Chapter 1

Extending Traditional Models

We assume the reader is familiar with non-normal group models (e.g., two samples from gamma distributions), basic generalized linear models, models with additive error (both linear and nonlinear), finite mixture models, and basic mixture and simple hierarchical models. The construction of these classes of models involved several concepts that can be extended to formulate more complex models.

1.1 Parameterized Link Functions

We may formulate a generalized linear model in the usual manner except that, rather than specifying a completely known link function such as $\log(\mu)$, we specify only that a link function be contained in some family of link functions that are indexed by a parameter, λ say. The link function then becomes $g(\mu|\lambda)$ and we may wish to estimate λ along with the other parameters of the model.

1.1.1 Historical Note

To the best of my knowledge, the concept of embedding a link function into an entire family of functions was introduced by Pregibon (1980). Pregibon used this idea primarily to develop a score test for a hypothesized link function. The idea was that we will typically have a some link function we're thinking about using and would like to assess that hypothesized link against a range of alternatives. Suppose that the hypothesized link function can be embedded in a parameterized family of link functions $g(\mu|\lambda)$ that contains both the hypothesized link and what we think of as the true link. For example, suppose that the parameterized family of link functions is the power family,

$$g(\mu|\lambda) = \begin{cases} \mu^\lambda & \lambda \neq 0 \\ \log(\mu) & \lambda = 0 \end{cases}$$

and the true (but unknown) link function also belongs to this family for a particular parameter value, say λ_*. We may be contemplating using a model with a link function in the above family with a particular value of the parameter, $\lambda_0 = 1$ for example would give an identity link and $\lambda_0 = 2$ would give the canonical link for a gamma random component. We might like a test for the hypothesis that $\lambda_* = \lambda_0$, and this is the problem Pregibon addressed.

Pregibon's solution made use of a first order Taylor series for the true link function expanded about the hypothesized link function, resulting in a model that could be "fitted" for one step using the hypothesized link (meaning no new software was needed, which was more important in 1980 than it is now). Pregibon also noted that this procedure was the first step of what could become an iterative solution for maximum likelihood estimation of λ which, although true, was perhaps unfortunate, because Pregibon's procedure has been used in an improper manner under the assumption that maximum like-

lihood estimates resulted (Kaiser 1997). But, there is an easy way to do it right, which we now give.

1.1.2 Maximum Likelihood Estimation

Let $Y_1, \ldots, Y_n$ represent independent response variables from exponential dispersion family distributions with density or mass functions,

$$f(y|\theta_i, \phi) = \exp\left[a(\phi)\{y\theta_i - b(\theta_i)\} + c(y, \phi)\right], \tag{1.1}$$

where $a(\phi) = \phi\, m_i$ for a known set of weights $\{m_i : i = 1, \ldots, n\}$. We have modified our standard form from class just a bit by using this function $a(\phi)$ in place of the simple ϕ; this will be useful in considering binomial random components, in which case the m_i become the binomial sample sizes. To complete the model, let the systematic model component be written as,

$$g(\mu_i|\lambda) = \boldsymbol{x}_i^T \beta = \eta_i \tag{1.2}$$

where $g(\mu|\lambda)$ is some family of link functions specified up to an unknown parameter λ, which may be either a scalar or vector. The i^{th} contribution to the log likelihood now is, for $i = 1, \ldots, n$,

$$\ell_i = a(\phi)\{y_i\theta_i - b(\theta_i)\} + c(y_i, \phi), \tag{1.3}$$

and the complete log likelihood is $\ell = \sum_{i=1}^n \ell_i$.

Here, we suppose that $\boldsymbol{x}_i^T = (x_{1,i}, \ldots, x_{p,i})$, $\beta = (\beta_1, \ldots, \beta_p)^T$, and $\lambda = (\lambda_1, \ldots, \lambda_q)^T$. Following the same type of progression we used in Stat 520 for developing a Fisher Scoring algorithm to locate maximum likelihood estimates of the regression parameters β, we have

$$\begin{aligned} \frac{\partial \ell_i}{\partial \beta_j} &= \frac{\partial \ell_i}{\partial \theta_i}\frac{d\theta_i}{d\mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} \\ \frac{\partial \ell_i}{\partial \lambda_k} &= \frac{\partial \ell_i}{\partial \theta_i}\frac{d\theta_i}{d\mu_i}\frac{\partial \mu_i}{\partial \lambda_k}. \end{aligned} \tag{1.4}$$

Note that in (1.4) the third right hand side terms are now partial derivatives rather than the ratio of differentials that we had for basic generalized linear models with fixed link functions.

In a manner similar to what was done for basic generalized linear models, note that

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \theta_i} &= a(\phi)\{y_i - b'(\theta_i)\} = a(\phi)(y_i - \mu_i) \\
\frac{d\theta_i}{d\mu_i} &= \left(\frac{d\mu_i}{d\theta_i}\right)^{-1} = \frac{1}{b''(\theta_i)} = V^{-1}(\mu_i) \\
\frac{\partial \mu_i}{\partial \eta_i} &= \left(\frac{\partial g(\mu_i|\lambda)}{\partial \mu_i}\right)^{-1} = \left(\frac{\partial \eta_i}{\partial \mu_i}\right)^{-1} \\
\frac{\partial \eta_i}{\partial \beta_j} &= x_{i,j} \\
\frac{\partial \mu_i}{\partial \lambda_k} &= \left(\frac{-\partial g(\mu_i|\lambda)}{\partial \lambda_k}\right)\left(\frac{\partial g(\mu_i|\lambda)}{\partial \mu_i}\right)^{-1} = \left(\frac{-\partial \eta_i}{\partial \lambda_k}\right)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)^{-1}
\end{aligned} \tag{1.5}$$

Notice that in the third and fifth lines of expression (1.5) we have applied an implicit function theorem as follows. For a basic generalized linear model with fixed link it is easy to write $\mu_i = g^{-1}(\eta_i)$ where $g^{-1}(\cdot)$ is the inverse of the link function $g(\cdot)$. When $g(\cdot)$ is a simple function of one argument this is not difficult (e.g., if $g(x) = \log(x)$ then $g^{-1}(x) = \exp(x)$). But when the link function is parameterized as $g(\cdot|\lambda)$ this is often not possible. Nevertheless, it remains true that $g(\mu_i|\lambda) - \eta_i = 0$ and then implicit functions immediately give lines three and five.

From the expressions (1.4) and (1.5) we can now write the i^{th} contribution to the first derivatives as,

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \beta_j} &= \phi\, m_i(y_i - \mu_i)\, V^{-1}(\mu_i) \left(\frac{\partial \eta_i}{\partial \mu_i}\right)^{-1} x_{i,j} \\
\frac{\partial \ell_i}{\partial \lambda_k} &= \phi\, m_i(y_i - \mu_i)\, V^{-1}(\mu_i) \left(\frac{\partial \eta_i}{\partial \mu_i}\right)^{-1} \left(\frac{-\partial \eta_i}{\partial \lambda_k}\right)
\end{aligned}$$

Now define the terms

$$w_i = m_i \left[\left(\frac{\partial \eta_i}{\partial \mu_i} \right)^2 V(\mu_i) \right]^{-1},$$

which allows (1.6) to be written as,

$$\begin{aligned} \frac{\partial \ell_i}{\partial \beta_j} &= \phi \, (y_i - \mu_i) \, w_i \left(\frac{\partial \eta_i}{\partial \mu_i} \right) x_{i,j} \\ \frac{\partial \ell_i}{\partial \lambda_k} &= \phi \, (y_i - \mu_i) \, w_i \left(\frac{\partial \eta_i}{\partial \mu_i} \right) \left(\frac{-\partial \eta_i}{\partial \lambda_k} \right) \end{aligned} \tag{1.6}$$

The point here is that the contribution of individual terms to the score functions (first derivatives) have been written in the same form for derivatives with respect to the link function parameters as for the regression parameters. That is, the only difference between the first and second lines of (1.6) is the final term on the right hand side.

