SDS7102: Linear Models and Extensions

Point Estimation

Qiang Sun, Ph.D. <qiang.sun@mbzuai.ac.ae>

These slides are due to Eric Moulines.

August 19, 2025

MBZUAI

Statistical Model

- Let $X_1, \cdots, X_n$ be random variables (or random vectors) and suppose that we observe $x_1, \cdots, x_n$, which can be thought of as outcomes of the random variables $X_1, \cdots, X_n$.
- Suppose that the joint distribution of $\boldsymbol{X} = (X_1, \cdots, X_n)$ is unknown but belongs to some particular family of distributions. Such a family of distributions is called a *statistical model*.
- It is convenient to index the distributions belonging to a statistical model by a parameter $\theta; \theta$ typically represents the unknown or unspecified part of the model. We can then write

$$\boldsymbol{X} = (X_1, \cdots, X_n) \sim F_\theta \quad \text{for} \quad \theta \in \Theta,$$

  where F_θ is the joint distribution function of $\boldsymbol{X}$ and Θ is the set of possible values for the parameter θ; we will call the set Θ the parameter space.

Statistical Model

- In general, θ can be either a single real-valued parameter or a vector of parameters; in this latter case, we will often write $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)$ to emphasize that we have a vector-valued parameter.
- We write $P_\theta(A), E_\theta(X)$, and $\mathrm{Var}_\theta(X)$ to denote (respectively) probability, expected value, and variance with respect to a distribution with unknown parameter θ.
- We usually assume that Θ is a subset of some Euclidean space; such a model is often called a parametric model.
- Models whose distributions cannot be indexed by a finite dimensional parameter are often (somewhat misleadingly) called non-parametric models.

Identifiability

- For a given parameter θ corresponds to a single distribution F_θ. However, this does not rule out the possibility that there may exist distinct parameter values θ_1 and θ_2 such that $F_{\theta_1} = F_{\theta_2}$.
- We often require that a given model, or more precisely, its parametrization be identifiable; a model is said to have an identifiable parametrization (or to be an identifiable model) if $F_{\theta_1} = F_{\theta_2}$ implies that $\theta_1 = \theta_2$.
- A nonidentifiable parametrization can lead to problems in estimation of the parameters in the model.

Example: Poisson Model

- Suppose that $X_1, \cdots, X_n$ are i.i.d. Poisson random variables with mean λ.
- The joint frequency function of $\boldsymbol{X} = (X_1, \cdots, X_n)$ is

$$f(\boldsymbol{x}; \lambda) = \prod_{i=1}^{n} \frac{\exp(-\lambda)\lambda^{x_i}}{x_i!}$$

  for $x_1, \cdots, x_n = 0, 1, 2, \cdots$.
- The parameter space for this parametric model is $\{\lambda : \lambda > 0\}$.

Example: non-parametric and semi-parametric model

- Suppose that $X_1, \cdots, X_n$ are i.i.d. random variables with a continuous distribution function F that is unknown.
- The parameter space for this model consists of all possible continuous distributions. These distributions cannot be indexed by a finite dimensional parameter and so this model is non-parametric.
- We may also assume that $F(x)$ has a density $f(x - \theta)$ where θ is an unknown parameter and f is an unknown density function satisfying $f(x) = f(-x)$.
- This model is also non-parametric but depends on the real-valued parameter θ. (This might be considered a semiparametric model because of the presence of θ.)

Example: linear Gaussian regression

- Suppose that $X_1, \cdots, X_n$ are independent Normal random variables with $E_{\boldsymbol{\theta}}(X_i) = \beta_0 + \beta_1 t_i + \beta_2 s_i$ (where $t_1, \cdots, t_n$ and $s_1, \cdots, s_n$ are known constants) and $\mathrm{Var}_{\boldsymbol{\theta}}(X_i) = \sigma^2$; the parameter space is

$$\left\{\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \sigma) : -\infty < \beta_0, \beta_1, \beta_2 < \infty, \sigma > 0\right\}.$$

- The parametrization for this model is identifiable if, and only if, the vectors

$$\boldsymbol{z}_0 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \boldsymbol{z}_1 = \begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix}, \quad \text{and} \quad \boldsymbol{z}_2 = \begin{pmatrix} s_1 \\ \vdots \\ s_n \end{pmatrix}$$

are linearly independent.

