8

Generalized Linear Models and Regression Trees

Indirect evidence is not the sole property of Bayesians. Regression models are the frequentist method of choice for incorporating the experience of "others." As an example, Figure 8.1 returns to the kidney fitness data of Section 1.1. A potential new donor, aged 55, has appeared, and we wish to assess his kidney fitness without subjecting him to an arduous series of medical tests. Only one of the 157 previously tested volunteers was age 55, his `tot` score being -0.01 (the upper large dot in Figure 8.1). Most applied statisticians, though, would prefer to read off the height of the least squares regression line at age $= 55$ (the green dot on the regression line), $\widehat{\texttt{tot}} = -1.46$. The former is the only direct evidence we have, while the

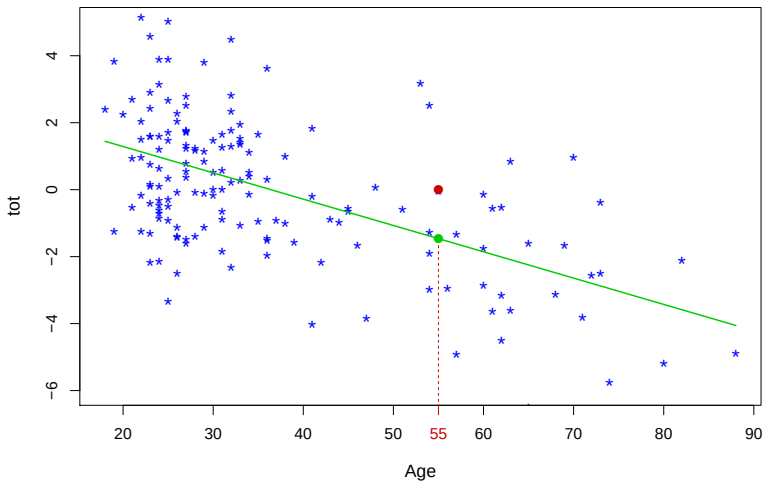

Figure 8.1 Kidney data; a new volunteer donor is aged 55. Which prediction is preferred for his kidney function?

regression line lets us incorporate indirect evidence for age 55 from all 157 previous cases.

Increasingly aggressive use of regression techniques is a hallmark of modern statistical practice, "aggressive" applying to the number and type of predictor variables, the coinage of new methodology, and the sheer size of the target data sets. Generalized linear models, this chapter's main topic, have been the most pervasively influential of the new methods. The chapter ends with a brief review of regression trees, a completely different regression methodology that will play an important role in the prediction algorithms of Chapter 17.

8.1 Logistic Regression

An experimental new anti-cancer drug called `Xilathon` is under development. Before human testing can begin, animal studies are needed to determine safe dosages. To this end, a *bioassay* or dose–response experiment was carried out: 11 groups of $n = 10$ mice each were injected with increasing amounts of `Xilathon`, dosages coded[1] $1, 2, \ldots, 11$.

Let

$$y_i = \#\text{ mice dying in } i\text{th group.} \tag{8.1}$$

The points in Figure 8.2 show the proportion of deaths

$$p_i = y_i/10, \tag{8.2}$$

lethality generally increasing with dose. The counts y_i are modeled as independent binomials,

$$y_i \overset{\text{ind}}{\sim} \text{Bi}(n_i, \pi_i) \qquad \text{for } i = 1, 2, \ldots, N, \tag{8.3}$$

$N = 11$ and all n_i equaling 10 here; π_i is the true death rate in group i, estimated unbiasedly by p_i, the direct evidence for π_i. The regression curve in Figure 8.2 uses *all* the doses to give a better picture of the true dose–response relation.

Logistic regression is a specialized technique for regression analysis of count or proportion data. The *logit* parameter λ is defined as

$$\lambda = \log\left\{\frac{\pi}{1-\pi}\right\}, \tag{8.4}$$

[1] Dose would usually be labeled on a log scale, each one, say, 50% larger than its predecessor.

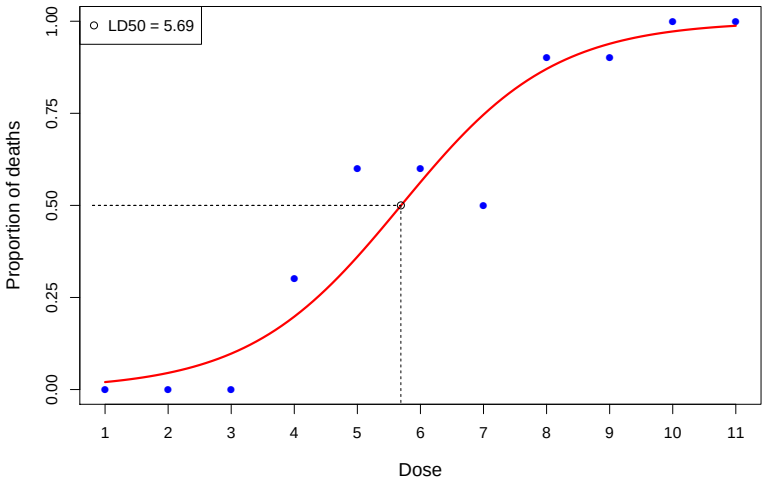

Figure 8.2 Dose–response study; groups of 10 mice exposed to increasing doses of experimental drug. The points are the observed proportions that died in each group. The fitted curve is the maximum-likelihoood estimate of the linear logistic regression model. The open circle on the curve is the LD50, the estimated dose for 50% mortality.

with λ increasing from $-\infty$ to ∞ as π increases from 0 to 1. A linear logistic regression dose–response analysis begins with binomial model (8.3), and assumes that the logit is a linear function of dose,

$$\lambda_i = \log\left\{\frac{\pi_i}{1-\pi_i}\right\} = \alpha_0 + \alpha_1 x_i. \tag{8.5}$$

Maximum likelihood gives estimates $(\hat{\alpha}_0, \hat{\alpha}_1)$, and fitted curve

$$\hat{\lambda}(x) = \hat{\alpha}_0 + \hat{\alpha}_1 x. \tag{8.6}$$

Since the inverse transformation of (8.4) is

$$\pi = \left(1 + e^{-\lambda}\right)^{-1} \tag{8.7}$$

we obtain from (8.6) the linear logistic regression curve

$$\hat{\pi}(x) = \left(1 + e^{-(\hat{\alpha}_0 + \hat{\alpha}_1 x)}\right)^{-1} \tag{8.8}$$

pictured in Figure 8.2.

Table 8.1 compares the standard deviation of the estimated regression

Table 8.1 *Standard deviation estimates for $\hat{\pi}(x)$ in Figure 8.1. The first row is for the linear logistic regression fit* (8.8)*; the second row is based on the individual binomial estimates p_i.*

x	1	2	3	4	5	6	7	8	9	10	11
sd $\hat{\pi}(x)$	.015	.027	.043	.061	.071	.072	.065	.050	.032	.019	.010
sd p_i	.045	.066	.094	.126	.152	.157	.138	.106	.076	.052	.035

curve (8.8) at $x = 1, 2, \ldots, 11$ (as discussed in the next section) with the usual binomial standard deviation estimate $[p_i(1-p_i)/10]^{1/2}$ obtained by considering the 11 doses separately.[2] Regression has reduced error by better than 50%, the price being possible bias if model (8.5) goes seriously wrong.

