Chapter 14
Generalized Linear and Nonlinear Regression

Multiple linear regression is a powerful method of exploring relationships between a response Y and a set of potential explanatory variables $x_1, \ldots, x_p$, but it has an obvious limitation: it assumes the predictive relationship is, on average, linear. In addition, in its standard form it assumes that the noise contributions are homogeneous and follow, roughly, a normal distribution. During the latter part of the twentieth century a great deal of attention was directed toward the development of generalized regression methods that could be applied to nonlinear relationships, with non-constant and non-normal noise variation. In this chapter and in Chapter 15 we discuss several of the most common techniques that come under the heading *modern regression.*

We alluded to modern regression in Chapter 12 by displaying diagram (12.4),

$$Y \leftarrow \begin{cases} \text{noise} \\ f(x_1, \ldots, x_p). \end{cases}$$

To be more specific about the models involved in modern regression let us write the multiple linear regression model (12.44) in the form

$$Y_i = \mu_i + \epsilon_i \tag{14.1}$$
$$\mu_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} \tag{14.2}$$

where $\epsilon_i \sim N(0, \sigma^2)$. In (14.1) and (14.2) we are separating two parts of the model. The deviations from the mean appear in (14.1) as additive noise while, according to Eq. (14.2), the mean itself is a linear function of the x variables. Modern regression models have the more general form

$$Y_i \sim f_{Y_i}(y_i|\theta_i) \tag{14.3}$$
$$\theta_i = f(x_{1i}, \ldots, x_{pi}) \tag{14.4}$$

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_14,

where $f_{Y_i}(y|\theta)$ is some family of pdfs that depend on a parameter θ, which[1] is related to $x_1, \ldots, x_p$ according to a function $f(x_1, \ldots, x_p)$. Here, not only is $f(x_1, \ldots, x_p)$ in (14.4) allowed to be nonlinear, but also the probabilistic representation of noise in (14.3) is more general than in (14.1). The family of pdfs $f_{Y_i}(y|\theta)$ must be specified. In Sections 14.1.1–14.1.3 and 14.1.4–14.1.5 we take the response distributions in (14.3) to be binomial and Poisson, respectively, but in applying (14.4) we retain the linear dependence on $x_1, \ldots, x_p$ for suitable parameters θ_i. In Section 14.1.6 we discuss the formal framework known as *generalized linear models* that encompasses methods based on normal, binomial, and Poisson distributions, along with several others. In Section 14.2 we describe the use of nonlinear functions $f(x_1, \ldots, x_p) = f(x_1, \ldots, x_p; \theta)$ that remain determined by a specified vector of parameters θ (such as $f(x; \theta) = \theta_1 \exp(-\theta_2 x)$).

Modern regression is also used to analyze spike trains, where it becomes *point process regression*. We discuss this in Chapter 19. We lay the foundation for point process regression with our description of Poisson regression, especially in Examples 14.4 and 14.5 in Section 14.2.2.

We hope that our presentation will make the generalization of the regression framework to (14.3) and (14.4) seem straightforward. From our current perspective, it is. Historically, however, the step from least squares to generalized linear models was huge: it required not only the advent of ML estimation, but also the recognition that some widely-used probability distributions had well-behaved likelihood functions (see Section 14.1.6) together with sufficient computational power to perform the fitting in a reasonable amount of time. All of this came together in the publication Nelder and Wedderburn (1972).

14.1 Logistic Regression, Poisson Regression, and Generalized Linear Models

14.1.1 Logistic regression may be used to analyze binary responses.

There are many situations where some y should be a noisy representation of some function of $x_1, \ldots, x_p$, but the response outcomes y are binary. For instance, behavioral responses are sometimes either correct or incorrect and we may wish to consider the probability of correct response as a function of some explanatory variable or variables, or across experimental conditions. Sometimes groups of binary responses are collected into proportions.

Example 5.5 (continued from p. 226) In Fig. 8.9 we displayed a sigmoidal curve fitted to the classic psychophysical data of Hecht et al. (1942) on perception of dim light. There, each response was binary and the 50 binary responses at a given light

[1] We apologize for the double use of f to mean both a pdf in $f_{Y_i}(y|\theta)$ and a general function in $f(x_1, \ldots, x_p)$. These two distinct uses of f are very common. We hope by pointing them out explicitly we will avoid confusion.

intensity could be collected into a proportion out of 50 that resulted in perception. We fit the data by applying maximum likelihood estimation to the logistic regression model in (8.43) and (8.44). This[2] is known as *logistic regression*. □

Example 2.1 (continued from p. 378) In Section 13.2.2 we discussed ANOVA interactions in the context of the study by Behrmann et al. (2002) on hemispatial neglect, where the response was saccadic reaction time and one of the explanatory variables was angle of the starting fixation point of the eyes away from "straight ahead." A second response variable of interest in that study was saccadic error, i.e., whether the patient failed to execute the saccade within a given time window. Errors may be coded as 0 and successful execution as 1. Behrmann et al. (2002) used logistic regression to analyze the error rate as a function of the same explanatory variables. They found, for example, that the probability of error increased as eyes fixated further to the right. □

From (14.1) and (14.2) together with normality, for a single explanatory variable x, in linear regression we assume

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2).$$

There are three problems in applying ordinary linear regression with binary responses to obtain fitted probabilities: (i) a line won't be constrained to (0, 1), (ii) the variances are not equal, and (iii) the responses are not normal (unless we have proportions among large samples, in which case the proportions would be binomial for large n and thus would be approximately normal, as in Section 5.2.2). The first problem, illustrated in Fig. 8.9, is that the linear regression may not make sense beyond a limited range of x values: if $y = a + bx$ and $b > 0$ then y must become infinitely large, or small, as x does. In many data sets with dichotomous or proportional responses there is a clear sigmoidal shape to the relationship with x. The second problem was discussed in the simpler context of estimating a mean, in Section 8.1.3. There we derived the best set of weights to be used for that problem, and showed that an estimator that omits weights can be very much more variable, effectively throwing away a substantial portion of the data. Much more generally it is also possible to solve

[2] The analysis of Hecht et al. (1942) was different, but related. They wished to obtain the minimum number of quanta, n, that would produce perception. Because quanta are considered to follow a Poisson distribution, in the notation we used above, they took $W \sim P(\lambda)$ and $c = n$, with λ, the mean number of quanta falling on the retina, being proportional to the intensity. This latter statement may be rewritten in the form $\log \lambda = \beta_0 + x$, with x again being the log intensity. Then $Y = 1$ (light is perceived) if $W \geq n$ which occurs with probability $p = 1 - P(W \leq n-1) = 1 - F(n-1|\lambda)$, where F is the Poisson cdf. This is a latent-variable model for the proportional data (similar to but different than the one on p. 399). It could be fitted by finding the MLE of β_0, though Hecht et al. apparently did the fitting by eye. Hecht et al. then determined the value of n that provided the best fit. They concluded that a very small number of quanta sufficed to produce perception, but see also Teich et al. (1982).

problem (ii) by using weighted least squares, as discussed surrounding Eq. (12.64), and such solutions apply to the logistic regression setting. The third problem can make distributional results (standard errors and p-values) suspect. The method of logistic regression, which applies maximum likelihood to the logistic regression model, fixes all three problems.

