Part 3

Generalized Linear Models

3.1 *Exponential family regression models* (pp 62–66) Natural parameter regression structure; MLE equations; geometric picture; two useful extensions

3.2 *Logistic regression* (pp 66–69) Binomial response models; transplant data; probit analysis; linkages

3.3 *Poisson regression* (pp 69–72) Poisson GLMs; galaxy data and truncation

3.4 *Lindsey's method* (pp 72–73) Densities as exponential families; discretization; Poisson solution for multinomial fitting

3.5 *Analysis of deviance* (pp 73–76) Nested GLMs; deviance additivity theorem; analysis of deviance tables; applied to prostate study data

3.6 *A survival analysis example* (pp 76–81) NCOG data; conesored data; hazard rates; life tables; Kaplan–Meier curves; Greenwood's formula; logistic regression and hazard rate analysis

3.7 *The proportional hazards model* (pp 81–87) Continuous hazard rates; the proportional hazards model; partial likelihood; pediatric abandonment study; risk sets; exponential family connections; multinomial GLMs

3.8 *Overdispersion and quasi-likelihood* (pp 87–92) Toxoplasmosis data; deviance and Pearson measures of overdispersion; extended GLMs; quasi-likelihood models

3.9 *Double exponential families* (pp 92–98) Double family density $f_{\mu,\theta,n}(\bar{y})$; constant $C(\mu,\theta,n)$; double Poisson and negative binomial; score function and MLEs; double family GLMs

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

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

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

$$\eta_i = x_i^\top \beta \qquad \text{for } i = 1, 2, \ldots, N,$$

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

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

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

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

where

- β is the $p \times 1$ natural parameter vector;
- $z = X^\top \boldsymbol{y}$ is the $p \times 1$ sufficient vector;
- $\phi(\beta) = \sum_{i=1}^N \psi(x_i^\top \beta)$ 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 $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$ and $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_N)^\top$; likewise $\boldsymbol{\mu}$ for the expectation vector $(\mu_1, \mu_2, \ldots, \mu_N)^\top$, $\mu_i = \dot{\psi}(\eta_i)$, or more succinctly, $\boldsymbol{\mu} = \dot{\psi}(\boldsymbol{\eta})$. Similarly, $\boldsymbol{V}$ will denote the $N \times N$ diagonal matrix of variances, written $\boldsymbol{V} = \ddot{\psi}(\boldsymbol{\eta}) = \text{diag}(\ddot{\psi}(\eta_i))$, with $\boldsymbol{V}_\beta = \text{diag}(\ddot{\psi}(x_i^\top \beta))$ indicating the variance matrix for parameter vectors $\boldsymbol{\eta} = X\beta$ in the GLM.

Homework 3.1. Show that

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

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

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

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

$$\dot{l}_\beta(\boldsymbol{y}) = z - E_\beta\{z\} = X^\top(\boldsymbol{y} - \boldsymbol{\mu}),$$

with

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

the $p \times p$ Fisher information matrix.

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

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

$\mathbf{0}$ 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

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

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

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

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,

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

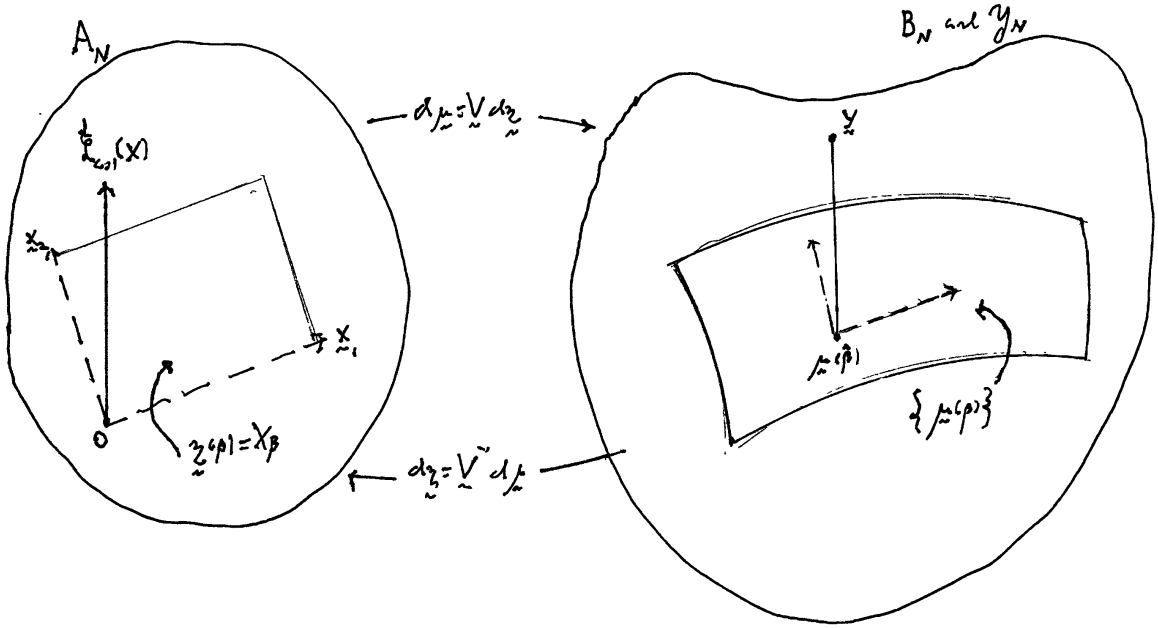
as pictured next. (In the linear regression situation, $y_i \overset{\text{ind}}{\sim} \mathcal{N}(x_i^\top\beta, \sigma^2)$, $\{\boldsymbol{\mu}(\beta)\}$ is flat, but this is essentially the only such case.)

The data vector $\boldsymbol{y}$ will usually *not* lie in the curved manifold $\{\boldsymbol{\mu}(\beta)\}$. From the MLE equations $X^\top(\boldsymbol{y} - \boldsymbol{\mu}(\hat{\beta})) = 0$ 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 $\mathcal{L}_{\rm col}(X)$ be the column space of X,

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

the residual $\boldsymbol{y} - X\hat{\beta}$ must lie in the space $\overset{\perp}{\mathcal{L}}_{\rm col}(X)$ of N-vectors orthogonal to $\mathcal{L}_{\rm col}(X)$. Notice that the *same* space $\overset{\perp}{\mathcal{L}}_{\rm col}(\boldsymbol{X})$ applies to MLE estimation for all choices of $\boldsymbol{y}$.



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

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

(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

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

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

Warning The resulting approximate confidence intervals, e.g., $\hat{\beta}_j \pm 1.96\widehat{\rm se}_j$, 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

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

Homework 3.2. Write the offset situation in the form $g_\beta(\boldsymbol{y}) = e^{\beta^\top z - \phi(\beta)} g_0(\boldsymbol{y})$. 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,

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

with

$$\eta_i = x_i^\top \beta, \qquad i = 1, 2, \ldots, N.$$

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

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

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 $\operatorname{diag}(n_i V_i(\beta))$.

(a) Show that

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

(b) Also show that

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

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, binomials with sample size 1),

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

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

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

The Fisher information matrix for β is

$$i_\beta = \boldsymbol{X}^\top \operatorname{diag}\left(\pi_i(1-\pi_i)\right) \boldsymbol{X}.$$

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.)

