Inference and Estimation

Author: Wenxiao Yang

Institute: Haas School of Business, University of California Berkeley

Date: 2023

All models are wrong, but some are useful.

Contents

Chapter 1 Statistics Basics

Objective: Using x to give (data-based) answers to questions about the distribution of X, i.e., P_0.

Probability vs. Statistics:

- Probability: Distribution known, outcome unknown;
- Statistics: Distribution unknown, outcome known.

Setting: $X_1, ..., X_n$ is a random sample from a discrete/continuous distribution with pmf/pdf $f(\cdot \mid \theta)$, where $\theta \in \Theta$ is unknown.

Types of Statistical Inference:

- Point estimation $\Rightarrow$ "What is θ?";
- Hypothesis testing $\Rightarrow$ "Is $\theta = \theta_0$?";
- Interval estimation $\Rightarrow$ "Which values of θ are 'plausible'?".

Example 1.1

Examples of Statistical Models

(1). $x_i \sim$ i.i.d. Bernoulli(p), where p is unknown.

(2). $x_i \sim$ i.i.d. $U(0, \theta)$, where $\theta > 0$ is unknown.

(3). $x_i \sim$ i.i.d. $N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown.

1.1 Random Sampling

Definition 1.1 (Random Sample)

A **random sample** is a collection $X_1, ..., X_n$ of random variables that are (mutually) independent and identical marginal distributions.

$X_1, ..., X_n$ are called "independent and identically distributed". The notation is $X_i \sim i.i.d.$

Definition 1.2 (Statistic)

A **statistic** (singular) or sample statistic is any quantity computed from values in a sample which is considered for a statistical purpose.

If $X_1, ..., X_n$ is a random sample and $T : \mathbb{R}^n \to \mathbb{R}^k$ (for some $k \in \mathbb{N}$), then $T(X_1, ..., X_n)$ is called a **statistic**.

1.1.1 Sample Mean and Sample Variance

Definition 1.3 (Sample Mean and Sample Variance)

The **sample mean** is $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$;

The **sample variance** is $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2)$

Note *We use "$X_i \sim i.i.d(\mu, \sigma^2)$" to denote a random sample from a distribution with mean μ and variance σ^2.*

Theorem 1.1 ($\mathbb{E}(\bar{X})$, Var$(\bar{X})$, $\mathbb{E}(S^2)$)

Suppose $X_1, ..., X_n$ is a random sample from a distribution with mean μ and variance σ^2 (denoted by $X_i \sim$ i.i.d(μ, σ^2)). Then,

(a). $\mathbb{E}(\bar{X}) = \mu$;

(b). Var$(\bar{X}) = \frac{\sigma^2}{n}$;

(c). $\mathbb{E}(S^2) = \sigma^2$.

1.1.2 Distributional Properties

Theorem 1.2 (Distributional Properties of Normal Distributions)

If $X_i \sim$ i.i.d. $N(\mu, \sigma^2)$, then

(a). $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$

(b). $\frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$

(c). $\bar{X} \perp S^2$

Theorem 1.3 ("Asymptotics")

If $X_i \sim$ i.i.d. (μ, σ^2) and if n is "large", then

(a). $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$ (converges in distribution) by CLT **??**;

(b). $S^2 = \sigma^2$ by LLN;

1.1.3 Order Statistics

Definition 1.4 (Order Statistics)

If $X_1, ..., X_n$ is a random sample, then the **characteristics** are the sample values placed in ascending order. Notation:

$$X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$$

Proposition 1.1 (Distribution of $X_n = \max_{i=1,\dots,n} X_i$)

If $X_1, \dots, X_n$ is a random sample form a distribution with cdf F (denoted by "$X_i \sim$ i.i.d. F"), then

$$F_{X_{(n)}}(x) = P(X_{(n)} \leq x) = F^n(x)$$

Proposition 1.2 (cdf and pdf)

More generally,

$$F_{X_{(r)}}(x) = \sum_{j=r}^{n} \binom{n}{j} [F_X(x)]^j [1 - F_X(x)]^{n-j}$$

$$f_{X_{(r)}}(x) = \frac{n!}{(r-1)!(n-r)!} f_X(x) [F_X(x)]^{r-1} [1 - F_X(x)]^{n-r}$$

Example 1.2

Order statistics sampled from a uniform distribution on unit interval (Unif$[0,1]$): Consider a random sample $U_1, \dots, U_n$ from the standard uniform distribution. Then,

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} u^{k-1}(1-u)^{n-k}$$

The k^{th} order statistic of the uniform distribution is a beta-distributed random variable.

$$U_{(k)} \sim \text{Beta}(k, n+1-k)$$

which has mean $\mathbb{E}[U_{(k)}] = \frac{k}{n+1}$.

Example 1.3

The joint distribution of the order statistics of the uniform distribution on unit interval (Unif$[0,1]$): Similarly, for $i < j$, the joint probability density function of the two order statistics $U_{(i)} < U_{(j)}$ can be shown to be

$$f_{U_{(i)},U_{(j)}}(u,v) = n! \frac{u^{i-1}}{(i-1)!} \frac{(v-u)^{j-i-1}}{(j-i-1)!} \frac{(1-v)^{n-j}}{(n-j)!}$$

The joint density of the n order statistics turns out to be constant:

$$f_{U_{(1)},U_{(2)},\dots,U_{(n)}}(u_1, u_2, \dots, u_n) = n!$$

For $n \geq k > j \geq 1$, $U_{(k)} - U_{(j)}$ also has a beta distribution:

$$U_{(k)} - U_{(j)} \sim \text{Beta}(k-j, n-(k-j)+1)$$

which has mean $\mathbb{E}[U_{(k)} - U_{(j)}] = \frac{k-j}{n+1}$.

1.2 Statistics Model (ECON 240B)

1.2.1 Model

A statistical model is a family of probability distributions over the data.

In statistics, we define *data* be a vector $x = (x_1, ..., x_n)' \in \Omega$ of numbers, where $x_i \in \mathbb{R}^d$. x is the realization of a random vector $X = (X_1, ..., X_n)'$. The X follows a distribution P_0, which is the *True Probability Generating Data (DGP)*. If P_0 is i.i.d., we have $P_0(X) = P_0(x_1)P(x_2)\cdots P_0(x_n)$.

Definition 1.5 (Model)

A model $P \subseteq \{\text{Probabilities over } \Omega\}$ and a i.i.d. model $P \subseteq \{\text{Probabilities over } \mathbb{R}^d\}$.

Definition 1.6 (Well-Specified Model)

A model is **well-specified** if $P \ni P_0$.

1.2.2 Parametric Model

Definition 1.7 (Parametric Model)

A non-parametric model $\bar{P} \cong \{\text{Probabilities over } \mathbb{R}^d\}$.

A parametric model $P = \{P_\theta : \theta \in \Theta \subseteq \mathbb{R}^v\}$.

A semi-parametric model: not parametric / non-parametric.

Example 1.4

Parametric model: $P = \{\Phi(\theta, 1) : \theta \in \mathbb{R}\}$, where Φ is the Gaussian c.d.f.

Example 1.5

Regression Models. $Z := (Y, X)$. P belongs to the model iff $\mathbb{E}_P[y^2] < \infty$ and $\mathbb{E}_P[XX^T]$ is non-singular and finite. The model gives $\mathbb{E}_P[Y|X] = h(X)$.

(A). Semi-parametric model: $h \in \{\text{linear functions}\}$ i.e., $h(X) = \beta^T X$ for some $\beta \in \mathbb{R}^d$.

(B). Non-parametric model: $h \in \{f : \mathbb{E}_p[f(x)^2] < \infty\}$.

1.2.3 Parameter

Example 1.6

Potential Outcome Model: $Z := (Y, D, X)$, where Y is the outcome, $D \in \{0, 1\}$ is the treatment, and X is the covariates.

- P belongs to the model iff $(y_{(0)}, y_{(1)})$ represents the potential outcome given different treatment $D \in \{0, 1\}$, $y = Dy_{(1)} + (1 - D)y_{(0)}$, and
- we study $e(x) := P(D = 1|x)$.
- Average Treatment Effect (ATE) is given by $\text{ATE}_{P_0} := \mathbb{E}_{P_0}[y_{(1)} - y_{(0)}]$, where P_0 is the DGP. It is impossible to estimate the ATE even if we have enough data, since $y_{(1)}$ and $y_{(0)}$ can't be observed at the same time. We need to link it to something we can estimate.

Definition 1.8 (Parameter)

A parameter is a ``feature'' of P_0: $v(P), P \in \mathcal{P}$. Specifically, $v(P_0)$ is the true parameter of the DGP.

Example 1.7

Linear Regression Model: $\mathbb{E}_{P_0}[Y|X] = \beta_0^T X$.

We solve β by $\min_\beta \mathbb{E}_{P_0}[(y - \beta^T x)^2]$. The F.O.C. gives $\mathbb{E}_{P_0}[YX^T] = \beta^T \mathbb{E}_{P_0}[XX^T]$. β_0 solves this.

Example 1.8

Linear Instrumental Variable Model: $\mathbb{E}_P[(Y - \beta_0^T X)|W] = 0$, where W is the instrumental variable. Look at $\mathbb{E}_{P_0}[(Y - \beta^T X)W] = 0$. Consider an estimator $\hat{\beta}$,

$$\begin{aligned} 0 &= \mathbb{E}_{P_0}[(Y - \beta^T X)W] \\ &= \mathbb{E}_{P_0}[(\hat{\beta} - \beta_0)^T XW] \\ &= \underbrace{(\hat{\beta} - \beta_0)^T}_{1\times m} \underbrace{\mathbb{E}_{P_0}[XW]}_{m\times k} \end{aligned}$$

which holds iff $\hat{\beta} = \beta_0$ given $\mathbb{E}_{P_0}[XW]$ has full rank.

Example 1.9

Identification of the ATE in the Potential Outcomes Model: To identify the ATE, we give two assumptions:

$$ATE := \mathbb{E}[Y(1) - Y(0)]$$

To identify the ATE, we give two assumptions:

1. A1 (Overlap): $e(X) := P(D=1|X) \in (0,1)$
2. A2 (Unconfoundednes): $(Y(0), Y(1)) \perp D|X$, i.e., $(Y(0), Y(1))$ are independent of D given X.

ATE $= \mathbb{E}[y(1) - y(0)] = \mathbb{E}[\mathbb{E}[y(1)|X] - \mathbb{E}[y(0)|X]]$. $\mathbb{E}[y|D=1,X] = \mathbb{E}[y(1)|D=1,X]$. Given Assumption A1: $y(1) \perp D|X$, $\mathbb{E}[y|D=1,X] = \mathbb{E}[y(1)|D=1,X] = \mathbb{E}[y(1)|X]$.

Example 1.10

Inference: For a parameter $\theta(P_0)$, we have an estimate $\hat{\theta}_m$ (with sample size m), which has C.D.F. $v(P_0)$. For all $t \in \mathbb{R}$, the C.D.F. is given by

$$v(P_0)(t) = \Pr_{P_0}(\hat{\theta}_m - \theta(P_0) \le t)$$

1.3 Model Estimation (ECON 240B)

1.3.1 Plug-In Estimation

For a model P, we have "identification" $v(P_0) := \theta_0$. How to estimate unknown P_0?

Definition 1.9 (Empirical Probability/CDF)

Empirical probability / CDF:

$$P_m(A) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{Z_i \in A\}$$

By the LLN, $P_m(A) \xrightarrow{P_0} P_0(A)$.

Definition 1.10 (Plug-in estimator)

A **Plug-in estimator** is an estimator based on the empirical CDF, which is given by

$$\hat{\theta}_m = v(P_m)$$

Note: The domain of v is $\mathcal{P}$. Is $v(P_m)$ well-defined? It might be $P_m \notin \mathcal{P}$.

Example 1.11

Examples of Plug-in estimators:

1. $\mathcal{P} = \{\text{all pdf with finite first moments}\}$. $v(P_0) = \mathbb{E}_{P_0}[Z]$, $v(P_m) = \frac{1}{m}\sum_{i=1}^{m} Z_i$.
2. $\mathcal{P}$ is the set of linear regression models. Define the

$$v(P_0) := \operatorname*{argmin}_b \mathbb{E}_{P_0}[(Y - b^T X)^2] = \mathbb{E}_{P_0}[XX^T]^{-1}\mathbb{E}_{P_0}[XY]$$

Then,

$$
\begin{aligned}
v(P_m) :&= \mathbb{E}_{P_m}[(Y - b^T X)^2] \\
&= \underset{b}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} \left(Y_i - b^T X_i\right)^2 = \left(\frac{1}{m} \sum_{i=1}^{m} X_i X_i^T\right)^{-1} \left(\frac{1}{m} \sum_{i=1}^{m} X_i Y_i\right)
\end{aligned}
$$

where $v(P_m)$ is OLS estimator.

3. **GMM**. $\forall P \in \mathcal{P} : \mathbb{E}_P[g(Z, v(p))] = 0$, where g is a known moment function.

$$
v(P_0) = \underset{\theta}{\operatorname{argmin}} \, \mathbb{E}_{P_0}[g(Z,\theta)]^T W \mathbb{E}_{P_0}[g(Z,\theta)]
$$

where W is a weighted matrix.

$$
v(P_m) = \underset{\theta}{\operatorname{argmin}} \left(\frac{1}{m}\sum_{i=1}^{m} g(Z_i,\theta)\right)^T W \left(\frac{1}{m}\sum_{i=1}^{m} g(Z_i,\theta)\right)
$$

The $v(P_m)$ is the **Gaussian Estimator**.

4. (When it doesn't work.) For the linear regression case, $v(P_m) = \underbrace{\left(\frac{1}{m}\sum_{i=1}^{m} X_i X_i^T\right)^{-1}}_{\text{well-defined?}} \left(\frac{1}{m}\sum_{i=1}^{m} X_i Y_i\right)$. If the # of Covariates $>$ m, the estimator is not well-defined.

5. (When it doesn't work.) $\mathcal{P}$ is the potential outcome model. ATE $= v(P_0) = \mathbb{E}_{P_0}[\mu_1(x) - \mu_0(x)]$ where $\mu_d(x) := \mathbb{E}_{P_0}[y|D = d, x], d = 0, 1$.

$$
v(P_m) = \frac{1}{m}\sum_{i=1}^{m}\left(\underbrace{\mathbb{E}_{P_m}[y|D=1,X_i]}_{\text{well-defined?}} - \mathbb{E}_{P_m}[y|D=0,X_i]\right)
$$

$\mathbb{E}_{P_m}[y|D = d, x]$ is "too complex" to define, (consider the example that x is continuous).

What is the solution when the Plug-in estimation doesn't work?

1. Propose a functional form restriction μ_d.
2. "Regularization": Kernel estimators and series estimators.

1.3.2 Bootstrap

Let $v(P_0)$ be the CDF of $\theta(P_m) - \theta(P_0)$, where $C(P_m, P_0) := \theta(P_m) - \theta(P_0)$.

$$
v(P_0)(t) = \Pr_{P_0}\left(C(P_m, P_0) \le t\right), \forall t
$$

Here, the data $\{Z_i\}_i$ is generated from P_0, which forms P_m.

Remark Sometimes, instead of $C(P_m, P_0)$, we may study

$$
v_A(P_0)(t) = \Pr_{P_0}\left(T(P_m, P_0) \le t\right), \forall t
$$

where $T(P_m, P_0) := \frac{C(P_m,P_0)}{\sqrt{\text{Var}_{P_0}(\theta(P_m))}}$.

Definition 1.11 (Bootstrap Estimator)

The Plug-in estimator $v(P_m)$ is a.k.a. the **Bootstrap estimator**. Now, we generate new data i.i.d. from P_m, $\{Z_i^*\}_i \overset{i.i.d.}{\sim} P_m$, which forms P_m^*.

$$v(P_m)(t) := \text{Pr}_{P_m}\left(\theta(P_m^*) - \theta(P_m) \leq t\right)$$

Computation of $v(P_m)$

(1). Draw $\{Z_i^*\}_i$ from P_m and forms P_m^*.

(2). Based on the new P_m^*, compute $C^{(b)}(P_m^*, P_m) = \theta(P_m^*) - \theta(P_m)$.

(3). Repeat (1) and (2):

$$\frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\{C^{(b)}(P_m^*, P_m) \leq t\} \xrightarrow{B\to\infty} v(P_m)(t)$$

Example 1.12 (Sample Mean)

Consider $\theta(P_0) = \mathbb{E}_{P_0}(Z)$, then $\theta(P_m) = \bar{Z}_m = \frac{1}{m}\sum_{i=1}^m Z_i$. $v(P_0)(t) = \text{Pr}_{P_0}\left(\frac{1}{m}\sum_{i=1}^m (Z_i - \mathbb{E}_{P_0}(Z)) \leq t\right)$. The Bootstrap estimator is given by

$$v(P_m)(t) = \text{Pr}_{P_m}\left(\frac{1}{m}\sum_{i=1}^m (Z_i^* - \bar{Z}_m) \leq t\right)$$

or

$$v_A(P_m)(t) = \text{Pr}_{P_m}\left(\sqrt{m}\frac{\frac{1}{m}\sum_{i=1}^m (Z_i^* - \bar{Z}_m)}{\sqrt{\text{Var}_{P_m}(\theta(P_m^*))}} \leq t\right)$$

where $Z_i^* \sim_{i.i.d.} P_m$, $Z_i^* \in \{Z_1, ..., Z_m\}, \forall i \in \{1, ..., m\}$. For the $v_A(P_0)$, $\text{Var}_{P_0}(\theta(P_m)) = \frac{1}{m}\sigma_{P_0}^2(Z)$ and $\text{Var}_{P_m}(\theta(P_m^*)) = \frac{1}{m}\sigma_{P_m}^2(Z) = \frac{1}{m}S_Z^2$, where S_Z^2 is the sample variance of Z.

It is equivalent to give a weight to each Z_i, $\sum_{i=1}^m Z_i^* = \sum_{i=1}^m W_{i,m}Z_i$, where

$$(W_{1,m}, ..., W_{m,m}) \sim \text{Multinomial}\left(\frac{1}{m}, ..., \frac{1}{m}, m\right), \; W_{i,m} \in \{0, 1, ..., m\}$$

Based on this, the Bootstrap estimator can be rewritten as

$$v(P_m)(t) = \text{Pr}\left(\frac{1}{m}\sum_{i=1}^m (W_{i,m} - 1)Z_i \leq t\right)$$

(Other Bootstrap procedure, $W_{i,m}$ is not restricted to be multinomial, $\mathbb{E}[W_{i,m}] = 1$.)

Consistency

Definition 1.12 (Consistency of Estimator)

The estimator $v(P_m)(t)$ is **consistent** if

$$\sup_t |v(P_m)(t) - v(P_0)(t)| = \underbrace{o_{P_0}(1)}_{\text{Goes to zero in probability}} \tag{*}$$

Bootstrap Confidence Intervals

Definition 1.13 (τ-th quantile)

Let $q_\tau(v(P))$ be the τ-th quantile of $v(P)$:

$$q_\tau(v(P)) = v(P)^{-1}(\tau),\ \tau \in (0,1)$$

"Ideal" Confidence Interval: Suppose you know $v(P_0)$, the ideal interval is

$$CI_\alpha^0 := \left[\theta(P_m) - q_{1-\frac{\alpha}{2}}(v(P_0)), \theta(P_m) - q_{\frac{\alpha}{2}}(v(P_0))\right]$$

The confidence interval of the Bootstrap estimator is given by

$$CI_\alpha^{\text{Bootstrap}} := \left[\theta(P_m) - q_{1-\frac{\alpha}{2}}(v(P_m)), \theta(P_m) - q_{\frac{\alpha}{2}}(v(P_m))\right]$$

Theorem 1.4

Assuming the consistency of the Bootstrap estimator, the confidence interval of it satisfies

$$\Pr_{P_0}\left(CI_\alpha^{\text{Bootstrap}} \ni \theta(P_0)\right) \geq 1 - \alpha + o_{P_0}(1)$$

Proof 1.1

By (*), we have

$$q_\tau(v(P_m)) = q_\tau(v(P_0)) + o_{P_0}(1)$$

Then,

$$\begin{aligned}
\Pr_{P_0}\left(CI_\alpha^{\text{Bootstrap}} \ni \theta(P_0)\right) &= \Pr_{P_0}\left[\theta(P_m) - q_{1-\frac{\alpha}{2}}(v(P_m)) \leq \theta(P_0) \leq \theta(P_m) - q_{\frac{\alpha}{2}}(v(P_m))\right] \\
&= \Pr_{P_0}\left[q_{1-\frac{\alpha}{2}}(v(P_m)) \geq C(P_m, P_0) \geq q_{\frac{\alpha}{2}}(v(P_m))\right] \\
&= v(P_0)\left(q_{1-\frac{\alpha}{2}}(v(P_m))\right) - v(P_0)\left(q_{\frac{\alpha}{2}}(v(P_m))\right) \\
&= v(P_0)\left(q_{1-\frac{\alpha}{2}}(v(P_0))\right) - v(P_0)\left(q_{\frac{\alpha}{2}}(v(P_0))\right) + o_{P_0}(1) \\
&= 1 - \alpha + o_{P_0}(1)
\end{aligned}$$

The second last equality holds by (*) and continuity of the c.d.f. $v(P_0)$ (assumed).

Remark

(1). Choice of quantiles:

(a). If you impose symmetry at 0: $-q_{1-\frac{\alpha}{2}}(v(P)) = q_{\frac{\alpha}{2}}(v(P))$.

(2). P-values: the same idea of using confidence intervals. By the consistency and the continuity of the c.d.f. $v(P)$, the p-value converges to the true p-value.

(3). "Bootstrap" standard errors can't be used.

Definition 1.14 (Bootstrap standard error)

The object of interest is $\sqrt{\text{Var}_{P_0}(\theta(P_m))}$. The bootstrap standard error is given by

$$\text{BSE}(P_m) = \sqrt{\text{Var}_{P_m}(\theta(P_m^*))}$$

Application:

1. For $b \in \{1, ..., B\}$

   For $b \in \{1, ..., B\}$, generate $Z_1^*, \ldots, Z_m^*$ from P_m and forms P_m^*.

   Compute $\theta_b(P_m^*)$

2. $\text{BSE}(P_m) \approx \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\theta_b(P_m^*) - \frac{1}{B}\sum_{i=1}^{B}\theta_i(P_m)\right)^2}$.

e.g. the bootstrap standard error for $\theta(P) = \mathbb{E}_P[Z]$ is

$$\text{BSE}(P_m) = \sqrt{\text{Var}_{P_m}(\bar{Z}_m^*)} \quad = \sqrt{\mathbb{E}_{P_m}\left[(\bar{Z}_m^* - \mathbb{E}_{P_m}[\bar{Z}_m^*])^2\right]}$$

As $\mathbb{E}_{P_m}[\bar{Z}_m^*] = \mathbb{E}_{P_m}[Z^*] = \bar{Z}_m$, we have

$$\begin{aligned}
\text{BSE}(P_m) &= \sqrt{\mathbb{E}_{P_m}\left[\left(\frac{1}{m}\sum_{i=1}^{m}(Z_i^* - \bar{Z}_m)\right)^2\right]} \\
&= \sqrt{\frac{1}{m}\mathbb{E}_{P_m}\left[\left(Z^* - \bar{Z}\right)^2\right]} \\
&= m^{-\frac{1}{2}}\sqrt{m^{-1}\sum_{i=1}^{m}\left(Z_i - \bar{Z}_m\right)^2} \\
&= m^{-\frac{1}{2}}S_Z
\end{aligned}$$

Inconsistency

We use bootstrap to approximate $v(P_m)$. It works to approximate $v(P_0)$ iff

$$v(P_m) \xrightarrow{P_0} v(P_0)$$

which may don't work if

1. $P_m \xrightarrow{P_0} P_0$ doesn't hold.
2. v is not continuous at P_0.

Example 1.13

Parameter at the Boundary (Andrew, 2000, ECTA)

Suppose the parameter of the interest is $\theta(P_0) := \mathbb{E}_{P_0}[Z]$, and we know $\mathbb{E}_{P_0}[Z] \geq 0$.

Z is i.i.d.; The set of models is $\mathcal{P} = \{\mathcal{N}(\theta, 1) : \theta \geq 0\}$. The plug-in estimator is given by $\theta(P_m) := \max\{\bar{Z}_m, 0\}$.

$$\begin{aligned} v(P_0)(t) : &= \Pr_{P_0}\left(\sqrt{m}\left(\max\{\bar{Z}_m, 0\} - \mathbb{E}_{P_0}[Z]\right) \leq t\right) \\ &= \Pr_{P_0}\left(\max\{\sqrt{m}(\bar{Z} - \mathbb{E}_{P_0}[Z]), -\sqrt{m}\mathbb{E}_{P_0}[Z]\} \leq t\right) \\ &= \Pr_{P_0}\left(\max\{\mathcal{Z}, -\sqrt{m}\mathbb{E}_{P_0}[Z]\} \leq t\right) \end{aligned}$$

where $\mathcal{Z} \sim \mathcal{N}(0, 1)$.

(a). If $\mathbb{E}_{P_0}[Z] = 0$, $v(P_0)(t) = \Pr_{P_0}(\max\{\mathcal{Z}, 0\} \leq t)$

(b). If $\mathbb{E}_{P_0}[Z] > 0$, $v(P_0)(t) \xrightarrow{m\to\infty} \Pr_{P_0}(\mathcal{Z} \leq t)$

Consider $P_0 = \mathcal{N}\left(\frac{c}{\sqrt{m}}, 1\right)$, where $c > 0$. We have $\mathcal{N}\left(\frac{c}{\sqrt{m}}, 1\right) \to \mathcal{N}(0, 1)$. However, $v(P_0)(t) = \Pr_{P_0}(\max\{\mathcal{Z}, -c\} \leq t) \neq \Pr_{P_0}(\max\{\mathcal{Z}, 0\} \leq t)$.

The bootstrap estimator is given by

$$v(P_m)(t) = \Pr_{P_m}\left(\sqrt{m}\left(\max\{\frac{1}{m}\sum_{i=1}^{m} Z_i^*, 0\} - \max\{\bar{Z}_m, 0\}\right) \leq t\right)$$

Consider the path of $(Z_i)_{i=1}^{\infty}$ such that $\sqrt{m}\bar{Z}_m \leq -c, c > 0$. $\frac{1}{m}\sum_{i=1}^{m}\left(Z_i - \bar{Z}_m\right)^2 = 1$.

To prove the inconsistency, we want to show

$$v(P_m)(t) \geq \Pr\left(\max\{\mathcal{Z} - c, 0\} \leq t\right) > v(P_0)(t)$$

We have

$$v(P_m)(t) = \Pr_{P_m}\left(\max\{\underbrace{\frac{1}{\sqrt{m}}\sum_{i=1}^{m}(Z_i^* - \bar{Z}_m)}_{(A)} + \underbrace{\sqrt{m}\bar{Z}_m}_{(B)}, 0\} - \underbrace{\max\{\sqrt{m}\bar{Z}_m, 0\}}_{(C)} \leq t\right)$$

Since

(A). $\frac{1}{\sqrt{m}}\sum_{i=1}^{m}(Z_i^* - \bar{Z}_m) \to \mathcal{N}(0, 1)$ given the data $(Z_i)_{i=1}^{\infty}$.

