STA3000: Note on transformation models

Introduction

Location models A one parameter family of distributions on R is said to be a location family if the density function $f(x;\theta)$ takes the form $f_0(x-\theta)$, for $-\infty < \theta < \infty$. The density $f_0(x)$ 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

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

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

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

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 $f(x;\theta)$ takes the form $\theta^{-1}f_0(x/\theta)$, for $0 < \theta < \infty$. The density $f_0(x)$ 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:

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

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

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

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 $f(x;\theta)$ takes the form $\theta_2^{-1}f_0((x-\theta_1)/\theta_2)$, for $-\infty < \theta_1 < \infty, 0 < \theta_2 < \infty$. The density $f_0(x)$ 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

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

and the logistic(θ_1, θ_2) density is

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

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 X is a member of the location-scale family, then it can be expressed as

$$X = \theta_2 Z + \theta_1$$

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 X has the density $f_0(x-\theta)$, then $Y = X + a$ has the density $f_0(y-a-\theta) = f_0(y-\theta')$ so the family of densities for X,$\{f_0(x-\theta); \theta \in R\}$, is is the same as that for $Y = X + a$. Similarly, the family of location-scale densities is unchanged under location and scale transformations: if $f(x;\theta) = \theta_2^{-1} f_0((x-\theta_1)/\theta_2)$ then $Y = cX+a$ has density $f(y;\theta') = \theta_2'^{-1} f_0((y-\theta_1')/\theta_2')$ 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 $\mathcal{X}$, and by g^* the induced transformation on the parameter space. Then the formalization of the above is the statement that

$$\mathrm{pr}(gX \in A; \theta) = \mathrm{pr}(X \in A; g^*\theta).$$

If $g^*\Theta = \Theta$, the family of densities indexed by $\theta \in \Theta$ in invariant under the transformation g on $\mathcal{X}$. Let $\mathcal{C}$ be a class of transformations on $\mathcal{X}$ 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 $R^p \times R^+$.

Dimension reduction in transformation models

A key feature of transformation models is that the model for a sample of size n permits a reduction in dimension of the sufficient statistic. This reduction is obtained by conditioning. Thus, there exist functions of the data, say $s(X)$ and $a(X)$, for which we can write

$$f(x;\theta) \propto f_1(s(x)|a(x);\theta)f_2(a(x))$$

where the marginal density of $a(X)$ does not depend on the parameter θ. More importantly, the conditional density f_1 is itself a transformation family density, with parameter θ and sample space variable $s(X)$.

Location family Suppose $X_1, \ldots, X_n$ are i.i.d. observations from the location family $f_0(x-\theta)$. Letting $s(x) = x_n$ and $a_i = x_i - x_n, i = 1, \ldots, n-1$, we can write

$$f(x;\theta) = \prod f_0(x_i - \theta) = f_1(x_n|a;\theta)f_2(a) \tag{1}$$

where

$$f_2(a) = \int f_0(a_1+t)\cdots f_0(a_{n-1}+t)f_0(t)\mathrm{d}t \tag{2}$$

and

$$f_1(x_n|a;\theta) = \frac{f_0(x_n+a_1-\theta)\cdots f_0(x_n+a_{n-1}-\theta)f_0(x_n-\theta)}{f_2(a)}. \tag{3}$$

The numerator is just a rewriting of $\prod f(x_i;\theta)$, and an expression equivalent to (3) is

$$f_1(x_n|a;\theta) = \frac{f_0(x_1-\theta)\cdots f_0(x_n-\theta)\mathrm{d}\theta}{\int f_0(x_1-\theta)\cdots f_0(x_n-\theta)\mathrm{d}\theta}. \tag{4}$$

The density of A does not depend on θ, and the conditional density of X_n given A is a location family density on R.

The functions $s(X)$ and $a(X)$ are not uniquely determined, but they are uniquely determined up to a location transformation. We could for example let $S = \bar{X}$ and $A_i = X_i - \bar{X}$. The vector A has n components but lies in R^{n-1} (as all its components must sum to 0, or in other words it is orthogonal to the 1-vector). The marginal density for a is again given by (2), and

$$f(x;\theta) \propto f(\bar{x}, a;\theta) = f_1(\bar{x}|a;\theta)f_2(a).$$

We might choose instead to let S be the maximum likelihood estimate of θ, and define $A_i = X_i - \hat{\theta}$.