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***

The logistic fit gave a predicted probability of infection

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

for each patient. The top panel of Figure 3.1 compares the predictions for the 199 patients who did 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 $\dot{h}(\beta) = (\cdots \partial h(\beta)/\partial\beta_j \cdots)^\top$. Then the approximate standard error assigned to the MLE $\hat{\zeta} = h(\hat{\beta})$ is usually

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

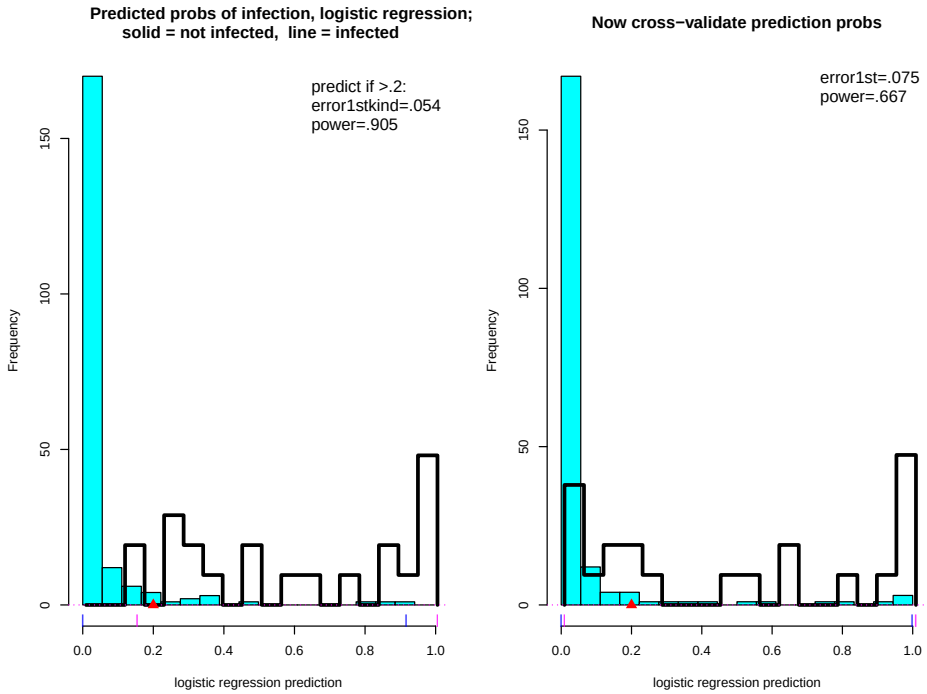


Figure 3.1: Predicted probabilities of infection, logistic regression; solid histogram represents not infected, line histogram represents infected.

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

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

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

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

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

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 Homework 3.3, with $\mu_i = \pi_i$. From Homework 3.3(b) we get

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

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, $i = 1, 2, \ldots, K$, and the proportion

p_i of deaths observed. (A customary goal is to estimate "LD50", the dose yielding 50% lethality.) The *probit model* is

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

where Φ is the standard normal cdf; maximum likelihood is used to solve for (β_0, β_1). Another way to say this is that each mouse has individual tolerance t for the drug, with $t \sim \mathcal{N}(-b_0/b_1, 1/b_1)$ for the population, and that dose d_i kills all mice with $t < d_i$.

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

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

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

$$\eta_i = a_i + x_i^\top \beta, \qquad i = 1, 2, \ldots, N,$$

where β is an unknown p-dimensional parameter vector, x_i a known p-dimensional covariate vector, and a_i a known scalar "offset". (Offsets are necessary if the counts y_i are obtained under different conditions — for example, using receptors of different, but known, sensitivities — in which case a_i would relate to the sensitivity of receptor i. Offsets are distinct from an intercept, which would be specified in x_i if desired.)

The MLE equation (3.3) is

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

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

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

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.

↑ Magnitude (dimmer)	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
10	1	0	0	2	2	2	1	2	0	0	0	1	2	0	0
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.

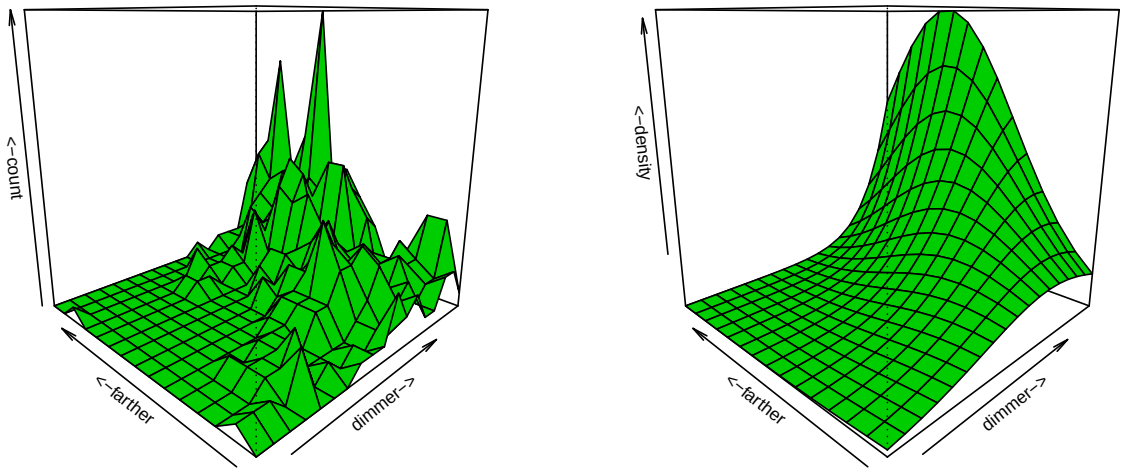


Figure 3.2: *Left panel*: galaxy data, binned counts. *Right panel*: Poisson GLM density estimate.

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,

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

Let $\boldsymbol{m}$ be the 270-vector listing the m_{ij} values in some order, say $\boldsymbol{m}$ in order $(18, 17, \ldots, 1)$ 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.

