Applied Topics for PhD

Monte Carlo & Bootstrap

Michael E. Kummer
Theoretical Slide Set, based on Stephen Kastoryano

OTIM, Nova School of Business and Economics

Table of Contents

SIMULATION TO UNDERSTAND ESTIMATORS BETTER

- Generally we always had $(y, \mathbf{x})$ are and i.i.d. draw from some F
- Then all "statistics", including estimators etc.

$$T_n = T_n((\boldsymbol{y}_1, \boldsymbol{x}_1), (\boldsymbol{y}_2, \boldsymbol{x}_2), ..., (\boldsymbol{y}_n, \boldsymbol{x}_n), \theta))$$

SIMULATION TO UNDERSTAND ESTIMATORS BETTER

- Generally we always had $(y, \mathbf{x})$ are and i.i.d. draw from some F
- Then all "statistics", including estimators etc.

$$T_n = T_n((\boldsymbol{y}_1, \boldsymbol{x}_1), (\boldsymbol{y}_2, \boldsymbol{x}_2), ..., (\boldsymbol{y}_n, \boldsymbol{x}_n), \boldsymbol{\theta}))$$

- e.g. t-ratio $(\hat{\theta} - \theta)/s(\hat{\theta})$
- or some estimator $\hat{\theta}$ (or $\hat{\beta}$)

SIMULATION TO UNDERSTAND ESTIMATORS BETTER

- Generally we always had $(y, \mathbf{x})$ are and i.i.d. draw from some F
- Then all "statistics", including estimators etc.

$$T_n = T_n((\boldsymbol{y}_1, \boldsymbol{x}_1), (\boldsymbol{y}_2, \boldsymbol{x}_2), ..., (\boldsymbol{y}_n, \boldsymbol{x}_n), \boldsymbol{\theta}))$$

- e.g. t-ratio $(\hat{\theta} - \theta)/s(\hat{\theta})$
- or some estimator $\hat{\theta}$ (or $\hat{\beta}$)

- depends on "the data" and F, it's distribution.

DEFINITION 2

- We know T_n (estimators, test statistics) are just another random variable
- With it's own expectation and distribution.
- So we model the c.d.f. as:

$$G_n(u, F) = Pr(T_n < u \,|F)$$

- but while the asymptotic behavior of T_n may be known, the one of G_n generally is not

LETS GO TO THE CASINO

- We can learn about the statistic, by simulation.
- Here's how:
  - We choose some F and then simulate data.
  - This nails down a true θ
  - Now we pass that data to our estimator...
  - ...and see if it "get's it right"

LETS GO TO THE CASINO

- We can learn about the statistic, by simulation.
- Here's how:
  - We choose some F and then simulate data.
  - This nails down a true θ
  - Now we pass that data to our estimator...
  - ...and see if it "get's it right"

- it'e like simulating it in a sandbox

HERE IS WHAT TO DO:

- Define some DGP (i.e. F) and draw n observations.
- we get n observations $(y_i^*, \mathbf{x_i}^*)$ $i = 1, ..., n$
- compute $T_n = T_n((\boldsymbol{y}_1, \boldsymbol{x}_1), (\boldsymbol{y}_2, \boldsymbol{x}_2), ..., (\boldsymbol{y_n}, \boldsymbol{x_n}), \theta))$ on the pseudo data.

- This generates **one** random draw of the unknown $G_n(u, F)$

NOTATION - B^{TH} EXPERIMENT

- repeat over and over.
- B times. (like 1000 or 5000 times)
- see how the estimator performs by varying n and F

NOTATION - B^{TH} EXPERIMENT

- repeat over and over.
- B times. (like 1000 or 5000 times)
- see how the estimator performs by varying n and F

- Notation:
  - T_{nb} ... is the outcome of the b^{th} experiment $(b = 1, \ldots, B)$
  - So Repeating B times gives:

  - a "random sample of size B"
  - coming from $G_n(u, F) = Pr(T_{nb} < u) = Pr(T_n < u \,|F)$

RESULTING TEST-STATISTICS

- e.g. Bias

$$\widehat{Bias}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B} T_{nb} = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_b - \theta$$

- MSE:

$$\widehat{MSE}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B} (T_{nb})^2 = \frac{1}{B}\sum_{b=1}^{B} (\hat{\theta}_b - \theta)^2$$

- or even the type I error of a two-sided t-test with $T_n = (\hat{\theta} - \theta)/s(\hat{\theta})$

$$\hat{P} = \frac{1}{B}\sum_{b=1}^{B} 1(T_{nb}) \leq 1.96)$$

- ...with 1() the indicator function, takes value 1 if condition true
- in other words simply counting.

COMPUTER TASK

1. Simulate some data for an interesting DGP (with a known Problem), absent other ideas, revisit the omitted variables bias problem
2. Now run this not only once but $B = 1000$ times.
3. Each time run the estimation procedure and store the result.
   - Note: You can either save 1000 datasets, or keep only the estimator.
4. Then compute the EDF of that estimation.
5. Compare with the CDF you know.

6. You can vary n and see whether the estimation gets better.

Table of Contents

INTRODUCTION

- Previous lectures were using well defined and understood estimator.
- Asymptotic behavior has been researched by earlier research.
- BUT: On high level research (as PhD-level applied research) the “straight forward”methods are not available

INTRODUCTION

- Previous lectures were using well defined and understood estimator.
- Asymptotic behavior has been researched by earlier research.
- BUT: On high level research (as PhD-level applied research) the “straight forward”methods are not available

- A useful non-parametric method is “the Bootstrap”
  - works in many settings.
  - frequently also for data or residuals from unknown distributions.

