Exponential Families in Theory and Practice

Bradley Efron
Stanford University

Introduction

The inner circle in Figure 1 represents normal theory, the preferred venue of classical applied statistics. Exact inference — t tests, F tests, chi-squared statistics, ANOVA, multivariate analysis — were feasible within the circle. Outside the circle was a general theory based on large-sample asymptotic approximation involving Taylor series and the central limit theorem.

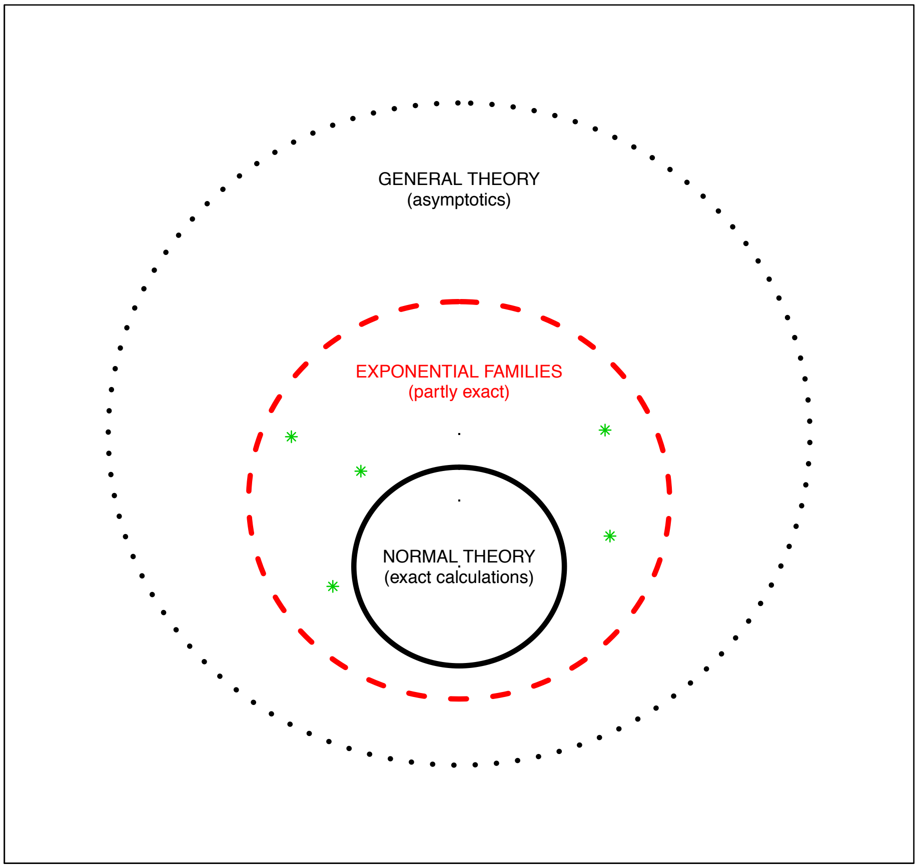


Figure 1: Three levels of statistical modeling

A few special exact results lay outside the normal circle, relating to specially tractable distributions such as the binomial, Poisson, gamma and beta families. These are the figure's green stars.

A happy surprise, though a slowly emerging one beginning in the 1930s, was that all the special cases were examples of a powerful general construction: *exponential families.* Within this superfamily, the intermediate circle in Figure 1, “almost exact” inferential theories such as generalized linear models (GLMs) are possible. This course will examine the theory of exponential families in a relaxed manner with an eye toward applications. A much more rigorous approach to the theory is found in Larry Brown’s 1986 monograph, *Fundamentals of Statistical Exponential Families*, IMS series volume 9.

Our title, “Exponential families in theory and practice,” might well be amended to “… *between* theory and practice.” These notes collect a large amount of material useful in statistical applications, but also of value to the theoretician trying to frame a new situation without immediate recourse to asymptotics. My own experience has been that when I can put a problem, applied or theoretical, into an exponential family framework, a solution is imminent. There are almost no proofs in what follows, but hopefully enough motivation and heuristics to make the results believable if not obvious. References are given when this doesn’t seem to be the case.

Part 1

One-parameter Exponential Families

One-parameter exponential families are the building blocks for the multiparameter theory developed in succeeding parts of this course. Useful and interesting in their own right, they unify a vast collection of special results from classical methodology. Part I develops their basic properties and relationships, with an eye toward their role in the general data-analytic methodology to follow.

1.1 Definitions and notation

This section reviews the basic definitions and properties of one-parameter exponential families. It also describes the most familiar examples — normal, Poisson, binomial, gamma — as well as some less familiar ones.

Basic definitions and notation

The fundamental unit of statistical inference is a family of probability densities $\mathcal{G}$, "density" here including the possibility of discrete atoms. A one-parameter exponential family has densities $g_\eta(y)$ of the form

$$\mathcal{G} = \{g_\eta(y) = e^{\eta y - \psi(\eta)} g_0(y),\ \eta \in A,\ y \in \mathcal{Y}\}, \tag{1.1}$$

A and $\mathcal{Y}$ subsets of the real line $\mathcal{R}^1$.

Terminology

- η is the *natural* or *canonical* parameter; in familiar families like the Poisson and binomial, it often isn't the parameter we are used to working with.

- y is the *sufficient statistic* or *natural statistic*, a name that will be more meaningful when we discuss repeated sampling situations; in many cases (the more interesting ones) $y = y(x)$ is a function of an observed data set x (as in the binomial example below).

- The densities in $\mathcal{G}$ are defined with respect to some *carrying measure* $m(dy)$, such as the uniform measure on $[-\infty, \infty]$ for the normal family, or the discrete measure putting weight 1 on the non-negative integers ("counting measure") for the Poisson family. Usually $m(dy)$ won't be indicated in our notation. We will call $g_0(y)$ the *carrying density.*

- $\psi(\eta)$ in (1.1) is the *normalizing function* or *cumulant generating function*; it scales the densities $g_\eta(y)$ to integrate to 1 over the sample space $\mathcal{Y}$,

$$\int_{\mathcal{Y}} g_\eta(y) m(dy) = \int_{\mathcal{Y}} e^{\eta y} g_0(y) m(dy) / e^{\psi(\eta)} = 1.$$

- The *natural parameter space* A consists of all η for which the integral on the left is finite,

$$A = \left\{ \eta : \int_{\mathcal{Y}} e^{\eta y} g_0(y) m(dy) < \infty \right\}.$$

Homework 1.1. Use convexity to prove that if η_1 and $\eta_2 \in A$ then so does any point in the interval $[\eta_1, \eta_2]$ (implying that A is a possibly infinite interval in $\mathcal{R}^1$).

Homework 1.2. We can reparameterize $\mathcal{G}$ in terms of $\tilde{\eta} = c\eta$ and $\tilde{y} = y/c$. Explicitly describe the reparameterized densities $\tilde{g}_{\tilde{\eta}}(\tilde{y})$.

We can construct an exponential family $\mathcal{G}$ through any given density $g_0(y)$ by "tilting" it exponentially,

$$g_\eta(y) \propto e^{\eta y} g_0(y)$$

and then renormalizing $g_\eta(y)$ to integrate to 1,

$$g_\eta(y) = e^{\eta y - \psi(\eta)} g_0(y) \qquad \left(e^{\psi(\eta)} = \int_{\mathcal{Y}} e^{\eta y} g_0(y) m(dy) \right).$$

It seems like we might employ other tilting functions, say

$$g_\eta(y) \propto \frac{1}{1+\eta|y|} g_0(y),$$

but only exponential tilting gives convenient properties under independent sampling.

If η_0 is any point on A we can write

$$g_\eta(y) = \frac{g_\eta(y)}{g_{\eta_0}(y)} g_{\eta_0}(y) = e^{(\eta-\eta_0)y - [\psi(\eta)-\psi(\eta_0)]} g_{\eta_0}(y).$$

This is the same exponential family, now represented with

$$\eta \longrightarrow \eta - \eta_0, \quad \psi \longrightarrow \psi(\eta) - \psi(\eta_0), \quad \text{and} \quad g_0 \longrightarrow g_{\eta_0}.$$

Any member $g_{\eta_0}(y)$ of $\mathcal{G}$ can be chosen as the carrier density, with all the other members as exponential tilts of g_{η_0}. Notice that the sample space $\mathcal{Y}$ is the *same* for all members of $\mathcal{G}$, and that all put positive probability on every point in $\mathcal{Y}$.

The Poisson family

As an important first example we consider the Poisson family. A Poisson random variable Y having mean μ takes values on the non-negative integers $\mathcal{Z}_+ = \{0, 1, 2, \dots\}$,

$$\Pr_\mu\{Y = y\} = e^{-\mu}\mu^y/y! \qquad (y \in \mathcal{Z}_+).$$

The densities $e^{-\mu}\mu^y/y!$, taken with respect to counting measure on $\mathcal{Y} = \mathcal{Z}_+$, can be written in exponential family form as

$$g_\eta(y) = e^{\eta y - \psi(\eta)} g_0(y) \begin{cases} \eta = \log(\mu) & (\mu = e^\eta) \\ \psi(\eta) = e^\eta & (= \mu) \\ g_0(y) = 1/y! & (\text{not a member of } \mathcal{G}). \end{cases}$$

Homework 1.3. (a) Rewrite $\mathcal{G}$ so that $g_0(y)$ corresponds to the Poisson distribution with $\mu = 1$. (b) Carry out the numerical calculations that tilt Poi(12), seen in Figure 1.1, into Poi(6).

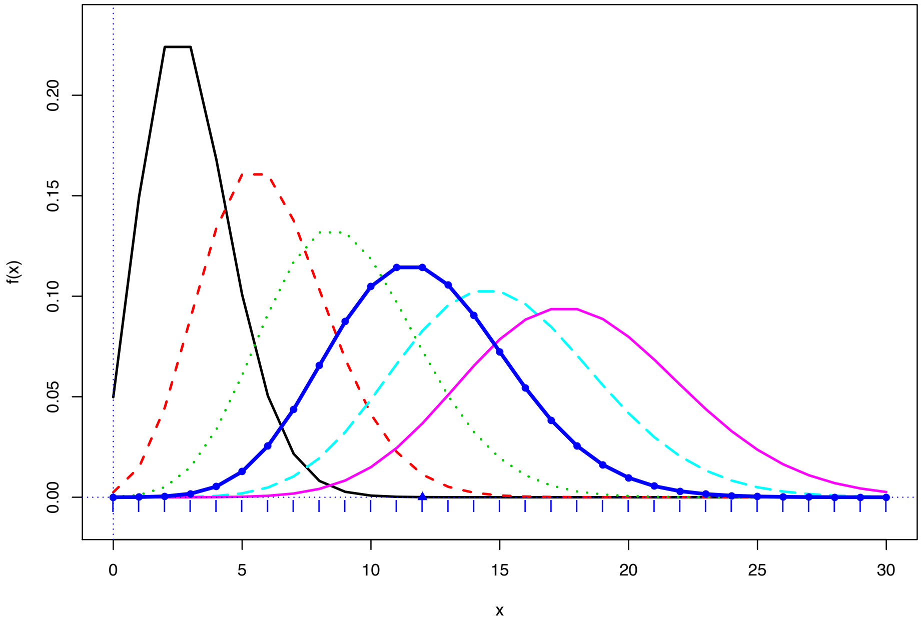


Figure 1.1: Poisson densities for $\mu = 3, 6, 9, 12, 15, 18$; heavy curve with dots for $\mu = 12$.

1.2 Moment relationships

Expectation and variance

Differentiating $\exp\{\psi(\eta)\} = \int_{\mathcal{Y}} e^{\eta y} g_0(y) m(dy)$ with respect to η, indicating differentiation by dots, gives

$$\dot{\psi}(\eta) e^{\psi(\eta)} = \int_{\mathcal{Y}} y e^{\eta y} g_0(y) m(dy)$$

and

$$\left[\ddot{\psi}(\eta) + \dot{\psi}(\eta)^2\right] e^{\psi(\eta)} = \int_{\mathcal{Y}} y^2 e^{\eta y} g_0(y) m(dy).$$

(The dominated convergence conditions for differentiating inside the integral are always satisfied in exponential families; see Theorem 2.2 of Brown, 1986.) Multiplying by $\exp\{-\psi(\eta)\}$ gives expressions for the mean and variance of Y,

$$\dot{\psi}(\eta) = E_\eta(y) \equiv \text{“}\mu_\eta\text{”}$$

and

$$\ddot{\psi}(\eta) = \text{Var}_\eta\{Y\} \equiv \text{“}V_\eta\text{”};$$

V_η is greater than 0, implying that $\psi(\eta)$ is a convex function of η.

Notice that

$$\dot{\mu} = \frac{d\mu}{d\eta} = V_\eta > 0.$$

The mapping from η to μ is $1:1$ increasing and infinitely differentiable. We can index the family $\mathcal{G}$ just as well with μ, the *expectation parameter*, as with η. Functions like $\psi(\eta)$, E_η, and V_η can just as well be thought of as functions of μ. We will sometimes write ψ, V, etc. when it's not necessary to specify the argument. Notations such as V_μ formally mean $V_{\eta(\mu)}$.

Note. Suppose

$$\zeta = h(\eta) = h\left(\eta(\mu)\right) = \text{“}H(\mu)\text{”}.$$

Let $\dot{h} = \partial h/\partial\eta$ and $H' = \partial H/\partial\mu$. Then

$$H' = \dot{h}\frac{d\eta}{d\mu} = \dot{h}/V.$$

Skewness and kurtosis

$\psi(\eta)$ is the *cumulant generating function* for g_0 and $\psi(\eta) - \psi(\eta_0)$ is the CGF for $g_{\eta_0}(y)$, i.e.,

$$e^{\psi(\eta)-\psi(\eta_0)} = \int_{\mathcal{Y}} e^{(\eta-\eta_0)y} g_{\eta_0}(y) m(dy).$$

By definition, the Taylor series for $\psi(\eta) - \psi(\eta_0)$ has the cumulants of $g_{\eta_0}(y)$ as its coefficients,

$$\psi(\eta) - \psi(\eta_0) = k_1(\eta - \eta_0) + \frac{k_2}{2}(\eta - \eta_0)^2 + \frac{k_3}{6}(\eta - \eta_0)^3 + \dots.$$

Equivalently,

$$\begin{array}{cccc} \dot{\psi}(\eta_0) = k_1, & \ddot{\psi}(\eta_0) = k_2, & \dddot{\psi}(\eta_0) = k_3, & \ddddot{\psi}(\eta_0) = k_4 \\ \Big[\quad = \mu_0 & = V_0 & = E_0\{y_0 - \mu_0\}^3 & = E_0\{y_0 - \mu_0\}^4 - 3V_0^2 \Big] \end{array}$$

etc., where $k_1, k_2, k_3, k_4, \dots$ are the *cumulants* of g_{η_0}.

By definition, for a real-valued random variable Y,

$$\text{SKEW}(Y) = \frac{E(Y - EY)^3}{[\text{Var}(Y)]^{3/2}} \equiv \text{“}\gamma\text{”} = \frac{k_3}{k_2^{3/2}}$$

and

$$\text{KURTOSIS}(Y) = \frac{E(Y - EY)^4}{[\text{Var}(Y)]^2} - 3 \equiv \text{“}\delta\text{”} = \frac{k_4}{k_2^2}.$$

Putting this all together, if $Y \sim g_\eta(\cdot)$ in an exponential family,

$$Y \sim \left[\begin{array}{cccc} \dot{\psi}, & \ddot{\psi}^{1/2}, & \dddot{\psi}/\ddot{\psi}^{3/2}, & \ddddot{\psi}/\ddot{\psi}^{2} \\ \uparrow & \uparrow & \uparrow & \uparrow \\ \text{expectation} & \text{standard} & \text{skewness} & \text{kurtosis} \\ & \text{deviation} & & \end{array} \right]$$

where the derivatives are taken at η.

For the Poisson family

$$\psi = e^\eta = \mu$$

so all the cumulants equal μ

$$\dot{\psi} = \ddot{\psi} = \dddot{\psi} = \ddddot{\psi} = \mu,$$

giving

$$Y \sim \left[\begin{array}{cccc} \mu, & \sqrt{\mu}, & 1/\sqrt{\mu}, & 1/\mu \\ \uparrow & \uparrow & \uparrow & \uparrow \\ \text{exp} & \text{st dev} & \text{skew} & \text{kurt} \end{array} \right]$$

A useful result

Continuing to use dots for derivatives with respect to η and primes for derivatives with μ, notice that

$$\gamma = \frac{\dddot{\psi}}{\ddot{\psi}^{3/2}} = \frac{\dot{V}}{V^{3/2}} = \frac{V'}{V^{1/2}}$$

(using $H' = \dot{h}/V$). Therefore

$$\gamma = 2(V^{1/2})' = 2\frac{d}{d\mu}\,\mathrm{sd}_\mu$$

where $\mathrm{sd}_\mu = V_\mu^{1/2}$ is the standard deviation of y. In other words, $\gamma/2$ is the rate of change of sd_μ with respect to μ.

Homework 1.4. Show that

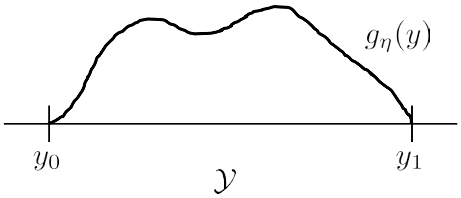
$$\text{(a)}\quad \delta = V'' + \gamma^2 \quad \text{and} \quad \text{(b)}\quad \gamma' = \left(\delta - \frac{3}{2}\gamma^2\right) \Big/ \mathrm{sd}\,.$$

Note. All of the classical exponential families — binomial, Poisson, normal, etc. — are those with closed form CGFs ψ. This yields neat expressions for means, variances, skewnesses, and kurtoses.

Unbiased estimate of η

By definition y is an unbiased estimate of μ (and in fact by completeness the only unbiased estimate of form $t(y)$). What about η?

- Let $l_0(y) = \log\{g_0(y)\}$ and $l_0'(y) = \frac{dl_0(y)}{dy}$.
- Suppose $\mathcal{Y} = [y_0, y_1]$ (both possibly infinite) and that $m(y) = 1$.



Lemma 1.

$$E_\eta\left\{-l_0'(y)\right\} = \eta - \left[g_\eta(y_1) - g_\eta(y_0)\right].$$

Homework 1.5. Prove the lemma. (*Hint*: integration by parts.)

So, if $g_\eta(y) = 0$ (or $\to 0$) at the extremes of $\mathcal{Y}$, then $-l_0'(y)$ is a unbiased estimate of η.

Homework 1.6. Numerically investigate how well $E_\eta\{-l_0'(y)\}$ approximates η in the Poisson family.

1.3 Repeated sampling

Suppose that $y_1, y_2, \ldots, y_n$ is an independent and identically distributed (i.i.d.) sample from an exponential family $\mathcal{G}$:

$$y_1, y_2, \ldots, y_n \overset{\text{iid}}{\sim} g_\eta(\cdot),$$

for an unknown value of the parameter $\eta \in A$. The density of $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ is

$$\begin{aligned}\prod_{i=1}^n g_\eta(y_i) &= e^{\sum_1^n(\eta y_i - \psi)}\\ &= e^{n(\eta\bar{y}-\psi)},\end{aligned}$$

where $\bar{y} = \sum_{i=1}^n y_i/n$. Letting $g_\eta^{\boldsymbol{Y}}(\boldsymbol{y})$ indicate the density of $\boldsymbol{y}$ with respect to $\prod_{i=1}^n m(dy_i)$,

$$g_\eta^{\boldsymbol{Y}}(\boldsymbol{y}) = e^{n[\eta\bar{y}-\psi(\eta)]}\prod_{i=1}^n g_0(y_i). \tag{1.2}$$

This is one-parameter exponential family, with:

- natural parameter $\eta^{(n)} = n\eta \quad (\text{so } \eta = \eta^{(n)}/n)$
- sufficient statistic $\bar{y} = \sum_1^n y_i/n \quad (\bar{\mu} = E_{\eta(n)}\{\bar{y}\} = \mu)$
- normalizing function $\psi^{(n)}(\eta^{(n)}) = n\psi(\eta^{(n)}/n)$
- carrier density $\prod_{i=1}^n g_0(y_i) \quad$ (with respect to $\prod m(dy_i)$)

Homework 1.7. Show that $\bar{y} \sim (\mu, \sqrt{V/n}, \gamma/\sqrt{n}, \delta/n)$.

Note. In what follows, we usually index the parameter space by η rather than $\eta^{(n)}$.

Notice that $\boldsymbol{y}$ is now a vector, and that the tilting factor $e^{\eta^{(n)}\bar{y}}$ is tilting the *multivariate* density $\prod_1^n g_0(y_i)$. This is still a one-parameter exponential family because the tilting is in a single direction, along $\mathbf{1} = (1, 1, \ldots, 1)$.

The sufficient statistic $\bar{y}$ also has a one-parameter exponential family of densities,

$$g_\eta^{\bar{Y}}(\bar{y}) = e^{n(\eta\bar{y}-\psi)} g_0^{\bar{Y}}(\bar{y}),$$

where $g_0^{\bar{Y}}(\bar{y})$ is the g_0 density with respect to $m^{\bar{Y}}(d\bar{y})$, the induced carrying measure.

The density (1.2) can also be written as

$$e^{\eta S - n\psi}, \qquad \text{where } S = \sum_{i=1}^n y_i.$$

This moves a factor of n from the definition of the natural parameter to the definition of the sufficient statistic. For any constant c we can re-express an exponential family $\{g_\eta(y) = \exp(\eta y - \psi)g_0(y)\}$ by mapping η to η/c and y to cy. This tactic will be useful when we consider multiparameter exponential families.

Homework 1.8. $y_1, y_2, \ldots, y_n \overset{\text{iid}}{\sim} \text{Poi}(\mu)$. Describe the distributions of $\bar{Y}$ and S, and say what are the exponential family quantities $(\eta, y, \psi, g_0, m, \mu, V)$ in both cases.

1.4 Some well-known one-parameter families

We've already examined the Poisson family. This section examines some other well-known (and not so well-known) examples.

Normal with variance 1

$\mathcal{G}$ is the normal family $Y \sim \mathcal{N}(\mu, 1)$, μ in $\mathcal{R}^1$. The densities, taken with respect to $m(dy) = dy$, Lebesque measure,

$$g_\mu(y) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}(y-\mu)^2}$$

can be written in exponential family form (1.1) with

$$\eta = \mu, \quad y = y, \quad \psi = \frac{1}{2}\mu^2 = \frac{1}{2}\eta^2, \quad g_0(y) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}y^2}.$$

Homework 1.9. Suppose $Y \sim \mathcal{N}(\mu, \sigma^2)$ with σ^2 known. Give η, y, ψ, and g_0.

Binomial

$Y \sim \text{Bi}(N, \pi)$, N known, so

$$g(y) = \binom{N}{y} \pi^y (1-\pi)^{n-y}$$

with respect to counting measure on $\{0, 1, \dots, n\}$. This can be written as

$$\binom{N}{y} e^{\left(\log \frac{\pi}{1-\pi}\right) y + N \log(1-\pi)},$$

a one-parameter exponential family, with:

- $\eta = \log[\pi/(1-\pi)]$ (so $\pi = 1/(1+e^{-\eta})$, $1-\pi = 1/(1+e^{\eta})$)
- $A = (-\infty, \infty)$
- $y = y$
- expectation parameter $\mu = N\pi = N/(1+e^{-\eta})$
- $\psi(\eta) = N \log(1+e^{\eta})$
- variance function $V = N\pi(1-\pi)$ $(= \mu(1-\mu)N)$
- $g_0(y) = \binom{N}{y}$

Homework 1.10. Show that for the binomial

$$\gamma = \frac{1-2\pi}{\sqrt{N\pi(1-\pi)}} \quad \text{and} \quad \delta = \frac{1-6\pi(1-\pi)}{N\pi(1-\pi)}.$$

Homework 1.11. Notice that $A = (-\infty, \infty)$ does *not* include the cases $\pi = 0$ or $\pi = 1$. Why not?

Gamma

$Y \sim \lambda G_N$ where G_N is a standard gamma variable, N known, and λ an unknown scale parameter,

$$g(y) = \frac{y^{N-1} e^{-y/\lambda}}{\lambda^N \Gamma(N)} \qquad [\mathcal{Y} = (0, \infty)].$$

This is a one-parameter exponential family with:

- $\eta = -1/\lambda$
- $\mu = N\lambda = -N/\eta$
- $V = N/\eta^2 = N\lambda^2 = \mu^2/N$
- $\psi = -N \log(-\eta)$
- $\gamma = 2/\sqrt{N}$
- $\delta = 6/N$

Negative binomial

A coin with probability of heads θ is flipped until exactly $k+1$ heads are observed. Let $Y = \#$ of tails observed. Then

$$\begin{aligned} g(y) &= \binom{y+k}{k}(1-\theta)^y\theta^{k+1} \\ &= \binom{y+k}{k}e^{[\log(1-\theta)]y+(k+1)\log\theta} \qquad [\mathcal{Y}=(0,1,2,\dots)]\,. \end{aligned}$$

This is a one-parameter exponential family with:

- $\eta = \log(1-\theta)$
- $\psi = -(k+1)\log(1-e^\eta)$

Homework 1.12. (a) Find μ, V, and γ as a function of θ. (b) Notice that $\psi = (k+1)\psi_0$ where ψ_0 is the normalizing function for $k=0$. Give a simple explanation for this. (c) How does it affect the formula for μ, V, and γ?

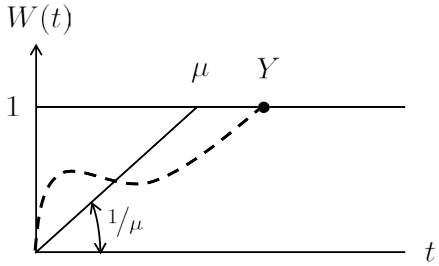
Inverse Gaussian

Let $W(t)$ be a Wiener process with drift $1/\mu$, so $W(t) \sim \mathcal{N}(t/\mu, t)$ $(\text{Cov}[W(t), W(t+d)] = t)$. Define Y as the first passage time to $W(t)=1$. Then Y has the "inverse Gaussian" or Wald density

$$g(y) = \frac{1}{\sqrt{2\pi y^3}}\,e^{-\frac{(y-\mu)^2}{2\mu^2 y}}\,.$$

This is an exponential family with:

- $\eta = -1/(2\mu^2)$
- $\psi = -\sqrt{-2\eta}$
- $V = \mu^3$



Reference Johnson and Kotz, *Continuous Univariate Densities Vol. 1*, Chapter 15

Homework 1.13. Show $Y \sim (\mu, \mu^{3/2}, 3\sqrt{\mu}, 15\mu)$ as the mean, standard deviation, skewness, and kurtosis, respectively.

Note. The early Generalized Linear Model literature was interested in the construction of nonstandard exponential families with relations such as $V = \mu^{1.5}$.

	Normal	Poisson	Gamma	Inverse normal
$V \propto$	constant	μ	μ^2	μ^3

2×2 table

Let $X = (x_1, x_2, x_3, x_4)$ be a multinomial sample of size N from a 4-category multinomial layout, where the categories form a double dichotomy as shown.

	Yes	*No*	Row totals
Men	X_1 π_1	X_2 π_2	r_1
Women	X_3 π_3	X_4 π_4	r_2
Column totals	c_1	c_2	N

$$(x_1, x_2, x_3, x_4) \sim \text{Mult}_4 \left[N, (\pi_1, \pi_2, \pi_3, \pi_4)\right]$$

with $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)$ the true probabilities, $\sum_1^4 \pi_i = 1$. Given the table's marginal totals (N, r_1, c_1), we need only know x_1 to fill in (x_1, x_2, x_3, x_4). (Fisher suggested analyzing the table with marginals thought of as fixed ancillaries, for reasons discussed next.)

The conditional density of x_1 given (N, r_1, c_1) depends only on the *log odds* parameter

$$\theta = \log\left(\frac{\pi_1}{\pi_2} \Big/ \frac{\pi_3}{\pi_4}\right),$$

so conditioning has reduced our four-parameter inferential problem to a simpler, one-parameter situation. Notice that $\theta = 0$ corresponds to $\pi_1/\pi_2 = \pi_3/\pi_4$, which is equivalent to independence between the two dichotomies.

The conditional density of $x_1 \mid (N, r_1, c_1)$, with respect to counting measure, is

$$\begin{aligned} g_\theta(x_1 \mid N, r_1, c_1) &= \binom{r_1}{x_1}\binom{r_2}{c_1 - x_1} e^{\theta x_1} / C(\theta), \\ C(\theta) &= \sum_{x_1} \binom{r_1}{x_1}\binom{r_2}{c_1 - x_1} e^{\theta x_1}, \end{aligned} \tag{1.3}$$

the sum being over the sample space of possible x_1 values,

$$\max(0, c_1 - r_2) \leq x_1 \leq \min(c_1, r_1).$$

Reference Lehmann, "Testing statistical hypotheses", Section 4.5

This is a one-parameter exponential family with:

- $\eta = \theta$
- $y = x$
- $\psi = \log(C)$

	Success	*Failure*	
Treatment	**9**	**12**	**21**
Control	**7**	**17**	**24**
	16	**29**	**45**

Example. The 14th experiment on `ulcdata` involved 45 patients in a clinical trial comparing a new

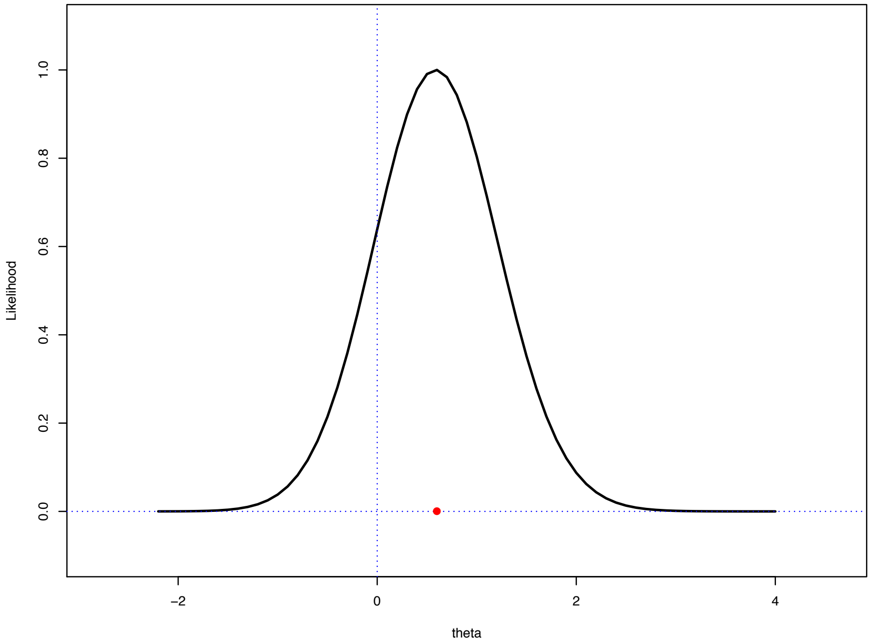


Figure 1.2: `ulcdata` #14; likelihood function for crossproduct ratio θ; max at $\theta = 0.600$; $-\ddot{l} = 2.56$

experimental surgery for stomach ulcers with the standard control procedure. The obvious estimate of θ is

$$\hat{\theta} = \log\left(\frac{9}{12}\Big/\frac{7}{17}\right) = 0.600.$$

Figure 1.2 graphs the likelihood, i.e., expression (1.3) as a function of θ, with the data held fixed as observed (normalized so that $\max\{L(\theta)\} = 1$).

Homework 1.14. (a) Compute the likelihood numerically and verify that it is maximized at $\hat{\theta} = 0.600$. (b) Verify numerically that

$$-\frac{d^2 \log L(\theta)}{d\theta^2}\bigg|_{\hat{\theta}} = 2.56.$$

(c) Using this result, guess the variance of $\hat{\theta}$.

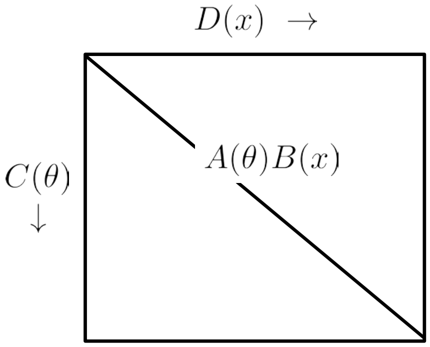
The structure of one-parameter exponential families

Suppose $f_\theta(x)$, θ and x possibly vectors, is a family of densities satisfying

$$\log f_\theta(x) = A(\theta)B(x) + C(\theta) + D(x),$$

A, B, C, D real. Then $\{f_\theta(x)\}$ is a one-parameter exponential family with:

- $\eta = A(\theta)$
- $y = B(x)$
- $\psi = -C(\theta)$
- $\log g_0 = D(x)$



A two-way table of $\log f_\theta(x)$ would have additive components $C(\theta) + D(x)$, and an interaction term $A(\theta)B(x)$.

Homework 1.15. I constructed a 14×9 matrix P with ijth element

$$p_{ij} = \mathrm{Bi}(x_i, \theta_j, 13),$$

the binomial probability of x_i for probability θ_j, sample size $n = 13$, where

$$\begin{aligned} &x_i = i \qquad \text{for } i = 0, 1, 2, \ldots, 13 \\ &\theta_j = 0.1, 0.2, \ldots, 0.9. \end{aligned}$$

Then I calculated the singular value decomposition (svd) of $\log P$. How many nonzero singular values did I see?

1.5 Bayes families

Suppose we observe $Y = y$ from

$$g_\eta(y) = e^{\eta y - \psi(\eta)} g_0(y), \tag{1.4}$$

where η itself has a prior density

$$\eta \sim \pi(\eta) \qquad \text{(with respect to Lebesgue measure on } A\text{)}.$$

Bayes rule gives posterior density for η

$$\pi(\eta \mid y) = \pi(\eta) g_\eta(y) / g(y),$$

where $g(y)$ is the *marginal density*

$$g(y) = \int_A \pi(\eta) g_\eta(y) \; d\eta.$$

(Note that $g_\eta(y)$ is the *likelihood function*, with y fixed and η varying.) Plugging in (1.4) gives

$$\pi(\eta \mid y) = e^{y\eta - \log[g(y)/g_0(y)]} \left[\pi(\eta) e^{-\psi(\eta)}\right]. \tag{1.5}$$

We recognize this as a one-parameter exponential family with:

- natural parameter $\eta = y$
- sufficient statistic $y = \eta$
- CGF $\psi = \log[g(y)/g_0(y)]$
- carrier $g_0 = \pi(\eta)e^{-\psi(\eta)}$

Homework 1.16. (a) Show that prior $\pi(\eta)$ for η corresponds to prior $\pi(\eta)/V_\eta$ for μ. (b) What is the posterior density $\pi(\mu \mid y)$ for μ?

Conjugate priors

Certain choices of $\pi(\eta)$ yield particularly simple forms for $\pi(\eta \mid y)$ or $\pi(\mu \mid y)$, and these are called *conjugate priors*. They play an important role in modern Bayesian applications. As an example, the conjugate prior for Poisson is the gamma.

Homework 1.17. (a) Suppose $y \sim \text{Poi}(\mu)$ and $\mu \sim mG_\nu$, a scale multiple of a gamma with ν degrees of freedom. Show that

$$\mu \mid y \sim \frac{m}{m+1}\, G_{y+\nu}.$$

(b) Then

$$E\{\mu \mid y\} = \frac{m}{m+1}\, y + \frac{1}{m+1}(m\nu)$$

(compared to $E\{\mu\} = m\nu$ *a priori*, so $E\{\mu \mid y\}$ is a linear combination of y and $E\{\mu\}$). (c) What is the posterior distribution of μ having observed $y_1, y_2 \dots, y_n \overset{\text{iid}}{\sim} \text{Poi}(\mu)$?

Diaconis and Ylvisaker (1979, *Ann. Statist.* 269–281) provide a general formulation of conjugacy:

$$y_1, y_2, \dots, y_n \overset{\text{iid}}{\sim} g_\eta(y) = e^{\eta y - \psi(\eta)} g_0(y);$$

the prior for μ wrt Lebesgue measure is

$$\pi_{n_0, y_0}(\mu) = c_0 e^{n_0[\eta y_0 - \psi(\eta)]} / V_\eta,$$

where y_0 is notionally the average of n_0 hypothetical prior observations of y (c_0 the constant making $\pi_{n_0,y_0}(\mu)$ integrate to 1).

Theorem 1.

$$\pi(\mu \mid y_1, y_2, \dots, y_n) = \pi_{n_+, y_+}(\mu),$$

where

$$n_+ = n_0 + n \quad \textit{and} \quad y_+ = \left(n_0 y_0 + \sum_1^n y_i \right) \Bigg/ n_+.$$

Moreover,

$$E\{\mu \mid y_1, y_2, \dots, y_n\} = y_+.$$

Binomial case

$y \sim \mathrm{Bi}(n, \pi)$, hypothetical prior observations y_0 successes out of n_0 tries. Assuming a "beta" prior (Part II) yields Bayes posterior expectation

$$\hat{\theta} = E\{\pi \mid y\} = \frac{y_0 + y}{n_0 + n}.$$

Current Bayes practice favors small amounts of hypothetical prior information, in the binomial case maybe

$$\hat{\theta} = \frac{1 + y}{2 + n},$$

pulling the MLE y/n a little toward $1/2$.

Tweedie's formula

Equation (1.5) gave

$$\pi(\eta \mid y) = e^{y\eta - \lambda(y)} \pi_0(y)$$

where

$$\pi_0(y) = \pi(\eta) e^{-\psi(\eta)} \quad \text{and} \quad \lambda(y) = \log\left[g(y)/g_0(y)\right],$$

$g(y)$ the marginal density of y. Define

$$l(y) = \log\left[g(y)\right] \quad \text{and} \quad l_0(y) = \log\left[g_0(y)\right].$$

We can now differentiate $\lambda(y)$ with respect to y to get the posterior moments (and cumulants) of η given y,

$$E\{\eta \mid y\} = \lambda'(y) = l'(y) - l_0'(y)$$

and

$$\mathrm{Var}\{\eta \mid y\} = \lambda''(y) = l''(y) - l_0''(y).$$

Homework 1.18. Suppose $y \sim \mathcal{N}(\mu, \sigma^2)$, σ^2 known, where μ has prior density $\pi(\mu)$. Show that the posterior mean and variance of μ given y is

$$\mu \mid y \sim \left\{y + \sigma^2 l'(y), \sigma^2 \left[1 + \sigma^2 l''(y)\right]\right\}. \tag{1.6}$$

Reference Efron (2012), "Tweedie's formula and selection bias", *JASA*

1.6 Empirical Bayes

With $y \sim \mathcal{N}(\mu, \sigma^2)$ and $\mu \sim \pi(\mu)$, Tweedie's formula gives posterior expectation

$$\hat{\theta} = E\{\mu \mid y\} = y + \sigma^2 l'(y);$$

y is the MLE of μ so we can think of this as

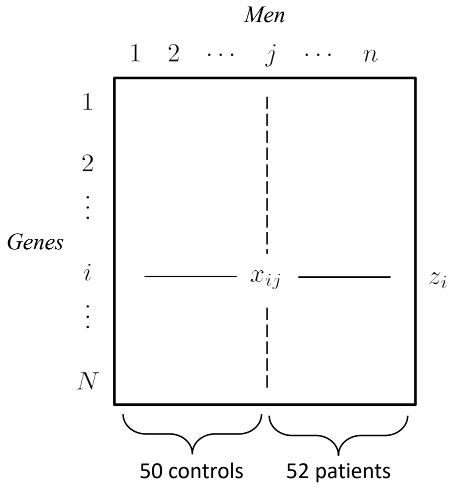
$$\hat{\theta} = \text{MLE} + \text{Bayes correction}.$$

That's fine if we know the prior $\pi(\mu)$, but what if not? In some situations, where we have many parallel experiments observed at the same time, we can effectively learn $\pi(\mu)$ from the data. This is the empirical Bayes approach, as illustrated next.

A microarray analysis

In a study of prostate cancer, $n = 102$ men each had his genetic expression level x_{ij} measured on $N = 6033$ genes,

$$x_{ij} = \begin{cases} i = 1, 2, \ldots, N & \text{genes}, \\ j = 1, 2, \ldots, n & \text{men}. \end{cases}$$



There were:

- $n_1 = 50$ healthy controls
- $n_2 = 52$ prostate cancer patients

For gene$_i$ let t_i = two-sample t statistic comparing patients with controls and

$$z_i = \Phi^{-1}\left[F_{100}(t_i)\right] \qquad (F_{100} \text{ cdf of } t_{100} \text{ distribution});$$

z_i is a *z-value*, i.e., a statistic having a $\mathcal{N}(0,1)$ distribution under the null hypothesis that there is no difference in gene$_i$ expression between patients and controls. (Note: in terms of our previous notation, $y = z_i$ and $\mu = \delta_i$.)

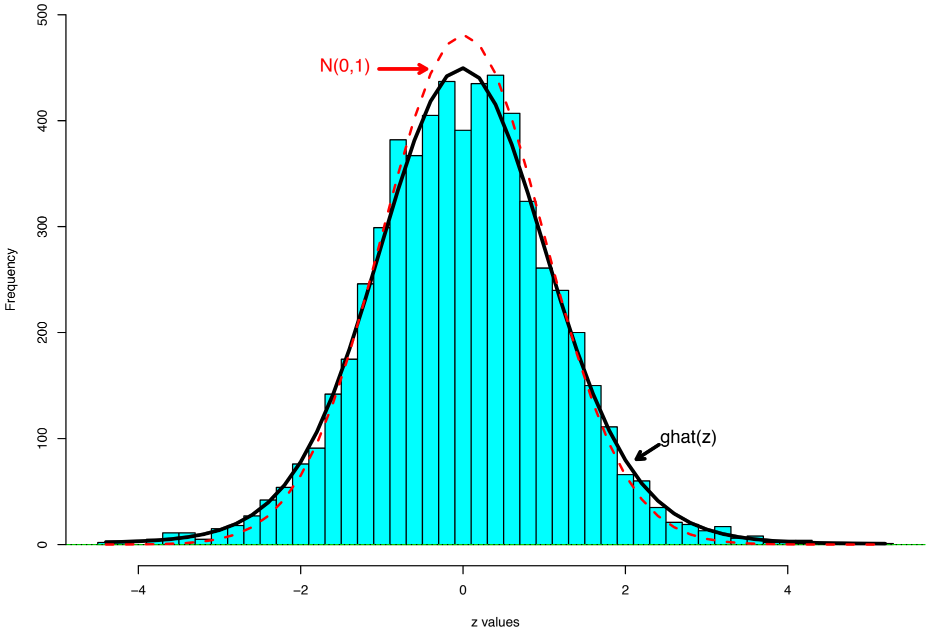


Figure 1.3: Prostate data microarray study. 6033 z-values; heavy curve is $\hat{g}(z)$ from GLM fit; dashed line is $\mathcal{N}(0,1)$.

A reasonable model is

$$z_i \sim \mathcal{N}(\delta_i, 1),$$

where δ_i is the *effect size* for gene i. The investigators were looking for genes with large values of δ_i, either positive or negative. Figure 1.3 shows the histogram of the 6033 z_i values. It is a little wider than a $\mathcal{N}(0,1)$ density, suggesting some non-null ($\delta_i = 0$) genes. Which ones and how much?

Empirical Bayes analysis

1.1 Compute $z_1, z_2, \dots, z_N$; $N = 6033$.

1.2 Fit a smooth parametric estimate $\hat{g}(z)$ to histogram (details in Part II).

1.3 Compute

$$\hat{\lambda}(z) = \log\left[\hat{g}(z)/g_0(z)\right] \qquad \left(\hat{g}_0(z) = \frac{1}{\sqrt{2\pi}}\, e^{-1/2 z^2}\right).$$

1.4 Differentiate $\hat{\lambda}(z)$ to give Tweedie estimates

$$\hat{E}\{\delta \mid z\} = \hat{\lambda}'(z) \quad \text{and} \quad \widehat{\text{Var}}\{\delta \mid z\} = \hat{\lambda}''(z).$$

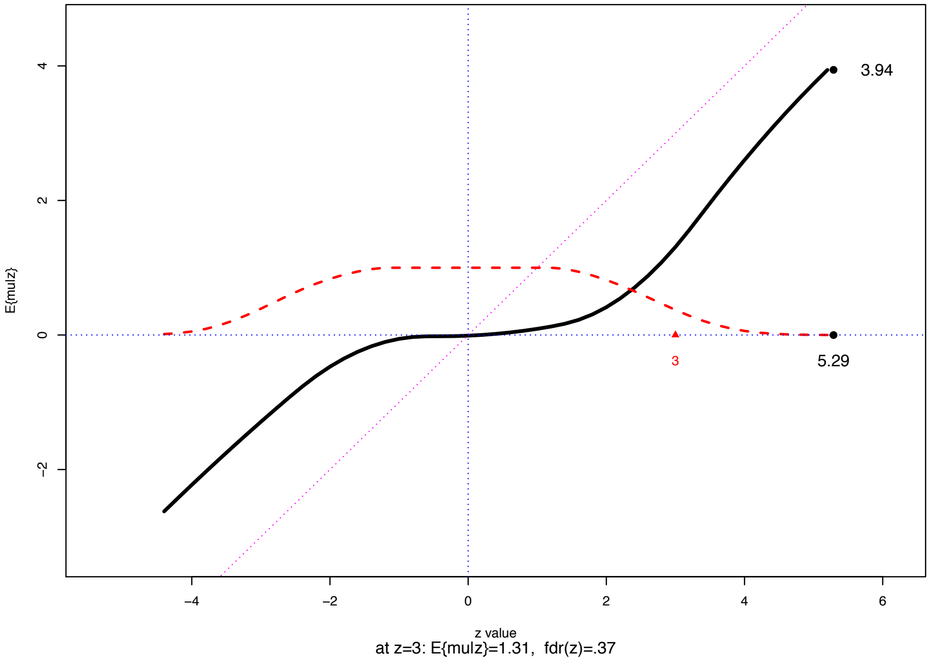


Figure 1.4: Tweedie estimate of $E\{\mu \mid z\}$, prostate study. Dashed curve is estimated local false discovery rate $\text{fdr}(z)$.