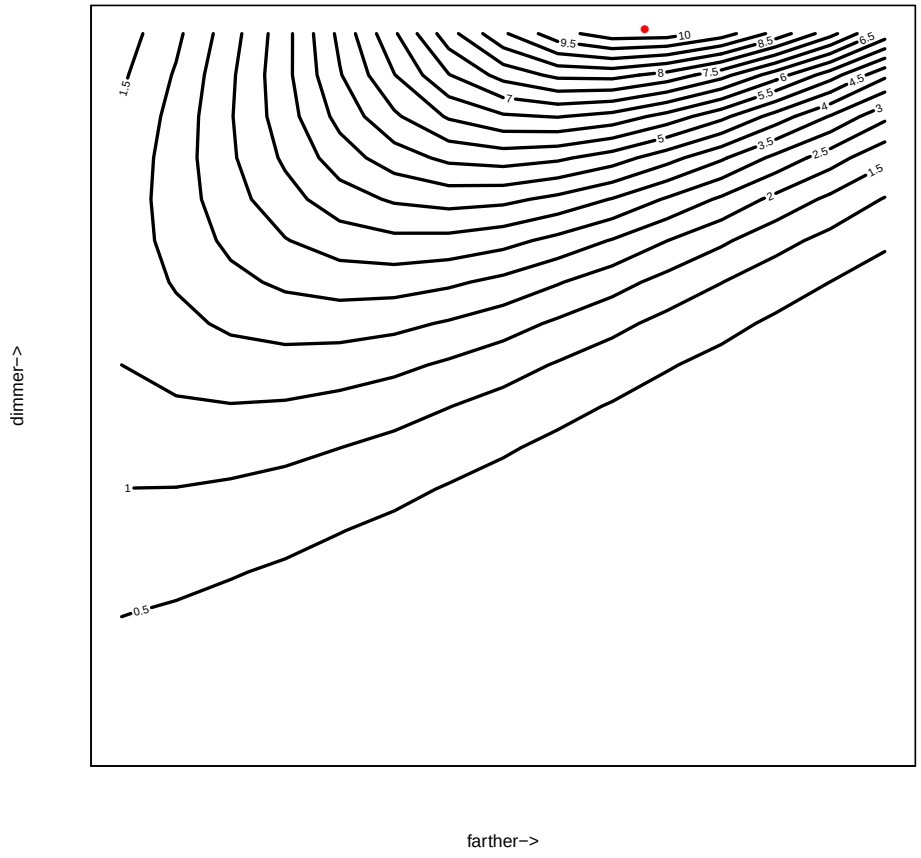


Figure 3.3: Contour curves for Poisson GLM density estimates for the galaxy data; dot shows point of maximum density.

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 estimated 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 (1.10). Can you detect any hints of poor fit?

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

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 $z_1, z_2, \ldots, z_N$ 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,

$$g_\beta(z) = e^{\beta^\top t(z) - \psi(\beta)} g_0(z), \tag{3.6}$$

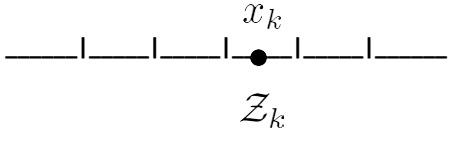
where β and $t(z)$ are in $\mathcal{R}^p$. In Figure 1.3, $t(z)$ was a fifth-degree polynomial

$$t(z) = (z, z^2, z^3, z^4, z^5).$$

(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

$$\pi_k(\beta) = \Pr_\beta\{z \in \mathcal{Z}_k\} = \int_{\mathcal{Z}_k} g_\beta(z)\, dz \doteq \Delta e^{\beta^\top t_k - \psi(\beta)}, \tag{3.7}$$

$t_k = t(x_k)$, and

$$\boldsymbol{\pi}(\beta) = (\pi_1(\beta), \pi_2(\beta), \ldots, \pi_K(\beta)).$$

Also let $\boldsymbol{y} = (y_1, y_2, \ldots, y_K)^\top$ 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

$$\mu_k(\beta_0, \beta) = e^{\beta_0 + \beta^\top t_k}, \tag{3.8}$$

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 invoke the Poisson trick and use standard GLM software to find the Poisson MLE $(\hat{\beta}_0, \hat{\beta})$ in model (3.8),

$$\boldsymbol{y} \sim \text{Poi}\left(\boldsymbol{\mu}(\beta_0, \beta)\right);$$

since $\log \mu_k(\beta_0, \beta) = \beta_0 + \beta^\top t_k$, this is a Poisson GLM, solvable directly in R.

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 GLMs

Suppose that in the original model

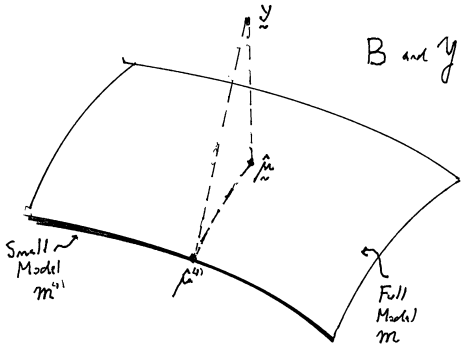
$$\underset{N\times 1}{\boldsymbol{\eta}} = \underset{N\times p}{X}\ \underset{p\times 1}{\beta},$$

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 $\boldsymbol{\eta} = X\beta$, say $\mathcal{M}^{(1)}$ as opposed to the full model $\mathcal{M}$, with

- $\hat{\beta}^{(1)}$ = 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)

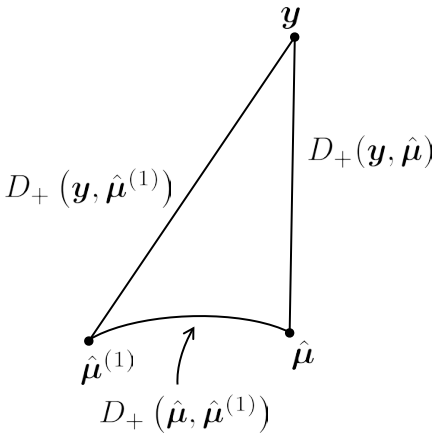
For standard normal regression theory (OLS), $\mathcal{M}$ and $\mathcal{M}^{(1)}$ are flat spaces, and deviance is Euclidean squared distance. Pythagoras' theorem says that

$$D_+(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}^{(1)}) = D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}^{(1)}) - D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}). \tag{3.9}$$

The *deviance additivity theorem* says that (3.9) holds for any GLM, as discussed next.

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 for $H_0 : \beta^{(2)} = 0$ If H_0 is true then Hoeffding's formula, Section 2.7, and Wilks' theorem say 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

$$D_+(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}^{(1)}) = \sum_{i=1}^{N} D(\hat{\mu}_i, \hat{\mu}_i^{(1)}),$$

if $D_+(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}^{(1)})$ is significantly too large 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 now β and X are divided into J parts,

$$\underset{p\times 1}{\beta} = \left(\beta^{(1)}_{p^{(1)}}, \beta^{(2)}_{p^{(2)}}, \ldots, \beta^{(J)}_{p^{(J)}}\right) \quad \text{and} \quad \underset{N\times p}{X} = \left(\underset{N\times p^{(1)}}{X^{(1)}}, \underset{N\times p^{(2)}}{X^{(2)}}, \ldots, \underset{N\times p^{(J)}}{X^{(J)}}\right).$$

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.

Table 3.3: Analysis of deviance table.

MLE		Twice max log like		Difference	Compare with
$\hat{\beta}(0) = \bar{y}$	$\longrightarrow$	$2l_{\hat{\beta}(0)}$			
			$\searrow$		
			$\nearrow$	$D_+(\hat{\mu}(1), \hat{\mu}(0))$	$\chi^2_{p^{(1)}}$
$\hat{\beta}(1)$	$\longrightarrow$	$2l_{\hat{\beta}(1)}$			
			$\searrow$		
			$\nearrow$	$D_+(\hat{\mu}(2), \hat{\mu}(1))$	$\chi^2_{p^{(2)}}$
$\hat{\beta}(2)$	$\longrightarrow$	$2l_{\hat{\beta}(2)}$			
			$\searrow$		
$\vdots$		$\vdots$		$\vdots$	$\vdots$
			$\nearrow$	$D_+(\hat{\mu}(J), \hat{\mu}(J-1))$	$\chi^2_{p^{(J)}}$
$\hat{\beta}(J) = \hat{\beta}$	$\longrightarrow$	$2l_{\hat{\beta}(J)}$			

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

$$\texttt{glm(}\boldsymbol{y} \sim \texttt{ poly(}\boldsymbol{x}, j\texttt{), poisson)}$$

for $j = 2, 3, \ldots, 8$, with $\boldsymbol{x}$ the vector of bin centers in Figure 1.3,

$$\boldsymbol{x} = (-4.4, -4.2, -4.0, \ldots, 5.0, 5.2),$$

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 $\texttt{glm(}\boldsymbol{y} \sim \texttt{ poly(}x\texttt{,df), poisson)}$.

