

Classical Least Squares Theory

CHUNG-MING KUAN

Department of Finance & CRETA

National Taiwan University

October 18, 2014

Lecture Outline

1 The Method of Ordinary Least Squares (OLS)

- Simple Linear Regression
- Multiple Linear Regression
- Geometric Interpretations
- Measures of Goodness of Fit
- Example: Analysis of Suicide Rate

2 Statistical Properties of the OLS Estimator

- Classical Conditions
- Without the Normality Condition
- With the Normality Condition

3 Hypothesis Testing

- Tests for Linear Hypotheses
- Power of the Tests
- Alternative Interpretation of the F Test
- Confidence Regions
- Example: Analysis of Suicide Rate

Lecture Outline (cont'd)

4 Multicollinearity

- Near Multicollinearity
- Regression with Dummy Variables
- Example: Analysis of Suicide Rate

5 Limitation of the Classical Conditions

6 The Method of Generalized Least Squares (GLS)

- The GLS Estimator
- Stochastic Properties of the GLS Estimator
- The Feasible GLS Estimator
- Heteroskedasticity
- Serial Correlation
- Application: Linear Probability Model
- Application: Seemingly Unrelated Regressions

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:

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

For the specification $\alpha + \beta x_t$, the objective is to find the “best” fit of the data (y_t, x_t) , $t = 1, \dots, T$.

- 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 $0 < \theta < 1$.

For the specification $\alpha + \beta x$, the objective is to find the “best” fit of the data (y_t, x_t) , $t = 1, \dots, T$.

- 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 $0 < \theta < 1$.

For the specification $\alpha + \beta x$, the objective is to find the “best” fit of the data (y_t, x_t) , $t = 1, \dots, T$.

- 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 OLS Estimators

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

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

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

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

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

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

- The estimated regression line is $\hat{y} = \hat{\alpha}_T + \hat{\beta}_T x$, which is the linear function evaluated at $\hat{\alpha}_T$ and $\hat{\beta}_T$, and $\hat{e} = y - \hat{y}$ is the error evaluated at $\hat{\alpha}_T$ and $\hat{\beta}_T$ and also known as **residual**.
 - The t -th **fitted value** of the regression line is $\hat{y}_t = \hat{\alpha}_T + \hat{\beta}_T x_t$.
 - The t -th **residual** is $\hat{e}_t = y_t - \hat{y}_t = e_t(\hat{\alpha}_T, \hat{\beta}_T)$.
 - **No** other linear functions of the form $a + bx$ can provide a better fit of the data in terms of sum of squared errors.
- $\hat{\beta}_T$ characterizes the the predicted change of y , given a change of one unit of x , whereas $\hat{\alpha}_T$ is the predicted y without (the information of) x .
- Note that the OLS method requires **no** assumption on the data, except that x_t can **not** be a constant.

Algebraic Properties

Substituting $\hat{\alpha}_T$ and $\hat{\beta}_T$ into the FOCs:

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

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 $\bar{y} = \bar{\hat{y}}$.
- $\bar{y} = \hat{\alpha}_T + \hat{\beta}_T \bar{x}$; that is, the estimated regression line must pass through the point (\bar{x}, \bar{y}) .

Example: Analysis of Suicide Rate

- Suppose we want to know how the suicide rate (s) in Taiwan can be explained by unemployment rate (u), GDP growth rate (g), or time (t). The suicide rate is 1/100000.
- Data (1981–2013): $\bar{s} = 12.05$ with s.d. 3.91; $\bar{g} = 5.64$ with s.d. 3.16; $\bar{u} = 3.09$ with s.d. 1.33.
- Estimation results:

$$\hat{s}_t = 15.11 - 0.54 g_t, \quad \bar{R}^2 = 0.17;$$

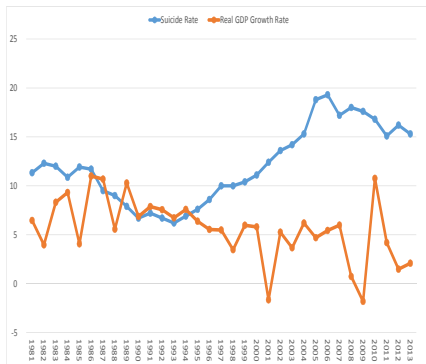
$$\hat{s}_t = 15.56 - 0.61 g_{t-1}, \quad \bar{R}^2 = 0.21;$$

$$\hat{s}_t = 4.40 + 2.48 u_t, \quad \bar{R}^2 = 0.70;$$

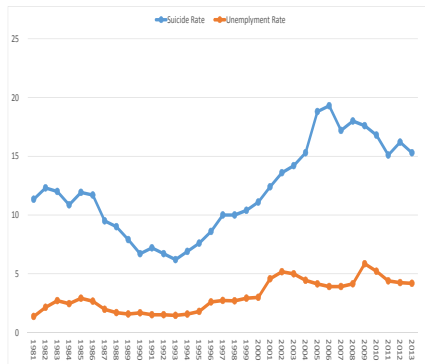
$$\hat{s}_t = 4.84 + 2.40 u_{t-1}, \quad \bar{R}^2 = 0.68;$$

$$\hat{s}_t = 7.40 + 0.27 t, \quad \bar{R}^2 = 0.44.$$

Suicide Rates 1981–2013



(a) Suicide & GDP growth rates



(b) Suicide and unemploy. rates

Multiple Linear Regression

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

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

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

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

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

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

- Least-squares criterion function:

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

- The FOCs of minimizing $Q_T(\beta)$ are $-2\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta)/T = \mathbf{0}$, leading to the **normal equations**:

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

- **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.
 - Intuition: \mathbf{X} does not contain **redundant** information.
 - When \mathbf{X} is **not** of full column rank, we say there exists **exact multicollinearity** among regressors.