The logistic regression model begins with the log-odds transformation. Recall that when p is a probability the associated *odds* are $p/(1-p)$. The number p lies in the range $(0, 1)$ while the associated odds is in the range $(0, \infty)$. If we then take logs, the number $\log(p/(1-p))$ will lie in the range $(-\infty, \infty)$, which corresponds to what we need for infinite straight lines. Therefore, instead of taking the expected value of Y to be linear in x ($E(Y_i) = \beta_0 + \beta_1 x_i$) we note that when $Y_i \sim B(n_i, p_i)$ we have $E(Y_i/n_i) = p_i$ and we apply $\log(p_i/(1-p_i)) = \beta_0 + \beta_1 x_i$. First, from the algebraic manipulations given in our discussion of Example 5.5 on p. 226, substituting z for u and w for p in (9.8) and (9.9), we have

$$z = \log\left(\frac{w}{1-w}\right) \Longleftrightarrow w = \frac{\exp(z)}{1+\exp(z)}. \tag{14.5}$$

In (14.5) we replace w with p_i and z with $\beta_0 + \beta_1 x_i$. The logistic regression model (8.43) and (8.44) may then be written in the form

$$\begin{aligned} Y_i &\sim B(n_i, p_i) \\ \log\frac{p_i}{1-p_i} &= \beta_0 + \beta_1 x_i. \end{aligned}$$

The log-odds (or *logit*) transformation is helpful in interpreting results. The log odds (of a response) are linear in x. Thus, β_1 is the change in the log odds for a unit change in x.

The log odds scale itself is a bit awkward to think about, though if the base of the logarithm is changed from e to 2 or 10 it becomes easier. It is often useful to transform back to the odds scale, where an increase of 1 unit in x is associated with an increase in the odds (that $Y = 1$) by a factor of $\exp(\beta_1)$. If we wish to interpret the change in probabilities, we must pick a particular probability p and conclude that a unit increase in x is associated with an increase from p to $expit(logit(p) + \beta_1)$, where $logit(z) = \log(z/(1-z))$ and $expit(w) = \exp(w)/(1+\exp(w))$. To illustrate, we provide some interpretation in the context of Example 5.5.

Example 5.5 (continued) On p. 213 we found $\hat{\beta}_1 = 10.7$ with standard error $SE = 1.2$. We interpret the fitted model as saying that, on average, for every increase of intensity by a factor of 10 (1 unit on the scale of the explanatory variable) there is a 10.7 ± 1.2 increase in the log odds of a response. To get an approximate 95 % CI for the factor by which the odds increase we exponentiate, $\exp(10.7 \pm 2(1.2))$, i.e., (4023, 489000). A more interpretable intensity change, perhaps, would be doubling. An increase in intensity by a factor of 2 corresponds to .30 units on the scale of the explanatory variable (because $\log_{10}(2) = .301$). For an increase of intensity by a

factor of 2 the log odds thus increase by $3.22 \pm .72$ (where $3.22 = (.301)(10.7)$ and $.72 = (.301)(2.4)$). This gives an approximate 95 % CI for the factor by which the odds increase, when the intensity doubles, of $\exp(3.22 \pm .72) = (12.2, 51.4)$.

We can go somewhat further by converting odds to the probability scale by inverting

$$\text{odds} = \frac{p}{1-p}$$

to get

$$p = \frac{\text{odds}}{1+\text{odds}}.$$

Let us pick $p = .5$, so that the odds are 1. If we increase the odds by a factor ranging from 12.2 to 51.4 then the probability would go from .5 to somewhere between .92 and .98 (where $.92 = 12.2/(1+12.2)$ and $.98 = 51.4/(1+51.4)$). Thus, if we begin at the x_{50} intensity (where $p = .5$) and then double the intensity, we would obtain a probability of perception between .92 and .98, with 95 % confidence. This kind of calculation may help indicate what the fitted model implies. □

Logistic regression extends immediately to multiple explanatory variables: for m variables $x_1, \ldots, x_m$ we write

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_m x_{mi}.$$

The multiple logistic regression model may be written in the form

$$\begin{aligned} Y_i &\sim B(n_i, p_i) \\ \log \frac{p_i}{1-p_i} &= x_i \beta \end{aligned} \tag{14.6}$$

where β is the coefficient vector and x_i is the $1 \times (m+1)$ vector of values of the several explanatory variables corresponding the ith unit under study.

14.1.2 In logistic regression, ML is used to estimate the regression coefficients and the likelihood ratio test is used to assess evidence of a logistic-linear trend with x.

It is not hard to write down the likelihood function for logistic regression. The responses Y_i are independent observations from $B(n_i, p_i)$ distributions, so each pdf has the form $\binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i}$ and the likelihood function is

Table 14.1 Linear regression results for data from subject S.S. in Example 5.5.

Variable	Coefficients	SE	t_{obs}	p-value
Intercept	−1.78	.30	−5.9	.0042
Intensity	1.20	.16	7.5	.0017

Table 14.2 Logistic regression results for data from subject S.S. in Example 5.5.

Variable	Coefficients	SE	t_{obs}	p-value
Intercept	−20.5	2.4	−8.6	$p < 10^{-6}$
Intensity	10.7	1.2	8.9	$p < 10^{-6}$

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{n_i - y_i}$$
$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

where the second equation is substituted into the first. Standard statistical software may be used to maximize this likelihood. The standard errors are obtained from the observed information matrix, as described in Section 8.3.2.