Figure 1.4 shows $\hat{E}\{\delta \mid z\}$. It is near zero ("nullness") for $|z| \leq 2$. At $z = 3$, $\hat{E}\{\delta \mid z\} = 1.31$. At $z = 5.29$, the largest observed z_i value (gene #610), $E\{\delta \mid z\} = 3.94$.

The "winner's curse" (regression to the mean)

Even though each z_i is unbiased for its δ_i, it isn't true that $z_{i_{\max}}$ is unbiased for $\delta_{i_{\max}}$ ($i_{\max} = 610$ here). The empirical Bayes estimates $\hat{\theta}_i = \hat{E}\{\delta_i \mid z_i\}$ help correct for the winner's curse ("selection bias"), having $|\hat{\theta}_i| < |z_i|$.

False discovery rates

Let π_0 be the *prior* probability of a null gene, i.e., $\delta = 0$. The "local false discovery rate" is the *posterior* null probability,

$$\text{fdr}(z) = \Pr\{\delta = 0 \mid z\}.$$

Homework 1.19. (a) Show that

$$\text{fdr}(z_i) = \pi_0 g_0(z_i)/g(z_i),$$

where $g(\cdot)$ is the marginal density. (b) In the normal case $z \sim \mathcal{N}(\delta, 1)$, what is the relationship between $\text{fdr}(z)$ and $E\{\delta \mid z\}$?

In practice, $\text{fdr}(z)$ is often estimated by

$$\widehat{\text{fdr}}(z) = g_0(z)/\hat{g}(z),$$

setting $\pi_0 = 1$, an upper bound. This is the dashed curve in Figure 1.4.

1.7 Some basic statistical results

This section briefly reviews some basic statistical results on estimation and testing. A good reference is Lehmann's *Theory of Point Estimation.*

Maximum likelihood and Fisher information

We observe a random sample $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$ from a member $g_\eta(y)$ of an exponential family $\mathcal{G}$,

$$y_i \overset{\text{iid}}{\sim} g_\eta(y), \qquad i = 1, 2, \dots, n.$$

According to (1.2) in Section 1.3, the density of $\boldsymbol{y}$ is

$$g_\eta^{\boldsymbol{Y}}(\boldsymbol{y}) = e^{n[\eta\bar{y} - \psi(\eta)]} \prod_{i=1}^{n} g_0(y_i),$$

where $\bar{y} = \sum_1^n y_i/n$. The log likelihood function $l_\eta(\boldsymbol{y}) = \log g_\eta^{\boldsymbol{Y}}(\boldsymbol{y})$, $\boldsymbol{y}$ fixed and η varying, is

$$l_\eta(\boldsymbol{y}) = n\left[\eta\bar{y} - \psi(\eta)\right],$$

giving *score function* $\dot{l}_\eta(\boldsymbol{y}) = {\partial}/{\partial\eta}\, l_\eta(y)$ equaling

$$\dot{l}_\eta(y) = n(\bar{y} - \mu) \tag{1.7}$$

(remembering that $\dot{\psi}(\eta) = {\partial}/{\partial\eta}\, \psi(\eta)$ equals μ, the expectation parameter).

The maximum likelihood estimate (MLE) of η is the value $\hat{\eta}$ satisfying

$$\dot{l}_{\hat{\eta}}(\boldsymbol{y}) = 0.$$

Looking at (1.7), $\hat{\eta}$ is that η such that $\mu = \dot{\psi}(\eta)$ equals $\bar{y}$,

$$\hat{\eta} : E_{\eta=\hat{\eta}}\{\bar{Y}\} = \bar{y}.$$

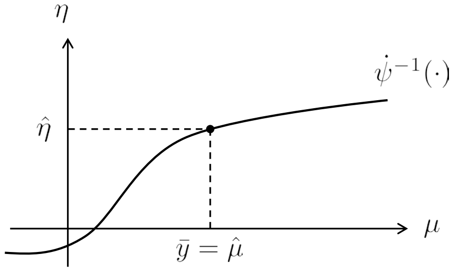
In other words, the MLE matches the theoretical expectation of $\bar{Y}$ to the observed mean $\bar{y}$.

If $\zeta = h(\eta)$ is any function of η then by definition, $\hat{\zeta} = h(\hat{\eta})$ is the MLE of η. If the function is 1-1 then $\hat{\eta} = h^{-1}(\hat{\zeta})$. Notice that $\hat{\mu} = \bar{y}$, so

$$\hat{\eta} = \dot{\psi}^{-1}(\hat{\mu}) = \dot{\psi}^{-1}(\bar{y}).$$

(If $\gamma = H(\mu)$ then $\hat{\gamma} = H(\hat{\mu}) = H(\bar{y})$.) For the Poisson, $\hat{\eta} = \log(\bar{y})$ and for the binomial, according to Section 1.4,

$$\hat{\eta} = \log \frac{\hat{\pi}}{1-\hat{\pi}} \qquad \text{where } \hat{\pi} = y/N.$$



We can also take score functions with respect to μ, say "at η wrt μ":

$$\begin{aligned}\frac{\partial}{\partial \mu}\, l_\eta(\boldsymbol{y}) &= \frac{{}^{\partial}\!/_{\partial\eta}\; l_\eta(\boldsymbol{y})}{{}^{\partial\mu}\!/_{\partial\eta}} = \dot{l}_\eta(\boldsymbol{y})/V \\ &= \frac{n(\bar{y}-\mu)}{V}.\end{aligned}$$

In general, the score function for parameter $\zeta = h(\eta)$, "at η wrt ζ", is

$$\partial/\partial\zeta\; l_\eta(\boldsymbol{y}) = \dot{l}_\eta(\boldsymbol{y})\big/\dot{h}(\eta) = n(\bar{y}-\mu)\big/\dot{h}(\eta).$$

The expected score function is

$$E_\eta \partial/\partial\zeta\; l_\eta(\boldsymbol{y}) = 0.$$

The *Fisher information* in $\boldsymbol{y}$, for ζ at η, sample size n, is $\mathrm{Var}\{\partial/\partial\zeta\, \dot{l}_\eta(\boldsymbol{y})\}$,

$$\begin{aligned} i_\eta^{(n)}(\zeta) &= E_\eta\left[\partial/\partial\eta\; l_\eta(\boldsymbol{y})\right]^2 = E_\eta\left[\dot{l}_\eta(\boldsymbol{y})\big/\dot{h}(\eta)\right]^2 \\ &= i_\eta^{(n)}\big/\dot{h}(\eta)^2. \end{aligned}$$

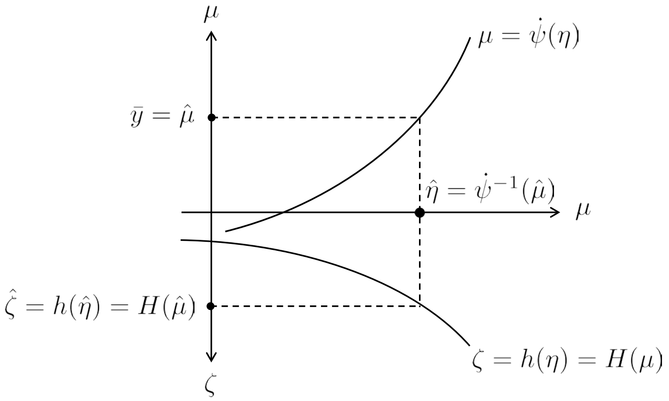
Here $i_\eta^{(n)}$ is the Fisher information in $\boldsymbol{y}$ at η, for η,

$$i_\eta^{(n)} = E_\eta\left[n(\bar{y}-\mu)\right]^2 = nV_\eta.$$

(Notice that $V_\eta = i_\eta$, the information in a single observation, omitting the "1" from the notation.) For parameter μ,

$$i_\eta^{(n)}(\mu) = \frac{nV_\eta}{V_\eta^2} = \frac{n}{V_\eta}.$$

(Always seems wrong!)


μ
$\mu = \dot\psi(\eta)$
$\bar y = \hat\mu$
$\hat\eta = \dot\psi^{-1}(\hat\mu)$
μ
$\hat\zeta = h(\hat\eta) = H(\hat\mu)$
$\zeta = h(\eta) = H(\mu)$
ζ

Score Functions	*Fisher Information*
$\eta: \quad \dot l_\eta(\boldsymbol y) = n(\bar y - \mu)$	$i_\eta^{(n)} = \mathrm{Var}_\eta\left[\dot l_\eta(\boldsymbol y)\right] = nV = ni_\eta$
$\mu: \quad \dfrac{\partial l_\eta(\boldsymbol y)}{\partial \mu} = \dfrac{n(\bar y - \mu)}{V}$	$i_\eta^{(n)}(\mu) = \dfrac{n}{V} = ni_\eta(\mu)$
$\zeta: \quad \dfrac{\partial l_\eta(\boldsymbol y)}{\partial \zeta} = \dfrac{n(\bar y - \mu)}{\dot h(\eta)}$	$i_\eta^{(n)}(\zeta) = \dfrac{nV}{\dot h(\eta)^2} = ni_\eta(\zeta)$

In general the Fisher information i_θ has two expressions, in terms of the 1st and 2nd derivatives of the log likelihood,

$$i_\theta = E\left\{\left(\frac{\partial l_\theta}{\partial \theta}\right)^2\right\} = -E\left\{\frac{\partial^2 l_\theta}{\partial \theta^2}\right\}.$$

For $i_\eta^{(n)}$, the Fisher information for η in $\boldsymbol y = (y_1, y_2, \ldots, y_n)$, we have

$$\begin{aligned} -\ddot l_\eta(\boldsymbol y) &= -\frac{\partial^2}{\partial \eta^2}\, n(\eta \bar y - \psi) = -\frac{\partial}{\partial \eta}\, n(\bar y - \mu) \\ &= nV_\eta = i_\eta^{(n)}, \end{aligned}$$

so in this case $-\ddot l_\eta(\boldsymbol y)$ gives $i_\eta^{(n)}$ without requiring an expectation over $\boldsymbol y$.

Homework 1.20. (a) Does

$$i_\eta^{(n)}(\mu) = -\frac{\partial^2}{\partial \mu^2}\, l_\eta(\boldsymbol y)?$$

(b) Does

$$i_{\eta=\hat\eta}^{(n)}(\mu) = -\frac{\partial}{\partial \mu^2}\, l_\eta(\boldsymbol y)\bigg|_{\eta=\hat\eta} ? \qquad (\hat\eta \text{ the MLE})$$

Cramér–Rao lower bound

The CRLB for an unbiased estimator $\bar{\zeta}$ for ζ is

$$\mathrm{Var}_\eta(\bar{\zeta}) \geq \frac{1}{i_\eta^{(n)}(\zeta)} = \dot{h}(\eta)^2/nV_\eta.$$

For $\zeta \equiv \mu$,

$$\mathrm{Var}(\bar{\mu}) \geq \frac{V_\eta^2}{nV_\eta} = \frac{V_\eta}{n}.$$

In this case the MLE $\hat{\mu} = \bar{y}$ is unbiased and achieves the CRLB. This happens only for μ or linear functions of μ, and not for η, for instance.

In general, the MLE $\hat{\zeta}$ is *not* unbiased for $\zeta = h(\eta)$, but the bias is of order $1/n$,

$$E_\eta\{\hat{\zeta}\} = \zeta + B(\eta)/n.$$

A more general form of the CRLB gives

$$\mathrm{Var}_\eta(\hat{\zeta}) \geq \frac{\left[\dot{h}(\eta) + \dot{B}(\eta)/n\right]^2}{nV_\eta} = \frac{\dot{h}(\eta)^2}{nV_\eta} + O\left(\frac{1}{n^2}\right).$$

Usually $\dot{h}(\eta)^2/(nV_\eta)$ is a reasonable approximation for $\mathrm{Var}_\eta(\hat{\zeta})$.

Delta method

If X has mean μ and variance σ^2, then $Y = H(X) \doteq H(\mu) + H'(\mu)(X - \mu)$ has approximate mean and variance

$$Y \dot{\sim} \left\{H(\mu), \sigma^2\left[H'(\mu)\right]^2\right\}.$$

Homework 1.21. Show that if $\zeta = h(\eta) = H(\mu)$, then the MLE $\hat{\zeta}$ has delta method approximate variance

$$\mathrm{Var}_\eta(\hat{\zeta}) \doteq \frac{\dot{h}(\eta)^2}{nV_\eta},$$

in accordance with the CRLB $1/i_\eta^{(n)}(\zeta)$. (In practice we must substitute $\hat{\eta}$ for η in order to estimate $\mathrm{Var}_\eta(\hat{\zeta})$.)

Hypothesis testing (Lehmann)

Suppose we wish to test

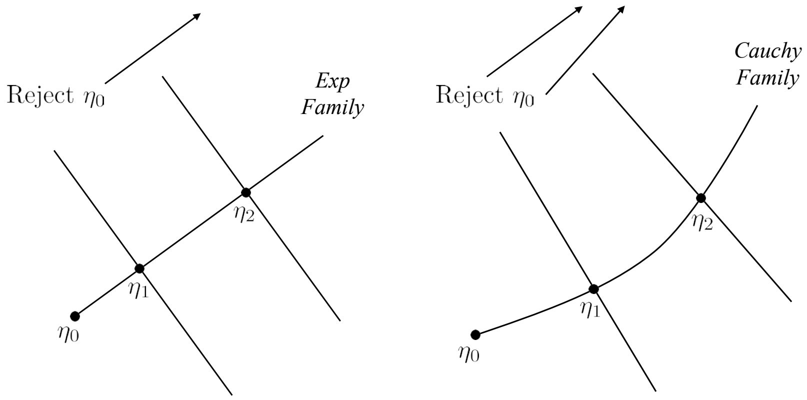
$$H_0: \quad \eta = \eta_0 \quad \text{versus} \quad H_A: \eta = \eta_1 \qquad (\eta_1 > \eta_0).$$

- $\log\{g_{\eta_1}(y)/g_{\eta_0}(y)\} = (\eta_1 - \eta_0)y - [\psi(\eta_1) - \psi(\eta_0)] \uparrow y$

- By the Neyman–Pearson lemma, MP$_\alpha$ test rejects for $y \geq Y_0^{(1-\alpha)}$ where $Y_0^{(1-\alpha)}$ is $(1-\alpha)$th quantile of Y under H_0.

- This doesn't depend on η_1, so the test is UMP$_\alpha$.

- For non-exponential families, such as Cauchy translation family, the MP$_\alpha$ test depends on η_1: "A one-parameter exponential family is a straight line through the space of probability distributions." (Efron 1975, *Ann. Statist.* pp. 1189-1281)



1.8 Deviance and Hoeffding's formula

Deviance is an analogue of Euclidean distance applied to exponential families $g_\eta(y) = e^{\eta y - \psi(\eta)} g_0(y)$. By definition the deviance $D(\eta_1, \eta_2)$ between g_{η_1} and g_{η_2} in family $\mathcal{G}$ is

$$\begin{aligned} D(\eta_1, \eta_2) &= 2E_{\eta_1}\left\{\log\left(\frac{g_{\eta_1}(y)}{g_{\eta_2}(y)}\right)\right\} \\ &= 2\int_{\mathcal{Y}} g_{\eta_1}(y) \log\left[g_{\eta_1}(y)/g_{\eta_2}(y)\right]\ m(dy). \end{aligned}$$

We will also write $D(\mu_1, \mu_2)$ or just $D(1,2)$; the deviance is the distance between the two densities, not their indices.

Homework 1.22. Show that $D(\eta_1, \eta_2) \geq 0$, with strict inequality unless the two densities are identical.

Note. In general, $D(\eta_1, \eta_2) \neq D(\eta_2, \eta_1)$.

Older name

The "Kullback–Leibler distance" equals $D(\eta_1, \eta_2)/2$. Information theory uses "mutual information" for $D[f(x,y), f(x)f(y)]/2$, where $f(x,y)$ is a bivariate density and $f(x)$ and $f(y)$ its marginals.

Homework 1.23. Verify these formulas for the deviance:

$$\begin{aligned}
\text{Poisson } Y \sim \text{Poi}(\mu): &\quad D(\mu_1,\mu_2) = 2\mu_1\left[\log\left(\frac{\mu_1}{\mu_2}\right) - \left(1-\frac{\mu_2}{\mu_1}\right)\right]\\
\text{Binomial } Y \sim \text{Bi}(N,\pi): &\quad D(\pi_1,\pi_2) = 2N\left[\pi_1\log\left(\frac{\pi_1}{\pi_2}\right) + (1-\pi_1)\log\left(\frac{1-\pi_1}{1-\pi_2}\right)\right]\\
\text{Normal } Y \sim \mathcal{N}(\mu,1): &\quad D(\mu_1,\mu_2) = (\mu_1-\mu_2)^2\\
\text{Gamma } Y \sim \lambda G_N: &\quad D(\lambda_1,\lambda_2) = 2N\left[\log\left(\frac{\lambda_2}{\lambda_1}\right) + \left(\frac{\lambda_1}{\lambda_2}-1\right)\right]\\
&\quad = 2N\left[\log\left(\frac{\mu_2}{\mu_1}\right) + \left(\frac{\mu_1}{\mu_2}-1\right)\right]
\end{aligned}$$

Hoeffding's formula

Let $\hat\eta$ be the MLE of η having observed y. Then

$$\boxed{g_\eta(y) = g_{\hat\eta}(y)e^{-D(\hat\eta,\eta)/2}.}$$

Indexing the family with the expectation parameter μ rather than η, and remembering that $\hat\mu = y$, we get a more memorable version of Hoeffding's formula,

$$\begin{aligned}
g_\mu(y) &= g_{\hat\mu}(y)e^{-D(\hat\mu,\mu)/2}\\
&= g_y(y)e^{-D(y,\mu)/2}.
\end{aligned} \tag{1.8}$$

This last says that a plot of the log likelihood $\log[g_\mu(y)]$ declines from its maximum at $\mu = y$ according to the deviance,

$$\log\left[g_\mu(y)\right] = \log\left[g_y(y)\right] - D(y,\mu)/2.$$

In our applications of the deviance, the first argument will always be the data, the second a proposed value of the unknown parameter.

Proof. The deviance in an exponential family is

$$\begin{aligned}
\frac{D(\eta_1,\eta_2)}{2} &= E_{\eta_1}\log\frac{g_{\eta_1}(y)}{g_{\eta_2}(y)} = E_{\eta_1}\left\{(\eta_1-\eta_2)y - \psi(\eta_1) + \psi(\eta_2)\right\}\\
&= (\eta_1-\eta_2)\mu_1 - \psi(\eta_1) + \psi(\eta_2).
\end{aligned}$$

Therefore

$$\frac{g_\eta(y)}{g_{\hat\eta}(y)} = \frac{e^{\eta y-\psi(\eta)}}{e^{\hat\eta y-\psi(\hat\eta)}} = e^{(\eta-\hat\eta)y-\psi(\eta)+\psi(\hat\eta)} = e^{(\eta-\hat\eta)\hat\mu-\psi(\eta)+\psi(\hat\eta)}.$$

Taking $\eta_1 = \eta$ and $\eta_2 = \hat\eta$, this last is $D(\hat\eta,\eta)$. ■

Repeated sampling

If $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ is an iid sample from $g_\eta(\cdot)$ then the deviance based on $\boldsymbol{y}$, say $D_n(\eta_1, \eta_2)$, is

$$\begin{aligned} D_n(\eta_1,\eta_2) = 2E_{\eta_1} \log\left[g_{\eta_1}^{\boldsymbol{Y}}(\boldsymbol{y})/g_{\eta_2}^{\boldsymbol{Y}}(\boldsymbol{y})\right] &= 2E_{\eta_1}\left\{\log \prod_{i=1}^n \left[\frac{g_{\eta_1}(y_i)}{g_{\eta_2}(y_i)}\right]\right\} \\ &= 2\sum_{i=1}^n \left\{E_{\eta_1} \log\left[\frac{g_{\eta_1}(y_i)}{g_{\eta_2}(y_i)}\right]\right\} = nD(\eta_1,\eta_2). \end{aligned}$$

(This fact shows up in the binomial, Poisson, and gamma cases of Homework 1.16.)

Note. We are indexing the possible distributions of $\boldsymbol{Y}$ with η, not $\eta^{(n)} = n\eta$.

Homework 1.24. What is the deviance formula for the negative binomial family?

Relationship with Fisher information

For η_2 near η, the deviance is related to the Fisher information $i_{\eta_1} = V_{\eta_1}$ (in a single observation y, for η_1 and at η_1):

$$\boxed{D(\eta_1,\eta_2) = i_{\eta_1}(\eta_2-\eta_1)^2 + O(\eta_2-\eta_1)^3.}$$

Proof.

$$\frac{\partial}{\partial\eta_2} D(\eta_1,\eta_2) = \frac{\partial}{\partial\eta_2} 2\left\{(\eta_1-\eta_2)\mu_1 - [\psi(\eta_1)-\psi(\eta_2)]\right\} = 2(-\mu_1+\mu_2) = 2(\mu_2-\mu_1).$$

Also

$$\frac{\partial^2}{\partial\eta_2^2} D(\eta_1,\eta_2) = 2\frac{\partial\mu_1}{\partial\eta_2} = 2V_{\eta_2}.$$

Therefore

$$\left.\frac{\partial}{\partial\eta_2} D(\eta_1,\eta_2)\right|_{\eta_2=\eta_1} = 0 \quad \text{and} \quad \left.\frac{\partial^2}{\partial\eta_2^2} D(\eta_1,\eta_2)\right|_{\eta_2=\eta_1} = 2V_{\eta_1},$$

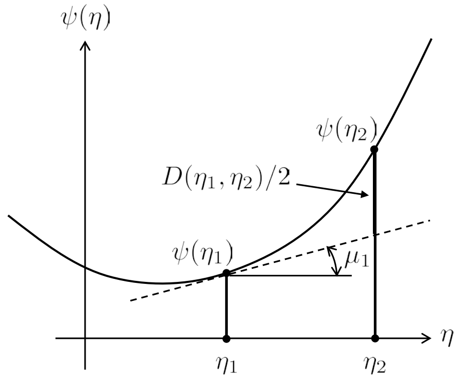
so

$$D(\eta_1,\eta_2) = 2V_{\eta_1}\frac{(\eta_2-\eta_1)^2}{2} + O(\eta_2-\eta_1)^3. \qquad \blacksquare$$

Homework 1.25. What is $\partial^3 D(\eta_1,\eta_2)/\partial\eta_2^3$?

An informative picture

$\psi(\eta)$ is a convex function of η since $\ddot{\psi}(\eta) = V_\eta > 0$. The picture shows $\pi(\eta)$ passing through $(\eta_1, \psi(\eta_1))$ at slope $\mu_1 = \dot{\psi}(\eta_1)$. The difference between $\psi(\eta_2)$ and the linear bounding line $\psi(\eta_1)+(\eta_2-\eta_1)\mu_1$ is $\psi(\eta_2)-\psi(\eta_1)+(\eta_1-\eta_2)\mu_1 = D(\eta_1,\eta_2)/2$.



The previous picture, unlike our other results, depends on parameterizing the deviance as $D(\eta_1,\eta_2)$. A version that uses $D(\mu_1,\mu_2)$ depends on the *dual function* $\phi(y)$ to $\psi(y)$,

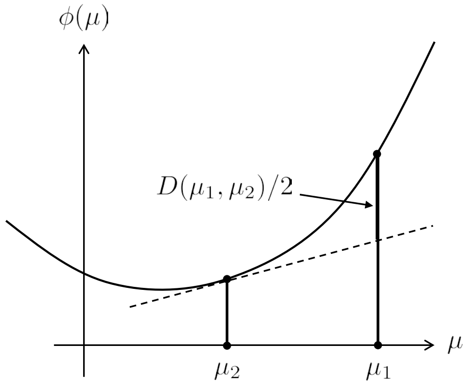
$$\phi(y) = \max_\eta \{\eta y - \psi(\eta)\}.$$

Reference Efron (1978), “Geometry of exponential families”, *Ann. Statist.*

Homework 1.26. Show that (a) $\phi(\mu) = \eta\mu - \psi(\eta)$, where $\mu = \dot{\psi}(\eta)$; (b) $\phi(\mu)$ is convex as a function of μ; and (c) $d\phi(\mu)/d\mu = \eta$. (d) Verify the picture at right.

Homework 1.27. Parametric bootstrap: we resample y^* from $g_{\hat{\eta}}(\cdot)$, $\hat{\eta}$ = MLE based on y. Show that

$$g_\eta(y^*) = g_{\hat{\eta}}(y^*) e^{(\eta-\hat{\eta})(y^*-y)-D(\hat{\eta},\eta)/2}.$$



Deviance residuals

The idea: if $D(y,\mu)$ is the analogue of $(y-\mu)^2$ in a normal model, then

$$\text{sign}(y-\mu)\sqrt{D(y,\mu)}$$

should be the exponential family analogue of a normal residual $y-\mu$.

We will work in the repeated sampling framework

$$y_i \overset{\text{iid}}{\sim} g_\mu(\cdot), \qquad i = 1,2,\dots,n,$$

with MLE $\hat{\mu} = \bar{y}$ and total deviance $D_i(\hat{\mu},\mu) = nD(\bar{y},\mu)$. The *deviance residual*, of $\hat{\mu} = \bar{y}$ from true mean μ, is defined to be

$$R = \text{sign}(\bar{y} - \mu)\sqrt{D(\bar{y}, \mu)}. \tag{1.9}$$

The hope is that R will be nearly $\mathcal{N}(0, 1)$, closer to normal than the obvious "Pearson residual"

$$R_P = \frac{\bar{y} - \mu}{\sqrt{V_\mu / n}}$$

(called "z_i" later). Our hope is bolstered by the following theorem, verified in Appendix C of McCullagh and Nelder, *Generalized Linear Models.*

Theorem 2. *The asymptotic distribution of R as $n \to \infty$ is*

$$R \dot{\sim} \mathcal{N}\left[-a_n, (1 + b_n)^2\right], \tag{1.10}$$

where a_n and b_n are defined in terms of the skewness and kurtosis of the original $(n = 1)$ exponential family,

$$a_n = (\gamma_\mu / 6)/\sqrt{n} \quad \text{and} \quad b_n = \left[(7/36)\, \gamma_\mu^2 - \delta_\mu\right]/n.$$

The normal approximation in (1.10) is accurate through $O_p\left(1/n\right)$, with errors of order $O_p\left(1/n^{3/2}\right)$, e.g.,

$$\Pr\left\{\frac{R + a_n}{1 + b_n} > 1.96\right\} = 0.025 + O\left(1/n^{3/2}\right)$$

(so-called "third order accuracy").

Corollary 1.

$$D_n(\bar{y}, \mu) = R^2 \dot{\sim} \left(1 + \frac{5\gamma_\mu^2 - 3\delta_\mu}{12n}\right) \cdot \chi_1^2,$$

χ_1^2 a chi-squared random variable with degrees of freedom 1. Since

$$D_n(\bar{y}, \mu) = 2 \log\left[g_{\hat{\mu}}^{\boldsymbol{Y}}(\boldsymbol{y}) / g_\mu^{\boldsymbol{Y}}(\boldsymbol{y})\right]$$

according to Hoeffding's formula, the corollary is an improved version of Wilks' theorem, i.e., $2\log(g_{\hat{\mu}}/g_\mu) \to \chi_1^2$ in one-parameter situations.

The constants a_n and b_n are called "Bartlett corrections". The theorem says that

$$R \dot{\sim} (Z + a_n)/(1 + b_n) \qquad \text{where } Z \sim \mathcal{N}(0, 1).$$

Since $a_n = O(1/\sqrt{n})$ and $b_n = O(1/n)$, the expectation correction in (1.10) is more important than the variance correction.

Homework 1.28. Gamma case, $y \sim \lambda G_N$ with N fixed (N can be thought of as n). (a) Show that the deviance residual $\text{sign}(y - \lambda N)\sqrt{D(y, \lambda N)}$ has the same distribution for all choices of λ. (b) What is the skewness of the Pearson residual $(y - \lambda N)/\lambda\sqrt{N}$?

Homework 1.29. Use our previous results to show that

$$D_n(\bar{y},\mu) \doteq R_P^2 + \frac{\gamma}{6\sqrt{n}}\, R_P^3 + O_P\left({}^1\!/_n\right).$$

An example

Figure 1.5 shows the results of 2000 replications of $y \sim G_5$ (or equivalently,

$$\bar{y} = \sum_1^5 y_i/5,$$

where y_i are independent G_1 variates, that is, standard one-sided exponentials). The qq-plot shows the deviance residuals (black) much closer to $\mathcal{N}(0,1)$ than the Pearson residuals (red).

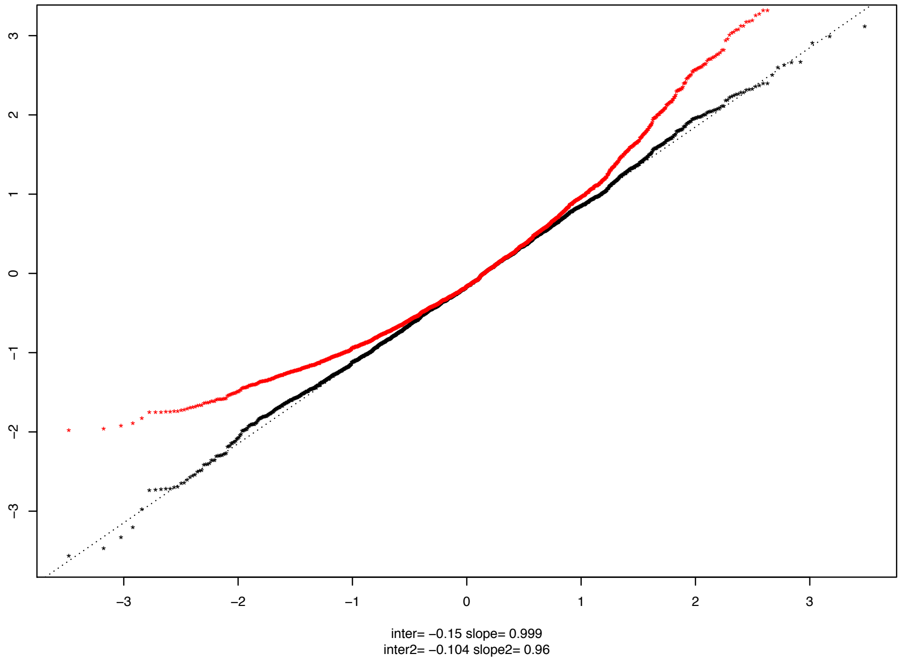


Figure 1.5: qq comparison of deviance residuals (black) with Pearson residuals (red); gamma $N = 1$, $\lambda = 1$, $n = 5$; $B = 2000$ simulations.

Homework 1.30. Compute a version of Figure 1.5 applying to $y \sim \text{Poi}(16)$.

An example of Poisson deviance analysis

Reference Thisted and Efron (1987), "Did Shakespeare write a newly discovered poem?", *Biometrika*

- A newly discovered poem is of total length 429 words, comprising 258 different words. An analysis is done to test the hypothesis that Shakespeare wrote the poem.

- 9 of the 258 works never appeared in the 884,647 total words of known Shakespeare; 7 of the 258 words appeared once each in the known Shakespeare, etc., as presented in column "y" of the table.

- A simple theory predicts 6.97 for the expected number of "new" words, given Shakespearean authorship, 4.21 "once before" words, etc., presented in column "ν" of the table. The theory also predicts independent Poisson distributions for the y values,

$$y_i \overset{\text{ind}}{\sim} \text{Poi}(\nu_i) \qquad \text{for } i = 1, 2, \dots, 11.$$

- "Dev" shows the Poisson deviances; the total deviance 19.98 is moderately large compared to a chi-squared distribution with 11 degrees of freedom, $P\{\chi^2_{11} > 19.98\} = 0.046$. This casts some moderate doubt on Shakespearan authorship.

- "R" is the signed square root of the deviance; "a_n" is the correction $1/6 \times \nu^{1/2}$ suggested by the theorem (1.10); "RR" is the corrected residual $R + a_n$. These should be approximately $\mathcal{N}(0, 1)$ under the hypothesis of Shakespearean authorship. The residual for 20–29 looks suspiciously large.

- 8 out of 11 of the RR's are positive, suggesting that the y's may be systematically larger than the ν's. Adding up the 11 cases,

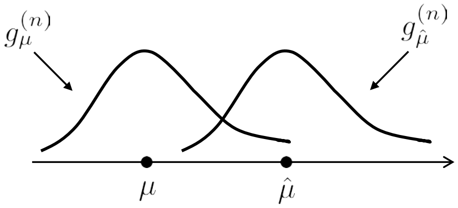
$$y^+ = 118, \qquad \nu^+ = 94.95.$$

This gives $D^+ = \text{Dev}(y^+, \nu^+) = 5.191$, $R^+ = 2.278$, and $RR^+ = 2.295$. The normal probability of exceeding 2.295 is 0.011, considerably stronger evidence (but see the paper). The actual probability is

$$\Pr\{\text{Poi}(94.95) \geq 118\} = 0.011.$$

# Prev	y	ν	Dev	R	a_n	RR
0	9	6.97	.5410	.736	.0631	.799
1	7	4.21	1.5383	1.240	.0812	1.321
2	5	3.33	.7247	.851	.0913	.943
3–4	8	5.36	1.1276	1.062	.0720	1.134
5–9	11	10.24	.0551	.235	.0521	.287
10–19	10	13.96	1.2478	−1.117	.0446	−1.072
20–29	21	10.77	7.5858	2.754	.0508	2.805
30–39	16	8.87	4.6172	2.149	.0560	2.205
40–59	18	13.77	1.1837	1.088	.0449	1.133
60–79	8	9.99	.4257	−.652	.0527	−.600
80–99	5	7.48	.9321	−.965	.0609	−.904

1.9 The saddlepoint approximation



We observe a random sample of size n from some member of an exponential family $\mathcal{G}$,

$$y_1, y_2, \dots, y_n \overset{\text{iid}}{\sim} g_\mu(\cdot)$$

(now indexed by expectation parameter μ), and wish to approximate the density of the sufficient statistic $\hat{\mu} = \bar{y}$ for some value of $\hat{\mu}$ perhaps far removed from μ. Let $g_\mu^{(n)}(\hat{\mu})$ denote this density.

The normal approximation

$$g_\mu^{(n)}(\hat{\mu}) \doteq \sqrt{\frac{n}{2\pi V_\mu}}\, e^{-\frac{1}{2}\frac{n}{V_\mu}(\hat{\mu}-\mu)^2}$$

is likely to be inaccurate if $\hat{\mu}$ is say several standard errors removed from μ. Hoeffding's formula gives a much better result, called the *saddlepoint approximation*:

$$\begin{aligned} g_\mu^{(n)}(\hat{\mu}) &= g_{\hat{\mu}}^{(n)}(\hat{\mu}) e^{-D_n(\hat{\mu},\mu)/2} \qquad [D_n(\hat{\mu},\mu) = nD(\hat{\mu},\mu)] \\ &\boxed{\doteq \sqrt{\frac{n}{2\pi V_{\hat{\mu}}}}\, e^{-D_n(\hat{\mu},\mu)/2}} \end{aligned} \tag{1.11}$$

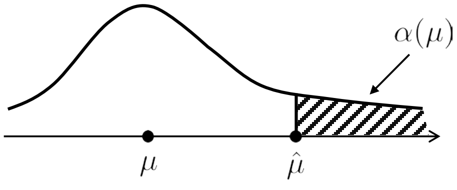
Here $V_{\hat{\mu}} = \ddot{\psi}(\hat{\eta})$, the variance of a single y_i if $\mu = \hat{\mu}$.

The approximation

$$g_{\hat{\mu}}^{(n)}(\hat{\mu}) \doteq \sqrt{\frac{n}{2\pi V_{\hat{\mu}}}}$$

comes from applying the central limit theorem *at the center of the* $g_{\hat{\mu}}^{(n)}(\cdot)$ *distribution*, just where it is most accurate. There is an enormous literature of extensions and improvements to the saddlepoint approximation: a good review article is Reid (1988) in *Statistical Science*.

The Lugananni–Rice formula



The saddlepoint formula can be integrated to give an approximation to $\alpha(\mu)$, the *attained significance level* or "p-value" of parameter value μ having observed $\bar{y} = \hat{\mu}$:

$$\alpha(\mu) = \int_{\hat{\mu}}^{\infty} g_\mu^{(n)}(t)\, dt.$$

Numerical integration is required to compute $\alpha(\mu)$ from the saddlepoint formula itself, but the *Lugananni–Rice formula* provides a highly accurate closed-form approximation:

$$\alpha(\mu) \doteq 1 - \Phi(R) - \varphi(R)\left(\frac{1}{R} - \frac{1}{Q}\right) + O\left(\frac{1}{n^{3/2}}\right),$$

where Φ and φ are the standard normal cdf and density,

$$R = \operatorname{sign}(\hat{\mu} - \mu)\sqrt{nD(\hat{\mu}, \mu)}$$

the deviance residual, and

$$Q = \sqrt{nV_{\hat{\mu}}} \cdot (\hat{\eta} - \eta)$$

the crude form of the Pearson residual based on the canonical parameter η, not on μ. (Remember that $\widehat{\mathrm{sd}}(\hat{\eta}) \doteq 1/\sqrt{n\hat{V}}$, so $Q = (\hat{\eta} - \eta)/\widehat{\mathrm{sd}}(\hat{\eta})$.) Reid (1988) is also an excellent reference here, giving versions of the L-R formula that apply not only to exponential family situations but also to general distributions of $\bar{y}$.

Homework 1.31. Suppose we observe $y \sim \lambda G_N$, G_N gamma df $= N$, with $N = 10$ and $\lambda = 1$. Use the L-R formula to calculate $\alpha(\mu)$ for $y = \hat{\mu} = 15, 20, 25, 30$, and compare the results with the exact values. (You can use any function R.)

Homework 1.32. Another version of the L-R formula is

$$1 - \alpha(\mu) \doteq \Phi(R'),$$

where

$$R' = R + \frac{1}{R}\log\left(\frac{Q}{R}\right).$$

How does this relate to the first form?

Large deviations and exponential tilting

In a generic "large deviations" problem, we observe an iid sample

$$y_1, y_2, \dots, y_n \overset{\text{iid}}{\sim} g_0(\cdot)$$

from a known density g_0 having mean and standard deviation

$$y_i \sim (\mu_0, \sigma_0).$$

We wish to compute

$$\alpha_n(\mu) = \Pr_0\{\bar{y} \geq \mu\}$$

for some fixed value $\mu > \mu_0$. As $n \to \infty$, the number of standard errors $\sqrt{n}(\mu - \mu_0)/\sigma_0$ gets big, rendering the central limit theorem useless.

Homework 1.33 ("Chernoff bound"). $g_\eta(y) = e^{\eta y - \psi(\eta)} g_0(y)$

(a) For any $\lambda > 0$ show that $\alpha_n(\mu) = \Pr_0\{\bar{y} \geq \mu\}$ satisfies

$$\alpha_n(\mu) \leq \beta_n(\mu) \equiv \int_{\mathcal{Y}} e^{n\lambda(\bar{y} - \mu)} g_0(y)\, dy.$$

(b) Show that $\beta_n(\mu)$ is minimized at $\lambda = \eta$.

(c) Finally, verify Chernoff's large deviation bound

$$\Pr_0\{\bar{y} \geq \mu\} \leq e^{-nD(\mu,0)},$$

where $D(\mu, 0)$ is the deviance between $g_\eta(y)$ and $g_0(y)$.

Notice that for fixed μ, $\alpha_n(\mu) \to 0$ exponentially fast, which is typical for large deviation results.

Homework 1.34. Extra credit: Suppose $g_0(y) = 1$ for y in $[0, 1]$ and 0 otherwise. Calculate the Chernoff bound for $\Pr_0\{\bar{y} \geq 0.9\}$.

1.10 Transformation theory

REFERENCE Hougaard (1982), *JRSS-B*; DiCiccio (1984) *Biometrika*; Efron (1982), *Ann. Statist.*

Power transformations are used to make exponential families more like the standard normal translation family $Y \sim \mathcal{N}(\mu, 1)$. For example, $Y \sim \text{Poi}(\mu)$ has variance $V_\mu = \mu$ depending on the expectation μ, while the transformation

$$Z = h(Y) = 2\sqrt{Y}$$

yields, approximately, $\text{Var}(Z) = 1$ for all μ. In a regression situation with Poisson responses $y_1, y_2, \ldots, y_n$, we might first change to $z_i = 2\sqrt{y_i}$ and then employ standard linear model methods. (That's *not* how we will proceed in Part II, where generalized linear model techniques are introduced.)

The following display summarizes an enormous number of transformation results for one-parameter exponential families. Let

$$\zeta = h(\mu)$$

and likewise $Z = h(Y)$ and $\hat{\zeta} = h(\hat{\mu})$. The choice of transformation $h(\cdot)$ satisfying

$$h'(\mu) = V_\mu^{\delta-1}$$

then results in:

δ	$1/3$	$1/2$	$2/3$
result	normal likelihood	stabilized variance	normal density

The stabilized variance result follows from the delta method:

$$\hat{\zeta} = h(\hat{\mu}) \qquad \text{with } h'(\mu) = \frac{1}{\sqrt{V_\mu}}$$

implies that

$$\mathrm{sd}_\mu(\hat{\zeta}) \doteq \frac{\mathrm{sd}_\mu(\hat{\mu})}{\sqrt{V_\mu}} = 1.$$

For the Poisson family, with $V_\mu = \mu$,

$$h'(\mu) = \frac{1}{\sqrt{\mu}}$$

gives

$$h(\mu) = 2\sqrt{\mu} + \text{any constant}$$

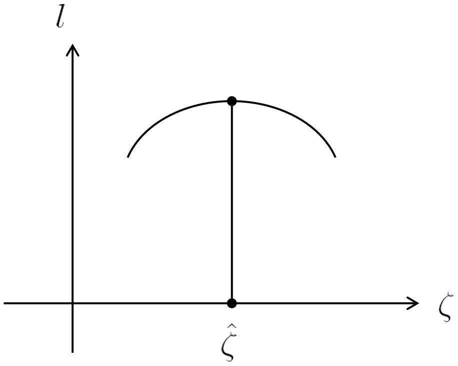
as above.

"Normal likelihood" means that the transformation $\hat{\zeta} = h(\hat{\mu})$ results in

$$\left.\frac{\partial^3 l_\eta(y)}{\partial \zeta^3}\right|_{\hat{\zeta}} = 0.$$

This makes the log likelihood look parabolic near its maximum at $\zeta = \hat{\zeta}$. For the Poisson the transformation is $h' = V^{-2/3} = \mu^{-2/3}$, or

$$h(\mu) = 3\mu^{1/3} + \text{constant}.$$



"Normal density" means that $\hat{\zeta} = h(\hat{\mu}) \dot{\sim} \mathcal{N}(0,1)$. For the Poisson $h' = \mu^{-1/3}$ or

$$h(\mu) = \frac{3}{2}\,\mu^{2/3} + \text{constant} \qquad (\text{makes skewness } \hat{\zeta} \doteq 0).$$

One sees all three transformations $2\mu^{1/2}$, $3\mu^{1/3}$, and $^3/_2\mu^{2/3}$ referred to as "the" transformation for the Poisson.

Homework 1.35. Numerically compare the three transformations for the Poisson for $n = 5, 10, 15,$ 20, and 25.

Our transformation results apply to any sample size n, with $V_\mu^{(n)} = V_\mu/n$. Verification of the normal density case relies on *Cornish–Fisher expansions*, which won't be presented here.

Homework 1.36. We observe independent χ^2 variables

$$\hat{\sigma}_i^2 \sim \sigma_i^2 \chi_{\nu_i}^2/\nu_i,$$

the ν_i being known degrees of freedom, and wish to regress $\hat{\sigma}_i^2$ versus some known covariates. Two frequently suggested transformations are $\log(\hat{\sigma}_i^2)$ and $(\hat{\sigma}_i^2)^{1/3}$, the latter being the "Wilson–Hilferty" transformation. Discuss the two transformations in terms of the previous results table.

Exponential Families in Theory and Practice

Bradley Efron
Stanford University

Part 2

Multiparameter Exponential Families

2.1 *Natural parameters, sufficient statistics, cgf* (pp 35–36) Natural parameter space A; carriers

2.2 *Expectation and covariance* (pp 36–37) $\mu = \dot{\psi}$, $V = \ddot{\psi}$; relationship of μ and η; expectation space B

2.3 *Review of transformations* (pp 37–38) Scalar function $h(\eta) = H(\mu)$; $H'(\mu) = D\dot{h}(\eta)$; $H''(\hat{\mu}) = D\ddot{h}(\hat{\eta})D'$

2.4 *Repeated sampling* (pp 38–39) $\bar{y} = \sum y_i/n$; $g_\eta^{(n)}(\bar{y}) = e^{n[\eta'\bar{y}-\psi(\eta)]}g_0^{(n)}(\bar{y})$; $A^{(n)} = nA$, $B^{(n)} = B$

2.5 *Likelihoods, score functions, Cramér–Rao lower bounds* (pp 39–41) Fisher information i_η; Fisher information for functions of η or μ; when can CRLB be attained?

