Convergence concept.

$\{X_n\}_{i=1}^n$ probability space (Ω, F, P).

$X_n: \Omega \to \mathbb{R}$. $X_n(\omega)$

Def: $X_n \to X$ a.s.

$P(\{\omega \in \Omega: \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1$

or: $P(\lim_{n\to\infty} X_n = X) = 1$

most generally: $P(\lim_{n\to\infty} d(X_n, X) = 0) = 1$ for random element X_n.

Def. Convergence in mean square:

$r \in (0, +\infty)$. $\lim_{n\to\infty} E[|X_n - X|^r] = 0$

for $r=2$: proposition:

$X_n \to C$ in mean square iff $\begin{cases} \lim_{n\to\infty} E[X_n] = C \\ \lim_{n\to\infty} \mathrm{var}(X_n) = 0 \end{cases}$.

Def. $X_n \xrightarrow{P} X$:

$\forall \varepsilon > 0$, $\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$

or $\lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1$

most generally: $\lim_{n\to\infty} P(d(X_n - X) \geq \varepsilon) = 0$

Def. $X_n \xrightarrow{d} X$:

if: $F_{X_n}(x) \to F_X(x)$ $\forall x \in C(F_X)$.

(F_X: limit distribution function)

Relationships:

a.s. $\Rightarrow$ in prob. $\Rightarrow$ in dist. $\quad X_n \xrightarrow{d} \alpha \Rightarrow X_n \xrightarrow{P} \alpha$

(in prob. $\Leftarrow$ in r-mean)

Theorem: Portmanteau Lemma:

① $X_n \xrightarrow{d} X$ <=> ② $\lim_{n\to\infty} P(X_n \in A) = P(X \in A)$.

Theorem: Algebra:

① $X_n \xrightarrow{P} X$, $Y_n \xrightarrow{P} Y \Rightarrow X_n + Y_n \xrightarrow{P} X + Y$

② $X_n \to X$ a.s., $Y_n \to Y$ a.s. $\Rightarrow X_n + Y_n \to X + Y$ a.s.

③ $X_n \to X$ in r-mean, $Y_n \to Y$ in r-mean $\Rightarrow$ same.

④ $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{d} Y$, $X_n \perp Y_n$ $\forall n$, $X \perp Y$:

$X_n + Y_n \xrightarrow{d} X + Y$

⑤ $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{d} \alpha$, $\alpha \in \mathbb{R} \Rightarrow X_n + Y_n \xrightarrow{d} X + \alpha$

⑥ $X_n \xrightarrow{P} X$, $Y_n \xrightarrow{P} Y$: then:

$X_n - Y_n \xrightarrow{P} X - Y$ $\quad X_n Y_n \xrightarrow{P} XY$

if $P(X_n \neq 0) = 1$ and $P(Y_n \neq 0) = 1$, then $\frac{X_n}{Y_n} \xrightarrow{P} \frac{X}{Y}$.

for r.vectors:

⑦ $X_n \xrightarrow{P} X$, $Y_n \xrightarrow{P} Y \Rightarrow (X_n, Y_n) \xrightarrow{P} (X, Y)$

⑧ $X_n \to X$ a.s., $Y_n \to Y$ a.s. $\Rightarrow (X_n, Y_n) \to (X, Y)$ a.s.

⑨ $X_n \xrightarrow{d} X$, $Y_n \to C$, $C \in \mathbb{R}$, $\Rightarrow (X_n, Y_n) \xrightarrow{d} (X, C)$.

The continuity theorem and continuous mapping theorem.

Theorem: Lévy's continuous theorem:

$\{X_n\}$: r.v. $\phi_{X_n} \to \phi_X$ pointwise on all $\mathbb{R}$ $\Leftrightarrow X_n \xrightarrow{d} X$

Theorem: continuous mapping theorem:

$g: \mathbb{R} \to \mathbb{R}$ continuous. $\{X_n\}$ r.v.

① $X_n \xrightarrow{P} X \Rightarrow g(X_n) \xrightarrow{P} g(X)$ ② $\to$ a.s. ③ $\xrightarrow{d}$.

Stochastic Order Notation: unfamiliar.

Def: $X_n = O_p(r_n)$:

$\{X_n\}$ r.v. $\{r_n\} \in \mathbb{R}$, $r_n > 0$

$\forall \varepsilon > 0$, $\exists M_\varepsilon$ and N_ε, for all $n \geq N_\varepsilon$:

$P(|\frac{X_n}{r_n}| \geq M_\varepsilon) = P(|X_n| \geq r_n M_\varepsilon) \leq \varepsilon$

$\Leftrightarrow P(|\frac{X_n}{r_n}| \leq M_\varepsilon) \geq 1 - \varepsilon$; $\frac{X_n}{r_n}$ is bounded in probability

Def: $X_n = o_p(r_n)$:

$\{X_n\}$: r.v. $\{r_n\} \in \mathbb{R}$, $r_n > 0$

$|\frac{X_n}{r_n}| \xrightarrow{P} 0$ as $n \to \infty$

r_n: $X_n \xrightarrow{P} 0$ at the rate r_n.

Theorem: $\{r_n\}$: r.v.

① $o_p(1) + o_p(1) = o_p(1)$

② $o_p(1) + O_p(1) = O_p(1)$.

③ $O_p(1) o_p(1) = o_p(1)$.

④ $(1 + o_p(1))^{-1} = O_p(1)$

⑤ $o_p(r_n) = r_n o_p(1)$

⑥ $O_p(r_n) = r_n O_p(1)$.

⑦ $o_p(O_p(1)) = o_p(1)$.

Theorem: $\hat{\theta}_n$: r.v. θ: scalar.

suppose $r_n(\hat{\theta}_n - \theta) \xrightarrow{d} Z$ for some $\uparrow$ $z_n > 0$, then $\hat{\theta}_n \xrightarrow{P} \theta$.

proof: $r_n(\hat{\theta}_n - \theta) \xrightarrow{d} Z \Rightarrow r_n(\hat{\theta}_n - \theta) = O_p(1)$, thus, $\hat{\theta}_n - \theta = \frac{1}{r_n} r_n(\hat{\theta}_n - \theta) = \frac{1}{r_n} O_p(1) = o_p(1) O_p(1) = o_p(1)$ $\therefore \hat{\theta}_n \xrightarrow{P} \theta$.

LLN:

Theorem: Bernoulli's LLN:

Let $X_1 \cdots X_n$ iid Bernoulli r.v. with p. $\bar{X}_n \xrightarrow{P} p$

Theorem: Weak LLN:

$\{X_i\}_{i=1}^N$ r.v. suppose: ① independent $\{X_i\}$ ② $E[X_i] = \mu < \infty, \forall i$ ③ $\exists M > 0$, $\mathrm{var}(X_i) \leq M, \forall i$

then: $\bar{X}_n \xrightarrow{P} \mu$

pf: $E[\bar{X}_n] = \mu$, $\mathrm{var}(\bar{X}_n) = \mathrm{var}(\frac{1}{n}\sum_{i=1}^N X_i) = \frac{1}{n^2}\mathrm{var}(\sum_{i=1}^N X_i) \leq \frac{M}{n}$. $P(|\bar{X}_n - E[\bar{X}_n]| \geq \varepsilon) \leq \frac{\mathrm{var}(\bar{X}_n)}{\varepsilon^2} \leq \frac{1}{n}\frac{M}{\varepsilon^2} \to 0$ as $n \to \infty$.

Theorem: Kolmogorov's 2nd strong LLN: most important, strongest.
$\{X_i\}$: iid r.v., then $\bar{X}_n \to \mu$ a.s. as $n \to \infty$ iff $\mathbb{E}[X_i] = \mu < \infty$ exists.

CLT.

Theorem: Lindeberg-Lévy CLT:
$\{X_j\}$ iid r.v. $\mathbb{E}[X_j] = \mu$, $\text{var}(X_j) = \sigma^2$, then: $\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{d} N(0,1)$.

Properties of Normal Distribution:
$X_i \sim N(\mu, \sigma^2)$ $\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$ $\frac{1}{n}\sum_{i=1}^{n}\frac{X_i-\mu}{\sigma} \sim N(0, \frac{1}{n})$ $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{X_i-\mu}{\sigma}\right) \sim N(0,1)$

Theorem: Cramér-Wold device
$\{X_n\}$ a sequence of random k-vectors. X random k-vector. Then the followings are equivalent
① $X_n \xrightarrow{d} X$ ② $\lambda' X_n \xrightarrow{d} \lambda' X$ for all $\lambda \in \mathbb{R}^k$.