Now, following the same progression for second derivatives and taking expected values to simplify the expressions, which is laid out in some detail for fixed link models in the Stat 520 notes, we end up with the following expressions.

$$\begin{aligned} -E \left(\frac{\partial^2 \ell_i}{\partial \beta_j \, \partial \beta_l} \right) &= \phi \, w_i x_{i,j} \, x_{i,l} \\ -E \left(\frac{\partial^2 \ell_i}{\partial \beta_j \, \partial \lambda_k} \right) &= \phi \, w_i \left(\frac{-\partial \eta_i}{\partial \lambda_k} \right) x_{i,j} \\ -E \left(\frac{\partial^2 \ell_i}{\partial \lambda_k \, \partial \lambda_h} \right) &= \phi \, w_i \left(\frac{-\partial \eta_i}{\partial \lambda_k} \right) \left(\frac{-\partial \eta_i}{\partial \lambda_h} \right) \end{aligned} \tag{1.7}$$

Let $\xi \equiv (\beta^T, \lambda^T)^T$ be the complete $(p+q)$ vector of systematic model component parameters. Summing over the expressions in (1.6) and collecting the score functions into a vector results in the gradient,

$$\nabla L = \left(\sum_{i=1}^n \frac{\partial \ell_i}{\partial \beta_1}, \ldots, \sum_{i=1}^n \frac{\partial \ell_i}{\partial \beta_p}, \sum_{i=1}^n \frac{\partial \ell_i}{\partial \lambda_1}, \ldots, \sum_{i=1}^n \frac{\partial \ell_i}{\partial \lambda_q} \right)^T \tag{1.8}$$

The (negative) expected second derivatives may be collected into a matrix H as

$$H = \begin{pmatrix} H_1 & H_2 \\ H_2^T & H_3 \end{pmatrix}, \tag{1.9}$$

where

$$\begin{aligned} H_1 \quad &\text{is} \quad p \times p \text{ with } jl^{th} \text{ element } \sum_{i=1}^{n} \phi w_i x_{i,j}\, x_{i,l} \\ H_2 \quad &\text{is} \quad p \times q \text{ with } jk^{th} \text{ element } \sum_{i=1}^{n} \phi w_i \left(\frac{-\partial \eta_i}{\partial \lambda_k}\right) x_{i,j} \\ H_3 \quad &\text{is} \quad q \times q \text{ with } kh^{th} \text{ element } \sum_{i=1}^{n} \phi w_i \left(\frac{-\partial \eta_i}{\partial \lambda_k}\right)\left(\frac{-\partial \eta_i}{\partial \lambda_h}\right) \end{aligned} \tag{1.10}$$

A Fisher Scoring algorithm may then be defined to update a current estimate $\xi^{(m)}$ to a new estimate $\xi^{(m+1)}$ as,

$$\xi^{(m+1)} = \xi^{(m)} + \left(H^{-1}\, \nabla L\right)|_{\xi=\xi^{(m)}}. \tag{1.11}$$

This is really all we need to locate maximum likelihood estimates of the elements of ξ but, as with basic generalized linear models we can perform some further manipulations to arrive at the form of an iteratively weighted least squares algorithm as follows.

Let X_A denote a matrix formed by augmenting the usual X matrix with q additional columns having elements given by the derivatives

$$\frac{-\partial \eta_i}{\partial \lambda_k}; \quad k = 1, \ldots, q$$

Then X_A is an $n \times (p+q)$ matrix with i^{th} row

$$\boldsymbol{x}_{A,i}^T = \left(x_{1,i}, \ldots, x_{p,i}, \frac{-\partial \eta_i}{\partial \lambda_1}, \ldots, \frac{-\partial \eta_i}{\partial \lambda_q}\right)$$

Let W be an $n \times n$ diagonal matrix with elements w_i as given immediately prior to expression (1.6), and let $\boldsymbol{z}^T = (z_1, \ldots, z_n)$ where

$$z_i = (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}.$$

Then we have that

$$\nabla L = \phi X_A^T W \boldsymbol{z} \quad \text{and} \quad H = \phi X_A^T W X_A, \tag{1.12}$$

and the Fisher Scoring algorithm of expression (1.11) becomes

$$\begin{aligned} \xi^{(m+1)} &= \xi^{(m)} + \left[\left(X_A^T W X_A \right)^{-1} X_A^T W \boldsymbol{z} \right] |_{\xi=\xi^{(m)}} \\ &= \left[\left(X_A^T W X_A \right)^{-1} \left(X_A^T W X_A \xi + X_A^T W \boldsymbol{z} \right) \right] |_{\xi=\xi^{(m)}} \\ &= \left[\left(X_A^T W X_A \right)^{-1} X_A^T W \boldsymbol{z}^* \right] |_{\xi=\xi^{(m)}}, \end{aligned} \tag{1.13}$$

where, $\boldsymbol{z}^* = (z_1^*, \ldots, z_n^*)^T$ with

$$z_i^* = \boldsymbol{x}_{A,i}^T \xi + z_i$$

The last line of (1.13) is in the form of an iteratively re-weighted least squares algorithm. Note here that the matrix X_A does not remain fixed in this algorithm as it needs to be evaluated at the current estimate $\xi^{(m)}$ at each iteration.

Useful Families of Link Functions

If we would like to estimate parameters of link functions we need useful families of such functions, and developing these is not a trivial task. Consider, for example, the power family given earlier,

$$g(\mu|\lambda) = \begin{cases} \mu^\lambda & \lambda \neq 0 \\ \log(\mu) & \lambda = 0 \end{cases} \tag{1.14}$$

This is a fine way to write a family of functions if our use is to select a power for a fixed link, but it is not so useful for estimation (of λ) if our desire is to separate a log link from some other power. For this, we might consider a family of link functions given by Pregibon (1980) as,

$$g(\mu|\lambda) = \frac{1}{\lambda_2} \left[(\mu + \lambda_1)^{\lambda_2} - 1 \right]. \tag{1.15}$$

This family includes, for example, the identity link which results from taking $\lambda_1 = \lambda_2 = 1$. It also includes the log link if we take $\lambda_1 = 0$ and let $\lambda_2 \to 0$,

$$\lim_{\lambda_2 \to 0} g(\mu|\lambda_1 = 0, \lambda_2) = \lim_{\lambda_2 \to 0} \frac{\mu^{\lambda_2} - 1}{\lambda_2} = \log(\mu).$$

Note that the difference with the power family (1.14) is that in (1.14) the link was *defined* as log for $\lambda = 0$ while in (1.15) we get the log link as the value of λ_2 goes to zero. This makes a difference if our objective is to estimate λ. What about other powers? Consider using $\lambda_1 = 0$ and $\lambda_2 = 2$ in (1.15). This gives

$$g(\mu|\lambda) = \frac{\mu^2 - 1}{2}.$$

Now, suppose that the linear predictor is $\eta_i = \beta_0 + \beta_1 x_i$. Then we would have the systematic model component,

$$g(\mu_i|\lambda) = \beta_0 + \beta_1 x_i \Rightarrow \frac{\mu^2 - 1}{2} = \beta_0 + \beta_1 x_i \Rightarrow \mu_i^2 = (2\,\beta_0 + 1) + 2\,\beta_1 x_i$$

or,

$$\mu_i^2 = \gamma_0 + \gamma_1 x_i$$

so that our model is one with an ordinary squared link function.