2.6 *Maximum likelihood estimation* (pp 41–45) $\hat{\mu} = \bar{y}$; mapping to $\hat{\eta}$ one-parameter subfamilies; $\dddot{\psi}$ and $\ddddot{\psi}$; Stein's least favorable family

2.7 *Deviance* (pp 45–45) Hoeffding's formula; relationship with Fisher information

2.8 *Examples of multiparameter exponential families* (pp 45–53) Beta; Dirichlet; univariate normal; multivariate normal; graph models; truncated data; conditional families; 2 × 2 tables; Gamma/Dirichlet; Wishart; the Poisson trick

2.9 *The multinomial as exponential family* (pp 53–56) Categorical data, $\pi_l = \Pr\{\text{category } l\}$; count vector s_l, proportion $p_l = s_l/n$; symmetric parameterization

2.1 Natural parameters, sufficient statistics, cgf

One-parameter families are too restrictive for most real data analysis problems. It is easy, however, to extend exponential families to multiparametric situations. A *p-parameter exponential family* is a collection of probability densities

$$\mathcal{G} = \{g_\eta(y), \eta \in A\}$$

of the form

$$g_\eta(y) = e^{\eta' y - \psi(\eta)} g_0(y). \tag{2.1}$$

- η is the $p \times 1$ *natural*, or *canonical*, parameter vector.
- y is the $p \times 1$ vector of *sufficient statistics*, range space $y \in \mathcal{Y} \subset \mathcal{R}^p$.
- $\psi(\eta)$ is the *carrying density*, defined with respect to some *carrying measure* $m(dy)$ on $\mathcal{Y}$.
- A is the *natural parameter space*: all η having $\int_{\mathcal{Y}} e^{\eta' y} g_0(y) m(dy) < \infty$.

For any point η_0 in A we can express $\mathcal{G}$ as

$$g_\eta(y) = e^{(\eta-\eta_0)'y - [\psi(\eta)-\psi(\eta_0)]} g_{\eta_0}(y).$$

$\mathcal{G}$ consists of exponential tilts of g_{η_0}. The log tilting functions are linear in the p sufficient statistics $y = (y(1), y(2), \ldots, y(p))'$.

Homework 2.1. Show that $x_1, x_2, \ldots, x_n \overset{\text{iid}}{\sim} \mathcal{N}(\lambda, \Gamma)$ can be written in form (2.1) with $y = (\bar{x}, \bar{x^2})$.

In most applications, y is a function of a complete data set $\boldsymbol{x}$, usually much more complicated than a p-vector. Then y earns the name "sufficient vector". The mapping $y = t(\boldsymbol{x})$ from the full data to the sufficient vector is crucial to the statistical analysis. It says which parts of the problem are important, and which can be ignored.

2.2 Expectation and covariance

The *expectation vector* $\mu = E_\eta\{y\}$ is given by

$$\mu = E_\eta\{y\} = \underset{p\times 1}{\dot{\psi}}(\eta) = \begin{pmatrix} \vdots \\ \partial\psi(\eta)/\partial\eta_i \\ \vdots \end{pmatrix}$$

while the *covariance matrix* equals the second derivative matrix of ψ,

$$V = \mathrm{Cov}_\eta\{y\} = \underset{p\times p}{\ddot{\psi}}(\eta) = \begin{pmatrix} \vdots \\ \partial^2\psi(\eta)/\partial^2\eta_i\partial\eta_j \\ \vdots \end{pmatrix}.$$

Both μ and V are functions of η, usually suppressed in our notation.

Relationship of μ and η

Notice that

$$\underset{p\times p}{d\mu/d\eta} = (\partial\mu_i/\partial\eta_j) = \ddot{\psi} = V,$$

so

$$\partial\eta/\partial\mu = (\partial\eta_j/\partial\mu_i) = V^{-1}.$$

Here we are assuming that the carrying density $g_0(y)$ in (2.1) is of full rank in the sense that it is not entirely supported on any lower-dimensional subspace of $\mathcal{R}^p$, in which case V will be positive definite for all choices of η.

The vector μ is a 1:1 function of η, usually nonlinear, the local equations of transformation being

$$d\mu = V d\eta \quad \text{and} \quad d\eta = V^{-1} d\mu$$

(remembering that V changes with η or μ). The set A of all η vectors for which $\int_{\mathcal{Y}} e^{\eta' y} g_0(y)$ is finite is convex, while the set B of all μ vectors is connected but not necessarily convex (though counterexamples are hard to construct),

$$B = \{\mu = E_\eta\{y\}, \eta \in A\}.$$

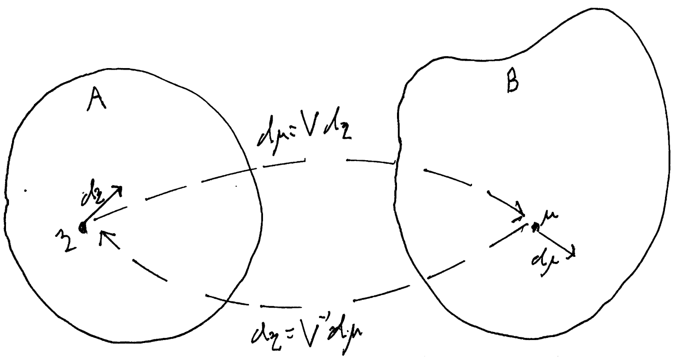


Figure 2.1

Note. $d\eta' d\mu = d\eta' V d\eta > 0$, so the angle between $d\eta$ and $d\mu$ is less than 90°; in this sense μ is an increasing function of η, and vice versa.

Homework 2.2. (a) Prove that A is convex. (b) If $\mathcal{Y}$ is the sample space of y, show that $B \subseteq$ convex hull of $\mathcal{Y}$. (c) Construct a one-parameter exponential family where the closure $\bar{B}$ is a proper subset of $\mathcal{Y}$.

Reference Efron (1978), "Geometry of exponential families", *Ann. Statist.*, Section 2, provides an example of non-convex B.

2.3 Review of transformations

This review applies generally, though we'll be interested in (η, μ) as previously defined. Suppose η and μ are vectors in $\mathcal{R}^p$, smoothly related to each other,

$$\eta \overset{1:1}{\longleftrightarrow} \mu,$$

and that $h(\eta) = H(\mu)$ is some smooth real-valued function. Let D be the $p \times p$ derivative matrix (Hessian)

$$D = \begin{pmatrix} & \vdots & \\ \cdots & \partial\eta_j/\partial\mu_i & \cdots \\ & \vdots & \end{pmatrix}, \qquad i \downarrow \quad \text{and} \quad j \rightarrow .$$

($D = V^{-1}$ in our case.) Letting $\cdot$ indicate derivatives with respect to η, and $'$ indicate derivatives with respect to μ, we have

$$H'(\mu) = D\dot{h}(\eta)$$

where $\dot{h}(\eta) = (\cdots \partial h/\partial\eta_j \cdots)^\top$ and $H'(\mu) = (\cdots \partial H/\partial\mu_i \cdots)^\top$ (here using $^\top$ to indicate transposition, as will be done whenever $'$ is used for differentiation).

The second derivative matrices of $h(\eta)$ and $H(\mu)$ are related by

$$\underset{(\partial^2 H/\partial\mu_i\partial\mu_j)}{\underset{\uparrow}{H''(\mu)}} \quad = \quad \underset{(\partial^2 h/\partial\eta_i\partial\eta_j)}{\underset{\uparrow}{D\ddot{h}(\eta)D^\top}} \quad + \quad D_2\dot{h}(\eta)$$

Here D_2 is the $p \times p \times p$ three-way array $(\partial^2\eta_j/\partial\mu_i\partial\mu_j)$.
Important: At a point where $\dot{h}(\eta) = 0$, $H''(\mu) = D\ddot{h}(\eta)D'$.

Homework 2.3. Prove the important statement above.

2.4 Repeated sampling

Suppose we observe repeated samples from (2.1),

$$\boldsymbol{y} = (y_1, y_2, \dots, y_n) \overset{\text{iid}}{\sim} g_\eta(y).$$

Let $\bar{y}$ denote the average of the vectors y_i,

$$\bar{y} = \sum_{i=1}^n y_i/n.$$

Then

$$g_\eta^{\boldsymbol{Y}}(\boldsymbol{y}) = e^{n[\eta'\bar{y} - \psi(\eta)]} g_0^{\boldsymbol{Y}}(\boldsymbol{y}), \tag{2.2}$$

as in (1.2), with $g_0^{\boldsymbol{Y}}(\boldsymbol{y}) = \prod_{i=1}^n g_0(y_i)$. This is a p-dimensional exponential family (2.1), with:

- natural parameter $\eta^{(n)} = n\eta$
- sufficient vector $y^{(n)} = \bar{y}$
- expectation vector $\mu^{(n)} = \mu$

- variance matrix $V^{(n)} = V/n$ [since $\text{Cov}(\bar{y}) = \text{Cov}(y)/n$]
- cgf $\psi^{(n)}(\eta^{(n)}) = n\psi(\eta^{(n)}/n)$
- sample space = product space $\mathcal{Y}_1 \otimes \mathcal{Y}_2 \otimes \cdots \otimes \mathcal{Y}_n$

Since $\bar{y}$ is sufficient, we can consider it on its own sample space, say $\mathcal{Y}^{(n)}$, with densities

$$g_\eta^{(n)}(\bar{y}) = e^{n[\eta'\bar{y}-\psi(\eta)]} g_0^{(n)}(\bar{y}), \tag{2.3}$$

where $g_0^{(n)}(\bar{y})$ is the density of $\bar{y}$ for $\eta = 0$.

From $\eta^{(n)} = n\eta$ and $\mu^{(n)} = \mu$ we see that

$$A^{(n)} = nA \quad \text{and} \quad B^{(n)} = B.$$

In what follows, we will parameterize family (2.3) with η rather than $\eta^{(n)} = n\eta$. Then we can use Figure 2.1 relating A and B *exactly as drawn.*

Homework 2.4. Is $d\mu^{(n)} = V^{(n)} d\eta^{(n)}$ the same as $d\mu = V d\eta$?

What *does* change is the distance of the sufficient statistic $\bar{y}$ from its expectation μ: since $V^{(n)} = V/n$, the "error" $\bar{y} - \mu$ decreases at order $O_p(1/\sqrt{n})$. As $n \to \infty$, $\bar{y} \to \mu$. This makes asymptotics easy to deal with in exponential families.

2.5 Likelihoods, score functions, Cramér–Rao lower bounds

From (2.2), the log likelihood function of $\boldsymbol{y} = (y_1, \ldots, y_n)$ is

$$l_\eta(\boldsymbol{y}) = n\left[\bar{y} - \psi(\eta)\right].$$

The *score function* is defined to be the component-wise derivative with respect to η,

$$\dot{l}_\eta(\boldsymbol{y}) = (\partial l_\eta(\boldsymbol{y})/\partial\eta_j) = n(\bar{y} - \mu)$$

(remembering that $\dot{\psi}(\eta) = \mu$), and the second derivative matrix is

$$\ddot{l}_\eta(\boldsymbol{y}) = -nV$$

(since $\dot{\mu} = \ddot{\psi} = V$). The *Fisher information* for η in $\boldsymbol{y}$ is the outer product

$$i_\eta^{(n)} = E_\eta\left\{\dot{l}_\eta(\boldsymbol{y})\dot{l}_\eta(\boldsymbol{y})'\right\} = nV,$$

(also equaling $E\{-\ddot{l}_\eta(\boldsymbol{y})\}$).

We can also consider the score function with respect to μ,

$$\begin{aligned}\frac{\partial l_\eta(\boldsymbol{y})}{\partial \mu} &= \frac{\partial \eta}{\partial \mu}\frac{\partial l_\eta(\boldsymbol{y})}{\partial \eta} = V^{-1}\dot{l}_\eta(\boldsymbol{y})\\ &= nV^{-1}(\bar{y}-\mu);\end{aligned}$$

the Fisher information for μ, denoted $i_\eta^{(n)}(\mu)$ ("at η, for μ, sample size n") is

$$\begin{aligned}i_\eta^{(n)}(\mu) &= E\left\{\frac{\partial l_\eta(\boldsymbol{y})}{\partial \mu}\frac{\partial l_\eta(\boldsymbol{y})'}{\partial \mu}\right\} = n^2V^{-1}\operatorname{Cov}(\bar{y})V^{-1}\\ &= nV^{-1}.\end{aligned}$$

Cramér–Rao lower bound (CRLB)

Suppose we have a multiparameter family $f_\alpha(z)$, score vector $\dot{l}_\alpha(z) = (\partial \log f_\alpha(z)/\partial \eta_j)$, and Fisher information i_α the expected outer product $E\{\dot{l}_\alpha(z)\dot{l}_\alpha(z)'\}$. Then the CRLB for α is i_α^{-1}: if $\bar{\alpha}$ is any unbiased estimate of vector α,

$$\operatorname{Cov}\{\bar{\alpha}\} \geq i_\alpha^{-1}.$$

(This is equivalent to $\operatorname{Var}\{c'\bar{\alpha}\} \geq c' i_\alpha^{-1} c$ for estimating any linear combination $c'\alpha$.)

The CRLB for μ is seen to be

$$\operatorname{CRLB}(\mu) = i_\eta^{(n)}(\mu)^{-1} = \frac{V}{n}.$$

But $\operatorname{Cov}(\bar{y}) = V/n$, so the MLE $\hat{\mu} = \bar{y}$ attains the CRLB. This only happens for μ, or linear transformations of μ. So for example, the MLE $\hat{\eta}$ does not attain

$$\operatorname{CRLB}(\eta) = \frac{V^{-1}}{n}.$$

Homework 2.5. Let ζ be a scalar function of η or μ, say

$$\zeta = t(\eta) = s(\mu),$$

with gradient vector $\dot{t}(\eta) = (\cdots \partial t/\partial \eta_j \cdots)$, and likewise $s'(\mu) = (\cdots \partial s/\partial \mu_j \cdots)$. Having observed $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$, show that the lower bound on the variance of an unbiased estimate of ζ is

$$\operatorname{CRLB}(\zeta) = \frac{\dot{t}(\eta)^\top V^{-1}\dot{t}(\eta)}{n} = s'(\mu)^\top V s'(\mu).$$

Note. In applications, $\hat{\eta}$ is substituted for η, or $\hat{\mu}$ for μ, to get an approximate variance for the MLE $\hat{\zeta} = t(\hat{\eta}) = s(\hat{\mu})$. Even though ζ is not generally unbiased for ζ, the variance approximation — which is equivalent to using the delta method — is usually pretty good.

2.6 Maximum likelihood estimation

From the score function for η, $\dot{l}_\eta(\boldsymbol{y}) = n(\bar{y} - \mu)$, we see that the maximum likelihood estimate $\hat{\eta}$ for η must satisfy

$$\mu_{\hat{\eta}} = \bar{y}.$$

That is, $\hat{\eta}$ is the value of η that makes the theoretical expectation μ equal the observed value $\bar{y}$. Moreover, $\ddot{l}_\eta(\boldsymbol{y}) = -nV$ shows that the log likelihood $l_\eta(\boldsymbol{y})$ is a *concave* function of η (since V is positive definite for all η) so there are no other local maxima.

From ${}^{\partial}\!/\!{}_{\partial\mu}\, l_\eta(\boldsymbol{y}) = nV^{-1}(\bar{y} - \mu)$, we see that the MLE of μ is $\hat{\mu} = \bar{y}$. This also follows from $\hat{\mu} = \mu(\hat{\eta})$ (that is, that MLE's map correctly) and $\mu_{\hat{\eta}} = \bar{y}$.

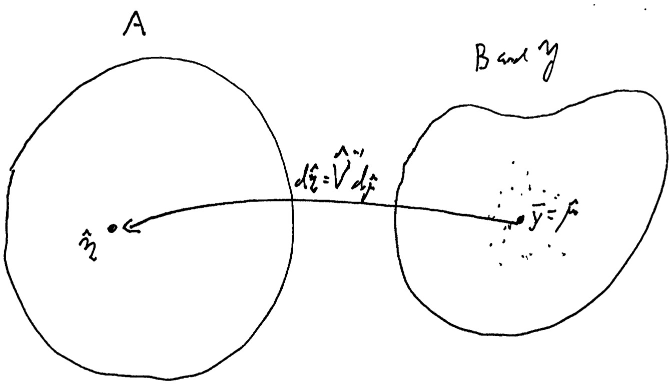


Figure 2.2

In Figure 2.2, the expectation space B and the sample space $\mathcal{Y}$ for individual y_i's are plotted over each other. The scattered dots represent the y_i's, with their average $\bar{y} = \hat{\mu}$ at the center. The nonlinear mapping $\hat{\eta} = \eta(\hat{\mu})$ has the local linear expression $d\hat{\eta} = \hat{V}^{-1} d\hat{\mu}$, where $\hat{V} = V_{\hat{\eta}}$, the variance matrix at the MLE point.

Homework 2.6. Use Figure 2.2 to get an approximation for $\mathrm{Cov}(\hat{\eta})$. How does it relate to $\mathrm{CRLB}(\hat{\eta})$?

Homework 2.7. Show that the $p \times p$ second derivative matrix with respect to μ is

$$\left.\frac{\partial^2 l_\eta(\boldsymbol{y})}{\partial \mu^2}\right|_{\hat{\mu}} = -n\hat{V}^{-1}.$$

From the central limit theorem,

$$\hat{\mu} = \bar{y} \mathrel{\dot{\sim}} \mathcal{N}_p\left(\mu, {}^{V}\!/\!{}_{n}\right),$$

where the normality is approximate but the expectation and covariance are exact. Then $d\eta =$

$V^{-1}d\mu$ suggests

$$\hat{\eta} \dot{\sim} \mathcal{N}_p\left(\eta, {}^{V^{-1}}\!/_{n}\right),$$

but now everything is approximate.

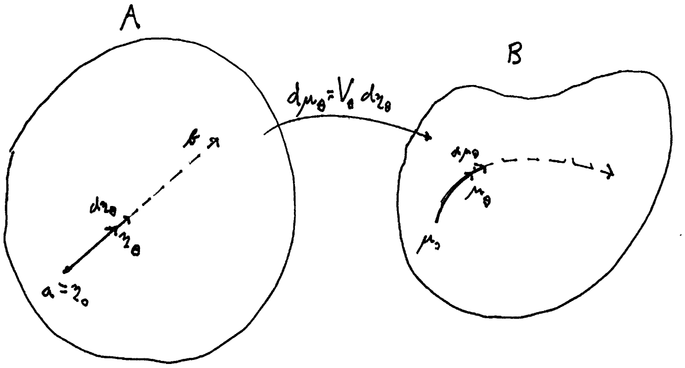


Figure 2.3

One-parameter subfamilies

Multiparameter exponential family calculations can sometimes be reduced to the one-parameter case by considering subfamilies of $g_\eta(y)$,

$$\eta_\theta = a + b\theta,$$

a and b fixed vectors in $\mathcal{R}^p$. This defines densities

$$f_\theta(y) = g_{\eta_\theta}(y) = e^{(a+b\theta)'y - \psi(a+b\theta)} g_0(y).$$

This is a one-parameter exponential family

$$f_\theta(y) = e^{\theta x - \phi(\theta)} f_0(y)$$

with

- natural parameter θ;
- sufficient statistic $x = b'y$;
- $\phi(\theta) = \psi(a + b\theta)$;
- $f_0(y) = e^{a'y} g_0(y)$.

For simple notation we write μ_θ for μ_{η_θ}, V_θ for V_{η_θ}, etc. As θ increases, η_θ moves in a straight line parallel to b, but μ_θ will usually follow a curve, the differential relationship being

$$d\mu_\theta = V_\theta d\eta_\theta.$$

Higher-order calculations are sometimes simplified by considering one-parameter subfamilies. For example we have

$$\begin{aligned}
d\phi/d\theta &= E_\theta\{x\} = b'\mu_\theta = b'\dot{\psi}(\eta_\theta) \qquad \left[\dot{\psi}(\eta_\theta) \equiv \dot{\psi}(\eta)\Big|_{\eta=\eta_\theta}\right]\\
d^2\phi/d\theta^2 &= \mathrm{Var}_\theta\{x\} = b'V_\theta b = b'\ddot{\psi}(\eta_\theta)b\\
d^3\phi/d\theta^3 &= E_\theta\left\{x - E_\theta(x)\right\}^3\\
&= \sum_i\sum_j\sum_k \dddot{\psi}_{ijk}(\eta_\theta)b_ib_jb_k \equiv \underset{p\times p\times p}{\dddot{\psi}(\eta_\theta)} \cdot b^3.
\end{aligned}$$

This last implies that

$$\dddot{\psi}_{ijk} = E_\eta(y_i - \mu_i)(y_j - \mu_j)(y_k - \mu_k).$$

Homework 2.8. Verify the result above. What is $\ddddot{\psi}_{ijkl}$?

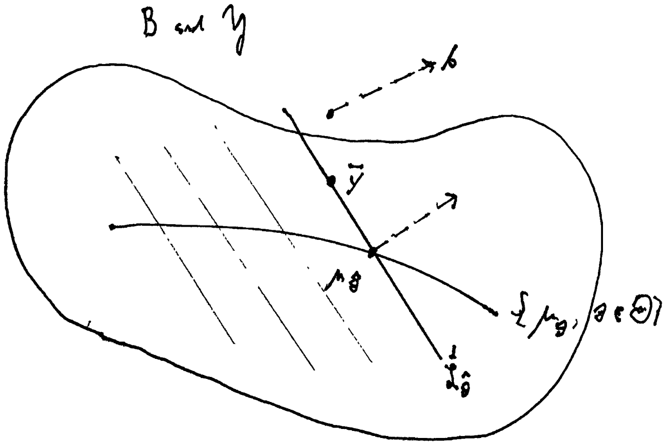


Figure 2.4

The subfamily $\mathcal{F} = \{f_\theta(y), \theta \in \Theta\}$ is defined for those values $\theta \in \Theta$ that keep η_θ within A, so

$$\mathcal{F} \subset \mathcal{G} = \{g_\eta(y), \eta \in A\}\,.$$

Suppose now that $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ is an iid sample from some member f_θ of $\mathcal{F}$.

Homework 2.9. (a) Show that the MLE $\hat{\theta}$ has $\mu_{\hat{\theta}}$ obtained by projection orthogonal to b, from $\bar{y}$ to $\{\mu_\theta, \theta \in \Theta\}$ as shown in Figure 2.4. (b) How would you estimate the standard deviation of $\hat{\theta}$?

Note. The set of $\bar{y}$ vectors having $\hat{\theta}$ as MLE is a $(p-1)$-dimensional hyperplane passing through $\mu_{\hat{\theta}}$ orthogonal to b. The hyperplanes are parallel for all values of $\hat{\theta}$.

Stein's least favorable family (another use of one-parameter subfamilies)

In a p-parameter exponential family $\mathcal{G} = \{g_\eta(y), \eta \in A\}$, we wish to estimate a real-valued parameter ζ which can be evaluated as

$$\zeta = t(\eta) = s(\mu).$$

Let the true value of η be η_0 and $\mu_0 = \dot{\psi}(\eta_0)$ the true value of μ. Then the true value of ζ is $\zeta_0 = t(\eta_0) = s(\mu_0)$.

Gradients Let $\dot{t}_0$ be the gradient of $t(\eta)$ at $\eta = \eta_0$, and likewise s_0' for the gradient of $s(\mu)$ at $\mu = \mu_0$,

$$\dot{t}(\eta) = \begin{pmatrix} \vdots \\ \partial t(\eta)/\partial \eta_j \\ \vdots \end{pmatrix}_{\eta_0} \quad \text{and} \quad s'(\mu) = \begin{pmatrix} \vdots \\ \partial s(\mu)/\partial \mu_j \\ \vdots \end{pmatrix}_{\mu_0},$$

both $\dot{t}_0$ and s_0' being p-vectors. As shown in Figure 2.5, $\dot{t}_0$ is orthogonal to the $(p-1)$-dimensional level surface of η vectors that give $t(\eta) = \zeta_0$, and s_0' is orthogonal to the level surface $s(\mu) = \zeta_0$.

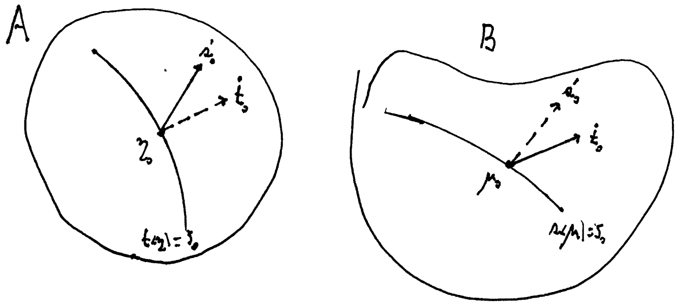


Figure 2.5

The *least favorable family* (LFF) is defined to be the one-parameter subfamily having

$$\eta_\theta = \eta_0 + s_0'\theta$$

(*not* $\eta_0 + \dot{t}_0\theta$) for θ in an open interval containing 0.

Homework 2.10. Show that the CRLB for estimating ζ in the LFF, evaluated at $\theta = 0$, is the same as the CRLB for estimating ζ in $\mathcal{G}$, evaluated at $\eta = \eta_0$; and that both equal $(s_0')^\top V_{\eta_0} s_0'$.

In other words, the reduction to the LFF does not make it any easier to estimate η. LFF operations are used in theoretical calculations where it is desired to extend a one-dimensional result to higher dimensions.

Extra Credit Any choice other than $b = s_0'$ in the family $\eta_\theta = \eta_0 + b\theta$ makes the one-parameter CRLB smaller.

2.7 Deviance

Everything said in Section 1.8 holds for deviance in multiparameter families:

$$\begin{aligned} D(\eta_1, \eta_2) &= 2E_{\eta_1}\left\{\log \frac{g_{\eta_1}(y)}{g_{\eta_2}(y)}\right\} \\ &= 2\left[(\eta_1 - \eta_2)'\mu_1 - \psi(\eta_1, \eta_2)\right] \geq 0. \end{aligned}$$

(Also denoted $D(\mu_1, \mu_2)$ or just $D(1, 2)$.) Under iid sampling, $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$,

$$D_n(\eta_1, \eta_2) = nD(\eta_1, \eta_2).$$

Usually, $D(\eta_1, \eta_2) \neq D(\eta_2, \eta_1)$.

Hoeffding's formula

Now indexing with μ rather than η,

$$\frac{g_\mu(\boldsymbol{y})}{g_{\hat{\mu}}(\boldsymbol{y})} = e^{-nD(\hat{\mu}, \mu)/2} = e^{-nD(\bar{y}, \mu)/2}.$$

Relationship with Fisher information

$$D(\eta_1, \eta_2) = (\eta_2 - \eta_1)'V_{\eta_1}(\eta_2 - \eta_1) + O(\eta_2 - \eta_1)^3$$

(remembering that $i_{\eta_1}^{(1)} = V_{\eta_1}$).

The tangency picture on page 26 in Part 1 remains valid: now $\psi(\eta)$ is a convex surface over the convex set A in $\mathcal{R}^p$. The tangent line on page 26 is now the tangent hyperplane to the surface, passing through $(\eta_1, \psi(\eta_1))$ in $\mathcal{R}^{p+1}$; $D(\eta_1, \eta_2)$ is the distance from $(\eta_2, \psi(\eta_2))$ projected down to the tangent hyperplane.

Homework 2.11. Prove the previous statement.

EXTRA CREDIT Draw a schematic picture analogous to the illustration at the top of page 26.

Homework 2.12. Show that

$$(\eta_2 - \eta_1)'(\mu_2 - \mu_1) = 1/2\left[D(1, 2) + D(2, 1)\right].$$

Since the right-hand side is positive, this proves that the relationship between η and μ is "globally monotonic": the angle between $(\eta_2 - \eta_1)$ and $(\mu_2 - \mu_1)$ is always less than 90°.

2.8 Examples of multiparameter exponential families

There is a short list of named multiparameter exponential families that show up frequently in applications. Later we will consider some important examples that don't have familiar names.

Beta

- X is univariate with density on $[0,1]$,

$$\frac{x^{\alpha_1-1}(1-x)^{\alpha_2-1}}{\text{be}(\alpha_1,\alpha_2)} \qquad \left[\text{be}(\alpha_1,\alpha_2)=\frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}\right],$$

α_1 and α_2 positive. The density is defined with respect to Lebesgue measure on $[0,1]$. This is written in exponential family form as

$$g_\alpha(x)=e^{\alpha' y-\psi(\alpha)}g_0(x)\begin{cases} y=[\log(x),\log(1-x)]' \text{ and } \mathcal{Y}=\mathcal{R}^2\\ \psi=\log\left[\text{be}(\alpha_1,\alpha_2)\right]\\ g_0(x)=1/[x(1-x)]. \end{cases}$$

Homework 2.13. (a) Verify directly that x has

$$\text{mean}\quad\frac{\alpha_1}{\alpha_1+\alpha_2}\qquad\text{and}\qquad\text{variance}\quad\frac{\alpha_1\alpha_2}{(\alpha_1+\alpha_2)^2(\alpha_1+\alpha_2+1)}.$$

(b) Find an expression for $E\{\log(x)\}$ in terms of the digamma function, digamma $(z)=\Gamma'(z)/\Gamma(z)$.

Dirichlet

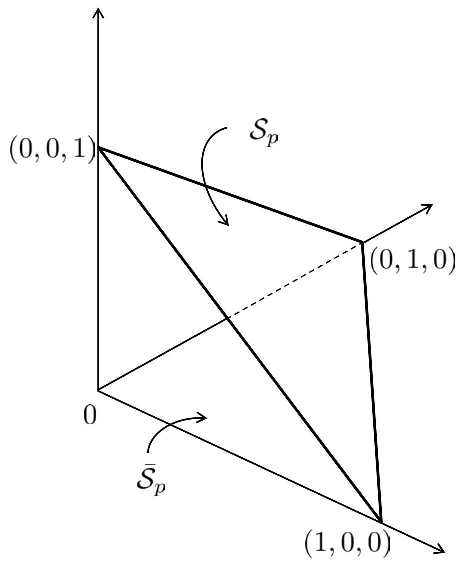
- Let $\mathcal{S}_p$ indicate the p-dimensional simplex

$$\mathcal{S}_p=\left\{x=(x_1,x_2,\dots,x_p): x_i\geq 0 \text{ and } \sum_1^p x_i=1\right\},$$

and $\bar{\mathcal{S}}_p$ the projected simplex

$$\bar{\mathcal{S}}_p=\left\{x: x_p=0, x_i\geq 0, \text{ and } \sum_1^{p-1} x_i\leq 1\right\}.$$

- *An almost obvious fact,* $\bar{\mathcal{S}}_p$ has $(p-1)$-dimensional "area" of $1/(p-1)!$ (the case $p=3$ *is* obvious).



Homework 2.14. Argue that $\mathcal{S}_p$ has $(p-1)$-dimensional area $\sqrt{p}/(p-1)!$.

- The Dirichlet is Beta's big brother for $p>2$. Let

$$\alpha=(\alpha_1,\alpha_2,\dots,\alpha_p)\qquad\text{with all } \alpha_i>0.$$

The corresponding Dirichlet density is

$$g_\alpha(x) = \prod_{i=1}^{p} x_i^{\alpha_i - 1} \Big/ \operatorname{di}(\alpha) \qquad \left[\operatorname{di}(\alpha) = \prod_{1}^{p} \Gamma(\alpha_i) \Big/ \Gamma\left(\sum \alpha_i\right)\right]$$

with respect to uniform measure on $\bar{\mathcal{S}}_p$, or uniform/$\sqrt{p}$ measure on $\mathcal{S}_p$. (Since $x_p = 1 - \sum_1^{p-1} x_i$, $g_\alpha(x)$ is actually $(p-1)$-dimensional.)

- $g_\alpha(x)$ can be written in exponential family form as

$$g_\alpha(x) = e^{\alpha' y - \psi(\alpha)} m(dx),$$

where

- α is the natural parameter vector ("η");
- $y = \log(x) = [\log(x_1), \log(x_2), \ldots, \log(x_p)]$;
- $\psi(\alpha) = \log[\operatorname{di}(\alpha)]$; $m(dx) =$ uniform/$\sqrt{p}$ measure on $\mathcal{S}_p$.

Homework 2.15. What are μ and V for the Dirichlet? Compare with Homework 2.13. What is rank(V)?

Univariate normal

- $X \sim \mathcal{N}(\lambda, \Gamma)$, univariate, with

$$g(x) = \frac{1}{\sqrt{2\pi\Gamma}} \, e^{-\frac{1}{2\Gamma}(x-\lambda)^2} = \frac{1}{\sqrt{2\pi\Gamma}} \, e^{\left\{-\frac{x^2}{2\Gamma} + \frac{\lambda}{\Gamma}x - \frac{\lambda^2}{2\Gamma}\right\}}.$$

- In exponential family form,

$$g_\eta(x) = e^{\eta_1 y_1 + \eta_2 y_2 - \psi(\eta)} g_0(x),$$

with

$$\eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \lambda/\Gamma \\ -1/2\Gamma \end{pmatrix}, \qquad \mu = \begin{pmatrix} \lambda \\ \lambda^2 + \Gamma \end{pmatrix};$$

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} x \\ x^2 \end{pmatrix}, \qquad \psi = \frac{1}{2}\left(\frac{\lambda^2}{\Gamma} + \log\Gamma\right);$$

$$g_0(x) = \frac{1}{\sqrt{2\pi}} \quad \text{with respect to uniform measure on } (-\infty, \infty).$$

Homework 2.16. Use $\dot{\psi}$ and $\ddot{\psi}$ to derive μ and V.

It seems like we have lost ground: our original univariate statistic x is now represented two-dimensionally by $y = (x, x^2)$. However, if we have an iid sample $x_1, x_2, \dots, x_n \overset{\text{iid}}{\sim} \mathcal{N}(\lambda, \Gamma)$, then

$$\bar{y} = \begin{pmatrix} \sum x_i/n \\ \sum x_i^2/n \end{pmatrix}$$

is still a two-dimensional sufficient statistic (though not the more usual form $(\bar{x}, \hat{\sigma}^2)$). Figure 2.6 presents diagrams of A and of B and $\mathcal{Y}$.

Homework 2.17. An iid sample $x_1, x_2, \dots, x_n \sim \mathcal{N}(\lambda, \Gamma)$ gives $y_1, y_2, \dots, y_n$, with $y_i = (x_i, x_i^2)$. Draw a schematic diagram of B and $\mathcal{Y}$ indicating the points y_i and the sufficient vector $\bar{y}$.

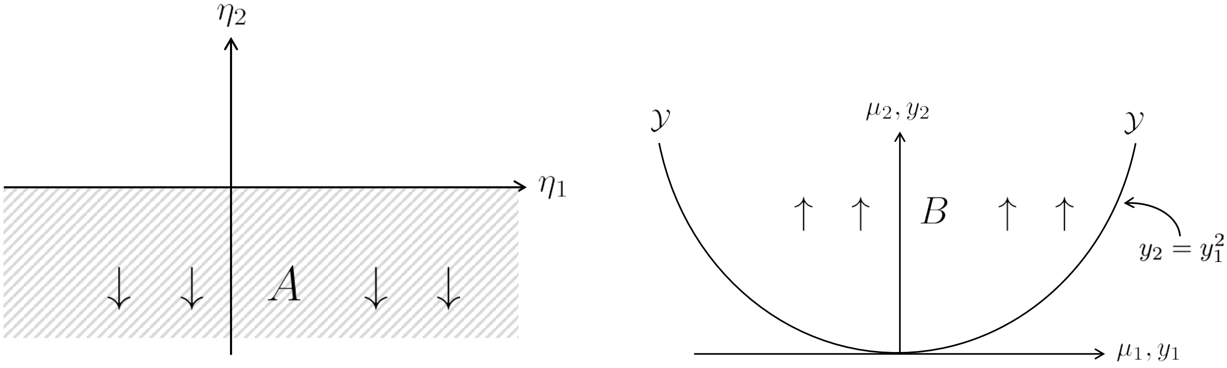


Figure 2.6

Multivariate normal

- Another big brother case: we observe n independent observations from a d-dimensional normal distribution with expectation vector λ and $d \times d$ covariance matrix Γ,

$$x_1, x_2, \dots, x_n \overset{\text{iid}}{\sim} \mathcal{N}_d(\lambda, \Gamma).$$

  (This is the generic beginning point for classical multivariate analysis.)

- We need some special notation: for H a $d \times d$ symmetric matrix, let $h = H^{(v)}$ be the $d(d+1)/2$ vector that strings out the on-or-above diagonal elements,

$$h' = (H_{11}, H_{12}, \dots, H_{1d}, H_{22}, H_{23}, \dots, H_{2d}, H_{31}, \dots, H_{dd}),$$

  and let $h^{(m)}$ be the inverse mapping from h back to H.

- Also let $\operatorname{diag}(H) =$ matrix with on-diagonal elements those of H and off-diagonal elements 0.

- Straightforward but tedious (and easily bungled) calculations show that the density of $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ forms a

$$p = d(d+3)/2$$

-dimensional exponential family

$$g_\eta(\boldsymbol{x}) = e^{n[\eta' y - \psi(\eta)]} g_0(\boldsymbol{x})$$

described as follows:

$$y' = (y1', y2') \quad \text{where} \quad y1 - \bar{x} \quad \text{and} \quad y2 = \left(\sum_{i=1}^n x_i x_i'/n \right)^{(v)},$$

dimensions d and $d(d+1)/2$, respectively;

$$\begin{aligned} &\eta' = (\eta1', \eta2') \quad \text{where} \quad \eta1 = n\Gamma^{-1}\lambda \quad \text{and} \quad \eta2 = n\left[\operatorname{diag}(\Gamma^{-1})/2 - \Gamma^{-1}\right]^{(v)}; \\ &\mu' = (\mu1', \mu2') \quad \text{where} \quad \mu1 = \lambda \quad \text{and} \quad \mu2 = (\lambda\lambda' + \Gamma)^{(v)}; \\ &\psi = \frac{1}{2}\left[\lambda'\Gamma^{-1}\lambda + \log(\Gamma)\right]. \end{aligned}$$

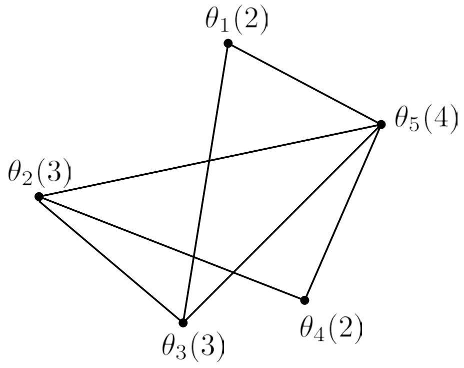
- We also have

$$\begin{aligned} \lambda &= \mu1 = \frac{1}{n}\Gamma \cdot \eta1; \\ \Gamma &= \mu2^{(m)} - \mu1\mu1' = -n\left[\operatorname{diag}(\eta2^{(m)}) + \eta2^{(m)}\right]^{-1}. \end{aligned}$$

Reference Efron and DiCiccio (1992), "More accurate confidence intervals in exponential families", *Biometrika*, Section 3.

Homework 2.18. Calculate the deviance $D[\mathcal{N}_p(\lambda_1, \Gamma_1), \mathcal{N}_p(\lambda_2, \Gamma_2)]$.

Graph models



Exponential families on graphs are now a booming topic in network theory. We have a graph with n nodes, each with its own parameter θ_i. Let

$$x_{ij} = \begin{cases} 1 & \text{if an edge between } \theta_1 \text{ and } \theta_2, \\ 0 & \text{if no edge between } \theta_1 \text{ and } \theta_2. \end{cases}$$

A simple model assumes that the x_{ij} are independent Bernoulli variates, with probability $x_{ij} = 1$,

$$\pi_{ij} = \frac{e^{\theta_i + \theta_j}}{1 + e^{\theta_i + \theta_j}}.$$

This is the "degree model".

Homework 2.19. (a) Show that the degree model is an n-parameter exponential family with sufficient vector $\boldsymbol{y}$,

$$y_i = \#\{\text{edges entering node } i\} \qquad (\text{“degree } i\text{”}).$$

(b) Describe η, μ, and V.

If all θ_i's are the same, the degree model reduces to a one-parameter exponential family, density $e^{\eta E-\psi(\eta)}$ with E the total number of edges. There are more complicated models, for instance

$$e^{\eta_1 E+\eta_2 T-\psi(\eta_1,\eta_2)},$$

where T is the total number of triangles. Very large graphs — the kinds of greatest interest these days — make it difficult to compute the normalizing factor $\psi(\eta_1, \eta_2)$, or to directly sample from the family–giving rise to Markov chain Monte Carlo (MCMC) techniques.

Truncated data

Suppose $y \sim g_\eta(y) = e^{\eta' y-\psi(\eta)} g_0(y)$, but we only get to observe y if it falls into a given subset $\mathcal{Y}_0$ of the sample space $\mathcal{Y}$. (This is the usual case in astronomy, where only sufficiently bright objects can be observed.) The conditional density of y given that it is observed is then

$$g_\eta(y \mid \mathcal{Y}_0) = e^{\eta' y-\psi(\eta)} g_0(y)/G_\eta(\mathcal{Y}_0)$$

where

$$G_\eta(\mathcal{Y}_0) = \int_{\mathcal{Y}_0} g_\eta(y) m(dy).$$

But this is still an exponential family,

$$g_\eta(y \mid \mathcal{Y}_0) = e^{\eta' y-\psi(\eta)-\log G_\eta(\mathcal{Y}_0)} g_0(y).$$

Homework 2.20. $x \sim \text{Poi}(\mu)$ but we only observe x if it is > 0. (a) Describe $g_\mu(x)$ in exponential family form. (b) Differentiate $\psi(\eta)$ to get μ and V.

Conditional families

Reference Lehmann and Romano (2005), *Testing Statistical Hypotheses*, 3rd edition, Section 2.7.

We have a p-parameter exponential family,

$$\mathcal{G} = \left\{ g_\eta(y) = e^{\eta' y-\psi(\eta)} dG_0(y), \eta \in A \right\},$$

where now we have represented the carrying density $g_0(y)m(dy)$ in Stiljes form “$dG_0(y)$”. We partition η and y into two parts, of dimensions p_1 and p_2,

$$\eta = (\eta_1, \eta_2) \quad \text{and} \quad y = (y_1, y_2).$$

Lemma 2. *The conditional distributions of y_1 given y_2 form a p_1-parameter exponential family, with densities*

$$g_{\eta_1}(y_1 \mid y_2) = e^{\eta_1' y_1 - \psi(\eta_1 \mid y_2)} dG_0(y_1 \mid y_2), \tag{2.4}$$

natural parameter vector η_1, sufficient statistic y_1, and cgf $\psi(\eta_1 \mid y_2)$ that depends on y_2 but not *on η_2. Here $dG_0(y_1 \mid y_2)$ represents the conditional distribution of y_1 given y_2 when $\eta_1 = 0$.*

Homework 2.21. Verify Lemma 2. (The fact that the carrier is $dG_0(y_1 \mid y_2)$ follows from a general probabilistic result, and may be assumed.)

The familiar uses of Lemma 2 often involve transformation of the original family $\mathcal{G}$: for M a $p \times p$ nonsingular matrix, let

$$\tilde{\eta} = M^{-1}\eta \quad \text{and} \quad \tilde{y} = My.$$

Since $\eta' \tilde{y} = \eta' y_1$, we see that the transformed densities $\tilde{g}_{\tilde{\eta}}(\tilde{y})$ also form an exponential family

$$\tilde{\mathcal{G}} = \left\{ \tilde{g}_{\tilde{\eta}}(\tilde{y}) = e^{\tilde{\eta}' \tilde{y} - \psi(M\tilde{\eta})} d\tilde{G}_0(\tilde{y}), \tilde{\eta} \in M^{-1}A \right\},$$

to which Lemma 2 can be applied. What follows are four useful examples.

Example 1 (2×2 tables). We start with

$$(x_1, x_2, x_3, x_4) \sim \text{Mult}_4\left[N, (\pi_1, \pi_2, \pi_3, \pi_4)\right]$$

as in Section 1.4. The conditional distribution of $x_1 \mid x_2, x_3, x_4$ is identical to $x_1 \mid r_1, c_1, N$. According to Lemma 2, x_1 conditionally follows a one-parameter exponential family, which turns out to be the "tilted hypergeometric" family (1.3) (applying to Fisher's exact test). The reason why $\eta_1 = \log(\pi_1\pi_4/\pi_2\pi_3)$, the log odds ratio, is connected to the exponential family representation of the multinomial, as discussed in the next section.

X_1 π_1	X_2 π_2	r_1
X_3 π_3	X_4 π_4	
	c_2	N

Homework 2.22. What is the matrix M being used here?

Example 2 (Gamma/Dirichlet). We have p independent gamma variates of possibly different degrees of freedom,

$$s_i \overset{\text{ind}}{\sim} G_{\nu_i}, \qquad i = 1, 2, \ldots, p,$$

so $\boldsymbol{s} = (s_1, s_2, \ldots, s_p)'$ follows a p-parameter exponential family. Let M have first row $(1, 1, \ldots, 1)'$, so that $\tilde{s} = Ms$ has first element $\tilde{s}_1 = \sum_1^p s_i = s_+$. Define

$$z = \boldsymbol{s}/s_+,$$

z taking its values in the simplex $\mathcal{S}_p$. Then

$$z \mid s_+ \sim \text{Dirichlet}(\boldsymbol{\nu}) \qquad [\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_p)],$$

a $(p-1)$-dimensional exponential family.

Example 3 (Wishart). Given a multivariate normal sample $x_1, x_2, \ldots, x_n \overset{\text{iid}}{\sim} \mathcal{N}_d(\lambda, \Gamma)$, let $y1 = \bar{x}$ and $y2 = \sum x_i x_i'/n$ as before. The conditional distribution of $y2 \mid y1$ is then a $p = d(d+1)/2$ exponential family. But the Wishart statistic

$$W = \sum (x_i - \bar{x})(x_i - \bar{x})'/n = \sum x_i x_i'/n - \bar{x}\bar{x}'$$

is, given $\bar{x}$, a function of $y2$. (In fact W is independent of $\bar{x}$.) This shows that W follows a $[d(d+1)/2]$-parameter exponential family.