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. Akaike's information criterion (AIC) suggests a penalty for increased model size,

$$\text{AIC}^{(j)} = D_+^{(j)} + 2j,$$

a nearly unbiased estimate of the true expected log likelihood for model j; 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.

Table 3.6: Arm A of NCOG head and neck cancer study, binned by month: n = number at risk at beginning of month, 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						

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,

$$y_i \mid n_i \sim \text{Bi}(n_i, h_i). \tag{3.11}$$

[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.

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

$$\hat{h}_i = y_i/n_i; \tag{3.12}$$

(3.10) then gives the survival estimate

$$\hat{S}_i = \prod_{j=1}^{i-1} \left(1 - \hat{h}_j\right) \tag{3.13}$$

(so $\hat{S}_1$ is the estimated probability of *not* dying in the first month following treatment).

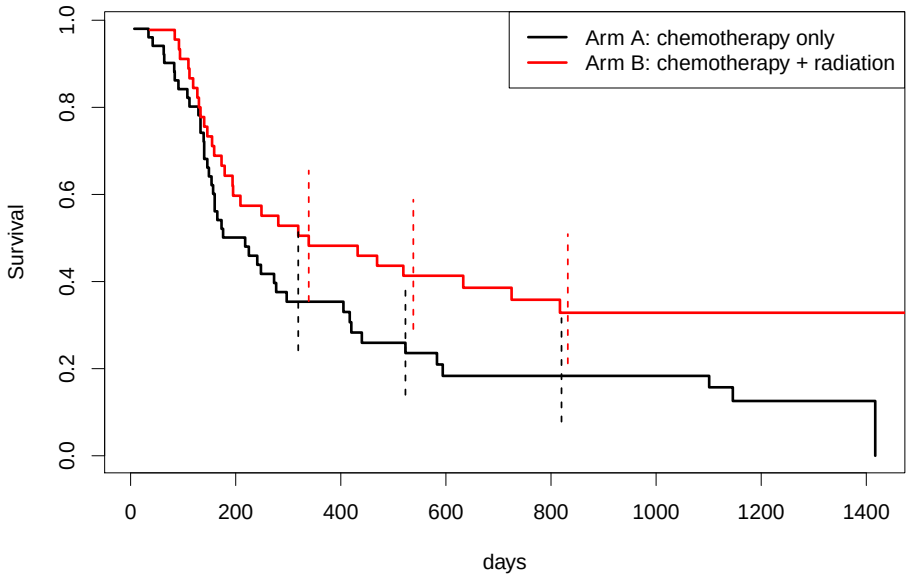


Figure 3.4: NCOG estimated survival curves; lower is Arm A (chemotherapy only); upper is Arm B (chemotherapy+radiation). Vertical lines indicate approximate 95% confidence intervals.

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 $t_{(i)}$?

The binomial model[2] (3.11) leads to *Greenwood's formula*, an approximate standard error for $\hat{S}_i$,

$$\mathrm{sd}\left\{\hat{S}_i\right\} \doteq \hat{S}_i \left(\sum_{j\le i} \frac{y_j}{n_j(n_j-y_j)}\right)^{1/2}.$$

The vertical bars in Figure 3.4 indicate $\pm 1.96\,\mathrm{sd}_i$, 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 model. 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 $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_N)^\top$ satisfies

$$\boldsymbol{\eta} = X\beta \tag{3.14}$$

as in Section 3.1–Section 3.2.

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

$$x_i = \left[1, i, (i-11)_+^2, (i-11)_+^3, \right]^\top, \tag{3.15}$$

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 parametric calculations in Figure 3.5, including the estimated standard errors.

[2] It is assumed that the conditional binomial distributions (3.11) are successively independent of the previous observations (n_j, y_j, l_j), $j < i$.

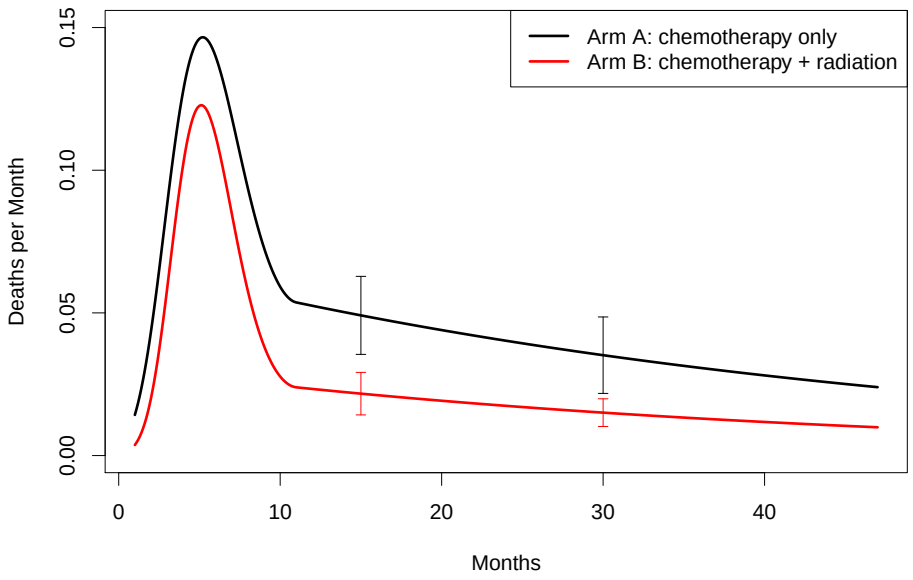


Figure 3.5: Parametric hazard rate estimates for NCOG study. Arm A (black curve) has about 2.5 times higher hazard than Arm B (red curve) for all times more than a year after treatment. Standard errors shown at 15 and 30 months.

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 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 that include covariate information. The data for subject i is a triple,[3] (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. 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.

Here we will work with *continuous*, as opposed to discrete, hazard rates. The lifetime T_i for subject i is assumed to follow density $f_i(t)$, $t \geq 0$, with survival function $S_i(t) = \int_t^\infty f_i(s)\,ds$ and hazard rate $h_i(t) = f_i(t)/S_i(t)$.

Homework 3.18. Show that $S_i(t) = e^{-H_i(t)}$ where $H_i(t) = \int_0^t h_i(s)\,ds$. How does this relate to formula (3.10)?

[3]Here we will use notation more standard in the survival analysis literature.