Another useful family of link functions that we will use in analysis of a short-term toxicity test is,

$$g(\mu|\lambda) = \log\left[\frac{(1-\mu)^{-\lambda} - 1}{\lambda}\right].$$

This family includes the logit link for $\lambda = 1$,

$$g(\mu|\lambda = 1) = \log\left[\frac{1}{(1-\mu)} - 1\right] = \log\left(\frac{\mu}{1-\mu}\right),$$

and the complementary log-log link as $\lambda \to 0$,

$$\begin{aligned}
\lim_{\lambda\to 0} g(\mu|\lambda) &= \lim_{\lambda\to 0} \log\left[\frac{(1-\mu)^{-\lambda}-1}{\lambda}\right] \\
&= \log\left[\lim_{\lambda\to 0}\frac{1}{\lambda}\left\{\frac{1}{(1-\mu)^{\lambda}}-1\right\}\right] \\
&= \log\left[\lim_{\lambda\to 0}\frac{1-(1-\mu)^{\lambda}}{\lambda(1-\mu)^{\lambda}}\right] \\
&= \log\left[\lim_{\lambda\to 0}\frac{-\log(1-\mu)}{1+\lambda\log(1-\mu)}\right] \\
&= \log\left[-\log(1-\mu)\right].
\end{aligned}$$

1.1.3 Analysis of Short-Term Toxicity Test Data

The design of a short-term toxicity test is quite simple. We have k concentrations of some potentially toxic substance and we expose groups of $n_1, n_2, \ldots, n_k$ organisms to these concentrations for a fixed period of time. At the end of that time the number of organisms in each group that have "responded" (usually died) is recorded. Such data are often called quantal response data.

The theoretical basis for the analysis of quantal response data is somewhat more complex than the experimental design. The fundamental elements of this theory are as follows.

1. It is supposed that for each individual organism there is a concentration of the toxicant, x say, such that the organism will respond for any con-

centration greater than or equal to x and the organism will not respond for any concentration less than x. This value is called the *tolerance* of the organism. That is, if R_j denotes the response of organism j, x_j its tolerance, and d the concentration to which it is exposed,

$$Pr(R_j = 1|d < x_j) = 0 \qquad Pr(R_j = 1|d \geq x_j) = 1$$

2. It is also supposed that, in the population of organisms, the tolerances x_j follow some distribution that is a location-scale family with distribution function G, mean μ_x and variance σ_x^2. Then the standardized tolerances are such that,

$$\tilde{x}_j = \frac{x_j - \mu_x}{\sigma_x} \sim iidG(0, 1)$$

3. Assume that at a given experimental concentrations (dose) d_i; $i = 1, \ldots, k$, there is a certain probability p_i; $i = 1, \ldots, k$ that the dose will exceed the tolerance of a randomly chosen organism so that, if Y_i is defined as the number of responses out of n_i organisms at dose i, the probability mass function of Y_i is,

$$f(y_i|p_i) = \frac{n_i!}{y_i!(n_i - y_i)!}\, p_i^{y_i}\, (1 - p_i)^{n_i - y_i}; \quad y_i = 0, 1, \ldots, n_i \qquad (1.16)$$

4. From the above we have that, at a given dose d_i,

$$p_i = \int_{-\infty}^{\delta_i} d\, G, \quad \text{where} \quad \delta_i = \frac{d_i - \mu_x}{\sigma_x}. \qquad (1.17)$$

That is, δ_i is the $p(100)\%-$tile of $G(0, 1)$, the distribution of standardized tolerances.

5. The objective is, given fixed $d_1, \ldots, d_k$, fixed $n_1, \ldots, n_k$ and observed $y_1, \ldots, y_k$, estimate μ_x and σ_x^2, the parameters of the tolerance distribution $G(\mu_x, \sigma_x^2)$. In particular, if the response is mortality and G is

assumed or chosen to be symmetric, μ_x is often called the "median effective dose" (MED), or the "lethal concentration that kills 50%" (LC_{50}).

Formulation as a Generalized Linear Model

To formulate this problem as a standard glm, take the response variables to be expressed as observed proportions rather than the counts of expression (1.16). The mass functions of these Y_i may be written in exponential dispersion family form as,

$$f(y_i|\theta_i) = \exp[a(\phi)\{y_i\theta_i - b(\theta_i)\} + c(y_i)], \tag{1.18}$$

where

$$\theta_i = \log\left(\frac{p_i}{1-p_i}\right); \quad b(\theta_i) = \log\{1+\exp(\theta_i)\}; \quad \text{and} \quad \phi \equiv 1; \;\; a(\phi) = n_i$$

Now, from (1.17) $G(\delta_i) = p_i$ so then,

$$G^{-1}(p_i) = \delta_i = \frac{d_i - \mu_x}{\sigma_x} = \frac{-\mu_x}{\sigma_x} + \frac{1}{\sigma_x}d_i, \tag{1.19}$$

which completes a standard generalized linear model with binomial random component, link function G^{-1}, and regression parameters $\beta_0 = -\mu_x/\sigma_x$ and $\beta_1 = 1/\sigma_x$.

Notice from this development that there is a one-to-one relation between distinct tolerance distributions $G(\mu_x, \sigma_x^2)$ and link functions. In particular some of the typical links and tolerance distributions are:

- Normal tolerance distribution, probit link
- Logistic tolerance distribution, logit link

- Extreme value tolerance distribution, complementary log-log link

Now, we would like to fit a model with random component (1.18), the linear predictor (1.19) and using the family of link functions,

$$g(\mu_i|\lambda) = \log\left[\frac{(1-\mu)^{-\lambda}-1}{\lambda}\right]. \tag{1.20}$$

Notice here that we are using μ_i for the expected value of Y_i. We have also used μ_x for the expected value of the tolerance distribution. It will be important to maintain this distinction in what is to come. Continuing to use $d_i : \ i = 1, \ldots, k$ as the experimental doses (i.e., the covariates in the glm), our systematic model component is,

$$g(\mu_i|\lambda) = \eta_i = \beta_0 + \beta_1 d_i; \quad i = 1, \ldots, k$$

The maximum likelihood algorithm presented earlier can be used to find estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. To do this, however, requires computing a number of quantities in a different manner than would be the case for a model with fixed link. I will not present derivations here, but will list the quantities that would need to be calculated in order to implement the algorithm given in expression (1.13).

