Econometrics

Introduction to Mathematical Statistics

Paul P. Momtaz

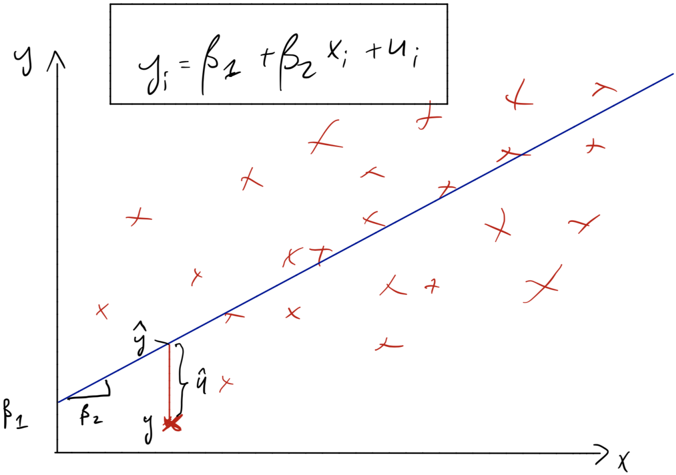
The Anderson School
UCLA

Outline

Outline

The basic OLS idea

Intuition



The basic OLS idea

Bivariate Model

Assumptions: $\mathbb{E}[u_i] = 0, Var[u_i] = \sigma^2, \mathbb{E}[u_i, u_j] = 0$ for $i \neq j$, x_i nonstochastic

From $y_i = \beta_1 + \beta_2 x_i + u_i$, want to minimize $RSS = \sum_{i=1}^{N}(y_i - \beta_1 - \beta_2 x_i)^2$

$$\text{FOCs: } \frac{\partial RSS}{\partial \beta_1} = -2\sum_{i=1}^{N} \tilde{u}_i = 0, \qquad \frac{\partial RSS}{\partial \beta_2} = -2\sum_{i=1}^{N} x_i \tilde{u}_i = 0$$

$$\text{From first FOC: } \hat{\beta}_1 = \sum_{i=1}^{N} \frac{y_i}{N} + \hat{\beta}_2 \sum_{i=1}^{N} \frac{x_i}{N} = \overline{y} - \hat{\beta}_2 \overline{x}$$

$$\text{Substituting into second FOC: } \hat{\beta}_2 = \frac{\sum_{i=1}^{N}(y_i - \overline{y}) \sum_{i=1}^{N}(x_i - \overline{x})}{\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

The basic OLS idea

Bivariate Model in Matrix Notation

$$y = X\beta + u$$

Assumptions: $\mathbb{E}[u] = 0 \qquad \mathbb{E}[uu'] = \sigma^2 I_N, \qquad X$ nonstochastic

$$\text{Minimize: } RSS = \tilde{u}'\tilde{u} = (y - X\hat{\beta})'(y - X\hat{\beta})$$

$$\text{FOC: } \frac{\partial RSS}{\partial \hat{\beta}} = -2x'y + 2X'X\hat{\beta} \Rightarrow \hat{\beta} = (X'X)^{-1}X'y$$

Outline

Gauss-Markov Assumptions

Given a linear mode, $\hat{\beta}_{OLS}$ is a consistent estimator of β if

(1)	No perfect collinearity in X	(drop redundant observations)
(2)	No selection bias	(a representative sample)
(3)	$\mathbb{E}[u] = 0$	(guaranteed if we include an intercept)
(4)	$\mathbb{E}[Xu] = 0$	(unbiased if $\mathbb{E}[u \| X] = 0$)

$$(3) + (4) \Rightarrow Var[\varepsilon] = \sigma^2 I_N.$$
$$(2) \Rightarrow \mathbb{E}[\varepsilon|x] = \mathbb{E}[\varepsilon] = 0 \text{ and } Var[\varepsilon|x] = Var[\varepsilon].$$

Assumption (4) is the focus of much attention, as we will discuss. But even without (4), we can always interpret $\hat{\beta}_{OLS}$ as summaries of correlations in the data, rather than as parameter estimates.

