

# Statistical Models

A statistical model often represents an asymmetric relationship. It has the form

$$E(Y) = f(x; \theta), \quad (1)$$

where  $Y$  is a random variable whose realizations are observable,  $x$  is a vector of  $m$  observable variables (“covariates”, “explanatory variables”, etc.) and  $\theta$  is a vector of  $k$  unobservable “parameters”.

Sometimes we can write this as a model with an additive “error” term:

$$Y = f(x; \theta) + E. \quad (2)$$

In this form, the expression “ $f(x; \theta)$ ” represents a systematic effect related to the values of “ $x$ ”, and “ $E$ ” represents a random effect or an unexplained effect.

Sometimes the function  $f$  can be decomposed additively:

$$Y = f_1(x; \theta) + \cdots + f_p(x; \theta) + E.$$

In this “additive model”, each function  $f_j$  may operate on only a subset of the covariates and parameters, so the functions are often written as  $f_j(x_j, \theta_j)$ .

These models have a “systematic component” that relates the covariates with the parameters and an additive “random component”

The important thing to note about these models is that we can model the additive random component  $E$  explicitly. We say, for example, that in a random sample of  $Y_i$ , the corresponding  $E_i$  are i.i.d.  $N(0, \sigma^2)$ .

In the following, we write the observed realizations of  $Y$  as  $y_i$ , the corresponding vector of covariates as  $x_i$  and the unobserved realizations of  $E$  as  $\epsilon_i$ .

We also write the vector of realizations of  $Y$  as  $y$ , and a matrix of observations of  $x$  as  $X$ , and the vector of unobserved realizations of  $E$  as  $\epsilon$ .

# Types of Models

## Linear model:

$$y_i = x_i^T \beta + \epsilon_i$$

or

$$y = X\beta + \epsilon.$$

## General linear model:

The phrase *general linear model*, or GLM, is usually taken to refer to a linear model in which one or more of the independent variables are categorical variables (or classification variables). This is an important type of model. It is, for example, the so-called analysis of variance model or the analysis of covariance model. In this case the  $X$  matrix is not of full rank. There are important consequences concerning what can be estimated or tested in the model, and much of the study of general linear models is concerned with estimability and testability.

## Nonlinear model:

Because the linear model is so widely used, we usually emphasize the distinction of the other case by calling it a “nonlinear model”. It is just the general form of (2). There are important differences in the variances of estimators in the nonlinear model from those in the linear model.

# Generalized Linear Models

For some types of applications, the models in which the response variable is represented as the sum of a systematic effect and a random effect do not serve our purposes well. In these applications, the response variable may take on only two distinct values, “live” or “die”, “success” or “failure”, and so on.

Nelder and Wedderburn (1972) developed a relationship between the models used in these applications and the linear model. The resulting model is called a *generalized linear model*, and is characterized by a *link function* that relates the expected value of the response variable to a linear combination of values of the independent variables plus the random component. Unfortunately the acronym for generalized linear model, GLM, has been used for the general linear model described above; nevertheless, the generalized linear model is sometimes called GLM, especially by those who are particularly interested in it.

# Generalized Models

“Generalized models” are formed by incorporating the covariates and the parameters, that is,  $f(x; \theta)$ , into a “link function”. The general form of the model in equation (1) has this link function as the expected value of the observable random variable of interest. Thus, the systematic component has a hierarchical structure.

## Generalized linear model:

$$\mu = X\beta$$

$$\eta = g(\mu)$$

$$E(Y) = \eta$$

## Generalized additive model:

There are different ways we could write a generalized additive model. If we keep the values of the link function separate, we can write it as

$$\mu_i = f_i(x; \theta_i)$$

$$\eta_i = g(\mu_i)$$

$$E(Y) = \eta_1 + \cdots + \eta_p.$$

We could also write it with a link function expressed as

$$\eta = g(f_1(x; \theta_1) + \cdots + f_p(x; \theta_p)),$$

or we could use different link functions.

# The Probability Distribution in Generalized Models

Notice in the generalized models, we key directly on the distribution of the random variable of interest.

This type of model is particularly useful when that distribution is some simple discrete distribution, such as Bernoulli, binomial, or Poisson.

Whatever the distribution, we will usually assume it to be of the univariate exponential class; that is, using  $\tau$  to represent the parameter, the density or probability function can be written in the form

$$p(y; \tau) = \exp(A(\tau)B(y) + C(y) + D(\tau)).$$

We often impose various restrictions on the support of the distribution and on the functions  $A$ ,  $B$ ,  $C$ , and  $D$ . We may also be able to write the density or probability function in the form

$$p(y; \theta, \phi) = \exp\left(\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\right).$$

In this form of the density of the exponential class,  $\theta$  is the “natural parameter”, and  $\phi$  is the “dispersion parameter”.

In the standard formulation of models with covariates, we often relate the natural parameter  $\theta$  to the specific value of the covariates; hence, with  $y_i$  we associate  $\theta_i$ , or to emphasize the role of the covariates,  $\theta(x_i)$ . You should expect to see either notation, sometimes used interchangeably.

As we have written the density, clearly we are assuming  $a$  is a scalar. We will continue with this assumption, although we could generalize it to be something like a variance-covariance matrix.

The function  $a$  is often the identity

$$a(\phi) = \phi.$$

or (corresponding to a weighted least squares formulation),

$$a(\phi) = \phi/w_i.$$

# The Likelihood

The objective will be to “fit” the model, that is, to estimate the parameters. The model parameters are usually determined either by a maximum likelihood method or by minimizing some function of the residuals.

