Lecture Notes on Mathematical Statistics

Shuyang Ling

December 23, 2020

Contents

This lecture note draft is prepared for MATH-SHU 234 Mathematical Statistics I am teaching at NYU Shanghai. It covers the basics of mathematical statistics at undergraduate level.

Chapter 1

Probability

1.1 Probability

Probability theory is the mathematical foundation of statistics. We will review the basics of concepts in probability before we proceed to discuss mathematical statistics.

The core idea of probability theory is studying the randomness. The randomness is described by random variable X, a function from sample space to a number. Each random variable X is associated with a distribution function.

We define the cumulative distribution function (cdf) of X as:

$$F_X(x) = \mathbb{P}(X \leq x). \tag{1.1.1}$$

The cdf satisfies three properties:

- $F_X(x)$ is non-decreasing
- $F_X(x)$ is right-continuous
- Limits at the infinity:

$$\lim_{x \to -\infty} F_X(x) = 0, \quad \lim_{x \to \infty} F_X(x) = 1.$$

A cdf uniquely determines a random variable; it can be used to compute the probability of X belonging to a certain range

$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a).$$

In many applications, we often encounter two important classes of random variables, discrete and continuous random variables.

We say X is a discrete random variable if X takes value from a countable set of numbers

$$\mathcal{X} = \{a_1, a_2, \cdots, a_n, \cdots\}.$$

The probability of X taking value a_i is given by

$$f_X(i) = p_i = \mathbb{P}(X = a_i)$$

and we require

$$p_i \geq 0, \quad \sum_{i=1}^{\infty} p_i = 1.$$

This discrete function $f_X(i)$ is called probability mass function (pmf). It is natural to see that the connection between cdf and pmf

$$F_X(x) = \sum_{i=1}^{\infty} p_i \cdot 1\{a_i \leq x\}$$

where $1\{a_i \leq x\}$ is called indicator function:

$$1\{a_i \leq x\} = \begin{cases} 1, & \text{if } a_i \leq x, \\ 0, & \text{otherwise.} \end{cases}$$

It is easily seen that the cdf of a discrete random variable is not continuous. In fact, it is piecewise continuous.

The expectation (mean) of a random variable is given by

$$\mathbb{E}\, X = \sum_{i=1}^{\infty} a_i p_i$$

given that

$$\sum_{i=1}^{\infty} |a_i| p_i < \infty.$$

More generally, suppose we have a function $\varphi(x) : \mathcal{X} \to \mathbb{R}$, then the expectation of $\varphi(X)$, a new random variable, is

$$\mathbb{E}\, \varphi(X) = \sum_{i=1}^{\infty} \varphi(a_i) \cdot p_i.$$

We say X is a continuous random variable if there exists a function $f_X(x)$ such that

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, \mathrm{d}t.$$

The function $f_X(x)$ is called the probability density function (pdf). To get pdf from cdf, we simply take the derivative of $F_X(x)$,

$$f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x).$$

- Continuous random variables takes uncountably many values.
- The probability of $X = a$, i.e., $\mathbb{P}(X = a) = 0$ since F_X is continuous.

What does pdf mean?

$$\lim_{\epsilon \to 0^+} \frac{\mathbb{P}(x - \epsilon \leq X \leq x + \epsilon)}{2\epsilon} = \lim_{\epsilon \to 0^+} \frac{F_X(x + \epsilon) - F_X(x - \epsilon)}{2\epsilon} = f_X(x).$$

The mean of X is defined as

$$\mathbb{E}\, X = \int_{\mathbb{R}} x f_X(x)\, \mathrm{d}x.$$

For a function $\varphi(x) : \mathcal{X} \to \mathbb{R}$, we have

$$\mathbb{E}\,\varphi(X) = \int_{\mathbb{R}} \varphi(x) f_X(x)\,\mathrm{d}x.$$

The variance, as a measure of uncertainty, is

$$\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}\,X)^2 = \mathbb{E}\,\varphi(X), \quad \varphi(x) = (x - \mathbb{E}\,X)^2.$$

We sometimes use another form

$$\mathrm{Var}(X) = \mathbb{E}\,X^2 - (\mathbb{E}\,X)^2.$$

Here $\mathbb{E}\,X^2$ is referred as the second moment. The p-th moment is defined as

$$\mathbb{E}\,X^p = \int_{\mathbb{R}} x^p\,\mathrm{d}F_X = \begin{cases} \sum_{i=1}^n p_i a_i^p, & \text{discrete} \\ \int_{\mathbb{R}} x^p f_X(x)\,\mathrm{d}x, & \text{continuous} \end{cases}$$

Independence: Independence is an important concept in probability. Two random variables X and Y are independent if

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B), \quad \forall A, B.$$

This is equivalent to

$$\mathbb{P}(X \le x, Y \le y) = \mathbb{P}(X \le x)\mathbb{P}(Y \le y) = F_X(x)F_Y(y), \quad \forall x, y,$$

i.e., the joint cdf of (X, Y) equals the product of its marginal distributions.

Suppose X and Y are independent, then $f(X)$ and $g(Y)$ are also independent for two functions f and g. As a result, we have

$$\mathbb{E}\,f(X)g(Y) = \mathbb{E}\,f(X)\,\mathbb{E}\,g(Y).$$

Given a sequence of n random variables $\{X_i\}_{i=1}^n$, they are independent if

$$\mathbb{P}(X_i \le x_i, 1 \le i \le n) = \prod_{i=1}^n \mathbb{P}(X_i \le x_i).$$

If X_i is discrete or continuous, then the independence can be characterized by using pmf and pdf:

$$f_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = \prod_{i=1}^n f_{X_i}(x_i).$$

The joint pdf/pmf is the product of individual pdf/pmf's (marginal distribution).

In probability and statistics, we often study the sum of i.i.d. (independent identically distributed) random variables $\sum_{i=1}^n X_i$.

Exercise: Denote $Z_n = \sum_{i=1}^n X_i$ as the sum of n i.i.d. random variables. Then $\mathbb{E}\,Z_n = n\mu$ and $\mathrm{Var}(Z_n) = n\sigma^2$.

We will see more in the next few sections.

1.2 Important distributions

1.2.1 Uniform distribution

If the pdf of a random variable X satisfies

$$f_X(x) = \frac{1}{b-a},\ a \le x \le b.$$

This is called the uniform distribution, denoted by $\mathrm{Unif}[a,b]$. Its cdf is

$$F_X(x) = \begin{cases} 0, & x \le a \\ \frac{x-a}{b-a}, & a < x < b \\ 1, & x \ge b. \end{cases}$$

Exercise: Show that $\mathbb{E}\, X = (a+b)/2$ and $\mathrm{Var}(X) = (b-a)^2/12$.

1.2.2 Normal distribution/Gaussian distribution

Normal distribution is the most important distribution in probability and statistics. It has extremely rich structures and connections with other distributions. A random variable X is Gaussian with mean μ and variable σ^2, denoted by $\mathcal{N}(\mu, \sigma^2)$, if its pdf is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad x \in \mathbb{R}.$$

In particular, if $\mu = 0$ and $\sigma = 1$, we say X is standard Gaussian. One can verify

$$\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-x^2/2}\, \mathrm{d}x = 1$$

by using the trick from multivariate calculus. Let's verify $\mathbb{E}\, X = 0$ and $\mathrm{Var}(X) = 1$.

$$\mathbb{E}\, X = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} x e^{-x^2/2}\, \mathrm{d}x = 0$$

since $xe^{-x^2/2}$ is an odd function. How about $\mathbb{E}\, X^2$?

$$\begin{aligned} \mathbb{E}\, X^2 &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} x^2 e^{-x^2/2}\, \mathrm{d}x \\ &= -\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} x\, \mathrm{d}e^{-x^2/2} \\ &= -\frac{1}{\sqrt{2\pi}} x e^{-x^2/2}\Big|_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-x^2/2}\, \mathrm{d}x = 1. \end{aligned}$$

Gaussian random variable is linearly invariant: suppose $X \sim \mathcal{N}(\mu, \sigma^2)$, then $aX + b$ is still Gaussian with mean $a\mu + b$ and variance $a^2\sigma^2$, i.e., $\mathcal{N}(a\mu + b, a^2\sigma^2)$

$$\mathbb{E}(aX + b) = a\mu + b, \quad \mathrm{Var}(aX + b) = \mathrm{Var}(aX) = a^2\, \mathrm{Var}(X) = a^2\sigma^2.$$

Moreover, suppose $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ are two independent random variables, then

$$X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

This can be extended to the sum of n independent Gaussian random variables. For example,

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(0, n)$$

if $X_i \sim \mathcal{N}(0,1)$ are i.i.d. random variables.

1.2.3 Moment-generating function (MGF)

Why does $\sum_{i=1}^n X_i \sim \mathcal{N}(0, n)$ hold? The moment generating function is

$$M(t) := \mathbb{E}\, e^{tX}.$$

Moment generating function is named after the following fact:

$$\mathbb{E}\, e^{tX} = 1 + t\,\mathbb{E}\, X + \frac{t^2\,\mathbb{E}\, X^2}{2!} + \cdots + \frac{t^n\,\mathbb{E}\, X^n}{n!} + \cdots$$

The coefficients of t^n corresponds to the nth moment of X. MGF does not always exist since it requires all the moments exist for a given random variable. We can compute moment of any order by differentiating mgf w.r.t. t and evaluating it at $t = 0$:

$$\mathbb{E}\, X^n = \frac{\mathrm{d}^n M(t)}{\mathrm{d}t^n}\Big|_{t=0}.$$

Suppose two random variables have the same moment generating functions, they are of the same distribution. MGF uniquely determines the distribution. For $X \sim \mathcal{N}(\mu, \sigma^2)$, it holds that

$$\begin{aligned} M(t) = \mathbb{E}\, e^{tX} &= e^{\mu t}\,\mathbb{E}\, e^{t(X-\mu)} \\ &= \frac{e^{\mu t}}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} e^{tz - \frac{z^2}{2\sigma^2}}\,\mathrm{d}z \\ &= \frac{e^{\mu t}}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} e^{-(z/\sigma - \sigma t)^2/2} e^{\sigma^2 t^2/2}\,\mathrm{d}z \\ &= \exp(\mu t + \sigma^2 t^2/2). \end{aligned}$$

Now for a sequence of n i.i.d. Gaussian random variables,

$$\mathbb{E}\, e^{t\sum_{i=1}^n X_i} = \prod_{i=1}^n \mathbb{E}\, e^{tX_i} = \prod_{i=1}^n \exp(\mu t + \sigma^2 t^2/2) = \exp(n\mu t + n\sigma^2 t^2/2)$$

This expression equals the moment generating function of $\mathcal{N}(n\mu, n\sigma^2)$.

Exercise: Show that $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, $1 \le i \le n$ is a sequence of n independent random variables, then

$$\sum_{i=1}^n X_i \sim \mathcal{N}\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right).$$

Exercise: Suppose we are able to generate uniform random samples, can we generate normal random variables?

1.2.4 Chi-squared distribution

In statistics, chi-squared distribution is frequently used in hypothesis testing. We say $X \sim \chi_n^2$, i.e., chi-squared distribution of degree n, if

$$f_X(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, \quad x > 0$$

where

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\,\mathrm{d}x.$$

In particular, if n is a positive integer, $\Gamma(n) = (n-1)!$ and $\Gamma(1/2) = \sqrt{\pi}$.

Chi-squared distribution is closely connected to normal distribution. Suppose $Z \sim \mathcal{N}(0,1)$. Now we take a look at $X = Z^2$:

$$\begin{aligned}\mathbb{P}(X \le x) &= \mathbb{P}(Z^2 \le x) = \mathbb{P}(-\sqrt{x} \le Z \le \sqrt{x}) \\ &= 2\mathbb{P}(0 \le Z \le \sqrt{x}) \\ &= \sqrt{\frac{2}{\pi}} \int_0^{\sqrt{x}} e^{-z^2/2}\,\mathrm{d}z.\end{aligned}$$

The pdf of X is obtained by differentiating the cdf,

$$f_X(x) = \sqrt{\frac{2}{\pi}} \cdot \frac{1}{2\sqrt{x}} \cdot e^{-x/2} = \frac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2}, \quad x > 0.$$

Now if $\{Z_i\}_{i=1}^n$ is a sequence of n independent standard normal random variables, then

$$X = \sum_{i=1}^n Z_i^2 \sim \chi_n^2.$$

Chi-squared distribution is a special family of Gamma distribution $\Gamma(\alpha, \beta)$.

$$f_X(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad x > 0.$$

If $\beta = 2$ and $\alpha = n/2$, then $\Gamma(n/2, 2) = \chi_n^2$.

Exercise: Show that $\mathbb{E}\, e^{tX} = (1-\beta t)^{-\alpha}$, for $t < 1/\beta$.

Exercise: Show that $\sum_{i=1}^n X_i \sim \Gamma(\sum_{i=1}^n \alpha_i, \beta)$ if $X_i \sim \Gamma(\alpha_i, \beta)$ are independent.

1.2.5 Exponential distribution

Exponential distribution: X has an exponential distribution with parameter β, i.e., $\mathcal{E}(\beta)$ if

$$f(x) = \beta^{-1} e^{-x/\beta}, \quad x \ge 0$$

where $\beta > 0$.

- The exponential distribution is used to model the waiting time of a certain event (lifetimes of electronic components).
- The waiting time of a bus arriving at the station.

It is also a special case of Gamma distribution $\Gamma(1, \beta)$. Exponential distribution satisfies the so-called memoryless property:

$$\mathbb{P}(X \ge t + s | X \ge t) = \mathbb{P}(X \ge s), \quad \forall s \ge 0.$$

Recall that the left side involves conditional probability. For two events A and B, the conditional probability of A given B is

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Here

$$\mathbb{P}(X \geq t+s | X \geq t) = \frac{\mathbb{P}(X \geq t+s, X \geq t)}{\mathbb{P}(X \geq t)} = \frac{\mathbb{P}(X \geq t+s)}{\mathbb{P}(X \geq t)}$$

since $\{X \geq t+s\}$ is contained in $\{X \geq t\}$

Exercise: Verify the memoryless properties and think about what does it mean?

Exercise: What is the distribution of $\sum_{i=1}^{n} X_i$ if $X_i \sim \mathcal{E}(\beta)$?

Exercise: Verify $\mathbb{E}\, X = \beta$ and $\mathrm{Var}(X) = \beta^2$ for $X \sim \mathcal{E}(\beta)$.

1.2.6 Bernoulli distributions

Let X represent the outcome of a binary coin flip. Then its pmf is

$$\mathbb{P}(X = 1) = p, \qquad \mathbb{P}(X = 0) = 1 - p.$$

Sometimes, we also write the pmf in this way:

$$f_X(x) = p^x (1-p)^{1-x}, \quad x \in \{0, 1\}.$$

The coin is fair if $p = 1/2$. The cdf is

$$F_X(x) = \begin{cases} 0, & x < 0, \\ 1-p, & 0 \leq x < 1, \\ 1, & x \geq 1. \end{cases}$$

In this case, we denote $X \sim$Bernoulli(p). The mean and variance of X is simple to obtain:

$$\mathbb{E}(X) = 1 \cdot \mathbb{P}(X = 1) + 0 \cdot \mathbb{P}(X = 0) = p$$

and

$$\mathrm{Var}(X) = \mathbb{E}\, X^2 - (\mathbb{E}\, X)^2 = \mathbb{E}\, X - p^2 = p(1-p).$$

1.2.7 Binomial distribution

Suppose we have a coin which falls heads up with probability p. Flip the coin n times and X is the number of heads. Each outcome is supposed to be independent.

If $X = k$, then there must be k heads and $n - k$ tails:

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Then its pmf is

$$f_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \{0, 1, \cdots, n\}.$$

In this case, we denote $X \sim$Binomial(n, p).

Exercise: Show $\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1$.

Binomial distribution is closely related to Bernoulli distribution. Suppose $\{X_i\}_{i=1}^n$ are n i.i.d. Bernoulli(p) random variables, then $X = \sum_{i=1}^n X_i \sim$Binomial$(n, p)$. In particular, if $X \sim$Binomial(n, p), $Y \sim$Binomial(m, p), and X is independent of Y, then $X + Y \sim$Binomial$(n + m, p)$.

Exercise: Use the idea of moment generating function to show that $\sum_{i=1}^n X_i \sim$ Binomial(n, p) if $X_i \sim$Bernoulli(p).

Exercise: What is the mean and variance of Binomial(n, p)?

Exercise: Use the mgf to obtain the mean and variance of Binomial(n, p).

1.2.8 Poisson distribution

Suppose we want to model the number of times an event occurs in an interval of time or space, we would want to use Poisson distribution.

- The number of volcanic eruptions in Hawaii in one year
- The number of customers arriving at McDonald's in the morning

Poisson distribution: A random variable satisfies Poisson distribution with parameter λ, i.e., $X \sim$Poisson(λ) if

$$f_X(k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k \in \mathbb{Z}^+$$

where $\lambda > 0$ is also referred to the intensity, the expected (average) number of occurrences.

Exercise: Verify $e^{-\lambda}\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = 1$.

Exercise: $\mathbb{E}(X) = \text{Var}(X) = \lambda$.

Poisson distribution can be derived from Binomial distribution. Suppose λ is the expected number of occurrences within a period of time. We divide the time interval into n equal sub-intervals. Each of them has an expected number of occurrences λ/n. Let's consider n as a large number. Now we assume the probability of one occurrence in an interval is λ/n, which is a Bernoulli random variable with parameter λ/n. Then the total number of occurrences is

$$\sum_{i=1}^{n} X_i = \text{Binomial}(n, \lambda/n)$$

where $X_i \sim$Bernoulli(λ/n). In fact, we can show that $\sum_{i=1}^n X_i$ converges to Poisson(λ) as $n \to \infty$.

Exercise: Show that for $X_i \sim$Bernoulli(λ/n),

$$\lim_{n\to\infty} \mathbb{P}\left(\sum_{i=1}^{n} X_i = k\right) = \frac{e^{-\lambda}\lambda^k}{k!}$$

for any given k.

Exercise: Suppose we are able to generate uniform random samples, can we generate Poisson random variables?

1.3 Limiting theorem

1.3.1 Law of large number

Law of large numbers, along with central limit theorem (CLT), plays fundamental roles in statistical inference and hypothesis testing.

Theorem 1.3.1 (Weak law of large number)**.** *Let X_i be a sequence of i.i.d. random variables,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu,$$

i.e., convergence in probability, where $\mu = \mathbb{E}\, X_i$.

We say a sequence of random variables X_n converge to X in probability if for any given $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0.$$

Law of large number basically says that the sample average converges to the expected value as the sample size grows to infinity.

We can prove the law of large number easily if assuming X_i has a finite second moment. The proof relies on Chebyshev's inequality.

Theorem 1.3.2 (Chebyshev's inequality)**.** *For a random variable X with finite second moment, then*

$$\mathbb{P}(|X - \mu| \geq \epsilon) \leq \frac{\mathbb{E}\,|X-\mu|^2}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2}.$$

Proof of WLLN. Consider $\overline{X}_n = n^{-1}\sum_{i=1}^{n} X_i$. We aim to prove $\overline{X}_n$ converges to μ in probability if $\mathrm{Var}(X_i) < \infty$. For $\epsilon > 0$, it holds that

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) \leq \frac{1}{\epsilon^2}\,\mathbb{E}\,|\overline{X}_n - \mu|^2.$$

It suffices to compute the variance of $\overline{X}_n$:

$$\begin{aligned}
\mathbb{E}\,|\overline{X}_n - \mu|^2 &= \mathbb{E}\left| n^{-1}\sum_{i=1}^{n} X_i - \mu \right|^2 \\
&= n^{-2}\,\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \mu)\right]^2 \\
&= n^{-2}\cdot \sum_{i=1}^{n}\mathbb{E}(X_i-\mu)^2 \\
&= n^{-2}\cdot n\sigma^2 = \frac{\sigma^2}{n}.
\end{aligned}$$

As a result,

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \to 0, \quad n\to\infty.$$

□

1.3.2 Central limit theorem

Theorem 1.3.3 (Central limit theorem)**.** *Let $X_i, 1 \le i \le n$ be a sequence of i.i.d. random variables with mean μ and finite variance σ^2, then*

$$Z_n := \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{d} \mathcal{N}(0,1),$$

i.e., convergence in distribution.

Sometimes, we also use

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}.$$

We say a sequence of random variable Z_n converges to Z in distribution if

$$\lim_{n\to\infty} \mathbb{P}(Z_n \le z) = \mathbb{P}(Z \le z)$$

for any $z \in \mathbb{R}$. In other words, the cdf of Z_n converges to that of Z pointwisely,

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x).$$

What does it mean?

$$\lim_{n\to\infty} \mathbb{P}(a \le Z_n \le b) = \mathbb{P}(a \le Z \le b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2}\,\mathrm{d}t.$$

In statistics, one useful choice of a and b are

$$a = z_{\alpha/2}, \quad b = z_{1-\alpha/2}.$$

Here z_α is defined as the α-quantile of normal random variable, i.e.,

$$\mathbb{P}(Z \le z_\alpha) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_\alpha} e^{-t^2/2}\,\mathrm{d}t = \alpha, \quad 0 \le \alpha \le 1$$

and by symmetry, we have

$$z_\alpha = -z_{1-\alpha}.$$

In particular,

$$z_{0.975} \approx 1.96.$$

In other words, as n is sufficiently large, with probability approximately $1-\alpha$, it holds that

$$z_{\alpha/2} \le \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \le z_{1-\alpha/2} \iff |\overline{X}_n - \mu| \le \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}}, \quad z_{\alpha/2} = -z_{1-\alpha/2}$$

which implies that the "error" decays at the rate of $1/\sqrt{n}$.

Theorem 1.3.4. *Convergence in probability implies convergence in distribution.*

Exercise: Show that $\mathbb{E}\, Z_n = 0$ and $\mathrm{Var}(Z_n) = 1$.

The proof of CLT relies on the moment generating function. We can show that the MGF of Z_n converges to that of a standard normal. Here we provide a sketch of the proof.

Proof: The mgf of Z_n is

$$\mathbb{E}\exp(tZ_n) = \mathbb{E}\exp\left(\frac{t}{\sqrt{n}\sigma}\sum_{i=1}^{n}(X_i-\mu)\right) = \prod_{i=1}^{n}\mathbb{E}\exp\left(\frac{t}{\sqrt{n}\sigma}(X_i-\mu)\right)$$

where the second equality uses independence of X_i.

For each i, we expand the mgf using Taylor approximation,

$$\begin{aligned}\mathbb{E}\exp\left(\frac{t}{\sqrt{n}\sigma}(X_i-\mu)\right) &= 1 + \frac{t}{\sqrt{n}\sigma}\,\mathbb{E}(X_i-\mu) + \frac{t^2}{2n\sigma^2}\cdot\mathbb{E}(X_i-\mu)^2 + o(n^{-1}) \\ &= 1 + \frac{t^2}{2n} + o(n^{-1})\end{aligned}$$

As a result,

$$\begin{aligned}\lim_{n\to\infty}\mathbb{E}\exp(tZ_n) &= \lim_{n\to\infty}\left(\mathbb{E}\exp\left(\frac{t}{\sqrt{n}\sigma}(X_i-\mu)\right)\right)^n \\ &= \lim_{n\to\infty}\left(1+\frac{t^2}{2n}+o(n^{-1})\right)^n \\ &= \exp(t^2/2).\end{aligned}$$

This is exactly the mgf of standard normal distribution. □

Exercise: (Wasserman 5.14) What is the limiting distribution of $\overline{X}_n^2$ if $X_i \overset{i.i.d.}{\sim} \text{Unif}[0,1]$?

Exercise: Draw m samples from Bernoulli distribution (or exponential, Gamma, Poisson, etc). Compute the normalized sum:

$$Y_m = \sqrt{n}(\overline{X}_n - \mu)/\sigma.$$

Repeat M times and collect the data. Plot the histogram or the empirical cdf. Here the empirical cdf is defined as

$$F_Y(y) = \frac{1}{M}\sum_{j=1}^{M} 1\{Y_i \le y\}.$$

Does it look like that of standard normal distribution?

Chapter 2

Introduction to statistics

2.1 Population

One core task of statistics is making inferences about an unknown parameter θ associated to a *population*. What is a population? In statistics, a population is a set consisting of the entire similar items we are interested in. For example, a population may refer to all the college students in Shanghai or all the residents in Shanghai. The choice of the population depends on the actual scientific problem.

Suppose we want to know the average height of all the college students in Shanghai or want to know the age distribution of residents in Shanghai. What should we do? Usually, a population is too large to deal with directly. Instead, we often draw samples from the population and then use the samples to estimate a population parameter θ such as mean, variance, median, or even the actual distribution.

This leads to several important questions in statistics:

- How to design a proper sampling procedure to collect data? Statistical/Experimental design.
- How to use the data to estimate a particular population parameter?
- How to evaluate the quality of an estimator?

2.1.1 Important statistics

If we get a dataset, we usually compute the basic statistics to roughly describe the dataset.

- Sample mean/average:

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Variance:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2$$

- Standard deviation:
$$S_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}$$
- Median:
$$\text{median}(x_i) = x_{[(n+1)/2]}$$
where $[(n+1)/2]$ means the closest integer to $(n+1)/2$. More generally, the α quantile is $x_{[\alpha(n+1)]}$.
- Range, max/min:
$$\text{Range} = \max_{1\le i\le n} x_i - \min_{1\le i\le n} x_i.$$
- Empirical cdf:
$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1\{x_i \le x\}.$$

Exercise: Show that $\overline{x}_n$ minimizes
$$f(z) = \sum_{i=1}^{n}(x_i - z)^2$$

Exercise: Show that median(x_i) minimizes
$$f(z) = \sum_{i=1}^{n}|x_i - z|$$
Is the global minimizer unique?

Notice that all the quantities above are based on samples $\{x_1, \cdots, x_n\}$. These quantities are called statistics.

Definition 2.1.1. *A statistic is a deterministic function of samples,*
$$y = T(x_1, \cdots, x_n),$$
which is used to estimate the value of a population parameter θ.

Question: How to evaluate the quality of these estimators?

2.1.2 Probabilistic assumption

We assume the population has a probability distribution F_X and each observed sample x_i is a realization of a random variable X_i obeying the population distribution F_X. A set of samples $\{x_1, \cdots, x_n\}$ are treated as one realization of a random sequence $\{X_1, \cdots, X_n\}$. From now on, we assume that all the random variables X_i are *i.i.d., independent identically distributed.*

In other words, $T(x_1, \ldots, x_n)$ is one copy of a random variable
$$\widehat{\theta}_n = T(X_1, \cdots, X_n),$$
which is a point estimator of θ. We ask several questions:

- Does $\widehat{\theta}_n$ well approximate the population parameter θ?
- How to evaluate the quality of the estimators?

2.2 Evaluation of estimators

There are several ways to evaluate the quality of point estimators.

Definition 2.2.1. *The bias of $\widehat{\theta}_n$ is*

$$bias(\widehat{\theta}_n) = \mathbb{E}\,\widehat{\theta}_n - \theta.$$

We say $\widehat{\theta}_n$ is an unbiased estimator of θ if the bias is zero.

Exercise: Show that $\overline{X}_n$ and S_n^2 are unbiased estimators of μ and σ^2 respectively.

For S_n^2, we know that

$$\begin{aligned} S_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu + \mu - \overline{X}_n)^2 \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}(X_i-\mu)^2 - n(\overline{X}_n - \mu)^2\right) \end{aligned}$$

Now by taking the expectation of S_n^2, we have

$$\mathbb{E}(X_i-\mu)^2 = \sigma^2, \quad \mathbb{E}(\overline{X}_n - \mu)^2 = \sigma^2/n.$$

As a result, it holds

$$\mathbb{E}\,S_n^2 = \frac{1}{n-1}\left(n\sigma^2 - \sigma^2\right) = \sigma^2.$$

However, bias is not usually a good measure of a statistic. It is likely that two unbiased estimators of the same parameter are of different variance. For example, both $T_1 = X_1$ and $T_2 = \overline{X}_n$ are both unbiased estimators of θ. However, the latter is definitely preferred as it uses all the samples: the sample mean is consistent.

Definition 2.2.2. *We say $\widehat{\theta}_n$ is a consistent estimator of θ if $\widehat{\theta}_n$ converges to θ in probability, i.e.,*

$$\lim_{n\to\infty} \mathbb{P}(|\widehat{\theta}_n - \theta| \geq \epsilon) = 0$$

for any fixed $\epsilon > 0$. The probability is taken w.r.t. the joint distribution of $(X_1, \cdots, X_n)$.

The consistency of sample mean is guaranteed by the law of large number. How about sample variance S_n^2?

Let's take a closer look at S_n^2:

$$S_n^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}_n^2\right).$$

We are interested in the limit of S_n^2 as $n \to \infty$. First note that by law of large number, we have

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} \mathbb{E}\,X^2, \quad \overline{X}_n \xrightarrow{p} \mathbb{E}\,X.$$

Recall that variance of X equals $\mathbb{E}\,X^2 - (\mathbb{E}\,X)^2$. As a result, we can finish the proof if $\overline{X}_n^2 \to \mu^2$. Does it hold?

Theorem 2.2.1 (Continuous mapping theorem)**.** *Suppose g is a continuous function and $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$. This also applies to convergence in distribution.*

Remark: this is also true for random vectors. Suppose $\boldsymbol{X}_n = (X_{n1}, \cdots, X_{nd}) \in \mathbb{R}^d$ is a random vector and $\boldsymbol{X}_i \xrightarrow{p} \boldsymbol{X}$, i.e., for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(\|\boldsymbol{X}_n - \boldsymbol{X}\| \geq \epsilon) = 0$$

where $\|\boldsymbol{X}_n - \boldsymbol{X}\|$ denotes the Euclidean distance between $\boldsymbol{X}_n$ and X, then $g(\boldsymbol{X}_n) \xrightarrow{p} g(\boldsymbol{X})$ for a continuous function g.

This justifies $\overline{X}_n^2 \xrightarrow{p} \mu^2$. Now, we have

$$\lim_{n\to\infty} \frac{n}{n-1} \cdot \left(n^{-1} \sum_{i=1}^{n} X_i^2 - \overline{X}_n^2 \right) = \mathbb{E}\, X^2 - (\mathbb{E}\, X)^2 = \sigma^2$$

convergence in probability.

Exercise: Complete the proof to show that S_n^2 and S_n are consistent estimators of σ^2 and σ.

Another commonly-used quantity to evaluate the quality of estimator is MSE (mean-squared-error).

Definition 2.2.3 (MSE: mean-squared-error)**.** *The mean squared error is defined as*

$$MSE(\widehat{\theta}_n) = \mathbb{E}(\widehat{\theta}_n - \theta)^2$$

where the expectation is taken w.r.t. the joint distribution of $(X_1, \cdots, X_n)$.

Recall that the pdf/pmf for $(X_1, \cdots, X_n)$ is

$$f_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

and the population parameter is associated to the actual distribution $f_X(x)$.

Note that by Chebyshev's inequality, convergence in MSE implies convergence in probability:

$$\mathbb{P}(|\widehat{\theta}_n - \theta| \geq \epsilon) \leq \frac{\mathbb{E}(\widehat{\theta}_n - \theta)^2}{\epsilon^2}.$$

The MSE is closely related to bias and variance of $\widehat{\theta}_n$. In fact, we have the following famous bias-variance decomposition

$$\mathrm{MSE}(\widehat{\theta}_n) = \mathrm{bias}(\widehat{\theta}_n)^2 + \mathrm{Var}(\widehat{\theta}_n).$$

Proof: The proof is quite straightforward:

$$\begin{aligned}
\mathrm{MSE}(\widehat{\theta}_n) &= \mathbb{E}(\widehat{\theta}_n - \theta)^2 \\
&= \mathbb{E}(\widehat{\theta}_n - \mu + \mu - \theta)^2 \\
&= \mathbb{E}(\widehat{\theta}_n - \mu)^2 + 2\,\mathbb{E}(\widehat{\theta}_n - \mu)(\mu - \theta) + (\mu - \theta)^2 \\
&= \underbrace{\mathbb{E}(\widehat{\theta}_n - \mu)^2}_{\mathrm{Var}(\widehat{\theta}_n)} + \underbrace{(\mu - \theta)^2}_{\mathrm{bias}(\widehat{\theta}_n)^2}
\end{aligned}$$

where $\mu = \mathbb{E}\,\widehat{\theta}_n$ and the second term equals 0. □

Lemma 2.2.2. *Convergence in MSE implies convergence in probability.*

The proof of this lemma directly follows from Chebyshev's inequality.

2.3 Confidence interval

All the aforementioned measures of estimators such as sample mean, variance, etc, are called point estimators. Can we provide an interval estimator for an unknown parameter? In other words, we are interested in finding a range of plausible values which contain an unknown parameter with reasonably large probability. This leads to the construction of confidence interval.

What is a confidence interval of θ?

Definition 2.3.1. *A $1-\alpha$ confidence interval for a parameter θ is an interval $C_n = (a, b)$ where $a = a(X_1, \cdots, X_n)$ and $b = b(X_1, \cdots, X_n)$ are two statistics of the data such that*

$$\mathbb{P}(\theta \in C_n) \geq 1 - \alpha,$$

i.e., the interval (a, b) contains θ with probability $1 - \alpha$.

Note that we cannot say the probability of θ falling inside (a, b) is $1 - \alpha$ since C_n is random while θ is a fixed value.

Question: How to construct a confidence interval?

Let's take a look at a simple yet important example. In fact, CLT is very useful in constructing a confidence interval for the mean.

We have shown that sample mean $\overline{X}_n$ is a consistent estimator of the population mean μ. By CLT, we have

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \to \mathcal{N}(0, 1)$$

where σ is the standard deviation.

For a sufficiently large n, CLT implies that

$$\mathbb{P}\left(|\overline{X}_n - \mu| \leq \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}\right) \approx 1 - \alpha$$

where z_α is the α-quantile of standard normal distribution.

Note that

$$|\overline{X}_n - \mu| \leq \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}} \iff \overline{X}_n - \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}_n + \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}$$

In other words, if σ is *known*, the random interval

$$\left(\overline{X}_n - \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}, \overline{X}_n + \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}\right)$$

covers μ with probability approximately $1 - \alpha$.

A few remarks:

- Suppose σ is known, then the confidence interval (CI) becomes small as the sample size n increase! Smaller interval is preferred since it means less uncertainty.

- As α decreases ($1-\alpha$ increases), $z_{1-\alpha/2}$ increases, making CI larger.
- Suppose we have the samples, the CI is

$$\left(\overline{x}_n - \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}, \overline{x}_n + \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}\right).$$

  The meaning of CI is: suppose we repeat this experiments many times, the frequency of this interval containing μ is close to $1-\alpha$.

Now, we focus on another question: what if σ is unknown? A simple remedy is: use sample standard deviation S_n to replace σ and we have the following potential candidate of CI:

$$\left(\overline{X}_n - \frac{z_{1-\alpha/2}S_n}{\sqrt{n}}, \overline{X}_n + \frac{z_{1-\alpha/2}S_n}{\sqrt{n}}\right)$$

Does it work? Does this interval cover μ with probability approximately $1-\alpha$. This question is equivalent to ask if

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \xrightarrow{d} \mathcal{N}(0,1)$$

as $n \to \infty$.

This is indeed true, which is guaranteed by Slutsky's theorem.

Theorem 2.3.1 (Slutsky's theorem)**.** *Suppose*

$$X_n \xrightarrow{d} X, \quad Y_n \xrightarrow{p} c$$

where c is a constant, then

$$X_n Y_n \xrightarrow{d} cX, \quad X_n + Y_n \xrightarrow{d} X + c.$$

Exercise: Construct a counterexample. Suppose $X_n \xrightarrow{d} X, \quad Y_n \xrightarrow{d} Y$, then $X_n + Y_n$ does not necessarily converge to $X + Y$ in distribution.

Exercise: Show that if $X_n \xrightarrow{d} c$, then $X_n \xrightarrow{p} c$ where c is a fixed value.

Exercise: Challenging! Prove Slutsky's theorem.

Note that $S_n \xrightarrow{p} \sigma$ (Why?). By Slutsky's theorem, we have

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} = \underbrace{\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}}_{\xrightarrow{d}\mathcal{N}(0,1)} \cdot \underbrace{\frac{\sigma}{S_n}}_{\xrightarrow{p}\sigma} \xrightarrow{d} \mathcal{N}(0,1).$$

The argument above justifies why $\left(\overline{X}_n - \frac{z_{1-\alpha/2}S_n}{\sqrt{n}}, \overline{X}_n + \frac{z_{1-\alpha/2}S_n}{\sqrt{n}}\right)$ is a $1-\alpha$ confidence interval of μ.

Chapter 3

Nonparametric inference

3.1 Cumulative distribution function

3.1.1 Estimation of CDF

Suppose we are interested in estimating the age distribution of residents in Shanghai with n samples. How to perform this statistical estimation? Naturally, it suffices to come up with an estimator of the cumulative distribution function $F_X(x) = \mathbb{P}(X \leq x)$. Once we have a good estimator of the cdf, we can use it to estimate other population parameters.

One common approach is to compute the empirical cdf. Let $X_1, \cdots, X_n$ be i.i.d. samples from F_X. The empirical cdf $F_{n,X}$ of F_X is

$$F_{n,X}(x) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \leq x\}$$

where

$$1\{X_i \leq x\} = \begin{cases} 1, & X_i \leq x, \\ 0, & X_i > x. \end{cases}$$

Why $F_{n,X}$ is a proper estimator of F_X for any x?

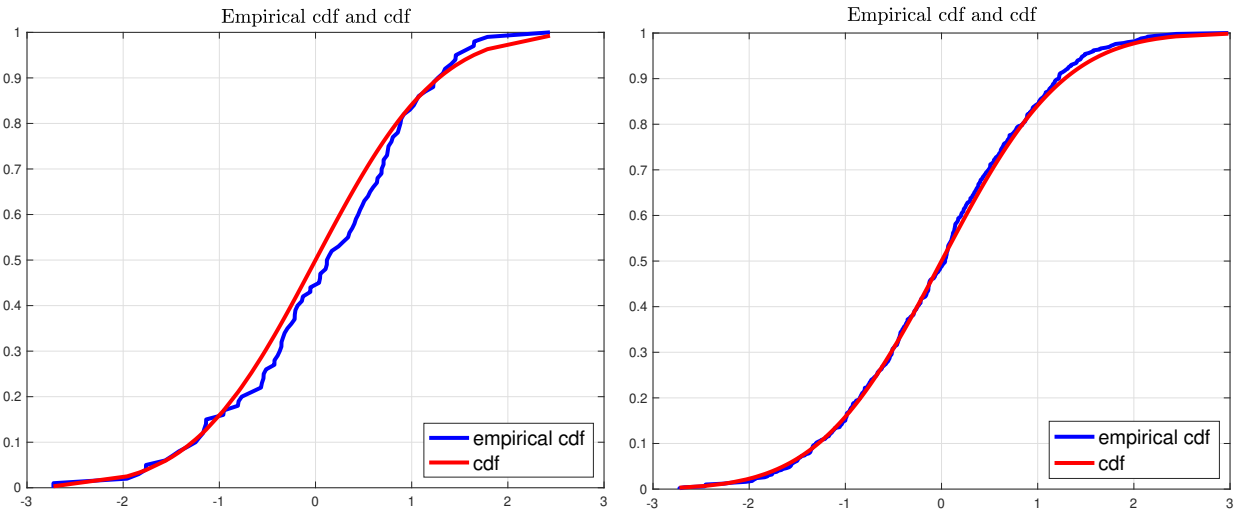


Figure 3.1: CDF v.s. empirical CDF for $n = 100$ and $n = 500$

Figure 3.1, the empirical cdf gets closer to cdf as the number of data increases. In fact, $F_{n,X}(x)$ is a consistent estimator of $F_X(x)$. Why?

Proof: Note that $\{X_i\}_{i=1}^n$ are i.i.d. random variables. Therefore, $1\{X_i \le x\}$ are also i.i.d. random variables.

$$F_{n,X}(x) = \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \le x\}$$

is the sample mean of $1\{X_i \le x\}$. By law of large number, it holds that

$$F_{n,X}(x) \xrightarrow{p} \mathbb{E}\, 1\{X_i \le x\}.$$

Since we have

$$\mathbb{E}\, 1\{X_i \le x\} = \mathbb{P}(X_i \le x) = F_X(x),$$

$F_{n,X}(x)$ is a consistent estimator of $F_X(x)$. □

Exercise: Construct a $1-\alpha$ confidence interval for $F_X(x)$. (See homework).

The consistency of $F_{n,X}(x)$ is pointwise for any fixed x. As we have seen the Figure 3.1, the convergence of $F_{n,X}(x)$ to $F_X(x)$ seems uniform for any x. This is indeed true and leads to the famous Glivenko-Cantelli theorem.

Theorem 3.1.1 (Glivenko-Cantelli theorem)**.** *For* $X_1, \cdots, X_n \overset{i.i.d.}{\sim} F_X$*, then*

$$\sup_x |\widehat{F}_{n,X}(x) - F_X(x)| \xrightarrow{p} 0$$

as $n \to \infty$.

In other words, the empirical cdf is a consistent estimator of the actual cdf.