Theorem: Multivariate CLT:
$\{X_i\}$: iid random k-vectors. $\mathbb{E}[X]$: common mean of X_i. $V = \text{var}(X_i)$ common covariance matrix (finite non-singular)
$\sqrt{n}(\bar{X}_n - \mathbb{E}[X]) \xrightarrow{d} N(0, V)$.

The Delta Method.

Theorem. Univariate delta method:
$\{Z_n\}$ r.v. $g: \mathbb{R} \to \mathbb{R}$. Let $\mu \in \mathbb{R}$.
Suppose that $g'(\mu)$ exists, $\neq 0$, continuous at μ.
Suppose: $\sqrt{n}(Z_n - \mu) \xrightarrow{d} N(0, \sigma^2)$ as $n \to \infty$
then $\sqrt{n}(g(Z_n) - g(\mu)) \xrightarrow{d} N(0, g'(\mu)^2\sigma^2)$

pf: Taylor's expansion near μ:
$g(Z_n) = g(\mu) + g'(\tilde{Z}_n)(Z_n - \mu)$, $\tilde{Z}_n$ between μ and Z_n
$g(Z_n) - g(\mu) = g'(\tilde{Z}_n)(Z_n - \mu)$ ($\tilde{Z}_n \xrightarrow{p} \mu$)
$\sqrt{n}(g(Z_n) - g(\mu)) = \sqrt{n}\, g'(\tilde{Z}_n)(Z_n - \mu)$
$\because \sqrt{n}(Z_n \to \mu) \xrightarrow{d} N(0, \sigma^2)$ $\therefore Z_n \to \mu$ $\therefore \tilde{Z}_n \to \mu$.
$\therefore \sqrt{n}(g(Z_n) - g(\mu)) \xrightarrow{d} N(0, g'(\mu)^2\sigma^2)$.

Theorem: Multivariate Delta Method.
$g: \mathbb{R}^k \to \mathbb{R}^l$ have non-zero derivative $\nabla g(a)$ at $a \in \mathbb{R}^k$.
$b > 0$ scalar. $\{Z_n\}$ r. k-vectors. Z r. k-vector
Suppose: $n^b(Z_n - a) \xrightarrow{d} Z$
then: $n^b(g(Z_n) - g(a)) \xrightarrow{d} [\nabla g(a)]^2 Z$
e.g. $\{X_i\}$: iid $f_X(x) = \alpha e^{-\alpha x}$, $\mathbb{E}[X] = \frac{1}{\alpha}$ $\alpha = \frac{1}{\mathbb{E}[X]}$
$\hat{\alpha} \equiv \frac{1}{\bar{X}_n}$. By CMT: $\hat{\alpha} \xrightarrow{p} \alpha$. $\text{var}(X) = \frac{1}{\alpha^2}$
$\hat{\alpha} = g(\bar{X}_n) = \frac{1}{\bar{X}_n}$ $\therefore g(t) = \frac{1}{t}$ $g'(t) = -\frac{1}{t^2}$
$g(\mathbb{E}[X]) = g(\frac{1}{\alpha}) = \alpha$ $g'(\mathbb{E}[X]) = g'(\frac{1}{\alpha}) = -\frac{1}{(1/\alpha)^2}$
$\therefore \sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{d} N(0, [-\frac{1}{(1/\alpha)^2}]^2 \cdot \frac{1}{\alpha^2})$
$\therefore \hat{\alpha} \xrightarrow{d} N(\alpha, \frac{\alpha^2}{n})$

Theory of Statistics.

$\mathcal{A}$: action space $\mathbb{P} \in \mathcal{P}$: state of the world. $U(\mathbb{P}, a)$: utility function $X = (X_1 \cdots X_n)$: observed realization
$d(X) \in \mathcal{A}$: our decision
Def. $d: \mathbb{R}^n \to \mathcal{A}$: a statistical decision rule.
Want to solve: $\max_{d(X)} U(\mathbb{P}, d(X))$

Subjective Bayesian Approach: ex post
① have prior belief $\pi(\cdot)$ about $\mathbb{P}$.
② update it to posterior belief $\pi(\cdot|x)$ after seeing X
③ $\max_{d(X)} \int U(\mathbb{P}, d(X))\, d\pi(\mathbb{P}|x)$.

Problems: ① $\pi(\cdot)$ come from?
② $\mathbb{P}$ may be non-parametric.

Frequentist's Approach: ex ante
Def: the distribution of d is the sampling distribution,
where $d: \mathbb{R}^n \to \mathcal{A}$. (dist. of $(X_1 \cdots X_n)$ induces it via the r.v. $d(X)$)
depends on $\mathbb{P}$ and d.
Def. welfare function: (ex ante) $W(\mathbb{P}, d) = \int U(\mathbb{P}, d(x))\, d\mathbb{P}(x)$ (unknown since $\mathbb{P}$ unknown.)
$= \mathbb{E}_{\mathbb{P}}[U(\mathbb{P}, d(X))$ eliminate X.

2 Ways to get rid of $\mathbb{P}$:
①
(i) suppose a prior dist. $\pi(\cdot)$ of $\mathbb{P}$: $B(d) = \int W(\mathbb{P}, d)\, d\pi(\mathbb{P})$
(ii) $\max_{d(\cdot)} B(d)$ → Bayesian optimal decision rule. d^*

② Worst case scenario:

(i) for each choice of d: $\text{Worstcase}(d) = \min_{\mathbb{P}\in\mathcal{P}} W(\mathbb{P}, d)$

(ii) $\max_{d(\cdot)} \text{Worstcase}(d) \cdot \longrightarrow$ maximin optimal decision rule d^*.

Point Estimation.

$\mathbb{P}$: probability measure of interest. $\theta = \theta(\mathbb{P})$: parameter of interest. Θ: parameter space (set of all logical possible values of θ).

Def. $X_1 \cdots X_n$ iid. A point estimator is a measurable function $\hat{\theta}: \mathbb{R}^n \to \Theta$.

$(\hat{\theta}_n) = \hat{\theta} = \hat{\theta}(X_1 \cdots X_n)$: estimator $\hat{\theta}(x_1 \cdots x_n)$: estimate (with realized $x_1 \cdots x_n$).

Good $\hat{\theta}(\cdot)$: ① close ② precise.

Finite Sample Version:

Closedness: $Bias(\hat{\theta}, \mathbb{P}) = \mathbb{E}[\hat{\theta}] - \theta = \mathbb{E}_{\mathbb{P}}(\hat{\theta}(X_1 \cdots X_n)) - \theta$

$= 0$, then unbiased.

Precision: $Var(\hat{\theta}) = Var_{\mathbb{P}}(\theta(X_1 \cdots X_n))$ as small as possible.

Balancing: $MSE(\hat{\theta}, \mathbb{P}) = \mathbb{E}_{\mathbb{P}}[(\hat{\theta}(X) - \theta)^2] = Bias^2(\hat{\theta}, \mathbb{P}) + Var(\hat{\theta}, \mathbb{P})$.

if $MSE(\hat{\theta}_1, \mathbb{P}) \le MSE(\hat{\theta}_2, \mathbb{P})$ $\forall n$ $\forall \mathbb{P}$, then $\hat{\theta}_1$ is relatively more efficient than $\hat{\theta}_2$.

Minimax optimal estimator:

$$\min_{\hat{\theta}(\cdot)} \max_{\mathbb{P}\in\mathcal{P}} MSE(\hat{\theta}, \mathbb{P}).$$

Asymptotic Version:

Closedness:

1. (Weakly) consistency: $\hat{\theta} \xrightarrow{p} \theta$ for all $\mathbb{P}$, $\theta(\mathbb{P})$. as $n \to \infty$.
2. Unbiased in the limit: $\mathbb{E}[\hat{\theta}] \to \theta$ as $n \to \infty$.
3. Asymptotically unbiased: $r_n(\hat{\theta} - \theta) \xrightarrow{d} Z$ if $\mathbb{E}[Z] = 0$, $r_n \to \infty$. Z non-degenerate r.v.

$3 \to 1$

Precision:

$Var(\hat{\theta}) \to 0$ in most cases. so, we look at:

1. $\lim_{n\to\infty} Var(r_n(\hat{\theta} - \theta))$, where $r_n(\hat{\theta} - \theta) \to Z$.
2. $r_n(\hat{\theta} - \theta) \to Z$, look at the dist. of Z. e.g. $var(z)$. $\begin{cases} \sqrt{n}(\hat{\theta}_1 - \theta) \to Z_1 \\ \sqrt{n}(\hat{\theta}_2 - \theta) \to Z_2 \end{cases}$ e.g. $var(z_1) > var(z_2)$ then $\hat{\theta}_2$ more precise.