Gauss-Markov Assumptions

Unbiasedness and Variance

Unbiasedness:

b=$\beta + (x'x)^{-1}x'\varepsilon$

$$\mathbb{E}[b|x] = \beta + \mathbb{E}[(x'x)^{-1}x'\varepsilon|x] \tag{1}$$

$$\overset{\text{by } 2}{=} \beta + \mathbb{E}[(x'x)^{-1}x'|x]\mathbb{E}[\varepsilon|x] \tag{2}$$

$$\overset{\text{by } 1}{=} \beta \tag{3}$$

Variance:

$$Var[b|x] = \mathbb{E}[(b - \mathbb{E}[b])(b - \mathbb{E}[b])'|x] \tag{4}$$

$$= \mathbb{E}[(b - \beta)(b - \beta)'|x] \text{ from unbiasedness} \tag{5}$$

$$= \mathbb{E}[(x'x)^{-1}x'\varepsilon\varepsilon'x(x'x)^{-1}|x] \tag{6}$$

$$= (x'x)^{-1}x'(\sigma^2 I_N)x(x'x)^{-1} \text{ by 3 and 4} \tag{7}$$

$$= \sigma^2(x'x)^{-1} \tag{8}$$

Gauss-Markov Assumptions

DF Adjustment, Consistency, and Asymptotics

Significance Tests:

$$\text{Assumption 5: } \varepsilon \sim N(0, \sigma^2 I_N) \Rightarrow b \sim N(\beta, \sigma^2 (x'x)^{-1})$$

$$\Rightarrow \text{ unbiased estimator of } \sigma^2 : S^2 = \frac{1}{N-k} \sum_{i=1}^{N} \varepsilon_i^2$$

Consistency:

$$\text{Plim} = \beta \text{ as } N \to \infty \text{ (Asymptotics)}$$
$$\text{Assumption 6: } \tfrac{1}{N} \textstyle\sum_{i=1}^{N} x_i x_i' \text{ converges to finite nonsingular matrix } \Sigma_{xx}.$$
$$\text{Assumption 7: } \mathbb{E}[x_i \varepsilon_i] = 0$$

Distribution:

$$\text{As } N \to \infty, \sqrt{N}(b - \beta) \to N(0, \sigma^2 \Sigma_{xx}^{-1})$$
$$\text{Approximate in finite samples } b \sim N(\beta, S^2 (x'x)^{-1})$$

Outline

Goodness-Of-Fit

$$R^2 = \frac{\widehat{Var}(\hat{y}_i)}{\widehat{Var}(y_i)} = \frac{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - \overline{y_i})^2}{\frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y_i})^2} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \overline{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y}_i)^2} = \frac{ESS}{TSS}$$

$$\widehat{Var}(y_i) = \widehat{Var}(\hat{y}_i) + \widehat{Var}(\varepsilon_i)$$

$$R^2 = 1 - \frac{\widehat{Var}(\varepsilon_i)}{\widehat{Var}(y_i)} = 1 - \frac{\sum_{i=1}^{N}\varepsilon_i^2}{\sum_{i=1}^{N}(y_i - \overline{y}_i)^2} = 1 - \frac{RSS}{TSS}$$

$$\text{Adjusted } R^2 = 1 - \frac{\frac{1}{N-k}\sum_{i=1}^{N}\varepsilon_i^2}{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{y_i})^2}$$

Outline

Hypothesis Testing

t test

$$\text{Standard Error of } b_j = SE(b_j) = \sqrt{Var(b_j)} = S\sqrt{(x'x)^{-1}_{jj}}$$

$$H_0 : \beta_j = \beta_j^0 \qquad t_j = \frac{b_j - \beta_j^0}{SE(b_j)} \sim t_{N-k}$$

$$\text{Two-sided: } \alpha = 5\% \rightarrow 1.96 < t_j,$$

$$\text{One sided: } \alpha = 5\% \rightarrow 1.64 < t_j$$

$$\text{Confidence Intervals: } b_j - T_{N-k,\frac{\alpha}{2}} SE(b_j) < \beta_j < b_j + t_{N-k,\frac{\alpha}{2}} SE(b_j)$$

Hypothesis Testing

F test

$$\text{H}_0 : \beta_{k-J+1} = \cdots = \beta_k = 0$$

Compare RSS of full and restricted model (S_1 and S_0)

$$f = \frac{(S_0 - S_1/J)}{S_1/(N-k)} \sim F^J_{N-k} \qquad \text{or} \qquad f = \frac{(R_1^2 - R_0^2)/J}{(1-R_1^2)/(N-k)}$$

Outline

Frisch-Waugh(-Lovell) Theorem

Partitioned Regressions

$$y = x_1\beta_1 + x_2\beta_2 + \varepsilon \qquad or \qquad \begin{bmatrix} x_1 & x_2 \\ x_1 & x_2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} + \begin{bmatrix} \varepsilon \\ \varepsilon \end{bmatrix} = \begin{bmatrix} y \\ y \end{bmatrix}$$

Premultiply by x_1' and x_2' yo get rid of ε.

