# Regression 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 $\theta$?";
- Hypothesis testing $\Rightarrow$ "Is $\theta = \theta_0$?";
- Interval estimation $\Rightarrow$ "Which values of $\theta$ 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)**

1. The **sample mean** is $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$;
2. 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 $\mu$ and variance $\sigma^2$.*

**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 $\mu$ and variance $\sigma^2$ (denoted by $X_i \sim$ i.i.d$(\mu, \sigma^2)$). 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**

*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. $(\mu, \sigma^2)$ and if $n$ is "large", then*

*(a).* $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$ *(converges in distribution) by CLT 4.2;*

*(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 $X_n = \max_{i=1,...,n} X_i$)**

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

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

♠

**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**

1. **Order statistics sampled from a uniform distribution on unit interval (Unif$[0,1]$):** Consider a random sample $U_1, ..., U_n$ 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 $\mathbb{E}[U_{(k)}] = \frac{k}{n+1}$.

2. **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:

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

For $n \ge k > j \ge 1$, $U_{(k)} - U_{(j)}$ 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 1.3**

1. Parametric model: $P = \{\Phi(\theta, 1) : \theta \in \mathbb{R}\}$, where $\Phi$ is the Gaussian c.d.f.
2. 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 1.4** 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 1.5**

1. Linear Regression Model: $\mathbb{E}_{P_0}[Y|X] = \beta_0^T X$.
   We solve $\beta$ by $\min_\beta \mathbb{E}_{P_0}[(y - \beta^T x)^2]$. The F.O.C. gives $\mathbb{E}_{P_0}[YX^T] = \beta^T \mathbb{E}_{P_0}[XX^T]$. $\beta_0$ solves this.
2. 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.
3. 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:
   (a). A1 (Overlap): $e(X) := P(D = 1|X) \in (0, 1)$
   (b). 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: $y(1) \perp D|X$, $\mathbb{E}[y|D = 1, X] = \mathbb{E}[y(1)|D = 1, X] = \mathbb{E}[y(1)|X]$.
4. 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
$$v(P_0)(t) = \Pr_{P_0}(\hat{\theta}_m - \theta(P_0) \leq t)$$

## 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 1.6

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. $v(P_0) = \operatorname{argmin}_b \mathbb{E}_{P_0}[(Y - b^T X)^2] = \mathbb{E}_{P_0}[XX^T]^{-1}\mathbb{E}_{P_0}[XY]$,
$$v(P_m) = \mathbb{E}_{P_m}[(Y - b^T X)^2] = \operatorname*{argmin}_b \frac{1}{m}\sum_{i=1}^{m}(Y_i - b^T X_i)^2 = \left(\frac{1}{m}\sum_{i=1}^{m} X_i X_i^T\right)^{-1}\left(\frac{1}{m}\sum_{i=1}^{m} X_i Y_i\right)$$
where $v(P_m)$ is OLS.
3. **GMM**. $\forall P \in \mathcal{P} : \mathbb{E}_P[g(Z, v(p))] = 0$, where $g$ is a known moment function.
$$v(P_0) = \operatorname*{argmin}_\theta \mathbb{E}_{P_0}[g(Z,\theta)]^T W \mathbb{E}_{P_0}[g(Z,\theta)]$$
where $W$ is a weighted matrix.
$$v(P_m) = \operatorname*{argmin}_\theta \left(\frac{1}{m}\sum_{i=1}^{m} g(Z_i,\theta)\right)^T W \left(\frac{1}{m}\sum_{i=1}^{m} g(Z_i,\theta)\right)$$
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 $\mu_d$.
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 $P_m^*$.

$$v(P_m)(t):=\Pr_{P_m}\left(\theta(P_m^*)-\theta(P_m)\le t\right)$$

♣

**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):

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

**Example 1.7 (Sample Mean)** Consider $\theta(P_0)=\mathbb{E}_{P_0}(Z)$, then $\theta(P_m)=\bar{Z}_m=\frac{1}{m}\sum_{i=1}^m Z_i$. $v(P_0)(t)=\Pr_{P_0}\left(\frac{1}{m}\sum_{i=1}^m(Z_i-\mathbb{E}_{P_0}(Z))\le t\right)$. The Bootstrap estimator is given by

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

or

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

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

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

(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 ($\tau$-th quantile)**

Let $q_\tau(v(P))$ be the $\tau$-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*

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

♡

**Proof 1.1**

*By (*), we have*

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

*Then,*

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

*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 $\sqrt{\mathrm{Var}_{P_0}(\theta(P_m))}$. The bootstrap standard error is given by

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

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

♣

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

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

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

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

## 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 1.8** 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 $\mathbb{E}_{P_0}[Z] = 0$, $v(P_0)(t) = \Pr_{P_0}\left(\max\{\mathcal{Z}, 0\} \leq t\right)$

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

   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, $v(P_0)(t) = \Pr_{P_0}\left(\max\{\mathcal{Z}, -c\} \leq t\right) \neq \Pr_{P_0}\left(\max\{\mathcal{Z}, 0\} \leq t\right)$.

   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

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

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

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

"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 1.8

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

**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 $X_1, ..., X_n$ 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 $\theta$) is a function of $(X_1, ..., X_n)$.

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

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 $\hat{\theta}_{MM}$ solves

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

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

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

♣

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

**Example 1.9**

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

$$\mu(p) = p$$

Then, the estimator is

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

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

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

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

   The moment function is

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

   Then, the estimator is

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

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

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

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

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

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 1.10**

Suppose $X_i \sim \text{i.i.d.} N(\mu, \sigma^2)$ 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

$$\mu'_1(\hat{\mu}_{MM}, \hat{\sigma}^2_{MM}) = \hat{\mu}_{MM} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$\mu'_2(\hat{\mu}_{MM}, \hat{\sigma}^2_{MM}) = \hat{\mu}_{MM} + \hat{\sigma}^2_{MM} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$
$$\Rightarrow \hat{\mu}_{MM} = \bar{X}$$
$$\hat{\sigma}^2_{MM} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

**Remark** $\bar{X}$ is the 'best' estimator of $\mu$; An alternative better estimator of $\sigma^2$ 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** $\hat{\theta}_{ML}$ solves

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

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 1.11**

1. Suppose $X_i \sim$ i.i.d. 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:

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

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

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

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

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

All in all,

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

**Remark** $\hat{p}_{ML} = \bar{X} = \hat{p}_{MM}$ 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

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

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

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

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

## 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 $\theta$ 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 $\theta$) 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 $\theta$) iff

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

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

**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 $\theta$

**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 $\theta$.

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

That is, the mutual information between $\theta$ and $T(X_1, ..., X_n)$ equals the mutual information between $\theta$ 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 $\theta$ and suppose $T$ is sufficient (for $\theta$). Then,

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

(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 $\theta$ 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*

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

*with strict inequality unless $Var(\hat{\theta}|T) = 0$ (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 1.12**

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 $\theta$. 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:

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

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

### 1.6.4 Minimal Sufficient Statistic

**Definition 1.24 (Minimal Sufficient Statistic)**

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

**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 $\theta$ if and only if $T(\vec{X}) = T(\vec{Y})$".*

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

**Example 1.13** Let $X_1, ..., X_n \sim$ 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

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

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 $\theta$ 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 $\theta$ 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 1.14** 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. $\hat{\theta}_{RB} = \mathbb{E}[\hat{\theta}_{MM}|X_{(n)}]$;
- $\mathbb{E}_\theta(X_{(n)}) = \frac{n}{n+1}\theta \Rightarrow$ unbiased $\frac{n+1}{n}X_{(n)}$ is UMVU ($= \hat{\theta}_{RB}$).

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

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

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,*

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

*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 1.15** 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 $(\mu, \sigma^2)$.

## 1.8 Fisher Information

### 1.8.1 Score Function

The score function is the derivative of the log likelihood function with respect to $\theta$.

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

**Lemma 1.1 ("Regularity" Condition ⇒ Mean of Score Function is Zero)**

Under "Regularity" condition and $X$ are continuous, the mean of score function, evaluated at the true parameter $\theta_0$, 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 $\theta$. ♡

### 1.8.2 Fisher Information

**Definition 1.28 (Fisher Information)**

The **Fisher information** is defined to be the variance of the score function at $\theta_0$.

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

♣

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

*Under "regularity" conditions, the* ***Fisher information*** *at $\theta_0$ can also be written as*

$$\mathcal{I}(\theta_0) = Var_{\theta_0}(u(\theta, \vec{X}))$$

♡

**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 $\hat{\theta}$*

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

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

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

**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. Ber$(p)$;
- (b). Estimating $\mu$ when $X \sim$ i.i.d. $N(\mu, \sigma^2)$.

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

- (a). Estimating $\sigma^2$ when $X \sim$ i.i.d. $N(\mu, \sigma^2)$: 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 $\theta$ when $X \sim$ i.i.d. $U[0, \theta]$: CR bound $\mathcal{I}(\theta)^{-1} = \frac{\theta^2}{n}$ and 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.*

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

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 $\Theta$ into two disjoint subsets $\Theta_0$ and $\Theta_1$.

(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 $\theta$;

(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 1.16** 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 1.17** 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 1.18** Heading Special Case: Making $P_\theta$[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

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

for corresponding statistic $T$ and critical value $c$.

- For $\theta \in \Theta_0$: $P_\theta$[Type I Error] $= P_\theta$[Reject $H_0$] $= \beta(\theta)$;
- For $\theta \in \Theta_1$: $P_\theta$[Type II Error] $= 1 - P_\theta$[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 $\alpha$)**

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

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

♣

**Example 1.19** 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 $\Phi$ 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 $\alpha$ 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 $\alpha$ Test)**

A test with level $\alpha$ and power function $\beta$ is a **uniformly most powerful (UMP) level $\alpha$ test** iff

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

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

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

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

## Example of Using NP Lemma

**Example 1.20** 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,

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

Decision Rule: Reject $H_0$ iff

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

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 $\alpha$) test of the $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$.

**Implications:** (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 $\sigma^2$ 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 $\beta$ 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**:

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

Motivation: Neyman-Pearson Lemma (1.11)

- **Lagrangian Multiplier Test Statistic**:
$$T_{LM}\left(X_1, ..., X_n\right) = \frac{\left(\frac{\partial}{\partial\theta} l(\theta_0 \mid X_1, ..., X_n)\right)^2}{-\frac{\partial^2}{\partial\theta^2} l(\theta_0 \mid X_1, ..., X_n)} = \frac{(u(\theta_0 \mid X_1, ..., X_n))^2}{\mathcal{I}(\theta_0 \mid X_1, ..., X_n)}$$
Motivation: $T_{LM}$ is approximate to $T_{LR}$; No estimation required.

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

**Claim 1.3**

*In general, for "large" $n$,*

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

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

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

(2). $T_W$:

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

Then,

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

## 1.11 Interval Estimation

**Definition 1.37**

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

1. An interval estimator of $\theta$ 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 converge probability (of the interval estimator) is the function (of $\theta$) given by

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

3. The confidence coefficient is $\inf_\theta P_\theta\left[L(X_1, ..., X_n) \leq \theta \leq U(X_1, ..., X_n)\right]$ ♣

**Example 1.21** Suppose $X_i \sim$ i.i.d. $N(\mu, 1)$, where $\mu$ is unknown.

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

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

Interpretation:

(I). Recall

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

(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 iff $|\bar{X} - \mu_0| > \frac{1.96}{\sqrt{n}}$

$\Leftrightarrow \bar{X} - \mu_0 > \frac{1.96}{\sqrt{n}}$ 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 M-Estimation

## 2.1 M-Estimation

### 2.1.1 Extremum Estimator and M-Estimator

Suppose there is a parameter of interest $\theta \in \mathbb{R}^d$. Data $Z$ is generated from $F_{\theta_0}$.

**Definition 2.1 (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 $\theta$ is of interest. $\hat{Q} : \Theta \to \mathbb{R}$ is a sample criterion. $\hat{\theta}$ is called **extremum estimator** of $\theta$ if

$$\hat{\theta}(\theta) = \underset{\theta\in\Theta}{\operatorname{argmin}}\, \hat{Q}(\theta)$$

♣

**Definition 2.2 (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 $\theta$.

♣

MLE is a special case of M-estimator.

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

**Example 2.1 (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 $Z$.

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

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

**Lemma 2.1 (Information Inequality: $\theta_0 = \text{argmin}_{\theta\in\Theta} -\mathbb{E}\log f(x|\theta)$)**

Given $\theta_0$ 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 2.1**

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

**Example 2.2 (Nonlinear Least Squares)** Consider the conditional restriction

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

where $g$ is known up to $\theta$ and differentiable in $\theta$. Then, the NLLS 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 2.2 (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 2.3**

Suppose the data $Z$ follows true distribution with parameter $\theta_0$.

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

### 2.1.2 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 2.4** $Q(\theta) = -\mathbf{1}\{\theta = 0\}$ 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 2.1 (Extremum Consistency)**

*Remind the definition of $\theta_0$ that $\theta_0 = \mathrm{argmin}_{\theta\in\Theta} Q(\theta)$. We give extra assumptions:*

*A1. Uniform Convergence, i.e., the worst-case distance converges to zero.*

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

*A2.* $\inf_{\|\theta-\theta_0\|>\epsilon} Q(\theta) > Q(\theta_0)$ *(Its **sufficient** condition: $Q(\theta)$ is continuous in $\theta$ on compact set $\Theta$.)*

*Suppose A1 and A2 hold. Then,*

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

### 2.1.3 Asymptotic Normality of M-estimators

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

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

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

**Theorem 2.2 (Asymptotic Normality of M-estimators)**

*Suppose*

*A1.* $\sup_{\theta\in\Theta} |\hat{Q}(\theta) - Q(\theta)| \xrightarrow{P} 0$.