Example 4 (The Poisson trick). Suppose that $\boldsymbol{s} = (s_1, s_2, \ldots, s_L)$ is a vector of L independent Poisson variates,

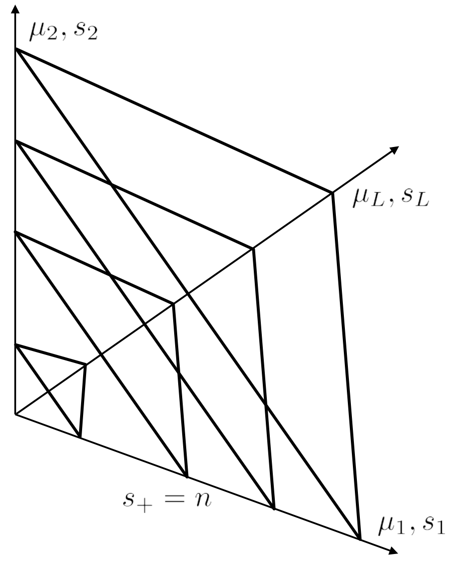
$$s_l \overset{\text{ind}}{\sim} \text{Poi}(\mu_l), \qquad l = 1, 2, \ldots, L.$$

Then the conditional distribution of $\boldsymbol{s} = (s_1, s_2, \ldots, s_L)$ — given that $\sum s_l$ equals n — is multinomial:

$$\boldsymbol{s} \mid n \sim \text{Mult}_L(n, \boldsymbol{\pi}), \tag{2.5}$$

the notation indicating an L category multinomial, n draws, true probabilities $\boldsymbol{\pi}$ where

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_L) \qquad \text{with } \pi_l = \mu_l \Big/ \sum_1^L \mu_i.$$



Homework 2.23. Directly verify (2.5).

Another way to say this: if

$$n \sim \text{Poi}(\mu_+) \quad \text{and} \quad \boldsymbol{s} \mid n \sim \text{Mult}_L(n, \boldsymbol{\pi}),$$

then

$$s_l \overset{\text{ind}}{\sim} \text{Poi}(\mu_+ \cdot \pi_l), \qquad l = 1, 2, \ldots, L. \tag{2.6}$$

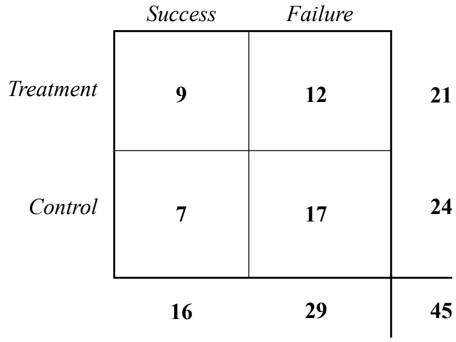
If you are having trouble with a multinomial calculation, usually because of multinomial correlation among the s_l, then thinking of sample size n as Poisson instead of fixed makes the components independent. This "Poisson trick" is often used for quick multinomial approximations.

Homework 2.24. For the 2×2 table of Example 1, or of Section 1.4, use the Poisson trick to find the delta-method approximation of the standard error for the log odds ratio, $\log(\pi_1\pi_4/\pi_2\pi_3)$.

2.9 The multinomial as exponential family

Traditional multivariate analysis focused on the multivariate normal distribution. More recently there has been increased attention to categorical data and the multinomial distribution. The multinomial's exponential family structure is just a little tricky.

We assume that n subjects have each been put into one of L categories. In the ulcer data example from Section 1.4, page 11, $n = 45$, $L = 4$, with categories *Treatment-Success*, *Treatment-Failure*, *Control-Success*, *Control-Failure*.



	Success	*Failure*	
Treatment	**9**	**12**	**21**
Control	**7**	**17**	**24**
	16	**29**	**45**

It is convenient to code the L categories with indicator vectors,

$$e_l = (0, 0, \ldots, 0, 1, 0, \ldots, 0)' \qquad \text{(1 in the lth place)},$$

indicating category l. The data can be written as

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_n)',$$

where

$$y_i = e_{l_i}, \qquad l_i \text{ the ith subject's category.}$$

The multinomial probability model says that each subject's category is chosen independently according to probability distribution $\pi = (\pi_1, \pi_2, \ldots, \pi_L)'$,

$$y_i \sim \{e_l \text{ with probability } \pi_l,\ l = 1, 2, \ldots, L\}$$

independently for $i = 1, 2, \ldots, n$. The probability density function is

$$g_{\boldsymbol{\pi}}(\boldsymbol{y}) = \prod_{l=1}^{L} \pi_l^{s_l} = e^{\eta' s},$$

where $s_l = \#\{y_i = e_l\}$, the count for category l, and

$$\eta_l = \log \pi_l \qquad \left[\eta = (\eta_1, \eta_2, \dots, \eta_L)'\right].$$

However, this isn't in exponential family form since $\eta = (\eta_1, \eta_2, \dots, \eta_L)'$ is constrained to lie in a *nonlinear* subset of $\mathcal{R}^L$,

$$\sum_{l=1}^{L} e^{\eta_l} = \sum_{1}^{L} \pi_l = 1.$$

To circumvent this difficulty we let η be *any* vector in $\mathcal{R}^L$, and define

$$\pi_l = e^{\eta_l} \Bigg/ \sum_{j=1}^{L} e^{\eta_j} \qquad \text{for } l = 1, 2, \dots, L, \tag{2.7}$$

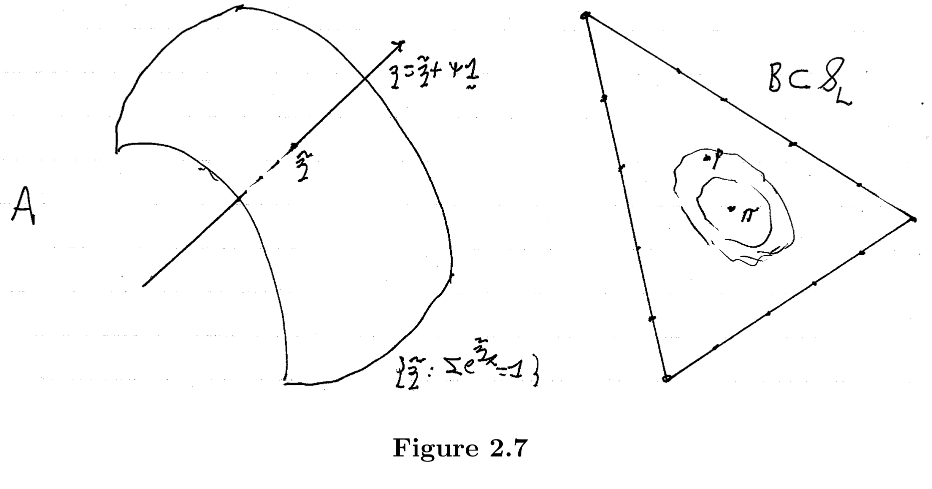
so

$$\log \pi_l = \eta_l - \log \sum_{j=1}^{L} e^{\eta_j}.$$

Now the multinomial density can be written in genuine exponential family form,

$$g_\eta(\boldsymbol{y}) = e^{\eta' s - n\psi(\eta)} \qquad \left[\psi(\eta) = \log \sum_{l=1}^{L} e^{\eta_l}\right].$$

The count vector $s = (s_1, s_2, \dots, s_L)$ is a sufficient statistic; the sample space $\mathcal{Y}$ for $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$ has L^n points, all with $g_0(\boldsymbol{y}) = 1$.



A
$\eta = \tilde{\eta} + \psi 1$
$\tilde{\eta}$
$\{\tilde{\eta} : \Sigma e^{\tilde{\eta}_l} = 1\}$
$B \subset S_L$
p
π

Figure 2.7

More familiarly, the multinomial is expressed in terms of the vector of observed proportions p,

$$p_l = \frac{s_l}{n} \qquad \left[p = (p_1, p_2, \ldots, p_n)'\right],$$

as

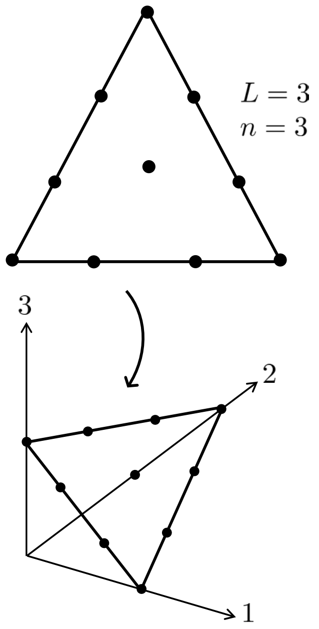
$$g_\eta(p) = e^{n[\eta' p - \psi(\eta)]} g_0(p). \tag{2.8}$$

Now $g_0(p)$ is the multinomial coefficient

$$g_0(p) = \binom{n}{np_1, np_2, \ldots, np_L} = n! \Big/ \prod_{l=1}^{L} s_l!$$

defined with respect to counting measure on the lattice of points $(s_1, s_2, \ldots, s_L)'/n$, the s_l being non-negative integers summing to n. This will be denoted

$$p \sim \mathrm{Mult}_L(n, \pi)/n. \tag{2.9}$$



To summarize, p is the vector of observed proportions, π the vector of category probabilities, parameterized in terms of the vector η at (2.7), and obeying density (2.8).

Homework 2.25. Re-express $g_\eta(p)$ so that $\eta = 0$ corresponds to $p \sim \mathrm{Mult}(n, \pi^0)/n$, where $\pi^0 = (1, 1, \ldots, 1)'/L$ (the centerpoint of the simplex $\mathcal{S}$).

Homework 2.26. Why is η_1 the log odds ratio $\log(\pi_1\pi_4/\pi_2\pi_3)$ in the 2×2 case of Example 1 of Section 2.8? Hint: Homework 2.22.

Homework 2.27. Differentiate $\psi(\eta)$ to show that p has expectation vector π ("$= \mu$" in our original notation) and covariance matrix $V = D_\pi - \pi\pi'$, where $D = \mathrm{diag}(\pi)$, the diagonal matrix with elements π_l, so that $V_{ij} = \pi_i(1 - \pi_i)$ if $i = j$, and $-\pi_i\pi_j$ otherwise.

As n gets large, p becomes approximately normal,

$$p \dot{\sim} \mathcal{N}_L(\pi, V),$$

but its L-dimensional distribution is confined to the $(L-1)$-dimensional flat set $\mathcal{S}_L$. The preceding diagram is a schematic depiction of the case $n = 3$ and $L = 3$.

In our parameterization, all η vectors on $\mathcal{R}^L$ of the form $\eta = \tilde{\eta} + \psi \cdot \mathbf{1}$, with ψ any number, $\mathbf{1} = (1, 1, \ldots, 1)'$, and $\sum e^{\eta_l} = 1$, map into the same expectation vector $\pi = (\cdots e^{\eta_l} \cdots)'$. As a result, the Hessian matrix $d\mu/d\eta = V$ is singular,

$$V \cdot \mathbf{1} = \pi - \pi = 0.$$

We can take the inverse matrix $d\eta/d\mu = V^{-1}$ to be any pseudoinverse of $D_\pi - \pi\pi'$, in particular

$$V^{-1} = \mathrm{diag}(1/\pi),$$

the diagonal matrix with entries $1/\pi_l$; but we won't need to do so here. The multinomial can be expressed in standard form by taking $\eta = (\log \pi_1, \log \pi_2, \dots, \log \pi_{L-1})'$ and $y = (p_1, p_2, \dots, p_{L-1})'$, but this often causes problems because of asymmetry.

Homework 2.28. In what way is the Poisson trick related to V^{-1} above?

There is one more thing to say about the multinomial family: it represents *all* distributions supported on exactly L points, while all other exponential families are subfamilies of more general probability models.

Exponential Families in Theory and Practice

Bradley Efron
Stanford University

Part 3

Generalized Linear Models

Normal theory linear regression, including the analysis of variance, has been a mainstay of statistical practice for nearly a century. Generalized linear models (GLMs) began their development in the 1960s, extending regression theory to situations where the response variables are binomial, Poisson, gamma, or any one-parameter exponential family. GLMs have turned out to be the great success story of exponential family techniques as applied to the world of statistical practice.

3.1 Exponential family regression models

Suppose that $y_1, y_2, \dots, y_N$ are independent observations from the same one-parameter exponential family, but having possibly different values of the natural parameter η,

$$y_i \overset{\text{ind}}{\sim} g_{\eta_i}(y_i) = e^{\eta_i y_i - \psi(\eta_i)} g_0(y_i) \qquad \text{for } i = 1, 2, \dots, N. \tag{3.1}$$

A generalized linear model expresses the η_i as linear functions of an unknown p-dimensional parameter vector β,

$$\eta_i = x_i'\beta \qquad \text{for } i = 1, 2, \dots, N,$$

where the x_i are known covariate vectors. We can write this all at once as

$$\underset{N\times 1}{\boldsymbol{\eta}} = \underset{N\times p}{X}\ \underset{p\times 1}{\beta} \qquad \left[X = (x_1, x_2, \dots, x_N)'\right].$$

The density for $\boldsymbol{y} = (y_1, y_2, \dots, y_N)'$ turns out to be a p-parameter exponential family. Multiplying factors (3.1) gives the density

$$\begin{aligned} g_\beta(\boldsymbol{y}) &= e^{\sum_i(\eta_i y_i - \psi(\eta_i))} \prod_i g_0(y_i) = e^{\beta' X' \boldsymbol{y} - \sum_i \psi(x_i'\beta)} \prod_i g_0(y_i) \\ &= e^{\beta' z - \phi(\beta)} g_0(\boldsymbol{y}) \end{aligned} \tag{3.2}$$

where

- β is the $p \times 1$ natural parameter vector;
- $z = X'\boldsymbol{y}$ is the $p \times 1$ sufficient vector;
- $\phi(\beta) = \sum_{i=1}^N \psi(x_i'\beta)$ is the cgf; and
- $g_0(\boldsymbol{y}) = \prod_{i=1}^N g_0(y_i)$ is the carrying density.

Notation Boldface vectors will be used for N-dimensional quantities, as with $\boldsymbol{y} = (y_1, y_2, \dots, y_n)'$ and $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_N)'$; likewise $\boldsymbol{\mu}$ for the expectation vector $(\mu_1, \mu_2, \dots, \mu_N)'$, $\mu_i = \dot\psi(\eta_i)$, or more succinctly, $\boldsymbol{\mu} = \dot\psi(\boldsymbol{\eta})$. Similarly, $\boldsymbol{V}$ will denote the $N \times N$ diagonal matrix of variances, written $\boldsymbol{V} = \ddot\psi(\boldsymbol{\eta}) = \operatorname{diag}(\ddot\psi(\eta_i))$, with $\boldsymbol{V}_\beta = \operatorname{diag}(\ddot\psi(x_i'\beta))$ indicating the variance matrix for parameter vectors $\boldsymbol{\eta} = X\beta$ in the GLM.

Homework 3.1. Show that

(a) $E_\beta\{z\} = \dot\phi(\beta) = \underset{p\times N}{X'}\ \underset{N\times 1}{\boldsymbol{\mu}}(\beta)$

(b) $\operatorname{Cov}_\beta\{z\} = \ddot\phi(\beta) = \underset{p\times N}{X'}\ \underset{N\times N}{\boldsymbol{V}_\beta}\ \underset{N\times p}{X} = i_\beta$ (i_β the Fisher information for β)

(c) $\underset{p\times N}{\dfrac{d\hat\beta}{d\boldsymbol{y}}} = (X'\boldsymbol{V}_{\hat\beta}X)^{-1}X'$ ("influence function of $\hat\beta$")

The score function for a GLM $\dot l_\beta(\boldsymbol{y}) = (\cdots \partial l(\boldsymbol{y})/\partial\beta_k \cdots)'$ is

$$\dot l_\beta(\boldsymbol{y}) = z - E_\beta\{z\} = X'(\boldsymbol{y} - \boldsymbol{\mu}),$$

with

$$-\ddot{l}_\beta(\boldsymbol{y}) = X'\boldsymbol{V}_\beta X = i_\beta,$$

the $p \times p$ Fisher information matrix.

The MLE equation is $z = E_{\beta=\hat{\beta}}\{Z\}$ or

$$X'\left(\boldsymbol{y} - \boldsymbol{\mu}\left(\hat{\beta}\right)\right) = \boldsymbol{0}, \tag{3.3}$$

$\boldsymbol{0}$ here being a p-vector of zeros. Asymptotically, as the Fisher information matrix i_β grows large, the general theory of maximum likelihood estimation yields the normal approximation

$$\hat{\beta} \dot{\sim} \mathcal{N}_\beta(\beta, i_\beta^{-1}) \tag{3.4}$$

(though a formal statement depends on the boundedness of the x_i vectors as $N \to \infty$). Solving (3.3) for $\hat{\beta}$ is usually carried out by some version of Newton–Raphson iteration, as discussed below.

The regression model $\{\eta_i = x_i'\beta\}$ is a p-parameter subfamily of the N-parameter family (3.1) that lets each η_i take on any value,

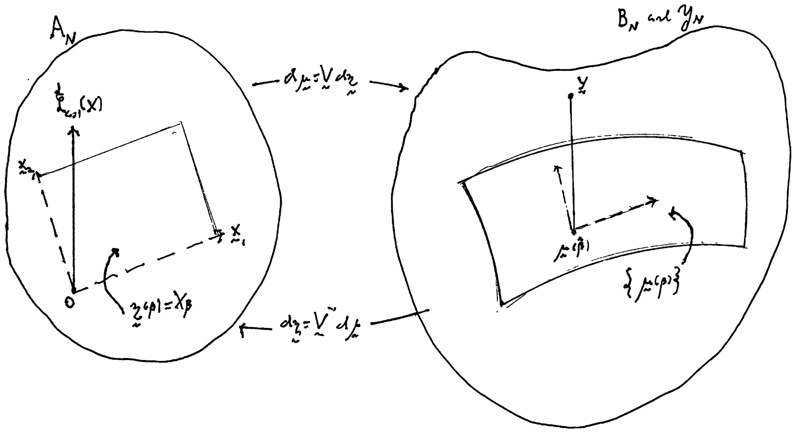
$$g_{\boldsymbol{\eta}}(\boldsymbol{y}) = e^{\boldsymbol{\eta}'\boldsymbol{y} - \sum_i \psi(\eta_i)} g_0(\boldsymbol{y}).$$

If the original one-parameter family $g_\eta(y)$ in (3.1) has natural space $\eta \in A$ and expectation space $\mu \in B$, then $g_{\boldsymbol{\eta}}(\boldsymbol{y})$ has spaces A_N and B_N as N-fold products, $A_N = A^N$ and $B_N = B^N$, and similarly $\mathcal{Y}_N = \mathcal{Y}^N$ for the sample spaces.

The natural parameter vectors for the GLM, $\boldsymbol{\eta} = X\beta$, lie in the p-dimensional linear subspace of A_N generated by the columns of $X = (\boldsymbol{x}_{(1)}, \boldsymbol{x}_{(2)}, \ldots, \boldsymbol{x}_{(p)})$. This flat space maps into a curved p-dimensional manifold in B_N,

$$\{\boldsymbol{\mu}(\beta)\} = \left\{\boldsymbol{\mu} = \dot{\psi}(\boldsymbol{\eta}), \boldsymbol{\eta} = X\beta\right\},$$

as pictured here.



(In the linear regression situation, $y_i \overset{\text{ind}}{\sim} \mathcal{N}(x_i'\beta, \sigma^2)$, $\{\boldsymbol{\mu}(\beta)\}$ is flat, but this is essentially the only such case.)

The data vector $\boldsymbol{y}$ will usually *not* lie in the curved manifold $\{\boldsymbol{\mu}(\beta)\}$. From the MLE equations $X'(\boldsymbol{y}-\boldsymbol{\mu}(\hat{\beta})) = 0$ we see that $\boldsymbol{y}-\boldsymbol{\mu}(\hat{\beta})$ must be orthogonal to the columns $\boldsymbol{x}_{(j)}$ of X. In other words, the maximum likelihood estimate $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}(\hat{\beta})$ is obtained by projecting $\boldsymbol{y}$ into $\{\boldsymbol{\mu}(\beta)\}$ orthogonally to the columns of X. Letting $\mathcal{L}_{\text{col}}(X)$ be the column space of X,

$$\mathcal{L}_{\text{col}}(X) = \{\boldsymbol{\eta} = X\beta, \beta \in \mathcal{R}^p\},$$

the residual $\boldsymbol{y} - X\hat{\beta}$ must lie in the space $\overset{\perp}{\mathcal{L}}_{\text{col}}(X)$ of N-vectors orthogonal to $\mathcal{L}_{\text{col}}(X)$.

In the normal case, where $\{\boldsymbol{\mu}(\beta)\}$ is flat, $\boldsymbol{\mu}(\hat{\beta})$ is obtained from the usual OLS (ordinary least squares) equations. Iterative methods are necessary for GLMs: if β^0 is an interim guess, we update to $\beta^1 = \beta^0 + d\beta$ where

$$d\beta = (X'\boldsymbol{V}_{\beta^0}X)^{-1}\left(\boldsymbol{y} - \boldsymbol{\mu}(\beta^0)\right)$$

(see Homework 3.1(c)), continuing until $d\beta$ is sufficiently close to 0. Because $g_\beta(\boldsymbol{y})$ is an exponential family (3.2) there are no local maxima to worry about.

Modern computer packages such as `glm` in R find $\hat{\beta}$ quickly and painlessly. Having found it they use the asymptotic normal approximation

$$\hat{\beta} \dot{\sim} \mathcal{N}_p\left(\beta, (X'\boldsymbol{V}_\beta X)^{-1}\right) \tag{3.5}$$

(from Homework 3.1(b)) to report approximate standard errors for the components of $\hat{\beta}$.

Warning The resulting approximate confidence intervals, e.g., $\hat{\beta}_j \pm 1.96\widehat{\text{se}}_j$, may be quite inaccurate, as shown by comparison with better bootstrap intervals.

Two useful extensions

We can add a known "offset" vector $\boldsymbol{a}$ to the definition of $\boldsymbol{\eta}$ in the GLM

$$\boldsymbol{\eta} = \boldsymbol{a} + X\beta.$$

Everything said previously remains valid, except now

$$\mu_i(\beta) = \dot{\psi}(a_i + x_i'\beta) \quad \text{and} \quad V_i(\beta) = \ddot{\psi}(a_i + x_i'\beta).$$

Homework 3.2. Write the offset situation in the form $g_\beta(\boldsymbol{y}) = e^{\beta' z - \phi(\beta)} g_0(\boldsymbol{y})$. Show that Homework 3.1(a) and (b) still hold true, with the changes just stated.

As a second extension, suppose that corresponding to each case i we have n_i iid observations,

$$y_{i1}, y_{i2}, \dots, y_{in_i} \overset{\text{iid}}{\sim} g_{\eta_i}(y) = e^{\eta_i y - \psi(\eta_i)} g_0(y),$$

with

$$\eta_i = x_i'\beta, \qquad i = 1, 2, \dots, N.$$

This is the same situation as before, now of size $\sum_1^N n_i$. However, we can reduce to the sufficient statistics

$$\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i,$$

having

$$\bar{y}_i \overset{\text{ind}}{\sim} e^{n_i[\eta_i \bar{y}_i - \psi(\eta_i)]} g_0(\bar{y}_i) \qquad (\eta_i = x_i'\beta),$$

reducing the size of the GLM from $\sum n_i$ to N.

Homework 3.3. Let $\boldsymbol{n\bar{y}}$ be the N-vector with elements $n_i\bar{y}_i$, similarly $\boldsymbol{n\mu}$ for the vector of elements $n_i\mu_i(\beta)$, and $\boldsymbol{nV_\beta}$ for $\operatorname{diag}(n_iV_i(\beta))$.

(a) Show that

$$g_\beta(\boldsymbol{y}) = e^{\beta' z - \phi(\beta)} g_0(\boldsymbol{y}) \begin{cases} z = X'(\boldsymbol{n\bar{y}}) \\ \phi = \sum_{i=1}^N n_i \psi(x_i'\beta). \end{cases}$$

(b) Also show that

$$\dot{l}_\beta(\boldsymbol{y}) = X'(\boldsymbol{n\bar{y}} - \boldsymbol{n\mu}), \quad -\ddot{l}_\beta(\boldsymbol{y}) = X'\boldsymbol{nV_\beta}X = i_\beta.$$

Note. Standard errors for the components of $\hat{\beta}$ are usually based on the approximation $\operatorname{Cov}(\hat{\beta}) = i_{\hat{\beta}}^{-1}$.

Homework 3.4 (“Self-grouping property”)**.** Suppose we *don’t* reduce to the sufficient statistics $\bar{y}_i$, instead doing a GLM with X having $\sum n_i$ rows. Show that we get the same estimates of $\hat{\beta}$ and $i_{\hat{\beta}}$.

Homework 3.5. Show that the solution to the MLE equation (3.3) minimizes the total deviance distance

$$\hat{\beta} = \arg\min_\beta \left\{ \sum_{i=1}^N D\left(y_i, \mu_i(\beta)\right) \right\}.$$

In other words, “least deviance” is the GLM analogue of least squares for normal regression.

3.2 Logistic regression

When the observations y_i are binomials we are in the realm of logistic regression, the most widely used of the generalized linear models. In the simplest formulation, the y_i’s are independent *Bernoulli* variables $\text{Ber}(\pi_i)$, that is $\text{Bi}(1, \pi_i)$,

$$y_i = \begin{cases} 1 & \text{with probability } \pi_i \\ 0 & \text{with probability } 1 - \pi_i, \end{cases}$$

where 1 or 0 code the outcomes of a dichotomy, perhaps male or female, or success or failure, etc.

The binomial natural parameter is the logistic transform

$$\eta = \log \frac{\pi}{1-\pi} \qquad \left(\pi = \frac{1}{1+e^{-\eta}}\right).$$

A logistic GLM is then of the form $\eta_i = x_i'\beta$ for $i = 1, 2, \ldots, N$ or $\boldsymbol{\eta}(\beta) = X\beta$. The vector of probabilities $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)'$ is $\boldsymbol{\mu}$ in this case; the MLE equations (3.3) can be written in vector notation as

$$X'\left(\boldsymbol{y} - \frac{\mathbf{1}}{\mathbf{1} + e^{-\boldsymbol{\eta}(\beta)}}\right) = \mathbf{0}.$$

Table 3.1: Output of logistic regression for the transplant data. Null deviance 139.189 on 222 degrees of freedom; residual deviance 54.465 on 210 df.

	Estimate	St. error	z-value	Pr($>|z|$)
inter	-6.76	1.48	-4.57	.00
age	$-.21$	.41	$-.52$	.60
gen	$-.61$	.42	-1.45	.15
diag	.57	.41	1.40	.16
donor	$-.68$	.46	-1.48	.14
start	$-.07$	.61	$-.12$	.91
date	.41	.49	.83	.41
datf	.12	.62	.19	.85
datl	-1.26	.66	-1.89	.06
vl1	.07	.49	.15	.88
vl2	$-.71$	.47	-1.49	.13
vl3	.21	.47	.44	.66
vl4	5.30	1.48	3.58	.00***

Example (The transplant data). Among $N = 223$ organ transplant patients, 21 subsequently suffered a severe viral infection. The investigators wished to predict infection ("$y = 1$"). There were 12 predictor variables, including age, gender, and initial diagnosis, and four viral load measurements taken during the first weeks after transplant. In this study, the covariate matrix X was 223×12, with response vector $\boldsymbol{y}$ of length 223. The R call `glm(`$y \sim X$`,binomial)` gave the results in Table 3.1. Only the final viral load measurement (vl4) was seen to be a significant predictor, but that doesn't mean that the others aren't informative. The total residual deviance from the MLE fit was $\sum D(y_i, \pi_i(\hat\beta)) = 54.465$, compared to $\sum D(y_i, \bar y) = 139.189$ for the model that only predicts the average response $\bar y = 21/220$. (More on GLM versions of F-tests later.)

The logistic fit gave a predicted probability of infection

$$\hat\pi_i = \pi_i\left(\hat\beta\right) = \frac{1}{1 + e^{-x_i'\hat\beta}}$$

for each patient. The top panel of Figure 3.1 compares the predictions for the 199 patients who did

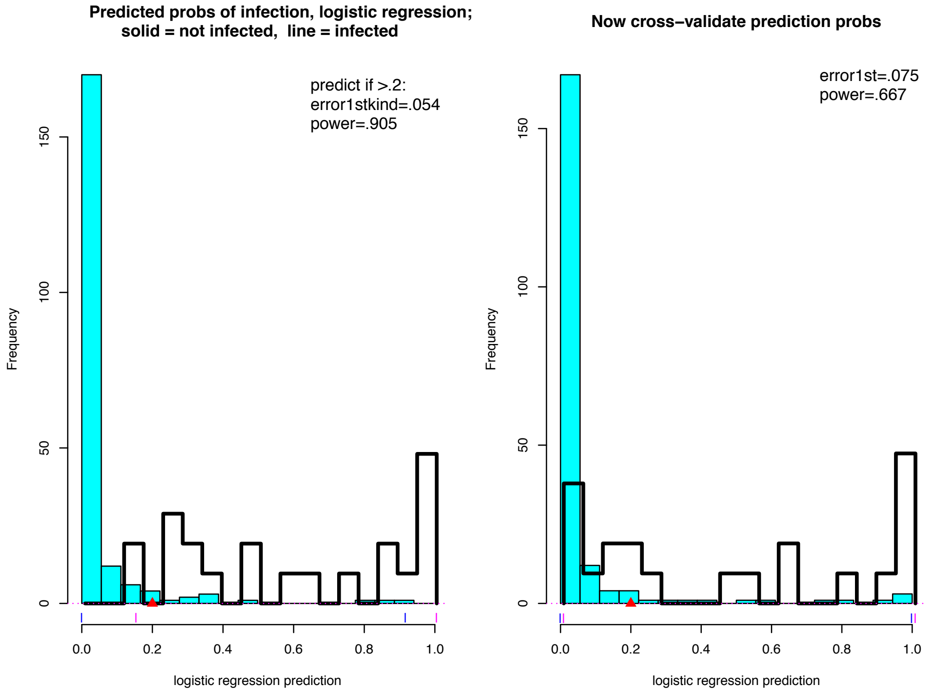


Figure 3.1: Predicted probabilities of infection, logistic regression; solid histogram represents not infected, line histogram represents infected.

not suffer an infection (solid blue histogram) with the 21 who did (line histogram). There seems to be considerable predictive power: the rule "predict infection if $\hat{\pi}_i$ exceeds 0.2" makes only 5% errors of the first kind (predicting an infection that doesn't happen) with 90% power (predicting infections that do happen).

This is likely to be optimistic since the MLE rule was fit to the data it is trying to predict. A cross-validation analysis split the 223 patients into 11 groups of 20 each (three of the patients were excluded). Each group was omitted in turn and a logistic regression fit to the reduced set of 200, then predictions $\tilde{\pi}_i$ made for the omitted patients, based on the reduced MLE. This gave more realistic prediction estimates, with 8% errors of the first kind and 67% power.

Homework 3.6. Repeat this analysis removing vl4 from the list of covariates. Comment on your findings.

Standard errors Suppose $\zeta = h(\beta)$ is a real-valued function of β, having gradient $\dot{h}(\beta) = (\cdots \partial h(\beta)/\partial\beta_j \cdots)'$. Then the approximate standard error assigned to the MLE $\hat{\zeta} = h(\hat{\beta})$ is usually

$$\mathrm{se}\left(\hat{\zeta}\right) \doteq \left\{\dot{h}\left(\hat{\beta}\right)' i_{\hat{\beta}}^{-1} \dot{h}\left(\hat{\beta}\right)\right\}^{1/2}.$$

In particular suppose we wish to estimate the probability of a 1 at a given covariate point x_0,

$$\pi_0 = \Pr\{y = 1 \mid x_0\} = 1/(1 + e^{-x_0'\beta}).$$

Then $\dot{h}(\beta) = x_0\pi_0(1-\pi_0)$, and

$$\mathrm{se}(\hat{\pi}_0) \doteq \hat{\pi}_0(1-\hat{\pi}_0)\left\{x_0' i_{\hat{\beta}}^{-1} x_0\right\}^{1/2},$$

$\hat{\pi}_0 = 1/(1+e^{-x_0'\hat{\beta}})$.

Logistic regression includes the situation where $y_i \overset{\text{ind}}{\sim} \mathrm{Bi}(n_i, \pi_i)$, n_i possibly greater than 1. Let $p_i = y_i/n_i$ denote the proportion of 1's,

$$p_i \overset{\text{ind}}{\sim} \mathrm{Bi}(n_i, \pi_i)/n_i,$$

so p_i is $\bar{y}_i$ in the notation of (3.4), with $\mu_i = \pi_i$. From Homework 3.3(b) we get

$$\dot{l}_\beta(\boldsymbol{y}) = X'(\boldsymbol{np} - \boldsymbol{n\pi}) \quad \text{and} \quad -\ddot{l}_\beta = X' \operatorname{diag}\left\{n_i\pi_i(\beta)\left[1-\pi_i(\beta)\right]\right\} X.$$

Probit analysis

The roots of logistic regression lie in *bioassay*: to establish the toxicity of a new drug, groups of n_i mice each are exposed to an increasing sequence of doses d_i, $i = 1, 2, \dots, K$, and the proportion p_i of deaths observed. (A customary goal is to estimate "LD50", the dose yielding 50% lethality.) The *probit model* is

$$\pi_i = \Phi(\beta_0 + \beta_1 d_i), \qquad i = 1, 2, \dots, K,$$

where Φ is the standard normal cdf; maximum likelihood is used to solve for (β_0, β_1).

Homework 3.7. Show that replacing $\Phi(x)$ above with the logistic cdf $\Lambda(x) = 1/(1+e^{-x})$ reduces the bioassay problem to logistic regression.

Linkages

The key idea of GLMs is to linearly model the natural parameters η_i. Since $\mu_i = \dot{\psi}(\eta_i)$, this is equivalent to linearly modeling $\dot{\psi}^{-1}(\mu_i)$. Other links appear in the literature. Probit analysis amounts to linearly modeling $\Phi^{-1}(\mu_i)$, sometimes called the *probit link*. But only the GLM "canonical link" allows one to make full use of exponential family theory.

3.3 Poisson regression

The second most familar of the GLMs — and for general purposes sometimes the most useful, as we will see — is Poisson regression. We observe independent Poisson variables

$$y_i \overset{\text{ind}}{\sim} \mathrm{Poi}(\mu_i) \qquad \text{for } i = 1, 2, \dots, N,$$

sometimes written $\boldsymbol{y} \sim \text{Poi}(\boldsymbol{\mu})$. A Poisson GLM is a linear model for the natural parameters $\eta_i = \log \mu_i$,

$$\eta_i = a_i + x_i'\beta, \qquad i = 1, 2, \dots, N,$$

where β is an unknown p-dimensional parameter vector, x_i a known p-dimensional covariate vector, and a_i a known scalar "offset". The MLE equation (3.3) is

$$X'(\boldsymbol{y} - e^{\boldsymbol{a}+X\hat{\beta}}) = \mathbf{0},$$

the exponential notation indicating the vector with components $e^{a_i+x_i'\hat{\beta}}$. Since the variance V_i equals $\mu_i = e^{\eta_i}$ for Poisson variates, the asymptotic approximation (3.4) is

$$\hat{\beta} \mathrel{\dot\sim} \mathcal{N}_p\left\{\beta, \left[X' \operatorname{diag}(e^{a_i+x_i'\beta})X\right]^{-1}\right\},$$

in practice with $\hat{\beta}$ substituted for β on the right.

Table 3.2: Counts for a truncated sample of 487 galaxies, binned by magnitude and redshift.

		Redshift (farther) ⟶														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	18	1	6	6	3	1	4	6	8	8	20	10	7	16	9	4
	17	3	2	3	4	0	5	7	6	6	7	5	7	6	8	5
	16	3	2	3	3	3	2	9	9	6	3	5	4	5	2	1
	15	1	1	4	3	4	3	2	3	8	9	4	3	4	1	1
	14	1	3	2	3	3	4	5	7	6	7	3	4	0	0	1
	13	3	2	4	5	3	6	4	3	2	2	5	1	0	0	0
	12	2	0	2	4	5	4	2	3	3	0	1	2	0	0	1
↑	11	4	1	1	4	7	3	3	1	2	0	1	1	0	0	0
Magnitude	10	1	0	0	2	2	2	1	2	0	0	0	1	2	0	0
(dimmer)	9	1	1	0	2	2	2	0	0	0	0	1	0	0	0	0
	8	1	0	0	0	1	1	0	0	0	0	1	1	0	0	0
	7	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0
	6	0	0	3	1	1	0	0	0	0	0	0	0	0	0	0
	5	0	3	1	1	0	0	0	0	0	0	0	0	0	0	0
	4	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0
	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

The galaxy data Table 3.2 shows counts of galaxies from a survey of a small portion of the sky: 487 galaxies have had their apparent magnitudes m and (log) redshifts r measured. Apparent brightness is a *decreasing* function of magnitude — stars of the 2nd magnitude are less bright than those of the first, etc. — while distance from Earth is an increasing function of r.

As in most astronomical studies, the galaxy data is *truncated*, very dim galaxies lying below the threshold of detection. In this study, attention was restricted to the intervals $17.2 \leq m \leq 21.5$ and $1.22 \leq r \leq 3.32$. The range of m has been divided into 18 equal intervals, and likewise 15 equal intervals for r. Table 3.2 gives the counts y_{ij} of the 487 galaxies in the $N = 270 = 18 \times 15$ bins.

The left panel of Figure 3.2 shows a perspective picture of the counts.

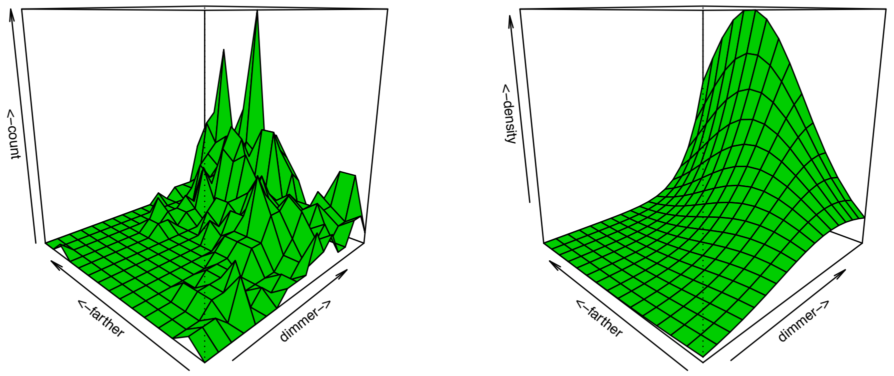


Figure 3.2: *Left panel*: galaxy data, binned counts. *Right panel*: Poisson GLM density estimate.

We can imagine Table 3.2 as the lower left corner of a much larger table we would see if the data were *not* truncated. We might then fit a bivariate normal density to the data. It seems awkward and difficult to fit part of a bivariate normal density to truncated data, but Poisson regression offers an easy solution.

We begin with the reasonable assumption that the counts are independent Poisson observations,

$$y_{ij} \overset{\text{ind}}{\sim} \text{Poi}(\mu_{ij}), \quad i = 1, 2, \dots, 18 \quad \text{and} \quad j = 1, 2, \dots, 15.$$

Let $\boldsymbol{m}$ be the 270-vector listing the m_{ij} values in some order, say $\boldsymbol{m} = (18, 17, \dots, 1)$ repeated 15 times, and likewise $\boldsymbol{r}$ for the 270 r_{ij} values. This defines the 270×15 structure matrix X,

$$X = (\boldsymbol{m}, \boldsymbol{r}, \boldsymbol{m}^2, \boldsymbol{mr}, \boldsymbol{r}^2),$$

where $\boldsymbol{m}^2$ is the 270-vector with components m_{ij}^2, etc.

Letting $\boldsymbol{y}$ denote the 270-vector of counts, the GLM call in R

$$\texttt{glm(}\boldsymbol{y} \sim X\texttt{, poisson),}$$

then produces an estimate of the best-fit truncated normal density. We can see the estimated contours of the fitted density in Figure 3.3. The estimate density itself is shown in the right panel of Figure 3.2.

Homework 3.8. Why does this choice of X for the Poisson regression produce an estimate of a truncated bivariate normal density?

Homework 3.9. (a) Reproduce the Poisson fit.

(b) Calculate the Poisson deviance residuals (1.9). Can you detect any hints of poor fit?

(c) How might you supplement the model we used to improve the fit?

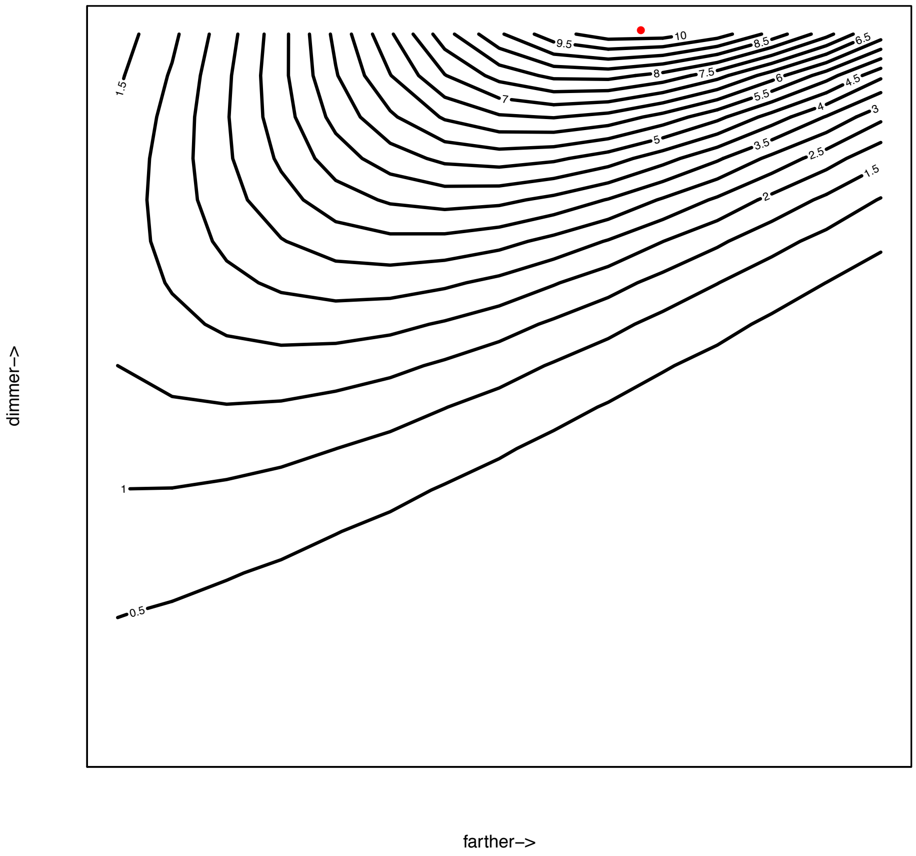


Figure 3.3: Contour curves for Poisson GLM density estimates for the galaxy data; dot shows point of maximum density.

3.4 Lindsey's method [Efron and Tibshirani (1996), *Ann. Statist.* 2431–2461]

Returning to the prostate study of Section 1.6, we have $N = 6033$ observations $z_1, z_2, \dots, z_N$ and wish to fit a density curve $\hat{g}(z)$ to their distribution. For a parametric analysis, we assume that the density is a member of a p-parameter exponential family,

$$g_\beta(z) = e^{\beta' t(z) - \psi(\beta)} g_0(z), \tag{3.6}$$

where β and $t(z)$ are in $\mathcal{R}^p$. In Figure 1.3, $t(z)$ was a fifth-degree polynomial

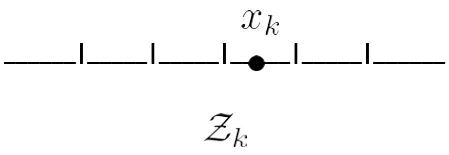
$$t(z) = \sum_{j=1}^{5} z^j.$$

(Choosing a second-degree polynomial, with $g_0(z)$ constant, would amount to fitting a normal density; going up to degree 5 permits us to accommodate non-normal tail behavior.)

How can we find the MLE $\hat{\beta}$ in family (3.6)? There is no closed form for $\psi(\beta)$ or $\mu = \dot{\psi}(\beta)$ except in a few special cases such as the normal and gamma families. This is where Lindsey's method comes in. As a first step we partition the sample space $\mathcal{Z}$ (an interval of $\mathcal{R}^1$) into K

subintervals $\mathcal{Z}_k$,

$$\mathcal{Z} = \bigcup_{k=1}^{K} \mathcal{Z}_k,$$



with $\mathcal{Z}_k$ having length Δ_k and centerpoint x_k. For simplicity we will take $\Delta_k = \Delta$ for all k, and $g_0(z) = 1$ in what follows.

Define

$$\pi_k(\beta) = \Pr_\beta\{z \in \mathcal{Z}_k\} = \int_{\mathcal{Z}_k} g_\beta(z)\ dz \doteq \Delta e^{\beta' t_k - \psi(\beta)}, \tag{3.7}$$

$t_k = t(x_k)$, and

$$\boldsymbol{\pi}(\beta) = (\pi_1(\beta), \pi_2(\beta), \dots, \pi_K(\beta)).$$

Also let $\boldsymbol{y} = (y_1, y_2, \dots, y_K)'$ be the count vector

$$y_k = \#\{z_i \in \mathcal{Z}_k\}.$$

If the z_i's are independent observations from $g_\beta(z)$ then $\boldsymbol{y}$ will be a multinomial sample of size N, Section 2.9,

$$\boldsymbol{y} \sim \text{Mult}_K\left(N, \boldsymbol{\pi}(\beta)\right).$$

For small values of Δ, the multinomial MLE will be nearly the same as the actual $\hat\beta$, but it doesn't seem any easier to find. Poisson regression and the Poisson trick come to the rescue.

Define

$$\mu_k(\beta_0, \beta) = e^{\beta_0 + \beta' t_k}, \tag{3.8}$$

where β_0 is a free parameter that absorbs Δ and $\psi(\beta)$ in (3.7), and let $\mu_+(\beta_0, \beta) = \sum_k \mu_k(\beta_0, \beta)$. Then

$$\frac{\mu_k(\beta_0, \beta)}{\mu_+(\beta_0, \beta)} = \pi_k(\beta).$$

We can now use standard GLM software to find the Poisson MLE $(\hat\beta_0, \hat\beta)$ in model (3.8),

$$\boldsymbol{y} \sim \text{Poi}\left(\boldsymbol{\mu}(\beta_0, \beta)\right).$$

Homework 3.10.

