CHAPTER 18

Generalized Linear Models (GLIM)

18.1. INTRODUCTION

In Section 9.2, we discussed *generalized least squares* (of which *weighted least squares* is a special case). That section concerned the application of least squares methods in situations where $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where the errors were normal, and $E(\boldsymbol{\epsilon}) = \mathbf{0}$, but the variance–covariance matrix of the errors was $\mathbf{V}(\mathbf{Y}) = \mathbf{V}\sigma^2$, not merely $\mathbf{I}\sigma^2$. The topic of this chapter is different from the Section 9.2 topic, although there are two points of confusion. One is the use of the word "generalized" in both titles. The other is that the maximum likelihood solution for estimating parameters in generalized linear models (this chapter!) involves the use of generalized (weighted) least squares in an iterative manner.

Generalized linear models (GLIM) analysis comes into play when the error distribution is not normal and/or when a vector of nonlinear *functions* of the responses, $\boldsymbol{\eta}(\mathbf{Y}) = \{\eta(Y_1), \eta(Y_2), \ldots, \eta(Y_n)\}'$, say, and not $\mathbf{Y}$ itself, has expection $\mathbf{X}\boldsymbol{\beta}$. GLIM analysis provides a larger framework for estimation, which *includes* the case of normally distributed errors. When the errors *are* normal and $\eta(Y_i) = Y_i$, the GLIM analysis applies perfectly well but simplifies to the least squares case with which we have become familiar. Because of limitations in the general GLIM procedure, direct use of least squares is most appropriate for regression data with normally distributed errors, employing techniques already described.

In this chapter we offer a brief self-contained introduction to GLIM. This is intended to set the least squares methods already discussed into a more general context and to prepare the reader for specialized GLIM textbooks. Actual implementation of GLIM on data requires a complicated calculation for which a number of standard programs are available. Becoming familiar with one of these is advisable for further study.

Acronym

We are going to use the acronym GLIM in this chapter. There is a slight confusion, when we do, with the GLIM computer package, which is only *one* of several possible packages that perform the appropriate calculations. (See Hilbe, 1994.) (However, both GLIM acronyms apply to the same thing, and this is not true of another possible acronym, GLM.)

18.2. THE EXPONENTIAL FAMILY OF DISTRIBUTIONS

We start by talking about a very general family of error distributions, which includes the normal distribution as a special case. A random variable u that belongs to the *exponential family* with a single parameter θ has a probability density function

$$f(u, \theta) = s(u)t(\theta)e^{a(u)b(\theta)}, \tag{18.2.1}$$

where the letters s, t, a, and b are all known functions. It is convenient to rewrite this as

$$f(u, \theta) = \exp\{a(u)b(\theta) + c(\theta) + d(u)\}, \tag{18.2.2}$$

where $d(u) = \ln s(u)$ and $c(\theta) = \ln t(\theta)$. If there were other parameters not of direct interest, they could occur within the functions a, b, c, and d.

Some Definitions

When $a(u) = u$ in (18.2.2), the distribution is said to be in *canonical form* and $b(\theta)$ is then called the *natural parameter* of the distribution. Parameters other than the parameter of interest θ can be involved in the functions a, b, c, and d. These are *nuisance* parameters and are regarded as known. For example, for the normal variable u in (18.2.4), given below, the natural parameter is μ/σ^2, and σ^2 is a nuisance parameter if μ is of interest. For the binomial u in (18.2.5), given below, the natural parameter is $\ln[p/(1-p)]$, the natural logarithm of the odds ratio $p/(1-p)$. Although the canonical parameterization has convenient computational and theoretical properties, it is not necessarily the best choice in applications.

[Nuisance parameters are usually scale parameters and, asymptotically in most cases, exactly in the normal case, do not need to be estimated along with the parameters of the linear model to be fitted in (18.3.2). However, when there is concern about the possibility of dependence in the observations, scale parameters may be estimated even when not required. This enables confidence intervals to be widened, for example, and so made more robust against the possible *overdispersion* of the observations due to dependence. We do not further discuss this complication.]