The “likelihood” is the density considered to be a function of the parameters, given realizations of the random variable:

$$L(\theta_i, \phi | y_i) = \exp \left( \frac{\theta_i y_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right).$$

Maximizing  $L$  is equivalent to maximizing  $l = \log(L)$ ,

$$\frac{\theta_i y_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi).$$

Some functionals of the likelihood are of interest, for example,

$$-\frac{\partial^2 l(y_i)}{\partial \theta^2},$$

which is called the *information* of  $y_i$  on  $\theta$ .

(Notice how we vary the notation:  $l$ ,  $l(y_i)$ , and  $l(\theta_i, \phi | y_i)$  all mean the same thing within context. It is important to note, however, that the argument to the function is the thing we normally think of as the *parameter*; that is  $\theta_i$ , or  $(\theta_i, \phi)$ . The issue of *nuisance parameters*, such as we will often take  $\phi$  to be leads to some interesting statistical theory and methods, such as those based on *quasilikelihood*.)

## Moments of $Y$

Note, that if

$$\frac{\partial}{\partial \theta} \mathbb{E}(l) = \mathbb{E} \left( \frac{\partial l}{\partial \theta} \right),$$

then

$$\mathbb{E} \left( \frac{\partial l}{\partial \theta} \right) = 0,$$

and

$$\mathbb{E} \left( \frac{\partial^2 l}{\partial \theta^2} \right) + \mathbb{E} \left( \frac{\partial l}{\partial \theta} \left( \frac{\partial l}{\partial \theta} \right)^{\text{T}} \right) = 0,$$

so for a random variable  $Y$  with this density,

$$\mathbb{E}(Y) = \frac{d}{d\theta} b(\theta)$$

and

$$\text{V}(Y) = \frac{d^2}{d\theta^2} b(\theta) a(\phi)$$

(remember we're assuming  $a(\phi)$  is a scalar).

In terms of the individual observations, it is convenient to write

$$\mu_i = \mathbb{E}(Y_i | x_i) = \frac{d}{d\theta} b(\theta_i).$$

# Model Fitting

Inference about the model involves estimation of the parameters  $\theta$ , tests of hypotheses about  $\theta$ , and inference about the probability distribution of  $\epsilon$ . It may also involve further consideration of the model,  $f$ , or about other issues relating to the population, e.g., whether it is homogeneous, whether certain observations are outliers, and so on.

A unified approach to model inference involves a method of estimation that allows for statements of confidence and that provides the basis for the subsequent inference regarding the distribution of  $\epsilon$  and the suitability of the model.

Fitting is usually done by maximum likelihood methods or by minimizing some function of the residuals,

$$r = y - f(X; \hat{\theta}).$$

# Fitting Generalized Linear Models

The model parameters are usually estimated either by a maximum likelihood method or by minimizing some function of the residuals. One approach is to use the link function and do a least squares fit of  $\eta$  using the residuals  $y_i - \mu_i$ . It is better, however, to maximize the likelihood or, alternatively, the log-likelihood,

$$l(\theta, \phi | \mathbf{y}) = \sum_{i=1}^n \left( \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right).$$

The most common method of optimizing this function is “Fisher scoring”, which is a method like Newton’s, except that some quantities are replaced by their expected values.

# Newton's Method in General

Problem: find the maximum of  $f(x)$ , a scalar function of a vector argument.

Minimum: maximum  $-f(x)$ .

What is  $f(\cdot)$  like? Suppose can expand in Taylor series:

$$f(x) = f(x_0) + (x - x_0)^T \nabla f(x_0) + \frac{1}{2!} (x - x_0)^T H_f(x_0) (x - x_0) + \dots$$

Now, suppose  $f(\cdot)$  is a quadratic (i.e., no  $\dots$  above).

Suppose its maximum (or minimum) is at  $x_1$ . Then

$$\nabla f(x)_{x=x_1} = 0, \text{ or}$$

$$\nabla \left( (x - x_0)^T \nabla f(x_0) \right)_{x=x_1} + \nabla \left( \frac{1}{2!} (x - x_0)^T H_f(x_0) (x - x_0) \right)_{x=x_1} = 0,$$

i.e.,

$$\nabla f(x_0) + H_f(x_0)(x_1 - x_0) = 0.$$

If  $H_f$  is nonsingular, we have

$$x_1 = x_0 - H_f^{-1}(x_0) \nabla f(x_0).$$

If  $f$  is not quadratic, we can build a sequence of approximations, by expanding  $f$  about some  $x^{(0)}$  getting  $x^{(1)}$ , getting  $x^{(2)}$  and so on, at each stage assuming a quadratic fit.

This is called Newton's method. (Note that in a straightforward application, we have to evaluate and invert  $H$  at each step. Invertible? Positive-definite? How to approximate it?)

# Fisher Scoring on the Log-Likelihood

In the maximum likelihood problem, the variable  $x$  in the previous discussion is the parameter to be estimated. Note therefore that the Hessian in Newton's method is the negative of the matrix of second partials in the loglikelihood evaluated at the observed values.

In the method called Fisher scoring, the Hessian in Newton's method is replaced by its expected value. The iterates then are

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} - \widetilde{H}_l^{-1}(\hat{\theta}^{(k)}|y) \nabla l(\hat{\theta}^{(k)}|y),$$

where  $\widetilde{H}_l$  is the expected value of the information matrix.

The reason this is good is because there is usually a term of the form  $E(Y - E(Y))$ , which is 0.