Premultiply lower equation by $(x_1'x_2)(x_2'x_2)^{-1}$ to get

$$(x_1'x_2)(x_2'x_2)^{-1}(x_2'x_1)\hat{\beta}_1 + (x_1'x_2)\hat{\beta}_2 = (x_1'x_2)(x_2'x_2)^{-1}x_2'y$$

Subtract this from upper equation to get rid of $\hat{\beta}_2$ and define $P_2 = x_2(x_2'x_2)^{-1}$.

$$\hat{\beta}_1^{OLS} = [x_1'(I - P_2)x_1]^{-1}x_1'(I - P_2)y$$

So, we purged x_1 off its correlation with x_2. The regression "controls" for x_2.

Frisch-Waugh(-Lovell) Theorem

Omitted Variable Bias

What did we do?

1. Regress x_1 on x_2.
2. Residual matrix $E_{1\cdot 2} = (I - P_2)x_1$.
3. $y \sim E_{1\cdot 2}$

Another way to see this:

$$\text{Assume } \mathbb{E}[x_2|x_1] = \pi x_1$$

$$y = \beta_1 x_1 + (\beta_2 x_2 + \varepsilon)$$
$$\mathbb{E}[y|x_1] = \beta_1 x_1 + \mathbb{E}[\beta_2 x_2|x_1] + \mathbb{E}[\varepsilon|x_1] = \beta_1 x_1 + \beta_2 \pi x_1$$

The multivariate regression decomposes the overall effect of changes in x_1 into a direct effect, β_1, and an indirect effect associated with changes in x_2, $\beta_2\pi$.

In contrast, a univariate regression would have given the overall effect $\hat{\beta}_1^* = \hat{\beta}_1 + \pi\hat{\beta}_1$.

Outline

Measurement Error

Also called Errors-in-the-variables (EIV), Attenuation Bias, and Asymptotic Bias

Measurement error in dependent variable not a problem for OLS consistency or validity, but problematic if in independent variable.

$$y_i = \beta x_i^* + \varepsilon_i$$

But we only observe $x_i = x_i^* + v_i, \qquad v_i \sim N(0, \sigma_v^2)$.

Errors flatten out OLS line due to scattered point estimates.

$$Plim(b) = plim\left(\frac{\sum x_i y_i / n}{\sum x_i^2 / n}\right) = \frac{Cov(x^* + v, \beta x^* + \varepsilon)}{Var(x)} \qquad \text{mean - zero variables}$$

$$= \beta \cdot \frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_v^2} = \beta \cdot \frac{1}{1 + \frac{\sigma_v^2}{\sigma_{x^*}^2}} \text{ Signal-to-noise-ratio}$$

Measurement Error

Alternative Intro

We observe $x_i = x_i^* + v_i$ where x_i^* is true value

$$\begin{aligned} y_i &= \alpha + \beta x_i^* + \varepsilon_i \\ &= \alpha + \beta(x_i - v_i) + \varepsilon_i \\ &= \alpha + \beta x_i + (\varepsilon_i - \beta v_i) \\ &= \alpha + \beta x_i + u_i \end{aligned}$$

Both x_i and u_i depend on v_i, so correlated $\Rightarrow$ OLS estimation downward biased.
Measurement error in y_i increase variance in error term but does not cause endogeneity.

Omitted Variables

$$y = x_1\beta_1^* + v \qquad \text{and} \qquad y = x_1\beta_1 + x_2\beta_2 + \varepsilon$$

$$\begin{aligned}\beta_2 &= (x_1'x_1)^{-1}x_1'(y - x_2\beta_2)\\ &= (x_1'x_1)^{-1}x_1'y - (x_1'x_1)^{-1}x_1'x_2\beta_2\\ &= \beta_1^* - (x_1'x_1)^{-1}x_1'x_2\beta_2\end{aligned}$$

$\Rightarrow \beta_1^* = \beta_1 + G\beta_2 \Rightarrow$ Asymptotic bias (unless $G = 0$ or $\beta_2 = 0$)

- No OVB if $corr(x_1, x_2) = 0$.
- OVB does not cause $corr(x_1, v) \neq 0$ as $\mathbb{E}[v|x_1] = 0$ in short regression.
- But if OVB x_1 corr to composite error term $u = x_2\beta_2 + \varepsilon$.
- If you do not control for x_2, you assume that x_2 varies with changes in x_1.