1. Linear Predictor
$$\eta_i = \beta_0 + \beta_1 d_i$$
2. Means (of Y_i)
$$\mu_i = 1 - \frac{1}{\{1+\lambda\exp(\eta_i)\}^{1/\lambda}}$$
3. Derivative of η_i wrt μ_i
$$\frac{\partial \eta_i}{\partial \mu_i} = \frac{\lambda}{(1-\mu_i)\{1-(1-\mu_i)^{\lambda}\}}$$

4. Weights

$$w_i = n_i \left[\left(\frac{\partial \eta_i}{\partial \mu_i} \right)^2 V(\mu_i) \right]^{-1}$$

where $V(\mu_i) = \mu_i(1 - \mu_i)$ as for any model with random component (16)

5. Derivative of η_i wrt λ

$$\frac{\partial \eta_i}{\partial \lambda} = \frac{-\log(1 - \mu_i)}{1 - (1 - \mu_i)^\lambda} - \frac{1}{\lambda}$$

Estimation of Tolerance Distributions

With maximum likelihood estimates of β_0 and β_1 in hand we have, from (1.19) and the invariance property of maximum likelihood, that maximum likelihood estimates of the tolerance function parameters are,

$$\hat{\mu}_x = \frac{-\hat{\beta}_0}{\hat{\beta}_1}; \qquad \hat{\sigma}_x = \frac{1}{\hat{\beta}_1} \tag{1.21}$$

Now, under the theory developed, link function is the inverse distribution function for standardized tolerances (denoted as $\tilde{x}_j$ previously). The distribution function and density function of the tolerances x_j may then be found by inverting expression (1.20) as follows. We wish to find $G(\tilde{x}) = p$ for some $0 < p < 1$ given that

$$\tilde{x} = G^{-1}(p|\lambda) = \log \left[\frac{1}{(1 - p)^\lambda} - 1 \right] - \log(\lambda)$$

Then,

$$\begin{aligned} & \frac{1-(1-p)^{\lambda}}{(1-p)^{\lambda}\,\lambda} = \exp(\tilde{x}) \\ \Rightarrow \quad & \exp(\tilde{x})\,\lambda(1-p)^{\lambda} = 1-(1-p)^{\lambda} \\ \Rightarrow \quad & (1-p)^{\lambda}\,\{\lambda\,\exp(\tilde{x})+1\} = 1 \\ \Rightarrow \quad & p = 1-\frac{1}{\{\lambda\,\exp(\tilde{x})+1\}^{1/\lambda}}. \end{aligned}$$

So the distribution function for standardized tolerances is then,

$$G(\tilde{x}|\lambda) = 1-\frac{1}{\{\lambda\,\exp(\tilde{x})+1\}^{1/\lambda}}. \tag{1.22}$$

Since tolerance x_j corresponds to standardized tolerance $\tilde{x}_j$ as $\tilde{x}_j = (x_j - \mu_x)/\sigma_x$, to obtain the actual distribution function of tolerances, we simply make the appropriate location and scale transformations as

$$x_j = \sigma_x\,\tilde{x} + \mu_x$$

and then,

$$G(x|\lambda,\mu_x,\sigma_x) = \left[1-\frac{1}{\{\lambda\,\exp\{(x-\mu_x)/\sigma_x\}+1\}^{1/\lambda}}\right] \tag{1.23}$$

which has density

$$g_t(x|\lambda,\mu_x,\sigma_x) = \frac{\exp\{(x-\mu_x)/\sigma_x\}}{\sigma_x\,[\lambda\exp\{(x-\mu_x)/\sigma_x\}+1]^{(1+1/\lambda)}}. \tag{1.24}$$

An estimated tolerance density then results from substitution of the estimates $\hat{\mu}_x$ and $\hat{\sigma}_x$ from (1.21) into (1.24).

Concentration (log)	No. Mortalities	No. Exposed
3.893	6	59
3.970	13	60
4.042	18	62
4.108	28	56
4.171	52	63
4.230	53	59
4.285	61	62
4.338	60	60

Table 1.1: Bliss Beetle Data

Example 1.1 – Bliss Beetle Data

One version of the famous Bliss beetle data is presented in Table 1.1. These data have been presented in a number of forms since there were two replicates that are sometimes combined and sometimes not, and concentration is reported on various scales. The data arose from a short-term toxicity test conducted with flour beetles and gaseous carbon disulphide exposure for 5 hours.

Estimates of the parameters in the systematic model component were located using a model with a fixed logit link, and a model using the family of links in expression (1.20). The results are given Table 1.2. Estimates of μ_x and σ_x were obtained from expression (1.21).

Because the model with a fixed logit link function is nested within the model having an estimated link function (by taking $\lambda = 1$) we may conduct a likelihood ratio test to compare a reduced model (logit link) with a full model (estimated link). The maximized log likelihood (sans constant terms) of the

	Estimates Under Model With	
Parameter	Logit Link	Link Family
β_0	-60.640	-39.352
β_1	14.865	9.518
λ	NA	-0.006
μ_x	4.079	4.134
σ_x	0.067	0.105

Table 1.2: Point Estimates for Bliss Beetle Data

logit link model was -186.1993, while that for the model with estimated link was -182.3464. This results in a likelihood ratio test statistic $T = -2(-186.1003 + 182.3464) = 7.7058$ and an associated p-value of $p = 0.0055$ (from comparison with a χ^2 distribution having 1 degree of freedom). Thus, we would prefer the model with an estimated link in this example. A plot of the observed responses and fitted curves from both the model with a logit link and the model with an estimated link is presented in the upper panel of Figure 1.1. A plot of the estimated tolerance density functions is presented in the lower panel of Figure 1.1, in which one can see the difference between the symmetric tolerance density dictated by the logit link (dashed curve) and the asymmetric density that results from the estimated link function.

Example 1.2 – Sub-lethal Exposure of Trout to Petroleum Hydrocarbons

We will apply the same methods used with the Bliss beetle data to another example, this involving the effect of sub-lethal exposure of Rainbow Trout to a petroleum hydrocarbon on response to lethal concentrations of the same