Location-scale model A version of S and A that works for the location-scale model is $s(X) = (\bar{X}, s_X)$, where $s_X^2 = (n-1)^{-1}\sum(X_i - \bar{X})^2$, and $a_i(X) = (X_i - \bar{X})/s_X, \quad i = 1, \ldots, n$. The vector A has n components, but is restricted to lie in R^{n-2}. To prove that the distribution of A is indeed free of θ, we write

$$\begin{aligned} f(x;\theta)\mathrm{d}x &= \theta_2^{-n}\prod\{f_0\left(\frac{x_i-\theta_1}{\theta_2}\right)\}\mathrm{d}x_1\ldots\mathrm{d}x_n \\ &= \theta_2^{-n}\prod f_0\left(\frac{a_i s+\bar{x}-\theta_1}{\theta_2}\right)|J|\mathrm{d}a_1\ldots\mathrm{d}a_n\mathrm{d}\bar{x}\mathrm{d}s \end{aligned} \tag{5}$$

where $|J|$ is the Jacobian of the transformation from x to $(a, \bar{x}, s)$. To compute $f(a)$ we need to integrate out $\bar{x}$ and s from this expression, so we need to figure out the dependence of $|J|$ on $\bar{x}$ and s. The computation is a little bit tricky, but by writing $x_i = a_i s + \bar{x}$ we can see that $\mathrm{d}x_i = s\mathrm{d}a_i$, and since a has $n-2$ free dimensions, the factor s^{n-2} will be part of the Jacobian. It turns out that this is the only part that depends on $\bar{x}$ and s. The details are presented in the next section. The result is

$$\begin{aligned}
f(a)\mathrm{d}a &= \int\int \frac{s^{n-2}}{\theta_2^n} f_0\left(\frac{a_1 s + \bar{x} - \theta_1}{\theta_2}\right) \ldots f_0\left(\frac{a_n s + \bar{x} - \theta_1}{\theta_2}\right) \mathrm{d}\bar{x}\mathrm{d}s \\
&= \int\int \frac{(\theta_2 v)^{n-2}}{\theta_2^n} f_0\left(\frac{a_1 \theta_2 v + \bar{x} - \theta_1}{\theta_2}\right) \ldots f_0\left(\frac{a_n \theta_2 v + \bar{x} - \theta_1}{\theta_2}\right) \mathrm{d}\bar{x}\theta_2\mathrm{d}v \\
&= \int\int \frac{v^{n-2}}{\theta_2} f_0\left(a_1 v + \frac{\bar{x} - \theta_1}{\theta_2}\right) \ldots f_0\left(a_n v + \frac{\bar{x} - \theta_1}{\theta_2}\right) \mathrm{d}\bar{x}\mathrm{d}v \\
&= \int\int v^{n-2} f_0(a_1 v + t) \ldots f_0(a_n v + t)\mathrm{d}t\mathrm{d}v \qquad (6)
\end{aligned}$$

which shows that the marginal distribution of A does not depend on θ. The conditional distribution of $s(X)$, given A, is simply the ratio of the joint density to this marginal density. Again, the ancillary is not uniquely determined, but it is unique up to choice of location and scale variable. We could use $(a', x_{(1)}, x_{(n)} - x_{(1)})$, where $a'_i = (x_{(i)} - x_{(1)}/(x_{(n)} - x_{(1)})$, instead of $(a, \bar{x}, s)$, or many other equivalent formulations.

Details on the Jacobian:

As mentioned above, it is necessary to compute the Jacobian in the transformation from x to $(\bar{x}, s, a)$, where $a = (a_1, \ldots a_n)$ and $a_i = (x_i - \bar{x})/s$. This will be done below both algebraically and geometrically. The algebraic derivation is due to Angelo Canty, and the geometric one is taken from Chapter 2 of *Inference and Linear Models*.

