STA 3000 Notes on models and likelihood

September 20, 2013

Statistical Models

"Y is a random variable with density function $f(y;\theta)$ or $f(y)$".

This is the starting point for most of the material to be covered. Typically Y will be a scalar random variable or a vector random variable of length n, and $f(y;\theta)$ will be a density function with respect to counting measure or Lebesgue measure. The problem is to reason from observed data y back to θ or $f(\cdot)$. In more complicated problems, such as observations taken in continuous time, the definitions of Y and its density may not be obvious.

In advanced texts this is often formulated from a measure-theoretic point of view. We have a so-called "probability triple" $(\mathcal{Y}, \mathcal{A}, \mathcal{P})$, where $\mathcal{Y}$ is the sample space, typically identified with $\mathbb{R}^n$ or $\mathbb{R}$, $\mathcal{A}$ is the Borel σ-field, and $\mathcal{P}$ is a probability measure which is absolutely continuous with respect to Lebesgue or counting measure. A very concise but helpful account is given in Chapter 1.2 and 1.3 of TPE. (I found two typos: on p.8, l.-2 "Example 2.1" should be "Example 2.2", p.16, 2nd display should be $E(T) = \int T(x)dP_Y(x)$.)

A simple set of examples of such models might be the following:

1. $Y_i = \alpha + \beta z_i + \epsilon_i, \quad i = 1,\dots n$, where z_i are known constants, and $\epsilon_1, \dots \epsilon_n$ are assumed to be independent, identically distributed, with each following a $N(0,1)$ distribution. Then

$$f(y;\theta) = \frac{1}{\sqrt{(2\pi)^n}} \exp -\frac{1}{2}\sum(y_i - \alpha - \beta z_i)^2$$

   and $\theta = (\alpha, \beta)$.

2. $Y_i = \alpha + \beta z_i + \epsilon_i, \quad i = 1,\dots n$, where z_i are known constants, and $\epsilon_1, \dots \epsilon_n$ are assumed to be independent, identically distributed, with each following the density $f_0(e)$ or $f_0(e;\nu)$ where the form of f_0 is known (possibly up to an unknown number of parameters).

3. $Y_i = \alpha + \beta z_i + \epsilon_i, \quad i = 1,\dots n$, where z_i are known constants, and $\epsilon_1, \dots \epsilon_n$ are assumed to be independent, identically distributed, with each following an unknown density f that satisfies some smoothness conditions

4. $Y_i = \alpha + \beta z_i + \epsilon_i, \quad i = 1,\dots n$, where z_i are known constants, and $\epsilon = (\epsilon_1, \dots \epsilon_n)$ follows a known (up to parameters) joint distribution $f_0(\epsilon;\nu)$

5. $Y_i = \mu(z_i) + \epsilon_i, \quad i = 1,\dots n$, where z_i are known constants, and $\epsilon_1, \dots \epsilon_n$ are assumed to be independent, identically distributed, with each following an unknown density f that satisfies some smoothness conditions

In general we would expect to require that the dimension of the data, n, be greater than the dimension of the parameter space, in order to construct inference about the parameter. This is not strictly true, as for example in cases 3 and 5 the dimension of the parameter space is a space of unknown functions, and hence effectively infinite, yet progress can be made via smoothing techniques. Model 3 is an example of a semiparametric model, and Model 5 a nonparametric model. Most models considered in this half of the course will be parametric models, with a parameter θ taking values in a p-dimensional space Θ, typically (a subset of) $\mathbb{R}^p$.

The likelihood function

Given a random variable Y and its density function $f(y;\theta)$, the *likelihood function* is defined to be

$$L(\theta; y) = c(y) f(y;\theta),$$

i.e. any function proportional to the density function evaluated at the observed value of the random variable. (There is really an equivalence class of likelihood functions, all differing by arbitrary multiplicative constants that may depend on y but may not depend on θ.)

The likelihood function is invariant under one to one transformations of Y not involving θ. Suppose $Z = g(Y)$, then

$$f_Z(z;\theta) = f_Y\{g^{-1}(z);\theta\}|dg^{-1}(z)/dz|$$

showing that the likelihood function based on Z is in the same equivalence class as the likelihood function based on Y.