3.1.2 Plug-in principle

One useful application regarding the consistency of empirical cdf is the plug-in principle, which gives us a simple way to construct statistical estimators. Suppose we aim to estimate a population parameter θ:

$$\theta = \mathbb{E}\, r(X) = \int_{\mathbb{R}} r(x)\, \mathrm{d}F_X(x).$$

Examples of $r(\cdot)$ include

- Population mean: $r(x) = x$
- Higher moment: $r(x) = x^p$

What would be a natural choice of estimator for θ? Simply replace $F_X(x)$ by $F_{n,X}(x)$!

$$\widehat{\theta}_n = \int_{\mathbb{R}} r(x)\, \mathrm{d}F_{n,X}(x) = \frac{1}{n}\sum_{i=1}^{n} r(X_i).$$

Here $F_{n,X}(x)$ is not continuous everywhere since it jumps at X_i. One way to understand this integral is to treat $F_{n,X}(x)$ as the cdf of a discrete random variable which takes value X_i with probability $1/n$. By law of large number, $\widehat{\theta}_n$ is a consistent estimator of θ if the population parameter θ exists.

The plug-in principle can be also extended to more complicated scenarios.

- α-quantile:
$$\theta = F_X^{-1}(\alpha) := \inf\{x : F_X(x) \geq \alpha\}.$$
The plug-in estimator for α-quantile is
$$F_{n,X}^{-1}(\alpha) := \inf\{x : F_{n,X}(x) \geq \alpha\}.$$
The empirical quantile is a consistent estimator of the population quantile. Why?

- Variance:
$$\theta = \mathbb{E}\, X^2 - (\mathbb{E}\, X)^2 = \int_{\mathbb{R}} x^2 \,\mathrm{d}F_X(x) - \left(\int_{\mathbb{R}} x \,\mathrm{d}F_X(x)\right)^2$$
Then the plug-in estimator for θ is
$$\begin{aligned}\widehat{\theta}_n &= \int_{\mathbb{R}} x^2 \,\mathrm{d}F_{n,X}(x) - \left(\int_{\mathbb{R}} x \,\mathrm{d}F_{n,X}(x)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2.\end{aligned}$$

- Covariance: if we have 2D samples, i.e., random vectors,
$$\sigma_{XY} = \mathrm{Cov}(X, Y) = \mathbb{E}\, XY - \mathbb{E}\, X\, \mathbb{E}\, Y.$$
What would be the plug-in estimator of σ_{XY}? The only difference here is: we need to use the 2D empirical cumulative distribution function.
$$F_n(x, y) = \frac{1}{n}\sum_{i=1}^n 1\{X_i \leq x, Y_i \leq y\}$$
which is a natural estimator of the 2D cdf:
$$F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y).$$
Note that
$$\sigma_{XY} = \int_{\mathbb{R}^2} xy \,\mathrm{d}F(x, y) - \int_{\mathbb{R}} x \,\mathrm{d}F(x, y) \int_{\mathbb{R}} y \,\mathrm{d}F(x, y).$$
and thus
$$\begin{aligned}\widehat{\sigma}_{XY,n} &= \int_{\mathbb{R}^2} xy \,\mathrm{d}F_n(x, y) - \int_{\mathbb{R}} x \,\mathrm{d}F_n(x, y) \int_{\mathbb{R}} y \,\mathrm{d}F_n(x, y) \\ &= \frac{1}{n}\sum_{i=1}^n X_i Y_i - \overline{X}_n \overline{Y}_n.\end{aligned}$$

Exercise: Show that
$$\frac{1}{n}\sum_{i=1}^n X_i Y_i - \overline{X}_n \overline{Y}_n = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n).$$

Exercise: Is $\widehat{\sigma}_{XY,n}$ an unbiased estimator of σ_{XY}? Why or why not?

Exercise: Show that $F_n(x, y)$ is an unbiased/consistent estimator of $F_{X,Y}(x, y)$.

Exercise: Show that
$$\lim_{y \to \infty} F_{n,XY}(x, y) = F_{n,X}(x),$$
i.e., sending y to ∞ gives the marginal empirical distribution of X.

Exercise: Compute the mean and variance of $F_n(x, y)$. What is the MSE(F_n)?

3.2 Bootstrap

We have discussed how to evaluate the quality of an estimator $\widehat{\theta}_n = T(X_1, \cdots, X_n)$ and construct a confidence interval for a population parameter. From the previous discussion, we may realize that the core problem in understanding $\widehat{\theta}_n$ is deriving its distribution. Once we know its distribution and connection to the population parameter θ, we can easily evaluate the its quality and construct confidence interval.

However, it is usually not easy to characterize the distribution of an estimator. Suppose we observe a set of samples $X_1, \cdots, X_n \sim F_X$. We want to estimate its mean, variance, and median, as well as a $1-\alpha$ confidence interval for these population parameters. What should we do? Our current toolbox is able to provide an interval estimation for μ; however, it is less clear for variance and median.

Let's start with providing estimators for μ, σ^2, and median.

$$\begin{aligned} T(F_n) &= \overline{X}_n, \\ T(F_n) &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2, \\ T(F_n) &= \text{median}(X_1, \cdots, X_n). \end{aligned}$$

All these estimators are consistent, i.e., as $n \to \infty$

$$T(F_n) \xrightarrow{p} T(F) = \theta.$$

Suppose we know the actual distribution of $\widehat{\theta}_n = T(F_n)$, it will be much easier to find a confidence interval for θ. Why? First we can find the $\alpha/2$ and $1-\alpha/2$ quantiles, denoted by $a = q_{\alpha/2}$ and $b = q_{1-\alpha/2}$, for $\widehat{\theta}_n - \theta$, then

$$\mathbb{P}(a \le \widehat{\theta}_n - \theta \le b) \ge 1 - \alpha.$$

Then a $1-\alpha$ confidence interval of θ is

$$\left(\widehat{\theta}_n - q_{1-\alpha/2},\ \widehat{\theta}_n - q_{\alpha/2}\right)$$

However, we don't know either the distribution of $\widehat{\theta}_n$ or θ. What is the solution?

Bootstrap was invented by Bradley Efron in 1979. It is widely used in various applications due to its simplicity and effectiveness. The idea of bootstrap is quite simple: we use the empirical distribution to approximate the actual population distribution plus resampling. Recall that

$$\theta = T(F), \quad \widehat{\theta}_n = T(F_n).$$

How to find the distribution of $\widehat{\theta}_n$? Assume we have access to a random number generator of F_X. Then we can sample $X_{1,k}, \cdots, X_{n,k}$, and compute

$$\widehat{\theta}_{n,k} = T(X_{1,k}, \cdots, X_{n,k}), \quad 1 \le k \le B.$$

Suppose we repeat this procedure many times (say B times), the distribution of $\widehat{\theta}_n$ will be well approximated by $\{\widehat{\theta}_{n,k}\}_{k=1}^{B}$.

However, we still haven't resolved the issue that F_X is unknown. The solution is: we replace F_X by $F_{n,X}$, i.e., instead of sampling data from F_X, we sample from $F_{n,X}$ which is known. Then use the obtained data to approximate the distribution of $\widehat{\theta}_n$. For the unknown parameter θ, we simply approximate it by $\widehat{\theta}_n$.

Essentially, the idea is summarized as follows:

The Annals of Statistics
1979, Vol. 7, No. 1, 1–26

THE 1977 RIETZ LECTURE

BOOTSTRAP METHODS: ANOTHER LOOK AT THE JACKKNIFE

BY B. EFRON

Stanford University

We discuss the following problem: given a random sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ from an unknown probability distribution F, estimate the sampling distribution of some prespecified random variable $R(\mathbf{X}, F)$, on the basis of the observed data x. (Standard jackknife theory gives an approximate mean and variance in the case $R(\mathbf{X}, F) = \theta(\hat{F}) - \theta(F)$, θ some parameter of interest.) A general method, called the "bootstrap," is introduced, and shown to work satisfactorily on a variety of estimation problems. The jackknife is shown to be a linear approximation method for the bootstrap. The exposition proceeds by a series of examples: variance of the sample median, error rates in a linear discriminant analysis, ratio estimation, estimating regression parameters, etc.

- If we know θ and F_X, to estimate the distribution of $\widehat{\theta}_n = T(F_n)$, we sample $(X_1, \cdots, X_n)$ from F_X and calculate $\widehat{\theta} = T(F_n)$. Repeat it many times and the obtained samples approximate the distribution of $\widehat{\theta}_n$.
- In reality, we only have one sample $x_1, \cdots, x_n$ without knowing the underlying F_X. Thus we approximate F_X by F_n from $x_1, \cdots, x_n$; use F_n to generate new data points $X_1^*, \cdots, X_n^*$ from $F_{n,X}$; and get $\widehat{\theta}_n^* = T(F_n^*)$ where F_n^* is the empirical cdf of $X_1^*, \cdots, X_n^*$. Use the simulated data $\widehat{\theta}_n^*$ to approximate the distribution of $\widehat{\theta}_n$.

Here we actually have a two-stage approximation:

- The approximation error due to $F_{n,X}$ and F_X may not be small.
- The approximation error due to the resampling is small if B is large; where B is the number of copies of $\widehat{\theta}_n^*$.

Now we are ready to present the bootstrap method for the construction of confidence interval.

- Step 1: Given the empirical cdf $F_{n,X}$ (one realization):

$$F_{n,X} = \frac{1}{n} \sum_{i=1}^{n} 1\{x_i \le x\}.$$

- Step 2: Generate n samples $X_{1,k}^*, \cdots, X_{n,k}^*$ from $F_{n,X}$ and compute

$$\widehat{\theta}_{n,k}^* = T(X_{1,k}^*, \cdots, X_{n,k}^*) = T(F_{n,X,k}^*), \quad 1 \le k \le B,$$

where $\{X_{i,k}^*\}_{i=1}^n$ are n independent samples from $F_{n,X}(x)$ and they form the empirical cdf $F_{n,X,k}^*$. Note that generating samples from $F_{n,X}$ is equivalent to uniformly picking data from $\{x_1, \cdots, x_n\}$ with replacement, i.e., it is the cdf of the following random variable Z:

$$\mathbb{P}(Z = x_i) = \frac{1}{n}$$

- Step 3a (Basic bootstrap): Compute $R_k = \widehat{\theta}^*_{n,k} - \widehat{\theta}_n$, and find the $\alpha/2$ and $1-\alpha/2$ empirical quantiles of $\{R_k\}_{k=1}^B$, i.e., $R(\alpha) = \widehat{\theta}^*_{n,k}(\alpha) - \widehat{\theta}_n$. A $(1-\alpha)$-confidence interval is given by

$$\widehat{\theta}^*_{n,k}\left(\frac{\alpha}{2}\right) - \widehat{\theta}_n \le \widehat{\theta}_n - \theta \le \widehat{\theta}^*_{n,k}\left(1-\frac{\alpha}{2}\right) - \widehat{\theta}_n$$

  which equals

$$2\widehat{\theta}_n - \widehat{\theta}^*_{n,k}\left(1-\frac{\alpha}{2}\right) \le \theta \le 2\widehat{\theta}_n - \widehat{\theta}^*_{n,k}\left(\frac{\alpha}{2}\right)$$

  In other words, we use the empirical quantile $R(\alpha) = \widehat{\theta}^*_{n,k}(\alpha) - \widehat{\theta}_n$ as an α-quantile estimator of the random variable $\widehat{\theta}_n - \theta$.

- Step 3b (Percentile intervals): Another way is to use the empirical quantile of $\{\widehat{\theta}^*_{n,k}\}_{k=1}^B$,

$$\left(\widehat{\theta}^*_{n,k}\left(\frac{\alpha}{2}\right), \widehat{\theta}^*_{n,k}\left(1-\frac{\alpha}{2}\right)\right)$$

  as a $1-\alpha$ confidence interval for θ.

- Estimation of standard deviation:

$$\widehat{\sigma}^2_* = \frac{1}{B-1}\sum_{k=1}^B \left(\widehat{\theta}^*_{n,k} - \frac{1}{B}\sum_{j=1}^n \widehat{\theta}^*_{n,j}\right)^2$$

Exercise: The empirical cdf $F_{n,X} = n^{-1}\sum_{i=1}^n 1\{x_i \le x\}$ defines a discrete random variable Z with pmf:

$$\mathbb{P}(Z = x_i) = \frac{1}{n}.$$

Exercise: Show that drawing i.i.d. samples $X_1^*, \cdots, X_n^*$ from $F_{n,X}$ is equivalent to draw n samples from $\{x_1, \cdots, x_n\}$ uniformly with replacement.

Exercise: Suppose $x_1, \cdots, x_n$ are observed i.i.d. data from F_X. Assume that they are distinct, i.e., $x_i \neq x_j$ for all $i \neq j$. Compute the mean and variance for the random variable X^* with cdf $F_{n,X} = n^{-1}\sum_{i=1}^n 1\{x_i \le x\}$.

Example: The number of traffic accidents in Berkeley, California, in 10 randomly chosen non-rainy days in 1998 is as follows:

$$4, 0, 6, 5, 2, 1, 2, 0, 4, 3$$

Can we find a 95% confidence interval for the mean?

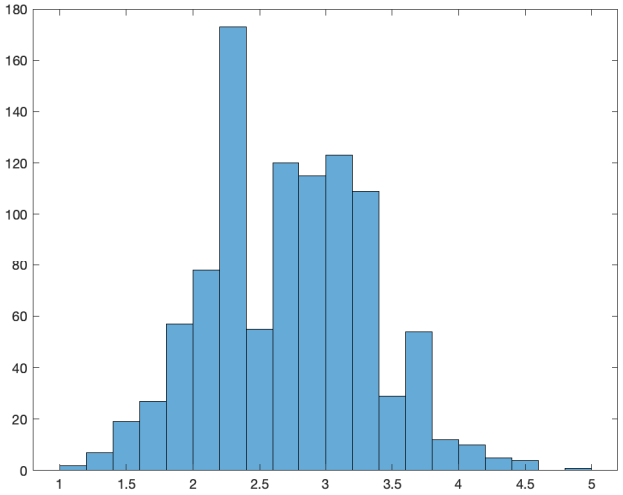
We know that

$$\overline{X}_n = 2.7, \quad S_n^2 = 4.2333.$$

We follow the bootstrap approach described above. Sampling n points from F_n

$$Y_k = T(X_1^*, \cdots, X_n^*)$$

by repeating B times. This is equivalent to picking n samples from $\{X_1, \cdots, X_n\}$ with replacement. Here is the histogram of $\{Y_j^*\}_{j=1}^B$ where $B = 1000$:



- The 95% CI via central limit theorem is $(1.4247, 3.9753)$:

$$\left(\overline{X}_n - 1.96\frac{S_n}{\sqrt{n}}, \overline{X}_n + 1.96\frac{S_n}{\sqrt{n}}\right)$$

  where $n = 10$.

- The 95% CI via Bootstrap methods is $(1.5, 3.9)$

Example: (Wasserman Example 8.6). Here is an example first used by Bradley Efron to illustrate the bootstrap. The data consist of the LSAT scores and GPA, and one wants to study the correlation between LSAT score and GPA.

The population correlation is

$$\rho = \frac{\mathbb{E}(X - \mu_X)(Y - \mu_Y)}{\sqrt{\mathbb{E}(X - \mu_X)^2}\sqrt{\mathbb{E}(Y - \mu_Y)^2}}$$

By using plug-in principle, we have a consistent estimator of ρ by using the empirical correlation.

$$\widehat{\rho} = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sqrt{\sum_{i=1}^n (X_i - \overline{X}_n)^2}\sqrt{\sum_{i=1}^n (Y_i - \overline{Y}_n)^2}}$$

Exercise: Show that $\widehat{\rho}$ is the plug-in estimator of ρ.

It is equal to

$$\widehat{\rho} = \frac{S_{XY}}{S_X S_Y}.$$

where S_{XY} is the sample covariance, and S_X and S_Y are the sample standard deviations of X and Y respectively.

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n),$$

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2,$$

$$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y}_n)^2.$$

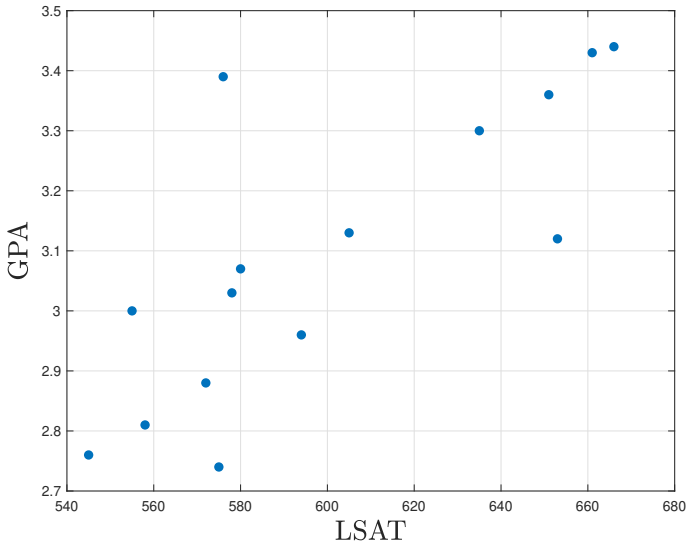


Figure 3.2: GPA v.s. LSAT

The scatterplot implies that higher LSAT score tends to give higher GPA. The empirical correlation is $\widehat{\rho} = 0.776$ which indicates high correlation between GPA and LSAT scores.

How to obtain a $1-\alpha$ confidence interval for ρ? We apply bootstrap method to obtain a confidence interval for ρ.

1. Independently sample (X_i^*, Y_i^*), $i = 1, 2, \cdots, n$, from $\{(X_i, Y_i)\}_{i=1}^n$ uniformly with replacement.
2. Let
$$\widehat{\rho}_k^* = \frac{\sum_{i=1}^n (X_i^* - \overline{X}_n^*)(Y_i - \overline{Y}_n^*)}{\sqrt{\sum_{i=1}^n (X_i^* - \overline{X}_n^*)^2}\sqrt{\sum_{i=1}^n (Y_i^* - \overline{Y}_n^*)^2}}$$
be an estimator of correlation of the resampled data. Repeat Step 1 and 2 $B = 1000$ times and obtain $\{\widehat{\rho}_k^*\}_{k=1}^B$.

- The estimated variance of the correlation is
$$\widehat{\sigma}_*^2 = \frac{1}{B-1}\sum_{k=1}^B (\widehat{\rho}_k - \overline{\widehat{\rho}}_B^*)^2$$
where
$$\overline{\widehat{\rho}}_B^* = \frac{1}{B}\sum_{k=1}^B \widehat{\rho}_k^*.$$
- The CI can be obtained via computing the empirical quantile of $\{\widehat{\rho}_k^*\}_{k=1}^B$.

Let $B = 1000$ and we have the histogram in Figure 3.3. A 95% confidence interval for the correlation $(0.4646, 0.9592)$.

Exercise: Show that drawing n i.i.d. samples from the 2D empirical cdf $F_n(x, y) = n^{-1}\sum_{i=1}^n 1\{X_i \le x, Y_i \le y\}$ is equivalent to sampling n points uniformly with replacement from $\{X_i, Y_i\}_{i=1}^n$.

In this note, we briefly introduce the bootstrap method and apply it to construct confidence interval. However, we didn't cover the theory for bootstrap. Under certain regularity condition, the CI from bootstrap covers the actual parameter θ with probability

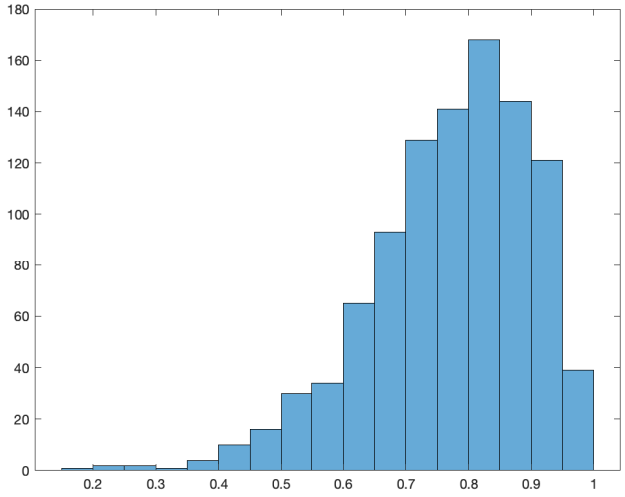


Figure 3.3: Histogram of $\{\widehat{\rho}_k^*\}_{k=1}^B$

approximately $1-\alpha$ as $n \to \infty$. For more details, you may refer to [3, Chapter 8] and [1, Chapter 6.5].

Chapter 4

Parametric inference

We have discussed the estimation of some common population parameters such as mean, variance, and median. You may have realized that we did not impose any assumption on the underlying population distribution: the analysis hold for quite general distributions. However, in many applications, we have more information about the underlying distribution, e.g., the population distribution may belong to a family of distributions $\mathcal{S} = \{f_\theta(x)|\theta \in \Theta\}$ where Θ is the parameter space and $f_\theta(x)$, also denoted by $f(x;\theta)$, is the pdf or pmf. This is referred to as the parametric models. The population distribution is uniquely determined by the hidden parameter θ. Of course, the validity of such an assumption remains verified: one may need to check if the population is indeed consistent with the assumed distribution. This is a separated important question which we will deal with later.

Here are several examples of parametric models.

Normal data model:

$$\mathcal{S} = \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} : \sigma > 0, \mu \in \mathbb{R} \right\}.$$

Poisson data model:

$$\mathcal{S} = \left\{ f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x \in \mathbb{Z}^+ \right\}.$$

Gamma distribution:

$$\mathcal{S} = \left\{ \frac{1}{\beta^\alpha \Gamma(a)} x^{\alpha-1} e^{-x/\beta},\ x > 0 : \theta \in \Theta \right\}, \quad \theta = (\alpha, \beta), \quad \Theta = \{(\alpha,\beta) : \alpha > 0, \beta > 0\}.$$

We are interested in several questions:

- Given i.i.d. samples $X_1, \cdots, X_n$ from $f(x;\theta)$, how to estimate θ?
- How to evaluate the quality of $\widehat{\theta}$?
- Construct a $1-\alpha$ confidence interval for θ?

4.1 Method of moments (M.O.M)

Method of moments is a convenient way to construct point estimators with wide applications in statistics and econometrics. Given a set of i.i.d. samples $X_1, \cdots, X_n$ from $f(x;\theta)$, how to find a suitable θ such that the data $X_1, \cdots, X_n$ well fit $f(x;\theta)$? Suppose two distributions are close, we would expect their mean, second moment/variance, and higher moments to be similar. This is the key idea of the method of moments: moment matching.

Suppose the distribution $f(x;\theta)$ depends on $\theta = (\theta_1, \cdots, \theta_k) \in \mathbb{R}^k$.

- First compute the jth moment of $F(x;\theta)$, i.e.,

$$\alpha_j(\theta) = \mathbb{E}\, X^j, \quad 1 \le j \le k.$$

- The jth moment can be estimated by sample moment

$$\widehat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$$

  where $\widehat{\alpha}_j$ converges to the population moment α_j in probability provided the higher order moment exists by the law of large number.

The method of moments estimator $\widehat{\theta}_n$ is the solution to a set of equations:

$$\alpha_j(\widehat{\theta}_n) = \widehat{\alpha}_j, \quad 1 \le j \le k.$$

In other words, the method of moments estimator matches sample moments. It is obtained by solving these equations.

Question: How to choose k? We usually choose the smallest k such that $\widehat{\theta}_n$ is uniquely determined, i.e., the solution to $\alpha_j(\theta) = \widehat{\alpha}_j$ is unique. We prefer lower moments since higher moment may not exist and also have a higher variance.

Example: Suppose $X_1, \cdots, X_n$ be i.i.d. samples from Poisson(λ). Find the method of moments estimators.

The first moment of Poisson(λ) is

$$\mathbb{E}\, X = \lambda$$

Thus

$$\widehat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

It is a consistent estimator and converges to λ in mean squared error.

However, the method of moments estimator is not unique:

$$\mathbb{E}\, X_i^2 = \mathrm{Var}(X_i) + (\mathbb{E}\, X_i)^2 = \lambda^2 + \lambda.$$

Now we can find the MOM estimator from

$$\lambda^2 + \lambda = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Another choice is

$$\widehat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$

since $\lambda = \mathrm{Var}(X_i)$. How to compare these three estimators? Which one is the best?

Example: For $X_1, \cdots, X_n \sim$Poisson(λ), $\overline{X}_n$ and S_n^2 are both unbiased estimators of λ. Which one has smaller MSE? What is the optimal MSE? You are encouraged to try but the calculation can be long and involved.

Example: Suppose $X_1, \cdots, X_n$ be i.i.d. samples from $\mathcal{N}(\mu, \sigma^2)$. Find the method of moments estimators.

The means of first two moments are

$$\alpha_1(\theta) = \mathbb{E}(X_i) = \mu$$

and

$$\alpha_2(\theta) = \mathbb{E}(X_i)^2 = \mathrm{Var}(X_i) + (\mathbb{E}(X_i))^2 = \mu^2 + \sigma^2.$$

Matching the quantities above with the sample moments gives

$$\mu = \frac{1}{n}\sum_{i=1}^n X_i = \overline{X}_n, \qquad \mu^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^n X_i^2.$$

Solving for μ and σ^2 gives

$$\widehat{\mu} = \overline{X}_n, \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2.$$

The estimator of σ^2 is biased but consistent.

Exercise: Find the method of moments estimator for Gamma(α, β).

Exercise: (Challenging if you are unfamiliar with multivariate probability) Find the method of moments estimator for bivariate Gaussian distribution. We say (X_i, Y_i) satisfies a bivariate Gaussian distribution with parameters $(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ if

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} \exp(-\frac{1}{2}[x-\mu_X, y-\mu_Y]\Sigma^{-1}[x-\mu_X, y-\mu_Y]^\top)$$

where $-1 \le \rho \le 1$ is the correlation between X and Y, and

$$\Sigma = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}.$$

Here μ_X and μ_Y are the mean of X and Y respectively; σ_X^2 and μ_Y^2 are the variance of X and Y respectively;

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}, \quad \mathrm{Cov}(X,Y) = \mathbb{E}\,XY - \mathbb{E}\,X\,\mathbb{E}\,Y.$$

You may try it now and we will focus more on multivariate Gaussian distributions.

Previous examples suggest that the method of moments is easy to use and also satisfies many useful properties. Is there any drawback for MOM? Yes, it has several weaknesses. First, some distributions do not have moments. For example, if $X_1, \cdots, X_n$ satisfy the shifted Cauchy distribution:

$$f_X(x,\theta) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}, \quad x \in \mathbb{R}.$$

Method of moment does not exist as $\mathbb{E}(X) = \infty$. It is also possible that the first moment exist while the second moment does not.

Secondly, the MOM estimators may not satisfy some natural constraints of the population distribution. Here is one such example.

Example: Suppose we have a sample $X_1, \cdots, X_n \sim \text{Binomial}(n, p)$ where n and p are unknown. Find the method of moment estimator for n and p.

First we compute the expectation and second moment.

$$\mathbb{E}(X) = np, \quad \mathbb{E}(X^2) = \text{Var}(X) + (\mathbb{E}(X))^2 = np(1-p) + n^2p^2.$$

We approximate the mean and second moment via its empirical mean and second moment:

$$np = \overline{X}_n, \quad np(1-p) + n^2p^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Solving for n and p:

$$\widehat{n} = \frac{\overline{X}_n^2}{\overline{X}_n - n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2}, \quad \widehat{p} = 1 - \frac{\sum_{i=1}^n (X_i - \overline{X}_n)^2}{n\overline{X}_n}$$

The estimation of n is not necessarily an integer, even negative if the rescaled empirical variance is larger than the empirical for a small sample.

4.2 Consistency and asymptotic normality of MM estimators

How about the quality of MOM estimators in general? We will discuss two important properties of MOM estimators under mild conditions.

- Consistency: $\widehat{\theta}_n \xrightarrow{p} \theta$. As the sample size increases to infinity, $\widehat{\theta}_n$ converges to θ in probability.
- Asymptotic normality:
$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$
for some variance σ^2. Asymptotic normality makes the construction of confidence interval possible.

Let's start with consistency. Remember that

$$\alpha_j(\theta) = \mathbb{E}\, X^j$$

form a set of functions w.r.t. variable θ. It is essentially a vector-valued function

$$h(\theta) = \begin{bmatrix} \alpha_1(\theta) \\ \vdots \\ \alpha_k(\theta) \end{bmatrix} \in \mathbb{R}^k.$$

Suppose $h(\theta)$ is a one-to-one function, then $h(\theta)$ is invertible, i.e., given $\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_1, \cdots, \widehat{\alpha}_k)^\top \in \mathbb{R}^k$,

$$\widehat{\theta}_n = h^{-1}(\widehat{\boldsymbol{\alpha}})$$

is uniquely determined. However, it could be tricky to determine if a function is one-to-one. For a single variable function, $h(\theta)$ is one-to-one if it is strictly increasing/decreasing.

Moreover, if $h(\theta)$ and $h^{-1}(\boldsymbol{\alpha})$ are also continuous at θ and $\boldsymbol{\alpha}$ respectively, then by continuous mapping theorem, it holds

$$\widehat{\theta}_n = h^{-1}(\widehat{\boldsymbol{\alpha}}) \xrightarrow{p} h^{-1}(\boldsymbol{\alpha}) = \theta.$$

since the law of large number guarantees

$$\widehat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j \xrightarrow{p} \alpha_j = \mathbb{E}\, X^j.$$

Theorem 4.2.1 (Consistency)**.** *Suppose $h(\theta)$ is one-to-one with its inverse function h^{-1} continuous at $\boldsymbol{\alpha}$, then $\widehat{\theta} = h^{-1}(\widehat{\boldsymbol{\alpha}})$ is a consistent estimator of θ provided that the corresponding moments exist.*

Question: Can we construct a confidence interval for θ? Yes, under certain mild conditions. Let's consider a simple case: the single variable case, and the analysis can be easily extended to multivariate case which will involve multivariate Gaussian distribution and CLT.

For $\theta \in \mathbb{R}$, we have

$$h(\theta) = \mathbb{E}\, X.$$

Suppose $h(\theta)$ is one-to-one, then the inverse function exists. The MOM estimator satisfies:

$$h(\widehat{\theta}_n) = \overline{X}_n.$$

In other words,

$$h(\widehat{\theta}_n) - h(\theta) = \overline{X}_n - \mathbb{E}\, X.$$

Recall that $\widehat{\theta}_n$ is a consistent estimator of θ, then $\widehat{\theta}_n$ is quite close to θ for large n and by linearization, we have

$$h'(\theta)(\widehat{\theta}_n - \theta) + R(\widehat{\theta}_n)(\widehat{\theta}_n - \theta) = \overline{X}_n - \mathbb{E}\, X$$

for some function $R(x)$ which goes to zero as $x \to \theta$. This is the Taylor's theorem with a remainder. Suppose $h'(\theta) \neq 0$ holds, then we have the following approximation:

$$\widehat{\theta}_n - \theta \approx \frac{1}{h'(\theta)}\left(\overline{X}_n - \mathbb{E}\, X\right)$$

Now we can see that

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2}{[h'(\theta)]^2}\right).$$

When constructing a confidence interval, we can use plug-in principle to estimate σ^2 and $h'(\theta)$ by S_n^2 and $h'(\widehat{\theta}_n)$ respectively.

By Slutsky's theorem, we have

$$\frac{\sqrt{n}h'(\widehat{\theta}_n)(\widehat{\theta}_n - \theta)}{S_n} \xrightarrow{d} \mathcal{N}(0,1).$$

A $1-\alpha$ confidence interval for θ is

$$\left|\theta - \widehat{\theta}_n\right| < \frac{z_{1-\alpha/2} S_n}{\sqrt{n}|h'(\widehat{\theta}_n)|}.$$

Example: Suppose $X_1, \cdots, X_n$ are samples from Geometric distribution with parameter p:

$$\mathbb{P}(X_i = k) = (1-p)^{k-1} p, \quad k \in \mathbb{Z}^+.$$

The MOM estimator is given by

$$\widehat{p} = 1/\overline{X}_n, \quad \mathbb{E}\, X = 1/p.$$

Can we derive a confidence interval for p? Let $h(p) = p^{-1}$ and then $h'(p) = -p^{-2}$. Thus a $1-\alpha$ CI for p is

$$|p - \widehat{p}| < \frac{z_{1-\alpha/2} S_n}{\sqrt{n} h'(\widehat{\theta}_n)} \Longrightarrow |p - \overline{X}_n^{-1}| < \frac{z_{1-\alpha/2} S_n}{\sqrt{n}} \cdot \overline{X}_n^2.$$

4.2.1 Generalized method of moments

We have seen that one issue of method of moments is the nonexistence of moments. Let's consider a simple scenario: it is possible that $\mathbb{E}\sqrt{|X|} < \infty$ while $\mathbb{E}\,|X|$ does not exist. More generally, what if we know $\mathbb{E}_\theta\, r(X)$ exists where $X \sim f(x;\theta)$? Can we extend the idea of method of moments to these general scenarios? The answer is yes. Assume

$$\alpha_r(\theta) = \mathbb{E}_\theta\, r(X) = \int_{\mathbb{R}} r(x) f(x;\theta)\, \mathrm{d}x$$

exists for some function $r(\cdot)$. Then we perform "moment matching" by solving $\widehat{\theta}_n$ from

$$\alpha_r(\theta) = \frac{1}{n}\sum_{i=1}^n r(X_i).$$

Intuitively, this approach also would work if $\alpha_r(\cdot)$ is one-to-one. This idea is actually called generalized method of moments. We can derive similar results for GMM by using the tools we have just covered.

4.3 Maximum likelihood estimation

Maximum likelihood estimation (MLE) is the most popular technique in statistical inference. Suppose we observe a set of data $x_1, \cdots, x_n \sim f(x;\theta)$. The idea of MLE is quite simple: we aim to find the θ such that the corresponding population distribution is most likely to generate the observed data. How to quantify the likelihood? Recall that the value of pdf/pmf $f(x;\theta)$ at x indicates the probability of the random variable X taking value around x. This leads to the definition of likelihood function.

Definition 4.3.1 (Likelihood function)**.** *If $X_1, \cdots, X_n$ are i.i.d. samples from their joint pdf/pmf $f(\boldsymbol{x};\theta)$, the likelihood function is defined by*

$$L(\theta | X_i = x_i, 1 \le i \le n) = \prod_{i=1}^n f(x_i;\theta).$$

Here we have a few remarks. Firstly, the likelihood function is just the joint density/pmf of the data, except that we treat it as a function of the parameter θ. Similarly, we can define the likelihood function for non i.i.d. samples. Suppose $X_1, \cdots, X_n$ are samples from a joint distribution $f_{X_1,\cdots,X_n}(\boldsymbol{x}|\theta)$, then

$$L(\theta|X_i = x_i, 1 \le i \le n) = f_{X_1,\cdots,X_n}(\boldsymbol{x};\theta).$$

The likelihood function for non-i.i.d. data will be very useful in linear regression. We will also briefly discuss an interesting example from matrix spike model later in this lecture. Another thing to bear in mind is: in general, $L(\theta|\boldsymbol{x})$ is not a density function of θ even though it looks like a conditional pdf of θ given $\boldsymbol{x}$.

Now it is natural to ask: what is maximum likelihood estimation? As the name has suggested, it means the maximizer of $L(\theta|\boldsymbol{x})$.

Definition 4.3.2 (Maximum likelihood estimation)**.** *The maximum likelihood estimator MLE denoted by $\widehat{\theta}_n$ is the value of θ which maximizes $L(\theta|x)$, i.e.,*

$$\widehat{\theta}_n = argmax_{\theta\in\Theta} L(\theta|\boldsymbol{x}).$$

This matches our original idea: we want to find a parameter such that the corresponding population distribution is most likely to produce the observed samples $\{x_1, \cdots, x_n\}$. In practice, we often maximize log-likelihood function instead of likelihood function. Log-likelihood function $\ell(\theta|\boldsymbol{x}) := \log L(\theta|\boldsymbol{x})$ is defined as the logarithm of likelihood $L(\theta|\boldsymbol{x})$.

There are two reasons to consider log-likelihood function:

- The log-transform will not change the maximizer since natural logarithm is a strictly increasing function.
- The log-transform enjoys the following property:

  $$\ell(\theta|\boldsymbol{x}) = \log \prod_{i=1}^{n} f_{X_i}(x_i;\theta) = \sum_{i=1}^{n} \log f_{X_i}(x_i;\theta)$$

  where X_i are independent with the pdf/pmf $f_{X_i}(\cdot;\theta)$. This simple equation usually makes the calculation and analysis much easier.

Example: Suppose that $x_1, \cdots, x_n \sim$ Bernoulli(p). The pmf is

$$f(x;p) = p^x(1-p)^{1-x}$$

where $x \in \{0, 1\}$.

First let's write down the likelihood function. Note that x_i is an independent copy of Bernoulli(p); their distribution equals the product of their marginal distributions.

$$\begin{aligned} L(p|\boldsymbol{x}) &:= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_i x_i}(1-p)^{n-\sum_i x_i} \\ &= p^{n\overline{x}_n}(1-p)^{n(1-\overline{x}_n)} \end{aligned}$$

where $\overline{x}_n = n^{-1}\sum_{i=1}^{n} x_i$. Next we take the logarithm of $L(p)$ (Here we omit $\boldsymbol{x}$ if there is no confusion) and we have

$$\ell(p) = n\overline{x}_n \log(p) + n\left(1 - \sum_{i=1}^{n} x_i\right)\log(1-p).$$

How to maximize it? We can differentiate it, find the critical point, and use tests to see if the solution is a global maximizer. We differentiate $\ell_n(p)$ w.r.t. p and obtain the critical point:

$$\frac{\mathrm{d}\ell(p)}{\mathrm{d}p} = \frac{n\overline{x}_n}{p} - \frac{n(1-\overline{x}_n)}{1-p} = 0 \Longrightarrow \widehat{p} = \overline{x}_n$$

Is $\widehat{p}$ a global maximizer?

$$\frac{\mathrm{d}^2\ell(p)}{\mathrm{d}p^2} = -\frac{n\overline{x}_n}{p^2} - \frac{n(1-\overline{x}_n)}{(1-p)^2} < 0.$$

This implies that the likelihood function is concave. All the local maximizers of a concave function are global. Therefore, $\widehat{p} = \overline{x}_n$ is the MLE. If we treat each x_i as a realization of X_i, then the statistic

$$\widehat{p} = \overline{X}_n$$

is a consistent estimator of p and enjoys asymptotic normality $\sqrt{n}(\widehat{p}-p) \xrightarrow{d} \mathcal{N}(0,1)$ by CLT. From now on, we replace x_i by X_i since x_i is a realization of X_i.

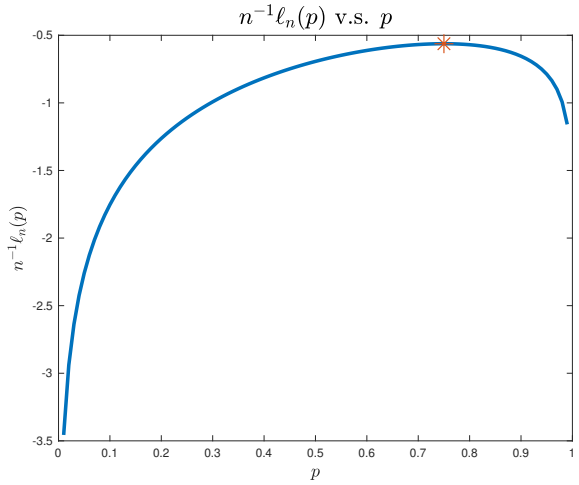


Figure 4.1: Plot of $\ell(p)$ with $n = 100$ and $\sum_{i=1}^n x_i = 75$.

Example: Suppose $X_1, \cdots, X_n$ are sampled from $\mathcal{N}(\mu, \sigma^2)$ where μ and $\sigma^2 > 0$ are unknown. What is the MLE of (μ, σ^2)?

$$\begin{aligned} L(\mu, \sigma^2 | X_1, \cdots, X_n) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i-\mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\sum_{i=1}^n \frac{(X_i-\mu)^2}{2\sigma^2}\right). \end{aligned}$$

Taking the logarithm of the likelihood function leads to

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\mu)^2 + C$$

where C contains no information about μ and σ^2. Taking the partial derivative w.r.t. μ and σ^2 (we treat σ^2 as a variable) gives

$$\begin{aligned} \frac{\partial\ell}{\partial\mu} &= -\frac{1}{\sigma^2}\sum_{i=1}^n (\mu - X_i) = 0 \\ \frac{\partial\ell}{\partial\sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (X_i-\mu)^2 = 0 \end{aligned}$$

Solving for μ and σ^2:

$$\widehat{\mu} = \overline{X}_n, \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{n-1}{n}S_n^2.$$

Are they the global maximizers of $\ell(\mu, \sigma^2)$?

Exercise: Show that $\left(\overline{X}_n, \frac{n-1}{n}S_n^2\right)$ is the global maximizer to $\ell(\mu, \sigma^2)$.

Exercise: Show that $(n-1)S_n^2/\sigma^2 \sim \chi^2_{n-1}$ if $X_i \sim \mathcal{N}(\mu, \sigma^2)$. (This may be challenging now. We will discuss this later. But it would be nice to think about it now.)

Exercise: Show that $\overline{X}_n$ and S_n^2 are independent. (Same comments as the previous exercise).

To show the global optimality could be slightly trickier. However, we can quickly verify its local optimality by checking it Hessian matrix.