Hypothesis Tests

$X_1 \cdots X_n$: iid $\mathbb{P}$, $\theta = \theta(\mathbb{P})$. $\Theta = \Theta_0 \cup (\Theta \setminus \Theta_0)$ $\theta \in \Theta_0$ or $\theta \in \Theta \setminus \Theta_0$

Null Alternative

Definition: A hypothesis test is a measurable function $\phi_n = \phi_n(X_1 \cdots X_n): \mathbb{R}^n \to \{0, 1\}$. 0: accepting Θ_0. 1: rejecting Θ_0.

Def. Power Function: Let ϕ_n be a test. Define the function power: $\mathcal{P} \to [0, 1]$ by: $Power(\mathbb{P}, \phi_n) = \mathbb{P}(\phi_n(X) = 1)$.

Def. Risk function of the test ϕ_n is:

$$R(\mathbb{P}, \phi_n) = L(\mathbb{P}, 1)\mathbb{P}(\phi_n(X) = 1) + L(\mathbb{P}, 0)\mathbb{P}(\phi_n(X) = 0)$$

$$= L(\mathbb{P}, 0) + \mathbb{P}(\phi_n(X) = 1)(L(\mathbb{P}, 1) - L(\mathbb{P}, 0)).$$

where $L(\mathbb{P}, a)$ is the loss function. $\phi_n(x) \to \mathbb{P}(\phi_n(X) = 1) \to R(\mathbb{P}, \phi_n)$

Def. 0-1 loss function (most standard one): $L(\mathbb{P}, a) = \begin{cases} 1 & \text{if } \theta(\mathbb{P}) \in \Theta_0 \text{ and } a = 1 \text{ (false positive)} \quad \text{or} \quad \theta(\mathbb{P}) \notin \Theta_0 \text{ and } a = 0 \text{ (false negative)} \\ 0 & \text{otherwise.} \end{cases}$

Under 0-1 loss function, $R(\mathbb{P}, \phi_n) = \begin{cases} \mathbb{P}(\phi_n(X)=1) & \text{if } \theta(\mathbb{P}) \in \mathbb{H}_0 \\ 1-\mathbb{P}(\phi_n(X)=1) & \text{if } \theta(\mathbb{P}) \notin \mathbb{H}_0 \end{cases}$

Best power function: $\text{power}(\mathbb{P}, \phi_n^*) = \begin{cases} 1 & \text{if } \theta(\mathbb{P}) \notin \mathbb{H}_0 \\ 0 & \text{if } \theta(\mathbb{P}) \in \mathbb{H}_0 \end{cases}$ not feasible! Instead, find "good one" in some sense.

The Neyman-Pearson Paradigm.

First, restrict attention to tests whose prob. of false positive is not too big.

Def. Size of a test ϕ_n: $\text{size}(\phi_n) = \sup_{\mathbb{P} \in \mathcal{P}: \theta(\mathbb{P}) \in \mathbb{H}_0} \mathbb{P}(\phi_n(X)=1)$, (largest prob. of false positive under all $\mathbb{P} \in \mathcal{P}$).

Def. ϕ_n is a level α test if $\text{size}(\phi_n) \leq \alpha$, $\alpha \in (0,1)$

Next, choose the one that has the largest power.

Def: ϕ_n^* is the uniformly most powerful (UMP) within $\mathfrak{C}$ if, for any $\phi_n \in \mathfrak{C}$,

$\text{Power}(\mathbb{P}, \phi_n^*) \geq \text{Power}(\mathbb{P}, \phi_n)$

for all $\mathbb{P}$ such that $\theta(\mathbb{P}) \notin \mathbb{H}_0$.

Problem & solution: see weekly summary.

Cutoff test. (Most hypothesis tests have this form):

$\phi_n(x) = \mathbb{1}[T(x) > c]$

$T(x)$: test statistics. c: critical value.

Asymptotic Criteria:

Def. the uniform asymptotic size of ϕ_n is: $\text{Asymptsize}(\phi_n) = \limsup_{n\to\infty} \sup_{\mathbb{P} \in \mathcal{P}: \theta(\mathbb{P}) \in \mathbb{H}_0} \mathbb{P}(\phi_n(X)=1)$.

Def. ϕ_n is pointwise consistent if $\lim_{n\to\infty} \mathbb{P}(\phi_n(X)=1) = 1$ for all $\mathbb{P} \in \mathcal{P}$ such that $\theta(\mathbb{P}) \notin \mathbb{H}_0$.

p-values.

Def. let $\mathfrak{C}$ be the set of all non-randomized tests of the null hypothesis $\mathbb{H}_0$. Let $x \in \text{supp}(X)$.

$$p(\mathbb{H}_0, x) = \inf_{\phi_n \in \mathfrak{C}: \phi_n(x)=1} \sup_{\mathbb{P} \in \mathcal{P}: \theta(\mathbb{P}) \in \mathbb{H}_0} \mathbb{P}(\phi_n(X)=1)$$

$$= \inf_{\phi_n \in \mathfrak{C}: \phi_n(x)=1} \text{size}(\phi_n)$$

The definition based on a specific cutoff test:

Let the test be $\phi_n(X) = \mathbb{1}[T(X) > c]$

let $c = c(1-\alpha)$. $c'(\cdot) > 0$. As $\alpha \uparrow$, $c \downarrow$. As $\alpha \downarrow$, $c \uparrow$. (size $\downarrow$, critical value $\uparrow$, harder to reject).

$$\phi_n(X) = \mathbb{1}[T(x) > c(1-\alpha)]$$

let $x \in \text{supp}(X)$ be a specific realization of the data, so $T(x)$ is fixed.

Suppose that we reject. $T(x) > c(1-\alpha)$.

this is not strict enough since $T(x) - c(1-\alpha) > 0$, so we let $\alpha \downarrow$ to p so that $c(1-\alpha) \uparrow$ to $c(1-p)$

$$T(x) = c(1-p)$$

$p = 1 - C^{-1}(T(x))$: this is the p-value within the restrict class of $\mathfrak{C}$.

① $p < \alpha$.

② as $\alpha \downarrow$ to p, still can reject $\mathbb{H}_0$ because $T(x) = c(1-p) > c(1-\alpha)$

③ as $\alpha \downarrow$ to be less than p, cannot reject $\mathbb{H}_0$ because $T(x) = c(1-p) < c(1-\alpha)$ now.

Add an assumption : $C^{-1}(t) = \mathbb{P}_0(T(X) \le t)$ *

$\mathbb{P}_0$: the dist. of X that $\theta(\mathbb{P}_0) = \theta$. $C^{-1}(t)$: cdf of $T(x)$ when true dist. is $\mathbb{P}_0$. $C(\cdot)$: inverse of cdf: quantile function.

Then: $p(x) = 1 - \mathbb{P}_0(T(X) \le T(x))$
$= \mathbb{P}_0(T(X) \ge T(x))$.

Most common def: p-value for a cutoff test with $T(\cdot)$ and * is the prob. of observing a test statistic's value more extreme than the one we actually saw in the data, assuming that the true dist is $\mathbb{P}_0$.

p-value depend on x, so it is a r.v.

Lemma. define $p(x) = \mathbb{P}_0(T(X) \ge T(x))$.
Suppose $X \sim \mathbb{P}_0$. Then $p \sim \text{Unif}[0,1]$. (ex-ante).

pf: $\mathbb{P}_0(p(X) \le p) = \mathbb{P}_0(1 - F_{T(X)}(T(X)) \le p)$
$= \mathbb{P}_0(1-p \le F_{T(X)}(T(X)))$
$= \mathbb{P}_0(F^{-1}_{T(X)}(1-p) \le T(X))$
$= 1 - F_{T(X)}(F^{-1}_{T(X)}(1-p))$
$= 1 - (1-p)$
$= p$.

Theorem. define $p(x) = \mathbb{P}_0(T(X) \ge T(x))$. Let $\alpha \in (0,1)$.
define: $\phi_\alpha(x) = \mathbb{1}[p(x) \le \alpha]$
Then, $\phi_\alpha(x)$ is a level-α test of the simple null $\mathbb{P}_0$ against the alternative that $\mathbb{P} \ne \mathbb{P}_0$.

pf: $\mathbb{P}(\phi_\alpha(X) = 1) = \mathbb{P}_0(p(x) \le \alpha)$
$= \alpha$ according to the lemma.

↑ using $p(x)$ to construct level-α test.

Confidence Set.

it depends on sample size!