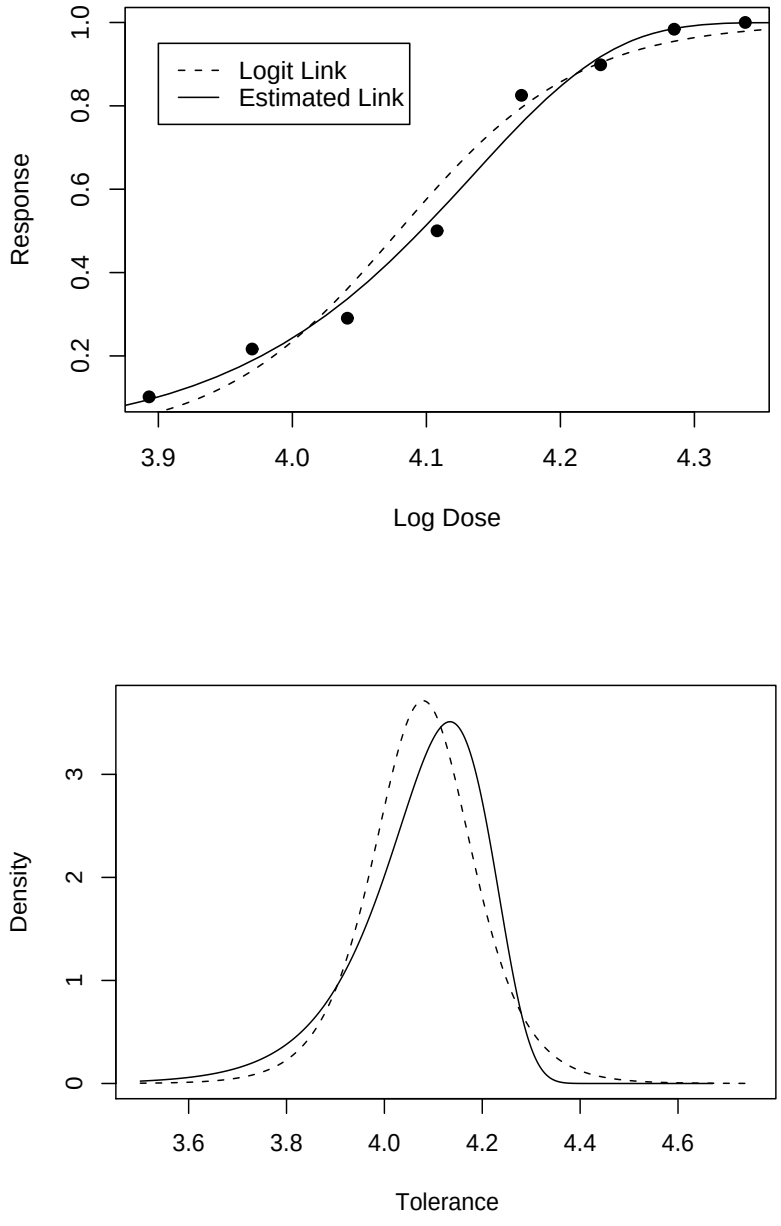

Figure 1.1: Estimated response curves (upper panel) and tolerance distributions (lower panel) for the Bliss beetle data.

substance. The scientific underpinnings of this study involved the fact that petroleum hydrocarbons (e.g., gas and fuel oils) are frequently spilled into aquatic environments. Once released into the environment, petroleum hydrocarbons bind to sediments and are slowly released back into the water column. Thus, in areas with natural hydrocarbon seeps (near natural deposits), around processing facilities (e.g., offshore oil platforms), or in areas that have been previously polluted, petroleum hydrocarbons are found at the level of μg/L, which are generally not fatal to fish. On the other hand, "major spills" do also occur with resulting concentrations in the range of mg/L (much higher). The question being investigated was whether a low level of "pre-exposure" makes Rainbow Trout more or less susceptible to higher levels of exposure, should they occur. There is some controversy about whether a low level of chronic exposure makes organisms *more resistant* or *less resistant* to particular toxicants.

In the study, test organisms (immature rainbow trout) were exposed to low levels of number 2 fuel oil (2FO) for 21 days. The concentrations used in this "pre-exposure period included 0, 25, and 50 mg/L of 2FO. After 21 days, fish were transferred to other tanks and portions of each pre-exposure group were exposed to 2FO at concentrations of 28.7, 57.4, 114.8, 229.6, 459.1 and 918.2 mg/L for a total of 335 hours. In the actual study, mortality was recorded every 4 or 8 hours and time to death was one of the response variables examined. Here, we will simply consider the data at 258 hours within the context of a typical dose-response analysis. The overall objective is to determine whether the pre-exposure groups differ in their response during the main toxicity test (at least at the time point of 258 hours). The data are presented in Table 1.3. We will analyze the data in the same way as for the Bliss beetle data, with a generalized linear model having binomial

	Pre-exposure concentration					
	0 μg/L		25 μg/L		50 μg/L	
Exposure (mg/L)	Y	N	Y	N	Y	N
28.7	0	10	0	12	1	9
57.4	0	10	1	10	0	10
114.8	2	10	1	10	2	10
229.6	4	10	4	10	5	10
459.1	8	9	8	10	7	10
918.2	10	10	9	9	10	10

Table 1.3: Mortalities (Y) and number exposed (N) in a toxicity test with 2FO.

random component and either logit or estimated link functions. Note that all of the results presented will come from models using log concentration as the covariate.

Parameter estimates for the control treatment (e.g., 0 μg/L of pre-exposure) are presented in Table 1.4 for models with fixed logit and estimated link functions. Parameter estimates for the pre-exposure treatment group of 25 μg/L are presented in Table 1.5 for models with fixed logit and estimated link functions. Parameter estimates for the pre-exposure treatment group of 50 μg/L are presented in Table 1.6 for models with fixed logit and estimated link functions.

Because a model with a fixed logit link is nested within a model with link family (1.20) we may conduct likelihood ratio tests between logit link and estimated link models for each pre-exposure group. Maximized log likelihoods and these likelihood ratio tests are presented in Table 1.7

	Model Link			
	Logit		Estimated	
Parameter	Estimate	Std. Error	Estimate	Std. Error
β_0	-15.820	4.172	-12.164	4.662
β_1	2.905	0.765	2.140	0.907
λ	NA	NA	0.129	0.811

Table 1.4: Parameter estimates for pre-exposure group 0 μg/L.

	Model Link			
	Logit		Estimated	
Parameter	Estimate	Std. Error	Estimate	Std. Error
β_0	-12.896	3.194	-9.422	2.943
β_1	2.338	0.580	1.591	0.553
λ	NA	NA	-0.199	0.567

Table 1.5: Parameter estimates for pre-exposure group 25 μg/L.

	Model Link			
	Logit		Estimated	
Parameter	Estimate	Std. Error	Estimate	Std. Error
β_0	-9.897	2.409	-6.580	2.132
β_1	1.816	0.442	1.069	0.404
λ	NA	NA	-0.482	0.386

Table 1.6: Parameter estimates for pre-exposure group 50 μg/L.

	Maximized Likelihoods			
Pre-exposure	Logit	Estimated	T	p-value
0 μg/L	-15.7426	-15.5222	1.0412	0.3075
25 μg/L	-19.3685	-18.7538	1.229	0.2675
50 μg/L	-23.7589	-22.7459	2.026	0.1546

Table 1.7: Likelihoods and LRT tests for logit versus estimated link models.

Pre-exposure	$\hat{\mu}_x$	95% Interval
0 μg/L	5.685	(5.049, 6.319)
25 μg/L	5.923	(5.304, 6.543)
50 μg/L	6.154	(5.303, 7.004)

Table 1.8: Estimates of tolerance distribution means.

Based on these results we would conclude that there is not sufficient evidence in the data to say that the tolerance distributions in any of the pre-exposure groups differ from a logistic. There is, however, an interesting pattern that suggests itself. The estimated values of the link function parameter λ appear to be decreasing as one moves from 0 to 25 to 50 μg/L pre-exposure (Tables 1.4, 1.5, 1.6). Concomitantly, the $p-$values for likelihood ratio tests are becoming smaller as well (Table 1.7). This might peek our curiosity as to whether anything is being "suggested" by the data in terms of a systematic pattern in the tolerance distributions.