- Given [ID-1], $\mathbf{X}'\mathbf{X}$ is positive definite and hence invertible. The **unique** solution to the normal equations is known as the OLS estimator of β :

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

- Under [ID-1], we have the second order condition:

$$\nabla_{\beta}^2 Q_T(\beta) = 2(\mathbf{X}'\mathbf{X})/T \text{ is p.d.}$$

- The result below requires only the identification requirement and does **not** depend on the statistical properties of \mathbf{y} and \mathbf{X} .

Theorem 3.1

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

- The magnitude of $\hat{\beta}_T$ is affected by the measurement units of the dependent and explanatory variables.
 - A larger coefficient does **not** imply that the associated regressor is more important.
 - The so-called “beta coefficients” (see homework) do not depend on the measurement units, and hence their magnitudes are comparable.
- Given $\hat{\beta}_T$, the vector of the OLS fitted values is $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}_T$, and the vector of the OLS residuals is $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{e}(\hat{\beta}_T)$.
- Plugging $\hat{\beta}_T$ into the FOCs: $\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta) = \mathbf{0}$, we have:
 - $\mathbf{X}'\hat{\mathbf{e}} = \mathbf{0}$.
 - When \mathbf{X} contains a vector of ones, $\sum_{t=1}^T \hat{e}_t = 0$.
 - $\hat{\mathbf{y}}'\hat{\mathbf{e}} = \hat{\beta}_T'\mathbf{X}'\hat{\mathbf{e}} = 0$.

These are all algebraic results under the OLS method.

Geometric Interpretations

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

- The vector of fitted values, $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}_T = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{P}\mathbf{y}$, is the **orthogonal projection** of \mathbf{y} onto $\text{span}(\mathbf{X})$.
- 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 $\text{span}(\mathbf{X})^\perp$.
- $\hat{\mathbf{e}}$ 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 $\text{span}(\mathbf{X})$, i.e., $\hat{\mathbf{y}}'\hat{\mathbf{e}} = 0$.

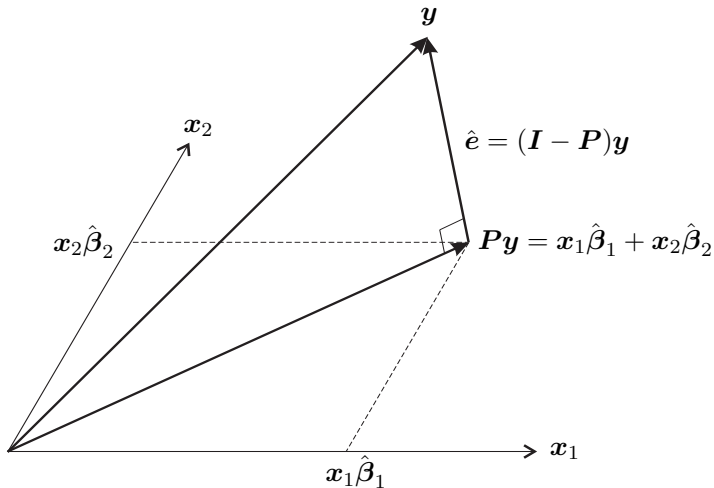


Figure: The orthogonal projection of \mathbf{y} onto $\text{span}(\mathbf{x}_1, \mathbf{x}_2)$.

Theorem 3.3 (Frisch-Waugh-Lovell)

Given $\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{e}$, the OLS estimators of β_1 and β_2 are

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

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

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

- $\hat{\beta}_{1,T}$ can also 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. That is, $\hat{\beta}_{1,T}$ characterizes the marginal effect of \mathbf{X}_1 on \mathbf{y} , after the effect of \mathbf{X}_2 is purged away.
- Similarly, regressing $(\mathbf{I} - \mathbf{P}_1)\mathbf{y}$ on $(\mathbf{I} - \mathbf{P}_1)\mathbf{X}_2$ yields $\hat{\beta}_{2,T}$. Thus, That is, $\hat{\beta}_{2,T}$ characterizes the marginal effect of \mathbf{X}_x on \mathbf{y} , after the effect of \mathbf{X}_1 is purged away.

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

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

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 $\hat{\beta}_{1,T}$.

Some Implications of the FWL Theorem

- The OLS estimator of regressing \mathbf{y} on \mathbf{X}_1 is **not** the same as $\hat{\beta}_{1,T}$, unless \mathbf{X}_1 and \mathbf{X}_2 are orthogonal to each other.
- Regressing \mathbf{y} on \mathbf{X}_1 and \mathbf{X}_2 enables us to assess the marginal effect of \mathbf{X}_1 on \mathbf{y} while “controlling” the effect of \mathbf{X}_2 and also the marginal effect of \mathbf{X}_2 on \mathbf{y} while “controlling” the effect of \mathbf{X}_1 .
- When one regresses \mathbf{y} on \mathbf{X}_1 only, the estimated marginal effect of \mathbf{X}_1 is affected by other relevant variables not included in the regression.

