Chapter 3

Generalized linear models (GLMs)

3.1 Introduction

I begin this chapter with the re-analysis of a small example from Finney (1978). Table 3.1 gives the data for the number of plates (out of five) on which growth of *Bacillus mesentericus* was successful in different dilutions of a potato flour suspension. Figure 3.1 shows a plot of the data. Several features of the experiment and the plot are worth noting. First, the way the experiment was conducted (positive/negative response for each of five independent trials) guarantees that the distribution of the response is binomial. Second, the form of the response is nonlinear, exhibiting somewhat of an S-shape. Third, extension of the range of the predictor in the positive direction would likely lead to more responses of 1.0 and extension in the negative direction would lead to additional responses of 0.0.

To account for these features, we hypothesize the following model:

$$\begin{aligned} Y_i &\sim \text{indep. binomial}[5, p(x_i)], \\ \log[p(x_i)/\{1-p(x_i)\}] &= \alpha + \beta x_i, \text{ or equivalently,} \\ p(x_i) &= 1/(1+\exp\{-[\alpha+\beta x_i]\}). \end{aligned} \tag{3.1}$$

Thus, p_i is modeled as an S-shaped function of the *linear predictor*, $\alpha + \beta x_i$. The log likelihood for this model is easily specified and is proportional to

$$\log L = \sum_i y_i(\alpha + \beta x_i) - \log[1 + \exp(\alpha + \beta x_i)]. \tag{3.2}$$

This can be easily maximized numerically as a function of α and β to find the maximum likelihood estimates of $\hat{\alpha} = 4.17$ and $\hat{\beta} = 1.62$. Figure 3.1 shows the fit of this model to the data.

TABLE 3.1.
Data for the potato flour example

	Spore Growth		
Dilution	Number of plates	Number positive	Proportion
1/128	5	0	0.0
1/64	5	0	0.0
1/32	5	2	0.4
1/16	5	2	0.4
1/8	5	3	0.6
1/4	5	4	0.8
1/2	5	5	1.0
1	5	5	1.0
2	5	5	1.0
4	5	5	1.0