In the generalized linear model, where the likelihood is linked to the parameters that are really of interest, this still must be cast in terms that will yield values for  $\hat{\beta}$ . The reformulation makes use of the chain rule.

# Models for Binary Data

Consider the situation in which several groups of subjects are each administered a given dose of a drug, and the number responding in each group is recorded. The data consist of the counts  $y_i$  responding in the  $i^{\text{th}}$  group, which received a level  $x_i$  of the drug.

A basic model is

$$P(Y_i = 0|x_i) = 1 - \pi_i$$

$$P(Y_i = 1|x_i) = \pi_i$$

The question is how does  $\pi$  depend on  $x$ ?

A linear dependence,  $\pi = \beta_0 + \beta_1 x$  does not fit well in this kind of situation – unless we impose restrictions,  $\pi$  would not be between 0 and 1.

We can try a transformation.

## Transforming to $[0, 1]$

Suppose we impose an invertable function on

$$\eta = \beta_0 + \beta_1 x$$

that will map it into  $[0, 1]$ :

$$\pi = h(\eta),$$

or

$$g(\pi) = \eta.$$

A link function!

The usual probit model is

$$\pi_x = \Phi(\beta_0 + \beta_1 x),$$

where  $\Phi$  is the normal cumulative distribution function, and  $\beta_0$  and  $\beta_1$  are unknown parameters to be estimated. The link function in this case is  $\Phi^{-1}$ .

The related logit model, in which the log odds ratio  $\log(\pi/(1 - \pi))$  is of interest, has as link function

$$\eta = \log\left(\frac{\pi}{1 - \pi}\right).$$

Other possibilities are the complementary log-log function

$$\eta = \log(-\log(1 - \pi)),$$

and the log-log function,

$$\eta = -\log(-\log(\pi)).$$

# Distributions and Link Functions

## Generalized Linear Models

| Distribution                                   | Link function   |
|--|---|
| <b>regression and analysis of variance</b>     |   |
| normal   | identity: $\mu$   |
| <b>inverse polynomials</b>                     |   |
| gamma  | inverse: $1/\mu$  |
| <b>loglinear models for contingency tables</b> |   |
| Poisson  | log: $\log(\mu)$  |
| <b>logistic regression</b>                     |   |
| binomial                                       | logit or logistic: $\log\left(\frac{\mu}{1-\mu}\right)$ |
| <b>probit analysis</b>                         |   |
| binomial                                       | probit: $\Phi^{-1}(\mu)$                                |
| <b>others</b>                                  |   |
|  | power family: $\mu^\lambda$                             |
|  | Box-Cox: $(\mu^\lambda - 1)/\lambda$                    |
|  | complementary log log: $\log(-\log(1 - \mu))$           |
|  | arcsine: $\sin^{-1}(2\mu - 1)$                          |

# Analysis of Deviance

Recall that our approach to modeling involved using the observations (including the realizations of the random variables) as fixed values and treating the parameters as variables (not random variables, however). The original model was then encapsulated into a likelihood function,  $L(\theta|y)$ , and the principle of fitting the model was maximization of the likelihood with respect to the parameters. The log likelihood,  $l(\theta|y)$ , is usually used.

In model fitting an important issue is how well does the model fit the data? How do we measure the fit? Maybe use residuals. (Remember, some methods of model fitting work this way; they minimize some function of the residuals.) We compare different models by means of the measure of the fit based on the residuals. We make inference about parameters based on changes in the measure of fit.

Using the likelihood approach, we make inference about parameters based on changes in the likelihood. Likelihood ratio tests are based on this principle.

# Comparing Fitted Models

A convenient way of comparing models or making inference about the parameters is with the *deviance function*, which is a likelihood ratio:

$$D(y|\hat{\theta}) = 2(l(\theta_{\max}|y) - l(\hat{\theta}|y)),$$

where  $\hat{\theta}$  is the fit of a potential model.

For generalized linear models the analysis of deviance plays a role similar to that of the analysis of sums of squares (analysis of “variance”) in linear models.

Under appropriate assumptions, when  $\theta_1$  is a subvector of  $\theta_2$ , the difference in deviances of two models,  $D(y|\hat{\theta}_2) - D(y|\hat{\theta}_1)$  has an asymptotic chi-squared distribution with degrees of freedom equal to the difference in the number of parameters.

# Residuals

For models with a binary response variable, we need a different measure of residuals. Because we are measuring the model fit in terms of the deviance,  $D$ , we may think of the observations as each contributing a quantity  $d_i$ , such that  $\sum d_i = D$ . (Exactly what that value is depends on the form of the systematic component and the link function that are in the likelihood.) The quantity

$$r_i = \text{sign}(y_i - \hat{\mu}_i) \sqrt{d_i}$$

increases in  $(y_i - \hat{\mu}_i)$  and  $\sum r_i^2 = D$ . We call  $r_i$  the *deviance residual*.

For the logit model,

$$r_i = \text{sign}(y_i - \hat{\mu}_i) \sqrt{-2[y_i \log(\hat{\pi}_i) + (1 - y_i) \log(1 - \hat{\pi}_i)]}$$

These are available in S-Plus via the `type="deviance"` option in `residuals` with an argument that is a fit object.

Another kind of residual is called the “working” residual. It is

$$r_i^W = (y_i - \hat{\mu}_i) \frac{\partial \hat{\eta}_i}{\partial \hat{\mu}_i},$$

where the derivatives are evaluated at the final iteration of the scoring algorithm. In S-Plus the working residuals can be obtained by `type="working"` in `residuals` with an argument that is a fit object.