Since $s^2 = \sum(x_i - \bar{x})^2$, we can see that the vector a, although it has n components, in fact lies in R^{n-2}, because $\sum a_i = a \cdot 1 = 0$ and $\sum a_i^2 = ||a||^2 = 1$. To compute the Jacobian we make the transformation one-to-one by letting

$$t_1 = \bar{x}, \quad t_2 = s, \quad t_i = a_i \qquad i = 3, \ldots, n;$$

note that we are explicitly using only $n-2$ components of a. To find the inverse transformation we have

$$x_i = t_1 + t_2 t_i \qquad i = 3, \ldots, n$$

and using the restrictions on a we have

$$\begin{aligned}
a_1 + a_2 &= -\sum_3^n t_i \\
1 - a_1 - a_2 &= \textstyle\sum_3^n t_i^2
\end{aligned}$$

from which we can write

$$\begin{aligned}
a_1 &= f_1(t_3, \ldots, t_n) = f_1(t_{(2)}), \text{say} \\
a_2 &= f_2(t_3, \ldots, t_n) = f_2(t_{(2)}), \text{say} \\
x_1 &= \bar{x} + s \cdot f_1(t_3, \ldots, t_n) \\
x_2 &= \bar{x} + s \cdot f_2(t_3, \ldots, t_n)
\end{aligned}$$

thus completing the transformation. We now have

$$\left|\frac{\partial x}{\partial t}\right| = \begin{vmatrix}
1 & f_1(t_{(2)}) & sg_{13}(t_{(2)}) & \ldots & sg_{1n}(t_{(2)}) \\
1 & f_2(t_{(2)}) & sg_{13}(t_{(2)}) & \ldots & sg_{1n}(t_{(2)}) \\
1 & t_3 & s & \ldots & 0 \\
1 & t_4 & 0 & s & \ldots \\
\vdots & \vdots & \vdots & & \vdots \\
1 & t_n & 0 & \ldots & s
\end{vmatrix}$$

where $g_{ij}(t_{(2)}) = \partial f_i(t_{(2)})/\partial t_j, i = 1, 2; j = 3, \ldots, n$. A well-known formula for the determinant of a partitioned matrix shows that the Jacobian is of the form

$$s^{n-2}h(t_{(2)}) = s^{n-2}h(a)$$

which is the result we were looking for.

In the above derivation the location and scale estimates were $\bar{x}$ and s, with the result that a is orthogonal to the 1-vector, and has length 1. However, the same derivation applies for a variety of other location and scale estimates. For example, suppose we wanted to use the maximum likelihood estimates of μ and σ, which are defined as the solutions to the equations $\partial \log f(x; \hat{\mu}, \hat{\sigma})/\partial\mu = 0, \quad \partial \log f(x; \hat{\mu}, \hat{\sigma})/\partial\sigma = 0$ i.e.

$$\begin{aligned}
\frac{-1}{\hat{\sigma}} \sum g'(\frac{x_i - \hat{\mu}}{\hat{\sigma}}) &= 0 \\
\frac{-n}{\hat{\sigma}} + \frac{x_i - \hat{\mu}}{\hat{\sigma}^2} \sum g'(\frac{x_i - \hat{\mu}}{\hat{\sigma}}) &= 0
\end{aligned}$$

where $g(x_i) = \log f(x_i)$. The ancillary statistic a is defined by $a_i = (x_i - \hat{\mu})/\hat{\sigma}$, so that $x_i = \hat{\sigma}a_i + \hat{\mu}$, and these two equations can be reexpressed as

$$\begin{aligned}
\sum g'(a_i) &= 0 \\
\sum a_i g'(a_i) &= n
\end{aligned}$$

which gives two restrictions on the a_i. Thus we can proceed as above and express $(\hat{\mu}, \hat{\sigma}, a_3, \ldots a_n)$ as a one-to-one function of x, and find the inverse transformation. It is of exactly the same form (with s replaced by $\hat{\sigma}$), but the functions called f_1 and f_2 in the above derivation are different.