DEFINITION

- Start from "some test-statistic."

$$T_n = T_n((\boldsymbol{y}_1, \boldsymbol{x}_1), (\boldsymbol{y}_2, \boldsymbol{x}_2), ..., (\boldsymbol{y}_n, \boldsymbol{x}_n), \boldsymbol{F}))$$

DEFINITION

- Start from "some test-statistic."

$$T_n = T_n((\boldsymbol{y}_1, \boldsymbol{x}_1), (\boldsymbol{y}_2, \boldsymbol{x}_2), ..., (\boldsymbol{y}_n, \boldsymbol{x}_n), \boldsymbol{F}))$$

- e.g. t-ratio $(\hat{\theta} - \theta)/s(\hat{\theta})$
- or some estimator $\hat{\theta}$ (or $\hat{\beta}$)

DEFINITION

- Start from "some test-statistic."

$$T_n = T_n((\boldsymbol{y}_1, \boldsymbol{x}_1), (\boldsymbol{y}_2, \boldsymbol{x}_2), ..., (\boldsymbol{y}_n, \boldsymbol{x}_n), \boldsymbol{F}))$$

- e.g. t-ratio $(\hat{\theta} - \theta)/s(\hat{\theta})$
- or some estimator $\hat{\theta}$ (or $\hat{\beta}$)

- depends on "the data" and F, it's distribution.

DEFINITION 2

- We know T_n (estimators, test statistics) are just another random variable
- With it's own expectation and distribution.
- So we model the c.d.f. as:

$$G_n(u, F) = Pr(T_n < u \,|F)$$

DEFINITION 3

- But we do not know F
- So what can we do?

DEFINITION 3

- But we do not know F
- So what can we do?
- We do not know F, but we have tons of data, that were generated by F
- Let's pull ourselves out of the swamp on our own Bootstraps

DEFINITION 3

- But we do not know F
- So what can we do?
- We do not know F, but we have tons of data, that were generated by F

- Let's pull ourselves out of the swamp on our own Bootstraps
  - by learning about the distribution from our data.

$$G_n^*(u) = G_n(u, F_n)$$

EMPIRICAL DISTRIBUTION FUNCTION 1

- Recall

$$\boldsymbol{F}(\boldsymbol{y}, \mathbf{x}) = Pr(y_i < y, \mathbf{x}_i < \mathbf{x}) = \mathrm{E}(1(y_i < y)1(\mathbf{x}_i < \mathbf{x}))$$

EMPIRICAL DISTRIBUTION FUNCTION 1

- Recall

$$\boldsymbol{F}(\boldsymbol{y},\mathbf{x}) = Pr(y_i < y, \mathbf{x}_i < \mathbf{x}) = \mathrm{E}(1(y_i < y)1(\mathbf{x}_i < \mathbf{x}))$$

- The corresponding method of moments estimator simply plugs in the empirical counterpart:
- (...as always.)
- Hence,

$$\boldsymbol{F_n}(\boldsymbol{y},\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} 1(y_i < y)1(\mathbf{x}_i < \mathbf{x})$$

- This is called the "Empirical Distribution Function" (EDF)

EMPIRICAL DISTRIBUTION FUNCTION - ASYMPTOTICS

- EDF consistently estimates CDF. Proof is based on
  - the insight that for any $(y, \mathbf{x})$, $1(y_i < y)1(\mathbf{x}_i < \mathbf{x})$ is an i.i.d. RV drawn from $F(y, \mathbf{x})$
  - WLLN
  - CLT

EMPIRICAL DISTRIBUTION FUNCTION - ASYMPTOTICS

- EDF consistently estimates CDF. Proof is based on
  - the insight that for any $(y, \mathbf{x})$, $1(y_i < y)1(\mathbf{x}_i < \mathbf{x})$ is an i.i.d. RV drawn from $F(y, \mathbf{x})$
  - WLLN
  - CLT

- Hence,

$$\sqrt{n}(\boldsymbol{F_n}(\boldsymbol{y}, \mathbf{x}) - \boldsymbol{F}(\boldsymbol{y}, \mathbf{x})) \xrightarrow{d} \boldsymbol{N}(0, F(y, \mathbf{x})(1 - F(y, \mathbf{x}))$$

EMPIRICAL DISTRIBUTION FUNCTION - ASYMPTOTICS

- And the EDF is a valid **discrete** prob. distribution.
- Probability mass is $1/n$ for each data-point $(y_i, \mathbf{x_i})$
- i.o.w. think of a random pair $(y_i^*, \mathbf{x_i}^*)$ as drawn from F_n

$$Pr(y_i^* < y, \mathbf{x}_i^* < \mathbf{x}) = \boldsymbol{F_n(y, x)}$$

EMPIRICAL DISTRIBUTION FUNCTION - ASYMPTOTICS

- Sample moments of $(y_i^*, \mathbf{x}_i^*)$

$$\begin{aligned}
\mathbf{E}h(\mathbf{y_i^*}, \mathbf{x_i^*}) &= \int (h(y_i, \mathbf{x}_i) \boldsymbol{F_n}(\boldsymbol{y}, \mathbf{x})) \\
&= \sum_{i=1}^{n} h(y_i, \mathbf{x}_i) Pr(y_i^* = y_i) 1(\mathbf{x}_i^* = \mathbf{x}_i) \\
&= \frac{1}{n} \sum_{i=1}^{n} h(y_i, \mathbf{x}_i)
\end{aligned}$$