For a single explanatory variable, the likelihood ratio test of Section 11.1.3 may be used to test $H_0 : \beta_1 = 0$. More generally, if there are variables $x_1, \ldots, x_p$ in model 1 and additional variable $x_{p+1}, \ldots, x_{p+m}$ in model 2, then the likelihood ratio test may again be applied to test $H_0 : \beta_{p+1} = \cdots = \beta_{p+m} = 0$. The log likelihood ratio has the form

$$-2 \log LR = -2[\log(\hat{L}_1) - \log(\hat{L}_2)]$$

where $\hat{L}_i$ is the maximum value of the likelihood under model i. For large samples, under H_0, $-2 \log LR$ follows the χ^2 distribution with m degrees of freedom.

In some software, the results are given in terms of "deviance." The *deviance* for a given model is $-2 \log(\hat{L})$. The *null deviance* is the deviance for the "intercept-only" model, and we denote it by $-2 \log \hat{L}(0)$. Often, the deviance from the full fitted model is called the *residual deviance*. In this terminology, the usual test of $H_0 : \beta_1 = 0$ is based on the difference between the null deviance and the residual deviance.

Example 5.5 (continued) The output from least-squares regression software is given in Table 14.1. The F statistic in this case is the square of t_{obs} and gives the $p = .0017$, as in Table 14.1. The results for logistic regression are given in Table 14.2. The null deviance was 257.3 on 5 degrees of freedom and the residual deviance was 2.9 on 4 degrees of freedom. The difference in deviance is

$$\text{null deviance} - \text{residual deviance} = 257.3 - 2.9 = 256.4$$

Table 14.3 Quadratic logistic regression results for data from subject S.S. in Example 5.5.

Variable	Coefficients	SE	t_{obs}	p-value
Intercept	−4.3	15.8	−.27	.78
Intensity	−6.6	17.0	−.39	.70
Intsq	4.6	4.6	1.0	.31

Table 14.4 Quadratic logistic regression results for data from subject S.S. in Example 5.5, after first centering the intensity variable.

Variable	Coefficient	SE	t_{obs}	p-value
Intercept	−20.3	2.3	−8.7	$p < 10^{-6}$
Intensity	10.5	1.2	8.6	$p < 10^{-6}$
Int2	4.6	4.6	1.0	.31

which should be compared to the chi-squared distribution on 1 degree of freedom. It is very highly significant, consistently with the result in Table 14.2. □

Polynomial terms in x may be handled in logistic regression just as they are in linear regression (Section 12.5.4).

Example 5.5 (continued) To consider whether an additional, nonlinear component might contribute usefully to the linear logistic regression model, we may square the intensity and try including it in a two-variable logistic regression model. In this case it is interesting to note that intensity and its square are highly correlated. To reduce the correlation it helps to subtract the mean before squaring. Thus, we define intsq $=$ (intensity)^2 and int2 $=$ (intensity $-$ mean(intensity))^2. The results using the alternative variables intsq and int2 are shown in Tables 14.3 and 14.4, respectively. Using either of these two logistic regression summaries we would conclude the quadratic term does not improve the fit. The results in Table 14.3 might, at first, be confusing because of the nonsignificant p-values. As we noted in Section 12.5.5, this is a fairly common occurrence with highly correlated explanatory variables, as x and x^2 often are. Recall that each nonsignificant p-value leads to the conclusion that its corresponding variable contributes little *in addition to* the other variable. Since we already found a very highly significant logistic linear relationship, we would conclude that the quadratic doesn't improve the fit. Again, though, the interpretation appears cleaner in the second formulation. □

In non-normal regression models there is no fully satisfactory generalization of the measure of fit R^2. One useful measure, proposed by Nagelkerke (1991) and usually called the *Nagelkerke* R^2, is defined by

$$R_N^2 = 1 - \left(\frac{\hat{L}(0)}{\hat{L}} \right)^{\frac{2}{n}}$$

where, again, $\hat{L}(0)$ is the maximized likelihood for the intercept-only model and $\hat{L}$ is the maximized likelihood for the model being considered. Because the maximum value of R_N^2 may be less than 1, a scaled version is often used:

$$R^2_{\text{scaled } N} = \frac{R_N^2}{R^2_{max}}$$

where

$$R^2_{max} = 1 - \left(\hat{L}(0)\right)^{\frac{2}{n}}.$$

14.1.3 The logit transformation is one among many that may be used for binomial responses, but it is the most commonly applied.

The *expit* function $\exp(x)/(1 + \exp(x))$, defined in Section 14.1.1, is one of many possible sigmoidal curves and thus logistic regression is only one of many possible models for binary or proportion data. In fact, expit(x) has an asymptote at 0 as $x \to -\infty$ and at 1 as $x \to \infty$, and is increasing, so it is a cumulative distribution function. The distribution having expit(x) as its cdf is called the *logistic distribution*, but the cdf of any continuous distribution could be used instead. One important alternative to logistic regression is the Probit regression model, which substitutes the normal cdf in place of the *expit*: specifically, the probit model is

$$\begin{aligned} Y_i &\sim B(n_i, p_i) \\ \Phi^{-1}(p_i) &= \beta_0 + \beta_1 x_i \end{aligned}$$

where $\Phi(z) = P(Z \le z)$, with $Z \sim N(0, 1)$. The fitted curve is then obtained from $y = \Phi(\hat{\beta}_0 + \hat{\beta}_1 x)$.

Example 5.5 (continued) Figure 14.1 displays the fitted curves from probit and logistic regression for the data shown previously in Fig. 8.9. The two models produce nearly identical fitted curves. □

As with the threshold data, the fitted curves from probit and logistic regression are generally very close to each other. This is because the graph of the logistic cdf (the *expit* function) is close to the graph of the normal cdf. Two things are special about the logistic regression model. First, it gives a nice interpretation of the coefficients in terms of log odds. Second, in the logistic regression model (but not the Probit or other versions) the loglikelihood function is necessarily concave (as long as there are at least two distinct values of x). This means that there is a unique MLE, which can be obtained from an arbitrary starting value in the iterative algorithm. Logistic

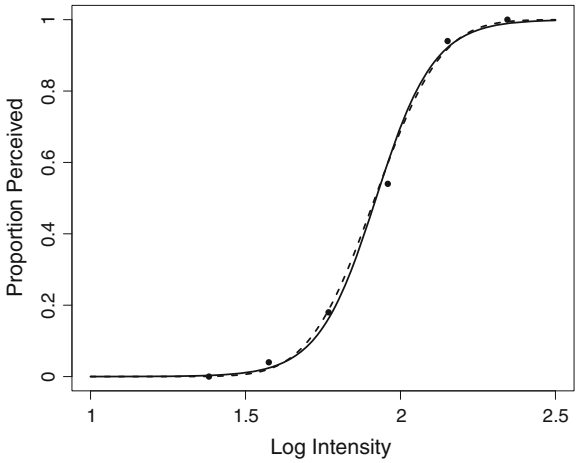

Fig. 14.1 Two curves fitted to the data in Fig. 8.9. The fitted curve from probit regression (*dashed line*) is shown together with the fitted curve from logistic regression. The fits are very close to each other.