Exponential families

- Suppose that $X_1, \cdots, X_n$ have a joint distribution F_θ where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)$ is an unknown parameter.
- We say that the family of distributions $\{F_\theta\}$ is a k-parameter exponential family if the joint density or joint frequency function of $(X_1, \cdots, X_n)$ is of the form

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \exp\left[\sum_{i=1}^{k} c_i(\boldsymbol{\theta}) T_i(\boldsymbol{x}) - d(\boldsymbol{\theta}) + S(\boldsymbol{x})\right]$$

  for $\boldsymbol{x} = (x_1, \cdots, x_n) \in A$ where A does not depend on the parameter $\boldsymbol{\theta}$.
- It is important to note that k need not equal p, the dimension of $\boldsymbol{\theta}$, although, in many cases, they are equal.

Binomial distribution

- Suppose that X has a Binomial distribution with parameters n and θ where θ is unknown.
- The frequency function of X is

$$\begin{aligned} f(x;\theta) &= \binom{n}{x}\theta^x(1-\theta)^{n-x} \\ &= \exp\left[\ln\left(\frac{\theta}{1-\theta}\right)x + n\ln(1-\theta) + \ln\binom{n}{x}\right] \end{aligned}$$

  for $x \in A = \{0, 1, \cdots, n\}$.
- The distribution of X has a one-parameter exponential family.

Gamma Distribution

- Suppose that $X_1, \cdots, X_n$ are i.i.d. Gamma random variables with unknown shape parameter α and unknown scale parameter λ.
- The joint density function of $\boldsymbol{X} = (X_1, \cdots, X_n)$ is

$$
\begin{aligned}
& f(\boldsymbol{x}; \alpha, \lambda) \\
& \quad = \prod_{i=1}^{n} \left[\frac{\lambda^{\alpha} x_i^{\alpha-1} \exp(-\lambda x_i)}{\Gamma(\alpha)} \right] \\
& \quad = \exp\left[(\alpha - 1) \sum_{i=1}^{n} \ln(x_i) - \lambda \sum_{i=1}^{n} x_i + n\alpha \ln(\lambda) - n \ln(\Gamma(\alpha)) \right]
\end{aligned}
$$

(for $x_1, \cdots, x_n > 0$) and so the distribution of $\boldsymbol{X}$ is a two-parameter exponential family.

Gaussian distribution

- Suppose that $X_1, \cdots, X_n$ are i.i.d. Normal random variables with mean θ and variance θ^2 where $\theta > 0$.
- The joint density function of $(X_1, \cdots, X_n)$ is

$$
\begin{aligned}
&f(\boldsymbol{x}; \theta) \\
&\quad = \prod_{i=1}^{n} \left[\frac{1}{\theta\sqrt{2\pi}} \exp\left(-\frac{1}{2\theta^2} (x_i - \theta)^2 \right) \right] \\
&\quad = \exp\left[-\frac{1}{2\theta^2} \sum_{i=1}^{n} x_i^2 + \frac{1}{\theta} \sum_{i=1}^{n} x_i - \frac{n}{2} \left(1 + \ln\left(\theta^2\right) + \ln(2\pi)\right) \right],
\end{aligned}
$$

and so $A = R^n$. Note that this is a two-parameter exponential family despite the fact that the parameter space is one-dimensional.

Poisson distribution

Suppose that $X_1, \cdots, X_n$ are independent Poisson random variables with $E(X_i) = \exp(\alpha + \beta t_i)$ where $t_1, \cdots, t_n$ are known constants. Setting $\boldsymbol{X} = (X_1, \cdots, X_n)$, the joint frequency function of $\boldsymbol{X}$ is

$$
\begin{aligned}
&f(\boldsymbol{x}; \alpha, \beta) \\
&\quad = \prod_{i=1}^{n} \left[\frac{\exp(-\exp(\alpha + \beta t_i)) \exp(\alpha x_i + \beta x_i t_i)}{x!} \right] \\
&\quad = \exp\left[\alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i t_i + \sum_{i=1}^{n} \exp(\alpha + \beta t_i) - \sum_{i=1}^{n} \ln(x_i!) \right].
\end{aligned}
$$

This is a two-parameter exponential family model; the set A is simply $\{0, 1, 2, 3, \cdots\}^n$.

Uniform distribution

Suppose that $X_1, \cdots, X_n$ are i.i.d. Uniform random variables on the interval $[0, \theta]$. The joint density function of $\boldsymbol{X} = (X_1, \cdots, X_n)$ is

$$f(\boldsymbol{x}; \theta) = \frac{1}{\theta^n} \quad \text{for } 0 \leq x_1, \cdots, x_n \leq \theta$$

The region on which $f(\boldsymbol{x}; \theta)$ is positive clearly depends on θ and so this model is not an exponential family model.

