sum_{i=0}^{\infty} \psi_1^{2i} \sigma_u^2 = \sigma_u^2/(1-\psi_1^2)$, and

  $$\operatorname{cov}(\epsilon_t, \epsilon_{t-1}) = \psi_1 \, \mathbb{E}(\epsilon_{t-1}^2) = \psi_1 \sigma_u^2/(1-\psi_1^2),$$

  so that $\operatorname{corr}(\epsilon_t, \epsilon_{t-1}) = \psi_1$.
- $\operatorname{cov}(\epsilon_t, \epsilon_{t-2}) = \psi_1 \operatorname{cov}(\epsilon_{t-1}, \epsilon_{t-2})$ so that $\operatorname{corr}(\epsilon_t, \epsilon_{t-2}) = \psi_1^2$. Thus,

  $$\operatorname{corr}(\epsilon_t, \epsilon_{t-i}) = \psi_1 \operatorname{corr}(\epsilon_{t-1}, \epsilon_{t-i}) = \psi_1^i,$$

  which depend only on i, but not on t.

The variance-covariance matrix var($\mathbf{y}$) is thus

$$\boldsymbol{\Sigma}_o = \sigma_o^2 \begin{bmatrix} 1 & \psi_1 & \psi_1^2 & \cdots & \psi_1^{T-1} \\ \psi_1 & 1 & \psi_1 & \cdots & \psi_1^{T-2} \\ \psi_1^2 & \psi_1 & 1 & \cdots & \psi_1^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \psi_1^{T-1} & \psi_1^{T-2} & \psi_1^{T-3} & \cdots & 1 \end{bmatrix},$$

with $\sigma_o^2 = \sigma_u^2/(1-\psi_1^2)$. Note that all off-diagonal elements of this matrix are non-zero, but there are only two unknown parameters.

A transformation matrix for GLS estimation is the following $\boldsymbol{\Sigma}_o^{-1/2}$:

$$\frac{1}{\sigma_o}\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -\frac{\psi_1}{\sqrt{1-\psi_1^2}} & \frac{1}{\sqrt{1-\psi_1^2}} & 0 & \cdots & 0 & 0 \\ 0 & -\frac{\psi_1}{\sqrt{1-\psi_1^2}} & \frac{1}{\sqrt{1-\psi_1^2}} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{\sqrt{1-\psi_1^2}} & 0 \\ 0 & 0 & 0 & \cdots & -\frac{\psi_1}{\sqrt{1-\psi_1^2}} & \frac{1}{\sqrt{1-\psi_1^2}} \end{bmatrix}.$$

Any matrix that is a constant proportion to $\boldsymbol{\Sigma}_o^{-1/2}$ can also serve as a legitimate transformation matrix for GLS estimation

The Cochrane-Orcutt Transformation is based on:

$$\mathbf{V}_o^{-1/2} = \sigma_o \sqrt{1-\psi_1^2}\, \boldsymbol{\Sigma}_o^{-1/2} = \begin{bmatrix} \sqrt{1-\psi_1^2} & 0 & 0 & \cdots & 0 & 0 \\ -\psi_1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\psi_1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & -\psi_1 & 1 \end{bmatrix},$$

which depends only on the single parameter ψ_1. The resulting transformed data are: $\mathbf{y}^* = \mathbf{V}_o^{-1/2}\mathbf{y}$ and $\mathbf{X}^* = \mathbf{V}_o^{-1/2}\mathbf{X}$ with

$$\begin{aligned} y_1^* &= (1-\psi_1^2)^{1/2} y_1, & \mathbf{x}_1^* &= (1-\psi_1^2)^{1/2}\mathbf{x}_1, & & \\ y_t^* &= y_t - \psi_1 y_{t-1}, & \mathbf{x}_t^* &= \mathbf{x}_t - \psi_1 \mathbf{x}_{t-1}, & & t = 2, \cdots, T, \end{aligned}$$

where $\mathbf{x}_t$ is the t th column of $\mathbf{X}'$.

Model Extensions

- Extension to AR(p) process:

$$\epsilon_t = \psi_1 \epsilon_{t-1} + \cdots + \psi_p \epsilon_{t-p} + u_t,$$

where $\psi_1, \ldots, \psi_p$ must be restricted to ensure weak stationarity.

- MA(1) (moving average of order 1) process:

$$\epsilon_t = u_t - \pi_1 u_{t-1}, \quad |\pi_1| < 1,$$

where $\{u_t\}$ is a white noise.