regression is the standard method for analyzing dichotomous or proportional data, though in some contexts probit regression remains popular.[3]

An interesting interpretation of binary phenomena involves the introduction of *latent variables*, meaning random variables that become part of the statistical model but are never observed (see the illustration on p. 216 and Section 16.2). Let us discuss this in terms of perception, and let us imagine that the binary experience of perception, as "perceived" or "not perceived" is controlled by an underlying continuous random variable, which we label W. We may think of W as summarizing the transduction process (from light striking the retina to firing rate among multiple ganglion cells), so that perception occurs whenever $W > c$ for some constant c. Neither the precise meaning of W, nor the units of c need concern us. Let us take W to be normally distributed and, because the units are arbitrary, we take its standard deviation to be 1. Finally, we take this latent transduction variable, on average, to be a linear function of the log intensity of light x and we write this in the form $\mu_W = c + \beta_0 + \beta_1 x$. We now have the probit regression model: $Y = 1$ when $W > c$ but, defining $-Z = W - \mu_W$ (so that $-Z \sim N(0, 1)$ and $Z \sim N(0, 1)$),

$$W > c \iff W - \mu_W > c - \mu_W \iff -Z > c - \mu_W \iff Z < \mu_W - c.$$

In other words, $Y = 1$ when $Z < \beta_0 + \beta_1 x$, which occurs with probability $p = \Phi(\beta_0 + \beta_1 x)$.

This latent-variable interpretation helps transfer the intuition of linear regression models over to the binary case, and provides an appealing way to think about many

[3] We have not discussed residual analysis here. It may be performed using *deviance residuals,* or other forms of residuals. See Agresti (1990) or McCullagh and Nelder (1989).

Table 14.5 Spike counts from an SEF neuron during directional saccades.

left	9	6	9	9	6	6	8	5	7	9	4	8	8	3	6
Up	2	0	6	4	4	0	0	0	5	2	1	0	3	0	
Right	4	8	2	2	4	0	3	4	1	1	0	3	4	0	2
Down	1	5	1	2	0	4	4	4	4	3	6	1	1	1	

phenomena. Note that logistic regression is obtained by taking W to have a *logistic distribution*,[4] having cdf

$$F(w) = \frac{1}{1+e^{-w}}.$$

14.1.4 The usual Poisson regression model transforms the mean λ to $\log \lambda$.

The simplest distribution for counts is Poisson, $Y \sim P(\lambda)$. Here, the Poisson mean must be positive and it is therefore natural to introduce dependence on explanatory variables through $\log \lambda$. In Section 14.1.6 we will note that models defined in terms $\log \lambda$ have special properties. The usual multiple Poisson regression model is

$$\begin{aligned} Y_i &\sim P(\lambda_i) \\ \log \lambda_i &= x_i \beta \end{aligned}$$

where β is the coefficient vector and x_i is the $1 \times (m+1)$ vector of values of the explanatory variables corresponding to the ith unit under study. Poisson regression is useful when we have counts depending on one or more explanatory variables.

Example 14.1 Directional sensitivity of an SEF neuron Olson et al. (2000) reported data collected from many individually-recorded neurons in the supplementary eye field (SEF). In this experiment, a monkey was trained to translate one of four possible icons displayed at the fixation point into an instruction of a location to which he was to move his eyes: either left, up, right, or down. SEF neurons tend to be directionally sensitive. To establish direction sensitivity, Olson et al. examined the number of spikes occurring 600–750 ms after presentation of the cue. The spike count data for one neuron across the various trials are given in Table 14.5. Is this neuron directionally sensitive?

By eye it appears that the firing rate is higher for the "left" condition than for the other conditions. There are various versions of ANOVA that may be used to check this. Analysis of spiking activity from these SEF neurons revealed that while the

[4] Probit regression was introduced by Chester Bliss in 1934, but the latent variable idea and normal cdf-transformation was part of Fechner's thinking about psychophysics in 1860; logistic regression was apparently discussed first by Ronald Fisher and Frank Yates in 1938. See Agresti (1990) for much more extensive discussion of the methods described briefly here.

spike counts deviated from that predicted by a Poisson distribution, the deviation was small (Ventura et al. 2002). Here we will use the data to illustrate a version of ANOVA based on Poisson regression. Note that in Table 14.5 there are a total of 58 spike counts, from 58 trials. □

The problem of fitting counts is analogous to, though less extreme than, that of fitting proportions. For proportions, the (0,1) range could make linear regression clearly inappropriate. Counts have a range of $(0, \infty)$. Because the ordinary regression line is not constrained, it will eventually go negative. The simple solution is to use a log transformation of the underlying mean. The usual Poisson regression model is

$$Y_i \sim P(\lambda_i) \tag{14.7}$$

$$\lambda_i = \exp(\beta_0 + \beta_1 x_i). \tag{14.8}$$

To interpret the model we use the log transformation:

$$\log \lambda_i = \beta_0 + \beta_1 x_i.$$

For example, in the SEF data of Example 14.1 Y_i is the spike count and x_i is the experimental condition (up, down, left, right) for the ith trial. The advantage of viewing ANOVA as a special case of regression is apparent: we immediately generalize Poisson ANOVA by applying our generalization of linear regression to the Poisson regression model above.

14.1.5 In Poisson regression, ML is used to estimate coefficients and the likelihood ratio test is used to examine trends.

As in logistic regression we use ML estimation and the likelihood ratio test ("analysis of deviance").

Example 14.1 (continued) We perform Poisson regression using indicator variables as described in Section 13.2.1 to achieve an ANOVA-like model. Specifically, we concatenate the data in Table 14.5 so that the counts form a 58×1 vector and define a variable *left* to be 1 for all data corresponding to the left saccade direction and 0 otherwise, and similarly define vectors *up* and *right*. The results from ordinary least-squares regression are shown in Table 14.6. The F-statistic was 18.76 on 3 and 54 degrees of freedom, giving $p < 10^{-6}$. The Poisson regression output, shown in Table 14.7 is similar in structure. Here the null Deviance was 149.8 on 57 degrees of freedom and the residual Deviance was 92.5 on 54 degrees of freedom. The difference in deviances is

$$\text{null deviance - residual deviance} = 149.8 - 92.5 = 57.3$$

Table 14.6 ANOVA Results for the SEF data in Table 14.5 shown in the form of regression output.