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \mu^2} &= -\frac{n}{\sigma^2}, \\
\frac{\partial^2 \ell}{\partial \mu \partial \sigma^2} &= \frac{1}{\sigma^4}\sum_{i=1}^{n}(\mu - X_i), \\
\frac{\partial^2 \ell}{\partial (\sigma^2)^2} &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(X_i - \mu)^2.
\end{aligned}$$

Let's evaluate the Hessian at $(\widehat{\mu}, \widehat{\sigma}^2)$:

$$\frac{\partial^2 \ell}{\partial \mu^2} = -\frac{n}{\widehat{\sigma}^2}, \quad \frac{\partial^2 \ell}{\partial \mu \partial \sigma^2} = 0$$

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial (\sigma^2)^2} &= \frac{n}{2\widehat{\sigma}^4} - \frac{1}{\widehat{\sigma}^6}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2 \\
&= \frac{n}{2\widehat{\sigma}^4} - \frac{n}{\widehat{\sigma}^4} = -\frac{n}{2\widehat{\sigma}^4}.
\end{aligned}$$

Thus the Hessian matrix is

$$\nabla^2 \ell(\widehat{\mu}, \widehat{\sigma}^2) = -\begin{bmatrix} \frac{n}{\widehat{\sigma}^2} & 0 \\ 0 & \frac{n}{2\widehat{\sigma}^4} \end{bmatrix}.$$

It is negative definite which is equivalent to the statement that all of its eigenvalues are negative. As a result, $(\widehat{\mu}, \widehat{\sigma}^2)$ is a local maximizer. The Hessian of log-likelihood function plays an important role in statistics, which is also called *Fisher information matrix*. Obviously, the MLE is a consistent estimator of (μ, σ^2).

Example: Let $X_1, \cdots, X_n \sim \text{Unif}(0, \theta)$ where θ is unknown. Recall that

$$f(x; \theta) = \begin{cases} \theta^{-1}, & \text{if } 0 \le x \le \theta, \\ 0, & \text{otherwise.} \end{cases}$$

First let's write down $L_n(\theta|\boldsymbol{X})$:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta) = \frac{1}{\theta^n}\prod_{i=1}^{n} 1\{X_i \le \theta\} = \frac{1}{\theta^n} 1\left\{\max_{1\le i\le n} X_i \le \theta\right\}.$$

Note $L_n(\theta)$ is zero if $\max_{1\le i\le n} X_i > \theta$ and is decreasing as θ increases. Therefore, the MLE is $\max_{1\le i\le n} X_i$.

Exercise: Is $\widehat{\theta} = \max_{1\le i\le n} X_i$ a consistent and unbiased estimator of θ?

Exercise: Does $\widehat{\theta} = \max_{1\le i\le n} X_i$ enjoy asymptotic normality? Or ask what is the distribution of $\widehat{\theta}$? We actually have derived the distribution in one homework problem.

All the examples we have seen have an MLE of closed form. However, it is often that we don't have an explicit formula for the MLE. Here is one such example.

Example: Consider

$$f(x;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad x > 0.$$

How to find out the MLE?

$$L(\alpha,\beta|X_i, 1\le i\le n) = \frac{1}{\Gamma(\alpha)^n \beta^{n\alpha}} \left(\prod_{i=1}^n X_i\right)^{\alpha-1} \exp(-\sum_{i=1}^n X_i/\beta), \quad x > 0.$$

After taking the log, it holds

$$\ell(\alpha,\beta) = -n\log\Gamma(\alpha) - n\alpha\log\beta + (\alpha-1)\log\left(\prod_{i=1}^n X_i\right) - \beta^{-1} n\overline{X}_n.$$

Maximizing the log-likelihood function over α is quite tricky as it will involve the derivative of Gamma function. On the other hand, it is quite simple to derive the method of moments estimator for Gamma(α,β).

Here is another very interesting but challenging problem related to the matrix spike model, a famous model in high dimensional statistics. Feel free to try it.

Challenging Exercise: Suppose we observe the data

$$Y_{ij} = \theta_i\theta_j + \sigma W_{ij},$$

where $\{W_{ij}\}_{i\le j}$ are independent Gaussian random variables and satisfy

$$W_{ij} = \begin{cases} \mathcal{N}(0,1), & i \ne j, \\ \mathcal{N}(0,2), & i = j. \end{cases}$$

and $W_{ij} = W_{ji}$. Here $\theta = (\theta_1, \cdots, \theta_n)^\top \in \mathbb{R}^n$ is the hidden parameter to be estimated and $\|\theta\| = 1$. Write down the log-likelihood function and show that the MLE of θ is the top eigenvector of $\boldsymbol{Y}$. (Note that Y_{ij} are independent but not identically distributed.)

In fact, finding the MLE in some statistical models is quite challenging and may even be NP-hard.

- Optimization tools are needed to maximize

  $$\max_{\theta\in\Theta} \ell_n(\theta).$$

  Algorithms include gradient descent, Newton's method, EM algorithm (expectation-maximization), etc...

- Does the likelihood function always have unique maximizer? In fact, this is not the case. The likelihood from some examples exhibit complicated landscape (location/number of local optima) which is an active research field.

4.4 Properties of MLE

We have discussed that MOM estimators satisfy consistency and asymptotic normality property under certain mild conditions. Does these two important properties also hold for MLE?

4.4.1 Consistency

First let's recall some examples:

1. For $X_i \sim$Bernoulli(p), the MLE is $\widehat{p} = \overline{X}_n$. By LLN, $\widehat{p}$ is a consistent estimator of p.
2. For $X_i \sim$Unif$[0, \theta]$, the MLE is $\widehat{\theta}_n = \max X_i \le \theta$. We have calculated the distribution of $\widehat{\theta}_n$
$$\mathbb{P}(\widehat{\theta}_n \le x) = \theta^{-n} x^n, \quad 0 \le x \le \theta.$$
Now $\mathbb{P}(|\theta - \widehat{\theta}_n| \ge \epsilon) = \mathbb{P}(\widehat{\theta}_n \le \theta - \epsilon) = (1 - \theta^{-1}\epsilon)^n \to 0$ as n goes to infinity.

Therefore, $\widehat{\theta}_n$ is a consistent estimator of θ in the two examples described above. Can we extend it to more general scenarios?

Theorem 4.4.1 (Consistency)**.** *Under certain regularity condition, the MLE is consistent, i.e.,*
$$\lim_{n\to\infty} \mathbb{P}(|\widehat{\theta}_n - \theta| \ge \epsilon) = 0$$
where θ is the population parameter.

These regularity conditions include:

- The pdf/pmf satisfies $f(x; \theta_1) = f(x; \theta_2)$ iff $\theta_1 = \theta_2$. In other words, we ensure the parameters determine the distribution uniquely.
- $\mathbb{E}_{\theta_0} \log f(X; \theta) < \infty$ for the true population parameter θ_0.

An informal proof. Suppose $X_1, \cdots, X_n \sim f(x; \theta_0)$ where θ_0 is the true parameter.

$$\ell(\theta) = \log \prod_{i=1}^n f(X_i; \theta) = \sum_{i=1}^n \log f(X_i; \theta).$$

Each $\log f(X_i; \theta)$ is an independent copy of the random variable $f(X; \theta)$. By the law of large number, it holds that
$$\frac{1}{n}\ell(\theta) \xrightarrow{p} \mathbb{E}_{\theta_0} \log f(X; \theta)$$
as $n \to \infty$ where
$$\mathbb{E}_{\theta_0} \log f(X; \theta) = \int_{\mathcal{X}} (\log f(x; \theta)) f(x; \theta_0) \, \mathrm{d}x.$$

On the other hand, we can show that

$$\theta_0 = \mathrm{argmax}_\theta \, \mathbb{E}_{\theta_0} \log f(X; \theta)$$

Why? Note that $\log t \le t - 1$ for $t > 0$.

$$\mathbb{E}_{\theta_0} \left(\log f(X; \theta) - \log f(X; \theta_0)\right) = \mathbb{E}_{\theta_0} \log \frac{f(X; \theta)}{f(X; \theta_0)} \le \mathbb{E}_{\theta_0} \left(\frac{f(X; \theta)}{f(X; \theta_0)} - 1 \right) = 0$$

since

$$\mathbb{E}_{\theta_0}\left(\frac{f(X;\theta)}{f(X;\theta_0)}-1\right)=\int_{\mathcal{X}}\left(f(x;\theta)-f(x;\theta_0)\right)\mathrm{d}x=1-1=0.$$

Now we summarize:

- $\widehat{\theta}_n$ is the MLE, i.e., the global maximizer of $\ell(\theta)$;
- θ_0 is the unique maximizer of $\mathbb{E}_{\theta_0}\log f(X;\theta)$ due to identifiability;
- For any θ, $n^{-1}\ell(\theta)$ converges to its expectation $\mathbb{E}_{\theta_0}\log f(X;\theta)$.

Since $\ell(\theta)$ and $\mathbb{E}_{\theta_0}\log f(X;\theta)$ are getting closer, their points of maximum should also get closer. □

The proof is not completely rigorous: in order to make the argument solid, we actually would ask the uniform convergence of $n^{-1}\ell(\theta)$ to $\mathbb{E}_{\theta_0}\log f(X;\theta)$.

4.4.2 Asymptotic normality

Like method of moments estimator, MLE also enjoys asymptotic normality, provided that certain regularity conditions holds. However, not every MLE enjoys asymptotic normality.

In order to discuss asymptotic distribution of $\widehat{\theta}_n$, we will introduce a very useful quantity called Fisher information. The Fisher information of a random variable $X\sim f(x;\theta)$ is given by

$$I(\theta)=-\,\mathbb{E}_\theta\left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(X;\theta)\right).$$

This is essentially the negative second derivative of log-likelihood function ($n=1$) in expectation.

Exercise: Show that

$$I(\theta)=-\,\mathbb{E}_\theta\left(\frac{\mathrm{d}^2\log f(X;\theta)}{\mathrm{d}\theta^2}\right)=\mathbb{E}_\theta\left(\frac{\mathrm{d}\log f(X;\theta)}{\mathrm{d}\theta}\right)^2.$$

(Hint: Suppose differentiation and integration can be exchanged. We have

$$\int_{\mathcal{X}}\frac{\mathrm{d}f(x;\theta)}{\mathrm{d}\theta}\mathrm{d}x=0,\quad \int_{\mathcal{X}}\frac{\mathrm{d}^2f(x;\theta)}{\mathrm{d}\theta^2}\mathrm{d}x=0,$$

since $\int_{\mathcal{X}}f(x;\theta)\,\mathrm{d}x=1$.)

Exercise: Find the Fisher information $I(\mu,\sigma^2)$ for $\mathcal{N}(\mu,\sigma^2)$.

In particular, for a sample of n i.i.d. data points $X_1,\cdots,X_n$, their Fisher information is

$$I_n(\theta)=-\,\mathbb{E}_\theta\,\ell''(\theta)$$

where

$$\ell(\theta)=\sum_{i=1}^n\log f(X_i;\theta).$$

Apparently, if $\{X_i\}_{i=1}^d$ are i.i.d., then

$$I_n(\theta)=-\,\mathbb{E}\sum_{i=1}^n\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(X_i;\theta)=nI(\theta)$$

provided that the differentiation and expectation are exchangeable.

Theorem 4.4.2. *Suppose $\widehat{\theta}_n$ is the MLE, i.e., the global maximizer of $\ell(\theta)$. Then*

$$\widehat{\theta}_n - \theta_0 \xrightarrow{d} \mathcal{N}(0, I_n(\theta_0)^{-1})$$

where θ_0 is the true parameter provided that $\ell(\theta)$ has a continuous second derivative in a neighborhood around θ_0 and $I_n(\theta) = nI(\theta)$ is continuous at θ_0.

Once we have the asymptotic normality of MLE, we can use it to construct a $(1-\alpha)$ confidence interval. Since $\widehat{\theta}_n$ is a consistent estimator of θ_0, Slutsky theorem implies that

$$\frac{\sqrt{n}(\widehat{\theta}_n - \theta_0)}{\sqrt{\left[I(\widehat{\theta}_n)\right]^{-1}}} \xrightarrow{d} \mathcal{N}(0,1)$$

since $I(\widehat{\theta}_n) \xrightarrow{p} I(\theta_0)$ if $I(\theta)$ is continuous at θ_0. Note that $I_n(\widehat{\theta}_n) = nI(\widehat{\theta}_n)$. Therefore, this asymptotic distribution can also be written into

$$\sqrt{I_n(\widehat{\theta}_n)} \cdot (\widehat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0,1) \quad \text{or} \quad \widehat{\theta}_n - \theta \xrightarrow{d} \mathcal{N}(0, I_n(\widehat{\theta}_n)^{-1}).$$

A $1-\alpha$ confidence interval could be

$$|\widehat{\theta}_n - \theta| \le \frac{z_{1-\alpha/2}}{\sqrt{nI(\widehat{\theta}_n)}}, \quad \text{or} \quad |\widehat{\theta}_n - \theta| \le \frac{z_{1-\alpha/2}}{\sqrt{I_n(\widehat{\theta}_n)}}.$$

Example: Let $X_1, \cdots, X_n \sim$Bernoulli(p). Its pmf is $f(x;p) = p^x(1-p)^{1-x}, \quad x \in \{0,1\}$. Then

$$\log f(x;p) = x \log p + (1-x)\log(1-p).$$

The first and second derivative are

$$\frac{\mathrm{d}\log f(x;p)}{\mathrm{d}p} = \frac{x}{p} - \frac{1-x}{1-p}, \quad \frac{\mathrm{d}^2 \log f(x;p)}{\mathrm{d}p^2} = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}$$

$$I(p) = -\mathbb{E}\,\frac{\mathrm{d}^2 \log f(X;p)}{\mathrm{d}p^2} = \frac{p}{p^2} + \frac{1-p}{(1-p)^2} = \frac{1}{p(1-p)}.$$

Therefore, we have

$$\sqrt{n}(\widehat{p}_n - p) \sim \mathcal{N}(0, p(1-p)).$$

This matches the result obtained from CLT.

Exercise: Let $X_1, \cdots, X_n \sim$Poisson(λ). First find the MLE of λ. Calculate the Fisher information $I(\lambda)$ and construct a $1-\alpha$ confidence interval for λ.

Example: Let $X_1, \cdots, X_n \sim$Unif$[0,\theta]$. We know that the MLE of θ is $\widehat{\theta} = \max X_i$. The cdf of $\widehat{\theta}_n$ is

$$\mathbb{P}_\theta\left(\widehat{\theta}_n \le x\right) = \frac{x^n}{\theta^n}, \quad 0 \le x \le \theta.$$

Now we can see that the distribution of $\sqrt{n}(\widehat{\theta}_n - \theta)$ does not satisfy normal distribution. The likelihood function $L(\theta)$ is discontinuous at $\widehat{\theta}_n$ and takes value 0 if $\theta > \widehat{\theta}_n$.

Proof of asymptotic normality. By Taylor theorem, it holds that

$$\ell'(\widehat{\theta}_n) = \ell'(\theta_0) + \ell''(\theta_0)(\theta_0 - \widehat{\theta}_n) + o(|\widehat{\theta}_n - \theta_0|).$$

Here $o(|\theta_0 - \widehat{\theta}_n|)$ means this term has a smaller order than $|\theta_0 - \widehat{\theta}_n|$.

Note that $\ell'(\widehat{\theta}_n) = 0$ since $\widehat{\theta}_n$ is the maximizer. We have

$$-\ell'(\theta_0) = \ell''(\theta_0)(\widehat{\theta}_n - \theta_0) + o(|\widehat{\theta}_n - \theta_0|) \Longrightarrow \widehat{\theta}_n - \theta_0 \approx -\frac{\ell'(\theta_0)}{\ell''(\theta_0)}$$

which implies

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = -\frac{\frac{1}{\sqrt{n}}\ell'(\theta_0)}{\frac{1}{n}\ell''(\theta_0)} + \text{small error}.$$

We will see why we need these terms containing n immediately.

Note that

$$\mathbb{E}_{\theta_0}\, \ell'(\theta_0) = 0$$

since θ_0 is global maximizer to $\mathbb{E}_{\theta_0} \log f(X;\theta)$. By CLT, we have

$$\frac{1}{\sqrt{n}}\ell'(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}\theta} \log f(X_i;\theta)\Big|_{\theta=\theta_0} \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log f(X_i;\theta)\right)^2 \Big|_{\theta=\theta_0}\right) = \mathcal{N}(0, I(\theta_0))$$

where each $\frac{\mathrm{d}}{\mathrm{d}\theta} \log f(X_i;\theta)$ is an i.i.d. random variable. On the other hand, by law of large number,

$$\frac{1}{n}\ell''(\theta_0) \xrightarrow{p} \mathbb{E}\, \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f(X;\theta_0) = -I(\theta_0).$$

in probability. Therefore, we have $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ converges in distribution:

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = -\frac{\frac{1}{\sqrt{n}}\ell'(\theta_0)}{\frac{1}{n}\ell''(\theta_0)} \xrightarrow{d} \mathcal{N}(0, I(\theta_0)^{-1})$$

which follows from Slutsky's theorem. □

4.4.3 Equivariance of MLE

Proposition 4.4.3. *If $\widehat{\theta}$ is the MLE of θ, then $g(\widehat{\theta})$ is the MLE of $g(\theta)$.*

Proof: Let $\tau = g(\theta)$ and $\widehat{\theta}$ be the MLE of $L(\theta|\boldsymbol{x})$. We prove the simplest case where g is one-to-one. Define

$$L^*(\tau|\boldsymbol{x}) := L(g^{-1}(\tau)|\boldsymbol{x}) \le L(\widehat{\theta}|\boldsymbol{x})$$

We call $L^*(\tau|\boldsymbol{x})$ the induced likelihood. It is easy to see that $\widehat{\tau} = g(\widehat{\theta})$ attains the maximum of $L^*(\tau|\boldsymbol{x})$:

$$L^*(\widehat{\tau}|\boldsymbol{x}) = L(g^{-1}(\widehat{\tau})|\boldsymbol{x}) = L(\widehat{\theta}|\boldsymbol{x}).$$

Thus $\widehat{\tau} = g(\widehat{\theta})$ is the MLE of $\tau = g(\theta)$.

For the non-one-to-one scenario, we define

$$L^*(\tau|\boldsymbol{x}) := \sup_{\{\theta: g(\theta)=\tau\}} L(\theta|\boldsymbol{x}).$$

It is the induced likelihood function of τ. This definition does not depend on whether g is one-to-one or not. Since $\widehat{\theta}$ is the MLE of $L(\theta|\boldsymbol{x})$, $L^*(\tau|\boldsymbol{x}) \leq L(\widehat{\theta}|\boldsymbol{x})$. On the other hand,

$$L^*(\widehat{\tau}|\boldsymbol{x}) = \sup_{\{\theta: g(\theta)=\widehat{\tau}\}} L(\theta|\boldsymbol{x}) \geq L(\widehat{\theta}|\boldsymbol{x}).$$

Thus $\widehat{\tau}$ is the MLE of $g(\theta)$. □

Example: Suppose $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ and the MLE of σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

By equivariance, the MLE of σ is

$$\widehat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}.$$

4.5 Delta method

Remember we know that $\widehat{\tau} = g(\widehat{\theta}_n)$ is the MLE of $\tau = g(\theta)$.

- Does $\widehat{\tau}$ converge to τ?
- How to construct an approximate confidence interval for τ?

If $g(\cdot)$ is continuous at θ, then by continuous mapping theorem, it holds that

$$g(\widehat{\theta}_n) \xrightarrow{p} g(\theta).$$

Regarding the construction of confidence interval for τ, we need to know the asymptotic distribution of $g(\widehat{\theta}_n)$.

Theorem 4.5.1 (Delta method)**.** *Suppose a sequence of random variables $\widehat{\theta}_n$ satisfies that $\sqrt{n}(\widehat{\theta}_n - \theta)$ converges to $\mathcal{N}(0, \sigma^2)$ in distribution. For a given function $g(x)$ such that $g'(\theta)$ exists and is nonzero, then*

$$\sqrt{n}(g(\widehat{\theta}_n) - g(\theta)) \xrightarrow{d} \mathcal{N}(0, [g'(\theta)]^2\sigma^2).$$

Proof: By Taylor's theorem, it holds that

$$\begin{aligned} g(\widehat{\theta}_n) - g(\theta) &= g'(\theta)(\widehat{\theta}_n - \theta) + R(\widehat{\theta}_n)(\widehat{\theta}_n - \theta) \\ &= (g'(\theta) + R(\widehat{\theta}_n))(\widehat{\theta}_n - \theta) \end{aligned}$$

where $R(\widehat{\theta}_n)$ is the remainder and goes to zero as $\widehat{\theta}_n \xrightarrow{p} \theta$. Note that $\widehat{\theta}_n$ converges to θ in probability and thus R vanishes as n approaches ∞. Therefore,

$$\sqrt{n}(g(\widehat{\theta}_n) - g(\theta)) = (g'(\theta) + R(\widehat{\theta}_n)) \cdot \sqrt{n}(\widehat{\theta}_n - \theta)$$

As $n \to \infty$, the Slutsky's theorem implies that

$$\sqrt{n}(g(\widehat{\theta}_n) - g(\theta)) \xrightarrow{d} \mathcal{N}(0, [g'(\theta)]^2\sigma^2)$$

since $\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ and $R(\widehat{\theta}_n) \xrightarrow{p} 0$. □

Exercise: Show that $\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0,1)$ implies that $\widehat{\theta}_n$ converges to θ in probability.

Following from Delta method, we immediately have:

Theorem 4.5.2. *Suppose $\widehat{\theta}_n$ is the MLE of θ, then $\widehat{\tau}$ satisfies*

$$\sqrt{n}(\widehat{\tau}_n - \tau) \to \mathcal{N}\left(0, \frac{[g'(\theta)]^2}{I(\theta)}\right)$$

where $I(\theta)$ is the Fisher information of θ and $g(\cdot)$ has non-vanishing derivative at θ.

Exercise: what if $g'(\theta) = 0$ but $g''(\theta)$ exists? Show that if $g'(\theta) = 0$ and $g''(\theta) \neq 0$, then

$$n(g(\widehat{\theta}_n) - g(\theta)) \xrightarrow{d} \frac{\sigma^2 g''(\theta)}{2}\chi_1^2$$

where

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

(Hint: Use 2nd-order Taylor approximation:

$$\begin{aligned} g(\widehat{\theta}_n) - g(\theta) &= g'(\theta)(\widehat{\theta}_n - \theta) + \frac{1}{2}g''(\theta)(\widehat{\theta}_n - \theta)^2 + R \\ &= \frac{1}{2}g''(\theta)(\widehat{\theta}_n - \theta)^2 + R \end{aligned}$$

Note that $Z = \frac{\sqrt{n}(\widehat{\theta}_n - \theta)}{\sigma}$ converges to a standard normal random variable. Z^2 is actually χ_1^2 distribution (chi-square distribution of degree 1). One can derive the corresponding distribution of $n(g(\widehat{\theta}_n) - g(\theta))$ by using this fact.)

Exercise: Show that the asymptotic distribution

$$\lambda(X_1, \cdots, X_n) = 2\log\frac{L(\widehat{\theta}_n)}{L(\theta_0)} = 2(\ell(\widehat{\theta}_n) - \ell(\theta_0)) \sim \chi_1^2$$

where $X_i \sim f(x;\theta_0)$, i.e., θ_0 is the true parameter, and $\widehat{\theta}_n$ is the MLE. This is called the likelihood ratio statistic. We will see this again in likelihood ratio test.

Example: Suppose we observe $X_1, \cdots, X_n \sim$Bernoulli(p) random variables. We are interested in the odds

$$\psi = \frac{p}{1-p}.$$

The MLE of ψ is $\widehat{\psi} = \widehat{p}/(1-\widehat{p})$ where $\widehat{p} = \overline{X}_n$.

The variance of $\widehat{\psi}$ is

$$\mathrm{Var}(\widehat{\psi}) \approx [g'(p)]^2 \,\mathrm{Var}(\widehat{p}) = \frac{1}{(1-p)^4} \cdot \frac{p(1-p)}{n} = \frac{p}{n(1-p)^3}$$

where

$$g'(p) = \frac{1}{(1-p)^2}.$$

The limiting distribution is

$$\sqrt{n}(\widehat{\psi} - \psi) \sim \mathcal{N}\left(0, \frac{p}{(1-p)^3}\right).$$

Example: Suppose $X_1, \cdots, X_n \sim$Geo(p). Find the MLE of p. Use Delta method to find an asymptotic distribution of $\sqrt{n}(\widehat{p} - p)$.

We start with the log-likelihood function:

$$L(p|X_i, 1 \le i \le n) = \prod_{i=1}^{n} (1-p)^{X_i - 1} p = (1-p)^{\sum_{i=1}^n X_i - n} p^n$$

and

$$\ell(p) = n(\overline{X}_n - 1)\log(1-p) + n \log p.$$

Let's compute the critical point and show it is a global maximizer.

$$\ell'(p) = -\frac{n(\overline{X}_n - 1)}{1-p} + \frac{n}{p} \Longrightarrow \widehat{p} = \frac{1}{\overline{X}_n}.$$

Note that by CLT, we have

$$\sqrt{n}(\overline{X} - 1/p) \xrightarrow{d} \mathcal{N}(0, (1-p)/p^2)$$

where $\mathrm{Var}(X_i) = (1-p)/p^2$.

By Delta method (letting $g(x) = 1/x$ and $x = \overline{X}_n$), it holds

$$\mathrm{Var}(g(\overline{X}_n)) \approx [g'(1/p)]^2 \cdot \mathrm{Var}(\overline{X}_n) = p^4 \cdot \frac{1-p}{np^2} = \frac{p^2(1-p)}{n}.$$

Therefore,

$$\sqrt{n}(\widehat{p} - p) \xrightarrow{d} \mathcal{N}\left(0, (1-p)p^2\right).$$

Now we try another approach:

$$\ell''(p) = -\frac{n(\overline{X}_n - 1)}{(1-p)^2} - \frac{n}{p^2}$$

So the Fisher information is

$$I_n(p) = \mathbb{E}_p\, \ell''(p) = \frac{n(1/p - 1)}{(1-p)^2} + \frac{n}{p^2} = \frac{n}{p(1-p)} + \frac{n}{p^2} = \frac{n}{p^2(1-p)}$$

where $\mathbb{E}\, X_i = 1/p$ and $\mathbb{E}\, \overline{X}_n = 1/p$. By asymptotic normality of MLE, we have

$$\sqrt{n}(\widehat{p} - p) \xrightarrow{d} \mathcal{N}\left(0, (1-p)p^2\right).$$

4.6 Cramer-Rao bound

Can we find an unbiased estimator of θ with variance as small as possible?

Theorem 4.6.1. *Suppose $\widehat{\theta}_n$ is an unbiased estimator of θ with finite variance. Then*

$$\mathrm{Var}_\theta(\widehat{\theta}_n) \ge \frac{1}{nI(\theta)}.$$

- Any unbiased estimator must have variance at least $1/I(\theta)$.
- MLE is asymptotically the best unbiased estimator of θ. (Efficiency)

Theorem 4.6.2 (Cramer-Rao inequality)**.** *Suppose $\widehat{\theta}_n$ is an estimator of θ with finite variance. Then*

$$\mathrm{Var}_\theta(\widehat{\theta}_n) \geq \frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\,\mathbb{E}_\theta\,\widehat{\theta}_n\right)^2}{nI(\theta)}.$$

where

$$I(\theta) = \int_{\mathbb{R}} \left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x;\theta)\right)^2 f(x;\theta)\,\mathrm{d}x$$

with $f(x;\theta)$ as the pdf.

Proof: Note that $\mathbb{E}_\theta\,\widehat{\theta}_n$ is a function of θ. Assume that the integration and differentiation can be exchanged, i.e.,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\,\mathbb{E}_\theta(\widehat{\theta}) = \int_{\mathcal{X}} \widehat{\theta}(\boldsymbol{x})\frac{\mathrm{d}}{\mathrm{d}\theta} f(\boldsymbol{x};\theta)\,\mathrm{d}\boldsymbol{x}$$

where $f(\boldsymbol{x};\theta) = \prod_{i=1}^n f(x_i;\theta)$ is the joint pdf of $X_1, \cdots, X_n$.)

Note that by definition we have

$$\int_{\mathcal{X}} (\widehat{\theta}(\boldsymbol{x}) - \mathbb{E}_\theta(\widehat{\theta}_n)) f(\boldsymbol{x};\theta)\,\mathrm{d}\boldsymbol{x} = 0.$$

Differentiating it w.r.t. θ:

$$\int_{\mathcal{X}} (\widehat{\theta}(\boldsymbol{x}) - \mathbb{E}_\theta(\widehat{\theta}_n))\frac{\mathrm{d}f(\boldsymbol{x};\theta)}{\mathrm{d}\theta}\,\mathrm{d}\boldsymbol{x} - \int_{\mathcal{X}} \frac{\mathrm{d}\,\mathbb{E}_\theta\,\widehat{\theta}_n}{\mathrm{d}\theta} f(\boldsymbol{x};\theta)\,\mathrm{d}\boldsymbol{x} = 0$$

which gives

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\,\mathbb{E}_\theta(\widehat{\theta}_n) = \int_{\mathcal{X}} (\widehat{\theta}(\boldsymbol{x}) - \mathbb{E}_\theta(\widehat{\theta}_n))\frac{\mathrm{d}f(\boldsymbol{x};\theta)}{\mathrm{d}\theta}\,\mathrm{d}\boldsymbol{x}.$$

Applying Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\left|\frac{\mathrm{d}}{\mathrm{d}\theta}\,\mathbb{E}_\theta(\widehat{\theta}_n)\right|^2 &= \left|\int_{\mathcal{X}} (\widehat{\theta}(\boldsymbol{x}) - \mathbb{E}_\theta(\widehat{\theta}_n))\sqrt{f(\boldsymbol{x};\theta)}\cdot\frac{1}{\sqrt{f(\boldsymbol{x};\theta)}}\frac{\mathrm{d}f(\boldsymbol{x};\theta)}{\mathrm{d}\theta}\,\mathrm{d}\boldsymbol{x}\right|^2 \\
&\leq \int_{\mathcal{X}} (\widehat{\theta}(\boldsymbol{x}) - \mathbb{E}_\theta(\widehat{\theta}_n))^2 f(\boldsymbol{x};\theta)\,\mathrm{d}\boldsymbol{x}\cdot\int_{\mathcal{X}}\frac{1}{f(\boldsymbol{x};\theta)}\left(\frac{\mathrm{d}f(\boldsymbol{x};\theta)}{\mathrm{d}\theta}\right)^2\mathrm{d}\boldsymbol{x} \\
&= \mathrm{Var}_\theta(\widehat{\theta}_n)\cdot\mathbb{E}_\theta\left(\frac{\mathrm{d}\log f(\boldsymbol{x};\theta)}{\mathrm{d}\theta}\right)^2 \\
&= \mathrm{Var}_\theta(\widehat{\theta}_n)\cdot nI(\theta).
\end{aligned}$$

This finishes the proof. □

Example: Suppose $X_1, \cdots, X_n \sim$Poisson(λ).

$$f(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x \in \{0, 1, \cdots\}.$$

The Fisher information is $I(\lambda) = \lambda^{-1}$. Note that the MLE of λ is $\overline{X}_n$.

$$\mathrm{Var}_\lambda(\overline{X}_n) = \frac{\lambda}{n}, \quad \frac{1}{nI(\lambda)} = \frac{\lambda}{n}.$$

Thus $\overline{X}_n$ is the best unbiased estimator of λ since $\text{Var}(\overline{X}_n) = \lambda/n$.

Exercise: Suppose $X_1, \cdots, X_n \sim \mathcal{N}(0, \sigma^2)$. Find the MLE of σ^2 and the Fisher information $I(\sigma^2)$. Show that

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2$$

is the unbiased estimator of σ^2 with the smallest possible variance.

Exercise: Suppose $X_1, \cdots, X_n \sim \mathcal{N}(0, \sigma^2)$. Find the MLE of σ and the Fisher information $I(\sigma)$. What is the actual and approximate variance of $\widehat{\sigma}$?

Exercise: Suppose $X_1, \cdots, X_n$ are samples from

$$f(x;\theta) = \begin{cases} \theta x^{\theta-1}, & 0 < x < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Recall that the MLE of θ is

$$\widehat{\theta} = -\frac{n}{\sum_{i=1}^{n} \log X_i}.$$

Use Delta method to find out the asymptotic distribution of $\widehat{\theta}$. Does it match the result obtained by directly applying the asymptotic normality of MLE?

4.7 Multiparameter models

In practice, we are often facing the problem of inferring multiple population parameters from a dataset. Therefore, we ask if it is possible to extend all the analysis we've done the multi-parameter scenario.

As a motivating example, let's recall the example of finding the MLE from $X_1, \cdots, X_n\mathcal{N}(\mu, \sigma^2)$ where μ and σ^2 are unknown. We know that

$$\widehat{\mu} = \overline{X}_n, \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

In probability and statistics, sometimes we are interested in the coefficient of variation (relative standard deviation) defined as

$$c_v = \frac{\sigma}{\mu}.$$

By equivariance, we know the MLE of c_v is

$$\widehat{c}_v = \frac{\widehat{\sigma}}{\widehat{\mu}}.$$

The question is: can we construct a $(1-\alpha)$ confidence interval for c_v?

This problem is closely related to finding the asymptotic distribution of $\widehat{c}_v$. We may ask whether $\widehat{c}_v$ satisfy certain asymptotic normality

$$\sqrt{n}(\widehat{c}_v - c_v) \xrightarrow{d} \mathcal{N}(0, \sigma_{c_v}^2)$$

for some variance $\sigma_{c_v}^2$? This requires us to find out what the joint distribution of $(\widehat{\mu}, \widehat{\sigma}^2)$, which is made more clearly later.

4.7.1 Multiparameter MLE

Suppose a family of distributions depends on several parameters $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)^\top$ (a column vector). By maximizing the likelihood function, we obtain the MLE:

$$\widehat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}).$$

Two questions:

- Is the MLE consistent?
- Does asymptotic normality hold?

Theorem 4.7.1. *The MLE is consistent under certain regularity condition.*

If you are interested in the rigorous proof of consistency, you may refer to advanced textbook such as [2].

Exercise: Can you generalize the argument in the single parameter scenario to multi-parameter scenario?

How about asymptotic normality of MLE? Note that the MLE is no longer a vector, we will naturally use multivariate normal distribution.

4.7.2 Bivariate normal distribution

We start with bivariate normal distribution. We say (X, Y) satisfies bivariate normal distribution $\mathcal{N}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ if its pdf is

$$f(x, y, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi\sqrt{\det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(x-\mu_X, y-\mu_Y)\boldsymbol{\Sigma}^{-1}(x-\mu_X, y-\mu_Y)^\top\right)$$

where

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}, \quad \boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_X^2\sigma_Y^2(1-\rho^2)} \begin{bmatrix} \sigma_Y^2 & -\rho\sigma_X\sigma_Y \\ -\rho\sigma_X\sigma_Y & \sigma_X^2 \end{bmatrix}$$

Here μ_X (σ_X^2) and μ_Y (σ_Y^2) are the mean (variance) of X and Y respectively. The parameter ρ is the correlation:

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X\sigma_Y}, \quad \text{Cov}(X, Y) = \rho\sigma_X\sigma_Y$$

which satisfies $|\rho| \leq 1$ by Cauchy-Schwarz inequality.

Exercise: Show that $\boldsymbol{\Sigma}$ is strictly positive definite.

Exercise: Verify that $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1} = \boldsymbol{I}_2$. Here $\boldsymbol{I}_2$ denotes 2×2 identity matrix.

Note that $\det(\boldsymbol{\Sigma}) = (1-\rho^2)\sigma_X^2\sigma_Y^2$ and

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{1-\rho^2} \begin{bmatrix} 1/\sigma_X^2 & -\rho/\sigma_X\sigma_Y \\ -\rho/\sigma_X\sigma_Y & 1/\sigma_Y^2 \end{bmatrix}$$

If written explicitly, the pdf is

$$f_{X,Y}(x, y; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi\sqrt{1-\rho^2}\sigma_X\sigma_Y} \cdot \exp\Big(-\frac{1}{2(1-\rho^2)}\Big[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y}\Big]\Big).$$

Example: If X and Y are two independent standard normal random variables, then their joint pdf is

$$f_{X,Y}(x,y) = \frac{1}{2\pi} \exp\left(-\frac{x^2+y^2}{2}\right).$$

This is essentially $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_2)$, i.e., $\mathcal{N}(0,0,1,1,0)$.

Example: No correlation implies independence for joint normal distribution. Suppose $(X,Y) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with zero correlation $\rho = 0$. Then it holds

$$\begin{aligned} f_{X,Y}(x,y;\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{1}{2\pi\sigma_X\sigma_Y} \exp\left(-\frac{1}{2}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y}\right]\right) \\ &= \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(-\frac{(x-\mu_X)^2}{2\sigma_X^2}\right) \cdot \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp\left(-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}\right) \\ &= f_X(x) f_Y(y). \end{aligned}$$

In other words, they are independent.

Question: Is $f_{X,Y}$ indeed a probability density function? To justify it, we need to show $f_{X,Y} \geq 0$ (obvious) and $\iint_{\mathbb{R}} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = 1$.

Proof: We will show that $\iint_{\mathbb{R}} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = 1$. To do this, we first introduce a few notations:

$$\boldsymbol{z} = \begin{bmatrix} x \\ y \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}$$

Then

$$\iint_{\mathbb{R}^2} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = \frac{1}{2\pi\sqrt{\det(\boldsymbol{\Sigma})}} \int_{\mathbb{R}^2} \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right) \mathrm{d}\boldsymbol{z}$$

where $\mathrm{d}\boldsymbol{z}$ equals $\mathrm{d}x\,\mathrm{d}y$. Now we perform a change of variable:

$$\boldsymbol{w} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{z}-\boldsymbol{\mu}), \quad \boldsymbol{\Sigma}^{1/2} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^\top = \boldsymbol{U}\begin{bmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{bmatrix}\boldsymbol{U}^\top$$

where $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ is the spectral decomposition (eigen-decomposition) of $\boldsymbol{\Sigma}$, i.e., $\boldsymbol{U}$ is orthogonal ($\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_2$) and

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}, \quad \lambda_1, \lambda_2 > 0$$

consists of all eigenvalues of $\boldsymbol{\Sigma}$. This change of variable maps $\mathbb{R}^2$ to $\mathbb{R}^2$ and is one-to-one. Now we substitute $\boldsymbol{z} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{w} + \boldsymbol{\mu}$ into the integral. Note that

$$\mathrm{d}\boldsymbol{z} = |\det(\boldsymbol{\Sigma}^{1/2})|\,\mathrm{d}\boldsymbol{w} = \sqrt{\lambda_1\lambda_2}\,\mathrm{d}\boldsymbol{w} = \sqrt{\det(\boldsymbol{\Sigma})}\,\mathrm{d}\boldsymbol{w}$$

where $\boldsymbol{\Sigma}^{1/2}$ is essentially the Jacobian matrix $\left[\frac{\partial z_i}{\partial w_j}\right]_{2\times 2}$ and

$$(\boldsymbol{z}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}) = \boldsymbol{w}^\top\boldsymbol{w} = w_1^2 + w_2^2$$

where $\boldsymbol{w} = (w_1, w_2)^\top$. Thus the integral equals

$$\begin{aligned} \iint_{\mathbb{R}^2} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y &= \frac{1}{2\pi\sqrt{\det(\boldsymbol{\Sigma})}} \int_{\mathbb{R}^2} \exp\left(-\frac{1}{2}\boldsymbol{w}^\top\boldsymbol{w}\right) \sqrt{\det(\boldsymbol{\Sigma})}\,\mathrm{d}\boldsymbol{w} \\ &= \frac{1}{2\pi} \int_{\mathbb{R}^2} \exp\left(-\frac{1}{2}\boldsymbol{w}^\top\boldsymbol{w}\right) \mathrm{d}\boldsymbol{w} \\ &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left(-\frac{1}{2}w_1^2\right) \mathrm{d}w_1 \cdot \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left(-\frac{1}{2}w_2^2\right) \mathrm{d}w_2 \\ &= 1. \end{aligned}$$

□

Essentially, the argument above proves the following statement.

Lemma 4.7.2. *Suppose* $\boldsymbol{Z} = (X, Y)^\top \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. *Then*

$$\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{Z} - \boldsymbol{\mu}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_2).$$

Exercise: Show that

$$\int_{\mathbb{R}} f_{X,Y}(x, y; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, \mathrm{d}y = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left(-\frac{(x-\mu_X)^2}{2\sigma_X^2}\right).$$

This exercise shows that the marginal distribution of X is still normal $\mathcal{N}(\mu_X, \sigma_X^2)$. For simplicity, you may try it by assuming $\mu_X = \mu_Y = 0$. Apply the similar technique used to verify that $f_{X,Y}$ is a pdf.

Exercise: Suppose $(X_i, Y_i), 1 \le i \le n$ are i.i.d. samples from $\mathcal{N}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$. Find the MLE for these five parameters.