*A2. $G(\theta)$ is continuous in $\Theta$.*

*A3.* $G := G(\theta_0) = \mathbb{E}_{\theta_0}\left[\frac{\partial^2}{\partial\theta\partial\theta^T} m(Z,\theta_0)\right]$ *is invertible.*

*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 2.2**

*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 2.1 (Asymptotic Normality of ML-estimator under correct specfication)**

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

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

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

♡

### 2.1.4 Efficiency of Asymptotically Linear Estimator

**Definition 2.4 (Efficient Asymptotically Linear Estimator)**

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

Use $\Omega_\beta$ 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. ♣

### 2.1.5 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 2.5 (Pseudo-true Parameter)**

**Pseudo-true parameter** is given by

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

$\beta_0$ 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, $G^{-1}\mathcal{I}G^{-1} \neq \mathcal{I}^{-1}$.

### Example of Misspecification

**Example 2.5** 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., $\theta$ may differ from $\theta_0$):

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

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

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

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 $\theta_*$ 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

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

## 2.2 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 2.6** The choice of link:

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

### 2.2.1 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

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

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

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

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.

## 2.2.2 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 $\beta$,

$$\arg\min_\beta Q(\beta) \equiv \arg\max_\beta \mathbb{E}_\beta \ln f(Y \mid X; \beta)$$

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

### 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.
$$\hat g(\beta)|_{\beta=\hat\beta_{\text{MLE}}}=\frac{1}{n}\sum_{i=1}^{n}g(Y_i\mid X_i;\beta)\big|_{\beta=\hat\beta_{\text{MLE}}}=0^d \tag{2.1}$$
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, (2.1) can be written as
$$\hat g(\beta)|_{\beta=\hat\beta_{\text{MLE}}}=\sum_{i=1}^{n}w(X_i,\hat\beta_{\text{MLE}})\cdot(Y_i-F(X_i^T\hat\beta_{\text{MLE}}))=0^d$$

## 2.2.3 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 $\beta$, pointwise convergence is enough, which follows from LLN.)*

By the Corollary 2.1,
$$\sqrt{n}\left(\hat\beta_{\text{MLE}}-\theta_0\right)\overset{d}{\Rightarrow}N\left(0,\mathcal{I}(\theta_0)^{-1}\right)$$

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 $\beta_0$).

**Example 2.7** 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 $0 < \kappa \leq F'(X^T\beta_0) \Leftrightarrow X^T\beta_0$ is not too large $\Leftrightarrow$ tails of $F'(X^T\beta)$ are not close to 0.

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

**Lemma 2.3**

*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 $\beta$.

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

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

## 2.3 Large Sample Testing

Let $\mathcal{I} := \mathcal{I}(\theta_0)$. By the Corollary 2.1,

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

We want to test

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

### 2.3.1 Wald Test: Distance on "$x$ axis"

The test statistic is

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

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 2.3**

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

### 2.3.2 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.,

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

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 2.4**

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

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

### 2.3.3 Likelihood Ratio Test

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

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

### 2.3.4 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 $H_0$:

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

## 2.4 Nonlinear Least Square

Suppose $Y$ is the outcome and $X$ are explanatory variables.

In previous "linear case," we use the form

$$\mathbb{E}[Y \mid X] = B(X)^T\beta, \quad B(X) = [1, X, X^2, ...]$$

Now, we consider a nonlinear expectation function

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

where $\rho$ is known up to $\beta$ and may not be linear in $\beta$

**Example 2.8**

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 $\beta_0$ s.t. $\beta_0 = \operatorname{argmin} Q(\beta)$ (sufficiency).

The sample criterion function is

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

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 2.1).

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.*

### 2.4.1 Efficient 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 2.1**

$\beta_0 = \operatorname{argmin} Q_w(\beta) := \frac{1}{2}\mathbb{E}w(x)(Y - \rho(B(X)^T\beta))^2$

**Proof 2.5**

*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 2.2**

*Optimal weight* $w^*(x) = \frac{1}{\text{Var}(Y|X)} = \frac{1}{\rho(B(X)^T\beta)(1-\rho(B(X)^T\beta))}$

**Proof 2.6**

$$Q_w(\beta) := \frac{1}{2}\mathbb{E}w(X)(Y - \rho(B(X)^T\beta))^2$$

$$G_w = \mathbb{E}\left[w(X)(\rho'(B(X)^T\beta))^2 B(X)B(X)^T\right]$$

$$\Omega_w = \mathbb{E}\left[w^2(X)\left(Y - \rho(B(X)^T\beta)\right)^2\left(\rho'(B(X)^T\beta)\right)^2 B(X)B(X)^T\right]$$

*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

$$\hat{\beta}_{\text{WNLS}} = \operatorname{argmin} \sum_{i=1}^{n} \frac{(Y - \rho(B(X)^T\beta))^2}{\rho(B(X)^T\beta)(1 - \rho(B(X)^T\beta))}$$

## 2.5 Quantile Regression

Let $\tau \in (0,1)$ be the quantile level and the $\tau$'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

$$q_Y(\tau) = \operatorname*{argmin}_q Q(q) := \mathbb{E}\rho_\tau(Y - q)$$

where $\rho_\tau(\cdot)$ is the check function defined as

$$\rho_\tau(u) := \{(1-\tau)\mathbf{1}\{u < 0\} + \tau\mathbf{1}\{u > 0\}\}|u|$$

### 2.5.1 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 2.6 (Linear Quantile Regression Model (LQR))**

$Q_{Y|X}(\tau \mid x) = X^T\beta_0(\tau)$ ♣

Consider

$$Y = X^T\gamma_0 + \epsilon$$

where $\epsilon$ is independent of $X$ (not $\mathbb{E}[\epsilon|X] = 0$, which is too weak).

**Assumption** *(Independence) $\epsilon$ is independent of $X$ (stronger than $\mathbb{E}[\epsilon|X] = 0$).*

**Lemma 2.4 (By Independence)**

$Q_{\epsilon|X}(\tau|X) = Q_\epsilon(\tau)$ a.s. in $X$ ♡

**Proof 2.7**

$$\begin{aligned} F_{\epsilon,X}(\epsilon, X) = F_\epsilon(\epsilon)F_X(X) &\Rightarrow F_{\epsilon|X}(\epsilon|X) = F_\epsilon(\epsilon) \\ &\Rightarrow Q_\epsilon(\tau) = F_\epsilon^{-1}(\epsilon) = Q_{\epsilon|X}(\tau|X) \end{aligned}$$

**Lemma 2.5 (Equivalence Property)**

*Let $T : \mathbb{R} \to \mathbb{R}$ be an increasing function. Then*

$$Q_{T(Y)}(\tau) = T(Q_Y(\tau))$$

♡

**Proof 2.8**

*Given $T$ is strictly increasing,*

$$\begin{aligned}\tau &= \Pr(Y < Q_Y(\tau)) \\ &= \Pr(T(Y) < T(Q_Y(\tau))) \\ &= F_{T(Y)}(T(Q_Y(\tau))) \\ \Rightarrow Q_{T(Y)}(\tau) &= T(Q_Y(\tau))\end{aligned}$$

**Example 2.9** The $T(\cdot)$ can be $T(y) = \min\{y, L\}, T(y) = ay + b$.

The quantile form of the LQR model is

$$Q_{Y|X}(\tau|X) = X^T\beta_0 + Q_\epsilon(\tau|X) = X^T\beta_0(\tau)$$

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 2.10 (Location-Scale Model)** $Y = X^T\gamma_0 + (X^T\delta_0)\epsilon$, where $X^T\delta_0 > 0$ a.s. in $X$. Then,

$$\begin{aligned}Q_{Y|X}(\tau|X) &= Q_{\epsilon|X}(\tau|X)(X^T\delta_0) + X^T\gamma_0 \\ \text{(by independence)} \quad &= X^T(Q_\epsilon(\tau)\delta_0) + X^T\gamma_0 \\ &= X^T\beta_0(\tau)\end{aligned}$$

where $\beta_0(\tau) = Q_\epsilon(\tau)\delta_0 + \gamma_0$.

### 2.5.2 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 $\tau$-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 2.3**

*If these three assumptions hold,*

$$Q_{Y(1)-Y(0)}(\tau) = Q_{Y|D=1}(\tau) - Q_{Y|D=0}(\tau)$$

♡

**Proof 2.9**

$$\begin{aligned}
Q_{Y(1)-Y(0)}(\tau) &= Q_{h(1,u)-h(0,u)}(\tau) \\
\text{(By equivalence property 2.5 and A3)} \quad &= h(1, Q_u(\tau)) - h(0, Q_u(\tau)) \\
\text{(By equivalence property 2.5 and A2)} \quad &= Q_{h(1,u)}(\tau) - Q_{h(0,u)}(\tau) \\
&= Q_{Y|D=1}(\tau) - Q_{Y|D=0}(\tau)
\end{aligned}$$

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* $x$*.*

# Chapter 3 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 $\theta$ of the distribution using an estimator $\hat{\theta}_n(\vec{Y})$. The distribution of the statistic $\theta$ 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:

$$G_n(t) = P(\hat{\theta}_n(\vec{Y}) < t), \forall t$$

## 3.1 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).

## 3.2 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 3.1 (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 $t \geq 0$,

$$\hat{G}_n^*(t) \geq P_{F_n}(Y_{(n)} = Y_{(n)}^*) = 1 - (1 - \frac{1}{n})^n \to 1 - e^{-1} \text{ as } n \to \infty$$

## 3.3 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**.

### 3.3.1 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

$$(\hat{a}_n, \hat{b}_n) = \underset{(a,b)}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - a - bs_i)^2$$

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 $1 \le i \le n$.
3. Compute LS estimator to the pseudo-data

$$\hat{\theta}_n^{(j)} = (\hat{a}_n^{(j)}, \hat{b}_n^{(j)}) = \underset{(a,b)}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i^{*(j)} - a - bs_i)^2$$

Then, we can evaluate bias

$$\widehat{Bias} = \frac{1}{k} \sum_{j=1}^{k} \hat{\theta}_n^{(j)} - \theta$$

### 3.3.2 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:

$$\hat{\theta}_n(\vec{Y}) = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - F_\theta(Y_{i-1}))^2$$

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

$$\hat{\theta}_n^{(j)} = \underset{(a,b)}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i^{*(j)} - F_{\hat{\theta}_n}(Y_{i-1}^{*}{}^{(j)}))^2$$

$\Rightarrow$ Evaluate bias

$$\widehat{Bias} = \frac{1}{k}\sum_{j=1}^{k}\hat{\theta}_n^{(j)} - \theta$$

## 3.4 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 3.1 (Steps to estimate $\beta$ 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 $\frac{1}{T}\sum_{t=1}^{T}\hat{\beta}(\vec{\theta}^{(t)})$.

♣

### 3.4.1 Dirichlet Distribution Prior

A convenient way to assign the prior distribution of $\vec{\theta}$ over $\Theta$ is to use Dirichlet distribution.

**Definition 3.2 (Dirichlet Distribution)**

A **Dirichlet distribution** with parameters $\vec{\alpha} = (\alpha_1, ..., \alpha_J)$, $J \geq 2$. It allocates mass on $\vec{\theta}$ over $\Theta$,

$$\pi(\vec{\theta}) = \frac{\Gamma(\sum_{j=1}^{J}\alpha_j)}{\sum_{j=1}^{N}\Gamma(\alpha_j)}\prod_{j=1}^{J}\theta_j^{\alpha_j - 1}$$

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

$$\pi(\vec{\theta} \mid \vec{Z}) = \frac{f(\vec{Z} \mid \vec{\theta})P(\vec{\theta})}{\int f(\vec{Z} \mid \vec{\theta}')P(\vec{\theta}')d\vec{\theta}'} = \frac{\Gamma(\sum_{j=1}^{J}(N_j + \alpha_j))}{\sum_{j=1}^{N}\Gamma(N_j + \alpha_j)}\prod_{j=1}^{J}\theta_j^{N_j + \alpha_j - 1}$$

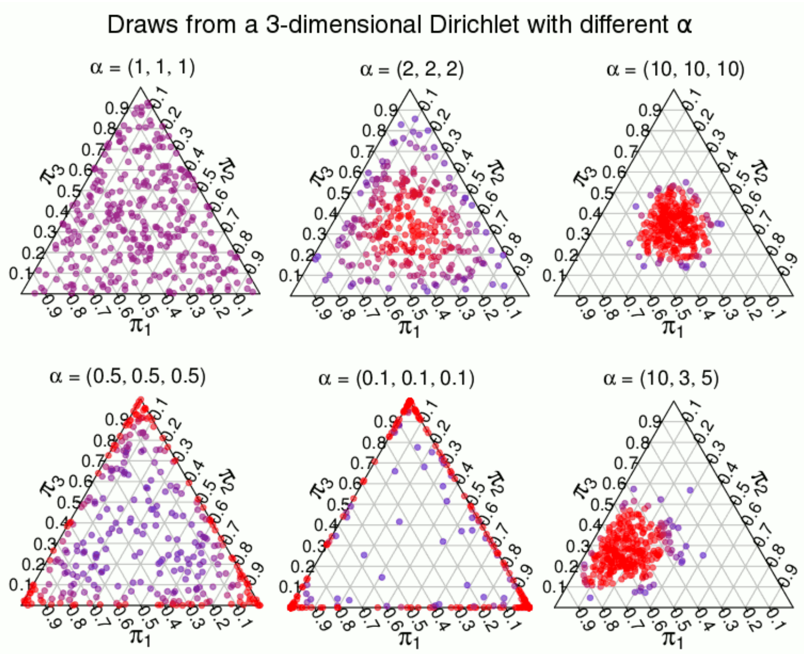


**Figure 3.1:** Dirichlet Distribution Examples

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 3.3 (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 $(v_1, ..., v_J) \sim \text{Dirichlet}(\alpha_1, ..., \alpha_J)$.