One advantage of the logit transformation is that λ isn't restricted to the range $[0, 1]$, so model (8.5) never verges on forbidden territory. A better reason has to do with the exploitation of exponential family properties. We can rewrite the density function for $\text{Bi}(n, y)$ as

$$\binom{n}{y}\pi^y(1-\pi)^{n-y} = e^{\lambda y - n\psi(\lambda)}\binom{n}{y} \tag{8.9}$$

with λ the logit parameter (8.4) and

$$\psi(\lambda) = \log\{1 + e^{\lambda}\}; \tag{8.10}$$

(8.9) is a one-parameter exponential family[3] as described in Section 5.5, with λ the natural parameter, called α there.

Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$ denote the full data set, $N = 11$ in Figure 8.2. Using (8.5), (8.9), and the independence of the y_i gives the probability density of $\boldsymbol{y}$ as a function of (α_0, α_1),

$$\begin{aligned} f_{\alpha_0,\alpha_1}(\boldsymbol{y}) &= \prod_{i=1}^{N} e^{\lambda_i y_i - n_i\psi(\lambda_i)}\binom{n_i}{y_i} \\ &= e^{\alpha_0 S_0 + \alpha_1 S_1} \cdot e^{-\sum_1^N n_i\psi(\alpha_0+\alpha_1 x_i)} \cdot \prod_{i=1}^{N}\binom{n_i}{y_i}, \end{aligned} \tag{8.11}$$

[2] For the separate-dose standard error, p_i was taken equal to the fitted value from the curve in Figure 8.2.

[3] It is not necessary for $f_{\mu_0}(x)$ in (5.46) on page 64 to be a probability density function, only that it not depend on the parameter μ.

where

$$S_0 = \sum_{i=1}^{N} y_i \quad \text{and} \quad S_1 = \sum_{i=1}^{N} x_i y_i. \tag{8.12}$$

Formula (8.11) expresses $f_{\alpha_0,\alpha_1}(\boldsymbol{y})$ as the product of three factors,

$$f_{\alpha_0,\alpha_1}(\boldsymbol{y}) = g_{\alpha_0,\alpha_1}(S_0, S_1)h(\alpha_0, \alpha_1)j(\boldsymbol{y}), \tag{8.13}$$

only the first of which involves both the parameters and the data. This implies that (S_0, S_1) is a *sufficient statistic*:† no matter how large N might be (later we will have N in the thousands), just the two numbers (S_0, S_1) contain all of the experiment's information. Only the logistic parameterization (8.4) makes this happen.[4]

†1

A more intuitive picture of logistic regression depends on $D(p_i, \hat{\pi}_i)$, the *deviance* between an observed proportion p_i (8.2) and an estimate $\hat{\pi}_i$,

$$D\left(p_i, \hat{\pi}_i\right) = 2n_i\left[p_i \log\left(\frac{p_i}{\hat{\pi}_i}\right) + (1-p_i)\log\left(\frac{1-p_i}{1-\hat{\pi}_i}\right)\right]. \tag{8.14}$$

The deviance[5] is zero if $\hat{\pi}_i = p_i$, otherwise it increases as $\hat{\pi}_i$ departs further from p_i.

The logistic regression MLE value $(\hat{\alpha}_0, \hat{\alpha}_1)$ also turns out to be the choice of (α_0, α_1) minimizing the total deviance between the N points p_i and their corresponding estimates $\hat{\pi}_i = \pi_{\hat{\alpha}_0,\hat{\alpha}_1}(x_i)$ (8.8):

$$(\hat{\alpha}_0, \hat{\alpha}_1) = \underset{(\alpha_0,\alpha_1)}{\arg\min} \sum_{i=1}^{N} D\left(p_i, \pi_{\alpha_0,\alpha_1}(x_i)\right). \tag{8.15}$$

The solid line in Figure 8.2 is the linear logistic curve coming closest to the 11 points, when distance is measured by total deviance. In this way the 200-year-old notion of least squares is generalized to binomial regression, as discussed in the next section. A more sophisticated notion of distance between data and models is one of the accomplishments of modern statistics.

Table 8.2 reports on the data for a more structured logistic regression analysis. Human muscle cell colonies were infused with mouse nuclei in five different ratios, cultured over time periods ranging from one to five

[4] Where the name "logistic regression" comes from is explained in the endnotes, along with a description of its nonexponential family predecessor *probit analysis*.

[5] Deviance is analogous to squared error in ordinary regression theory, as discussed in what follows. It is twice the "Kullback–Leibler distance," the preferred name in the information-theory literature.

Table 8.2 *Cell infusion data; human cell colonies infused with mouse nuclei in five ratios over 1 to 5 days and observed to see whether they did or did not thrive. Green numbers are estimates $\hat{\pi}_{ij}$ from the logistic regression model. For example, 5 of 31 colonies in the lowest ratio/days category thrived, with observed proportion $5/31 = 0.16$, and logistic regression estimate $\hat{\pi}_{11} = 0.11$.*

Ratio	Time 1	2	3	4	5
1	5/31 **.11**	3/28 **.25**	20/45 **.42**	24/47 **.54**	29/35 **.75**
2	15/77 **.24**	36/78 **.45**	43/71 **.64**	56/71 **.74**	66/74 **.88**
3	48/126 **.38**	68/116 **.62**	145/171 **.77**	98/119 **.85**	114/129 **.93**
4	29/92 **.32**	35/52 **.56**	57/85 **.73**	38/50 **.81**	72/77 **.92**
5	11/53 **.18**	20/52 **.37**	20/48 **.55**	40/55 **.67**	52/61 **.84**

days, and observed to see whether they thrived. For example, of the 126 colonies having the third ratio and shortest time period, 48 thrived.

Let π_{ij} denote the true probability of thriving for ratio i during time period j, and λ_{ij} its logit $\log\{\pi_{ij}/(1-\pi_{ij})\}$. A two-way additive logistic regression was fit to the data,[6]

$$\lambda_{ij} = \mu + \alpha_i + \beta_j, \qquad i = 1, 2, \ldots, 5,\ j = 1, 2, \ldots, 5. \tag{8.16}$$

The green numbers in Table 8.2 show the maximum likelihood estimates

$$\hat{\pi}_{ij} = 1 \Big/ \left[1 + e^{-\left(\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_j\right)}\right]. \tag{8.17}$$

Model (8.16) has nine free parameters (taking into account the constraints $\sum \alpha_i = \sum \beta_j = 0$ necessary to avoid definitional difficulties) compared with just two in the dose–response experiment. The count can easily go much higher these days.

Table 8.3 reports on a 57-variable logistic regression applied to the `spam` data. A researcher (named George) labeled $N = 4601$ of his email mes-

[6] Using the statistical computing language `R`; see the endnotes.