Theorem 4.7.3. *Suppose* $(X, Y) \sim \mathcal{N}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$. *Then*

$$aX + bY \sim \mathcal{N}(a\mu_X + b\mu_Y, a^2\sigma_X^2 + 2ab\rho\sigma_X\sigma_Y + b^2\sigma_Y^2).$$

Proof: First note that $aX + bY$ is normal. Thus it suffices to determines its mean and variance.

$$\mathbb{E}(aX + bY) = a\,\mathbb{E}\,X + b\,\mathbb{E}\,Y = a\mu_X + b\mu_Y$$

and

$$\begin{aligned}
\mathrm{Var}(aX + bY) &= \mathbb{E}\left[a(X - \mathbb{E}\,X) + b(Y - \mathbb{E}\,Y)\right]\left[a(X - \mathbb{E}\,X) + b(Y - \mathbb{E}\,Y)\right] \\
&= a^2\,\mathbb{E}(X - \mathbb{E}\,X)^2 + 2ab\,\mathbb{E}(X - \mathbb{E}\,X)(Y - \mathbb{E}\,Y) + b^2\,\mathbb{E}(Y - \mathbb{E}\,Y)^2 \\
&= a^2\,\mathrm{Var}(X) + 2ab\,\mathrm{Cov}(X, Y) + b^2\,\mathrm{Var}(Y) \\
&= a^2\sigma_X^2 + 2ab\rho\sigma_X\sigma_Y + b^2\sigma_Y^2.
\end{aligned}$$

Thus we have the result. □

Question: Why is $aX + bY$ normal? We can use the moment generating function. For simplicity, we can let μ_X and μ_Y be zero. We let

$$\boldsymbol{\lambda} = \begin{bmatrix} a \\ b \end{bmatrix}, \quad \boldsymbol{z} = \begin{bmatrix} x \\ y \end{bmatrix} \quad \boldsymbol{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{z}.$$

Then

$$M(t) := \mathbb{E}\, e^{t(aX+bY)} = \mathbb{E}\, e^{t\boldsymbol{\lambda}^\top \boldsymbol{z}} = \frac{1}{2\pi\sqrt{\det(\boldsymbol{\Sigma})}} \int_{\mathbb{R}^2} e^{t\boldsymbol{\lambda}^\top \boldsymbol{z}} \exp\left(-\frac{1}{2}\boldsymbol{z}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{z}\right) \mathrm{d}\boldsymbol{z}$$

Note that $\boldsymbol{z} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{w}$ and thus $\mathrm{d}\boldsymbol{z} = |\det(\boldsymbol{\Sigma}^{1/2})|\, \mathrm{d}\boldsymbol{w}$.

$$\begin{aligned}
M(t) &:= \frac{1}{2\pi\sqrt{\det(\boldsymbol{\Sigma})}} \int_{\mathbb{R}^2} e^{t\boldsymbol{\lambda}^\top \boldsymbol{z}} \exp\left(-\frac{1}{2}\boldsymbol{z}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{z}\right) \mathrm{d}\boldsymbol{z} \\
&= \frac{1}{2\pi\sqrt{\det(\boldsymbol{\Sigma})}} \int_{\mathbb{R}^2} e^{t\boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{1/2}\boldsymbol{w}} \exp\left(-\frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w}\right) |\det(\boldsymbol{\Sigma}^{1/2})|\, \mathrm{d}\boldsymbol{w} \\
&= \frac{1}{2\pi} \int_{\mathbb{R}^2} \exp\left(-\frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + t\boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{1/2}\boldsymbol{w}\right) \mathrm{d}\boldsymbol{w} \\
&= \exp\left(-\frac{1}{2}\boldsymbol{\lambda}^\top \boldsymbol{\Sigma}\boldsymbol{\lambda}\right) \cdot \frac{1}{2\pi} \int_{\mathbb{R}^2} \exp\left(-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{\Sigma}^{1/2}\boldsymbol{\lambda})^\top (\boldsymbol{w} - \boldsymbol{\Sigma}^{1/2}\boldsymbol{\lambda})\right) \mathrm{d}\boldsymbol{w}
\end{aligned}$$

Note that the integral is 1 since it is a pdf, i.e., the pdf of $\mathcal{N}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\lambda}, \boldsymbol{I}_2)$. Thus

$$M(t) = \exp\left(-\frac{1}{2}\boldsymbol{\lambda}^\top \boldsymbol{\Sigma}\boldsymbol{\lambda}\right).$$

It is the mgf of $\mathcal{N}(0, \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}\boldsymbol{\lambda})$ where

$$\boldsymbol{\lambda}^\top \boldsymbol{\Sigma}\boldsymbol{\lambda} = a^2\sigma_X^2 + 2ab\rho\sigma_X\sigma_Y + b^2\sigma_Y^2.$$

4.7.3 Asymptotic normality of MLE

There is a natural extension of asymptotic normality from single parameter to multiple parameters scenario. Consider the likelihood function

$$\ell(\boldsymbol{\theta}; \boldsymbol{X}_1, \cdots, \boldsymbol{X}_n) = \sum_{i=1}^{n} \log f(\boldsymbol{X}_i; \boldsymbol{\theta})$$

where $\boldsymbol{X}_i \in \mathbb{R}^k$ are i.i.d. random variable (vectors). Define the Fisher information matrix as

$$I_n(\boldsymbol{\theta}) = -\left[\mathbb{E}_{\boldsymbol{\theta}} \frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\boldsymbol{\theta})\right] = -\begin{bmatrix} \mathbb{E}_{\boldsymbol{\theta}} \frac{\partial^2}{\partial\theta_1^2}\ell(\boldsymbol{\theta}) & \mathbb{E}_{\boldsymbol{\theta}} \frac{\partial^2}{\partial\theta_1\partial\theta_2}\ell(\boldsymbol{\theta}) & \cdots & \mathbb{E}_{\boldsymbol{\theta}} \frac{\partial^2}{\partial\theta_1\partial\theta_k}\ell(\boldsymbol{\theta}) \\ \mathbb{E}_{\boldsymbol{\theta}} \frac{\partial^2}{\partial\theta_1\partial\theta_2}\ell(\boldsymbol{\theta}) & \mathbb{E}_{\boldsymbol{\theta}} \frac{\partial^2}{\partial\theta_2^2}\ell(\boldsymbol{\theta}) & \cdots & \mathbb{E}_{\boldsymbol{\theta}} \frac{\partial^2}{\partial\theta_2\partial\theta_k}\ell(\boldsymbol{\theta}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}_{\boldsymbol{\theta}} \frac{\partial^2}{\partial\theta_1\partial\theta_k}\ell(\boldsymbol{\theta}) & \mathbb{E}_{\boldsymbol{\theta}} \frac{\partial^2}{\partial\theta_2\partial\theta_k}\ell(\boldsymbol{\theta}) & \cdots & \mathbb{E}_{\boldsymbol{\theta}} \frac{\partial^2}{\partial\theta_k^2}\ell(\boldsymbol{\theta}) \end{bmatrix}$$

Fisher information matrix is equal to the Hessian matrix of $-\ell(\boldsymbol{\theta}|\boldsymbol{X})$ under expectation. Is it always positive semidefinite?

Theorem 4.7.4 (Asymptotic normality of MLE). *Under certain regularity condition, it holds that*

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \xrightarrow{d} \mathcal{N}(0, I_n^{-1}(\boldsymbol{\theta}))$$

where $I_n^{-1}(\boldsymbol{\theta})$ is the inverse of Fisher information matrix.

If $\boldsymbol{X}_i \in \mathbb{R}^k$ are i.i.d. random vectors, then

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{I}^{-1}(\boldsymbol{\theta}))$$

where $\boldsymbol{I}(\boldsymbol{\theta})$ is given by

$$[\boldsymbol{I}(\boldsymbol{\theta})]_{ij} = -\,\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(\boldsymbol{X}; \boldsymbol{\theta})\right].$$

Obviously, $I_n(\boldsymbol{\theta}) = nI(\boldsymbol{\theta})$ follows from the linearity of expectation.

We will not give the proof here but the idea is similar to that of the single parameter scenario.

Example: Suppose $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. samples. Let's compute the Fisher information:

$$\log f(X; \mu, \sigma^2) = \log \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}} = -\frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(X-\mu)^2$$

The second order derivative of log-pdf is

$$\frac{\partial \ell}{\partial \mu} = -\frac{1}{\sigma^2}(\mu - X), \quad \frac{\partial \ell}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(X-\mu)^2$$

and

$$\frac{\partial^2 \ell}{\partial \mu^2} = -\frac{1}{\sigma^2}, \quad \frac{\partial^2 \ell}{\partial (\sigma^2)^2} = \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}(X-\mu)^2, \quad \frac{\partial^2 \ell}{\partial \mu \partial \sigma^2} = \frac{1}{\sigma^4}(\mu - X)$$

The Fisher information is

$$I(\mu, \sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}$$

and its inverse is

$$I(\mu, \sigma^2)^{-1} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix}$$

By asymptotic normality of MLE, we have

$$\sqrt{n} \begin{bmatrix} \widehat{\mu} - \mu \\ \widehat{\sigma}^2 - \sigma^2 \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix}\right)$$

where $\widehat{\mu} = \overline{X}_n$ and $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$.

Still by Slutsky theorem, we can replace σ^2 by $\widehat{\sigma}^2$. A $(1-\alpha)$ confidence interval for σ^2 is given by

$$|\widehat{\sigma}^2 - \sigma^2| \le z_{1-\alpha/2} \frac{\sqrt{2}\widehat{\sigma}^2}{\sqrt{n}}.$$

Exercise: Find the Fisher information matrix $I(\mu, \sigma)$ (not $I(\mu, \sigma^2)$). Derive the asymptotic distribution $\sqrt{n}(\widehat{\sigma} - \sigma)$. Then find a $(1-\alpha)$ confidence interval for σ.

4.7.4 Multiparameter Delta method

Theorem 4.7.5. *Suppose* $\nabla g(\boldsymbol{\theta}) = (\partial g/\partial \theta_1, \cdots, \partial g/\partial \theta_k)^\top$ *is not 0 at* $\boldsymbol{\theta}$*, then*

$$\sqrt{n}(g(\widehat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta})) \xrightarrow{d} \mathcal{N}(0, (\nabla g(\boldsymbol{\theta}))^\top I^{-1}(\boldsymbol{\theta}) \nabla g(\boldsymbol{\theta}))$$

where $\widehat{\boldsymbol{\theta}}$ *is the MLE of* $\boldsymbol{\theta}$*.*

Proof: The proof follows from

$$g(\widehat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta}) = \sum_{i=1}^k \frac{\partial g}{\partial \theta_i}(\boldsymbol{\theta})(\widehat{\theta}_i - \theta_i) + \text{error} = \langle \nabla g(\boldsymbol{\theta}), \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \rangle + \text{error}$$

where the error diminishes to 0 as $n \to \infty$. Here $\langle \cdot, \cdot \rangle$ denotes the inner product between two vectors. This is approximately Gaussian since $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normal $\mathcal{N}(0, I^{-1}(\boldsymbol{\theta}))$. The variance of $\sqrt{n}(g(\widehat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta}))$ is given by $(\nabla g(\boldsymbol{\theta}))^\top I^{-1}(\boldsymbol{\theta}) \nabla g(\boldsymbol{\theta})$. □

Example Consider $\tau = g(\mu, \sigma^2) = \sigma/\mu$ where the samples are drawn from Gaussian $X_i \sim \mathcal{N}(\mu, \sigma^2)$. The goal is to find out the asymptotic distribution of MLE of τ.

Note that the MLE of τ is given by $\widehat{\tau} = \widehat{\sigma}/\widehat{\mu}$ where $\widehat{\mu} = \overline{X}_n$ and $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$.

$$\frac{\partial g}{\partial \mu} = -\frac{\sigma}{\mu^2}, \quad \frac{\partial g}{\partial \sigma^2} = \frac{1}{2\mu\sigma}.$$

Thus

$$\sqrt{n}(\widehat{\tau}-\tau) \sim \mathcal{N}\left(0, \frac{\sigma^4}{\mu^4}+\frac{\sigma^2}{2\mu^2}\right)$$

where

$$(\nabla g(\boldsymbol{\theta}))^\top I^{-1}(\boldsymbol{\theta}) \nabla g(\boldsymbol{\theta}) = \frac{\sigma^2}{\mu^4}\cdot\sigma^2 + \frac{1}{4\mu^2\sigma^2}\cdot 2\sigma^4 = \frac{\sigma^4}{\mu^4}+\frac{\sigma^2}{2\mu^2}$$

4.7.5 Multiparameter normal distribution

We have discussed bivariate normal distribution. In practice, we often encounter multi-parameter normal distribution with dimension greater than 2. What is multiparameter normal distribution? A multi-parameter normal distribution is characterized by its mean vector $\boldsymbol{\mu} \in \mathbb{R}^k$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{k\times k}$. In particular, we require the covariance matrix $\boldsymbol{\Sigma}$ (symmetric) is positive definite.

Theorem 4.7.6. *A matrix $\boldsymbol{\Sigma}$ is positive definite, i.e., $\boldsymbol{\Sigma} \succ 0$, if one of the following equivalent statements is true:*

1. *All the eigenvalues of $\boldsymbol{\Sigma}$ are positive;*
2. *There exists a full-rank lower triangle matrix $\boldsymbol{L}$ such that $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top$;*
3. *$\boldsymbol{x}^\top\boldsymbol{\Sigma}\boldsymbol{x} > 0$ for any nonzero $\boldsymbol{x} \in \mathbb{R}^k$;*
4. *The determinant of all the leading principal submatrices are positive.*

In particular, if a matrix is 2×2, then

$$\boldsymbol{\Sigma} = \begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{12} & \sigma_{22}\end{bmatrix} \succeq 0 \Longleftrightarrow \sigma_{11} > 0,\ \sigma_{11}\sigma_{22} - \sigma_{12}^2 > 0.$$

Definition 4.7.1. *A random vector $\boldsymbol{X} = (X_1, \cdots, X_k)^\top$ satisfies multivariate normal $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if its probability density function is*

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(\sqrt{2\pi})^k\sqrt{\det\boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

We can see that the pdf involves the inverse of covariance matrix. Usually, finding the matrix inverse is tricky. In some special cases, the inverse is easy to obtain.

1. If a matrix $\boldsymbol{\Sigma}$ is 2×2, then its inverse is

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_{11}\sigma_{22}-\sigma_{12}^2}\begin{bmatrix}\sigma_{22} & -\sigma_{12}\\ -\sigma_{12} & \sigma_{11}\end{bmatrix}$$

2. If $\boldsymbol{\Sigma}$ is a diagonal matrix

$$\boldsymbol{\Sigma}_{ij} = \begin{cases}\sigma_{ii}, & i=j,\\ 0, & i\neq j,\end{cases}$$

then its inverse is

$$[\boldsymbol{\Sigma}^{-1}]_{ij} = \begin{cases}\sigma_{ii}^{-1}, & i=j,\\ 0, & i\neq j,\end{cases}$$

In fact, we have seen multivariate normal distribution before. For example, if $X_1, \cdots, X_k$ are independent standard normal random variables, then the random vector $\boldsymbol{X} = (X_1, \cdots, X_k)^\top$ has a joint distribution

$$\begin{aligned} f(\boldsymbol{x}; 0; \boldsymbol{I}_n) &= \frac{1}{\sqrt{2\pi}^k \sqrt{\det(\boldsymbol{I}_n)}} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{I}_n^{-1} \boldsymbol{x}\right) \\ &= \frac{1}{\sqrt{2\pi}^k} \exp\left(-\frac{1}{2}\|\boldsymbol{x}\|^2\right) \\ &= \prod_{i=1}^k \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_i^2}{2}\right) \end{aligned}$$

which is essentially $\mathcal{N}(0, \boldsymbol{I}_k)$ where $\boldsymbol{I}_k$ is the $k \times k$ identity matrix.

Exercise: Verify that $f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is indeed a pdf, i.e.,

$$\int_{\boldsymbol{x} \in \mathbb{R}^k} f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \, \mathrm{d}\boldsymbol{x} = 1.$$

Exercise: If $\boldsymbol{\Sigma}$ is diagonal, then all $X_1, \cdots, X_k$ are independent.

Exercise: If $\sigma_{ij} = 0$, then X_i and X_j are independent. In other words, if (X_i, X_j) is jointly normal and $\mathrm{Cov}(X_i, X_j) = 0$, then X_i and X_j are independent.

Exercise: Show that

$$\mathbb{E}\, \boldsymbol{X} = \boldsymbol{\mu}, \quad \mathbb{E}(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\top = \boldsymbol{\Sigma}.$$

In other words, $\Sigma_{ij} = \mathbb{E}(X_i - \mu_i)(X_j - \mu_j)$ is the covariance between X_i and X_j and Σ_{ii} is the variance of X_i.

An important property of joint normal distribution is that: the linear combination $\sum_{i=1}^n a_i X_i$ is still normal for any deterministic $\boldsymbol{a} = (a_1, \cdots, a_k)^\top$. How to find its distribution? Since it is normal, we only need to compute its mean and variance.

Theorem 4.7.7. *Suppose $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\sum_{i=1}^k a_i X_i$ obeys*

$$\sum_{i=1}^k a_i X_i \sim \mathcal{N}\left(\sum_{i=1}^k a_i \mu_i, \sum_{i,j} \Sigma_{ij} a_i a_j\right)$$

or equivalently,

$$\langle \boldsymbol{a}, \boldsymbol{X} \rangle \sim \mathcal{N}\left(\langle \boldsymbol{a}, \boldsymbol{\mu} \rangle, \boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a}\right)$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of two vectors.

Proof: It suffices to compute its mean and variance:

$$\mathbb{E} \sum_{i=1}^k a_i X_i = \sum_{i=1}^k a_i \, \mathbb{E}\, X_i = \sum_{i=1}^k a_i \mu_i.$$

Its variance is

$$\begin{aligned}\operatorname{Var}\left(\sum_{i=1}^k a_i X_i\right) &= \mathbb{E}\left(\sum_{i=1}^k a_i(X_i - \mathbb{E}\,X_i)\sum_{j=1}^k a_j(X_j - \mathbb{E}\,X_j)\right)\\ &= \sum_{i=1}^k\sum_{j=1}^k a_i a_j\,\mathbb{E}(X_i - \mathbb{E}\,X_i)(X_j - \mathbb{E}\,X_j)\\ &= \sum_{i=1}^k\sum_{j=1}^k a_i a_j\operatorname{Cov}(X_i, X_j) = \sum_{i=1}^k\sum_{j=1}^k a_i a_j\Sigma_{ij} = \boldsymbol{a}^\top\boldsymbol{\Sigma}\boldsymbol{a}.\end{aligned}$$

□

Exercise: Moment-generating function for $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. What is the moment generating function for $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,

$$M(\boldsymbol{t}) = \mathbb{E}\,e^{\sum_{i=1}^k t_i X_i}$$

where $\boldsymbol{t} = (t_1, \cdots, t_k)^\top$? Hints: by definition, it holds

$$\begin{aligned}M(\boldsymbol{t}) &= \mathbb{E}\exp(\boldsymbol{t}^\top\boldsymbol{X})\\ &= \frac{1}{(\sqrt{2\pi})^{n/2}\sqrt{\det(\boldsymbol{\Sigma})}}\int_{\mathbb{R}^n}\exp\left(\boldsymbol{t}^\top\boldsymbol{x} - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)\mathrm{d}\boldsymbol{x}.\end{aligned}$$

Let's perform a change of variable: $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu})$. Then

$$\mathrm{d}\boldsymbol{x} = \mathrm{d}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{z} + \boldsymbol{\mu}) = |\det(\boldsymbol{\Sigma}^{1/2})|\,\mathrm{d}\boldsymbol{z} = \sqrt{\det(\boldsymbol{\Sigma})}\,\mathrm{d}\boldsymbol{z}.$$

Then

$$\begin{aligned}M(\boldsymbol{t}) &= \frac{1}{(\sqrt{2\pi})^{n/2}\sqrt{\det(\boldsymbol{\Sigma})}}\int_{\mathbb{R}^n}\exp\left(\boldsymbol{t}^\top(\boldsymbol{\Sigma}^{1/2}\boldsymbol{z} + \boldsymbol{\mu}) - \frac{1}{2}\boldsymbol{z}^\top\boldsymbol{z}\right)\sqrt{\det(\boldsymbol{\Sigma})}\,\mathrm{d}\boldsymbol{z}\\ &= \frac{\exp(\boldsymbol{t}^\top\boldsymbol{\mu})}{(\sqrt{2\pi})^{n/2}}\int_{\mathbb{R}^n}\exp\left(\boldsymbol{t}^\top\boldsymbol{\Sigma}^{1/2}\boldsymbol{z} - \frac{1}{2}\boldsymbol{z}^\top\boldsymbol{z}\right)\mathrm{d}\boldsymbol{z}\\ &= \exp\left(\boldsymbol{t}^\top\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t}\right)\cdot\frac{1}{(\sqrt{2\pi})^{n/2}}\int_{\mathbb{R}^n}\exp\left(-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{\Sigma}^{1/2}\boldsymbol{t})^\top(\boldsymbol{z} - \boldsymbol{\Sigma}^{1/2}\boldsymbol{t})\right)\mathrm{d}\boldsymbol{z}\\ &= \exp\left(\boldsymbol{t}^\top\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t}\right)\end{aligned}$$

In fact, multivariate normal distribution is still multivariate norm under linear transform.

Lemma 4.7.8. *Suppose $\boldsymbol{A}$ is any deterministic matrix in $\mathbb{R}^{l\times k}$. Then*

$$\boldsymbol{A}\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)$$

for $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

It suffices to compute the mean and covariance. Note that the ith entry of $\boldsymbol{A}\boldsymbol{X}$ in expectation is

$$\mathbb{E}[\boldsymbol{A}\boldsymbol{X}]_i = \mathbb{E}\sum_{j=1}^k a_{ij}X_j = \sum_{j=1}^k a_{ij}\,\mathbb{E}\,X_j = \sum_{j=1}^k a_{ij}\mu_j = [\boldsymbol{A}\boldsymbol{\mu}]_i, \quad 1 \le i \le l.$$

Thus $\mathbb{E}\,\boldsymbol{AX} = \boldsymbol{A}\,\mathbb{E}\,\boldsymbol{X}$. For the covariance, by definition, we have

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{AX}) &= \mathbb{E}(\boldsymbol{AX} - \mathbb{E}\,\boldsymbol{AX})(\boldsymbol{AX} - \mathbb{E}\,\boldsymbol{AX})^\top \\ &= \mathbb{E}\,\boldsymbol{A}(\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X})(\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X})^\top \boldsymbol{A}^\top \\ &= \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top.\end{aligned}$$

For a special case, if $\boldsymbol{A}$ is orthogonal, i.e. $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{I}_k$, then for $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k)$, then

$$\boldsymbol{AX} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{I}_k\boldsymbol{A}^\top) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k).$$

Exercise: Show that the variance of $\boldsymbol{a}^\top\boldsymbol{X}$ (subject to $\|\boldsymbol{a}\| = 1$) is maximized if $\boldsymbol{a}$ is leading eigenvector of $\boldsymbol{\Sigma}$, and the variance is largest eigenvalue.

4.7.6 Independence between sample mean and variance

Recall that for $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$

$$\sqrt{n}\begin{bmatrix}\widehat{\mu} - \mu \\ \widehat{\sigma}^2 - \sigma^2\end{bmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{bmatrix}\sigma^2 & 0 \\ 0 & 2\sigma^4\end{bmatrix}\right)$$

It seems that $\widehat{\mu} = \overline{X}_n$ and $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ are "near" independent. Is it true?

Theorem 4.7.9. *The sample mean $\widehat{\mu} = \overline{X}_n$ and variance $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ are independent. Moreover,*

$$n\widehat{\sigma}^2 = \sum_{i=1}^n (X_i - \overline{X}_n)^2 \sim \sigma^2\chi^2_{n-1}.$$

But how to justify this theorem? Now we let $\boldsymbol{X} = (X_1, \cdots, X_n)^\top \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, i.e.,

$$f_{\boldsymbol{X}}(\boldsymbol{x}; 0, \boldsymbol{I}_n) = \frac{1}{(2\pi)^{n/2}}\left(-\frac{1}{2}\sum_{i=1}^n x_i^2\right).$$

We define a vector $\boldsymbol{v}$ and a matrix $\boldsymbol{P} \in \mathbb{R}^{n\times n}$:

$$\boldsymbol{v} = \begin{bmatrix}1 \\ \vdots \\ 1\end{bmatrix} \in \mathbb{R}^n, \quad \boldsymbol{P} = \boldsymbol{I}_n - \frac{1}{n}\boldsymbol{v}\boldsymbol{v}^\top = \begin{bmatrix}\frac{n-1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & \frac{n-1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & \frac{n-1}{n}\end{bmatrix}.$$

Exercise: Show that $\boldsymbol{P}\boldsymbol{v} = 0$ and $\boldsymbol{P}^2 = \boldsymbol{P}$. ($\boldsymbol{P}$ is called projection matrix).

Exercise: The eigenvalues of $\boldsymbol{P}$ is 1 with multiplicities $n - 1$ and 0 with multiplicity 1.

With $\boldsymbol{u}$ and $\boldsymbol{P}$, we have

$$\begin{aligned}
\overline{X}_n &= \frac{1}{n}\boldsymbol{v}^\top \boldsymbol{X} = \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{v} \\
\widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}_n^2\right) \\
&= \frac{1}{n}\left(\boldsymbol{X}^\top \boldsymbol{X} - \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{v}\boldsymbol{v}^\top \boldsymbol{X}\right) \\
&= \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{P}\boldsymbol{X}.
\end{aligned}$$

Assume $\boldsymbol{U} \in \mathbb{R}^{n\times(n-1)}$ consists of $n-1$ orthonormal eigenvectors of $\boldsymbol{P}$ w.r.t. the eigenvalue 1. Then we know that $\boldsymbol{U}^\top \boldsymbol{v} = 0$ and moreover $\boldsymbol{P} = \boldsymbol{U}\boldsymbol{U}^\top$, i.e.,

$$\boldsymbol{P} = \sum_{i=1}^{n-1} \boldsymbol{u}_i \boldsymbol{u}_i^\top, \quad \boldsymbol{u}_i \perp \boldsymbol{u}_j, i \neq j$$

where $\boldsymbol{u}_i$ is the ith column of $\boldsymbol{U}$. Also $\boldsymbol{u}_i \perp \boldsymbol{v}$ holds since they belong to eigenvectors w.r.t. different eigenvectors.

Now

$$\widehat{\sigma}^2 = \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{X} = \frac{1}{n}\|\boldsymbol{U}^\top \boldsymbol{X}\|^2 = \frac{1}{2}\sum_{i=1}^{n-1} |\boldsymbol{u}_i^\top \boldsymbol{X}|^2$$

where $\boldsymbol{U}^\top \boldsymbol{X} \in \mathbb{R}^{n-1}$.

Key: If we are able to show that $\boldsymbol{v}^\top \boldsymbol{X}$ and $\boldsymbol{U}^\top \boldsymbol{X}$ are independent, then $\overline{X}_n$ and $\widehat{\sigma}^2$ are independent.

What is the joint distribution of $\boldsymbol{v}^\top \boldsymbol{X}$ and $\boldsymbol{U}^\top \boldsymbol{X}$? Consider

$$\boldsymbol{\Pi} = \begin{bmatrix} n^{-1/2}\boldsymbol{v}^\top \\ \boldsymbol{U}^\top \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

The term $n^{-1/2}$ is to ensure that $\|n^{-1/2}\boldsymbol{v}\| = 1$. By linear invariance of normal distribution, $\boldsymbol{\Pi}\boldsymbol{X}$ is also jointly normal. It is not hard to see that

$$\boldsymbol{\Pi}\boldsymbol{\Pi}^\top = \begin{bmatrix} n^{-1}\boldsymbol{v}^\top \boldsymbol{v} & -n^{-1/2}\boldsymbol{v}^\top \boldsymbol{U} \\ -n^{-1/2}\boldsymbol{U}\boldsymbol{v}^\top & \boldsymbol{U}^\top \boldsymbol{U} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \boldsymbol{I}_{n-1} \end{bmatrix} = \boldsymbol{I}_n.$$

In other words, the covariance matrix of $(n^{-1/2}\boldsymbol{v}^\top \boldsymbol{X}, \boldsymbol{U}^\top \boldsymbol{X})$ equals

$$\Sigma_{\boldsymbol{\Pi}\boldsymbol{X}} = \boldsymbol{\Pi}\,\mathrm{Cov}(\boldsymbol{X})\boldsymbol{\Pi}^\top = \sigma^2 \boldsymbol{\Pi}\boldsymbol{I}_n\boldsymbol{\Pi}^\top = \sigma^2 \boldsymbol{I}_n.$$

which implies that $\boldsymbol{v}^\top \boldsymbol{X}$ and $\boldsymbol{U}^\top \boldsymbol{X}$ are independent. Now let's look at the distribution of $[\boldsymbol{U}^\top \boldsymbol{X}]_i = \boldsymbol{u}_i^\top \boldsymbol{X}$:

$$\boldsymbol{u}_i^\top \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{u}_i^\top \cdot \mu\boldsymbol{v}, \sigma^2 \boldsymbol{u}_i^\top \boldsymbol{I}_n \boldsymbol{u}_i) = \mathcal{N}(0, \sigma^2)$$

where $\mathbb{E}\,\boldsymbol{X} = \mu\boldsymbol{v}$ and $\boldsymbol{u}_i \perp \boldsymbol{v}$. Therefore, $\boldsymbol{u}_i^\top \boldsymbol{X}/\sigma$ are independent standard normal and we have

$$\boldsymbol{X}^\top \boldsymbol{P}\boldsymbol{X} = \sum_{i=1}^{n-1} [\boldsymbol{U}^\top \boldsymbol{X}]_i^2 \sim \sigma^2 \chi_{n-1}^2.$$

Chapter 5

Hypothesis testing

5.1 Motivation

In many applications, we are often facing questions like these:

1. Motivation: In 1000 tosses of a coin, 560 heads and 440 tails appear. Is the coin fair?
2. Whether two datasets come from the same distribution? Do the data satisfy normal distribution?
3. Clinical trials: does the medicine work well for one certain type of disease?

These questions are called hypothesis testing problem.

5.1.1 Hypothesis

The first question is: what is a hypothesis?

Definition 5.1.1. *A hypothesis is a statement about a population parameter.*

For example, $X_1, \cdots, X_n$

The two complementary hypothesis in a hypothesis testing problem are

- the null hypothesis, denoted by H_0
- the alternative hypothesis, denoted by H_1

Example: Suppose $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ with known σ^2. We are interested in testing if $\mu = \mu_0$. The hypothesis $H_0 : \mu = \mu_0$ is called the null hypothesis, i.e.,

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu \neq \mu_0.$$

Example: Suppose $X_1, \cdots, X_n$ are outcomes from Bernoulli(θ), which is a natural model of coin tossing. If we want to know whether the coin is fair, we are essentially testing:

$$H_0 : \theta = \frac{1}{2}, \qquad H_1 : \theta \neq \frac{1}{2}.$$

In statistical practice, we usually treat these two hypotheses unequally. When we perform a testing, we actually design a procedure to decide if we should reject the null hypothesis

or retain (not to reject) the null hypothesis. It is important to note that rejecting the null hypothesis does not mean we should accept the alternative hypothesis.

Now, let's discuss how to design a test procedure. Apparently, this procedure will depend on the observed data $X_1, \cdots, X_n$. We focus on the example in which $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ and want to test if $\mu = 0$. Naturally, we can first obtain an estimator of μ from the data and see if it is close to $\mu = 0$ (compared with the standard deviation as well).

1. Compute the sample average $T(X) = \overline{X}_n$.
2. If $T(X)$ is far away from $\mu = 0$, we should reject H_0; if $T(X)$ is close to $\mu = 0$, we choose not to reject H_0. Namely, we reject H_0 if

$$|\overline{X}_n - \mu_0| \geq c$$

for some properly chosen c.

5.1.2 Test statistics and rejection region

This leads to two concepts: test statistics and rejection region. Let $X_1, \cdots, X_n \sim f(x; \theta)$, and $T(X)$ is a statistic. Suppose we have designed a decision rule: reject the H_0 if $T(X) \in R$ where R is a region, then $T(X)$ is called the test statistic and R is the rejection region.

Example: In the example of testing $\mu \neq 0$ for $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$,

$$T(X) = \overline{X}_n, \quad R = \{x : |x - \mu_0| \geq c\}.$$

However, is it possible that $T(X)$ and the choice of R give you a wrong answer?

5.1.3 Type I and II error

There are two types of errors, often called Type I and Ii error.

- Type I error: we reject H_0 but H_0 is the truth.
- Type II error: we retain H_0 but H_1 is the truth.

Table 5.1: Summary of outcomes of hypothesis testing

	Retain Null H_0	Reject Null H_0
H_0 true	✓	Type I error
H_1 true	Type II error	✓

How to control the error level?

Definition 5.1.2. *The power function of a test with rejection region R is*

$$\beta(\theta) = \mathbb{P}_\theta(T(X) \in R)$$

where $X_1, \cdots, X_n$ are samples from $f(x; \theta)$.

Remark: The power function $\beta(\theta_0)$ is the probability of rejecting $\theta = \theta_0$.

Example - continued: For $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ with unknown σ^2. Let's compute the power function:

$$\begin{aligned}
\beta(\mu) &= \mathbb{P}_\mu(|\overline{X}_n - \mu_0| \geq c) \\
&= \mathbb{P}_\mu(\overline{X}_n - \mu_0 \geq c) + \mathbb{P}_\mu(\overline{X}_n - \mu_0 \leq -c) \\
&= \mathbb{P}_\mu(\overline{X}_n - \mu \geq c - \mu + \mu_0) + \mathbb{P}_\mu(\overline{X}_n - \mu \leq -c - \mu + \mu_0) \\
&= \mathbb{P}_\mu\left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \geq \frac{\sqrt{n}(c - \mu + \mu_0)}{\sigma}\right) \\
&\quad + \mathbb{P}_\mu\left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \leq \frac{\sqrt{n}(-c - \mu + \mu_0)}{\sigma}\right) \\
&= 1 - \Phi\left(\frac{\sqrt{n}(c - \mu + \mu_0)}{\sigma}\right) + \Phi\left(\frac{\sqrt{n}(-c - \mu + \mu_0)}{\sigma}\right)
\end{aligned}$$

where $\Phi(\cdot)$ is the cdf of standard normal distribution.

How to quantify Type-I error?

$$\begin{aligned}
\beta(\mu_0) &= 1 - \Phi\left(\frac{\sqrt{n}c}{\sigma}\right) + \Phi\left(-\frac{\sqrt{n}c}{\sigma}\right) \\
&= 2\left(1 - \Phi\left(\frac{\sqrt{n}c}{\sigma}\right)\right)
\end{aligned}$$

where $\Phi(x)+\Phi(-x) = 1$ for any $x > 0$. To make the Type-I error under α, we require

$$c = \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}.$$

How about Type-II error? By definition, Type-II error is the probability of retaining the null (not rejecting the null, i.e., $T(X) \notin R$) while the alternative is true. Suppose the true parameter is $\mu_A \neq \mu_0$, then

$$\text{Type-II error at } \mu_A = 1 - \beta(\mu_A).$$

Is it possible to control both Type I and II error? It might be tricky sometimes. Here since we don't know the actual true parameter, the conservative way to control the Type-II error is to find a uniform bound of the Type-II error for any $\mu \neq \mu_0$:

$$\sup_{\mu \neq \mu_0}(1 - \beta(\mu)) = 1 - \beta(\mu_0)$$

which is actually given by $\mu = \mu_0$. In other words, we cannot make both Type-I and Type II simultaneously small in this case. In fact, in most testing problem, the asymmetry between H_0 and H_1 is natural. We usually put a tighter control on the more serious error such as Type-I error.

Exercise: Show that $\sup_{\mu \neq \mu_0}(1 - \beta(\mu)) = 1 - \beta(\mu_0)$.

How to control the Type-I error? Sometimes, the parameter space Θ_0 of H_0 is not a singleton (e.g. $H_0 : \theta = \theta_0$). To overcome this issue, we introduce the size of a test:

Definition 5.1.3 (Size of a test)**.** *The size of a test is defined to be*

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$$

where Θ_0 consists of all parameters in the null hypothesis.

A test is said to have level α if its size is less than or equal to α.

- The size of a test is the maximal probability of rejecting the null hypothesis when the null hypothesis is true.
- If the level α is small, it means type I error is small.

Example - continued: If $\Theta_0 = \{\mu_0\}$, then the size equals $\beta(\mu_0)$. To make the size smaller than a given α, we need to have

$$\beta(\mu_0) = 2\left(1 - \Phi\left(\frac{\sqrt{n}c}{\sigma}\right)\right) = \alpha$$

which gives

$$\frac{\sqrt{n}c}{\sigma} = z_{1-\alpha/2} \Longleftrightarrow \sqrt{n}c = z_{1-\alpha/2}\sigma.$$

Given the number of samples n, σ, and size α, we can determine c such that the Type I error is at most α :

$$c = \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}.$$

In other words, we reject the null hypothesis $H_0 : \mu = \mu_0$ if

$$|\overline{X}_n - \mu_0| \geq \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}} \Longleftrightarrow \frac{\sqrt{n}|\overline{X}_m - \mu_0|}{\sigma} \geq z_{1-\alpha/2}.$$

This is called z-test, a test for the mean of a distribution.

Example: t-test Consider $X_1, \cdots, X_n$ are samples from $\mathcal{N}(\mu, \sigma^2)$ with unknown μ and σ^2. Consider

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu \neq \mu_0.$$

We may choose the following estimator

$$T(X) = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{S_n}$$

since it is asymptotically normal distribution if n is large and the null is true. The rejection region could be

$$R = \{X : |T(X)| \geq z_{1-\alpha/2}\}.$$

What if n is small? In fact, $T_n(X)$ satisfies t-distribution of degree $n-1$.

Student-t distribution: The following random variable

$$T = \frac{Z}{\sqrt{Y/n}}$$

satisfies t-distribution of degree n if

- Z is a standard normal random variable;
- Y is a χ^2_n random variable with degree n; if $Y_1, \cdots, Y_n$ are i.i.d. $\mathcal{N}(0,1)$, then $Y = \sum_{i=1}^n Y_i^2$ is χ^2_n, which is basically $\Gamma(n/2, 1)$ distribution;
- Y and Z are independent.

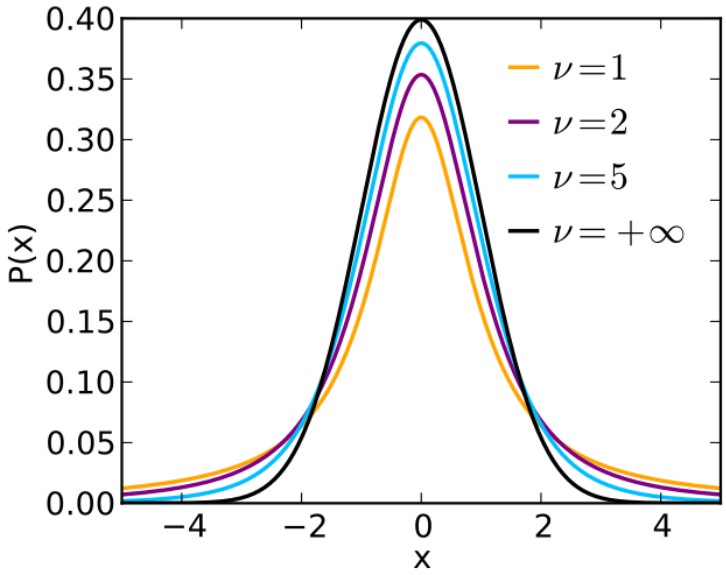


Figure 5.1: The pdf of Student-t distribution. Source: wikipedia

Exercise: Verify that

$$T(X) = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{S_n}$$

satisfies t_{n-1}.

What is the pdf of t-distribution of degree n?

$$f_T(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

where $\Gamma(\alpha)$ is called Gamma function:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,\mathrm{d}x.$$

In particular, $\Gamma(1/2) = \sqrt{\pi}$. If $\alpha = n$ for some positive integer n, $\Gamma(n) = (n-1)!$

- if $n = 1$, T is Cauchy distribution:

$$f_T(x) = \frac{1}{\pi}\cdot\frac{1}{1+x^2}.$$

- if $n \to \infty$, we have

$$f_T(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}.$$

Exercise: Show that as $n \to \infty$,

$$t_n \xrightarrow{d} \mathcal{N}(0,1): \quad \lim_{n\to\infty}\frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}.$$

It is an interpolation between Cauchy and Gaussian distribution.

Thus we can reject $H_0 : \theta = \theta_0$

$$|T_n(X)| > t_{n-1,1-\alpha/2}$$

with size α (type-I error) where $t_{n-1,1-\alpha/2}$ is the $1-\alpha/2$ quantile of Student-t distribution of degree $n-1$.

5.2 More on hypothesis testing

5.2.1 Composite hypothesis testing

So far, we have discussed simple null hypothesis, i.e., $|\Theta_0| = 1$. On the other hand, we often encounter *composite* hypothesis, the parameter space Θ_0 contains multiple or even infinitely many parameters.

Example: Let $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ where σ is known. We want to test

$$H_0 : \mu \leq \mu_0, \qquad H_1 : \mu > \mu_0$$

Hence

$$\Theta_0 = (-\infty, \mu_0] \text{ versus } \Theta_1 = (\mu_0, \infty).$$

Note that

$$T(X) = \overline{X}_n$$

is the MLE of μ. We reject H_0 if $T(X) > c$ where c is a number.

We reject H_0 if $T(X) > c$. The power function is

$$\begin{aligned} \beta(\mu) &= \mathbb{P}_\mu(T(X) > c) = \mathbb{P}_\mu(\overline{X}_n > c) \\ &= \mathbb{P}_\mu\left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} > \frac{\sqrt{n}(c-\mu)}{\sigma}\right) \\ &= 1 - \Phi\left(\frac{\sqrt{n}(c-\mu)}{\sigma}\right) \end{aligned}$$

What is the size? Note that $\beta(\mu)$ is increasing!

$$\sup_{\mu \leq \mu_0} \beta(\mu) = \beta(\mu_0) = 1 - \Phi\left(\frac{\sqrt{n}(c-\mu_0)}{\sigma}\right).$$

To have a size α test, we set $\beta(\mu_0) = \alpha$:

$$c = \mu_0 + \frac{\sigma z_{1-\alpha}}{\sqrt{n}}.$$

In other words, the Type-I error is at most α if we reject $\overline{X}_n \geq \mu_0 + \frac{\sigma z_{1-\alpha}}{\sqrt{n}}$.