Simultaneity

Given two structural equations $y_i = \beta_1 x_i + \gamma_1 w_i + u_i$ and $w_i = \beta_2 x_i + \gamma_2 y_i + v_i$.

$$w_i = \frac{\beta_2 + \gamma_2 \beta_1}{1 - \gamma_1 \gamma_2} x_i + \frac{1}{1 - \gamma_1 \gamma_2} v_i + \frac{\gamma_2}{1 - \gamma_1 \gamma_2} u_i$$

Assuming

$$Corr(x_i, u_i) = Corr(v_i, u_i) = 0 \Rightarrow \mathbb{E}[w_i u_i] = \frac{\gamma_2}{1 - \gamma_1 \gamma_2} \mathbb{E}[u_i u_i] \neq 0 \Rightarrow \mathbb{E}[u|w] \neq 0$$

Generalized Least Squares (GLS)

$y_t = x_t'\beta + \varepsilon_t$ where $\varepsilon_t = u_t + u_{t-1} \Rightarrow$ error structure not homoskedastic.

$$\mathbb{E}[\varepsilon\varepsilon'] = \Sigma = \sigma^2 D = \sigma^2 \begin{bmatrix} 2 & 1 & 0 & \dots & 0 \\ 1 & 2 & 1 & \dots & \vdots \\ 0 & 1 & 1 & \dots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & \dots & 2 \end{bmatrix}$$

Outline

Generalized Least Squares (GLS)

Cholesky Decomposition

To demonstrate properties of GLS transform to CLRM and show that BLUE.

1. Transform

$$\text{Cholesky Factorization: } D = (L')^{-1}L^{-1} \Rightarrow L'DL = I_t, \text{ given } D \text{ symm, pos - def}$$

$$\text{So: } L'y = L'x\beta + L'\varepsilon \Rightarrow y^* + x^*\beta + \varepsilon^* \Rightarrow \text{CLRM}$$

2. Show Properties

$$\text{By Gauss - Markov, } \hat{\beta}^*_{OLS} = (x^{*'}x^*)^{-1}x^{*'}y^* \text{ is BLUE, so}$$

$$\begin{aligned}
\hat{\beta}_{GLS} &= (x'LL'x)^{-1}x'LL'y \\
&= (x'D^{-1}x)^{-1}x'D^{-1}y \\
&= (x'(\sigma^2 D)^{-1}x)^{-1}x'(\sigma^2 D)^{-1}y \\
&= (x'\Sigma^{-1}x)^{-1}x'\Sigma^{-1}y \\
&\text{Note That} \\
&= (x'\Sigma^{-1}x)^{-1}x'\Sigma^{-1}(x\beta + \varepsilon) \\
&= \beta + (x'\Sigma^{-1}x)x'\Sigma^{-1}\varepsilon
\end{aligned}$$

Generalized Least Squares (GLS)

So:

$$Var[\hat{\beta}_{GLS}] = \mathbb{E}[(\hat{\beta}_{GLS} - \beta)(\hat{\beta}_{GLS} - \beta)'] = (x'\Sigma^{-1}x)^{-1}$$

Intuition: GLS weights observations by the inverse square root of the variance of the error term since $L' = D^{\frac{-1}{2}} \propto \Sigma^{-\frac{1}{2}}$

Outline

Feasible Generalized Least Squares (Feasible GLS)

Often Σ is unknown but $\hat{\Sigma}$ exists.
Sufficient conditions for consistency and asymptotic normality of FGLS:

1. $plim \frac{1}{N} x' \hat{\Sigma}^{-1} x \to Q_{\Sigma} > 0.$
2. $plim \frac{1}{N} x' \hat{\Sigma}^{-1} \varepsilon \to 0$

why?