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,

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

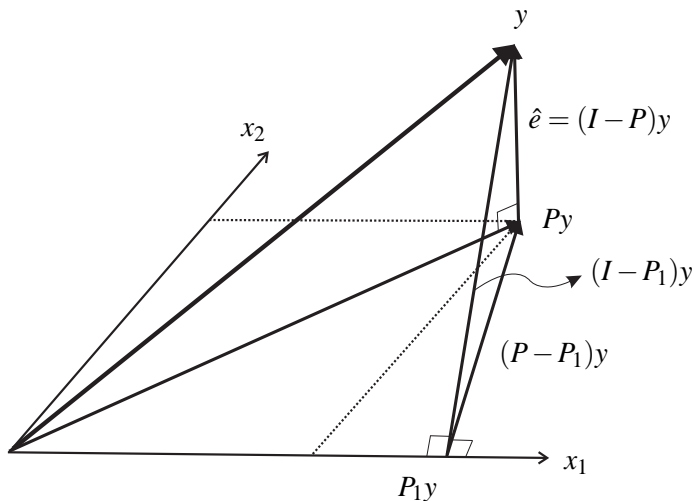


Figure: An illustration of the Frisch-Waugh-Lovell Theorem.

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),

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

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,

$$\underbrace{\sum_{t=1}^T (y_t - \bar{y})^2}_{\text{centered TSS}} = \underbrace{\sum_{t=1}^T (\hat{y}_t - \bar{\hat{y}})^2}_{\text{centered RSS}} + \underbrace{\sum_{t=1}^T \hat{e}_t^2}_{\text{ESS}}$$

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

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

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 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 .

Example: Analysis of Suicide Rate

- Q: How the suicide rate (s) can be explained by unemployment rate (u), GDP growth rate (g), and time (t) during 1981–2010?
- Estimation results with g_t and u_t :

$$\hat{s}_t = 15.11 - 0.54 g_t, \quad \bar{R}^2 = 0.17;$$

$$\hat{s}_t = 4.40 + 2.48 u_t, \quad \bar{R}^2 = 0.70;$$

$$\hat{s}_t = 4.20 + 2.50 u_t + 0.02 g_t, \quad \bar{R}^2 = 0.69.$$

Estimation results with g_{t-1} and u_{t-1} :

$$\hat{s}_t = 15.56 - 0.61 g_{t-1}, \quad \bar{R}^2 = 0.21;$$

$$\hat{s}_t = 4.84 + 2.40 u_{t-1}, \quad \bar{R}^2 = 0.68;$$

$$\hat{s}_t = 5.56 + 2.31 u_{t-1} - 0.08 g_{t-1}, \quad \bar{R}^2 = 0.67.$$

- Estimation results with t but without g :

$$\hat{s}_t = 4.40 + 2.48 u_t, \quad \bar{R}^2 = 0.70;$$

$$\hat{s}_t = 4.41 + 2.41 u_t + 0.01 t, \quad \bar{R}^2 = 0.69;$$

$$\hat{s}_t = 4.84 + 2.40 u_{t-1}, \quad \bar{R}^2 = 0.68;$$

$$\hat{s}_t = 4.84 + 2.35 u_{t-1} + 0.01 t, \quad \bar{R}^2 = 0.67.$$

- Estimation results with t and g :

$$\hat{s}_t = 4.20 + 2.50 u_t + 0.02 g_t, \quad \bar{R}^2 = 0.69;$$

$$\hat{s}_t = 4.16 + 2.43 u_t + 0.03 g_t + 0.01 t, \quad \bar{R}^2 = 0.68;$$

$$\hat{s}_t = 5.56 + 2.31 u_{t-1} - 0.08 g_{t-1}, \quad \bar{R}^2 = 0.67;$$

$$\hat{s}_t = 5.55 + 2.29 u_{t-1} - 0.08 g_{t-1} + 0.004 t, \quad \bar{R}^2 = 0.66.$$

- Estimation results with t and t^2 :

$$\hat{s}_t = 7.40 + 0.27 t, \quad \bar{R}^2 = 0.44;$$

$$\hat{s}_t = 11.60 - 0.45 t + 0.02 t^2, \quad \bar{R}^2 = 0.63;$$

$$\hat{s}_t = 8.05 + 1.93 u_t - 0.46 t + 0.02 t^2, \quad \bar{R}^2 = 0.78;$$

$$\hat{s}_t = 12.27 - 0.09 g_t - 0.45 t + 0.02 t^2, \quad \bar{R}^2 = 0.62;$$

$$\hat{s}_t = 7.68 + 1.96 u_t + 0.04 g_t - 0.46 t + 0.02 t^2, \quad \bar{R}^2 = 0.77;$$

$$\hat{s}_t = 8.37 + 1.80 u_{t-1} - 0.43 t + 0.01 t^2, \quad \bar{R}^2 = 0.75;$$

$$\hat{s}_t = 13.02 - 0.18 g_{t-1} - 0.45 t + 0.02 t^2, \quad \bar{R}^2 = 0.64;$$

$$\hat{s}_t = 9.03 + 1.74 u_{t-1} - 0.07 g_{t-1} - 0.43 t + 0.01 t^2, \quad \bar{R}^2 = 0.74.$$

- As far as \bar{R}^2 is concerned, a specification with t , t^2 , and u seems to provide good fit of data and reasonable interpretation.

Q: Is there any other way to determine if a specification is “good”?

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) $\mathbb{E}(\mathbf{y}) = \mathbf{X}\beta_o$ for some β_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. $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta_o, \sigma_o^2 \mathbf{I}_T)$ for some β_o and $\sigma_o^2 > 0$.

- The specification (1) with [A1] and [A2] is known as the **classical linear model**, whereas (1) with [A1] and [A3] is the **classical normal linear model**.
- When $\text{var}(\mathbf{y}) = \sigma_o^2 \mathbf{I}_T$, the elements of \mathbf{y} are **homoskedastic** and (serially) **uncorrelated**.