Def. A level α confidence set for θ is a function $C(X) = C(X_1 \cdots X_n)$ from the sample to the subset of (Θ) such that
1. $\{x \in \text{supp}(X) : C(x) \ni \theta(\mathbb{P})\}$ is $\mathbb{P}$-measurable and 2. $\mathbb{P}(C(x) \ni \theta(\mathbb{P})) \ge 1-\alpha$ for all $\mathbb{P} \in \mathcal{P}$.

Interpretation: $\theta(\mathbb{P})$ is fixed. throw sets at it and hit it at least $100(1-\alpha)\%$ of the times.

Asymptotic version: $\liminf_{n\to\infty} \inf_{\mathbb{P}\in\mathcal{P}} \mathbb{P}(C(X) \ni \theta(\mathbb{P})) \ge 1-\alpha$.

Duality with hypothesis tests: (equivalency)

Rejection region $R = \{x \in \text{supp}(X) : \phi_{\theta_0}(x) = 1\}$ flip the inequality!
Test inversion confidence set $C(x) = \{\theta \in (\Theta) : \phi_\theta(x) = 0\}$

① For each $\theta \in (\Theta)$, let $\phi_\theta(x)$ be a level α test of $H_0: \theta(\mathbb{P}) = \theta$. Let $C(x) = \{\theta \in (\Theta) : \phi_\theta(x) = 0\}$, then $C(x)$ is a level α confidence set.

② Let $C(x)$ be a level α confidence set. For each $\theta \in (\Theta)$, define $\phi_\theta(x) = \mathbb{1}[C(x) \not\ni \theta]$, then for each $\theta \in (\Theta)$, $\phi_\theta(x)$ is a level α test of $H_0: \theta(\mathbb{P}) = \theta$.

Regression Mechanism.

describing the conditional mean of an observed r.v. $Y \mid X = x$. Study $\mathbb{E}[Y \mid X = x]$

Basic Variables X
Derived Variables $p(X) = \begin{bmatrix} p_1(x) \\ \vdots \\ p_J(x) \end{bmatrix}$

Suppose $m(x) = \mathbb{E}[Y \mid X = x] = p(x)'\beta$, then.

Theorem: Y: r.v. X: r.vector. Assume:
① linearity: $m(x) = \mathbb{E}[Y \mid X = x] = p'(x)\beta$
② finite moments: $\mathbb{E}[p(x)Y]$ and $\mathbb{E}[p(x)p(x)']$ finite
③ sufficient variation: $\mathbb{E}[p(x)p(x)']$ invertible.
Then: $\beta = \mathbb{E}[p(x)p(x)']^{-1}\mathbb{E}[p(x)Y]$

proof: $\mathbb{E}[Y \mid X = x] = p(x)'\beta$
$p(x)\,\mathbb{E}[Y \mid X = x] = p(x)p(x)'\beta$
$\mathbb{E}[p(x)\mathbb{E}[Y \mid X = x]] = \mathbb{E}[p(x)p(x)']\beta$
$\mathbb{E}[p(x)Y] = \mathbb{E}[p(x)p(x)']\beta$
$\beta = \mathbb{E}[p(x)p(x)']^{-1}\mathbb{E}[p(x)Y]$

proof: $p(x)p(x)' = \begin{bmatrix}1\\ x\end{bmatrix}\begin{bmatrix}1 & x\end{bmatrix} = \begin{bmatrix}1 & x\\ x & x^2\end{bmatrix}$ $\mathbb{E}[p(x)p(x)'] = \begin{bmatrix}1 & \mathbb{E}[x]\\ \mathbb{E}[x] & \mathbb{E}[x^2]\end{bmatrix}$

$\mathbb{E}[p(x)p(x)']^{-1} = \frac{1}{\mathbb{E}[x^2] - \mathbb{E}[x]^2}\begin{bmatrix}\mathbb{E}[x^2] & -\mathbb{E}[x]\\ -\mathbb{E}[x] & 1\end{bmatrix}$

$\mathbb{E}[p(x)p(x)']^{-1}\mathbb{E}[p(x)Y] = \frac{1}{\mathbb{E}[x^2] - \mathbb{E}[x]^2}\begin{bmatrix}\mathbb{E}[x^2] & -\mathbb{E}[x]\\ -\mathbb{E}[x] & 1\end{bmatrix}\begin{bmatrix}\mathbb{E}[Y]\\ \mathbb{E}[xY]\end{bmatrix}$

$= \frac{1}{\mathbb{E}[x^2] - \mathbb{E}[x]^2}\begin{bmatrix}\mathbb{E}[x^2]\mathbb{E}[Y] - \mathbb{E}[x]\mathbb{E}[xY]\\ \mathbb{E}[xY] - \mathbb{E}[x]\mathbb{E}[Y]\end{bmatrix}$

Proposition: Let Y and X be random variables. $p = (1, X)'$.
$\beta^{ols} = \begin{pmatrix}\beta_0^{ols}\\ \beta_1^{ols}\end{pmatrix} = \mathbb{E}[p(x)p(x)']^{-1}\mathbb{E}[p(x)Y]$

Then: $\beta_1^{ols} = \frac{\text{cov}(Y,X)}{\text{var}(X)}$, $\beta_0^{ols} = \mathbb{E}[Y] - \beta_1^{ols}\mathbb{E}[X]$.
sufficient variation: $\text{var}(X) > 0$.

The residual zero correlation property.

Proposition: Define the residual $E \equiv Y - p(x)'\beta^{ols}$, then.

① $\mathbb{E}[p(x)E] = 0$.

If $p(X)$ contains a constant, then $\mathbb{E}(E)=0$, $cov(p_j(X), E) = 0$, $j = 1, 2, \cdots, J$

② $cov(Y, E) = Var(E)$.

proof: ① $\mathbb{E}[p(X)E] = \mathbb{E}[p(X)(Y - p(X)'\beta^{ols})]$

$= \mathbb{E}[p(X)Y] - \mathbb{E}[p(X)p(X)']\beta^{ols}$

$= \mathbb{E}[p(X)Y] - \mathbb{E}[p(X)p(X)']\mathbb{E}[p(X)p(X)']^{-1}\mathbb{E}[p(X)Y]$

$= 0$

② $cov(Y, E) = cov(p(X)'\beta^{ols} + E, E)$

$= \sum_{j=1}^{J} cov(p_j(X), E)\beta^{ols} + Cov(E, E)$

$= Var(E)$

Residual Regression: the FWL theorem.

Let $X_{-k} = (1, X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_K)'$

Let $\tilde{\beta}_k = \mathbb{E}[X_{-k}X_{-k}']^{-1}\mathbb{E}[X_{-k}X_k]$, $X_k^{\perp X_{-k}} = X_k - X_{-k}'\tilde{\beta}_k$

Theorem: FWL theorem: Let Y be r.v., X r.k-vector.

let $p(X) = (1, X)'$. Define $\beta^{ols} = \mathbb{E}[p(x)p(x)']^{-1}\mathbb{E}[p(x)Y]$.

Then: $\beta_k^{ols} = \frac{cov(Y, X_k^{\perp X_{-k}})}{Var(X_k^{\perp X_{-k}})}$

proof: define $E = Y - p(X)'\beta^{ols}$

Then $Y = p(X)\beta^{ols} + E$

$cov(Y, X_k^{\perp X_{-k}}) = cov(\beta_0^{ols} + \beta_1^{ols}X_1 + \cdots + \beta_K^{ols}X_K + E, X_k^{\perp X_{-k}})$

$= \beta_1^{ols}cov(X_1, X_k^{\perp X_{-k}}) + \beta_2^{ols}cov(X_2, X_k^{\perp X_{-k}}) + \cdots + cov(E, X_k^{\perp X_{-k}})$

$cov(E, X_k^{\perp X_{-k}}) = cov(E, X_k - \sum_{j \neq k} X_j\tilde{\beta}_{kj})$

$= cov(E, X_k) - \sum_{j \neq k}\tilde{\beta}_{kj}cov(E, X_j)$

$= 0$

For $j \neq k$: $cov(X_j, X_k^{\perp X_{-k}}) = 0$ follows ① of proposition

For $j = k$: $cov(X_k, X_k^{\perp X_{-k}}) = Var(X_k^{\perp X_{-k}})$ ② of proposition.

$\therefore \beta_k^{ols} = \frac{cov(Y, X_k^{\perp X_{-k}})}{Var(X_k^{\perp X_{-k}})}$

The long and short regressions.

Analyze why and when coefficients change when adding variables.