$$\hat{\beta}_{FGLS} = (x'\hat{\Sigma}^{-1}x)^{-1} x'\hat{\Sigma}^{-1} y = \beta + (x'\hat{\Sigma}^{-1}x)^{-1} x'\hat{\Sigma}^{-1} \varepsilon$$
$$Plim(\hat{\beta}_{FGLS} - \beta) = Plim(x'\hat{\Sigma}^{-1}x)^{-1} plim(x'\hat{\Sigma}^{-1}\varepsilon) \to (Q_{\Sigma})^{-1} \cdot 0 = 0$$

If heteroskedasticity of unknown form, i.e. $\mathbb{E}[\varepsilon_t^2] = \sigma_t^2$ and $\Sigma = diag(\sigma_t^2)$,
We can estimate $\hat{\sigma}_t^2 = \hat{\varepsilon}_t^2$ from a first - stage regression. But $\hat{\Sigma} \nrightarrow \Sigma$.

Note that $\frac{1}{N} x' \hat{\Sigma}^{-1} x = \frac{1}{N} \Sigma_t \frac{1}{\hat{\sigma}_t^2} x_t x_t'$ where $\hat{\Sigma}$ is $k \times k$ fixed since $N \to \infty$.
So: $plim \frac{1}{N} x' \hat{\Sigma}^{-1} x \to Q_{\Sigma} > 0$ and $plim \frac{1}{N} x' \hat{\Sigma}^{-1} \hat{\varepsilon} \to 0.$

Outline

MLE Introduction

Intuition: Find θ that maximizes the likelihood of observations.

$$\max_\theta L(\theta; y) = \max_\theta \prod_{i=1}^{N} f(y_i; \theta) \text{ or } \max_\theta \log L(\theta; y) = \max_\theta \sum_{i=1}^{N} \log f(y_i; \theta)$$

Efficient score: $S(\theta) = \frac{\partial \log L}{\partial \theta}$. MLE is solution to $S(\hat{\theta}) = 0$.

Example:

$$y_i = \alpha + \beta x_i + \varepsilon_i, \qquad \varepsilon_i \sim IN(0, \sigma^2), \qquad y_i \sim IN(\alpha + \beta x_i, \sigma^2)$$

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2\right]$$

$$\log L(\alpha, \beta, \sigma^2; data) = \sum_{i=1}^{N}\left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2\right]$$

MLE Introduction

FOC's:

$$[\alpha] \sum_{i=1}^{N} (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0$$

$$[\beta] \sum_{i=1}^{N} x_i (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0$$

$$[\sigma^2] \underbrace{N\hat{\sigma}^2}_{\text{not } (N-2)!} = \sum_{i=1}^{N} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

Plug in $\hat{\sigma}^2$ in $\log L(\alpha, \beta, \sigma^2; data)$ to see maximized $\log L$ is

$$\log L(\hat{\theta}) = \text{constant} - \frac{N}{2} \log \hat{\sigma}^2 = \text{constant} - \frac{N}{2} \log \left(\frac{RSS}{N} \right)$$

$$\text{Individual Score: } S_i(\theta) = \frac{\partial \log L_i(\theta)}{\partial \theta}$$

Outline

MLE Properties

$$\text{Average information matrix : } \bar{J}_N(\theta) = -\mathbb{E}\left[\frac{1}{N}\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}\right]$$

$$\text{Limiting information matrix: } J(\theta) \equiv \lim_{N\to\infty} \bar{J}_N(\theta)$$

$MLE, \hat{\theta}$, is

- Consistent
- Asymptotically normal since $\sqrt{N}(\hat{\theta}-\theta) \to N(0,[J(\theta)]^{-1})$.
- Asymptotically efficient (Cramer - Rao lower bound).
- Invariant (continuous function theorem: MLE of $g(\theta)$ for any $g(\hat{\theta})$).
- $Var[S_i(\theta)] = \mathbb{E}[S_i(\theta)S_i(\theta)'] = J_i(\theta)$.

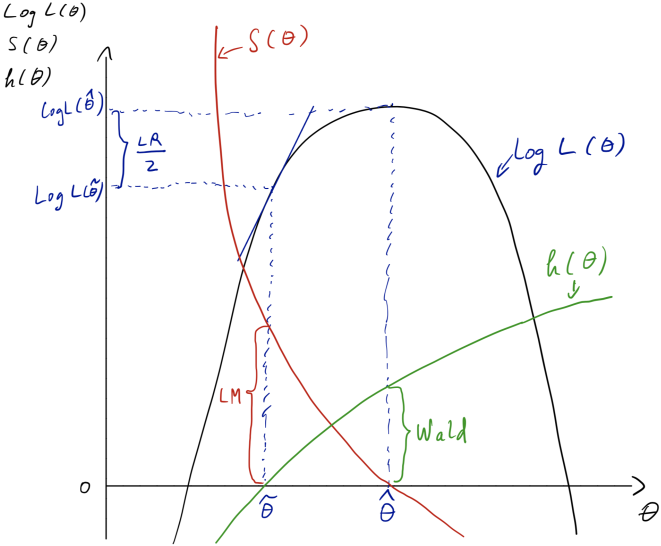
Can be estimated outer product of gradients (G)