Variable	Coefficient	SE	t_{obs}	p-value
Intercept	3.49	.26	13.2	$p < 10^{-6}$
Left	2.11	.37	5.6	$p < 10^{-6}$
Up	−.74	.21	−3.5	.0011
Right	−.52	0.15	−3.4	.0014

Table 14.7 Poisson regression results for the SEF data in Table 14.5. The form of the results is similar to that given in Table 14.6.

Variable	Coefficients	SE	t_{obs}	p-value
Intercept	1.12	.079	14.2	$p < 10^{-6}$
Left	.475	.096	4.9	3×10^{-6}
Up	−.173	.063	−2.76	.0039
Right	−.155	.052	−2.96	.0023

which should be compared to the chi-squared distribution on 3 degrees of freedom. It is very highly significant. □

In Example 14.1 the results from Poisson regression were the same as with ordinary linear regression (standard ANOVA), but the details are different. In some situations the conclusions drawn from the two methods could be different.

14.1.6 Generalized linear models extend regression methods to response distributions from exponential families.

We began this chapter by saying that modern regression models have the form given by (14.3) and (14.4), which for convenience we repeat:

$$\begin{aligned} Y_i &\sim p(y_i|\theta_i) \\ \theta_i &= f(x_i). \end{aligned}$$

The simple logistic regression model may be put into this form by writing

$$\begin{aligned} Y_i &\sim B(n_i, p_i) \\ \theta_i &= \beta_0 + x_i\beta_1 \end{aligned}$$

where

$$\theta_i = \log \frac{p_i}{1 - p_i}$$

or, more succinctly,

$$Y_i \sim B(n_i, p_i)$$
$$\log \frac{p_i}{1-p_i} = \beta_0 + x_i\beta_1.$$

Similarly, the simple Poisson regression model may be written

$$Y_i \sim P(\lambda_i)$$
$$\log \lambda_i = \beta_0 + x_i\beta_1.$$

Logistic and Poisson regression are special cases of *generalized linear models*. These generalize linear regression by allowing the response variable to follow a distribution from a certain class known as *exponential families*. They also use a *link* function that links the expected value (the mean) μ_i of the data with the linear model $\beta_0 + \beta_1 x_i$. For example, the usual link functions for binomial and Poisson data are the log odds and the log, respectively, as shown above.

Exponential families have pdfs of the form

$$f_Y(y|\eta(\theta)) = h(y)\exp(\eta(\theta)T(y) - B(\theta)). \tag{14.9}$$

For instance, in the Poisson case $Y \sim P(\lambda)$, the pdf (from Chapter 5, p. 112) is

$$P(Y = y) = \frac{1}{y!}\lambda^y e^{-\lambda}.$$

We can rewrite this in the form

$$\frac{1}{y!}\lambda^y e^{-\lambda} = \frac{1}{y!}\exp(y\log\lambda - \lambda).$$

If we let $\theta = \lambda$, $\eta(\theta) = \log\lambda$, $B(\lambda) = \lambda$, $T(y) = y$ and $h(y) = 1/y!$ we obtain (14.9). Now, with $\mu = \lambda$, if we define the link function to be

$$g(\mu) = \log\mu \tag{14.10}$$

the simple Poisson regression model becomes

$$g(\mu) = \beta_0 + \beta_1 x_i.$$

Here, the log provides the link in the sense that it is the function by which the mean is transformed before being equated to the linear model.

We may rewrite (14.9) in the form

$$f_Y(y|\eta) = h(y)\exp(\eta T(y) - A(\eta))$$

in which case $\eta = \eta(\theta)$ is called the *natural parameter* (or *canonical parameter*). In the Poisson case the natural parameter is $\log \lambda$. The logarithmic link function is thus often called *the canonical link*. In the binomial case the log odds function becomes the canonical link. The statistic $T(y)$ is *sufficient* in the sense described on p. 200. The extension to the multiparameter case, in which η and $T(y)$ are vectors, is immediate:

$$f_Y(y|\eta) = h(y) \exp(\eta^T T(y) - A(\eta)). \tag{14.11}$$

Assuming that Y_i comes from an exponential family, we obtain a generalized linear model by writing

$$g(\mu_i) = \beta_0 + \beta_1 x_i, \tag{14.12}$$

where $\mu_i = E(Y_i)$. Equation (14.10) provided an example in the Poisson case, but in (14.12) $g(\mu)$ may be any link function.

Common distributions forming exponential families include binomial, multinomial, Poisson, normal, inverse Gaussian, gamma, and beta. The introduction of generalized linear models allowed regression methods to be extended immediately to all of these families, and a multiple-variable generalized linear model may be written

$$\begin{aligned} Y_i &\sim f_{Y_i}(y_i|\eta_i) \\ g(\mu_i) &= x_i\beta \end{aligned} \tag{14.13}$$

where $f_{Y_i}(y_i|\eta_i)$ is an exponential family pdf as in (14.11), $\mu_i = E(Y_i)$, and $g(\mu)$ is the link function. The unification of mathematical form meant that implementation of maximum likelihood, and likelihood ratio tests, could use the same algorithms with only minor changes in each particular case. Furthermore, for the canonical link it turns out (under relatively mild conditions on the x and y variables[5]) that the loglikelihood function is concave so that the MLE is unique. This guarantees that the maximum of the loglikelihood function will be found by the function maximizer (using *Newton's method*, i.e., iterative quadratic approximation) beginning with any starting value, and convergence will tend to be fast. Generalized linear models are part of most statistical software.

In addition to the canonical link, several other link functions are usually available in software. For example, it is usually possible to perform binomial regression using the probit link instead of the log odds, or logit link. Similarly, a Poisson regression could be performed using the identity link so that

$$\log \lambda_i = \beta_0 + \beta_1 x_i$$

is replaced by

[5] The regularity conditions insure non-degeneracy. For example, if there is only one x variable, it must take on at least two distinct values so that a line may be fitted. The y observations also must correspond to values that are possible according to the model; in dealing with proportions, for instance, the observed proportions can not all be zero.

$$\lambda_i = \beta_0 + \beta_1 x_i.$$

Occasionally, the identity link provides a better description of the data than the canonical link, as in Example 14.3 on p. 406.