5.2.2 Wald test

Consider testing

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

Assume that $\widehat{\theta}_n$ is a consistent estimator of θ and is asymptotically normal:

$$W := \frac{\widehat{\theta}_n - \theta_0}{\widehat{\text{se}}(\widehat{\theta}_n)} \sim \mathcal{N}(0, 1)$$

where θ_0 is the true parameter and $\widehat{\text{se}}(\widehat{\theta}_n)$ is an estimation of the standard deviation of $\widehat{\theta}_n$. Then a size-α Wald test is: reject H_0 if $|W| \geq z_{1-\frac{\alpha}{2}}$.

This can be extended to one-sided hypothesis as well:

1. For $H_0 : \theta < \theta_0$ and $H_1 : \theta \geq \theta_0$, we reject the null hypothesis if

$$\frac{\widehat{\theta}_n - \theta_0}{\widehat{\mathrm{se}}(\widehat{\theta}_n)} > z_{1-\alpha}$$

2. For $H_0 : \theta > \theta_0$ and $H_1 : \theta \leq \theta_0$, we reject the null hypothesis if

$$\frac{\widehat{\theta}_n - \theta_0}{\widehat{\mathrm{se}}(\widehat{\theta}_n)} < -z_{1-\alpha}$$

Exercise: Show that the size is approximately α for the two one-sided hypothesis testing problems above.

Why? Let's focus on the first one-sided hypothesis. We reject the null hypothesis if $\widehat{\theta}_n > c$ for some c. How to compute its power function?

$$\begin{aligned}
\beta(\theta) &= \mathbb{P}_\theta\left(\widehat{\theta}_n \geq c\right) \\
&= \mathbb{P}_\theta\left(\frac{\widehat{\theta}_n - \theta}{\widehat{\mathrm{se}}(\widehat{\theta}_n)} \geq \frac{c - \theta}{\widehat{\mathrm{se}}(\widehat{\theta}_n)}\right) \\
&\approx 1 - \Phi\left(\frac{c - \theta}{\widehat{\mathrm{se}}(\widehat{\theta}_n)}\right).
\end{aligned}$$

Note that the power function is an increasing function of θ; thus the maximum is assumed when $\theta = \theta_0$.

$$\sup_{\theta \in \Theta_0} \beta(\theta) = 1 - \Phi\left(\frac{c - \theta_0}{\widehat{\mathrm{se}}(\widehat{\theta}_n)}\right) = \alpha$$

which gives

$$c = \theta_0 + z_{1-\alpha}\widehat{\mathrm{se}}(\widehat{\theta}_n).$$

Thus we reject H_0 if

$$\widehat{\theta}_n \geq c \Longleftrightarrow \frac{\widehat{\theta}_n - \theta_0}{\widehat{\mathrm{se}}(\widehat{\theta}_n)} > z_{1-\alpha}.$$

Example: Suppose $X_1, \cdots, X_n$ are samples from Bernoulli(θ). We want to test if

$$H_0 : \theta = \theta_0 \qquad H_1 : \theta \neq \theta_0.$$

Equivalently, we reject $H_0 : \theta = \theta_0$ if

$$\left| \frac{\overline{x}_n - \theta_0}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}} \right| > z_{1-\frac{\alpha}{2}}.$$

where $\overline{x}_n$ is the observed value of $\overline{X}_n$.

First, the MLE of θ is $\overline{X}_n$. Suppose θ_0 is the true parameter, then by CLT, we have

$$\frac{\sqrt{n}(\overline{X}_n - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}} \xrightarrow{d} \mathcal{N}(0, 1).$$

We reject the null hypothesis if

$$\left| \frac{\sqrt{n}(\overline{X}_n - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}} \right| >_{1-\alpha/2} .$$

Another alternative test statistic is

$$T(X) = \frac{\sqrt{n}(\overline{X}_n - \theta_0)}{\sqrt{\overline{X}_n(1-\overline{X}_n)}},$$

and we reject the null if $|T(X)| > z_{1-\alpha/2}$.

Recall the example: if we observe 560 heads and 440 tails, is the coin fair?

In this case, we let $\theta_0 = \frac{1}{2}$, $n = 1000$, $\alpha = 0.05$, and $z_{0.975} = 1.96$. Suppose the null is true, then

$$T_1(X) = \frac{\sqrt{n}(\overline{X}_n - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}} = 3.7947 > z_{0.975} = 1.96.$$

We can conclude that with size 0.05, we reject the null hypothesis. If we choose the alternative statistics, the same conclusion holds

$$\frac{\sqrt{n}(\overline{X}_n - \theta_0)}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} = 3.8224 > z_{0.975}.$$

Question: What is the connection between confidence interval and rejection region?

The confidence interval is closely related to the rejection region. The size-α Wald test rejects $H_0 : \theta = \theta_0$ v.s. $H_1 : \theta \neq \theta_0$ if and only if $\theta_0 \notin C$ where

$$C = \left(\widehat{\theta}_n - z_{1-\frac{\alpha}{2}} \cdot \widehat{\text{se}}(\widehat{\theta}_n), \quad \widehat{\theta}_n + z_{1-\frac{\alpha}{2}} \cdot \widehat{\text{se}}(\widehat{\theta}_n) \right).$$

which is equivalent to

$$|\widehat{\theta}_n - \theta_0| \geq z_{1-\frac{\alpha}{2}} \cdot \widehat{\text{se}}(\widehat{\theta}_n).$$

Thus, testing the null hypothesis is equivalent to checking whether the null value is in the confidence interval for this simple hypothesis. When we reject H_0, we say that the result is statistically significant.

5.2.3 p-value

What is p-value? Let's first give the formal definition of p-value and then discuss its meaning.

Definition 5.2.1 (p-value)**.** *Suppose that for every $\alpha \in (0,1)$, we have a size α test with rejection region R_α. Then*

$$p\text{-}value = \inf\{\alpha : T(x) \in R_\alpha\}$$

where $T(x)$ is the observed value of $T(X)$. The p-value is the smallest level at which we can reject H_0. The smaller α is, the smaller the rejection region is.

The definition of p-value does not look obvious at the first glance. We try to do a concrete example.

Example: Suppose $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. Consider the testing problem $H_0 : \mu = \mu_0$ v.s. $H_1 : \mu \neq \mu_0$. We reject the null hypothesis with size α if

$$\left| \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\sigma} \right| > z_{1-\alpha/2} \Longleftrightarrow |\overline{X}_n - \mu_0| > \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}.$$

As we can see that, if α decreases, then $z_{1-\alpha/2}$ increases to infinity and the rejection region shrinks. Now suppose we observe the data and calculate the test statistic: $T(x) = \overline{x}_n$. We try to find the smallest possibly α^* such that the rejection region includes $\overline{x}_n$, i.e.,

$$|\overline{x}_n - \mu_0| = \frac{z_{1-\alpha^*/2}\sigma}{\sqrt{n}}$$

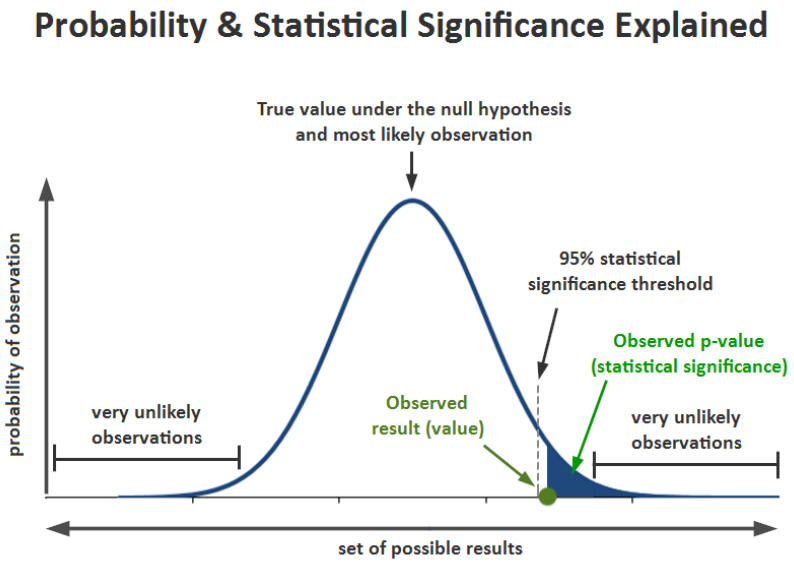
which means

$$\Phi\left(\frac{\sqrt{n}|\overline{x}_n - \mu_0|}{\sigma}\right) = 1 - \frac{\alpha^*}{2} \quad \Longleftrightarrow \quad \alpha^* = 2\left(1 - \Phi\left(\frac{\sqrt{n}|\overline{x}_n - \mu_0|}{\sigma}\right)\right).$$

This gives an example of how to compute the p-value (i.e., equal to the value of α^* by definition) for an outcome of the statistic $T(x) = \overline{x}_n$. But what does it mean? It becomes more clear if we write this α^* in another form:

$$\begin{aligned} p\text{-value} &= \alpha^* \\ &= \mathbb{P}_{\mu_0}\left(\left|\frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\sigma}\right| \geq \left|\frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\sigma}\right|\right) \\ &= \mathbb{P}_{\mu_0}\left(|\overline{X}_n - \mu_0| \geq |\overline{x}_n - \mu_0|\right). \end{aligned}$$

In other words, p-value equals the probability under H_0 of observing a value of the test statistic the *same as or more extreme* than what was actually observed.



What is the point of computing p-value? If p-value is small, say smaller than 0.05, we say the result is significant: the evidence is strong against H_0 and we should reject the null hypothesis.

p-value	evidence
< 0.01	very strong evidence against H_0
$[0.01, 0.05]$	strong evidence against H_0
$[0.05, 0.1]$	weak evidence against H_0
> 0.1	little or no evidence against H_0

Example: How to compute the p-value for general tests, e.g.,

$$H_0 : \theta \in \Theta_0 \quad \text{v.s.} \quad H_1 : \theta \notin \Theta_0.$$

Suppose that a size-α test is of the form

$$\text{reject } H_0 : \theta \in \Theta_0 \quad \text{ if and only if } \quad T(X) \geq c_\alpha$$

where c_α depends on α and c_α increases as α decreases since the rejection region shrink as the size/level α decreases.

For the observed data x, we find the smallest α^* such that

$$T(x) = c_{\alpha^*}$$

where the rejection region is $R_\alpha = [c_\alpha, \infty)$.

Then

$$p\text{-value} = \alpha^* = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T(X) \geq c_{\alpha^*}) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T(X) \geq T(x))$$

where x is the observed value of X. In particular, if $\Theta_0 = \{\theta_0\}$, then

$$p\text{-value} = \mathbb{P}_{\theta_0}(T(X) \geq T(x)).$$

Example: The p-value of Wald statistics. Suppose

$$T(x) = \left| \frac{\widehat{\theta}(x) - \theta_0}{\widehat{\text{se}}(x)} \right| = |W(x)|$$

denotes the observed absolute value of the Wald statistics $W(X)$. Then p-value is given by

$$p\text{-value} = \mathbb{P}_{\theta_0}(|W(X)| \geq |W(x)|) \approx \mathbb{P}(|Z| \geq |w|) = 2\Phi(-|w|)$$

where $Z \sim \mathcal{N}(0, 1)$. In other words, $|w| = z_{1-\alpha^*/2}$ and α^* is the p-value.

For example, in the coin tossing example,

$$W(x) = \frac{\widehat{\theta}(x) - \theta_0}{\widehat{\text{se}}(x)} = \frac{0.56 - 0.5}{\sqrt{\frac{0.56(1-0.56)}{1000}}} = 3.8224$$

and the p-value is equal to $\mathbb{P}(|Z| \geq 3.8224)$ which is approximately 0.0001 ¡ 0.01. In other words, the observed data are strongly against the null hypothesis $H_0 : \theta = 1/2$ and we should reject the null.

5.3 Likelihood ratio test

We have spent a lot of time discussing the basics of hypothesis testing. You may have realized that the key components in hypothesis testing are: (a) find a proper testing statistic; (b) identify the rejection region with a given size/level α. In all the examples we have covered so far, the construction of the rejection region and testing statistics rely on our intuition. In this lecture, we will introduce a systematic way to tackle the two aforementioned problems, which is based on the likelihood function.

Now let's consider the following testing problem:

$$H_0 : \theta \in \Theta_0 \text{ versus } H_1 : \theta \notin \Theta_0.$$

Suppose we observe samples $X_1, \cdots, X_n$ from $f(X;\theta)$. We define the likelihood ratio statistic as

$$\lambda(X) = 2\log\left(\frac{\sup_{\theta\in\Theta} L(\theta)}{\sup_{\theta\in\Theta_0} L(\theta)}\right) = 2(\ell(\widehat{\theta}) - \ell(\widehat{\theta}_0))$$

where $\widehat{\theta}$ is the MLE, $\widehat{\theta}_0$ is the MLE when θ is restricted to Θ_0, and $\ell(\theta) = \log L(\theta)$ is the log-likelihood function.

What properties does $\lambda(X)$ satisfy?

1. Since Θ (the natural parameter space) contains Θ_0, thus $\sup_{\theta\in\Theta} L(\theta) \geq \sup_{\theta\in\Theta_0} L(\theta)$, implying that $\lambda(X) \geq 0$.
2. Suppose H_0 is true, then the MLE is likely to fall into Θ_0 if n is sufficiently large because of the consistency of MLE.

Therefore, we can use $\lambda(X)$ as a testing statistic: if $\lambda(x_1, \cdots, x_n)$ is close to 0, we should retain H_0; if $\lambda(x_1, \cdots, x_n)$ is large, we reject H_0.

Example: Testing of normal mean with known variance. Suppose $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ and we want to test

$$H_0 : \mu = \mu_0 \quad \text{v.s.} \quad H_1 : \mu \neq \mu_0.$$

Now we consider the likelihood ratio statistics. Note that the likelihood function is

$$\ell(\mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 + C$$

where C is a scalar which does not depend on μ. There is no need to take σ^2 into consideration since it is assumed known.

Next, we compute the MLE of $\ell(\mu, \sigma^2)$ on Θ_0 and Θ:

$$\Theta_0 := \{\mu : \mu = \mu_0\}, \quad \Theta = \{\mu : \mu \in \mathbb{R}\}.$$

On the Θ_0, the maximum is simply

$$\ell(\mu_0) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu_0)^2$$

and on the Θ, the maximum is attained when $\mu = \overline{X}_n$:

$$\ell(\widehat{\mu}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Note that

$$\sum_{i=1}^{n}(X_i-\mu_0)^2=\sum_{i=1}^{n}(X_i-\overline{X}_n)^2+n(\overline{X}_n-\mu_0)^2.$$

Thus

$$\lambda(X)=\frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(X_i-\mu_0)^2-\sum_{i=1}^{n}(X_i-\overline{X}_n)^2\right)=\frac{n(\overline{X}_n-\mu_0)^2}{\sigma^2}.$$

Since we reject the null hypothesis if $\lambda(X)$ is large, it is equivalent to rejecting the null if $|\overline{X}_n-\mu_0|$ is large which matches our previous discussion. Now, how to evaluate this test? What is the power function?

$$\beta(\mu)=\mathbb{P}_\mu(\lambda(X)>c)$$

for some c. Under the null hypothesis, $\lambda(X)\sim\chi^2_1$ since

$$\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma}\sim\mathcal{N}(0,1).$$

Therefore, we reject the null hypothesis if

$$\lambda(X)>\chi^2_{1,1-\alpha}$$

where $\chi^2_{1-\alpha}$ is the $(1-\alpha)$-quantile of χ^2_1 distribution.

Exercise: Derive the one-sided test for the mean from normal data with known mean, using likelihood ratio statistics.

Exercise: Derive the one-sided test for the mean from normal data with unknown mean, using likelihood ratio statistics.

Example: Binomial likelihood ratio test. Recall the coin tossing problem:

$$H_0:\theta=\frac{1}{2}\quad\text{v.s.}\quad H_1:\theta\neq\frac{1}{2}.$$

We observe $\overline{x}_n=0.56$ (560 heads plus 440 tails out of $n=1000$ trials).

The likelihood function for $X_1,\cdots,X_n\sim$Bernoulli(θ) is

$$L(\theta)=\theta^{\sum_{i=1}^n X_i}(1-\theta)^{n-\sum_{i=1}^n X_i}=\theta^{n\overline{X}_n}(1-\theta)^{n(1-\overline{X}_n)}$$

and $L(\theta)$ is maximized by $\overline{X}_n$.

$$\lambda(X)=2\log\left(\frac{\sup_{\theta\in[0,1]}L(\theta)}{\sup_{\theta\in\Theta_0}L(\theta)}\right)=2\log\left(\frac{\overline{X}_n^{n\overline{X}_n}(1-\overline{X}_n)^{n(1-\overline{X}_n)}}{2^{-n}}\right)$$

Simplify the expression and we have

$$\lambda(X)=2n\left(\log 2+\overline{X}_n\log\overline{X}_n+(1-\overline{X}_n)\log(1-\overline{X}_n)\right).$$

For $\overline{x}_n=0.56$, the value of the statistic is $\lambda(x)=14.4348$.

However, it is unclear what the distribution of $\lambda(X)$ is.

5.3.1 Asymptotics of LRT

Theorem 5.3.1 (Asymptotic behavior). *For simple test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Suppose $X_1, \cdots, X_n$ are i.i.d. $f(x;\theta)$ and $\widehat{\theta}$ is the MLE of θ. Under H_0, as $n \to \infty$,*

$$\lambda(X) \xrightarrow{d} \chi_1^2$$

in distribution. The p-value is given by $\mathbb{P}(\chi_1^2 \geq \lambda(X))$.

Proof: Let $\widehat{\theta}_n$ be the MLE and θ_0 is the parameter under null hypothesis:

$$\ell(\theta_0) - \ell(\widehat{\theta}_n) = \ell'(\widehat{\theta}_n)(\theta_0 - \widehat{\theta}_n) + \frac{\ell''(\widehat{\theta}_n)}{2}(\widehat{\theta}_n - \theta_0)^2 + \text{remainder}.$$

Note that $\ell'(\widehat{\theta}_n) = 0$. Thus

$$\begin{aligned}\lambda(X) &= 2(\ell(\widehat{\theta}_n) - \ell(\theta_0)) \approx -\ell''(\widehat{\theta}_n)(\widehat{\theta}_n - \theta_0)^2 \\ &= -\frac{\ell''(\widehat{\theta}_n)}{n} \cdot [\sqrt{n}(\widehat{\theta}_n - \theta_0)]^2\end{aligned}$$

If the null is true, $\ell''(\widehat{\theta}_n) \approx -nI(\theta_0)$ where $I(\theta_0)$ is Fisher information. By asymptotical normality of MLE, $\sqrt{n}(\widehat{\theta}_n - \theta_0) \to \mathcal{N}(0, 1/I(\theta_0))$. Therefore, by Slutsky's theorem, it holds that $\lambda(X) \xrightarrow{d} \chi_1^2$ in distribution. □

Therefore, a size-α test is: reject H_0 if

$$\lambda(X) > \chi^2_{1,1-\alpha}$$

where $\chi^2_{1,1-\alpha}$ is the $(1-\alpha)$-quantile of χ_1^2.

How to compute the p-value? Note that p-value equals the probability of observing a new value which is equal to or more extreme than the observed one. Thus

$$p\text{-value} = \mathbb{P}(\chi_1^2 \geq \lambda(x_1, \cdots, x_n))$$

where $\lambda(x_1, \cdots, x_n)$ is the observed value of the likelihood ratio statistic. Or we can follow another definition: p-value is the smallest size/level at which we reject the null hypothesis. Thus the p-value equals α^* which satisfies

$$\lambda(x_1, \cdots, x_n) = \chi^2_{1,1-\alpha^*}$$

since the rejection region is

$$R_\alpha = (\chi^2_{1,1-\alpha}, \infty).$$

In other words,

$$1 - \alpha^* = \mathbb{P}(\chi_1^2 \leq \lambda(x_1, \cdots, x_n)) \Longleftrightarrow \alpha^* = \mathbb{P}(\chi_1^2 \geq \lambda(x_1, \cdots, x_n)).$$

In the previous example where $\lambda(x_1, \cdots, x_n) = 14.4348$, the p-value is

$$\mathbb{P}(\chi_1^2 \geq 14.4348) \approx 0.0001.$$

So we reject the null hypothesis since the evidence is very strong against the null hypothesis.

5.3.2 General LRT and asymptotics

Theorem 5.3.2. *Under certain regularity condition for the pdf/pmf of the sample* $(X_1, \cdots, X_n)$*, it holds that under* $H_0 : \theta \in \Theta_0$*,*

$$\lambda(X) \xrightarrow{d} \chi^2_{r-q}$$

in distribution as $n \to \infty$ *where*

- r *is the number of free parameters specified by* $\theta \in \Theta$*;*
- q *is the number of free parameter specified by* $\theta \in \Theta_0$

Example: Suppose $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_q, \theta_{q+1}, \cdots, \theta_r) \subseteq \mathbb{R}^r$. Let

$$\Theta_0 = \{\boldsymbol{\theta} \in \mathbb{R}^r : (\theta_{q+1}, \cdots, \theta_r) = (\theta_{q+1,0}, \cdots, \theta_{r,0})\}$$

which means the degree of freedom of Θ_0 is q. Then under H_0, it holds that

$$\lambda(X) \to \chi^2_{r-q}.$$

Question: How to calculating the degree of freedom? Suppose the parameter space Θ contains an open subset in $\mathbb{R}^r$ and Θ_0 contains an open subset in $\mathbb{R}^q$. Then $r - q$ is the degree of freedom for the test statistic.

We reject H_0 at the size α if

$$\lambda(X) > \chi^2_{r-q,1-\alpha}$$

where $\chi^2_{r-q,1-\alpha}$ is $1-\alpha$ quantile of χ^2_{r-q} distribution.

The p-value is calculated via

$$p\text{-value} = \mathbb{P}_{\theta_0}(\chi^2_{r-q} \geq \lambda(\boldsymbol{x}))$$

where $\lambda(\boldsymbol{x})$ is the observed data.

Example: Mendel's pea. Mendel bred peas with round yellow and wrinkled green seeds. The progeny has four possible outcomes:

$$\{\text{Yellow, Green}\} \times \{\text{Wrinkled, Round}\}$$

His theory of inheritance implies that $\mathbb{P}(\text{Yellow}) = 3/4$, $\mathbb{P}(\text{Green}) = 1/4$, $\mathbb{P}(\text{Round}) = 3/4$, and $\mathbb{P}(\text{Wrinkled}) = 1/4$:

$$\boldsymbol{p}_0 = \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right).$$

What he observed in the trials are $X = (315, 101, 108, 32)$ with $n = 556$.

We want to test if Mendel's conjecture is valid:

$$H_0 : \boldsymbol{p} = \boldsymbol{p}_0, \quad H_1 : \boldsymbol{p} \neq \boldsymbol{p}_0.$$

It means:

$$\Theta = \{\boldsymbol{p} : \boldsymbol{p} \geq 0, \sum_{i=1}^4 p_i = 1\}, \quad \Theta_0 := \{\boldsymbol{p} : \boldsymbol{p} = \boldsymbol{p}_0\}.$$

The number of each type follows *multinomial distribution*. The multinomial distribution has its pmf as

$$f(x_1, \cdots, x_k|\boldsymbol{p}) = \binom{n}{x_1, \cdots, x_k} p_1^{x_1} \cdots p_k^{x_k}, \quad x_i \in \mathbb{Z}^+, \sum_{i=1}^k x_i = n, \quad \sum_{i=1}^k p_i = 1.$$

- Suppose there are k different types of coupons. The probability of getting coupon i is p_i. Collect n coupons and ask what is the distribution for the count of each coupon X_i?
- It satisfies the multinomial distribution with parameter n and $\boldsymbol{p}$, denoted by $\mathcal{M}(n, \boldsymbol{p})$.
- It is a generalization of binomial distribution.

What is the MLE of $\boldsymbol{p}$? We have shown in the homework that

$$\widehat{p}_i = \frac{x_i}{n}.$$

Therefore,

$$\widehat{p} = (0.5665, 0.1817, 0.1942, 0.0576).$$

Consider the LRT:

$$\begin{aligned}
\lambda(X) &= 2\log \frac{\sup_{p:p\geq 0, \sum_i p_i = 1} L(p)}{\sup_{p:p=p_0} L(p)} \\
&= 2\log \frac{\prod_{i=1}^k \widehat{p}_i^{x_i}}{\prod_{i=1}^k p_{0i}^{x_i}} \\
&= 2\sum_{i=1}^k x_i \log \frac{\widehat{p}_i}{p_{0i}}
\end{aligned}$$

If the null is true, then it holds that $\lambda(X) \xrightarrow{d} \chi^2_{(4-1)-0} = \chi^2_3$. In this case, $\lambda(X) = 0.4754$ and p-value is

$$\mathbb{P}_{p_0}(\chi^2_3 \geq 0.4754) = 0.9243$$

which is not significant. Thus we retain the null hypothesis.

5.4 Goodness-of-fit test

Recall that when we discuss the parametric inference, we assume the data are drawn from a certain family of distribution. One hidden unanswered question is: are the data indeed samples from a specific distribution? More precisely, suppose we observe a set of samples $X_1, \cdots, X_n$, we ask if the underlying population distribution equals to another F_0. This is a quite fundamental problem since it will provide an approach to verify if our assumption on the family of distribution is valid. This is actually a hypothesis testing problem: we want to know if

$$H_0 : F_X = F_0 \quad \text{versus} \quad H_0 : F_X \neq F_0$$

where F_0 is the cdf of a specific distribution. This problem gives rise to an important topic in hypothesis testing: Goodness-of-fit test. In this lecture, we will discuss a few approaches to test the goodness-of-fit.

5.4.1 Likelihood ratio tests

In fact, likelihood ratio tests can be smartly used in goodness-of-fit tests. Suppose $X_1, \cdots, X_n$ are samples from F_X. We first divide the range of the data into k disjoint subintervals:

$$I_i = (a_i, a_{i+1}], \quad 1 \leq i \leq k, \quad \mathbb{R} = \cup_{i=1}^k I_i$$

where a_0 and a_{k+1} are $-\infty$ and ∞ respectively. Then we count

$$n_i = |\{X_k : X_k \in (a_i, a_{i+1}],\ 1 \le k \le n\}|, \quad n = \sum_{i=1}^{k} n_i.$$

Note that the probability of a random variable taking value I_i equals

$$p_i = \mathbb{P}(X \in I_i) = F_X(a_{i+1}) - F_X(a_i).$$

Then the random variable $(n_1, \cdots, n_k)$ is actually Multinomial distribution $\mathcal{M}(n, \boldsymbol{p})$ where $\boldsymbol{p} = (p_1, \cdots, p_k)^\top$ is a nonnegative vector with $\sum_{i=1}^k p_i = 1$. Now how to use this fact to do the following testing problem

$$H_0 : F_X = F_0 \quad \text{versus } H_1 : F_X \neq F_0.$$

One natural way is to first compute $p_{i0} = F_0(a_{i+1}) - F_0(a_i), 1 \le i \le k$; then test if

$$H_0 : p_i = p_{i0},\ \forall 1 \le i \le k \quad \text{versus} \quad H_1 : p_i \neq p_{i0} \text{ for some } i,$$

where $H_0 : p_i = p_{i0}$ is a *necessary* condition of $F_X = F_0$.

This is similar to the example we have covered before. We can use likelihood ratio test which is equal to

$$\lambda(X) = 2\sum_{i=1}^{k} n_i \log \frac{\widehat{p}_i}{p_{i0}} = 2n \sum_{i=1}^{k} \widehat{p}_i \log \frac{\widehat{p}_i}{p_{i0}}$$

where $\widehat{p}_i = n_i/n$. If the null is true, then

$$\lambda(X) \xrightarrow{d} \chi^2_{k-1}, \quad n \to \infty.$$

Therefore, we can reject the null hypothesis at level α if the observed value $\lambda(x)$ of LRT is too large:

$$\lambda(x) > \chi^2_{k-1,1-\alpha}.$$

Exercise: Verify that the LRT is

$$\lambda(X) = 2n \sum_{i=1}^{k} \widehat{p}_i \log \frac{\widehat{p}_i}{p_{i0}}.$$

In fact, we have derived it in the example of Mendel's peas.

Remember that the LRT is actually depending on how we partition the data. In practice, we usually perform the partitioning such that each interval has a substantial amount of data (say 10∼20) or each interval contains an approximately same amount of data.

5.4.2 Pearson χ^2-test

Another famous and commonly-used test for the goodness-of-fit is the Pearson χ^2-test. Let's first state the main idea. Similar to our previous discussion, instead of testing $F_X = F_0$ directly, we put the data points into several categories and then test $H_0 : p_i = \widehat{p}_{i0}$ for all $1 \le i \le k$ where

$$p_i = F_X(a_{i+1}) - F_X(a_i), \qquad p_{i0} = F_0(a_{i+1}) - F_0(a_i).$$

Let $O_i = n\widehat{p}_i$ be the number of observed samples in the ith category and $E_i = np_i$ is the expected number of samples in the ith category. The Pearson χ^2-test uses the following statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = n \sum_{i=1}^{k} \frac{(\widehat{p}_i - p_i)^2}{p_i}.$$

Now the question is: what is the distribution of χ^2 if the null is true?

Theorem 5.4.1. *Under null hypothesis, it holds that*

$$\chi^2 = n \sum_{i=1}^{k} \frac{(\widehat{p}_i - p_{i0})^2}{p_{i0}} \xrightarrow{d} \chi^2_{k-1}.$$

Example: Mendel's peas. The observed number of four possible outcomes is $\boldsymbol{x} = (316, 101, 108, 32)$ and

$$\widehat{\boldsymbol{p}} = (0.5665, 0.1817, 0.1942, 0.0576)$$

and

$$\boldsymbol{p}_0 = \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right).$$

The χ^2-statistic is

$$\chi^2 = n \sum_{i=1}^{4} \frac{(\widehat{p}_i - p_{i0})^2}{p_{i0}} = 0.4849$$

where $n = 556$. The p-value is

$$p - \text{value} = \mathbb{P}(\chi^2_3 > 0.4849) = 0.9222.$$

The result is not significant enough to reject the null hypothesis.

Question: Can we provide a justification for this theorem? In fact, the proof simply follows from our homework and multivariate normal distribution. Suppose $(X_1, \cdots, X_k)$ is a random variable of multinomial distribution $\mathcal{M}(n, \boldsymbol{p})$. Then the MLE $\widehat{p}_i$ of $p_i = X_i/n$. Recall that the MLE satisfies asymptotic normality:

$$\sqrt{n}(\widehat{\boldsymbol{p}} - \boldsymbol{p}) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \quad \Sigma_{ij} = \begin{cases} p_i(1 - p_i), & i = j, \\ -p_i p_j, & i \neq j. \end{cases}$$

In fact, the covariance matrix is equivalent to the following form

$$\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top$$

Note that this covariance matrix is not strictly positive semidefinite since one eigenvalue is zero. It is called the *degenerate* multivariate normal distribution, i.e., one coordinate in the random vector can be represented as a linear combination of other coordinates. Now let's study why

$$\chi^2 = \sum_{i=1}^{k} \frac{(\widehat{p}_i - p_i)^2}{p_i} \sim \chi^2_{k-1}$$

holds. First of all, χ^2-statistic can be written as the squared norm of a vector:

$$\chi^2 = \|\sqrt{n} \operatorname{diag}(\boldsymbol{p})^{-1/2}(\widehat{\boldsymbol{p}} - \boldsymbol{p})\|^2$$

where $\mathrm{diag}(\boldsymbol{p})$ is a diagonal matrix whose diagonal entries are given by $\boldsymbol{p}$.

What is the distribution of $\sqrt{n}[\mathrm{diag}(\boldsymbol{p})]^{-1/2}(\widehat{\boldsymbol{p}}-\boldsymbol{p})$? Since $\sqrt{n}(\widehat{\boldsymbol{p}}-\boldsymbol{p})$ is asymptotically normal, so is $\sqrt{n}[\mathrm{diag}(\boldsymbol{p})]^{-1/2}(\widehat{\boldsymbol{p}}-\boldsymbol{p})$. Its covariance matrix is given by

$$\mathrm{Cov}(\sqrt{n}[\mathrm{diag}(\boldsymbol{p})]^{-1/2}(\widehat{\boldsymbol{p}}-\boldsymbol{p})) = [\mathrm{diag}(\boldsymbol{p})]^{-1/2}\boldsymbol{\Sigma}[\mathrm{diag}(\boldsymbol{p})]^{-1/2} = \boldsymbol{I}_k - \sqrt{\boldsymbol{p}}\sqrt{\boldsymbol{p}}^\top.$$

Thus

$$\sqrt{n}[\mathrm{diag}(\boldsymbol{p})]^{-1/2}(\widehat{\boldsymbol{p}}-\boldsymbol{p}) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{I}_k - \sqrt{\boldsymbol{p}}\sqrt{\boldsymbol{p}}^\top).$$

Now we can see that $\boldsymbol{I}_k - \sqrt{\boldsymbol{p}}\sqrt{\boldsymbol{p}}^\top$ is actually a projection matrix. Since it is a projection matrix of rank $k-1$, we can find a matrix $\boldsymbol{U}$ (eigenvectors of $\boldsymbol{I}_k - \sqrt{\boldsymbol{p}}\sqrt{\boldsymbol{p}}^\top$ w.r.t. eigenvalue 1) of size $k \times (k-1)$ such that $\boldsymbol{I}_k - \sqrt{\boldsymbol{p}}\sqrt{\boldsymbol{p}}^\top = \boldsymbol{U}\boldsymbol{U}^\top$ and $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_{k-1}$.

Exercise: Show that $\boldsymbol{I}_k - \sqrt{\boldsymbol{p}}\sqrt{\boldsymbol{p}}^\top$ is an orthogonal projection matrix and compute its eigenvalues (with multiplicities). Remember that $\sum_{i=1}^k p_i = 1$.

Suppose $\boldsymbol{Z}$ is $\mathcal{N}(0, \boldsymbol{I}_{k-1})$. Then

$$\boldsymbol{U}\boldsymbol{Z} \sim \mathcal{N}(0, \boldsymbol{U}\boldsymbol{U}^\top)$$

which matches the asymptotic distribution of $\sqrt{n}[\mathrm{diag}(\boldsymbol{p})]^{-1/2}(\widehat{\boldsymbol{p}}-\boldsymbol{p})$. Therefore, its squared Euclidean norm is chi-squared distribution with degree of freedom $k-1$.

5.4.3 Test on Independence

Suppose $(X_i, Y_i), 1 \le i \le n$ are i.i.d. samples from $F_{X,Y}$. Can we design a procedure to test if X and Y are independent? In fact, we can use χ^2 Pearson test and likelihood ratio test. How? We still start with discretization by partitioning the range of X and Y respectively:

$$\mathrm{Range}(X) = \bigcup_{i=1}^r A_i, \quad \mathrm{Range}(Y) = \bigcup_{j=1}^c B_j$$

where $\{A_i\}_{i=1}^r$ and $\{B_j\}_{j=1}^c$ are partitions of $\mathrm{Range}(X)$ and $\mathrm{Range}(Y)$ respectively (mutually disjoint).

Denote

$$p_i = \mathbb{P}(X \in A_i), \quad q_j = \mathbb{P}(Y \in B_j), \quad \theta_{ij} = \mathbb{P}(X \in A_i, Y \in B_j).$$

To test the independence, we are actually interested in testing the following hypothesis

$$H_0 : \theta_{ij} = p_i q_j, \ \forall i, j, \quad \text{versus} \quad H_1 : \theta_{ij} \neq p_i q_j, \ \text{for some } i, j.$$

The null hypothesis H_0 is a necessary condition for the independence of X and Y.

How to construct a testing statistic? We can first use LRT. Note that $\{n_{ij}\}_{1\le i\le r, 1\le j\le c}$ satisfies multinomial distribution $\mathcal{M}(n, \{\theta_{ij}\}_{1\le i\le r, 1\le j\le c})$, whose pmf is

$$f(n_{11}, n_{12}, \cdots, n_{rc}; n, \theta_{ij}) = \binom{n}{n_{11}, n_{12}, \cdots, n_{rc}} \prod_{i=1}^r \prod_{j=1}^c \theta_{ij}^{n_{ij}}$$

where $n = \sum_{i,j} n_{ij}$.

Exercise: Show the MLE for $\boldsymbol{p}$ and $\boldsymbol{q}$ under $H_0 : \theta_{ij} = p_i q_j$ is given by

$$\widehat{\boldsymbol{p}}_i = \frac{1}{n}\sum_{j=1}^c n_{ij}, \quad \widehat{\boldsymbol{q}}_j = \frac{1}{n}\sum_{i=1}^r n_{ij}.$$

Exercise: Show the likelihood ratio statistic for this hypothesis testing problem is

$$\lambda(X) = 2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \log \frac{\widehat{\theta}_{ij}}{\widehat{p}_i \widehat{q}_j} = 2n \sum_{i=1}^{r}\sum_{j=1}^{c} \widehat{\theta}_{ij} \log \frac{\widehat{\theta}_{ij}}{\widehat{p}_i \widehat{q}_j}.$$

What is the asymptotic distribution of $\lambda(X)$ under null hypothesis?

First note that we don't know p_i, q_j, and θ_{ij} since the joint distribution $F_{X,Y}$ is unknown. Therefore, we need to estimate these parameters:

$$\widehat{p}_i = \frac{n_{i\cdot}}{n} = \frac{1}{n}\sum_{j=1}^{c} n_{ij}, \quad \widehat{q}_j = \frac{n_{\cdot j}}{n} = \frac{1}{n}\sum_{i=1}^{r} n_{ij}, \quad \widehat{\theta}_{ij} = \frac{n_{ij}}{n}$$

where

$$n_{i\cdot} = |\{k : X_k \in A_i\}|, \quad n_{\cdot j} = |\{k : Y_k \in B_j\}|, \quad n_{ij} = |\{k : (X_k, Y_k) \in A_i \times B_j\}|$$

is the number of observed samples belonging to $A_i \times B_j$. This gives a contingency table:

	B_1	B_2	$\cdots$	B_c	Row total
A_1	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r\cdot}$
Column total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot c}$	n

Table 5.2: Contingency table

If $\widehat{\theta}_{ij}$ deviates from $\widehat{p}_i\widehat{q}_j$ by a large margin, we are likely to reject the null hypothesis. Therefore, we introduce the following χ^2-statistic:

$$\chi^2 = n\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(\widehat{\theta}_{ij} - \widehat{p}_i\widehat{q}_j)^2}{\widehat{p}_i\widehat{q}_j}$$

Exercise: Show that χ^2 statistic is approximately equal to likelihood ratio statistic (Hint: use Taylor approximation).

How to construct a rejection region? To do that, we need to know the asymptotic distribution of χ^2 under the null.

Theorem 5.4.2. *Under the null hypothesis,*

$$\chi^2 \xrightarrow{d} \chi^2_{(r-1)(c-1)}.$$

As a result, we reject the null at size α if $\chi^2 > \chi^2_{(r-1)(c-1),1-\alpha}$.

How to understand this degree of freedom in the chi-squared distribution? The degree of freedom of Θ_0 is

$$(r-1) + (c-1) = r + c - 2$$

since $\sum_{i=1}^{r} p_i = \sum_{j=1}^{c} q_j = 1$. The degree of freedom of Θ is $rc - 1$. Thus the difference is

$$rc - 1 - (r + c - 2) = rc - r - c + 1 = (r-1)(c-1).$$

Example: (DeGroot and Schervish, Ex. 10.3.5) Suppose that 300 persons are selected at random from a large population, and each person in the sample is classified according to blood type, O, A, B, or AB, and also according to Rh (Rhesus, a type of protein on the surface of red blood cells. Positive is the most common blood type), positive or negative. The observed numbers are given in Table 10.18. Test the hypothesis that the two classifications of blood types are independent.

Table 5.3: Data for the two classifications of blood types

	O	A	B	AB	Total
Rh positive	82	89	54	19	244
Rh negative	13	27	7	9	56
Total	95	116	61	28	300

Here we have $n = 300$, and we compute $\widehat{p}_i$, $\widehat{q}_i$, and $\widehat{\theta}_{ij}$: and

Table 5.4: Estimation of θ_{ij}, p_i, and q_j

$\widehat{\theta}_{ij}$	O	A	B	AB	$\widehat{p}_i$
Rh positive	0.2733	0.2967	0.1800	0.0633	0.8133
Rh negative	0.0433	0.0900	0.0233	0.0300	0.1867
$\widehat{q}_j$	0.3167	0.3867	0.2033	0.0933	1

Table 5.5: Estimation of $p_i q_j$

$\widehat{p}_i\widehat{q}_j$	O	A	B	AB
Rh positive	0.2576	0.3145	0.1654	0.0759
Rh negative	0.0591	0.0722	0.0380	0.0174

The result is $\chi^2 = 8.6037$. Under the null hypothesis, $\chi^2 \sim \chi^2_3$, with $r = 4$ and $c = 2$. The p-value is

$$\mathbb{P}(\chi^2_3 > 8.6037) = 0.0351$$

We reject the null hypothesis that Rh factor and ABO system are independent at the level 0.05.

Exercise: Perform the independence test by using LRT and show if you can get a similar result.

5.5 Kolmogorov-Smirnov test

All the tests described above rely on discretization. Is it possible to use the cdf to construct a testing statistic? At the beginning of the course, we focus a lot on the properties of empirical cdf and we know the empirical cdf is a consistent estimator of the population cdf.