Without Normality

The OLS estimator of the parameter σ_o^2 is an average of squared residuals:

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

Theorem 3.4

Consider the linear specification (1).

- (a) Given [A1] and [A2](i), $\hat{\beta}_T$ is unbiased for β_o .
- (b) Given [A1] and [A2], $\hat{\sigma}_T^2$ is unbiased for σ_o^2 .
- (c) Given [A1] and [A2], $\text{var}(\hat{\beta}_T) = \sigma_o^2(\mathbf{X}'\mathbf{X})^{-1}$.

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

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

proving unbiasedness. Given $\hat{\mathbf{e}} = (\mathbf{I}_T - \mathbf{P})\mathbf{y} = (\mathbf{I}_T - \mathbf{P})(\mathbf{y} - \mathbf{X}\beta_o)$,

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

where the 4-th equality follows from [A2](ii) that $\text{var}(\mathbf{y}) = \sigma_o^2 \mathbf{I}_T$.

Proof (cont'd): 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

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

proving (b). By [A1] and [A2](ii),

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

This establishes the assertion of (c).

- Theorem 3.4 establishes unbiasedness of the OLS estimators $\hat{\beta}_T$ and $\hat{\sigma}_T^2$ but does **not** address the issue of efficiency.
- By Theorem 3.4(c), the elements of $\hat{\beta}_T$ can be more precisely estimated (i.e., with a smaller variance) when \mathbf{X} has larger variation. To see this, consider the simple linear regression: $y = \alpha + \beta x + e$, it can be verified that

$$\text{var}(\hat{\beta}_T) = \sigma_o^2 \frac{1}{\sum_{t=1}^T (x_t - \bar{x})^2}.$$

Thus, the larger the (squared) variation of x_t (i.e., $\sum_{t=1}^T (x_t - \bar{x})^2$), the smaller is the variance of $\hat{\beta}_T$.

The result below establishes efficiency of $\hat{\beta}_T$ among all unbiased estimators of β_o that are linear in \mathbf{y} .

Theorem 3.5 (Gauss-Markov)

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

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

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

By [A1] and [A2](i), $\mathbb{E}(\check{\beta}_T) = \beta_o + \mathbf{C}\mathbf{X}\beta_o$, which is unbiased iff $\mathbf{C}\mathbf{X} = \mathbf{0}$.

Proof (cont'd): The condition $\mathbf{CX} = \mathbf{0}$ implies $\text{cov}(\hat{\beta}_T, \mathbf{Cy}) = \mathbf{0}$. Thus,

$$\text{var}(\check{\beta}_T) = \text{var}(\hat{\beta}_T) + \text{var}(\mathbf{Cy}) = \text{var}(\hat{\beta}_T) + \sigma_o^2 \mathbf{CC}'.$$

This shows that $\text{var}(\check{\beta}_T) - \text{var}(\hat{\beta}_T)$ is a p.s.d. matrix $\sigma_o^2 \mathbf{CC}'$, so that $\hat{\beta}_T$ is more efficient than any linear unbiased estimator $\check{\beta}_T$.

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

$$\mathbf{y} = \mathbf{X}_1 \beta_1 + \mathbf{e}.$$

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

$$\mathbf{y} = \mathbf{X} \beta + \mathbf{e} = \mathbf{X}_1 \beta_1 + \mathbf{X}_2 \beta_2 + \mathbf{e}.$$

with the OLS estimator $\hat{\beta}_T = (\hat{\beta}'_{1,T} \hat{\beta}'_{2,T})'$. 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,

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

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

That is, $\hat{\beta}_T$ 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., it follows that

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

is p.s.d. This shows that $\hat{\mathbf{b}}_{1,T}$ 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

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

With the normality condition on \mathbf{y} , a lot more can be said about the OLS estimators.

Theorem 3.7

Given the linear specification (1), suppose that [A1] and [A3] hold.

- (a) $\hat{\beta}_T \sim \mathcal{N}(\beta_o, \sigma_o^2(\mathbf{X}'\mathbf{X})^{-1})$.
- (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 $\hat{\beta}_T$ is a linear transformation of $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta_o, \sigma_o^2\mathbf{I}_T)$ and hence also a normal random vector. As for (b), writing $\hat{\mathbf{e}} = (\mathbf{I}_T - \mathbf{P})(\mathbf{y} - \mathbf{X}\beta_o)$, we have

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

where $\mathbf{y}^* = (\mathbf{y} - \mathbf{X}\beta_o)/\sigma_o \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_T)$ 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,

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

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

$$\mathbf{y}^{*'}(\mathbf{I}_T - \mathbf{P})\mathbf{y}^* = \sum_{i=1}^{T-k} \eta_i^2 \sim \chi^2(T - k),$$

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 [A1] and [A3] hold. Then the OLS estimators $\hat{\beta}_T$ and $\hat{\sigma}_o^2$ are the **best unbiased** estimators (BUE) for β_o and σ_o^2 , 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], $\mathbb{E}[\mathbf{s}(\beta_o, \sigma_o^2)] = \mathbf{0}$ and

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

Proof (cont'd):

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

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

is the **Cramér-Rao lower bound**.