(B). $\sqrt{m}\bar{Z}_m \leq -c$ based on the assumption.

(C). $\max\{\sqrt{m}\bar{Z}_m, 0\} \geq 0$.

Hence, $v(P_m)(t) \geq \Pr\left(\max\{\mathcal{Z} - c, 0\} \leq t\right) > v(P_0)(t)$.

Sub-Sampling / k-out-of m Bootstrap

Idea: We sample k (not m) observations.

- without replacement: Sub-Sampling
- with replacement: k-out-of-m Bootstrap

The bootstrap estimator is given by

$$v_k(P_m)(t) = \Pr_{P_m}\left(\sqrt{k}\,(\theta(P_k^*) - \theta(P_m)) \le t\right)$$

where P_k^* is the empirical probability using $Z_1^*, \ldots, Z_k^*$.

Suppose P_0 is know, the difference between the estimator and the true value is

$$\sup_t |v_k(P_m)(t) - v(P_0)(t)| \le \underbrace{\sup_t |v_k(P_m)(t) - v_k(P_0)(t)|}_{\text{“Sampling Error”}} + \underbrace{\sup_t |v_k(P_0)(t) - v(P_0)(t)|}_{\text{“Bias”}}$$

“Sampling Error” is small when k is small ($k << m$), while “Bias” is small when k is large ($k \approx m$).

For a $k(m)$ such that $k(m) \to \infty$ as $m \to \infty$, but $\frac{k(m)}{m} \to 0$. Intuition: consider the previous example 1.13

$$\begin{aligned} v_k(P_m)(t) &= \Pr_{P_m}\left(\sqrt{k}\left(\max\{\frac{1}{k}\sum_{i=1}^{k} Z_i^*, 0\} - \max\{\bar{Z}_m, 0\}\right) \le t\right) \\ &= \Pr_{P_m}\left(\max\{\underbrace{\frac{1}{\sqrt{k}}\sum_{i=1}^{k}(Z_i^* - \bar{Z}_m)}_{\to\mathcal{N}(0,1)} + \underbrace{\sqrt{k}\bar{Z}_m}_{\xrightarrow{P} 0 \text{ since } k<m}, 0\} - \underbrace{\max\{\sqrt{m}\bar{Z}_m, 0\}}_{\xrightarrow{P} 0 \text{ since } k<m} \le t\right) \end{aligned}$$

Theorem 1.5

The c.d.f. $v(P_0)(t) = \Pr_{P_0}(C(P_m, P_0) \le t)$ converges to $F(P_0)(t)$ if $F(P_0)$ is continuous. Then, the sub-sampling estimator is consistent.

1.4 Point Estimation

Suppose $X_1, \ldots, X_n$ is a random sample from a discrete/continuous distribution with pmf/pdf $f(\cdot \mid \theta)$, where $\theta \in \Theta$ is unknown.

Definition 1.15 (Point Estimator)

A **point estimator** (of θ) is a function of $(X_1, \ldots, X_n)$.

Notation: $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$.

Agenda

(1). Constructing point estimators

- Method of moments;

- Maximum likelihood.

(2). Comparing estimators

- Pairwise comparisons;
- Finding 'optimal' estimators.

1.4.1 Method of Moments (MM)

Definition 1.16 (Method of Moments in $\mathbb{R}^1$)

Suppose $\Theta \subseteq \mathbb{R}^1$. A **method of moments** estimator $\hat{\theta}_{\text{MM}}$ solves

$$\mu(\hat{\theta}_{\text{MM}}) = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

where $\mu : \Theta \to \mathbb{R}$ is given by

$$\mu(\theta) = \begin{cases} \sum_{x\in\mathbb{R}} xf(x \mid \theta), & \text{if } X_i \text{ are discrete} \\ \int_{-\infty}^{\infty} xf(x \mid \theta)dx, & \text{if } X_i \text{ are continuous} \end{cases}$$

Remark Existence of $\mu(\cdot)$ is assumed; Existence (and uniqueness) of $\hat{\theta}_{\text{MM}}$ is assumed.

Example 1.14

1. Suppose $X_i \sim$ i.i.d. Ber(p) where $p \in [0, 1]$ is unknown. The moment function is

$$\mu(p) = p$$

Then, the estimator is

$$\hat{p}_{\text{MM}} = \mu(\hat{p}_{\text{MM}}) = \bar{X}$$

Remark $\hat{p}_{\text{MM}} = \bar{X}$ is the 'best' estimator of p.

2. Suppose $X_i \sim$ i.i.d.$U(0, \theta)$ where $\theta > 0$ is unknown.

Remark Non-regular statistical model: parameter dependent support, where supp$X = [0, \theta]$.

The moment function is

$$\mu(\theta) = \frac{\theta}{2}$$

Then, the estimator is

$$\hat{\theta}_{\text{MM}} = 2\mu(\hat{\theta}_{\text{MM}}) = 2\bar{X}$$

Remark $\hat{\theta}_{\text{MM}}$ is not a very good estimator of θ. Concern $X_i > \hat{\theta}_{\text{MM}}$ could happen. So, $\max\{\hat{\theta}_{\text{MM}}, X_{(n)}\}$ can be better.

Definition 1.17 (Method of Moments in $\mathbb{R}^k$)

Suppose $\Theta \subseteq \mathbb{R}^k$. A **method of moments** estimator $\hat{\theta}_{\text{MM}}$ solves

$$\mu_j'(\hat{\theta}_{\text{MM}}) = \frac{1}{n}\sum_{i=1}^{n} X_i^j, \quad (j = 1, ..., k)$$

where $\mu_j' : \Theta \to \mathbb{R}$ is given by

$$\mu_j'(\theta) = \begin{cases} \sum_{x\in\mathbb{R}} x^j f(x \mid \theta), & \text{if } X_i \text{ are discrete} \\ \int_{-\infty}^{\infty} x^j f(x \mid \theta) dx, & \text{if } X_i \text{ are continuous} \end{cases}$$

Example 1.15

Suppose $X_i \sim$ i.i.d.$N(\mu, \sigma^2)$ where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown. The moment function is

$$\mu_1'(\mu, \sigma^2) = \mu$$

$$\mu_2'(\mu, \sigma^2) = \mu^2 + \sigma^2$$

Then, the estimator is

$$\mu_1'(\hat{\mu}_{\text{MM}}, \hat{\sigma}^2_{\text{MM}}) = \hat{\mu}_{\text{MM}} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\mu_2'(\hat{\mu}_{\text{MM}}, \hat{\sigma}^2_{\text{MM}}) = \hat{\mu}_{\text{MM}} + \hat{\sigma}^2_{\text{MM}} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

$$\Rightarrow \hat{\mu}_{\text{MM}} = \bar{X}$$

$$\hat{\sigma}^2_{\text{MM}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Remark $\bar{X}$ is the 'best' estimator of μ; An alternative better estimator of σ^2 is $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

1.4.2 Maximum Likelihood (ML)

Definition 1.18 (Maximum Likelihood)

A **maximum likelihood estimator** $\hat{\theta}_{\text{ML}}$ solves

$$L(\hat{\theta}_{\text{ML}} \mid X_1, ..., X_n) = \max_{\theta\in\Theta} L(\theta \mid X_1, ..., X_n)$$

where $L(\cdot \mid X_1, ..., X_n) : \Theta \to \mathbb{R}_+$ is given by

$$L(\theta \mid X_1, ..., X_n) = \prod_{i=1}^{n} f_{X_i}(X_i \mid \theta),\ \theta \in \Theta$$

Remark $L(\cdot \mid X_1, ..., X_n)$ is called the likelihood function.

Definition 1.19 (Log-Likelihood)

The **log-likelihood** function is

$$l(\theta \mid X_1, ..., X_n) = \log L(\theta \mid X_1, ..., X_n) = \sum_{i=1}^{n} \log f_{X_i}(X_i \mid \theta),\ \theta \in \Theta$$

Example 1.16

1. Suppose $X_i \sim$ i.i.d. $\mathrm{Ber}(p)$ where $p \in [0,1]$ is unknown. The marginal pmf is

$$f(x \mid p) = \begin{cases} p, & x = 1 \\ 1-p, & x = 0 \\ 0, & \text{otherwise} \end{cases} = p^x(1-p)^{1-x}\mathbf{1}_{\{x\in\{0,1\}\}}$$

Then, the likelihood function is

$$\begin{aligned} L(p \mid X_1, ..., X_n) &= \prod_{i=1}^{n}\left\{p^{X_i}(1-p)^{1-X_i}\underbrace{\mathbf{1}_{\{X_i\in\{0,1\}\}}}_{=1}\right\} \\ &= p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i},\ p \in [0,1] \end{aligned}$$

and the log-likelihood function is

$$l(p \mid X_1, ..., X_n) = (\sum_{i=1}^{n} X_i)\log p + (n - \sum_{i=1}^{n} X_i)\log(1-p),\ p \in (0,1)$$

Maximization:

(a). Suppose $0 < \sum_{i=1}^n X_i < n$, we can give the first-order condition:

$$\begin{aligned} \frac{\partial l(p \mid X_1, ..., X_n)}{\partial p}\Big|_{p=\hat{p}_{\mathrm{ML}}} &= \frac{\sum_{i=1}^n X_i}{\hat{p}_{\mathrm{ML}}} - \frac{n-\sum_{i=1}^n X_i}{n-\hat{p}_{\mathrm{ML}}} = 0 \\ \Rightarrow \hat{p}_{\mathrm{ML}} &= \frac{\sum_{i=1}^n X_i}{n} = \bar{X} \end{aligned}$$

(b). Suppose $\sum_{i=1}^n X_i = 0$, then

$$l(p \mid X_1, ..., X_n) = n\log(1-p),\ p \in [0,1) \Rightarrow \hat{p}_{\mathrm{ML}} = 0$$

(c). Suppose $\sum_{i=1}^n X_i = n$, then

$$l(p \mid X_1, ..., X_n) = n\log p,\ p \in (0,1] \Rightarrow \hat{p}_{\mathrm{ML}} = 1$$

All in all,

$$\hat{p}_{\mathrm{ML}} = \bar{X}$$

Remark $\hat{p}_{\mathrm{ML}} = \bar{X} = \hat{p}_{\mathrm{MM}}$ is the 'best' estimator of p.

2. Suppose $X_i \sim$ i.i.d. $U[0, \theta]$ where $\theta > 0$ is unknown. The marginal pdf is

$$f(x \mid \theta) = \begin{cases} \frac{1}{\theta}, & x \in [0, \theta] \\ 0, & \text{otherwise} \end{cases} = \frac{1}{\theta}\mathbf{1}_{\{x \in [0,\theta]\}}$$

and the likelihood function is

$$L(\theta \mid X_1, ..., X_n) = \prod_{i=1}^{n} \left\{\frac{1}{\theta}\mathbf{1}_{\{x \in [0,\theta]\}}\right\} = \begin{cases} \frac{1}{\theta^n}, & \theta \geq X_{(n)} \\ 0, & \text{otherwise} \end{cases}$$

$$\Rightarrow \hat{\theta}_{\text{ML}} = X_{(n)}$$

Remark $\hat{\theta}_{\text{ML}} = X_{(n)} \neq 2\bar{X} = \hat{\theta}_{\text{MM}}$; $\hat{\theta}_{\text{ML}} < X_i$ can't occur, which is good news; $\hat{\theta}_{\text{ML}} \leq \theta$ (low) must occur, which is bad news.

3. Suppose $X_i \sim$ i.i.d.$N(\mu, \sigma^2)$ where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown. Then,

$$\hat{\mu}_{\text{ML}} = \hat{\mu}_{\text{MM}} = \bar{X},\ \hat{\sigma}^2_{\text{ML}} = \hat{\sigma}^2_{\text{MM}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

1.5 Comparing Estimators: Mean Squared Error

1.5.1 Mean Squared Error = Bias^2 + Variance

General Approach

- Statistical Decision Theory

  Leading Special Case: Mean Squared Error.

Definition 1.20 (Mean Squared Error)

The **mean squared error** (MSE) of one estimator $\hat{\theta}$ of θ is defined as

$$\text{MSE}_\theta(\hat{\theta}) = \mathbb{E}_\theta[(\hat{\theta} - \theta)^2],\ \theta \in \Theta \subseteq \mathbb{R}$$

Definition 1.21 (Bias)

The **bias** of $\hat{\theta}$ is (the function of θ) given by

$$\text{Bias}_\theta(\hat{\theta}) = \mathbb{E}_\theta(\hat{\theta}) - \theta,\ \theta \in \Theta$$

$\hat{\theta}$ is **unbiased** iff $\text{Bias}_\theta(\hat{\theta}) = 0$ $(\forall \theta \in \Theta)$

Decomposition:

$$\text{MSE}_\theta(\hat{\theta}) = \text{Bias}_\theta(\hat{\theta})^2 + \text{Var}_\theta(\hat{\theta})$$

which is given by $\mathbb{E}[X^2] = \mathbb{E}[X]^2 + \text{Var}(X)$. Hence, if $\hat{\theta}$ is unbiased ($\text{Bias}_\theta(\hat{\theta}) = 0$), $\text{MSE}_\theta(\hat{\theta}) = \text{Var}_\theta(\hat{\theta})$.

1.5.2 Uniform Minimum Variance Unbiased (UMVU)

Definition 1.22 (Uniform Minimum Variance Unbiased (UMVU))

An unbiased estimator $\hat{\theta}$ is a **uniform minimum variance unbiased (UMVU)** estimator (of θ) iff

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathrm{Var}_\theta(\hat{\theta}) \leq \mathrm{Var}_\theta(\tilde{\theta}) = \mathrm{MSE}_\theta(\tilde{\theta})$$

whenever $\tilde{\theta}$ is an unbiased estimator of θ.

Remark UMVU estimators often exist; UMVU estimators are based on sufficient statistics.

1.6 Sufficient Statistics

1.6.1 Sufficient Statistic: contains all information of θ

Definition 1.23 (Sufficient Statistic)

A statistic $T = T(X_1, ..., X_n)$ is **sufficient** iff the conditional distribution of $(X_1, ..., X_n)$ given T, $(X_1, ..., X_n)|T$, doesn't depend on θ.

$$f_X(x \mid T(X_1, ..., X_n) = t; \theta) = f_X(x \mid T(X_1, ..., X_n) = t), \ \forall x$$

That is, the mutual information between θ and $T(X_1, ..., X_n)$ equals the mutual information between θ and $\{X_1, ..., X_n\}$,

$$\mathcal{I}(\theta; T(X_1, ..., X_n)) = \mathcal{I}(\theta; \{X_1, ..., X_n\})$$

1.6.2 Rao-Blackwell Theorem

Theorem 1.6 (Rao-Blackwell Theorem)

Suppose $\tilde{\theta}$ is an unbiased estimator of θ and suppose T is sufficient (for θ). Then,

(a). $\hat{\theta} = \mathbb{E}[\tilde{\theta}|T]$ is an unbiased estimator of θ.

(b). $\mathrm{Var}_\theta(\hat{\theta}) \leq \mathrm{Var}_\theta(\tilde{\theta})$, $\forall \theta \in \Theta$.

Proof 1.2

(a). Estimator: $\hat{\theta} = \mathbb{E}[\tilde{\theta} \mid T]$ doesn't depend on θ because T is sufficient. By the Law of Iterative Expectation, we have

$$\mathbb{E}_\theta(\hat{\theta}) = \mathbb{E}_\theta[\mathbb{E}[\tilde{\theta} \mid T]] = \mathbb{E}_\theta[\tilde{\theta}] = \theta$$

(b). Variance Reduction: By the Law of Total Variance

$$\mathrm{Var}(\hat{\theta}) = \mathrm{Var}_\theta[\mathbb{E}[\tilde{\theta} \mid T]] \leq \mathrm{Var}_\theta(\tilde{\theta}),\ \forall \theta \in \Theta$$

with strict inequality unless $\mathrm{Var}(\hat{\theta}|T) = 0$ (which also makes $\hat{\theta} = \tilde{\theta}$).

$\hat{\theta} = \mathbb{E}[\tilde{\theta}|T]$ is based on more information than $\tilde{\theta}$, which gives lower variance.

1.6.3 Fisher-Neyman Factorization Theorem

Finding sufficient statistics

- Apply ”definition”;
- Apply factorization criterion.

Proposition 1.3 (Fisher-Neyman Factorization Criterion)

A statistic $T = T(X_1, ..., X_n)$ is sufficient if and only if $\exists g(\cdot|\cdot)$ and $h(\cdot)$ such that

$$\begin{aligned} f_X((X_1, ..., X_n) \mid \theta) &= \prod_{i=1}^n f(X_i \mid \theta) \\ &= g[T(X_1, ..., X_n)|\theta]h(X_1, ..., X_n) \end{aligned}$$

Example 1.17

1. Suppose $\{X_i\}_{i=1}^n$ be a random sample from $Poisson(\theta)$. Then, show $T(X_1, ..., X_n) = \sum_{i=1}^n X_i$ is a sufficient statistic.

   (a). **Prove by Definition:** The sum of independent Poisson random variables are Poisson random variable, so we have $T = \sum_{i=1}^n X_i \sim Pois(n\theta)$. Then the conditional distribution of $X_1, ..., X_n$ given T is

   $$f(X_1, ..., X_n \mid T) = \frac{\prod_{i=1}^n \frac{\theta^{X_i} e^{-\theta}}{X_i!}}{\frac{(n\theta)^T e^{-n\theta}}{T!}} = \frac{T!}{n^T \prod_{i=1}^n X_i!}$$

   which is independent of θ. So, $T(X_1, ..., X_n)$ is sufficient statistic by definition.

   (b). **Prove by Factorization Theorem:**

   $$\prod_{i=1}^n f(X_i \mid \theta) = \prod_{i=1}^n \frac{\theta^{X_i} e^{-\theta}}{X_i!} = \frac{\theta^{T(X_1,...,X_n)} e^{-n\theta}}{\prod_{i=1}^n X_i!} = g(T(X_1, ..., X_n) \mid \theta) h(X_1, ..., X_n)$$

   where $g(T(X_1, ..., X_n) \mid \theta) = \theta^{T(X_1,...,X_n)} e^{-n\theta}$ and $h(X_1, ..., X_n) = \frac{1}{\prod_{i=1}^n X_i!}$. Hence, $T(X_1, ..., X_n)$ is sufficient statistic by Fisher-Neyman Factorization Criterion.

(c). **Prove by Exponential Family:**

$$f(X \mid \theta)=\frac{\theta^X e^{-\theta}}{X!}=\frac{e^{-\theta+X\ln\theta}}{X!}$$

Hence, the distribution is a member of the exponential family, where $c(\theta)=1, h(X)=\frac{1}{X!}, w_1(\theta)=-\theta, w_2(\theta)=\ln\theta, t_1(X)=1, t_2(X)=X$. By theorem 1.9, $\sum_{i=1}^n X_i$ is sufficient because $\{w_1(\theta)=-\theta, w_2(\theta)=\ln\theta\}$ is non-empty.

2. Suppose $X_i \sim$ i.i.d. $U[0,\theta]$ where $\theta>0$ is unknown. The marginal pdf is

$$f(x \mid \theta)=\begin{cases}\frac{1}{\theta}, & x\in[0,\theta]\\ 0, & \text{otherwise}\end{cases}=\frac{1}{\theta}\mathbf{1}_{\{x\in[0,\theta]\}}$$

Factorization:

$$\prod_{i=1}^n f(X_i \mid \theta)=\underbrace{\frac{1}{\theta^n}\mathbf{1}_{\{X_{(n)}\le\theta\}}}_{g(X_{(n)}|\theta)}\underbrace{\mathbf{1}_{\{X_{(1)}\ge 0\}}}_{h(X_1,\ldots,X_n)}$$

Hence, we have shown that $X_{(n)}$ is sufficient $\Rightarrow \hat{\theta}_{\text{MM}}=2\bar{X}$ cannot be UMVU and $\hat{\theta}_{RB}=\mathbb{E}[\hat{\theta}_{\text{MM}}|X_{(n)}]$ is better.

1.6.4 Minimal Sufficient Statistic

Definition 1.24 (Minimal Sufficient Statistic)

A sufficient statistic $T(X_1,\ldots,X_n)$ is called a **minimal sufficient statistic** if, for any other sufficient statistic $T'(X_1,\ldots,X_n)$, $T(X_1,\ldots,X_n)$ is a function of $T'(X_1,\ldots,X_n)$.

Theorem 1.7 (Theorem to Check Minimal Sufficient Statistic)

Let $f(\vec{X})$ be the pmf or pdf of a sample $\vec{X}$. Suppose there exists a function $T(\vec{X})$ such that,

"for every sample points $\vec{X}$ and $\vec{Y}$, the ratio $\frac{f(\vec{X}|\theta)}{f(\vec{Y}|\theta)}$ is constant for any θ if and only if $T(\vec{X})=T(\vec{Y})$".

Then $T(\vec{X})$ is a **minimal sufficient statistic** for θ.

Example 1.18

Let $X_1,\ldots,X_n \sim$ i.i.d. $U[\theta-\frac{1}{2},\theta+\frac{1}{2}]$, with $\theta\in\mathbb{R}$ unknown.

By $f(X \mid \theta)=\mathbf{1}_{\{X\in[\theta-\frac{1}{2},\theta+\frac{1}{2}]\}}$, we have

$$\prod_{i=1}^n f(X_i \mid \theta)=\underbrace{\mathbf{1}_{\{X_{(1)}\ge\theta-\frac{1}{2}\}}\mathbf{1}_{\{X_{(n)}\le\theta+\frac{1}{2}\}}}_{g[T(X_1,\ldots,X_n)|\theta]}\underbrace{1}_{h(X_1,\ldots,X_n)}$$

By the Fisher-Neyman Factorization Criterion, $T(X_1, ..., X_n) = \{X_{(1)}, X_{(n)}\}$ is a sufficient statistic.

We can prove $T(X_1, ..., X_n) = \{X_{(1)}, X_{(n)}\}$ is a minimal sufficient statistic by proving "for every sample points $(X_1, ..., X_n)$ and $(Y_1, ..., Y_n)$, $\frac{f(X_1,...,X_n|\theta)}{f(Y_1,...,Y_n|\theta)}$ is constant as a function of θ if and only if $T(X_1, ..., X_n) = T(Y_1, ..., Y_n)$."

$$\frac{f(X_1, ..., X_n \mid \theta)}{f(Y_1, ..., Y_n \mid \theta)} = \frac{\mathbf{1}_{\{X_{(1)} \geq \theta - \frac{1}{2}\}} \mathbf{1}_{\{X_{(n)} \leq \theta + \frac{1}{2}\}}}{\mathbf{1}_{\{Y_{(1)} \geq \theta - \frac{1}{2}\}} \mathbf{1}_{\{Y_{(n)} \leq \theta + \frac{1}{2}\}}}$$

Hence, for every sample points $(X_1, ..., X_n)$ and $(Y_1, ..., Y_n)$, $\frac{f(X_1,...,X_n|\theta)}{f(Y_1,...,Y_n|\theta)}$ is constant for all θ if and only if $X_{(1)} = Y_{(1)}$ and $X_{(n)} = Y_{(n)}$. That is, $T(X_1, ..., X_n) = T(Y_1, ..., Y_n)$. Hence, $T(X_1, ..., X_n) = \{X_{(1)}, X_{(n)}\}$ is a **minimal sufficient statistic**.

Consider $g(T) = X_{(n)} - X_{(1)} - \frac{n-1}{n+1}$, it has $\mathbb{E}[g(T)] = 0$ but $P_\theta[g(T) = 0] < 1$. Hence, T is not a complete statistic by definition.

1.7 Complete Statistic

1.7.1 Complete Statistic

Suppose T is sufficient and then $\hat{\theta} = \hat{\theta}(T)$ is unbiased. Under what conditions (on T) is $\hat{\theta}$ UMVU?

Answers: If "only one" estimator based on T is unbiased. (T is complete.)

Definition 1.25 (Complete Statistic)

A statistic T is **complete** if and only if

$$P_\theta[g(T) = 0] = 1, \ \forall \theta \in \Theta$$

whenever $g(\cdot)$ is such that

$$\mathbb{E}_\theta[g(T)] = 0, \ \forall \theta \in \Theta$$

(whenever the mean is zero, it can only equal to zero).

Recall: A matrix $A_{m \times k}$ has rank k iff $Ax = 0 \Rightarrow x = 0$.

Theorem 1.8 (Lehmann-Scheffé Theorem)

If T is complete and if $\hat{\theta} = \hat{\theta}(T)$ and $\tilde{\theta} = \tilde{\theta}(T)$ are unbiased, then

$$\mathbb{E}_\theta[\hat{\theta} - \tilde{\theta}] = 0 \Rightarrow P(\hat{\theta} - \tilde{\theta} = 0) = P(\hat{\theta} = \tilde{\theta}) = 1$$

1.7.2 Unbiased $\hat{\theta}(T)$ with sufficient and complete T is UMVU

Implication:

Corollary 1.1 (Unbiased $\hat{\theta}(T)$ with sufficient and complete T is UMVU)

If T is sufficient and complete and if $\hat{\theta} = \hat{\theta}(T)$ is unbiased, then $\hat{\theta}$ is UMVU (let $\tilde{\theta}$ be an UMVU).

Example 1.19

Suppose $X_i \sim$ i.i.d. $U[0, \theta]$ where $\theta > 0$ is unknown.

Facts:

- $X_{(n)}$ is sufficient and complete $\Rightarrow$ Any unbiased estimator given $X_{(n)}$ is UMVU, e.g. $\hat{\theta}_{RB} = \mathbb{E}[\hat{\theta}_{\text{MM}} | X_{(n)}]$;
- $\mathbb{E}_\theta(X_{(n)}) = \frac{n}{n+1}\theta \Rightarrow$ unbiased $\frac{n+1}{n} X_{(n)}$ is UMVU ($= \hat{\theta}_{RB}$).

Remark The cdf of $X_{(n)}$ is

$$F_{X_{(n)}}(x \mid \theta) = F(x \mid \theta)^n = \begin{cases} 0, & \text{if } x < 0 \\ \left(\frac{x}{\theta}\right)^n & \text{if } 0 \leq x \leq \theta \\ 1, & \text{if } x > \theta \end{cases}$$

so $X_{(n)}$ is continuous with pdf

$$f_{X_{(n)}}(x \mid \theta) = \begin{cases} \frac{n}{\theta^n} x^{n-1} & \text{if } x \in [0, \theta] \\ 0, & \text{otherwise} \end{cases}$$

Hence, $\mathbb{E}_\theta X_{(n)} = \int_0^\theta \frac{n}{\theta^n} x^{n-1} x dx = \frac{n}{n+1}\theta$.