♣

## 3.4.2 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 $N_j = \sum_{i=1}^{N} \mathbf{1}\{Z_i = z_j\}$.

## 3.4.3 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 $\mathbb{E}^*[Y \mid X = x] = x'\beta$.

### 3.4.4 Bernoulli Case

Consider the problem of Example **??**. Given $N$ random sample $\{Z_1, ..., Z_N\}$ from a Bernoulli distribution with parameter $\theta$ 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 **??** 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 $\mathbb{E}[\theta^{(t)} | \{Z_1, ..., Z_n\}]$.

# Chapter 4 Linear Predictors / Regression

## 4.1 Best Linear Predictor

Consider a prediction problem that the distribution $F_{X,Y}$ is known, we observe $X = \begin{pmatrix} 1 \\ R \end{pmatrix} \in \mathbb{R}^{K\times 1}$ and predict $Y \in \mathbb{R}$. Only linear functions of $X$ are allowed $\mathcal{L} = \{X'b : b \in \mathbb{R}^K\}$. We use square experience loss $(Y - X'b)^2$. We want to minimze Risk (mean squared error)

$$\mathbb{E}_{X,Y}[(Y - X'b)^2] = \int_{x,y} (y - x'b)^2 f_{x,y}(x,y)dxdy$$

**Assumption** *Following inference is based on assumptions:*

*(i).* $\mathbb{E}[Y^2] < \infty$*;*

*(ii).* $\mathbb{E}[||X||^2] < \infty$ *(Frobenius norm);*

*(iii).* $\mathbb{E}[(\alpha' X)^2] > 0$ *for any non-zero* $\alpha \in \mathbb{R}^K$.

Let $\beta_0 = \arg\min_{b\in\mathbb{R}^k} \mathbb{E}_{X,Y}[(Y - X'b)^2]$. By the F.O.C.

$$\mathbb{E}[X(Y - X'\beta_0)] = 0$$

$$\mathbb{E}[XY] - \mathbb{E}[XX']\beta_0 = 0$$

$$\mathbb{E}[XY] = \underbrace{\mathbb{E}[XX']}_{non-singular} \beta_0$$

$$\beta_0 = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$$

**Proposition 4.1 (Best Linear Predictor)**

*Hence, the mean-squared error minimizing linear predictor of* $Y$ *given* $X$ *is*

$$\mathbb{E}^*[Y|X] = X'\beta_0, \text{ where } \beta_0 = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$$

$$\mathbb{E}_{X,Y}[X(\underbrace{Y - X'\beta_0}_{\triangleq u})] = \begin{pmatrix} \mathbb{E}[u] \\ \mathbb{E}[uR] \end{pmatrix} = \mathbf{0}$$

Hence, we have $\mathbb{E}[u] = 0$, then $\mathbb{E}[uR] = 0 = \text{Cov}(u, R)$.

**Lemma 4.1**

$\mathbb{E}[u] = \mathbb{E}[uR] = Cov(u, R) = 0$, *where* $u = Y - \mathbb{E}^*[Y|X]$.

If $u > 0$, it is underpredicting and if $u < 0$, it is overpredicting.

**Result 1 (ure Partitioned Inverse Formula)**

When we separate the constant term from other variables, we can write the Best Linear Predictor as:

**Proposition 4.2 (Best Linear Predictor (ure Partitioned Inverse Formula))**

$X = \begin{pmatrix} 1 \\ R \end{pmatrix}$, $\beta_0 = \begin{pmatrix} \alpha_0 \\ \beta_* \end{pmatrix}$, $\mathbb{E}[XX']^{-1} = \begin{bmatrix} 1 & \mathbb{E}[R]' \\ \mathbb{E}[R] & \mathbb{E}[RR'] \end{bmatrix}^{-1}$, $\mathbb{E}[XY] = \begin{pmatrix} \mathbb{E}[Y] \\ \mathbb{E}[RY] \end{pmatrix}$. *Then,*

$$\alpha_0 = \mathbb{E}[Y] - \mathbb{E}[R]'\beta_*$$

$$\beta_* = \underbrace{Var(R)^{-1}}_{(K-1)\times(K-1)} \times \underbrace{Cov(R,Y)}_{(K-1)\times 1}$$

## 4.2 Convergence of OLS

### 4.2.1 Approximation

OLS Fit is

$$\hat{\beta} = \left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right]^{-1} \left[\frac{1}{N}\sum_{i=1}^{N} X_i Y_i\right]$$

**Theorem 4.1 (Weak Law of Large Numbers (wLLN))**

*The weak law of large numbers (also called Khinchin's law) states that the sample average converges in probability towards the expected value.*

$$\overline{X}_n \xrightarrow{P} \mu \qquad \text{when } n \to \infty.$$

*That is, for any positive number $\varepsilon$,*

$$\lim_{n\to\infty} \Pr\left(|\overline{X}_n - \mu| < \varepsilon\right) = 1.$$

1. By LLN: $\frac{1}{N}\sum_{i=1}^{N} X_i Y_i \xrightarrow{P} \mathbb{E}[XY]$
2. By LLN and $f(X) = X^{-1}$ is continuous, $\left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right] \xrightarrow{P} \mathbb{E}[XX']^{-1}$
3. Hence,

$$\hat{\beta} = \left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right]^{-1} \left[\frac{1}{N}\sum_{i=1}^{N} X_i Y_i\right] \xrightarrow{P} \mathbb{E}[XX']^{-1}\mathbb{E}[XY] = \beta_0$$

**Theorem 4.2 (Central Limit Theorem (CLT))**

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{D} N(0,1) \; when \; n \to \infty$$

*$Z$ converges in distribution to $N(0,1)$ as $n \to \infty$*

*(converges in distribution: $P(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq a) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{x^2}{2}} dx$)*

♡

Application to OLS: Let $u = Y - X'\beta_0$. Then,

$$\begin{aligned}
\hat{\beta} &= \left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right]^{-1} \left[\frac{1}{N}\sum_{i=1}^{N} X_i Y_i\right] \\
&= \left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right]^{-1} \left[\frac{1}{N}\sum_{i=1}^{N} X_i (u_i + X_i'\beta_0)\right] \\
&= \beta_0 + \left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right]^{-1} \left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N} X_i u_i\right]
\end{aligned}$$

Then,

$$\sqrt{N}(\hat{\beta} - \beta_0) = \left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right]^{-1} \left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N} X_i u_i\right]$$

1. By LLN, $\left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right]^{-1} \xrightarrow{P} \mathbb{E}[XX']^{-1} \triangleq \Gamma_0^{-1}$.
2. By CLT, $\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N} X_i u_i\right] \sim \mathcal{N}(0, \Omega_0)$, where

$$\Omega_0 = Var[X_i u_i] = \mathbb{E}[\|X_i u_i\|^2] = \mathbb{E}[\|x_i\|^2 u_i^2] \leq \left(\mathbb{E}[\|x_i\|^4]\right)^{\frac{1}{2}} \mathbb{E}[u_i^4]^{\frac{1}{2}}$$

Hence,

$$\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{D} N\left(0, \Gamma_0^{-1}\Omega_0\Gamma_0^{-1}\right)$$

The estimation of $\Gamma_0$ and $\Omega_0$:

$$\begin{aligned}
\hat{\Gamma} &= \frac{1}{N}\sum_{i=1}^{N} X_i X_i' \\
\hat{\Omega} &= \frac{1}{N}\sum_{i=1}^{N} X_i \hat{u}_i \hat{u}_i' X_i', \quad \text{where } \hat{u}_i = Y_i - X_i'\hat{\beta}
\end{aligned}$$

We have

$$\hat{\Gamma}^{-1}\hat{\Omega}\hat{\Gamma}^{-1} \xrightarrow{P} \Gamma_0^{-1}\Omega_0\Gamma_0^{-1}$$

Then,

$$\hat{\beta} \xrightarrow{approx} N\left(\beta_0, \frac{\hat{\Gamma}^{-1}\hat{\Omega}\hat{\Gamma}^{-1}}{N}\right)$$

### 4.2.2 Testing and Confidence Interval

Let $\hat{\Lambda} = \hat{\Gamma}^{-1}\hat{\Omega}\hat{\Gamma}^{-1}$, $\Lambda = \Gamma_0^{-1}\Omega_0\Gamma_0^{-1}$, $\sqrt{N}(\hat{\beta}_k - \beta_k) \xrightarrow{D} N(0, \Lambda_{kk})$. Hence,

$$T_N \triangleq \sqrt{N}\Lambda_{kk}^{-\frac{1}{2}}\left(\hat{\beta}_k - \beta_k\right) \xrightarrow{D} N(0,1)$$

Consider the event $A = \mathbf{1}\{|T_N| \leq 1.96\}$. We have

$$\Pr(A = 1) = \Phi(1.96) - \Phi(-1.96) = 0.95$$

Specifically,

$$\begin{aligned} A &= \mathbf{1}\{|T_N| \leq 1.96\} \\ &= \mathbf{1}\left\{\hat{\beta}_k - 1.96\frac{\Lambda_{kk}^{\frac{1}{2}}}{\sqrt{N}} \leq \beta_k \leq \hat{\beta}_k + 1.96\frac{\Lambda_{kk}^{\frac{1}{2}}}{\sqrt{N}}\right\} \end{aligned}$$

The “Random Interval” is

$$\left[\hat{\beta}_k - 1.96\frac{\Lambda_{kk}^{\frac{1}{2}}}{\sqrt{N}}, \hat{\beta}_k + 1.96\frac{\Lambda_{kk}^{\frac{1}{2}}}{\sqrt{N}}\right]$$

#### Testing Linear Restrictions

Let $\theta = H\beta$, where $H$ is $p \times k$ and $\beta$ is $k \times 1$.

$$H_0 : \theta = \theta_0; \quad H_1 : \theta \neq \theta_0$$

We have

$$\sqrt{N}(\hat{\theta} - \theta_0) = H\sqrt{N}\left(\hat{\beta} - \beta_0\right) \xrightarrow[H_0]{D} N(0, H\Lambda_0 H')$$

Moreover,

$$W_0 = N\left(\hat{\theta} - \theta_0\right)(H\Lambda_0 H')^{-1}\left(\hat{\theta} - \theta_0\right) \xrightarrow[H_0]{D} \chi_p^2$$

where $\mathbb{E}[\chi_p^2] = p$.

## 4.3 Long, Short, Auxilary Regression

$Y \in \mathbb{R}^1$, $X \in \mathbb{R}^K$, $K \in \mathbb{R}^J$. Consider a researcher interested in the conditional distribution of the logarithm of weekly wages ($Y \in \mathbb{R}^1$) given years of competed schooling ($X \in \mathbb{R}^K$) and vector of additional worker attributes. This vector could include variables such as age, childhood test scores, and race. Let $W$ be this $J \times 1$ vector of additional variables.

We can run regression by two ways:

1. Long regression: $\mathbb{E}^*[Y|X, W] = X'\beta_0 + W'\gamma_0$.

2. Short regression: $\mathbb{E}^*[Y|X] = X'b_0$.

**Proposition 4.3 (Long Regression)**

*Long regression is another form of best linear predictor.*

$$\begin{aligned}\mathbb{E}^*[Y|X,W] &= \mathbb{E}^*[Y|Z] \\ &= Z'\left(\mathbb{E}[ZZ']^{-1}\mathbb{E}[ZY]\right) \\ &= X'\beta_0 + W'\gamma_0\end{aligned}$$

*where* $\begin{pmatrix}\beta_0 \\ \gamma_0\end{pmatrix} = \mathbb{E}[ZZ']^{-1}\mathbb{E}[ZY]$, $Z = \begin{pmatrix}X \\ W\end{pmatrix}$.

**Proposition 4.4 (Auxiliary Regression)**

$$\mathbb{E}^*[W|X] = \Pi_0 X$$

*which is multivariate regression. For each row* $j = 1, ..., J$,

$$\mathbb{E}^*[W_j|X] = X'\Pi_{j0}$$

*where* $\Pi_{j0} = \mathbb{E}[XX']^{-1}\mathbb{E}[XW_j]$ *and* $\Pi_0 = \begin{pmatrix}\Pi'_{10} \\ \vdots \\ \Pi'_{J0}\end{pmatrix} = \mathbb{E}[WX']\mathbb{E}[XX']^{-1}$.

**Theorem 4.3 (Law of Iterated Linear Predictors (LILP))**

$$\mathbb{E}^*[Y|X] = \mathbb{E}^*[\mathbb{E}^*[Y|X,W]|X]$$

Facts: Linear predictor is linear operator, $\mathbb{E}^*[X+Y|W] = \mathbb{E}^*[X|W] + \mathbb{E}^*[Y|W]$.

Let $Y = \mathbb{E}^*[Y|X,W] + u = X'\beta_0 + W'\gamma_0 + u$. Then,

$$\begin{aligned}\mathbb{E}^*[Y|X] &= \mathbb{E}^*[X'\beta_0 + W'\gamma_0 + u|X] \\ &= \mathbb{E}^*[X'\beta_0|X] + \mathbb{E}^*[W'\gamma_0|X] + \mathbb{E}^*[u|X] \\ &= X'\beta_0 + (\Pi_0 X)'\gamma_0 + 0 \\ &= X'\underbrace{(\beta_0 + \Pi_0'\gamma_0)}_{b_0}\end{aligned}$$

**Proposition 4.5 (Short Regression)**

$$\mathbb{E}^*[Y|X] = X'b_0$$

*where* $b_0 = \beta_0 + \Pi_0' \gamma_0$.