Using results from the models with estimated link functions, values for μ_x and σ_x were arrived at through the use of expression (1.21) and their standard errors were computed using the delta method in the usual manner. Point and 95% interval estimates of μ_x are given in Table 1.8

These intervals certainly overlap to a great extent, and the same is true for intervals computed under the model with a fixed logit link (not shown). Thus, we are led to the belief that the data do not provide sufficient support for claiming any difference at all between the pre-exposure groups. There is insufficient evidence in the data that pre-exposure of trout to sublethal levels of 2FO changes the response to lethal concentrations at all. What happens if we plot the estimated tolerance distributions for the pre-exposure groups? The estimated densities are presented in Figure 1.2

This figure does suggest a systematic change in the tolerance distributions as the level of pre-exposure to 2FO increases from 0 to 25 to 50 μg/L, but is that suggestion one of sensitization (i.e., becoming more susceptible) or acclimation (i.e., becoming less susceptible) to the toxicant. Our eye is drawn to the left tails of these estimated densities, which are apparently becoming heavier as the level of pre-exposure increases. But note also the upper portions of the densities which also contain more and more probability as pre-exposure increases. Recall that examination of cumulative densities can often aid in interpretation. The cumulative densities corresponding to the distributions of Figure 3 are presented in Figure 1.3.

These estimated cumulative densities suggest an effect of acclimation since the rate at which probability (of mortality) is accumulating in these distributions is slower for pre-exposure of 50 μg/L than it is for pre-exposure of 25 μg/L which is in turn slower than for no pre-exposure to 2FO. For example, the cumulative probability in these distributions at a log concentration of 6.0 is 0.862 for the 0 μg/L group, 0.642 for the 25 μg/L group, and 0.432 for the 50 μg/L group.

The overall suggestion from fitting these models is that the effect of pre-exposure, if there in fact is one, is an acclimation effect. Recall that we are

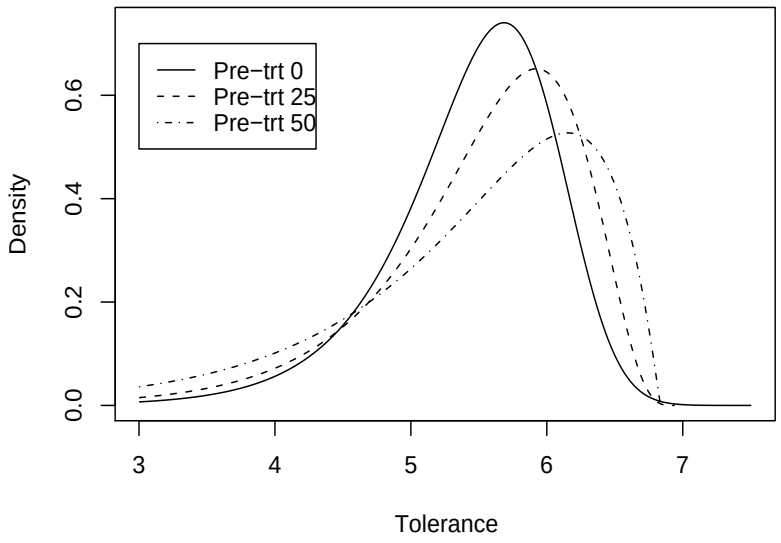

Figure 1.2: Estimated tolerance densities for Pre-exposure groups.

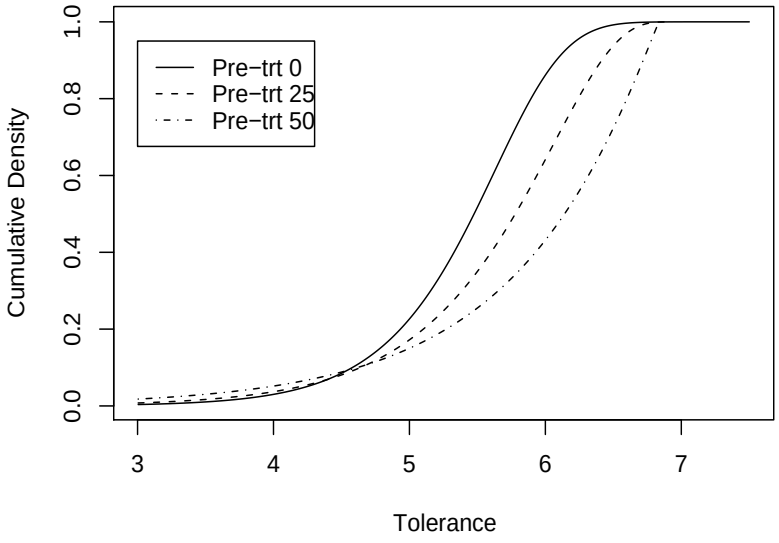

Figure 1.3: Estimated cumulative tolerance densities for Pre-exposure groups.

not able to conclude there is sufficient evidence in the data to proclaim any effect at all. But consider whether a conclusion that there is, in fact, no effect of pre-exposure is warranted. The study contained small sample sizes and we might also suggest that in order to detect this type of difference, if indeed one exists, one needs increased information. The point is that, although we cannot conclude from this study that there is any effect of pre-exposure, we are able to suggest (1) if such an effect is real it is likely to be acclimation rather than sensitization, and (2) in order to detect such differences a larger sample size is needed.

1.2 Lessons from Generalized Linear Models

While the unified algorithm for estimation of basic generalized linear models is nice, of greater value are the lessons for model construction that result from the development of this class of models. In particular, recall that writing distributions in the form of exponential dispersion families allowed us to isolate a parameter that governs expected values. That parameter, or a simple function of it, was allowed to vary across a set of random variables while the other parameter, if present, was held constant. Along the way, each distribution dictated a certain relation between means and variances, which assisted in selection of the model random component. There are distributions that cannot be put into the form of exponential dispersion families but to which we can apply these same ideas.

1.2.1 Example 1.3 – Storage Time of Meat

The production of what is called "case-ready" meat by processing plants has become quite popular, especially in grocery stores and supermarkets that are part of large national or regional chains. Such case-ready meats are delivered to the store in pre-packaged containers and reduce or remove the need for the store to have a butcher on staff. A major concern with this process is shelf-life of the packaged meats, particularly with respect to red meats such as beef. The aging process of beef results in color changing from pink to brown, and this occurs prior to spoilage. And, color is one of the primary characteristics used by consumers in choosing whether or not to buy beef. As a result, many grocers end up discarding large amounts of discolored beef, even when that beef is perfectly safe and healthy to eat.

It has been discovered that a packaging process that uses vacuum packaging with a tiny amount of carbon monoxide inserted into the package helps retard discoloration of beef, making the product look fresher for longer, and thus extending the effective shelf-life of case-ready beef products. This practice has, of course, become quite controversial, with any number of consumer advocate groups claiming it amounts to false advertising and is unsafe (even if a product is safe when purchased, if it looks fresher than it is, a consumer might store it for longer than they otherwise would before eating it, increasing the risk that it is spoiled when consumed). In March 2006, the City of Chicago considered a ban on carbon monoxide packaging (I don't know what they decided) and in mid-2006 the FDA was petitioned by a number of groups to ban the practice nationwide (I think that's still pending).

In a major study to examine the effect of carbon monoxide packaging on the appeal of beef steak to consumers, a food science department conducted

the following study. Over a period of 25 days, fresh beef was obtained every 12 hours from a packing plant. Two steaks were chosen at each point in time, chosen to be the same "grade" by meat experts; grade involves the amount of fat marbling, quality of trim to remove excess fat around the edges, and so forth. One steak was packaged in the traditional manner while the other was packaged using the carbon monoxide process. The total of 100 steaks were then judged in terms of how "desirable" each was on a scale of 1 to 10 by a large panel of 2000 consumers. The data reported are in the form of a proportional score; the total score for each steak was the total score for that steak divided by the score of the "most desirable" steak (the steak receiving the highest total score).