Examples:

1. $Y_1, \ldots, Y_n$ i.i.d. Bernoulli (p):

$$L(p;y) = \prod p^{y_i}(1-p)^{1-y_i} = p^{\sum y_i}(1-p)^{n-\sum y_i}$$

2. R follows Binomial(n, p):

$$L(p;r) = \binom{n}{r} p^r (1-p)^{n-r}$$

3. Y follows a Negative Binomial(r, p):

$$L(p;n) = \binom{n-1}{r-1} p^r (1-p)^{n-r}$$

Note that in (a) $\sum y_i$ is a sufficient statistic (determines the likelihood function), and that the likelihood functions in (a), (b) and (c) are in the same equivalence class, showing in particular that the likelihood function is unaffected by the sampling rule. This is true in general, and has generated much discussion in the literature. However if we are interested in the distribution of the random function $L(\theta; Y)$ or quantities derived from it, this distribution will depend on the sampling rule.

4. $Y_1, \ldots Y_n$ i.i.d. $N(\mu, \sigma^2)$:

$$\begin{aligned} L(\mu, \sigma^2; y) &= \left(\frac{1}{\surd(2\pi)\sigma}\right)^n \exp\{-\frac{1}{2\sigma^2}\sum(y_i - \mu)/\sigma^2\} \\ &= \sigma^{-n}\exp-(\frac{\sum y_i^2}{2\sigma^2} + \frac{\mu\sum y_i}{\sigma^2} - \frac{n\mu^2}{2\sigma^2}) \\ &= \sigma^{-n}\exp\{-\frac{\sum(y_i - \bar{y})^2}{2\sigma^2} - \frac{n(\bar{y} - \mu)^2}{2\sigma^2}\} \end{aligned}$$

Note that the likelihood function depends only on $(\sum y_i^2, \sum y_i)$, or equivalently $(\bar{y}, \sum(y_i - \bar{y})^2)$, and in particular if σ^2 is assumed to be known, that the likelihood function for μ depends only on $\bar{y}$.

5. Nonhomogeneous Poisson process

Consider a Poisson process with rate function $\rho(t)$ observed over the interval $[0, t_0)$. Denoting the process by $\{N(t), 0 \leq t \leq t_0\}$, by definition

(a) $\text{pr}[N(t + h) = N(t) + 1|\{N(t), t \geq 0\}] = \rho(t)h + o(h)$

(b) $\text{pr}[N(t + h) = N(t)|\{N(t), t \geq 0\}] = 1 - \rho(t)h + o(h)$

(c) $\text{pr}[N(t + h) \geq N(t) + 2|\{N(t), t \geq 0\}] = o(h)$

Thus if we observe $\{N(t)\}$ to increase by 1 at points $y_1, \ldots y_n$, and to be constant at all other points in the interval $[0, t_0)$ the joint probability of this is

$$\begin{aligned} L\{\rho(\cdot)\} &= \prod_{i=1}^n \rho(y_i)h \prod^* \{1 - \rho(a_j)h\}\{1 + o(1)\} \\ &= h^n \prod_{i=1}^n \rho(y_i) \exp \sum^* \log\{1 - \rho(a_j)h\}\{1 + o(1)\} \\ &= h^n \prod_{i=1}^n \rho(y_i) \exp \sum^* \{-\rho(a_j)h\}\{1 + o(1)\} \\ &\approx \prod_{i=1}^n \rho(y_i) \exp\{-\int_0^{t_0} \rho(y)dy\} \end{aligned}$$

where the approximation in the last line comes from taking the limit as $h \to 0$, and we drop the factor h^n without changing the equivalence class.

The properties i.–iii. are the defining properties of the nonhomogeneous Poisson process, and also define the rate function $\rho(\cdot)$. This function would normally be parametrized by a relatively small number of parameters, such as $\rho(t) = \exp(\theta_0 + \theta_1 t)$ or $\rho(t) = \theta$. In the latter case we get

$$L(\theta) = \theta^n \exp(-\theta t_0)$$

which is the same likelihood function as that for a sample of n independent observations from the exponential density

$$f(y;\theta) = \theta \exp(-\theta y)$$

with observed total failure time $t_0 = \sum y_i$.

Exercise: Verify the last remark. Give the expression for the likelihood function when $\rho(t) = \exp(\theta_0 + \theta_1 t)$.

Exponential Families

The family $\{\mathcal{F}_\theta; \theta \in \Theta \subset R^m\}$ is called an exponential family if the distributions have densities of the form

$$\begin{aligned} f(y;\theta) &= \exp\{\phi(\theta)^T t(y) - c(\theta) - d(y)\} \\ &= \exp\{\sum_1^m \phi_j(\theta) t_j(y) - c(\theta) - d(y)\} \qquad (1) \end{aligned}$$

with respect to a common measure μ (usually Lebesgue measure). In canonical form we have the family $\{\mathcal{F}_\phi; \phi \in \Phi\}$ with densities

$$f(y;\phi) = \exp\{\phi^T t(y) - c(\phi) - d(y)\} \qquad (2)$$

where $c(\phi) = c\{\phi(\theta)\}$. If Θ is such that Φ is the set

$$\{\phi : \int \exp\{\phi^T t(y) - d(y)\} dy < \infty\},$$

then the exponential family has full rank. If Φ is also open then the (full) exponential family is called regular.