Table 8.3 *Logistic regression analysis of the* `spam` *data, model* (8.17)*; estimated regression coefficients, standard errors, and* z = *estimate*/*se, for 57 keyword predictors. The notation* `char$` *means the relative number of times* `$` *appears, etc. The last three entries measure characteristics such as length of capital-letter strings. The word* `george` *is special, since the recipient of the email is named George, and the goal here is to build a customized spam filter.*

	Estimate	se	z-value		Estimate	se	z-value
`intercept`	−12.27	1.99	−6.16	`lab`	−1.48	.89	−1.66
`make`	−.12	.07	−1.68	`labs`	−.15	.14	−1.05
`address`	−.19	.09	−2.10	`telnet`	−.07	.19	− .35
`all`	.06	.06	1.03	`857`	.84	1.08	.78
`3d`	3.14	2.10	1.49	`data`	−.41	.17	−2.37
`our`	.38	.07	5.52	`415`	.22	.53	.42
`over`	.24	.07	3.53	`85`	−1.09	.42	−2.61
`remove`	.89	.13	6.85	`technology`	.37	.12	2.99
`internet`	.23	.07	3.39	`1999`	.02	.07	.26
`order`	.20	.08	2.58	`parts`	−.13	.09	−1.41
`mail`	.08	.05	1.75	`pm`	−.38	.17	−2.26
`receive`	−.05	.06	− .86	`direct`	−.11	.13	− .84
`will`	−.12	.06	−1.87	`cs`	−16.27	9.61	−1.69
`people`	−.02	.07	− .35	`meeting`	−2.06	.64	−3.21
`report`	.05	.05	1.06	`original`	−.28	.18	−1.55
`addresses`	.32	.19	1.70	`project`	−.98	.33	−2.97
`free`	.86	.12	7.13	`re`	−.80	.16	−5.09
`business`	.43	.10	4.26	`edu`	−1.33	.24	−5.43
`email`	.06	.06	1.03	`table`	−.18	.13	−1.40
`you`	.14	.06	2.32	`conference`	−1.15	.46	−2.49
`credit`	.53	.27	1.95	`char;`	−.31	.11	−2.92
`your`	.29	.06	4.62	`char(`	−.05	.07	− .75
`font`	.21	.17	1.24	`char_`	−.07	.09	− .78
`000`	.79	.16	4.76	`char!`	.28	.07	3.89
`money`	.19	.07	2.63	`char$`	1.31	.17	7.55
`hp`	−3.21	.52	−6.14	`char#`	1.03	.48	2.16
`hpl`	−.92	.39	−2.37	`cap.ave`	.38	.60	.64
`george`	−39.62	7.12	−5.57	`cap.long`	1.78	.49	3.62
`650`	.24	.11	2.24	`cap.tot`	.51	.14	3.75

sages as either `spam` or `ham` (nonspam[7]), say

$$y_i = \begin{cases} 1 & \text{if email } i \text{ is } \texttt{spam} \\ 0 & \text{if email } i \text{ is } \texttt{ham} \end{cases} \tag{8.18}$$

[7] "Ham" refers to "nonspam" or good email; this is a playful connection to the processed

(40% of the messages were **spam**). The $p = 57$ predictor variables represent the most frequently used words and tokens in George's corpus of email (excluding trivial words such as articles), and are in fact the relative frequencies of these chosen words in each email (standardized by the length of the email). The goal of the study was to predict whether future emails are **spam** or **ham** using these keywords; that is, to build a customized *spam filter*.

Let x_{ij} denote the relative frequency of keyword j in email i, and π_i represent the probability that email i is **spam**. Letting λ_i be the logit transform $\log\{\pi_i/(1-\pi_i)\}$, we fit the additive logistic model

$$\lambda_i = \alpha_0 + \sum_{j=1}^{57} \alpha_i x_{ij}. \tag{8.19}$$

Table 8.3 shows $\hat{\alpha}_i$ for each word—for example, -0.12 for **make**—as well as the estimated standard error and the *z-value*: estimate/se.

It looks like certain words, such as **free** and **your**, are good **spam** predictors. However, the table as a whole has an unstable appearance, with occasional very large estimates $\hat{\alpha}_i$ accompanied by very large standard deviations.[8] The dangers of high-dimensional maximum likelihood estimation are apparent here. Some sort of shrinkage estimation is called for, as discussed in Chapter 16.

Regression analysis, either in its classical form or in modern formulations, requires covariate information x to put the various cases into some sort of geometrical relationship. Given such information, regression is the statistician's most powerful tool for bringing "other" results to bear on a case of primary interest: for instance, the age-55 volunteer in Figure 8.1.

Empirical Bayes methods do not require covariate information but may be improvable if it exists. If, for example, the player's age were an important covariate in the baseball example of Table 7.1, we might first regress the MLE values on age, and then shrink them toward the regression line rather than toward the grand mean $\bar{p}$ as in (7.20). In this way, two different sorts of indirect evidence would be brought to bear on the estimation of each player's ability.

spam that was fake ham during WWII, and has been adopted by the machine-learning community.

[8] The 4601×57 $\boldsymbol{X}$ matrix (x_{ij}) was standardized, so disparate scalings are not the cause of these discrepancies. Some of the features have mostly "zero" observations, which may account for their unstable estimation.

8.2 Generalized Linear Models[9]

Logistic regression is a special case of *generalized linear models* (GLMs), a key 1970s methodology having both algorithmic and inferential influence. GLMs extend ordinary linear regression, that is least squares curve-fitting, to situations where the response variables are binomial, Poisson, gamma, beta, or in fact any exponential family form.

We begin with a one-parameter exponential family,

$$\left\{ f_\lambda(y) = e^{\lambda y - \gamma(\lambda)} f_0(y),\ \lambda \in \Lambda \right\}, \tag{8.20}$$

as in (5.46) (now with α and x replaced by λ and y, and $\psi(\alpha)$ replaced by $\gamma(\lambda)$, for clearer notation in what follows). Here λ is the *natural parameter* and y the *sufficient statistic*, both being one-dimensional in usual applications; λ takes its values in an interval of the real line. Each coordinate y_i of an observed data set $\boldsymbol{y} = (y_1, y_2, \ldots, y_i, \ldots, y_N)'$ is assumed to come from a member of family (8.20),

$$y_i \sim f_{\lambda_i}(\cdot) \text{ independently for } i = 1, 2, \ldots, N. \tag{8.21}$$

Table 8.4 lists λ and y for the first four families in Table 5.1, as well as their deviance and normalizing functions.

By itself, model (8.21) requires N parameters $\lambda_1, \lambda_2, \ldots, \lambda_N$, usually too many for effective individual estimation. A key GLM tactic is to specify the λs in terms of a linear regression equation. Let $\boldsymbol{X}$ be an $N \times p$ "structure matrix," with ith row say x_i', and α an unknown vector of p parameters; the N-vector $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_N)'$ is then specified by

$$\boldsymbol{\lambda} = \boldsymbol{X}\alpha. \tag{8.22}$$

In the dose–response experiment of Figure 8.2 and model (8.5), $\boldsymbol{X}$ is $N \times 2$ with ith row $(1, x_i)$ and parameter vector $\alpha = (\alpha_0, \alpha_1)$.

