Classical Linear Regression Model

- Normality Assumption
- Hypothesis Testing Under Normality
- Maximum Likelihood Estimator
- Generalized Least Squares

Normality Assumption

- Assumption 5
  $\boldsymbol{\varepsilon}|\mathbf{X} \sim N(\mathbf{0},\sigma^2\mathbf{I}_n)$
- Implications of Normality Assumption
  – $(\mathbf{b}-\boldsymbol{\beta})|\mathbf{X} \sim N(\mathbf{0},\sigma^2(\mathbf{X}'\mathbf{X})^{-1})$
  – $(b_k-\beta_k)|\mathbf{X} \sim N(0, \sigma^2([\mathbf{X}'\mathbf{X})^{-1}]_{kk})$

$$z_k = \frac{b_k - \beta_k}{\sqrt{\sigma^2[(\mathbf{X}'\mathbf{X})^{-1}]_{kk}}} \sim N(0,1)$$

Hypothesis Testing under Normality

- Implications of Normality Assumption
  - Because $\varepsilon_i/\sigma \sim N(0,1)$,

$$\frac{(n-K)s^2}{\sigma^2} = \frac{\mathbf{e}'\mathbf{e}}{\sigma^2} = \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)'\mathbf{M}\left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right) \sim \chi^2(trace(\mathbf{M}))$$

where $\mathbf{M} = \mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $trace(\mathbf{M}) = n-K$.

Hypothesis Testing under Normality

- If σ^2 is not known, replace it with s^2. The standard error of the OLS estimator β_k is $SE(b_k) = \sqrt{s^2[(\mathbf{X}'\mathbf{X})^{-1}]_{kk}}$

- Suppose A.1-5 hold. Under H_0: $\beta_k = \bar{\beta}_k$ the t-statistic defined as

$$t_k = \frac{b_k - \bar{\beta}_k}{SE(b_k)} = \frac{b_k - \bar{\beta}_k}{\sqrt{s^2[(\mathbf{X}'\mathbf{X})^{-1}]_{kk}}} \sim t(\text{n-K}).$$

Hypothesis Testing under Normality

- Proof of

$$t_k = \frac{b_k - \overline{\beta}_k}{SE(b_k)} = \frac{b_k - \overline{\beta}_k}{\sqrt{s^2[(\mathbf{X}'\mathbf{X})^{-1}]_{kk}}} \sim t(\text{n-K})$$

$$\frac{\dfrac{b_k - \overline{\beta}_k}{\sqrt{\sigma^2[(\mathbf{X}'\mathbf{X})^{-1}]_{kk}}}}{\sqrt{\dfrac{(n-K)s^2/\sigma^2}{n-K}}} \sim \frac{N(0,1)}{\sqrt{\dfrac{\chi^2(n-K)}{n-K}}} = t(n-K)$$

Hypothesis Testing under Normality

- Testing Hypothesis about Individual Regression Coefficient, H_0: $\beta_k = \bar{\beta}_k$
  - If σ^2 is known, use $z_k \sim N(0,1)$.
  - If σ^2 is not known, use $t_k \sim t(n\text{-}K)$.
    Given a level of significance α,
    $\text{Prob}(-t_{\alpha/2}(n\text{-}K) < t < t_{\alpha/2}(n\text{-}K)) = 1\text{-}\alpha$

$$-t_{\alpha/2}(n-K) < \frac{b_k - \bar{\beta}_k}{SE(b_k)} < t_{\alpha/2}(n-K)$$

Hypothesis Testing under Normality

– Confidence Interval

$$b_k - SE(b_k)t_{\alpha/2}(n-K) < \overline{\beta}_k < b_k + SE(b_k)t_{\alpha/2}(n-K)$$

$[b_k\text{-}SE(b_k)\ t_{\alpha/2}(n\text{-}K),\ b_k\text{+}SE(b_k)\ t_{\alpha/2}(n\text{-}K)]$

– p-Value: $p = Prob(t>|t_k|)\times 2$

$Prob(-|t_k|<t<|t_k|) = 1-p$

since $Prob(t>|t_k|) = Prob(t<-|t_k|)$.