Verifying Completeness

- Apply definition:
  - Example: $\sum_{i=1}^n X_i$ is complete when $X_i \sim i.i.d.$ Ber(p) - compute rank of the matrix to check completeness
- Show that $\{f(\cdot|\theta) : \theta \in \Theta\}$ is on exponential family and apply theorem 1.9.

Theorem 1.9 (Sufficient and Complete Statistic for Exponential Family)

If the distribution is a member of the exponential family, that is,

$$f(x|\theta) = c(\theta)h(x)\text{exp}\left\{\sum_{j=1}^k w_j(\theta)t_j(x)\right\}$$

then

$$T = \left(\sum_{i=1}^n t_1(x_i), ..., \sum_{i=1}^n t_k(x_i)\right)$$

is sufficient and complete if $\{\{w_1(\theta), ..., w_k(\theta)\} : \theta \in \Theta\}$ contains an open set.

Example 1.20

Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$ for some $\mu \in \mathbb{R}$ and some $\sigma^2 > 0$. Then, $\theta = (\mu, \sigma^2)$ and $\Theta = \mathbb{R} \times \mathbb{R}_{++}$. The pdf can be written as

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\mu^2}{2\sigma^2}} e^{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2}$$

We can have $h(x) = 1, c(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\mu^2}{2\sigma^2}}, t_1(x) = x, w_1(\mu, \sigma^2) = \frac{\mu}{\sigma^2}, t_2(x) = x^2, w_2(\mu, \sigma^2) = -\frac{1}{2\sigma^2}$.

That is, $T = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is sufficient and complete.

And $\left(\bar{X}, S^2\right) = \left(\frac{1}{n}\sum_{i=1}^n X_i, \frac{1}{n-1}\sum_{i=1}^n \left[X_i^2 - \frac{(\sum_{i=1}^n X_i^2)^2}{n}\right]\right)$ is UMVU estimator of (μ, σ^2).

1.8 Fisher Information

1.8.1 Score Function

The score function is the derivative of the log likelihood function with respect to θ.

Definition 1.26 (Score Function)

The **score function** is

$$u(\theta, \vec{X}) = \frac{\partial}{\partial\theta} \log f_{\vec{X}}(\vec{X} \mid \theta)$$

where $f_{\vec{X}}(\vec{X} \mid \theta) = L(\theta \mid X_1, ..., X_n) = \prod_{i=1}^n f_{X_i}(X_i \mid \theta)$.

Definition 1.27 (``Regularity" Condition)

The regularity conditions are as follows:

1. The partial derivative of $f_{\vec{X}}(\vec{X} \mid \theta)$ with respect to θ exists almost everywhere. (It can fail to exist on a null set, as long as this set does not depend on θ.)
2. The integral of $f_{\vec{X}}(\vec{X} \mid \theta)$ can be differentiated under the integral sign with respect to θ.
3. The support of $f_{\vec{X}}(\vec{X} \mid \theta)$ does not depend on θ.

Lemma 1.1 (``Regularity" Condition $\Rightarrow$ Mean of Score Function is Zero)

Under ``Regularity" condition and X are continuous, the mean of score function, evaluated at the true parameter θ_0, is zero:

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left[u(\theta_0, \vec{X})\right] &= \int_{\vec{X}} \left[\frac{\partial}{\partial\theta}\log f_{\vec{X}}(\vec{X} \mid \theta_0)\right] f_{\vec{X}}(\vec{X} \mid \theta_0)d\vec{X} \\
&= \int_{\vec{X}} \left[\frac{\partial}{\partial\theta} f_{\vec{X}}(\vec{X} \mid \theta_0)\right] d\vec{X} \\
(*) \qquad &= \frac{\partial}{\partial\theta}\underbrace{\int_{\vec{X}} f_{\vec{X}}(\vec{X} \mid \theta_0)d\vec{X}}_{=1} = 0
\end{aligned}$$

(*): Moving the derivative outside the integral can be done as long as the limits of integration are fixed, i.e. they do not depend on θ.

1.8.2 Fisher Information

Definition 1.28 (Fisher Information)

The **Fisher information** is defined to be the variance of the score function at θ_0.

$$\mathcal{I}(\theta_0) = \mathbb{E}_{\theta_0}[u(\theta_0, \vec{X})u(\theta_0, \vec{X})^T] = \mathbb{E}_{\theta_0}\left[\left(\frac{\partial}{\partial\theta}\log f_{\vec{X}}(\vec{X} \mid \theta_0)\right)^2\right]$$

Lemma 1.2 (Fisher Information with ``Regularity" Condition)

Under ``regularity" conditions, the **Fisher information** at θ_0 can also be written as

$$\mathcal{I}(\theta_0) = \text{Var}_{\theta_0}(u(\theta, \vec{X}))$$

Lemma 1.3 (Second Information Equality)

Under ``Regularity" condition, the Fisher information is equal to the minus Hessian matrix,

$$\mathcal{I}(\theta_0) = -\mathbb{E}_{\theta}\left[\frac{\partial^2}{\partial\theta^2}\log f_{\vec{X}}(\vec{X} \mid \theta_0)\right]$$

Proof 1.3

$$\begin{aligned}
\frac{\partial^2}{\partial\theta^2}\log f_{\vec{X}}(\vec{X} \mid \theta) &= \frac{\frac{\partial^2}{\partial\theta^2} f_{\vec{X}}(\vec{X} \mid \theta)}{f_{\vec{X}}(\vec{X} \mid \theta)} - \left(\frac{\frac{\partial}{\partial\theta} f_{\vec{X}}(\vec{X} \mid \theta)}{f_{\vec{X}}(\vec{X} \mid \theta)}\right)^2 \\
&= \frac{\frac{\partial^2}{\partial\theta^2} f_{\vec{X}}(\vec{X} \mid \theta)}{f_{\vec{X}}(\vec{X} \mid \theta)} - \left(\frac{\partial}{\partial\theta}\log f_{\vec{X}}(\vec{X} \mid \theta)\right)^2
\end{aligned}$$

where

$$\mathbb{E}_{\theta}\left[\frac{\frac{\partial^2}{\partial\theta^2} f_{\vec{X}}(\vec{X} \mid \theta)}{f_{\vec{X}}(\vec{X} \mid \theta)} \mid \theta\right] = \frac{\partial^2}{\partial\theta^2}\int_x f_{\vec{X}}(\vec{X} \mid \theta)dx = 0$$

1.8.3 Cramér-Rao Lower Bound

Proposition 1.4 (Cramér-Rao Lower Bound)

Under ``regularity'' conditions, for every estimator $\hat{\theta}$

$$\mathrm{Var}_\theta[\hat{\theta}(\vec{X})] \geq \frac{\left(\frac{d}{d\theta}\mathbb{E}_\theta[\hat{\theta}(\vec{X})]\right)^2}{\mathcal{I}(\theta)} \equiv \mathrm{CRLB}(\theta)$$

Specifically, if the estimator $\hat{\theta}$ is unbiased,

$$\mathrm{CRLB}(\theta) = \mathcal{I}(\theta)^{-1}$$

Remark $\mathcal{I}(\theta)$ is called the **Fisher Information**; "Regularity" conditions are satisfied by "smooth" exponential families; Proof uses Cauchy-Schwarz inequality.

3 Possibilities

(1). CR inequality is applicable and attainable:

- (a). Estimating p when $X \sim$ i.i.d. $\mathrm{Ber}(p)$;
- (b). Estimating μ when $X \sim$ i.i.d. $N(\mu, \sigma^2)$.

(2). CR inequality is applicable, but not attainable:

- (a). Estimating σ^2 when $X \sim$ i.i.d. $N(\mu, \sigma^2)$: $\mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n} = \mathcal{I}(\theta)^{-1}$ (CR bound).

(3). CR inequality is not applicable:

- (a). Estimating θ when $X \sim$ i.i.d. $U[0, \theta]$: CR bound $\mathcal{I}(\theta)^{-1} = \frac{\theta^2}{n}$ and $\mathrm{Var}(\hat{\theta}_{UMVU}) = \frac{\theta^2}{n(n+2)}$

Theorem 1.10 (MLE Covariance $\xrightarrow{n\to\infty}$ Cramér-Rao Lower Bound)

Suppose the sample $\{X_i\}_{i=1}^n$ is i.i.d. The Maximum likelihood estimator (MLE) $\hat{\theta} = \arg\max_\theta L(\theta \mid X_1, ..., X_n)$, under ``regularity'' conditions, as $n \to \infty$

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \to N(0, \mathcal{I}(\theta)^{-1})$$

Proposition 1.5 (Approximation of MLE Covariance Matrix)

When the sample x is made up of i.i.d. observations, the covariance matrix of the maximum likelihood estimator $\hat{\theta}$ is approximately equal to the inverse of the information matrix.

$$\mathrm{Cov}(\hat{\theta}) \approx (\mathcal{I}(\theta))^{-1}$$

Hence, the covariance matrix can be estimated as $(\mathcal{I}(\hat{\theta}))^{-1}$. Similarly, *SE* is estimated by $\sqrt{(\mathcal{I}(\hat{\theta}))^{-1}}$.

1.9 Hypothesis Testing

$X_1, ..., X_n$ is a random sample from a discrete/continuous distribution with pmf/pdf $f(\cdot \mid \theta)$, where $\theta \in \Theta$ is unknown.

Ingredients of Hypothesis Test

(1). Formulation of Testing Problem:

- Partioning of Θ into two disjoint subsets Θ_0 and Θ_1.

(2). Testing Procedure:

- Rule for choosing the two subsets specified in (1).

1.9.1 Formulation of Testing Problem

Formulating a Testing Procedure

- Terminology:

Definition 1.29 (Hypothesis)

(a). A hypothesis is a statement about θ;

(b). Null hypothesis: $H_0 : \theta \in \Theta_0$;

(c). Alternative hypothesis: $H_1 : \theta \in \Theta_1 = \Theta \backslash \Theta_0$;

(d). Maintained hypothesis: $\theta \in \Theta$ (always correct).

(e). *Typical Formulation*:

$$H_0 : \theta \in \Theta_0 \text{ vs. } H_1 : \theta \in \Theta_1$$

Example 1.21

Suppose $X \sim$ i.i.d. $N(\mu, 1)$, where $\mu \geq 0$ is unknown.

Objective: Determine whether $\mu = 0$.

Two possible formulation: $H_0 : \mu = 0$ vs. $H_1 : \mu > 0$ (or vice versa).

- Testing Procedure:

Consider the problem of testing $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$.

Definition 1.30 (Testing Procedure with Critical Region)

A testing procedure is a (data-based) rule for choosing between H_0 and H_1.

The rule:

$$\text{"Reject } H_0 \text{ iff } (X_1, ..., X_n) \in C\text{" (for some } C \in \mathbb{R}^n)$$

is a testing procedure with critical region C.

Example 1.22

Suppose $X \sim$ i.i.d. $N(\mu, 1)$, where $\mu \geq 0$ is unknown. The decision rule "Reject H_0 iff $\frac{\sum_{i=1}^{n} X_i}{n} = \bar{X} \geq \frac{1.645}{\sqrt{n}}$", where the critical region is $C = \{(X_1, ..., X_n) : \frac{\sum_{i=1}^{n} X_i}{n} \geq \frac{1.645}{\sqrt{n}}\}$

Proposition 1.6 (Critical Region ⇔ Test Statistic and Critical Value)

Any set $C \in \mathbb{R}^n$ can be written as

$$C = \{(X_1, ..., X_n) : T(X_1, ..., X_n) > c\}$$

for some $T : \mathbb{R}^n \to \mathbb{R}$ and some $c \in \mathbb{R}$.

Definition 1.31 (Test Statistic and Critical Value)

$T(X_1, ..., X_n)$ is called a test statistic and c is called the critical value (of the test).

1.9.2 Errors, Power Function, and Agenda

Agenda

1. Choosing critical value (given test statistic).
2. Choosing test statistic.

Definition 1.32 (Type I and Type II Errors)

Decision vs. Truth	H_0 (True)	H_1 (False)
H_0 (Fail to Reject)		Type II Error
H_1 (Reject)	Type I Error	

where

1. Type I Error: mistaken rejection of a null hypothesis that is actually true;
2. Type II Error: failure to reject a null hypothesis that is actually false.

There is a trade-off between Type I and Type II errors. The general approach is *statistical decision theory*.

Example 1.23

Heading Special Case: Making P_θ[Type I Error] "small".

Definition 1.33 (Power Function)

The **power function** of a test unit critical region $C \subseteq \mathbb{R}^n$ is the function $\beta : \Theta \to [0,1]$ given by

$$\begin{aligned} \beta(\theta) &= P_\theta[\text{Reject } H_0] \\ &= P_\theta[(X_1, ..., X_n)' \in C] \\ \text{(equivalently)} &= P_\theta[T(X_1, ..., X_n) > c] \end{aligned}$$

for corresponding statistic T and critical value c.

- For $\theta \in \Theta_0$: $P_\theta[\text{Type I Error}] = P_\theta[\text{Reject } H_0] = \beta(\theta)$;
- For $\theta \in \Theta_1$: $P_\theta[\text{Type II Error}] = 1 - P_\theta[\text{Reject } H_0] = 1 - \beta(\theta)$;
- Hence, the ideal power function is $\beta(\theta) = \begin{cases} 1, & \theta \in \Theta_1 \\ 0, & \theta \in \Theta_0 \end{cases}$;
- "Good" Power Function: $\beta(\theta)$ is "low" ("high") when $\theta \in \Theta_0$ ($\theta \in \Theta_1$).

Standard:

(1). Given $T(\cdot)$, choose critical value c such that $\beta(\theta) = P_\theta[T(X_1, ..., X_n) > c] \leq 5\%$ when $\theta \in \Theta_0$ (i.e., $\sup_{\theta \in \Theta_0} \beta(\theta) \leq 5\%$);

(2). Choose test statistic such that $\beta(\theta) = P_\theta[T(X_1, ..., X_n) > c(T)]$ is "large" for $\theta \in \Theta_1$. (Main Tool: Neyman-Pearson Lemma).

1.9.3 Choice of Critical Value

Given $T(\cdot)$, choose critical value c such that $\beta(\theta) = P_\theta[T(X_1, ..., X_n) > c] \leq 5\%$ when $\theta \in \Theta_0$ (i.e., $\sup_{\theta \in \Theta_0} \beta(\theta) \leq 5\%$).

Definition 1.34 (Test Size and Level α)

The **size** of a test (with power function β) is $\sup_{\theta \in \Theta_0} \beta(\theta)$.

A test is of **level** α **(**$\in [0,1]$**)** if and only if its size is $\leq \alpha$. (Standard choice $\alpha = 0.05$).

Example 1.24

Suppose $X \sim$ i.i.d. $N(\mu, 1)$, where $\mu \geq 0$ is unknown.

Consider the decision rule "Reject H_0 iff $\frac{\sum_{i=1}^n X_i}{n} = \bar{X} \geq \frac{1.645}{\sqrt{n}}$". The power function is $\beta(\mu) = P_\mu[\text{Reject } H_0] = P_\mu(\bar{X} \geq \frac{1.645}{\sqrt{n}})$

Recall: $\bar{X} \sim \mathcal{N}(\mu, \frac{1}{n}) \Rightarrow \sqrt{n}(\bar{X} - \mu) \sim \mathcal{N}(0, 1)$.

$$\begin{aligned}\beta(\mu) = P_\mu[\text{Reject } H_0] &= P_\mu(\bar{X} \geq \frac{1.645}{\sqrt{n}}) \\ &= P_\mu(\sqrt{n}(\bar{X} - \mu) \geq 1.645 - \sqrt{n}\mu) \\ &= 1 - \Phi(1.645 - \sqrt{n}\mu)\end{aligned}$$

where Φ is the standard normal cdf.

Size$= \beta(0) = 1 - \Phi(1.645) \approx 0.05$.

1.9.4 Choice of Test Statistic: Uniformly Most Powerful (UMP) Level α Test

Choose test statistic such that $\beta(\theta) = P_\theta[T(X_1, ..., X_n) > c(T)]$ is "large" for $\theta \in \Theta_1$. (Main Tool: Neyman-Pearson Lemma).

Definition 1.35 (Uniformly Most Powerful (UMP) Level α Test)

A test with level α and power function β is a **uniformly most powerful (UMP) level α test** iff

$$\beta(\theta) \geq \tilde{\beta}(\theta), \ \forall \theta \in \Theta_1$$

where $\tilde{\beta}$ is the power function of some (other) level α test.

Consider the problem of testing $H_0 : \theta = \theta_0 \in \mathbb{R}$

- UMP level α test always $\exists$ if $H_1 : \theta = \theta_1$ (Proven by Neyman-Pearson Lemma);
- UMP level α test often $\exists$ if $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$ (Proven by Karlin-Rubin Theorem);
- UMP level α test often $\nexists$ if $H_1 : \theta \neq \theta_0$; UMP "unbiased" level α test often $\exists$.

Theorem 1.11 (Neyman-Pearson Lemma)

Consider the problem of testing,

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta = \theta_1$$

For any $k \geq 0$, the test which

$$\text{Rejects } H_0 \text{ iff } L(\theta_1 \mid X_1, ..., X_n) \geq kL(\theta_0 \mid X_1, ..., X_n)$$

is a UMP level α test, where

$$\alpha = P_{\theta_0}[L(\theta_1 \mid X_1, ..., X_n) \geq kL(\theta_0 \mid X_1, ..., X_n)]$$

and where $L(\theta \mid X_1, ..., X_n) = \prod_{i=1}^n f(X_i \mid \theta)$.

Remark

- UMP level α test exists if $\alpha \in \{P_{\theta_0}[L(\theta_1 \mid X_1, ..., X_n) \geq kL(\theta_0 \mid X_1, ..., X_n)] : k \geq 0\}$.

- The Neyman-Pearson Lemma rejects the H_0 iff

$$L(\theta_1 \mid X_1, ..., X_n) \geq kL(\theta_0 \mid X_1, ..., X_n) \Leftrightarrow \frac{L(\theta_1 \mid X_1, ..., X_n)}{L(\theta_0 \mid X_1, ..., X_n)} \geq k$$

  $(L(\theta_0 \mid X_1, ..., X_n) \neq 0)$

- Hence, it is called **"Likelihood Ratio" test**.
- Converse: Any UMP level α test is of "NP type."

Example of Using NP Lemma

Example 1.25

Suppose $X \sim$ i.i.d. $N(\mu, 1)$, where $\mu \geq 0$ is unknown.

Let $\mu_1 = 0$ be given and consider the problem of testing

$$H_0 : \mu = 0 \text{ vs. } H_1 : \mu = \mu_1 > 0$$

We have $L(\mu \mid X_1, ..., X_n) = \prod_{i=1}^n \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i-\mu)^2}{2}} \right) = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^n X_i^2} e^{\mu \sum_{i=1}^n X_i} e^{-\frac{n\mu^2}{2}}$. Then,

$$\frac{L(\mu = \mu_1 \mid X_1, ..., X_n)}{L(\mu = 0 \mid X_1, ..., X_n)} = e^{\mu_1 \sum_{i=1}^n X_i} e^{-\frac{n\mu_1^2}{2}}$$

<u>Decision Rule:</u> Reject H_0 iff

$$\begin{aligned} \frac{L(\mu = \mu_1 \mid X_1, ..., X_n)}{L(\mu = 0 \mid X_1, ..., X_n)} &= e^{\mu_1 \sum_{i=1}^n X_i} e^{-\frac{n\mu_1^2}{2}} \geq k \\ \Leftrightarrow \quad & -\frac{n\mu_1^2}{2} + \mu_1 \sum_{i=1}^n X_i \geq \log k \\ \Leftrightarrow \quad & \bar{X} \geq \frac{\log k}{n\mu_1} + \frac{\mu_1}{2} \end{aligned}$$

The NP test reject for large values of $\bar{X}$.

Optimality Theorem for One-sided Testing Problem

Consider

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu > \mu_0$$

For any $\theta_1 > \theta_0$, use NP Lemma to find optimal test of $H_0 : \mu = \theta_0$ vs. $H_1 : \mu = \mu_1$.

- If the NP tests coincide, then the test is the UMP test of $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$;
- Otherwise, $\nexists$ UMP (level α) test of the $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$.

<u>**Implications:**</u> (The previous $N(\mu, 1)$ example)

(i). The UMP 5% test of $H_0 : \mu = 0$ vs. $H_1 : \mu > 0$ rejects H_0 iff $\bar{X} > \frac{1.645}{\sqrt{n}}$.

(ii). The UMP 5% test of $H_0 : \mu = 0$ vs. $H_1 : \mu < 0$ rejects H_0 iff $-\bar{X} > \frac{1.645}{\sqrt{n}}$.

(iii). $\nexists$ UMP 5% test of $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$.

Definition 1.36 (Unbiased Test)

A test of

$$H_0 : \theta \in \Theta_0 \text{ vs. } H_1 : \theta \in \Theta_1$$

is **unbiased** iff its power function $\beta(\cdot)$ satisfies $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \inf_{\theta \in \Theta_1} \beta(\theta)$

Claim 1.1

The UMP unbiased 5% test of $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$: Rejects H_0 iff $|\bar{X}| > \frac{1.96}{\sqrt{n}}$.

Corollary 1.2

Suppose $X_i \sim$ i.i.d. $N(\mu, \sigma^2)$, where σ^2 is known. Then, the UMP unbiased 5% test of the $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$: Rejects H_0 if $|\frac{\bar{X} - \mu_0}{\sigma}| > \frac{1.96}{\sqrt{n}}$.

Claim 1.2

"In general", "Natural" test statistics are (approximately) optimal and critical values can be find.

1.9.5 Generalized Neyman-Pearson Lemma

NP Lemma: $\max \beta(\theta_1)$ s.t. $\beta(\theta_0) \leq \alpha$;

Generalized NP Lemma: How to optimize a function with infinity constraints.

Observation: If β is differentiable, then an unbiased test of the $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ satisfies $\beta'(\theta_0) = 0$

Theorem 1.12 (Generalized Neyman-Pearson Lemma)

1.10 Trinity of Classical Tests

- Likelihood Ratio Test
- Lagrangian Multiplier Test (Score Test)
- Wald Test

Properties: Deliver optimal test in motivating example; closely related (and "approximately" optimal) in general.

1.10.1 Test Statistics

Settings: $X_1, ..., X_n$ is a random sample from a discrete/continuous distribution with pmf/pdf $f(\cdot \mid \epsilon)$, where $\theta \in \Theta \subseteq \mathbb{R}$ is unknown.

Testing Problem: $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ for some $\theta_0 \in \Theta$.

Recall the log likelihood function is given by

$$l(\theta \mid X_1, ..., X_n) = \sum_{i=1}^{n} \log f(X_i \mid \theta)$$

The (sample) score function is

$$u(\theta \mid X_1, ..., X_n) = \frac{\partial}{\partial\theta} l(\theta \mid X_1, ..., X_n)$$

and the (sample) fisher information is

$$\mathcal{I}(\theta \mid X_1, ..., X_n) = -\frac{\partial^2}{\partial\theta^2} l(\theta_0 \mid X_1, ..., X_n)$$

- **Likelihood Ratio Test Statistic**:

$$\begin{aligned} T_{LR}(X_1, ..., X_n) &= 2\left\{\max_{\theta\in\Theta} l(\theta \mid X_1, ..., X_n) - \max_{\theta\in\Theta_0} l(\theta \mid X_1, ..., X_n)\right\} \text{ (general form)} \\ &= 2\left\{l(\hat{\theta}_{\text{ML}} \mid X_1, ..., X_n) - l(\theta_0 \mid X_1, ..., X_n)\right\} \\ &= 2\log\left\{\frac{l(\hat{\theta}_{\text{ML}} \mid X_1, ..., X_n)}{l(\theta_0 \mid X_1, ..., X_n)}\right\} \end{aligned}$$

  Motivation: Neyman-Pearson Lemma (1.11)

- **Lagrangian Multiplier Test Statistic**:

$$T_{LM}(X_1, ..., X_n) = \frac{\left(\frac{\partial}{\partial\theta} l(\theta_0 \mid X_1, ..., X_n)\right)^2}{-\frac{\partial^2}{\partial\theta^2} l(\theta_0 \mid X_1, ..., X_n)} = \frac{(u(\theta_0 \mid X_1, ..., X_n))^2}{\mathcal{I}(\theta_0 \mid X_1, ..., X_n)}$$

  Motivation: T_{LM} is approximate to T_{LR}; No estimation required.

- **Wald Test Statistic**:

$$T_W(X_1, ..., X_n) = \frac{(\hat{\theta}_{\text{ML}} - \theta_0)^2}{\left\{-\frac{\partial^2}{\partial\theta^2} l(\hat{\theta}_{\text{ML}} \mid X_1, ..., X_n)\right\}^{-1}} = \frac{(\hat{\theta}_{\text{ML}} - \theta_0)^2}{\left(\mathcal{I}(\hat{\theta}_{\text{ML}} \mid X_1, ..., X_n)\right)^{-1}}$$

  Motivation: T_W is approximate to T_{LR};

  Generalization: Reject the $H_0 : \theta = \theta_0$ if $|\hat{\theta} - \theta_0|$ is "large", when $\hat{\theta}$ is some estimator of θ.

Claim 1.3

In general, for "large" n,

$$T_{LR} \approx T_{LM} \approx T_W \sim \chi^2(1) = N(0,1)^2 \text{ under } H_0 : \theta = \theta_0$$

- Approximate 5%critical value is $(1.96)^2 = 3.84$.
- $T_{LR} = T_{LM} = T_W \sim \chi^2(1) = N(0,1)^2$ under $H_0 : \theta = \theta_0$ when $X_i \sim$ i.i.d. $N(\mu, 1)$.

1.10.2 Approximation to T_{LR}

In this part as $n \to \infty$, we use $l(\theta), l'(\theta), l''(\theta)$ to denote $l(\theta \mid X_1, ..., X_n), l'(\theta \mid X_1, ..., X_n) \triangleq u(\theta \mid X_1, ..., X_n), l''(\theta \mid X_1, ..., X_n) \triangleq -\mathcal{I}(\theta \mid X_1, ..., X_n)$.