## 4.4 Residual Regression

Let the variation in $W$ unexplained by $X$.

$$\underbrace{V}_{J\times 1} = \underbrace{W}_{J\times 1} - \underbrace{\mathbb{E}^*[W|X]}_{J\times 1} = W - \Pi_0 X$$

**Proposition 4.6 (Residual Regression)**

*Let* $\tilde{Y} = Y - \mathbb{E}^*[Y|X]$,

$$\mathbb{E}^*[\tilde{Y}|V] = V'\gamma_0$$

**Proof 4.1**

$$\begin{aligned}
Y &= X'\beta_0 + W'\gamma_0 + u \\
\tilde{Y} &= X'\beta_0 - \mathbb{E}^*[Y|X] + W'\gamma_0 + u \\
&= -X'(\Pi_0'\gamma_0) + W'\gamma_0 + u \\
&= V'\gamma_0 + u \\
\mathbb{E}^*[\tilde{Y}|V] &= V'\gamma_0
\end{aligned}$$

By long regression,

$$\begin{aligned}
\mathbb{E}^*[Y|X,W] &= X'\beta_0 + W'\gamma_0 \\
&= X'b_0 - X'(\Pi_0'\gamma_0) + W'\gamma_0 \\
&= X'b_0 + V'\gamma_0 \\
&= \mathbb{E}^*[Y|X] + \mathbb{E}^*[\tilde{Y}|V]
\end{aligned}$$

**Theorem 4.4 (Frisch-Waugh Theorem)**

$$\begin{aligned}
\mathbb{E}^*[Y|X,V] &= \mathbb{E}^*[Y|X] + \mathbb{E}^*[Y|V] - \mathbb{E}[Y] \\
&= \mathbb{E}^*[Y|X,W]
\end{aligned}$$

**Lemma 4.2**

*If* $Cov(X,W) = 0$, *then*

$$\mathbb{E}^*[Y|X,W] = \mathbb{E}^*[Y|X] + \mathbb{E}^*[Y|W] - \mathbb{E}[Y]$$

**Proof 4.2**

*Let* $u = Y - \mathbb{E}^*[Y|X, W]$.

$$\begin{aligned} 0 &= \mathbb{E}[uW] \\ &= \mathbb{E}[(Y - \mathbb{E}^*[Y|X] - \mathbb{E}^*[Y|W] + \mathbb{E}[Y])W] \\ &= \underbrace{\mathbb{E}[(Y - \mathbb{E}^*[Y|W])W]}_{=0 \text{ by F.O.C.}} - \underbrace{\mathbb{E}[\mathbb{E}^*[Y|X]]}_{=\mathbb{E}[Y]}\mathbb{E}[W] + \mathbb{E}[Y]\mathbb{E}[W] \end{aligned}$$

## 4.5 Card-Krueger Model

Consider a model about log-learning based on schooling, ability, luck.

$$Y(s) = \alpha_0 + \beta_0 \underbrace{s}_{\text{schooling } s \in \mathbb{S}} + \underbrace{A}_{\text{ability}} + \underbrace{V}_{\text{luck}}$$

Given a cost function about $s$:

$$C(s) = \underbrace{C}_{\text{cost heterogeneity}} s + \frac{k_0}{2}s^2$$

**Assumption** *We assume*

1. *Information set* $I_0 = (C, A)$ *are known by agent when choosing schooling.*
2. $V$ *is independent of* $C, A$*:* $V|C, A \triangleq V$.

Then, the observed schooling $s$ should satsify

$$\begin{aligned} s &= \arg\max_s \mathbb{E}[Y(s) - C(s) \mid I_0] \\ &= \arg\max_s \alpha_0 + \beta_0 s + A - Cs - \frac{k_0}{2}s^2 \end{aligned}$$

By F.O.C.

$$\beta_0 - C - k_0 s = 0 \Rightarrow s = \frac{\beta_0 - C}{k_0}$$

1. **Long Regression**:

$$\mathbb{E}^*[Y|s, A] = \alpha_0 + \beta_0 s + A \tag{LR}$$

2. Short Regression:

$$\mathbb{E}^*[Y|s] = a_0 + b_0 s$$

3. **Auxillary Regression**: By the best linear predictor, the $\mathbb{E}^*[A|s]$ can be written as

$$\begin{aligned} \mathbb{E}^*[A|s] &= \mathbb{E}[A] - \frac{\text{Cov}(A, s)}{\text{Var}(s)}\mathbb{E}[s] + \frac{\text{Cov}(A, s)}{\text{Var}(s)}s \\ &= \mathbb{E}[A] - \eta_0\mathbb{E}[s] + \eta_0 s \end{aligned} \tag{AR}$$

where $\eta_0 = \frac{\text{Cov}(A,s)}{\text{Var}(s)}$ and $s = \frac{\beta_0 - C}{k_0}$ and $\mathbb{E}[s] = \frac{\beta_0 - \mu_C}{k_0}$,

$$\text{Cov}(A,s) = \text{Cov}\left(A, \frac{\beta_0 - C}{k_0}\right) = -\frac{\text{Cov}(A,C)}{k_0} = -\frac{\sigma_{AC}}{k_0}$$

$$\text{Var}(s) = \text{Var}\left(\frac{\beta_0 - C}{k_0}\right) = \frac{\sigma_C^2}{k_0^2}$$

$$\eta_0 = -k_0 \frac{\sigma_{AC}}{\sigma_C^2} = -k_0 \frac{\sigma_{AC}}{\sigma_A \sigma_C} \frac{\sigma_A}{\sigma_C} = -k_0 \rho_{AC} \frac{\sigma_A}{\sigma_C}$$

The Auxillary Regression is written as

$$\begin{aligned} \mathbb{E}^*[A|s] &= \mathbb{E}[A] + k_0 \rho_{AC} \frac{\sigma_A}{\sigma_C} \frac{\beta_0 - \mu_C}{k_0} - k_0 \rho_{AC} \frac{\sigma_A}{\sigma_C} s \\ &= \mathbb{E}[A] + \rho_{AC} \frac{\sigma_A}{\sigma_C} (\beta_0 - \mu_C) - k_0 \rho_{AC} \frac{\sigma_A}{\sigma_C} s \end{aligned} \tag{AR-1}$$

Hence, the **Short Regression**

$$\begin{aligned} \mathbb{E}^*[Y|s] &= \mathbb{E}^*\left[\mathbb{E}^*[Y|s,A]|s\right] \\ &= \mathbb{E}^*\left[\alpha_0 + \beta_0 s + A|s\right] \\ &= \alpha_0 + \beta_0 s + \mathbb{E}^*[A|s] \\ &= \underbrace{\alpha_0 + \mathbb{E}[A] + \rho_{AC} \frac{\sigma_A}{\sigma_C} (\beta_0 - \mu_C)}_{a_0} + \underbrace{\left(\beta_0 - k_0 \rho_{AC} \frac{\sigma_A}{\sigma_C}\right)}_{b_0} s \end{aligned} \tag{SR}$$

### 4.5.1 Proxy Variable Regression

What if we don't observe $A$ or $C$. We observe some observed variables $W$ (**proxy variable**) instead.

**Assumption** *We assume*

1. *Redundancy:* $\mathbb{E}^*[Y|s,A,W] = \mathbb{E}^*[Y|s,A]$ *($W$ doesn't give extra information).*
2. *Conditional Uncorrelatedness:* $\mathbb{E}^*[A|s,W] = \mathbb{E}^*[A|W] = \Pi_0 + W'\Pi_W$ *(Auxillary Regression).*
3. *Conditional Independence:* $C \perp A|W = w$.

The **Proxy Variable Regression** is given by

$$\begin{aligned} \mathbb{E}^*[Y|s,W] &= \mathbb{E}^*\left[\mathbb{E}^*[Y|s,A,W]|s,W\right] \\ &= \mathbb{E}^*\left[\mathbb{E}^*[Y|s,A]|s,W\right] \\ &= \mathbb{E}^*[\alpha_0 + \beta_0 s + A|s,W] \\ &= \alpha_0 + \beta_0 s + (\Pi_0 + W'\Pi_W) \\ &= (\alpha_0 + \Pi_0) + \beta_0 s + W'\Pi_W \end{aligned} \tag{PVR}$$

A general form of **Proxy Variable Regression** with

1. Long Regression: $\mathbb{E}^*[Y|X,A] = X'\beta_0 + A'\gamma_0$
2. Redundancy: $\mathbb{E}^*[Y|X,A,W] = \mathbb{E}^*[Y|X,A]$

3. Conditional Uncorrelatedness: $\mathbb{E}^*[A|X,W] = \mathbb{E}^*[A|W] = \Pi_0 W$

   where $\Pi_0$ is $P \times J$, $W$ is $J \times 1$, and $A$ is $P \times 1$.

$$\begin{aligned}
\mathbb{E}^*[Y|X,W] &= \mathbb{E}^*\left[\mathbb{E}^*[Y|X,A,W]|X,W\right] \\
&= \mathbb{E}^*\left[\mathbb{E}^*[Y|X,A]|X,W\right] \\
&= \mathbb{E}^*\left[X'\beta_0 + A'\gamma_0|X,W\right] \\
&= X'\beta_0 + \mathbb{E}^*[A|X,W]'\gamma_0 \\
&= X'\beta_0 + W'\Pi_0'\gamma_0
\end{aligned}$$

## 4.6 Instrumental Variables

### 4.6.1 Motivation

Suppose we want to estimate an OLS model $y = \beta^T x + e$, where $x \in \mathbb{R}^k$. The OLS estimator is given by

$$\hat{\beta}_{\text{OLS}} = \left(\frac{1}{m}\sum_{i=1}^m X_i X_i^T\right)^{-1}\left(\frac{1}{m}\sum_{i=1}^m X_i Y_i\right)$$

which converges (in probability) to

$$\mathbb{E}_{P_0}[XX^T]^{-1}\mathbb{E}_{P_0}[XY] = \beta + \mathbb{E}_{P_0}[XX^T]^{-1}\underbrace{\mathbb{E}_{P_0}[Xe]}_{\text{assumed to be 0 (Exogeneity)}}$$

What if the exogeneity doesn't hold?

**Example 4.1**

1. $y = \beta x^* + e$, where $\mathbb{E}[x^* e] = 0$. However, we don't have $x^*$ and we only have a noisy variable $x = x^* + v$ (with $\mathbb{E}[v] = 0$). Then, $y = \beta(x - v) + e = \beta x + \epsilon$, where $\epsilon := e - \beta v$. The probability limits of the OLS estimator satisfies

$$\hat{\beta}_{\text{OLS}} - \beta = \frac{\mathbb{E}_{P_0}[x\epsilon]}{\mathbb{E}_{P_0}[x^2]} = \frac{\mathbb{E}_{P_0}[(x^* + v)(e - \beta v)]}{\mathbb{E}_{P_0}[(x^* + v)^2]} = -\frac{\beta\mathbb{E}_{P_0}[v^2]}{\mathbb{E}_{P_0}[(x^* + v)^2]}$$

   Hence, it is impossible to let the estimator converge to the true $\beta$.

2. Returns to Schooling: Consider a model

$$\ln \text{Wage} = \beta_0 + \beta_1 \text{EDUC} + e$$

   Suppose the $e$ is correlated to both the wage and the education. Given $e$ is positively correlated to the education, the OLS estimator is over-estimating.

### 4.6.2 I.V. Model

Consider a model $Y = X^T\beta + e$, where $X \in \mathbb{R}^k$ and $\mathbb{E}_{P_0}[xe] \neq 0$.

**Definition 4.1 (Instrumental Variable)**

A variable $Z \in \mathbb{R}^l$ is an **instrumental variable** if it satisfies

(1). $\mathbb{E}_{P_0}[Ze] = 0$ (exogeneity).

(2). $\mathbb{E}_{P_0}[ZZ^T]$ is non-singular (tech).

(3). $\text{Rank}(\mathbb{E}_{P_0}(ZX^T)) = k$ (relevance), which requires $l \geq k$. ♣

**Remark** Exogeneity implies "exclusion restriction", which means the $Z$ can't directly affect $Y$ without affecting $X$.

**Implementation:**

- Outcome Equation:

$$Y = X^T \beta + e$$

- $1^{st}$ Stage Equation (no economic meaning, just for mathematical use):

$$X = \Gamma^T Z + u$$

  where $X$ and $u$ are $k \times 1$, $\Gamma$ are $l \times k$, and $Z$ is $l \times 1$. $Z \perp u$ and $\Gamma = \mathbb{E}[ZZ^T]^{-1}\mathbb{E}[ZX^T]$.

- Reduced Form Equation:

$$\begin{aligned} Y &= \beta^T X + e \\ &= \beta^T(\Gamma^T Z + u) + e \\ &= \lambda^T Z + v \end{aligned}$$

  where $\lambda = \Gamma\beta$ and $v = \beta^T u + e$.

  Note that $\mathbb{E}[Zv] = 0$, which satisfies exogeneity. Hence, we can use OLS to estimate $\lambda$.

**Identification:** Suppose $\lambda$ and $\Gamma$ are known, we want to recover $\beta$.

$$\lambda = \Gamma\beta$$

1. Case 1: $l = k$,

$$\beta = \Gamma^{-1}\lambda$$

   where $\Gamma^{-1}$ exists by relevance.

2. Case 2: $l > k$,

$$\Gamma^T\lambda = (\Gamma^T\Gamma)\beta \Rightarrow \beta = (\Gamma^T\Gamma)^{-1}\Gamma^T\lambda$$

**Estimation of $\Gamma$ and $\lambda$:**

(A). "Plug In"

(a). The estimation of $\Gamma$ is given by

$$\hat{\Gamma} = \left(\frac{1}{m}\sum_{i=1}^{m} Z_i Z_i^T\right)^{-1} \left(\frac{1}{m}\sum_{i=1}^{m} Z_i X_i^T\right) \tag{hG}$$

The OLS estimator of regressing $X$ on $Z$ should converge to $\Gamma$ in probability.