# Standardizing Residuals

They can be standardized by taking into account their different standard deviations that result from the influence, using `lm.influence`.

This is the same kind of concept as influence in linear models. Here, however, we have

$$\hat{\beta}^{(k+1)} = (X^T W^{(k)} X)^{-1} X^T W^{(k)} y^{(k)},$$

where the weights are

$$m_i \hat{\pi}_i^{(k)} (1 - \hat{\pi}_i^{(k)}).$$

One measure is the diagonal of the hat matrix:

$$W^{\frac{1}{2}} X (X^T W X)^{-1} X^T W^{\frac{1}{2}}$$

In the case of generalized linear models, the hat matrix is only the prediction transformation matrix for the linear, systematic component.

In S-Plus, `lm.influence` operating on an object of type `lm` gives various types of influence measures.

# Partial Residual Plots

Partial residual plots (also called “component-plus-residual”, or “C+R”, plots) are often used to assess the usefulness of a predictor variable and to determine whether it may be better to transform a predictor variable.

The idea is to look at the sum of the fitted effect of the given predictor and the residual from the fit (with all variables in the model).

In the linear model, for example,

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi},$$

we fit the full model (called the “working model”), and then look at, say,  $\hat{\beta}_p x_{pi} + r_i^W$ . A plot of this against  $x_{pi}$  is called a partial residual plot. It can be informative in assessing whether  $x_{pi}$  belong in the model, or perhaps some transform,  $t(x_{pi})$  would work better. The idea is to inspect the model

$$\hat{t}(x_{pi}) = t(x_{pi}) + \epsilon_i.$$

# Generalized Linear Models in S-Plus

S-Plus handles generalized linear models in the `glm` function.

Produces an object of class “glm” which is a generalized linear fit of the data.

The link function is specified through the `family` keyword.

`family`

Generates a family object containing a list of functions and expressions used by `glm` and `gam`.

```
family(object)
binomial(link=logit)
gaussian()
Gamma(link=inverse)
inverse.gaussian()
poisson(link=log)
quasi(link=identity, variance=constant)
```

From the S-Plus documentation:

The choices of link functions are logit, probit, cloglog, identity, inverse, log,  $1/\mu^{**2}$ , and sqrt. Not all links are suitable for all families.

Each of the names, except for quasi and the family extractor function family, are associated with a member of the exponential family of distributions. As such, they have a fixed variance function. There is typically a choice of link functions, with the default corresponding to the canonical link for that family. The quasi name represents Quasi-likelihood and need not correspond to any particular distribution; rather quasi can be used to combine any available link and variance function.

The following table summarizes the suitable pairings:

|               | binomial | gaussian | Gamma | inverse.gaussian | poisson | quasi |
|---------------|----------|----------|-------|------------------|---------|-------|
| logit         | *        |          |       |                  |         | *     |
| probit        | *        |          |       |                  |         | *     |
| cloglog       | *        |          |       |                  |         | *     |
| identity      |          | *        | *     |                  | *       | *     |
| inverse       |          |          | *     |                  |         | *     |
| log           |          |          | *     |                  | *       | *     |
| $1/\mu^{**2}$ |          |          |       | *                |         | *     |
| sqrt          |          |          |       |                  | *       | *     |

The function power can also be used to generate a power link function object for use with quasi; power takes an argument lambda.

Users can construct their own families, as long as they have compatible components having the same names as those.

## An Example

The data are from E. T. Lee (1980), *Statistical Methods for Survival Analysis*. The data are measurements of six pre-treatment variables on 51 patients with acute myeloblastic leukemia. Three post-treatment variables are available. These data were also analyzed by Everitt (1994). The data are in `cancer.dat` on the class homepage.

We bring them into S-Plus with

```
cancer.dat<-read.table("cancer.dat",  
  col.names=c("Age", "Smear", "Infil", "Index",  
  "Blasts", "Temp", "Resp", "Time", "Status"))
```

And we get a summary of the data with

```
summary(cancer.dat)
```

| Age           | Smear         | Infil        |
|---------------|---------------|--------------|
| Min. :20.00   | Min. :26.00   | Min. : 8.0   |
| 1st Qu.:35.00 | 1st Qu.:53.50 | 1st Qu.:37.5 |
| Median :50.00 | Median :69.00 | Median :61.0 |
| Mean :49.86   | Mean :65.94   | Mean :58.2   |
| 3rd Qu.:61.00 | 3rd Qu.:81.00 | 3rd Qu.:72.5 |
| Max. :80.00   | Max. :97.00   | Max. :95.0   |

| Index          | Blasts         | Temp           |
|----------------|----------------|----------------|
| Min. : 1.000   | Min. : 0.000   | Min. : 980.0   |
| 1st Qu.: 6.000 | 1st Qu.: 1.000 | 1st Qu.: 986.0 |
| Median : 9.000 | Median : 2.600 | Median : 990.0 |
| Mean : 9.804   | Mean : 7.339   | Mean : 996.1   |
| 3rd Qu.:13.500 | 3rd Qu.: 9.950 | 3rd Qu.:1005.0 |
| Max. :20.000   | Max. :38.000   | Max. :1038.0   |

| Resp           | Time          | Status         |
|----------------|---------------|----------------|
| Min. :0.0000   | Min. : 0.00   | Min. :0.0000   |
| 1st Qu.:0.0000 | 1st Qu.: 1.50 | 1st Qu.:0.0000 |
| Median :0.0000 | Median : 7.00 | Median :0.0000 |
| Mean :0.4706   | Mean :11.75   | Mean :0.1176   |
| 3rd Qu.:1.0000 | 3rd Qu.:18.00 | 3rd Qu.:0.0000 |
| Max. :1.0000   | Max. :45.00   | Max. :1.0000   |

The question of interest is whether the response to the treatment appeared to be related to the pre-treatment variables. There are three response variables: Resp, which is whether or not any response was evident (0,1); Time, which is survival time, in months, following the treatment; and Status, which is whether or not the patient was still alive.