The probability density function $f_\alpha(\boldsymbol{y})$ of the data vector $\boldsymbol{y}$ is

$$f_\alpha(\boldsymbol{y}) = \prod_{i=1}^{N} f_{\lambda_i}(y_i) = e^{\sum_1^N (\lambda_i y_i - \gamma(\lambda_i))} \prod_{i=1}^{N} f_0(y_i), \tag{8.23}$$

which can be written as

$$f_\alpha(\boldsymbol{y}) = e^{\alpha' z - \psi(\alpha)} f_0(\boldsymbol{y}), \tag{8.24}$$

[9] Some of the more technical points raised in this section are referred to in later chapters, and can be scanned or omitted at first reading.

Table 8.4 *Exponential family form for first four cases in Table 5.1; natural parameter λ, sufficient statistic y, deviance* (8.31) *between family members f_1 and f_2, $D(f_1, f_2)$, and normalizing function $\gamma(\lambda)$.*

		λ	y	$D(f_1, f_2)$	$\gamma(\lambda)$
1.	*Normal* $\mathcal{N}(\mu, \sigma^2)$, σ^2 known	μ/σ^2	x	$\left(\frac{\mu_1-\mu_2}{\sigma}\right)^2$	$\sigma^2\lambda^2/2$
2.	*Poisson* Poi(μ)	$\log\mu$	x	$2\mu_1\left[\left(\frac{\mu_2}{\mu_1}-1\right)-\log\frac{\mu_2}{\mu_1}\right]$	e^λ
3.	*binomial* Bi(n, π)	$\log\frac{\pi}{1-\pi}$	x	$2n\left[\pi_1\log\frac{\pi_1}{\pi_2}+(1-\pi_1)\log\frac{1-\pi_1}{1-\pi_2}\right]$	$n\log(1+e^\lambda)$
4.	*Gamma* Gam(ν, σ), ν known	$-1/\sigma$	x	$2\nu\left[\left(\frac{\sigma_1}{\sigma_2}-1\right)-\log\frac{\sigma_1}{\sigma_2}\right]$	$-\nu\log(-\lambda)$

where

$$z = \boldsymbol{X}'\boldsymbol{y} \quad \text{and} \quad \psi(\alpha) = \sum_{i=1}^{N} \gamma(\boldsymbol{x}_i'\alpha), \tag{8.25}$$

a p-parameter exponential family (5.50), with natural parameter vector α and sufficient statistic vector z. The main point is that all the information from a p-parameter GLM is summarized in the p-dimensional vector z, no matter how large N may be, making it easier both to understand and to analyze.

We have now reduced the N-parameter model (8.20)–(8.21) to the p-parameter exponential family (8.24), with p usually much smaller than N, in this way avoiding the difficulties of high-dimensional estimation. The moments of the one-parameter constituents (8.20) determine the estimation properties in model (8.22)–(8.24). Let $(\mu_\lambda, \sigma_\lambda^2)$ denote the expectation and variance of univariate density $f_\lambda(y)$ (8.20),

$$y \sim (\mu_\lambda, \sigma_\lambda^2), \tag{8.26}$$

for instance $(\mu_\lambda, \sigma_\lambda^2) = (e^\lambda, e^\lambda)$ for the Poisson. The N-vector $\boldsymbol{y}$ obtained from GLM (8.22) then has mean vector and covariance matrix

$$\boldsymbol{y} \sim (\boldsymbol{\mu}(\alpha), \boldsymbol{\Sigma}(\alpha)), \tag{8.27}$$

where $\boldsymbol{\mu}(\alpha)$ is the vector with ith component μ_{λ_i} with $\lambda_i = x_i'\alpha$, and $\boldsymbol{\Sigma}(\alpha)$ is the $N \times N$ diagonal matrix having diagonal elements $\sigma^2_{\lambda_i}$.

The maximum likelihood estimate $\hat{\alpha}$ of the parameter vector α can be shown to satisfy the simple equation†

$$\boldsymbol{X}'\left[\boldsymbol{y} - \boldsymbol{\mu}\left(\hat{\alpha}\right)\right] = 0. \tag{8.28}$$

For the normal case where $y_i \sim \mathcal{N}(\mu_i, \sigma^2)$ in (8.21), that is, for ordinary linear regression, $\boldsymbol{\mu}(\hat{\alpha}) = \boldsymbol{X}\hat{\alpha}$ and (8.28) becomes $\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\alpha}) = 0$, with the familiar solution

$$\hat{\alpha} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}; \tag{8.29}$$

otherwise, $\boldsymbol{\mu}(\alpha)$ is a nonlinear function of α, and (8.28) must be solved by numerical iteration. This is made easier by the fact that, for GLMs, $\log f_\alpha(\boldsymbol{y})$, the likelihood function we wish to maximize, is a *concave function of* α. The MLE $\hat{\alpha}$ has approximate expectation and covariance†

$$\hat{\alpha} \doteq (\alpha, \left(\boldsymbol{X}'\boldsymbol{\Sigma}(\alpha)\boldsymbol{X}\right)^{-1}), \tag{8.30}$$

similar to the exact OLS result $\hat{\alpha} \sim (\alpha, \sigma^{-2}(\boldsymbol{X}'\boldsymbol{X})^{-1})$.†

Generalizing the binomial definition (8.14), the *deviance* between densities $f_1(y)$ and $f_2(y)$ is defined to be

$$D(f_1, f_2) = 2\int_{\mathcal{Y}} f_1(y)\log\left\{\frac{f_1(y)}{f_2(y)}\right\}\, dy, \tag{8.31}$$

the integral (or sum for discrete distributions) being over their common sample space $\mathcal{Y}$. $D(f_1, f_2)$ is always nonnegative, equaling zero only if f_1 and f_2 are the same; in general $D(f_1, f_2)$ does not equal $D(f_2, f_1)$. Deviance does not depend on how the two densities are named, for example (8.14) having the same expression as the *Binomial* entry in Table 8.4.

In what follows it will sometimes be useful to label the family (8.20) by its *expectation parameter* $\mu = E_\lambda\{y\}$ rather than by the natural parameter λ:

$$f_\mu(y) = e^{\lambda y - \gamma(\lambda)} f_0(y), \tag{8.32}$$

meaning the same thing as (8.20), only the names attached to the individual family members being changed. In this notation it is easy to show a fundamental result sometimes known as

Hoeffding's Lemma† *The maximum likelihood estimate of μ given y is y itself, and the log likelihood* $\log f_\mu(y)$ *decreases from its maximum* $\log f_y(y)$ *by an amount that depends on the deviance* $D(y, \mu)$,

$$f_\mu(y) = f_y(y)e^{-D(y,\mu)/2}. \tag{8.33}$$

Returning to the GLM framework (8.21)–(8.22), parameter vector α gives $\boldsymbol{\lambda}(\alpha) = \boldsymbol{X}\alpha$, which in turn gives the vector of expectation parameters

$$\boldsymbol{\mu}(\alpha) = (\ldots \mu_i(\alpha) \ldots)', \tag{8.34}$$

for instance $\mu_i(\alpha) = \exp\{\lambda_i(\alpha)\}$ for the Poisson family. Multiplying Hoeffding's lemma (8.33) over the N cases $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)'$ yields

$$f_\alpha(\boldsymbol{y}) = \prod_{i=1}^{N} f_{\mu_i(\alpha)}(y_i) = \left[\prod_{i=1}^{N} f_{y_i}(y_i)\right] e^{-\sum_1^N D(y_i, \mu_i(\alpha))}. \tag{8.35}$$

This has an important consequence: *the MLE $\hat{\alpha}$ is the choice of α that minimizes the total deviance* $\sum_1^N D(y_i, \mu_i(\alpha))$. As in Figure 8.2, GLM maximum likelihood fitting is "least total deviance" in the same way that ordinary linear regression is least sum of squares.