- $\text{var}(\hat{\beta}_T)$ achieves this lower bound (the upper-left block) so that $\hat{\beta}_T$ is the best unbiased estimator for β_o . This conclusion is much stronger than the Gauss-Markov Theorem.
- 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: $\mathbf{R}\beta_o = \mathbf{r}$, 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 $\mathbf{R}\hat{\beta}_T$ and \mathbf{r} ; we reject the null if their difference is too “large.”
- Given [A1] and [A3], Theorem 3.7(a) states:

$$\hat{\beta}_T \sim \mathcal{N}(\beta_o, \sigma_o^2(\mathbf{X}'\mathbf{X})^{-1}),$$

so that

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

The comparison between $\mathbf{R}\hat{\beta}_T$ and \mathbf{r} is based on this distribution.

When $q = 1$, $\mathbf{R}\hat{\beta}_T$ and $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$ are scalars. Under the null hypothesis,

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

As σ_o^2 is unknown, the statistic above is not operational. By replacing σ_o with $\hat{\sigma}_T$, we obtain the so-called ***t* statistic**:

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

Theorem 3.9

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

Note: The normality condition [A3] is crucial for this *t* distribution result.

Proof: We write the statistic τ as

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

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 $\hat{\beta}_T$ and $\hat{\mathbf{e}}$ are jointly normally distributed with

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

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, the resulting t statistic is:

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

where $m^{ii} = \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$. 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 t statistic reads:

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

When \mathbf{R} is a $q \times k$ matrix with full row rank q ($q > 1$), we have under the null hypothesis: $[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1/2}(\mathbf{R}\hat{\beta}_T - \mathbf{r})/\sigma_o \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q)$. Hence,

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

The so-called F test statistic is:

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

It is easy to verify that

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

where both the numerator and denominator are χ^2 variables divided by their corresponding degrees of freedom.

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.

Notes:

- 1 When $q = 1$, $\varphi \sim F(1, T - k)$, and this distribution is the same as that of τ^2 .
- 2 Note that t distribution is symmetric about zero. Hence one may consider one- or two-sided t test. On the other hand, F distribution is non-negative and asymmetric, it is more typical to consider only one-sided F test.

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 $\beta_3 = 0, \dots$ and $\beta_k = 0$,

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

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 $\boldsymbol{\delta} \neq \mathbf{0}$.

Theorem 3.11

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

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

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

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

a non-central χ^2 distribution with the non-centrality parameter $\delta'\mathbf{D}^{-1}\delta$. 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 $\delta' \mathbf{D}^{-1} \delta$, where δ signifies the deviation from the null. When $\mathbf{R}\beta_o$ deviates farther from the hypothetical value \mathbf{r} (i.e., δ is “large”), the non-centrality parameter $\delta' \mathbf{D}^{-1} \delta$ 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**:

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

where λ is the $q \times 1$ vector of **Lagrangian multipliers**, we have

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

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

- The constrained OLS residuals are

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

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

- 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 $\text{ESS}_c = \ddot{\mathbf{e}}'\ddot{\mathbf{e}}$ and $\text{ESS}_u = \hat{\mathbf{e}}'\hat{\mathbf{e}}$ we have

$$\varphi = \frac{\ddot{\mathbf{e}}'\ddot{\mathbf{e}} - \hat{\mathbf{e}}'\hat{\mathbf{e}}}{q\hat{\sigma}_T^2} = \frac{(\text{ESS}_c - \text{ESS}_u)/q}{\text{ESS}_u/(T - k)},$$

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 $\text{ESS}_c = \ddot{\mathbf{e}}'\ddot{\mathbf{e}}$ and $\text{ESS}_u = \hat{\mathbf{e}}'\hat{\mathbf{e}}$ we have

$$\varphi = \frac{\ddot{\mathbf{e}}'\ddot{\mathbf{e}} - \hat{\mathbf{e}}'\hat{\mathbf{e}}}{q\hat{\sigma}_T^2} = \frac{(\text{ESS}_c - \text{ESS}_u)/q}{\text{ESS}_u/(T-k)},$$

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.

Confidence Regions

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

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

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

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

$$\begin{aligned} & \mathbb{P}\left\{ \left| (\hat{\beta}_{i,T} - \beta_{i,o}) / (\hat{\sigma}_T \sqrt{m^{ii}}) \right| \leq c_{\alpha/2} \right\} \\ &= \mathbb{P}\left\{ \hat{\beta}_{i,T} - c_{\alpha/2} \hat{\sigma}_T \sqrt{m^{ii}} \leq \beta_{i,o} \leq \hat{\beta}_{i,T} + c_{\alpha/2} \hat{\sigma}_T \sqrt{m^{ii}} \right\} \\ &= 1 - \alpha. \end{aligned}$$

- 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.

Example: Analysis of Suicide Rate

Part I: Estimation results with t

const	u_t	u_{t-1}	t	\bar{R}^2	Reg F
4.40 (4.57**)	2.48 (8.63**)			0.70	74.51**
4.41 (4.50**)	2.41 (5.04**)		0.01 (0.18)	0.69	36.11**
4.84 (5.09**)		2.40 (8.30**)		0.68	68.92**
4.84 (5.01**)		2.35 (4.74**)	0.01 (0.13)	0.67	33.37**

Note: The numbers in parentheses are t -ratios; ** and * stand for significance of a two-sided test at 1% and 5% levels, respectively.

Part II: Estimation results with t and g

const	u_t	u_{t-1}	g_t	g_{t-1}	t	\bar{R}^2	Reg F
4.20 (2.47*)	2.50 (7.25**)		0.02 (0.14)			0.69	36.09**
4.16 (2.39*)	2.43 (4.85**)		0.03 (0.18)		0.01 (0.21)	0.68	23.31**
5.56 (3.25**)		2.31 (6.70**)		-0.08 (-0.51)		0.67	33.77**
5.55 (3.18**)		2.29 (4.40**)		-0.08 (-0.49)	0.004 (0.06)	0.66	21.77**

F tests for the joint significance of the coefficients of g and t : 0.03 (Model 2) and 0.13 (Model 4), which are insignificant even at 10% level.