Accept H_0 if $p > \alpha$. Reject otherwise.

Hypothesis Testing under Normality

- Linear Hypotheses H_0: $\mathbf{R\beta} = \mathbf{q}$

$$\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1K} \\ r_{21} & r_{22} & \cdots & r_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ r_{J1} & r_{J2} & \cdots & r_{JK} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_J \end{bmatrix}$$

Hypothesis Testing under Normality

- Let $\mathbf{m} = \mathbf{Rb\text{-}q}$, where $\mathbf{b}$ is the unrestricted least squares estimator of $\boldsymbol{\beta}$.
  - $E(\mathbf{m}|\mathbf{X}) = E(\mathbf{Rb\text{-}q}|\mathbf{X}) = \mathbf{R}\boldsymbol{\beta}\mathbf{\text{-}q} = \mathbf{0}$
  - $Var(\mathbf{m}|\mathbf{X}) = Var(\mathbf{Rb\text{-}q}|\mathbf{X}) = \mathbf{R}Var(\mathbf{b}|\mathbf{X})\mathbf{R}' = \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$
- Wald Principle
  $W = \mathbf{m}'Var(\mathbf{m}|\mathbf{X})^{-1}\mathbf{m} = (\mathbf{Rb\text{-}q})'[\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb\text{-}q})$
  $\sim \chi^2(J)$, where J is the number of restrictions
- Define $F = (W/J)/(s^2/\sigma^2)$
  $= (\mathbf{Rb\text{-}q})'[s^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb\text{-}q})/J$

Hypothesis Testing under Normality

- Suppose A.1-5 holds. Under H_0: $\mathbf{R\beta} = \mathbf{q}$, where **R** is J×K with rank(**R**)=J, the **F-statistic** defined as

$$F = \frac{(\mathbf{Rb}-\mathbf{q})'[\mathbf{R}(\mathbf{X'X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb}-\mathbf{q})/J}{s^2}$$

$$= (\mathbf{Rb}-\mathbf{q})'\{\mathbf{R}[Est\,Var(\mathbf{b}\,|\,\mathbf{X})]\mathbf{R}'\}^{-1}(\mathbf{Rb}-\mathbf{q})/J$$

is distributed as F(J,n-K), the F distribution with J and n-K degrees of freedom.

Discussions

- Residuals $\mathbf{e} = \mathbf{y}\text{-}\mathbf{Xb} \sim N(0,\sigma^2\mathbf{M})$ if σ^2 is known and $\mathbf{M} = \mathbf{I}\text{-}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$
- If σ^2 is unknown and estimated by s^2,

$$\frac{e_i}{\sqrt{s^2[1-\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i]}} \sim t(n-K)$$

$$i = 1, 2, ..., n$$

Discussions

- Wald Principle vs. Likelihood Principle: By comparing the restricted (R) and unrestricted (UR) least squares, the **F-statistic** is shown

$$F = \frac{(SSR_R - SSR_{UR})/J}{SSR_{UR}/(n-K)} = \frac{(R_{UR}^2 - R_R^2)/J}{(1-R_{UR}^2)/(n-K)}$$

Discussions

- Testing $R^2 = 0$:
  Equivalently, H_0: $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, where

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

  J = K-1, and β_1 is the unrestricted constant term.
  The F-statistic follows F(K-1,n-K).