If the components of ϕ satisfy a linear constraint, then it will clearly be possible to re-write the family of densities in terms of a parameter of $m-1$ components and a set of $m-1$ sufficient statistics, so we usually assume this is not possible.

A *curved exponential family* is formed when the components of ϕ satisfy one or more nonlinear constraints. In that case the dimension of Φ will be smaller than m, and could most conveniently be specified by writing $\phi = \phi(\theta)$, $\theta \in \Theta \subset R^d$, $d < m$.

Sampling from an exponential family

Suppose $Y_1, \ldots, Y_n$ are independent, identically distributed, each with a distribution of the form (2). Then the joint density of $Y = (Y_1, \ldots, Y_n)$ at $y = (y_1, \ldots, y_n)$ is

$$f(y;\phi) = \exp\{\phi_1 \sum_i t_1(y_i) + \cdots + \phi_m \sum_i t_m(y_i) - nc(\phi) - \sum d(y_i)\}$$

showing that $T = t(Y) = \{\sum_i t_1(y_i), \ldots, \sum_i t_m(y_i)\}$ is minimal sufficient for ϕ in a full exponential family (assuming there is no linear constraint on the ϕ).

The marginal density of T is again of exponential family form:

$$f(t;\phi) = \exp\{\phi_1 t_1 + \cdots + \phi_m t_m - nc(\phi) - h(t)\} = \exp\{\phi^T t - nc(\phi) - h(t)\}$$

which can be verified by writing

$$\begin{aligned} f(t;\phi) &= \int_{\{x:t(y)=t\}} f(y;\phi)dy \\ &= \int_{\{x:t(y)=t\}} \exp\{\phi_1 \sum_i t_1(y_i) + \cdots + \phi_m \sum_i t_m(y_i) - c(\phi) - d(y)\}dy \\ &= \exp\{\phi^T t - nc(\phi)\} \int_{\{x:t(y)=t\}} \exp\{d(y)\}dy \\ &= \exp\{\phi^T t - nc(\phi) - h(t)\} \qquad (3) \end{aligned}$$

showing also that the cumulant generating function for T is $K_T(\alpha) = nK_Y(\alpha) = n\{c(\alpha+\phi) - c(\phi)\}$.

Example of a curved exponential family

Suppose Y follows a bivariate normal distribution with mean 0, and variance-covariance matrix

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

The density is thus

$$\begin{aligned} f(y;\rho) &= (2\pi)^{-1}(1-\rho^2)^{-1/2} \exp\{-\frac{1}{2(1-\rho^2)}(y_1^2 - 2\rho y_1 y_2 + y_2^2)\} \\ &= \exp\{-\frac{1}{2}(1-\rho^2)^{-1}(y_1^2 + y_2^2) + \rho(1-\rho^2)^{-1} y_1 y_2 - \frac{1}{2}\log(1-\rho^2) - \log(2\pi)\} \end{aligned}$$

which is a $(2,1)$ exponential family. In i.i.d. sampling the statistic $T = t(Y) = \{\sum(Y_{1i}^2 + Y_{2i}^2), \sum(Y_{1i}Y_{2i})\}$ is minimal sufficient for the parameter ρ.

References

TPE has a discussion of curved exponential families, but it is not very detailed, and in particular they allow 'curved' exponential families to be defined by linear relationships among the ϕ_i, as in Example 5.5, which would normally be considered a full exponential family with a two-dimensional sufficient statistic. BNC has an introduction in Chapter 1. SM §5.2 defines exponential families by 'exponential tilting' of a base density $f_0(\cdot)$, but then quickly moves to the general definition, given at equation (5.9).

A measure-theoretic treatment can be found in the book *Fundamentals of Statistical Exponential Families*, by L.D. Brown, IMS Lecture Notes-Monograph Series Volume 9 (1986), and Schervish, Ch. 2.2.

Transformation models

Location models A one parameter family of distributions on R is said to be a location family if the density function $f(y;\theta)$ takes the form $f_0(y-\theta)$, for $-\infty < \theta < \infty$. The density $f_0(y)$ is the standard form of the density and θ is the location parameter.
Examples The normal distribution with mean θ and known variance is a location family. The t_ν distribution with density function given by

$$f(y;\theta) = c\{1 + \frac{(y-\theta)^2}{\nu}\}^{-(\nu+1)/2}$$

is also a location family. The standard form of the density is just the usual t_ν density. The Cauchy distribution is a special case of this. The exponential location density is

$$f(y;\theta) = e^{-(y-\theta)}\,, \qquad y - \theta \geq 0;$$

note that the support of the density is the interval (θ, ∞), although θ can be any real value. Members of the same location family are simply shifted along the axis, relative to each other, and all have the same shape.