Part III: Estimation results with t and t^2

const	u_t	u_{t-1}	g_t	g_{t-1}	t	t^2	\bar{R}^2/F
11.60 (8.82**)					-0.45 (-2.50*)	0.02 (4.16**)	0.63 28.61**
8.05 (6.32**)	1.93 (4.62**)				-0.46 (-3.34**)	0.02 (3.73**)	0.78 39.10**
7.68 (4.42**)	1.96 (4.50**)		0.04 (0.32)		-0.46 (-3.27**)	0.02 (3.68**)	0.77 28.44**
8.37 (6.11**)		1.80 (3.87**)			-0.43 (-2.92**)	0.01 (3.27**)	0.75 32.97**
9.03 (4.84**)		1.74 (3.60**)		-0.07 (-0.53)	-0.43 (-2.90**)	0.01 (3.22**)	0.74 24.18**

F tests for the joint significance of the coefficients of g and t : 5.45* (Model 3) and 4.29* (Model 5), which are significant at 5% level.

Selected estimation results (with more precise estimates):

$$\hat{s}_t = 8.05 + 1.93 u_t - 0.460 t + 0.0154 t^2, \quad \bar{R}^2 = 0.78;$$

$$\hat{s}_t = 8.37 + 1.80 u_{t-1} - 0.430 t + 0.0147 t^2, \quad \bar{R}^2 = 0.75.$$

- For the second regression, the marginal effect of u_{t-1} on s_t is 1.80, approx. 410 persons. That is, the suicide rate would increase by 1.8 when there is one percent increase of last year's unemployment rate.
- The time effect of the second regression is $-0.43 + 0.0294 t$, which changes with t . At 2013, the time effect is 0.54, approx. 120 persons.
- Since 1994 (about 14.6 years after 1980), there has been a natural increase of the suicide rate in Taiwan; lowering unemployment rate would help cancel out the time effect to some extent.
- The predicted and actual suicide rates in 2013 are, respectively, 17.86 and 15.3; the difference between them is 2.56, approx. 590 persons.

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 denote the wage of the t th individual and x_t the working experience (in years). Consider the following specification:

$$y_t = \alpha_0 + \alpha_1 D_t + \beta_0 x_t + e_t,$$

where D_t is a **dummy variable** such that $D_t = 1$ if t is a male and $D_t = 0$ otherwise. This specification puts together two regressions:

- Regression for female: $D_t = 0$, and intercept is α_0 .
- Regression for male: $D_t = 1$, and intercept is $\alpha_0 + \alpha_1$.

These two regressions coincide if $\alpha_1 = 0$. Testing no wage discrimination against female amounts to testing the hypothesis of $\alpha_1 = 0$.

We may also consider the specification with a dummy variable and its interaction with a regressor:

$$y_t = \alpha_0 + \alpha_1 D_t + \beta_0 x_t + \beta_1 (x_t D_t) + e_t.$$

Then, the slopes of the regressions for female and male are, respectively, β_0 and $\beta_0 + \beta_1$. These two regressions coincide if $\alpha_1 = 0$ and $\beta_1 = 0$. In this case, testing no wage discrimination against female amounts to testing the joint hypothesis of $\alpha_1 = 0$ and $\beta_1 = 0$.

Example: Consider two dummy variables:

$D_{1,t} = 1$ if high school is t 's highest degree and $D_{1,t=0}$ otherwise;

$D_{2,t} = 1$ if college or graduate is t 's highest degree and $D_{2,t=0}$ otherwise.

The specification below in effect puts together 3 regressions:

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

where below-high-school regression has intercepts α_0 , high-school regression has intercept $\alpha_0 + \alpha_1$, college regression has intercept $\alpha_0 + \alpha_2$. Similar to the previous example, we may also consider a more general specification in which x interacts with D_1 and D_2 .

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

Example: Analysis of Suicide Rate

Let $D_t = 1$ for $t = T^* + 1, \dots, T$ and $D_t = 0$ otherwise, where T^* is the year of **structure change**. Consider the specification:

$$s_t = \alpha_0 + \delta D_t + \beta_0 u_{t-1} + \gamma u_{t-1} D_t + e_t.$$

The “before-change” regression has the intercept α_0 and slope β_0 , and the “after-change” regression has the intercept $\alpha_0 + \delta$ and slope $\beta_0 + \gamma$.

Testing a structure change at T^* amounts to testing $\delta = 0$ and $\gamma = 0$ (**Chow test**).

Alternatively, we can estimate the specification:

$$s_t = \alpha_0(1 - D_t) + \alpha_1 D_t + \beta_0 u_{t-1}(1 - D_t) + \beta_1 u_{t-1} D_t + e_t,$$

and test a structure change at T^* by checking if $\alpha_0 = \alpha_1$ and $\beta_0 = \beta_1$.

Part I: Estimation results with a known change: Without t

T^*	const	D_t	u_{t-1}	$u_{t-1}D_t$	$\bar{R}^2/\text{Reg } F$	Chow
1992	6.94 (2.85**)	-2.82 (-1.00)	1.42 (1.20)	1.17 (0.94)	0.67 22.56**	0.50
1993	6.07 (2.58*)	-1.39 (-0.49)	1.75 (1.51)	0.71 (0.57)	0.66 21.86**	0.17
1994	5.58 (2.48*)	-0.35 (-0.12)	1.94 (1.72)	0.39 (0.32)	0.66 21.80**	0.14
1995	5.36 (2.46*)	0.51 (0.17)	2.02 (1.83)	0.16 (0.13)	0.66 22.05**	0.26

Chow test is the F test of the coefficients of D_t and $u_{t-1}D_t$ being zero.