(1). T_{LM}:

Suppose

$$l(\theta) \approx l(\theta_0) + l'(\theta_0)(\theta - \theta_0) + \frac{1}{2}l''(\theta_0)(\theta - \theta_0)^2 \triangleq \tilde{l}(\theta)$$

Then

$$\hat{\theta}_{\text{ML}} = \underset{\theta}{\operatorname{argmax}}\, l(\theta) \approx \underset{\theta}{\operatorname{argmax}}\, \tilde{l}(\theta) = \theta_0 - \frac{l'(\theta_0)}{l''(\theta_0)} \triangleq \tilde{\theta}_{\text{ML}}$$

Hence,

$$T_{LR} = 2\left\{l(\hat{\theta}_{\text{ML}}) - l(\theta_0)\right\} \approx 2\left\{\tilde{l}(\tilde{\theta}_{\text{ML}}) - \tilde{l}(\theta_0)\right\} = -\frac{l'(\theta_0)^2}{l''(\theta_0)} = T_{LM}$$

(2). T_W:

Suppose

$$l(\theta) \approx l(\hat{\theta}_{\text{ML}}) + l'(\hat{\theta}_{\text{ML}})(\theta - \hat{\theta}_{\text{ML}}) + \frac{1}{2}l''(\hat{\theta}_{\text{ML}})(\theta - \hat{\theta}_{\text{ML}})^2 \triangleq \hat{l}(\theta)$$

Then,

$$T_{LR} = 2\left\{l(\hat{\theta}_{\text{ML}}) - l(\theta_0)\right\} \approx 2\left\{\tilde{l}(\hat{\theta}_{\text{ML}}) - \hat{l}(\theta_0)\right\} = \frac{(\hat{\theta}_{\text{ML}} - \theta_0)^2}{(-l''(\hat{\theta}_M L))^{-1}} = T_W$$

1.11 Interval Estimation

Definition 1.37

Suupose $\theta \in \mathbb{R}$.

1. An <u>interval estimator</u> of θ is an interval $[L(X_1, ..., X_n), U(X_1, ..., X_n)]$, where $L(X_1, ..., X_n)$ and $U(X_1, ..., X_n)$ are statistics.
2. The <u>converge probability</u> (of the interval estimator) is the function (of θ) given by

$$P_\theta\left[L(X_1, ..., X_n) \le \theta \le U(X_1, ..., X_n)\right]$$

3. The <u>confidence coefficient</u> is $\inf_\theta P_\theta\left[L(X_1, ..., X_n) \le \theta \le U(X_1, ..., X_n)\right]$

Example 1.26

Suppose $X_i \sim$ i.i.d. $N(\mu, 1)$, where μ is unknown.

<u>Interval estimator:</u> $\left[\bar{X} - \frac{1.96}{\sqrt{n}}, \bar{X} + \frac{1.96}{\sqrt{n}}\right]$.

<u>Converge probability:</u> $P_\mu\left[\bar{X} - \frac{1.96}{\sqrt{n}} \le \mu \le \bar{X} + \frac{1.96}{\sqrt{n}}\right] = P_\mu\left[-1.96 \le \sqrt{n}(\bar{X} - \mu) \le 1.96\right] = \Phi(1.96) - \Phi(-1.96) \approx 0.95$.

<u>Interpretation:</u>

(I). Recall

(i). $\bar{X} = \hat{\mu}_{\text{MM}} = \hat{\mu}_{\text{ML}} = \hat{\mu}_{UMVU}$;

(ii). $\bar{X} \sim \mathcal{N}(\mu, \frac{1}{n}) \Rightarrow \frac{1}{\sqrt{n}} = \sqrt{\text{Var}(\bar{x})}$.

Hence, $\left[\bar{X} - \frac{1.96}{\sqrt{n}}, \bar{X} + \frac{1.96}{\sqrt{n}}\right] = \left[\bar{X} - 1.96\sqrt{\text{Var}(\bar{x})}, \bar{X} + 1.96\sqrt{\text{Var}(\bar{x})}\right]$. $\frac{\bar{X}-\mu}{\text{Var}(\bar{x})} \sim \mathcal{N}(0,1)$.

(II). Recall: The "optimal" two-sided 5% of the $\mu = \mu_0$ rejects <u>iff</u> $|\bar{X} - \mu_0| > \frac{1.96}{\sqrt{n}}$

$\Leftrightarrow \bar{X} - \mu_0 > \frac{1.96}{\sqrt{n}}$ or $\bar{X} - \mu_0 < -\frac{1.96}{\sqrt{n}}$

$\Leftrightarrow \mu_0 < \bar{X} - \frac{1.96}{\sqrt{n}}$ or $\mu_0 > \bar{X} + \frac{1.96}{\sqrt{n}}$

Hence, the test "accepts" H_0 iff

$$\bar{X} - \frac{1.96}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + \frac{1.96}{\sqrt{n}}$$

Chapter 2 Decision Rule Based Statistical Inference

2.1 Decision Rule

Given an observation $x \in X$, we want to estimate an unknown state $\theta \in S$ (not necessarily random). The θ can form x with $P_\theta(x)$. We use decision rule $\delta(x)$ to form an action (estimation of θ) $a = \hat{\theta}$.

Example:

(1) Binary hypothesis testing (detection) when $S = \{0, 1\}$ e.g. $P_0 \sim \mathcal{N}(0, \sigma^2), P_1 \sim \mathcal{N}(\mu, \sigma^2)$

(2) Multiple hypothesis testing (classification) when $S = \{1, 2, ..., n\}$

(3) (Estimation) when $S = \mathbb{R}$ e.g. $P_\theta \in N(\theta, \sigma^2)$

Example 2.1 (Binary HT)

For the example Binary HT, $P_0 \sim \mathcal{N}(0, \sigma^2), P_1 \sim \mathcal{N}(\mu, \sigma^2)$: decision rule $\delta : \mathbb{R} \to \{0, 1\}$

We can find a τ such that $\delta(x) = \begin{cases} 1, & x \geq \tau \\ 0, & else \end{cases} = \mathbf{1}_{x \geq \tau}$. Howe to choose τ?

Type-I error probability: probability that θ is 0 but receive $\delta(x) = 1$.

$$P_I = P_0\{\delta(x) = 1\} = P_0\{x \geq \tau\} = Q\left(\frac{\tau}{\sigma}\right)$$

Type-II error probability: probability that θ is 1 but receive $\delta(x) = 0$.

$$P_{II} = P_1\{\delta(x) = 0\} = P_1(x < \tau) = Q(\frac{\mu - \tau}{\sigma})$$

Both P_I and P_{II} depends on τ. $Q(t) = \int_t^\infty \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx$

For $\tau = \frac{\mu}{2}$, $P_I = P_{II} = Q\left(\frac{\mu}{2\sigma}\right)$

Example 2.2 (Multiple HT)

Consider three state $S = \{1, 2, 3\}$. We can find a τ such that $\delta(x) = \begin{cases} 1, & x < \tau_1 \\ 2, & \tau_1 \leq x \leq \tau_2 \\ 3, & x > \tau_2 \end{cases} = \mathbf{1}_{x \geq \tau}$.

Conditional Error Probabilities: probability that θ is i but receive $\delta(x) = j$ (6 types in this example)

$$P_i\{\delta(x) = j\}, \forall i \neq j$$

2.2 Maximum-Likelihood Principle (state is norandom)

Maximum-Likelihood Principle

$$\hat{\theta} = \underset{\theta \in S}{\operatorname{argmax}}\, P_\theta(x) = \underset{\theta \in S}{\operatorname{argmax}} \ln P_\theta(x)$$

Applied to the binary example: $P_0 \sim \mathcal{N}(0, \sigma^2), P_1 \sim \mathcal{N}(\mu, \sigma^2)$.

$P_0(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}, P_1(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. $\ln P_0(x) = c - \frac{x^2}{2\sigma^2}, \ln P_1(x) = c - \frac{(x-\mu)^2}{2\sigma^2}$.

Then, the rule can become

$$\hat{\theta} = \begin{cases} 0, & x^2 < (x-\mu)^2 \\ 1, & else \end{cases} = \mathbf{1}_{x^2 \geq (x-\mu)^2} = \mathbf{1}_{x \geq \frac{\mu}{2}}$$

Vector Observations

Observations $X = (x_1, x_2, ..., x_n)$, where i.i.d. $x_i \sim P_\theta$. Then

$$P_\theta(X) = \prod_{i=1}^{n} P_\theta(x_i), \ \ln P_\theta(X) = \sum_{i=1}^{n} \ln P_\theta(x_i)$$

$\ln P_0(x) = cn - \frac{\sum_{i=1}^n x_i^2}{2\sigma^2}, \ln P_1(x) = cn - \frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}$.

Then, the rule can become

$$\hat{\theta} = \begin{cases} 0, & \sum_{i=1}^n x_i^2 < \sum_{i=1}^n (x_i-\mu)^2 \\ 1, & else \end{cases} = \mathbf{1}_{\sum_{i=1}^n x_i^2 \geq \sum_{i=1}^n (x_i-\mu)^2} = \mathbf{1}_{\bar{x} \geq \frac{\mu}{2}}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$. Under both H_0 and H_1, $\bar{x} \sim \mathcal{N}(0, \frac{\sigma^2}{n})$.

Then, type I error prob and type II error prob are the same

$$P_I = P_0\{\bar{x} \geq \frac{\mu}{2}\} = P_{II} = P_1\{\bar{x} < \frac{\mu}{2}\} = Q\left(\frac{\mu\sqrt{n}}{2\sigma}\right)$$

Estimation $S = \mathbb{R}$

To estimate θ when $S = \mathbb{R}$

$$\begin{aligned} &\max_{\theta \in \mathbb{R}} \sum_{i=1}^{n} \ln P_\theta(x_i) \\ \Leftrightarrow &\max_{\theta \in \mathbb{R}} \left[cn - \frac{\sum_{i=1}^n (x_i - \theta)^2}{2\sigma^2} \right] \\ \Leftrightarrow &\max_{\theta \in \mathbb{R}} \sum_{i=1}^{n} (x_i - \theta)^2 \Rightarrow \hat{\theta} = \bar{x} \end{aligned}$$

Then, with $\bar{x} \sim \mathcal{N}(\theta, \frac{\sigma^2}{n})$, the

$$MSE_\theta = \mathbb{E}_\theta \left(\bar{x} - \theta\right)^2 = \frac{\sigma^2}{n}$$

2.3 Bayesian Decision Rule (state is random)

2.3.1 Rules

Prior probability distribution of θ is $\pi(\theta)$.

Loss/cost function with action (estimation) a is $l(a, \theta)$. e.g.

1. (binary HT) Hamming/zero-one loss $l(a = \hat{\theta}, \theta) = \mathbf{1}_{a \neq \theta}$
2. (estimation) Squared error loss $l(a = \hat{\theta}, \theta) = (a - \theta)^2$; Absolute error loss $l(a, \theta) = |a - \theta|$.

Definition 2.1 (Risk)

Risk of decision rule δ on θ:

$$R(\delta, \theta) = \mathbb{E}_{X \sim \pi(\theta)} \left[l(\delta(X), \theta) \right]$$

where X are random with prob $P(\cdot|\theta)$.

Risk of decision rule δ:

$$\begin{aligned} R(\delta) &= \mathbb{E}_{\theta \sim P_\theta} [R(\delta, \theta)] \\ &= \mathbb{E}_{\theta \sim P_\theta} \mathbb{E}_{X \sim \pi(\theta)} \left[l(\delta(X), \theta) \right] \end{aligned}$$

where (X, θ) are random with joint probability distribution

$$P(X, \theta) = P(X)\pi(\theta \mid X)$$

Note *In machine learning, we normally use y to substitute θ.*

Example 2.3 (Hamming/zero-one Loss)

The risk of decision $\delta(x)$ in Hamming/zero-one loss $l(a = \hat{y}, y) = \mathbf{1}_{a \neq y}$

$$\begin{aligned} R(\delta) = \mathbb{E}(\mathbf{1}_{\delta(x) \neq y}) &= \mathbb{E}[\delta(x) \neq y] \\ &= P(y = 0)P[\delta(x) \neq 0 | y = 0] + P(y = 1)P[\delta(x) \neq 1 | y = 1] \\ &= P(y = 0)P[\delta(x) = 1 | y = 0] + P(y = 1)P[\delta(x) = 0 | y = 1] \end{aligned}$$

2.3.2 Optimization Problem in Bayes Form

We want to compute the optimal rule that minimizes the risk:

$$\delta_B = \underset{\delta}{\operatorname{argmin}}\, R(\delta)$$

Derive Bayes rule

$$\begin{aligned} R(\delta) &= \int_x \int_\theta P(x, \theta) l(\delta(x), \theta) d\theta dx \\ &= \int_x P(x) \int_\theta \pi(\theta | x) l(\delta(x), \theta) d\theta dx \end{aligned}$$

δ_B is given by solving the optimization problem:

$$\min_{\delta} \int_x P(x) \int_\theta \pi(\theta|x) l(\delta(x), \theta) d\theta dx$$

Hence,

Proposition 2.1

$\delta_B = \text{argmin}_\delta R(\delta)$ can be transformed into optimization problems for each $x \in S$

$$\min_{\delta(x)} \int_\theta \pi(\theta \mid x) l(\delta(x), \theta) d\theta$$

The problem becomes to compute $\pi(\theta|x)$, which is computed by

$$\pi(\theta \mid x) = \frac{\pi(\theta) P(x \mid \theta)}{P(x)}$$

Example 2.4 (Square Loss)

Consider a Bernoulli Distribution with parameter θ. We have data $\{Z_1, ..., Z_n\}$ and we want to predict the next sample Z.

Note that Z is what we want to predict, so the Sqaure error loss given estimation t is

$$\begin{aligned} R(t, Z; \theta) &= \mathbb{E}_\theta[(t - Z)^2] \\ &= (t - \theta)^2 + \theta(1 - \theta) \end{aligned}$$

Then, the optimal estimation is $t^* = \theta$. And we say the minimal risk is **oracle risk**

$$R(t^*, Z \mid \theta) = \theta(1 - \theta)$$

However, we don't have θ. What if we use the MLE $t(\{Z_1, ..., Z_n\}) = \hat{\theta}_{\text{MLE}} = \bar{Z} = \frac{1}{n} \sum_{i=1} n Z_i$ instead?

$$\begin{aligned} R(\hat{\theta}_{\text{MLE}}, Z \mid \theta) &= \mathbb{E}_\theta[(\bar{Z} - Z)^2] \\ &= \mathbb{E}_\theta[(\bar{Z} - \theta)^2] + \theta(1 - \theta) \\ &= \underbrace{\frac{\theta(1 - \theta)}{n}}_{\text{Sample Uncertainty}} + \underbrace{\theta(1 - \theta)}_{\text{Oracle Risk}} \end{aligned}$$

The average risk with prior belief $P(\theta) = 1_{\{\theta \in [0,1]\}}$ ($\theta \sim \text{Beta}(1, 1) = \text{Unif}[0, 1]$) is

$$R(t) = \int_\theta R(t, Z \mid \theta) \pi(\theta|\{Z_1, ..., Z_n\}) d\theta$$

$\pi(\theta|\{Z_1, ..., Z_n\})$ is the posterior beliefs about the θ:

$$\pi(\theta \mid \{Z_1, ..., Z_n\}) = \frac{f(\{Z_1, ..., Z_n\} \mid \theta) P(\theta)}{\int f(\{Z_1, ..., Z_n\} \mid \theta') P(\theta') d\theta'}$$

As $Z_i \sim \text{Bernoulli}(\theta)$, we have

$$\theta \mid \{Z_1, ..., Z_n\} \sim \text{Beta}\left(\sum_{i=1}^{n} Z_i + 1, n - \sum_{i=1}^{n} Z_i + 1\right)$$

which has mean $\mathbb{E}[\theta \mid \{Z_1, ..., Z_n\}] = \frac{\sum_{i=1}^{n} Z_i + 1}{n+2} = \frac{\hat{\theta}_{MLE} + \frac{1}{2}}{1 + \frac{2}{n}}$. Then,

$$\begin{aligned} R^*(t) &= \min_t \int_\theta R(t, Z \mid \theta)\pi(\theta|\{Z_1, ..., Z_n\})d\theta \\ &= \min_t \int_\theta (t-\theta)^2 \pi(\theta|\{Z_1, ..., Z_n\})d\theta + \int_\theta \theta(1-\theta)\pi(\theta|\{Z_1, ..., Z_n\})d\theta \\ \Rightarrow t^* &= \mathbb{E}[\theta \mid \{Z_1, ..., Z_n\}] \end{aligned}$$

2.3.3 Maximum A Posteriori (MAP) Decision Rule (Binary example)

Example 2.5

Hamming/zero-one loss $l(a, y) = \mathbf{1}_{a \neq y}$

Maximum A Posteriori (MAP) Decision Rule:

Optimization problem is

$$\delta(x) = \operatorname*{argmin}_a \sum_{y=0,1} \pi(y|x)\mathbf{1}_{a\neq y}dy = \operatorname*{argmax}_{y\in\{0,1\}} \pi(y|x)$$

$$\Rightarrow \sum_{y=0,1} \pi(y|x)\mathbf{1}_{\delta(x)\neq y}dx = \min_a \sum_{y=0,1} \pi(y|x)\mathbf{1}_{a\neq y}dy = \min\{\pi(1|x), \pi(0|x)\}$$

Likelihood ratio: $L(x) = \frac{P_1(x)}{P_0(x)}$

Likelihood ratio test: threshold $\tau = \frac{\pi(0)}{\pi(1)}$. If $L(x) > \tau$ accept H_1 (equivalent to $P_1(x)\pi(1) > P_0(x)\pi(0)$ which is also equivalent to comparing $\pi(y|x)$).

In this rule the whole optimization problem also goes to

$$\begin{aligned} R(\delta_{MAP}) &= \int_x P(x) \sum_{y=0,1} \pi(y|x)\mathbf{1}_{\delta(x)\neq y}dx \\ &= \int_x P(x) \min\{\pi(1|x), \pi(0|x)\}dx \end{aligned}$$

2.3.4 Minimum Mean Squared Error (MMSE) Rule ($\mathbb{R}^n$ example)

Example 2.6 (Estimation)

Squared error loss $l(a, y) = (a - y)^2$.

Minimum Mean Squared Error (MMSE) Rule:

Optimization problem is $\delta(x) = \text{argmin}_a \int_y \pi(y|x)(a-y)^2 dy$

$$0 = \int_y \pi(y|x)(\delta_B(x) - y)dy = \delta_B(x) - \mathbb{E}\left[Y|X=x\right]$$

$$\Rightarrow \delta_B(x) = \mathbb{E}\left[Y|X=x\right]$$

which is called **conditional mean estimation**.

In this rule the whole optimization problem also goes to

$$R(\delta_{MMSE}) = \int_x P(x) \int_y \pi(y|x)(y - \mathbb{E}\left[Y|X=x\right])^2 dydx = \mathbb{E}_X Var\left[Y|X=x\right]$$

Gaussian case: If $X \in \mathbb{R}^n$ and (Y, X) are jointly Gaussian, then the conditional mean is a linear function of x, also called linear MMSE estimator.

$$\mathbb{E}\left[Y|X=x\right] = \mathbb{E}[Y] + Cov(Y,X)Cov(X)^{-1}(x - \mathbb{E}[X])$$

and the posterior risk is independent of x:

$$Var\left[Y|X=x\right] = Var[Y] - Cov(Y,X)Cov(X)^{-1}Cov(X,Y)$$

<u>**Note:**</u> MMSE estimator coincides with the MAP estimator for Gaussian Variables.

2.4 Comparison

Maximum-Likelihood Principle (state is nonrandom):

$$\delta_{ML}(x) = \underset{y}{\text{argmax}}\, P_y(x)$$

Maximum A Posteriori (MAP) Decision Rule (state is random):

$$\delta_{MAP}(x) = \underset{y}{\text{argmax}}\, \pi(y|x) = \underset{y}{\text{argmax}}\{\pi(y|x), P_y(x)\}$$

Chapter 3 Non-parameteric Prediction Problem

Problem

There are J non-stochastic treats $X \in \mathbb{X} \subseteq \mathbb{R}^J$, and we want to predict a related outcome $Y \in \mathbb{Y} \subseteq \mathbb{R}$. Given a sample X_i,

$$Y_i = m(X_i) + \sigma u_i$$

where $m(\cdot)$ is an unknown function and $u_i | X_i \sim \mathcal{N}(0,1)$.

Goal:

- Predict Y given a new X;
- Learn $m(\cdot)$.

Decision Rule

Given $\vec{X} = \{X_1, ..., X_N\}$, we want to derive a decision rule $d(\vec{Y})$ given corresponding $\vec{Y}$ of $\vec{X}$.

Define $\boldsymbol{m} \triangleq [m(X_1), ..., m(X_N)]'$, its estimation is denoted by $\hat{\boldsymbol{m}}$, which is based on the decision rule.

Sum of Squared Residual

Proposition 3.1 (Sum of Squared Residual)

Sum of Squared Residual (SSR) of an estimation $\hat{\boldsymbol{m}}$ is given by

$$\mathbb{E}\left[\|\vec{Y} - \hat{\boldsymbol{m}}\|^2\right] = N\sigma^2 + \sum_{i=1}^{N} (\hat{m}(X_i) - m(X_i))^2 - 2\sigma^2 df(\hat{\boldsymbol{m}}) \tag{SSR}$$

where the norm is $\|\vec{X}\| = \left[\sum_{i=1}^{N} X_i^2\right]^{1/2}$, and the degree of freedom $df(\hat{\boldsymbol{m}}) = \frac{\sum_{i=1}^{N} \text{Cov}(Y_i, \hat{m}_i)}{\sigma^2}$.

Proof 3.1

$$\begin{aligned}
\mathbb{E}\left[\|\vec{Y} - \hat{\boldsymbol{m}}\|^2\right] &= \mathbb{E}\left[\|(\vec{Y} - \boldsymbol{m}) + (\boldsymbol{m} - \hat{\boldsymbol{m}})\|^2\right] \\
&= \mathbb{E}\left[\|\vec{Y} - \boldsymbol{m}\|^2\right] + \mathbb{E}\left[\|\hat{\boldsymbol{m}} - \boldsymbol{m}\|^2\right] - 2\mathbb{E}\left[(\vec{Y} - \boldsymbol{m})'(\hat{\boldsymbol{m}} - \boldsymbol{m})\right] \\
&= N\sigma^2 + \sum_{i=1}^{N} (\hat{m}(X_i) - m(X_i))^2 - 2\sigma^2 df(\hat{\boldsymbol{m}})
\end{aligned}$$

The second equality is because

$$\begin{aligned}
\mathbb{E}\left[(\vec{Y}-\boldsymbol{m})'(\hat{\boldsymbol{m}}-\boldsymbol{m})\right] &= \sum_{i=1}^{N}\mathbb{E}\left[(Y_i-m_i)(\hat{m}_i-m_i)\right] \\
&= \sum_{i=1}^{N}\mathbb{E}\left[(Y_i-m_i)\hat{m}_i\right] \\
&= \sum_{i=1}^{N}\mathbb{E}\left[(Y_i-m_i)(\hat{m}_i-\mathbb{E}m_i)\right] \\
&= \sum_{i=1}^{N}\text{Cov}(Y_i,\hat{m}_i)
\end{aligned}$$

We can represent the risk of estimation $\hat{\boldsymbol{m}}$ by rewriting SSR

$$\mathbb{E}\left[\|\hat{\boldsymbol{m}}-\boldsymbol{m}\|^2\right] = \mathbb{E}\left[\|\vec{Y}-\hat{\boldsymbol{m}}\|^2\right] - N\sigma^2 + 2\sigma^2 df(\hat{\boldsymbol{m}})$$

3.1 K-normal Means Probelm

3.1.1 Assumptions

Assumption

1. *Linear Combination: Suppose the $m(\cdot)$ can be wrritten as a liner combination of bases functions:*

$$\boldsymbol{m}(X) = \sum_{k=1}^{K}\alpha_k g_k(X)$$

2. *Gram-Schmidt Orthonormalization (How to processe raw data):*

Definition 3.1 (Gram-Schmidt Orthonormalization)

$\phi_k(X), k=1,...,K$ such that

(1). $\frac{1}{N}\sum_{i=1}^{N}\phi_k^2(X_i)=1$ and

(2). $\frac{1}{N}\sum_{i=1}^{N}\phi_k(X_i)\phi_k(X_j)=0$.

Suppose the $\boldsymbol{m}(\cdot)$ is a liner combination of Gram-Schmidt orthonormalizations:

$$\boldsymbol{m}(X) = \sum_{k=1}^{K}\theta_k\phi_k(X) \quad \triangleq W\boldsymbol{\theta}$$

where $W=(w(X_1),...,w(X_N))^T \in \mathbb{R}^{N\times K}$, $w(X_i)=(\phi_1(X_i),...,\phi_K(X_i))^T \in \mathbb{R}^{K\times 1}$, and $\boldsymbol{\theta}=(\theta_1,...,\theta_K)^T$.

That is, given X, let

$$\underbrace{\boldsymbol{Y}}_{N\times 1} = \underbrace{W\boldsymbol{\theta}}_{(N\times K)(K\times 1)} + \underbrace{\sigma^2 U}_{N\times 1}, \text{ where } U\sim\mathcal{N}(0,I_N)$$

3.1.2 Maximum Likelihood Estimator

Based on these assumptions, the conditional distribution is

$$\boldsymbol{Y} \mid X \sim \mathcal{N}(W\boldsymbol{\theta}, \sigma^2 I_N)$$

and the log-likelihood function is

$$l(Y \mid X, \boldsymbol{\theta}) = -\frac{N}{2}\ln 2\pi - N\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(Y_i - w(X_i)^T\boldsymbol{\theta}\right)^2$$

Then, we get the maximum likelihood estimator (MLE),

$$\hat{\theta}_{MLE} = \left[\frac{1}{N}\sum_{i=1}^{N} w(X_i)w(X_i)^T\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^{N} w(X_i)Y_i\right]$$

By the orthonormalization assumption 3.1, $\frac{1}{N}\sum_{i=1}^{N} w(X_i)w(X_i)^T = I_K$. Hence,

$$\hat{\theta}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} w(X_i)Y_i = \frac{W^T\boldsymbol{Y}}{N} \in \mathbb{R}^{K\times 1}$$

We can observe that it is a **conditionally unbiased** estimate of $\boldsymbol{\theta}$:

$$\mathbb{E}[\hat{\theta}_{MLE} \mid X] = \frac{1}{N}\sum_{i=1}^{N} w(X_i)\mathbb{E}[Y_i \mid X] \quad = \left(\frac{1}{N}\sum_{i=1}^{N} w_i w_i^T\right)\boldsymbol{\theta} = \boldsymbol{\theta}$$

The **variance** of k^{th} item of $\hat{\theta}_{MLE}$, $Z_k \triangleq \frac{1}{N}\sum_{i=1}^{N}\phi_k(X_i)Y_i \mid X$, is

$$\text{Var}(Z_k) = \text{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\phi_k(X_i)Y_i \mid X\right) = \frac{1}{N^2}\sum_{i=1}^{N}\phi_k^2(X_i)\text{Var}\left(Y_i \mid X\right) = \frac{\sigma^2}{N}$$

Hence,

$$\hat{\theta}_{MLE} \mid X \sim \mathcal{N}\left(\boldsymbol{\theta}, \frac{\sigma^2}{N} I_K\right)$$

3.1.3 Risk of MLE

All in all, we estimate the $\boldsymbol{m}(\cdot)$ by the unbiased estimator

$$\hat{\boldsymbol{m}} = W\hat{\theta}_{MLE}$$

and then the loss is given by

$$\|\hat{\boldsymbol{m}} - \boldsymbol{m}\|^2 = \|W(\hat{\theta}_{MLE} - \boldsymbol{\theta})\|^2 = \sum_{k=1}^{K}(Z_k - \theta_k)^2 \triangleq L(\hat{\theta}_{MLE}, \boldsymbol{\theta})$$

where we consider the square loss. Hence, the risk of MLE estimation is

$$\begin{aligned} R(d_{MLE}, \boldsymbol{\theta}) &= \mathbb{E}[\|\hat{\boldsymbol{m}} - \boldsymbol{m}\|^2] \\ &= \sum_{k=1}^{K}\mathbb{E}[(Z_k - \theta_k)^2] = \frac{K}{N}\sigma^2 \end{aligned}$$

3.1.4 James-Stein Type Estimator

MLE is a member of the class of estimators $\mathcal{L} = \{C\mathbb{Z} : C = \{c_1, ..., c_K\}, c_k \in [0,1]\}$. Here, we consider a estimartor in the class:

$$\begin{aligned}\mathbb{E}\left[\sum_{k=1}^{K}(c_k Z_k - \theta_k)^2\right] &= \mathbb{E}\left[\sum_{k=1}^{K}(c_k(Z_k - \theta_k) - (1-c_k)\theta_k)^2\right] \\ &= \frac{\sigma^2}{N}\sum_{i=1}^{K} c_k^2 + \sum_{i=1}^{K}(1-c_k)^2\theta_k^2\end{aligned}$$

By F.O.C., the optimal estimator minimizing the risk is

$$c_k^* = \frac{\theta_k^2}{\frac{\sigma^2}{N} + \theta_k^2},\ k = 1, ..., K$$

Then, the oracle bound is

$$\inf_{d_{\mathcal{L}}} R(d_{\mathcal{L}}, \boldsymbol{\theta}) = \underbrace{\frac{\sigma^2}{N}\left(\sum_{k=1}^{K}\frac{\theta_k^2}{\frac{\sigma^2}{N} + \theta_k^2}\right)}_{\text{oracle bound}} < \frac{K}{N}\sigma^2$$

(we can't achieve it as we don't know $\boldsymbol{\theta}$).