The objective in a statistical analysis of these data is to relate proportional score for desirability to storage time, and to determine if there is a systematic difference between traditional packaging and packaging with carbon monoxide. The relation at 14 days is of particular interest, because 14 days is typically taken as the "effective" shelf-life for steak packaged with the traditional method (i.e., after 14 days, traditionally packaged steak is often marked as "reduced for quick sale" or discarded by grocers).

The example just described involves a situation in which response variables might be taken as independent, the objective being to formulate a regression of responses on covariates (time in the example) and in which the possible values of response variables would most correctly be taken as the unit interval for each response. One could certainly take a transformation of the responses to make a model with additive errors more palatable (e.g., the so-called angular transformation) but doing so would certainly complicate making inference on the original scale. And in the example we already know that desirability score is going to decrease with time, so finding that there is

an inverse relation between the freshness of meat and time in package is not going to earn a paycheck. The natural random component to choose in the example would be to take the responses as having beta distributions but this places us solidly outside the realm of basic generalized linear models because the beta density cannot be coerced into the form of an exponential dispersion family.

1.2.2 A Model with Beta Response Distributions

To determine whether we can formulate a useful regression model with beta response distributions we can begin by attempting to achieve the same ends that result from exponential dispersion family random components. These are the isolation of expected values that can vary over observations, another parameter that can be assumed constant over observations, and the identification of a relation between expected values and variances. A first step is to determine a mean value parameterization of a beta density. A standard form for a beta density is, for $\alpha > 0$ and $\beta > 0$,

$$f(y|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}; \quad 0 < y < 1. \tag{1.25}$$

For a random variable Y that follows this density, the expected value is

$$\mu = E(Y) = \frac{\alpha}{\alpha+\beta}.$$

In the parameterization of (1.25) the variance becomes

$$var(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2\,(\alpha+\beta+1)} = \mu(1-\mu)\frac{1}{\alpha+\beta+1}.$$

So now let $\phi = 1/(\alpha+\beta+1)$.

To formulate a regression then let $Y_1, \ldots, Y_n$ be random variables with beta distributions, expected values $E(Y_i) = \mu_i$ and constant dispersion parameter ϕ. The variances of these variables are $var(Y_i) = \phi\,\mu_i\,(1-\mu_i)$. The

model would be completed by modeling the μ_i in terms of covariates $\boldsymbol{x}_i$ and parameters $\boldsymbol{\gamma}$, say

$$\mu_i = h(\boldsymbol{x}_i, \boldsymbol{\gamma}) \tag{1.26}$$

If we wanted to mimic generalized linear models we might take $h(\cdot)$ to correspond to a simple link function such as $h(\boldsymbol{x}_i, \boldsymbol{\beta}) = \exp(\gamma_0 + \gamma_1 x_i)/[1 + \exp(\gamma_0 + \gamma_1 x_i)]$ but there is really no need to do so.

A question about this model is how much the assumption that ϕ is constant over observations restricts distributional shapes across levels of the covariate. Note that both α and β in (1.25) can vary across covariate values, but only in a manner such that ϕ remains constant.

Estimation and Inference

Estimation and inference can proceed according to a likelihood or Bayesian approach. In either case, it is probably more convenient to write the likelihood in terms of α_i and β_i and use the relations

$$\begin{aligned} \alpha_i &= \left(\frac{1}{\phi} - 1\right)\mu_i \\ \beta_i &= \left(\frac{1}{\phi} - 1\right)(1 - \mu_i) \end{aligned} \tag{1.27}$$

to translate between (μ_i, ϕ) and (α_i, β_i). The expected values $\{\mu_i : i = 1, \ldots, n\}$ are of course written as functions of the elements of $\boldsymbol{\gamma}$ as in (1.26) so the focus of estimation is actually $\boldsymbol{\gamma}$ and ϕ, and derivatives for a likelihood analysis are most easily derived using the chain rule. The parameter space of ϕ is the unit interval $\phi \in (0, 1)$ so some type of beta prior is natural to use for this parameter in a Bayesian analysis. Assuming that the systematic model component given by $h(\cdot)$ in (1.26) allows elements of $\boldsymbol{\gamma}$ to assume values on

the entire line, diffuse normal priors would be a naive choice for nearly any particular model.

1.3 Focusing on Random Model Components

A device we relied on in selecting random components in basic generalized linear models, and powers for power of the mean models in additive error regressions, was to relate response variances to expected values. In formulating models that are not members of these classes, mean-variance relations can still be valuable, although what they tell us about a problem may be different than what we are familiar with.

1.3.1 Example 1.4 – Soil Respiration and Temperature

In this era of climate change there is great interest in "sources and sinks" of carbon in the environment due to the role of carbon compounds in global warming. For example, the Amazon rain forest is estimated to have somewhere around 100 billion tons (or 75 billion tonnes) of carbon stored in trees. Trees do respire, and this releases carbon on a regular basis, as well as isoprene, a catalyst in the production of ozone (this is believed to be the reason for the "gloriously stupid concept" espoused by President Ronald Reagan in 1981 that trees 'cause more pollution than automobiles do')

`https://rationalwiki.org/wiki/Trees_cause_pollution.`

Certainly, when the forest is cleared (often by burning) huge amounts of carbon dioxide are released into the atmosphere, so destruction of forest causes substantial pollution and reduces the capacity of the environment to capture additional carbon. Although this role for the Amazon forest is impressive, it

pales in comparison with the amount of carbon in soils, estimated at about 2,500 gigatons. Soil, too, respires. More correctly, roots, bacteria, fungi, and subterranean animals respire. So soils also contribute to the carbon dioxide load of the atmosphere, as well as being a major carbon sink. It is generally accepted that soil respiration increases with soil temperature. This has apparently caused some to wonder if there might not be a "positive feed-back loop" in which a warming climate leads to greater soil respiration which, in turn, leads to more carbon being released to the atmosphere and, therefore, more warming. In a study that investigated the relation between soil respiration, soil temperature, soil moisture, and other factors, Raich *et al.* (2021) used observations from four locations that included temperate grasslands, northern forests, and tropical forest plantations. A scatterplot of soil respiration versus soil temperature from one of those locations is reproduced in Figure 1.4. The variable Rsoil is soil respiration and has units of grams of carbon per square meter per day. Soil temperature is measured in degrees celcius at a depth of 5 cm.