Short regression: Y on $(1, X)$: get $(\alpha^{short}, \beta^{short})$

Long regression: Y on $(1, X, W)$: get $(\alpha^{long}, \beta^{long}, \gamma^{long})$

Then: $\beta^{short} = \frac{cov(Y, X)}{Var(X)}$ $\beta^{long} = \frac{cov(Y, X^{\perp W})}{Var(X^{\perp W})}$, $\beta^{short} = \beta^{long} + \gamma^{long}\frac{cov(W, X)}{Var(X)}$

proof: define $E = Y - (\alpha^{long} + \beta^{long}X + \gamma^{long}W)$

Then: $cov(Y, X) = cov(\alpha^{long} + \beta^{long}X + \gamma^{long}W + E, X)$

$= \beta^{long}Var(X) + \gamma^{long}cov(W, X) + cov(E, X)$

$= \beta^{long}Var(X) + \gamma^{long}cov(W, X)$. follows ① of proposition.

$\therefore \frac{cov(Y, X)}{Var(X)} = \beta^{short} = \beta^{long} + \gamma^{long}\frac{cov(W, X)}{Var(X)}$.

Saturated regression:

$m(x) = \mathbb{E}[Y|X=x] = p(x)'\beta$

If we pick the right $p(x)'$, the linearity is guaranteed to hold when X is discrete.

Let $\mathcal{X} = \{0, 1\}$. $p(X) = (1, \mathbb{1}(X=1))'$

Define $\beta_0 = \mathbb{E}[Y|X=0]$, $\beta_1 = \mathbb{E}[Y|X=1] - \mathbb{E}[Y|X=0]$

Then: $m(x) = \mathbb{E}[Y|X=x] = p(x)'\beta = \mathbb{E}[Y|X=0] + \mathbb{1}[X=1](\mathbb{E}[Y|X=1] - \mathbb{E}[Y|X=0])$

$= \beta_0 + \beta_1 x$

Characterize OLS coefficients.

① Apply FWL theorem

② show that the conditional mean function is linear.

Regression Reinterpretation: interpreting ols estimand when the true conditional mean function is not linear.

As a best linear approximation to the true non-linear conditional mean function

As a best linear predictor of the outcome variable

As a weighted average derivative of the true non-linear conditional mean function.

OLS estimator

Ways to come up with estimator:

① Analogy principle: want to know $\theta = \theta(F_X)$, use the sample cdf to estimate $\hat{\theta} \equiv \theta(\hat{F}_X)$

$F_X(x) = \mathbb{E}[\mathbb{1}[X \le x]]$, $\hat{F}_X(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[X_i \le x]$. $\theta = \mathbb{E}[X]$, $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$

② MLE ③ GMM

Obtaining OLS as a sample analogue:

$$\beta \equiv \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$$

$$\hat{\beta}_{ols} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right)$$

For $\beta = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$:

① asymptotic variance matrix without hetero:
$V_1 = \sigma^2 \mathbb{E}[XX']^{-1}$

② with hetero: $V_2 = \mathbb{E}[XX']^{-1}\mathbb{E}[\varepsilon^2 XX']\mathbb{E}[XX']^{-1}$

Obtaining OLS via least squares:

$$\beta \equiv \underset{b \in \mathbb{R}^k}{\operatorname{argmin}}\ \mathbb{E}[(Y - X'b)^2]$$

$$\hat{\beta}_{ols} = \underset{b \in \mathbb{R}^k}{\operatorname{argmin}}\ \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'b)^2.$$

Observe: $\{(y_i, x_i)\}_{i=1}^{n}$

where $x_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ik} \end{bmatrix}_{k\times 1}$ $X = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1K} \\ & & \\ x_{n1} & \cdots & x_{nK} \end{bmatrix}_{n\times k}$ $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}$

for an $n\times 1$ matrix: $a'a = \sum_{i=1}^{n} a_i^2$

for a matrix X: $X'X = [x_1 \cdots x_n]\begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix} = \sum_{i=1}^{n} x_i x_i'$

For each i: $e_i = y_i - x_i'\beta$ $\quad y_i = x_i'\beta + e_i$

Stack up from $i=1$ to n: $y = X\beta + e$ $\quad e = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$

→ Therefore, equivalent as minimizing $\sum_{i=1}^{n}(y_i - x_i'\beta)^2 = (y - X\beta)'(y - X\beta) = y'y - b'X'y - y'Xb + b'X'Xb$

$\frac{d(a'b)}{db} = \frac{d(b'a)}{db} = a$

$\frac{d(b'Ab)}{db} = (A + A')b = 2Ab$ if A symmetric.

$$\therefore \frac{d\sum_{i=1}^{n}(y_i - x_i'\beta)^2}{d\beta} = -X'y - X'y + 2X'Xb = 0$$

$$\therefore \hat{\beta}_{ols} = (X'X)^{-1}X'y \quad (\hat{\beta})$$

$$\hat{\beta}_{ols} = (X'X)^{-1}(X'Y) = \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1}\left(\sum_{i=1}^{n} x_i y_i\right) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right)$$ same as analogue!

Finite sample residuals.

$-X'y + X'Xb = 0 \Rightarrow X'(y - Xb) = 0$

define finite sample residual for the i^{th} observation by: $\hat{e}_i \equiv y_i - x_i'\hat{\beta}$ (an estimation of population level residual $e_i = y_i - x_i'\beta$)

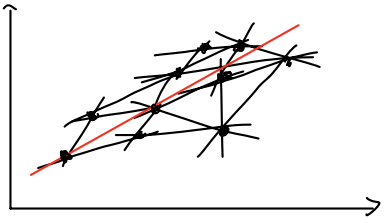
Stacking up: $\hat{e} \equiv y - X\hat{\beta}$

then: $X'\hat{e} = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix}'\hat{e} = (x_1 \cdots x_n)\hat{e} = 0$, a system of $(x_{1k} \cdots x_{nk})\hat{e} = \langle (x_{1k} \cdots x_{nk})', \hat{e} \rangle = 0$

$\therefore \hat{e} \perp X$, $\hat{e} \perp$ every column of X $((x_{1k} \cdots x_{nk})' \perp \hat{e})$

(population: $E \perp X$, $\mathbb{E}[X_k E] = 0 \Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} x_{ik}\hat{e}_i = 0$. when x_i contains a constant, $\frac{1}{n}\sum_{i=1}^{n}\hat{e} = 0 \Leftrightarrow \mathbb{E}[E] = 0$)

Geometric interpretation of OLS



Obtaining OLS as an MLE.

Assume $Y|X = x \sim N(x'\beta, \sigma^2)$, then $\hat{\beta}_{ols} = \hat{\beta}_{mle}$

OLS as a weighted average of pairwise slope.

OLS (finite sample)

Unbiasedness:

Proposition: Unbiased under linearity.

If $\mathbb{E}[Y|X=x] = x'\beta$, $\{(y_i, x_i)\}_{i=1}^n$ iid. from $(Y\ X)$

then $\mathbb{E}[\hat{\beta}_{ols}] = \beta$

proof: $\mathbb{E}[\hat{\beta}_{ols}|X] = \mathbb{E}[(X'X)^{-1}X'y|X]$

$= (X'X)^{-1}X'\mathbb{E}[y|X]$ ← $\mathbb{E}[y_i|X] = \mathbb{E}[y_i|x_1 \cdots x_n] = \mathbb{E}[y_i|x_i] = x_i'\beta$ stack up

$= (X'X)^{-1}X'X\beta$

$= \beta$

$\mathbb{E}[\hat{\beta}_{ols}] = \mathbb{E}[\mathbb{E}[\hat{\beta}_{ols}|X]] = \beta$

Matrix Analogue:

$\mathbb{E}[A \cdot y] = A \cdot \mathbb{E}[y]$

$var(Ay) = A\ var(y)\ A'$

$$var(y) = \begin{bmatrix} var(y_1) & cov(y_1,y_2) & \cdots & cov(y_1,y_n) \\ cov(y_1,y_2) & var(y_2) & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ cov(y_1,y_n) & cov(y_2,y_n) & \cdots & var(y_n) \end{bmatrix}$$

Variance of OLS:

Def: conditional variance function: $\sigma^2(x) = Var(Y|X=x)$

Proposition: Assume: ① linear $\mathbb{E}[Y|X=x] = x'\beta$ ② iid $\{(y_i, x_i)\}_{i=1}^n$

then: $var(\hat{\beta}_{ols}|X) = (X'X)^{-1}X'\Omega X(X'X)^{-1}$

$$\Omega = \begin{bmatrix} \sigma^2(x_1) & & 0 \\ & \ddots & \\ 0 & & \sigma^2(x_n) \end{bmatrix} = var(y|X)$$