Some Members of the Exponential Family

1. The normal distribution $N(\mu, \sigma^2)$.

$$f(u, \mu) = (2\pi\sigma^2)^{-1/2}\exp\{-(u-\mu)^2/(2\sigma^2)\}, \qquad -\infty \le u \le \infty, \tag{18.2.3}$$

where μ is the parameter of interest and σ^2 is regarded as a nuisance parameter (or as known). By the contortional rewrite

$$f(u, \mu) = \exp\{u(\mu/\sigma^2) + [-\mu^2/(2\sigma^2) - \tfrac{1}{2}\ln(2\pi\sigma^2)] - u^2/(2\sigma^2)\}, \tag{18.2.4}$$

we recognize that with $a(u) = u$, $b(\theta) = \mu/\sigma^2$, and so on, the form (18.2.2) appears.

2. The binomial distribution $B(n, p)$.

$$\begin{aligned} f(u, p) &= \binom{n}{u} p^u(1-p)^{n-u}, \qquad u = 0, 1, 2, \ldots, n, \\ &= \exp\left\{u \ln[p/(1-p)] + n\ln(1-p) + \ln\binom{n}{u}\right\}. \end{aligned} \tag{18.2.5}$$

3. The Poisson distribution.

$$\begin{aligned} f(u, \lambda) &= \frac{\lambda^u e^{-\lambda}}{u!}, \qquad u = 0, 1, 2, \ldots \\ &= \exp\{u \ln \lambda - \lambda - \ln(y!)\}. \end{aligned} \tag{18.2.6}$$

The natural parameter is $\ln \lambda$.

4. The gamma distribution with parameter θ of interest and ϕ as nuisance parameter.

$$\begin{aligned} f(u, \theta) &= \{u^{\phi-1}\theta^{\phi} \exp(-u\theta)\}/\Gamma(\phi), \qquad 0 \le u \le 1 \\ &= \exp\{-u\theta + [\phi \ln u - \ln \Gamma(\phi)] + (\phi - 1)\ln u\}. \end{aligned} \tag{18.2.7}$$

5. The Pareto distribution.

$$\begin{aligned} f(u, \theta) &= \theta u^{-(1+\theta)}, \qquad u \ge 0 \\ &= \exp\{-(1 + \theta)\ln u + \ln \theta\}. \end{aligned} \tag{18.2.8}$$

This is not of canonical form.

6. The exponential distribution.

$$\begin{aligned} f(u, \theta) &= \theta \exp(-u\theta), \qquad u \ge 0 \\ &= \exp\{-u\theta + \ln \theta\}. \end{aligned} \tag{18.2.9}$$

The natural parameter is θ. This distribution is a special case of (18.2.7) when $\phi = 1$.

7. The negative binomial distribution. The variable u is the number of failures observed to attain a given number (r) of successes in binomial trials with probability θ of success. In the GLIM context, the distribution more often arises via an overdispersed Poisson variable (Feller, 1957).

$$\begin{aligned} f(u, \theta) &= \binom{u + r - 1}{r - 1} \theta^r (1 - \theta)^u, \qquad u = 1, 2, \ldots \\ &= \exp\left\{u \ln(1 - \theta) + r \ln \theta + \ln \binom{u + r - 1}{r - 1}\right\}. \end{aligned} \tag{18.2.10}$$

The natural parameter is $\ln(1 - \theta)$.

Expected Value and Variance of $a(u)$

The following results can be obtained:

$$\begin{aligned} E[a(u)] &= -c'(\theta)/b'(\theta), \\ V[a(u)] &= [b''(\theta)c'(\theta) - c''(\theta)b'(\theta)]/[b'(\theta)]^3, \end{aligned} \tag{18.2.11}$$

where the prime indicates differentiation with respect to θ. For the binomial, for example, $a(u) = u$, $\theta = p$, and

$$\begin{aligned} &b(\theta) = \ln[p/(1 - p)], \qquad b'(\theta) = 1/[p(1 - p)], \qquad b''(\theta) = (2p - 1)/[p(1 - p)]^2, \\ &c(\theta) = n \ln(1 - p), \qquad c'(\theta) = -n/(1 - p), \qquad c''(\theta) = -n/(1 - p)^2, \end{aligned} \tag{18.2.12}$$

and so

$$E(u) = np, \qquad V(u) = np(1-p). \tag{18.2.13}$$

For the normal distribution the formulas (18.2.11) lead to μ and σ^2, of course.