5.5.1 KS test for Goodness-of-fit

Now suppose we observe $X_i, 1 \leq i \leq n$ and want to test if the data come from F_0.

$$H_0 : F_X = F_0, \quad H_1 : F_X \neq F_0$$

Can we design a testing procedure purely based on the empirical cdf? Recall that the empirical cdf equals:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \le x\}.$$

Under the null hypothesis, we know that $F_n(x)$ converges to $F_0(x)$ for any fixed x as $n \to \infty$, even uniformly, as implied by Glivenko-Cantelli theorem.

Therefore, if $F_n(x)$ is far away from $F_0(x)$, we are more likely to reject the null hypothesis. One intuitive way to measure the difference between $F_n(x)$ and F_0 is to use the supremum of $|F_n(x) - F_0(x)|$, i.e., Kolmogorov-Smirnov statistic:

$$T_{KS} = \sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)|.$$

If T_{KS} is large, we are likely to reject the null hypothesis. The tricky part is: what is the distribution of T_{KS} under the null?

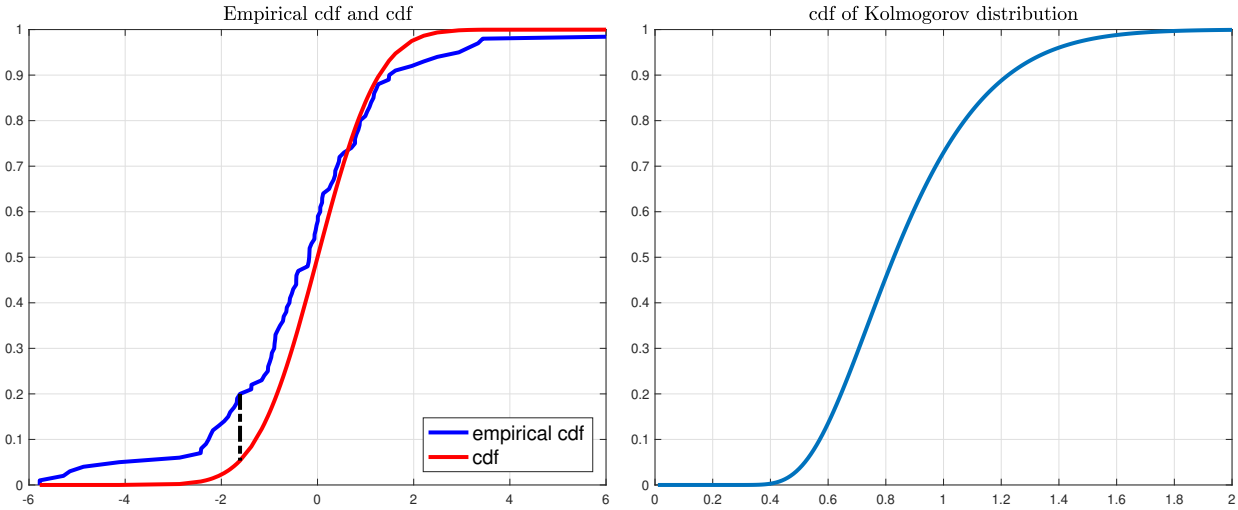


Figure 5.2: Left: empirical cdf v.s. population cdf; Right: cdf of Kolmogorov distribution

Under null hypothesis, $K = \sqrt{n}T_{KS}$ satisfies the Kolmogorov distribution

$$\mathbb{P}(K \le x) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2x^2}.$$

We reject the null hypothesis if

$$\sqrt{n}T_{KS} > K_{1-\alpha},$$

where $K_{1-\alpha}$ is the $(1-\alpha)$ quantile of Kolmogorov distribution.

5.5.2 Two-sample test

Let $X_1, \cdots, X_n \sim F_X$ and $Y_1, \cdots, Y_m \sim F_Y$. Can we test $H_0 : F_X = F_Y$?

$$H_0 : F_X = F_Y, \qquad H_1 : F_X \neq F_Y.$$

We construct the KS statistic as

$$T_{KS} = \sup_{x\in\mathbb{R}} |F_{X,n}(x) - F_{Y,m}(x)|$$

We reject the null hypothesis if

$$\sqrt{\frac{mn}{m+n}}T_{KS} > K_{1-\alpha}$$

where $\sqrt{\frac{mn}{m+n}}T_{KS}$ satisfies the Kolmogorov distribution under H_0.

Exercise: Under null hypothesis, show that

$$\text{Var}\left(\frac{mn}{m+n}(F_{X,n} - F_{Y,m})\right) = F_X(1 - F_X).$$

Chapter 6

Linear and logistic regression

6.1 What is regression?

Suppose we observe a set of data (X_i, Y_i), $1 \leq i \leq n$. They are not necessarily identically distributed. Now we ask this question: assume you are given a new data point X_{new}, can you make a prediction of X_{new}? This is actually the core problem in machine learning: learn a model from the data and use the model to make predictions.

- Image classification
- Credit card fraud detection
- Life expectancy prediction based on health data (blood pressure, age, weight, height, location, diet, etc)

Here X_i is called the predictor and Y_i is the response. In most cases, X_i can be a vector consisting of many different features. How is it related to statistics? Let's first look at a simple case in which (X_i, Y_i) is a 2D vector. Using the probabilistic assumption, we treat the observed data as realizations of a bivariate joint distribution $f_{X,Y}(x, y)$ of (X, Y). Our goal is to make a prediction of Y based on X. Mathematically, this prediction is a function of X, say, $g(X)$.

How to evaluate the quality of this prediction $f(X)$? We hope to have a prediction $f(X)$ such that it is close to Y. Their "distance" function is often measured by using a loss function. Some common loss functions include

- Quadratic loss (ℓ_2-loss): $\ell(x, y) = (x - y)^2$
- Absolute loss (ℓ_1-loss): $\ell(x, y) = |x - y|$
- Logistic loss, hinge loss, KL divergence, Wasserstein distance, etc

Different loss functions may lead to vastly different results. Choosing a proper function often heavily relies on specific applications and there is no universally prefect loss function.

6.1.1 Global minimizer under quadratic loss

Let's focus on the most widely used ℓ_2-loss function. Note that X and Y are both random variables. Therefore, it is natural to find out a prediction function $g(\cdot)$ such that

the average loss, i.e., *risk* function, is as small as possible:

$$\min_{g} \ \mathbb{E}_{(X,Y)\sim f_{X,Y}}(X - g(X))^2$$

Can we find the global minimizer to this risk function? If so, what is it? Surprisingly, the global minimizer is the conditional expectation of Y given X, i.e., the optimal choice of g is

$$g(x) = \mathbb{E}(Y|X = x) = \int_{\mathbb{R}} f_{Y|X}(y|x)\,\mathrm{d}y$$

where $f(y|x)$ denotes the condition pdf of Y given X,

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

Here $\mathbb{E}(Y|X)$ is called the *regression* of Y on X, the "best" predictor of Y conditioned on X, which is also a function of X. We provide the proof to justify why conditional expectation is the best predictor of Y under quadratic loss function.

Theorem 6.1.1. *Suppose the random vector (X, Y) has finite second moments, then*

$$\mathbb{E}(Y|X) = \mathit{argmin}_g \ \mathbb{E}_{(X,Y)\sim f_{X,Y}}(Y - g(X))^2$$

Proof: The proof follows from the property of conditional expectation:

$$\begin{aligned}
&\mathbb{E}_{X,Y}(Y - g(X))^2 - \mathbb{E}_{X,Y}(Y - \mathbb{E}(Y|X))^2 \\
=&\mathbb{E}_{X,Y}[(2Y - g(X) - \mathbb{E}(Y|X))(\mathbb{E}(Y|X) - g(X))] \\
=&\mathbb{E}_X\,\mathbb{E}_Y\left([(2Y - g(X) - \mathbb{E}(Y|X))(\mathbb{E}(Y|X) - g(X))]\Big|X\right)
\end{aligned}$$

Note that given X, $\mathbb{E}(Y|X)$ and $g(X)$ are both known. Thus

$$\begin{aligned}
&\mathbb{E}_Y\left([(2Y - g(X) - \mathbb{E}(Y|X))(\mathbb{E}(Y|X) - g(X))]\Big|X\right) \\
&= (\mathbb{E}(Y|X) - g(X))\,\mathbb{E}_Y\left((2Y - g(X) - \mathbb{E}(Y|X))\Big|X\right) \\
&= (\mathbb{E}(Y|X) - g(X))(2\,\mathbb{E}(Y|X) - g(X) - \mathbb{E}(Y|X)) \\
&= (\mathbb{E}(Y|X) - g(X))^2 \geq 0.
\end{aligned}$$

Therefore, $\mathbb{E}(Y|X)$ is the global minimizer as it holds that

$$\mathbb{E}_{X,Y}(Y - g(X))^2 \geq \mathbb{E}_{X,Y}(Y - \mathbb{E}(Y|X))^2.$$

□

It seems that we have already found the solution to the aforementioned prediction problem: the answer is conditional expectation. However, the reality is not so easy. What is the main issue here? Apparently, in practice, we do not know the actual population distribution $f_{X,Y}$ directly. Instead, we only have an access to a set of n data points (X_i, Y_i), which can be viewed as a realization from a bivariate distribution. Therefore, we cannot write down the loss function explicitly since it involves an expectation taken w.r.t. the joint distribution $f_{X,Y}$.

6.1.2 Empirical risk minimization

How to resolve this problem? Recall that in bootstrap, we replace the population cdf by the empirical cdf. It seems that we can follow a similar idea here:

$$\mathbb{E}_{(X,Y)\sim F_n(x,y)}\,\mathbb{E}(Y-g(X))^2.$$

Note that the empirical cdf $F_n(x,y)$ equals

$$F_n(x,y)=\frac{1}{n}\sum_{i=1}^{n}1\{X_i\le x, Y_i\le y\}$$

whose corresponding pmf is $\mathbb{P}(X^*=X_i,Y^*=Y_i)=1/n$.

Then the expectation is taken w.r.t. this discrete distribution

$$\ell(g)=\mathbb{E}_{(X,Y)\sim F_n}(Y-g(X))^2=\sum_{i=1}^{n}(Y_i-g(X_i))^2.$$

It suffices to minimize this loss function by a proper function g. The value of this loss function is called *training error* and $\ell(g)$ is called the empirical loss (risk) function. However, this cannot work directly. Why? In fact, there are infinitely many ways to minimize this loss function. Let's look at a simple case: suppose (X_i,Y_i) are mutually different, we can always find a curve (a function g) going through all the observed data.

6.1.3 Occam's razor and bias-variance tradeoff

From our discussion in the previous section, we have seen that without any constraints on $g(\cdot)$, it is easy to make the training error become zero. However, this does not provide a meaningful result since it will create overfitting issue. Suppose (X,Y) is a new sample from $f_{X,Y}$, then the *generalization error* associated to $\widehat{g}_n$ (the trained function from data (X_i,Y_i), $1\le i\le n$) is

$$\mathbb{E}_{(X,Y)\sim f_{X,Y}}(Y-\widehat{g}_n(X))^2=\int_{\mathbb{R}^2}(y-\widehat{g}_n(x))^2 f(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

can be large.

Therefore, in practice, we usually would restrict g to a class of functions. For example, the function class $\mathcal{G}$ could be

- Polynomial: $\mathcal{G}=\{a_nx^n+\cdots+a_1x+a_0|a_i\in\mathbb{R}\}$;
- Logistic function:
$$\mathcal{G}=\left\{\frac{1}{\exp(\theta x)+1}\Big|\theta\in\mathbb{R}\right\}$$
- Neural network with multiple layers and nodes

Ideally, we would like to keep this class of functions as simple as possible, by following Occam's razor. What is Occam's razor? It is the principle that states a preference for simple theories, or "accept the simplest explanation that fits the data". The simplest class of function is *linear* function. In other words, we assume $\mathbb{E}(Y|X)$ as a function of X, i.e.,

$$\mathbb{E}(Y|X=x)=\alpha x+\beta$$

or

$$\mathcal{G} = \{g(x) = \alpha x + \beta : \alpha, \beta \in \mathbb{R}\}.$$

If $\mathcal{G}$ is linear function, we say the regressor is linear, which is the focus of this chapter *linear regression*. You may ask what is the point of studying linear regression? Linear regression is seemingly too naive. In fact, linear regression is extremely useful in statistics. Statisticians have developed comprehensive theory for linear models. In practice, linear models have already provided a satisfactory explanation for many datasets. Moreover, for data which are close to multivariate normal distribution, using linear model is sufficient.

Exercise: For $(X, Y) \sim \mathcal{N}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, $\mathbb{E}(Y|X)$ is a linear function of X.

6.2 Simple linear regression

6.2.1 Data fitting using LS estimator

Starting from this lecture, we will study simple linear regression in which we have a single predictor and the expected response given the predictor is a linear function

$$\mathbb{E}(Y_i|X_i = x_i) = \beta_0 + \beta_1 x_i,$$

or we often use another equivalent form:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where ϵ_i is the noise term.

Now suppose observe a set of data samples (X_i, Y_i), how to find (β_0, β_1) such that the model fits the data? Recalling in our last lecture, we use the quadratic loss function to fit the data by empirical risk minimization:

$$\min_{g \in \mathcal{G}} \sum_{i=1}^{n} (g(X_i) - Y_i)^2 = \min_{\beta_0, \beta_1} \sum_{i=1}^{n} (\beta_0 + \beta_1 X_i - Y_i)^2$$

where

$$\mathcal{G} = \{g(x) : g(x) = \beta_0 + \beta_1 x, \quad \beta_0, \beta_1 \in \mathbb{R}\}.$$

This approach is the well-known *l*inear least squares method. The minimizers are called linear least squares estimator.

Definition 6.2.1 (The least squares estimates)**.** *The least squares estimates are the values $(\widehat{\beta}_0, \widehat{\beta}_1)$ such that the residual sum of squares or RSS is minimized, i.e.,*

$$(\widehat{\beta}_0, \widehat{\beta}_1) = argmin_{\beta_0, \beta_1} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2.$$

Exercise: The empirical risk function $R(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$ is a convex function of β_0 and β_1.

Lemma 6.2.1. *The least squares estimator is given by*

$$\widehat{\beta}_0 = \overline{Y}_n - \widehat{\beta}_1 \overline{X}_n, \quad \widehat{\beta}_1 = \frac{\sum_i X_i Y_i - n\overline{X}_n \overline{Y}_n}{\sum_i X_i^2 - n\overline{X}_n^2} = \frac{\sum_i (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_i (X_i - \overline{X}_n)^2}$$

Proof: Define $R(\beta_0, \beta_1) = \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_i))^2$.

$$\frac{\partial R}{\partial \beta_0} = 2\sum_{i=1}^n (\beta_0 + \beta_1 X_i - Y_i) = 0, \quad \frac{\partial R}{\partial \beta_1} = 2\sum_{i=1}^n X_i(\beta_0 + \beta_1 X_i - Y_i) = 0$$

Simplifying the equations:

$$\beta_0 + \beta_1 \overline{X}_n = \overline{Y}_n, \quad \beta_0 n\overline{X}_n + \beta_1 \sum_{i=1}^n X_i^2 = \sum_{i=1}^n X_i Y_i \tag{6.2.1}$$

Solving for (α, β) gives the linear squares estimator. Substitute $\beta_0 = \overline{Y}_n - \beta_1 \overline{X}_n$ into the second equation:

$$n(\overline{Y}_n - \beta_1 \overline{X}_n)\overline{X}_n + \beta_1 \sum_{i=1}^n X_i^2 = \sum_{i=1}^n X_i Y_i$$

It gives

$$\left(\sum_{i=1}^n X_i^2 - n\overline{X}_n^2\right)\beta_1 = \sum_{i=1}^n X_i Y_i - n\overline{X}_n \overline{Y}_n \Longleftrightarrow \widehat{\beta}_1 = \frac{\sum_i (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_i (X_i - \overline{X}_n)^2}$$

Then $\widehat{\beta}_0 = \overline{Y}_n - \widehat{\beta}_1 \overline{X}_n$, which finishes the proof. □

Suppose we have a set of $(\widehat{\beta}_0, \widehat{\beta}_1)$. The predicted value at X_i is given by

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i,$$

where

- $\widehat{\beta}_0$ is related to the sample mean of X_i and Y_i.
- $\widehat{\beta}_1$ is the ratio of sample covariance of (X_i, Y_i) over the sample variance of X_i.

The residuals (prediction error) are

$$\widehat{\epsilon}_i = Y_i - \widehat{Y}_i = Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i)$$

In other words, the least squares approach aims to minimize the sum of squares of the prediction error.

Example: Predict grape crops: the grape vines produce clusters of berries and a count of these clusters can be used to predict the final crop yield at harvest time.

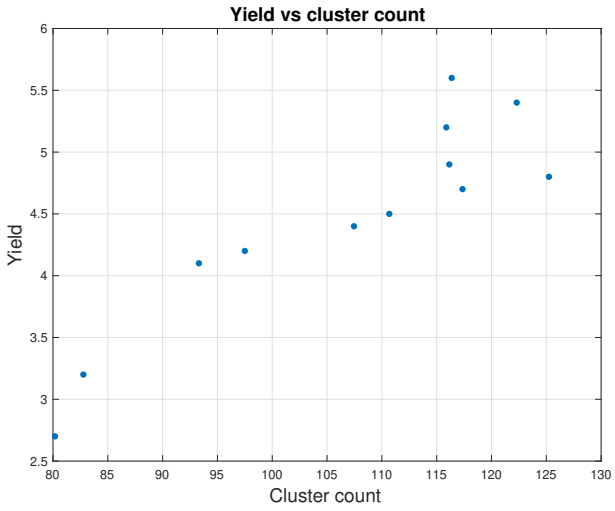
- Predictor: cluster count, $\{X_i\}_{i=1}^n$
- Response: yields $\{Y_i\}_{i=1}^n$

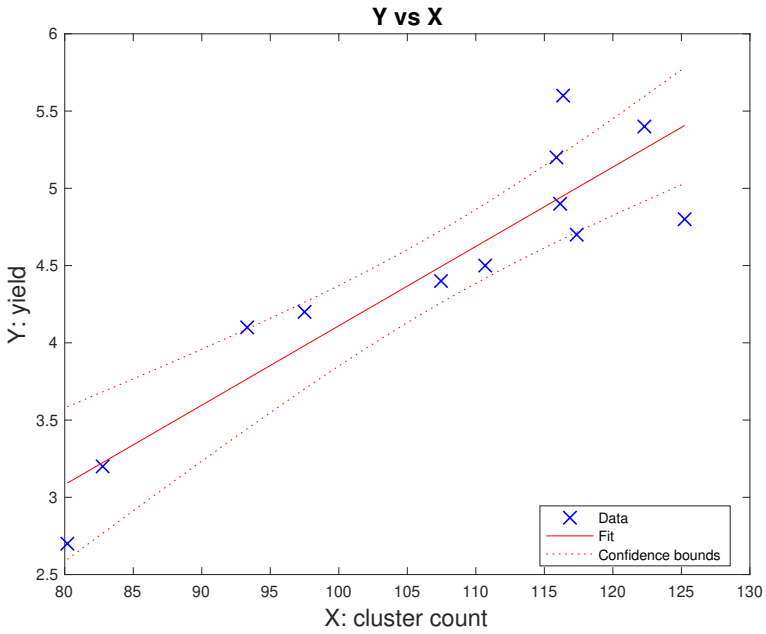
The LS estimator of (β_0, β_1) is

$$\widehat{\beta}_0 = -1.0279, \quad \widehat{\beta}_1 = 0.0514$$

Year	Cluster count (X)	Yields (Y)
1971	116.37	5.6
1973	82.77	3.2
1974	110.68	4.5
1975	97.5	4.2
1976	115.88	5.2
1977	80.19	2.7
1978	125.24	4.8
1979	116.15	4.9
1980	117.36	4.7
1981	93.31	4.1
1982	107.46	4.4
1983	122.3	5.4

Table 6.1: Data source: [Casella, Berger, 2001]. The data in 1972 is missing due to Hurricane.





6.2.2 Best linear unbiased estimator

All the analyses so far do not involve any statistical noise. Here we ask such a question: is the least squares estimator optimal under statistical models? Here we introduce a fairly

general model. Consider Y_i is the value of the response variable in the ith case and X_i is the value of the predictor variable. We assume that

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad 1 \le i \le n$$

where ϵ_i are uncorrelated, zero-mean, and equal variance random variables:

$$\mathbb{E}(\epsilon_i) = 0, \quad \mathrm{Var}(\epsilon_i) = \sigma^2, \quad 1 \le i \le n$$

and

$$\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0, \quad 1 \le i \ne j \le n.$$

The following parameters are unknown and to be estimated:

- β_0: intercept
- β_1: slope
- σ^2: unknown variance

An estimator of $\boldsymbol{\beta}$ is essentially a function of the response $Y_1, \cdots, Y_n$. Now let's only focus on a small set of estimators: the linear estimators of (β_0, β_1), i.e., these estimators of the following form

$$\left\{ \sum_{i=1}^n \alpha_i Y_i : \alpha_i \in \mathbb{R},\ 1 \le i \le n \right\}.$$

In other words, we are looking at the estimators which are given by the linear combination of Y_i where the coefficients α_i are to be determined. In particular, we are interested in finding an unbiased linear estimator for (β_0, β_1) with the smallest variance. What is it?

We take β_1 as an example as the same argument applies to β_0 accordingly. In order to ensure unbiasedness of the estimator, we need to have

$$\begin{aligned} \mathbb{E}\left(\sum_{i=1}^n \alpha_i Y_i \right) &= \sum_{i=1}^n \alpha_i \, \mathbb{E}\, Y_i = \sum_{i=1}^n \alpha_i (\beta_0 + \beta_1 X_i) \\ &= \beta_0 \sum_{i=1}^n \alpha_i + \beta_1 \sum_{i=1}^n \alpha_i X_i = \beta_1. \end{aligned}$$

This gives

$$\sum_{i=1}^n \alpha_i = 0, \qquad \sum_{i=1}^n \alpha_i X_i = 1.$$

What is the variance of $\sum_{i=1}^n \alpha_i Y_i$? Note that Y_i are uncorrelated and then

$$\mathrm{Var}\left(\sum_{i=1}^n \alpha_i Y_i \right) = \sum_{i=1}^n \alpha_i^2 \, \mathrm{Var}(Y_i) + \sum_{i<j} \alpha_i \alpha_j \, \mathrm{Cov}(Y_i, Y_j) = \sigma^2 \sum_{i=1}^n \alpha_i^2.$$

Now to find the best linear unbiased estimator (BLUE), it suffices to minimize:

$$\min_{\alpha_i} \ \frac{1}{2} \sum_{i=1}^n \alpha_i^2 \quad \text{s.t.} \quad \sum_{i=1}^n \alpha_i = 0, \ \sum_{i=1}^n \alpha_i X_i = 1.$$

We resort to the method of Lagrangian multiplier to find the optimal α_i. Let λ and μ be the multiplier, and then

$$L(\alpha_i, \lambda, \mu) = \frac{1}{2}\sum_{i=1}^{n} \alpha_i^2 - \lambda \sum_{i=1}^{n} \alpha_i - \mu \left(\sum_{i=1}^{n} \alpha_i X_i - 1 \right).$$

The optimal solution satisfies:

$$\begin{aligned}
&\frac{\partial L}{\partial \alpha_i} = \alpha_i - \lambda - \mu X_i = 0,\ 1 \le i \le n, \\
&\sum_{i=1}^{n} \alpha_i = 0, \\
&\sum_{i=1}^{n} \alpha_i X_i = 1.
\end{aligned}$$

Substituting $\alpha_i = \lambda + \mu X_i$ into the second and third equations:

$$\begin{aligned}
\lambda + \overline{X}_n \mu &= 0, \\
n\overline{X}_n \lambda + \sum_{i=1}^{n} X_i^2 \mu &= 1.
\end{aligned}$$

Solving the equation gives:

$$\mu = \frac{1}{\sum_{i=1}^{n} X_i^2 - n\overline{X}_n^2}, \quad \lambda = -\frac{\overline{X}_n}{\sum_{i=1}^{n} X_i^2 - n\overline{X}_n^2}.$$

Now we have the BLUE:

$$\begin{aligned}
\widehat{\beta}_1 &= \sum_{i=1}^{n} \alpha_i Y_i = \sum_{i=1}^{n} (\mu X_i + \lambda) Y_i \\
&= \frac{1}{S_{xx}} \sum_{i=1}^{n} (X_i - \overline{X}_n) Y_i \\
&= \frac{1}{S_{xx}} \sum_{i=1}^{n} (X_i - \overline{X}_n)(Y_i - \overline{Y}_n) = \widehat{\beta}_{1,LS}
\end{aligned}$$

where $\sum_{i=1}^{n} (X_i - \overline{X}_n)\overline{Y}_n = 0$. In other words, the least squares estimator is the BLUE!

Exercise: Find the mean and variance of $\widehat{\beta}_0$ and $\widehat{\beta}_1$. Are $\widehat{\beta}_0$ and $\widehat{\beta}_1$ uncorrelated?

6.2.3 Matrix form

We have derived the LS estimator of β_0 and β_1. Now we incorporate linear algebra into LS estimation, which will be very useful for our future study. First, we let

$$\boldsymbol{X} = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}, \quad \boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Then the data equation becomes

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Recall that the least squares estimator of β_0 and β_1 satisfy (6.2.1)

$$\beta_0 + \beta_1 \overline{X}_n = \overline{Y}_n, \quad \beta_0 n\overline{X}_n + \beta_1 \sum_{i=1}^n X_i^2 = \sum_{i=1}^n X_i Y_i$$

Rewriting them into matrix form, we have

$$\begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i Y_i \end{bmatrix} \Longleftrightarrow \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^\top \boldsymbol{Y}.$$

Therefore, we have

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y} = \frac{1}{n\sum_{i=1}^n X_i^2 - (\sum_i X_i)^2} \begin{bmatrix} \sum_{i=1}^n X_i^2 & -\sum_{i=1}^n X_i \\ -\sum_{i=1}^n X_i & n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i Y_i \end{bmatrix}.$$

Exercise: Verify that $\widehat{\boldsymbol{\beta}}$ above matches our previous derivation.

Question: What is the mean and covariance of $\widehat{\boldsymbol{\beta}}$?

Lemma 6.2.2. *Under the assumption of simple linear regression, i.e., ϵ_i are uncorrelated, equal variance, and zero-mean random variables. The LS estimator $\widehat{\boldsymbol{\beta}}$ has mean and variance as follows:*

$$\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}, \quad \mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \frac{\sigma^2}{n} \frac{1}{\sum_i X_i^2 - n\overline{X}_n^2} \begin{bmatrix} \sum_i X_i^2 & -n\overline{X}_n \\ -n\overline{X}_n & n \end{bmatrix}.$$

Proof: Note that the LS estimator is

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \boldsymbol{\beta} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon}.$$

By linearity of expectation, $\mathbb{E}\,\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ since $\mathbb{E}\,\boldsymbol{\epsilon} = 0$. For its covariance, we know that the covariance equals

$$\begin{aligned} \mathrm{Cov}(\widehat{\boldsymbol{\beta}}) &= \mathbb{E}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \\ &= \mathbb{E}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top) \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\ &= \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\ &= \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}. \end{aligned}$$

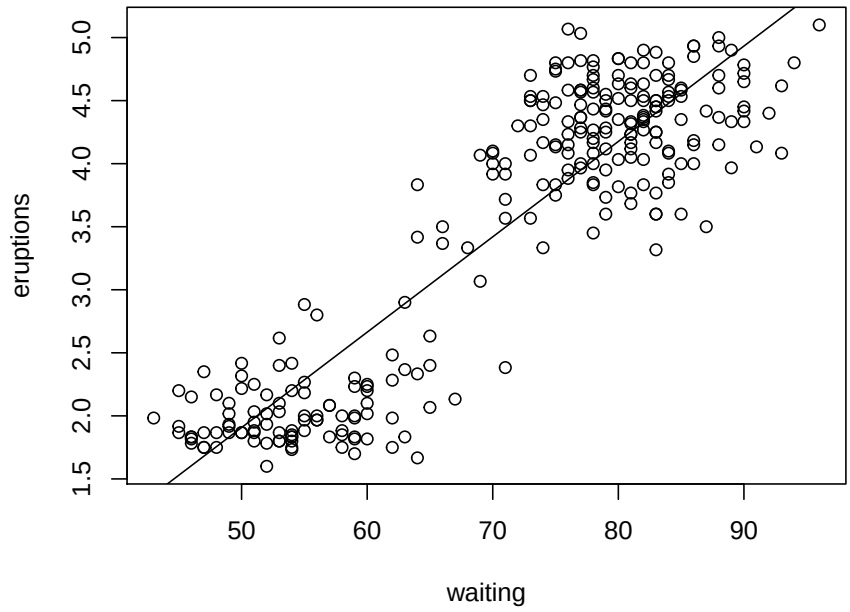
□

6.3 Simple linear regression under normal error model

Recall the data of old faithful geyser. We want to understand the relation between eruption and waiting time.

Using the formula discussed before, it is easy to fit a linear model to the data, using R.

```
> data
    eruptions waiting
1       3.600      79
2       1.800      54
```




```
3       3.333      74
4       2.283      62
> lmfit <- lm(eruptions~waiting, data = data)
> lmfit$coefficients
(Intercept)     waiting
-1.87401599  0.07562795
```

We obtain $\widehat{\beta}_0 = -1.874$ and $\widehat{\beta}_1 = 0.0756$. Then we have a linear model:

$$Y = -1.874 + 0.0756X + \epsilon$$

where ϵ is the error.

Now we ask the following questions: suppose we observe the waiting time is $X = 70$, what is the predicted value of the eruption time? Can you construct a confidence interval? Or can we perform a hypothesis testing such as whether $\beta_1 = 0$ or not?

To construct a confidence interval, we need to assume more on the linear model, especially the distribution of the error. From now on, we will discuss a commonly-used model in linear regression: normal error model.

Under normal error model, we are assuming that each response Y_i given the predictor X_i is a normal random variable:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad 1 \le i \le n$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. normal random variables with unknown σ^2. In other words, $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$ are independent random variables with equal variance. Apparently, the least squares estimator in this case is the BLUE, i.e., best linear unbiased estimator. The reason is simple: all the noise ϵ_i are i.i.d., and thus their covariance satisfies:

$$\mathbb{E}\,\epsilon_i\epsilon_j = \begin{cases} 0, & i \neq j, \\ \sigma^2, & i = j, \end{cases} \quad \Longleftrightarrow \quad \mathbb{E}\,\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top = \sigma^2 \boldsymbol{I}_n.$$

6.3.1 MLE under normal error model

Since we observe a set of random variables Y_i with unknown parameters β_0, β_1, and σ^2, it is natural to use maximal likelihood estimation to estimate the parameter.

Lemma 6.3.1. *The MLE of $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$ matches the least square estimators. The MLE of σ^2 is*

$$\widehat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^n \widehat{\epsilon}_i^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2$$

where $\widehat{\epsilon}_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i$.

Proof: Because of the independence among Y_i, the likelihood function is

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - (\beta_0 + \beta_1 X_i))^2}{2\sigma^2}\right)$$

The log-likelihood function is

$$\ell(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_i))^2.$$

To maximize the log-likelihood function, we first maximize over β_0 and β_1 whose maximizer equals the minimizer to the least squares risk function:

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \text{argmin}_{(\beta_0,\beta_1)} \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_i))^2$$

How about the MLE of σ^2? Once $\boldsymbol{\beta}$ is fixed, we just need to take the derivative w.r.t. σ^2:

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2 = 0$$

and the MLE of σ^2 is

$$\widehat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^n (Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i))^2 = \frac{1}{n}\sum_{i=1}^n \widehat{\epsilon}_i^2.$$

□

Question: What is the distribution of $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1)^\top$?

Knowing its distribution is crucial in constructing a confidence interval. Note that the least squares estimator $\widehat{\boldsymbol{\beta}}$ is obtained by solving a linear equation $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^\top\boldsymbol{Y}$.

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} \\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}^\top\boldsymbol{\epsilon}) \\
&= \boldsymbol{\beta} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\epsilon}
\end{aligned}$$

where

$$\boldsymbol{X}^\top\boldsymbol{X} = \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix}, \quad \boldsymbol{X}^\top\boldsymbol{Y} = \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i Y_i \end{bmatrix}.$$

Note that $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$ is a normal random vector in $\mathbb{R}^n$. By linearity of multivariate normal distribution, $\widehat{\boldsymbol{\beta}}$ is also multivariate normal, with mean and variance:

$$\begin{aligned}
\mathbb{E}\,\widehat{\boldsymbol{\beta}} &= \boldsymbol{\beta}, \\
\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) &= \mathbb{E}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \\
&= \mathbb{E}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\
&= \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}.
\end{aligned}$$

As a result, $\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1})$ holds where

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix}^{-1} = \frac{1}{n} \frac{1}{\sum_i X_i^2 - n\overline{X}_n^2} \begin{bmatrix} \sum_i X_i^2 & -n\overline{X}_n \\ -n\overline{X}_n & n \end{bmatrix}$$

From the result above, we have

$$\widehat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^n X_i^2/n}{\sum_{i=1}^n X_i^2 - n\overline{X}_n^2}\right), \quad \widehat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n X_i^2 - n\overline{X}_n^2}\right).$$

Now, can we construct a confidence interval for β_0 and β_1? The parameter σ^2 is unknown. We could replace the unknown σ^2 by its MLE $\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^n \widehat{\epsilon}_i^2$. Is the MLE an unbiased and consistent estimator of σ^2?

Definition 6.3.1. *The error sum of squares is*

$$SSE = \sum_{i=1}^n \widehat{\epsilon}_i^2$$

where $\widehat{\epsilon}_i = Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i)$ are the residuals. The mean squared error (MSE) is

$$MSE = \frac{SSE}{n-2}$$

Theorem 6.3.2. *Under normal error model, the error sum of squares SSE satisfies*

- $SSE/\sigma^2 \sim \chi^2_{n-2}$.
- *SSE is independent of $\widehat{\boldsymbol{\beta}}$.*

As a result, MSE is a consistent and unbiased estimator of σ^2:

$$\widehat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \widehat{\epsilon}_i^2$$

since $\sum_{i=1}^n \widehat{\epsilon}_i^2 \sim \chi^2_{n-2}$ and the consistency follows from law of large number.

Proof: Now let's prove the theorem. The proof is very straightforward. Note that

$$\begin{aligned}
\widehat{\boldsymbol{\epsilon}} &= \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} \\
&= \boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y} \\
&= (\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top)(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\
&= (\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top)\boldsymbol{\epsilon}
\end{aligned}$$

where $(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{X} = 0$ and $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$.

In fact, the matrix $\boldsymbol{P} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ is a projection matrix with rank $n-2$. Therefore,

$$SSE = \widehat{\boldsymbol{\epsilon}}^\top\widehat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}^\top(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{\epsilon} \sim \sigma^2\chi^2_{n-2}.$$

We leave it as an exercise to show $\boldsymbol{P}$ is a projection matrix. The rank of P follows from

$$\mathrm{Tr}(\boldsymbol{P}) = \mathrm{Tr}(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top) = \mathrm{Tr}(\boldsymbol{I}_n) - \mathrm{Tr}((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}) = n - \mathrm{Tr}(\boldsymbol{I}_2) = n-2$$

where $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}$ is a 2 by 2 identity matrix. Why SSE is independent of $\widehat{\boldsymbol{\beta}}$? This is because

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\epsilon}$$

and we can show that $\widehat{\boldsymbol{\beta}}$ is uncorrelated with $\widehat{\boldsymbol{\epsilon}}$.

$$\begin{aligned}\mathrm{Cov}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\epsilon}}) &= \mathbb{E}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\widehat{\boldsymbol{\epsilon}}^\top \\ &= \mathbb{E}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top) \\ &= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top) = 0.\end{aligned}$$

For two random Gaussian vectors which are uncorrelated, they are independent. □

Exercise: Show that $\boldsymbol{P}$ is a symmetric projection matrix, i.e., $\boldsymbol{P}^2 = \boldsymbol{P}$ with rank $n-2$.

Exercise: Show that $(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{X} = 0$.

6.3.2 Confidence interval

Note that MSE is a consistent estimator of σ^2. Therefore, after replacing σ^2 with MSE, we have

$$\widehat{\beta}_0 - \beta_0 \sim \mathcal{N}\left(0, \frac{MSE}{n}\frac{\sum_i X_i^2}{S_{xx}}\right), \quad \widehat{\beta}_1 - \beta_1 \sim \mathcal{N}\left(0, \frac{MSE}{S_{xx}}\right)$$

where $S_{xx} = \sum_{i=1}^n X_i^2 - n\overline{X}_n^2$.

Note that an estimation of the standard deviation $\mathrm{se}(\widehat{\beta}_0)$ and $\mathrm{se}(\widehat{\beta}_1)$ is

$$\widehat{\mathrm{se}}(\widehat{\beta}_0) = \sqrt{\frac{MSE\sum_{i=1}^n X_i^2}{nS_{xx}}}, \quad \widehat{\mathrm{se}}(\widehat{\beta}_0) = \sqrt{\frac{MSE}{S_{xx}}}.$$

Then we have the following result.

Lemma 6.3.3. *Under normal error model, the LS estimator $\widehat{\boldsymbol{\beta}}$ satisfies*

$$\frac{\widehat{\beta}_0 - \beta_0}{\widehat{\mathrm{se}}(\widehat{\beta}_0)} \sim t_{n-2}, \quad \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\mathrm{se}}(\widehat{\beta}_0)} \sim t_{n-2}.$$

Proof: We just prove for $\widehat{\beta}_0$ since the justification applies to $\widehat{\beta}_1$ similarly.

$$\frac{\widehat{\beta}_0 - \beta_0}{\widehat{\mathrm{se}}(\widehat{\beta}_0)} = \left(\frac{\widehat{\beta}_0 - \beta_0}{\mathrm{se}(\widehat{\beta}_0)}\right) / \left(\frac{\widehat{\mathrm{se}}(\widehat{\beta}_0)}{\mathrm{se}(\widehat{\beta}_0)}\right)$$

where

$$\frac{\widehat{\beta}_0 - \beta_0}{\mathrm{se}(\widehat{\beta}_0)} \sim \mathcal{N}(0,1), \quad \frac{\widehat{\mathrm{se}}(\widehat{\beta}_0)}{\mathrm{se}(\widehat{\beta}_0)} = \sqrt{\frac{MSE}{\sigma^2}} = \sqrt{\frac{SSE}{(n-2)\sigma^2}}, \quad \frac{SSE}{\sigma^2} \sim \chi^2_{n-2}.$$

By definition of Student t-distribution, i.e., $t_\nu = \frac{Z}{\sqrt{\chi^2_\nu/\nu}}$, where $Z \sim \mathcal{N}(0,1)$ is independent of χ^2_ν, we have our result.

□

Theorem 6.3.4 (Hypothesis testing for $\boldsymbol{\beta}$)**.** *Under normal error model, we have*

- $(1-\alpha)$ *confidence interval for* β_0 *and* β_1 *is*

$$\widehat{\beta}_0 \pm t_{n-2,1-\frac{\alpha}{2}} \widehat{\mathrm{se}}(\widehat{\beta}_0), \qquad \widehat{\beta}_1 \pm t_{n-2,1-\frac{\alpha}{2}} \widehat{\mathrm{se}}(\widehat{\beta}_1)$$

- *Test* $H_0 : \beta_1 = 0$ *v.s.* $H_1 : \beta_1 \neq 0$*. A test of size* α *is to reject* H_0 *if*

$$|w| = \left| \frac{\widehat{\beta}_1}{\widehat{\mathrm{se}}(\widehat{\beta}_1)} \right| > t_{n-2,1-\frac{\alpha}{2}}$$

The p-value is $\mathbb{P}(|t_{n-2}| > |w|)$ *where* t_{n-2} *is a Student-t distribution of degree* $n-2$.

If n is large, we have

$$\frac{\widehat{\beta}_0 - \beta_0}{\widehat{\mathrm{se}}(\widehat{\beta}_0)} \sim \mathcal{N}(0,1), \quad \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\mathrm{se}}(\widehat{\beta}_0)} \sim \mathcal{N}(0,1)$$

Thus with large samples, it is safe to replace t_{n-2} by $\mathcal{N}(0,1)$ (e.g. $t_{n-2,1-\frac{\alpha}{2}}$ by $z_{1-\frac{\alpha}{2}}$.)

6.3.3 Prediction interval of the mean response

Note that if we have $\widehat{\boldsymbol{\beta}}$, we can predict the mean response value Y_* for a newly given X^*. The population expected mean response is $Y_* = \beta_0 + \beta_1 X_*$ and a natural choice of predicted value of Y_* is given by

$$\widehat{Y}_* = \widehat{\beta}_0 + \widehat{\beta}_1 X_*$$

How to get a confidence interval for Y_*?