Scale models A one parameter family of distributions on R is said to be a scale family if the density function $f(y;\theta)$ takes the form $\theta^{-1}f_0(y/\theta)$, for $0 < \theta < \infty$. The density $f_0(y)$ is the standard form of the density and θ is the scale parameter.
Examples The normal distribution with known mean and unknown variance is a scale family. The gamma distribution with known shape parameter is a scale family. A special case of this is the simple exponential distribution:

$$f(y;\theta) = \theta^{-1}\exp(-y/\theta); \qquad y > 0.$$

Note that $Z = \log(Y)$ has the density function

$$g(z;\eta) = \exp\{z - \eta - e^{(z-\eta)}\}; \qquad -\infty < z < \infty$$

where $\eta = \log\theta$; this is a location family.

Location-scale models A one parameter family of distributions on R is said to be a location-scale family if the density function $f(y;\theta)$ takes the form $\theta_2^{-1}f_0((y-\theta_1)/\theta_2)$, for $-\infty < \theta_1 < \infty, 0 < \theta_2 < \infty$. The density $f_0(y)$ is the standard form of the density, θ_1 is the location parameter and θ_2 is the scale parameter.
Examples The normal distribution with mean θ_1 and variance θ_2^2 is a location scale family, and the standard normal is the standard form. The $t_\nu(\theta_1, \theta_2)$ is

$$f(t;\theta) = c\{1 + (y-\theta_1)^2/\nu\theta_2\}^{-(\nu+1)/2}$$

and the logistic(θ_1, θ_2) density is

$$f(y;\theta) = \frac{e^{-(y-\theta_1)/\theta_2}}{\{1 + e^{-(y-\theta_1)/\theta_2}\}^2}.$$

It is more conventional to use μ and σ for the location and scale parameter, although they do not always correspond to the mean and variance of the distribution.

Any continuous density on R can be embedded in a location-scale family, and in fact most location-scale families are constructed this way. It is easily proved that if the distribution of the random variable Y is a member of the location-scale family, then it can be expressed as

$$Y = \theta_2 Z + \theta_1$$

where Z has the standard distribution with density function $f_0(z)$.

Note that discrete distributions are not members of the location-scale family, essentially because the parameter space and the variable must take values on the same space.

Transformation families The location, scale and location-scale families are examples of transformation families. The basic idea is that a transformation on the sample space has a corresponding transformation on the parameter space that leaves the density function unchanged. For example, if Y has the density $f_0(y - \theta)$, then $Z = Y + a$ has the density $f_0(z - a - \theta) = f_0(z - \theta')$ so the family of densities for Y,$\{f_0(y - \theta); \theta \in R\}$, is is the same as that for $Z = Y + a$. Similarly, the family of location-scale densities is unchanged under location and scale transformations: if $f(y;\theta) = \theta_2^{-1} f_0((y-\theta_1)/\theta_2)$ then $Z = cY+a$ has density $f(z;\theta') = \theta_2'^{-1} f_0((z-\theta_1')/\theta_2')$ where $\theta_1' = c\theta_1 + a$ and $\theta_2' = c\theta_2$.

In general, we denote by g a transformation on the sample space $\mathcal{Y}$, and by g^* the induced transformation on the parameter space. Then the formalization of the above is the statement that

$$\text{pr}(gY \in A; \theta) = \text{pr}(Y \in A; g^*\theta).$$

If $g^*\Theta = \Theta$, the family of densities indexed by $\theta \in \Theta$ in invariant under the transformation g on $\mathcal{Y}$. Let $\mathcal{C}$ be a class of transformations on $\mathcal{Y}$ satisfying this condition, and $\mathcal{G}$ the smallest class containing $\mathcal{C}$ that is a group. Then $\mathcal{G}^* = \{g^* \text{ induced by } g \in \mathcal{G}\}$ is a group on Θ.

Example: linear regression A generalization of the location-scale model is the regression model

$$y = X\beta + \sigma\epsilon$$

where y is a vector of length n, X is a known $n \times p$ matrix, β is a vector of unknown parameters of length p, ϵ is a vector of length n that follows a known distribution $f_0(\cdot)$. If we let $y^* = Xb + cy$, where $b \in R^p$ and $c > 0$, then we can write

$$y^* = X(b + c\beta) + c\sigma\epsilon$$

which is a member of the same family, as long as the parameter space is $R^p \times R^+$.