EMPIRICAL DISTRIBUTION FUNCTION - ASYMPTOTICS

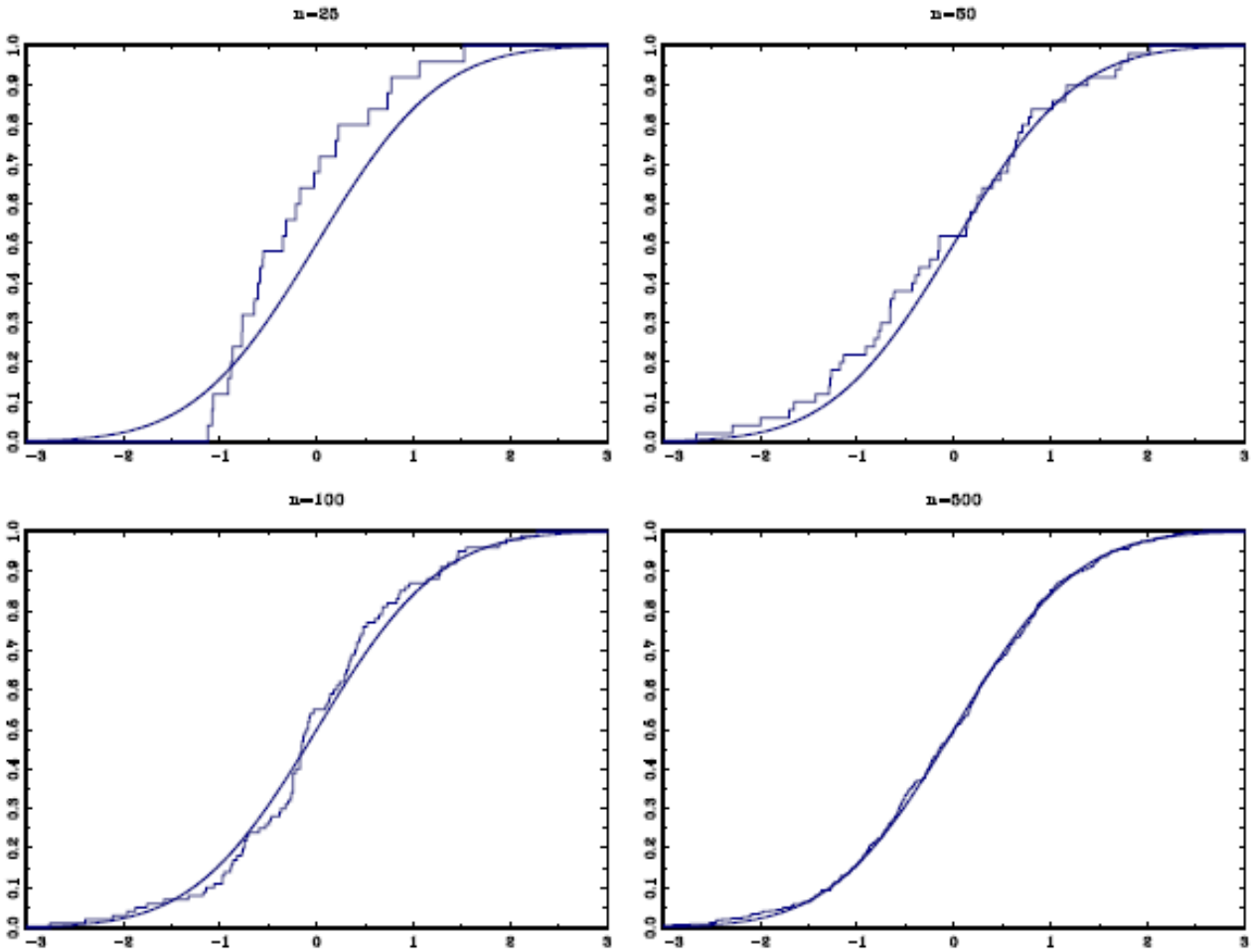

Figure 10.1: Empirical Distribution Functions

EMPIRICAL DISTRIBUTION FUNCTION - ASYMPTOTICS

- Sample moments of $(y_i^*, \mathbf{x}_i^*)$

$$\begin{aligned}
\mathbf{E}h(\mathbf{y_i^*}, \mathbf{x_i^*}) &= \int (h(y_i, \mathbf{x}_i) \boldsymbol{F_n}(\boldsymbol{y}, \mathbf{x})) \\
&= \sum_{i=1}^{n} h(y_i, \mathbf{x}_i) Pr(y_i^* = y_i) 1(\mathbf{x}_i^* = \mathbf{x}_i) \\
&= \frac{1}{n} \sum_{i=1}^{n} h(y_i, \mathbf{x}_i)
\end{aligned}$$

Table of Contents

NONPARAMETRIC BOOTSTRAP

- Use EDF as the estimate $\boldsymbol{F_n}$ of $\boldsymbol{F}$
- Use Monte-Carlo simulations to approximate $\boldsymbol{G_n^*}$

- Note:
  - The sample size in the simulation should be the same as the sample size
  - $(y_i^*, \mathbf{x}_i^*)$ are drawn from the data, i.e.
  - You randomly sample (WITH REPLACEMENT!) n observations $(y, \mathbf{x})$

NONPARAMETRIC BOOTSTRAP

- Use EDF as the estimate $\boldsymbol{F_n}$ of $\boldsymbol{F}$
- Use Monte-Carlo simulations to approximate $\boldsymbol{G_n^*}$
- Note:
  - The sample size in the simulation should be the same as the sample size
  - $(y_i^*, \mathbf{x}_i^*)$ are drawn from the data, i.e.
  - You randomly sample (WITH REPLACEMENT!) n observations $(y, \mathbf{x})$
- And you get your **first instance** of T_n^*:

$$T_n^* = T_n((\boldsymbol{y_1^*}, \boldsymbol{x_1^*}), (\boldsymbol{y_2^*}, \boldsymbol{x_2^*}), ..., (\boldsymbol{y_n^*}, \boldsymbol{x_n^*}), \boldsymbol{F_n}))$$

- ...that's simply the statistic computed with the simulated sample you have just drawn.

NONPARAMETRIC BOOTSTRAP

- Now we need to get the distribution (confidence interval)
- Hence repeat this B times.

- Note:
  - B is the number of bootstrap replications.
  - Large B is better, but takes longer to compute
  - A theory about that was developed (Andrews and Buchinsky, 2000), but
  - $B = 1000$ is typically sufficient and reasonable to compute.

EXAMPLE: T-RATIO

- Now we need to get the distribution (confidence interval)
- Hence repeat this B times.
- Note:
  - B is the number of bootstrap replications.
  - Large B is better, but takes longer to compute
  - A theory about that was developed (Andrews and Buchinsky, 2000), but
  - $B = 1000$ is typically sufficient and reasonable to compute.
- Finally: The general test statistic T_n usually is a function of F
  - e.g. t-ratio $(\hat{\theta} - \theta)/s(\hat{\theta})$ depends on θ:
- Bootstrap replaces F with F_n and similarly θ with θ_n
  - ...usually that's simply the newly estimated $\hat{\theta}$

ESTIMATING BIAS

- Denote the Bias $\mathrm{E}(\hat{\theta} - \theta_0) = \tau_n$
- Let $T_n(\boldsymbol{\theta}) = (\hat{\theta} - \theta)$ (*...for any theta*), hence $\tau_n = \mathrm{E}(T_n(\boldsymbol{\theta}_0))$
- The Bootstrap estimate of $\boldsymbol{\tau_n}$ is given by