The mean of $\widehat{Y}_*$ is

$$\mathbb{E}(\widehat{Y}_*) = \beta_0 + \beta_1 X_*.$$

For the variance, we have

$$\begin{aligned}
\mathrm{Var}(\widehat{Y}_*) &= \mathrm{Var}(\widehat{\beta}_0) + X_*^2 \,\mathrm{Var}(\widehat{\beta}_1) + 2X_* \,\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) \\
&= \frac{\sigma^2}{n(\sum_i X_i^2 - n\overline{X}_n^2)} \left(\sum_i X_i^2 + nX_*^2 - 2nX_*\overline{X}_n \right) \\
&= \frac{\sigma^2}{n(\sum_i X_i^2 - n\overline{X}_n^2)} \left(\sum_i X_i^2 - n\overline{X}_n^2 + n\overline{X}_n^2 + nX_*^2 - 2nX_*\overline{X}_n \right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(\overline{X}_n - X_*)^2}{\sum_i X_i^2 - n\overline{X}_n^2} \right)
\end{aligned}$$

What is the distribution of $\widehat{Y}_*$?

$$\widehat{Y}_* \sim \mathcal{N}\left(Y_*, \mathrm{Var}(\widehat{Y}_*)\right).$$

Replacing σ^2 by $\widehat{\sigma}^2$ in $\mathrm{Var}(\widehat{Y}_*)$, we have

$$\frac{\widehat{Y}_* - Y^*}{\widehat{\mathrm{se}}(\widehat{Y}_*)} \sim t_{n-2}.$$

where

$$\widehat{\mathrm{se}}(\widehat{Y}_*) = \sqrt{MSE}\sqrt{\frac{1}{n} + \frac{(\overline{X}_n - X_*)^2}{\sum_i X_i^2 - n\overline{X}_n^2}}$$

Therefore, an approximate $1-\alpha$ confidence interval of the mean response value is

$$(\widehat{Y}_* - t_{n-2,1-\frac{\alpha}{2}}\widehat{\mathrm{se}}(\widehat{Y}_*), \widehat{Y}_* + t_{n-2,1-\frac{\alpha}{2}}\widehat{\mathrm{se}}(\widehat{Y}_*))$$

6.4 Multiple regression

Simple linear regression only involves only one predictor. In practice, we often have multiple features to predict a response. If we want to predict a person's weight, we can use his/her height, gender, diet, age, etc. Here we generalize simple linear regression to multiple linear regression with more than one variables.

Definition 6.4.1 (Multiple linear regression model).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots \beta_{p-1} X_{i,p-1} + \epsilon_i,$$

where

- Y_i*: value of the response variable* Y *in the* i*th case*
- $X_{i1}, \cdots, X_{i,p-1}$ *: values of the variables* $X_1, \cdots, X_{p-1}$
- $\beta_0, \cdots, \beta_{p-1}$ *: regression coefficients.* p*: the number of regression coefficients; in simple regression* $p = 2$
- *Error term:* $\mathbb{E}(\epsilon_i) = 0$*,* $\mathrm{Var}(\epsilon_i) = \sigma^2$ *and* $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$

The mean response is

$$\mathbb{E}(Y|X_1, \cdots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_p.$$

The model equation can be written into

$$\underbrace{\boldsymbol{Y}}_{n\times 1} = \underbrace{\boldsymbol{X}}_{n\times p}\underbrace{\boldsymbol{\beta}}_{p\times 1} + \underbrace{\boldsymbol{\epsilon}}_{n\times 1}$$

where

$$\boldsymbol{X} := \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & X_{i1} & X_{i2} & \cdots & X_{i,p-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}, \quad \boldsymbol{\beta} := \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$$

Under model assumption:

$$\mathbb{E}(\boldsymbol{\epsilon}) = 0, \quad \mathrm{Cov}(\boldsymbol{\epsilon}) = \mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top) = \sigma^2 \boldsymbol{I}_n.$$

Now let's consider the least squares estimator for the multiple regression. The LS estimator is given by

$$\widehat{\boldsymbol{\beta}}_{LS} = \mathrm{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Y}\|^2$$

where $\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Y}\|^2 = (\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Y})^\top(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Y})$.

How to find the global minimizer to this program? A few facts on how to take gradients of vector-valued functions:

Lemma 6.4.1. *Let $\langle \boldsymbol{x}, \boldsymbol{v}\rangle = \boldsymbol{v}^\top \boldsymbol{x}$ denote the inner product between two column vectors $\boldsymbol{x}$ and $\boldsymbol{v}$ and $\boldsymbol{A}$ is a symmetric matrix:*

$$\begin{aligned} f_1(\boldsymbol{x}) &= \boldsymbol{x}^\top \boldsymbol{v}, & \frac{\partial f_1}{\partial \boldsymbol{x}} &= \boldsymbol{v}, \\ f_2(\boldsymbol{x}) &= \boldsymbol{x}^\top \boldsymbol{x}, & \frac{\partial f_2}{\partial \boldsymbol{x}} &= 2\boldsymbol{x}, \\ f_3(\boldsymbol{x}) &= \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}, & \frac{\partial f_3}{\partial \boldsymbol{x}} &= 2\boldsymbol{A}\boldsymbol{x}, \frac{\partial^2 f_3}{\partial \boldsymbol{x}^2} = 2\boldsymbol{A}. \end{aligned}$$

Proof: For $f_1(\boldsymbol{x})$, we know $f_1(\boldsymbol{x}) = \sum_{i=1}^n v_i x_i$ is a linear function of $\boldsymbol{x}$:

$$\frac{\partial f_1}{\partial x_i} = v_i, \quad \frac{\partial f_1}{\partial \boldsymbol{x}} = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = \boldsymbol{v}.$$

For $f_2(\boldsymbol{x})$, $f_2(\boldsymbol{x}) = \sum_{i=1}^n x_i^2$ is a quadratic function.

$$\frac{\partial f_2}{\partial x_i} = 2x_i, \quad \frac{\partial f_2}{\partial \boldsymbol{x}} = \begin{bmatrix} 2x_1 \\ \vdots \\ 2x_n \end{bmatrix} = 2\boldsymbol{x}.$$

For $f_3(\boldsymbol{x})$, $f_3(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = \sum_{i,j} a_{ij} x_i x_j$.

$$\begin{aligned} \frac{\partial f_3}{\partial x_i} &= \frac{\partial}{\partial x_i}\left(\sum_{j\neq i}^n a_{ij}x_j + \sum_{j\neq i}^n a_{ji}x_j + a_{ii}x_i^2\right) x_i \\ &= \sum_{j\neq i}^n a_{ij}x_j + \sum_{j\neq i}^n a_{ji}x_j + 2a_{ii}x_i \\ &= 2\sum_{j=1}^n a_{ij}x_j = 2[\boldsymbol{A}\boldsymbol{x}]_i, \end{aligned}$$

where $a_{ij} = a_{ji}$ follows from symmetry of $\boldsymbol{A}$.

Therefore,

$$\frac{\partial f_3}{\partial \boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{x}.$$

The Hessian of f_3 is defined by

$$\nabla^2 f_3 = \left[\frac{\partial^2 f_3}{\partial x_i \partial x_j}\right]_{1\le i,j\le n}.$$

Then

$$\frac{\partial^2 f_3}{\partial x_i \partial x_j} = \frac{\partial}{\partial x_j}\frac{\partial f_3}{\partial x_i} = \frac{\partial}{\partial x_j}\left(2\sum_{j=1}^{n} a_{ij}x_j\right) = 2a_{ij}.$$

□

Theorem 6.4.2. *Suppose $\boldsymbol{X}$ is of rank p, i.e., all the columns are linearly independent, then the least squares estimator is*

$$\widehat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}.$$

Proof: Define $f(\boldsymbol{\beta}) = \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Y}\|^2 = \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{Y} + \|\boldsymbol{Y}\|^2$:

$$\frac{\partial f}{\partial \boldsymbol{\beta}} = 2(\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}^\top \boldsymbol{Y}) = 0$$

which gives

$$\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^\top \boldsymbol{Y}.$$

Suppose $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is of rank p, then $\boldsymbol{X}^\top \boldsymbol{X} \in \mathbb{R}^{p\times p}$ is also rank p and invertible. Thus the LS estimator is

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}$$

given that the rank of $\boldsymbol{X}$ equals p. Now let's check its optimality:

$$\frac{\partial^2 f}{\partial \boldsymbol{\beta}^2} = 2\boldsymbol{X}^\top \boldsymbol{X} \succ 0.$$

Why $\boldsymbol{X}^\top \boldsymbol{X} \succ 0$? Its quadratic form $\boldsymbol{v}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{v} = \|\boldsymbol{X}\boldsymbol{v}\|^2 > 0$ for any $\boldsymbol{v} \neq 0$. □

Example: Polynomial regression. Given a set of distinct samples (X_i, Y_i), $1 \le i \le n$, we are seeking for a polynomial $f(x) = \beta_p x^p + \cdots + \beta_1 x + \beta_0$ of degree p that fits the data. We want to find a curve such that the squared distance from each point to the curve is minimized:

$$\min_{\beta_k, 0\le k\le p} \sum_{i=1}^{n} (Y_i - f(X_i))^2$$

How is it related to linear least squares estimation? If $p + 1 < n$, it is impossible to find a curve to go through every sample point. Therefore, we allow some error for each X_i,

$$Y_i = \sum_{k=0}^{p} \beta_k X_i^k + \epsilon_i.$$

This fits into the framework of multiple linear regression. Define

$$\boldsymbol{X} = \begin{bmatrix} 1 & X_1 & \cdots & X_1^p \\ 1 & X_2 & \cdots & X_2^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_n & \cdots & X_n^p \end{bmatrix} \in \mathbb{R}^{n\times(p+1)}$$

and

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\beta} \in \mathbb{R}^{p+1}.$$

From the result discussed before, we know that the least-squares estimator for $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

Why is $\boldsymbol{X}^\top \boldsymbol{X}$ invertible? $\boldsymbol{X}$ is called Vandermonde matrix and is full rank as long as there are $p+1$ distinct value of X_i.

6.4.1 Statistical properties of LS estimator

Under the multiple linear regression with uncorrelated noise of equal variance, i.e.,

$$\mathrm{Cov}(\epsilon_i, \epsilon_j) = \begin{cases} \sigma^2, & i = j, \\ 0, & i \neq j, \end{cases}$$

we have the famous Gauss-Markov theorem. Here we explain it informally. It is a generalization of the scenario we discussed in the simple linear regression. Consider all the linear estimator of $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}}_C = \boldsymbol{C}\boldsymbol{Y}$$

where $\boldsymbol{C}$ is a matrix of size $p \times n$. Now we are looking for unbiased estimator with smallest variance for $\boldsymbol{\beta}$, i.e., requiring $\boldsymbol{C}$ such that $\boldsymbol{C}\,\mathbb{E}\,\boldsymbol{Y} = \boldsymbol{\beta}$ and ensure $\mathrm{Var}(\boldsymbol{C}\boldsymbol{Y})$ is as small as possible. We need to generalize the "order" for positive semidefinite covariance matrix. We say two positive semidefinite matrix $\boldsymbol{A}$ and $\boldsymbol{B}$ satisfies $\boldsymbol{A} \succeq \boldsymbol{B}$ if $\boldsymbol{A} - \boldsymbol{B} \succeq 0$. The Gauss-Markov theorem tells us that this is given by the least squares estimator, i.e.,

$$\widehat{\boldsymbol{C}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top,$$

satisfies

$$\mathrm{Var}(\widehat{\boldsymbol{C}}\boldsymbol{Y}) = \widehat{\boldsymbol{C}}\,\mathrm{Cov}(\boldsymbol{Y})\widehat{\boldsymbol{C}}^\top \preceq \boldsymbol{C}\,\mathrm{Cov}(\boldsymbol{Y})\boldsymbol{C}^\top = \mathrm{Var}(\boldsymbol{C}\boldsymbol{Y})$$

for any $\boldsymbol{C}$ such that $\boldsymbol{C}\,\mathbb{E}\,\boldsymbol{Y} = \boldsymbol{\beta}$.

In practice, we often use the multiple regression model under normal error, i.e.,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$.

Lemma 6.4.3. *Under the assumption that $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$, it holds that*

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}).$$

Proof: Suppose $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$. Note that $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

$$\mathbb{E}\,\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \mathbb{E}(\boldsymbol{Y}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

What is the covariance of $\widehat{\boldsymbol{\beta}}$?

$$\begin{aligned} \mathrm{Var}(\widehat{\boldsymbol{\beta}}) &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top) \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\ &= \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \end{aligned}$$

where $\mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top) = \sigma^2 \boldsymbol{I}_p$. Therefore,

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1})$$

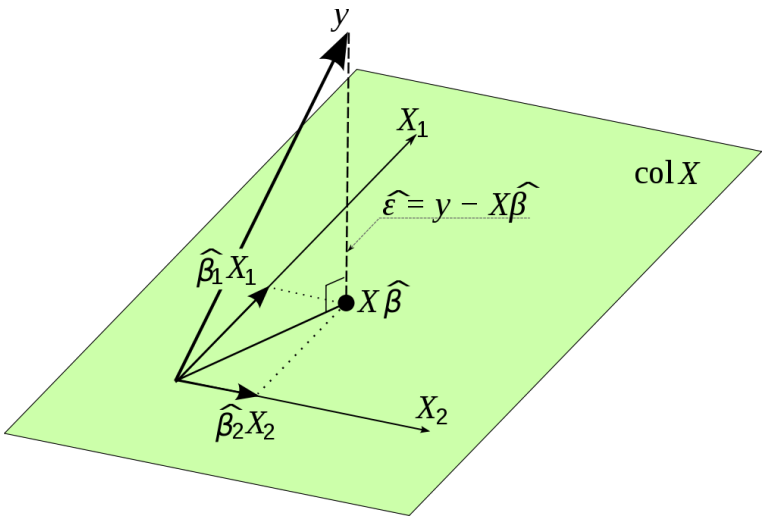
□

6.4.2 Geometric meaning of least squares estimator

Here we assume $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is rank-p, i.e., all the columns are linear independent. We want to under understand what least squares mean geometrically?

Let's first understand the meaning of minimizing $\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Y}\|^2$ mean.

- Find a vector in the range of $\boldsymbol{X}$ such that its distance to $\boldsymbol{Y}$ is minimized
- Equivalent to projecting $\boldsymbol{Y}$ onto the linear subspace spanned by the columns of $\boldsymbol{X}$.
- The residue $\widehat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ is perpendicular to any vectors in the column space of $\boldsymbol{X}$, i.e., $\boldsymbol{X}^\top\widehat{\boldsymbol{\epsilon}} = 0 \iff \boldsymbol{X}^\top\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^\top\boldsymbol{Y}$.



Projection matrix

Note that $\widehat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$. The fitted value $\widehat{\boldsymbol{Y}}$ is

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{LS} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$$

The matrix $\boldsymbol{H} := \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ is called hat matrix, i.e., projection matrix. The projection matrix $\boldsymbol{H}$ (projection onto the column (range) space of $\boldsymbol{X}$) has the following properties:

- Symmetric: $\boldsymbol{H}^\top = \boldsymbol{H}$
- Idempotent: $\boldsymbol{H}^2 = \boldsymbol{H}$. Applying the projection twice won't change the outcome.
- All of its eigenvalues are either 0 (multiplicity $n - p$) or 1 (multiplicity p)

Exercise: show that $\boldsymbol{I} - \boldsymbol{H}$ is also a projection matrix.

Question: how to representing fitted value and residuals by using $\boldsymbol{H}$?

- The fitted data $\widehat{\boldsymbol{Y}}$ is the projection of $\boldsymbol{Y}$ on the range of $\boldsymbol{X}$, i.e.,

$$\widehat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$$

- The residue $\widehat{\epsilon}$ is equal to

$$\widehat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\epsilon}$$

  which is the projection of $\boldsymbol{Y}$ on the complement of $\mathrm{Ran}(\boldsymbol{X})$.

6.4.3 Inference under normal error bound

Question: how to representing SSE and MSE by using $\boldsymbol{H}$?

The SSE and MSE are defined by

$$\begin{aligned} SSE &:= \sum_{i=1}^{n} \widehat{\epsilon}_i^2 = \|\widehat{\boldsymbol{\epsilon}}\|^2 = \|(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\epsilon}\|^2 \\ &= \boldsymbol{\epsilon}^\top(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\epsilon} = \boldsymbol{\epsilon}^\top(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\epsilon}, \\ MSE &:= \frac{SSE}{n-p}, \end{aligned}$$

where $\boldsymbol{I}_n - \boldsymbol{H}$ is a projection matrix.

Question: What is $\mathbb{E}(SSE)$ and $\mathbb{E}(MSE)$ under normal error model, i.e., $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$?

In this part, we use the fact that all eigenvalues of projection matrices are either 0 or 1, and its trace is

$$\mathrm{Tr}(\boldsymbol{H}) = \mathrm{Tr}(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top) = \mathrm{Tr}(\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}) = \mathrm{Tr}(\boldsymbol{I}_p) = p.$$

$$\begin{aligned} \mathbb{E}(SSE) &= \mathbb{E}\,\boldsymbol{\epsilon}^\top(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\epsilon} = \mathbb{E}\,\mathrm{Tr}((\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top) \\ &= \mathrm{Tr}((\boldsymbol{I}-\boldsymbol{H})\,\mathbb{E}\,\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top) = \sigma^2\mathrm{Tr}(\boldsymbol{I}-\boldsymbol{H}) = \sigma^2(n-p). \end{aligned}$$

Therefore, $\mathbb{E}(MSE) = \sigma^2$ is an unbiased estimator of σ^2.

Lemma 6.4.4. *Under normal error model, the distribution of SSE is*

$$SSE \sim \sigma^2\chi^2_{n-p}$$

and is independent of $\widehat{\boldsymbol{\beta}}$.

Proof: Note that

$$SSE = \boldsymbol{\epsilon}^\top(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0,\sigma^2\boldsymbol{I}_n).$$

Applying spectral decomposition to $\boldsymbol{I}-\boldsymbol{H} = \boldsymbol{U}\boldsymbol{U}^\top$ where $\boldsymbol{U} \in \mathbb{R}^{n\times(n-p)}$ and $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_{n-p}$ (identity matrix of size $n-p$). Thus

$$SSE = \boldsymbol{\epsilon}^\top\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{\epsilon} = \|\boldsymbol{U}^\top\boldsymbol{\epsilon}\|^2$$

Note that $\boldsymbol{U}^\top\boldsymbol{\epsilon}$ is $\mathcal{N}(0,\sigma^2\boldsymbol{I}_{n-p})$. Therefore, SSE/σ^2 is the sum of $n-p$ independent squared standard normal random variables, i.e., χ^2_{n-p}.

On the other hand,

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}) \\ &= \boldsymbol{\beta} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\epsilon}. \end{aligned}$$

We can see that $\widehat{\boldsymbol{\epsilon}} = (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\epsilon}$ and $\widehat{\boldsymbol{\beta}}$ are jointly normal. Why? Simply speaking, $\widehat{\boldsymbol{\epsilon}}$ and $\widehat{\boldsymbol{\beta}}$ can be obtained by applying a linear transform to $\boldsymbol{\epsilon}$:

$$\begin{bmatrix} \widehat{\boldsymbol{\epsilon}} \\ \widehat{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} (\boldsymbol{I}-\boldsymbol{H}) \\ (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top \end{bmatrix}\boldsymbol{\epsilon} + \begin{bmatrix} 0 \\ \boldsymbol{\beta} \end{bmatrix}$$

By the invariance of normal random vectors under linear transform, we know that $\widehat{\boldsymbol{\epsilon}}$ and $\widehat{\boldsymbol{\beta}}$ are jointly normal.

Moreover, they are independent since they are uncorrelated:

$$\begin{aligned}\mathrm{Cov}(\widehat{\boldsymbol{\epsilon}}, \widehat{\boldsymbol{\beta}}) &= \mathbb{E}\,\widehat{\boldsymbol{\epsilon}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \\ &= \mathbb{E}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = 0.\end{aligned}$$

□

Exercise: Are $\widehat{\epsilon}_i$ mutually independent?

Note that under normal error model, MSE is an unbiased and consistent estimator of σ^2. Therefore, we can have the following asymptotic distribution of $\widehat{\boldsymbol{\beta}}$.

Lemma 6.4.5. *Under normal error model, it holds approximately that*

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, MSE(\boldsymbol{X}^\top \boldsymbol{X})^{-1}).$$

For each $\widehat{\beta}_i$, we can derive a similar result for the distribution of each $\widehat{\beta}_i$:

$$\frac{\widehat{\beta}_i - \beta_i}{\sqrt{MSE[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}]_{ii}}} \sim t_{n-p}.$$

Proof: We will use a similar argument as the scenario for $p = 2$:

$$\begin{aligned}\frac{\widehat{\beta}_i - \beta_i}{\sqrt{MSE[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}]_{ii}}} &= \frac{(\widehat{\beta}_i - \beta_i)/\sqrt{\sigma^2[\boldsymbol{X}^\top \boldsymbol{X}]_{ii}^{-1}}}{\sqrt{MSE[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}]_{ii}}/\sqrt{\sigma^2[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}]_{ii}}} \\ &= \frac{(\widehat{\beta}_i - \beta_i)/\sqrt{\sigma^2[\boldsymbol{X}^\top \boldsymbol{X}]_{ii}^{-1}}}{\sqrt{MSE/\sigma^2}} \\ &= \frac{(\widehat{\beta}_i - \beta_i)/\sqrt{\sigma^2[\boldsymbol{X}^\top \boldsymbol{X}]_{ii}^{-1}}}{\sqrt{SSE/(n-p)\sigma^2}}\end{aligned}$$

Note that SSE is independent of $\widehat{\boldsymbol{\beta}}$ and $SSE \sim \sigma^2 \chi^2_{n-p}$. By definition of Student t-distribution,

$$\frac{\widehat{\beta}_i - \beta_i}{\sqrt{MSE[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}]_{ii}}} \sim t_{n-p}.$$

□

With that, we are able to construct confidence interval for β_i and perform hypothesis testing.

- An approximately $1 - \alpha$ confidence interval of each β_i is

$$\left(\widehat{\beta}_i - t_{n-p,1-\frac{\alpha}{2}}\sqrt{MSE(\boldsymbol{X}^\top \boldsymbol{X})_{ii}^{-1}}, \widehat{\beta}_i + t_{n-p,1-\frac{\alpha}{2}}\sqrt{MSE(\boldsymbol{X}^\top \boldsymbol{X})_{ii}^{-1}}\right)$$

- A test of size α for $H_0 : \beta_i = 0$ v.s. $H_1 : \beta_i \neq 0$ is to reject H_0 if

$$\left|\frac{\widehat{\beta}_i}{\sqrt{MSE[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}]_{ii}}}\right| > t_{n-p,1-\frac{\alpha}{2}}$$

6.5 Model diagnostics

We have discussed statistical inferences for linear models: perform interval estimation and hypothesis testing for coefficients. The key ingredient is that we assume the underlying model is a linear model with normal error. Let's recall the simple linear model with normal error:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), \quad 1 \leq i \leq n.$$

All stated assumptions are crucial: linear regression relation and i.i.d. normal error. However, in practice, the data are unlikely to satisfy these assumptions which may cause inaccuracy in our inference. This is the reason why we need to consider model diagnostics and come up with remedies to solve these issues.

There are a few typical violations for the simple linear model with normal error.

- Nonlinearity of the regression relation
- Nonconstant variance of the error terms
- Non-normality of the error terms
- Independence of the error terms
- Existence of outliers

In this lecture, we will briefly discuss model diagnostics by using residual plots. Due to the time limitation, we will focus on the first three issues and propose solutions. We will not discuss how to handle outliers and influential cases, which are also very important in practice.

6.5.1 Nonlinearity in the regression relation:

Let's consider a synthetic data:

$$Y_i = 10X_i + 0.2X_i^2 + \epsilon_i, \quad 1 \leq i \leq 100 \tag{6.5.1}$$

where

$$X_i \sim \mathcal{N}(10, 1), \quad \epsilon_i \sim \mathcal{N}(0, 1)$$

If we simply look at Figure 6.1, the scatterplot of (X_i, Y_i), it seems linear model is quite adequate. Let's fit a linear model to the data. Under the simple regression with normal error, the residuals $\widehat{\epsilon_i} = Y_i - \widehat{Y_i}$ are multivariate normal and independent of the fitted value $\widehat{Y_i}$,

$$\widehat{\boldsymbol{\epsilon}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\epsilon}, \quad \widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}.$$

Therefore, the residuals v.s. fitted values plot should have no distinct patterns, i.e., residuals should spread equally around a horizontal line.

What do we observe in Figure 6.2? We can see a nonlinear pattern between the residuals and fitted values, which is not captured by a linear model. Therefore, one needs to take nonlinearity into consideration, such as adding nonlinear predictor, i.e., consider polynomial regression,

$$Y_i = \beta_0 + \beta_1 X_i + \cdots \beta_p X_i^p + \epsilon_i, \quad 1 \leq i \leq n,$$

which is an important example of multiple regression.

Adding the extra quadratic term to the simple linear model indeed makes the pattern disappear, i.e., the red line becomes horizontal and the residuals spread equally around the fitted values.

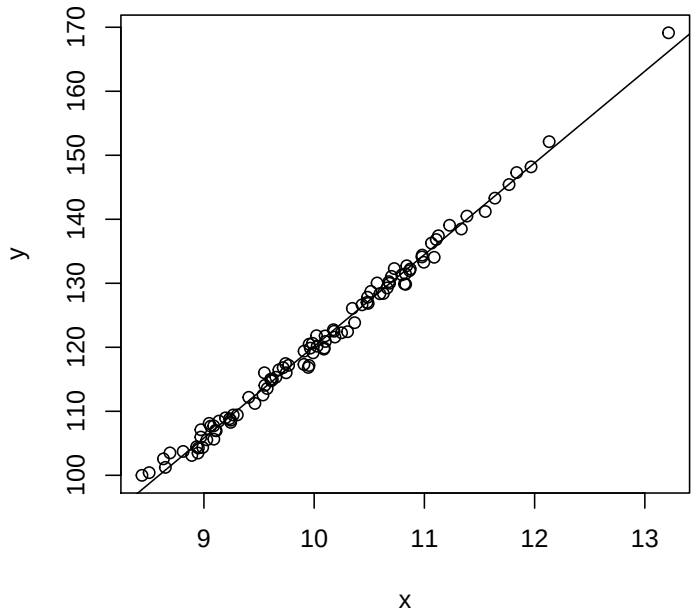


Figure 6.1: Scatterplot for (X_i, Y_i) drawn from $Y_i = 10X_i + 0.2X_i^2 + \epsilon_i$

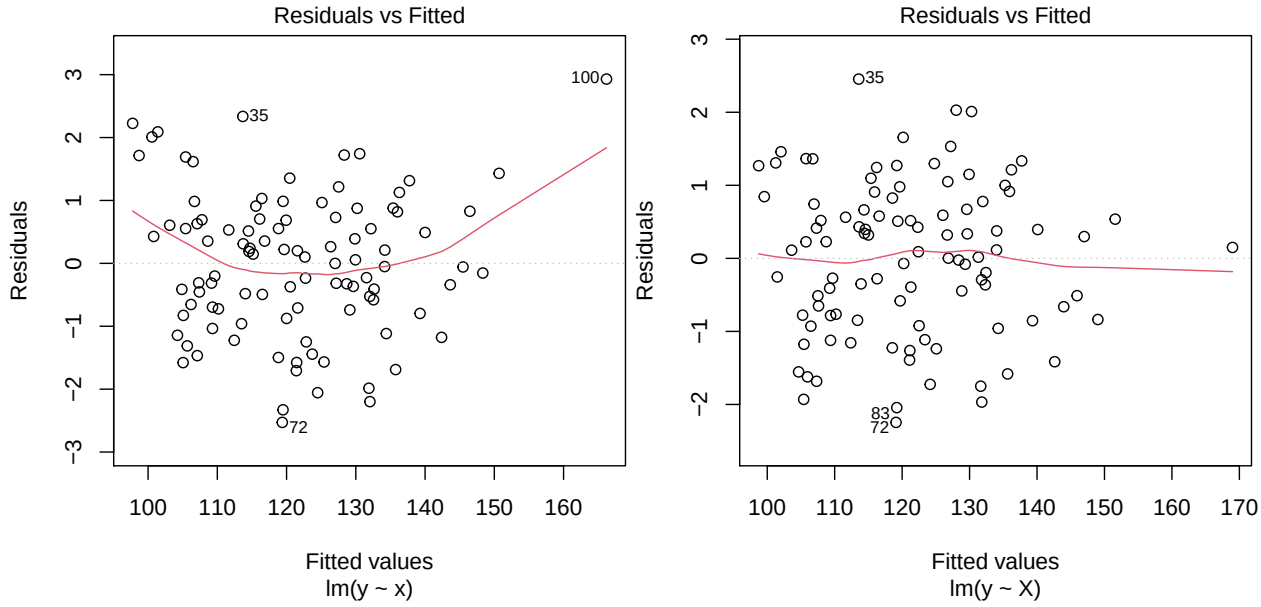


Figure 6.2: Residuals v.s. fitted values for $Y_i = 10X_i + 0.2X_i^2 + \epsilon_i$. Left: without adding quadratic term; right: after adding quadratic term

6.5.2 Error terms with non-constant variance

Consider another synthetic data model:

$$Y_i = 2 + 3X_i + e^{0.1+0.2X}\epsilon_i \tag{6.5.2}$$

where $\epsilon_i \sim \mathcal{N}(0, 1)$. In the simulation, we let $X_i \sim$Unif[0,10]. We first obtain the scatterplot in Figure 6.3, and also fit a linear model and get the residuals v.s. fitted plot. We

can see that the residuals are spreading out more as the fitted value increases.

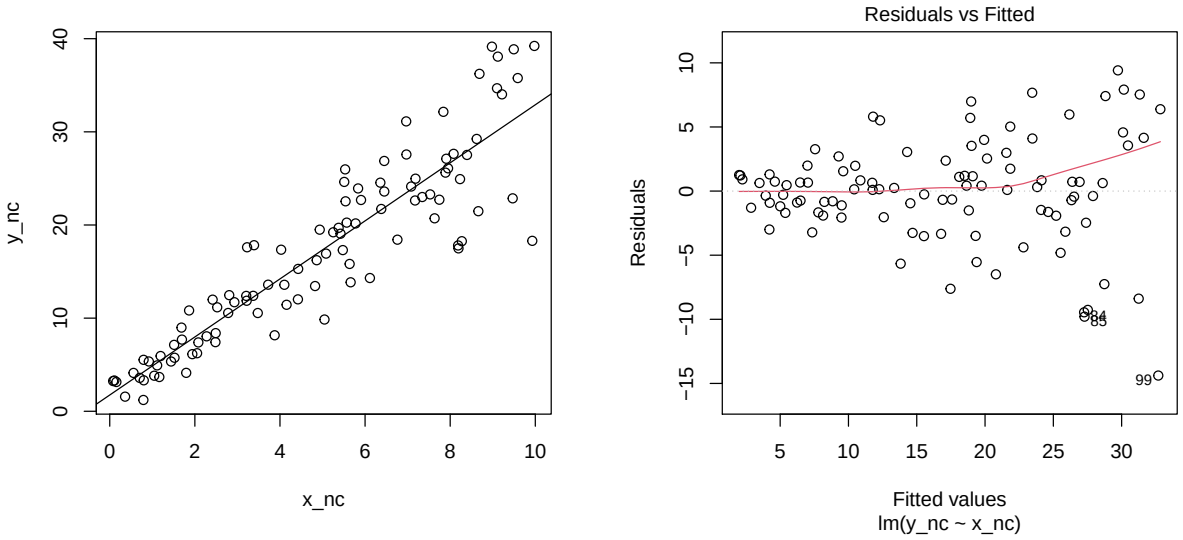


Figure 6.3: Scatterplot for $Y_i = 2 + 3X_i + e^{0.1+0.2X}\epsilon_i$

6.5.3 QQ-plot: Non-normality of error terms

How to verify if the error terms are normal? A direct way is to look at the residuals: if the error terms are i.i.d. normal, then the residuals are "close" to "i.i.d." normal. In practice, we simply test if the residuals are samples from normal distributions. In hypothesis testing, we actually have learnt how to test the normality of a given dataset, such as Kolmogorov-Smirnov tests, goodness-of-fit χ^2 tests, and likelihood ratio tests. These tests are of course applicable to our setting. In practice, one would simply use the normal quantile plot, also known as QQ-plot, to test if the residuals are normal.

- Suppose we observe $\{x_i\}_{i=1}^n$ and sort them in an increasing order,

  $$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}.$$

  Moreover, we standardize them by updating each x_i with $(x_i - \overline{x}_n)/S_n$.

- An approximation of the expected value for the k-th smallest $x_{(k)}$ (order statistics) under $\mathcal{N}(0, \widehat{\sigma}^2)$ is approximately

  $$z_{(k)} = \Phi^{-1}\left(\frac{k}{n+1}\right), \quad k = 1, 2, \cdots, n,$$

  where $\Phi^{-1}(\alpha)$ is the α-quantile of $\mathcal{N}(0,1)$ distribution.

- Then plot $(z_{(k)}, x_{(k)})$, $1 \le k \le n$.

If the data are indeed normal, the plot should be close to a straight line $y = x$, see Figure 6.4. Why? Since if $\{x_k\}$ are samples from standard normal, then

$$x_{(k)} = \Phi_n^{-1}\left(\frac{k}{n}\right)$$

where Φ_n is the empirical cdf of $\{x_k\}_{k=1}^n$. Note that Φ_n converges to Φ in probability. As a result, $x_{(k)}$ should be very close to $z_{(k)}$ and thus $(z_{(k)}, x_{(k)})$ is approximately on the line $y = x$.

To test the normality of error terms in the linear models, we can simply obtain the residuals and plot their QQ-plot. The line in the QQ-plot passes the first and third quartiles.

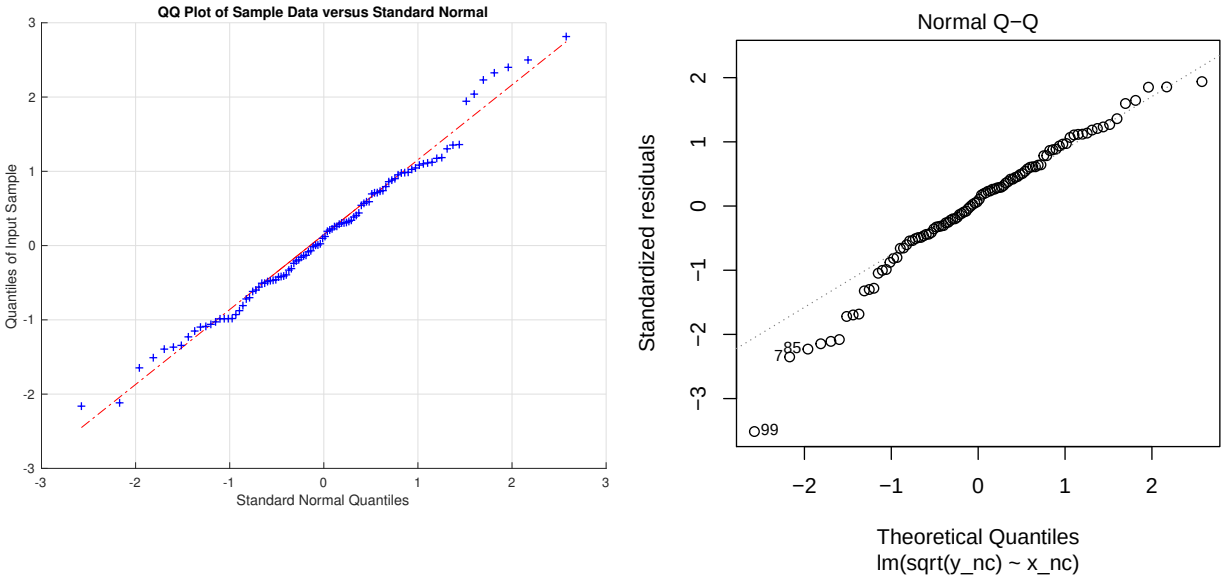
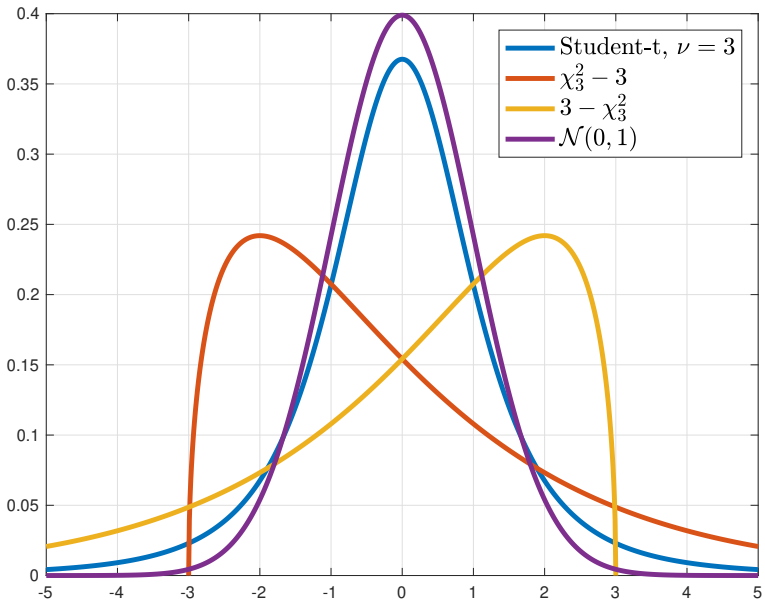


Figure 6.4: QQ-plot for normal data and for the residuals in Model 6.5.2

Question: How to read QQ-plots?

Let's investigate four common distributions and their corresponding QQ plots.



- Student-t: heavy-tailed
- $\chi^2_3 - 3$: right skewed, the right tail is longer; the mass of the distribution is concentrated on the left of the figure
- $3 - \chi^2_3$: left skewed, the left tail is longer; the mass of the distribution is concentrated on the right of the figure.
- $\mathcal{N}(0,1)$: standard normal distribution.

QQ-plot Example 1:

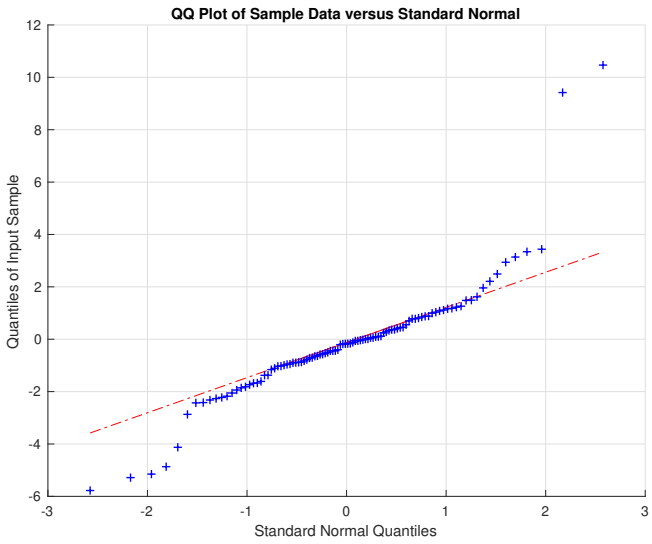


Figure 6.5: QQ plot for Student t distribution

QQ plot shows more probabilities in the tails than a normal distribution. It is student-t distribution.

QQ-plot Example 2:

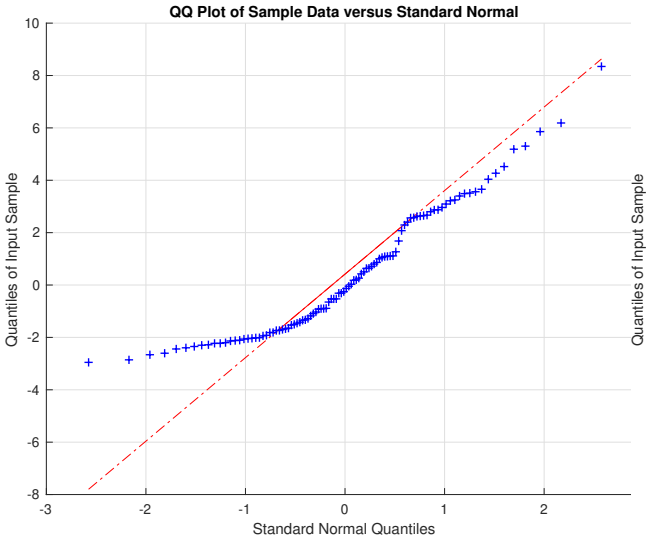


Figure 6.6: QQ plot for $\chi^2_3 - 3$

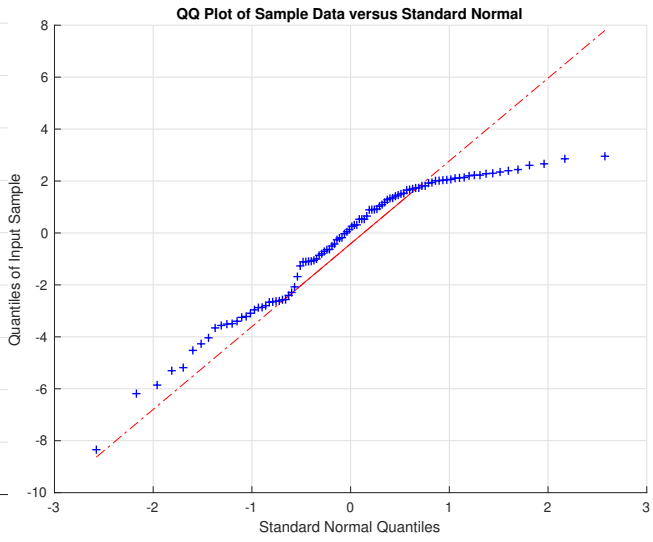


Figure 6.7: QQ plot for $3 - \chi^2_3$

Figure 6.6 left: QQ plot shows more probabilities in the right tail and less probabilities in the left tail. It is $\chi^2_3 - 3$ distribution; Figure 6.7 right: Q-Q plot shows more probabilities in the left tail and less probabilities in the right tail. It is $3 - \chi^2_3$ distribution.

6.5.4 Box-Cox transform

Question: How to solve the issue of non-constant and non-normal error?

In applied statistics, the Box-Cox procedure, a.k.a. power transformation, provides a family of transformations on the response variable Y (also may apply to the predictors)

such that the resulting model is close to a linear model with normal error. More precisely, we perform a power transform on the response Y_i using the following form:

$$Y' = Y^\lambda$$

where λ is a parameter to be determined. Here we first assume the response is positive for simplicity. In particular, if $\lambda = 0$, $Y' = \log(Y)$.