Discussions

- Testing $\beta_k = 0$:
  Equivalently, H_0: $\mathbf{R\beta} = \mathbf{q}$, where

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

  $F(1,n\text{-}K) = b_k[\text{Est Var}(\mathbf{b})]^{-1}{}_{kk} b_k$
  t-ratio: $t(n\text{-}K) = b_k / SE(b_k)$

Discussions

- t vs. F:
  - $t^2(n\text{-}K) = F(1,n\text{-}K)$ under H_0: $\mathbf{R}\boldsymbol{\beta}=\mathbf{q}$ when J=1
  - For J > 1, the F test is preferred to multiple t tests
- Durbin-Watson Test Statistic for Time Series Model:

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

  - The conditional distribution, and hence the critical values, of DW depends on X…

Maximum Likelihood

- Assumption 1, 2, 4, and 5 imply
  $\mathbf{y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}_n)$
- The conditional density or likelihood of **y** given **X** is

$$f(\mathbf{y} \mid \mathbf{X};\boldsymbol{\beta},\sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right]$$

Maximum Likelihood

- Likelihood Function
  $L(\boldsymbol{\beta},\sigma^2) = f(y|X;\boldsymbol{\beta},\sigma^2)$
- Log Likelihood Function

$$\log L(\boldsymbol{\beta},\sigma^2)$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\mathrm{SSR}(\boldsymbol{\beta})$$

Maximum Likelihood

- ML estimator of $(\boldsymbol{\beta},\sigma^2)$
  $= \text{argmax}_{(\boldsymbol{\beta},\gamma)} \log L(\boldsymbol{\beta},\gamma)$, where we set $\gamma = \sigma^2$

$$\frac{\partial \log L(\boldsymbol{\beta},\gamma)}{\partial \boldsymbol{\beta}} = -\frac{1}{2\gamma}\frac{\partial \mathrm{SSR}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0}$$

$$\frac{\partial \log L(\boldsymbol{\beta},\gamma)}{\partial \gamma} = -\frac{n}{2\gamma} + \frac{1}{2\gamma^2}\mathrm{SSR}(\boldsymbol{\beta}) = 0$$

Maximum Likelihood

- Suppose Assumptions 1-5 hold. Then the ML estimator of $\boldsymbol{\beta}$ is the OLS estimator **b** and ML estimator of γ or σ^2 is

$$\frac{SSR}{n} = \frac{\mathbf{e}'\mathbf{e}}{n} = \frac{n-K}{n} s^2$$

Maximum Likelihood

- Maximum Likelihood Principle
  - Let $\boldsymbol{\theta} = (\boldsymbol{\beta},\gamma)$
  - Score: $s(\boldsymbol{\theta}) = \partial \log L(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$
  - Information Matrix: $I(\boldsymbol{\theta}) = E(s(\boldsymbol{\theta})s(\boldsymbol{\theta})'|\mathbf{X})$
  - Information Matrix Equality:

$$I(\boldsymbol{\theta}) = -E\left[\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] = \begin{bmatrix} \frac{1}{\sigma^2}\mathbf{X}'\mathbf{X} & \mathbf{0} \\ \mathbf{0} & \frac{n}{2\sigma^4} \end{bmatrix}$$

Maximum Likelihood

- Maximum Likelihood Principle
  - Cramer-Rao Bound: $I(\boldsymbol{\theta})^{-1}$
    That is, for an unbiased estimator of $\boldsymbol{\theta}$ with a finite variance-covariance matrix:

$$\operatorname{Var}(\hat{\boldsymbol{\theta}}) \geq \mathrm{I}(\boldsymbol{\theta})^{-1} = \begin{bmatrix} \sigma^2(\mathbf{X'X})^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{2\sigma^4}{n} \end{bmatrix}$$

Maximum Likelihood

- Under Assumptions 1-5, the ML or OLS estimator **b** of **β** with variance $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ attains the Cramer-Rao bound.
- ML estimator of σ^2 is biased, so the Cramer-Rao bound does not apply.
- OLS estimator of σ^2, $s^2 = \mathbf{e}'\mathbf{e}/(n\text{-}K)$ with $E(s^2|\mathbf{X}) = \sigma^2$ and $Var(s^2|\mathbf{X}) = 2\sigma^4/(n\text{-}K)$, does not attain the Cramer-Rao bound $2\sigma^4/n$.