ESTIMATING BIAS

- Denote the Bias $\mathrm{E}(\hat{\theta} - \theta_0) = \tau_n$
- Let $T_n(\boldsymbol{\theta}) = (\hat{\theta} - \theta)$ (*...for any theta*), hence $\tau_n = \mathrm{E}(T_n(\boldsymbol{\theta}_0))$
- The Bootstrap estimate of $\boldsymbol{\tau_n}$ is given by

$$\tau_n^* = \mathrm{E}(T_n^*(\boldsymbol{\theta}_0))$$

- ...it's all in the stars

COMPUTATION OF THE ESTIMATE

- Denote the Bias $\mathrm{E}(\hat{\theta} - \theta_0) = \tau_n$
- Let $T_n(\boldsymbol{\theta}) = (\hat{\theta} - \theta)$ (*...for any theta*), hence $\tau_n = \mathrm{E}(T_n(\boldsymbol{\theta}_0))$
- Using the simulation, the estimate of $\boldsymbol{\tau_n}$ is given by

COMPUTATION OF THE ESTIMATE

- Denote the Bias $\mathrm{E}(\hat{\theta} - \theta_0) = \tau_n$
- Let $T_n(\boldsymbol{\theta}) = (\hat{\theta} - \theta)$ (*...for any theta*), hence $\tau_n = \mathrm{E}(T_n(\boldsymbol{\theta}_0))$
- Using the simulation, the estimate of $\boldsymbol{\tau_n}$ is given by

$$\begin{aligned}\hat{\tau_n^*} &= \frac{1}{B}\sum_{b=1}^{B} T_{nb}^* \\ &= \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^* - \hat{\theta} \\ &= \overline{\hat{\theta}^*} - \hat{\theta}\end{aligned}$$

COMPUTATION OF THE ESTIMATE

- Denote the Bias $\mathrm{E}(\hat{\theta} - \theta_0) = \tau_n$
- Let $T_n(\boldsymbol{\theta}) = (\hat{\theta} - \theta)$ (*...for any theta*), hence $\tau_n = \mathrm{E}(T_n(\boldsymbol{\theta}_0))$
- Using the simulation, the estimate of $\boldsymbol{\tau_n}$ is given by

$$\begin{aligned}\hat{\tau_n^*} &= \frac{1}{B}\sum_{b=1}^{B} T_{nb}^* \\ &= \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^* - \hat{\theta} \\ &= \overline{\hat{\theta}^*} - \hat{\theta}\end{aligned}$$

- ...that is, subtract your "real" estimate from the BS average estimate. Similarly the variance:

$$\hat{V_n^*} = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b^* - \overline{\hat{\theta}^*})^2$$

- This procedure also can afford a bias-correction and Percentile Interval Estimation, Symmetric Percentile-t Intervals, Asymptotic Expansion etc. (cf. Hansen, ch. 10)

***

Table of Contents

NON-LINEAR IN ERRORS

- Models which can not be transformed into linear models. For example, models with limited dependent variables
  - Binary variables: employed vs. not employed outcome.
  - Categorical variables: Choose between 5 political candidates.
  - Nonnegative variables: wages, prices, interest rates.
  - Nonnegative variables with excess zeros: labor supply, doctor visits.
  - Count variables: the number of cigarettes smoked per day.
  - Censored variables: unemployment durations.

NON-LINEAR IN ERRORS

- Models which can not be transformed into linear models. For example, models with limited dependent variables
  - Binary variables: employed vs. not employed outcome.
  - Categorical variables: Choose between 5 political candidates.
  - Nonnegative variables: wages, prices, interest rates.
  - Nonnegative variables with excess zeros: labor supply, doctor visits.
  - Count variables: the number of cigarettes smoked per day.
  - Censored variables: unemployment durations.
- The partial effects no longer straightforward

$$\frac{\partial \mathrm{E}(y_i|\mathbf{x}_i)}{\partial \mathbf{x}_i} \neq \beta$$

- Now the effects depend upon the level of the variables $\mathbf{x}_i$.
- The average partial effects (APE): $\mathrm{E}\frac{\partial \mathrm{E}(y_i|\mathbf{x}_i)}{\partial \mathbf{x}_i}$ estimated by $\frac{1}{n}\sum_i^n \frac{\partial \mathrm{E}(y_i|\mathbf{x}_i)}{\partial \mathbf{x}_i}$.
- The partial effects at the average (PEA): Partial effects $\frac{\partial \mathrm{E}(y_i|\mathbf{x}_i)}{\partial \mathbf{x}_i}$ when fixing $\mathbf{x}_i = \mathrm{E}(\mathbf{x}_i)$ estimated by $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$.

LIKELIHOOD PRINCIPLE

- *Likelihood principle* (R. A. Fisher, 1922): Choose as estimator of parameter vector θ_0 the value of θ which maximizes likelihood of observing the actual data in sample within proposed model.
  - In discrete case: likelihood is the probability obtained from the probability mass function.
  - In continuous case: likelihood is the is the density (since probability of observing single realization of continuous variable is always 0).
- Example in discrete case: If one value of θ implies that probability of the observed data occurring is 0.0010, and other value of θ gives a higher probability of 0.0015, then second value better estimator.

Table of Contents

LIKELIHOOD AND LOG LIKELIHOOD FUNCTION

- Joint probability mass function or density $f(\mathbf{y}, \mathbf{X}|\theta)$ is viewed here as a function of θ given the data $(\mathbf{y}, \mathbf{X})$.
- This is called the likelihood function and is denoted by $= L(\theta|\mathbf{y}, \mathbf{X})$.
- Maximizing $L(\theta)$ is equivalent to maximizing the log-likelihood function

$$\mathcal{L}(\theta) = lnL(\theta)$$

- ML estimator special property: it is most efficient estimator among consistent and asymptotically normal estimators.

LIKELIHOOD FUNCTION AND CONDITIONAL LIKELIHOOD

- Given data $(y_i, \mathbf{x}_i)$, the likelihood $L(\boldsymbol{\theta}) = f(\mathbf{y}, \mathbf{X}|\boldsymbol{\theta}) = f(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) f(\mathbf{X}|\boldsymbol{\theta})$ requires specification of conditional density of $\mathbf{y}$ given $\mathbf{X}$ and the marginal density of $\mathbf{X}$.
- Often we assume data generating process (DGP) of $\mathbf{y}$ given $\mathbf{X}$ and of $\mathbf{X}$ depend on mutually exclusive set of parameters, $f(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\eta}) = f(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$ and $f(\mathbf{X}|\boldsymbol{\theta}, \boldsymbol{\eta}) = f(\mathbf{X}|\boldsymbol{\eta})$.
- Since our primary interest is in the relationship between y_i and $\mathbf{x}_i$ we usually focus on *conditional likelihood function*: $L(\boldsymbol{\theta}) = f(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$
- Can you think about situation where this would not be appropriate?