If homoskedasticity, $var(\hat{\beta}_{ols}|X) = \sigma^2(X'X)^{-1}$ $(V = \sigma^2\mathbb{E}[xx']^{-1})$

For either case, $var(\hat{\beta}_{ols}) = \mathbb{E}[var(\hat{\beta}_{ols}|X)]$

The law of iterated variance: $var(Y) = var(\mathbb{E}[Y|X]) + \mathbb{E}[var(Y|X)]$

$\therefore var(\hat{\beta}_{ols}) = \mathbb{E}[var(\hat{\beta}_{ols}|X)] + var(\mathbb{E}(\hat{\beta}_{ols}|X))$

$= \mathbb{E}[var(\hat{\beta}_{ols}|X)]$ ← $\mathbb{E}(\hat{\beta}_{ols}|X)$ constant

proof: $var(\hat{\beta}|X) = var((X'X)^{-1}X'y|X)$

$= (X'X)^{-1}X'\ var(y|x)\ X(X'X)^{-1}$

for $var(y|X)$: its diagonal entries are:

$var(y_i|X) = var(y_i|x_1 \cdots x_n)$

$= var(y_i|x_i)$

$= \sigma^2(x_i)$

for off-diagonal entries:

$cov(y_i, y_j|X) = \mathbb{E}[y_iy_j|X] - \mathbb{E}[y_i|X]\mathbb{E}[y_j|X]$ (*)

$\mathbb{E}[y_iy_j|X] = \mathbb{E}[y_iy_j|x_1 \cdots x_n]$

$= \mathbb{E}[y_iy_j|x_i, x_j]$

$= \mathbb{E}[\mathbb{E}[y_iy_j|x_i, x_j, y_j]|x_i, x_j]$

$= \mathbb{E}[y_j\mathbb{E}[y_i|x_i, x_j, y_j]|x_i, x_j]$

$= \mathbb{E}[y_j\mathbb{E}[y_i|x_i]|x_i, x_j]$

$= \mathbb{E}[y_i|x_i]\mathbb{E}[y_j|x_i, x_j]$

$= \mathbb{E}[y_i|x_i]\mathbb{E}[y_j|x_j]$

$\therefore (*) = \mathbb{E}[y_i|x_i]\mathbb{E}[y_j|x_j] - \mathbb{E}[y_i|x_i]\mathbb{E}[y_j|x_j]$

$= 0$.

The variance in the scalar regressor case.

OLS of Y on X: design matrix is: $X = (\iota, x)$, ι is a column vector of ones.

$$X'X = \begin{bmatrix} \iota' \\ x' \end{bmatrix}[\iota\ x] = \begin{bmatrix} \iota'\iota & \iota'x \\ x'\iota & x'x \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}$$

$$\therefore (X'X)^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}\begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix}$$

$\sum_i (x_i - \frac{1}{n}\sum x_i)^2 = \sum_i \left(x_i^2 - 2x_i\frac{1}{n}\sum_i x_i + \left(\frac{1}{n}\sum_i x_i\right)^2\right)$

$= \sum_i x_i^2 - \frac{2}{n}\sum_i x_i \cdot \sum_i x_i + \frac{1}{n}(\sum_i x_i)^2$

$= \sum_i x_i^2 - \frac{1}{n}(\sum_i x_i)^2$

suppose homoskedasticity. then

$$var\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} | X = \frac{\sigma^2}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}\begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix}$$

$$var(\hat{\beta}_1|X) = \frac{n\sigma^2}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{\sigma^2}{\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i)^2}$$

$$= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{n\ \hat{var}(X)}$$

① $\sigma^2 \to 0$: $var(\hat{\beta}_{ols}|X) \to 0$.

② $n \to \infty$: $var(\hat{\beta}_{ols}|X) \to 0$.

③ $\hat{var}(x) \to \infty$: $var(\hat{\beta}_{ols}|X) \to 0$.

OLS is the best linear unbiased estimator of β (BLUE).

Estimating σ^2 to do inference on $\hat{\beta}$

$\sigma^2 = var(Y) = var(E) = \mathbb{E}[E^2] - (\mathbb{E}[E])^2 = \mathbb{E}[E^2]$ (homoskedasticity) $var(\hat{\beta}|X) = \sigma^2(X'X)^{-1}$

population: $E \equiv Y - X'\beta$

infeasible: $e_i = Y_i - x_i'\beta = y_i - x_i'\beta$ $\quad \hat{e} = \begin{bmatrix} y_1 - x_1'\hat{\beta} \\ \vdots \\ y_n - x_n\hat{\beta} \end{bmatrix}$

feasible: $\hat{e}_i = y_i - x_i'\hat{\beta}_{ols}$

$\therefore$ from sample analogue: $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \hat{e}_i^2 = \frac{1}{n}\hat{e}'\hat{e}$. (biased).

proposition. Assume: ① linearity $\mathbb{E}[Y|X=x]=x'\beta$ ② iid $\{(y_i,x_i)\}_{i=1}^n$ ③ $\sigma^2(x)=\mathrm{var}(Y|X=x)=\sigma^2>0$

Then: $\mathbb{E}[\hat\sigma^2|X]=\mathbb{E}[\frac{1}{n}\hat e'\hat e|X]=\frac{n-k}{n}\sigma^2$ $\quad\mathbb{E}[\hat\sigma^2]=\frac{n-k}{n}\sigma^2$ Biased.

proof: $M=I_n-P$ where $P=X(X'X)^{-1}X'$

$\hat e=y-X\hat\beta$
$=y-X(X'X)^{-1}X'y$
$=(I_n-X(X'X)^{-1}X')y$
$=My$
$=M(X\beta+e)$
$=Me$

$\hat e'\hat e=(Me)'(Me)$
$=e'M'Me$
$=e'MMe$ (symmetric M)
$=e'Me$ (idempotent M)
$=\mathrm{trace}(e'Me)$ ($e'Me$ is a scalar)
$=\mathrm{trace}(Mee')$

$\mathbb{E}[\hat e'\hat e|X]=\mathbb{E}[\mathrm{trace}(Mee')|X]$
$=\mathrm{trace}(\mathbb{E}[Mee'|X])$
$=\mathrm{trace}(M\mathbb{E}[ee'|X])$ M is constant given X
$=\mathrm{trace}(M\sigma^2I_n)$ homoskedasticity & iid
$=\sigma^2\,\mathrm{trace}(M)$
$=\sigma^2\,\mathrm{rank}(M)$ M idempotent

$\mathrm{rank}(M)=\mathrm{trace}(M)$
$=\mathrm{trace}(I_n-X(X'X)^{-1}X')$
$=\mathrm{trace}(I_n)-\mathrm{trace}(X(X'X)^{-1}X')$
$=n-\mathrm{trace}(X'X(X'X)^{-1})$
$=n-\mathrm{trace}(I_k)$
$=n-k$

$\therefore\ \mathbb{E}[\hat e'\hat e|X]=\sigma^2\cdot(n-k)$
$\therefore\ \mathbb{E}[\hat\sigma^2|X]=\frac{n-k}{n}\sigma^2=\mathbb{E}[\hat\sigma^2]$

Unbiased estimator of σ^2: $\boxed{s^2=\frac{1}{n-k}\sum_{i=1}^n\hat e_i^2=\frac{1}{n-k}\hat e'\hat e}$.

proposition: $\mathbb{E}[\hat V]=V$.

proof: $\mathbb{E}[\hat V]=\mathbb{E}[s^2(X'X)^{-1}]$
$=\mathbb{E}[\mathbb{E}[s^2(X'X)^{-1}|X]]$
$=\mathbb{E}[\mathbb{E}[s^2|X](X'X)^{-1}]$
$=\mathbb{E}[\sigma^2(X'X)^{-1}]$
$=V$

Then: $V=\mathrm{var}(\hat\beta_{ols}|X_1\cdots X_n)=\sigma^2(X'X)^{-1}$. non stochastic
$\hat V=\widehat{\mathrm{var}}(\hat\beta_{ols})=s^2(X'X)^{-1}=\widehat{\mathrm{var}}(\hat\beta_{ols}|X)$. stochastic.

Inference on $\hat\beta_{ols}$ (finite sample):

Assumption: $Y|X=x\sim N(x'\beta,\sigma^2)$ for all $x\in\mathrm{supp}(X)$
$\Rightarrow\mathbb{E}[Y|X=x]=x'\beta$, $\mathrm{var}(Y|X=x)=\sigma^2$, $E=Y-X'\beta$, $E\perp X$, $E\sim N(0,\sigma^2)$

Then: $y|X\sim N(X\beta,\sigma^2I_n)$
define $e_i=y_i-x_i'\beta$, then $e_i\sim N(0,\sigma^2)$, $e\sim N(0,\sigma^2I_n)$.