Discussions

- Concentrated Log Likelihood Function

$$\log L_c(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}, SSR(\boldsymbol{\beta})/n)$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(SSR(\boldsymbol{\beta})/n) - \frac{n}{2}$$

$$= -\frac{n}{2}\left[\log(2\pi/n) + 1\right] - \frac{n}{2}\log(SSR(\boldsymbol{\beta}))$$

- Therefore, $\text{argmax}_{\beta} \log L(\boldsymbol{\beta}) = \text{argmin}_{\beta}\, SSR(\boldsymbol{\beta})$

Discussions

- Hypothesis Testing H_0: $\mathbf{R\beta = q}$
  - Likelihood Ratio Test

$$\lambda = \frac{L_{UR}}{L_R} = \left(\frac{SSR_R}{SSR_{UR}}\right)^{n/2}$$

  - F Test as a Likelihood Ratio Test

$$F = \frac{(SSR_R - SSR_{UR})/J}{SSR_{UR}/(n-K)}$$

$$= \frac{n-K}{J}\left(\frac{SSR_R}{SSR_{UR}} - 1\right) = \frac{n-K}{J}\left(\lambda^{2/n} - 1\right)$$

Discussions

- Quasi-Maximum Likelihood
  - Without normality (Assumption 5), there is no guarantee that ML estimator of **β** is OLS or that the OLS estimator **b** achieves the Cramer-Rao bound.
  - However, **b** is a **quasi-** (or **pseudo-**) maximum likelihood estimator, an estimator that maximizes a misspecified (normal) likelihood function.

Generalized Least Squares

- Assumption 4 Revisited:
  $E(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}|\mathbf{X}) = Var(\boldsymbol{\varepsilon}|\mathbf{X}) = \sigma^2\mathbf{I}_n$
- Assumption 4 Relaxed (Assumption 4'):
  $E(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}|\mathbf{X}) = Var(\boldsymbol{\varepsilon}|\mathbf{X}) = \sigma^2\mathbf{V}(\mathbf{X})$, with nonsingular and known $\mathbf{V}(\mathbf{X})$.
  - OLS estimator of $\boldsymbol{\beta}$, $\mathbf{b}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, is not efficient although it is still unbiased.
  - t-test and F-test are no longer valid.

Generalized Least Squares

- Since $\mathbf{V}=\mathbf{V}(\mathbf{X})$ is known, $\mathbf{V}^{-1} = \mathbf{C}'\mathbf{C}$
- Let $\mathbf{y}^* = \mathbf{C}\mathbf{y}$, $\mathbf{X}^* = \mathbf{C}\mathbf{X}$, $\boldsymbol{\varepsilon}^* = \mathbf{C}\boldsymbol{\varepsilon}$
- $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \Rightarrow \mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$
  - Checking A.2: $E(\boldsymbol{\varepsilon}^*|\mathbf{X}^*) = E(\boldsymbol{\varepsilon}^*|\mathbf{X}) = \mathbf{0}$
  - Checking A.4: $E(\boldsymbol{\varepsilon}^*\boldsymbol{\varepsilon}^{*\prime}|\mathbf{X}^*) = E(\boldsymbol{\varepsilon}^*\boldsymbol{\varepsilon}^{*\prime}|\mathbf{X}) = \sigma^2\mathbf{C}\mathbf{V}\mathbf{C}' = \sigma^2\mathbf{I}_n$
- GLS: OLS for the transformed model $\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$