The normal error regression model with the response variable a member of the family of power transformations becomes:

$$Y_i^\lambda = \beta_0 + \beta_1 X_i + \epsilon_i.$$

Essentially, we treat λ as an additional variable and hope to identify the best λ (as well as $\boldsymbol{\beta}$ and σ^2) such that the model fits the data.

Question: How to identify a suitable λ?

Denote $Y_i(\lambda) = Y_i^\lambda$. Then under the normal error model, the joint pdf for $Y_i(\lambda)$ is

$$f_{\boldsymbol{Y}(\lambda)}(Y_i(\lambda); \boldsymbol{\beta}, \lambda, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \left(-\frac{1}{2\sigma^2} \left(\beta_0 + \beta_1 X_i - Y_i(\lambda)\right)^2 \right).$$

Recall that we only observe Y_i instead of $Y_i(\lambda)$; to derive a likelihood function based on Y_i, we need to perform a change of variable, i.e., obtaining the joint pdf of Y_i. This can be done by introducing a Jacobian factor.

$$\begin{aligned} f_{\boldsymbol{Y}}(Y_i; \boldsymbol{\beta}, \lambda, \sigma^2) &= f_{\boldsymbol{Y}(\lambda)}(Y_i(\lambda); \boldsymbol{\beta}, \lambda, \sigma^2) \cdot \prod_{i=1}^n \frac{\mathrm{d}Y_i(\lambda)}{\mathrm{d}Y_i} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \left(-\frac{1}{2\sigma^2} \left(\beta_0 + \beta_1 X_i - Y_i(\lambda)\right)^2 \right) \prod_{i=1}^n \lambda Y_i^{\lambda-1} \end{aligned}$$

where

$$\frac{\mathrm{d}Y_i(\lambda)}{\mathrm{d}Y_i} = \frac{\mathrm{d}}{\mathrm{d}Y_i} Y_i^\lambda = \lambda Y_i^{\lambda-1}.$$

Denote $K = (\prod_i Y_i)^{1/n}$ as the geometric mean of $\{Y_i\}_{i=1}^n$. The log-likelihood function is

$$\ell(\lambda, \boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Y}(\lambda)\|^2 + n \log(K^{\lambda-1}\lambda).$$

Note that the MLE for $\boldsymbol{\beta}$ is $(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}(\lambda)$ and the MLE for σ^2 is

$$\widehat{\sigma}^2(\lambda) = \frac{1}{n} \|(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}(\lambda)\|^2.$$

Therefore, we have

$$\begin{aligned} \ell(\lambda) &:= \sup_{\boldsymbol{\beta}, \sigma^2} \ell(\lambda, \boldsymbol{\beta}, \sigma^2) \\ &= -\frac{n}{2} \log \|(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}(\lambda)\|^2 - \frac{1}{2\widehat{\sigma}^2(\lambda)} \|(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}(\lambda)\|^2 + n \log(K^{\lambda-1}\lambda) \\ &= -\frac{n}{2} \log \|(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}(\lambda)\|^2 + n \log(K^{\lambda-1}\lambda) + C \\ &= -\frac{n}{2} \log \left\| \frac{(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}(\lambda)}{K^{\lambda-1}\lambda} \right\|^2 \end{aligned}$$

where C is a constant. For each λ, the log-likelihood function equals to

$$\ell(\lambda) = -\frac{n}{2}\log SSE(\lambda)$$

where

$$SSE(\lambda) = \left\| \frac{(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}(\lambda)}{K^{\lambda-1}\lambda} \right\|^2.$$

If we choose to normalize Y_i^λ via

$$Y_i'(\lambda) = \begin{cases} K^{1-\lambda}(Y_i^\lambda - 1)/\lambda, & \lambda \neq 0, \\ K\log(Y_i), & \lambda = 0, \end{cases}$$

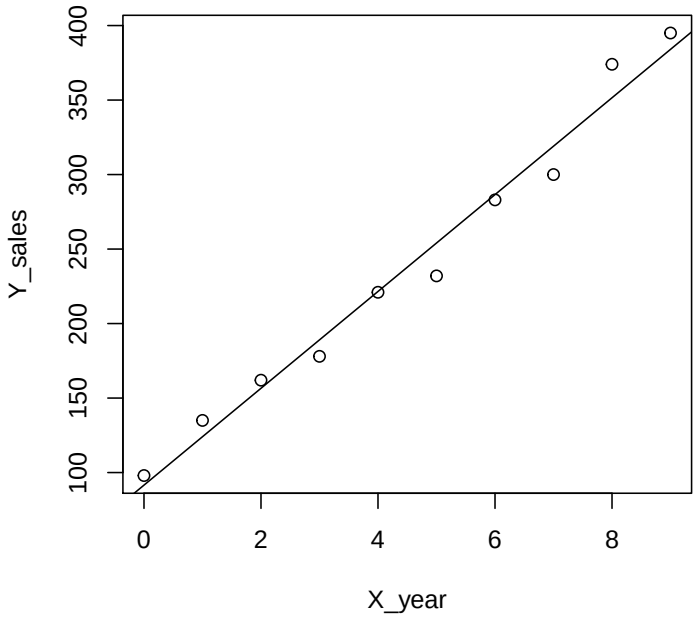
where $K = (\prod_i Y_i)^{1/n}$ is the geometric mean of $\{Y_i\}_{i=1}^n$. Then we try to identify a suitable λ such that the SSE associated with $Y_i'(\lambda)$, i.e., $SSE(\lambda)$, is the smallest.

Many statistical packages provides existing built-in function to compute this likelihood function w.r.t. λ. In R, the function is boxCox().

Example: A marketing researcher studied annual sales of a product that had been introduced 10 years ago. If we fit a linear model, we may feel that a linear model is quite

Year	0	1	2	3	4	5	6	7	8	9
Sales (thousands of unit)	98	135	162	178	221	232	283	300	374	395

adequate for the data. Now let's perform model diagnostics by plotting its residuals v.s. fitted and QQ plot.



What do you observe? There is a pattern (nonlinearity) in the residuals v.s. fitted plot. How to deal with it? We perform a box-cox transform.

Figure 6.10 indicates $\lambda = 1/2$ is the best choice. We use $\lambda = 1/2$, i.e., $Y' = \sqrt{Y}$ and use the following new model:

$$\sqrt{Y_i} = \beta_0 + \beta_1 X_i + \epsilon_i.$$

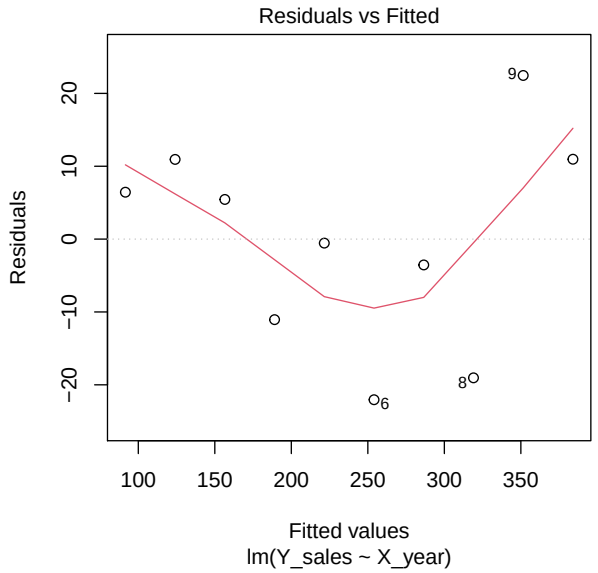


Figure 6.8: Residuals v.s. fitted

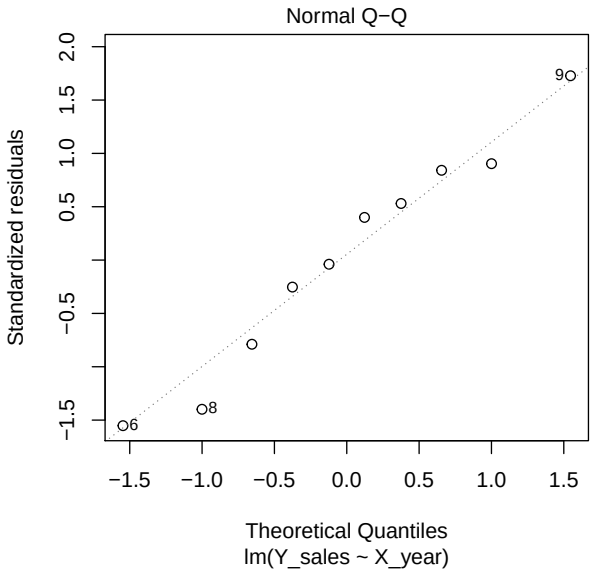


Figure 6.9: QQ plot

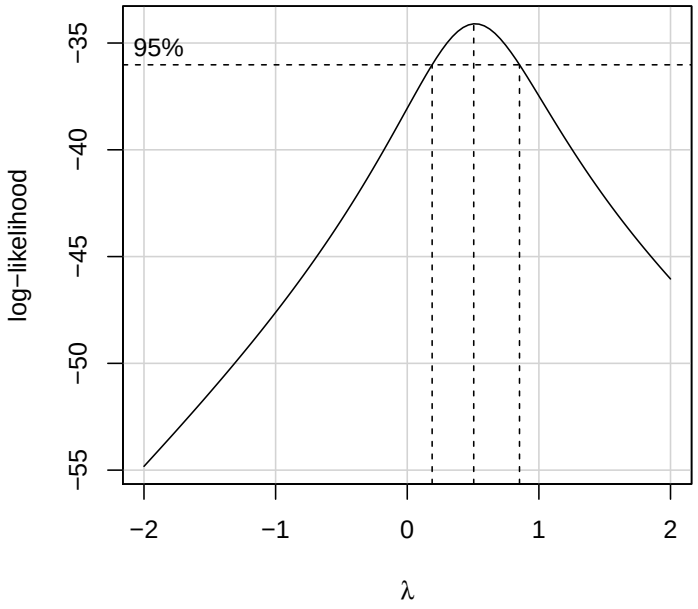


Figure 6.10: Log-likelihood function $\ell(\lambda)$

Written in the original form, the response Y_i satisfies

$$Y_i = (\beta_0 + \beta_1 X_i + \epsilon_i)^2.$$

We fit the model again and perform a model diagnostic. Figure 6.11 implies that the model fits the data well: the residuals are approximately normal and also spread equally around the fitted values.

Finally, we briefly discuss what if some of the data Y_i are negative. In fact, a more general

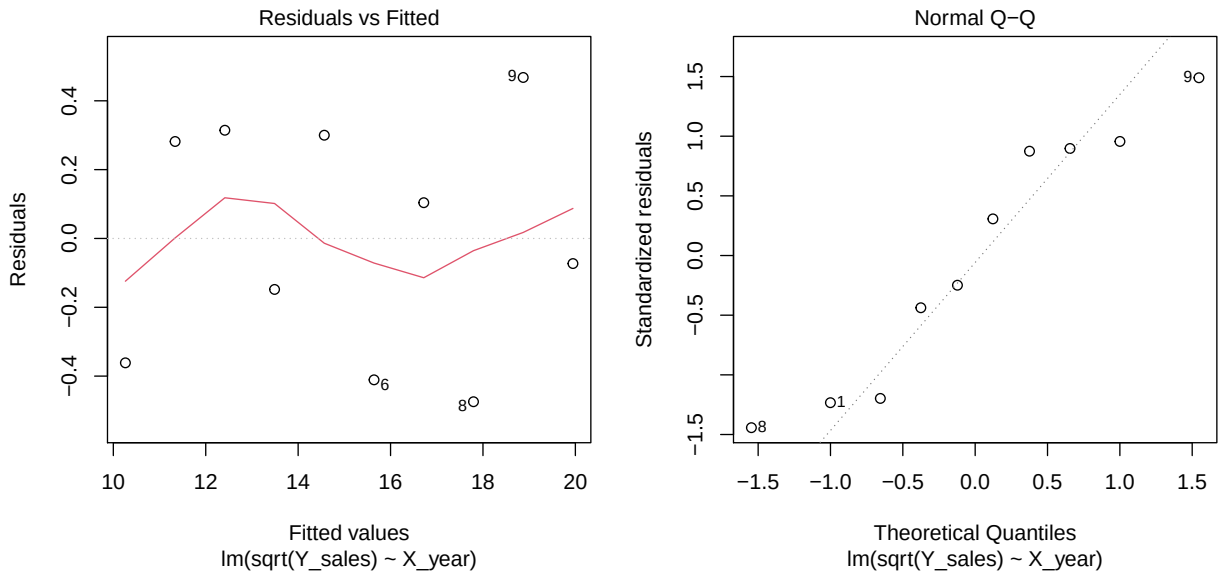


Figure 6.11: Residuals v.s. fitted and QQ plot for the transformed model

version of Box-Cox transform involves one more shift parameter:

$$Y' = (Y + \lambda_2)^{\lambda_1},$$

which is a family of transform with two parameters. Then we can follow the similar procedure to obtain (λ_1, λ_2) which maximizes the likelihood function, and use them to transform the response.

6.6 Logistic regression

In 1986, the space shuttle Challenger exploded during take off, killing seven astronauts aboard. The explosion was the result of an O-ring failure, a splitting of a ring of rubber that seals the parts of the ship together. It is believed that the failure of the rubber depends on the temperature.

Flight	14	9	23	10	1	5	13	15	4	3	8	17	2	11	6	7	16	21	19	22	12	20	18
Failure	1	1	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0
Temp.	53	57	58	63	66	67	67	67	68	69	70	70	70	70	72	73	75	75	76	76	78	79	81

Figure 6.12: Data from Example 1.13 [Robert, Casella, 2004]: 1 stands for failure, 0 for success.

Goal: What is the relationship between the predictor and response?

- The accident was believed to be caused by the cold weather at the time of launch
- The probability of failure increases as the temperature decreases

Question: Can we use linear regression? No, the response is discrete. However, we believe that the probability of failure may be a *linear* function of temperature X_i.

Suppose we observe (X_i, Y_i), $1 \leq i \leq n$ with binary response Y_i is 0 or 1. The logistic

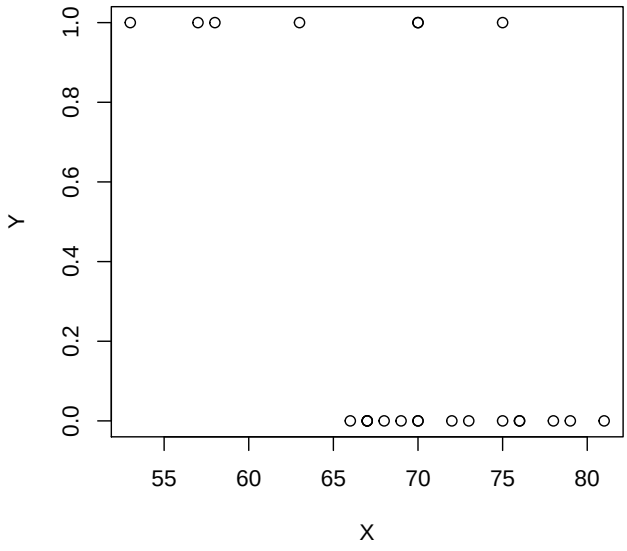


Figure 6.13: Plot of failure v.s. temperature

regression is: the response is a binary random variable depending on the predictor,

$$\mathbb{P}(Y_i = 1|X_i) = \frac{e^{\beta_0+\beta_1 X_i}}{1 + e^{\beta_0+\beta_1 X_i}},$$
$$\mathbb{P}(Y_i = 0|X_i) = \frac{1}{1 + e^{\beta_0+\beta_1 X_i}}.$$

What does it mean? The distribution of Y given X is Bernoulli(p) where

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

The following function is called *logistic function*:

$$f(x) = \frac{e^x}{1 + e^x}$$

In other words, $p(X) = f(\beta_0 + \beta_1 X)$. How does $f(x)$ look like?

- $f(x)$ is increasing,
- $f(x)$ takes value between 0 and 1.

Logistic regression is an example of *generalized linear model*:

$$\beta_0 + \beta_1 X_i = \log\left(\frac{p_i}{1 - p_i}\right) = \text{logit}(p_i)$$

where

$$\text{logit}(x) = \log\left(\frac{x}{1 - x}\right), \quad 0 < x < 1$$

is called logit function, which is a canonical *link* function for binary response.

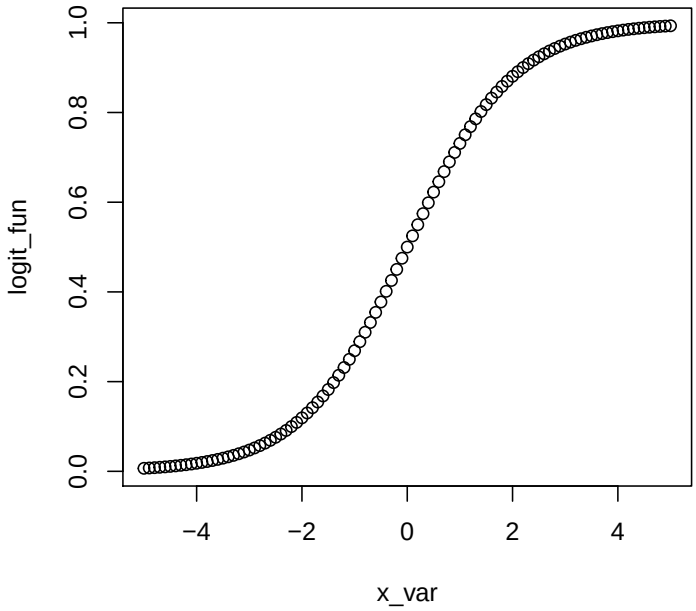


Figure 6.14: Logistic function

6.6.1 Maximum likelihood estimation

Suppose we observe $Y_i \sim$Bernoulli(p_i) where $\text{logit}(p_i) = \beta_0 + \beta_1 X_i$, how to obtain an estimation of $\boldsymbol{\beta}$? The common approach is the maximum likelihood estimation. The joint distribution of Y_i under (β_0, β_1) is

$$f(\boldsymbol{Y}; \beta_0, \beta_1) = \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i}, \quad p_i = \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}$$

The log-likelihood function is

$$\begin{aligned}
\ell(\beta_0, \beta_1) &= \sum_{i=1}^{n} Y_i \log p_i + (1-Y_i)\log(1-p_i) \\
&= \sum_{i=1}^{n} Y_i \log \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}} + (1-Y_i)\log \frac{1}{1+e^{\beta_0+\beta_1 X_i}} \\
&= \sum_{i=1}^{n} Y_i(\beta_0+\beta_1 X_i) - \log(1+e^{\beta_0+\beta_1 X_i})
\end{aligned}$$

The MLE is the maximizer to $\ell(\beta_0, \beta_1)$:

$$\widehat{\boldsymbol{\beta}}_{MLE} = \text{argmax}_{\boldsymbol{\beta}\in\mathbb{R}^2}\, \ell(\boldsymbol{\beta}), \quad \boldsymbol{\beta} = [\beta_0, \beta_1]^\top.$$

Question: Can we find the global maximizer to $\ell(\boldsymbol{\beta})$?

The log-likelihood function $\ell(\boldsymbol{\beta})$ is concave. It is very straightforward to verify the concavity of $\ell(\boldsymbol{\beta})$. The gradient is

$$\frac{\partial \ell}{\partial \beta_0} = \sum_{i=1}^{n}\left(Y_i - 1 + \frac{1}{1+e^{\beta_0+\beta_1 X_i}}\right), \quad \frac{\partial \ell}{\partial \beta_1} = \sum_{i=1}^{n}\left(Y_i - 1 + \frac{1}{1+e^{\beta_0+\beta_1 X_i}}\right) X_i$$

For the Hessian matrix, we have

$$\frac{\partial^2 \ell}{\partial \beta_0^2} = -\sum_{i=1}^{n} \frac{e^{\beta_0+\beta_1 X_i}}{(1+e^{\beta_0+\beta_1 X_i})^2}, \quad \frac{\partial^2 \ell}{\partial \beta_1^2} = -\sum_{i=1}^{n} \frac{e^{\beta_0+\beta_1 X_i}}{(1+e^{\beta_0+\beta_1 X_i})^2} X_i^2$$

and

$$\frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_1} = -\sum_{i=1}^{n} \frac{e^{\beta_0+\beta_1 X_i}}{(1+e^{\beta_0+\beta_1 X_i})^2} X_i$$

Therefore, the Hessian matrix equals

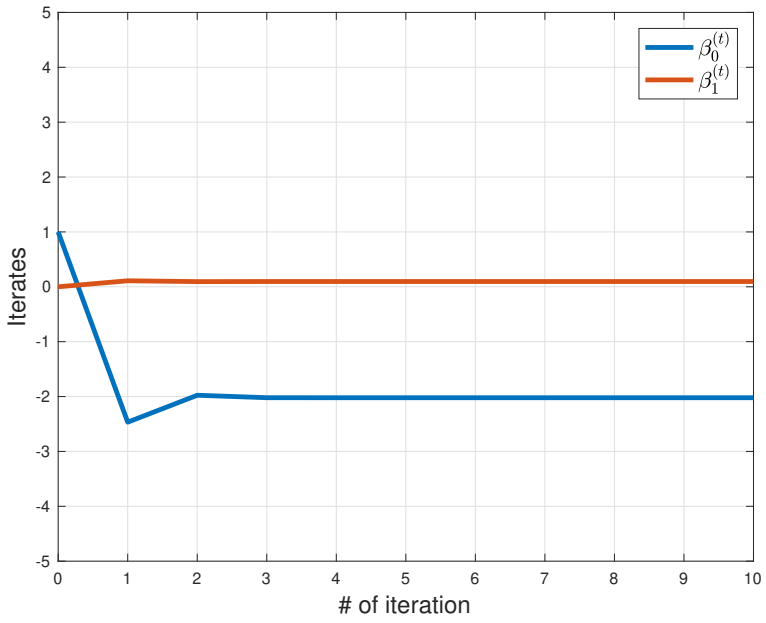
$$\nabla^2 \ell = -\begin{bmatrix} \sum_{i=1}^n p_i(1-p_i) & \sum_{i=1}^n p_i(1-p_i)X_i \\ \sum_{i=1}^n p_i(1-p_i)X_i & \sum_{i=1}^n p_i(1-p_i)X_i^2 \end{bmatrix}$$

Exercise: Show that the Hessian is negative semidefinite.

Question: How to optimize this function? One can use Newton's method:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - [\nabla^2 \ell(\boldsymbol{\beta}^{(t)})]^{-1} \nabla \ell(\boldsymbol{\beta}^{(t)})$$

Stop until $\boldsymbol{\beta}^{(t)}$ stablizes.



Most programs have build-in packages to get an estimation of $\boldsymbol{\beta}$. Call the "*glm*" function in R.

```
glm_fit <- glm(Y~X, data = data, family = binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  15.0429     7.3786   2.039   0.0415 *
X            -0.2322     0.1082  -2.145   0.0320 *
```

The estimated coefficients of Challenger data are

$$\widehat{\beta}_0 = 15.0429, \quad \widehat{\beta}_1 = -0.2322.$$

The predicted value of probability p_i at X_i is

$$\widehat{p}_i = \frac{e^{\widehat{\beta}_0+\widehat{\beta}_1 X_i}}{1+e^{\widehat{\beta}_0+\widehat{\beta}_1 X_i}}.$$

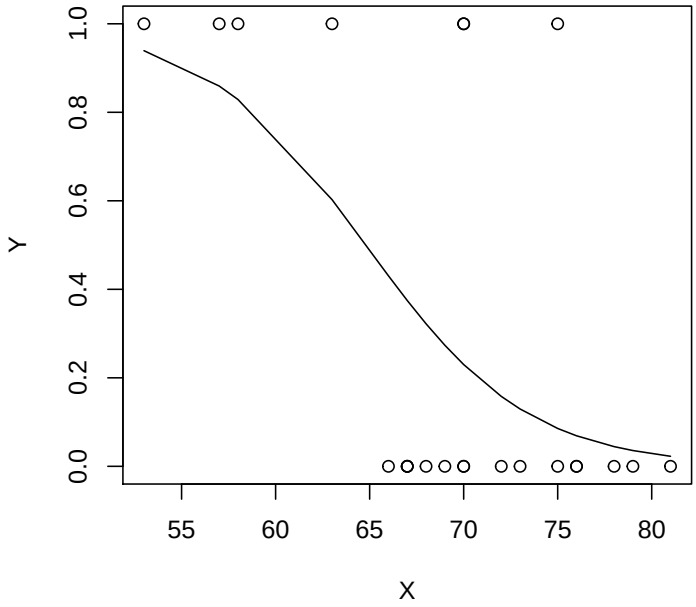


Figure 6.15: Fitted value v.s. predictor X_i

We plot $(X_i, \widehat{p}_i)$ in Figure 6.15 where $\widehat{p}_i$ is the fitted value of probability and the curve "decreases" as X gets larger.

How to interpret β_0 and β_1? The odds is given by

$$\text{odds}(X) = \frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}.$$

Now let's consider

$$\frac{\text{odds}(X+1)}{\text{odds}(X)} = \frac{\frac{p(X+1)}{1-p(X+1)}}{\frac{p(X)}{1-p(X)}} = \frac{e^{\beta_0+\beta_1(X+1)}}{e^{\beta_0+\beta_1 X}} = e^{\beta_1}.$$

In other words, the odds at $X+1$ is about e^{β_1} times the odds at X. In the example of Shuttle Challenger, $\widehat{\beta}_1$ is -0.232 and

$$e^{\widehat{\beta}_1} = e^{-0.232} \approx 0.793.$$

The odds decreases by a factor of 0.793 when the temperature increases by 1 ^oF.

6.6.2 Inference in logistic regression

Question: Can we conclude that a decreased temperature leads to higher failure probability? This leads to a hypothesis testing problem:

$$H_0 : \beta_1 < 0 \quad \text{versus} \quad H_1 : \beta_1 \geq 0.$$

To derive the rejection region, we need to know the distribution of the MLE.

Theorem 6.6.1. *The MLE enjoys*

- *consistency: as* $n \to \infty$,
$$\widehat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}$$
where $\boldsymbol{\beta}$ *is the underlying parameter.*

- *asymptotic normality:*
$$\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \xrightarrow{d} \mathcal{N}(0, [\boldsymbol{I}(\boldsymbol{\beta})]^{-1})$$
where $\boldsymbol{I}(\boldsymbol{\beta})$ *is the Fisher information.*

The Fisher information matrix is the negative inverse of the second order derivative of log-likelihood function evaluated at $\boldsymbol{\beta}$, i.e.,

$$\boldsymbol{I}(\boldsymbol{\beta}) = -\,\mathbb{E}\,\nabla^2 \ell(\boldsymbol{\beta}) = -\nabla^2 \ell(\boldsymbol{\beta})$$

where
$$\nabla^2 \ell(\boldsymbol{\beta}) = -\begin{bmatrix} \sum_{i=1}^n p_i(1-p_i) & \sum_{i=1}^n p_i(1-p_i)X_i \\ \sum_{i=1}^n p_i(1-p_i)X_i & \sum_{i=1}^n p_i(1-p_i)X_i^2 \end{bmatrix}$$
and $p_i = \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}$. However, the exact value of $\boldsymbol{\beta} = [\beta_0, \beta_1]^\top$ is unknown. In practice, we can approximate $\boldsymbol{\beta}$ via $\widehat{\boldsymbol{\beta}}$ and get an estimate of the information matrix:

$$I(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} \sum_{i=1}^n \widehat{p}_i(1-\widehat{p}_i) & \sum_{i=1}^n \widehat{p}_i(1-\widehat{p}_i)X_i \\ \sum_{i=1}^n \widehat{p}_i(1-\widehat{p}_i)X_i & \sum_{i=1}^n \widehat{p}_i(1-\widehat{p}_i)X_i^2 \end{bmatrix}$$

where $\widehat{p}_i = \frac{e^{\widehat{\beta}_0+\widehat{\beta}_1 X_i}}{1+e^{\widehat{\beta}_0+\widehat{\beta}_1 X_i}}$.

Question: How to obtain a confidence interval for $\boldsymbol{\beta}$?

The standard deviation of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are

$$\widehat{se}(\widehat{\beta}_0) = [I(\widehat{\boldsymbol{\beta}})^{-1}]_{11}, \quad \widehat{se}(\widehat{\beta}_1) = [I(\widehat{\boldsymbol{\beta}})^{-1}]_{22}$$

The asymptotic distribution of β_0 and β_1 is

$$\frac{\widehat{\beta}_0 - \beta_0}{\widehat{se}(\widehat{\beta}_0)} \sim \mathcal{N}(0,1), \quad \frac{\widehat{\beta}_1 - \beta_1}{\widehat{se}(\widehat{\beta}_1)} \sim \mathcal{N}(0,1).$$

An approximate 95% confidence interval of β_0 and β_1 are

$$\widehat{\beta}_i \pm 1.96 \cdot \widehat{se}(\widehat{\beta}_i), \quad i = 0, 1, \quad 1.96 \approx z_{0.975}.$$

Return to the example of Shuttle Challenger,

$$\nabla^2 \ell(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} -3.26 & -221.80 \\ -221.80 & -15163.10 \end{bmatrix}, \quad [\boldsymbol{I}(\widehat{\boldsymbol{\beta}})]^{-1} = \begin{bmatrix} 54.44 & -0.80 \\ -0.80 & 0.0117 \end{bmatrix}$$

An estimation of standard deviation is

$$\widehat{se}(\widehat{\beta}_0) = 7.3786, \quad \widehat{se}(\widehat{\beta}_1) = 0.108.$$

The confidence interval for β_1 is

$$\widehat{\beta}_i \pm 1.96 \cdot \widehat{se}(\widehat{\beta}_i) \iff -0.2322 \pm 1.96 \cdot 0.108 \implies (-0.44, -0.02)$$

Therefore, the confidence interval supports the conclusion that $\beta_1 < 0$.

6.6.3 Hypothesis testing

Question: How to perform hypothesis testing?

z-test: Consider the hypothesis testing problem,

$$H_0 : \beta_i = 0 \quad \text{v.s.} \quad H_1 : \beta_i \neq 0.$$

For each i, we define the z-value as

$$z = \frac{\widehat{\beta}_i}{\widehat{se}(\widehat{\beta}_i)}$$

Why do we use z-value? The z-value is used as a test statistic: recall under H_0,

$$z = \frac{\widehat{\beta}_i}{\widehat{se}(\widehat{\beta}_i)} \sim \mathcal{N}(0,1), \quad i = 0 \text{ or } 1$$

Let's compute the z-value:

$$z = \frac{\widehat{\beta}_1}{\widehat{se}(\widehat{\beta}_1)} = \frac{-0.2322}{0.1082} = -2.144$$

What is the p-value? We reject the null if $|z|$ is too large, the p-value equals

$$\mathbb{P}(|Z| \geq |z|) = 0.032 < 0.05.$$

We reject the null hypothesis $H_0 : \beta_1 = 0$.

Likelihood ratio test: The other alternative way to perform testing is to use LRT. Still we consider

$$H_0 : \beta_1 = 0 \quad \text{v.s.} \quad H_1 : \beta_1 \neq 0.$$

as an example. Under null hypothesis, it holds that

$$\lambda(\boldsymbol{X}) = 2(\sup_{\beta_0,\beta_1} \ell(\beta_0,\beta_1) - \sup_{\beta_0,\beta_1=0} \ell(\beta_0,0)) \to \chi_1^2$$

The log-likelihood under full model is

$$\ell(\widehat{\beta}_0,\widehat{\beta}_1) = -10.16, \quad \ell(\widehat{\beta}_0,0) = -14.78.$$

Therefore,

$$\lambda(\boldsymbol{X}) = 2(-10.16 + 14.13) = 7.95.$$

The p-value is

$$\mathbb{P}(\chi_1^2 \geq 7.95) = 0.005 < 0.05.$$

Therefore, we should reject the H_0.

6.6.4 Repeated observations - Binomial outcomes

Motivation: in some experiments, a number of repeat observations are obtained at several levels of the predictor variable X.

By computing the proportion of coupons redeemed, we have

$$\widehat{\boldsymbol{p}} = (0.15, 0.275, 0.345, 0.5, 0.6875)^\top.$$

Level i	Price reduction X_i	# of households n_i	# of coupons redeemed
1	5	100	15
2	10	120	33
3	15	110	38
4	20	200	100
5	30	160	110

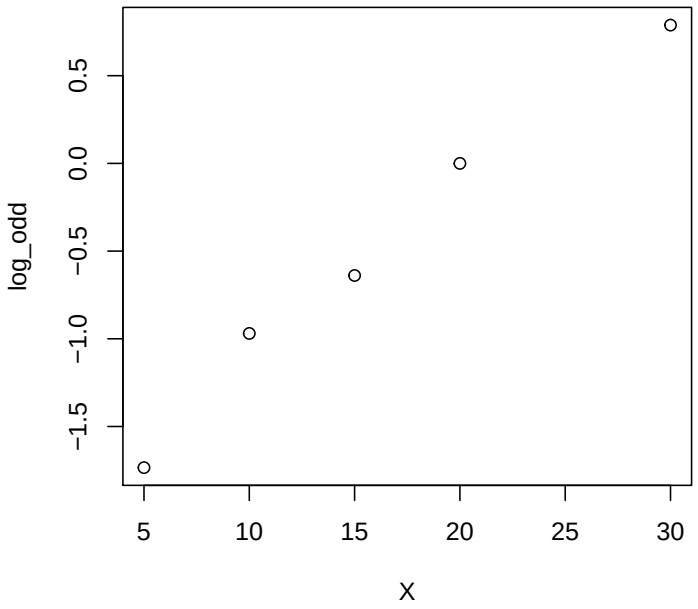


Figure 6.16: $\widehat{p}_i$ v.s. X_i

We plot $\widehat{p}_i$ w.r.t. X_i:

Observation: It seems that $\log(p_i/(1-p_i))$ depends linearly on X_i.

Question: can we fit a logistic regression model to this dataset?

Let Y_{ij} be the jth case at the level i, $1 \le j \le n_i$.

$$Y_{ij} \sim \text{Bernoulli}\left(\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}\right)$$

and we only observe $\sum_{j=1}^{n_i} Y_{ij}$ and n_i.

First, let's derive the log-likelihood function:

$$\begin{aligned}
\ell(\boldsymbol{\beta}) &= \log \prod_{i=1}^{m} \prod_{j=1}^{n_i} p_i^{Y_{ij}} (1-p_i)^{1-Y_{ij}} \\
&= \sum_{i=1}^{m} \left(\sum_{j=1}^{n_i} Y_{ij} \log p_i + (1 - Y_{ij}) \log(1-p_i) \right) \\
&= \sum_{i=1}^{m} \left(\sum_{j=1}^{n_i} Y_{ij} \log p_i + \left(n_i - \sum_{j=1}^{n_i} Y_{ij} \right) \log(1-p_i) \right) \\
&= \sum_{i=1}^{m} n_i (\widehat{p}_i \log(p_i) + (1-\widehat{p}_i) \log(1-p_i))
\end{aligned}$$

where

$$\widehat{p}_i = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}.$$

```
Call:
glm(formula = cbind(N_s, N_f) ~ X, family = "binomial", data = mydata)

Deviance Residuals:
      1        2        3        4        5
-0.7105   0.4334  -0.3098   0.6766  -0.4593

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.02150    0.20908  -9.669   <2e-16 ***
X            0.09629    0.01046   9.203   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

fitted.values(glm_fit)
```

The estimation $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is

$$\widehat{\beta}_0 = -2.02, \quad \widehat{\beta}_1 = 0.096$$

Price reduction X_i	$\widehat{p}_i$	fitted p_i
5	0.15	0.1765
10	0.275	0.2575
15	0.345	0.3595
20	0.5	0.4761
30	0.6875	0.7041

The Y-axis is $\log\left(\frac{p}{1-p}\right)$, i.e., the log of odds. It is almost linear w.r.t. X_i.

6.6.5 General logistic regression

Suppose we have predictors X_{ij}: $1 \le i \le n, 0 \le j \le p-1$. The outcome Y_i is 0 or 1.

- X_{ij}: ith case of predictor j;
- Y_i is binary

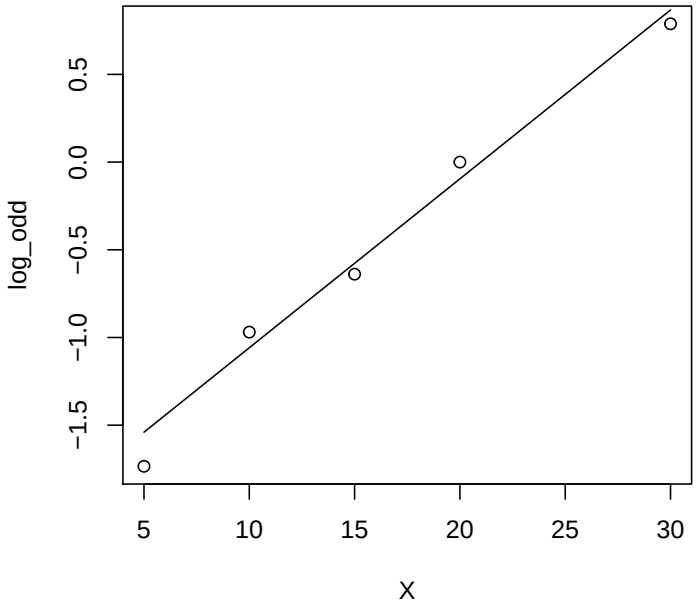


Figure 6.17: $\widehat{p_i}$ v.s. fitted value p_i

A commonly-used model is called logistic regression

$$\mathbb{P}(Y_i = 1|\boldsymbol{X}_i) = \frac{e^{\boldsymbol{X}_i^\top \boldsymbol{\beta}}}{1 + e^{\boldsymbol{X}_i^\top \boldsymbol{\beta}}}$$
$$\mathbb{P}(Y_i = 0|\boldsymbol{X}_i) = \frac{1}{1 + e^{\boldsymbol{X}_i^\top \boldsymbol{\beta}}}$$

where $\boldsymbol{X}_i$ is the i-th case:

$$\boldsymbol{X}_i = (X_{i0}, \cdots, X_{i,p-1})^\top \in \mathbb{R}^p, \quad \boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_{p-1})^\top \in \mathbb{R}^p.$$

Let $f(x) = \frac{e^x}{1+e^x}$ be the logistic function, then

$$Y_i \sim \text{Bernoulli}(f(\boldsymbol{X}_i^\top \boldsymbol{\beta})).$$

In other words, the sampling probability p_i of Y_i is

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{X}_i^\top \boldsymbol{\beta}, \quad p_i = \frac{e^{\boldsymbol{X}_i^\top \boldsymbol{\beta}}}{1 + e^{\boldsymbol{X}_i^\top \boldsymbol{\beta}}}$$

To estimate $\boldsymbol{\beta}$, we maximize the likelihood function:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i}.$$

The log-likelihood function is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i \log p_i + (1 - Y_i)\log(1 - p_i).$$

Recall the likelihood function is

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} Y_i \log \frac{e^{\boldsymbol{X}_i^\top \boldsymbol{\beta}}}{1+e^{\boldsymbol{X}_i^\top \boldsymbol{\beta}}} + (1-Y_i)\log \frac{1}{1+e^{\boldsymbol{X}_i^\top \boldsymbol{\beta}}} \\ &= \sum_{i=1}^{n} \left(\boldsymbol{X}_i^\top \boldsymbol{\beta} \cdot Y_i - \log(1+e^{\boldsymbol{X}_i^\top \boldsymbol{\beta}}) \right)\end{aligned}$$

How to find out the minimizer of this program?

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left(Y_i \boldsymbol{X}_i - \frac{e^{\boldsymbol{X}_i^\top \boldsymbol{\beta}}}{1+e^{\boldsymbol{X}_i^\top \boldsymbol{\beta}}} \boldsymbol{X}_i \right) = \sum_{i=1}^{n} (Y_i - 1)\boldsymbol{X}_i + \frac{1}{1+e^{\boldsymbol{X}_i^\top \boldsymbol{\beta}}} \boldsymbol{X}_i$$

and

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = -\sum_{i=1}^{n} \frac{e^{\boldsymbol{X}_i^\top \boldsymbol{\beta}}}{(1+e^{\boldsymbol{X}_i^\top \boldsymbol{\beta}})^2} \boldsymbol{X}_i \boldsymbol{X}_i^\top \preceq 0$$

This is actually a concave function.

Example: Continuation of Shuttle Challenger data. We want to see if adding more predictors helps fit the data better. Now we consider

$$Y_i \sim \text{Bernoulli}\left(\frac{e^{\beta_0 + \beta_1 X_i + \beta_2 X_i^2}}{1+e^{\beta_0 + \beta_1 X_i + \beta_2 X_i^2}} \right).$$

Consider the testing problem:

$$H_0 : \beta_2 = 0 \text{ v.s. } H_1 : \beta_2 \neq 0.$$

Under the full model (including three parameters) and the null hypothesis, the log-likelihood value is

$$\ell(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2) = -9.694, \quad \ell(\widehat{\beta}_0, \widehat{\beta}_1) = -10.158$$

$$\lambda(X) = 2(\ell(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2) - \ell(\widehat{\beta}_0, \widehat{\beta}_1)) = 2(-9.694 + 10.158) = 0.9266.$$

Compute the p-value,

$$\mathbb{P}(\chi_1^2 \geq 0.9266) = 0.336.$$

It is not significant enough to reject H_0. We retain the null hypothesis.

Bibliography

[1] E. L. Lehmann. *Elements of Large-Sample Theory*. Springer Science & Business Media, 2004.

[2] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer Science & Business Media, 2006.

[3] L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Science & Business Media, 2013.