(b). The estimation of $\lambda$ is given by

$$\hat{\lambda} = \left(\frac{1}{m}\sum_{i=1}^{m} Z_i Z_i^T\right)^{-1} \left(\frac{1}{m}\sum_{i=1}^{m} Z_i Y_i\right)$$

which converges to $\lambda$ in probability.

(B). "2SLS"

The reduced form can also be written as

$$\begin{aligned} Y &= \beta^T X + e \\ &= \beta^T(\Gamma^T Z + u) + e \\ &= \beta^T \underbrace{(\Gamma^T Z)}_{W} + v \end{aligned} \tag{hl}$$

Assuming $\Gamma$ is known, we can regress $Y$ on $W$:

$$\begin{aligned} \tilde{\beta} &= \left(\frac{1}{m}\sum_{i=1}^{m} W_i W_i^T\right)^{-1} \left(\frac{1}{m}\sum_{i=1}^{m} W_i Y_i\right) \\ &= \left(\Gamma^T \left(\frac{1}{m}\sum_{i=1}^{m} Z_i Z_i^T\right)\Gamma\right)^{-1} \Gamma^T \left(\frac{1}{m}\sum_{i=1}^{m} Z_i Y_i\right) \end{aligned}$$

Hence, we can estimate $\beta$ based on

$$\hat{\beta}_{\text{2SLS}} = \left(\hat{\Gamma}^T \left(\frac{1}{m}\sum_{i=1}^{m} Z_i Z_i^T\right)\hat{\Gamma}\right)^{-1} \hat{\Gamma}^T \left(\frac{1}{m}\sum_{i=1}^{m} Z_i Y_i\right)$$

where $\hat{\Gamma}$ is given by (4.1). Specifically, in the case of $l = k$, $\hat{\beta}_{\text{2SLS}} = \left(\frac{1}{m}\sum_{i=1}^{m} Z_i X_i^T\right)^{-1}\left(\frac{1}{m}\sum_{i=1}^{m} Z_i Y_i\right)$.

**Remark** Why not use the following steps?

(a). Regress $X$ on $Z$ to construct $\hat{W} := \hat{\Gamma}^T Z$.

(b). Regress $Y$ on $\hat{W}$.

(Note that the mathematical foundation of OLS doesn't hold here because $\hat{W}$ is not i.i.d.)

### 4.6.3 Weak I.V.

The "relevance" of the IV doesn't hold: $\mathbb{E}[ZX^T] \approx 0$. Why this is a problem?

Let's begin with a simple case that $l = k = 1$. The 2SLS estimator is given by

$$\hat{\beta}_{\text{2SLS}} = \frac{\frac{1}{m}\sum_{i=1}^{m} Z_i Y_i}{\frac{1}{m}\sum_{i=1}^{m} Z_i X_i} = \beta + \frac{\frac{1}{m}\sum_{i=1}^{m} Z_i e_i}{\frac{1}{m}\sum_{i=1}^{m} Z_i X_i}$$

where the small $Z_i X_i$ may lead to a large bias.

Consider the $\mathbb{E}[ZX] = \frac{c}{\sqrt{m}}, c \neq 0$. Then, the 2SLS estimator can be written as

$$\hat{\beta}_{\text{2SLS}} = \beta + \frac{\frac{1}{m}\sum_{i=1}^m Z_i e_i}{\frac{c}{\sqrt{m}}\frac{1}{m}\sum_{i=1}^m Z_i^2 + \frac{1}{m}\sum_{i=1}^m Z_i v_i} = \beta + \frac{\frac{1}{\sqrt{m}}\sum_{i=1}^m Z_i e_i}{c\frac{1}{m}\sum_{i=1}^m Z_i^2 + \frac{1}{\sqrt{m}}\sum_{i=1}^m Z_i u_i}$$

where the $\lim_{m\to\infty} \frac{1}{\sqrt{m}}\sum_{i=1}^m Z_i e_i \sim \mathcal{N}(0, \sigma^2)$ and $\lim_{m\to\infty} \frac{1}{\sqrt{m}}\sum_{i=1}^m Z_i u_i \sim \mathcal{N}(0, r^2)$ by LLN, and $\frac{1}{m}\sum_{i=1}^m Z_i^2 \to 1 + 0_P(1)$ with normalized $Z$. Hence, As $m \to \infty$,

$$\hat{\beta}_{\text{2SLS}} \approx \beta + \frac{\mathcal{N}(0, \sigma^S)}{\mathcal{N}(c, r^2)}$$

which gives that $\hat{\beta}_{\text{2SLS}}$ is not good for nonzero $\mathbb{E}[ZX]$.

## 4.7 Linear Generalized Method of Moments (Linear GMM)

### 4.7.1 Generalized Method of Moments (GMM)

**Assumption** *GMM model assumes that, given the true probability of data $P_0$, there exists a unique parameter $\beta$ such that*

$$\mathbb{E}_{P_0}[g(\text{Data}, \beta_0)] = 0$$

*where $g(\cdot)$ is a residual function.*

$\beta_0$ is given by

$$\beta_0 = \underset{\beta}{\text{argmin}}\, J(\beta, P_0)$$

where

$$J(\beta, P_0) := \left(\mathbb{E}_{P_0}[g(Y, X, Z, \beta)]\right)^T W \left(\mathbb{E}_{P_0}[g(Y, X, Z, \beta)]\right)$$

and the weight matrix $W \succ 0$ (is positive definite and symmetric).

The GMM estimator is given by

$$\hat{\beta}_{\text{GMM}} = \underset{\beta}{\text{argmin}}\, J(\beta, P_m)$$

Using this for

1. Linear Regression: $g(Y, X, \beta) := (Y - X^T\beta)X$;
2. IV Model: $g(Y, X, Z, \beta) = Z(Y - X^T\beta)$, which is called Linear GMM.

### 4.7.2 Linear GMM

**Definition 4.2 (Linear GMM)**

A **Linear GMM** is defined as

$$\mathbb{E}_{P_0}[\underbrace{Z}_{l\times 1}(\underbrace{Y}_{1\times 1}-\beta_0^T\underbrace{X}_{k\times 1})]=0$$

If Rank $\left(\mathbb{E}_{P_0}[ZX^T]\right)=k$, there is a unique $\beta_0=$ minimizes $J(\beta,P_0)$ with

$$J(\beta,P_0):=\left(\mathbb{E}_{P_0}[Z(Y-X^T\beta)]\right)^T W\left(\mathbb{E}_{P_0}[Z(Y-X^T\beta)]\right)$$

$$J(\hat{\beta},P_0):=\left(\frac{1}{m}\sum_{i=1}^m Z_i(Y_i-X_i^T\beta)\right)^T W\left(\frac{1}{m}\sum_{i=1}^m Z_i(Y_i-X_i^T\beta)\right)$$

The GMM estimator is given by

$$\hat{\beta}_{\text{GMM}}=\underset{\beta}{\operatorname{argmin}}\left(\frac{1}{m}\sum_{i=1}^m Z_i(Y_i-X_i^T\beta)\right)^T W\left(\frac{1}{m}\sum_{i=1}^m Z_i(Y_i-X_i^T\beta)\right) \tag{4.1}$$

**Remark** $W$ matters for $\hat{\beta}_{\text{GMM}}$.

The FOC of (4.1) is given by

$$\left(\frac{1}{m}\sum_{i=1}^m Z_iX_i^T\right)^T W\left(\frac{1}{m}\sum_{i=1}^m Z_iY_i-\left(\frac{1}{m}\sum_{i=1}^m Z_iX_i^T\right)\hat{\beta}_{\text{GMM}}\right)=0$$

Let $\hat{Q}:=\frac{1}{m}\sum_{i=1}^m Z_iX_i^T\in\mathbb{R}^{l\times k}$. Then,

$$\hat{\beta}_{\text{GMM}}=\left(\hat{Q}^TW\hat{Q}\right)^{-1}\hat{Q}^TW\frac{1}{m}\sum_{i=1}^m Z_iY_i$$

**Lemma 4.3**

*If $W=(\frac{1}{m}\sum_{i=1}^m Z_iZ_i^T)^{-1}$, then $\hat{\beta}_{\text{GMM}}=\hat{\beta}_{\text{2SLS}}$*

**Proof 4.3**

*With $W^T=W$,*

$$\begin{aligned}\hat{\beta}_{\text{GMM}}&=\left(\hat{Q}^TW\hat{Q}\right)^{-1}\hat{Q}^TW\frac{1}{m}\sum_{i=1}^m Z_iY_i\\ &\left(\hat{Q}^TWW^{-1}W\hat{Q}\right)^{-1}\hat{Q}^TW\frac{1}{m}\sum_{i=1}^m Z_iY_i\\ &=\left((W\hat{Q})^TW^{-1}(W\hat{Q})\right)^{-1}(W\hat{Q})^T\frac{1}{m}\sum_{i=1}^m Z_iY_i\end{aligned}$$

*Substitute $W$ by $W=(\frac{1}{m}\sum_{i=1}^m Z_iZ_i^T)^{-1}$. We have $W\hat{Q}=\hat{\Gamma}$. The lemma is proved.*

### 4.7.3 Properties of Linear GMM Estimator

**Theorem 4.5 (Asymptotic)**

$$\sqrt{m}\left(\hat{\beta}_{\text{GMM}} - \beta_0\right) \to \mathcal{N}(0, V_{P_0}).$$

**Proof 4.4**

$$\begin{aligned}
\hat{\beta}_{\text{GMM}} &= \left(\hat{Q}^T W \hat{Q}\right)^{-1} \hat{Q}^T W \frac{1}{m}\sum_{i=1}^m Z_i \underbrace{Y_i}_{X_i^T\beta_0 + e_i} \\
&= \left(\hat{Q}^T W \hat{Q}\right)^{-1} \hat{Q}^T W \left( \underbrace{(\frac{1}{m}\sum_{i=1}^m Z_i X_i^T)}_{\hat{Q}} \beta_0 + \frac{1}{m}\sum_{i=1}^m Z_i e_i \right) \\
&= \beta_0 + \left(\hat{Q}^T W \hat{Q}\right)^{-1} \hat{Q}^T W \frac{1}{m}\sum_{i=1}^m Z_i e_i
\end{aligned}$$

*By LLN, $\hat{Q} \xrightarrow{P} Q := \mathbb{E}[ZX^T]$. Then we have, $\hat{Q}^T W \hat{Q} \xrightarrow{P} Q^T W Q$. Because $Q^T W Q$ is invertible, $(\hat{Q}^T W \hat{Q})^{-1} \xrightarrow{P} (Q^T W Q)^{-1}$. So, $(\hat{Q}^T W \hat{Q})^{-1} = (Q^T W Q)^{-1} + o_{P_0}(1)$. Hence,*

$$\begin{aligned}
\hat{\beta}_{\text{GMM}} &= \beta_0 + \left((Q^T W Q)^{-1} + o_{P_0}(1)\right)\left(Q^T W + o_{P_0}(1)\right) \frac{1}{m}\sum_{i=1}^m Z_i e_i \\
&= \beta_0 + ((Q^T W Q)^{-1} Q^T W + o_{P_0}(1)) \frac{1}{m}\sum_{i=1}^m Z_i e_i \\
&= \beta_0 + (Q^T W Q)^{-1} Q^T W \frac{1}{m}\sum_{i=1}^m Z_i e_i + o_{P_0}(1)\frac{1}{m}\sum_{i=1}^m Z_i e_i
\end{aligned}$$

*By orthogonality condition, $\mathbb{E}_{P_0}[Ze] = 0$. And by central limit theorem, we have $\sqrt{m}\frac{1}{m}\sum_{i=1}^m Z_i e_i \to \mathcal{N}(0, \Omega_{P_0})$. Then, we represent $\hat{\beta}_{\text{GMM}}$ as*

$$\hat{\beta}_{\text{GMM}} = \beta_0 + (Q^T W Q)^{-1} Q^T W \frac{1}{m}\sum_{i=1}^m Z_i e_i + o_{P_0}(\frac{1}{\sqrt{m}}) \tag{4.2}$$

*which is called **asymptotic linear representation**.*

*Multiplying $\sqrt{m}$,*

$$\begin{aligned}
\sqrt{m}(\hat{\beta}_{\text{GMM}} - \beta_0) &= (Q^T W Q)^{-1} Q^T W \underbrace{\frac{1}{\sqrt{m}}\sum_{i=1}^m Z_i e_i}_{\to \mathcal{N}(0, \Omega_{P_0})} + o_{P_0}(1) \\
&\to \mathcal{N}\left(0, \underbrace{(Q^T W Q)^{-1} Q^T W \Omega_{P_0} W Q (Q^T W Q)^{-1}}_{\triangleq V_{P_0}}\right)
\end{aligned}$$

**Corollary 4.1**

$\hat{\beta}_{\text{GMM}} \xrightarrow{P} \beta_0$.

**Proof 4.5**

$\hat{\beta}_{\text{GMM}} - \beta_0 = O_{P_0}(\frac{1}{\sqrt{m}}) \to o_{P_0}(1)$.

**Efficiency Consideration** We want to choose the weight matrix to minimize the asymptotic variance within GMM estimator, $W^* = \text{argmin}_W V_{P_0}$.

**Theorem 4.6**

$W^* = \Omega_{P_0}^{-1}$. *That is,* $V_{P_0}^* := \left(Q^T \Omega_{P_0}^{-1} Q\right)^{-1} \leq V_{P_0}, \forall W$.