Generalized Least Squares

- $\mathbf{b}_{GLS} = (\mathbf{X}^{*\prime}\mathbf{X}^{*})^{-1}\mathbf{X}^{*\prime}\mathbf{y}^{*} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$
- $\text{Var}(\mathbf{b}_{GLS}|\mathbf{X}) = \sigma^2(\mathbf{X}^{*\prime}\mathbf{X}^{*})^{-1} = \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$
- If $\mathbf{V} = \mathbf{V}(\mathbf{X}) = \text{Var}(\boldsymbol{\varepsilon}|\mathbf{X})/\sigma^2$ is known,
  - $\mathbf{b}_{GLS} = (\mathbf{X}'[\text{Var}(\boldsymbol{\varepsilon}|\mathbf{X})]^{-1}\mathbf{X})^{-1}\mathbf{X}'[\text{Var}(\boldsymbol{\varepsilon}|\mathbf{X})]^{-1}\mathbf{y}$
  - $\text{Var}(\mathbf{b}_{GLS}|\mathbf{X}) = (\mathbf{X}'[\text{Var}(\boldsymbol{\varepsilon}|\mathbf{X})]^{-1}\mathbf{X})^{-1}$
  - GLS estimator $\mathbf{b}_{GLS}$ of $\boldsymbol{\beta}$ is BLUE.

Generalized Least Squares

- Under Assumption 1-3, $E(\mathbf{b}_{GLS}|\mathbf{X}) = \boldsymbol{\beta}$.
- Under Assumption 1-3, and 4', $Var(\mathbf{b}_{GLS}|\mathbf{X}) = \sigma^2 (\mathbf{X}'\mathbf{V}(\mathbf{X})^{-1}\mathbf{X})^{-1}$
- Under Assumption 1-3, and 4', the GLS estimator is efficient in that the conditional variance of any unbiased estimator that is linear in y is greater than or equal to $[Var(\mathbf{b}_{GLS}|\mathbf{X})]$.

Discussions

- Weighted Least Squares (WLS)
  - Assumption 4”: **V**(**X**) is a diagonal matrix, or $E(\varepsilon_i^2|\mathbf{X}) = Var(\varepsilon_i|\mathbf{X}) = \sigma^2 v_i(\mathbf{X})$

    Then $$y_i^* = \frac{y_i}{\sqrt{v_i(\mathbf{X})}}, \quad \mathbf{x}_i^* = \frac{\mathbf{x}_i}{\sqrt{v_i(\mathbf{X})}}$$

    $$(i = 1,2,\dots,n)$$
  - WLS is a special case of GLS.

Discussions

- If **V** = **V**(**X**) is not known, we can estimate its functional form from the sample. This approach is called the **Feasible** GLS. **V** becomes a random variable, then very little is known about the distribution and finite sample properties of the GLS estimator.

Example

- Cobb-Douglas Cost Function for Electricity Generation (Christensen and Greene [1976])
- Data: Greene's Table F4.3
  - Id = Observation, 123 + 35 holding companies
  - Year = 1970 for all observations
  - Cost = Total cost,
  - Q = Total output,
  - Pl = Wage rate,
  - Sl = Cost share for labor ,
  - Pk = Capital price index,
  - Sk = Cost share for capital,
  - Pf = Fuel price,
  - Sf = Cost share for fuel

Example

- Cobb-Douglas Cost Function for Electricity Generation (Christensen and Greene [1976])
  - $\ln(Cost) = \beta_1 + \beta_2\ln(PL) + \beta_3\ln(PK) + \beta_4\ln(PF) + \beta_5\ln(Q) + ½\beta_6\ln(Q)^2 + \beta_7\ln(Q)*\ln(PL) + \beta_8\ln(Q)*\ln(PK) + \beta_9\ln(Q)*\ln(PF) + \varepsilon$
  - Linear Homogeneity in Prices:
    - $\beta_2+\beta_3+\beta_4=1$, $\beta_7+\beta_8+\beta_9=0$
  - Imposing Restrictions:
    - $\ln(Cost/PF) = \beta_1 + \beta_2\ln(PL/PF) + \beta_3\ln(PK/PF) + \beta_5\ln(Q) + ½\beta_6\ln(Q)^2 + \beta_7\ln(Q)*\ln(PL/PF) + \beta_8\ln(Q)*\ln(PK/PF) + \varepsilon$