The *proportional hazards model* assumes that each $h_i(t)$ is proportional to a "baseline hazard rate" $h_0(t)$ multiplied by a factor that depends on the covariate vector x_i,

$$h_i(t) = h_0(t)e^{x_i^\top \beta}. \tag{3.16}$$

Here β is an unknown $p \times 1$ parameter vector. It will turn out that $h_0(t)$ does not need to be specified in a proportional hazards analysis, leaving only the regression coefficient vector β to be estimated.

Homework 3.19. Denoting $e^{x_i^\top \beta} = \alpha_i$, show that $S_i(t) = S_0(t)^{\alpha_i}$ (a relationship known as "Lehmann alternatives"), where $S_0(t)$ is the baseline survival function.

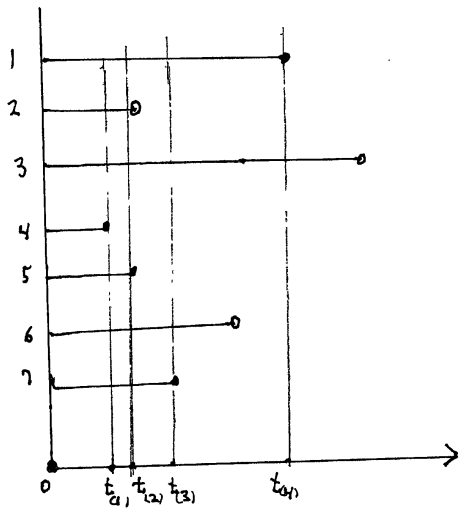
The key idea of proportional hazards analysis is to condition the occurrence of each observed event on the *risk set* of subjects under observation just before the event occurred. Let J be the total number of deaths observed; that is, cases with $d_i = 1$, say at times $t_{(1)} < t_{(2)} < \cdots < t_{(j)} < \cdots < t_{(J)}$ (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 3. *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*

$$\pi_i(\beta \mid \mathcal{R}_j) = \frac{e^{x_i^\top \beta}}{\sum_{k \in \mathcal{R}_j} e^{x_k^\top \beta}}. \tag{3.17}$$

Homework 3.20. Verify Lemma 3.

Partial likelihood

Cox (1972, 1975) suggested using the *partial likelihood*

$$L(\beta) = \prod_{j=1}^{J} \frac{e^{x_{i_j}^\top \beta}}{\sum_{\mathcal{R}_j} e^{x_k^\top \beta}} = \prod_{j=1}^{J} \pi_{i_j}(\beta \mid \mathcal{R}_j) \tag{3.18}$$

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 (Efron 1977, *JASA* 557–565).

The log partial likelihood $l(\beta) = \log L(\beta)$ is

$$l(\beta) = \sum_{j=1}^{J} \left(x_{i_j}^\top \beta - \log \sum_{\mathcal{R}_j} e^{x_i^\top \beta} \right).$$

Taking derivatives with respect to β gives, in the notation of (3.17),

$$\begin{aligned} \underset{p\times 1}{\dot{l}(\beta)} &= \sum_{j=1}^{J} \left(x_{i_j} - E_j(\beta) \right) \qquad \text{where } E_j(\beta) = \sum_{\mathcal{R}_j} x_i \pi_i(\beta \mid \mathcal{R}_j); \\ \underset{p\times p}{-\ddot{l}(\beta)} &= \sum_{j=1}^{J} V_j(\beta) \qquad \text{where } V_j(\beta) = \sum_{\mathcal{R}_j} \pi_i(\beta \mid \mathcal{R}_j) \left(x_i - E_j(\beta) \right) \left(x_i - E_j(\beta) \right)^\top. \end{aligned} \tag{3.19}$$

Homework 3.21. (a) Verify (3.19). (b) Show that $l(\beta)$ is a concave function of β.

The partial likelihood estimate 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

$$\hat{I} = -\ddot{l}\left(\hat{\beta}\right).$$

Considerable theoretical effort has gone into verifying the asymptotic normal approximation

$$\hat{\beta} \dot{\sim} \mathcal{N}_p\left(\beta, \hat{I}^{-1}\right).$$

Table 3.7 shows a small portion of the data from the *pediatric abandonment study*. Over a twelve-year period, $n = 1620$ children were treated for cancer and other serious diseases at a medical facility in a developing country. The investigators wished to identify the factors impacting treatment abandonment. The response variable was time, the number of days from entrance to last observation. Only one-tenth of the cases were abandoned ($d = 1$), which was a good thing for the children but meant that nine-tenths of the data was censored.

Table 3.7: 40 randomly selected children from 1620 in pediatric abandonment study. Sex: female = 1, male = 2; race: Ladino = 1, Indigina = 2; diag: leuk = 1, lymph = 2, solid = 3; age: at admission; enter: entry date in days since 01/01/2000; far: distance in kms from home to study facility; time: days from entrance to last observation; d: abandonment observed = 1, not observed = 0.

child	sex	race	diag	age	enter	far	time	d
1272	1	1	1	15.4	1712	118	2319	0
916	2	1	1	3.92	223	0	3626	0
180	2	1	1	10.9	848	74	3530	0
663	1	1	3	3.92	2856	126	562	0
541	2	1	3	6.17	2238	123	2218	0
1026	2	1	3	6.67	742	74	3115	0
892	1	1	1	4.83	3022	18	1454	0
1182	1	2	1	5.33	1191	32	234	1
18	1	1	3	10.8	2807	42	1616	0
1255	1	2	3	1.92	1273	32	182	1
249	2	1	3	3.33	1041	151	2441	0
1037	1	1	1	11.4	127	92	152	0
673	1	1	1	7.92	2734	108	1690	0
836	2	1	1	11.6	2351	52	1678	0
894	1	1	1	7.5	195	4	150	0
748	2	2	1	12.8	2918	52	1583	0
1197	1	1	1	10.9	1826	136	155	0
30	1	1	2	6.33	1666	0	373	1
1074	1	1	3	1	1506	208	467	0
971	2	1	1	16.2	2031	110	303	0
1038	1	2	3	3	2716	78	600	0
1113	1	1	3	16.1	2709	78	388	0
682	2	1	1	3.58	2182	0	2021	0
449	1	1	3	7.25	2975	34	1181	0
724	1	1	1	9.42	278	4	3158	0
698	1	1	3	4.17	2442	88	2023	0
485	2	1	3	13.6	1751	12	1060	0
51	2	1	3	15	1656	142	1936	0
1118	2	1	1	16.3	2399	207	1339	0
1397	2	1	1	9.33	891	6	3008	0
629	2	2	1	7.17	2232	112	1237	0
145	2	1	2	3.92	2833	60	1563	0
635	1	1	1	12.1	1462	4	2696	0
840	2	1	3	5.83	270	98	3366	0
816	1	1	3	4.5	2206	88	1306	0
1105	1	2	3	1.42	2153	91	431	0
699	1	1	3	6.17	2594	52	181	0
810	2	1	2	6.75	1882	98	499	0
999	2	2	1	5.75	2594	70	520	0
44	2	1	3	2.25	899	108	128	1

Six possible explanatory variables are listed in Table 3.7:

- sex: female = 1; male = 2
- race: Ladino = 1; Indigina = 2
- diag: diagnosis leukemia = 1; lymphoma = 2; other = ?
- age: at admission, in years
- enter: entry date in days since 01/01/2000
- far: distance in kilometers from child's home to the medical facility

Standardized versions of the explanatory variables were used in the analysis, for instance

$$\text{Age} = (\text{age} - \text{mean(age)}/\,\text{sd(age)},$$

and similarly for Sex, Race, Diag, Enter, Far.