Mean and variance of exponential distribution

Proposition

Suppose that $\boldsymbol{X} = (X_1, \cdots, X_n)$ has a one-parameter exponential family distribution with density or frequency function

$$f(\boldsymbol{x};\theta) = \exp[c(\theta)T(\boldsymbol{x}) - d(\theta) + S(\boldsymbol{x})]$$

for $\boldsymbol{x} \in A$ where

(a) *the parameter space Θ is open,*

(b) *$c(\theta)$ is a one-to-one function on Θ,*

(c) *$c(\theta), d(\theta)$ are twice differentiable functions on Θ.*

Then

$$E_\theta[T(\boldsymbol{X})] = \frac{d'(\theta)}{c'(\theta)}$$

and $$\mathrm{Var}_\theta[T(\boldsymbol{X})] = \frac{d''(\theta)c'(\theta) - d'(\theta)c''(\theta)}{[c'(\theta)]^3}$$

Statistics

- Suppose that the model for $\boldsymbol{X} = (X_1, \cdots, X_n)$ has a parameter space Θ.
- Since the true value of the parameter θ (or, equivalently, the true distribution of $\boldsymbol{X}$) is unknown, we would like to summarize the available information in $\boldsymbol{X}$ without losing too much information about the unknown parameter θ.
- At this point, we are not interested in estimating θ per se but rather in determining how to best use the information in $\boldsymbol{X}$.

Statistics

- Define a statistic $T = T(\boldsymbol{X})$ to be a function of $\boldsymbol{X}$ that does not depend on any unknown parameter; that is, the statistic T depends only on observable random variables and known constants.
- A statistic can be real- or vector-valued.

Example

$T(\boldsymbol{X}) = \bar{X} = n^{-1}\sum_{i=1}^{n} X_i$. Since n (the sample size) is known, T is a statistic.

Example

$T(\boldsymbol{X}) = \left(X_{(1)}, \cdots, X_{(n)}\right)$ where $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ are the order statistics of $\boldsymbol{X}$. Since T depends only on the values of $\boldsymbol{X}, T$ is a statistic.

Statistics

- It is important to note that any statistic is itself a random variable and so has its own probability distribution; this distribution may or may not depend on the parameter θ.
- Ideally, a statistic $T = T(\boldsymbol{X})$ should contain as much information about θ as $\boldsymbol{X}$ does.
- However, this raises several questions.
  - For example, how does one determine if T and $\boldsymbol{X}$ contain the same information about θ ?

Ancillary Statistics

Definition (Ancillary statistics)

A statistic T is an ancillary statistic (for θ) if its distribution is independent of θ; that is, for all $\theta \in \Theta, T$ has the same distribution.

Example: ancillary statistics in normal sample

- Suppose that X_1 and X_2 are independent Normal random variables each with mean μ and variance σ^2 (where σ^2 is known).
- Let $T = X_1 - X_2$; then T has a Normal distribution with mean 0 and variance $2\sigma^2$. Thus T is ancillary for the unknown parameter μ.
- However, if both μ and σ^2 were unknown, T would not be ancillary for $\theta = (\mu, \sigma^2)$. (The distribution of T depends on σ^2 so T contains some information about σ^2.)

Example: ancilarity range w.r.t translation parameter

- Suppose that $X_1, \cdots, X_n$ are i.i.d. random variables with density function
$$f(x;\mu,\eta) = \frac{1}{2\eta} \quad \text{for } \mu - \eta \leq x \leq \mu + \eta.$$
- Define a statistic $R = X_{(n)} - X_{(1)}$, which is the sample range of $X_1, \cdots, X_n$.
- The density function of R is
$$f_R(r) = \frac{n(n-1)r^{n-2}}{(2\eta)^{n-1}} \left(1 - \frac{r}{2\eta}\right) \quad \text{for } 0 \leq r \leq 2\eta$$
which depends on η but not μ. Thus R is ancillary for μ.

Uniform distribution

Suppose that $X_1, \cdots, X_n$ are i.i.d. Uniform random variables on the interval $[0, \theta]$ where $\theta > 0$ is an unknown parameter. Define two statistics, $S = \min(X_1, \cdots, X_n)$ and $T = \max(X_1, \cdots, X_n)$. The density of S is

$$f_S(x;\theta) = \frac{n}{\theta}\left(1 - \frac{x}{\theta}\right)^{n-1} \quad \text{for} \quad 0 \leq x \leq \theta,$$

while the density of T is

$$f_T(x;\theta) = \frac{n}{\theta}\left(\frac{x}{\theta}\right)^{n-1} \quad \text{for} \quad 0 \leq x \leq \theta.$$

- Note that the densities of both S and T depend on θ and so neither is ancillary for θ. However, as n increases, it becomes clear that the density of S is concentrated around 0 for all possible values of θ while the density of T is concentrated around θ.