(a) Show that $\hat\beta$ is the MLE in the multinomial model above. What does $e^{\hat\beta_0}$ equal?

(b) How is Lindsey's method applied if the Δ_k are unequal or $g_0(z)$ is not constant?

3.5 Analysis of deviance

Idea We fit an increasing sequence of GLMs to the data, at each stage measuring the lack of fit by the total residual deviance. Then we use the residual deviances to construct an ANOVA-like table.

Total deviance $y_i \overset{\text{ind}}{\sim} g_{\eta_i}(\cdot)$ for $i = 1, 2, \ldots, N$, as in (3.1). The total deviance of $\boldsymbol{y}$ from $\boldsymbol{\eta}$ (or from $\boldsymbol{\mu}$) is

$$D_+(\boldsymbol{y}, \boldsymbol{\mu}) = \sum_{i=1}^{N} D(y_i, \mu_i).$$

Homework 3.11. Verify Hoeffding's formula,

$$\frac{g_{\boldsymbol{y}}^{\boldsymbol{Y}}(\boldsymbol{y})}{g_{\boldsymbol{\mu}}^{\boldsymbol{Y}}(\boldsymbol{y})} \equiv \prod_{i=1}^{N} \frac{g_{y_i}(y_i)}{g_{\mu_i}(y_i)} = e^{D_+(\boldsymbol{y},\boldsymbol{\mu})/2}.$$

Nested linear models

Suppose that in the original model

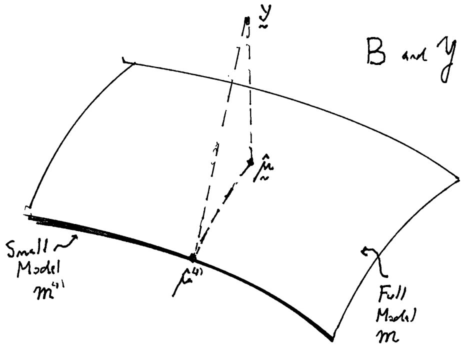
$$\underset{N\times 1}{\boldsymbol{\eta}} = \underset{N\times p}{X} \ \underset{p\times 1}{\beta},$$

β is divided into $(\beta^{(1)}, \beta^{(2)})$ of dimensions $p^{(1)}$ and $p^{(2)}$, $X = (X_1, X_2)$ with X_1 $N \times p^{(1)}$ and X_2 $N \times p^{(2)}$. Then

$$\boldsymbol{\eta} = X^{(1)}\beta^{(1)}$$

is a $p^{(1)}$-parameter GLM submodel of $\boldsymbol{\eta} = X\beta$. Let

- $\hat{\beta}^{(1)}$ = MLE of $\beta^{(1)}$ in the smaller model;
- $\hat{\boldsymbol{\mu}}^{(1)}$ = corresponding expectation vector $\dot{\psi}(X^{(1)}\hat{\beta}^{(1)})$.



The MLEs $\hat{\beta}$ and $\hat{\beta}^{(1)}$ are both obtained by projection as in Section 3.1, with the projections into the curved manifolds

$$\mathcal{M} = \left\{\boldsymbol{\mu} = \dot{\psi}(X\beta)\right\} \quad \text{and} \quad \mathcal{M}^{(1)} = \left\{\boldsymbol{\mu} = \dot{\psi}\left(X^{(1)}, \beta^{(1)}\right)\right\},$$

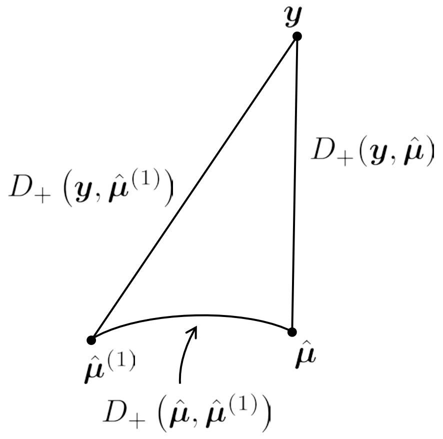
along directions $\overset{\perp}{\mathcal{L}}_{\text{col}}(X)$ and $\overset{\perp}{\mathcal{L}}_{\text{col}}(X^{(1)})$.

The deviance additivity theorem [G. Simon; Efron (1978), *Ann. Statist.* p. 362 Sect. 4]

For standard normal regression theory (OLS), $\mathcal{M}$ and $\mathcal{M}^{(1)}$ are flat spaces, and deviance is Euclidean squared distance. Pythagoras' theorem says that

$$D_+(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}^{(1)}) = D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}^{(1)}) - D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}). \tag{3.9}$$

The *deviance additivity theorem* says that (3.9) holds for any GLM, as discussed next.


$\boldsymbol{y}$
$D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$
$D_+\left(\boldsymbol{y}, \hat{\boldsymbol{\mu}}^{(1)}\right)$
$\hat{\boldsymbol{\mu}}^{(1)}$
$\hat{\boldsymbol{\mu}}$
$D_+\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}^{(1)}\right)$

Hoeffding's formula can be written as differences between total log likelihoods $l_{\boldsymbol{\mu}}(\boldsymbol{y}) = \sum_{i=1}^N l_{\mu_i}(y_i)$,

- $D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2\left[l_{\boldsymbol{y}}(\boldsymbol{y}) - l_{\hat{\boldsymbol{\mu}}}(\boldsymbol{y})\right]$ and
- $D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}^{(1)}) = 2\left[l_{\boldsymbol{y}}(\boldsymbol{y}) - l_{\hat{\boldsymbol{\mu}}^{(1)}}(\boldsymbol{y})\right]$.

Taking differences gives

$$2\left[l_{\hat{\boldsymbol{\mu}}}(\boldsymbol{y}) - l_{\hat{\boldsymbol{\mu}}^{(1)}}(\boldsymbol{y})\right] = D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}^{(1)}) - D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}).$$

Homework 3.12. Show that the left side equals $D_+(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}^{(1)})$.

Testing $H_0 : \beta^{(2)} = 0$ If H_0 is true then Wilks' theorem says that

$$D_+(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}^{(1)}) = 2 \log \left(\frac{g_{\hat{\boldsymbol{\mu}}}(\boldsymbol{y})}{g_{\hat{\boldsymbol{\mu}}^{(1)}}(\boldsymbol{y})} \right) \dot{\sim} \chi^2_{p^{(2)}},$$

so we reject H_0 if $D_+(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}^{(1)})$ exceeds $\chi^{2(\alpha)}_{p^{(2)}}$ for $\alpha = 0.95$ or 0.975, etc. Since

$$D_+(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}^{(1)}) = \sum_{i=1}^N D(\hat{\mu}_i, \hat{\mu}_i^{(1)}),$$

we can examine the individual components to see if any one point is causing a bad fit to the smaller model.

Analysis of deviance table

Suppose β and X are divided into J parts,

$$\underset{p\times 1}{\beta} = \left(\beta^{(1)}_{p^{(1)}}, \beta^{(2)}_{p^{(2)}}, \dots, \beta^{(J)}_{p^{(J)}}\right) \quad \text{and} \quad \underset{N\times p}{X} = \left(\underset{N\times p^{(1)}}{X^{(1)}}, \underset{N\times p^{(2)}}{X^{(2)}}, \dots, \underset{N\times p^{(J)}}{X^{(J)}}\right).$$

Let $\hat{\beta}(j)$ be the MLE for β assuming that $\beta^{(j+1)} = \beta^{(j+2)} = \cdots = \beta^{(J)} = 0$. An analysis of deviance table is obtained by differencing the successive maximized log likelihoods $l_{\hat{\beta}(j)}(\boldsymbol{y}) = \sum_{i=1}^N \log g_{\mu_i(\hat{\beta}(j))}(y_i)$; see Table 3.3.

Note. It is common to adjoin a "zero-th column" of all ones to X, in which case $\hat{\beta}(0)$ is taken to be the value making $\hat{\boldsymbol{\mu}}(0)$ a vector with all entries $\bar{y}$.

Table 3.4 shows $D_+(\boldsymbol{y}_0, \hat{\boldsymbol{\mu}}^{(j)})$ for the prostate data, where $\hat{\boldsymbol{\mu}}^{(j)}$ is the fitted expectation vector from the R call

```
glm(y ~ poly(x,j), poisson)
```

Table 3.3: Analysis of deviance table.

MLE		Twice max log like		Difference	Compare with
$\hat{\beta}(0) = 0$	$\longrightarrow$	$2l_{\hat{\beta}(0)}$			
			$\searrow$		
			$\nearrow$	$D_+(\hat{\mu}(1), \hat{\mu}(0))$	$\chi^2_{p^{(1)}}$
$\hat{\beta}(1)$	$\longrightarrow$	$2l_{\hat{\beta}(1)}$			
			$\searrow$		
			$\nearrow$	$D_+(\hat{\mu}(2), \hat{\mu}(2))$	$\chi^2_{p^{(2)}}$
$\hat{\beta}(2)$	$\longrightarrow$	$2l_{\hat{\beta}(2)}$			
			$\searrow$		
$\vdots$		$\vdots$		$\vdots$	$\vdots$
			$\nearrow$	$D_+(\hat{\mu}(J), \hat{\mu}(J-1))$	$\chi^2_{p^{(J)}}$
$\hat{\beta}(J) = \hat{\beta}$	$\longrightarrow$	$2l_{\hat{\beta}(J)}$			

for $j = 2, 3, \dots, 8$, with $\boldsymbol{y}$ the vector of bin centers in Figure 1.3,

$$\boldsymbol{x} = (-4.4, -4.2, -4.0, \dots, 5.0, 5.2),$$

length $K = 49$. In other words, we used Lindsey's method to fit log polymial models of degrees 2 through 8 to the 6033 z-values.

Table 3.4: Deviance and AIC for prostate data fits `glm(`$\boldsymbol{y} \sim$ `poly(`x`,df), poisson)`.

df	2	3	4	5	6	7	8
Dev	139	137	65.3	64.3	63.8	63.8	59.6
AIC	143	143	73.3	74.3	75.8	77.8	75.6

Because the models are successively bigger, the deviance $D_+^{(j)}$ must decrease with increasing j. It cannot be that bigger models are always better, they just appear so. AIC, Akaike's information criterion, suggests a penalty for increased model size,

$$\text{AIC}^{(j)} = D_+^{(j)} + 2j;$$

see Efron (1986) *JASA*, Remark R. We see that $j = 4$ minimizes AIC for the prostate data.

Homework 3.13.

(a) Construct the deviance table and give the significance levels for the chi-square tests.

(b) Construct the analogous table using natural splines instead of polynomials,

$$\texttt{glm(}\boldsymbol{y}\sim\texttt{ ns(}\boldsymbol{x},j\texttt{), poisson)}.$$

3.6 A survival analysis example

A randomized clinical trial conducted by the Northern California Oncology Group (NCOG) compared two treatments for head and neck cancer: chemotherapy (Arm A of the trial, $n = 51$ patients) and chemotherapy plus radiation (Arm B, $n = 45$ patients). The results are reported in Table 3.5 in terms of the survival time in number of days past treatment. The numbers followed by + indicate patients still alive on their final day of observation. For example, the sixth patient in Arm A was alive on day 74 after his treatment, and then "lost to follow-up"; we only know that his survival time *exceeded* 74 days.

Table 3.5: Censored survival times in days, from the two arms of NCOG study of head and neck cancer.

Arm A: Chemotherapy

7	34	42	63	64	74+	83	84	91	108	112
129	133	133	139	140	140	146	149	154	157	160
160	165	173	176	185+	218	225	241	248	273	277
279+	297	319+	405	417	420	440	523	523+	583	594
1101	1116+	1146	1226+	1349+	1412+	1417				

Arm B: Chemotherapy+Radiation

37	84	92	94	110	112	119	127	130	133	140
146	155	159	169+	173	179	194	195	209	249	281
319	339	432	469	519	528+	547+	613+	633	725	759+
817	1092+	1245+	1331+	1557	1642+	1771+	1776	1897+	2023+	2146+
2297+										

This is a case of *censored data*, an endemic problem in medical survival studies. A powerful methodology for the statistical analysis of censored data was developed between 1955 and 1975. Here we will discuss only a bit of the theory, concerning its connection with generalized linear models. A survey of survival analysis appears in Chapter 9 of Efron and Hastie's 2016 book, *Computer Age Statistical Inference.*

Hazard rates

Survival analysis theory requires stating probability distribution in terms of hazard rates rather then densities. Suppose X is a nonnegative discrete random variable, with probability density

$$f_i = \Pr\{X = i\} \qquad \text{for } i = 1, 2, 3, \ldots,$$

and *survival function*

$$S_i = \Pr\{X \geq i\} = \sum_{j \geq i} f_j.$$

Then h_i, the *hazard rate* at time i,

$$h_i = f_i/S_i = \Pr\{X = i \mid X \geq i\}.$$

In words, h_i is the probability of dying at time i after having survived up until time i. Notice that

$$S_i = \prod_{j=1}^{i-1}(1 - h_j). \tag{3.10}$$

Homework 3.14. Prove (3.10) and give an intuitive explanation.

Life tables

Table 3.6 presents the Arm A data in *life table* form. Now the time unit is months rather than days. Three statistics are given for each month:

- n_i = number of patients under observation at the beginning of month i;
- y_i = number of patients observed to die during month i;
- l_i = number of patients lost to follow-up at the end of month i.

So for instance $n_{10} = 19$ patients were under observation ("at risk") at the beginning of month 10, $y_{10} = 2$ died, $l_{10} = 1$ was lost to follow-up,[1] leaving $n_{11} = 16$ at risk for month 11.

The key assumption of survival analysis is that, given n_i, the number of deaths y_i is binomial with probability of death the hazard rate h_i,

$$y_i \mid n_i \sim \text{Bi}(n_i, h_i). \tag{3.11}$$

This amounts to saying that drop-outs before time i are uninformative for inference except in their effect on n_i.

Homework 3.15. Suppose patients can sense when the end is near, and drop out of the study just before they die. How would this affect model (3.11)?

The unbiased hazard rate estimate based on (3.11) is

$$\hat{h}_i = y_i/n_i; \tag{3.12}$$

[1] Patients can be lost to follow-up for various reasons — moving away, dropping out of the study, etc. — but most often because they entered the study late and were still alive when it closed.

Table 3.6: Arm A of NCOG head and neck cancer study, binned by month: n = number at risk, y = number deaths, l = lost to follow-up, $\hat{h}$ = hazard rate y/n; $\hat{S}$ = life table survival estimate.

month	n	y	l	$\hat{h}$	$\hat{S}$	month	n	y	l	$\hat{h}$	$\hat{S}$
1	51	1	0	.020	.980	25	7	0	0	.000	.184
2	50	2	0	.040	.941	26	7	0	0	.000	.184
3	48	5	1	.104	.843	27	7	0	0	.000	.184
4	42	2	0	.048	.803	28	7	0	0	.000	.184
5	40	8	0	.200	.642	29	7	0	0	.000	.184
6	32	7	0	.219	.502	30	7	0	0	.000	.184
7	25	0	1	.000	.502	31	7	0	0	.000	.184
8	24	3	0	.125	.439	32	7	0	0	.000	.184
9	21	2	0	.095	.397	33	7	0	0	.000	.184
10	19	2	1	.105	.355	34	7	0	0	.000	.184
11	16	0	1	.000	.355	35	7	0	0	.000	.184
12	15	0	0	.000	.355	36	7	0	0	.000	.184
13	15	0	0	.000	.355	37	7	1	1	.143	.158
14	15	3	0	.200	.284	38	5	1	0	.200	.126
15	12	1	0	.083	.261	39	4	0	0	.000	.126
16	11	0	0	.000	.261	40	4	0	0	.000	.126
17	11	0	0	.000	.261	41	4	0	1	.000	.126
18	11	1	1	.091	.237	42	3	0	0	.000	.126
19	9	0	0	.000	.237	43	3	0	0	.000	.126
20	9	2	0	.222	.184	44	3	0	0	.000	.126
21	7	0	0	.000	.184	45	3	0	1	.000	.126
22	7	0	0	.000	.184	46	2	0	0	.000	.126
23	7	0	0	.000	.184	47	2	1	1	.500	.063
24	7	0	0	.000	.184						

(3.10) then gives the survival estimate

$$\hat{S}_i = \prod_{j=1}^{i-1} \left(1 - \hat{h}_j\right). \tag{3.13}$$

Figure 3.4 compares the estimated survival curves for the two arms of the NCOG study. The more aggressive treatment seems better: the Arm B one-year survival rate estimate is about 50%, compared with 35% for Arm A.

Note. Estimated survival curves are customarily called *Kaplan–Meier curves* in the literature. Formally speaking, the name applies to estimates (3.13) where the time unit, months in our example, is decreased to zero. Suppose the observed death times are

$$t_{(1)} < t_{(2)} < t_{(3)} < \cdots < t_{(m)}$$

(assuming no ties). Then the Kaplan–Meier curve $\hat{S}(t)$ is flat between death times, with downward jumps at the observed $t_{(i)}$ values.

Homework 3.16. What is the downward jump at $t_{(i)}$?

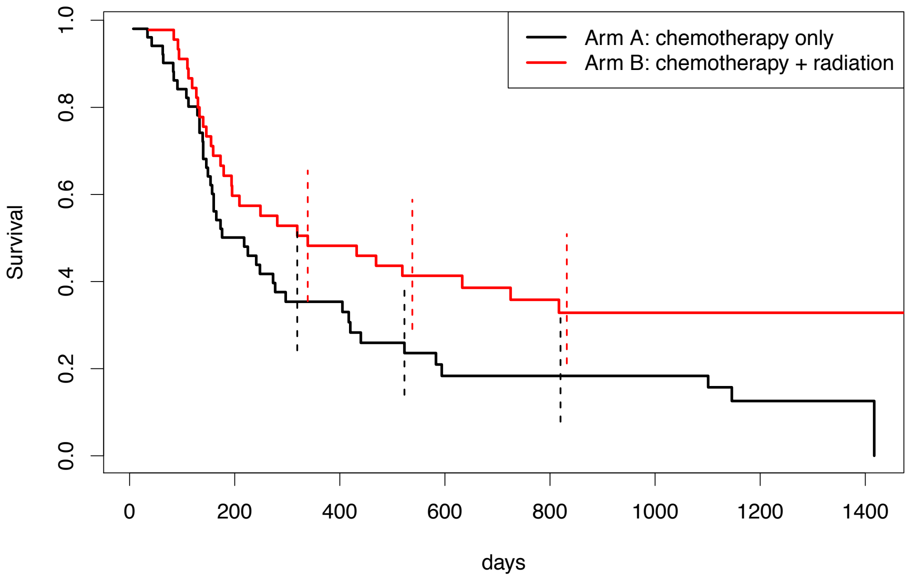


Figure 3.4: NCOG estimated survival curves; lower is Arm A (chemotherapy only); upper is Arm B (chemotherapy+radiation). Vertical lines indicate approximate 95% confidence intervals.

The binomial model[2] (3.11) leads to *Greenwood's formula*, an approximate standard error for $\hat{S}_i$,

$$\text{sd}\left\{\hat{S}_i\right\} \doteq \hat{S}_i \left(\sum_{j \le i} \frac{y_j}{n_j(n_j - y_j)}\right)^{1/2}.$$

The vertical bars in Figure 3.4 indicate $\pm 1.96\,\text{sd}_i$, approximate 95% confidence limits for S_i. There is overlap between the bars for the two curves; at no one time point can we say that Arm B is significantly better than Arm A (though more sophisticated two-sample tests do in fact show B's superiority).

Parametric survival analysis

Life table survival curves are nonparametric in the sense that the true hazard rates h_i are not assumed to follow any particular pattern. A parametric approach can greatly improve the estimation accuracy of the curves. In particular, we can use a logistic GLM: letting η_i be the logistic transform of h_i,

$$\eta_i = \log \frac{h_i}{1 - h_i},$$

and assuming that $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_N)'$ satisfies

$$\boldsymbol{\eta} = X\beta \tag{3.14}$$

[2] It is assumed that the conditional binomial distributions (3.11) are successively independent of the previous observations (n_j, y_j, l_j), $j < i$.

as in Section 3.1–Section 3.2.

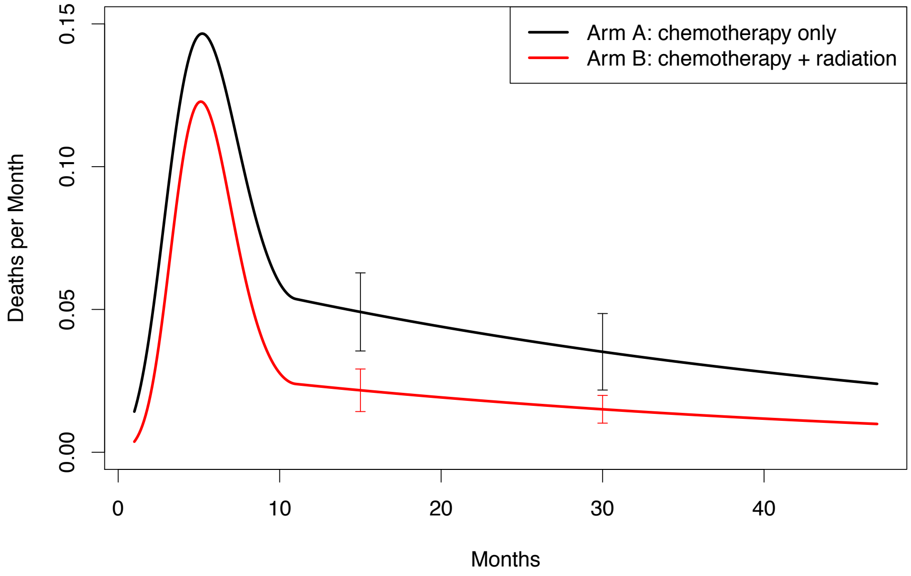


Figure 3.5: Parametric hazard rate estimates for NCOG study. Arm A (black curve) has about 2.5 times higher hazard than Arm B (red curve) for all times more than a year after treatment. Standard errors shown at 15 and 30 months.

Consider the Arm A data of Table 3.6, providing $N = 47$ binomial observations $y_i \sim \mathrm{Bi}(n_i, h_i)$, assumed independent as in Greenwood's formula. For the analysis in Figure 3.5, we took X in (3.14) to be the 47×4 matrix having ith row

$$x_i = \left[1, i, (i-11)_-^2, (i-11)_-^3, \right]', \tag{3.15}$$

where $(i-11)_-$ equals $i-11$ for $i \le 11$ and 0 for $i > 11$. Then $\boldsymbol{\eta} = X\beta$ describes a cubic-linear spline with the knot at 11. This choice allows for more detailed modeling of the early months, when there is the most data and the greatest variation in response, as well as allowing stable estimation in the low-data right tail.

Homework 3.17. Repeat the Arm A calculation in Figure 3.5, including the estimated standard errors.

The comparison of estimated hazard rate in Figure 3.5 is more informative than the survival curve comparison of Figure 3.4. Both arms show a peak in hazard rates at five months, a swift decline, and then a long slow decline after one year, reflecting to some extent the spline model (3.15). Arm B hazard is always below Arm A by a factor of about 2.5.

3.7 Overdispersion and quasi-likelihood

Applications of binomial or Poisson generalized linear models often encounter difficulties with *overdispersion*: after fitting the best GLM we can find, the residual errors are still too large by the standards of binomial or Poisson variability. *Quasilikelihood* is a simple method for dealing with overdispersion while staying within the GLM framework. A more detailed technique, *double exponential families*, is developed in the next section.

Table 3.7: Toxoplasmosis data: rainfall, #sampled and #positive in 34 cities in El Salvador; $p = s/n$, $\hat{\pi} =$ fit from cubic logistic regression in rainfall; R binomial dev residual, Rp Pearson residual; $\sum(R\mathrm{p}^2)/30 = 1.94$, estimated overdispersion factor.

City	r	n	s	p	$\hat{\pi}$	R	Rp
1	1735	4	2	.500	.539	−.16	−.16
2	1936	10	3	.300	.506	−1.32	−1.30
3	2000	5	1	.200	.461	−1.22	−1.17
4	1973	10	3	.300	.480	−1.16	−1.14
5	1750	2	2	1.000	.549	1.55	1.28
6	1800	5	3	.600	.563	.17	.17
7	1750	8	2	.250	.549	−1.72	−1.70
8	2077	19	7	.368	.422	−.47	−.47
9	1920	6	3	.500	.517	−.08	−.08
10	1800	10	8	.800	.563	1.58	1.51
11	2050	24	7	.292	.432	−1.42	−1.39
12	1830	1	0	.000	.560	−1.28	−1.13
13	1650	30	15	.500	.421	.87	.88
14	2200	22	4	.182	.454	−2.69	−2.57
15	2000	1	0	.000	.461	−1.11	−.92
16	1770	11	6	.545	.558	−.08	−.08
17	1920	1	0	.000	.517	−1.21	−1.03
18	1770	54	33	.611	.558	.79	.79
19	2240	9	4	.444	.506	−.37	−.37
20	1620	18	5	.278	.353	−.68	−.67
21	1756	12	2	.167	.552	−2.76	−2.69
22	1650	1	0	.000	.421	−1.04	−.85
23	2250	11	8	.727	.523	1.39	1.36
24	1796	77	41	.532	.563	−.54	−.54
25	1890	51	24	.471	.536	−.93	−.93
26	1871	16	7	.438	.546	−.87	−.87
27	2063	82	46	.561	.427	2.44	2.46
28	2100	13	9	.692	.417	2.00	2.01
29	1918	43	23	.535	.518	.22	.22
30	1834	75	53	.707	.559	2.62	2.57
31	1780	13	8	.615	.561	.40	.40
32	1900	10	3	.300	.530	−1.47	−1.46
33	1976	6	1	.167	.477	−1.60	−1.52
34	2292	37	23	.622	.611	.13	.13

As an example, Table 3.7 reports on the prevalence of toxoplasmosis, an endemic blood infection, in 34 cities of El Salvador; Efron (1986), *JASA* 709–721. The data consists of triplets (r_i, n_i, s_i), $i = 1, 2, \dots, 34$, where

- $r_i =$ annual rainfall in city i;
- $n_i =$ number of people sampled;
- $s_i =$ number testing positive for toxoplasmosis.

Table 3.8: Cubic logistic regression of toxoplasmosis data: `glm(formula` $= p \sim$ `poly(r,3), family` $=$ `binomial, weights` $= n$`)`. Overdispersion: deviance $62.635/30 = 2.09$; Pearson 1.94. Null deviance 74.212 on 33 degrees of freedom; residual deviance 62.635 on 30 df.

Coefficients	Estimate	St. error	z-value	Pr($>|z|$)
(Intercept)	.0243	.0769	.32	.75240
poly(r,3)1	−.0861	.4587	−.19	.85117
poly(r,3)2	−.1927	.4674	−.41	.68014
poly(r,3)3	1.3787	.4115	3.35	.00081 ***

Let $p_i = s_i/n_i$ be the observed proportion positive in city i. A cubic logistic regression of p_i on r_i was run,

$$\texttt{glm(}p \sim \texttt{poly(}r\texttt{,3), binomial, weight } = n\texttt{)},$$

with p, r, and n indicating their respective 34-vectors. Part of the output appears in Table 3.8. We see that the cubic regression coefficient 1.3787 is strongly positive, z-value 3.35, two-sided p-value less than 0.001.

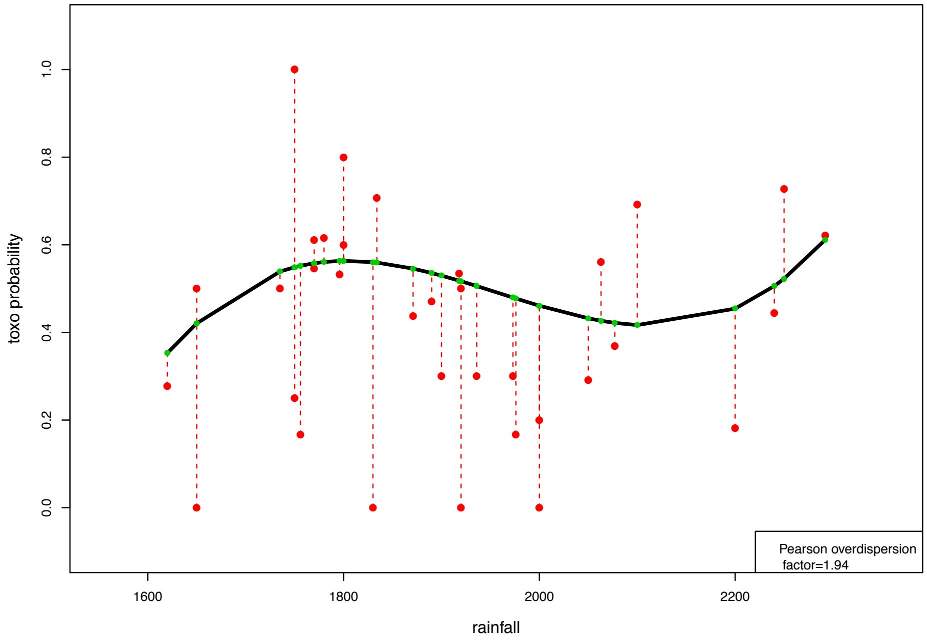


Figure 3.6: Observed proportions of toxoplasmosis, 34 cities in El Salvador; curve is cubic logistic regression.

The points (r_i, p_i) are shown in Figure 3.6, along with the fitted cubic regression curve. Each p_i is connected to its fitted value $\hat{\pi}_i$ by a dashed line. We will see that the points are too far from the curve by the standard of binomial variability. This is what overdispersion looks like.

The middle two columns of Table 3.7 show the observed proportions p_i and the fitted values $\hat{\pi}_i$ from the cubic logistic regression. Two measures of discrepancy are shown in the last two columns:

the binomial deviance residual

$$R_i = \operatorname{sign}(p_i - \hat{\pi}_i)\sqrt{n_i D(p_i, \hat{\pi}_i)},$$

(1.9) and Homework 1.23, and the Pearson residual

$$R\mathrm{p}_i = (p_i - \hat{\pi}_i)\Big/\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)/n_i},$$

$\hat{\pi}_i = 1/(1+e^{-x_i'\hat{\beta}})$.

In the absence of overdispersion, we would expect both

$$\sum_1^{34} R_i^2/30 \quad \text{and} \quad \sum_1^{34} R\mathrm{p}_i^2/30$$

to be close to 1 ($30 = 34 - 4$ is the added degrees of freedom in going from cubic regression to the model allowing a separate estimate π_i for each city). Instead we have

$$\sum_1^{34} R_i^2/30 = 2.09 \quad \text{and} \quad \sum_1^{34} R\mathrm{p}_i^2/30 = 1.94;$$

the points in Figure 3.6 are about $\sqrt{2}$ farther from the fitted curve than as suggested by binomial variability.

Homework 3.18. Compute the p-value for Wilks' likelihood ratio test of the null hypothesis that there is no overdispersion around the cubic regression curve.

Table 3.9: Toxoplasmosis data matrix; c_j city residence for subject j, r_j rainfall in that subject's city, z_j either 0 or 1 indicating positive test for toxoplasmosis or not.

	City	Rainfall	Response
1	$\vdots$	$\vdots$	$\vdots$
2	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	c_j	r_j	z_j
$\vdots$	$\vdots$	$\vdots$	$\vdots$
697	$\vdots$	$\vdots$	$\vdots$

The toxoplasmosis study comprised 697 subjects. It was originally presented as a 697 by 3 matrix, such as that suggested by Table 3.9, with c_j the city residence for subject j, r_j the rainfall in that subject's city, and z_j either 1 or 0 if the test for toxoplasmosis was positive or not.

Homework 3.19. Run logistic regressions for these three models: (1) $z \sim 1$, i.e., only a single constant fit; (2) $z \sim$ `poly(`r`,3)`; (3) $z \sim$ `as.factor(city)`. Compute the analysis of deviance table, using the `anova` function, and interpret the results.

There is nothing mysterious about overdispersion. Overly large residuals mean that our model is deficient. In the toxoplasmosis example there are certainly other predictors — age, gender, neighborhood, etc. — that would reduce residual error if included in the logistic regression model– if we knew them. We don't, but we can at least assess the degree of overdispersion, and account for its effect on the accuracy of estimates $\hat{\beta}_i$ such as those in Table 3.8. This is what the theory of quasilikelihood does.

Quasilikelihood A normal-theory GLM, that is, ordinary least squares, has no problem with overdispersion. The usual model,

$$\boldsymbol{y} \sim X\beta + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}_N(\mathbf{0}, \sigma^2 I),$$

gives MLE $\hat{\beta} = (X'X)^{-1}X'y$, with

$$\hat{\beta} \sim \mathcal{N}_p\left(\beta, \sigma^2(X'X)^{-1}\right);$$

σ^2 is estimated from the residual sum of squared errors. A significance test for the kth coefficient is based on $\hat{\beta}_k/\hat{\sigma}$, automatically accounting for dispersion. Notice that the point estimate $\hat{\beta}_k$ doesn't depend on $\hat{\sigma}$, while its variability does.

Homework 3.20. Suppose we observe independent 0/1 random variables $y_1, y_2, \dots, y_N$, with unknown expectations $E\{y_i\} = \pi_i$, and wish to estimate $\theta = \sum_1^N \pi_i/N$. An unbiased estimate is $\hat{\theta} = \bar{y}$. What can we learn from

$$\hat{\sigma}^2 = \sum_{i=1}^N (y_i - \bar{y})^2/N?$$

The advantage of the normal-theory GLM

$$y_i \overset{\text{ind}}{\sim} \mathcal{N}(x_i'\beta, \sigma^2), \qquad i = 1, 2, \dots, N,$$

is that it incorporates a dispersion parameter σ^2 without leaving the world of exponential families. This isn't possible for other GLMs (however, see Section 3.8). Quasilikelihood theory says that we can act as if it *is* possible.

We begin by considering an extension of the GLM structure. As in (3.1), the observations y_i are obtained from possibly different members of a one-parameter exponential family. Denote the mean and variance of y_i by $y_i \sim (\mu_i, v_i)$, where μ_i is a function $\mu_i(\beta)$ of an unknown p-dimensional parameter vector β; $v_i = v(\mu_i(\beta))$ is determined by the variance function $v(\mu)$ of the family, e.g.,

$v(\mu) = \mu$ for the Poisson family. We assume that the function $\boldsymbol{\mu}_\beta = (\cdots \mu_i(\beta) \cdots)'$ is smoothly defined, with $N \times p$ derivative matrix

$$\boldsymbol{w}_\beta = \left(\frac{\partial \mu_i}{\partial \beta_j}\right).$$

Lemma 3. *The score function and information matrix for an extended GLM family are*

$$\dot{l}_\beta(\boldsymbol{y}) = \boldsymbol{w}'_\beta \boldsymbol{v}_\beta^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_\beta) \quad \textit{and} \quad i_\beta = \boldsymbol{w}'_\beta \boldsymbol{v}_\beta^{-1} \boldsymbol{w}_\beta,$$

where $\boldsymbol{v}_\beta$ is the diagonal matrix with elements $v_i(\beta)$.

The proof of Lemma 3 begins by differentiating $l_\beta(\boldsymbol{y}) = \boldsymbol{\eta}'_\beta \boldsymbol{y} - \sum \psi(\eta_i(\beta))$, using $d\boldsymbol{\eta}_\beta / d\beta = \boldsymbol{v}_\beta^{-1} \boldsymbol{w}_\beta$.

Homework 3.21. Complete the proof.

Note. In a standard unextended GLM, $\boldsymbol{\eta}_\beta = X\beta$, we get

$$\boldsymbol{w}_\beta = \frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}} X = \boldsymbol{V}_\beta X.$$

Setting $\boldsymbol{v}_\beta = \boldsymbol{V}_\beta$ in Lemma 3 gives

$$\dot{l}_\beta(\boldsymbol{y}) = X'(\boldsymbol{y} - \boldsymbol{\mu}_\beta) \quad \text{and} \quad i_\beta = X' \boldsymbol{V}_\beta X,$$

the same as in Section 3.1.

The quasilikelihood approach to overdispersion is simply to assume that

$$v(\mu) = \sigma^2 V(\mu) \qquad \text{(for } \sigma^2 \text{ an unknown positive constant)}, \tag{3.16}$$

where $V(\mu)$ is the variance function in the original family. For instance,

$$v(\mu) = \sigma^2 \mu (1 - \mu) = \sigma^2 \pi (1 - \pi)$$

for the binomial family, or $v(\mu) = \sigma^2 \mu$ for the Poisson family. Applied formally, Lemma 3 gives

$$\dot{l}_\beta(\boldsymbol{y}) = \boldsymbol{w}'_\beta \boldsymbol{V}_\beta^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_\beta)/\sigma^2 \quad \text{and} \quad i_\beta = \boldsymbol{w}'_\beta \boldsymbol{V}_\beta^{-1} \boldsymbol{w}_\beta / \sigma^2. \tag{3.17}$$

Chapter 9 of McCullagh and Nelder (1989) shows that under reasonable asymptotic conditions, the MLE $\hat{\beta}$, i.e., the solution to $\dot{l}_\beta(\boldsymbol{y}) = \boldsymbol{0}$, satisfies

$$\hat{\beta} \mathrel{\dot{\sim}} \mathcal{N}_p(\beta, i_\beta^{-1}).$$

Applied to the original model $\boldsymbol{\eta}_\beta = X\beta$, (3.17) gives

$$\dot{l}_\beta(\boldsymbol{y}) = X'(\boldsymbol{y} - \boldsymbol{\mu}_\beta)/\sigma^2 \quad \text{and} \quad i_\beta = X'\boldsymbol{V}_\beta X/\sigma^2. \tag{3.18}$$

The MLE equation $\dot{l}_\beta(\boldsymbol{y}) = \mathbf{0}$ gives the same estimate $\hat{\beta}$. However the estimated covariance matrix for $\hat{\beta}$ is now multiplied by σ^2,

$$\hat{\beta} \mathrel{\dot{\sim}} \mathcal{N}_p\left(\beta, \sigma^2 (X'\boldsymbol{V}_\beta X)^{-1}\right),$$

compared with (3.1) in Section 3.1.

The toxoplasmosis data was rerun using a quasibinomial model, as shown in Table 3.10. It estimated σ^2 as 1.94, the Pearson residual overdispersion estimate from Table 3.7. Comparing the results with the standard binomial GLM in Table 3.8 we see that:

- The estimated coefficient vector $\hat{\beta}$ is the same.
- The estimated standard errors are multipled by $\sqrt{1.94} = 1.39$.
- The estimated t-values are divided by 1.39.

This last item results in a two-sided p-value for the cubic coefficient of 0.023, compared with 0.00081 previously.

Table 3.10: Quasibinomial logistic regression for toxoplasmosis data: `glm(formula` $= p \sim$ `poly(r,3), family` $=$ `quasibinomial, weights` $= n$`)`.

Coefficients	Estimate	St. error	t-value	Pr(> $\lvert t \rvert$)
(Intercept)	.0243	.1072	.23	.822
`poly(`r`,3)1`	−.0861	.6390	−.13	.894
`poly(`r`,3)2`	−.1927	.6511	−.30	.769
`poly(`r`,3)3`	1.3787	.5732	2.41	.023 ***

3.8 Double exponential families [Efron (1986), *JASA* 709–721]

The quasilikelihood analysis of the toxoplasmosis data proceeded *as if* the observed proportions p_i were obtained from a one-parameter exponential family with expectation π_i and variance $\sigma^2\pi_i(1-\pi_i)/n$. There is no such family, but it turns out we can come close using the *double exponential family* construction.

Forgetting about GLMs for now, suppose we have a single random sample $y_1, y_2, \ldots, y_n$ from a one-parameter exponential family $g_\mu(y)$ having expectation μ and variance function $V(\mu)$. The average $\bar{y}$ is then a sufficient statistic, with density say

$$g_{\mu,n}(\bar{y}) = e^{n(\eta\bar{y} - \psi(\eta))} g_{0,n}(\bar{y})$$

as in Section 1.3, and expectation and variance

$$\bar{y} \sim \left(\mu, \frac{V(\mu)}{n}\right). \tag{3.19}$$

Hoeffding's formula, Section 1.8, expresses $g_{\mu,n}(\bar{y})$ in terms of deviance,

$$g_{\mu,n}(\bar{y}) = g_{\bar{y},n}(\bar{y})e^{-nD(\bar{y},\mu)/2},$$

with $D(\mu_1, \mu_2)$ the deviance function for $n = 1$.

The double exponential family corresponding to $g_{\mu,n}(\bar{y})$ (3.19) is the two-parameter family

$$\boxed{f_{\mu,\theta,n}(\bar{y}) = C\,\theta^{1/2} g_{\bar{y},n}(\bar{y})e^{-n\theta D(\bar{y},\mu)/2},} \tag{3.20}$$

μ in the interval of allowable expectations for $g_{\mu,n}(\bar{y})$, and $\theta > 0$. An important point is that the carrier measure $m(d\bar{y})$ for $f_{\mu,\theta,n}(\bar{y})$, suppressed in our notation, is the same as that for $g_{\mu,n}(\bar{y})$. This is crucial for discrete distributions like the Poisson where the support stays the same — counting measure on $0, 1, 2, \ldots$ — for all choices of θ.

What follows is a list of salient facts concerning $f_{\mu,\theta,n}(\bar{y})$, as verified in Efron (1986).

Fact 1 The constant $C = C(\mu, \theta, n)$ that makes $f_{\mu,\theta,n}(\bar{y})$ integrate to 1 is close to 1.0. Standard Edgeworth calculations give

$$C(\mu, \theta, n) \doteq 1 + \frac{1}{n}\left(\frac{1-\theta}{\theta}\,\frac{9\delta_\mu - 15\gamma_\mu^2}{72}\right) + O\left(\frac{1}{n^2}\right),$$

with γ_u and δ_u the skewness and kurtosis of $g_{\mu,1}(\bar{y})$. Taking $C = 1$ in (3.20) is convenient, and usually accurate enough for applications.

Fact 2 Formula (3.20) can also be written as

$$f_{\mu,\theta,n}(\bar{y}) = C\,\theta^{1/2} g_{\mu,n}(\bar{y})^{\theta} g_{\bar{y},n}(\bar{y})^{1-\theta},$$

which says that $\log f_{\mu,\theta,n}(\bar{y})$ is essentially a linear combination of $\log g_{\mu,n}(\bar{y})$ and $\log g_{\bar{y},n}(\bar{y})$.

Homework 3.22. Verify Fact 2.

Fact 3 The expectation and variance of $\bar{y} \sim f_{\mu,\theta,n}$ are, to excellent approximations,

$$\bar{y} \dot{\sim} \left(\mu, \frac{V(\mu)}{n\theta}\right),$$

with errors of order $1/n^2$ for both terms. Comparison with (3.16) shows that $1/\theta$ measures dispersion,

$$\sigma^2 = 1/\theta.$$

Homework 3.23. Suppose $g_{\mu,n}(\bar{y})$ represents a normal mean, $\bar{y} \sim \mathcal{N}(\mu, 1/n)$. Show that $f_{\mu,\theta,n}(\bar{y})$ has $\bar{y} \sim \mathcal{N}(\mu, 1/(n\theta))$.

Fact 4 $f_{\mu,\theta,n}(\bar{y})$ is a two-parameter exponential family having natural parameter "η" equal $n(\theta\eta, \theta)$, and sufficient vector "y" equal $(\bar{y}, -\bar{\eta}\bar{\psi})$, where $\bar{\eta} = \dot{\psi}^{-1}(\bar{y})$. Moreover, with θ and n fixed, $f_{\mu,\theta,n}(\bar{y})$ is a one-parameter exponential family with natural parameter $n\theta\eta$ and sufficient statistic $\bar{y}$; with μ and n fixed, $f_{\mu,\theta,n}(\bar{y})$ is a one-parameter exponential family with natural parameter θ and sufficient statistic $-nD(\bar{y}, \mu)/2$.

Homework 3.24. Verify Fact 4.

Together, Facts 3 and 4 say that $f_{\mu,\theta,n}(\bar{y})$, with θ fixed, is a one-parameter exponential family having expectation and variance nearly μ and $V_\mu/(n\theta)$, respectively. This is just what was required for the notional quasilikelihood families.

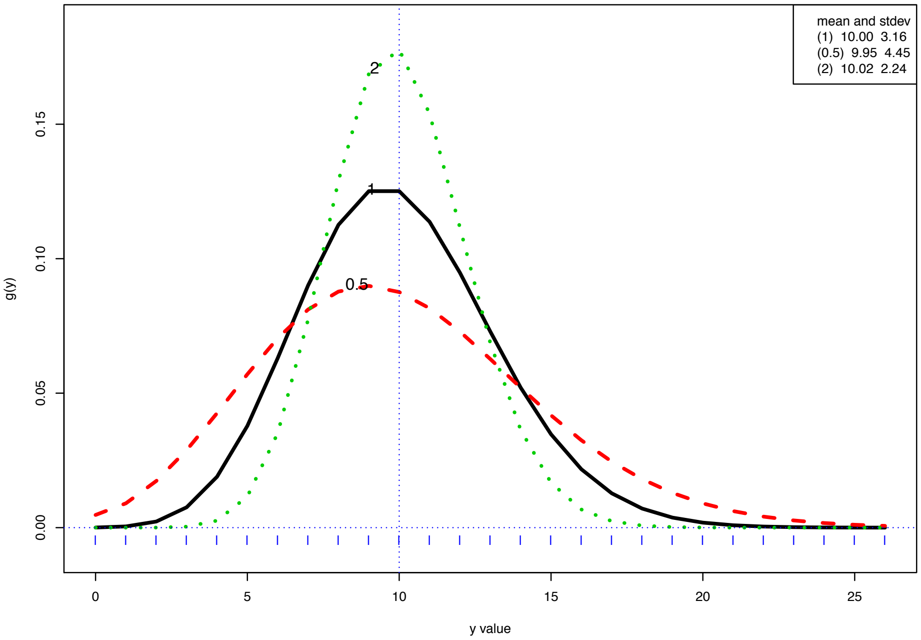


Figure 3.7: Double Poisson densities for $\mu = 10$ and $\theta = 1$ (solid), $= 1/2$ (dashed), or $= 2$ (dotted).