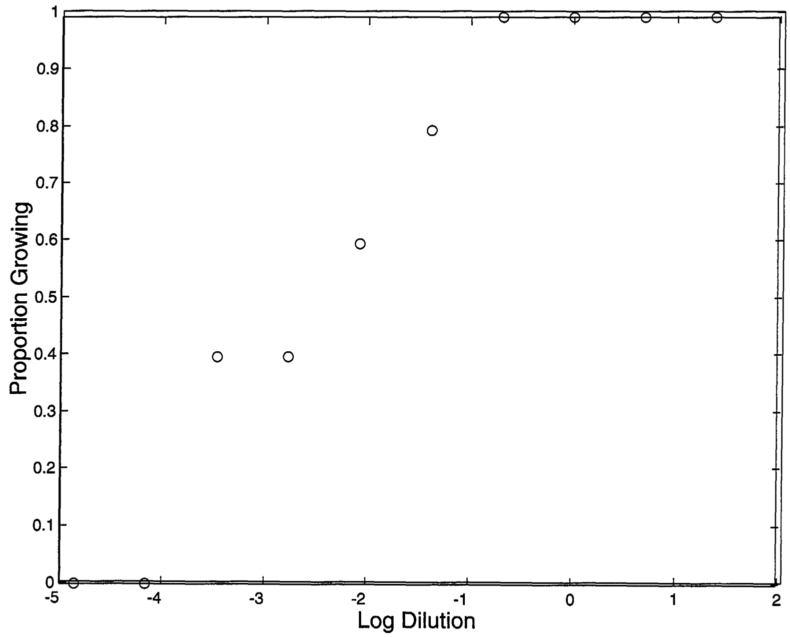

FIG. 3.1. *Plot of proportion of positive plates versus log dilution.*

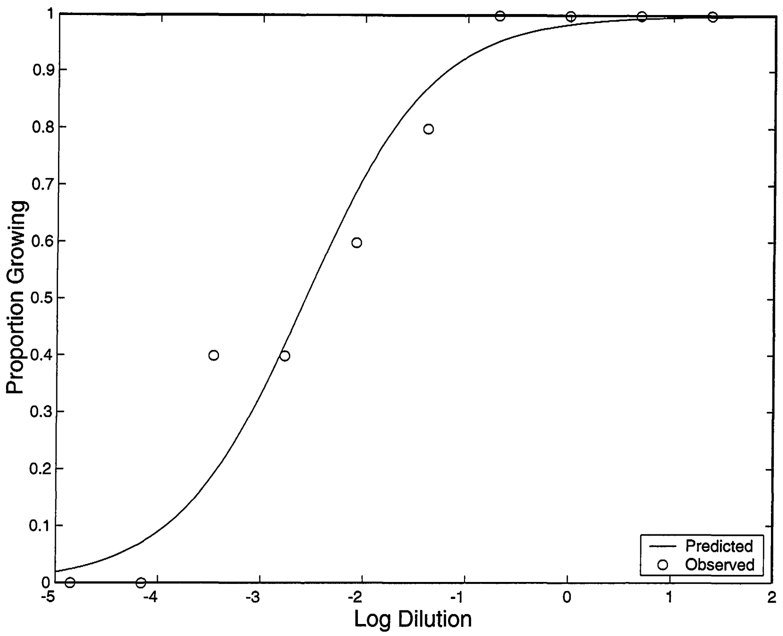

FIG. 3.2. *Plot of data and maximum likelihood fit versus log dilution.*

3.2 The modeling process

Modeling using generalized linear models (GLMs) involves answering three questions:

1. What is the distribution of the data?

2. What aspect of the problem will be modeled?

3. What will the predictors be?

In our example, the distribution was necessarily binomial as governed by the experimental design (we are not usually so lucky in having the proper distribution handed to us!). We chose a logit function to model the S-shaped response (others are possible, for example the probit model of Chapter 1), and we have only a single predictor, log dilution.

In the general case of building a GLM we assign to the data, Y, a distribution, often from the exponential family. Letting μ represent the mean of Y, we then chose a function $g(\cdot)$ such that $g(\mu) = \mathbf{X}\boldsymbol{\beta}$. This function is called the *link* function and it relates, or links, the mean to a linear (in the parameters) combination of the predictor variables. In our example the chosen distribution was binomial, with $\mu = 5p(x_i)$ and $g(\mu) = \log\{(\mu/5)/[1 - (\mu/5)]\}$.