  - $\mathbb{E}(\epsilon_t) = 0$, $\text{var}(\epsilon_t) = (1 + \pi_1^2)\sigma_u^2$.
  - $\text{cov}(\epsilon_t, \epsilon_{t-1}) = -\pi_1 \sigma_u^2$, and $\text{cov}(\epsilon_t, \epsilon_{t-i}) = 0$ for $i \geq 2$.

- MA(q) Process: $\epsilon_t = u_t - \pi_1 u_{t-1} - \cdots - \pi_q u_{t-q}$.

Tests for AR(1) Disturbances

Under AR(1), the null hypothesis is $\psi_1 = 0$. A natural estimator of ψ_1 is the OLS estimator of regressing $\hat{e}_t$ on $\hat{e}_{t-1}$:

$$\hat{\psi}_T = \frac{\sum_{t=2}^{T} \hat{e}_t \hat{e}_{t-1}}{\sum_{t=2}^{T} \hat{e}_{t-1}^2}.$$

- The Durbin-Watson statistic is

$$d = \frac{\sum_{t=2}^{T} (\hat{e}_t - \hat{e}_{t-1})^2}{\sum_{t=1}^{T} \hat{e}_t^2}.$$

- When the sample size T is large, it can be seen that

$$d = 2 - 2\hat{\psi}_T \frac{\sum_{t=2}^{T} \hat{e}_{t-1}^2}{\sum_{t=1}^{T} \hat{e}_t^2} - \frac{\hat{e}_1^2 + \hat{e}_T^2}{\sum_{t=1}^{T} \hat{e}_t^2} \approx 2(1 - \hat{\psi}_T).$$

- For $0 < \hat{\psi}_T \leq 1$ ($-1 \leq \hat{\psi}_T < 0$), $0 \leq d < 2$ ($2 < d \leq 4$), there may be positive (negative) serial correlation. Hence, d essentially checks whether $\hat{\psi}_T$ is "close" to zero (i.e., d is "close" to 2).
- Difficulty: The exact null distribution of d holds only under the classical conditions [A1] and [A3] and depends on the data matrix $\mathbf{X}$. Thus, the critical values for d can not be tabulated, and this test is not pivotal.
- The null distribution of d lies between a lower bound (d_L) and an upper bound (d_U):

$$d^*_{L,\alpha} < d^*_\alpha < d^*_{U,\alpha}.$$

The distributions of d_L and d_U are not data dependent, so that their critical values $d^*_{L,\alpha}$ and $d^*_{U,\alpha}$ can be tabulated.

- Durbin-Watson test:
  (1) Reject the null if $d < d^*_{L,\alpha}$ ($d > 4 - d^*_{L,\alpha}$).
  (2) Do not reject the null if $d > d^*_{U,\alpha}$ ($d < 4 - d^*_{U,\alpha}$).
  (3) Test is inconclusive if $d^*_{L,\alpha} < d < d^*_{U,\alpha}$ ($4 - d^*_{L,\alpha} > d > 4 - d^*_{U,\alpha}$).
- For the specification $y_t = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + \gamma y_{t-1} + e_t$, Durbin's h statistic is

$$h = \hat{\gamma}_T \sqrt{\frac{T}{1 - T\widehat{\text{var}}(\hat{\gamma}_T)}} \approx \mathcal{N}(0,1),$$

where $\hat{\gamma}_T$ is the OLS estimate of γ with $\widehat{\text{var}}(\hat{\gamma}_T)$ the OLS estimate of $\text{var}(\hat{\gamma}_T)$.

Note: $\widehat{\text{var}}(\hat{\gamma}_T)$ can not be greater $1/T$. (Why?)