The next thing to do is to look at some plots of the data.

With data for which the usual linear model works well we would do some scatterplots. In a general linear model, or an analysis of variance model, we would probably do some boxplots of the response variable at the different levels of the factors.

For these data we may look at boxplots of the predictor variables at various levels of the response variables using statements such as

```
plot.factor(as.factor(cancer.dat$Resp),  
            cancer.dat$Age,xlab='Response',ylab='Age')
```

Doing these plots, we do not see anything very unusual. There are no real outliers. No particular predictor variable seems closely related to the outcome variable. (Remember, of course, these are two-way plots.)

The data have information about initial response, survival time, and ultimate survival. Even if the patient's final outcome is not good, the initial response may provide information for development of better treatments.

Our next step is to analyze relationship of the pre-treatment variables and the initial response with a generalized linear model. Letting  $\pi$  be the probability of an initial response, we may use the log odds ratio,  $\log(\pi/(1 - \pi))$ , to map the interval  $(0, 1)$  to  $(-\infty, \infty)$ , and then write the logistic regression model:

$$\log(\pi/(1 - \pi)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_6 x_6$$

# Fitting the Model

The response variable is Bernoulli (or binomial). We model the log odds ratios as

$$\begin{aligned}\log\left(\frac{\pi_i}{1-\pi_i}\right) &= \eta_i \\ &= \beta_0 + \beta_1 x_{1i} + \cdots + \beta_6 x_{6i} \\ &= \mathbf{x}_i^T \boldsymbol{\beta}.\end{aligned}$$

For a binomial with number  $m_i$ , we write the log-likelihood,

$$l(\boldsymbol{\pi}|\mathbf{y}) = \sum_{i=1}^n (y_i \log(\pi_i/(1-\pi_i)) + m_i \log(1-\pi_i)),$$

where a constant involving  $m_i$  and  $y_i$  has been omitted. Substituting, we have,

$$l(\boldsymbol{\beta}|\mathbf{y}) = \sum_{i=1}^n y_i \mathbf{x}_i^T \boldsymbol{\beta} - \sum_{i=1}^n m_i \log(1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})).$$

The log likelihood depends on  $\mathbf{y}$  only through  $\mathbf{X}^T \mathbf{y}$ .

# The Likelihood

$$\frac{\partial l}{\partial \pi_i} = \frac{y_i - m_i \pi_i}{\pi_i(1 - \pi_i)}$$

Using the chain rule, we have

$$\begin{aligned} \frac{\partial l}{\partial \beta_j} &= \sum_{i=1}^n \frac{y_i - m_i \pi_i}{\pi_i(1 - \pi_i)} \frac{\partial \pi_i}{\partial \beta_j} \\ &= \sum_{i=1}^n \frac{y_i - m_i \pi_i}{\pi_i(1 - \pi_i)} \frac{d\pi_i}{d\eta_i} x_{ij} \end{aligned}$$

The Fisher information is

$$\begin{aligned} -\mathbb{E} \left( \frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right) &= \sum_{i=1}^n \frac{m_i}{\pi_i(1 - \pi_i)} \frac{\partial \pi_i}{\partial \beta_j} \frac{\partial \pi_i}{\partial \beta_k} \\ &= \sum_{i=1}^n \frac{m_i (d\pi_i/d\eta_i)^2}{\pi_i(1 - \pi_i)} x_{ij} x_{ik} \\ &= (X^T W X)_{jk}, \end{aligned}$$

where  $W$  is a diagonal matrix of weights,

$$\frac{m_i (d\pi_i/d\eta_i)^2}{\pi_i(1 - \pi_i)}$$

Notice

$$\frac{d\pi_i}{d\eta_i} = \pi_i(1 - \pi_i),$$

so we have the simple expression,

$$\frac{\partial l}{\partial \beta} = X^T(y - m\pi)$$

in matrix notation, and for the weights we have,

$$m_i\pi_i(1 - \pi_i)$$

# Maximizing the Likelihood

Use a Newton approach,

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} - H_l^{-1}(\hat{\beta}^{(k)}) \nabla l(\hat{\beta}^{(k)}),$$

in which  $H_l$  is replaced by

$$\mathbb{E} \left( - \frac{\partial^2 l}{\partial \beta \partial \beta^T} \right)$$

Using  $\hat{\beta}^{(k)}$ , we form  $\hat{\pi}^{(k)}$  and  $\hat{\eta}^{(k)}$ , and then, an adjusted  $y^{(k)}$ ,

$$y_i^{(k)} = \hat{\eta}^{(k)} + \frac{(y - m_i \hat{\pi}_i^{(k)})}{m_i} \frac{d\eta_i}{d\pi_i}$$

This leads to

$$\hat{\beta}^{(k+1)} = (X^T W^{(k)} X)^{-1} X^T W^{(k)} y^{(k)},$$

and it suggests an iteratively reweighted least squares (IRLS) algorithm.