The terminology "generalized linear model" should not to be confused with "the general linear model," which is the matrix form of regression and includes ANOVA. Both have the acronym GLM. Also, the "linear" part of the terminology is misleading because the framework really includes *nonlinear* and *nonparametric* models, as well. Specifically, while linear models with the canonical link have especially nice properties, more generally in Equation (14.4) $f(x_i)$ does not need to be linear. See Examples 14.3, 14.4, and 14.5 in Section 14.2.1 and 14.2.2.

14.2 Nonlinear Regression

14.2.1 Nonlinear regression models may be fitted by least squares.

In Section 12.5.4 we pointed out that when $f(x)$ is a polynomial in x, linear regression could be used to fit a function of the form $y = f(x)$ to (x, y) data. This involved the "trick" of starting with an initial definition of x, relabeling it as x_1 and then defining the new variable $x_2 = x_1^2$, and so on for higher-order polynomials. The resulting expectation of Y,

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

followed the form required in the linear regression model. In particular, although the relationship of Y and x, on average, was nonlinear, the *coefficients* entered linearly into the model and therefore—as in any linear regression model—the likelihood equations could be solved easily by linear algebra. A similar trick was used to fit directional tuning data with a cosine function.

There are, however, many nonlinear relationships where this sort of manipulation does not apply. For example, if

$$E(Y) = \theta_1 e^{-\theta_2 x}$$

it is not possible to redefine the x variable so that the form becomes linear in the parameters. Instead, we have the *nonlinear regression model*,

$$Y_i = f(x_i; \theta) + \epsilon_i \tag{14.14}$$

$$f(x_i; \theta) = \theta_1 e^{\theta_2 x_i}. \tag{14.15}$$

Here, the usual assumption is $\epsilon_i \sim N(0, \sigma^2)$, independently (though, again, normality is not crucial).

Models of the form (14.14)–(14.15) may still be fit by least-squares and, in fact, least squares remains a special case of ML estimation. What is different is that the equations defining the least-squares solution (the likelihood equations) are no longer solved by a single linear algebraic step. Instead, they must be solved iteratively. The problem is thus usually called *nonlinear least squares*. Example 1.6 on p. 14 provided an illustration, with the nonlinear function given by (1.6) and the fit based on nonlinear squares given in Fig. 1.5.

Example 14.2 Magnesium block of NMDA receptors NMDA receptors, which are ubiquitous in the vertebrate central nervous system, may be blocked by Magnesium ions (Mg^{2+}). To investigate the quantitative dependence of NMDA-receptor currents on the concentration of Mg^{2+}, Qian et al. (2005) measured currents at various concentrations, then summarized the data using the equation

$$\frac{I}{I_0} = \frac{1}{1 + (\frac{[Mg^{2+}]}{IC_{50}})^{n_H}}$$

where the measurements are the current I and the Magnesium concentration $[Mg^{2+}]$, I_0 being the current in the absence of Mg^{2+}. The free parameters are the "Hill constant" n_H and the 50 % inhibition concentration IC_{50} (when $[Mg^{2+}] = IC_{50}$ we get $I/I_0 = .5$). The authors estimated these constants using nonlinear least squares, and they examined IC_{50} across voltages, and across receptor subunit types. □

The term "nonlinear regression" usually refers to models of the form (14.14). However, similar models may be used with binomial or Poisson responses, and may be fit using ML. The next example illustrates nonlinear regression models using both normal and Poisson distributions.

Example 14.3 Non-cosine directional tuning of motor cortical neurons Amirikian and Georgopoulos (2000) investigated cosine and non-cosine directional tuning for 2-dimensional hand movement among motor cortical neurons. In Section 12.5.4 we considered the cosine tuning model given by (12.67) and (12.68) where, according to (12.67), a neuron's firing rate $\mu(v)$ when the movement is in direction v was linear in the components v_1 and v_2 and the model could be fit using linear regression. To investigate departures from cosine tuning, Amirikian and Georgopoulos used a class of functions involving exponentials that are not amenable to reconfiguration in a linear model and, as a result, reported that the tuning curves in motor cortical neurons, for 2-dimensional hand movement, tend to be substantially narrower than cosine tuning curves.

Examples of nonlinear fits to data from two neurons are shown in Fig. 14.2. The functions fitted were

$$\mu(v) = \mu + \beta \exp(\kappa \cos(\theta - \tau + \eta \cos(\theta - \tau))) \tag{14.16}$$

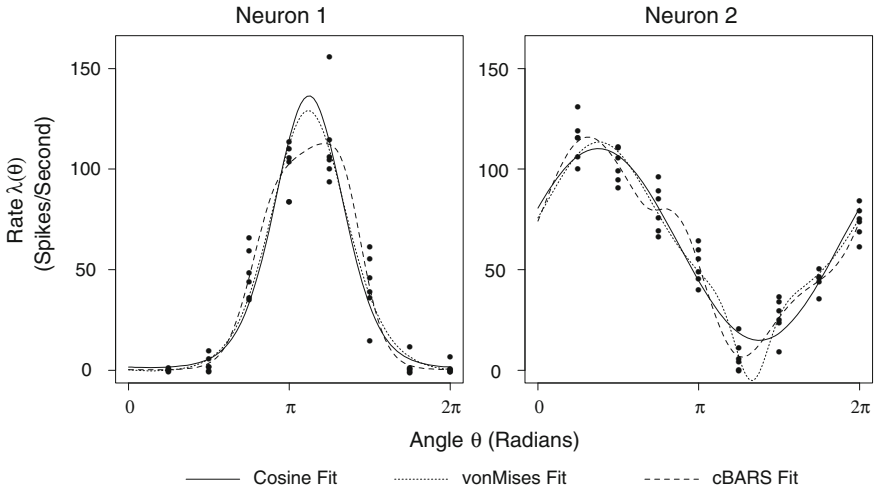

Fig. 14.2 Fits to activity of two neurons in primate motor cortex (reprinted with permission from Kaufman et al. 2005). Each datapoint represents the observed firing rate of a neuron in the motor cortex of a monkey during one repetition of a wrist movement to a particular target. The cosine fits use the cosine function in Eq. (12.67) and the von Mises fits use more complicated parametric forms given by Eq. (14.16), for Neuron 1, and Eq. (14.17) for Neuron 2. The cosine and von Mises parametric fits use Poisson maximum likelihood for Neuron 1 and least squares for Neuron 2. Also shown is the fit from a nonparametric regression method called cBARS, described by Kaufman et al. (2005).

for the first neuron, where $\theta = \arctan(v_2/v_1)$, and

$$\mu(v) = \mu + \beta_1 \exp(\kappa_1 \cos(\theta - \tau_1)) + \beta_2 \exp(\kappa_2 \cos(\theta - \tau_2)) \tag{14.17}$$

for the second neuron. These results come from Kaufman et al. (2005), who also considered nonparametric methods, discussed in Chapter 15. The function in (14.16) includes parameters corresponding roughly to the baseline firing rate, the amplitude, width, and location of the mode, and the skewness about the mode. The function in (14.17) includes parameters corresponding to two modes, one of which is constrained to be in the positive direction and the other in the negative direction. This is of use in fitting the data for the Neuron 2 in Fig. 14.2. For both neurons the data indicate mild but noticeable departures from cosine tuning.