Is there a feasible estimator which uniformly improves upon MLE? Yes!

3.1.4.1 Stein's Unbiased Risk Estimate (SURE)

Consider sample $\mathcal{Z} \sim \mathcal{N}(\theta, \sigma^2 I_K)$, and a estimator based on $\mathcal{Z}$, $\hat{\theta} = \hat{\theta}(\mathcal{Z})$. Let $g(\hat{\theta}) = \hat{\theta} - \mathcal{Z}$. Then, the estimate of risk is given by

$$\hat{R}_{\text{SURE}}(\mathcal{Z}) = K\sigma^2 + 2\sigma^2\sum_{k=1}^{K}\frac{\partial g_k(\mathcal{Z})}{\partial \mathcal{Z}_k} + \underbrace{\sum_{k=1}^{K}\left(\hat{\theta}_k - \mathcal{Z}_k\right)^2}_{\|g(\hat{\theta})\|^2_{\text{Frobenius}}}$$

It is an unbiased estimate of the mean-squared error:

$$\mathbb{E}[\hat{R}_{\text{SURE}}] = \mathbb{E}[\|\hat{\theta} - \theta\|^2]$$

Proof 3.2 (for the unbiased property)

According to (SSR),

$$\begin{aligned}&\mathbb{E}[\|\hat{\theta} - Z\|^2] = \mathbb{E}[\|\hat{\theta} - \theta\|^2] + K\sigma^2 - 2\sum_{k=1}^{K}\text{Cov}(\mathcal{Z}_k, \hat{\theta}_k) \\ \Rightarrow\ &\mathbb{E}[\|\hat{\theta} - \theta\|^2] = -K\sigma^2 + 2\sum_{k=1}^{K}\text{Cov}(\mathcal{Z}_k, \hat{\theta}_k) + \mathbb{E}[\|\hat{\theta} - \mathcal{Z}\|^2]\end{aligned}$$

where

$$\begin{aligned}\mathrm{Cov}(\mathcal{Z}_k, \hat{\theta}_k) &= \mathbb{E}[\hat{\theta}_k(\mathcal{Z}_k - \theta_k)] \\ &= \mathbb{E}[(\hat{\theta}_k - Z_k)(\mathcal{Z}_k - \theta_k)] + \mathbb{E}[(\mathcal{Z}_k - \theta_k)^2] \\ &= \mathbb{E}[g_k(\mathcal{Z})(\mathcal{Z}_k - \theta_k)] + \sigma^2\end{aligned}$$

Note $g(\hat{\theta}) = \hat{\theta} - \mathcal{Z}$.

Claim 3.1

$\mathbb{E}[g_k(\mathcal{Z})(\mathcal{Z}_k - \theta_k)] = \sigma^2 \mathbb{E}[\frac{\partial g_k(\mathcal{Z})}{\partial \mathcal{Z}_k}]$

$$\begin{aligned}\sigma^2 \mathbb{E}[\nabla_Z g(\mathbf{Z})] &= \sigma^2 \int_z f_Z(z) \nabla_z g(z) dz \\ &= \sigma^2 \int_z f_Z(z) dg(z) \\ &= \sigma^2 \left(f_Z(z) g(z) \Big|_{\partial \mathbb{R}^K} - \int_z g(z) \frac{\partial f_Z(z)}{\partial z} dz \right) \\ &= \sigma^2 \int_z g(z) \left(\frac{1}{\sigma^2} (z - \theta)' f_Z(z) \right) dz \\ &= \int_z f_Z(z) g(z) (z - \theta)' dz \\ &= \mathbb{E}[g(\mathbf{Z})(\mathbf{Z} - \theta)']\end{aligned}$$

Hence,

$$\mathbb{E}[\|\hat{\theta} - \theta\|^2] = K\sigma^2 + 2\sigma^2 \sum_{k=1}^{K} \mathbb{E}[\frac{\partial g_k(\mathcal{Z})}{\partial \mathcal{Z}_k}] + \mathbb{E}[\|\hat{\theta} - \mathcal{Z}\|^2] = \mathbb{E}[\hat{R}_{\text{SURE}}]$$

3.1.4.2 James and Stein Estimator

Note: Here, we consider $\mathcal{Z} \sim \mathcal{N}(\theta, \frac{\sigma^2}{N} I_K)$

Theorem 3.1 (James and Stein (1961))

$$\hat{\theta}_{JS}(\mathcal{Z}) = \left(1 - \frac{(K-2)}{\mathcal{Z}'\mathcal{Z}} \frac{\sigma^2}{N} \right) \mathcal{Z}$$

We have, $g_{JS}(\mathcal{Z}) = -\frac{(K-2)}{\mathcal{Z}'\mathcal{Z}} \frac{\sigma^2}{N} \mathcal{Z}$ and

$$\sum_{k=1}^{K} \frac{\partial g_k(\mathcal{Z})}{\partial \mathcal{Z}_k} = -\frac{(K-2)}{\mathcal{Z}'\mathcal{Z}} \frac{\sigma^2}{N} \sum_{k=1}^{K} \left(1 - \frac{2\mathcal{Z}_k^2}{\mathcal{Z}'\mathcal{Z}} \right) = -\frac{(K-2)^2}{\mathcal{Z}'\mathcal{Z}} \frac{\sigma^2}{N}$$

Hence, the corresponding SURE is

$$\begin{aligned}\hat{R}_{\text{SURE}}(\mathcal{Z}) &= \frac{K}{N}\sigma^2 - \frac{2\sigma^2}{N}\frac{(K-2)^2}{\mathcal{Z}'\mathcal{Z}}\frac{\sigma^2}{N} + \frac{(K-2)^2}{(\mathcal{Z}'\mathcal{Z})^2}\left(\frac{\sigma^2}{N}\right)^2\sum_{k=1}^{K}\mathcal{Z}_k^2 \\ &= \frac{K}{N}\sigma^2 - \frac{(K-2)^2}{\mathcal{Z}'\mathcal{Z}}\frac{\sigma^4}{N^2}\end{aligned}$$

Then,

$$\begin{aligned}R(\hat{\theta}_{JS},\theta) &= \mathbb{E}[\hat{R}_{\text{SURE}}(\mathcal{Z})] \\ &= \frac{K}{N}\sigma^2 - (K-2)^2\frac{\sigma^4}{N^2}\mathbb{E}\left[\frac{1}{\mathcal{Z}'\mathcal{Z}}\right]\end{aligned} \tag{RJS}$$

As $\mathcal{Z}_k \sim \mathcal{N}(\theta_k, \frac{\sigma^2}{N})$, $\mathcal{Z}'\mathcal{Z} = \sum_{k=1}^{K}\mathcal{Z}_k^2 \sim \frac{\sigma^2}{N}V$ such that $V \sim \chi^2_{K+2W}$ where $W \sim \text{Poisson}(\frac{\rho}{2})$ and $\rho = N\sum_{k=1}^{K}\frac{\theta_K^2}{\sigma^2}$. So,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\mathcal{Z}'\mathcal{Z}}\right] &= \frac{N}{\sigma^2}\mathbb{E}\left[\frac{1}{V}\right] \\ &= \frac{N}{\sigma^2}\mathbb{E}\left[\frac{1}{K-2+2W}\right] \text{ (by the identity of chi-square distribution)} \\ &\geq \frac{N}{\sigma^2}\frac{1}{K-2+\rho} \text{ (by Jensen's inequality)} \\ &= \frac{1}{(K-2)\frac{\sigma^2}{N} + \|\theta\|_2^2}\end{aligned}$$

Substitute it into RJS,

$$R(\hat{\theta}_{JS},\theta) \leq \frac{K}{N}\sigma^2 - \frac{(K-2)^2\frac{\sigma^4}{N^2}}{(K-2)\frac{\sigma^2}{N} + \|\theta\|_2^2}$$

Hence,

$$R(\hat{\theta}_{MLE},\theta) - R(\hat{\theta}_{JS},\theta) \geq \frac{(K-2)^2\frac{\sigma^4}{N^2}}{(K-2)\frac{\sigma^2}{N} + \|\theta\|_2^2} \geq 0$$

which shows that $\underline{\hat{\theta}_{JS} \text{ works better than } \hat{\theta}_{MLE}}$ (ML is inadmissable) under squared error loss.

3.1.4.3 A more general form of estimator $\mathcal{L} = \{C\mathbf{Z} : C = \textbf{diag}\vec{c}, \vec{c} \in [0,1]^K\}$

Consider a new estimator $\hat{\theta} \in \mathcal{L} = \{C\mathbf{Z} : C = \text{diag}\vec{c}, \vec{c} \in [0,1]^K\}$. The SURE is

$$\hat{R}_{\text{SURE}}(\mathbf{Z}, \vec{c}) = \frac{\sigma^2}{N}\sum_{k=1}^{K}c_k^2 + \sum_{k=1}^{K}\left(\mathcal{Z}_k^2 - \frac{\sigma^2}{N}\right)(1-c_k)^2$$

Empirical Risk Minimization: taking F.O.C.

$$\hat{c}_k = \left(1 - \frac{\sigma^2}{N}\frac{1}{\mathcal{Z}_k^2}\right), \ k = 1, ..., K$$

Chapter 4 M-Estimation

Suppose there is a parameter of interest $\theta \in \mathbb{R}^d$. Data Z is generated from F_{θ_0}.

Definition 4.1 (Extremum Estimator)

Extremum estimators are a wide class of estimators for parametric models that are calculated through maximization (or minimization) of a certain objective function, which depends on the data.

Suppose the true parameter $\theta_0 = \operatorname{argmin}_{\theta \in \Theta} Q(\theta)$, where $Q : \Theta \to \mathbb{R}$ is criterion (objective) function (unknown). In estimation, $\{Z_i\}_{i=1}^n$ are i.i.d. sample, where $Z_i \sim F_Z$ whose parameter θ is of interest.

$\hat{Q} : \Theta \to \mathbb{R}$ is a sample criterion. $\hat{\theta}$ is called **extremum estimator** of θ if

$$\hat{\theta}(\theta) = \operatorname*{argmin}_{\theta \in \Theta} \hat{Q}(\theta)$$

Definition 4.2 (M-Estimator)

M-estimators are a broad class of extremum estimators for which the objective function is a sample average. Specifically, Q is in the form of $\mathbb{E}m(Z, \theta)$, where $m(Z, \theta)$ is called M-estimator loss that only depends on one data sample and the parameter. Then, $\hat{Q}$ is in the form of

$$\hat{Q} = \frac{1}{n} \sum_{i=1}^{n} m(Z_i, \theta)$$

we call the $\hat{\theta} = \operatorname{argmin}_{\theta \in \Theta} \hat{Q}(\theta)$ be the **M-estimator** of θ.

MLE is a special case of M-estimator.

$$\text{Maximum Likelihood Estimators} \subseteq \text{M-Estimators} \subseteq \text{Extremum Estimators}$$

4.1 Consistency and Asymptotic Normality of M-estimator

4.1.1 Identification of M-estimator: $\theta_0 = \mathrm{argmin}_{\theta \in \Theta} Q(\theta)$

Example 4.1 (ML Identification)

Take $m(Z, \theta) = -\ln f(Z|\theta)$, where $z \to f(z|\theta)$ is the parametric density function such that $z \to f(z|\theta_0)$ is the true density function of Z.

$$\theta_0 = \underset{\theta \in \Theta}{\mathrm{argmin}}\, Q(\theta) := -\mathbb{E} \log f(x|\theta)$$

Why this is feasible? We can show that $Q(\theta) \geq Q(\theta_0), \forall \theta \in \Theta$.

Lemma 4.1 (Information Inequality: $\theta_0 = \mathrm{argmin}_{\theta \in \Theta} -\mathbb{E} \log f(x|\theta)$)

Given θ_0 be the true parameter, we have

$$Q(\theta) - Q(\theta_0) = -\mathbb{E}\left[\log f(x|\theta) - \log f(x|\theta_0)\right] > 0, \forall \theta \neq \theta_0$$

Proof 4.1

$$\begin{aligned}
Q(\theta) - Q(\theta_0) &= -\mathbb{E}_{\theta_0}\left[\log f(x|\theta) - \log f(x|\theta_0)\right] \\
&= -\mathbb{E}_{\theta_0}\left[\log \frac{f(x|\theta)}{f(x|\theta_0)}\right], \text{ where } \log(z) \text{ is concave} \\
\text{by Jensen's inequality } &> -\log \mathbb{E}_{\theta_0} \frac{f(x|\theta)}{f(x|\theta_0)} \\
&= -\log \int \frac{f(x|\theta)}{f(x|\theta_0)} f(x|\theta_0) dx \\
&= -\log 1 = 0
\end{aligned}$$

Example 4.2 (Nonlinear Least Squares)

Consider the conditional restriction

$$\mathbb{E}[Y|X = x] = g(x, \theta_0)$$

where g is known up to θ and differentiable in θ. Then, the NLS criterion function is

$$Q(\theta) = \mathbb{E}[Y - g(X, \theta)]^2$$

We can show that $Q(\theta_0) \leq Q(\theta), \forall \theta \in \Theta$.

Lemma 4.2 (NLS Identification)

$$
\begin{aligned}
Q(\theta) &= \mathbb{E}[Y - g(X,\theta)]^2 \\
&= \mathbb{E}[Y - g(X,\theta_0) - (g(X,\theta) - g(X,\theta_0))]^2 \\
&= \mathbb{E}[Y - g(X,\theta_0)]^2 + \mathbb{E}[g(X,\theta) - g(X,\theta_0)]^2 \\
&= Q(\theta_0) + \mathbb{E}[g(X,\theta) - g(X,\theta_0)]^2 \geq Q(\theta_0)
\end{aligned}
$$

Notations

Define $g(Z,\theta) := \frac{\partial}{\partial\theta} m(Z,\theta) \in \mathbb{R}^d$ and $G(Z,\theta) := \frac{\partial^2}{\partial\theta\partial\theta^T} m(Z,\theta) \in \mathbb{R}^{d\times d}$.

Definition 4.3

Suppose the data Z follows true distribution with parameter θ_0.

1. Loss: $Q(\theta) := \mathbb{E}_{\theta_0} m(Z,\theta)$.
2. Score: $g(\theta) := \mathbb{E}_{\theta_0} g(Z,\theta)$.
3. Hessian: $G(\theta) := \mathbb{E}_{\theta_0} G(Z,\theta) = \mathbb{E}_{\theta_0}\left[\frac{\partial^2}{\partial\theta\partial\theta^T} m(Z,\theta)\right]$. (We use G denote the true population Hessian, $G := G(\theta_0)$).

In the MLE $m(Z,\theta) = \ln f(Z;\theta)$, we also use Information Matrix $\mathcal{I}(\theta) := \mathbb{E}[g(Z,\theta)g(Z,\theta)^T]$.

Example 4.3 (Poisson Distribution)

A Poisson distribution with rate parameter λ has p.m.f. $f(Z;\lambda) = \frac{\lambda^Z}{Z!}e^{-\lambda}$. Then, in MLE, we have $g(Z;\lambda) = \frac{Z}{\lambda} - 1 \Rightarrow \lambda_0 = \mathbb{E}Z = \mathrm{Var}Z$. $I(\lambda_0) = \frac{1}{\lambda_0}, G(\lambda_0) = -\frac{1}{\lambda_0}$.

4.1.2 Consistency of M-estimators

Consistency means: $\hat\theta \xrightarrow{P_0} \theta_0$ as $n \to \infty$.

Can $\hat Q(\theta) \xrightarrow{P_0} Q(\theta)$ give the consistency of the M-estimator ($\hat\theta \xrightarrow{P_0} \theta_0$)? No.

Example 4.4

$Q(\theta) = -\mathbf{1}\{\theta = 0\}$ and $Q_n(\theta) = -\mathbf{1}\{\theta = 0\} - 2\mathbf{1}\{\theta = n\}$. $\theta_n \nrightarrow \theta_0$ but $Q_n(\theta) - Q(\theta) \to 0$.

Theorem 4.1 (Extremum Consistency)

Given three assumptions

A0. <u>Global Identification</u>: $\theta_0 = \mathrm{argmin}_{\theta\in\Theta} Q(\theta)$.

A1. Uniform Convergence: the worst-case distance converges to zero.

$$\sup_{\theta \in \Theta} |\hat{Q}(\theta) - Q(\theta)| \xrightarrow{P} 0$$

(if $Q(\beta)$ is convex in β, pointwise convergence is enough,

$$\hat{Q}(\theta) - Q(\theta) \xrightarrow{P} 0$$

which follows from LLN.)

A2. Continuity:

$$\inf_{\|\theta - \theta_0\| > \epsilon} Q(\theta) > Q(\theta_0)$$

(Its **sufficient** condition: $Q(\theta)$ is continuous in θ on compact set Θ.)

Suppose A0, A1 and A2 hold. Then,

$$\hat{\theta} \xrightarrow{P} \theta_0$$

Theorem 4.2 (Uniform Law of Large Numbers (ULLN), Theorem 22.2 (Hansen, 2022))

Suppose

1. (Y_i, X_i) are i.i.d.
2. $\mathbb{E}[m(Z, \theta)] < \infty$ for all $\theta \in \Theta$.
3. Θ is bounded.
4. For some $A < \infty$ and $\alpha > 0$, $\mathbb{E}|m(Z, \theta_1) - m(Z, \theta_2)| \leq A\|\theta_1 - \theta_2\|^\alpha$ for all $\theta_1, \theta_2 \in \Theta$.

Then Uniform Convergence holds:

$$\sup_{\theta \in \Theta} |\hat{Q}(\theta) - Q(\theta)| \xrightarrow{P} 0$$

4.1.3 Asymptotic Normality of M-estimators

Review: By the Taylor expansion for any $f - n$, the $h : \Theta \to \mathbb{R}^d$,

$$h(\theta) - h(\theta_0) = \underbrace{\left(\frac{\partial h}{\partial \theta}|_{\theta = \bar{\theta}}\right)}_{\in \mathbb{R}^{d \times d}} \cdot \underbrace{(\theta - \theta_0)}_{\in \mathbb{R}^d}$$

where $\bar{\theta} = \alpha\theta + (1 - \alpha)\theta_0$ for some $\alpha \in (0, 1)$.

Theorem 4.3 (Asymptotic Normality of M-estimators)

Suppose A0, A1 and A2 hold. With extra assumptions: **Theorem 22.3 (Hansen, 2022):**

A3. $\mathbb{E}\|g(\theta_0)\|^2 < \infty$.

A4. Continuous Hessian: $G(\theta)$ is continuous in Θ.

A5. For some $A < \infty$ and $\alpha > 0$, $\mathbb{E}|g(\theta_1) - g(\theta_2)| \leq A\|\theta_1 - \theta_2\|^\alpha$ for all $\theta_1, \theta_2 \in \Theta$.

A6. $G := G(\theta_0)$ is invertible (i.e., full rank).

A7. θ_0 is in the interior of Θ.

Or, assumptions in ECON 715 lecture note (Shi, Xiaoxia)

A3. Differentiability: $\hat{Q}_n(\theta)$ is twice continuously differentiable in on some neighborhood $\Theta_0 \subset \Theta$ of θ_0 (with probability one).

A4. Asymptotic normality of the gradient: $\sqrt{n}\hat{g}(\theta_0) := \sqrt{n}\frac{\partial}{\partial\theta}\hat{Q}_n(\theta) \xrightarrow{d} N(0, \Omega)$.

A5. Continuous convergence of the Hessian: for any sequence $\tilde{\theta}_n \xrightarrow{P} \theta_0$, $\hat{G}(\tilde{\theta}_n) := \frac{\partial^2}{\partial\theta\partial\theta^T}\hat{Q}_n(\tilde{\theta}_n) \xrightarrow{P} B_0$ for some non-stochastic $d \times d$ matrix (usually we check G_0).

Then,

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \overset{d}{\Rightarrow} N\left(0, G^{-1}\Omega G^{-1}\right)$$

where

$$\Omega = \text{Var}(\sqrt{n}\hat{g}(\theta_0)) = \text{Var}(g(Z, \theta_0)), \quad \hat{g}(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} g(Z_i, \theta_0)$$

Proof 4.2

By the optimality of $\hat{\theta}$,

$$\hat{g}(\hat{\theta}) = 0$$

where $\hat{g}(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} g(Z_i, \theta_0)$,

$$\mathbb{E}\hat{g}(\theta_0) = \mathbb{E}g(Z, \theta_0) = 0$$

$$\text{Var}(\hat{g}(\theta_0)) = \frac{1}{n}\underbrace{\text{Var}(g(Z, \theta_0))}_{:=\mathcal{I}(\theta_0)}$$

By Taylor,

$$\hat{g}(\hat{\theta}) - \hat{g}(\theta_0) = \hat{G}(\bar{\theta})(\hat{\theta} - \theta_0)$$

for some $\hat{\theta}$. By assumptions and results above

$$-\hat{g}(\theta_0) = \hat{g}(\hat{\theta}) - \hat{g}(\theta_0) \approx G(\hat{\theta} - \theta_0)$$

$$\hat{\theta} - \theta_0 \approx -G^{-1}\hat{g}(\theta_0)$$

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \overset{d}{\Rightarrow} N\left(0, G^{-1}\underbrace{\text{Var}(\sqrt{n}\hat{g}(\theta_0))}_{=\text{Var}(g(Z,\theta_0))}G^{-1}\right)$$

Corollary 4.1 (Asymptotic Normality of ML-estimator under correct specfication)

For MLE, under ``Regularity" condition, $\mathcal{I}(\theta_0) = -G(\theta_0)$,

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \overset{d}{\Rightarrow} N\left(0, \mathcal{I}(\theta_0)^{-1}\right)$$

$$\sqrt{n}\hat{g}(\theta_0) \overset{d}{\Rightarrow} N\left(0, \mathcal{I}(\theta_0)\right)$$

4.2 Efficiency and Misspecification

4.2.1 Efficiency of Asymptotically Linear Estimator

Definition 4.4 (Efficient Asymptotically Linear Estimator)

An asymptotically linear estimator is called **efficient** if it attains the smallest variance among the class of asymptotic estimators.

Use Ω_β denote the variance of $\hat{\beta}$.

$\hat{\beta}_1$ is more efficient than $\hat{\beta}_2$ if both of them are asymptotic normal

- $\Omega_{\hat{\beta}_2} - \Omega_{\hat{\beta}_1} \succeq 0$ in matrix sense.
- Standard errors of $\hat{\beta}_1$ are smaller in large sample.

$\hat{\beta}$ is **efficient** if for any other $\hat{\beta}_2$, $\Omega_{\hat{\beta}_2} - \Omega_{\hat{\beta}_1} \succeq 0$ in matrix sense.

4.2.2 Misspecification and Pseudo-true Parameter

Misspecification: Sometimes, the true density of the data distribution is unknown. We minimize a criterion function (or a density function we assume for MLE) to approximate the true parameter. This assumed function loses the original interpretation.

Definition 4.5 (Pseudo-true Parameter)

Pseudo-true parameter is given by

$$\beta_0 \equiv \arg\min_{\beta} Q(\beta)$$

β_0 s.t. $g(\beta_0) = 0 = \mathbb{E}[g(Y|X, \beta_0)] = 0$.

In MLE case, because the density function used in the criterion function is different to the true density function of data, the pseudo-true parameter doesn't satisfy the second information equality, $G^{-1}\mathcal{I}G^{-1} \neq \mathcal{I}^{-1}$.

4.2.3 Example of Misspecification

Example 4.5

Consider a linear exponential density of the form

$$f(y;\theta) = exp\left(A(\theta) + B(y) + C(\theta)y\right)$$
$$\theta = \int y f(y;\theta) dy$$

(a). What is $\mathbb{E} \ln f(y;\theta)$ when y has PDF $f(y;\theta_0)$ (i.e., θ may differ from θ_0):

$$\begin{aligned}\mathbb{E} \ln f(y;\theta) &= \int f(y;\theta_0)\left(A(\theta) + B(y) + C(\theta)y\right) dy \\ &= A(\theta) + \int f(y;\theta_0) B(y) dy + C(\theta)\theta_0\end{aligned}$$

(b). By information inequality, for any other θ, $\mathbb{E}_{\theta_0}[\ln(y;\theta_0)] > \mathbb{E}_{\theta_0}[\ln(y;\theta)]$. That is,

$$A(\theta_0) + \int f(y;\theta_0) B(y) dy + C(\theta_0)\theta_0 > A(\theta) + \int f(y;\theta_0) B(y) dy + C(\theta)\theta_0$$

$$A(\theta_0) + C(\theta_0)\theta_0 > A(\theta) + C(\theta)\theta_0$$

i.e., $A(\theta) + C(\theta)\theta_0$ is maximized at $\theta = \theta_0$.