# Logistic Regression in S-Plus

```
cancer.fit<-  
  glm(Resp~Age+Smear+Infil+Index+Blasts+Temp,  
      data=cancer.dat,family=binomial)  
summary(cancer.fit)
```

Call: glm(formula = Resp ~ Age + Smear + Infil + Index + Blasts + Temp,  
family = binomial, data = cancer.dat)

Deviance Residuals:

| Min      | 1Q        | Median      | 3Q        | Max      |
|----------|-----------|-------------|-----------|----------|
| -1.73878 | -0.581007 | -0.05505222 | 0.6261866 | 2.284218 |

Coefficients:

|             | Value        | Std. Error  | t value    |
|-------------|--------------|-------------|------------|
| (Intercept) | 98.520952735 | 40.72037076 | 2.4194513  |
| Age         | -0.060290736 | 0.02719599  | -2.2168979 |
| Smear       | -0.004799158 | 0.04098958  | -0.1170824 |
| Infil       | 0.036211819  | 0.03921805  | 0.9233457  |
| Index       | 0.398436871  | 0.13221744  | 3.0134971  |
| Blasts      | 0.013434989  | 0.05769707  | 0.2328539  |
| Temp        | -0.102225888 | 0.04167088  | -2.4531731 |

(Dispersion Parameter for Binomial family taken to be 1 )

Null Deviance: 70.52444 on 50 degrees of freedom

Residual Deviance: 40.05991 on 44 degrees of freedom

Number of Fisher Scoring Iterations: 5

Correlation of Coefficients:

|        | (Intercept) | Age        | Smear      | Infil      | Index      | Blasts     |
|--------|-------------|------------|------------|------------|------------|------------|
| Age    | -0.3527213  |            |            |            |            |            |
| Smear  | 0.0455059   | 0.1730232  |            |            |            |            |
| Infil  | 0.0366861   | -0.2439339 | -0.8298836 |            |            |            |
| Index  | 0.5429021   | -0.4015964 | 0.0868805  | 0.0833581  |            |            |
| Blasts | 0.3337598   | 0.0808698  | 0.1169333  | -0.2633919 | -0.1796423 |            |
| Temp   | -0.9985338  | 0.3302739  | -0.0732672 | -0.0306282 | -0.5630481 | -0.3275219 |

In S-Plus, the glm object is appropriate for handling by anova.

```
anova(cancer.fit)
```

Yields

Analysis of Deviance Table

Binomial model

Response: Resp

Terms added sequentially (first to last)

|        | Df | Deviance | Resid. | Df | Resid. Dev |
|--------|----|----------|--------|----|------------|
| NULL   |    |          |        | 50 | 70.52444   |
| Age    | 1  | 6.52068  |        | 49 | 64.00376   |
| Smear  | 1  | 1.25490  |        | 48 | 62.74885   |
| Infil  | 1  | 1.80467  |        | 47 | 60.94418   |
| Index  | 1  | 12.12508 |        | 46 | 48.81910   |
| Blasts | 1  | 0.54165  |        | 45 | 48.27745   |
| Temp   | 1  | 8.21754  |        | 44 | 40.05991   |

# How to Do Modeling

There are many issues in modeling relationships. Two classes of (related) questions to be addressed for the systematic component are what independent (predictor) variables are relevant, and what is the functional form of the relationship.

Building models is an iterative process. We begin with a reasonable (and simple) functional form that includes many (all) independent variables that may be relevant.

We fit and assess.

The three variables Age, Index, and Temp seem to be most important in our preliminary fit. We will now study these three further.

As in Everitt (1994), let's consider them sequentially:

```
cancer.fit1<-glm(Resp~Age,  
  data=cancer.dat,family=binomial)  
cancer.fit2<-glm(Resp~Age+Index,  
  data=cancer.dat,family=binomial)  
cancer.fit3<-glm(Resp~Age+Index+Temp,  
  data=cancer.dat,family=binomial)
```

Now,

```
anova(cancer.fit1,cancer.fit2,cancer.fit3)
```

yields

Analysis of Deviance Table

Response: Resp

|   | Terms          | Resid. Df | Resid. Dev | Test   | Df | Deviance |
|---|----------------|-----------|------------|--------|----|----------|
| 1 | Age            | 49        | 64.00374   |        |    |          |
| 2 | Age+Index      | 48        | 51.38696   | +Index | 1  | 12.61677 |
| 3 | Age+Index+Temp | 47        | 43.26538   | +Temp  | 1  | 8.12158  |

For the final model

```
summary(cancer.fit3)
```

yields

```
Call: glm(formula = Resp ~ Age + Index + Temp,
          family = binomial, data = cancer.dat)
```

Deviance Residuals:

| Min       | 1Q         | Median      | 3Q        | Max      |
|-----------|------------|-------------|-----------|----------|
| -1.761036 | -0.6868326 | -0.09746944 | 0.6738808 | 2.165102 |

Coefficients:

|             | Value       | Std. Error  | t value   |
|-------------|-------------|-------------|-----------|
| (Intercept) | 87.38781928 | 35.42139859 | 2.467091  |
| Age         | -0.05850149 | 0.02555483  | -2.289254 |
| Index       | 0.38492521  | 0.12136186  | 3.171715  |
| Temp        | -0.08897297 | 0.03603080  | -2.469359 |

(Dispersion Parameter for Binomial family taken to be 1 )

Null Deviance: 70.52444 on 50 degrees of freedom

Residual Deviance: 43.26538 on 47 degrees of freedom

Number of Fisher Scoring Iterations: 5

Correlation of Coefficients:

|       | (Intercept) | Age        | Index      |
|-------|-------------|------------|------------|
| Age   | -0.3414968  |            |            |
| Index | 0.6470771   | -0.3676001 |            |
| Temp  | -0.9992598  | 0.3162529  | -0.6616655 |

# Interpreting the Model

The estimated coefficient for Age, -0.059, applies to the log of the odds, when Index and Temp are held fixed. For the odds, we use

```
exp(cancer.fit3$coef[2])
```

and get

```
Age  
0.9431768
```

This means that, other things being the same, the odds for a patient who is one year older than the other, in favor of responding to the treatment, is 94%.