Theorem: $\hat\beta_{ols}|X\sim N(\beta,\sigma^2(X'X)^{-1})$

$\begin{pmatrix}\hat\beta_{ols}-\beta\\ \hat e\end{pmatrix}\sim N\left(0,\begin{pmatrix}\sigma^2(X'X)^{-1} & 0\\ 0 & \sigma^2M\end{pmatrix}\right)$. $\quad M=I_n-X(X'X)^{-1}X'$

Corollary: $\hat\beta_{ols}\perp\hat e\ |X$, $\hat\beta_{ols}\perp s^2|X$.

Corollary: $\hat\beta_k|X\sim N(\beta_k,\sigma^2[(X'X)^{-1}]_{kk})$

proof: conditional on X: $\begin{pmatrix}\hat\beta_{ols}-\beta\\ \hat e\end{pmatrix}=\begin{bmatrix}(X'X)^{-1}X'\\ M\end{bmatrix}e$

$\sim\begin{bmatrix}(X'X)^{-1}X'\\ M\end{bmatrix}N(0,\sigma^2I_n)$

remember that $A\cdot N(0,V)=N(0,AVA')$

$\therefore\ \mathrm{var}\left(\begin{bmatrix}\hat\beta_{ols}-\beta\\ \hat e\end{bmatrix}|X\right)=\begin{bmatrix}(X'X)^{-1}X'\\ M\end{bmatrix}\sigma^2I_n\begin{bmatrix}X(X'X)^{-1} & M'\end{bmatrix}$

$=\sigma^2\begin{bmatrix}(X'X)^{-1} & 0\\ 0 & M\end{bmatrix}$

Get CIs for β_k assuming σ^2 is known: marginal

$H_{null}:\beta_k=b_k$ $\quad H_{alt}:\beta_k\neq b_k$. ($b_k=0$ most common).

Define the t-statistic:

$$T_k(\beta_k)=\frac{\hat\beta_k-b_k}{\mathrm{stderr}(\hat\beta_k)}=\frac{\hat\beta_k-b_k}{\sqrt{\sigma^2[(X'X)^{-1}]_{kk}}}$$

proposition. under the null, $T_k(\beta_k)\sim N(0,1)$ unconditionally.

Conducting a level α hypothesis test:

① compute T_k
② compute $z_{\alpha/2}$ and $z_{1-\alpha/2}$ under $N(0,1)$. $\Leftrightarrow$
③ reject H_{null} if $T_k<z_{\alpha/2}$ or $T_k>z_{1-\alpha/2}$.

$N(0,1)$ symmetric.
$-z_{\alpha/2}=z_{1-\alpha/2}$
let $C(\alpha)=z_{1-\alpha/2}$. then:

$\phi_{t\text{-test}}=\mathbb{1}[|T_k|>C(\alpha)]$

Functions of standard normal distributions.

1. χ^2_n: let $z\sim N(0,\sigma^2I_n)$, then
$z'z=\sum_{i=1}^n z_i^2\sim\chi^2_n$

2. F-distribution: let $W_1\sim\chi^2_m$, $W_2\sim\chi^2_n$, $W_1\perp W_2$
$V=\frac{W_1/m}{W_2/n}\sim F(m,n)$

3. Student t-distribution: let $Z\sim N(0,1)$, $W\sim\chi^2_n$, $Z\perp W$
$U=Z/\sqrt{\frac{W}{n}}\sim t_n$ $\quad U^2\sim F(1,n)$

proposition: y: r.n-vector, $y\sim N(0,\Sigma)$, Σ symmetric positive semidefinite full rank. then $y'\Sigma^{-1}y\sim\chi^2_n$

lemma: A,B: r.vector. if $F_{A|B}(a|b)=F_A(a)$ for all $(a,b)\in\mathbb{R}^{\dim(A)}\times\mathbb{R}^{\dim(B)}$, $f_B(b)>0$, then $A\perp B$.

proposition: Exact size control.
Suppose σ^2 known. Let $\mathbb{P}_0$ be the dist. of $(Y, X_1, \cdots, X_k)$ such that H_{null} is true. then $\mathbb{P}_0(\phi_{t\text{-}test}=1)=\alpha$

proof: $\mathbb{P}_0(\phi_{t\text{-}test}=1)=\mathbb{P}_0(T_k<z_{\alpha/2} \text{ or } T_k>z_{1-\frac{\alpha}{2}})$
$=\mathbb{P}_0(T_k<z_{\alpha/2}) \text{ or } \mathbb{P}_0(T_k>z_{1-\alpha/2})$
$=\Phi(z_{\alpha/2})+1-\Phi(z_{1-\alpha/2})$
$=\frac{\alpha}{2}+(1-(1-\alpha/2))$
$=\alpha$.

Constructing a $100(1-\alpha)\%$ confidence interval for β_k by inverting the test.

$$T_k(b_k)=\frac{\hat{\beta}_k-b_k}{\text{stderr}(\hat{\beta}_k)}$$

$|T_k|<c(\alpha)$.

$$|T_k|<c(\alpha) \quad -c(\alpha)\le\frac{\hat{\beta}_k-b_k}{\text{stderr}(\hat{\beta}_k)}\le c(\alpha)$$
$$-c(\alpha)\le\frac{b_k-\hat{\beta}_k}{\text{stderr}(\hat{\beta}_k)}\le c(\alpha)$$
$$\hat{\beta}_k-c(\alpha)\,\text{stderr}(\hat{\beta}_k)\le b_k\le\hat{\beta}_k+c(\alpha)\,\text{stderr}(\hat{\beta}_k)$$
$$\hat{\beta}_k-z_{1-\alpha/2}\,\text{stderr}(\hat{\beta}_k)\le b_k\le\hat{\beta}_k-z_{\alpha/2}\,\text{stderr}(\hat{\beta}_k)$$
$(+)$ $(-)$

$$CI_n(1-\alpha)=\{b_k\in\mathbb{R}: z_{\alpha/2}\le T_k(b_k)\le z_{1-\alpha/2}\}$$
$$=[\hat{\beta}_k-z_{1-\alpha/2}\,\text{stderr}(\hat{\beta}_k),\ \hat{\beta}_k-z_{\alpha/2}\,\text{stderr}(\hat{\beta}_k)]$$
$$=[\hat{\beta}_k-c(\alpha)\,\text{stderr}(\hat{\beta}_k),\ \hat{\beta}_k+c(\alpha)\,\text{stderr}(\hat{\beta}_k)]$$
$$\mathbb{P}(CI_n(1-\alpha)\ni\beta_k)=1-\alpha$$

Get CIs for β assuming σ^2 is known. joint.
Simple idea: suppose $k=2\Rightarrow\beta_1,\beta_2$; then: $CI^{joint}=CI_1(1-\alpha)\times CI_2(1-\alpha)$ get a box.
suppose that $\{CI_1(1-\alpha)\ni\beta_1\}$ and $\{CI_2(1-\alpha)\ni\beta_2\}$ are independent. then: $\mathbb{P}(CI_1\times CI_2\ni(\beta_1,\beta_2))=\mathbb{P}(CI_1\ni\beta_1)\cdot\mathbb{P}(CI_2\ni\beta_2)$
$=(1-\alpha)^2$
$\neq(1-\alpha)$
Marginal → Joint WRONG!

The Wald statistic based confidence set.

$$\hat{\beta}-\beta|X\sim N(0,\sigma^2(X'X)^{-1})=\sigma(X'X)^{-\frac{1}{2}}N(0,I_n)$$
$$\therefore\ \sigma^{-1}(X'X)^{\frac{1}{2}}(\hat{\beta}-\beta)|X\sim N(0,I_n)$$
$$\therefore \text{ define } W(\beta)\equiv(\hat{\beta}-\beta)'\sigma^{-2}(X'X)(\hat{\beta}-\beta)\sim\chi^2_k$$
define $c(\alpha)$ to be $1-\alpha$ quantile of χ^2_k. $G(c(\alpha))=1-\alpha$, $G(\cdot)$ is the cdf of χ^2_k.
define $CI_n^W(1-\alpha)=\{b\in\mathbb{R}^k: W(b)\le c(\alpha)\}$
proposition: Exact coverage: $\mathbb{P}(CI_n^W(1-\alpha)\ni\beta)=1-\alpha$ shape: confidence ellipsoids.

proof: $\mathbb{P}(CI_n^W(1-\alpha)\ni\beta)=\mathbb{P}(W(\beta)\le c(\alpha))$
$=G(c(\alpha))$
$=1-\alpha$.