The inner circle of Figure 8.3 represents normal theory, the preferred venue of classical applied statistics. Exact inferences—t-tests, F distributions, most of multivariate analysis—were feasible within the circle. Outside the circle was a general theory based mainly on asymptotic (large-sample) approximations involving Taylor expansions and the central limit theorem.

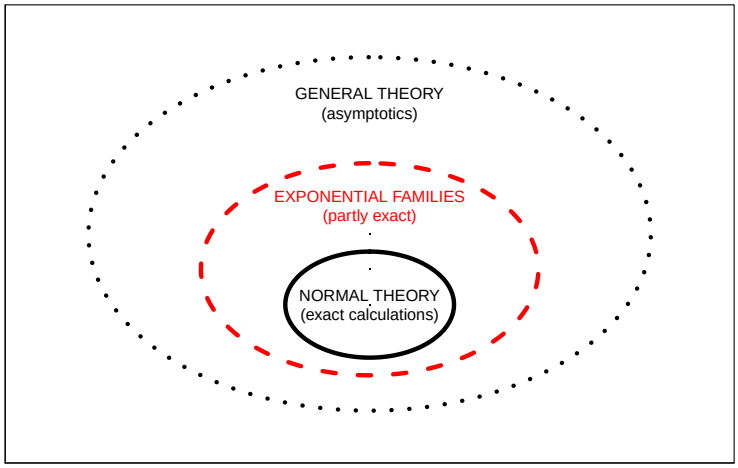

Figure 8.3 Three levels of statistical modeling.

A few useful exact results lay outside the normal theory circle, relating

to a few special families: the binomial, Poisson, gamma, beta, and others less well known. Exponential family theory, the second circle in Figure 8.3, unified the special cases into a coherent whole. It has a "partly exact" flavor, with some ideal counterparts to normal theory—convex likelihood surfaces, least deviance regression—but with some approximations necessary, as in (8.30). Even the approximations, though, are often more convincing than those of general theory, exponential families' fixed-dimension sufficient statistics making the asymptotics more transparent.

Logistic regression has banished its predecessors (such as probit analysis) almost entirely from the field, and not only because of estimating efficiencies and computational advantages (which are actually rather modest), but also because it is seen as a clearer analogue to ordinary least squares, our 200-year-old dependable standby. GLM research development has been mostly frequentist, but with a substantial admixture of likelihood-based reasoning, and a hint of Fisher's "logic of inductive inference."

Helping the statistician choose between competing methodologies is the job of statistical inference. In the case of generalized linear models the choice has been made, at least partly, in terms of aesthetics as well as philosophy.

8.3 Poisson Regression

The third most-used member of the GLM family, after normal theory least squares and logistic regression, is Poisson regression. N independent Poisson variates are observed,

$$y_i \overset{\text{ind}}{\sim} \text{Poi}(\mu_i), \qquad i = 1, 2, \ldots, N, \tag{8.36}$$

where $\lambda_i = \log \mu_i$ is assumed to follow a linear model,

$$\boldsymbol{\lambda}(\alpha) = \boldsymbol{X}\alpha, \tag{8.37}$$

where $\boldsymbol{X}$ is a known $N \times p$ structure matrix and α an unknown p-vector of regression coefficients. That is, $\lambda_i = x_i'\alpha$ for $i = 1, 2, \ldots, N$, where x_i' is the ith row of $\boldsymbol{X}$.

In the chapters that follow we will see Poisson regression come to the rescue in what at first appear to be awkward data-analytic situations. Here we will settle for an example involving density estimation from a spatially truncated sample.

†6 Table 8.5 shows galaxy counts† from a small portion of the sky: 486 galaxies have had their redshifts r and apparent magnitudes m measured.

Table 8.5 *Counts for a truncated sample of 486 galaxies, binned by redshift and magnitude.*

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

Distance from earth is an increasing function of r, while apparent brightness is a decreasing function[10] of m. In this survey, counts were limited to galaxies having

$$1.22 \le r \le 3.32 \quad \text{and} \quad 17.2 \le m \le 21.5, \tag{8.38}$$

the upper limit reflecting the difficulty of measuring very dim galaxies.

The range of $\log r$ has been divided into 15 equal intervals and likewise 18 equal intervals for m. Table 8.5 gives the counts of the 486 galaxies in the $18 \times 15 = 270$ bins. (The lower right corner of the table is empty because distant galaxies always appear dim.) The multinomial/Poisson connection (5.44) helps motivate model (8.36), picturing the table as a multinomial observation on 270 categories, in which the sample size N was itself Poisson.

We can imagine Table 8.5 as a small portion of a much more extensive table, hypothetically available if the data were *not* truncated. Experience suggests that we might then fit an appropriate bivariate normal density to the data, as in Figure 5.3. It seems like it might be awkward to fit part of a bivariate normal density to truncated data, but Poisson regression offers an easy solution.

[10] An object of the second magnitude is less bright than one of the first, and so on, a classification system owing to the Greeks.

Let $\boldsymbol{r}$ be the 270-vector listing the values of r in each bin of the table (in column order), and likewise $\boldsymbol{m}$ for the 270 m values—for instance $\boldsymbol{m} = (18, 17, \ldots, 1)$ repeated 15 times—and define the 270×5 matrix $\boldsymbol{X}$ as

$$\boldsymbol{X} = [\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{r}^2, \boldsymbol{rm}, \boldsymbol{m}^2], \tag{8.39}$$

where $\boldsymbol{r}^2$ is the vector whose components are the square of $\boldsymbol{r}$'s, etc. The log density of a bivariate normal distribution in (r, m) is of the form $\alpha_1 r + \alpha_2 m + \alpha_3 r^2 + \alpha_4 rm + \alpha_5 m^2$, agreeing with $\log \mu_i = x_i' \alpha$ as specified by (8.39). We can use a Poisson GLM, with y_i the ith bin's count, to estimate the portion of our hypothesized bivariate normal distribution in the truncation region (8.38).

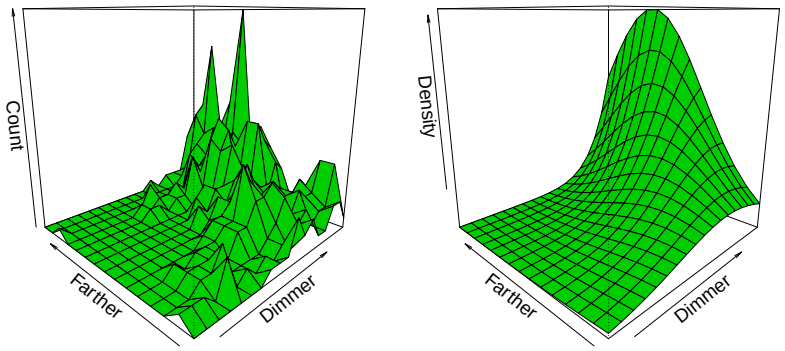

Figure 8.4 *Left* galaxy data; binned counts. *Right* Poisson GLM density estimate.