Table 3.8: Results of proportional hazards analysis of pediatric abandonment data. Six explanatory variables standardized to have mean 0, variance 1.

	coef	sterr	z-value	p-value	exp(coef)
Sex	−.008	.079	−.097	.923	.992
Race	.134	.075	1.772	.076	1.143
Diag	.150	.081	1.866	.062	1.162
Age	−.201	.088	−2.284	.022*	.818
Enter	−.454	.079	−5.756	.000***	.635
Far	.295	.072	4.115	.000***	1.343

An excellent proportional hazards program `coxph` is available in the R package `survival`. Setting

$$S = \texttt{Surv}(\text{time}, d),$$

the call

$$\texttt{coxph}(S \sim \text{Sex} + \text{Race} + \text{Diag} + \text{Age} + \text{Enter} + \text{Far})$$

gave the results shown in Table 3.8. Sex, Race, and Diag are insignificant as predictors of abandonment. Age is mildly interesting, with two-sided p-value 0.022. The two dramatic predictors are Enter and Far: children entering the study later suffered less abandonment (as indicated by the negative regression coefficient $\hat{\beta}_{\text{Entry}} = -0.454$, which reduces the hazard rate in model (3.16)) while those living farther away had a greater hazard rate for abandonment.

Homework 3.22. Run $\texttt{coxph}(S \sim \text{Sex} + \text{Race} + \text{Diag} + \text{Age})$ and $\texttt{coxph}(S \sim \text{Age} + \text{Enter} + \text{Far})$. Comment.

Homework 3.23. Use `coxph` to test the null hypothesis that Arm B is no better than Arm A for the NCOG data listed at the beginning of Section 3.6; data is in the file "ncogdata". *Hint*: The only explanatory variable is the Arm indicator.

Proportional hazards as an exponential family model

The partial likelihood function (3.18) can be written as

$$L(\beta) = \prod_{j=1}^{J} e^{\beta^\top x_{i_j} - \psi_j(\beta)} = e^{\beta^\top \sum_j x_{i_j} - \psi_+(\beta)},$$

where

$$\psi_+(\beta) = \sum_{j=1}^{J} \left(\log \sum_{\mathcal{R}_j} e^{\beta^\top x_i} \right).$$

This is the likelihood function of a p-parameter exponential family $f_\beta(\boldsymbol{y})$, having

- natural parameter β;
- sufficient statistic $\boldsymbol{y} = \sum_{j=1}^{J} x_{i_j}$;
- cgf $\psi_+(\beta)$.

Homework 3.24.

(a) Differentiate $\psi_+(\beta)$ to get the expectation vector and covariance matrix of $\boldsymbol{y}$.

(b) Apply the general theory of maximum likelihood estimation to family $f_\beta(\boldsymbol{y})$ to get $\dot{l}(\beta)$ and $-\ddot{l}(\beta)$ as in (3.19).

Let $X(j)$ denote the $L_j \times p$ matrix ($L_j = |\mathcal{R}_j|$) having the covariate vectors x_i in the risk set $\mathcal{R}_j$ as rows, and consider the model

$$\eta(j) = X(j)\beta. \tag{3.20}$$

As in Section 2.9, this defines a p-dimensional exponential family of probability vectors $\pi_\beta(j)$ on the L_j-dimensional simplex,

$$\pi_\beta(j) = e^{\eta(j)} \Big/ \sum_{k=1}^{L_j} e^{\eta_k(j)}; \tag{3.21}$$

(3.20)–(3.21) can be thought of as a multinomial GLM, constituting a p-parameter subexponential family of the full L_j-parameter unrestricted multinomial family. Partial likelihood analysis (3.18) amounts to considering J notionally independent such subfamilies.

3.8 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 3.9: 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; $\sum(R\text{p}^2)/30 = 1.94$, 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 3.9 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), $i = 1, 2, \ldots, 34$, where

Table 3.10: 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 ***

- r_i = annual rainfall in city i;
- n_i = number of people sampled;
- s_i = number testing positive for toxoplasmosis.

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,

$$\texttt{glm}(p \sim \texttt{poly}(r,3), \texttt{binomial, weight} = n),$$

with p, r, and n indicating their respective 34-vectors. Part of the output appears in Table 3.10. We see that the cubic regression coefficient 1.3787 is strongly positive, z-value 3.35, two-sided p-value less than 0.001.

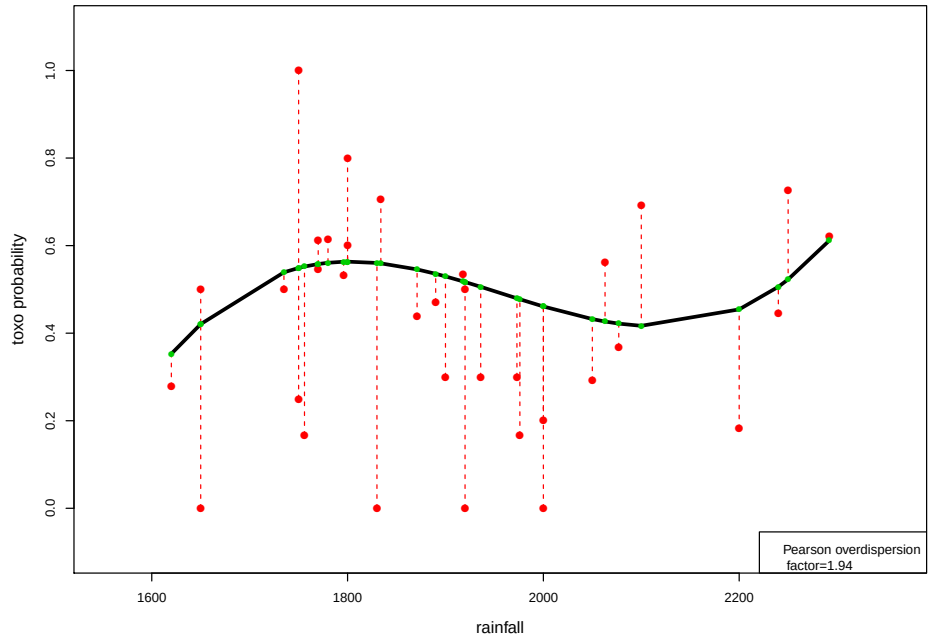


Figure 3.6: Observed proportions of toxoplasmosis, 34 cities in El Salvador; curve is cubic logistic regression.

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 3.9 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

$$R_i = \text{sign}(p_i - \hat{\pi}_i)\sqrt{n_i D(p_i, \hat{\pi}_i)},$$

(1.10) and Homework 1.23, and the Pearson residual

$$R\text{p}_i = \text{sign}(p_i - \hat{\pi}_i)\Big/\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)/n_i},$$

$\hat{\pi}_i = 1/(1 + e^{-x_i^\top \hat{\beta}})$.