$$\hat{V}_G = \left[\frac{1}{N}\sum_{i=1}^{N} S_i(\hat{\theta})S_i(\hat{\theta})'\right]^{-1} \quad \text{or Hessian} \quad \hat{V}_H = \left[-\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2 \log L_i(\theta)}{\partial\theta\partial\theta'}\bigg|_{\hat{\theta}}\right]^{-1}$$

Outline

The Three Classical Tests of MLE

Idea: Set of restrictions to be tested $H_0 : h(\theta) = 0$
$\hat{\theta}$ and $\tilde{\theta}$ are unrestricted and restricted MLE, respectively.



Wald Test

- Approximate $Var[h(\hat{\theta})] = G(\theta)' Var[\hat{\theta}] G(\theta)$ where $G(\theta) = \frac{\partial h(\theta)'}{\partial \theta}$.
- Test Statistic
  $\xi_w = N h(\hat{\theta})' [G(\theta)' [J(\theta)]^{-1} G(\theta)]^{-1} h(\hat{\theta}), \xi_w \sim_a \chi^2(d)$.
- Estimate $[J(\theta)]^{-1}$ by $\hat{V}_H$ or $\hat{V}_G$ evaluated at $\hat{\theta}$, evaluate $G(\theta)$ at $\hat{\theta}$.
- Shortcoming: Wald test not invariant to how restrictions (linear/non-linear) are formulated.
- Overall question: How close is $h(\hat{\theta})$ to zero since $h(\tilde{\theta}) = 0$?

Likelihood Ratio Test

- How close are $L(\hat{\theta})$ and $L(\tilde{\theta})$?
- $\lambda = L(\tilde{\theta})/L(\hat{\theta})$.
- $\xi_{LR} = -2\left\{\log L(\tilde{\theta}) - \log L(\hat{\theta})\right\} = 2\left\{\log L(\hat{\theta}) - \log L(\tilde{\theta})\right\}$ under the null, $\xi_{LR} \sim_a \chi^2(d)$

Lagrange Multiplier or Score Test

- How close is $S(\tilde{\theta})$ to zero given $S(\hat{\theta}) = 0$?
- $\xi_{LM} = N^{-1} S(\tilde{\theta}) [J(\tilde{\theta})]^{-1} S(\tilde{\theta}) \sim_a \chi^2(d)$

Lagrange Multiplier or Score Test

Alternative Computation

Alternative for ξ_{LM}

- Supplementary regression $\tilde{u} \sim x$
- $RSS = N\hat{\sigma}^2$.
- $TSS = N\tilde{\sigma}^2$

$$\Rightarrow R_u^2 = 1 - \frac{\hat{\sigma}^2}{\tilde{\sigma}^2}, \qquad \xi_{LM} = NR_u^2$$

The Three Classical Tests of MLE

Application

$$y = x\beta + u \qquad N(0, \sigma^2 I)$$

$$\text{Unrestricted Estimation:} \quad \log L(\hat{\beta}, \hat{\sigma}^2) = \text{constant} - \frac{N}{2} \log \hat{\sigma}^2$$

$$\text{Restricted Estimation:} \quad \log L(\tilde{\beta}, \tilde{\sigma}^2) = \text{constant} - \frac{N}{2} \log \tilde{\sigma}^2$$

$$\xi_{LR} = N \log \left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2} \right), \qquad \xi_N = N \left(\frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2} \right), \qquad \xi_{LM} = N \left(\frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\tilde{\sigma}^2} \right)$$

Note that test statistics are functions of each other

$$\xi_{LR} = N \log \left(1 + \frac{\xi_w}{N} \right), \qquad \xi_{LM} = \frac{\xi_w}{1 + \frac{\xi_w}{N}}$$

Linear Case: $\xi_w \geq \xi_{LR} \geq \xi_{LM}$