The left panel of Figure 8.4 is a perspective picture of the raw counts in Table 8.5. On the right is the fitted density from the Poisson regression. Irrespective of density estimation, Poisson regression has done a useful job of smoothing the raw bin counts.

Contours of equal value of the fitted log density

$$\hat{\alpha}_0 + \hat{\alpha}_1 r + \hat{\alpha}_2 m + \hat{\alpha}_3 r^2 + \hat{\alpha}_4 rm + \hat{\alpha}_5 m^2 \tag{8.40}$$

are shown in Figure 8.5. One can imagine the contours as truncated portions of ellipsoids, of the type shown in Figure 5.3. The right panel of Figure 8.4 makes it clear that we are nowhere near the center of the hypothetical bivariate normal density, which must lie well beyond our dimness limit.

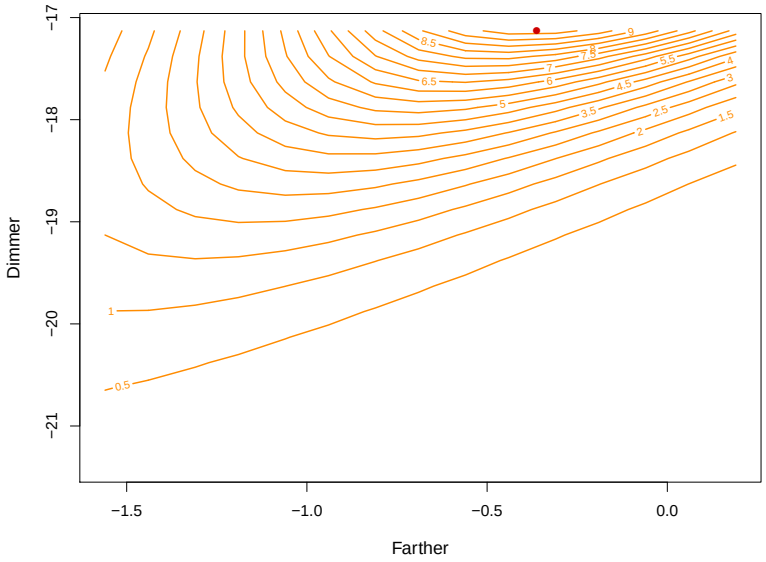

Figure 8.5 Contour curves for Poisson GLM density estimate for the galaxy data. The red dot shows the point of maximum density.

The Poisson *deviance residual* Z between an observed count y and a fitted value $\hat{\mu}$ is

$$Z = \text{sign}(y - \hat{\mu})D(y, \hat{\mu})^{1/2}, \tag{8.41}$$

with D the Poisson deviance from Table 8.4. Z_{jk}, the deviance residual between the count y_{ij} in the ijth bin of Table 8.5 and the fitted value $\hat{\mu}_{jk}$ from the Poisson GLM, was calculated for all 270 bins. Standard frequentist GLM theory says that $S = \sum_{jk} Z_{jk}^2$ should be about 270 if the bivariate normal model (8.39) is correct.[11] Actually the fit was poor: $S = 610$.

In practice we might try adding columns to $\boldsymbol{X}$ in (8.39), e.g., $\boldsymbol{rm}^2$ or $\boldsymbol{r}^2\boldsymbol{m}^2$, improving the fit where it was worst, near the boundaries of the table. Chapter 12 demonstrates some other examples of Poisson density estimation. In general, Poisson GLMs reduce density estimation to regression model fitting, a familiar and flexible inferential technology.

[11] This is a modern version of the classic chi-squared goodness-of-fit test.

8.4 Regression Trees

The data set $\boldsymbol{d}$ for a regression problem typically consists of N pairs (x_i, y_i),

$$\boldsymbol{d} = \{(x_i, y_i),\ i = 1, 2, \ldots, N\}, \tag{8.42}$$

where x_i is a vector of *predictors*, or "covariates," taking its value in some space $\mathcal{X}$, and y_i is the *response*, assumed to be univariate in what follows. The regression algorithm, perhaps a Poisson GLM, inputs $\boldsymbol{d}$ and outputs a *rule* $r_{\boldsymbol{d}}(x)$: for any value of x in $\mathcal{X}$, $r_{\boldsymbol{d}}(x)$ produces an estimate $\hat{y}$ for a possible future value of y,

$$\hat{y} = r_{\boldsymbol{d}}(x). \tag{8.43}$$

In the logistic regression example (8.8), $r_{\boldsymbol{d}}(x)$ is $\hat{\pi}(x)$.

There are three principal uses for the rule $r_{\boldsymbol{d}}(x)$.

1 For *prediction*: Given a new observation of x, but not of its corresponding y, we use $\hat{y} = r_{\boldsymbol{d}}(x)$ to predict y. In the **spam** example, the 57 keywords of an incoming message could be used to predict whether or not it is spam.[12] (See Chapter 12.)

2 For *estimation*: The rule $r_{\boldsymbol{d}}(x)$ describes a "regression surface" $\hat{S}$ over $\mathcal{X}$,

$$\hat{S} = \{r_{\boldsymbol{d}}(x),\ x \in \mathcal{X}\}. \tag{8.44}$$

The right panel of Figure 8.4 shows $\hat{S}$ for the galaxy example. $\hat{S}$ can be thought of as estimating S, the *true* regression surface, often defined in the form of conditional expectation,

$$S = \{E\{y|x\},\ x \in \mathcal{X}\}. \tag{8.45}$$

(In a dichotomous situation where y is coded as 0 or 1, $S = \{\Pr\{y = 1|x\},\ x \in \mathcal{X}\}$.)

For estimation, but not necessarily for prediction, we want $\hat{S}$ to accurately portray S. The right panel of Figure 8.4 shows the estimated galaxy density still increasing monotonically in **dimmer** at the top end of the truncation region, but not so in **farther**, perhaps an important clue for directing future search counts.[13] The flat region in the kidney function regression curve of Figure 1.2 makes almost no difference to prediction, but is of scientific interest if accurate.

[12] Prediction of dichotomous outcomes is often called "classification."

[13] Physicists call a regression-based search for new objects "bump hunting."

3 For *explanation*: The 10 predictors for the diabetes data of Section 7.3, **age**, **sex**, **bmi**,..., were selected by the researcher in the hope of explaining the etiology of diabetes progression. The relative contribution of the different predictors to $r_{\boldsymbol{d}}(x)$ is then of interest. *How* the regression surface is composed is of prime concern in this use, but not in use 1 or 2 above.

The three different uses of $r_{\boldsymbol{d}}(x)$ raise different inferential questions. Use 1 calls for estimates of prediction error. In a dichotomous situation such as the **spam** study, we would want to know both error probabilities

$$\Pr\{\hat{y} = \texttt{spam} | y = \texttt{ham}\} \quad \text{and} \quad \Pr\{\hat{y} = \texttt{ham} | y = \texttt{spam}\}. \tag{8.46}$$

For estimation, the accuracy of $r_{\boldsymbol{d}}(x)$ as a function of x, perhaps in standard deviation terms,

$$\text{sd}(x) = \text{sd}(\hat{y}|x), \tag{8.47}$$

would tell how closely $\hat{S}$ approximates S. Use 3, explanation, requires more elaborate inferential tools, saying for example which of the regression coefficients α_i in (8.19) can safely be set to zero.