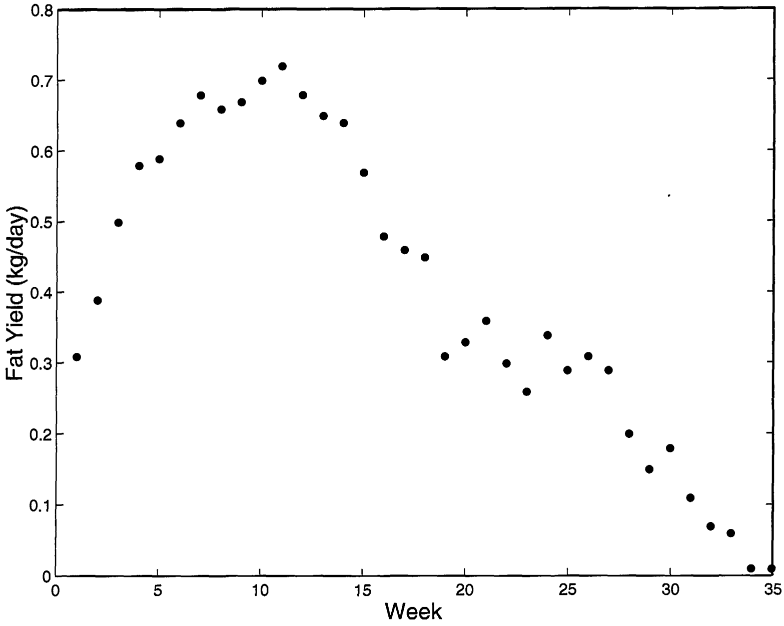

FIG. 3.3. *Plot of fat yield data versus week.*

3.3 Transforming versus linking

Using a link function to transform the mean response is *not* the same as transforming the data itself. An easy situation in which to understand this is Poisson distributed data, for which the log link is a common one. Whereas it is always possible to take the log of the mean response (under the reasonable assumption that the mean is positive), it would often be the case that zero counts would be generated, making it impossible to log transform the actual data.

Here is another example that illustrates the difference. Figure 3.3 gives the fat yields in milk (in kg/day) from a single cow over 35 weeks of lactation (Henderson and McCulloch, 1989). A typical curve that is fit to such data is a nonlinear regression equation called Wood's curve (Wood, 1967), given by

$$\text{FAT}(t) = \alpha t^{\beta} e^{\gamma t} \tag{3.3}$$

where, for the moment, I have avoided specifying either an error term or expectation.

Taking the logarithm of both sides of (3.3) and defining $\alpha^* = \log \alpha$ gives the convenient-looking equation

$$\log \text{FAT}(t) = \alpha^* + \beta \log(t) + \gamma t, \tag{3.4}$$

which can be fit by ordinary least squares regression. This would be a transformation approach and it would be congruent with the assumption of homoscedasticity on the log scale, that is, the following model:

$$\begin{aligned} \log \text{FAT}_i &\sim \text{indep. } \mathcal{N}(\nu_i, \tau^2), \\ \nu_i &= \alpha_T^* + \beta_T \log(t) + \gamma_T t, \end{aligned} \tag{3.5}$$

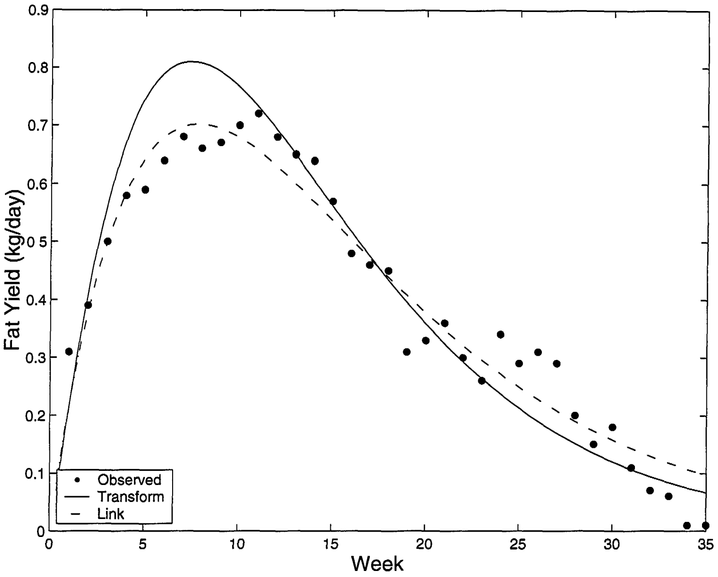

FIG. 3.4. *Fat yield data with fit to transformed data and generalized linear model fit with log link.*

where the subscript T indicates the parameters under the transformation approach.

An alternative approach is to assume homoscedasticity on the original (kg/day) scale and fit the data directly using a log link. This would correspond to a model

$$\begin{aligned} &\text{FAT}_i \sim \text{indep. } \mathcal{N}(\mu_i, \sigma^2), \\ &\log \mu_i = \alpha_L^* + \beta_L \log(t) + \gamma_L t. \end{aligned} \tag{3.6}$$

Figure 3.3 shows the data with each of these fits superimposed. Clearly, they are quite different. Why is this so? The reason for the differences is revealed by looking at the plots of the same predicted values on the log FAT scale, given in Figure 3.3. The two small fat values of 0.01 in weeks 34 and 35 are outliers on the log scale. The log fit attempts to accommodate them at the cost of the fit around week 7. On the original scale these do not influence the fit as dramatically.