FGLS Estimation

- Notations: Write $\boldsymbol{\Sigma}(\sigma^2, \psi)$ and $\mathbf{V}(\psi)$, so that $\boldsymbol{\Sigma}_o = \boldsymbol{\Sigma}(\sigma_o^2, \psi_1)$ and $\mathbf{V}_o = \mathbf{V}(\psi_1)$. Based on $\mathbf{V}(\psi)^{-1/2}$, we have

$$\begin{aligned} y_1(\psi) &= (1-\psi^2)^{1/2} y_1, \quad & \mathbf{x}_1(\psi) &= (1-\psi^2)^{1/2}\mathbf{x}_1, \\ y_t(\psi) &= y_t - \psi y_{t-1}, & \mathbf{x}_t(\psi) &= \mathbf{x}_t - \psi \mathbf{x}_{t-1}, \qquad t = 2, \cdots, T. \end{aligned}$$

- Iterative FGLS Estimation:
  - (1) Perform OLS estimation and compute $\hat{\psi}_T$ using the OLS residuals $\hat{e}_t$.
  - (2) Perform the Cochrane-Orcutt transformation based on $\hat{\psi}_T$ and compute the resulting FGLS estimate $\hat{\boldsymbol{\beta}}_{\mathrm{FGLS}}$ by regressing $y_t(\hat{\psi}_T)$ on $\mathbf{x}_t(\hat{\psi}_T)$.
  - (3) Compute a new $\hat{\psi}_T$ with $\hat{e}_t$ replaced by $\hat{e}_{t,\mathrm{FGLS}} = y_t - \mathbf{x}_t' \hat{\boldsymbol{\beta}}_{\mathrm{FGLS}}$.
  - (4) Repeat steps (2) and (3) until $\hat{\psi}_T$ converges numerically.

  Steps (1) and (2) suffice for FGLS estimation; more iterations may improve the performance in finite samples.

Instead of estimating $\hat{\psi}_T$ based on OLS residuals, the Hildreth-Lu procedure adopts grid search to find a suitable $\psi \in (-1, 1)$.

- For a ψ in $(-1, 1)$, conduct the Cochrane-Orcutt transformation and compute the resulting FGLS estimate (by regressing $y_t(\psi)$ on $\mathbf{x}_t(\psi)$) and the ESS based on the FGLS residuals.
- Try every ψ on the grid; a ψ is chosen if the corresponding ESS is the smallest.
- The results depend on the grid.

Note: This method is computationally intensive and difficult to apply when ϵ_t follow an AR(p) process with $p > 2$.

Application: Linear Probability Model

Consider binary y with $y = 1$ or 0.

- Under [A1] and [A2](i), $\mathbb{E}(y_t) = \mathbb{P}(y_t = 1) = \mathbf{x}_t'\boldsymbol{\beta}_o$; this is known as the linear probability model.
- Problems with the linear probability model:
  - Under [A1] and [A2](i), there is heteroskedasticity:

    $$\text{var}(y_t) = \mathbf{x}_t'\boldsymbol{\beta}_o(1 - \mathbf{x}_t'\boldsymbol{\beta}_o),$$

    and hence the OLS estimator is not the BLUE for $\boldsymbol{\beta}_o$.
  - The OLS fitted values $\mathbf{x}_t'\hat{\boldsymbol{\beta}}_T$ need not be bounded between 0 and 1.

Application: Linear Probability Model

Consider binary y with $y = 1$ or 0.

- Under [A1] and [A2](i), $\mathbb{E}(y_t) = \mathbb{P}(y_t = 1) = \mathbf{x}_t'\boldsymbol{\beta}_o$; this is known as the linear probability model.
- Problems with the linear probability model:
  - Under [A1] and [A2](i), there is heteroskedasticity:

    $$\text{var}(y_t) = \mathbf{x}_t'\boldsymbol{\beta}_o(1 - \mathbf{x}_t'\boldsymbol{\beta}_o),$$

    and hence the OLS estimator is not the BLUE for $\boldsymbol{\beta}_o$.
  - The OLS fitted values $\mathbf{x}_t'\hat{\boldsymbol{\beta}}_T$ need not be bounded between 0 and 1.

- An FGLS estimator may be obtained using

$$\widehat{\boldsymbol{\Sigma}}_T^{-1/2} = \text{diag}\Big[[\mathbf{x}_1'\hat{\boldsymbol{\beta}}_T(1 - \mathbf{x}_1'\hat{\boldsymbol{\beta}}_T)]^{-1/2}, \ldots, \\ [\mathbf{x}_T'\hat{\boldsymbol{\beta}}_T(1 - \mathbf{x}_T'\hat{\boldsymbol{\beta}}_T)]^{-1/2}\Big].$$

- Problems with FGLS estimation:
  - $\widehat{\boldsymbol{\Sigma}}_T^{-1/2}$ can not be computed if $\mathbf{x}_t'\hat{\boldsymbol{\beta}}_T$ is not bounded between 0 and 1.
  - Even when $\widehat{\boldsymbol{\Sigma}}_T^{-1/2}$ is available, there is no guarantee that the FGLS fitted values are bounded between 0 and 1.
  - The finite-sample properties of the FGLS estimator are unknown.
- A key issue: A linear model here fails to take into account data characteristics.