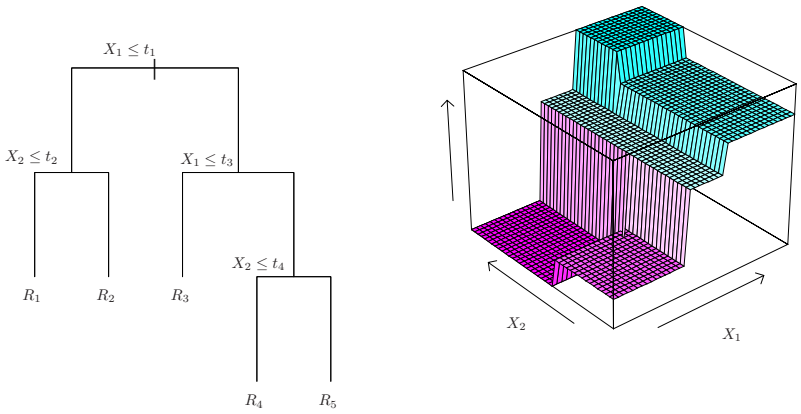

Figure 8.6 *Left* a hypothetical regression tree based on two predictors X_1 and X_2. *Right* corresponding regression surface.

Regression trees use a simple but intuitively appealing technique to form a regression surface: recursive partitioning. The left panel of Figure 8.6 illustrates the method for a hypothetical situation involving two predictor variables, X_1 and X_2 (e.g., r and m in the galaxy example). At the top of

the tree, the sample population of N cases has been split into two groups: those with X_1 equal to or less than value t_1 go to the left, those with $X_1 > t_1$ to the right. The leftward group is itself then divided into two groups depending on whether or not $X_2 \le t_2$. The division stops there, leaving two *terminal nodes* R_1 and R_2. On the tree's right side, two other splits give terminal nodes R_3, R_4, and R_5.

A prediction value $\hat{y}_{R_j}$ is attached to each terminal node R_j. The prediction $\hat{y}$ applying to a new observation $x = (x_1, x_2)$ is calculated by starting x at the top of the tree and following the splits downward until a terminal node, and its attached prediction $\hat{y}_{R_j}$, is reached. The corresponding regression surface $\hat{S}$ is shown in the right panel of Figure 8.6 (here the $\hat{y}_{R_j}$ happen to be in ascending order).

Various algorithmic rules are used to decide which variable to split and which splitting value t to take at each step of the tree's construction. Here is the most common method: suppose at step k of the algorithm, group_k of N_k cases remains to be split, those cases having mean and sum of squares

$$m_k = \sum_{i \in \text{group}_k} y_i / N_k \quad \text{and} \quad s_k^2 = \sum_{i \in \text{group}_k} (y_i - m_k)^2. \tag{8.48}$$

Dividing group_k into $\text{group}_{k,\text{left}}$ and $\text{group}_{k,\text{right}}$ produces means $m_{k,\text{left}}$ and $m_{k,\text{right}}$, and corresponding sums of squares $s^2_{k,\text{left}}$ and $s^2_{k,\text{right}}$. The algorithm proceeds by choosing the splitting variable X_k and the threshold t_k to minimize

$$s^2_{k,\text{left}} + s^2_{k,\text{right}}. \tag{8.49}$$

In other words, it splits group_k into two groups that are as different from each other as possible.† †7

Cross-validation estimates of prediction error, Chapter 12, are used to decide when the splitting process should stop. If group_k is not to be further divided, it becomes terminal node R_k, with prediction value $\hat{y}_{R_k} = m_k$. None of this would be feasible without electronic computation, but even quite large prediction problems can be short work for modern computers.

Figure 8.7 shows a regression tree analysis[14] of the **spam** data, Table 8.3. There are seven terminal nodes, labeled 0 or 1 for decision **ham** or **spam**. The leftmost node, say R_1, is a 0, and contains 2462 **ham** cases and 275 **spam** (compared with 2788 and 1813 in the full data set). Starting at the top of the tree, R_1 is reached if it has a low proportion of \$ symbols

[14] Using the R program **rpart**, in classification mode, employing a different splitting rule than the version based on (8.49).

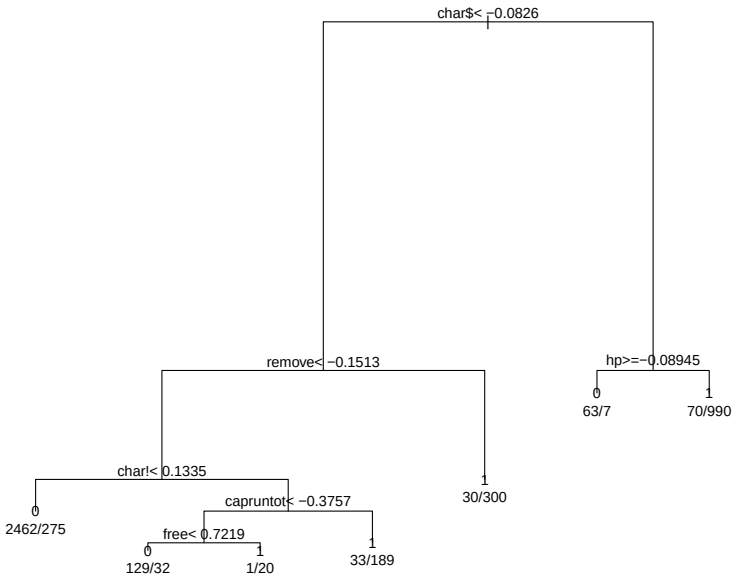

Figure 8.7 Regression tree on the **spam** data; 0 = **ham**, 1 = **spam**. Error rates: **ham** 5.2%, **spam** 17.4%. Captions indicate leftward (ham) moves.

char$, a low proportion of the word **remove**, and a low proportion of exclamation marks **char!**.

Regression trees are easy to interpret ("Too many dollar signs means spam!") seemingly suiting them for use 3, explanation. Unfortunately, they are also easy to overinterpret, with a reputation for being unstable in practice. Discontinuous regression surfaces $\hat{S}$, as in Figure 8.6, disqualify them for use 2, estimation. Their principal use in what follows will be as key parts of prediction algorithms, use 1. The tree in Figure 8.6 has apparent error rates (8.46) of 5.2% and 17.4%. This can be much improved upon by "bagging" (bootstrap aggregation), Chapters 17 and 20, and by other computer-intensive techniques.

Compared with generalized linear models, regression trees represent a break from classical methodology that is more stark. First of all, they are totally nonparametric; bigger but less structured data sets have promoted nonparametrics in twenty-first-century statistics. Regression trees are more computer-intensive and less efficient than GLMs but, as will be seen in Part III, the availability massive data sets and modern computational equipment

has diminished the appeal of efficiency in favor of easy assumption-free application.