It is possible, but quite rare, for soil respiration to be negative at temperatures of less than 0 degrees C. There are a few such values evident in the scatterplot of Figure 1.4 and it would not be unreasonable to delete these values and assume that response variables are strictly positive. Even if we would use response distributions that accommodate negative values, we would want the left tail of those distributions to be 'short' relative to the right tail. The scatterplot also hints at right skew response distributions for larger values of the covariate of soil temperature. Taking all of this into consideration, we might decide that a regression having gamma random component would be a reasonable place to start in developing a model for these data after dropping negative responses. Figure 1.4 also suggests that a log link might not be a

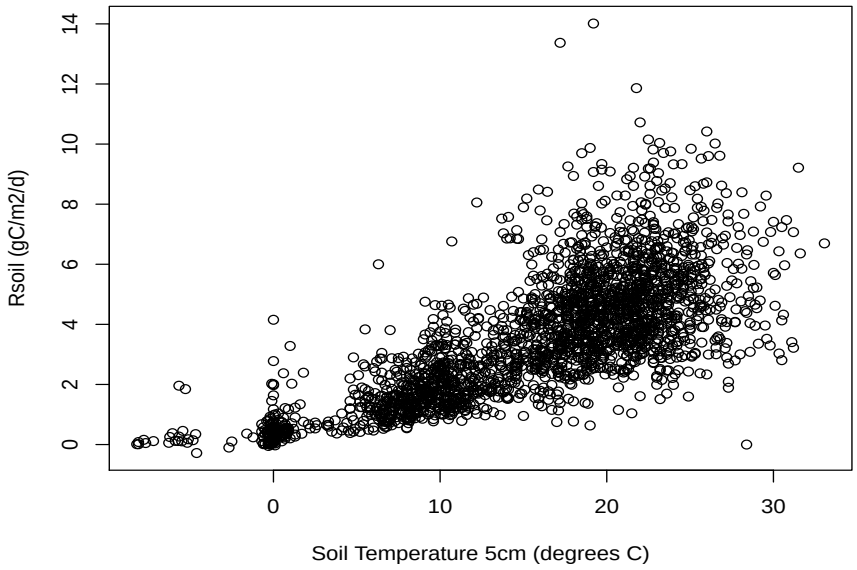

Figure 1.4: A scatterplot of Rsoil against Tsoil.

bad choice (the leftmost "tail" in the scatterplot is what argues against an identity link). So, as a fist try to model these data we might fit a basic generalized linear model with log link and gamma random component. If we do so, we find that the deviance residual plot in Figure 1.5 indicates that the model implies that response variances are given as too large a power of the expected values. That is, the data do not support a model in which $var(Y) \propto \mu^2$, as implied by the basic glm with gamma random component. Box-Cox plots for these data (not shown) suggest that there is a relation between variances and expected values in the form of $var(Y) \propto \mu^d$, but indicate that d should be chosen somewhere around 1, not 2. We are now in a situation for which we would like to (at least initially) maintain a gamma random component, but would also like to model variances as proportional to the expected values.

1.3.2 A Gamma Model

A basic glm with gamma random component is no longer an attractive possibility. In the development of a basic glm with a gamma random component, expected values and variances are related as follows. Begin with potentially different gamma parameters, α_i and β_i for each response random variable Y_i; $i = 1, \ldots, n$. To obtain variances as proportional to some power of the expected values we need,

$$\frac{\alpha_i}{\beta_i^2} \propto \left(\frac{\alpha_i}{\beta_i}\right)^c.$$

A basic glm takes the factor of proportionality to be $1/\alpha_i$ and then makes this constant across values of Y_i, resulting in,

$$\frac{\alpha}{\beta_i^2} = \frac{1}{\alpha}\left(\frac{\alpha}{\beta_i}\right)^2.$$

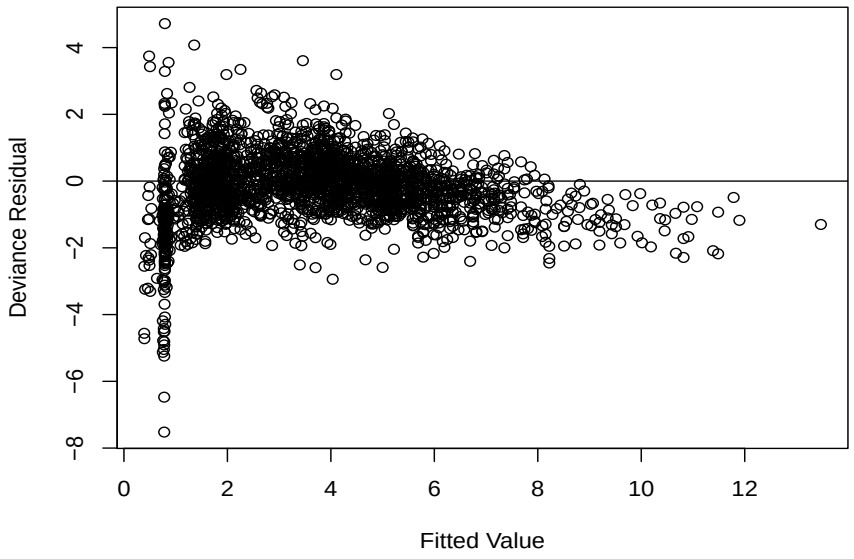

Figure 1.5: Deviance residuals for a standard glm with gamma random component and log link.

But if we would take the proportionality factor to be $1/\beta_i$, and then make this constant across values of Y_i we would arrive at,

$$\frac{\alpha_i}{\beta^2} = \frac{1}{\beta}\left(\frac{\alpha_i}{\beta}\right).$$

We could then formulate a model with gamma response distributions such that

$$\begin{aligned} E(Y_i) &= \mu_i = \frac{\alpha_i}{\beta} \\ var(Y_i) &= \phi V(\mu_i) = \frac{1}{\beta}\mu_i \\ \log(\mu_i) &= \gamma_0 + \gamma_1 x_i \end{aligned}$$

Note that this is no longer a basic generalized linear model, as gamma densities with parameters α_i and constant β can no longer be written in the form of an exponential dispersion family. This is however, of no great consequence unless one is dependent on software packages that only will deal with certain classes of models, such as glms.

1.3.3 An Extreme Value Model

Consider, again, the decision to eliminate negative values of soil respiration from the data before developing a regression model. The whole point of such a modeling exercise is to arrive at a model that captures the *distributions* of responses (soil respiration) as soil temperatures vary. It is, after all, well accepted that respiration will increase with increasing temperature, so determining that there is a positive relation between these variables is of no consequence; actually, this is well accepted for temperatures up to *some point* after which it is not clear that respiration continues to increase. We might want to see if we can determine a model that basically captures the

same features of the data as the gamma model developed previously, but that can also accommodate negative responses. One response distribution that suggests itself is that of a right-skew extreme value random variable. This extreme value distribution is a location-scale family, so an additive error model formulation would be natural. At the same time, we also would like to model response variances as proportional to expected values. While this is not easily achieved, we can do something quite similar.

The right-skew version of the extreme value distribution has probability density function, for parameters $-\infty < \xi < \infty$ and $\theta > 0$,

$$f(y|\xi,\theta) = \frac{1}{\theta}\,\exp\left(-\left\{\frac{y-\xi}{\theta}\right\}\right)\,\exp\left[-\exp\left(-\left\{\frac{y-\xi}{\theta}\right\}\right)\right];\quad -\infty < y < \infty. \tag{1.28}$$

The density (1.28) defines a location-scale family of distributions in which the location parameter ξ is equal to the *mode* of the distribution (rather than the expected value) and the variance is given by $(\pi^2/6)\theta^2$. We could then formulate a 'power of the mode' model as,

$$Y_i = \xi_i + \sigma\,\xi_i^{\phi}\,\epsilon_i, \tag{1.29}$$

where

$$\log(\xi_i) = \beta_0 + \beta_1 x_i,$$

and the ϵ_i are assumed to be independent and identically distributed with densities

$$f(\epsilon) = \exp(-\epsilon)\,\exp[-\exp(-\epsilon)];\quad -\infty < \epsilon < \infty.$$