LIKELIHOOD FUNCTION AND CONDITIONAL LIKELIHOOD

- Given data $(y_i, \mathbf{x}_i)$, the likelihood $L(\theta) = f(\mathbf{y}, \mathbf{X}|\theta) = f(\mathbf{y}|\mathbf{X}, \theta) f(\mathbf{X}|\theta)$ requires specification of conditional density of $\mathbf{y}$ given $\mathbf{X}$ and the marginal density of $\mathbf{X}$.
- Often we assume data generating process (DGP) of $\mathbf{y}$ given $\mathbf{X}$ and of $\mathbf{X}$ depend on mutually exclusive set of parameters, $f(\mathbf{y}|\mathbf{X}, \theta, \eta) = f(\mathbf{y}|\mathbf{X}, \theta)$ and $f(\mathbf{X}|\theta, \eta) = f(\mathbf{X}|\eta)$.
- Since our primary interest is in the relationship between y_i and $\mathbf{x}_i$ we usually focus on *conditional likelihood function*: $L(\theta) = f(\mathbf{y}|\mathbf{X}, \theta)$
- Can you think about situation where this would not be appropriate?
- For endogenous sampling, consistent estimation requires full density.
- For time series case (leaving out $\mathbf{X}$ for simplicity) we might have for $t = 0, \ldots, T$,

$$
\begin{aligned}
f(\mathbf{y}|\theta) &= f(y_T|y_{T-1}, \ldots, y_1, \theta) \cdot f(y_{T-1}|y_{T-2}, \ldots, y_1, \theta) \cdots f(y_2|y_1, \theta) \cdot f(y_1|\theta) \\
&= \prod_{t=1}^{T} f(y_t|y_{t-1}, \ldots, y_1, \theta)
\end{aligned}
$$

MAXIMUM LIKELIHOOD ESTIMATOR

- Joint density of IID sample $(y_1, \ldots, y_n)$ given $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is,

$$L(\theta) = \prod_{i=1}^{n} f(y_i|\mathbf{x}_i, \theta)$$

- The log-likelihood in turn can be written,

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} lnf(y_i|\mathbf{x}_i, \theta)$$

- Note that likelihood-based models should not be specified by making assumptions on the distribution of an error term.
- Assuming linear CEF model and $y_i|\mathbf{x}_i \sim \mathcal{N}(\mathbf{x}_i'\beta, \sigma^2)$, and independence of u_i and x_i implies $u_i|\mathbf{x}_i \sim \mathcal{N}(0, \sigma^2)$.
- So it is simpler to say errors are normally distributed $u_i|\mathbf{x}_i \sim \mathcal{N}(0, \sigma^2)$, but this is because of assumed additivity of errors.

MAXIMUM LIKELIHOOD ESTIMATOR

- The maximum likelihood estimator (MLE) $\hat{\theta}_{ML}$ is parameter value which maximizes the log-likelihood,

$$\hat{\theta}_{ML} = \arg\max_{\theta} \mathcal{L}(\theta)$$

- We can view $\hat{\theta}_{ML}$ as estimator within class of extremum estimators.
- Notice that $\frac{1}{n}\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} lnf(y_i|\mathrm{x}_i,\theta) \xrightarrow{p} \mathrm{E}(lnf(y_i|\mathbf{x}_i,\theta))$.
- The f.o.c. imply that $\hat{\theta}_{ML}$ maximises population function at $\theta = \theta_0$:

$$\frac{\partial}{\partial \theta}\mathbf{E}(lnf(y_i|\mathbf{x}_i,\theta))\Big|_{\theta=\theta_0} = \mathbf{0}$$

- Sample counterpart f.o.c. imply that $\hat{\theta}_{ML}$ solves,

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial lnf(y_i|\mathbf{x}_i,\theta)}{\partial \theta} = \mathbf{0}$$

- In some simple cases, explicit expression exists for $\hat{\theta}_{ML}$ as a function of data, but typically $\hat{\theta}_{ML}$ must be estimated using numerical methods.

ML IN NORMAL REGRESSION MODEL

- Normal regression model assumes linear CEF model and $y_i|\mathbf{x}_i \sim \mathcal{N}(\mathbf{x}_i'\beta, \sigma^2)$, and independence of u_i and x_i (homoskedasticity)
- The log-likelihood of the normal regression model is,

$$\begin{aligned}\mathcal{L}(\beta, \sigma^2) &= \sum_{i=1}^{n} ln\left(\frac{1}{(2\pi\sigma^2)^{1/2}} exp(-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i'\beta)^2)\right) \\ &= -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \frac{n}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)\end{aligned}$$

ML IN NORMAL REGRESSION MODEL

- Maximization with respect to β yields f.o.c.

$$\frac{\partial}{\partial \beta}\mathcal{L}(\beta, \sigma^2) = \frac{1}{\sigma^2}(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\hat{\beta}_{ML}) = \mathbf{0}$$

$$\Leftrightarrow \qquad \hat{\beta}_{ML} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- The equivalence of $\hat{\beta}_{ML} = \hat{\beta}_{OLS}$ should not be surprising since maximizing $\mathcal{L}(\beta, \sigma^2)$ with respect to β is equivalent to minimizing sum of squared error criterion function for OLS.

ML IN NORMAL REGRESSION MODEL

- Maximization with respect to β yields f.o.c.

$$\frac{\partial}{\partial \beta}\mathcal{L}(\beta, \sigma^2) = \frac{1}{\sigma^2}(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\hat{\beta}_{ML}) = \mathbf{0}$$
$$\Leftrightarrow \qquad \hat{\beta}_{ML} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- The equivalence of $\hat{\beta}_{ML} = \hat{\beta}_{OLS}$ should not be surprising since maximizing $\mathcal{L}(\beta, \sigma^2)$ with respect to β is equivalent to minimizing sum of squared error criterion function for OLS.