In the absence of overdispersion, we would expect both

$$\sum_1^{34} R_i^2/30 \quad \text{and} \quad \sum_1^{34} R\text{p}_i^2/30$$

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

$$\sum_1^{34} R_i^2/30 = 2.09 \quad \text{and} \quad \sum_1^{34} R\text{p}_i^2/30 = 1.94;$$

the points in Figure 3.6 are about $\sqrt{2}$ farther from the fitted curve than as suggested by binomial variability.

Homework 3.25. Compute the p-value for Wilks' likelihood ratio test of the null hypothesis that there is no overdispersion around the cubic regression curve.

Table 3.11: 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	⋮	⋮	⋮
2	⋮	⋮	⋮
⋮	⋮	⋮	⋮
j	c_j	r_j	z_j
⋮	⋮	⋮	⋮
697	⋮	⋮	⋮

The toxoplasmosis study comprised 697 subjects. It was originally presented as a 697 by 3

matrix, such as that suggested by Table 3.11, 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 3.26. 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 3.10. 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 $\hat{\beta} = (X^\top X)^{-1} X^\top \boldsymbol{y}$, with

$$\hat{\beta} \sim \mathcal{N}_p\left(\beta, \sigma^2 (X^\top X)^{-1}\right);$$

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 3.27. Suppose we observe independent 0/1 random variables $y_1, y_2, \ldots, y_N$, 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

$$y_i \overset{\text{ind}}{\sim} \mathcal{N}(x_i^\top \beta, \sigma^2), \qquad i = 1, 2, \ldots, N,$$

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 3.9). 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, $v(\mu) = \ddot{\psi}(\eta)$ (but with $v(\mu)$ not necessarily of the familiar forms we have been dealing with, e.g., $V(\mu) = \mu$ for the Poisson.)

Generalizing the GLM assumption $\boldsymbol{\mu}_\beta = \dot{\psi}(X\beta)$, we assume only that the function $\boldsymbol{\mu}_\beta = (\cdots \mu_i(\beta) \cdots)^\top$ is smoothly defined, with $N \times p$ derivative matrix, say

$$\boldsymbol{w}_\beta = \left(\frac{\partial \mu_i}{\partial \beta_j}\right).$$

Lemma 4. *The score function and information matrix for an extended GLM family are*

$$\dot{l}_\beta(\boldsymbol{y}) = \boldsymbol{w}_\beta^\top \boldsymbol{v}_\beta^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_\beta) \quad \textit{and} \quad i_\beta = \boldsymbol{w}_\beta^\top \boldsymbol{v}_\beta^{-1} \boldsymbol{w}_\beta,$$

where $\boldsymbol{v}_\beta$ is the diagonal matrix with elements $v_i(\beta)$.

The proof of Lemma 4 begins by differentiating $l_\beta(\boldsymbol{y}) = \boldsymbol{\eta}_\beta^\top \boldsymbol{y} - \sum \psi(\eta_i(\beta))$, using $d\boldsymbol{\eta}_\beta/d\beta = \boldsymbol{v}_\beta^{-1}\boldsymbol{w}_\beta$.

Homework 3.28. 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 4 gives

$$\dot{l}_\beta(\boldsymbol{y}) = X^\top(\boldsymbol{y} - \boldsymbol{\mu}_\beta) \quad \text{and} \quad i_\beta = X^\top \boldsymbol{V}_\beta X,$$

the same as in Section 3.1.

The quasilikelihood approach to overdispersion is simply to assume that

$$v(\mu) = \sigma^2 V(\mu) \qquad (\text{for } \sigma^2 \text{ an unknown positive constant}), \tag{3.22}$$

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 4 gives

$$\dot{l}_\beta(\boldsymbol{y}) = \boldsymbol{w}_\beta^\top \boldsymbol{V}_\beta^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_\beta)/\sigma^2 \quad \text{and} \quad i_\beta = \boldsymbol{w}_\beta^\top \boldsymbol{V}_\beta^{-1} \boldsymbol{w}_\beta/\sigma^2. \tag{3.23}$$

Chapter 9 of McCullagh and Nelder (1989) shows that under reasonable asymptotic conditions, the

MLE $\hat{\beta}$, i.e., the solution to $\dot{l}_\beta(\boldsymbol{y}) = \mathbf{0}$, satisfies

$$\hat{\beta} \doteq \mathcal{N}_p(\beta, i_\beta^{-1}).$$

Applied to the original model $\boldsymbol{\eta}_\beta = X\beta$, (3.23) gives

$$\dot{l}_\beta(\boldsymbol{y}) = X^\top(\boldsymbol{y} - \boldsymbol{\mu}_\beta)/\sigma^2 \quad \text{and} \quad i_\beta = X^\top \boldsymbol{V}_\beta X/\sigma^2. \tag{3.24}$$

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,

$$\hat{\beta} \doteq \mathcal{N}_p\left(\beta, \sigma^2(X^\top \boldsymbol{V}_\beta X)^{-1}\right),$$

compared with (3.1) in Section 3.1.

The toxoplasmosis data was rerun using a quasibinomial model, as shown in Table 3.12. It estimated σ^2 as 1.94, the Pearson residual overdispersion estimate from Table 3.9. Comparing the results with the standard binomial GLM in Table 3.10 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 3.12: Quasibinomial logistic regression for toxoplasmosis data: `glm(formula` $= p \sim$ `poly(r,3), family = quasibinomial, weights` $= n$`)`.

Coefficients	Estimate	St. error	t-value	Pr($>\|t\|$)
(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 ***

3.9 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

$$\bar{y} \sim \left(\mu, \frac{V(\mu)}{n}\right). \tag{3.25}$$

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 $g_{\mu,n}(\bar{y})$ (3.25) is the two-parameter family

$$\boxed{f_{\mu,\theta,n}(\bar{y}) = C\,\theta^{1/2}g_{\bar{y},n}(\bar{y})e^{-n\theta D(\bar{y},\mu)/2},} \tag{3.26}$$

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 (3.26) is convenient, and usually accurate enough for applications.

Fact 2 Formula (3.26) 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 3.29. Verify Fact 2.

Fact 3 The expectation and variance of $\bar{y} \sim f_{\mu,\theta,n}$ are, to excellent approximations,

$$\bar{y} \doteq \left(\mu, \frac{V(\mu)}{n\theta}\right),$$

with errors of order $1/n^2$ for both terms. Comparison with (3.22) shows that $1/\theta$ measures dispersion,

$$\sigma^2 = 1/\theta.$$

Homework 3.30. 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 $(\bar{y}, -\psi(\bar{\eta}))$, 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 3.31. 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.