A nice discussion of the proper scale on which to fit data (in the context of fitting the Michaelis–Menten equation) is given in Ruppert et al. (1989).

3.4 Potato flour revisited

I now return to the potato flour example of Section 3.1 and discuss inference in more detail. As suggested in that section, a common method of fitting these models is by maximum likelihood. That leads naturally to likelihood ratio tests and a measure of model fit, called the deviance. Roughly speaking, the deviance is the difference between twice the maximum possible log likelihood and twice the log

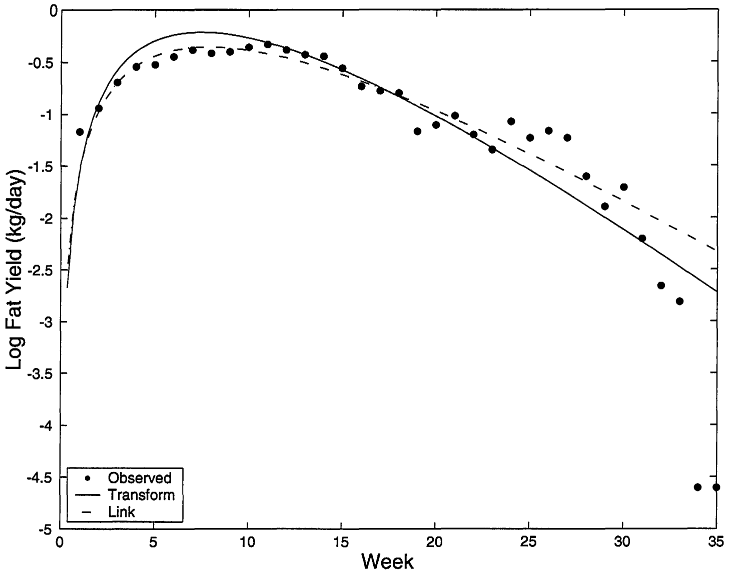

FIG. 3.5. *Fat yield data on log scale with transform and link fits.*

likelihood of the fitted model. The maximum possible log likelihood is achieved when a saturated model is fit. For the potato flour data, that corresponds to fitting a separate parameter for each level of dilution (10 parameters). Large values of the deviance indicate a model that fits poorly. In cases in which there are sufficient replicates for each pattern of covariates represented in the saturated model, the usual chi-square approximations can be used (McCullagh and Nelder, 1989) to form a formal lack-of-fit test. In other situations those approximations break down.

For the potato flour data, we consider three models: (1) the saturated model with 10 parameters, (2) the model of Section 3.1, with 2 parameters (slope and intercept) and (3) a reduced model with only an intercept (1 parameter). Respectively their log likelihood values are: $-12.597, -14.214$ and -33.203. The deviance for the slope and intercept model is $2(-12.597 + 14.214) = 3.234$ on $10 - 2 = 8$ degrees of freedom. Performing a formal chi-square test gives a p-value of 0.92, indicating that the saturated model is no better than the slope and intercept model, or, equivalently, that there is not evidence for lack of fit.

For the intercept-only model, the deviance is $2(-12.597 + 33.203) = 41.212$ on 9 degrees of freedom, which is statistically significant with a p-value of approximately 0. The deviance thus indicates that the intercept-only model is inadequate (pretty obvious from Figure 3.1!).

The deviance statistics can also be used to form likelihood ratio tests since the difference in the deviance of two models is the usual likelihood ratio statistic (the log likelihood of the saturated model cancelling out). Using the deviance statistics to compare the intercept-only and the slope and intercept models gives a likelihood ratio statistic of $41.212 - 3.234 = 37.978$ on 1 degree of freedom with a p-value of approximately zero. As expected, we reject the intercept-only model in favor of the slope and intercept model.

A somewhat different technique, called maximum quasi-likelihood can also be used to fit these models. That will be described in Chapter 7.