In fact, the data in Fig. 14.2 coming from Neuron 1 exhibited roughly Poisson variation. The fits shown there were based on $Y_i \sim P(\mu_i)$ with $\mu_i = \mu(v)$ given by Eq. (14.16). This is a Poisson nonlinear regression model (with the identity link, as defined in Section 14.1.6). □

Another example of nonlinear least squares has been discussed in earlier chapters. We provide some more details here.

Example 8.2 (continued from p. 241) In presenting this example on p. 193 we said the model took Y to be the spike width and x the preceding ISI length, and assumed there was an ISI length τ such that, on average, Y is quadratic for $x < \tau$ and constant for all $x \geq \tau$. As we noted, τ is called a change point. Specifically, the statistical model was

$$Y_i \sim N(\mu(x_i), \sigma^2) \tag{14.18}$$

independently for $i = 1, \ldots, n$ where

$$\mu(x; \beta_0, \beta_1, \tau) = \begin{cases} \beta_0 + \beta_1 (x - \tau)^2 & \text{if } x < \tau \\ \beta_0 & \text{if } x \geq \tau \end{cases} \tag{14.19}$$

and the least-squares estimate $(\hat{\beta}_1, \hat{\beta}_1, \hat{\tau})$ becomes defined by

$$\sum_{i=1}^{n} \left(y_i - \mu(x_i; \hat{\beta}_0, \hat{\beta}_1, \hat{\tau}) \right)^2 = \min_{\beta_0, \beta_1, \tau} \sum_{i=1}^{n} (y_i - \mu(x_i; \beta_0, \beta_1, \tau))^2 . \tag{14.20}$$

The parameter τ enters nonlinearly into the statistical model, and this makes (14.20) a nonlinear least squares problem. However, for every value of τ we may formulate a simple linear regression problem as follows. Let us define new values $u_1(\tau), \ldots, u_n(\tau)$ by

$$u_i(\tau) = \begin{cases} (x_i - \tau)^2 & \text{if } x_i < \tau \\ 0 & \text{if } x \geq \tau \end{cases}$$

so that $\mu(x_i)$ in (14.19) may be rewritten as

$$\mu(x_i; \beta_0, \beta_1, \tau) = \beta_0(\tau) + \beta_1(\tau) u_i(\tau).$$

We then define $(\hat{\beta}_0(\tau), \hat{\beta}_1(\tau))$ by

$$\sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0(\tau) + \hat{\beta}_1(\tau) u_i) \right)^2 = \min_{\beta_0(\tau), \beta_1(\tau)} \sum_{i=1}^{n} (y_i - (\beta_0(\tau) + \beta_1(\tau) u_i))^2$$

which has the form of the simple least-squares regression problem on p. 12 and thus is easily solved. Finally, defining

$$g(\tau) = \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0(\tau) + \hat{\beta}_1(\tau) u_i) \right)^2 ,$$

the nonlinear least squares problem in (14.20) is found by minimizing $g(\tau)$. This can be achieved in software (e.g., in Matlab) with one-dimensional nonlinear minimization. Therefore, it was easy to implement nonlinear least squares for this change-point problem. □

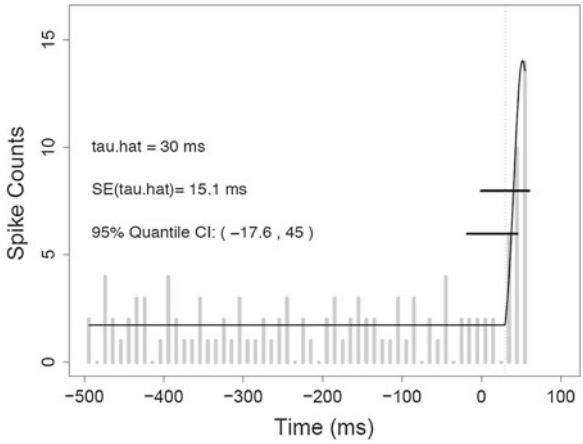

Fig. 14.3 Initiation of firing in a neuron from the basal ganglia: change-point and bootstrap confidence intervals when a quadratic model is used for the post-change-point firing rate. Two forms of approximate 95 % confidence intervals are shown. The first is the usual estimate $\pm 2SE$ interval. The second is the interval formed by the .025 and .975 quantiles among the bootstrap samples. The latter typically performs somewhat better, in the sense of having coverage probability closer to .95. See Sect. 9.2.2.

14.2.2 Generalized nonlinear models may be fitted using maximum likelihood.

Nonlinear relationships also arise in the presence of non-normal noise. We use the term *generalized nonlinear model* to refer to a model in which the linear function $g(\mu_i)$ in (14.13) is replaced by a nonlinear function. We give two examples of nonlinear Poisson regression. The first involves determination of a change-point, and is similar to Example 8.2 in Section 14.2.1.

Example 14.4 Onset latency in a basal ganglia neuron An unfortunate symptom of Parkinson's disease (PD) is muscular rigidity. This has been associated with increased gain and inappropriate timing of the long latency component of the stretch reflex, which is a muscular response to sudden perturbations of limb position. One of the important components of the stretch reflex is mediated by a trans-cortical reflex, probably via corticospinal neurons in primary motor cortex that are sensitive to kinesthetic input. To investigate the neural correlates of degradation in stretch reflex, Dr. Robert Turner and colleagues at the University of Pittsburgh have recorded neurons in primary motor cortex of monkeys before and after experimental production of PD-like symptoms. One part of this line of work aims at characterizing neuronal response latency following a limb perturbation (see Turner and DeLong 2000). Figure 14.3 displays a PSTH from one neuron prior to drug-induced PD symptoms. The statistical problem is to identify the time at which the neuron begins to increase

its firing rate, with the goal being to compare these latencies in the population of neurons before and after induction of PD.