The geometric derivation of the result is actually very similar, but a little more elegant. It again uses $\bar{x}$ and s as coordinates to get the result, and then argues that this choice of coordinates is arbitrary. Although it's not necessary, it's a little bit easier to first define $z_i = (x_i - \mu)/\sigma$; we want to construct the conditional distribution of $\bar{z}, s(z)$, where $\bar{z} = \sum z_i$, and $s^2(\mathbf{z}) = \sum(z_i - \bar{z})^2$, given $\mathbf{d}(\mathbf{z}) = (\mathbf{z} - \bar{z}\,\mathbf{1})/||\mathbf{z} - z\,\mathbf{1}||$. Note that in terms of the original variables $\bar{z} = (\bar{x} - \mu)/\sigma$, $s(\mathbf{z}) = \sigma s$, and $\mathbf{d}(\mathbf{z}) = \mathbf{a}$. (Bold font is used for vectors here to try to clarify the geometric argument.)

Now we compute the Jacobian of the transformation from $\mathbf{z}$ to $(\bar{z}, s(\mathbf{z}), \mathbf{d}(\mathbf{z}))$ by figuring out what the differential element $d\mathbf{z}$ is in the new coordinates. That is, in the joint density of $\mathbf{z}$,

$$f(\mathbf{z})d\mathbf{z} = \prod f(z_i)dz_i$$

we consider the differential element $\prod dz_i$ as giving the volume of a small box at the point $\mathbf{z}$. We want to express this volume in the new coordinates. The coordinates $(\bar{z}, s(\mathbf{z}), \mathbf{d}(\mathbf{z}))$ provide locally orthogonal coordinates at the point $\mathbf{z} = \bar{z}\,1 + s(\mathbf{z})\mathbf{d}(\mathbf{z})$, and we want to know how they change as we change to point $\mathbf{z}$ to $\mathbf{z} + d\mathbf{z}$. The coordinate specified by $\bar{z}$ lies on the $\mathbf{1}$-vector, so a small change in the coordinates $\mathbf{z}$ cause a change in $\bar{z}$ of $\sqrt{n}d\bar{z}$. Since $s(\mathbf{z})$ measures the length of $\mathbf{z} - \bar{z}\,\mathbf{1}$, its rate of change is simply ds. (Think of the picture in R^2.) Now $\mathbf{d}(\mathbf{z})$ is orthogonal to the $\mathbf{1}$-vector, and lies on a unit sphere in the $n-1$-dimensional subspace of R^n that is orthogonal to the $\mathbf{1}$-vector. Thus the volume element is the surface volume on the sphere defined by $s(\mathbf{z})\mathbf{d}(\mathbf{z})$, i.e. the sphere of radius $s(\mathbf{z})$. This volume is $s(\mathbf{z})^{n-1}du$, where du is surface volume on the unit sphere in R^{n-1} (which is $n-2$-dimensional). (In fact an explicit expression for the surface area of the unit sphere in R^d is $(2\pi)^{d/2}/\Gamma(d/2)$.)

The coordinates in this development are orthogonal, so the volume element is the product of the three pieces. For this reason these coordinates are a convenient choice for computing the differential. However, if we choose to coordinatize the point using other location and scale functions, the result is unchanged. The only thing we need to make sure of is that the location coordinate, say $\tilde{\mu}(\mathbf{z})$, satisfies the property $\tilde{\mu}(a\mathbf{z} + b\mathbf{1}) = a\tilde{\mu}(\mathbf{z}) + b$, the scale coordinate, say $\tilde{\sigma}(\mathbf{z})$, satisfies $\tilde{\sigma}(a\mathbf{z} + b\mathbf{1}) = a\tilde{\sigma}(\mathbf{z})$, and $\mathbf{d}$ is appropriately defined in terms of these two coordinates. We can show that such location and scale coordinates must themselves be location scale transformations of $\bar{z}$ and $s(\mathbf{z})$, so that we can convert the above result to a more general one. (Although $\hat{\mu}\mathbf{1}$ and $\hat{\sigma}$ will not give orthogonal coordinates, so that in these two dimensions the 'box' is a parallelogram, we can still figure out the volume by multiplying the base by the height!) Using the more general coordinates will not provide an explicit expression for the normalizing constant in terms of the surface area on the sphere, because the new vector $\mathbf{d}$ isn't forced to lie in the orthogonal complement of the $\mathbf{1}$-vector.

This latter part of the argument is formalized in Chapter 2 of Fraser's *The Structure of Inference.*