Then, we want to compute the efficient GMM by $\Omega_{P_0} := \mathbb{E}[e^2 Z Z^T]$.

$$\hat{W}^* = \left(\hat{\Omega}\right)^{-1}$$

where $\hat{\Omega} = \frac{1}{m}\sum_{i=1}^m \hat{e}_i^2 Z Z^T$ and $\hat{e}_i$ is given by

$$\hat{e}_i := Y_i - X_i^T \hat{\beta}$$

where $\hat{\beta}$ can be any GMM estimator, e.g., $W = I$ or a 2SLS estimator. As long as we can make sure $\hat{\Omega} \xrightarrow{P} \Omega_{P_0}$. Finally, we have $\hat{\beta}_{\text{EFFI}} := \hat{W}^* = W^* + o_{P_0}(1)$,

$$\sqrt{m}\left(\hat{\beta}_{\text{EFFI}} - \beta_0\right) \to \mathcal{N}(0, \left(Q^T \Omega_{P_0}^{-1} Q\right)^{-1})$$

**Remark** If $\mathbb{E}_{P_0}[e^2|Z] = \sigma_e^2$, then 2SLS is efficient.

$$\Omega^{-1} = \left(\mathbb{E}_{P_0}[e^2 Z Z^T]\right)^{-1} = \frac{1}{\sigma_e^2} \underbrace{\left(\mathbb{E}_{P_0}[Z Z^T]\right)^{-1}}_{W \text{ used in 2SLS}}$$

### 4.7.4 Alternative: Continuous Updating Estimator

Based on the idea of efficiency, we may use

$$\hat{\beta}_{\text{CUE}} = \underset{\beta}{\text{argmin}} \left(\frac{1}{m}\sum_{i=1}^m g(\text{Data}_i, \beta)\right)^T \left(\frac{1}{m}\sum_{i=1}^m \hat{e}_i^2 Z Z^T\right) \left(\frac{1}{m}\sum_{i=1}^m g(\text{Data}_i, \beta)\right)$$

However, it may not be convex.

### 4.7.5 Inference

Suppose we want test $H_0 : \Gamma(\beta_0) = \theta_0 = 0$ or $H_0 : \theta_0 = \Gamma(\beta_0) \neq \hat{\theta} = \Gamma(\hat{\beta})$.

**Theorem 4.7 (Construct Chi-square)**

*By using the asymptotic variance of GMM, $V_{P_0}$,*

$$m(\hat{\theta}-\theta)^T\left(R(\beta_0)^T V_{P_0} R(\beta_0)\right)^{-1}(\hat{\theta}-\theta) \Rightarrow \chi^2_l$$

*where $R(\beta_0) := \frac{d\Gamma(\beta_0)}{d\beta} \in \mathbb{R}^{k\times l}$.*

**Proof 4.6**

*Let*

$$\overbrace{m(\hat{\theta}-\theta)^T \underbrace{\left(R(\beta_0)^T V_{P_0} R(\beta_0)\right)^{-1}}_{\triangleq\Omega}(\hat{\theta}-\theta)}^{\mathcal{W}} \Rightarrow \chi^2_l$$

*We have*

$$\hat{\theta}-\theta_0 = \Gamma(\hat{\beta}) - \Gamma(\beta_0) = \underbrace{\frac{d\Gamma(\beta_0)}{d\beta}}_{R(\beta_0)}(\hat{\beta}-\beta_0) + o_{P_0}(m^{-\frac{1}{2}})$$

$$\mathcal{W} = \left(\sqrt{m}R(\beta_0)(\hat{\beta}-\beta_0) + o_{P_0}(1)\right)^T \Omega \left(\sqrt{m}R(\beta_0)(\hat{\beta}-\beta_0) + o_{P_0}(1)\right)$$

*As $\sqrt{m}\left(\hat{\beta}-\beta_0\right) \Rightarrow \mathcal{N}(0, V_{P_0})$, by continuous mapping theorem, we have*

$$\mathcal{W} \Rightarrow \left(\mathcal{N}(0, R(\beta_0)V_{P_0}R(\beta_0)^T)\right)^T \Omega \left(\mathcal{N}(0, R(\beta_0)V_{P_0}R(\beta_0)^T)\right)$$

*Let $M := R(\beta_0)V_{P_0}R(\beta_0)^T$. Since $M$ is symmetric, it can be decomposed by $M = LL^T$. Then, $M^{-1} = (L^T)^{-1}L^{-1}$. We have $L^{-1}M(L^T)^{-1} = I$.*

*Since $\Omega = M^{-1} = (L^{-1})^T L^{-1}$,*

$$\mathcal{W} \Rightarrow (\mathcal{N}(0, I))^T(\mathcal{N}(0, I)) = \chi^2_l$$

Based on this theorem, we have the "real" Wald test for $H_0 : \Gamma(\beta_0) = \theta_0 = 0$.

$$\mathcal{W} = m(\hat{\theta}-\theta)^T\left(R(\hat{\beta})^T \hat{V}_{P_0} R(\hat{\beta})\right)^{-1}(\hat{\theta}-\theta) \Rightarrow \chi^2_l$$

### 4.7.6 OVER-ID Test

Remind that

$$J(\beta, P_0) := \left(\mathbb{E}_{P_0}[Z(Y - X^T\beta)]\right)^T W \left(\mathbb{E}_{P_0}[Z(Y - X^T\beta)]\right)$$

We want to test

$$H_0 : J(\beta, P_0) = 0$$

which is equivalent to $\mathbb{E}[Ze] = 0$. $H_1 : J(\beta, P_0) > 0$, which is equivalent to $\mathbb{E}[Ze] \neq 0$.

**Theorem 4.8**

*If $W$ is efficient weighting matrix ($W = \hat{\Omega}^{-1}$), then $mJ(\hat{\beta}, P_m) \Rightarrow \chi^2_{l-k}$*

**Proof 4.7**

*Remind* (4.2) *that* $\hat{\beta} = \beta_0 + (Q^T W Q)^{-1} Q^T W \frac{1}{m} \sum_{i=1}^m Z_i e_i + o_{P_0}(\frac{1}{\sqrt{m}})$ *and* $Q := \mathbb{E}[ZX^T]$*. Then,*

$$
\begin{aligned}
Z_i(Y_i - X_i^T \hat{\beta}) &= Z_i(X_i^T \beta_0 + e_i - X_i^T \hat{\beta}) \\
&= -Q(\hat{\beta} - \beta_0) + \frac{1}{m} \sum_{i=1}^m Z_i e_i + o_{P_0}(\frac{1}{\sqrt{m}})
\end{aligned}
$$

*which gives*

$$
\frac{1}{m} \sum_{i=1}^m Z_i(Y_i - X_i^T \hat{\beta}) = \left(I - Q(Q^T W Q)^{-1} Q^T W\right) \frac{1}{m} \sum_{i=1}^m Z_i e_i + o_{P_0}(\frac{1}{\sqrt{m}})
$$

*By decomposing $W$ by $W := LL^T$,*

$$
mJ(\hat{\beta}, P_m) = \left(L^T \frac{1}{\sqrt{m}} \sum_{i=1}^m Z_i(Y_i - X_i^T \hat{\beta})\right)^T \left(L^T \frac{1}{\sqrt{m}} \sum_{i=1}^m Z_i(Y_i - X_i^T \hat{\beta})\right)
$$

*where*

$$
\begin{aligned}
L^T \frac{1}{\sqrt{m}} \sum_{i=1}^m Z_i(Y_i - X_i^T \hat{\beta}) &= \left(L^T - \underbrace{L^T Q}_{:=M}((L^T Q)^T (L^T Q))^{-1} (L^T Q)^T L^T\right) \frac{1}{\sqrt{m}} \sum_{i=1}^m Z_i e_i + o_{P_0}(1) \\
&= \underbrace{(I - M(M^T M)^{-1} M^T)}_{:=R_M} \left(L^T \left(\frac{1}{\sqrt{m}} \sum_{i=1}^m Z_i e_i\right)\right) + o_{P_0}(1)
\end{aligned}
$$

*where $R_M$ satisfies $R_M = R_M^T R_M$, which shows $R_M$ has eigenvalues $\in \{0, 1\}$ and its number of eigenvalues equal to $1$ is $l - k$.*

*Hence,*

$$
mJ(\hat{\beta}, P_m) = \left(L^T \left(\frac{1}{\sqrt{m}} \sum_{i=1}^m Z_i e_i\right)\right)^T R_M \left(L^T \left(\frac{1}{\sqrt{m}} \sum_{i=1}^m Z_i e_i\right)\right) + o_{P_0}(1)
$$

*As* $\left(L^T \left(\frac{1}{\sqrt{m}} \sum_{i=1}^m Z_i e_i\right)\right) \Rightarrow \xi \sim \mathcal{N}(0, L^T \Omega L)$*. So,*

$$
mJ(\hat{\beta}, P_m) \Rightarrow \xi^T R_m \xi
$$

*If $W = \Omega^{-1}$, then $L^T \Omega L = I$, which gives*

$$
\begin{aligned}
mJ(\hat{\beta}, P_m) &\Rightarrow \xi_*^T R_m \xi_*, \ \xi_* \sim \mathcal{N}(0, I) \\
&= \sum_{j=1}^{l-k} \omega_j^2, \omega_j \sim \mathcal{N}(0, 1) \\
&\sim \chi^2_{l-k}
\end{aligned}
$$

**Remark**

1. Test by $c_\alpha$, which gives $\Pr(\chi^2_{l-k} \geq c_\alpha) = \alpha \in (0, 1)$.

2. Only make sense for $l > k$.
   (a). You "spent" $k$ degrees of freedom estimating $\beta_0$.
   (b). The rest $(l-k)$ is "spent" on testing.

### 4.7.7 Bootstrap GMM

Now, we gives estimator by using bootstrap data,

$$\hat{\beta}^* = \underset{\beta}{\operatorname{argmin}}\, J(\beta, P_m^*)$$

where

$$J(\beta, P_m^*) := \left(\frac{1}{m}\sum_{i=1}^m Z_i^*(Y_i^* - X_i^{*T}\beta) - \mathbb{E}_{P_m}[Z(Y - X^T\hat{\beta})]\right)^T W \left(\frac{1}{m}\sum_{i=1}^m Z_i^*(Y_i^* - X_i^{*T}\beta) - \mathbb{E}_{P_m}[Z(Y - X^T\hat{\beta})]\right)$$

where $\mathbb{E}_{P_m}[Z(Y - X^T\hat{\beta})] = \frac{1}{m}\sum_{i=1}^m Z_i\hat{e}_i$, which is used to debias. Then,

$$\hat{\beta}_{\text{GMM}} = \left(\hat{Q}^{*T} W \hat{Q}^*\right)^{-1} \hat{Q}^{*T} W \left(\frac{1}{m}\sum_{i=1}^m (Z_i^* Y_i^* - Z_i\hat{e}_i)\right)$$

**Bootstrap OVER-ID Test** The distribution $mJ(\hat{\beta}^*, P_m^*)$ is the <u>same</u> as $mJ(\hat{\beta}, P_m)$ regardless of $W$.

## 4.8 Panel Data Models

**Definition 4.3 (Panel Data)**

For each unit $i$, it has time $\{1, ..., T\}$.

| | |
|---|---|
| | $t=1$ |
| $i=1$ | $\vdots$ |
| | $t=T$ |
| | $t=1$ |
| $i=2$ | $\vdots$ |
| | $t=T$ |
| $\vdots$ | $\vdots$ |

♣

The typical model is given by

$$Y_{i_t} = \underbrace{\alpha_i}_{\text{Fixed Effect}} + X_{i_t}^T\beta + \epsilon_{i_t}$$

$\alpha_i$ is a fixed effect, which is unobserved, random, and time invariant.

**Assumption**

1. *$\{\alpha_i, (X_{i_t})_{t=1}^T, (Y_{i_t})_{t=1}^T, (\epsilon_{i_t})_{t=1}^T\}$ is i.i.d. for all $i \in \{1, ..., N\}$. (Within a unit, data at different time can be dependent, which means there are no estimators within units.)*
2. *$N \to \infty$, $T$ is fixed.*

### 4.8.1 Pooled OLS

$$Y_{i_t} = X_{i_t}^T \beta_0 + \underbrace{e_{i_t}}_{:=\alpha_i + \epsilon_{i_t}}$$

Use the notations of vectors $\vec{Y}_i := \begin{bmatrix} Y_{i_1} \\ \vdots \\ Y_{i_T} \end{bmatrix}$, $\vec{X}_i := \begin{bmatrix} X_{i_1} \\ \vdots \\ X_{i_T} \end{bmatrix}$, $\vec{e}_i := \mathbf{1}\alpha_i + \vec{\epsilon}_i$, where $\mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$. Then, the equation can be written as

$$\vec{Y}_i = \vec{X}_i \beta_0 + \vec{e}_i$$

The pooled OLS estimator is

$$\hat{\beta}_{\text{pool}} := \left( \frac{1}{N} \sum_{i=1}^N \vec{X}_i^T \vec{X}_i \right)^{-1} \left( \frac{1}{N} \sum_{i=1}^N \vec{X}_i^T \vec{Y}_i \right)$$

**Properties**

$$\hat{\beta}_{\text{pool}} = \beta_0 + \left( \frac{1}{N} \sum_{i=1}^N \vec{X}_i^T \vec{X}_i \right)^{-1} \left( \frac{1}{N} \sum_{i=1}^N \vec{X}_i^T \vec{e}_i \right)$$

For consistency:

1. $\frac{1}{N} \sum_{i=1}^N \vec{X}_i^T \vec{X}_i \xrightarrow{P} \mathbb{E}[\vec{X}^T \vec{X}]$, which is required to be non singular.
2. $\frac{1}{N} \sum_{i=1}^N \vec{X}_i^T \vec{e}_i \xrightarrow{P} \mathbb{E}[\vec{X}^T \vec{e}]$, where

$$\mathbb{E}[\vec{X}^T \vec{e}] = \underbrace{\mathbb{E}[\vec{X}^T \mathbf{1}\alpha]}_{\text{need assumed to be 0}} + \underbrace{\mathbb{E}[\vec{X}^T \vec{\epsilon}]}_{:=0, \text{ by assumption}}$$

The pooled OLS estimator is inconsistent if $X_{it}$ is correlated with $\alpha_i$.