To solve this problem we used a change-point model similar to that used in Example 8.2 on p. 408. In this case, we assume the counts within the PSTH time bins—after pooling the data across trials—follow Poisson distributions. Let Y_t be the pooled spike count in the bin centered at time t and let $\mu(t)$ be its mean. The change-point model assumes the mean counts are constant up until time $t = \tau$, at which time they increase. For simplicity, we assume the count increases as a quadratic. This gives us the Poisson change-point model

$$Y_t \sim P(\mu(t))$$

with

$$\mu(t) = \begin{cases} \beta_0 & \text{if } t \leq \tau \\ \beta_0 + \beta_1(t-\tau)^2 & \text{if } t > \tau. \end{cases}$$

The value τ is the change point. For any fixed τ the change-point model becomes simply a Poisson regression model. Specifically, for a given τ we define

$$x = \begin{cases} 0 & \text{if } t \leq \tau \\ (t-\tau)^2 & \text{if } t > \tau. \end{cases}$$

We then apply Poisson regression with the regression variable x.

However, the parameter τ is unknown and is, in fact, the object of interest. We may maximize the likelihood function iteratively over τ. That is, in software such as *R* or Matlab we set up a loop within which, for a fixed τ, we perform Poisson regression and obtain the value of the loglikelihood. We then iterate until we maximize the loglikelihood across values of τ. This gives us the MLE of τ. We may then obtain a SE for τ by applying a parametric bootstrap. Results are given in Fig. 14.3. □

Here is another example of a nonlinear model for spike counts.

Example 14.5 A Poisson regression model for a hippocampal place cell Neurons in rodent hippocampus have spatially specific firing properties, whereby the spiking intensity is highest when the animal is at a specific location in an environment, and falls off as the animal moves further away from that point (e.g., Brown et al., 1998). Such receptive fields are called *place fields*, and neurons that have such firing properties are called *place cells*. The left panel of Fig. 14.4 shows an example of the spiking activity of one such place cell, as a rat executes a free-foraging task in a circular environment. The rat's path through this environment is shown, and the location of the animal at spike times is overlain as dark dots. It is clear that the firing intensity is highest slightly to the southwest of the center of the environment, and decreases when the rat moves away from this point.

One very simple way to describe this hippocampal neural activity is to use a Poisson generalized linear model for spike counts in successive time bins while the rat forages, and to assume that the spike count depends on location in the environment

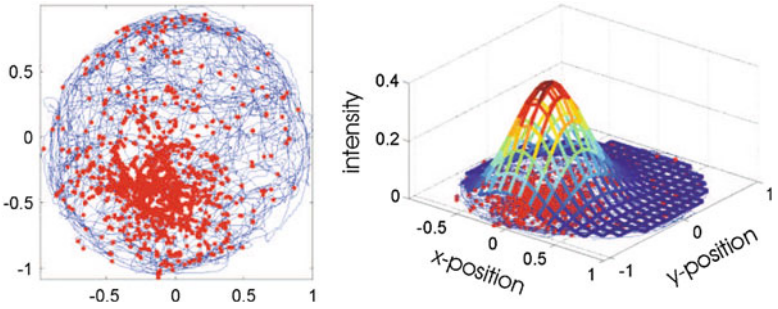

Fig. 14.4 Spiking activity of a rat Hippocampal place cell during a free-foraging task in a circular environment. *Left* Visualization of animal's path and locations of spikes. *Right* Place field model for this neuron, with parameters fit by the method of maximum likelihood.

based on a 2-dimensional bell-shaped curve. For this purpose of specifying the dependence of spiking activity on location a normal pdf may be used. Let us take $Y_t \sim P(\lambda_t)$, with t signifying time, and then define

$$\lambda_t = \exp\left\{\alpha - \frac{1}{2}\left(x(t) - \mu_x \; y(t) - \mu_y\right)\begin{pmatrix}\sigma_x^2 & \sigma_{xy}\\ \sigma_{xy} & \sigma_y^2\end{pmatrix}^{-1}\begin{pmatrix}x(t)-\mu_x\\ y(t)-\mu_y\end{pmatrix}\right\}. \quad (14.21)$$

The explanatory variables in this model are $x(t)$ and $y(t)$, the animal's x and y-position. The model parameters are $(\alpha, \mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \sigma_{xy})$, where (μ_x, μ_y) is the center of the place field, $\exp\alpha$ is the maximum firing intensity at that point, and σ_x^2, σ_y^2, and σ_{xy} express how the intensity drops off away from the center. Note that it is the shape of the place field that is assumed normal, not the distribution of the spiking activity. The right panel of Fig. 14.4 displays a fit of the place field to the data in the left panel. We will discuss models of this sort when we discuss point processes in Chapter 19. □

14.2.3 In solving nonlinear optimization problems, good starting values are important, and it can be helpful to reparameterize.

As in maximization of any likelihood, use of the numerical procedures requires care. Two important issues are the choice of initial values, and of parameterization. Both of these may be illustrated with the exponential model (14.15).

Illustration: Exponential regression To fit the exponential model (14.15) a first step is to reparameterized from θ to ω using $\omega_1 = \log(\theta_1)$ and $\omega_2 = \theta_2$ so that the expected values have the form

$$E(Y) = \exp(\omega_1 + \omega_2 x).$$

The loglikelihood is typically closer to being quadratic as a function of ω than as a function of θ. Taking logs of both sides of this expectation equation gives

$$\log E(Y) = \omega_1 + \omega_2 x.$$

This suggests we may define $U_i = \log(Y_i)$ and apply the linear model,

$$U_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \tag{14.22}$$

The resulting fitted values $\hat{\beta}_0$ $\hat{\beta}_1$ make good starting values for the iterative procedure used to obtain ω_1 and ω_2. □

It is important to recognize the distinction between the exponential model in (14.14) and (14.15) and the linearized version (14.22). Either could be used to fit data, but they make different assumptions about the way the noise contributes. In many examples, the fits based on (14.14) and (14.22) would be very close, but sometimes the resulting inferences would be different. It is an empirical question which model does a better job of describing the data. The point here, however, is that if the exponential form is preferred, the log-linear form may still be used to obtain starting values for the parameters. The linearization method of obtaining starting values is frequently used in fitting nonlinear models. (See Bates and Watts (1988) for further discussion.) These issues also arise in generalized nonlinear models.