(c). In general, if the distribution of y is not in the form $f(y \mid \theta)$ and we only know $\mathbb{E}[y]$, we can show that $\mathbb{E}[\ln f(y;\theta)]$ is maximized at $\mathbb{E}[y]$:

$$\underset{\theta}{\operatorname{argmax}}\, \mathbb{E}[\ln f(y;\theta)] = \underset{\theta}{\operatorname{argmax}} \left(A(\theta) + C(\theta)\mathbb{E}[y]\right) = \mathbb{E}[y]$$

The last equality is given by the previous result.

(d). Hence, when the likelihood is not correctly specified, the pseudo-true parameter is given by

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{i=1}^{n} \ln f(y_i;\theta) \xrightarrow{P} \underset{\theta}{\operatorname{argmax}}\, \mathbb{E}\left[\ln f(y_i;\theta)\right] = \mathbb{E}[y]$$

(e). Now, suppose we use the following density function as the criterion

$$f(y \mid x, \beta, \gamma) = exp\left(A(h(x,\beta), x, \gamma) + B(y, x, \gamma) + C(h(x,\beta), x, \gamma)y\right)$$

$$\mathbb{E} \ln f(y \mid x, \beta, \gamma) = A(h(x,\beta), x, \gamma) + \mathbb{E}[B(y, x, \gamma) \mid x, \beta, \gamma] + C(h(x,\beta), x, \gamma)\mathbb{E}[y \mid x, \beta, \gamma]$$

- If specified correctly, i.e., the $y \mid x$ has the form $f(y \mid x, \beta_0, \gamma)$ and $\beta_0 = \mathbb{E}[y \mid x, \beta_0, \gamma]$: By information inequality,

$$\beta_0 = \underset{\beta}{\operatorname{argmax}}\, \mathbb{E} \ln f(y \mid x, \beta, \gamma) = \underset{\beta}{\operatorname{argmax}}\, A(h(x,\beta), x, \gamma) + C(h(x,\beta), x, \gamma)\mathbb{E}[y \mid x, \beta_0, \gamma]$$

- If misspecified, i.e., the $y \mid x$ has expectation $\mathbb{E}[y \mid x]$ but we still maximize $\mathbb{E} \ln f(y \mid x, \beta, \gamma)$:

$$\mathbb{E}[y \mid x] = \underset{\beta}{\operatorname{argmax}}\, \mathbb{E} \ln f(y \mid x, \beta, \gamma) = \underset{\beta}{\operatorname{argmax}}\, A(h(x,\beta), x, \gamma) + C(h(x,\beta), x, \gamma)\mathbb{E}[y \mid x]$$

Suppose you are interested in firms' applications for patents. You estimate the conditional mean parameters using a Poisson regression model:

$$\log \lambda = \log\left(\mathbb{E}[Y \mid X]\right) = X^T\beta$$
$$\Rightarrow f(y \mid x) = \frac{\lambda^Y}{Y!}e^{-\lambda} = \frac{[exp(X^T\beta)]^Y}{Y!}exp(-exp(X^T\beta))$$

However, the truth (unbeknownst to you) is that patents actually follow a negative binomial model (which permits the variance to differ from the mean), but the mean is correctly specified.

1. Will your estimator be consistent? Yes. This is directly given by the result above.
2. Will your estimator be asymptotically normal? Yes. The data are iid and the estimator is consistent, so the CLT holds under regularity conditions on the existence of second moments.
3. The information matrix equality **does not hold** if the likelihood is not correct.
4. An estimator of the asymptotic variance of the quasi-maximum likelihood estimator of the Poisson regression model that **is consistent** even if the Poisson assumption is incorrect:

$$\sqrt{n}\left(\hat{\theta} - \theta_*\right) \overset{d}{\Rightarrow} N\left(0, G^{-1}\Omega G^{-1}\right)$$

where θ_* is the pseudo-true parameter that estimated by the Poisson regression model.

$$\Omega = \mathbb{E}[s(z,\theta_*)s(z,\theta_*)^T],\ G = \mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta^T}f(Z;\theta_0)\right]$$

where $s(\cdot)$ is the score function. To obtain a consistent estimator, we would use $\hat{G}^{-1}\hat{\Omega}\hat{G}^{-1}$, where

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n}[s(z_i,\hat{\theta})s(z_i,\hat{\theta})^T],\ \hat{G} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial^2}{\partial\theta\partial\theta^T}f(z_i;\hat{\theta})\right]$$

4.3 Binary Choice

The goal in binary choice analysis is estimation of the **conditional or response probability**, $\Pr(Y = 1 \mid X)$, given a set of regressors X. We may be interested in the response probability or some transformation such as its derivative - the **marginal effect**, $\frac{\partial}{\partial X}\Pr(Y = 1 \mid X)$.

$Y \in \{0, 1\}$, $X \in \mathbb{R}^d$ (is assumed to) affects Y via $X^T\beta_0$, where $\beta_0 \in \mathbb{R}^d$.

The conditional probability of $Y = 1$ is represented by a link function $F : \mathbb{R} \to [0, 1]$.

$$\Pr(Y = 1 \mid X) = F(X^T\beta_0)$$

In other words, the model assumes that $Y \mid X$ is a coin flip (i.e., Bernoulli) with the parameter $F(X^T\beta_0)$:

$$Y \mid X \sim \text{Bernoulli}(F(X^T\beta_0)) \text{ a.s. in } X$$

Example 4.6

The choice of link:

1. Linear Probability Model (LPM): $F(t) = t\mathbf{1}\{t \in [0,1]\} = \begin{cases} 0, & t \leq 0 \\ t, & t \in [0,1] \\ 1, & t \geq 1 \end{cases}$ (projection).
2. Logit Model: $F(t) = \Lambda(t) = \frac{e^t}{1+e^t}$
3. Probit Model: $F(t) = \Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$

4.3.1 Latent Utility Models (structural motivation for probit model)

An agent makes a binary choice $d \in \{0,1\}$. The utility of each choice is given by

$$Y^*(d) = X^T\gamma_d + \epsilon(d), d \in \{0,1\}$$

where $X^T\gamma_d$ is the predicted/explained part of utility and $\epsilon(d)$ is the "taste shock" unobservable part of utility, $\mathbb{E}\epsilon(0) = \mathbb{E}\epsilon(1) = 0$. The key difference from RCT is the Y^* is not randomly assigned.

After observing X and $\epsilon(1), \epsilon(0)$, the agent makes a utility-maximizing choice

$$Y = 1\{Y^*(1) \geq Y^*(0)\}$$

The conditional probability of $Y = 1$ given X is

$$\begin{aligned}
\text{Pr}(Y = 1|X) &= \text{Pr}(Y^*(1) \geq Y^*(0) \mid X) \\
&= \text{Pr}(X^T\gamma_1 + \epsilon(1) \geq X^T\gamma_0 + \epsilon(0)) \\
&= \text{Pr}\left(\frac{\epsilon(0) - \epsilon(1)}{\sqrt{\text{Var}(\epsilon(0) - \epsilon(1))}} \leq X^T\left(\frac{\gamma_1 - \gamma_0}{\sqrt{\text{Var}(\epsilon(0) - \epsilon(1))}}\right)\right) \\
&= F\left(X^T\left(\frac{\gamma_1 - \gamma_0}{\sigma_{\epsilon(1)-\epsilon(0)}}\right)\right)
\end{aligned}$$

where $F(\cdot)$ is the CDF of $\frac{\epsilon(1)-\epsilon(0)}{\sigma_{\epsilon(1)-\epsilon(0)}}$. If $\epsilon(1), \epsilon(0)$ are jointly normal, then $F(\cdot) = \Phi(\cdot)$ is the CDF of the standard normal. It gives probit link function by leting $\beta = \frac{\gamma_1-\gamma_0}{\sigma_{\epsilon(1)-\epsilon(0)}} \in \mathbb{R}^d$.

The relative importance of X_j relative to X_k is $\frac{\beta_j}{\beta_k} = \frac{(\gamma_1-\gamma_0)_j}{(\gamma_1-\gamma_0)_k}, \forall j,k \in \{1,...,d\}$.

Marginal Effect

The marginal effect of change on X_j is

$$\frac{\partial}{\partial X_j}\text{Pr}(Y = 1|X = X) = F'(X^T\beta_0) \cdot \beta_j$$

The "average marginal effect" (AME) is given by

$$\text{AME} = \mathbb{E}_X F'(X^T\beta_0) \cdot \beta_j$$

The marginal effect for an "average person" (MEA) (may not make sense if X is discrete).

$$\text{MEA} = F'\left((\mathbb{E}X)'\beta_0\right)\beta_j$$

When $F'(\cdot)$ is nonlinear, AME $\neq$ MEA.

4.3.2 Estimation: Binary Regression

From joint to conditional likelihood

Denote the joint distribution of Y and X

$$f(Y, X; \beta) = f(Y \mid X; \beta) \cdot f_X(X)$$

Then,

$$\ln f(Y, X; \beta) = \ln f(Y \mid X; \beta) + \ln f_X(X)$$

Define the conditional likelihood criterion function,

$$Q(\beta) := -\mathbb{E}_\beta \ln f(Y, X; \beta) = -\mathbb{E}_\beta \ln f(Y \mid X; \beta) - \mathbb{E}_\beta \ln f_X(X)$$

The sample criterion function is given by

$$\hat{Q}_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \ln f(Y_i, X_i; \beta)$$

Since $\ln f_X(X)$ doesn't depend on β,

$$\arg\min_\beta Q(\beta) \equiv \arg\max_\beta \mathbb{E}_\beta \ln f(Y \mid X; \beta)$$
$$\hat{\theta} = \arg\min_\beta \hat{Q}_n(\beta) \equiv \arg\max_\beta \frac{1}{n}\sum_{i=1}^{n} \ln f(Y_i \mid X_i; \beta)$$

Binary Regression

1. $\Pr(Y = 1|X; \beta) = F(X^T\beta)$.
2. Log-likelihood
$$\ln f(Y \mid X; \beta) = Y \cdot \ln F(X^T\beta) + (1 - Y) \cdot \ln(1 - F(X^T\beta))$$
3. Take the derivative, the score is
$$\begin{aligned} g(Y \mid X; \beta) :&= \frac{\partial \ln f(Y|X; \beta)}{\partial \beta} = \frac{\partial \ln f(Y|X, \beta)}{\partial F(X^T\beta)}\frac{\partial F(X^T\beta)}{\partial \beta} \\ &= \frac{Y - F(X^T\beta)}{F(X^T\beta)(1 - F(X^T\beta))} \cdot \left(F'(X^T\beta) \cdot X\right) \end{aligned}$$

Note that the score function obeys conditional mean zero restriction at the true value $\beta = \beta_0$: $\mathbb{E}[Y - F(X^T\beta_0) \mid X] = 0 \Rightarrow \mathbb{E}g(Y \mid X; \beta_0) = 0$

The MLE ($\hat{\beta}_{\text{MLE}}$) is given by solving F.O.C.

$$\hat{g}(\beta)|_{\beta=\hat{\beta}_{\text{MLE}}} = \frac{1}{n}\sum_{i=1}^{n} g(Y_i \mid X_i;\beta)\big|_{\beta=\hat{\beta}_{\text{MLE}}} = 0^d \tag{4.1}$$

which is a system of (non)linear equations.

Let the weight of observation i be $w(X_i,\beta) := \frac{F'(X_i^T\beta)}{F(X_i^T\beta)(1-F(X_i^T\beta))} \cdot X_i$. Then, (4.1) can be written as

$$\hat{g}(\beta)|_{\beta=\hat{\beta}_{\text{MLE}}} = \sum_{i=1}^{n} w(X_i,\hat{\beta}_{\text{MLE}}) \cdot (Y_i - F(X_i^T\hat{\beta}_{\text{MLE}})) = 0^d$$

4.3.3 Consistency and Asymptotic Normality

Remind that $\hat{\beta}_{\text{MLE}}$ is M-estimator.

Assumption *The consistency theorem requires assumptions:*

(A1). $Q(\beta)$ is uniquely minimized at $\beta = \beta_0$.

(A2). $Q(\beta)$ is continuous on a compact subset of $\mathbb{R}$. ($Q(\beta)$ is continuous if the link $F(\cdot)$ is continuous.)

(A3). Uniform Convergence (if $Q(\beta)$ is convex in β, pointwise convergence is enough, which follows from LLN.)

By the Corollary 4.1,

$$\sqrt{n}\left(\hat{\beta}_{\text{MLE}} - \theta_0\right) \overset{d}{\Rightarrow} N\left(0, \mathcal{I}(\theta_0)^{-1}\right)$$

Since $Y \mid X \sim \text{Bernoulli}(F(X^T\beta_0))$, $\text{Var}(Y|X) = F(X^T\beta_0)\cdot(1-F(X^T\beta_0))$,

$$\begin{aligned}
\mathcal{I}(\theta_0) = G(\theta_0) &= \text{Var}\left(g(Y \mid X;\theta_0)\right) \\
&= \mathbb{E}\frac{\text{Var}\left(Y \mid X;\theta_0\right)}{F(X^T\beta_0)^2(1-F(X^T\beta_0))^2} \cdot \left(F'(X^T\beta_0)\cdot X\right)\cdot\left(F'(X^T\beta_0)\cdot X\right)^T \\
&= \mathbb{E}\frac{(F'(X^T\beta_0))^2}{F(X^T\beta_0)(1-F(X^T\beta_0))} \cdot XX^T
\end{aligned}$$

We want to find the "sufficient conditions" for A1 (to ensure that $Q(\beta)$ is uniquely minimized at β_0).

Example 4.7

Consider the example $F(t) = \frac{e^t}{1+e^t}$. The Hessian is

$$G(\beta) = \mathbb{E}\frac{\partial g(Y|X,\beta)}{\partial\beta} = \mathbb{E}\frac{\partial X\cdot(Y-F(X^T\beta))}{\partial\beta} = -\mathbb{E}F'(X^T\beta)X\cdot X^T$$

The sufficient condition for (A1) ($\mathbb{E}XX^T$ is positive definite) is $0 < \kappa \le F'(X^T\beta_0) \Leftrightarrow X^T\beta_0$ is not too large $\Leftrightarrow$ tails of $F'(X^T\beta)$ are not close to 0.

4.3.4 Example: Logistic Regression $F(t) = \frac{e^t}{1+e^t}$

Lemma 4.3

Given the link function $F(t) = \frac{e^t}{1+e^t}$,

$$F'(t) = \frac{e^t(1+e^t) - e^t \cdot e^t}{(1+e^t)^2} = \frac{e^t}{1+e^t} \cdot \frac{1}{1+e^t} = F(t) \cdot (1 - F(t))$$

It implies that

$$g(Y \mid X; \beta) = \left(Y - F(X^T\beta)\right) X$$

In this case, $w(X_i, \beta) = X_i$ doesn't depend on β.

The information matrix is

$$\mathcal{I}(\beta_0) = \mathbb{E}F(X^T\beta_0) \cdot (1 - F(X^T\beta_0)) \cdot XX^T$$

The asymptotic normality is

$$\sqrt{n}\left(\hat{\theta}_{MLE} - \theta_0\right) \overset{d}{\Rightarrow} N\left(0, [\mathcal{I}(\beta_0)]^{-1}\right)$$

The standard errors can be computed by

$$se(\hat{\theta}_{MLE}) = \text{diagonal}\left(\frac{1}{n}\hat{\mathcal{I}}(\theta_{MLE})^{-1}\right)^{\frac{1}{2}}$$

4.4 Large Sample Testing

Let $\mathcal{I} := \mathcal{I}(\theta_0)$. By the Corollary 4.1,

$$\sqrt{n}\left(\hat{\theta}_{\text{MLE}} - \theta_0\right) \overset{d}{\Rightarrow} N\left(0, \mathcal{I}^{-1}\right)$$
$$\sqrt{n}\hat{g}(\theta_0) \overset{d}{\Rightarrow} N\left(0, \mathcal{I}\right)$$

We want to test

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0$$

4.4.1 Wald Test: Distance on "x axis"

Definition 4.6 (Wald Test Statistic)

The test statistic is

$$W = n\left(\hat{\theta}_{\text{MLE}} - \theta_0\right)^T \hat{\mathcal{I}}\left(\hat{\theta}_{\text{MLE}} - \theta_0\right)$$

where $\hat{\mathcal{I}}$ is an estimator of $\mathcal{I}(\theta_0)$, $\hat{\mathcal{I}} := \mathcal{I}(\hat{\theta}_{\text{MLE}})^{-1}$.

Under H_0:

$$W \sim \chi^2(d), \text{ where } d = \dim(\theta)$$

The rejection region (RR) is RR $= \{W \geq C_{1-\alpha}\}$, where $C_{1-\alpha}$ is the $1-\alpha$ quantile of $\chi^2(d)$.

Proof 4.3

$\sqrt{n}\mathcal{I}^{\frac{1}{2}}\left(\hat{\theta}_{\text{MLE}} - \theta_0\right) \xrightarrow{d} N(0, I_d)$, where I_d is the identity matrix.

4.4.2 Lagrange Multiplier Test: Distance using "gradient"

Consider the optimization problem:

$$\max -\hat{Q}(\theta) \text{ s.t. } \theta = \theta_0$$

Note $\hat{g}(\theta) = -\frac{\partial \hat{Q}(\theta)}{\partial \theta}$. By the F.O.C.,

$$\left.\begin{aligned} \hat{g}(\hat{\theta}) + \lambda = 0 \\ \hat{\theta} = \theta_0 \end{aligned}\right\} \Rightarrow \hat{\lambda} = -\hat{g}(\theta_0)$$

Definition 4.7 (Lagrange Multiplier Test Statistic)

The Lagrange Multiplier test statistic is

$$\text{LM} = n\hat{g}(\theta_0)\mathcal{I}^{-1}\hat{g}(\theta_0), \text{ where } \mathcal{I}^{-1} \text{ is calculated by hypothetical value}$$

Under H_0:

$$W \sim \chi^2(d), \text{ where } d = \dim(\theta)$$

The rejection region (RR) is RR $= \{\text{LM} \geq C_{1-\alpha}\}$, where $C_{1-\alpha}$ is the $1-\alpha$ quantile of $\chi^2(d)$.

Proof 4.4

$\sqrt{n}\mathcal{I}^{-\frac{1}{2}}\hat{g}(\theta_0) \xrightarrow{d} N(0, I_d)$, where I_d is the identity matrix.

Note *In most distribution, $W \geq$ LM. (Use Wald if you want to reject.)*

4.4.3 Likelihood Ratio Test

For MLE, we also use the Likelihood Ratio test statistic is

$$\text{LR} = -2n\left(\hat{Q}(\theta_0) - \hat{Q}(\hat{\theta}_{\text{MLE}})\right) \geq 0$$

By Taylor expansion

$$\hat{Q}(\theta_0) - \hat{Q}(\hat{\theta}_{\text{MLE}}) = \underbrace{\frac{\partial}{\partial\theta}\hat{Q}(\hat{\theta}_{\text{MLE}})}_{=0}\left(\theta_0 - \hat{\theta}_{\text{MLE}}\right) + \frac{1}{2}\left(\theta_0 - \hat{\theta}_{\text{MLE}}\right)^T \frac{\partial^2}{\partial\theta^2}\hat{Q}(\theta)|_{\theta=\bar{\theta}}\left(\theta_0 - \hat{\theta}_{\text{MLE}}\right)$$

Then, under H_0

$$LR = -2n\left(\hat{Q}(\theta_0) - \hat{Q}(\hat{\theta}_{\text{MLE}})\right) = n\left(\theta_0 - \hat{\theta}_{\text{MLE}}\right)^T \underbrace{\left(-\frac{\partial^2}{\partial\theta^2}\hat{Q}(\theta)|_{\theta=\bar{\theta}}\right)}_{\triangleq\hat{\mathcal{I}}}\left(\theta_0 - \hat{\theta}_{\text{MLE}}\right) \sim \chi^2(d)$$

4.4.4 Wald is not invariant to parametrization

Consider the hypothesis $H_0 : \beta = 1$ vs. $H_1 : \beta \neq 1$ $(\beta > 0)$. The Wald test statistic is

$$W = n\left(\hat{\beta}_{\text{MLE}} - 1\right)^T \hat{\mathcal{I}}\left(\hat{\beta} - 1\right)$$

Parametrization: an equivalent form, $H_0 : \tau(\beta) = \tau(1)$ vs. $H_1 : \tau(\beta) \neq \tau(1)$ $(\beta > 0)$.

By first order continuously differentiable,

$$\tau(\hat{\beta}) - \tau(1) = \tau'(1)(\hat{\beta} - 1) + \frac{1}{2}\tau''(\bar{\beta})(\hat{\beta} - 1)^2$$
$$\sqrt{n}\left(\tau(\hat{\beta}) - \tau(1)\right) = \sqrt{n}\tau'(1)(\hat{\beta} - 1) + \sqrt{n}\frac{1}{2}\tau''(\bar{\beta})(\hat{\beta} - 1)^2$$

where $\bar{\beta} \in [1, \hat{\beta}]$. Then, under H_0:

$$\sqrt{n}\left(\tau(\hat{\beta}) - \tau(1)\right) \xrightarrow{d} N(0, \tau'(1)\text{Var}(\hat{\beta})\tau'(1))$$

4.5 Nonlinear Least Square

Suppose Y is the outcome and X are explanatory variables.

In previous "linear case," we use the form $\mathbb{E}[Y \mid X] = B(X)^T\beta, B(X) = [1, X, X^2, ...]$. Now, we consider a nonlinear expectation function

$$\mathbb{E}[Y \mid X] = \rho(X, \beta_0)$$

where ρ is known up to β and may not be linear in β

Example 4.8

1. Binary case, $\mathbb{E}[Y \mid X] = \Pr(Y = 1 \mid X)$ $Y \in \{0, 1\}$

$$Y \mid X \propto \text{Bernoulli}(\rho(X, \beta_0))$$

2. Exponential case, $\mathbb{E}[Y \mid X] = \lambda(X) := \exp(B(X)^T\beta)$

$$Y \mid X \propto \text{Poisson}(\lambda(X))$$

Consider the nonlinear expectation

$$\mathbb{E}[Y \mid X] = \rho(X, \beta_0) = \rho(B(X)^T\beta)$$

Then, a criterion function can be given

$$Q(\beta) = \mathbb{E}[Y - \rho(B(X)^T\beta)]^2, \quad Q(\beta) \geq 0, \forall\beta$$

Necessary: $\mathbb{E}[Y \mid X] = \operatorname{argmin}_f \mathbb{E}[Y - f(X)]^2$; We want to find the β_0 s.t. $\beta_0 = \operatorname{argmin} Q(\beta)$ (sufficiency).

The sample criterion function is

$$\hat{Q}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}[Y_i - \rho(B(X_i)^T\beta)]^2$$

The NLS estimator is given by

$$\hat{\beta}_{\text{NLS}} = \operatorname{argmin} \hat{Q}_n(\beta)$$

NLS estimator is also M-estimator, which satisfies consistency and asymptotic normality under some conditions (see Section 4.1).

Let $m(Z \mid \beta) = \frac{1}{2}(Y - \rho(B(X)^T\beta))^2$. The score function is

$$g(Z \mid \beta) = \frac{\partial \frac{1}{2}(Y - \rho(B(X)^T\beta))^2}{\partial\beta} = -\left[Y - \rho(B(X)^T\beta)\right]\rho'(B(X)^T\beta)B(X)$$

where $\mathbb{E}g(Z \mid \beta_0) = 0$ because $\mathbb{E}[Y|X] = \rho(B(X)^T\beta_0)$.

The Hessian matrix is given by

$$\begin{aligned} G(Z \mid \beta) = \frac{\partial}{\partial\beta^T}g(Z \mid \beta) = &-\left[Y - \rho(B(X)^T\beta)\right]\rho''(B(X)^T\beta)B(X)B(X)^T \\ &+ \rho'(B(X)^T\beta)\rho'(B(X)^T\beta)B(X)B(X)^T \end{aligned}$$

The Hessian matrix function at $\beta = \beta_0$ is

$$G = \mathbb{E}G(Z \mid \beta_0) = \mathbb{E}\left[(\rho'(B(X)^T\beta))^2 B(X)B(X)^T\right]$$

The variance of $g(Z \mid \beta)$ can be computed by Law of total variance,

$$\begin{aligned} \Omega = \text{Var}(g(Z \mid \beta)) &= \mathbb{E}_X\text{Var}(g(Z \mid \beta) \mid X) + \text{Var}\,\underbrace{\mathbb{E}[g(Z \mid \beta) \mid X]}_{=0} \\ &= \mathbb{E}\left[\left(Y - \rho(B(X)^T\beta)\right)^2\left(\rho'(B(X)^T\beta)\right)^2 B(X)B(X)^T\right] \end{aligned}$$

The asymptotic normality gives

$$\sqrt{n}\left(\hat{\beta}_{\text{NLS}} - \beta_0\right) \Rightarrow N\left(0, G^{-1}\Omega G^{-1}\right)$$

We can find the second information equality doesn't hold, $G \neq \Omega \Rightarrow G^{-1}\Omega G^{-1} \neq G^{-1}$.

Note *Second information equality gives $I = -G$ for maximization problem (e.g. MLE) and $I = G$ for minimization problem.*

4.5.1 Efficient NLS: Weighted NLS

In binary case, $m(Z \mid \beta) = \frac{1}{2}(Y - \rho(B(X)^T\beta))^2$ is the simplest criterion but $G \neq \Omega \Rightarrow$ NLS may not be efficient. The inefficiency can be fixed by

$$m_w(Z \mid \beta) = \frac{1}{2}w(x)(Y - \rho(B(X)^T\beta))^2$$

where $w(x)$ is a non-negative weight.