# Continuing with the Logistic Regression Example

The data are from Lee (1980) are measurements on 51 cancer patients of six pre-treatment variables and three response variables. The data are in `cancer.dat`.

After some preliminary analyses, the three variables Age, Index, and Temp seem to be most important and we used `glm` in S-Plus to fit the logistic model.

```
cancer.fit3<-glm(Resp~Age+Index+Temp,  
  data=cancer.dat,family=binomial)  
anova(cancer.fit3)  
summary(cancer.fit3)
```

We got

Analysis of Deviance Table

Binomial model

Response: Resp

Terms added sequentially (first to last)

|       | Df | Deviance | Resid. | Df | Resid. | Dev      |
|-------|----|----------|--------|----|--------|----------|
| NULL  |    |          |        | 50 |        | 70.52444 |
| Age   | 1  | 6.52068  |        | 49 |        | 64.00376 |
| Index | 1  | 12.61677 |        | 48 |        | 51.38699 |
| Temp  | 1  | 8.12161  |        | 47 |        | 43.26538 |

and

Call: glm(formula = Resp ~ Age + Index + Temp,  
family = binomial, data = cancer.dat)

Deviance Residuals:

| Min       | 1Q         | Median      | 3Q        | Max      |
|-----------|------------|-------------|-----------|----------|
| -1.761036 | -0.6868326 | -0.09746944 | 0.6738808 | 2.165102 |

Coefficients:

|             | Value       | Std. Error  | t value   |
|-------------|-------------|-------------|-----------|
| (Intercept) | 87.38781928 | 35.42139859 | 2.467091  |
| Age         | -0.05850149 | 0.02555483  | -2.289254 |
| Index       | 0.38492521  | 0.12136186  | 3.171715  |
| Temp        | -0.08897297 | 0.03603080  | -2.469359 |

(Dispersion Parameter for Binomial family taken to be 1 )

Null Deviance: 70.52444 on 50 degrees of freedom

Residual Deviance: 43.26538 on 47 degrees of freedom

Number of Fisher Scoring Iterations: 5

Correlation of Coefficients:

|       | (Intercept) | Age        | Index      |
|-------|-------------|------------|------------|
| Age   | -0.3414968  |            |            |
| Index | 0.6470771   | -0.3676001 |            |
| Temp  | -0.9992598  | 0.3162529  | -0.6616655 |

# Logistic Regression in SAS

```
options linesize=78;
data;
  infile "cancer.dat";
  input age smear infil index blasts temp resp time status;
proc logistic;
  model resp = age smear infil index blasts temp
    /selection = stepwise;
run;
```

# The SAS System

## The LOGISTIC Procedure

Data Set: WORK.DATA1  
Response Variable: RESP  
Response Levels: 2  
Number of Observations: 51  
Link Function: Logit

### Response Profile

| Ordered Value | RESP | Count |
|---------------|------|-------|
| 1             | 0    | 27    |
| 2             | 1    | 24    |

### Stepwise Selection Procedure

Step 0. Intercept entered:

Residual Chi-Square = 21.8168 with 6 DF (p=0.0013)

Step 1. Variable INDEX entered:

### Criteria for Assessing Model Fit

| Criterion      | Intercept Only | Intercept and Covariates | Chi-Square for Covariates   |
|----------------|----------------|--------------------------|-----------------------------|
| AIC            | 72.524         | 61.119                   | .                           |
| SC             | 74.456         | 64.982                   | .                           |
| -2 LOG L Score | 70.524         | 57.119                   | 13.406 with 1 DF (p=0.0003) |
|                | .              | .                        | 12.155 with 1 DF (p=0.0005) |

Residual Chi-Square = 13.3917 with 5 DF (p=0.0200)

Step 2. Variable TEMP entered:

Criteria for Assessing Model Fit

| Criterion | Intercept<br>Only | Intercept<br>and<br>Covariates | Chi-Square for Covariates   |
|-----------|-------------------|--------------------------------|-----------------------------|
| AIC       | 72.524            | 55.645                         | .                           |
| SC        | 74.456            | 61.441                         | .                           |
| -2 LOG L  | 70.524            | 49.645                         | 20.879 with 2 DF (p=0.0001) |
| Score     | .                 | .                              | 16.744 with 2 DF (p=0.0002) |

Residual Chi-Square = 8.5414 with 4 DF (p=0.0736)

Step 3. Variable AGE entered:

Criteria for Assessing Model Fit

| Criterion | Intercept<br>Only | Intercept<br>and<br>Covariates | Chi-Square for Covariates   |
|-----------|-------------------|--------------------------------|-----------------------------|
| AIC       | 72.524            | 51.265                         | .                           |
| SC        | 74.456            | 58.993                         | .                           |
| -2 LOG L  | 70.524            | 43.265                         | 27.259 with 3 DF (p=0.0001) |
| Score     | .                 | .                              | 20.315 with 3 DF (p=0.0001) |

Residual Chi-Square = 2.9539 with 3 DF (p=0.3988)

NOTE: No (additional) variables met the 0.05 significance level for entry into the model.