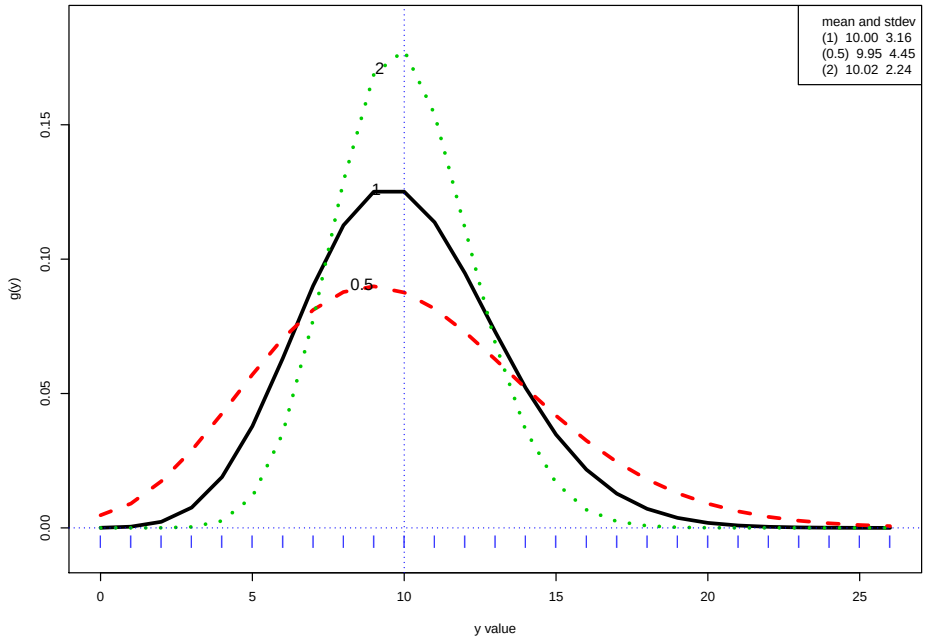


Figure 3.7: Double Poisson densities for $\mu = 10$ and $\theta = 1$ (solid), $= 1/2$ (dashed), or $= 2$ (dotted).

Figure 3.7 illustrates the double Poisson distribution: we have taken[4] $n = 1$, $\mu = 10$, 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 $0, 1, 2, \ldots$.

[4]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 3.32.

(a) For the Poisson family ($n = 1$) show that (3.26), 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 $C = 1$?

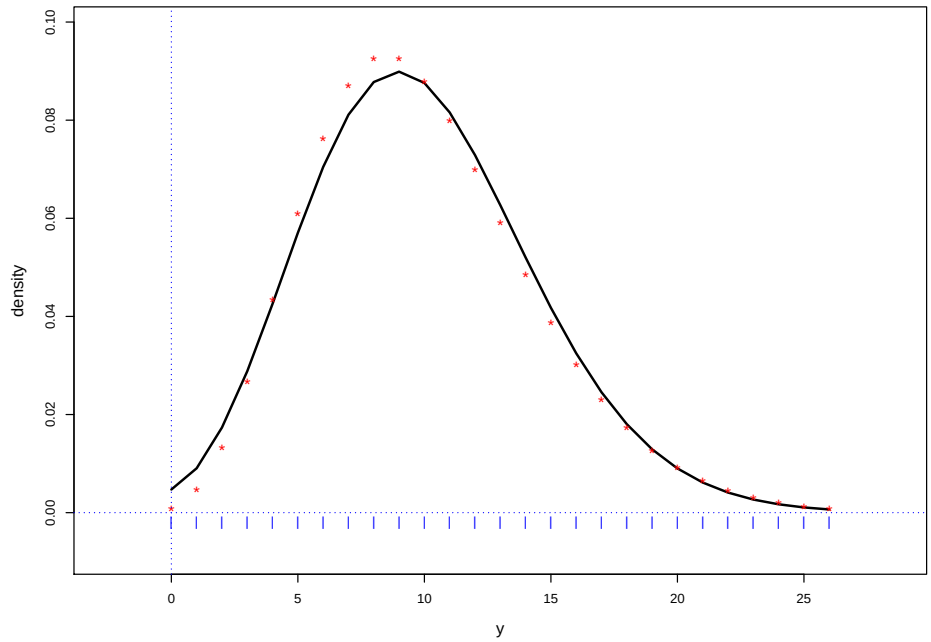


Figure 3.8: Comparison of double Poisson density, $\mu = 10$, $\theta = 0.5$ (solid), with negative binomial density, $\mu = 10$, variance $= 20$ (points).

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 3.30.

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 $0, {}^1\!/_n, {}^2\!/_n, \ldots, 1$, while $m_{n\theta}(p)$ is counting measure on $0, {}^1\!/_{n\theta}, {}^2\!/_{n\theta}, \ldots, 1$. This assumes $n\theta$ is an integer, and shows the limitations of Fact 5 in discrete families.

Homework 3.33. In the binomial case, numerically compare the cumulative distribution function of $f_{\mu,\theta,n}(p)$, i.e., $(\mu,\theta,n) = (0.4, 0.5, 16)$, with that of $g_{\mu,n}(p)$, $(\mu,n) = (0.4, 8)$.

Homework 3.34. 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 3.35.

(a) Show that

$$f_{\mu,\theta,n}(\bar{y}) = \frac{g_{\mu,n\theta}(\bar{y})}{C(\mu,\theta,n)},$$

where

$$C(\mu,\theta,n) = \theta^{-1/2}\ \frac{\Gamma(n)}{(n/e)^n} \Big/ \frac{\Gamma(n\theta)}{(n\theta/e)^{n\theta}}.$$

(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

$$\bar{y}_i \overset{\text{ind}}{\sim} f_{\mu_i,\theta,n_i} \qquad \text{for } i = 1, 2, \ldots, N, \tag{3.27}$$

with the GLM model $\boldsymbol{\eta}(\beta) = X\beta$ giving $\boldsymbol{\mu}(\beta) = (\cdots \mu_i = \dot{\psi}(\eta_i(\beta)) \cdots)^\top$. The score functions for the full data set $\boldsymbol{y} = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_N)^\top$ are

$$\frac{\partial}{\partial \beta}\, l_{\beta,\theta}(\boldsymbol{y}) = \theta X^\top \operatorname{diag}(\boldsymbol{n})\, (\boldsymbol{y} - \boldsymbol{\mu}(\beta))\,,$$

$\operatorname{diag}(\boldsymbol{n})$ 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 3.36. 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

$$\hat{\theta} = N \Big/ \sum_{i=1}^{N} n_i D\left(\bar{y}_i, \mu_i\left(\hat{\beta}\right)\right).$$

Homework 3.37. How does this estimate relate to the overdispersion estimates $\hat{\sigma}^2$ for the toxoplasmosis data of Section 3.8?

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 $p_i \sim f_{\pi_i,\theta_i,n_i}$, $i = 1, 2, \ldots, 34$. 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 (3.26).

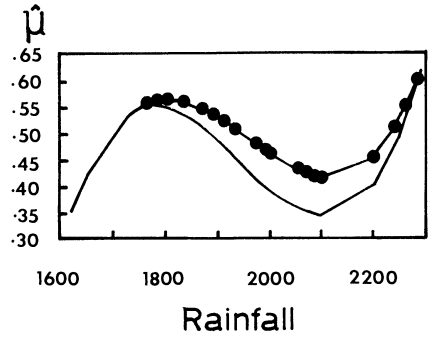
My model let π_i be a cubic polynomial function of rainfall, as in Section 3.8; θ_i was modeled as a function of n_i, the number of people sampled in city i. Let

$$\tilde{n}_i = \frac{n_i - \bar{n}}{\text{sd}_n},$$

$\bar{n}$ 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 3.38. 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 3.10, 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.