Claim 4.1

$\beta_0 = \operatorname{argmin} Q_w(\beta) := \frac{1}{2}\mathbb{E}w(x)(Y - \rho(B(X)^T\beta))^2$

Proof 4.5

Notice that by definition

$$\rho(B(X)^T\beta_0) := \mathbb{E}[Y \mid X = x] = \operatorname*{argmin}_{f(x)} \mathbb{E}[(Y - f(x))^2 \mid X = x]$$

Then,

$$\beta_0 = \operatorname*{argmin}_{\beta} \mathbb{E}[Y - \rho(B(X)^T\beta) \mid X]w(x)$$

$$\Rightarrow \beta_0 = \operatorname*{argmin}_{\beta} \int_x \mathbb{E}[Y - \rho(B(X)^T\beta) \mid X]w(x)f_X(x)dx$$

Claim 4.2

Optimal weight $w^*(x) = \frac{1}{\text{Var}(Y|X)} = \frac{1}{\rho(B(X)^T\beta)(1-\rho(B(X)^T\beta))}$

Proof 4.6

$$\begin{aligned}
Q_w(\beta) &:= \frac{1}{2}\mathbb{E}w(X)(Y - \rho(B(X)^T\beta))^2 \\
G_w &= \mathbb{E}\left[w(X)(\rho'(B(X)^T\beta))^2 B(X)B(X)^T\right] \\
\Omega_w &= \mathbb{E}\left[w^2(X)\left(Y - \rho(B(X)^T\beta)\right)^2 \left(\rho'(B(X)^T\beta)\right)^2 B(X)B(X)^T\right]
\end{aligned}$$

The efficient choice of $w^*(x)$ is to make $G_w = \Omega_w$

$$w^*(X) = \frac{1}{\mathbb{E}\left(Y - \rho(B(X)^T\beta) \mid X\right)^2} = \frac{1}{\text{Var}(Y \mid X)}$$

Two-Step NLS

1. Estimate $\hat{\beta}_{\text{NLS}}$ by (regular) NLS.
2. Estimate $\hat{\beta}_{\text{WNLS}}$ by

$$\hat{\beta}_{\text{WNLS}} = \operatorname{argmin} \sum_{i=1}^{n} \frac{(Y - \rho(B(X)^T\beta))^2}{\rho(B(X)^T\beta)(1 - \rho(B(X)^T\beta))}$$

4.6 (Linear) Quantile Regression

Let $\tau \in (0,1)$ be the quantile level and the τ'th quantile $q_Y(\tau) \in \mathbb{R}$ is defined as

$$F_Y(q_Y(\tau)) = \tau$$

Given $Y \sim F_Y$ (CDF, continuous without point mass), we construct a criterion $Q(\tau)$ such that

$$q_Y(\tau) = \underset{q}{\operatorname{argmin}}\, Q(q) := \mathbb{E}\rho_\tau(Y-q)$$

where $\rho_\tau(\cdot)$ is the check function defined as

$$\rho_\tau(u) := \{(1-\tau)\mathbf{1}\{u<0\} + \tau\mathbf{1}\{u>0\}\}|u|$$

4.6.1 Linear Quantile Regression Model

Given (Y,X), let $F_{Y|X}(y \mid x)$ be the conditional CDF, which is strictly monotone a.s. in X (for all values of X).

Define $Q_{Y|X}(\tau \mid x)$ be the conditional quantile, where

$$F_Y(Q_{Y|X}(\tau \mid x)) = \tau \text{ a.s. in } X$$

Definition 4.8 (Linear Quantile Regression Model (LQR))

$Q_{Y|X}(\tau \mid x) = X^T\beta_0(\tau)$

Consider

$$Y = X^T\gamma_0 + \epsilon$$

where ϵ is independent of X (not $\mathbb{E}[\epsilon|X]=0$, which is too weak).

Assumption *(Independence) ϵ is independent of X (stronger than $\mathbb{E}[\epsilon|X]=0$).*

Lemma 4.4 (By Independence)

$Q_{\epsilon|X}(\tau|X) = Q_\epsilon(\tau)$ a.s. in X

Proof 4.7

$$\begin{aligned} F_{\epsilon,X}(\epsilon,X) = F_\epsilon(\epsilon)F_X(X) \Rightarrow F_{\epsilon|X}(\epsilon|X) &= F_\epsilon(\epsilon) \\ \Rightarrow Q_\epsilon(\tau) = F_\epsilon^{-1}(\epsilon) &= Q_{\epsilon|X}(\tau|X) \end{aligned}$$

Lemma 4.5 (Equivalence Property)

Let $T : \mathbb{R} \to \mathbb{R}$ be an increasing function. Then

$$Q_{T(Y)}(\tau) = T(Q_Y(\tau))$$

Example 4.9

The $T(\cdot)$ can be $T(y) = \min\{y, L\}, T(y) = ay + b$.

Proof 4.8

Given T is strictly increasing,

$$\begin{aligned} \tau &= \Pr(Y < Q_Y(\tau)) \\ &= \Pr(T(Y) < T(Q_Y(\tau))) \\ &= F_{T(Y)}(T(Q_Y(\tau))) \\ \Rightarrow Q_{T(Y)}(\tau) &= T(Q_Y(\tau)) \end{aligned}$$

The quantile form of the LQR model is

$$Q_{Y|X}(\tau|X) = X^T\beta_0 + Q_\epsilon(\tau|X) = X^T\beta_0(\tau) \tag{4.2}$$

as $X = (1, X_1, ..., X_n)$, where

$$\begin{aligned} (\beta_0(\tau))_1 &= (\beta_0)_1 + Q_\epsilon(\tau) \\ (\beta_0(\tau))_{2:d} &= (\beta_0)_{2:d} \end{aligned}$$

Example 4.10 (Location-Scale Model)

$Y = X^T\gamma_0 + (X^T\delta_0)\epsilon$, where $X^T\delta_0 > 0$ a.s. in X. Then,

$$\begin{aligned} Q_{Y|X}(\tau|X) &= Q_{\epsilon|X}(\tau|X)(X^T\delta_0) + X^T\gamma_0 \\ \text{(by independence)} &= X^T(Q_\epsilon(\tau)\delta_0) + X^T\gamma_0 \\ &= X^T\beta_0(\tau) \end{aligned}$$

where $\beta_0(\tau) = Q_\epsilon(\tau)\delta_0 + \gamma_0$.

4.6.2 Quantile Causal Effects

$Z = (D, Y)$, there is no covariate X for now.

$$Y = h(D, u)$$

where $D \in \{0,1\}$ is binary treatment and $u \in \mathbb{R}$ is unobservable.

The treatment effect is

$$Y(1) - Y(0) = h(1,u) - h(0,u)$$

Suppose $D \perp (Y(1), Y(0))$ by random assignment. ATE $= \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0]$.

Instead of considering the ATE, we care about the τ-quantile of TE

$$Q_{Y(1)-Y(0)}(\tau)$$

It can be identified without further assumptions

Assumption

A1. $D \perp (Y(1), Y(0))$

A2. $h(1,u)$ *and* $h(0,u)$ *are increasing in* u.

A3. $h(1,u) - h(0,u)$ *is also increasing in* u.

Theorem 4.4

If these three assumptions hold,

$$Q_{Y(1)-Y(0)}(\tau) = Q_{Y|D=1}(\tau) - Q_{Y|D=0}(\tau)$$

Proof 4.9

$$\begin{aligned} Q_{Y(1)-Y(0)}(\tau) &= Q_{h(1,u)-h(0,u)}(\tau) \\ \text{(By equivalence property 4.5 and A3)} \quad &= h(1, Q_u(\tau)) - h(0, Q_u(\tau)) \\ \text{(By equivalence property 4.5 and A2)} \quad &= Q_{h(1,u)}(\tau) - Q_{h(0,u)}(\tau) \\ &= Q_{Y|D=1}(\tau) - Q_{Y|D=0}(\tau) \end{aligned}$$

With covariate X, the assumptions needed for identification change to

Assumption

A1. $D \perp (Y(1), Y(0)) \mid X$

A2. $h(1,x,u)$ *and* $h(0,x,u)$ *are increasing in* u *for each* x.

A3. $h(1,x,u) - h(0,x,u)$ *is also increasing in* u *for each* x.

4.6.3 M-estimator of Quantile

We want to find the optimal function $q(\cdot)$ that minimizes the criterion function given any data x.

$$Q_{Y|X}(\tau \mid x) = \underset{q}{\operatorname{argmin}}\, \mathbb{E}[\rho_\tau(Y - q) \mid X = x], \forall x$$

$$\Leftrightarrow Q_{Y|X}(\tau \mid x) = \underset{q(x)}{\operatorname{argmin}}\, \mathbb{E}\rho_\tau(Y - q(x))$$

The linearity form (4.2) motivates the following M-estimator loss function

$$\begin{aligned} m(Z, b) &= \rho_\tau(Y - X^T b) \\ &= \{(1-\tau)\mathbf{1}\{Y - X^T b < 0\} + \tau\mathbf{1}\{Y - X^T b > 0\}\}|Y - X^T b| \end{aligned}$$

The problem is

$$\beta_0(\tau) = \underset{b}{\operatorname{argmin}}\, \mathbb{E}[\rho_\tau(Y - X^T b)]$$

The score function is

$$g(Z, b) = -(\tau - \mathbf{1}\{Y - X^T b \le 0\})X$$

By the definition of quantile,

$$\begin{aligned} \mathbb{E}_{Y|X}[g(Z, \beta_0(\tau)) \mid X] &= -\mathbb{E}_{Y|X}[(\tau - \mathbf{1}\{Y - X^T\beta_0(\tau) \le 0\})X] \\ &= -(\tau - F_{Y|X}(X^T\beta_0(\tau)|X))X = 0 \end{aligned}$$

Since $\mathbf{1}\{Y - X^T\beta_0(\tau) \le 0\} \mid X \sim \text{Bernoulli}(\tau)$,

$$\text{Var}_{Y|X}[g(Z, \beta_0(\tau)) \mid X] = \text{Var}_{Y|X}[-(\tau - \mathbf{1}\{Y - X^T\beta_0(\tau) \le 0\})X] = \tau(1-\tau)XX^T$$

By the Law of Total Variance, the variance of $g(Z, \beta_0(\tau))$ is

$$\begin{aligned} \Omega_\tau :&= \text{Var}(g(Z, \beta_0(\tau))) \\ &= \mathbb{E}[\text{Var}(g(Z, \beta_0(\tau))|X)] + \text{Var}(\mathbb{E}[g(Z, \beta_0(\tau))|X]) \\ &= \tau(1-\tau)\mathbb{E}[XX^T] + 0 \\ &= \tau(1-\tau)\mathbb{E}[XX^T] \end{aligned}$$

The Hessian matrix at $\beta_0(\tau)$ is

$$G_\tau = \mathbb{E}_X \frac{\partial g(Z, \beta_0(\tau))}{\partial \beta_0(\tau)^T} = \mathbb{E}_X \frac{\partial - (\tau - F_{Y|X}(X^T\beta_0(\tau)|X))X}{\partial \beta_0(\tau)^T} = \mathbb{E}_X\left[f_{Y|X}(X^T\beta_0(\tau)|X)XX^T\right]$$

By the consistency and asymptotic normality,

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N\left(0, G^{-1}\Omega G^{-1}\right)$$

$G^{-1}\Omega G^{-1}$ can be estimated by $\hat{G}^{-1}\hat{\Omega}\hat{G}^{-1}$, where $\hat{G} := \frac{1}{n}\sum_{i=1}^n \hat{f}_{Y|X}\left(Y_i - X_i^T(\hat{\beta}(\tau))\right) X_i X_i^T$ can be estimated by *kernel density estimation* and $\hat{\Omega} := \tau(1-\tau)\frac{1}{n}\sum_{i=1}^n X_i X_i^T$.

Efficient sample size is τn.

4.7 Example of M-estimator

4.7.1 Optimal Weighting Example: $y_i = x_i^{\beta_0} + \epsilon_i$

An empirical study uses a nonlinear regression equation

$$y_i = x_i^{\beta_0} + \epsilon_i$$

where it is assumed that $\mathbb{E}[\epsilon_i \mid x_i] = 0$ and $\mathrm{Var}[\epsilon_i \mid x_i] = \sigma^2(x_i)$. The regressor x_i has support on $[c, \frac{1}{c}]$ for some small positive c.

(a).

Given a random sample of size n from this model, the NLS estimator is given by

$$\hat{\beta}_{\text{NLS}} = \operatorname*{argmin}_{\beta} \hat{Q}_n(\beta) := \frac{1}{n}\sum_{i=1}^{n} m(x_i, y_i \mid \beta)$$

where $m(X, Y \mid \beta) = \frac{1}{2}(Y - X^{\beta})^2$. The score function is

$$g(X, Y \mid \beta) = -\ln X \cdot X^{\beta}(Y - X^{\beta})$$

where $\mathbb{E}g(X, Y \mid \beta_0) = 0$ and the variance is

$$\Omega = \mathbb{E}[g(X, Y \mid \beta_0)^2] = \mathbb{E}[(\ln X)^2 X^{2\beta_0}(Y - X^{\beta_0})^2] = \mathbb{E}[(\ln X)^2 X^{2\beta_0}\sigma^2(X)]$$

The Hessian matrix is

$$G(X, Y \mid \beta) = \frac{\partial g(X, Y \mid \beta)}{\partial \beta} = -(\ln X)^2 X^{2\beta}(Y - X^{\beta}) + (\ln X)^2 X^{2\beta}$$

The Hessian matrix function at $\beta = \beta_0$ is

$$G = \mathbb{E}G(X, Y \mid \beta_0) = \mathbb{E}[(\ln X)^2 X^{2\beta_0}]$$

On the premise that identification and the relevant regularity conditions are satisfied, we know from our results on M-estimators and our discussion of NLS in lecture that:

$$\sqrt{n}\left(\hat{\beta}_{\text{NLS}} - \beta_0\right) \Rightarrow N\left(0, G^{-1}\Omega G^{-1}\right)$$

The consistent estimator of its asymptotic variance

$$\hat{G}^{-1}\hat{\Omega}\hat{G}^{-1} = \frac{\frac{1}{N}\sum_{i=1}^{n}(\ln x_i)^2 (x_i)^{2\hat{\beta}_{\text{NLS}}}\left(y_i - x_i^{\hat{\beta}_{\text{NLS}}}\right)^2}{\left(\frac{1}{N}\sum_{i=1}^{N}(\ln x_i)^2 (x_i)^{2\hat{\beta}_{\text{NLS}}}\right)^2}$$

(b).

The nonlinear weighted least-squares (NL-WLS) estimator is defined as the minimizer of

$$\hat{Q}_n(b) := \frac{1}{n}\sum_{i=1}^{n} w(x_i)(y_i - x_i^b)^2$$

the optimal weighting function $w^*(X) = \frac{1}{\text{Var}(Y|X)} = \frac{1}{\sigma^2(X)}$. The asymptotic variance of NL-WLS is

$$\frac{1}{\mathbb{E}[w^*(X)(\ln X)^2 X^{2\beta_0}]} = \frac{1}{\mathbb{E}[\frac{1}{\sigma^2(X)}(\ln X)^2 X^{2\beta_0}]}$$

4.7.2 Conditional Beta Distribution $Beta(\alpha, 1)$

A random variable U is said to have a $Beta(\alpha, 1)$ distribution with parameter $\alpha > 0$, denotes as $U \sim Beta(\alpha, 1)$ if it is continuously distributed on the interval $(0, 1)$ with c.d.f.

$$\Pr(U \le u) \equiv F(u; \alpha) = \min\{\max\{0, u^\alpha\}, 1\}$$

and p.d.f.

$$f(u; \alpha) = \alpha u^{\alpha-1}$$

for $u \in (0, 1)$ and zero elsewhere. For this distribution, $\mathbb{E}[U^j] = \frac{\alpha}{\alpha+j}$ and $\mathbb{E}[\log(U)] = -\frac{1}{\alpha}$, $\text{Var}[\log(U)] = \frac{1}{\alpha^2}$. Suppose you have a random sample $\{(y_i; x_i)\}_{i=1}^N$ where the conditional distribution of y_i given the p-dimensional vector x_i is $Beta(\alpha_i, 1)$ with parameter $\alpha_i = x_i^T\beta_0$, i.e.,

$$y_i \mid x_i \sim Beta(x_i^T\beta_0, 1)$$

Also, suppose the marginal distribution of the p-dimensional regressors x_i is unspecified and, as usual, β_0 is unknown, except that the distribution of x_i and the (compact) parameter space $B \ni \beta_0$ has $\Pr(x_i^T\beta \ge c)$ for some small positive number c and for all $\beta \in B$.

> **A.** Derive the average log-likelihood function $L_N(\beta)$ for this problem, and show that the first-order condition for the maximum likelihood (ML) estimator $\hat{\beta}$ can be rewritten in the form
>
> $$0 = \frac{1}{N}\sum_{i=1}^{N} u_i(\hat{\beta}) \cdot x_i$$
>
> for some "pseudo-residual" function $u_i(\beta)$ which satisfies $\mathbb{E}[u_i(\beta_0) \mid x_i] = 0$.

The average log-likelihood takes the form

$$L_N(\beta) = \frac{1}{N}\sum_{i=1}^{N} \log f(y_i; x_i^T\beta) = \frac{1}{N}\sum_{i=1}^{N}[(x_i^T\beta - 1)\log y_i + \log x_i^T\beta]$$

and the (interior) ML estimator $\hat{\beta}$ satisfies the first-order condition

$$0 = \frac{\partial L_N(\hat{\beta})}{\partial \beta} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{x_i^T\hat{\beta}} + \log y_i\right] x_i = \frac{1}{N}\sum_{i=1}^{N} u_i(\hat{\beta}) \cdot x_i$$

Here $u_i(\beta) \equiv (x_i^T\beta)^{-1} + \log(y_i)$ satisfies

$$\mathbb{E}[u_i(\beta_0) \mid x_i] = \frac{1}{x_i^T\beta_0} + \mathbb{E}[\log y_i \mid x_i] = \frac{1}{x_i^T\beta_0} - \frac{1}{x_i^T\beta_0} = 0$$

B. Derive an expression for the asymptotic distribution of the ML estimator $\hat{\beta}$, including an expression for its asymptotic covariance matrix.

The Hessian of the average log-likelihood is

$$\frac{\partial^2 L_N(\beta)}{\partial\beta\partial\beta^T} = -\frac{1}{N}\sum_{i=1}^N \left(\frac{1}{x_i^T\beta}\right)^2 x_i x_i^T$$

The second information equality holds in MLE, $I_0 = -\mathbb{E}\left[\frac{\partial^2 L_N(\beta_0)}{\partial\beta\partial\beta^T}\right] = \mathbb{E}\left[\left(\frac{1}{x_i^T\beta}\right)^2 x_i x_i^T\right]$.

$$\sqrt{N}(\hat{\beta}_{\text{ML}} - \beta_0) \xrightarrow{a} N(0, I_0^{-1})$$

C. Suppose that the first component of x_i is one and you want to test the null hypothesis that the intercept of β_0 is one and the slope coefficient are all zero, i.e.,

$$H_0 : \beta_0 = (1, 0, ..., 0)^T$$

which implies that $x_i^T\beta_0 = 1$ (w.p.1) and that y_i has a Unif$(0,1)$ distribution, independent of x_i. Derive the form of the Lagrange multiplier or "Score" (LM) test statistic for this null hypothesis, and state the form of its asymptotic distribution under H_0.

Under H_0, the regression function $x_i^T\beta_0 = 1$,

$$\hat{g}(\theta_0) = \frac{\partial L_N(\beta_0)}{\partial\beta} = \frac{1}{N}\sum_{i=1}^N (1 + \log y_i)\, x_i$$

$$\hat{I} = -\frac{\partial^2 L_N(\beta_0)}{\partial\beta\partial\beta^T} = \frac{1}{N}\sum_{i=1}^N x_i x_i^T$$

Then, the LM test statistic is

$$LM = n\hat{g}(\theta_0)\hat{I}^{-1}\hat{g}(\theta_0) = \left(\sum_{i=1}^N (1 + \log y_i)\, x_i\right)\left(\sum_{i=1}^N x_i x_i^T\right)^{-1}\left(\sum_{i=1}^N (1 + \log y_i)\, x_i\right)$$

Under H_0, $LM \sim \chi^2_p$.

D. For general $K \geq 1$, we can use NLLS instead of MLE based on the known parametric form of $\mathbb{E}[y_i \mid x_i]$: $\mathbb{E}[y_i \mid x_i] = \frac{x_i^T\beta_0}{x_i^T\beta_0 + 1}$.

Let $m(Z,\beta) = \frac{1}{2}\left(Y - \frac{X^T\beta}{X^T\beta+1}\right)^2$. The score function is

$$g(Z,\beta) = -\left(Y - \frac{X^T\beta}{X^T\beta + 1}\right)\frac{X}{(X^T\beta+1)^2}$$

The Hessian at β_0 is

$$G = \mathbb{E}_Z\left[\frac{\partial g(Z,\beta_0)}{\partial\beta}\right] = \mathbb{E}\left[\frac{XX^T}{(X^T\beta+1)^4}\right]$$

and the variance of the score at β_0 is

$$\begin{aligned}
\Omega = \mathrm{Var}_Z[g(Z,\beta_0)] &= \mathbb{E}_X\left[\mathrm{Var}_{Y|X}[g(Z,\beta_0)\mid X]\right] + \mathrm{Var}_X\mathbb{E}_{Y|X}[g(Z,\beta_0)|X] \\
&= \mathbb{E}_X\left[\mathrm{Var}_{Y|X}\left[-\left(Y-\frac{X^T\beta_0}{X^T\beta_0+1}\right)\frac{X}{(X^T\beta_0+1)^2}\right]\right]+0 \\
&= \mathbb{E}_X\left[\left(\frac{X^T\beta}{X^T\beta+2}-\left(\frac{X^T\beta}{X^T\beta+1}\right)^2\right)\frac{XX^T}{(X^T\beta_0+1)^4}\right] \\
&= \frac{X^T\beta XX^T}{(X^T\beta+2)(X^T\beta_0+1)^6}
\end{aligned}$$

where $\mathrm{Var}_{Y|X}[Y|X] = \mathbb{E}[Y^2|X]-\mathbb{E}^2[Y|X] = \frac{X^T\beta}{X^T\beta+2}-\left(\frac{X^T\beta}{X^T\beta+1}\right)^2$.

Then, $\sqrt{n}(\hat{\beta}_{NLLS}-\beta_0) \overset{d}{\Rightarrow} N(0, G^{-1}\Omega G^{-1})$.

4.7.3 "Two-Sided" Censored Regression Model

Consider a censored regression model in which a linear latent dependent variable $y_i^* = x_i^T\beta+\epsilon_i$ is related to an observable dependent variable y_i by the following transformation:

$$y_i = \max\{c, |y_i^*|\}\mathrm{sgn}\{y_i^*\}$$

where $c > 0$ is a known constant, i.e.,

$$y_i = \begin{cases} x_i^T\beta+\epsilon_i, & \text{if } x_i^T\beta+\epsilon_i < -c \\ -c, & \text{if } -c \le x_i^T\beta+\epsilon_i \le 0 \\ c, & \text{if } 0 < x_i^T\beta+\epsilon_i \le c \\ x_i^T\beta+\epsilon_i, & \text{if } c < x_i^T\beta+\epsilon_i \end{cases}$$

A. Assuming the error term ϵ_i is normally distributed, $\epsilon_i \sim N(0,\sigma^2)$, and is independent of the regressor vector x_i, derive an expression for the average log likelihood function $L_n(\beta,\sigma^2)$ for a sample of N i.i.d. observations on y_i and x_i.

Given $y_i = c$, $P(y_i = c\mid x_i;\beta,\sigma^2) = \Pr(\epsilon_i \in (-x_i^T\beta, c-x_i^T\beta]) = \Phi\left(\frac{c-x_i^T\beta}{\sigma}\right)-\Phi\left(\frac{-x_i^T\beta}{\sigma}\right)$.

Given $y_i = -c$, $P(y_i = -c\mid x_i;\beta,\sigma^2) = \Pr(\epsilon_i \in [-c-x_i^T\beta, -x_i^T\beta]) = \Phi\left(\frac{-x_i^T\beta}{\sigma}\right)-\Phi\left(\frac{-c-x_i^T\beta}{\sigma}\right)$.

Given $|y_i| > c$, $Y\mid x_i \sim N(x_i^T\beta,\sigma^2) \Rightarrow \frac{Y-x_i^T\beta}{\sigma}\mid x_i \sim N(0,1)$,

$$\Pr(Y\le y\mid x_i;\beta,\sigma^2) = \Pr\left(\frac{Y-x_i^T\beta}{\sigma}\le\frac{y-x_i^T\beta}{\sigma}\mid x_i;\beta,\sigma^2\right) = \Phi\left(\frac{y-x_i^T\beta}{\sigma}\right)$$

$$\Rightarrow p(y_i\mid x_i;\beta,\sigma^2) = \frac{\partial\Phi\left(\frac{y-x_i^T\beta}{\sigma}\right)}{\partial y}\Big|_{y=y_i} = \frac{1}{\sigma}\phi\left(\frac{y_i-x_i^T\beta}{\sigma}\right)$$

So, the average log likelihood function is

$$L_N(\beta,\sigma^2) = \frac{1}{N}\sum_{i=1}^{N}\left\{\mathbf{1}\{y_i = c\}\log\left(\Phi\left(\frac{c - x_i^T\beta}{\sigma}\right) - \Phi\left(\frac{-x_i^T\beta}{\sigma}\right)\right)\right.$$
$$\left. + \mathbf{1}\{y_i = -c\}\log\left(\Phi\left(\frac{-x_i^T\beta}{\sigma}\right) - \Phi\left(\frac{-c - x_i^T\beta}{\sigma}\right)\right) + \mathbf{1}\{|y_i| > c\}\log\left(\frac{1}{\sigma}\phi\left(\frac{y_i - x_i^T\beta}{\sigma}\right)\right)\right\}$$

B. Now suppose that you only observe x_i and $D_i = \mathbf{1}\{y_i \neq c\}$, and do not observe y_i, again derive the average log likelihood function.

The binary response model is

$$\begin{aligned}\Pr\{D_i = 1 \mid x_i\} &= 1 - \Pr\{y_i = c \mid x_i\} \\ &= 1 - \left[\Phi\left(\frac{c - x_i^T\beta}{\sigma}\right) - \Phi\left(\frac{-x_i^T\beta}{\sigma}\right)\right]\end{aligned}$$

Then, the average log likelihood function is

$$L_N(\beta,\sigma^2) = \frac{1}{N}\sum_{i=1}^{N}\left\{D_i\log\left(1 - \Phi\left(\frac{c - x_i^T\beta}{\sigma}\right) + \Phi\left(\frac{-x_i^T\beta}{\sigma}\right)\right)\right.$$
$$\left. +(1 - D_i)\log\left(\Phi\left(\frac{c - x_i^T\beta}{\sigma}\right) - \Phi\left(\frac{-x_i^T\beta}{\sigma}\right)\right)\right\}$$

4.7.4 Regression Example: $y_i = exp\{x_i^T\beta_0\} + \epsilon_i$

Consider $y_i = exp\{x_i^T\beta_0\} + \epsilon_i$, where $\mathbb{E}[\epsilon_i \mid x_i] = 0$ and $\mathrm{Var}[\epsilon_i \mid x_i] = \sigma^2(x_i)$.
$\sqrt{n}\left(\hat{\beta}_{\text{NLS}} - \beta_0\right) \Rightarrow N\left(0, G^{-1}\Omega G^{-1}\right)$, where $g(z_i,\beta) = -(y_i - exp\{x_i^T\beta\})exp\{x_i^T\beta\}x_i$, $G = \mathbb{E}\left[\frac{\partial g(x_i,\beta_0)}{\partial\beta}\right] = \mathbb{E}[exp\{2x_i^T\beta_0\}x_ix_i^T]$ and $\Omega = \mathrm{Var}(g(x_i,\beta_0)) = \mathbb{E}[(y_i - exp\{x_i^T\beta_0\})^2 exp\{2x_i^T\beta_0\}x_ix_i^T] = \mathbb{E}[\sigma^2(x_i)exp\{2x_i^T\beta_0\}x_ix_i^T]$.