### Summary of Stepwise Procedure

| Step | Variable |         | Number<br>In | Score<br>Chi-Square | Wald<br>Chi-Square | Pr ><br>Chi-Square |
|------|----------|---------|--------------|---------------------|--------------------|--------------------|
|      | Entered  | Removed |              |                     |                    |                    |
| 1    | INDEX    |         | 1            | 12.1555             | .                  | 0.0005             |
| 2    | TEMP     |         | 2            | 6.4453              | .                  | 0.0111             |
| 3    | AGE      |         | 3            | 6.0313              | .                  | 0.0141             |

### Analysis of Maximum Likelihood Estimates

| Variable | DF | Parameter<br>Estimate | Standard<br>Error | Wald<br>Chi-Square | Pr ><br>Chi-Square | Standardized<br>Estimate | Odds<br>Ratio |
|----------|----|-----------------------|-------------------|--------------------|--------------------|--------------------------|---------------|
| INTERCPT | 1  | -87.3880              | 35.4581           | 6.0740             | 0.0137             | .                        | 0.000         |
| AGE      | 1  | 0.0585                | 0.0256            | 5.2319             | 0.0222             | 0.532528                 | 1.060         |
| INDEX    | 1  | -0.3849               | 0.1215            | 10.0340            | 0.0015             | -1.012469                | 0.681         |
| TEMP     | 1  | 0.0890                | 0.0361            | 6.0851             | 0.0136             | 0.744581                 | 1.093         |

### Association of Predicted Probabilities and Observed Responses

|                    |                   |
|--------------------|-------------------|
| Concordant = 87.8% | Somers' D = 0.756 |
| Discordant = 12.2% | Gamma = 0.756     |
| Tied = 0.0%        | Tau-a = 0.384     |
| (648 pairs)        | c = 0.878         |

# Other Topics

The mechanical process of dealing with generalized additive models parallels that of dealing with generalized linear models. There are some very important differences, however. The most important is probably that the distribution of the deviances is not worked out.

The meaning of degrees of freedom is also somewhat different.

Notice that the mean of the binomial and the Poisson distributions determine the variance.

In practice the variance of discrete response data, such as binomial or Poisson data, is observed to exceed the nominal variance that would be determined by the mean.

This phenomenon is referred to as “over-dispersion”. There may be logical explanations for over-dispersion, such as additional heterogeneity over and above what is accounted for by the covariates, or some more complicated variance structure arising from correlations among the responses.

Dean (1992) has proposed tests for overdispersion in binomial and Poisson models.

## Quasilikelihood Methods

Over-dispersion can often be accounted for by the nuisance parameter  $\phi$  in the likelihood. For example, we modify the simple binomial model so the variance is

$$V(y_i|x_i) = \phi \frac{\pi_i(1 - \pi_i)}{n_i}.$$

Notice the multiplier  $\phi$  is constant, while  $\pi$  depends on the covariates and  $n$  depends on the group size. This of course leads to a more complicated likelihood function, but it may not be necessary to use the actual likelihood.

Wedderburn (1974) introduced a quasilikelihood function to allow

$$E(y|x) = \mu = h(x^T\beta)$$

and

$$V(y|x) = \sigma^2(\mu) = \phi v(\mu),$$

where  $\phi$  is the (nuisance) dispersion parameter in the likelihood and  $v(\mu)$  is a variance function that is entirely separate from the likelihood.

McCullagh (1983) extended the concept of quasilikelihood to allow for a variance-covariance matrix  $V$ , and derived an asymptotic theory for the resulting estimators. There may or may not be a true likelihood function with corresponding mean and variance (see Morris, 1982).

Quasilikelihood methods require only specification of a relationship between the mean and variance of the response.

Fahrmeir and Tutz (1994), following ideas of Gourieroux, Montfort, and Trognon (1984), assume that the mean is indeed given by  $\mu = h(x^T\beta)$ , but that the variance is

$$V(y|x) = \sigma_0^2(\mu),$$

which may be different from  $\sigma^2(\mu) = \phi v(\mu)$ . They take  $\phi v(\mu)$  to be a “working variance”. Assuming that the responses are independent, they write a “quasi-score function”,

$$s(\beta) = \sum_i x_i D_i(\beta) \sigma_i^2(\beta) (y_i - \mu_i(\beta)),$$

where  $\mu_i(\beta)$  is the correct mean, i.e.,  $h(x_i^T\beta)$  and  $D_i(\beta)$  is the derivative of  $h$ , but  $\sigma_i^2(\beta)$  is a working variance,  $\phi v(\mu(\beta))$ , with  $v$  arbitrary. Obviously, to use the quasi-score function for computing estimates,  $v$  must be somewhat close to the true variance of the data.

In the ordinary quasiliikelihood methods, the variance function is assumed known (or arbitrary, but not estimated directly). Nelder and Pregibon (1987) developed an extended quasiliikelihood approach, in which the variance function is also studied.

Green and Silverman (1994), using a roughness penalty approach, developed quasiliikelihood methods for semiparametric generalized linear models, in which the penalized quasiliikelihood estimates are penalized least squares estimates with a weight function corresponding to the inverse of the variance-covariance matrix,  $V$ .

Morgenthaler (1992) and Jung (1996) considered quasiliikelihood methods of estimation for generalized linear models using least absolute deviations and other robust fitting methods.