8.5 Notes and Details

Computer-age algorithms depend for their utility on statistical computing languages. After a period of evolution, the language **S** (Becker *et al.*, 1988) and its open-source successor **R** (R Core Team, 2015), have come to dominate applied practice.[15] Generalized linear models are available from a single **R** command, e.g.,

```
glm(y~X,family=binomial)
```

for logistic regression (Chambers and Hastie, 1993), and similarly for regression trees and hundreds of other applications.

The classic version of bioassay, *probit analysis*, assumes that each test animal has its own lethal dose level X, and that the population distribution of X is normal,

$$\Pr\{X \le x\} = \Phi(\alpha_0 + \alpha_1 x) \tag{8.50}$$

for unknown parameters (α_0, α_1) and standard normal cdf Φ. Then the number of animals dying at dose x is binomial $\mathrm{Bi}(n_x, \pi_x)$ as in (8.3), with $\pi_x = \Phi(\alpha_0 + \alpha_1 x)$, or

$$\Phi^{-1}(\pi_x) = \alpha_0 + \alpha_1 x. \tag{8.51}$$

Replacing the standard normal cdf $\Phi(z)$ with the logistic cdf $1/(1 + e^{-z})$ (which resembles Φ), changes (8.51) into logistic regression (8.5). The usual goal of bioassay was to estimate "LD50," the dose lethal to 50% of the test population; it is indicated by the open circle in Figure 8.2.

Cox (1970), the classic text on logistic regression, lists Berkson (1944) as an early practitioner. Wedderburn (1974) is credited with generalized linear models in McCullagh and Nelder's influential text of that name, first edition 1983; Birch (1964) developed an important and suggestive special case of GLM theory.

The twenty-first century has seen an efflorescence of computer-based regression techniques, as described extensively in Hastie *et al.* (2009). The discussion of regression trees here is taken from their Section 9.2, including our Figure 8.6. They use the **spam** data as a central example; it is publicly

[15] Previous computer packages such as SAS and SPSS continue to play a major role in application areas such as the social sciences, biomedical statistics, and the pharmaceutical industry.

available at `ftp.ics.uci.edu`. Breiman *et al.* (1984) propelled regression trees into wide use with their CART algorithm.

†1 [p. 112] *Sufficiency as in* (8.13). The Fisher–Neyman criterion says that if $f_\alpha(\boldsymbol{x}) = h_\alpha(S(\boldsymbol{x}))g(\boldsymbol{x})$, when $g(\cdot)$ does not depend on α, then $S(\boldsymbol{x})$ is sufficient for α.

†2 [p. 118] *Equation* (8.28). From (8.24)–(8.25) we have the log likelihood function

$$l_\alpha(\boldsymbol{y}) = \alpha' z - \psi(\alpha) \tag{8.52}$$

with sufficient statistic $z = \boldsymbol{X}'\boldsymbol{y}$ and $\psi(\alpha) = \sum_{i=1}^{N} \gamma(x_i'\alpha)$. Differentiating with respect to α,

$$\dot{l}_\alpha(\boldsymbol{y}) = z - \dot{\psi}(\alpha) = \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{\mu}(\alpha), \tag{8.53}$$

where we have used $d\gamma/d\lambda = \mu_\lambda$ (5.55), so $\dot{\gamma}(x_i'\alpha) = x_i'\mu_i(\alpha)$. But (8.53) says $\dot{l}_\alpha(\boldsymbol{y}) = \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{\mu}(\alpha))$, verifying the MLE equation (8.28).

†3 [p. 118] *Concavity of the log likelihood.* From (8.53), the second derivative matrix $\ddot{l}_\alpha(\boldsymbol{y})$ with respect to α is

$$-\ddot{\psi}(\alpha) = -\operatorname{cov}_\alpha(z), \tag{8.54}$$

(5.57)–(5.59). But $z = \boldsymbol{X}'\boldsymbol{y}$ has

$$\operatorname{cov}_\alpha(z) = \boldsymbol{X}'\boldsymbol{\Sigma}(\alpha)\boldsymbol{X}, \tag{8.55}$$

a positive definite $p \times p$ matrix, verifying the concavity of $l_\alpha(\boldsymbol{y})$ (which in fact applies to any exponential family, not only GLMs).

†4 [p. 118] *Formula* (8.30). The sufficient statistic z has mean vector and covariance matrix

$$z \sim (\beta, V_\alpha), \tag{8.56}$$

with $\beta = E_\alpha\{z\}$ (5.58) and $V_\alpha = \boldsymbol{X}'\boldsymbol{\Sigma}(\alpha)\boldsymbol{X}$ (8.55). Using (5.60), the first-order Taylor series for $\hat{\alpha}$ as a function of z is

$$\hat{\alpha} \doteq \alpha + V_\alpha^{-1}(z - \beta). \tag{8.57}$$

Taken literally, (8.57) gives (8.30). In the OLS formula, we have σ^{-2} rather than σ^2 since the natural parameter α for the Normal entry in Table 8.4 is μ/σ^2.

†5 [p. 118] *Formula* (8.33). This formula, attributed to Hoeffding (1965), is a key result in the interpretation of GLM fitting. Applying definition (8.31)

to family (8.32) gives

$$\begin{aligned}\frac{1}{2}D(\lambda_1,\lambda_2) &= E_{\lambda_1}\{(\lambda_1-\lambda_2)y-[\gamma(\lambda_1)-\gamma(\lambda_2)]\}\\ &= (\lambda_1-\lambda_2)\mu_1-[\gamma(\lambda_1)-\gamma(\lambda_2)]\,.\end{aligned}\tag{8.58}$$

If λ_1 is the MLE $\hat{\lambda}$ then $\mu_1 = y$ (from the maximum likelihood equation $0 = d[\log f_\lambda(y)]/d\lambda = y - \dot{\gamma}(\lambda) = y - \mu_\lambda$), giving[16]

$$\frac{1}{2}D\left(\hat{\lambda},\lambda\right) = \left(\hat{\lambda}-\lambda\right)y - \left[\gamma\left(\hat{\lambda}\right)-\gamma(\lambda)\right] \tag{8.59}$$

for any choice of λ. But the right-hand side of (8.59) is $-\log[f_\lambda(y)/f_y(y)]$, verifying (8.33).

†6 [p. 120] *Table 8.5.* The galaxy counts are from Loh and Spillar's 1988 redshift survey, as discussed in Efron and Petrosian (1992).

†7 [p. 126] *Criteria* (8.49). Abbreviating "left" and "right" by l and r, we have

$$s_k^2 = s_{kl}^2 + s_{kr}^2 + \frac{N_{kl}N_{kr}}{N_k}(m_{kl}-m_{kr})^2, \tag{8.60}$$

with N_{kl} and N_{kr} the subgroup sizes, showing that minimizing (8.49) is the same as maximizing the last term in (8.60). Intuitively, a *good* split is one that makes the left and right groups as different as possible, the ideal being all 0s on the left and all 1s on the right, making the terminal nodes "pure."

[16] In some cases $\hat{\lambda}$ is undefined; for example, when $y = 0$ for a Poisson response, $\hat{\lambda} = \log(y)$ which is undefined. But, in (8.59), we assume that $\hat{\lambda}y = 0$. Similarly for binary y and the binomial family.