- Maximization with respect to σ^2 yields f.o.c.

$$\frac{\partial}{\partial \sigma^2}\mathcal{L}(\beta, \sigma^2) = -\frac{n}{2\hat{\sigma}^2_{ML}} + \frac{1}{2(\hat{\sigma}^2_{ML})^2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) = 0$$

- Plugging in $\hat{\beta}_{ML}$ for β and solving for $\hat{\sigma}^2_{ML}$ results in

$$\hat{\sigma}^2_{ML} = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\beta}_{ML})'(\mathbf{y} - \mathbf{X}\hat{\beta}_{ML})$$

Table of Contents

SCORE AND HESSIAN

- The likelihood score is the gradient (vector of partial derivatives) of the log-likelihood evaluated at true parameter θ_0,

$$\mathbf{S}_i = \frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L}_i(\boldsymbol{\theta}_0) = \frac{\partial}{\partial \boldsymbol{\theta}} lnf(y_i|\mathbf{x}_i, \boldsymbol{\theta}_0)$$

- The Hessian is the matrix of partial derivatives of the score,

$$\mathbf{H}_i(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \mathcal{L}_i(\boldsymbol{\theta}_0) = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} lnf(y_i|\mathbf{x}_i, \boldsymbol{\theta}_0)$$

- We use here following shorthand notation,

$$\mathbf{S}_i(\tilde{\boldsymbol{\theta}}) = \frac{\partial}{\partial \boldsymbol{\theta}} lnf(y_i|\mathbf{x}_i, \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} = \frac{\partial}{\partial \boldsymbol{\theta}} lnf(y_i|\mathbf{x}_i, \tilde{\boldsymbol{\theta}})$$

$$\mathbf{H}_i(\tilde{\boldsymbol{\theta}}) = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} lnf(y_i|\mathbf{x}_i, \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} lnf(y_i|\mathbf{x}_i, \tilde{\boldsymbol{\theta}})$$

PROPERTIES OF LIKELIHOOD

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{E}_{\theta_0}(lnf(y_i|\mathbf{x}_i,\boldsymbol{\theta}))\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \mathbf{0}$$

where $\mathrm{E}_{\boldsymbol{\theta}_0}(\cdot)$ means we are taking expectation under the DGP characterized by $\boldsymbol{\theta}_0$.

- Proof:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{E}_{\boldsymbol{\theta}_0}(lnf(y_i|\mathbf{x}_i,\boldsymbol{\theta}))\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} &= \frac{\partial}{\partial \boldsymbol{\theta}} \int lnf(y_i|\mathbf{x}_i,\boldsymbol{\theta}) f(y_i|\mathbf{x}_i,\boldsymbol{\theta}_0) dy_i \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\
&= \int \frac{\partial}{\partial \boldsymbol{\theta}} f(y_i|\mathbf{x}_i,\boldsymbol{\theta}) \frac{f(y_i|\mathbf{x}_i,\boldsymbol{\theta}_0)}{f(y_i|\mathbf{x}_i,\boldsymbol{\theta})} dy_i \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\
&= \int \frac{\partial}{\partial \boldsymbol{\theta}} f(y_i|\mathbf{x}_i,\boldsymbol{\theta}_0) \frac{f(y_i|\mathbf{x}_i,\boldsymbol{\theta}_0)}{f(y_i|\mathbf{x}_i,\boldsymbol{\theta}_0)} dy_i \\
&= \int \frac{\partial}{\partial \boldsymbol{\theta}} f(y_i|\mathbf{x}_i,\boldsymbol{\theta}) dy_i \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\
&= \frac{\partial}{\partial \boldsymbol{\theta}} \int f(y_i|\mathbf{x}_i,\boldsymbol{\theta}) dy_i \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \frac{\partial}{\partial \boldsymbol{\theta}} 1 \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \mathbf{0}
\end{aligned}$$

PROPERTIES OF LIKELIHOOD

$$\mathrm{E}_{\boldsymbol{\theta}_0}(\mathbf{S}_i(\boldsymbol{\theta}_0)) = \mathbf{0}$$

- Proof:

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{\theta}_0}(\mathbf{S}_i(\boldsymbol{\theta}_0)) &= \mathrm{E}_{\boldsymbol{\theta}_0}\left(\frac{\partial}{\partial \boldsymbol{\theta}} lnf(y_i|\mathbf{x}_i, \boldsymbol{\theta}_0)\right) \\
&= \frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{E}_{\boldsymbol{\theta}_0}(lnf(y_i|\mathbf{x}_i, \boldsymbol{\theta}))\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\
&= \mathbf{0}
\end{aligned}$$

- and similarly,

$$\mathrm{E}_{\boldsymbol{\theta}_0}\left(\frac{\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} f(y_i|\mathbf{x}_i, \boldsymbol{\theta}_0)}{f(y_i|\mathbf{x}_i, \boldsymbol{\theta}_0)}\right) = \mathbf{0}$$

THE HESSIAN AND THE EXPECTED OUTER PRODUCT SCORE

$$-\mathrm{E}_{\boldsymbol{\theta}_0}(\mathbf{H}_i(\boldsymbol{\theta}_0)) = \mathrm{E}_{\boldsymbol{\theta}_0}(\mathbf{S}_i(\boldsymbol{\theta}_0)\mathbf{S}_i'(\boldsymbol{\theta}_0))$$

- Proof:

$$\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} lnf(y_i|\mathbf{x}_i,\boldsymbol{\theta}_0) = \frac{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} f(y_i|\mathbf{x}_i,\boldsymbol{\theta}_0)}{f(y_i|\mathbf{x}_i,\boldsymbol{\theta}_0)} - \frac{\frac{\partial}{\partial\boldsymbol{\theta}} f(y_i|\mathbf{x}_i,\boldsymbol{\theta}_0)\frac{\partial}{\partial\boldsymbol{\theta}} f(y_i|\mathbf{x}_i,\boldsymbol{\theta}_0)'}{f(y_i|\mathbf{x}_i,\boldsymbol{\theta}_0)^2}$$