4.7.5 Regression Example: $y_i = (\beta_0)^{x_i} + \epsilon_i$

Consider $y_i = exp\{x_i^T\beta_0\} + \epsilon_i$, where $\mathbb{E}[\epsilon_i \mid x_i] = 0$ and $\mathrm{Var}[\epsilon_i \mid x_i] = \sigma^2(x_i)$.
$\sqrt{n}\left(\hat{\beta}_{\text{NLS}} - \beta_0\right) \Rightarrow N\left(0, G^{-1}\Omega G^{-1}\right)$, where $g(z_i,\beta) = -(y_i - (\beta)^{x_i})\beta^{x_i-1}x_i$, $G = \mathbb{E}\left[\frac{\partial g(x_i,\beta_0)}{\partial\beta}\right] = \mathbb{E}[x_i^2(\beta_0)^{2(x_i-1)}]$ and $\Omega = \mathrm{Var}(g(x_i,\beta_0)) = \mathbb{E}[(y_i - (\beta_0)^{x_i})^2\beta_0^{2(x_i-1)}x_i^2] = \mathbb{E}[\sigma^2(x_i)\beta_0^{2(x_i-1)}x_i^2]$.

4.7.6 Regression Example: $y_i = \log(x_i^T\beta_0) + \epsilon_i$

Consider $y_i = \log(x_i^T\beta_0) + \epsilon_i$, where $\mathbb{E}[\epsilon_i \mid x_i] = 0$ and $\mathrm{Var}[\epsilon_i \mid x_i] = \tau_0^2(x_i^T\beta_0)^2$.
$\sqrt{n}\left(\hat{\beta}_{\text{NLS}} - \beta_0\right) \Rightarrow N\left(0, G^{-1}\Omega G^{-1}\right)$, where $g(z_i,\beta) = -\frac{y_i - \log(x_i^T\beta_0)}{x_i^T\beta}x_i$, $G = \mathbb{E}\left[\frac{\partial g(x_i,\beta_0)}{\partial\beta}\right] = \mathbb{E}[\frac{x_ix_i^T}{(x_i^T\beta_0)^2}]$ and $\Omega = \mathrm{Var}(g(z_i,\beta_0)) = \mathbb{E}[\left(\frac{y_i - \log(x_i^T\beta_0)}{x_i^T\beta}\right)^2 x_ix_i^T] = \tau_0^2\mathbb{E}[x_ix_i^T]$. The optimal weight in WNLS is $w^*(x_i) = \frac{1}{\mathrm{Var}(y_i|x_i)} = \frac{1}{\tau_0^2\cdot(x_i^T\beta_0)^2}$, which gives asymptotic variance of the optimal NLWLS estimator $(\Omega^*)^{-1} = \tau_0^2\left[\mathbb{E}\left(\frac{x_ix_i^T}{(x_i^T\beta_0)^4}\right)\right]^{-1}$.

4.7.7 $Geometric(exp\{x_i^T\beta_0\})$ Distribution: $f(y; x_i^T\beta_0) = (1 - exp\{x_i^T\beta_0\})exp\{y \cdot (x_i^T\beta_0)\}$

$f(y; x_i^T\beta_0) = (1 - exp\{x_i^T\beta_0\})exp\{y \cdot (x_i^T\beta_0)\}$, where $x_i^T\beta < 0$ with probability one for all possible values of β; but the distribution of x_i does not otherwise involve β. We have $\mathbb{E}[y; x_i^T\beta_0] = \frac{exp\{x_i^T\beta_0\}}{1-exp\{x_i^T\beta_0\}}$ and $\text{Var}[y; x_i^T\beta_0] = \frac{exp\{x_i^T\beta_0\}}{(1-exp\{x_i^T\beta_0\})^2}$.

The average conditional log-likelihood function can be written as

$$L_N(\beta) = \frac{1}{N}\sum_{i=1}^{N}\log f(y_i; x_i^T\beta) = \frac{1}{N}\sum_{i=1}^{N}[\log(1 - exp\{x_i^T\beta\}) + y_i(x_i^T\beta)]$$

and the (interior) ML estimator $\hat{\beta}$ satisfies the first-order condition

$$0 = \frac{\partial L_N(\hat{\beta})}{\partial \beta} = \frac{1}{N}\sum_{i=1}^{N}\left[y_i - \frac{exp\{x_i^T\beta\}}{1 - exp\{x_i^T\beta\}}\right]x_i = \frac{1}{N}\sum_{i=1}^{N}u_i(\hat{\beta}) \cdot x_i$$

Here $u_i(\beta) \equiv y_i - \frac{exp\{x_i^T\beta\}}{1-exp\{x_i^T\beta\}}$ satisfies

$$\mathbb{E}[u_i(\beta_0) \mid x_i] = \mathbb{E}[y_i \mid x_i] - \frac{exp\{x_i^T\beta_0\}}{1 - exp\{x_i^T\beta_0\}} = 0$$

The score function is $g(z_i, \beta) = \left[y_i - \frac{exp\{x_i^T\beta\}}{1-exp\{x_i^T\beta\}}\right]x_i$. The fisher information matrix is

$$I_0 = \text{Var}_Z\left[g(z_i, \beta_0)\right] = \mathbb{E}_X\text{Var}_{Y|X}\left[g(z_i, \beta_0) \mid x_i\right] + \text{Var}_X\mathbb{E}_{Y|X}[g(z_i, \beta_0) \mid x_i] = \mathbb{E}\left[\frac{exp\{x_i^T\beta_0\}}{(1 - exp\{x_i^T\beta_0\})^2}x_ix_i^T\right]$$

NLS form: Hessian $H_0 = \mathbb{E}\left[\left[\frac{exp\{x_i^T\beta_0\}}{(1-exp\{x_i^T\beta_0\})^2}\right]^2 x_ix_i^T\right]$, Variance $V_0 = \mathbb{E}\left[\left[\frac{exp\{x_i^T\beta_0\}}{(1-exp\{x_i^T\beta_0\})^3}\right]^2 x_ix_i^T\right]$.

4.7.8 $Exponential(x_i^T\beta_0)$ Distribution: $f(y; x_i^T\beta_0) = x_i^T\beta_0 \cdot exp\{-(x_i^T\beta_0)y\}$

$f(y; x_i^T\beta_0) = x_i^T\beta_0 \cdot exp\{-(x_i^T\beta_0)y\}$ where $\mathbb{E}[y_i; x_i^T\beta_0] = \frac{1}{x_i^T\beta_0}$ and $\text{Var}[y; x_i^T\beta_0] = \frac{1}{(x_i^T\beta_0)^2}$.

The average log-likelihood takes the form

$$L_N(\beta) = \frac{1}{N}\sum_{i=1}^{N}\log f(y_i; x_i^T\beta) = \frac{1}{N}\sum_{i=1}^{N}[\log(x_i^T\beta) - (x_i^T\beta \cdot y_i)]$$

and the (interior) ML estimator $\hat{\beta}$ satisfies the first-order condition

$$0 = \frac{\partial L_N(\hat{\beta})}{\partial \beta} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{x_i^T\hat{\beta}} - y_i\right]x_i = \frac{1}{N}\sum_{i=1}^{N}u_i(\hat{\beta}) \cdot x_i$$

Here $u_i(\beta) \equiv (x_i^T\beta)^{-1} - y_i$ satisfies

$$\mathbb{E}[u_i(\beta_0) \mid x_i] = \frac{1}{x_i^T\beta_0} - \mathbb{E}[y_i \mid x_i] = \frac{1}{x_i^T\beta_0} - \frac{1}{x_i^T\beta_0} = 0$$

The information matrix is given by

$$I_0 = \text{Var}_Z\left[\left[\frac{1}{x_i^T\beta_0} - y_i\right]x_i\right] = \mathbb{E}_X\text{Var}_{Y|X}[y_ix_i] + \text{Var}_X\mathbb{E}_{Y|X}\left[\left[\frac{1}{x_i^T\beta_0} - y_i\right]x_i\right] = \mathbb{E}\left[\frac{1}{(x_i^T\beta_0)^2} \cdot x_ix_i^T\right]$$

$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \overset{d}{\Rightarrow} N(0, I_0^{-1})$. The estimator of the variance is $\hat{I} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{(x_i^T\beta_0)^2} \cdot x_ix_i^T$.

$H_0 : \beta_0 = (1, 0, ..., 0)^T$ (all slope coefficients are zero, the intercept term is one). The LM test statistic is

$$LM = n\hat{g}(\theta_0)\hat{I}^{-1}\hat{g}(\theta_0) = \left(\sum_{i=1}^{N} (1 - y_i)\, x_i\right) \left(\sum_{i=1}^{N} x_i x_i^T\right)^{-1} \left(\sum_{i=1}^{N} (1 - y_i)\, x_i\right)$$

Under H_0, $LM \sim \chi^2_p$.

Chapter 5 Bootstrap

Bootstrap is a procedure to compute properties of an estimator by random re-sampling with replacement from the data. It was first introduced by Efron (1979).

Suppose we have i.i.d. sample $\vec{Y} = (Y_1, Y_2, ..., Y_n)$ taken i.i.d. from a distribution with cdf F and we want to compute a statistic θ of the distribution using an estimator $\hat{\theta}_n(\vec{Y})$. The distribution of the statistic θ has cdf G. While the estimator $\hat{\theta}_n(\vec{Y})$ may not be optimal in any sense, it is often the case that $\hat{\theta}_n(\vec{Y})$ is consistent in probability, i.e., $\hat{\theta}_n(\vec{Y}) \xrightarrow{p} \theta$ as $n \to \infty$. We want to analyze the performance of the estimartor $\hat{\theta}_n(\vec{Y})$ in terms of the following quantities:

(1). Bias:

$$\text{Bias}(\hat{\theta}_n) = \mathbb{E}_\theta[\hat{\theta}_n(\vec{Y})] - \theta$$

(2). Variance:

$$\text{Var}(\hat{\theta}_n) = \mathbb{E}_\theta[\hat{\theta}_n^2(\vec{Y})] - \mathbb{E}_\theta^2[\hat{\theta}_n(\vec{Y})]$$

(3). CDF:

$$G_n(t) = P(\hat{\theta}_n(\vec{Y}) < t), \forall t$$

5.1 Traditional Monte-Carlo Approach

Generate k vectors $\vec{Y}^{(i)}, i = 1, 2, ..., k$ (total kn random variables)

(1). Bias:

$$\widehat{\text{Bias}}(\hat{\theta}_n) = \frac{1}{k}\sum_{j=1}^{k} \hat{\theta}_n(\vec{Y}^{(j)}) - \theta$$

By the strong law of large number, the mean $\frac{1}{k}\sum_{j=1}^{k} \hat{\theta}_n(\vec{Y}^{(j)})$ converges almost surely to the expected value $\mathbb{E}_\theta[\hat{\theta}_n(\vec{Y})]$, so $\widehat{\text{Bias}}(\hat{\theta}_n) \xrightarrow{a.s.} \text{Bias}(\hat{\theta}_n)$.

(2). Variance:

$$\widehat{\text{Var}}(\hat{\theta}_n) = \frac{1}{k}\sum_{j=1}^{k} \hat{\theta}_n^2(\vec{Y}^{(j)}) - \left(\frac{1}{k}\sum_{j=1}^{k} \hat{\theta}_n(\vec{Y}^{(j)})\right)^2$$

Still by the strong law of large number, the mean $\frac{1}{k}\sum_{j=1}^{k} \hat{\theta}_n(\vec{Y}^{(j)})$ converges almost surely to the expected value $\mathbb{E}_\theta[\hat{\theta}_n(\vec{Y})]$ and the mean $\frac{1}{k}\sum_{j=1}^{k} \hat{\theta}_n^2(\vec{Y}^{(j)})$ converges almost surely to the expected value $\mathbb{E}_\theta[\hat{\theta}_n^2(\vec{Y})]$, so $\widehat{\text{Var}}(\hat{\theta}_n) \xrightarrow{a.s.} \text{Var}(\hat{\theta}_n)$.

(3). Empirical Distribution Function (CDF):

$$\hat{G}_n(t) = \frac{1}{k}\sum_{j=1}^{k} \mathbf{1}\{\hat{\theta}_n(\vec{Y}^{(j)}) < t\}, \forall t$$

By law of large numbers, we have $\hat{G}_n(x) \xrightarrow{a.s.} G_n(x), \forall t \in \mathbb{R}$ as $k \to \infty$.

By Glivenko-Cantelli Theorem, we have $\sup_{t\in\mathbb{R}|\mathbb{R}} \hat{G}_n(x) - G_n(x)| \xrightarrow{a.s.} 0$ as $k \to \infty$. (Stronger result).

5.2 Bootstrap (When data is not enough)

Suppose we only have data $\vec{Y} = (Y_1, ..., Y_n)$ and we can't draw new samples from the real distribution anymore. We reuse $Y_1, ...Y_n$ to obtain resamples $\vec{Y}^* = (Y_1^*, ..., Y_n^*)$ (drawing from $\{Y_1, ...Y_n\}$ uniformly, equivalently drawing from the empirical distribution with cdf $F_n(y) \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Y_i = y\}$). We get k resamples, denoted by $\vec{Y^*}^{(1)}, ..., \vec{Y^*}^{(k)}$.

1. Bias:

$$\text{Bias}^*(\hat{\theta}_n) \triangleq \frac{1}{k}\sum_{j=1}^{k} \hat{\theta}_n(\vec{Y^*}^{(j)}) - \theta$$

2. Variance:

$$\text{Var}^*(\hat{\theta}_n) \triangleq \frac{1}{k}\sum_{j=1}^{k} \hat{\theta}_n^2(\vec{Y^*}^{(j)}) - \left(\frac{1}{k}\sum_{j=1}^{k} \hat{\theta}_n(\vec{Y^*}^{(j)})\right)^2$$

3. CDF:

$$\hat{G}_n^*(t) = \frac{1}{k}\sum_{j=1}^{k} \mathbf{1}_{\hat{\theta}_n(\vec{Y^*}^{(j)})<t}, \forall t$$

Note *$\hat{G}_n^*(t)$ may not always converges to G_n as $n \to \infty$.*

Example 5.1 (Bootstrap Fail Example)

Suppose $Y \sim$ i.i.d. $[0, \theta]$ and consider the estimator $\hat{\theta}_n(\vec{Y}) = \max_i Y_i \triangleq Y_{(n)}$. Then, for all $t \geq 0$,

$$G_n(t) \to 1 - e^{-\frac{t}{\theta_F}} \text{ as } n \to \infty$$

But for all $t \geq 0$,

$$\hat{G}_n^*(t) \geq P_{F_n}(Y_{(n)} = Y_{(n)}^*) = 1 - (1 - \frac{1}{n})^n \to 1 - e^{-1} \text{ as } n \to \infty$$

5.3 Residual Bootstrap (for problem with not i.i.d. data)

The bootstrap principle is quite general and may also be used in problems where the data Y_i, $1 \leq i \leq n$, **are not i.i.d**.

5.3.1 Example: Linear

Consider the model

$$Y_i = a + bs_i + Z_i,\ i = 1, 2, ..., n$$

where $\theta = (a, b)$ is the parameter to be estimated, $\vec{s} = (s_1, ..., s_n)$ is a known signal, and $Z_i \sim \mathcal{N}(0, \sigma^2)$ (i.i.d.).

The Linear Least Square Estimator is

$$(\hat{a}_n, \hat{b}_n) = \underset{(a,b)}{\text{argmin}} \sum_{i=1}^{n} (Y_i - a - bs_i)^2$$

Given $\vec{Y}$ and estimator $\hat{\theta}_n = (\hat{a}_n, \hat{b}_n)$, define the residual errors (not i.i.d.)

$$E_i = Y_i - \hat{a}_n - \hat{b}_n s_i \approx Z_i$$

Then, we use bootstrap to generate k resamples of $\vec{E} = (E_1, E_2, ..., E_n)$.

For $j = 1, ..., k$, do the following:

1. Obtain $\vec{E^*}^{(j)}$ by uniformly resampling from $\vec{E}$.
2. Compute pseudo-data $Y_i^{*(j)} = \hat{a}_n + \hat{b}_n s_i + E_i^{*(j)}$ for $1 \leq i \leq n$.
3. Compute LS estimator to the pseudo-data

$$\hat{\theta}_n^{(j)} = (\hat{a}_n^{(j)}, \hat{b}_n^{(j)}) = \underset{(a,b)}{\text{argmin}} \sum_{i=1}^{n} (Y_i^{*(j)} - a - bs_i)^2$$

Then, we can evaluate bias

$$\widehat{Bias} = \frac{1}{k} \sum_{j=1}^{k} \hat{\theta}_n^{(j)} - \theta$$

5.3.2 Example: Nonlinear Markov Process

Consider the model $Y_i = F_\theta(Y_{i-1}) + Z_i$, where $Z_i \sim \mathcal{N}(0, \sigma^2)$ (i.i.d.) for $i = 1, 2, ..., n$

Parameter $\theta = (a, b)$. Linear Least Square Estimator:

$$\hat{\theta}_n(\vec{Y}) = \underset{\theta}{\text{argmin}} \sum_{i=1}^{n} (Y_i - F_\theta(Y_{i-1}))^2$$

Given $\vec{Y}$, the residual (not i.i.d.)

$$E_i = Y_i = \hat{a}_n - F_{\hat{\theta}_n}(Y_{i-1}) \approx Z_i$$

Generate k resamples of $\vec{E} = (E_1, E_2, ..., E_n)$

$\Rightarrow$ obtain $\vec{E^*}^{(1)}, \vec{E^*}^{(2)}, ..., \vec{E^*}^{(k)}$ by resampling

$\Rightarrow$ Fix $Y_0^{*(j)} = Y_0$, compute pseudo-data $Y_i^{*(j)} = F_{\hat{\theta}_n}(Y_{i-1}^{*}{}^{(j)}) + E_i^{*(j)}$

$\Rightarrow$ Compute LS estimator

$$\hat{\theta}_n^{(j)} = \underset{(a,b)}{\text{argmin}} \sum_{i=1}^{n} (Y_i^{*(j)} - F_{\hat{\theta}_n}(Y_{i-1}^{*}{}^{(j)}))^2$$

$\Rightarrow$ Evaluate bias

$$\widehat{Bias} = \frac{1}{k}\sum_{j=1}^{k}\hat{\theta}_n^{(j)} - \theta$$

5.4 Posterior Simulation / Bayesian (Weighted) Bootstrap

Assumption *Bootstrap makes a strong assumption: The data is discrete and values not seen in the data are impossible.*

Consider $Z \in \mathbb{Z} = \{z_1, ..., z_J\}$ with parameter $\vec{\theta} = \{\theta_1, ..., \theta_J\} \in \Theta = \mathbb{S}^{J-1} = \{\vec{\theta} \in \mathbb{R}^J : \sum_{j=1}^{J}\theta_j = 1, \theta_j \geq 0, j = 1, ..., J\}$ such that $P(Z = z_j \mid \vec{\theta}) = \theta_j$.

Given a sample $\vec{Z} = (Z_1, ..., Z_N)$. Define $N_j = \sum_{i=1}^{N}\mathbf{1}\{Z_i = z_j\}, j = 1, 2, ..., J$, the number of observations that have value z_j. Then, the conditional pmf of $\vec{Z} \mid \vec{\theta}$ is

$$f(\vec{Z} \mid \vec{\theta}) = \prod_{j=1}^{J}\theta_j^{N_j}$$

Definition 5.1 (Steps to estimate β by Bayesian Bootstrap)

(1). We have prior $\pi(\vec{\theta})$.

(2). Given $\vec{Z}$, calculate posterior distribution $\pi(\vec{\theta} \mid \vec{Z})$.

(3). Draw samples $\vec{\theta}^{(t)}, t = 1, ..., T$ from $\pi(\vec{\theta} \mid \vec{Z})$.

(4). Then compute $\frac{1}{T}\sum_{t=1}^{T}\hat{\beta}(\vec{\theta}^{(t)})$.

5.4.1 Dirichlet Distribution Prior

A convenient way to assign the prior distribution of $\vec{\theta}$ over Θ is to use Dirichlet distribution.

Definition 5.2 (Dirichlet Distribution)

A **Dirichlet distribution** with parameters $\vec{\alpha} = (\alpha_1, ..., \alpha_J)$, $J \geq 2$. It allocates mass on $\vec{\theta}$ over Θ,

$$\pi(\vec{\theta}) = \frac{\Gamma(\sum_{j=1}^{J}\alpha_j)}{\sum_{j=1}^{N}\Gamma(\alpha_j)}\prod_{j=1}^{J}\theta_j^{\alpha_j-1}$$

where $\Gamma(z) \triangleq \int_0^{\infty} t^{z-1}e^{-t}dt$ is Gamma function (if z is positive integer, $\Gamma(z) = (z-1)!$).

Note *Dirichlet distribution generalizes Beta distribution.*

Now let's use Dirichlet distribution with parameters $\vec{\alpha} = (\alpha_1, ..., \alpha_J)$ to estimate $\mathbb{E}[\vec{\theta} \mid \vec{Z}]$.

As $f(\vec{Z} \mid \vec{\theta}) = \prod_{j=1}^{J}\theta_j^{N_j}$, we can compute the posterior beliefs

$$\pi(\vec{\theta} \mid \vec{Z}) = \frac{f(\vec{Z} \mid \vec{\theta})P(\vec{\theta})}{\int f(\vec{Z} \mid \vec{\theta}')P(\vec{\theta}')d\vec{\theta}'} = \frac{\Gamma(\sum_{j=1}^{J}(N_j + \alpha_j))}{\sum_{j=1}^{N}\Gamma(N_j + \alpha_j)}\prod_{j=1}^{J}\theta_j^{N_j+\alpha_j-1}$$

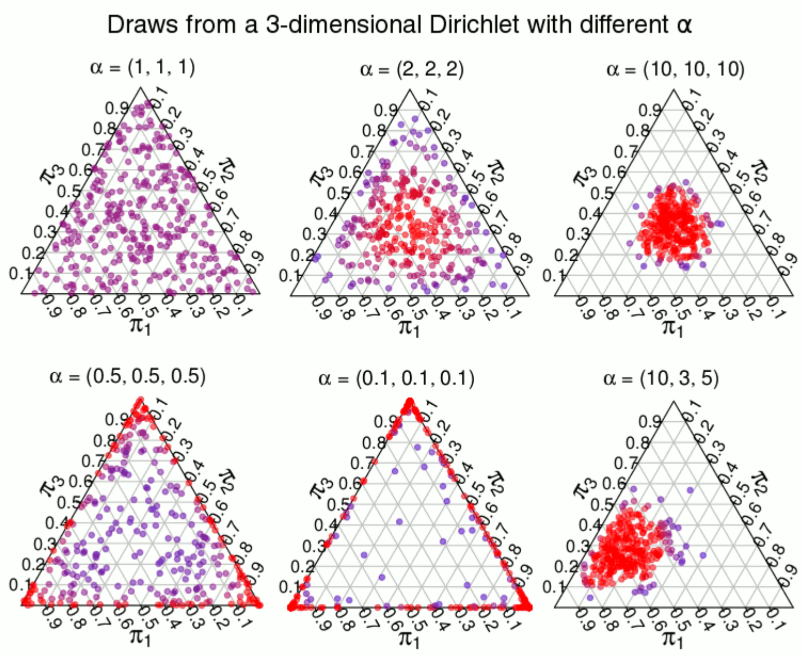

Figure 5.1: Dirichlet Distribution Examples

That is

$$\theta \mid \vec{Z} \sim \text{Dirichlet}(\bar{\alpha}), \text{ where } \bar{\alpha}_j = \alpha_j + N_j, \forall j$$

Simulate samples from Dirichlet distribution

Definition 5.3 (Simulate samples from Dirichlet($\vec{\alpha}$))

1. Consider a series of independent Gamma random variable $w_i \sim \text{Gamma}(\alpha_i, 1), i = 1, ..., J$;
2. Define $v_i = \frac{w_i}{\sum_{j=1}^{J} w_j}$;
3. We have $(v_1, ..., v_J) \sim \text{Dirichlet}(\alpha_1, ..., \alpha_J)$.

5.4.2 Haldane Prior

We may also begin with an uninformative prior, an improper prior, Dirichlet($\vec{\alpha}$), where $\vec{\alpha} \to 0$. $\pi(\theta) \propto \frac{1}{\theta_1 \theta_2 \cdots \theta_J}$. Under this prior, the posterior is Dirichlet$(N_1, ..., N_J)$, where $N_j = \sum_{i=1}^{N} \mathbf{1}\{Z_i = z_j\}$.

5.4.3 Linear Model Case

Each sample is $Z_i = (1, X_{1,i}, X_{2,i}, X_{3,i}, X_{4,i})$. The linear regression coefficient is $\beta = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$, and $\mathbb{E}^*[Y \mid X = x] = x'\beta$.

5.4.4 Bernoulli Case

Consider the problem of Example 2.4. Given N random sample $\{Z_1, ..., Z_N\}$ from a Bernoulli distribution with parameter θ and the sum $\sum_{i=1}^N Z_i = S$.

Consider a series of Gamma random variable $w_i^{(t)} \sim \text{Gamma}(1,1)$ from time $t = 1, ..., T$. Then, we have

$$\sum_{i=1}^N w_i^{(t)} \mathbf{1}_{\{Z_i=1\}} \sim \text{Gamma}(S,1)$$
$$\sum_{i=1}^N w_i^{(t)} \mathbf{1}_{\{Z_i=0\}} \sim \text{Gamma}(N-S,1)$$

Define $v_i^{(t)} = \frac{w_i^{(t)}}{\sum_{j=1}^N w_j^{(t)}}$. Based on the property of Gamma distribution, we have $\mathbb{E}[w_i^{(t)}] = \text{Var}[w_i^{(t)}] = 1$ and $\mathbb{E}[v_i^{(t)}] = \frac{1}{N}$.

As the relation between Gamma distribution and Beta distribution, we have

$$\frac{\text{Gamma}(S,1)}{\text{Gamma}(S,1) + \text{Gamma}(N-S,1)} \sim \text{Beta}(S, N-S)$$

Hence, we can define

$$\begin{aligned} \hat{\theta}^{(t)} &= \sum_{i=1}^N v_i^{(t)} Z_i \\ &= \sum_{i=1^N} \frac{w_i^{(t)} Z_i}{\sum_{j=1}^N w_j^{(t)}} \sim \text{Beta}(S, N-S) \end{aligned}$$

which is close to the posterior beliefs in Example 2.4 and can be seen as the posterior beliefs drawn from an improper prior: $\theta \sim \text{Beta}(\epsilon, \epsilon), \epsilon \to 0$, which has p.d.f. $\pi(\theta) = \frac{1}{\theta(1-\theta)}$.

We use

$$\frac{1}{T} \sum_{t=1}^T \hat{\theta}^{(t)} \approx \mathbb{E}[\theta^{(t)} | \{Z_1, ..., Z_n\}]$$

to estimate $\mathbb{E}[\theta^{(t)} | \{Z_1, ..., Z_n\}]$.