Application: Seemingly Unrelated Regressions

To study the joint behavior of several dependent variables, consider a system of N equations, each with k_i explanatory variables and T obs:

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta}_i + \mathbf{e}_i, \qquad i = 1, 2, \ldots, N.$$

Stacking these equations yields Seemingly unrelated regressions (SUR):

$$\underbrace{\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_N \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X}_N \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_N \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \mathbf{e}_N \end{bmatrix}}_{\mathbf{e}}.$$

where $\mathbf{y}$ is $TN \times 1$, $\mathbf{X}$ is $TN \times \sum_{i=1}^{N} k_i$, and $\boldsymbol{\beta}$ is $\sum_{i=1}^{N} k_i \times 1$.

- Suppose y_{it} and y_{jt} are contemporaneously correlated, but y_{it} and $y_{j\tau}$ are serially uncorrelated, i.e., $\operatorname{cov}(\mathbf{y}_i, \mathbf{y}_j) = \sigma_{ij}\mathbf{I}_T$.
- For this system, $\boldsymbol{\Sigma}_o = \mathbf{S}_o \otimes \mathbf{I}_T$ with

$$\mathbf{S}_o = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1N} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \cdots & \sigma_N^2 \end{bmatrix};$$

that is, the SUR system has both serial and spatial correlations.

- As $\boldsymbol{\Sigma}_o^{-1} = \mathbf{S}_o^{-1} \otimes \mathbf{I}_T$, then

$$\hat{\beta}_{\mathrm{GLS}} = [\mathbf{X}'(\mathbf{S}_o^{-1} \otimes \mathbf{I}_T)\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{S}_o^{-1} \otimes \mathbf{I}_T)\mathbf{y},$$

and its covariance matrix is $[\mathbf{X}'(\mathbf{S}_o^{-1} \otimes \mathbf{I}_T)\mathbf{X}]^{-1}$.

- Remarks:
  - When $\sigma_{ij} = 0$ for $i \neq j$, $\mathbf{S}_o$ is diagonal, and so is $\boldsymbol{\Sigma}_o$. Then, the GLS estimator for each $\boldsymbol{\beta}_i$ reduces to the corresponding OLS estimator, so that joint estimation of N equations is not necessary.
  - If all equations in the system have the same regressors, i.e., $\mathbf{X}_i = \mathbf{X}_0$ (say) and $\mathbf{X} = \mathbf{I}_N \otimes \mathbf{X}_0$, the GLS estimator is also the same as the OLS estimator.
  - More generally, there would not be much efficiency gain for GLS estimation if $\mathbf{y}_i$ and $\mathbf{y}_j$ are less correlated and/or $\mathbf{X}_i$ and $\mathbf{X}_j$ are highly correlated.
- The FGLS estimator can be computed as

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = [\mathbf{X}'(\widehat{\mathbf{S}}_{TN}^{-1} \otimes \mathbf{I}_T)\mathbf{X}]^{-1}\mathbf{X}'(\widehat{\mathbf{S}}_{TN}^{-1} \otimes \mathbf{I}_T)\mathbf{y}.$$

- $\widehat{\mathbf{S}}_{TN}$ is an $N \times N$ matrix:

$$\widehat{\mathbf{S}}_{TN} = \frac{1}{T} \begin{bmatrix} \hat{\mathbf{e}}_1' \\ \hat{\mathbf{e}}_2' \\ \vdots \\ \hat{\mathbf{e}}_N' \end{bmatrix} \begin{bmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \ldots & \hat{\mathbf{e}}_N \end{bmatrix},$$

  where $\hat{\mathbf{e}}_i$ is the OLS residual vector of the i th equation.

- The estimator $\widehat{\mathbf{S}}_{TN}$ is valid provided that $\text{var}(\mathbf{y}_i) = \sigma_i^2 \mathbf{I}_T$ and $\text{cov}(\mathbf{y}_i, \mathbf{y}_j) = \sigma_{ij} \mathbf{I}_T$. Without these assumptions, FGLS estimation would be more complicated.

- Again, the finite-sample properties of the FGLS estimator are unknown.