**Assumption** *$X_{it}$ is uncorrelated with $\alpha_i$, $\mathbb{E}[X_{it}\alpha_i] = 0$.*

Asymptotic Normality:

$$\sqrt{N}\left(\hat{\beta}_{\text{pool}} - \beta_0\right) = \underbrace{\left( \frac{1}{N} \sum_{i=1}^N \vec{X}_i^T \vec{X}_i \right)^{-1}}_{\mathbb{E}[\vec{X}^T \vec{X}] + o_{P_0}(1)} \underbrace{\left( \frac{1}{\sqrt{N}} \sum_{i=1}^N \vec{X}_i^T \vec{e}_i \right)}_{\text{by CLT:} \Rightarrow N(0, \mathbb{E}[\vec{X}^T \vec{e}\vec{e}^T \vec{X}])}$$

$$\Rightarrow N\left(0, \mathbb{E}[\vec{X}^T \vec{X}]^{-1} \mathbb{E}[\vec{X}^T \vec{e}\vec{e}^T \vec{X}] \mathbb{E}[\vec{X}^T \vec{X}]^{-1}\right)$$

where $\mathbb{E}[\vec{X}^T \vec{e}\vec{e}^T \vec{X}] = \vec{X}^T \mathbb{E}[\vec{e}\vec{e}^T \mid \vec{X}] \vec{X}$. Specifically, $\mathbb{E}[e_s e_t \mid \vec{X}] = \mathbb{E}[\alpha^2 + \epsilon_s \epsilon_t \mid \vec{X}] \neq 0, \forall s \neq t$. Hence, the variance of the normal distribution is not identical matrix. We need to compute the variance:

$$[\frac{1}{N}\sum_{i=1}^{N}\vec{X}_i^T\vec{X}_i]^{-1}[\frac{1}{N}\sum_{i=1}^{N}\vec{X}_i^T\hat{\vec{e}}_i\hat{\vec{e}}_i^T\vec{X}_i][\vec{X}_i^T\vec{X}_i]^{-1}$$

where $\hat{\vec{e}}_i = \vec{Y}_i - \vec{X}_i\hat{\beta}_{\text{pool}}$.

### 4.8.2 Fixed Effect Model

$$Y_{i_t} = \underbrace{\alpha_i}_{\text{Fixed Effect}} + X_{i_t}^T\beta + \epsilon_{i_t}$$

where is no assumption over $\alpha$ and $\vec{X}_i$.

**"Naive" Time Difference** (losing many data, inefficient):

$$\Delta Y_i = Y_{i_t} - Y_{i_{t-1}}, \text{ for some } t$$

$$\Delta Y_i = \Delta X_i\beta_0 + \Delta\epsilon_i$$

We get OLS estimator

$$\hat{\beta}_{\text{Diff}} = \frac{\sum_{i=1}^{n}\Delta X_i\Delta Y_i}{\sum_{i=1}^{n}\Delta X_i^2}$$

With assumptions $\mathbb{E}[X_t\epsilon_t] = \mathbb{E}[X_t\epsilon_{t-1}] = \mathbb{E}[X_{t-1}\epsilon_t] = \mathbb{E}[X_{t-1}\epsilon_{t-1}] = 0$, we have $\mathbb{E}[\Delta X\Delta\epsilon] = 0$, which gives the consistency.

**Fixed Effect Estimator** (most used): Let

$$\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T}Y_{i_t} = \alpha_i + \bar{X}_i\beta + \bar{\epsilon}_i$$

"Dot" Model:

$$\dot{Y}_{i_t} = Y_{i_t} - \bar{Y}_i = \dot{X}_{i_t}\beta_0 + \dot{\epsilon}_{i_t}$$

Use the notations of vectors $\dot{\vec{Y}}_i := \begin{bmatrix}\dot{Y}_{i_1}\\ \vdots \\ \dot{Y}_{i_T}\end{bmatrix} = \vec{Y}_i - \mathbf{1}\left(\mathbf{1}^T\mathbf{1}\right)^{-1}\mathbf{1}^T\vec{Y}_i =: Q\vec{Y}_i$, where $Q := I - \mathbf{1}\left(\mathbf{1}^T\mathbf{1}\right)^{-1}\mathbf{1}^T$ (notice that $QQ = Q$).

Then, the equation $\dot{\vec{Y}}_i = \dot{\vec{X}}_i\beta_0 + \dot{\vec{\epsilon}}_i$ can be written as

$$Q\vec{Y}_i = Q\vec{X}_i\beta_0 + Q\vec{\epsilon}_i$$

Run OLS

$$\hat{\beta}_{FE} = \left(\frac{1}{N}\sum_{i=1}^{N}\vec{X}_i^TQ\vec{X}_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\vec{X}_i^TQ\vec{Y}_i\right)$$

**Assumption** *We assume* $\mathbb{E}[\vec{X}^T Q\vec{\epsilon}] = 0$*, which is equivalent to* $\mathbb{E}[\vec{X}_i^T \vec{\epsilon}_i] = 0$.

**Note** *"Strict exogeneity" is sufficient for above assumption:* $\mathbb{E}[X_s \epsilon_t] = 0, \forall s, t$ *(*$\epsilon$ *is uncorrelated with past, present, and future* $X$*'s).*

Consistency:

$$\hat{\beta}_{FE} = \beta_0 + \left(\frac{1}{N}\sum_{i=1}^{N} \vec{X}_i^T Q \vec{X}_i\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} \vec{X}_i^T Q \vec{\epsilon}_i\right)$$

The sufficient condition is $\mathbb{E}[\vec{X}^T Q\vec{\epsilon}] = 0$, that is the motivation of giving the above assumption.

**Theorem 4.9**

$$\sqrt{N}(\hat{\beta}_{FE} - \beta_0) \Rightarrow N\left(0, (\mathbb{E}[\vec{X}^T Q \vec{X}])^{-1}\mathbb{E}[\vec{X}^T Q\vec{\epsilon}\vec{\epsilon}^T Q\vec{X}](\mathbb{E}[\vec{X}^T Q\vec{X}])^{-1}\right)$$

**Remark**

1. Actually, all we want to do is constructing a matrix $Q$ such that $Q\alpha_i = 0$, so that we can get rid of fixed effect. Another example of this kind of matrix is $D = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ & & & \cdots & & \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}$.
2. Time invariant covariant? No.
3. Dummy interpretation:

$$Y_{i_t} = \gamma_1 D1_{i_t} + \gamma_2 D2_{i_t} + \vdots + \gamma_N DN_{i_t} + X_{i_t}\beta + \epsilon_{i_t}$$

   where $Dj_{i_t} = 1$ if $i = j$ and $Dj_{i_t} = 0$ if $i \neq j$.
4. Fixed effect can't be estimated.

### 4.8.3 Random Effect Model

(Based on many assumptions, but more efficient than fixed effect. However, still not suggested.)

**Assumption** $\alpha_i$ *is orthogonal to* $X_{it}$, $\mathrm{Cov}(\alpha_i X_{i_t}) = 0$.

$$Y_{i_t} = X_{i_t}\beta_0 + e_{i_t},\ e_{i_t} = \alpha_i + \epsilon_{i_t}$$

which can be written as the form of vector

$$\vec{Y}_i = \vec{X}_i\beta_0 + \vec{e}_i, \vec{e}_i = \alpha_i \mathbf{1} + \vec{\epsilon}_i \tag{4.3}$$

The R.E. estimator is the OLS estimator for (4.3). The pooled OLS estimator:

$$\sqrt{N}\left(\hat{\beta}_{\text{pool}} - \beta_0\right) \Rightarrow N\left(0, \mathbb{E}[\vec{X}^T\vec{X}]^{-1}\mathbb{E}[\vec{X}^T\vec{e}\vec{e}^T\vec{X}]\mathbb{E}[\vec{X}^T\vec{X}]^{-1}\right)$$

where $\mathbb{E}[\vec{X}^T\vec{e}\vec{e}^T\vec{X}] = \vec{X}^T\mathbb{E}[\vec{e}\vec{e}^T \mid \vec{X}]\vec{X}$. Specifically, $\mathbb{E}[e_s e_t \mid \vec{X}] = \mathbb{E}[\alpha^2 + \epsilon_s\epsilon_t \mid \vec{X}] \neq 0, \forall s \neq t$.

$$\begin{aligned}
\mathbb{E}[\vec{e}\vec{e}^T \mid \vec{X}] &= \mathbb{E}[(\alpha\mathbf{1} + \vec{\epsilon})(\alpha\mathbf{1} + \vec{\epsilon})^T \mid \vec{X}] \\
\text{(assuming } \alpha \perp \vec{\epsilon}\text{)} \quad &= \mathbb{E}[\alpha^2\mathbf{1}\mathbf{1}^T \mid \vec{X}] + \mathbb{E}[\vec{\epsilon}\vec{\epsilon}^T \mid \vec{X}] \\
\text{(assuming homoscedasticity and } \mathrm{Cov}(\epsilon_s, \epsilon_t) = 0\text{)} \quad &= \sigma_\alpha^2\mathbf{1}\mathbf{1}^T + \sigma_\epsilon^2 I \\
&:= \Omega
\end{aligned}$$

Given $\Omega$ (or $\hat{\Omega}$),

$$\hat{\beta}_{RE} = \left(\frac{1}{N}\sum_{i=1}^{N}\vec{X}_i^T\Omega^{-1}\vec{X}_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\vec{X}_i^T\Omega^{-1}\vec{Y}_i\right)$$

So,

$$\sqrt{N}\left(\hat{\beta}_{RE} - \beta_0\right) \Rightarrow N\left(0, \underbrace{(\mathbb{E}[\vec{X}^T\Omega^{-1}\vec{X}])^{-1}}_{V_{RE}}\right)$$

**Hausmon Test** We want to test $H_0 : \mathrm{Cov}(\alpha_i, X_{i_t}) = 0$. Under $H_0$:

$$\begin{aligned}
\sqrt{N}\left(\hat{\beta}_{RE} - \beta_0\right) &\Rightarrow N(0, V_{RE}) \\
\sqrt{N}\left(\hat{\beta}_{FE} - \beta_0\right) &\Rightarrow N(0, V_{FE}) \\
\text{where } V_{FE} &\geq V_{RE}
\end{aligned}$$

**Theorem 4.10**

*Under $H_0$, $\hat{H} := N\left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right)^T(V_{FE} - V_{RE})^{-1}\left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right) \Rightarrow \chi^2_{\dim(\beta_0)}$.* ♡

### 4.8.4 Two-Way Fixed Effect Model

In this model, we consider an extra "time fixed effect" $V_t$.

$$Y_{i_t} = \alpha_i + V_t + X_{i_t}\beta_0 + \epsilon_{i_t}$$

1. Principle of deleting fixed effect:

$$\dot{Y}_{i_t} = Y_{i_t} - \bar{Y}_i - \bar{Y}_t + \bar{Y}$$

where $\bar{Y}_t := \frac{1}{N}\sum_{i=1}^{N} Y_{i_t}$ and $\bar{Y} := \frac{1}{NT}\sum_{t,i} Y_{it}$. Then,

$$\dot{Y}_{i_t} = \dot{X}_{i_t}\beta_0 + \dot{\epsilon}_{i_t}$$

where $\dot{X}_{i_t}$ and $\dot{\epsilon}_{i_t}$ are given in the same way.

2. Hybrid Model (better?):

$$Y_{i_t} = \alpha_i + \gamma_2\delta 2_t + \gamma_3\delta 3_t + \cdots + \gamma_T\delta T_t + X_{i_t}\beta_0 + \epsilon_{i_t}$$

where $\delta s_t = \begin{cases} 1, & s = t \\ 0, & s \neq t \end{cases}$. Then, in the matrix form,

$$Y_{i_t} = \alpha_i + Z_{i_t}^T \Theta + \epsilon_{i_t}, \text{ where } Z_{i_t}^T = \begin{bmatrix} X \\ \delta 2 \\ \vdots \\ \delta T \end{bmatrix}$$

### 4.8.5 Arellano Bond Approach

1. "Strict exogeneity": $\mathbb{E}[X_s \epsilon_t] = 0, \forall s, t$ ($\epsilon$ is uncorrelated with past, present, and future $X$'s).
2. "Sequential exogeneity": $\mathbb{E}[X_s \epsilon_t] = 0, \forall t \geq s$ ($\epsilon$ is uncorrelated with past $X$'s).

Reminds that Fixed Effect model has assumption $\mathbb{E}[\vec{X}_i \vec{\epsilon}_i] = 0$, which can be given by "strict exogeneity". However, the assumption of "strict exogeneity" is too strong.

**Example 4.2** $Y_{i_t} = \alpha_i + \rho \underbrace{Y_{i_{t-1}}}_{X_{i_t}} + \epsilon_{i_t}$, which doesn't satisfy the "strict exogeneity": $\mathbb{E}[X_{i_{t+1}} \epsilon_{i_t}] = \mathbb{E}[Y_{i_t} \epsilon_{i_t}] \neq 0$.

Instead of using the "strict exogeneity" assumption, <u>we can use "sequential exogeneity" assumption</u>.