Figure 3.7 illustrates the double Poisson distribution: we have taken[3] $n = 1$, $\mu = 10$, and $\theta = 1, 1/2$, or 2 (using $C(\mu, \theta, n)$ from Fact 1). The case $\theta = 1$, which is the standard Poisson distribution, has $\mu = 10$ and $V_\mu = \sqrt{10} = 3.16$. As claimed, the variance doubles for $\theta = 1/2$ and halves for $\theta = 2$, while μ stays near 10. All three distributions are supported on $0, 1, 2, \dots$.

[3] Because the Poisson family is closed under convolutions, the double Poisson family turns out to be essentially the same for any choice of n.

Homework 3.25.

(a) For the Poisson family ($n = 1$) show that (3.20), with $C = 1$, gives

$$f_{\mu,\theta}(y) = \theta^{1/2} e^{-\theta\mu} \left(\frac{e^{-y} y^y}{y!}\right) \left(\frac{e\mu}{y}\right)^{\theta y}$$

(y is $\bar{y}$ here).

(b) Compute $f_{\mu,\theta}(y)$ for $\theta = 1/3$ and $\theta = 3$, and numerically calculate the expectations and variances.

(c) Use Fact 1 to give an expression for $C(\mu, \theta, n)$.

(d) What are the expectations and variances now using $C(\mu, \theta, n)$ instead of $C = 1$?

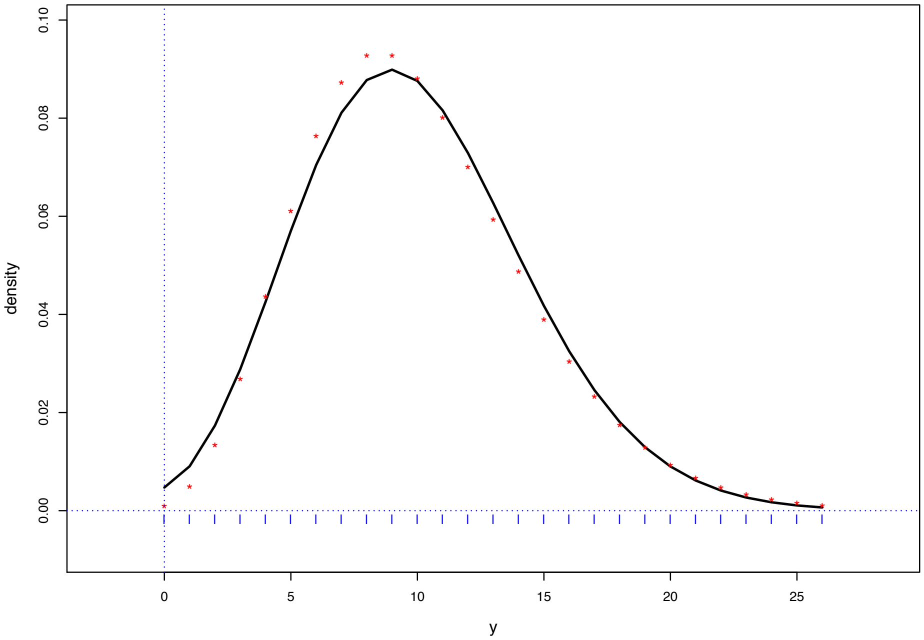


Figure 3.8: Comparison of double Poisson density, $\mu = 10$, $\theta = 0.5$ (solid), with negative binomial density, $\mu = 10$, variance $= 20$ (points).

Count statistics that appear to be overdispersed Poissons are often modeled by negative binomial distributions, Section 1.4. Figure 3.8 compares $f_{10,.5}(y)$ with the negative binomial density having expectation 10 and variance 20, showing a striking similarity. Negative binomials form a one-parameter family in θ, but the auxiliary parameter k cannot be incorporated into a two-parameter exponential family.

The fact that $\bar{y} \sim f_{\mu,\theta,n}$ has expectation and variance approximately μ and $V_\mu/(n\theta)$ suggests that the density $f_{\mu,\theta,n}(\bar{y})$ is similar to $g_{\mu,n\theta}(\bar{y})$. Choosing $\theta = 1/2$, say, effectively reduces the sample size in the original family from n to $n/2$. This was exactly true in the normal case of Homework 3.23.

Fact 5 For any interval I of the real line,

$$\int_I f_{\mu,\theta,n}(\bar{y})\ m_n(d\bar{y}) = \int_I g_{\mu,n\theta}(\bar{y})\ m_{n\theta}(d\bar{y}) + O\left(\frac{1}{n}\right).$$

Here m_n and $m_{n\theta}$ are the carrying measures in the original family, for sample sizes n and $n\theta$.

For the binomial family $p \sim \text{Bi}(n,\pi)/n$ — where we are thinking of $\bar{y} = p$ as the average of n Bernoulli variates $y_i \sim \text{Bi}(1,\pi)$ — $m_n(p)$ is counting measure on $0, {}^1\!/_n, {}^2\!/_n, \dots, 1$, while $m_{n\theta}(p)$ is counting measure on $0, {}^1\!/_{n\theta}, {}^2\!/_{n\theta}, \dots, 1$. This assumes $n\theta$ is an integer, and shows the limitations of Fact 5 in discrete families.

Homework 3.26. In the binomial case, numerically compare the cumulative distribution function of $f_{\mu,\theta,n}(p)$, $(\mu,\theta,n) = (0.4, 0.5, 16)$, with that of $g_{\mu,n}(p)$, $(\mu,n) = (0.4, 8)$.

Homework 3.27. What would be the comparisons suggested by Fact 5 for the Poisson distributions in Figure 3.7?

The double family constant $C(\mu,\theta,n)$ can be calculated explicitly for the gamma family $\bar{y} \sim \mu G_n/n$,

$$g_{\mu,n}(\bar{y}) = \frac{\bar{y}^{n-1}e^{-(n\bar{y}/\mu)}}{(\mu/n)^n\Gamma(n)} \qquad (\bar{y} > 0).$$

Homework 3.28.

(a) Show that

$$f_{\mu,\theta,n}(\bar{y}) = \frac{g_{\mu,n\theta}(\bar{y})}{C(\mu,\theta,n)},$$

where

$$C(\mu,\theta,n) = \theta^{-1/2}\ \frac{\Gamma(n)}{(n/e)^n} \Bigg/ \frac{\Gamma(n\theta)}{(n\theta/e)^{n\theta}}.$$

(b) Using Stirling's formula

$$\Gamma(z) \doteq \sqrt{2\pi}\, z^{z-1/2} \exp\left(-z + {}^1\!/_{12z}\right),$$

compare $C(\mu,\theta,n)$ with the approximation of Fact 1.

Differentiating the log likelihood function $l_{\mu,\theta,n}(\bar{y}) = \log f_{\mu,\theta,n}(\bar{y})$,

$$l_{\mu,\theta,n}(\bar{y}) = -n\theta\ \frac{D(\bar{y},\mu)}{2} + \frac{1}{2}\log\theta + \log g_{\bar{y},\mu}(\bar{y}),$$

and using $\partial D(\bar{y},\mu)/\partial\mu = 2(\mu - \bar{y})/V_\mu$, gives the next fact.

Fact 6 The score functions for $f_{\mu,\theta,n}(\bar{y})$ are

$$\frac{\partial l_{\mu,\theta,n}(\bar{y})}{\partial\mu} = \frac{\bar{y}-\mu}{V_\mu/(n\theta)} \quad \text{and} \quad \frac{\partial l_{\mu,\theta,n}(\bar{y})}{\partial\theta} = \frac{1}{2\theta} - \frac{nD(\bar{y},\mu)}{2}.$$

Double family GLMs Suppose we have a generalized regression setup, observations

$$\bar{y}_i \overset{\text{ind}}{\sim} f_{\mu_i,\theta,n_i} \qquad \text{for } i = 1, 2, \dots, N, \tag{3.21}$$

with the GLM model $\boldsymbol{\eta}(\beta) = X\beta$ giving $\boldsymbol{\mu}(\beta) = (\cdots \mu_i = \dot{\psi}(\eta_i(\beta)) \cdots)'$. The score functions for the full data set $\boldsymbol{y} = (\bar{y}_1, \bar{y}_2, \dots, \bar{y}_N)'$ are

$$\frac{\partial}{\partial \beta}\, l_{\beta,\theta}(\boldsymbol{y}) = \theta X' \operatorname{diag}(\boldsymbol{n})\, (\boldsymbol{y} - \boldsymbol{\mu}(\beta))\,,$$

$\operatorname{diag}(\boldsymbol{n})$ the diagonal matrix with entries n_i, and

$$\frac{\partial}{\partial \theta}\, l_{\beta,\theta}(\boldsymbol{y}) = \frac{N}{2\theta} - \sum_{i=1}^{N} \frac{n_i D\left(\bar{y}_i, \mu_i(\beta)\right)}{2}.$$

Homework 3.29. Verify the score functions.

We see that the MLE $\hat{\beta}$ does not depend on θ, which only enters the β score function as a constant multiple. The MLE for $\hat{\theta}$ is

$$\hat{\theta} = N \Bigg/ \sum_{i=1}^{N} n_i D\left(\bar{y}_i, \mu_i\left(\hat{\beta}\right)\right).$$

Homework 3.30. How does this estimate relate to the overdispersion estimates $\hat{\sigma}^2$ for the toxoplasmosis data of Section 3.7?

I ran a more ambitious GLM for the toxoplasmosis data, where θ_i as well as p_i was modeled. It used the double binomial model $p_i \sim f_{\pi_i,\theta_i,n_i}$, $i = 1, 2, \dots, 34$. Here p_i and π_i, the observed and true proportion positive in city i, play the roles of y_i and μ_i in (3.20).

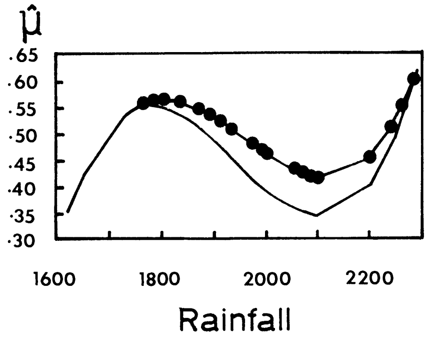
My model let π_i be a cubic polynomial function of rainfall, as in Section 3.7; θ_i was modeled as a function of n_i, the number of people sampled in city i. Let

$$\tilde{n}_i = \frac{n_i - \bar{n}}{\text{sd}_n},$$

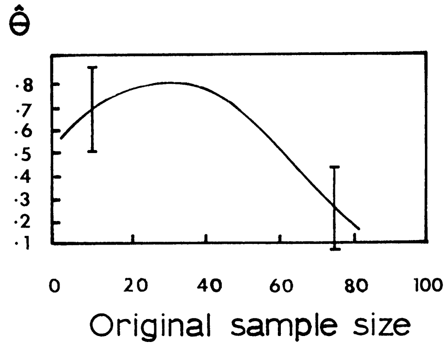
$\bar{n}$ and sd_i the mean and standard deviation of the n_i values. I took $\theta_i = 1.25 \cdot (1 + e^{-\lambda_i})$, where $\lambda_i = \gamma_0 + \gamma_1 \tilde{n}_i + \gamma_2 \tilde{n}_i^2$. This allowed the θ_i to range from 1.25 (mild underdispersion) all the way down to zero. All together the model had seven parameters, four for the cubic rainfall regression and three for the θ regression. The seven-parameter MLE was found using the R function nonlinear maximizer `nlm`.

Homework 3.31. What was the function I minimized?

The resulting cubic regression of π_i as a function of r_i (solid curve) was somewhat more extreme than the original GLM in Table 3.8, the latter shown as the dotted curve here.



Perhaps more interesting was the fitted regression for the dispersion parameter $\hat{\theta}_i$ as a function of number sampled n_i. It peaked at about $\hat{\theta}_i = 0.8$ at $n_i = 30$, declining to 0.2 for $n_i = 70$. Rather unintuitively, overdispersion *increased* in the largest samples.



Exponential Families in Theory and Practice

Bradley Efron
Stanford University

Part 4

Curved Exponential Families, Empirical Bayes, Missing Data, and the EM Algorithm

Not every problem can be reduced to exponential family form. This chapter concerns situations that start out as exponential families but then suffer from some complicating factor that makes them less tractable. Exponential family ideas can still be useful in such problems. In terms of the three-circles diagram in Figure 1 of the Introduction, page iii, we are venturing out of the exponential families circle into the realm of General Theory, in hopes of reducing — if not removing — our dependence on abstract asymptotic arguments.

4.1 Curved exponential families: Definitions and first results

We begin with a p-parameter exponential family $\mathcal{G}$, as in Part 2,

$$\mathcal{G} = \left\{ g_\eta(y) = e^{\eta' y - \psi(\eta)} g_0(y) \text{ for } \eta \in A, y \in \mathcal{Y} \right\}.$$

In some situations though, we may know that η is restricted to lie in a q-parameter curved subspace of A, say

$$\eta = \eta_\theta \text{ for } \theta \in \Theta \subset \mathcal{R}^q,$$

$q < p$. The mapping $\eta = \eta_\theta$ from Θ into A will be assumed to have continuous second derivatives. We can express the q-parameter subfamily of densities as

$$\mathcal{F} = \left\{ f_\theta(y) = g_{\eta_\theta}(y) \text{ for } \theta \in \Theta \right\},$$

so

$$f_\theta(y) = e^{\eta_\theta' y - \psi(\eta_\theta)} g_0(y).$$

If the mapping $\Theta \to A$ is linear, say

$$\eta = X\theta,$$

then

$$f_\theta(y) = e^{\theta' X' y - \psi(X\theta)} f_0(y)$$

is a q-parameter exponential family with natural parameter vector θ, as in the generalized linear models of Part 3. Here we will be interested in *curved exponential familes*, where η_θ is a nonlinear function of θ.

Some convenient notation We will write μ_θ for $\mu(\eta_\theta)$, and likewise V_θ for $V_{\eta_\theta} = \mathrm{Cov}_\theta(y)$, ψ_θ for ψ_{η_θ}, etc. Derivatives with respect to θ will be indicated by dot notation,

$$\underset{p\times q}{\dot\eta_\theta} = \left(\frac{\partial \eta_i}{\partial \theta_j}\right)_{\eta=\eta_\theta}, \qquad \underset{p\times q}{\dot\mu_\theta} = \left(\frac{\partial \mu_i}{\partial \theta_j}\right)_{\mu=\mu_\theta} = V_\theta \dot\eta_\theta,$$

and

$$\ddot\eta_\theta = \left(\frac{\partial^2 \eta_i}{\partial \theta_j \partial \theta_k}\right)_{\eta=\eta_\theta},$$

a $p \times p \times q$ array. Family $\mathcal{F}$ has representations in both the η and μ ("A" and "B") spaces,

$$\mathcal{F}_A = \{\eta_\theta, \theta \in \Theta\} \quad \text{and} \quad \mathcal{F}_B = \{\mu_\theta, \theta \in \Theta\},$$

both $\mathcal{F}_A$ and $\mathcal{F}_B$ being q-dimensional manifolds in $\mathcal{R}^p$, usually curved.

The log likelihood function The log likelihood function is

$$l_\theta(y) = \log f_\theta(y) = \eta_\theta' y - \psi(\eta_\theta).$$

Taking derivatives with respect to θ gives the *score function*

$$\underset{q\times 1}{\dot{l}_\theta(y)} = \left(\frac{\partial l_\theta}{\partial \theta_j}\right) = \dot{\eta}_\theta'(y - \mu_\theta), \tag{4.1}$$

where we have used

$$\dot{\psi}_\theta = \dot{\eta}_\theta' \mu_\theta;$$

(4.1) gives the *Fisher information matrix*

$$\underset{p\times p}{i_\theta} = E_\theta\left\{\dot{l}_\theta(y)\dot{l}_\theta(y)'\right\} = \dot{\eta}_\theta' V_\theta \dot{\eta}_\theta. \tag{4.2}$$

Taking derivatives again yields a simple but important result:

Lemma 4 (Second Derivative Lemma)**.** *Minus the second derivative matrix of the log likelihood is*

$$-\underset{q\times q}{\ddot{l}_\theta(y)} = \left(-\frac{\partial^2 l_\theta}{\partial \theta_j \partial \theta_k}\right) = i_\theta - \ddot{\eta}_\theta'(y - \mu_\theta). \tag{4.3}$$

Here $\ddot{\eta}_\theta'(y - \mu_\theta)$ has (j,k)th element

$$\sum_{i=1}^{p} \frac{\partial^2 \eta_i}{\partial \theta_j \partial \theta_k}(y - \mu_\theta)_i.$$

Homework 4.1. (a) Verify the formulas for $\dot{l}_\theta$, $\dot{\psi}_\theta$, i_θ, and $-\ddot{l}_\theta$.

(b) Show that $\dot{\mu}_\theta = V_\theta \dot{\eta}_\theta$ and so

$$i_\theta = \dot{\eta}_\theta' i_\theta = \dot{\eta}_\theta' V_\theta \dot{\eta}_\theta = 0. \tag{4.4}$$

(In the one-dimensional case $q = 1$, (4.4) shows that the curves η_θ and μ_θ through p-space have directional derivates within 90° of each other.)

4.2 Two pictures of the MLE

4.2.1 Fisher's picture

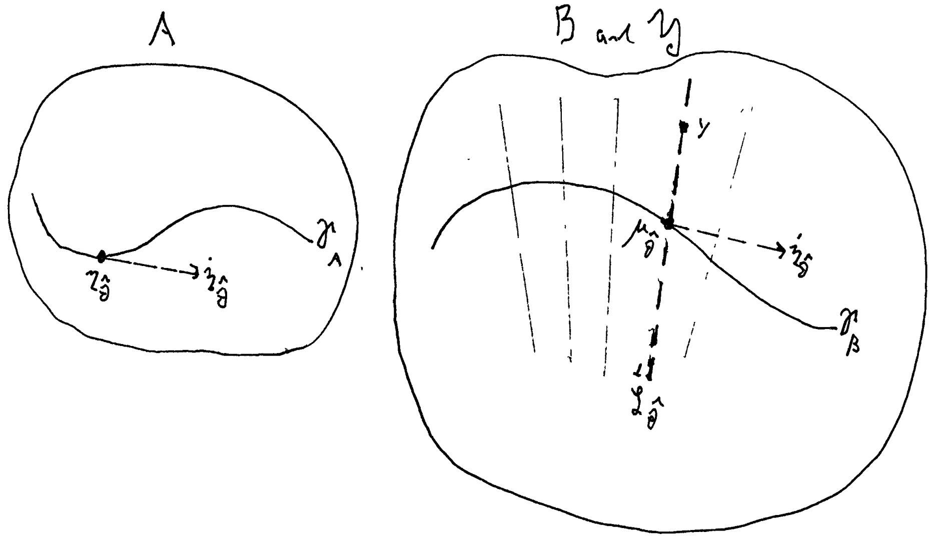


Figure 4.1: Fisher's picture.

The maximum likelihood estimate $\hat\theta$ satisfies $\dot{l}_{\hat\theta}(y) = 0$, which according to (4.1) is the local orthogonality condition

$$\dot\eta'_{\hat\theta}(y - \mu_{\hat\theta}) = 0. \tag{4.5}$$

Here 0 represents a vector of q zeros. We obtain a pleasing geometric visualization of MLE estimation from Figure 4.1. Notice that:

- The expectation vector $\mu_{\hat\theta}$ corresponding to $\hat\theta$ is obtained by projecting the data vector y onto $\mathcal{F}_B$ orthogonally to $\dot\eta_{\hat\theta}$, the tangent space to $\mathcal{F}_A$ at $\eta_{\hat\theta}$.
- If η_θ is linear in θ, say
$$\eta_\theta = a + b\theta,$$
with a $p\times 1$ and b $p \times q$ (that is, if $\mathcal{F}_A$ is a flat space), then $\dot\eta_\theta = b$ for all θ, and the projection direction is always orthogonal to b,
$$b'(y - \mu_{\hat\theta}) = 0.$$
- Otherwise, the projection orthogonals $\dot\eta_{\hat\theta}$ change with $\hat\theta$.
- The set of y vectors having MLE equaling some particular value $\hat\theta$ is $(p-q)$-dimensional flat space in $\mathcal{Y}$, passing through $\mu_{\hat\theta}$ orthogonally to $\dot\eta_{\hat\theta}$, say
$$\overset{\perp}{\mathcal{L}}_{\hat\theta} = \left\{ y : \dot\eta'_{\hat\theta}(y - \mu_{\hat\theta}) = 0 \right\}. \tag{4.6}$$

4.2.2 Hoeffding's picture

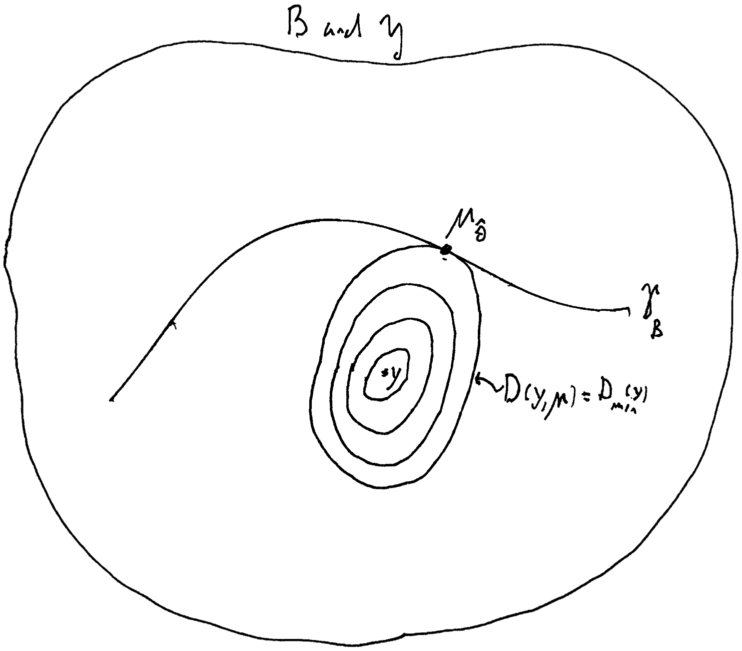


Figure 4.2: Hoeffding's picture.

Hoeffding's formula, Section 2.7 (now indexing $\mathcal{G}$ by μ rather than η), is

$$\begin{aligned} f_\theta(y) &= g_{\mu_\theta}(y) \\ &= g_y(y) e^{-D(y,\mu_\theta)/2}. \end{aligned}$$

Therefore the MLE $\hat\theta$ must minimize the deviance $D(y, \mu_\theta)$ between y and a point μ_θ on $\mathcal{F}_B$,

$$\hat\theta : D(y, \mu_{\hat\theta}) = \min_{\theta\in\Theta} D(y, \mu_\theta) \equiv D_{\text{min}}(y).$$

Another way to say this: as d is increased from 0, the level curves $D(y, \mu) = d$ touch $\mathcal{F}_B$ for the first time when $d = D_{\text{min}}$. Figure 4.2 is the visualization.

Homework 4.2. A normal theory linear model $y \sim \mathcal{N}_p(\mu, I)$ has

$$\eta_\theta = \mu_\theta = X\theta$$

for X a known $p \times q$ matrix. What do the two pictures look like? What are their usual names?

Homework 4.3. Show that the level curves of constant deviance,

$$\mathcal{C}_d = \{\mu : D(y, \mu) = d\},$$

become ellipsoidal as $d \downarrow 0$. What determines the shape of the ellipsoids?

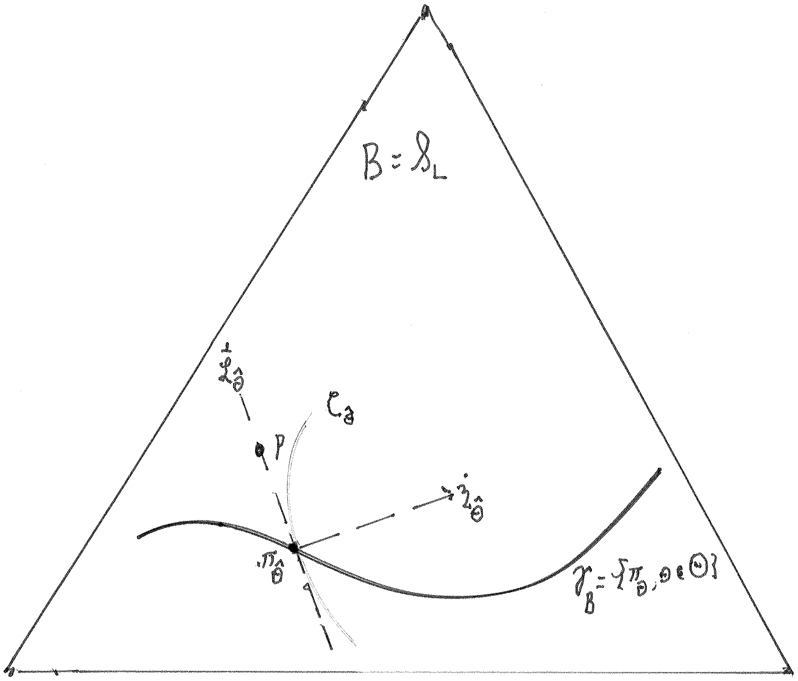

$B = \mathcal{S}_L$
$\mathcal{L}^{\perp}_{\hat\theta}$
p
$\mathcal{C}_{\hat\theta}$
$\dot\eta_{\hat\theta}$
$\pi_{\hat\theta}$
$\mathcal{F}_B = \{\pi_\theta, \theta \in \Theta\}$

Figure 4.3: See text.

Fisher never drew "Fisher's picture" but he did carry out an incisive analysis of the multinomial case

$$p \sim \text{Mult}_L(n, \pi_\theta)/n$$

with $q = 1$, that is, with θ one-dimensional. (As in Section 2.9, L or $L - 1$ is the dimension p, p is now y, and π_θ is μ_θ.)

Homework 4.4. Show that we can take $\dot\eta_{\theta j} = \dot\pi_{\theta j}/\pi_{\theta j}$.

Fisher argued that the MLE $\hat\theta$ was superior to other possible smooth estimations $\tilde\theta = t(p)$, such as minimum chi-squared. He made three points:

[1] *Consistency* In order for $\tilde\theta = t(p)$ to be a consistent estimator of θ as $n \to \infty$, it must satisfy

$$t(\pi_\theta) = \theta \qquad \text{("Fisher consistency").}$$

Homework 4.5. What was Fisher's argument?

[2] *Efficiency* In order to achieve the Fisher information bound for estimating θ, a Fisher consistent estimator must have $\mathcal{C}(\hat\theta)$, the level surface $\{p : t(p) = \hat\theta\}$, and must intersect $\mathcal{F}_B$ at π_θ *orthogonally* to $\dot\eta_\theta$.

Homework 4.6. Verify [2] above (not worrying about rigor).

[3] *Second-order efficiency* If $\mathcal{F}_B$ represents a one-parameter exponential family ($\eta_\theta = a+b\theta$) then the MLE $\hat{\theta}$ is a sufficient statistic. Even if not, $\hat{\theta}$ loses less information than any competitor $\tilde{\theta}$. In other words, the flat level surfaces of MLE estimation are superior to any curved one of the type suggested in Figure 4.3. This was a controversial claim, but we will see it supported later.

4.3 Repeated sampling and the influence function of the MLE

Suppose $y_1, y_2, \ldots, y_n$ is an i.i.d. sample from some member of a curved family $\mathcal{F} = \{f_\theta(y) = g_{\eta_\theta}(y)\}$. Then $\bar{y}$ is a sufficient statistic in $\mathcal{F}$, since it is so in the larger family $\mathcal{G}$. The family of distributions of $\bar{y}$, $\mathcal{F}_n$, is the curved exponential family

$$\mathcal{F}_n = \left\{f_{\theta,n}(\bar{y}) = g_{\eta_{\theta,n}}(\bar{y}) = e^{n(\eta_\theta'\bar{y}-\psi_\theta)}, \theta \in \Theta\right\}.$$

$\mathcal{F}_n$ is essentially the same family as $\mathcal{F}$: B and $\mathcal{F}_B$ stay the same, while $A_n = nA$, $\mathcal{F}_{A,n} = n\mathcal{F}_A$, and $\dot{\eta}_{\theta,n} = n\dot{\eta}_\theta$, as in Section 2.4. Fisher's and Hoeffding's pictures stay exactly as shown, with $\bar{y}$ replacing y. The covariance matrix of $\bar{y}$, $V_{\theta,n}$, equals V_θ/n. This means that typically $\bar{y}$ is closer than y to $\mathcal{F}_B$ by a factor of $1/\sqrt{n}$, this applying to both pictures. Asymptotic calculations in curved exponential families are greatly simplified by the unchanging geometry.

The log likelihood $l_{\theta,n}(\bar{y})$ in $\mathcal{F}_A$ is

$$l_{\theta,n}(\bar{y}) = n(\eta_\theta'\bar{y} - \psi_\theta)$$

(with the case $n = 1$ denoted $l_\theta(y)$ as before). The first and second derivatives with respect to θ are

$$\begin{aligned} \dot{l}_{\theta,n}(\bar{y}) &= n\dot{\eta}_\theta'(\bar{y} - \mu_\theta) \\ \text{and} \quad -\ddot{l}_{\theta,n}(\bar{y}) &= n\left[i_\theta - \ddot{\eta}_\theta'(\bar{y} - \mu_\theta)\right], \end{aligned} \tag{4.7}$$

according to (4.1) and (4.3) in Section 4.1.

Looking at Fisher's picture, we can think of the MLE $\hat{\theta}$ as a function of $\bar{y}$,

$$\hat{\theta} = t(\bar{y}),$$

$t(\cdot)$ being a mapping from $\mathcal{R}^p$ into $\mathcal{R}^q$. A small change $\bar{y} \to \bar{y} + d\bar{y}$ produces a small change in the MLE, $\hat{\theta} \to \hat{\theta} + d\hat{\theta}$, according to the *influence function* (or derivative matrix)

$$\underset{q\times p}{dt/d\bar{y}} = \left(\frac{\partial\hat{\theta}_i}{\partial\bar{y}_j}\right) \equiv \frac{d\hat{\theta}}{d\bar{y}},$$

approximately

$$d\hat{\theta} \doteq \frac{d\hat{\theta}}{d\bar{y}}\, d\bar{y}.$$

The MLE influence function takes a simple but useful form:

Lemma 5 (Influence Lemma)**.** *The influence function of the MLE is*

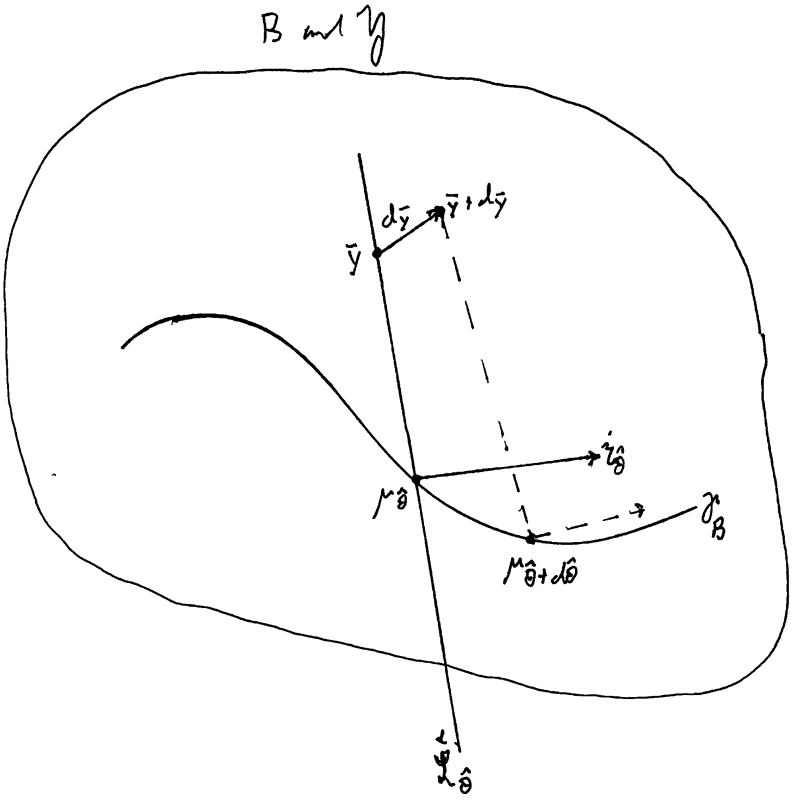
$$\underset{q\times p}{d\hat\theta/d\bar y} = \underset{q\times q}{\left[-\ddot l_{\hat\theta}(\bar y)\right]^{-1}} \underset{q\times p}{\dot\eta'_{\hat\theta}} \tag{4.8}$$

where

$$-\ddot l_{\hat\theta}(\bar y) = i_{\hat\theta} - \ddot\eta'_{\hat\theta}(\bar y - \mu_{\hat\theta}).$$

(Note (4.8) uses $-\ddot l_{\hat\theta}(\bar y)$, not $-\ddot l_{\hat\theta,n}(\bar y)$; it can also be expressed as $[-\ddot l_{\hat\theta,n}(\bar y)]^{-1}(n\dot\eta'_{\hat\theta})$.)

Homework 4.7. Verify Lemma 5. *Hint*: $(\dot\eta_{\hat\theta+d\hat\theta})'(\bar y + d\bar y - \mu_{\hat\theta+d\hat\theta}) = 0$.



$\mathcal{B}$ and $\mathcal{Y}$
$d\bar y$
$\bar y + d\bar y$
$\bar y$
$\dot\eta_{\hat\theta}$
$\mu_{\hat\theta}$
$\mathcal{F}_{\mathcal{B}}$
$\mu_{\hat\theta+d\hat\theta}$
$\overset{\perp}{\mathcal{L}}_{\hat\theta}$

Figure 4.4: See text.

4.4 Variance calculations for the MLE

Corresponding to a given true value θ in $\mathcal{F}$ is $\overset{\perp}{\mathcal{L}}_\theta = \{\bar y : \dot\eta'_\theta(\bar y - \mu_\theta) = 0\}$, the set of observations $\bar y$ such that the MLE $t(\bar y) = \theta$. In particular, $\bar y = \mu_\theta$ is on $\overset{\perp}{\mathcal{L}}_\theta$, i.e.,

$$t(\mu_\theta) = \theta;$$

so the MLE is *Fisher consistent.* Since $\bar{y}$ converges almost surely to μ_θ as $n \to \infty$, and $t(\bar{y})$ is a nicely continuous function, Fisher consistency implies ordinary consistency: $\hat{\theta} \to \theta$ a.s.

The quantity

$$I_n(\bar{y}) = -\ddot{l}_{\theta,n}(\bar{y}) = n\left[i_{\hat{\theta}} - \ddot{\eta}'_{\hat{\theta}}(\bar{y} - \mu_{\hat{\theta}})\right] \tag{4.9}$$

is called the *observed Fisher information* for θ, with the case $n = 1$ denoted $I(y)$. For reasons that will be discussed in Section 4.6, Fisher claimed that the covariance of the MLE is better assessed by $I_n(\bar{y})^{-1}$ rather than $(ni_{\hat{\theta}})^{-1}$. Notice that $I(\mu_\theta) = i_\theta$, so that $I(\bar{y})$ is itself Fisher consistent as an estimate of i_θ,

$$-\ddot{l}_\theta(\mu_\theta) = i_\theta.$$

The variance matrix of $\hat{\theta}$ can be approximated using Fisher consistency and the MLE influence function (4.8):

$$\begin{aligned}\hat{\theta} = t(\bar{y}) = t(\mu_\theta + \bar{y} - \mu_\theta) &\doteq t(\mu_\theta) + \left.\frac{dt}{d\bar{y}}\right|_{\mu_\theta}(\bar{y} - \mu_\theta)\\ &\doteq \theta + \left[-\ddot{l}_\theta(\mu_\theta)\right]^{-1}\dot{\eta}'_\theta(\bar{y} - \mu_\theta)\\ &\doteq \theta + i_\theta^{-1}\dot{\eta}'_\theta(\bar{y} - \mu_\theta),\end{aligned}$$

a Taylor series statement of how $\hat{\theta}$ behaves as a function of $\bar{y}$ when $n \to \infty$. This yields the covariance approximation

$$\mathrm{Cov}_\theta\left(\hat{\theta}\right) \doteq i_\theta^{-1}\dot{\eta}'_\theta\,\frac{V_\theta}{n}\,\dot{\eta}_\theta i_\theta^{-1} = i_\theta^{-1}/n,$$

the Cramér–Rao lower bound for estimating θ (Section 2.5), where we have used $\dot{\eta}'_\theta V_\theta \dot{\eta}_\theta = i_\theta$. Except in special cases, $\mathrm{Cov}_\theta(\hat{\theta})$ will exceed i_θ^{-1}; nevertheless, $i_{\hat{\theta}}^{-1}$ is usually a reasonable estimate of $\mathrm{Cov}(\hat{\theta})$, as the derivation above suggests.

Variance if our model is wrong Suppose we were wrong in assuming that $\mu \in \mathcal{F}_B$, and that actually μ lies somewhere else in B, as in Figure 4.5. Let

$$\tilde{\theta} = t(\mu),$$

so $\mu_{\tilde{\theta}}$ is the closest point to μ on $\mathcal{F}_B$ in terms of deviance distance (Hoeffding's picture). This means that μ, the true expectation of $\bar{y}$, lies on $\overset{\perp}{\mathcal{L}}_{\tilde{\theta}}$, passing through $\mathcal{F}_B$ at $\mu_{\tilde{\theta}}$, orthogonally to $\dot{\eta}_{\tilde{\theta}}$.

As $n \to \infty$, $\bar{y}$ goes to μ and

$$\hat{\theta} \longrightarrow t(\mu) = \tilde{\theta}.$$

We can still use the MLE influence function

$$\left.\frac{d\hat{\theta}}{d\bar{y}}\right|_{\bar{y}=\mu} = \left[-\ddot{l}_{\tilde{\theta}}(\mu)\right]^{-1}\dot{\eta}'_{\tilde{\theta}} = \left[i_{\tilde{\theta}} - \ddot{\eta}_{\tilde{\theta}}(\mu - \mu_{\tilde{\theta}})\right]^{-1}\dot{\eta}_{\tilde{\theta}}$$

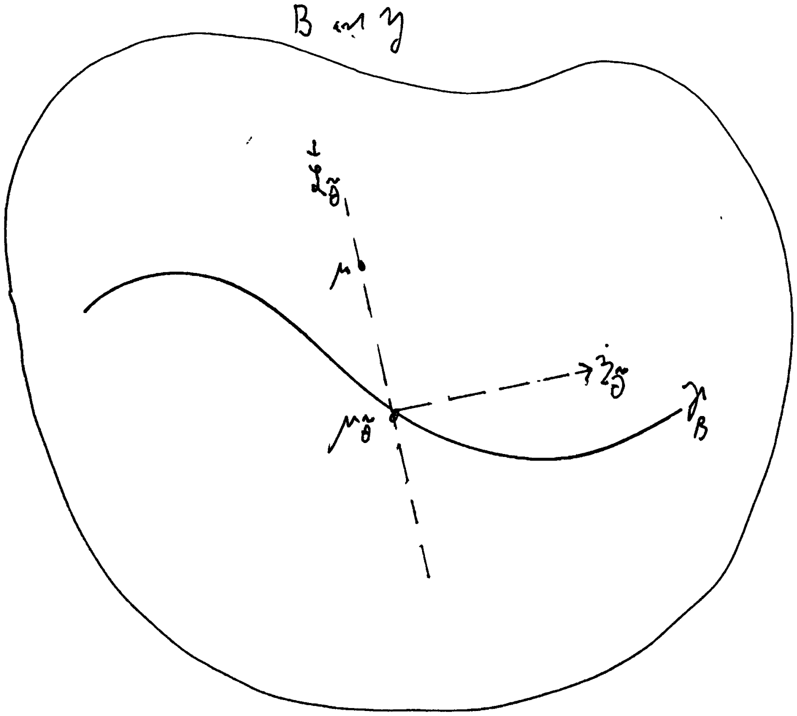

B ⇝ y
$\overset{\perp}{\dot{l}}_{\tilde\theta}$
μ
$\dot\eta_{\tilde\theta}$
$\mu_{\tilde\theta}$
γ_B

Figure 4.5: See text.

to approximate the covariance matrix $\mathrm{Var}_\mu\{\hat\theta\}$,

$$\mathrm{Var}_\mu\left\{\hat\theta\right\} \doteq \mathrm{Var}_\mu\left\{\tilde\theta + \left.\frac{d\hat\theta}{d\bar y}\right|_\mu (\bar y - \mu)\right\} = \left.\frac{d\hat\theta}{d\bar y}\right|_\mu^{\prime} \frac{V_\mu}{n} \left.\frac{d\hat\theta}{d\bar y}\right|_\mu$$

or

$$\mathrm{Var}_\mu\left\{\hat\theta\right\} \doteq \left[-\ddot l_{\tilde\theta}(\mu)\right]^{-1} \left(\dot\eta_{\tilde\theta}^{\prime} \frac{V_\mu}{n} \dot\eta_{\tilde\theta}\right) \left[-\ddot l_{\tilde\theta}(\mu)\right]^{-1}.$$

In applied use, we would estimate the covariance of $\hat\theta$ by substituting $\bar y$ for μ and $\hat\theta$ for $\tilde\theta$, giving

$$\mathrm{Var}\left\{\hat\theta\right\} \doteq \left[-\ddot l_{\hat\theta}(\bar y)\right]^{-1} \left(\dot\eta_{\hat\theta}^{\prime} \frac{V_{\bar y}}{n} \dot\eta_{\hat\theta}\right) \left[-\ddot l_{\hat\theta}(\bar y)\right]^{-1}, \tag{4.10}$$

sometimes called "Huber's sandwich estimator".

The covariance matrix $V_{\mu=\bar y}$ is the only probabilistic element of the exponential family $\mathcal{G}$ entering into (4.10). Everything else depends only on the geometry of $\mathcal{F}$'s location inside $\mathcal{G}$.

Homework 4.8. Suppose $q = 1$, i.e., θ is real-valued, and that both $\mathcal{F}_A$ and $\mathcal{F}_B$ are specified. How might you numerically evaluate (4.10)?

We can retreat further from $\mathcal{G}$ by using the nonparametric covariance estimate

$$\overline{V} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})'$$

in place of $V_{\bar{y}}$. This results in the *nonparametric delta method* estimate of covariance,

$$\widehat{\mathrm{Var}}\left(\hat{\theta}\right) = \left[-\ddot{l}_{\hat{\theta}}(\bar{y})\right]^{-1}\left(\frac{\dot{\eta}'_{\hat{\theta}}\overline{V}\dot{\eta}_{\hat{\theta}}}{i}\right)\left[-\ddot{l}_{\hat{\theta}}(\bar{y})\right]^{-1}. \tag{4.11}$$

Homework 4.9. A nonparametric bootstrap sample $y_1^*, y_2^*, \ldots, y_n^*$ consists of a random sample of size n drawn *with* replacement from $(y_1, y_2, \ldots, y_n)$, and gives a bootstrap replication $\hat{\theta}^* = t(\sum_{i=1}^{n} y_i^*/n)$ of $\hat{\theta}$. The bootstrap estimate of covariance for $\hat{\theta}$ is $\mathrm{Cov}_*\{\hat{\theta}^*\}$, Cov_* indicating the covariance of $\hat{\theta}^*$ under bootstrap sampling, with $(y_1, y_2, \ldots, y_n)$ held fixed. Give an argument suggesting that $\mathrm{Cov}_*\{\hat{\theta}^*\}$ is approximately the same as (4.11). *Hint*: $\bar{y}^* \underset{*}{\sim} (\bar{y}, \overline{V}/n)$.

Note. In complicated models it is often easiest to evaluate $-\ddot{l}_{\hat{\theta}}(\bar{y})$ and $\dot{\eta}_{\hat{\theta}}$ numerically, by computing the likelihood and $\dot{\eta}_\theta$ at the $2p$ points $\hat{\theta} \pm \epsilon e_i$, where $e_i = (0, 0, \ldots, 1, 0, \ldots, 0)$, 1 in the ith place.

Homework 4.10. Explain how you would carry out the evaluation described above.

Table 4.1: The spatial test data.

	1	2	3	4	5	6	7	8	9	10	11	12	13
A	48	36	20	29	42	42	20	42	22	41	45	14	6
B	42	33	16	39	38	36	15	33	NA	NA	NA	NA	7

	14	15	16	17	18	19	20	21	22	23	24	25	26
A	0	33	28	34	4	32	24	47	41	24	26	30	41
B	15	NA	NA	NA	NA	NA	NA	NA	41	28	14	NA	NA

4.5 Missing data and the EM algorithm [Little and Rubin (1987), *Statistical Analysis with Missing Data* Chap. 7]

Exponential family structure is lost if some of a data set is missing. In the *spatial test data* in Table 4.1 and Figure 4.6, the original data set consisted of $n = 26$ pairs (A_i, B_i), each pair being two different measurements of spatial ability on a neurologically impaired child. However the B score was lost for 13 of the 26 children, as indicated by the Not Available abbreviation NA.