Inference with unknown σ^2

$H_{null}: \beta_k=b_k$ $H_{alt}: \beta_k\neq b_k$

$$T_k=\frac{\hat{\beta}_k-b_k}{\text{stderr}(\hat{\beta}_k)}=\frac{\hat{\beta}_k-b_k}{\sqrt{s^2[(X'X)^{-1}]_{kk}}}$$

Theorem: under the null $\beta_k=b_k$, $T_k\sim t_{n-k}$

$$W(\beta)\equiv(\hat{\beta}-\beta)'s^{-2}(X'X)(\hat{\beta}-\beta)$$
$$W(\beta)\sim\frac{\chi^2_k}{\chi^2_{n-k}/(n-k)}\qquad F(\beta)=\frac{W(\beta)}{k}\sim\frac{\chi^2_k/k}{\chi^2_{n-k}/n-k}$$

proposition. under the null $\beta=b$: $F(b)\sim F(K,n-k)$

WLS

$$\hat{\beta}_{wls}(W)=\underset{b\in\mathbb{R}^k}{\text{argmin}}\ \frac{1}{n}\sum_{i=1}^n w_i(y_i-x_i'b)^2\qquad w_i=w(x_i)$$
$$W=\begin{bmatrix}w_1 & & 0\\ & \ddots & \\ 0 & & w_n\end{bmatrix}$$
$$\hat{\beta}_{wls}(W)=\underset{b\in\mathbb{R}^k}{\text{argmin}}\,(y-Xb)'W(y-Xb)$$

proof: objective function: $(y-Xb)'W(y-Xb)=(y'-b'X')W(y-Xb)$
$=y'Wy-y'WXb-b'X'Wy-b'X'WXb$
take derivative wrt b: $-2X'Wy+2X'WX=0$
$\therefore\hat{\beta}_{wls}=(X'WX)^{-1}X'WY$

gives: $\hat{\beta}_{wls}(w) = (X'WX)^{-1}X'WY$

Theorem: $\hat{\beta}_{wls} \xrightarrow{p} \mathbb{E}[w(X)XX']^{-1}\mathbb{E}[XY] \equiv \tilde{\beta}$ It is a generalization of OLS estimand

Theorem: If $\mathbb{E}[Y|X=x] = x'\beta$, then $\tilde{\beta} = \beta$

Finite Sample Properties of WLS.

Proposition unbiased under linearity & variance

Suppose ① linearity: $\mathbb{E}[Y|X=x] = x'\beta$

② iid $\{(y_i, x_i')\}_{i=1}^n$ ③ $w_i = w(x_i)$.

then: $\mathbb{E}[\hat{\beta}_{wls}] = \beta$

$var(\hat{\beta}_{wls}|X) = (X'WX)^{-1}X'W\Omega WX(X'WX)^{-1}$

$\Omega = \begin{bmatrix} \sigma^2(x_1) & & 0 \\ & \ddots & \\ 0 & & \sigma^2(x_n) \end{bmatrix}$ where $\sigma^2(x) = var(Y|X=x)$.

If we pick $W = \Omega^{-1}$: then $var(\hat{\beta}_{wls}|X) = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\Omega\Omega^{-1}X(X'\Omega^{-1}X)^{-1}$

$= (X'\Omega^{-1}X)^{-1}$.

$W = \Omega^{-1} = \begin{bmatrix} 1/\sigma^2(x_1) & & 0 \\ & \ddots & \\ 0 & & 1/\sigma^2(x_n) \end{bmatrix}$

(infeasible) GLS: $\hat{\beta}_{GLS} = \hat{\beta}_{wls}(\Omega^{-1})$

★ $var(\hat{\beta}_{GLS}) \le var(\tilde{\beta})$ $\forall \tilde{\beta}$ that is linear and unbiased conditional on X.

$\hat{\beta}_{GLS} = \hat{\beta}_{wls}(w_i = \frac{1}{\sigma^2(x_i)}) = \underset{b \in \mathbb{R}^k}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^n \frac{1}{\sigma^2(x_i)}(y_i - x_i'b)^2$

feasible GLS: FGLS:

$\sigma^2(x) = var(\varepsilon|X=x) = \mathbb{E}[\varepsilon^2|X=x]$

Assume $\sigma^2(x) = \exp(\alpha'x)$

① get $\hat{e}_i = y_i - x_i'\hat{\beta}_{ols}$

② get $\hat{\alpha}$ from wls of $\hat{e}^2$ on x: $\hat{\alpha}_{wls} = \underset{a \in \mathbb{R}^k}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^n (\hat{e}_i^2 - \exp(a'x_i))^2$

③ compute $\hat{\sigma}^2(x_i) = \exp(\hat{\alpha}'x_i)$

Problem: ① requires of estimation of $\sigma^2(x)$, functional form not correct.

Asymptotic Properties of OLS Estimator

Theorem: Consistency.

① iid sample ② finite moments ③ sufficient variation.

then: $\hat{\beta}_{ols} \xrightarrow{p} \beta_{ols}$

proof: $\hat{\beta}_{ols} = (\frac{1}{n}\sum_{i=1}^n x_i x_i')^{-1}(\frac{1}{n}\sum_{i=1}^n x_i y_i)$ by CMT $\xrightarrow{p} \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$

($\frac{1}{n}\sum x_i x_i' \xrightarrow{p} \mathbb{E}[XX']$, $\frac{1}{n}\sum x_i y_i \xrightarrow{p} \mathbb{E}[XY]$)

proof: $\hat{\beta}_{ols} = (\frac{1}{n}\sum_{i=1}^n x_i x_i')^{-1}(\frac{1}{n}\sum_{i=1}^n x_i y_i)$

$= (\frac{1}{n}\sum_{i=1}^n x_i x_i')^{-1}(\frac{1}{n}\sum_{i=1}^n x_i(x_i'\beta_{ols} + e_i))$

$= \beta_{ols} + (\frac{1}{n}\sum_{i=1}^n x_i x_i')^{-1}(\frac{1}{n}\sum_{i=1}^n x_i e_i)$

$\therefore \hat{\beta}_{ols} - \beta_{ols} = (\frac{1}{n}\sum_{i=1}^n x_i x_i')^{-1}(\frac{1}{n}\sum_{i=1}^n x_i e_i)$

$\sqrt{n}(\hat{\beta}_{ols} - \beta_{ols}) = (\frac{1}{n}\sum_{i=1}^n x_i x_i')^{-1}[\sqrt{n}(\frac{1}{n}\sum_{i=1}^n x_i e_i)]$

$= (\frac{1}{n}\sum_{i=1}^n x_i x_i')^{-1}\underbrace{[\sqrt{n}(\frac{1}{n}\sum_{i=1}^n x_i e_i - \mathbb{E}[x_i e_i])]}_{①}$

① $\xrightarrow{d} N(0, var(xe)) = N(0, C)$

$\therefore \sqrt{n}(\hat{\beta}_{ols} - \beta_{ols}) \xrightarrow{d} D^{-1}N(0, C)$ by CMT.

$\sqrt{n}(\hat{\beta}_{ols} - \beta_{ols}) \xrightarrow{d} N(0, D^{-1}CD^{-1})$

Estimating the asymptotic covariance matrix:

$\sqrt{n}(\hat{\beta}_{ols} - \beta) \xrightarrow{d} N(0, D^{-1}CD^{-1})$

$V \equiv D^{-1}CD^{-1} = \mathbb{E}[XX']^{-1}\mathbb{E}[\varepsilon^2 XX']\mathbb{E}[XX']^{-1}$

$\hat{V} \equiv (\frac{1}{n}\sum_{i=1}^n x_i x_i')^{-1}(\frac{1}{n}\sum_{i=1}^n \hat{e}_i^2 x_i x_i')(\frac{1}{n}\sum_{i=1}^n x_i x_i')^{-1}$

↑ heteroskedasticity robust covariance matrix (std. err) estimator.

Why this result is important?

Remember. $\operatorname{var}(\hat{\beta}_{OLS}|X) = (X'X)^{-1}X'\Omega X(X'X)^{-1}$ $\quad \Omega = \begin{bmatrix} \sigma^2(x_1) & & 0 \\ & \ddots & \\ 0 & & \sigma^2(x_n) \end{bmatrix}$

$$n\operatorname{var}(\hat{\beta}_{OLS}|X) = n(X'X)^{-1}X'\Omega X(X'X)^{-1}$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2 x_i x_i'\right)\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}$$