- Take expectations on both sides under the DGP characterized by $\boldsymbol{\theta}_0$,

$$-\mathrm{E}_{\boldsymbol{\theta}_0}(\mathbf{H}_i(\boldsymbol{\theta}_0)) = \mathrm{E}_{\boldsymbol{\theta}_0}(\mathbf{S}_i(\boldsymbol{\theta}_0)\mathbf{S}_i(\boldsymbol{\theta}_0)')$$

Table of Contents

INFORMATION MATRIX EQUALITY

- Define $\mathcal{H} = -\mathrm{E}_{\boldsymbol{\theta}_0}(\mathbf{H}_i(\boldsymbol{\theta}_0))$ and $\dot{} = \mathrm{E}_{\boldsymbol{\theta}_0}(\mathbf{S}_i(\boldsymbol{\theta}_0)\mathbf{S}_i(\boldsymbol{\theta}_0)')$ then

$$\mathcal{H} = \dot{} \equiv \mathcal{I}$$

- $\mathcal{I}$ is the Fisher information matrix, it is a way of measuring the amount of information that observable random variables $(y_i, \mathbf{x}_i)$ carry about $\boldsymbol{\theta}_0$.
- The relation $\mathcal{H} = \dot{} \equiv \mathcal{I}$ is called the information matrix equality.

ML ASYMPTOTIC DISTRIBUTION

- We will not go extensively through the proof since it is very similar to that of a GMM estimator.
- Check Newey and McFadden (1994) for thorough proof and B. Hansen (appendix B.11) for proof sketch.
- Under regularity conditions,

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \quad \xrightarrow{d} \quad \mathcal{H}^{-1}\mathcal{N}(\mathbf{0}, \dot{}) = \mathcal{N}(\mathbf{0}, \mathcal{H}^{-1}\dot{}\mathcal{H}^{-1}) = \mathcal{N}(\mathbf{0}, \mathcal{I}^{-1})$$

- This is a beautiful result since it shows that the asymptotic variance of the MLE estimator is

$$\mathbf{V}_{\theta_{ML}} = \mathcal{H}^{-1} = \dot{}^{-1} = \mathcal{I}^{-1}$$

INTUITION BEHIND $\mathbf{V}_{\theta_{ML}} = \mathcal{H}^{-1} = {}^{\bullet -1} = \mathcal{I}^{-1}$

- Intuitively, think of asymptotic variance $\mathbf{V}_{\theta_{ML}}$ as measure of precision of θ_{ML}. It is expected value of (scaled) square of 'average' *spread* of θ_{ML} when we have infinite data.
- Right side of information matrix equality says that the precision of $\hat{\theta}_{ML}$ depends on shape of log-likelihood function near $\hat{\theta}_{ML}$.
- If log-likelihood very *curved* or *steep* around $\hat{\theta}_{ML}$, then θ_0 will be precisely estimated. In this case, we say that we have a lot of information about θ_0.

INTUITION BEHIND $\mathbf{V}_{\theta_{ML}} = \mathcal{H}^{-1} = {}^{\bullet -1} = \mathcal{I}^{-1}$

- Intuitively, think of asymptotic variance $\mathbf{V}_{\theta_{ML}}$ as measure of precision of θ_{ML}. It is expected value of (scaled) square of 'average' *spread* of θ_{ML} when we have infinite data.
- Right side of information matrix equality says that the precision of $\hat{\theta}_{ML}$ depends on shape of log-likelihood function near $\hat{\theta}_{ML}$.
- If log-likelihood very *curved* or *steep* around $\hat{\theta}_{ML}$, then θ_0 will be precisely estimated. In this case, we say that we have a lot of information about θ_0.
- *Steepness* of log-likelihood captured by first derivative which relates to $\mathbf{S}_i(\theta_0)$.
- *Curvature* of log-likelihood captured by its second derivative. Information in sample therefore also relates to minus the Hessian $-\mathbf{H}_i(\theta_0)$ (since the Hessian is negative semi-definite).
- Information matrix equality says when sample size 'plays no role' in increasing efficiency, all three measures have a simple equivalence.

$\hat{\beta}_{ML}$ AND $\widehat{\mathbf{V}}_{\beta_{ML}}$ ESTIMATORS FOR NORMAL REGRESSION UNDER HOMOSKEDASTICITY

- Under homoskedasticity assumption, asymptotic distribution of MLE estimators is

$$\begin{bmatrix} \hat{\beta}_{ML} \\ \hat{\sigma}^2_{ML} \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{bmatrix} \beta_0 \\ \sigma^2_0 \end{bmatrix}, \begin{bmatrix} \sigma^2(\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ 0 & \frac{2}{n}\sigma^4 \end{bmatrix} \right)$$

- Additive separability and normality of the errors implies here that covariance terms are 0.

Table of Contents

CRAMER-RAO LOWER BOUND AND INFORMATION MATRIX

- Cramer-Rao Lower Bound theorem shows that asymptotic variance of the MLE estimator is smallest possible within a certain class of estimators.
- Cramer-Rao Lower Bound (CRLB): If $\tilde{\theta}$ is an unbiased regression estimator of θ, then $\mathbf{V}_{\tilde{\theta}} \geq (n\mathcal{I})^{-1}$.
- Implies asymptotic variance of the standardized estimator $\sqrt{n}(\hat{\theta}_{ML} - \theta_0)$ is bounded below by $\mathcal{I}^{-1}$.
- $\hat{\theta}_{ML}$ is therefore asymptotically efficient.
- Efficiency bounds: CRLB is to likelihood framework what Gauss-Markov theorem (or GMM extension) is to second moment context.