Part II: Estimation results with a known change: **With t**

T^*	const	D_t	u_{t-1}	t	tD_t	$\bar{R}^2/\text{Reg } F$
1992	11.11 (8.30**)	-10.41 (-5.09**)	1.08 (2.52*)	-0.54 (-4.00**)	0.92 (5.45**)	0.83 40.21**
1993	11.11 (8.47**)	-9.97 (-4.79**)	1.05 (2.47*)	-0.52 (-4.42**)	0.90 (5.63**)	0.83 41.09**
1994	11.05 (8.42**)	-9.46 (-4.41**)	0.97 (2.25*)	-0.48 (-4.48**)	0.85 (5.53**)	0.83 40.60**
1995	10.92 (8.19**)	-8.78 (-3.92**)	0.89 (1.98)	-0.43 (-4.32**)	0.79 (5.26**)	0.83 38.98**

F test of the coefficients of D_t and tD_t being zero: 15.28** ('92); 15.83** ('93); 15.52** ('94); 14.51** ('95)

We do not know T^* , the year of change, and hence tried estimating with different T^* :

$$T^* = 1992 : \hat{s}_t = 11.11 - 10.41 D_t + 1.08 u_{t-1} - 0.54 t + 0.92 tD_t;$$

$$T^* = 1993 : \hat{s}_t = 11.11 - 9.97 D_t + 1.05 u_{t-1} - 0.52 t + 0.90 tD_t;$$

$$T^* = 1994 : \hat{s}_t = 11.05 - 9.46 D_t + 0.97 u_{t-1} - 0.48 t + 0.85 tD_t;$$

$$T^* = 1995 : \hat{s}_t = 10.92 - 8.78 D_t + 0.89 u_{t-1} - 0.43 t + 0.79 tD_t.$$

- Consider the regression with $T^* = 1994$. The “before-change” and “after-change” slopes are -0.48 (a decrease over time) and 0.37 (an increase over time), respectively. The marginal effect of u_{t-1} on s_t is significant at 5% level.
- The regression line with $T^* = 1994$ predicts the suicide rate in 2013 as 17.8; the difference between the predicted and actual suicide rates is 2.5, approximately 570 persons.

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) $\mathbb{E}(\mathbf{y}) = \mathbf{X}\beta_o$: $\mathbb{E}(\mathbf{y})$ 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 $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, suppose, in addition to [A1] and [A2](i), $\text{var}(\mathbf{y}) = \boldsymbol{\Sigma}_o \neq \sigma_o^2 \mathbf{I}_T$, where $\boldsymbol{\Sigma}_o$ 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

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

- $\hat{\boldsymbol{\beta}}_T$ 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 $\text{var}(\mathbf{Gy}) = \mathbf{G}\Sigma_o\mathbf{G}'$.
- \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 $\mathbf{G}\Sigma_o\mathbf{G}' = \sigma_o^2\mathbf{I}_T$.

- Setting $\mathbf{G} = \Sigma_o^{-1/2}$, where $\Sigma_o^{-1/2} = \mathbf{C}\Lambda^{-1/2}\mathbf{C}'$ and \mathbf{C} orthogonally diagonalizes Σ_o : $\mathbf{C}'\Sigma_o\mathbf{C} = \Lambda$, we have $\Sigma_o^{-1/2}\Sigma_o\Sigma_o^{-1/2'} = \mathbf{I}_T$.

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

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

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

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

- The vector of GLS fitted values, $\hat{\mathbf{y}}_{\text{GLS}} = \mathbf{X}(\mathbf{X}' \boldsymbol{\Sigma}_o^{-1} \mathbf{X})^{-1} (\mathbf{X}' \boldsymbol{\Sigma}_o^{-1} \mathbf{y})$, is an **oblique** projection of \mathbf{y} onto $\text{span}(\mathbf{X})$, because $\mathbf{X}(\mathbf{X}' \boldsymbol{\Sigma}_o^{-1} \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\Sigma}_o^{-1}$ is idempotent but asymmetric. The GLS residual vector is $\hat{\mathbf{e}}_{\text{GLS}} = \mathbf{y} - \hat{\mathbf{y}}_{\text{GLS}}$.
- 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 $\text{var}(\mathbf{y}) = \mathbf{\Sigma}_o$ is positive definite. Then, $\hat{\beta}_{\text{GLS}}$ is the BLUE for β_o .

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

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

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

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

with the FOC: $\mathbf{X}'\mathbf{\Sigma}_o^{-1}(\mathbf{y} - \mathbf{X}\beta) = \mathbf{0}$. 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

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

- 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

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

- 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

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

where $\hat{\Sigma}_T$ is an estimator of Σ_o .

- Further difficulties in FGLS estimation:
 - The number of parameters in Σ_o is $T(T+1)/2$. Estimating Σ_o without some prior restrictions on Σ_o is practically infeasible.
 - Even when an estimator $\hat{\Sigma}_T$ is available under certain assumptions, $\hat{\beta}_{\text{FGLS}}$ is a complex function of the data \mathbf{y} and \mathbf{X} . As such, the finite-sample properties of the FGLS estimator are typically difficult to derive.