Consider model

$$\Delta Y_{i_t} = \Delta X_{i_t} \beta_0 + \Delta \epsilon_{i_t}$$

we have

$$\mathbb{E}[X_s(\Delta \epsilon_t)] = \underbrace{\mathbb{E}[X_s \epsilon_t]}_{=0, \forall s \leq t} - \underbrace{\mathbb{E}[X_s \epsilon_{t-1}]}_{=0, \forall s \leq t-1}$$

Moreover, we suppose $\mathbb{E}[X_s \Delta X_t] \neq 0$, then $\{X_s, s \leq t-1\}$ are I.V. for the model above!

$\mathbb{E}[X_s (\Delta Y_t - \Delta X_t \beta_0)] = 0, \forall t, s : s \leq t - 1$.

| | |
|---|---|
| $t = 2$ | $\mathbb{E}[X_1 (\Delta Y_2 - \Delta X_2 \beta_0)]$ |
| $t = 3$ | $\mathbb{E}[X_1 (\Delta Y_3 - \Delta X_3 \beta_0)]$ |
| | $\mathbb{E}[X_2 (\Delta Y_3 - \Delta X_3 \beta_0)]$ |
| $\vdots$ | $\vdots$ |

All in all, we have

$$\mathbb{E}[g(\Delta \vec{Y}, \Delta \vec{X}, \vec{X}, \beta_0)] = \begin{bmatrix} \mathbb{E}[X_1 (\Delta Y_2 - \Delta X_2 \beta_0)] \\ \mathbb{E}[X_1 (\Delta Y_3 - \Delta X_3 \beta_0)] \\ \mathbb{E}[X_2 (\Delta Y_3 - \Delta X_3 \beta_0)] \\ \vdots \end{bmatrix} = 0$$

We can use GMM to estimate the parameters:

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \left( \frac{1}{N} \sum_{i=1}^{N} g(\Delta \vec{Y}_i, \Delta \vec{X}_i, \vec{X}_i, \beta_0) \right)^T W \left( \frac{1}{N} \sum_{i=1}^{N} g(\Delta \vec{Y}_i, \Delta \vec{X}_i, \vec{X}_i, \beta_0) \right)$$

Arellano Bond estimator is GMM estimator over I.D.

## 4.9 Control Function Approach (another approach to handle endogenieity)

Another approach to handle endogenieity.

Suppose we are facing the problem of endogenieity that

$$Y_i = X_i \beta_i + U_i,\ \mathbb{E}[U|X] \neq 0$$

Suppose $W$ is a variable that

$$\mathbb{E}[U|X, W] = \varphi(W)$$

which is only a function of $W$. That is, the relationship between $X$ and $U$ can only be determined by $W$: $X \to W \to U$.

**Definition 4.4 (Control Variable)**

W is a **Control Variable**.

♣

A control variable doesn't have to be an I.V.

**Example 4.3** $X = Z\gamma + V$, where $Z$ is I.V. that $\mathbb{E}[ZU] = 0$. $\mathbb{E}[U|X, V] = \varphi(V)$.

Based on the control variable, we can write the regression as

$$Y_i = X_i \beta_0 + \gamma W_i + U_i$$

$$Y_i = X_i \beta_0 + \gamma W_i + \varphi(W_i) + \underbrace{U_i - \varphi(W_i)}_{\xi_i}$$

where $\mathbb{E}[\xi_i | X_i, W_i] = 0$.

To implement this, we can decompose $\varphi(W_i) := \sum_{l=1}^{L} \pi_l \phi_l(W_i)$ (e.g. polynomial).

**Note** *We may get inconsistent* $\gamma$.

**Example 4.4** Suppose $\varphi(W) = \Pi W$, then $Y_i = X_i \beta_0 + \underbrace{(\gamma + \Pi)}_{\beta_1} W_i + \xi_i$. Hence, in OLS, $\hat{\beta}_0 \xrightarrow{P} \beta_0$ and $\hat{\beta}_1 \xrightarrow{P} \beta_1 = \gamma + \Pi$.

## 4.10 LATE (Local ATE): Application of I.V. on Potential Outcomes

(Application of I.V.)

Consider the potential outcome framework: $X \in \{0, 1\}, Y(0), Y(1) : Y := XY(1) + (1 - X)Y(0)$.

The Average treatment effect (ATE) is

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

Consider another variable $Z \in \{0, 1\}$.

1. $X$: the assigned treatment of an agent.
2. $Z$: the intended treatment of an agent. (instrument)

Suppose $X(Z)$ be the potential treatment status $X(0), X(1)$. $X = ZX(1) + (1 - Z)X(0)$.

**Example 4.5** Some people are suggested to stay at home, but they don't.

We have $Z \to X \to Y$ and $Z$ doesn't have a direct effect on $Y$.

There are four possible cases:

1. Never Treated (NT): $X(0) = X(1) = 0$.
2. Always Treated (AT): $X(0) = X(1) = 1$.
3. Complies (C): $X(0) = 0, X(1) = 1$.
4. Defiers (D): $X(0) = 1, X(1) = 0$.

Usually, we assume the instruments are relevant and rule out the defiers.

**Assumption** $X_i(0) \leq X_i(1), \forall i$ *and* $X_j(0) < X_j(1)$ *for some* $j$.

$\hat{\beta}_{2SLS} = \frac{\widehat{\text{Cov}(Y,Z)}}{\widehat{\text{Cov}(X,Z)}} \xrightarrow{P} \frac{\text{Cov}(Y,Z)}{\text{Cov}(X,Z)}$

**Theorem 4.11**

$\frac{\text{Cov}(Y,Z)}{\text{Cov}(X,Z)} = \frac{\mathbb{E}[Y|Z=1]-\mathbb{E}[Y|Z=0]}{\mathbb{E}[X|Z=1]-\mathbb{E}[X|Z=0]}$

♡

**Proof 4.8**

$$\begin{aligned}
\text{Cov}(Y, Z) &= \mathbb{E}[YZ] - \mathbb{E}[Y]P(Z = 1) \\
&= \mathbb{E}[Y|Z = 1]P(Z = 1) - (\mathbb{E}[Y|Z = 1]P(Z = 1) + \mathbb{E}[Y|Z = 0]P(Z = 0))P(Z = 1) \\
&= P(Z = 1)\left(\mathbb{E}[Y|Z = 1](1 - P(Z = 1)) - \mathbb{E}[Y|Z = 0]P(Z = 0)\right) \\
&= P(Z = 1)P(Z = 0)\left(\mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0]\right)
\end{aligned}$$

*Similarly,*

$$\text{Cov}(X, Z) = P(Z = 1)P(Z = 0)\left(\mathbb{E}[X|Z = 1] - \mathbb{E}[X|Z = 0]\right)$$

Since we rule out the possible of $(D)$, we can write

$$\begin{aligned}
&\mathbb{E}[Y|Z = 1] \\
=&\mathbb{E}[Y|AT, Z = 1]\text{Pr}(AT|Z = 1) + \mathbb{E}[Y|NT, Z = 1]\text{Pr}(NT|Z = 1) + \mathbb{E}[Y|C, Z = 1]\text{Pr}(C|Z = 1) \\
=&\mathbb{E}[Y(1)|AT]\text{Pr}(AT) + \mathbb{E}[Y(0)|NT]\text{Pr}(NT) + \mathbb{E}[Y(1)|C]\text{Pr}(C)
\end{aligned}$$

We can also decompose the $\mathbb{E}[Y|Z=1]$.

$$\begin{cases} \mathbb{E}[Y|Z=1] & = \mathbb{E}[Y(1)|AT]\text{Pr}(AT) + \mathbb{E}[Y(0)|NT]\text{Pr}(NT) + \mathbb{E}[Y(1)|C]\text{Pr}(C) \\ \mathbb{E}[Y|Z=0] & = \mathbb{E}[Y(1)|AT]\text{Pr}(AT) + \mathbb{E}[Y(0)|NT]\text{Pr}(NT) + \mathbb{E}[Y(0)|C]\text{Pr}(C) \end{cases}$$

Then, we have

$$\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0] = \text{Pr}(C)\left(\mathbb{E}[Y(1)|C] - \mathbb{E}[Y(0)|C]\right)$$

We also have $\mathbb{E}[X|Z=1] = \text{Pr}(AT) + \text{Pr}(C)$ and $\mathbb{E}[X|Z=0] = \text{Pr}(AT)$. Hence,

$$\begin{aligned} \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[X|Z=1] - \mathbb{E}[X|Z=0]} &= \frac{\text{Pr}(C)\left(\mathbb{E}[Y(1)|C] - \mathbb{E}[Y(0)|C]\right)}{\text{Pr}(C)} \\ &= \mathbb{E}[Y(1)|C] - \mathbb{E}[Y(0)|C] \\ &= \mathbb{E}[Y(1) - Y(0)|C] \end{aligned}$$

which is called **LATE**.

**Proposition 4.7**

*With Assumption 4.10, the* ***LATE*** *is given by*

$$\mathbb{E}[Y(1) - Y(0)|C] = \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[X|Z=1] - \mathbb{E}[X|Z=0]} = \frac{\text{Cov}(Y,Z)}{\text{Cov}(X,Z)}$$

♠

**Remark**

1. In RCT, $\text{Pr}(C) = 1$, in which case ATE=LATE.

## 4.11 Difference in Difference (DiD)

The setup is the potential outcomes in Panel data.

Consider a two-way fixed effect model on the potential outcomes. For $D_{i_t} \in \{0,1\}$, $Y_{i_t}$ is given by

$$Y_{i_t}(0) = \alpha_i + \delta_t + \gamma X_{i_t} + \epsilon_{i_t}(0)$$

$$Y_{i_t}(1) = \alpha_i + \delta_t + \gamma X_{i_t} + \epsilon_{i_t}(1) + \theta$$

**Assumption** *We use following assumptions:*

1. $\epsilon_{i_t}(0) = \epsilon_{i_t}(1) := \epsilon_{i_t}$
2. $\mathbb{E}[\epsilon_{i_t}|X_{i_t}] = 0$

The ATE is given by

$$ATE := \mathbb{E}[Y_t(1) - Y_t(0)] = \theta + \underbrace{\mathbb{E}[\epsilon_{i_t}(1) - \epsilon_{i_t}(0)]}_{\text{by assumption } = 0}$$

**Lemma 4.4**

*With Assumption 4.11, $ATE = \theta$.*

♡

$$Y_{i_t} = D_{i_t} Y_{i_t}(1) + (1 - D_{i_t}) Y_{i_t}(0) = \alpha_i + \delta_t + \theta D_{i_t} + \gamma X_{i_t} + \epsilon_{i_t}$$

### 4.11.1 After OLS Regression

Let $T = 2$, we have

$$Y_{i_2} = \delta_2 + \theta D_{i_2} + \gamma X_{i_2} + e_{i_2}, \text{ where } e_{i_2} = \alpha_i + \epsilon_{i_2}$$

**Theorem 4.12**

*If* $\mathbb{E}[e_{i_2} | X_{i_2}, D_{i_2}] = \Pi_0 + \Pi_1 X_{i_2}$, *then the control function estimator (OLS) is consistent:*

$$\hat{\theta}_{\text{CF}} \xrightarrow{P} ATE = \theta$$

♡

However, what if $\alpha_i < \alpha_j$, the assumption $\mathbb{E}[e_{i_2} | X_{i_2}, D_{i_2}] = \Pi_0 + \Pi_1 X_{i_2}$ doesn't hold.

### 4.11.2 Difference in Difference

$$\Delta Y_i := Y_{i_2} - Y_{i_1} = \underbrace{\delta_2 - \delta_1}_{\delta} + \theta \Delta D_i + \gamma \Delta X_i + \Delta \epsilon_i$$

**Case without covariate** ($\gamma = 0$)

$$\Delta Y_i = \delta + \theta D_{i_2} + \Delta \epsilon_i$$

**Assumption** *[Parallel Trends Assumption]* $\mathbb{E}[\Delta\epsilon | D_2 = 1] = \mathbb{E}[\Delta\epsilon | D_2 = 0]$.

**Theorem 4.13**

*Parallel Trends Assumption is equivalent to each of following conditions.*

$$\begin{aligned} PT &\Leftrightarrow \mathbb{E}[\Delta Y(1) | D_2 = 1] = \mathbb{E}[\Delta Y(1) | D_2 = 0] \\ &\Leftrightarrow \mathbb{E}[\Delta Y(0) | D_2 = 1] = \mathbb{E}[\Delta Y(0) | D_2 = 0] \\ &\Leftrightarrow \text{Cov}(D_2, \Delta\epsilon) = 0 \end{aligned}$$

♡

The DiD estimator is numerically same with OLS:

$$\hat{\theta}_{\text{DiD}} = \frac{\frac{1}{N}\sum_{i=1}^{N} \Delta Y_i D_{i_2}}{\frac{1}{N}\sum_{i=1}^{N} D_{i_2}} - \frac{\frac{1}{N}\sum_{i=1}^{N} \Delta Y_i (1 - D_{i_2})}{1 - \frac{1}{N}\sum_{i=1}^{N} D_{i_2}} \tag{DiD}$$

**Case with covariates**

$$\Delta Y_i = \delta + \theta D_{i_2} + \gamma \Delta X_i + \Delta \epsilon_i$$

**Assumption** $\mathbb{E}[\Delta\epsilon|D_2 = 1, \Delta X] = \mathbb{E}[\Delta\epsilon|D_2 = 0, \Delta X]$, *which is equivalent to* $\mathbb{E}[\Delta Y(d)|D_2 = 1, \Delta X] = \mathbb{E}[\Delta Y(d)|D_2 = 0, \Delta X], \forall d \in \{0, 1\}$.

**Remark** The DiD estimator (DiD) is no longer consistent:

$$\hat{\theta}_{\text{DiD}} \xrightarrow{P} \theta + \underbrace{\gamma \left(\mathbb{E}[\Delta X|D_2 = 1] - \mathbb{E}[\Delta X|D_2 = 0]\right)}_{\text{“selection on observables”}}$$