CRAMER-RAO LOWER BOUND PROOF

1. Define $\mathbf{S}(\boldsymbol{\theta}_0) = \frac{\partial}{\partial \boldsymbol{\theta}} lnf(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}_0) = \sum_{i=1}^{n} \mathbf{S}_i(\boldsymbol{\theta}_0)$ which from slide 14-15 we know has mean $\mathrm{E}(\mathbf{S}(\boldsymbol{\theta}_0)) = \mathbf{0}$ and variance $\mathrm{E}(\mathbf{S}(\boldsymbol{\theta}_0)\mathbf{S}'(\boldsymbol{\theta}_0)) = n\mathcal{I}$.
2. We assume $\tilde{\boldsymbol{\theta}}$ is unbiased estimator for any $\boldsymbol{\theta}$ and writing $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{X})$ as function of $(\mathbf{y}, \mathbf{X})$ we have,

$$\boldsymbol{\theta} = \mathrm{E}(\tilde{\boldsymbol{\theta}}) = \int \tilde{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{X}) f(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) d\mathbf{y}$$

CRAMER-RAO LOWER BOUND PROOF

1. Define $\mathbf{S}(\boldsymbol{\theta}_0) = \frac{\partial}{\partial \boldsymbol{\theta}} lnf(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}_0) = \sum_{i=1}^{n} \mathbf{S}_i(\boldsymbol{\theta}_0)$ which from slide 14-15 we know has mean $\mathrm{E}(\mathbf{S}(\boldsymbol{\theta}_0)) = \mathbf{0}$ and variance $\mathrm{E}(\mathbf{S}(\boldsymbol{\theta}_0)\mathbf{S}'(\boldsymbol{\theta}_0)) = n\mathcal{I}$.
2. We assume $\tilde{\boldsymbol{\theta}}$ is unbiased estimator for any $\boldsymbol{\theta}$ and writing $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{X})$ as function of $(\mathbf{y}, \mathbf{X})$ we have,
$$\boldsymbol{\theta} = \mathrm{E}(\tilde{\boldsymbol{\theta}}) = \int \tilde{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{X}) f(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) d\mathbf{y}$$
3. Differentiating both sides with respect to $\boldsymbol{\theta}'$ and evaluating at $\boldsymbol{\theta}_0$,
$$\begin{aligned} \mathbf{I} &= \int \tilde{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{X}) \frac{\partial}{\partial \boldsymbol{\theta}'} f(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) d\mathbf{y} \\ &= \int \tilde{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{X}) \mathbf{S}(\boldsymbol{\theta}_0)' f(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}_0) d\mathbf{y} \qquad \text{(see slide 13)} \\ &= \mathrm{E}((\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\mathbf{S}(\boldsymbol{\theta}_0)') \qquad (\text{from } \mathrm{E}(\boldsymbol{\theta}_0 \mathbf{S}(\boldsymbol{\theta}_0)) = \boldsymbol{\theta}_0 \mathrm{E}(\mathbf{S}(\boldsymbol{\theta}_0)) = \mathbf{0}) \end{aligned}$$

CRAMER-RAO LOWER BOUND PROOF

1. Define $\mathbf{S}(\theta_0) = \frac{\partial}{\partial \boldsymbol{\theta}} lnf(\mathbf{y}|\mathbf{X}, \theta_0) = \sum_{i=1}^{n} \mathbf{S}_i(\theta_0)$ which from slide 14-15 we know has mean $\mathrm{E}(\mathbf{S}(\theta_0)) = \mathbf{0}$ and variance $\mathrm{E}(\mathbf{S}(\theta_0)\mathbf{S}'(\theta_0)) = n\mathcal{I}$.
2. We assume $\tilde{\theta}$ is unbiased estimator for any θ and writing $\tilde{\theta} = \tilde{\theta}(\mathbf{y}, \mathbf{X})$ as function of $(\mathbf{y}, \mathbf{X})$ we have,
$$\theta = \mathrm{E}(\tilde{\theta}) = \int \tilde{\theta}(\mathbf{y}, \mathbf{X}) f(\mathbf{y}|\mathbf{X}, \theta) d\mathbf{y}$$
3. Differentiating both sides with respect to θ' and evaluating at θ_0,
$$\begin{aligned} \mathbf{I} &= \int \tilde{\theta}(\mathbf{y}, \mathbf{X}) \frac{\partial}{\partial \theta'} f(\mathbf{y}|\mathbf{X}, \theta) d\mathbf{y} \\ &= \int \tilde{\theta}(\mathbf{y}, \mathbf{X}) \mathbf{S}(\theta_0)' f(\mathbf{y}|\mathbf{X}, \theta_0) d\mathbf{y} \qquad \text{(see slide 13)} \\ &= \mathrm{E}((\tilde{\theta} - \theta_0)\mathbf{S}(\theta_0)') \qquad (\text{from } \mathrm{E}(\theta_0 \mathbf{S}(\theta_0)) = \theta_0 \mathrm{E}(\mathbf{S}(\theta_0)) = \mathbf{0}) \end{aligned}$$
4. Last, by Cauchy-Schwartz inequality for matrices we have,
$$\begin{aligned} \mathbf{V}_{\tilde{\theta}} &= \mathrm{E}((\tilde{\theta} - \theta_0)(\tilde{\theta} - \theta_0)') \\ &\geq \mathrm{E}((\tilde{\theta} - \theta_0)\mathbf{S}(\theta_0)')\mathrm{E}(\mathbf{S}(\theta_0)\mathbf{S}(\theta_0)')^{-1}\mathrm{E}(\mathbf{S}(\theta_0)(\tilde{\theta} - \theta_0)') \\ &= \mathrm{E}(\mathbf{S}(\theta_0)\mathbf{S}(\theta_0)')^{-1} \\ &= (n\mathcal{I})^{-1} \end{aligned}$$

CRAMER-RAO LOWER BOUND AND INVARIANCE PRINCIPLE

- Finally, consider functions of parameters. If $\psi = \mathbf{g}(\theta)$ then the MLE of $\hat{\psi}_{ML}$ is $\mathbf{g}(\hat{\theta}_{ML})$.
- This is because maximization of objective function is unaffected by parameterization and transformation.

$$\sqrt{n}(\hat{\psi}_{ML} - \psi) \simeq \mathbf{G}_0 \sqrt{n}(\hat{\theta}_{ML} - \theta_0) \quad \xrightarrow{d} \quad \mathcal{H}^{-1}\mathcal{N}(\mathbf{0}, \mathbf{G}_0' \mathcal{I}^{-1} \mathbf{G}_0)$$

  with $\mathbf{G}_0 = \frac{\partial}{\partial \theta} \mathbf{g}(\theta_0)$.

- A useful extension is that Cramer-Rao lower bound for $\psi = \mathbf{g}(\hat{\theta})$ is $\mathbf{G}_0' \mathcal{I}^{-1} \mathbf{G}_0$, and the MLE $\widehat{\psi}_{ML} = \mathbf{g}(\hat{\theta}_{ML})$ is asymptotically efficient.