The Feasible GLS Estimator

- The **Feasible GLS** (FGLS) estimator is

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

where $\hat{\Sigma}_T$ is an estimator of Σ_o .

- Further difficulties in FGLS estimation:
 - The number of parameters in Σ_o is $T(T+1)/2$. Estimating Σ_o without some prior restrictions on Σ_o is practically infeasible.
 - Even when an estimator $\hat{\Sigma}_T$ is available under certain assumptions, $\hat{\beta}_{\text{FGLS}}$ is a complex function of the data \mathbf{y} and \mathbf{X} . As such, the finite-sample properties of the FGLS estimator are typically difficult to derive.

The Feasible GLS Estimator

- The **Feasible GLS** (FGLS) estimator is

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

where $\hat{\Sigma}_T$ is an estimator of Σ_o .

- Further difficulties in FGLS estimation:
 - The number of parameters in Σ_o is $T(T+1)/2$. Estimating Σ_o without some prior restrictions on Σ_o is practically infeasible.
 - Even when an estimator $\hat{\Sigma}_T$ is available under certain assumptions, $\hat{\beta}_{\text{FGLS}}$ is a complex function of the data \mathbf{y} and \mathbf{X} . As such, the finite-sample properties of the FGLS estimator are typically difficult to derive.

Some Remarks on the Feasible GLS Estimation

- In the classical literature, the feasible GLS estimation is based on some assumptions on Σ_o . These assumptions are needed to reduce the number of parameters and hence render the estimation of Σ_o feasible.
- The subsections below deal with the estimation under some special assumptions, which are imposed on either the diagonal or off-diagonal terms of Σ_o but not both.
- Although the assumptions on Σ_o are weaker than the condition of $\Sigma_o = \sigma_o^2 \mathbf{I}_T$, they are very **ad hoc** and can **not** be easily generalized.

Tests for Heteroskedasticity

A simple form of Σ_o is

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

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 $\hat{\sigma}_{T_1}^2$ and $\hat{\sigma}_{T_2}^2$.
- Under [A1] and [A3'], the F test is:

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

- 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:

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

- 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 Σ_ϵ 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 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:

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

- 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 Σ_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 $\hat{\sigma}_{T_1}^2$ and $\hat{\sigma}_{T_2}^2$ from separate regressions, an estimator of Σ_o is

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

The FGLS estimator is:

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

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 $\text{var}(y_t/x_{tj}) = c := \sigma_o^2$. 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 Σ_o take multiple values (so that there are more than 2 groups)?
- A general form of heteroskedasticity: $\sigma_t^2 = h(\alpha_0 + \mathbf{z}_t' \alpha_1)$, 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 Σ_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 Σ_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**: $\epsilon := \mathbf{y} - \mathbf{X}\beta_o$ so that $\text{var}(\mathbf{y}) = \text{var}(\epsilon) = \mathbb{E}(\epsilon\epsilon')$.
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:

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

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

so that $\text{corr}(\epsilon_t, \epsilon_{t-1}) = \psi_1$.

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

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

which depend only on i , but not on t .

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

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

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 $\Sigma_o^{-1/2}$:

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

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

The **Cochrane-Orcutt Transformation** is based on:

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

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

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

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 ψ_1, \dots, ψ_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 $\Sigma(\sigma^2, \psi)$ and $\mathbf{V}(\psi)$, so that $\Sigma_o = \Sigma(\sigma_o^2, \psi_1)$ and $\mathbf{V}_o = \mathbf{V}(\psi_1)$. Based on $\mathbf{V}(\psi)^{-1/2}$, we have

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

- **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 $\hat{\beta}_{\text{FGLS}}$ 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 $\hat{e}_{t,\text{FGLS}} = y_t - \mathbf{x}_t' \hat{\beta}_{\text{FGLS}}$.
- (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), $\mathbb{E}(y_t) = \mathbb{P}(y_t = 1) = \mathbf{x}_t' \beta_o$; this is known as the **linear probability model**.
- Problems with the linear probability model:
 - Under [A1] and [A2](i), there is heteroskedasticity:

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

and hence the OLS estimator is not the BLUE for β_o .

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

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

and hence the OLS estimator is not the BLUE for β_o .

- The OLS fitted values $\mathbf{x}'_t \hat{\beta}_T$ need not be bounded between 0 and 1.

- An FGLS estimator may be obtained using

$$\hat{\Sigma}_T^{-1/2} = \text{diag} \left[[\mathbf{x}'_1 \hat{\beta}_T (1 - \mathbf{x}'_1 \hat{\beta}_T)]^{-1/2}, \dots, [\mathbf{x}'_T \hat{\beta}_T (1 - \mathbf{x}'_T \hat{\beta}_T)]^{-1/2} \right].$$

- Problems with FGLS estimation:
 - $\hat{\Sigma}_T^{-1/2}$ can not be computed if $\mathbf{x}'_t \hat{\beta}_T$ is not bounded between 0 and 1.
 - Even when $\hat{\Sigma}_T^{-1/2}$ 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:

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

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., $\text{cov}(\mathbf{y}_i, \mathbf{y}_j) = \sigma_{ij} \mathbf{I}_T$.
- For this system, $\mathbf{\Sigma}_o = \mathbf{S}_o \otimes \mathbf{I}_T$ 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 $\mathbf{\Sigma}_o^{-1} = \mathbf{S}_o^{-1} \otimes \mathbf{I}_T$, then

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

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 $\mathbf{\Sigma}_o$. Then, the GLS estimator for each β_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

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

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

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

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

- The estimator $\hat{\mathbf{S}}_{TN}$ 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.