Suppose we assume, perhaps not very wisely, that the (A_i, B_i) were originally bivariate normal

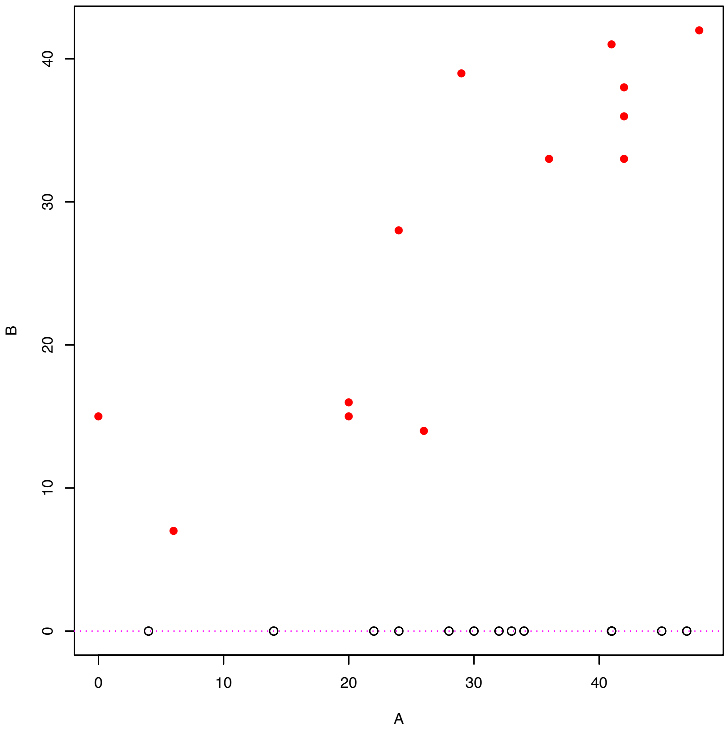


Figure 4.6: Spatial test data; open circles represent 13 children missing B measurement.

pairs, obtained independently,

$$\begin{pmatrix} A_i \\ B_i \end{pmatrix} \overset{\text{ind}}{\sim} \mathcal{N}_2(\lambda, \Gamma), \qquad i = 1, 2, \ldots, 26. \tag{4.12}$$

This is a five-parameter exponential family (Section 2.8). Somehow though, the B measurements have been lost for 13 of the children.[1] The data we get to see, 13 (A, B) pairs and 13 A values, is not from an exponential family.

Homework 4.11. In addition to (4.12), assume that $A_i \overset{\text{ind}}{\sim} \mathcal{N}(\lambda_1, \Gamma_{11})$ for the B open circle points in Figure 4.6. Show that this situation represents a curved exponential family. (The normal assumptions make this a special case; usually we wouldn't obtain even a curved exponential family.)

In the absence of missing data — if all 26 (A_i, B_i) pairs were observable — it would be straightforward to find the MLE $(\hat{\lambda}, \hat{\Gamma})$. With data missing, the naïve approach, simply using the 13 observable complete pairs, is both inefficient and biased. What follows is a brief discussion of an expansive theory that deals correctly with missing data situations. The theory is not confined to exponential family situations, but takes on a neater form in such situations.

[1] It is crucial in what follows that missingness should be random, in the sense that the chance B is missing should not depend on the value of A.

The Fisher–Louis expressions

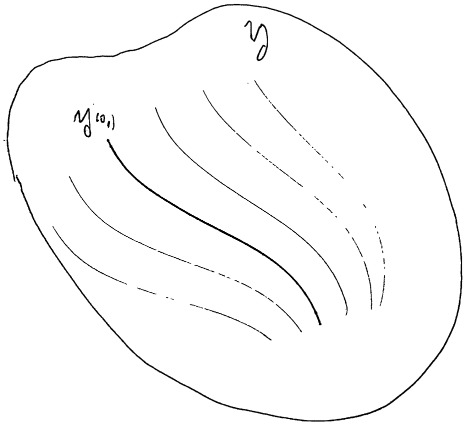
Let y be a data vector, of which we observe part o, with part u unobserved, and let $\mathcal{Y}(o_1)$ denote the set of y vectors having a certain value o_1 of o,

$$\mathcal{Y}(o_1) = \{y = (o, u) : o = o_1\}.$$

Fisher derived a simple but evocative expression for the score function based on observing only o:

Lemma 6 (Fisher)**.**

$$\dot{l}_\theta(o) = E_\theta\left\{\dot{l}_\theta(y) \mid o\right\}. \tag{4.13}$$



Fisher's lemma applies to all smoothly defined q-parameter families, not just exponential families. Its proof is particularly transparent when the sample space $\mathcal{Y}$ of y is discrete. The discrete density of o is then

$$f_\theta(o) = \sum_{\mathcal{Y}(o)} f_\theta(y).$$

Letting $\dot{f}_\theta$ indicate the gradient $(\partial f_\theta/\partial\theta_j)$, a q-vector,

$$\dot{f}_\theta(o) = \sum_{\mathcal{Y}(o)} \dot{f}_\theta(y) = \sum_{\mathcal{Y}(o)} \left(\frac{\dot{f}_\theta(y)}{f_\theta(y)}\right) f_\theta(y)$$

or

$$\frac{\dot{f}_\theta(o)}{f_\theta(o)} = \sum_{\mathcal{Y}(o)} \frac{\dot{f}_\theta(y)}{f_\theta(y)} \frac{f_\theta(y)}{f_\theta(o)},$$

which is (4.13).

There is also a slightly less simple expression for the observed Fisher information, attributed in its multiparameter version to Tom Louis:

Lemma 7 (Louis)**.**

$$-\ddot{l}_\theta(o) = E_\theta\left\{-\ddot{l}_\theta(y) \mid o\right\} - \operatorname{Cov}_\theta\left\{\dot{l}_\theta(y) \mid o\right\}. \tag{4.14}$$

Together the Fisher–Louis expressions say that the score function for o is that based on data y averaged over $\mathcal{Y}(o)$, while the observed Fisher information $I(o)$ is the corresponding average of $I(y)$ *minus* the covariance matrix of $\dot{l}_\theta(y)$ given $\mathcal{Y}(o)$. We lose Fisher information (and increase the variability of the MLE) when data is partly missing.

Homework 4.12. What happens if o is sufficient for y?

Homework 4.13. Prove (4.14) (assuming discreteness if you wish).

Suppose now that $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$ is an i.i.d. sample from a curved exponential family

$$f_\theta(y_i) = g_{\eta_\theta}(y_i) = e^{\eta_\theta' y_i - \psi(\eta_\theta)} g_0(y_i),$$

but we only observe part o_i of each y_i, $\boldsymbol{o} = (o_1, o_2, \dots, o_n)$. Let $y_i(\theta)$ and $V_i(\theta)$ indicate the conditional expectation and covariance of y_i given o_i,

$$y_i(\theta) = E_\theta\{y_i \mid o_i\} \quad \text{and} \quad V_i(\theta) = \mathrm{Cov}_\theta\{y_i \mid o_i\}.$$

Since

$$\dot{l}_\theta(y_i) = \dot{\eta}_\theta'(y_i - \theta) \quad \text{and} \quad -\ddot{l}_\theta(y_i) = i_\theta - \ddot{\eta}_\theta'(y_i - \mu_\theta),$$

the Fisher–Louis expressions become

$$\begin{aligned} \dot{l}_\theta(o_i) &= \dot{\eta}_\theta' \left(y_i(\theta) - \mu_\theta\right), \\ -\ddot{l}_\theta(o_i) &= i_\theta - \ddot{\eta}_\theta' \left(y_i(\theta) - \mu_\theta\right) - \dot{\eta}_\theta' V_i(\theta) \dot{\eta}_\theta. \end{aligned}$$

If we also assume that the missing data mechanism $y_i \to o_i$ operates independently for $i = 1, 2, \dots, n$, then the o_i are independent, $\dot{l}_\theta(\boldsymbol{o}) = \sum_1^n \dot{l}_\theta(o_i)$ and $\ddot{l}_\theta(\boldsymbol{o}) = \sum_1^n \ddot{l}_\theta(o_i)$, putting the Fisher–Louis expressions for $\boldsymbol{o}$ into compact form:

$$\begin{aligned} \dot{l}_\theta(\boldsymbol{o}) &= n\left[\dot{\eta}_\theta' \left(\bar{y}(\theta) - \mu_\theta\right)\right] \qquad \left(\bar{y}(\theta) = \sum_{i=1}^n y_i(\theta)/n\right), \\ -\ddot{l}_\theta(\boldsymbol{o}) &= n\left[i_\theta - \ddot{\eta}_\theta' \left(\bar{y}(\theta) - \mu_\theta\right) - \dot{\eta}_\theta' \overline{V}(\theta) \dot{\eta}_\theta\right] \qquad \left(\overline{V}(\theta) = \sum_{i=1}^n V_i(\theta)/n\right). \end{aligned} \tag{4.15}$$

Fisher's picture with missing data The MLE equation is now

$$\dot{\eta}_{\hat{\theta}}' \left(\bar{y}(\theta) - \mu_\theta\right) = 0.$$

Fisher's picture is the same as in Figure 4.1 (with $\bar{y}$ for y) *except* that $\bar{y}$ itself is now $\bar{y}(\theta)$, a function of the unknown parameter θ. This induces extra variability in $\hat{\theta}$, as suggested by the decrease of $n\dot{\eta}_{\hat{\theta}}' \overline{V}(\hat{\theta}) \dot{\eta}_{\hat{\theta}}$ in the observed Fisher information and depicted in Figure 4.7.

Homework 4.14. Suppose $o_i = y_i$ for $i = 1, 2, \dots, n_1$, and $o_i = \text{NA}$ for $i = n_{1+1}, n_{1+2}, \dots, n$. What does Figure 4.7 look like?

The EM algorithm [Dempster, Laird and Rubin (1977), *JRSS-B* 1–38]

One way to solve for $\hat{\theta}$ is to evaluate $\dot{l}_\theta$ and $\ddot{l}_\theta$ as in (4.15) and then iterate toward the solution using Newton–Raphson,

$$\theta^{(j+1)} - \theta^{(j)} = \left[-\ddot{l}_{\theta^{(j)}}(\boldsymbol{o})\right]^{-1} \left[\dot{l}_{\theta^{(j)}}(\boldsymbol{o})\right].$$

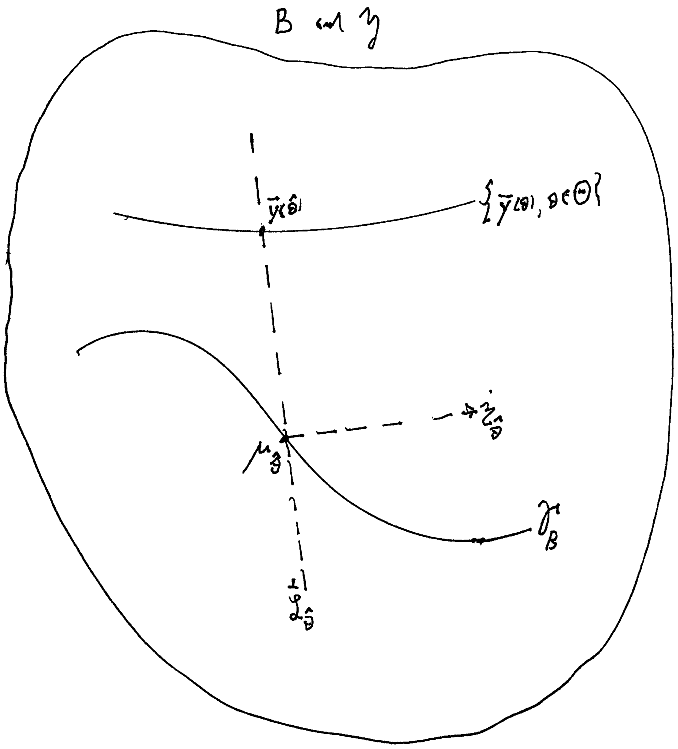


Figure 4.7: See text.

This requires the evaluation of $\bar{y}(\theta^{(j)})$ and $\overline{V}(\theta^{(j)})$ at each step.

The EM algorithm lets you get by with just $\bar{y}(\theta^{(j)})$:

$$\theta^{(j)} \xrightarrow{\text{“E”}} \bar{y}\left(\theta^{(j)}\right) \xrightarrow{\text{“M”}} \theta^{(j+1)} : \dot{\eta}'_{\theta^{(j+1)}}\left(\bar{y}\left(\theta^{(j)}\right) - \mu_{\theta^{(j+1)}}\right) = 0.$$

In words: given $\theta^{(j)}$, we use it to *Estimate* $\bar{y}$, the sufficient statistic if there were no missing data; and then we *Maximize* the full-data log likelihood $\eta'_\theta \bar{y} - \psi(\eta_\theta)$ to get $\theta^{(j+1)}$. See Figure 4.8.

The EM algorithm has some notable advantages:

- It is often easy to program.
- It is conservative; a theorem shows that $l_{\theta^{(j+1)}}(\boldsymbol{o}) \geq l_{\theta^{(j)}}(\boldsymbol{o})$ at every step.
- It is a good way to begin searching for the MLE even if you eventually switch over to Newton–Raphson.

There is a significant disadvantage:

- Convergence is slow, especially near the solution $\hat{\theta}$, with

$$\left\|\theta^{(j+1)} - \hat{\theta}\right\| \approx c_1 \left\|\theta^{(j)} - \hat{\theta}\right\|$$

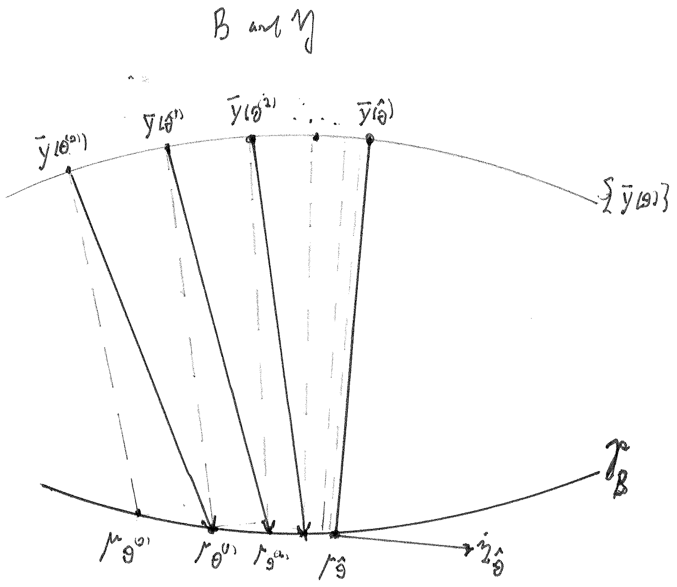


Figure 4.8: See text.

compared to the Newton–Raphson asymptotic convergence rate

$$\left\|\theta^{(j+1)} - \hat{\theta}\right\| \approx c_2 \left\|\theta^{(j)} - \hat{\theta}\right\|^2,$$

c_1 and c_2 small constants.

Homework 4.15. In the spatial test data (4.12), we wish to estimate $\theta = (\lambda_1, \lambda_2, \Gamma_{11}, \Gamma_{12}, \Gamma_{22})$.

(a) Program the EM algorithm to find the MLE $\hat{\theta}$. *Hint*: $E_\theta\{B_i \mid A_i\} = \lambda_2 + (\Gamma_{12}/\Gamma_{11})(A_i - \lambda_2)$.

(b) Program a Newton–Raphson search for $\hat{\theta}$ and compare the two convergence rates.

4.6 Statistical curvature[2]

We begin with a one-dimensional curved exponential family, $q = 1$, where $\mathcal{F}_A$ and $\mathcal{F}_B$ are curves through $\mathcal{R}^p$ and with sample size $n = 1$, $\bar{y} = y$; this last being no restriction since the geometry in Fisher's and Hoeffding's pictures stays the same for all n.

If $\mathcal{F}$ were an (uncurved) one-parameter subfamily of $\mathcal{G}$ then the MLE $\hat{\theta}$ would be a sufficient statistic for θ. This is not the case for curved families, where different values of y on $\overset{\perp}{\mathcal{L}}_{\hat{\theta}}$ yield the

[2] Derivations and details for the material in this section are in these references:

1. Efron (1975), Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.* **3**, 1189–1242.
2. Efron (1978). The geometry of exponential families. *Ann. Statist.* **6**, 362–376.
3. Efron (2018). Curvature and inference for maximum likelihood estimates. To appear *Ann. Statist.*
4. Efron and Hinkley (1978). Assessing the accuracy of the MLE: Observed versus expected Fisher information. *Biometrika* **65**, 457–487.

same MLE $\hat\theta = t(y)$ but different amounts of observed information,

$$I(y) = -\ddot{l}_{\hat\theta}(y) = i_{\hat\theta} - \ddot{\eta}'_{\hat\theta}(y - \mu_{\hat\theta}).$$

This happens because $\hat\theta$ is *not* sufficient in curved families.

Homework 4.16. Show the following:

(a) If $\hat\theta$ were sufficient, $-\ddot{l}_{\hat\theta}(y)$ would be constant on $\overset{\perp}{\mathcal{L}}_{\hat\theta}$.

(b) If $\ddot\eta_\theta = c_\theta \dot\eta_\theta$ for all θ and for some scalar function c_θ then $-\ddot{l}_{\hat\theta}(y) \equiv i_{\hat\theta}$.

(c) Under the same condition, $\mathcal{F}_A$ is a flat subset of A (so $\mathcal{F}$ is an exponential family).

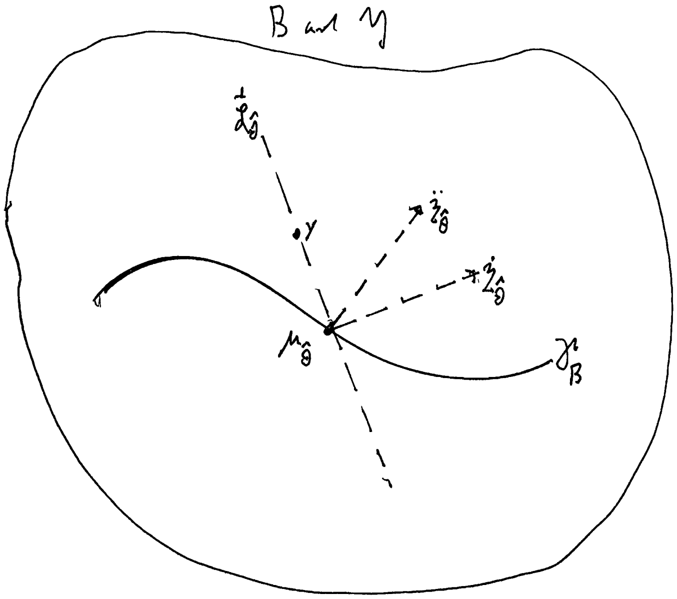


Figure 4.9: See text.

When $q = 1$, $\ddot\eta_{\hat\theta}$ as well as $\dot\eta_{\hat\theta}$ is a p-dimensional vector, and we can draw a helpful schematic diagram like that shown in Figure 4.9. Fisher argued for $I(y)^{-1}$ rather than $i_{\hat\theta}^{-1}$ as an estimate of $\mathrm{Cov}(\hat\theta)$. From Figure 4.9 we see $I(y) = i_{\hat\theta} - \ddot\eta'_{\hat\theta}(y - \mu_{\hat\theta})$ is *less* than $i_{\hat\theta}$ for y above $\mu_{\hat\theta}$ ("above" meaning in the direction best aligned with $\ddot\eta_{\hat\theta}$), presumably making $\hat\theta$ *more* variable; and conversely for y below $\mu_{\hat\theta}$. Curvature theory is intended to quantify this argument.

For a given value of θ, denote the 2×2 covariance matrix of $\dot{l}_\theta(y) = \dot\eta'_\theta(y - \mu_\theta)$ and $\ddot{l}_\theta(y) = \ddot\eta'_\theta(y - \mu_\theta) - i_\theta$ by

$$\begin{pmatrix} \nu_{11\theta} & \nu_{12\theta} \\ \nu_{21\theta} & \nu_{22\theta} \end{pmatrix} = \begin{pmatrix} \dot\eta'_\theta V_\theta \dot\eta_\theta & \dot\eta'_\theta V_\theta \ddot\eta_\theta \\ \ddot\eta'_\theta V_\theta \dot\eta_\theta & \ddot\eta'_\theta V_\theta \ddot\eta_\theta \end{pmatrix} \tag{4.16}$$

($\nu_{11\theta} = i_\theta = E_\theta\{-\ddot{l}_\theta(y)\}$). The residual of $\ddot{l}_\theta(y)$ after linear regression on $\dot{l}_\theta(y)$,

$$\overset{\perp}{l}_\theta(y) = \ddot{l}_\theta(y) - \frac{\nu_{12\theta}}{\nu_{11\theta}}\,\dot{l}_\theta(y)$$

has variance

$$\mathrm{Var}_\theta\left\{\overset{\perp}{l}(y)\right\} = \nu_{22\theta} - \frac{\nu_{12\theta}^2}{\nu_{11\theta}}.$$

Definition. The *statistical curvature* of $\mathcal{F}$ at θ is

$$\gamma_\theta = \left(\frac{\nu_{22\theta}}{\nu_{11\theta}^2} - \frac{\nu_{21\theta}^2}{\nu_{11\theta}^3}\right)^{1/2} = \frac{\mathrm{sd}_\theta\left\{-\overset{\perp}{l}_\theta(y)\right\}}{E_\theta\left\{-\ddot{l}_\theta(y)\right\}}. \tag{4.17}$$

(In classical differential geometry terms, γ_θ is the curvature of $\mathcal{F}_A$ in the metric V_θ.)

Curvature can be computed for general one-parameter families $\mathcal{F} = \{f_\theta(y), \theta \in \Theta\}$, not necessarily of exponential family form. Having calculated

$$\dot{l}_\theta(y) = \frac{\partial}{\partial\theta}\log f_\theta(y) \quad \text{and} \quad \ddot{l}_\theta(y) = \frac{\partial^2}{\partial\theta^2}\log f_\theta(y),$$

we define the 2×2 matrix (4.16) as the covariance matrix of $(\dot{l}_\theta(y), \ddot{l}_\theta(y))$ and then compute γ_θ as in (4.17).

Some important properties follow:

- The value of the curvature is invariant under smooth monotonic transformations of θ and y.
- $\gamma_\theta = 0$ for all $\theta \in \Theta$ if and only if $\mathcal{F}$ is a one-parameter exponential family.
- Large values of γ_θ indicate a breakdown of exponential family properties; for instance, locally most powerful tests of $H_0 : \theta = \theta_0$ won't be globally most powerful.
- In repeated sampling situations, $\gamma_{\theta,n} = \gamma_\theta/\sqrt{n}$ (so increased sample sizes make $\mathcal{F}_n$ more exponential family-like).
- In repeated sampling situations, with θ the true value,

  $$\frac{-\ddot{l}_{\hat\theta,n}(\bar{y})}{i_{\hat\theta,n}} = \frac{-\ddot{l}_{\hat\theta}(\bar{y})}{i_{\hat\theta}} \longrightarrow \mathcal{N}(1, \gamma_{\hat\theta}^2/n).$$

  This says that $\gamma_{\theta,n} = \gamma_\theta/\sqrt{n}$ determines the variability of the observed Fisher information $I_n(\bar{y})$ around the expected Fisher information $i_{\hat\theta,n}$.

Efron and Hinkley (1978) show that there is reason to take $1/I_n(\bar{y})^{1/2}$ rather than $1/i_{\hat\theta,n}^{1/2}$ as the estimate of $\mathrm{sd}\{\hat\theta\}$. If $\gamma_{\theta,n}^2$ is large, say $\geq 1/8$, then the two estimates can differ substantially.

Homework 4.17. Suppose $\gamma_{\hat{\theta},n} = 1/8$. What is a rough 68% interval for the ratio of the two estimates of $\mathrm{sd}\{\hat{\theta}\}$?

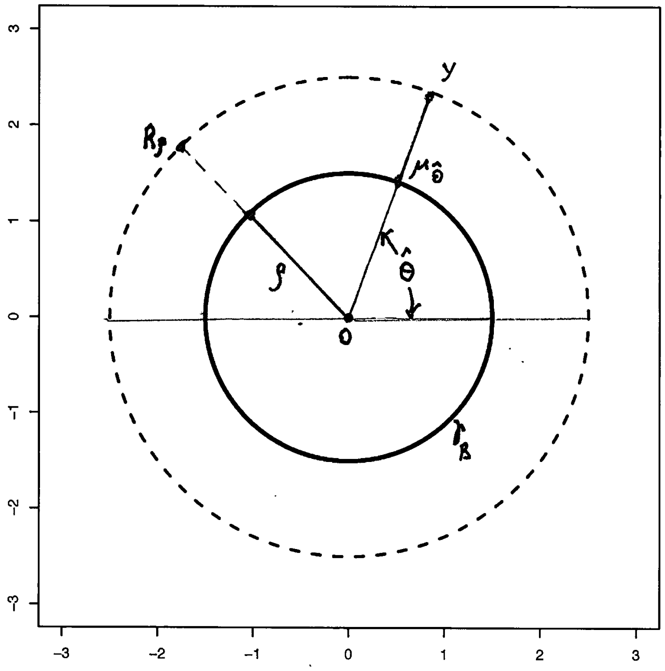


Figure 4.10: See text.

Fisher's circle model

Fisher provided a salient example of what we have been calling a curved exponential family, designed to show why $I(y)^{-1/2}$ is better than $i_{\hat{\theta}}^{-1/2}$ as an assessment of $\mathrm{sd}\{\hat{\theta}\}$. The example has $p = 2$, and y bivariate normal with identity covariance matrix $y \sim \mathcal{N}_2(\mu, I)$, where μ lies on a circle of known radius ρ:

$$\mathcal{F}_B = \left\{ \mu_\theta = \rho \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}, -\pi < \theta \leq \pi \right\}.$$

Homework 4.18. Show that

(a) $i_\theta = \rho^2$ for all θ;

(b) $\gamma_\theta = 1/\rho$ for all θ;

(c) $\mu_{\hat{\theta}}$ is the point on $\mathcal{F}_B$ nearest y;

(d) $I(y) = -\ddot{l}_{\hat{\theta}}(y) = Ri_\theta$ if y lies on a circle of radius $R\rho$ ($\|y\| = R\rho$).

Finally, give a heuristic argument supporting Fisher's preference for $I(y)$ in place of $i_{\hat{\theta}}$. *Hint*: Consider the case $\hat{\theta} = 0$.

R is an *ancillary statistic*: a random variable whose distribution does not depend on θ, but whose value determines the accuracy of $\hat{\theta}$ as an estimate of θ.

Homework 4.19. What is the distribution of R?

Bayesians criticize frequentist calculations for averaging over possible data sets that are different from the one actually observed: $\text{sd}(\hat{\theta}) = 1/i_{\hat{\theta}}^{1/2}$ is correct on average, but is misleading for R very large or very small. Conditioning on ancillary statistics was Fisher's method for reconciling frequentist and Bayesian calculations (though he wouldn't have said it that way). Like the MLE itself, $I(y)$ depends only on the observed likelihood function and not its frequentist framework, moving it closer to the Bayesian perspective. From a practical point of view, $I(y)$ can be numerically calculated directly from the likelihood, obviating the need for theoretical derivation of i_θ.

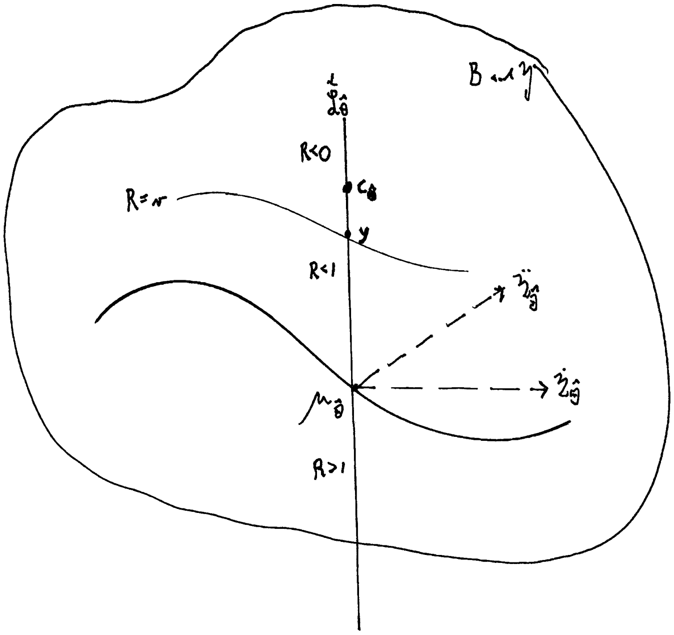

$B \sim \mathcal{Y}$
$\overset{\perp}{\mathcal{L}}_{\hat\theta}$
$R<0$
$c_{\hat\theta}$
$R=r$
y
$R<1$
$\ddot{\eta}_{\hat\theta}$
$\dot{\eta}_{\hat\theta}$
$\mu_{\hat\theta}$
$R>1$

Figure 4.11: See text.

Approximate versions of ancillarity hold more generally. This is illustrated in Figure 4.11, drawn for the situation $p = 2$ and $q = 1$. From $I(y) = i_{\hat{\theta}} - \ddot{\eta}_{\hat{\theta}}'(y - \mu_{\hat{\theta}}) = -\ddot{l}_{\hat{\theta}}(y)$, we see that there must be a critical point $c_{\hat{\theta}}$ above $\mu_{\hat{\theta}}$ on $\overset{\perp}{\mathcal{L}}_{\hat{\theta}}$ such that

$$I(c_{\hat{\theta}}) = -\ddot{l}_{\hat{\theta}}(c_{\hat{\theta}}) = 0.$$

Homework 4.20. Show that $i_{\hat{\theta}} = \ddot{\eta}_{\hat{\theta}}'(c_{\hat{\theta}} - \mu_{\hat{\theta}})$.

Let $\rho_{\hat{\theta}}$ be the *radius of curvature* at $\hat{\theta}$, $\rho_{\hat{\theta}} = 1/\gamma_{\hat{\theta}}$. Some results from Efron (1978) are:

- $\rho_{\hat{\theta}} = [(c_{\hat{\theta}} - \mu_{\hat{\theta}})' V_{\hat{\theta}}^{-1} (c_{\hat{\theta}} - \mu_{\hat{\theta}})]^{1/2}$, that is, $\rho_{\hat{\theta}}$ equals the Mahalanobis distance from $c_{\hat{\theta}}$ to $\mu_{\hat{\theta}}$.

- Let R be the proportional distance of y from $c_{\hat\theta}$ to $\mu_{\hat\theta}$,

  $$R = \frac{\|c_{\hat\theta} - y\|}{\|c_{\hat\theta} - \mu_{\hat\theta}\|}$$

  (in any metric since $\overset{\perp}{\mathcal{L}}_{\hat\theta}$ is one-dimensional). Then

  $$R = -\ddot{l}_{\hat\theta}(y)/i_{\hat\theta} = I(y)/i_{\hat\theta},$$

  the repeated sampling version $-\ddot{l}_{\hat\theta}(\bar{y})/i_{\hat\theta}$ being asymptotically $\mathcal{N}(1, \gamma_\theta^2/n)$. R is > 1 below $\mathcal{F}_B$ and < 1 above it.

- Beyond the critical point $c_{\hat\theta}$, R is less than 0 and $\hat\theta$ is a local *minimum* rather than *maximum* of the likelihood.

- Let $\mathcal{C}_r$ be the curve in $\mathcal{Y}$ representing those y having R equal to some particular value r. Then the conditional standard deviation of $\hat\theta$ given that y is on $\mathcal{C}_r$ is approximately

  $$\text{sd}\left\{\hat\theta \mid R = r\right\} = 1\Big/\sqrt{r i_\theta},$$

  leading to the data-based estimate

  $$\text{sd}\left\{\hat\theta \mid R\right\} \doteq \sqrt{R i_{\hat\theta}}^{\,-1} = \sqrt{I(y)}^{\,-1}. \tag{4.18}$$

  R is approximately ancillary for θ — i.e., its distribution (almost) doesn't depend on θ — strengthening the rationale for (4.18).[3] Note that $I_n(\bar{y})/i_{\hat\theta,n} = I(\bar{y})/i_{\hat\theta}$ so the curve $\mathcal{C}_r$ stays the same under repeated sampling: the difference being that $\bar{y}$ tends closer to $\mathcal{F}_B$, making R closer to 1.

Homework 4.21. (a) In Figure 4.11, identify the elements of Fisher's circle model: $c_{\hat\theta}$, $\dot\eta_{\hat\theta}$, etc.

(b) Give a heuristic argument justifying (4.18) in this case.

Large values of the curvature move the critical points c_θ nearer to $\mathcal{F}_B$. This can destabilize the MLE. In addition to increased variance, there is the possibility of false roots, the probability of falling on the far side of the critical boundary being approximately

$$\Pr_\theta\{R < 0\} \doteq \Phi(-1/\gamma_\theta),$$

Φ the standard normal cdf.

[3] A better approximate ancillary is $\tilde{R} = (R-1)/\gamma_{\hat\theta}$, better in the sense of having its distribution depend less on θ, but conditioning on $\tilde{R}$ lends again to $\sqrt{I(y)}^{\,-1}$ as the preferred estimate of $\text{sd}\{\hat\theta\}$. The underlying rationale for (4.18) is geometrical, as exemplified in Fisher's circle model.

Curvature theory also has something to say about Fisher's claim that the MLE is a superior estimate of θ. As in Section 4.2, suppose that $\tilde{\theta} = t(\bar{y})$ is a locally unbiased and first-order efficient estimator of the true value θ, so that asymptotically

$$\lim_{n\to\infty} n \operatorname{Var}_{\theta,n}\left\{\tilde{\theta}\right\} = i_\theta.$$

The theorem in Section 10 of Efron (1975) says that

$$\operatorname{Var}_\theta\left\{\tilde{\theta}\right\} \doteq \frac{1}{ni_\theta} + \frac{1}{n^2 i_\theta}\{\gamma_\theta^2 + C_1 + C_2\},$$

where

$$C_1 \text{ is the same positive constant for all choices of the estimator } \tilde{\theta},$$

while

$$C_2 \geq 0, \text{ equaling zero only for the MLE.}$$

This result demonstrates the "second-order efficiency of the MLE": using an estimator other than the MLE increases asymptotic variance.

Some examples of one-parameter curved families

Normal with known coefficient of variation Suppose $x \sim \mathcal{N}(\theta, \Gamma)$ as in Section 2.8, but with $\Gamma = \tau\theta^2$, τ known (so the coefficient of variation of x is $\sqrt{\Gamma}/|\theta| = \sqrt{\tau}$). This is a one-parameter curved family in $p = 2$ dimensions, with $y = (x, x^2)$ and

$$\eta_\theta = \frac{1}{\tau}\left(\frac{1}{\theta}, -\frac{1}{2\theta^2}\right)'.$$

Homework 4.22. (a) Augment Figure 2.6 to include $\mathcal{F}_A$ and $\mathcal{F}_B$.

(b) What is γ_θ?

Autoregressive process $x_0, x_1, \dots, x_T$ are observations of an autoregressive process

$$y_0 = u_0, \quad y_{t+1} = \theta_{y_t} + (1-\theta^2)^{1/2} u_{t+1} \qquad (t = 1, 2, \dots, T),$$

where $u_t \overset{\text{ind}}{\sim} \mathcal{N}(0,1)$ for $t = 0, 1, \dots, T$, and $\Theta = (-1, 1)$.

Homework 4.23. Show that the preceding is a one-parameter curved family in $p = 3$ dimensions, and give expressions for y and η_θ. (Efron 1975 shows that $\gamma_0^2 = (8T-6)/T^2$.)

A genetics linkage model [Dempster, Laird and Rubin (1977), *JRSS-B*] Four phenotypes occur in proportions

$$\pi_\theta = \left(\frac{1}{2} + \frac{\theta}{4}, \frac{1}{4} - \frac{\theta}{4}, \frac{1}{4} - \frac{\theta}{4}, \frac{\theta}{4}\right), \tag{4.19}$$

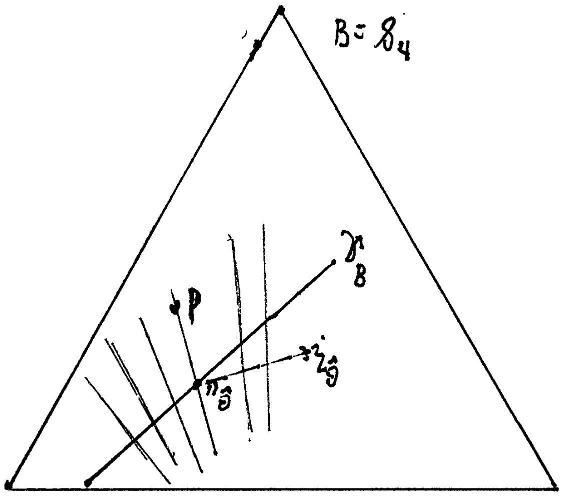


Figure 4.12: See text.

where θ is an unknown parameter between 0 and 1. A sample of $n = 197$ animals gave observed proportions

$$p = (125, 18, 20, 34)/197,$$

$p \sim \text{Mult}_4(n, \pi_\theta)/n$ as in Section 2.9. Model (4.19) is linear in the B space, as shown in Figure 4.12, but curved in the B space, making it a one-parameter curved family: $\mathcal{F}_B$ is the line

$$\pi_\theta = a + b\theta \qquad \begin{cases} a = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}, 0\right) \\ b = \left(\frac{1}{4}, -\frac{1}{4}, -\frac{1}{4}, \frac{1}{4}\right), \end{cases}$$

with η_θ the curve $\log \pi_\theta$.

Homework 4.24. (a) Write an iterative program to find the MLE $\hat{\theta}$.

(b) Provide an estimated standard error of $\hat{\theta}$.

(c) Show, numerically, that $\gamma_{\hat{\theta},n} = 0.0632$.

Cauchy translation family A Cauchy observation y has density

$$g_\theta(y) = \frac{1}{\pi} \frac{1}{1 + (y - \theta)^2} \qquad (y \text{ and } \theta \in \mathcal{R}^p),$$

the translation parameter θ being the centerpoint of the symmetric density. Though *not* an exponential family, we can differentiate the log likelihood function $l_\theta(y) = -\log[1 + (y - \theta)^2] - \log \pi$ to get $\dot{l}_\theta(y)$ and $\ddot{l}_\theta(y)$, and then compute their covariance matrix (4.16) to get the curvature (4.17), $\gamma_\theta = \sqrt{2.5}$ (the same for all θ because this is a translation family). A Cauchy sample of say $n = 10$ observations $y_1, y_2, \dots, y_n$ would have $\gamma_{\theta,n} = \sqrt{0.25} = 0.50$: a large curvature implying, correctly, difficulties with multiple roots to the MLE equation $\dot{l}_{\hat{\theta}}(\boldsymbol{y}) = 0$.

For θ near 0 we can expand $l_\theta(y)$ in a Taylor series,

$$l_\theta(y) = l_0(y) + \dot{l}_0(y)\theta + \ddot{l}_0(y)\theta^2/2 + \dddot{l}_0(y)\theta^3/6 + \dots,$$

and think of this as a one-parameter curved exponential family having sufficient vector $(\dot{l}_0(y), \ddot{l}_0(y), \dddot{l}_0(y), \dots)$ and

$$\eta_\theta = (\theta, \theta^2/2, \theta^3/6, \dots).$$

The point here is that curved exponential families can be thought of as an approximating framework for all smoothly defined parametric families, with the amount of curvature indicating favorable or unfavorable properties of the MLE. Section 4.8 discusses an example of multiparameter curved families ($q > 1$) and generalizations of Figure 4.11.

4.7 Regions of stability for the MLE[4]

Figure 4.11 showed the MLE $\hat{\theta}$ growing more variable as the observation vector y moved along $\overset{\perp}{\mathcal{L}}_{\hat{\theta}}$ toward the critical point $c_{\hat{\theta}}$—becoming unstable beyond $c_{\hat{\theta}}$, where $\hat{\theta}$ is a minimum rather than maximum of the likelihood $l_\theta(y)$. Figure 4.11 applied to the case $p = 2$ and $q = 1$, but a theory of stability applies to general (p, q) curved exponential families. Only a brief description of the theory is developed here.

A transformation of coordinates simplifies the description. Let η_0 be any point in A, the parameter space of the full p-parameter exponential family $\mathcal{G}$ (Section 4.1). At η_0, y has expectation vector μ_0 and covariance matrix V_0. Also let M be a symmetric $p \times p$ square root matrix of V_0, $M^2 = V_0$. (If V_0 has eigenvalue-eigenvector representation $V_0 = \Gamma D \Gamma'$, D the diagonal matrix of eigenvalues, we can set $M = \Gamma D^{1/2} \Gamma'$.)

Transform η and y to

$$\tilde{\eta} = M\eta \quad \text{and} \quad \tilde{y} = M^{-1}(y - \mu_0).$$

Then $\tilde{y}$ has expectation 0 and covariance matrix

$$M^{-1} V_0 M^{-1} = I_p,$$

the $p \times p$ identity matrix at $\eta = \eta_0$. The family $\mathcal{G}$ transforms into an equivalent exponential family,

$$\widetilde{\mathcal{G}} = \left\{ \tilde{g}_{\tilde{\eta}}(\tilde{y}) = e^{\tilde{\eta}'\tilde{y} - \tilde{\psi}(\tilde{\eta})} \tilde{g}_0(\tilde{y}), \tilde{\eta} \in \tilde{A}, \tilde{y} \in \widetilde{\mathcal{Y}} \right\}, \tag{4.20}$$

with $\tilde{y} \sim (0, I_p)$ at $\tilde{\eta} = M\eta_0$.

Homework 4.25. Verify (4.20).

In what follows, we will assume that (η, y) have *already been transformed*, with η_0 chosen to be

[4] From Efron (2018), "Curvature and inference for maximum likelihood estimates", to appear *Ann. Statist.*

the MLE point $\eta_{\hat\theta}$, so that

$$\mu_{\hat\theta} = 0 \quad \text{and} \quad V_{\hat\theta} = I_p \tag{4.21}$$

in the curved family $\mathcal{F}$. All the calculations will be conditional on $\theta = \hat\theta$, making (4.21) legitimate for the purpose of easy calculation.

Returning to one-parameter curved families as in Section 4.6, let $\hat\nu_{11} = \nu_{11\hat\theta}$, $\hat\nu_{12} = \nu_{12\hat\theta}$, and $\hat\nu_{22} = \nu_{22\hat\theta}$ in (4.16), and define

$$\overset{\perp}{\eta}_{\hat\theta} = \ddot\eta_{\hat\theta} - \frac{\hat\nu_{12}}{\hat\nu_{11}}\dot\eta_{\hat\theta},$$

the component of $\ddot\eta_{\hat\theta}$ orthogonal to $\dot\eta_{\hat\theta}$.

Homework 4.26. Show that, in terms of the definitions in Section 4.6,

(a) $\|\overset{\perp}{\eta}_{\hat\theta}\| = i_{\hat\theta}\gamma_{\hat\theta}$;

(b) $-\ddot l_{\hat\theta}(y) = i_{\hat\theta} - \overset{\perp}{\eta}{}'_{\hat\theta}(y - \mu_{\hat\theta})$.

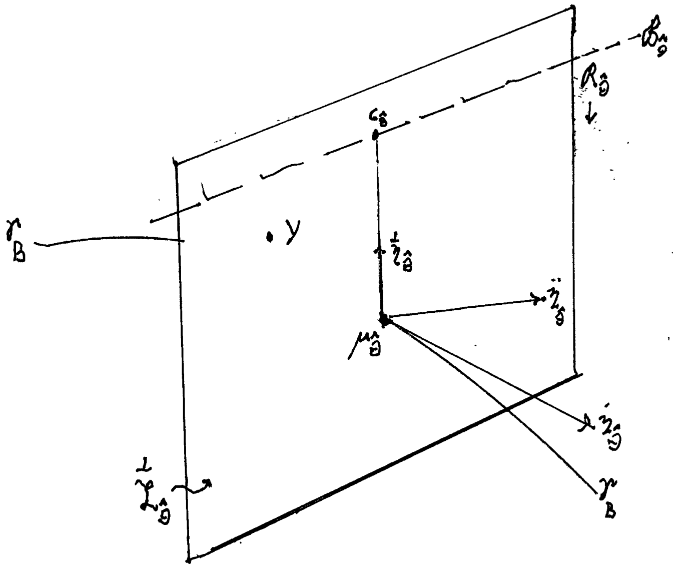

$\mathcal{B}_{\hat\theta}$
$\mathcal{R}_{\hat\theta}$
$c_{\hat\theta}$
$\mathcal{F}_B$
y
$\overset{\perp}{\eta}_{\hat\theta}$
$\ddot\eta_{\hat\theta}$
$\mu_{\hat\theta}$
$\dot\eta_{\hat\theta}$
$\overset{\perp}{\mathcal{L}}_{\hat\theta}$
$\mathcal{F}_B$

Figure 4.13: See text.

Figure 4.13 illustrates maximum likelihood estimation for one-parameter curved families, generalizing Figure 4.11 to $p \geq 2$, so that $\overset{\perp}{\mathcal{L}}_{\hat\theta} = \{y : \text{MLE} = \hat\theta\}$ has dimension $p-1$. The critical point $c_{\hat\theta}$ is the nearest point to $\mu_{\hat\theta}$ such that the observed Fisher information $I(y) = -\ddot l_{\hat\theta}(y)$ equals 0.

Homework 4.27. Let $v_{\hat\theta} = \overset{\perp}{\eta}_{\hat\theta}/i_{\hat\theta}\gamma_{\hat\theta}$ (so $\|v_{\hat\theta}\| = 1$). Show that

(a) $c_{\hat\theta} = v_{\hat\theta}/\gamma_{\hat\theta}$;

(b) $I(y) = i_{\hat\theta}(1 - b\gamma_{\hat\theta})$, where $y = \mu_{\hat\theta} + bv_{\hat\theta} + r$, with r in $\overset{\perp}{\mathcal{L}}_{\hat\theta}$ and $v'_{\hat\theta}r = 0$.

The critical boundary

$$\mathcal{B}_{\hat\theta} = \{y = \mu_{\hat\theta} + v_{\hat\theta}/\gamma_{\hat\theta} + r\}$$

separates $\overset{\perp}{\mathcal{L}}_{\hat\theta}$ into halves having $I(y) > 0$ or < 0. By definition, the *stable region* $\mathcal{R}_{\hat\theta}$ is the good half,

$$\mathcal{R}_{\hat\theta} = \{y : y = \mu_{\hat\theta} + bv_{\hat\theta} + r, b < 1/\gamma_{\hat\theta}\}.$$

Large values of the curvature move $\mathcal{B}_{\hat\theta}$ closer to $\mu_{\hat\theta}$, destabilizing the MLE, allowing $I(y)$ to vary more from $i_{\hat\theta}$ and even to go negative. Since $\mathrm{Cov}_{\hat\theta}(y) = I_p$ in our transformed coordinates, we have the approximation

$$\Pr_{\hat\theta}\{I(y) < 0\} \doteq \Phi(-1/\gamma_{\hat\theta}),$$

Φ the standard normal cdf.