Sufficiency

- The first mention of sufficiency was made by Fisher (1920) in which he considered the estimation of the variance σ^2 of a Normal distribution based on i.i.d. observations $X_1, \cdots, X_n$.
- In particular, he considered estimating σ^2 based on the statistics

$$T_1 = \sum_{i=1}^{n} \left| X_i - \bar{X} \right| \quad \text{and} \quad T_2 = \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2$$

  where $\bar{X}$ is the average of $X_1, \cdots, X_n$.
- Fisher showed that the distribution of T_1 conditional on $T_2 = t$ does not depend on the parameter σ while the distribution of T_2 conditional on $T_1 = t$ does depend on σ.
- He concluded that all the information about σ^2 in the sample was contained in the statistic T_2 and that any estimate of σ^2 should be based on T_2;
- Any estimate of σ^2 based on T_1 could be improved by using the information in T_2 while T_2 could not be improved by using T_1.

Sufficient statistics

Definition (Sufficient statistics)

A statistic $T = T(\boldsymbol{X})$ is a sufficient statistic for a parameter θ if for all sets A, $P_\theta[\boldsymbol{X} \in A \mid T = t]$ is independent of θ for all t in the range of T.

- Sufficient statistics are not unique; from the definition of sufficiency, it follows that if g is a one-to-one function over the range of the statistic T then $g(T)$ is also sufficient.
- It also follows that if T is sufficient for θ then the distribution of any other statistic $S = S(\boldsymbol{X})$ conditional on T is independent of θ.

Sufficient statistics in binomial model

- Suppose that $X_1, \cdots, X_k$ are independent random variables where X_i has a Binomial distribution with parameters n_i (known) and θ (unknown).
- Let $T = X_1 + \cdots + X_k$; T will also have a Binomial distribution with parameters $m = n_1 + \cdots + n_k$ and θ.
- Show that T is sufficient.

Neyman factorization Lemma

Theorem (Neyman Factorization Criterion)

Suppose that $\boldsymbol{X} = (X_1, \cdots, X_n)$ *has a joint density or frequency function* $f(\boldsymbol{x}; \theta)(\theta \in \Theta)$. *Then* $T = T(\boldsymbol{X})$ *is sufficient for* θ *if, and only if,*

$$f(\boldsymbol{x}; \theta) = g(T(\boldsymbol{x}); \theta)h(\boldsymbol{x}).$$

(Both T *and* θ *can be vector-valued.)*

Sufficiency in uniform model

Suppose that $X_1, \cdots, X_n$ are i.i.d. random variables with density function

$$f(x;\theta) = \frac{1}{\theta} \quad \text{for } 0 \leq x \leq \theta$$

- Show that $X_{(n)}$ is sufficient.

Sufficient statistics in exponential model

Suppose that $\boldsymbol{X} = (X_1, \cdots, X_n)$ have a distribution belonging to a k-parameter exponential family with joint density or frequency function satisfying

$$f(\boldsymbol{x};\theta) = \exp\left[\sum_{i=1}^{k} c_i(\theta)T_i(\boldsymbol{x}) - d(\theta) + S(\boldsymbol{x})\right] I(\boldsymbol{x} \in A)$$

- Show that the statistic

$$T = (T_1(\boldsymbol{X}), \cdots, T_k(\boldsymbol{X}))$$

  is sufficient for θ.

Minimal sufficient statistics

- There are two notions of what is meant by the ”best possible” reduction of the data.
- The first of these is minimal sufficiency; a sufficient statistic T is minimal sufficient if for any other sufficient statistic S, there exists a function g such that $T = g(S)$.
- Thus a minimal sufficient statistic is the sufficient statistic that represents the maximal reduction of the data that contains as much information about the unknown parameter as the data itself.

Complete sufficient statistics

- A second (and stronger) notion is completeness. If $\boldsymbol{X} \sim F_\theta$ then a statistic $T = T(\boldsymbol{X})$ is complete if $E_\theta(g(T)) = 0$ for all $\theta \in \Theta$ implies that $P_\theta(g(T) = 0) = 1$ for all $\theta \in \Theta$.
- In particular, if T is complete then $g(T)$ is ancillary for θ only if $g(T)$ is constant; thus a complete statistic T contains no ancillary information.