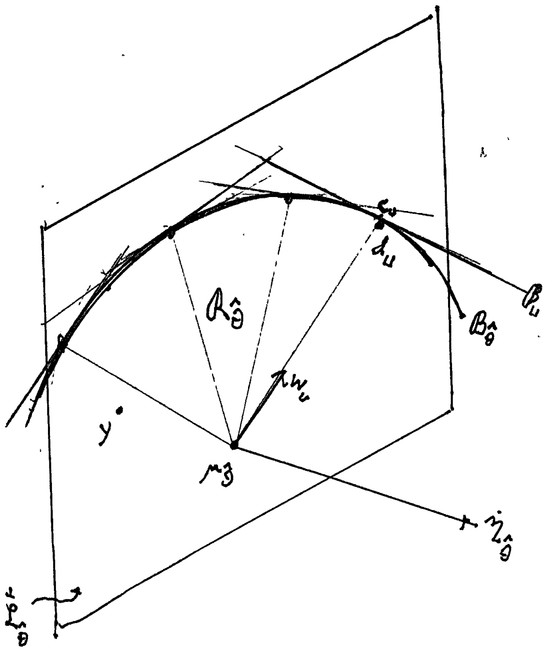

c_u
a_u
$\mathcal{R}_{\hat\theta}$
$\mathcal{B}_u$
$\mathcal{B}_{\hat\theta}$
w_u
y
$\mu_{\hat\theta}$
$\dot\eta_{\hat\theta}$
$\overset{\perp}{\mathcal{L}}_{\hat\theta}$

Figure 4.14: See text.

What happens in multiparameter curved exponential families, that is, those with $q > 1$? A schematic answer appears in Figure 4.14. The stable region $\mathcal{R}_{\hat\theta}$ is now a convex subset of $\overset{\perp}{\mathcal{L}}_{\hat\theta}$. For a given unit vector u in p-dimensions, let

$$\mathcal{F}_u = \left\{f_{\hat\theta+\lambda u}, \lambda \in \Lambda\right\}$$

be a one-parameter subfamily of $\mathcal{F}$, as in Section 2.6. Here Λ is an interval of R^1 containing 0 as an interior point. $\mathcal{F}_u$ determines Fisher information i_u and curvature γ_u at $\lambda = 0$, and also direction vector $\overset{\perp}{\eta}_u/i_u\gamma_u$.

Finally, let a_u be the angle between $\overset{\perp}{\eta}_u$ and $\overset{\perp}{\mathcal{L}}_{\hat\theta}$ ($a_u = 0$ if $q = 1$), and w_u the unit projection vector of $\overset{\perp}{\eta}_u$ into $\overset{\perp}{\mathcal{L}}_{\hat\theta}$. The $q = 1$ theory yields a half-space $\mathcal{R}_u$ of $\overset{\perp}{\mathcal{L}}_{\hat\theta}$ as the stable region for family $\mathcal{F}_u$. Its boundary $\mathcal{B}_u$ lies at distance d_u from $\mu_{\hat\theta}$ and it can be shown that

$$d_u = 1 \big/ \left[\cos(a_u) \cdot \gamma_u\right]. \tag{4.22}$$

Theorem 3. *The observed Fisher information matrix $I(y) = -\ddot{l}_{\hat\theta}(y)$ is positive definite if and only if y is in the* stable region

$$\mathcal{R}_{\hat\theta} = \bigcap_u \mathcal{R}_u, \tag{4.23}$$

where the intersection is over all p-dimensional unit vectors.

All of this is verified in Efron (2018). An example having $p = 37$ and $q = 8$ (dimension $\overset{\perp}{\mathcal{L}}_{\hat\theta} = 37 - 8 = 29$) appears in Section 4.8.

Homework 4.28. In family $\mathcal{F}_u$, show that at $\lambda = 0$,

(a) The expected Fisher information i_u is $u' i_{\hat\theta} u$;

(b) the observed Fisher information $I_u(y) = u' I(y) u$;

(c) $u' I(y) u = u' i_{\hat\theta} u \cdot (1 - b/d_u)$, where $y = \mu_{\hat\theta} + b w_u + r$, with r in $\overset{\perp}{\mathcal{L}}_{\hat\theta}$ and orthogonal to w_u.

Once again, large curvatures move the critical boundary $\mathcal{B}_{\hat\theta}$ closer to $\mu_{\hat\theta}$, increasing conditioning effects and potentially destabilizing the MLE. Note that not all choices of u contribute to $\mathcal{B}_{\hat\theta}$. In our examples, some choices give enormous values of d_u, so that $\mathcal{B}_u$ lies far beyond $\mathcal{B}_{\hat\theta}$.

Figure 4.15 shows a toy example from Efron (2018). The observed data y comprises seven independent Poisson observations,

$$y_i \overset{\text{ind}}{\sim} \text{Poi}(\mu_i), \qquad i = 1, 2, \dots, 7, \tag{4.24}$$

$y = (1, 1, 6, 11, 7, 14, 15)$. A hypothesized model supposes that

$$\mu_i = \theta_1 + \theta_2 x_i \qquad \text{for } x_i = -3, -2, -1, 0, 1, 2, 3. \tag{4.25}$$

In this case, $\mathcal{G}$ is a $(p = 7)$-parameter exponential family, having $\mathcal{F}$ as a two-parameter curved subfamily. (If instead we had taken $\log \mu_i = \theta_1 + \theta_2 x_i$, $\mathcal{F}$ would a two-parameter GLM.)

Homework 4.29. Write an explicit expression of $\mathcal{F}$ for this $\mathcal{G}$.

Direct numerical calculation gave MLE $\hat\theta = (\hat\theta_1, \hat\theta_2) = (7.86, 2.38)$ in model (4.24)–(4.25), illustrated by the straight line fit $\hat\theta_1 + \hat\theta_2 x_i$ in Figure 4.15. The observed and expected Fisher information matrices were calculated to be

$$i_{\hat\theta} = \begin{pmatrix} 2.63 & -5.75 \\ -5.75 & 14.98 \end{pmatrix} \quad \text{and} \quad I(y) = \begin{pmatrix} 3.00 & -6.97 \\ -6.97 & 23.00 \end{pmatrix}.$$

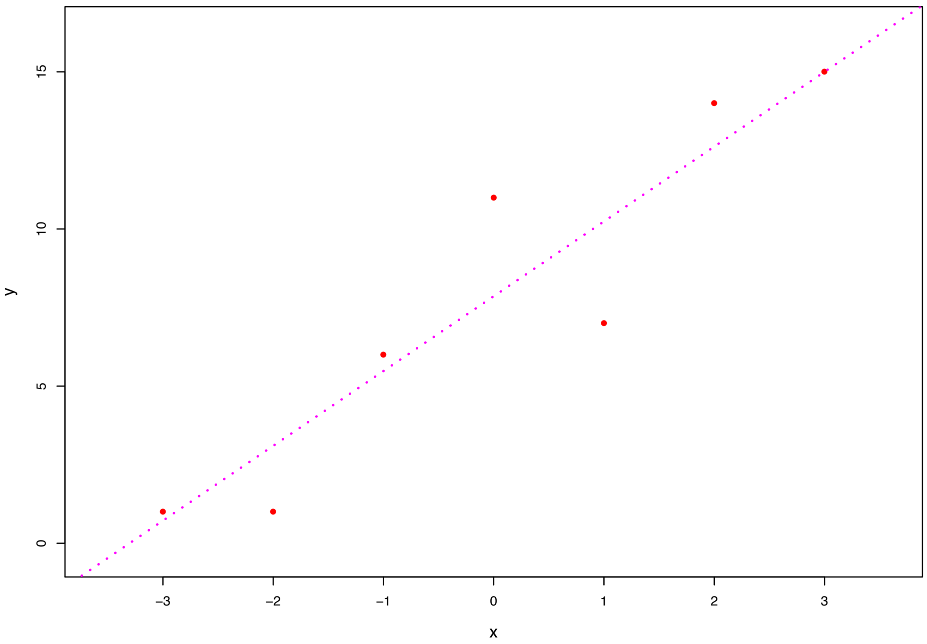


Figure 4.15: Toy example. $y_i \sim \text{Poi}(\mu_i)$ for $i = 1, 2, \ldots, 7$; curved model $\mu_i = \theta_0 + \theta_1 x_i$; MLE line $\hat{\theta} = (7.86, 2.38)$.

Model (4.24)–(4.25) is a $(p, q) = (7, 2)$ curved family, making $\overset{\perp}{\mathcal{L}}_{\hat{\theta}}$ in Figure 4.15 a five-dimensional flat space. Direction-distance pairs (w_u, d_u) were calculated for 101 choices of u, $u(t) = (\cos t, \sin t)$, $t = k\pi/100$ for $k = 0, 1, \ldots, 100$. Somewhat surprisingly, w_u was always the same,

$$w = (-0.390, 0.712, 0.392, 0.145, -0.049, -0.210, -0.340); \tag{4.26}$$

d_u took on its mininum value $d_{\min} = 2.85$ at $u = (0, 1)$, with the other d_u's ranging up to 1133. In this case, the stable region $\mathcal{R}_u$ was a half-space of $\overset{\perp}{\mathcal{L}}_{\hat{\theta}}$, with boundary $\mathcal{B}_{\hat{\theta}}$ minimum distance 2.85 from $\mu_{\hat{\theta}}$ (after transformation to standard coordinates (4.21)).

The observed information matrix $I(bw)$, w as in (4.26), decreases toward singularity as b increases; it becomes singular at $b = 2.85$, at which point its lower right corner equals 0. Further increases of b reduce other quadratic forms $u(t)'I(bw)u(t)$ to zero, as in Homework 4.28(c).

Homework 4.30. (a) Why was it the lower right corner?

(b) Given a list of $d_{u(t)}$ versus t, which choice $u(t)'I(bw)u(t)$ would give the second zero?

From $y \sim (0, I_p)$ in our transformed coordinates, we get

$$b = (y - \mu_{\hat{\theta}})'w \sim (0, 1)$$

for the projection of y along w. Using Homework 4.28(a), applied to $u = (0, 1)$, the lower right corners of $I(y)$ and $i_{\hat{\theta}}$ have

$$I_{11}(y)/i_{\hat{\theta}11} \sim (1, 1/2.85^2) = (1, 0.35^2),$$

so we can expect conditioning effects on the order of 35%.

Penalized maximum likelihood estimation The performance of maximum likelihood estimates can often be improved by "regularization", that is, by adding a penalty term to the log likelihood in order to tamp down volatile behavior of $\hat{\theta}$, this being especially true when the number of parameters p is large.

We define the penalized log likelihood function $m_\theta(y)$ to be

$$m_\theta(y) = l_\theta(y) - s_\theta,$$

where s_θ is a non-negative *penalty function.* Two familiar choices are $s_\theta = c\sum_1^p \theta_j^2$ ("ridge regression") and $s_\theta = c\sum_1^p |\theta_j|$ ("the lasso"), c a positive constant. The "g-modeling" example of the next section uses $c(\sum_1^p \theta_j^2)^{1/2}$. Larger values of c pull the penalized maximum likelihood estimate (pMLE) more strongly toward 0 (using definitions such that $\theta = 0$ is a plausible choice in the absence of much data).

By definition, the pMLE is the value of θ maximizing $m_\theta(y)$. In a (p, q) curved exponential family $\mathcal{F}$, the equivalent of the score function condition $\dot{l}_\theta(y) = 0$ is $\dot{m}_\theta(y) = 0$ or, as in (4.5),

$$\dot{\eta}_{\hat{\theta}}(y - \mu_{\hat{\theta}}) - \dot{s}_{\hat{\theta}} = 0,$$

where $\dot{s}_\theta$ is the q-dimensional gradient vector $(\partial s/\partial\theta_j)$. For a given value of $\hat{\theta}$, the set of y vectors satisfying this is the (p, q)-dimensional hyperplane $\overset{\perp}{\mathcal{M}}_{\hat{\theta}}$,

$$\overset{\perp}{\mathcal{M}}_{\hat{\theta}} = \{y : \dot{\eta}'_{\hat{\theta}}(y - \mu_{\hat{\theta}}) = \dot{s}_{\hat{\theta}}\}. \tag{4.27}$$

$\overset{\perp}{\mathcal{M}}_{\hat{\theta}}$ lies parallel to $\overset{\perp}{\mathcal{L}}_{\hat{\theta}}$ (4.6), but offset from $\mu_{\hat{\theta}}$.

In Euclidean distance, the nearest point in $\overset{\perp}{\mathcal{M}}_{\hat{\theta}}$ to $\mu_{\hat{\theta}}$ is

$$\nu_{\hat{\theta}} = \mu_{\hat{\theta}} + \dot{\eta}_{\hat{\theta}}(\dot{\eta}'_{\hat{\theta}}\dot{\eta}_{\hat{\theta}})^{-1}\dot{s}_{\hat{\theta}},$$

having squared distance

$$\|\nu_{\hat{\theta}} - \mu_{\hat{\theta}}\|^2 = \dot{s}'_{\hat{\theta}}(\dot{\eta}'_{\hat{\theta}}\dot{\eta}_{\hat{\theta}})^{-1}\dot{s}_{\hat{\theta}},$$

these being standard projection calculations.

Homework 4.31. Draw the pMLE equivalent of Figure 4.1.

By analogy with the observed information matrix $I(y) = -\ddot{l}_{\hat{\theta}}(y)$, we define

$$J(y) = -\ddot{m}_{\hat{\theta}}(y) = I(y) + \ddot{s}_{\hat{\theta}},$$

$\ddot{s}$ the $q \times q$ matrix $(\partial^2 s/\partial\theta_j \partial\theta_k)$. $J(y)$ plays a central role in the accuracy and stability of the pMLE, as discussed in Efron (2018). For instance the influence function (4.8) becomes

$$\frac{\partial\hat{\theta}}{\partial y} = J(y)^{-1}\dot{\eta}'_{\hat{\theta}}.$$

Versions of Figure 4.13 and Figure 4.14 apply to the pMLE as well, with $J(y)$ and $\nu_{\hat{\theta}}$ playing the roles of $I(y)$ and $\mu_{\hat{\theta}}$.

4.8 Empirical Bayes theory[5]

A typical empirical Bayes estimation problem begins with a collection $\theta_1, \theta_2, \ldots, \theta_N$ of real-valued, unobserved parameters sampled from an unknown probability density,

$$\theta_i \overset{\text{ind}}{\sim} g(\theta) \qquad \text{for } i = 1, 2, \ldots, N. \tag{4.28}$$

Each θ_i independently produces an observation x_i according to a known probability density function $p(x \mid \theta)$,

$$x_i \mid \theta_i \sim p(x_i \mid \theta_i), \tag{4.29}$$

for instance $x_i \sim \mathcal{N}(\theta_i, 1)$ or $x_i \sim \text{Poi}(\theta_i)$.

We wish to estimate the θ_i, perhaps a specific one of them or perhaps all. If the prior density $g(\theta)$ were known, Bayes rule would yield ideal inferences based on the posterior density $g(\theta_i \mid x_i)$. It came as a pleasant surprise to statisticians in the 1950s[6] that, in situation (4.28)–(4.29), it is often possible to do nearly as well *without* knowledge of $g(\cdot)$.

How this can be done is the subject of empirical Bayes theory. A key element is the *marginal density* $f(x)$,

$$f(x) = \int_{\mathcal{T}} p(x \mid \theta) g(\theta)\, d\theta, \tag{4.30}$$

$\mathcal{T}$ the sample space for the θ's. The observed data $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$ is a random sample from $f(x)$,

$$x_i \overset{\text{iid}}{\sim} f(\cdot), \qquad i = 1, 2, \ldots, N,$$

which is all the statistician gets to see.

In what follows, we will bin the data as in Section 3.4, partitioning the x-axis into K bins $\mathcal{Z}_k$,

[5] The notation in this section is specialized to empirical Bayes calculations, and differs from that in Section 1.6. The main reference again is Efron (2018), "Curvature and inference for maximum likelihood estimates."

[6] This work was pioneered by Herbert Robbins in key papers in the early 1950s; he coined the name "empirical Bayes".

and letting $y_k = \#\{x_i \in \mathcal{Z}_k\}$ be the count in bin k. The count vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_K)$ is a sufficient statistic for the discretized data. It has a multinomial distribution (Section 2.9)

$$\boldsymbol{y} \sim \text{Mult}_K(N, \boldsymbol{f}),$$

where $\boldsymbol{f} = (f_1, f_2, \ldots, f_K)$ is a discretized version of $f(x)$,

$$f_k = \int_{\mathcal{Z}_k} f(x)\ dx.$$

Here is a schematic diagram of the empirical Bayes model (4.28)–(4.30):

$$g \longrightarrow f \longrightarrow \boldsymbol{y} \sim \text{Mult}_K(N, \boldsymbol{f}). \tag{4.31}$$

We observe $\boldsymbol{y}$ and wish to estimate various functions of g, perhaps

$$E\{\theta \mid x\} = \int_{\mathcal{T}} \theta p(x \mid \theta) g(\theta)\ d\theta \bigg/ f(x).$$

Efficient estimation requires some form of parametric modeling. There are two basic choices, modeling g or modeling f, each with advantages and disadvantages.

We begin with an example of f-modeling. A diffusion tensor imaging (DTI) study compared six dyslexic children with six normal controls at 15,443 brain locations, or voxels. Here we will analyze only the $N = 477$ voxels located at the extreme back of the brain. Each voxel produced a statistic z_i comparing dyslexics with normals — z_i playing the role of x_i in (4.29) — and a reasonable model for this is

$$z_i \sim \mathcal{N}(\theta_i, 1), \qquad i = 1, 2, \ldots, N = 477,$$

with θ_i the true "effect size" for voxel i. The investigators were hoping to pinpoint voxels having θ_i much different than 0.

The left panel of Figure 4.16 shows a histogram of the 477 z-values, and a smooth fitted curve obtained from a Poisson GLM regression, as in Section 3.3. Letting c_k be the centerpoint of bin $\mathcal{Z}_k$ and $\boldsymbol{c} = (c_1, c_2, \ldots, c_K)$,

$$\hat{\boldsymbol{f}} = \texttt{glm(}\boldsymbol{y} \sim \texttt{poly(}\boldsymbol{c}\texttt{,5),Poisson)\$fit}/N$$

estimates the marginal density $f(z)$; the smooth curve is $N \cdot \hat{\boldsymbol{f}}$.

Some familiar empirical Bayes results can be computed directly from $\hat{f}$ without any need to estimate the prior density $g(\theta)$. Two of these appear in the right panel of Figure 4.16: the solid black curve shows Tweedie's estimate (Section 1.5)

$$E\{\theta \mid z\} = z + \frac{d}{dz} \log \hat{f}(z),$$

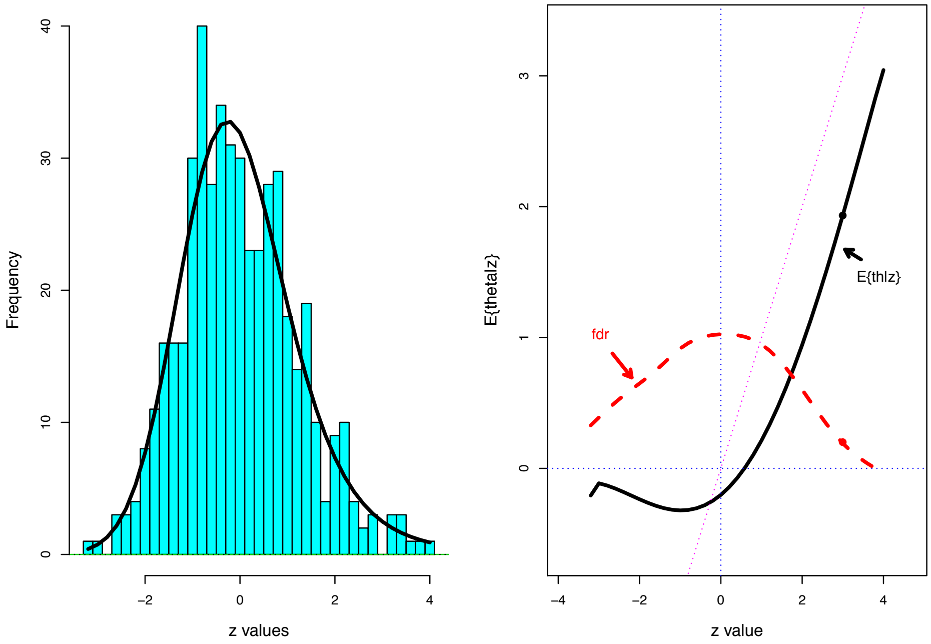


Figure 4.16: *Left panel* DTI data, $N = 477$; `poly(df=5)` fit. *Right panel* Tweedie estimate $E\{\theta \mid z\}$ at $z = 3 : E\{\theta \mid z\} = 1.93$, fdr $= 0.20$.

while the dashed red curve is the local false discovery rate (Section 1.6)

$$\widehat{\text{fdr}}(z) = \widehat{\Pr}\{\theta = 0 \mid z\} \doteq \phi(z)/\hat{f}(z),$$

$\phi(z) = \exp\{-z^2/2\}/\sqrt{2\pi}$.

Homework 4.32. What is an estimate of $E\{\theta \mid z = 3 \text{ and } \theta \neq 0\}$?

But what if we need to estimate something *not* in the small catalog of those having direct expressions in terms of $\hat{f}$? Reverting to the (θ_i, x_i) notation of (4.28)–(4.29), some examples might be

$$\Pr\{|\theta| > 2\} \quad \text{or} \quad E\{e^{-\theta} \mid x\}. \tag{4.32}$$

This is where g-modeling becomes essential. We assume that the prior density $g(\theta)$ belongs to a multiparameter exponential family, say $g_\beta(\theta)$, β an unknown p-dimensional parameter vector. Corresponding to each choice of β is a marginal density $f_\beta(x)$ (4.30). The schematic diagram (4.31) becomes

$$\beta \longrightarrow g_\beta \longrightarrow f_\beta \longrightarrow \boldsymbol{y} \sim \text{Mult}_K(N, \boldsymbol{f}_\beta). \tag{4.33}$$

Now we are dealing with a "hidden exponential family", $\theta \sim g_\beta(\cdot)$.

The disadvantage of g-modeling is that $x \sim f_\beta(\cdot)$ is *not* an exponential family,[7] making it more difficult to find the MLE $\hat{\beta}$. The advantage is that having found $\hat{\beta}$ we have a direct estimate $g_{\hat{\beta}}(\cdot)$ of the prior, and of all quantities such as (4.32).

[7] Except in the case where g_β represents a normal regression model and $x \mid \theta$ is also normal.

Here is a brief description of the MLE calculations, taken from Efron (2016), "Empirical Bayes deconvolution estimates", *Biometrika* 1–23. It simplifies the discussion to take $\mathcal{T}$, the θ sample space, to be discrete, say

$$\mathcal{T} = \{\theta_{(1)}, \theta_{(2)}, \dots, \theta_{(m)}\},$$

so that the prior $g(\theta)$ can be represented as an m-vector $\boldsymbol{g} = (g_1, g_2, \dots, g_m)$, $g_j = \Pr\{\theta = \theta_{(j)}\}$. As before, the x's will also be discretized, with sample space

$$\mathcal{X} = \{x_{(1)}, x_{(2)}, \dots, x_{(K)}\}$$

and marginal density $\boldsymbol{f} = (f_1, f_2, \dots, f_K)$. Defining the $K \times m$ matrix $\boldsymbol{P}$,

$$\boldsymbol{P} = (p_{kj}), \qquad p_{kj} = \Pr\{x = x_{(k)} \mid \theta = \theta_{(j)}\},$$

we have $f_k = \sum_j p_{kj} g_j$ or $\boldsymbol{f} = \boldsymbol{P}\boldsymbol{g}$.

The p-parameter exponential family for $\boldsymbol{g}_\beta$ is written as

$$\boldsymbol{g}_\beta = e^{\boldsymbol{Q}\beta - \phi(\beta)}, \tag{4.34}$$

$\boldsymbol{Q}$ an $m \times p$ matrix having jth row q_j'; that is,

$$g_{j\beta} = e^{g_j'\beta - \phi(\beta)} \qquad \left(\phi(\beta) = \log \sum_{l=1}^{m} e^{q_l'\beta}\right).$$

Then $\boldsymbol{y} \sim \text{Mult}_K(N, \boldsymbol{f}_\beta)$ will be a (K, p) *curved* exponential family,[8] with $\boldsymbol{y}$ in the simplex $\mathcal{S}_K$, as in Figure 4.3.

Model (4.33)–(4.34) yields simple expressions for the score function and Fisher information. Define $w_{kj}(\beta) = g_{j\beta}\{p_{kj}/f_{k\beta} - 1\}$ and let $W_k(\beta)$ be the m-vector

$$W_k(\beta) = (w_{k1}(\beta), w_{k2}(\beta), \dots, w_{km}(\beta))',$$

with $W_+(\beta) = \sum_1^K W_k(\beta)$. It turns out that

$$\begin{aligned} \dot{l}_\beta(\boldsymbol{y}) &= \boldsymbol{Q}' W_+(\beta) \boldsymbol{Q}, \\ I(\boldsymbol{y}) = -\ddot{l}_{\hat\beta}(\boldsymbol{y}) &= \boldsymbol{Q}' \left(\sum_{k=1}^{K} W_k\left(\hat\beta\right) y_k W_k\left(\hat\beta\right)' - \operatorname{diag}\left\{W_+\left(\hat\beta\right)\right\} \right) \boldsymbol{Q}, \\ \text{and} \qquad i_\beta &= \boldsymbol{Q}' \left(\sum_{i=1}^{K} W_k(\beta) \{N f_{k\beta}\} W_k(\beta)' \right) \boldsymbol{Q}, \end{aligned}$$

$\operatorname{diag}\{W_+(\hat\beta)\}$ denoting the diagonal matrix having entries $W_{+j}(\hat\beta)$.

[8] Notice the change of notation to (K, p) from (p, q) used previously.

Homework 4.33. Show that $w_{kj}(\beta) = g(\theta = \theta_{(j)} \mid x = x_{(k)}) - g(\theta = \theta_{(j)})$.

The MLE $\hat{\beta}$ can be found by direct numerical maximization of the log likelihood, employing say, the R algorithm `nlm`. (Making use of the previous expression for $\dot{\boldsymbol{l}}_\beta(\boldsymbol{y})$ accelerates the convergence rate.) Because $\boldsymbol{y} \sim \mathrm{Mult}_K(N, \boldsymbol{f}_\beta)$ is not an exponential family there is the possibility of multiple local maxima. Regularization, as in Section 4.7, greatly improves the performance of $\hat{\beta}$. In what follows, the penalty function for the pMLE will be

$$s_\beta = \left(\sum_{j=1}^{P} \beta_j^2 \right)^{1/2}.$$

Homework 4.34. Show that

$$\dot{s}_\beta = \frac{\beta}{\|\beta\|} \quad \text{and} \quad \ddot{s}_\beta = \frac{I - \frac{\beta\beta'}{\|\beta\|^2}}{\|\beta\|}.$$

G-model (4.34) was applied to the DTI data of Figure 4.16, $N = 477$, taking the θ sample space to be $\mathcal{T} = (-2.4, -2.2, \dots, 3.6)$, $m = 31$. The structure matrix $\boldsymbol{Q}$ for the hidden glm had $p = 8$ degrees of freedom,

$$\boldsymbol{Q} = [\delta_0, \texttt{poly}(\mathcal{T}, 7)], \tag{4.35}$$

δ_0 indicating a delta function at $\theta = 0$ — vector $(0, 0, \dots, 1, 0, \dots, 0)$ having 1 in the 13th place — and $\texttt{poly}(\mathcal{T}, 7)$ the $m \times 7$ matrix provided by the R function `poly`.

Specification (4.35) represents a "spike and slab" prior $g(\theta)$. It allows a spike of "null" cases at $\theta = 0$, i.e., zero difference betwen dyslexics and controls, and a smooth polynomial distribution for the non-null cases. The MLE $\hat{\beta}$ put weight 0.644 on the prior probability of 0, and estimated a mildly long-tailed density $\hat{g}(\theta)$ to the right of zero. This is indicated by the solid red curve in panel A of Figure 4.17, pictured against a histogram of the z-values.

An important question: is the pMLE estimation process stable for g-modeling? To this end, a calculation of the stable region $\mathcal{R}_{\hat{\beta}}$ was carried out for the DTI data. The computations were done after transformation to standardized coordinates (4.21),

$$\mu_{\hat{\beta}} = 0 \quad \text{and} \quad V_{\hat{\beta}} = I_8.$$

We wish to compute the boundary $\mathcal{B}_{\hat{\beta}}$ as in Figure 4.14, now in $\overset{\perp}{\mathcal{M}}_{\hat{\beta}}$ (4.28), a 29-dimensional hyperplane. ($K = 37$, the number of bins in the panel A histogram, minus $p = 8$ equals 29.)

With $p = 8$, as opposed to $p = 2$ in the toy example of Figure 4.15, choosing the vectors u for the one-parameter bounding families $\mathcal{F}_u$ becomes challenging. For this analysis, 5000 u vectors were chosen randomly and uniformly from the surface of the unit sphere $\mathcal{S}^8$ in $\mathcal{R}^8$. Each u yielded a direction vector w_u and a bounding distance d_u, as portrayed in Figure 4.15. Panel B of Figure 4.17

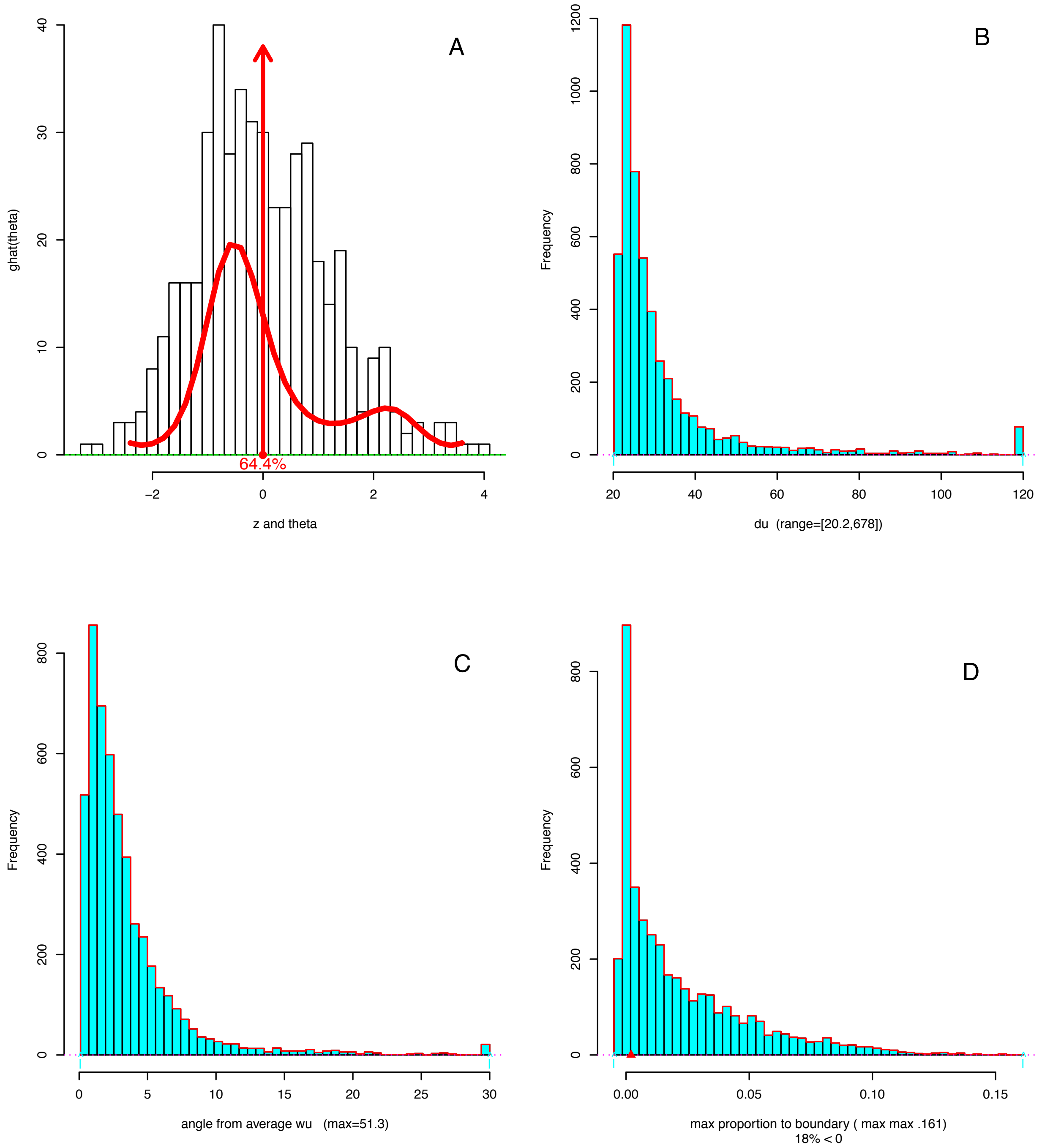


Figure 4.17: Empirical Bayes g-modeling analysis of the DTI data of Figure 4.16. Panel A indicates fitted slab and spike prior in red. Panels B, C, and D refer to stability calculations for the g-model fitting algorithm as described in the text.

is a truncated histogram of the 5000 d_u values,

$$20.2 \leq d_u \leq 678. \tag{4.36}$$

The minimum of d_u over *all* vectors on $\mathcal{S}^8$ — found by direct numerical minimization — was 20.02, suggesting that our 5000 random choices are enough to provide a good estimate of $\mathcal{B}_{\hat{\beta}}$.

Even though the u vectors pointed in all directions on $\mathcal{S}^8$, the direction vectors w_u clustered closely around their average vector $\bar{w}$. Panel C shows the histogram of angular distances in degrees between the 5000 w_u's and $\bar{w}$: 95% of them are less than 10 degrees.[9]

The stable region $\mathcal{R}_{\hat{\beta}}$ has its boundary more than 20 standard Mahalanobis units from $\mu_{\hat{\beta}}$. Is this sufficiently distant to rule out unstable behavior? As a check, 4000 parametric bootstrap Y^* were generated,

$$Y_i^* \sim \text{Mult}_{37}(477, f_{\hat{\beta}}), \qquad i = 1, 2, \ldots, 4000,$$

and then standardized and projected into vectors y_i^* in the 29-dimensional space $\overset{\perp}{\mathcal{M}}_{\hat{\beta}}$. Each y_i^* yielded a worst-case value,

$$m_i^* = \max_h \{ y_i^{*'} w_{uh} / d_{uh}, h = 1, 2, \ldots, 5000 \};$$

$m_i^* > 1$ would indicate that y_i^* fell outside the stable region $\mathcal{R}_{\hat{\beta}}$.

Homework 4.35. Why is that last statement true?

In fact, none of the 4000 m_i^*'s exceeded 0.16, so none of the y_i^*'s fell anywhere near the boundary of $\mathcal{R}_{\hat{\beta}}$. For the actual observation $\boldsymbol{y}$, m equaled 0.002, locating it quite near to $\mu_{\hat{\beta}}$. Observed and expected Fisher information are almost the same, and probably would not vary much for other possible observation vectors y^*.

Panel D of Figure 4.17 shows the histogram of the 4000 m_i^* values; 18% of them had $m_i^* < 0$. These represent y_i^* vectors having negative correlation with all 5000 w_{uh} direction vectors, implying that $\mathcal{R}_{\hat{\beta}}$ is open (going off to infinity) in some direction, just as suggested in Figure 4.14.

Homework 4.36. Draw a schematic picture supporting this last conclusion.

4.9 The proportional hazards model D.R. Cox (1972), *JRSS-B* 187–220; (1975), *Biometrika* 269–276

We return to the analysis of censored data (Section 3.6) but now in the more useful context of regression models. The data for subject i is a triple set,[10] (T_i, d_i, x_i), $i = 1, 2, \ldots, N$, where T_i is a non-negative observed lifetime, x_i is a p-vector of observed covariates, and d_i equals 1 or 0 as subject i was or was not observed to die.

[9] A spherical cap of radius 10° on the surface of a 29-dimensional sphere covers only proportion 3.9×10^{-23} of the "area" of the full sphere.

[10] Here we will use notation more standard in the survival analysis literature.

Table 4.2: Gehan survival data (partly artificial) from Chapter 12 of Venables and Ripley (1998).

	T	d	age	first	treat
1	1	1	71	2	0
2	10	1	56	22	1
3	22	1	38	6	0
4	7	1	9	36	1
5	3	1	37	28	0
6	32	0	48	39	1
7	12	1	36	25	0
8	23	1	43	33	1
9	8	1	52	40	0
10	22	1	34	32	1
11	17	1	27	34	0
12	6	1	30	3	1
13	2	1	6	24	0
14	16	1	60	26	1
15	11	1	21	7	0
16	34	0	56	11	1
17	8	1	24	19	0
18	32	0	82	18	1
19	12	1	44	31	0
20	25	0	37	1	1
21	2	1	14	5	0
22	11	0	33	15	1
23	5	1	69	9	0
24	20	0	50	27	1
25	4	1	44	30	0
26	19	0	27	14	1
27	15	1	29	16	0
28	6	1	28	38	1
29	8	1	72	10	0
30	17	0	63	8	1
31	23	1	61	42	0
32	35	0	55	29	1
33	5	1	21	4	0
34	6	1	12	13	1
35	11	1	13	37	0
36	13	1	25	21	1
37	4	1	38	12	0
38	9	0	35	17	1
39	1	1	9	23	0
40	6	0	24	20	1
41	8	1	76	41	0
42	10	0	14	35	1

If all the d_i equaled 1, that is, if there was no censored data, we could do a standard regression analysis of T on x. The proportional hazards model allows us to proceed in the face of censoring. There is a connection with generalized linear models, but that won't be apparent for a while.

Table 4.2 shows the Gehan data, a partially artificial leukemia-survival data set. Survival time T is in months; x_i includes three covariates: **age** in years, **first** the month of first treatment (measured from when the study began), and **treat** the treatment indicator, 1 for the new treatment and 0 for the control. We wish to learn the efficacy of the new treatment, as well as the possible effects of **age** and **first**.

The lifetime T_i of subject i is assumed to follow density $f_i(t)$, $t \geq 0$, with cdf $F_i(t) = \int_0^t f_i(s)\,ds$ and *hazard rate*

$$h_i(t) = f_i(t) \big/ \left[1 - F_i(t)\right]$$

so $\Pr\{T_i \leq t + dt \mid T_i \geq t\} \doteq h(t)dt$. The proportional hazards model assumes that the hazard rates are related as follows:

$$h_i(t) = h_0(t) e^{x_i'\beta}. \tag{4.37}$$

Here β is an unknown $p \times 1$ parameter vector, while $h_0(t)$ is a *baseline hazard rate* that does not need to be specified.

Homework 4.37. Let $H_i(t)$ be the *cumulative hazard rate* $\int_0^t h_i(s)\,ds$, and $S_i(t) = 1 - F_i(t)$, the *survival function.*

(a) Show that $S_i(t) = e^{-H_i(t)}$.

(b) Denoting $e^{x_i\beta} = \alpha_i$, show that $S_i(t) = S_0(t)^{\alpha_i}$ (a relationship known as "Lehmann alternatives").

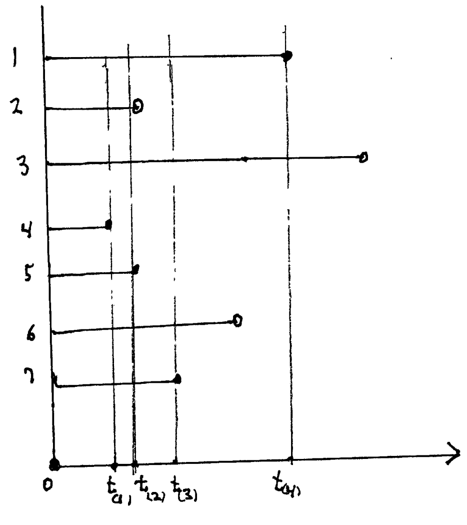
Let J be the total number of deaths observed, say at times $t_{(1)} < t_{(2)} < \cdots < t_{(j)} < \cdots < t_{(J)}$, and assuming no ties for convenience. The *risk set* $\mathcal{R}_j$ for event j is

$$\mathcal{R}_j = \{\text{subjects under observation at time } t_{(j)}\}.$$

In this little example there are $N = 7$ subjects, $J = 4$ of whom were observed to die and 3 were lost to follow-up (the open circles). $\mathcal{R}_3$ equals $\{1, 3, 6, 7\}$, for instance. We also denote

$$i_j = \{\text{index of subject who died at time } t_{(j)}\},$$

$i_1 = 4$, $i_2 = 5$, $i_3 = 7$, $i_4 = 1$ in the example.



A simple but crucial result underlies the proportional hazards method.

Lemma 8. *Under the proportional hazards model, the probability that $i_j = i$, i.e., that the death occurred to member i of $\mathcal{R}_j$, is*

$$\pi_i(\beta \mid \mathcal{R}_j) = \frac{e^{x_i'\beta}}{\sum_{k\in\mathcal{R}_j} e^{x_k'\beta}}. \tag{4.38}$$

Homework 4.38. Verify Lemma 8.

Table 4.3: `coxph` output for the Gehan data.

	$\hat\beta$	se($\hat\beta$)	z-value	p-value
age	$-.021$	.012	-1.75	.081
first	$-.006$	.014	$-.43$	.660
treat	-1.52	.420	-3.63	.000 $**$

Partial likelihood

Cox (1972, 1975) suggested using the *partial likelihood*

$$L(\beta) = \prod_{j=1}^{J} \frac{e^{x_{i_j}'\beta}}{\sum_{\mathcal{R}_j} e^{x_k'\beta}} = \prod_{j=1}^{J} \pi_{ij}(\beta \mid \mathcal{R}_j) \tag{4.39}$$

as if it were the true likelihood for the unknown parameter β. It is "partial" because it ignores all the non-events, times when nothing happened or there were losses to follow-up. Nevertheless, it can be shown to be quite efficient under reasonable assumptions.

An excellent program, `coxph`, is available in the R package `survival`. To apply `coxph` to the Gehan data of Table 4.2, first apply `Surv`, $S = \texttt{Surv}(TT, d)$ ($TT = T$ as T is not an allowable variable name) then execute

$$\text{result} = \texttt{coxph}(S \sim \text{age} + \text{first} + \text{treat}).$$

This produced the results in Table 4.3. **Treat** was strongly significant. Its negative regression coefficient $\hat\beta = -1.52$ shows that the new treatment, **treat** $= 1$, decreases the hazard rate and increases survival as in Homework 4.37(a). **Age** is only borderline significant, and **first** is insignificant.

Homework 4.39. Run $\texttt{coxph}(S \sim \text{age} + \text{treat})$, $\texttt{coxph}(S \sim \text{treat})$, and $\texttt{coxph}(S \sim \text{age})$. What are your conclusions?

The log partial likelihood $l(\beta) = \log L(\beta)$ is

$$l(\beta) = \sum_{j=1}^{J} \left(x_{ij}'\beta - \log \sum_{\mathcal{R}_j} e^{x_i'\beta} \right).$$

Taking derivatives with respect to β gives, in the notation of (4.38),

$$\begin{aligned}
\underset{p\times 1}{\dot{l}(\beta)} &= \sum_{j=1}^{J} \left(x_{ij} - E_j(\beta)\right) \qquad \text{where } E_j(\beta) = \sum_{\mathcal{R}_j} x_i \pi_i(\beta \mid \mathcal{R}_j); \\
-\underset{p\times p}{\ddot{l}(\beta)} &= \sum_{j=1}^{J} V_j(\beta) \qquad \text{where } V_j(\beta) = \sum_{\mathcal{R}_j} \pi_i(\beta \mid \mathcal{R}_j)\left(x_i - E_j(\beta)\right)\left(x_i - E_j(\beta)\right)'.
\end{aligned} \tag{4.40}$$

Homework 4.40. (a) Verify (4.40). (b) Show that $l(\beta)$ is a concave function of β.

The partial likelihood estimnate of β is defined by

$$\hat{\beta} : \dot{l}\left(\hat{\beta}\right) = \mathbf{0},$$

$\mathbf{0}$ a vector of p zeros, with approximate observed information matrix

$$\hat{I} = -\ddot{l}\left(\hat{\beta}\right).$$

These gave the estimates in Table 4.3, the square roots of the diagonal elements of $\hat{I}$ being the tabled standard errors. Considerable theoretical effort has gone into verifying the asymptotic normal approximation

$$\hat{\beta} \dot{\sim} \mathcal{N}_p\left(\beta, \hat{I}^{-1}\right).$$

GLM connection

Let $n_j = |\mathcal{R}_j|$, the number of subjects in risk set $\mathcal{R}_j$. Expression (4.38) represents a multinomial exponential family on n_j categories, with $\eta_l = x_l'\beta$ in (2.7). The proportional hazards model amounts to a generalized linear model (Part 3) where the component observations are multinomials of varying sizes n_j.

Multicategory logistic regression A multicategory version of logistic regression (Section 3.2) involves independent observations y_i, $i = 1, 2, \dots, N$, each of which falls into one of L categories: for instance, $L = 3$ with categories {Democrat, Republican, Independent}. A GLM model takes

$$\Pr\{y_i \text{ in category } l\} = \pi_i(\beta, l) = \frac{e^{x_{il}'\beta}}{\sum_{k=1}^{L} e^{x_{ik}'\beta}}, \tag{4.41}$$

the x_{il} being known covariate vectors.

Homework 4.41. (a) Describe the MLE analysis of model (4.41).

(b) Show how this reduces to the logistic regression model of Section 3.2.