Joint Probability Density of a Sample

If we have a sample $Y_1, Y_2, \ldots, Y_n$ of independent observations from (18.2.2), the joint probability density function is

$$\exp\left\{b(\theta)\sum_{i=1}^{n} a(Y_i) + nc(\theta) + \sum_{i=1}^{n} d(Y_i)\right\}. \tag{18.2.14}$$

The quantity $\Sigma_{i=1}^{n} a(Y_i)$ is a *sufficient statistic* for $b(\theta)$. This implies that it carries all the information the sample provides about $b(\theta)$; knowing the individual Y_i, for example, would not tell us any more about $b(\theta)$.

18.3. FITTING GENERALIZED LINEAR MODELS (GLIM)

The fitting of generalized linear models via iteratively reweighted least squares, which has proved a useful practical technique over the years, stems from work of Nelder and Wedderburn (1972). Suppose we have a set of independent observations $Y_1, Y_2, \ldots, Y_n$ each from the same exponential type distribution (e.g., all binomials) of canonical form [i.e., $a(Y) = Y$]. Then the joint probability density function is

$$f(Y_i, \boldsymbol{\theta}, \boldsymbol{\phi}) = \exp\{\Sigma Y_i b(\theta_i) + \Sigma c(\theta_i) + \Sigma d(Y_i)\}, \tag{18.3.1}$$

where all summations are from $i = 1, \ldots, n$ unless otherwise specified, where $\boldsymbol{\phi}$ is a vector of nuisance parameters that occur within $b(\theta)$, $c(\theta)$, and $d(\theta)$ although not explicitly shown therein, and where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)'$ is a vector of parameters of interest. Note that the θ_i *could* be all different. In general, however, we typically consider a smaller set of parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$, which relate some function $g(\mu_i)$ of μ_i [which is $E(Y_i)$] to a linear combination of β's via

$$g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}, \tag{18.3.2}$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ is a p by 1 vector of predictor variables observed along with Y_i. The function $g(\cdot)$ is monotone, differentiable, and called the *link function.*

Example: Binomial Distributions, Indices n_i, Parameters p_i

Suppose we have data $Y_i, \mathbf{x}_i'$ from a binomial distribution with index n_i and parameter p_i. The single observation Y_i is of the form r_i/n_i, where r_i is the number of successes in n_i trials, each having probability p_i of success, and

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})' \tag{18.3.3}$$

is a set of observations of p predictor variables associated with Y_i. As we have seen, the binomial distribution is a member of the exponential family. The probability distribution function of n such sets of data for $i = 1, 2, \ldots, n$ is then

$$\exp\left\{\sum Y_i \ln[p_i/(1-p_i)] + \sum n_i \ln(1-p_i) + \sum \ln\binom{n_i}{Y_i}\right\}, \tag{18.3.4}$$

with summations over $i = 1, 2, \ldots, n$. We would typically hope that the variation in the p_i values could be explained in terms of the $\mathbf{x}_i$ values; that is, we would hope that we could find some suitable link function $g(\cdot)$ such that the model

$$g(p_i) = \mathbf{x}_i'\boldsymbol{\beta} \tag{18.3.5}$$

held, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ was a set of parameters, which was the same for each i value. A link function that is often regarded as a sensible one for binomial data is the natural parameter, that is, the *log odds* or *logit* given by

$$g(p_i) = \ln\{p_i/(1 - p_i)\}, \tag{18.3.6}$$

and this would imply that we wish to fit a model in which

$$\ln\{p_i/(1 - p_i)\} = \mathbf{x}_i'\boldsymbol{\beta} = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}. \tag{18.3.7}$$

Note that this is equivalent to saying that p_i is representable as

$$p_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})/\{1 + \exp(\mathbf{x}_i'\boldsymbol{\beta})\}. \tag{18.3.8}$$

[When $\mathbf{x}_i'\boldsymbol{\beta} = \beta_1 + x_{i2}\beta_2$, (18.3.8) is sometimes called the logistic function.]

Estimation Via Maximum Likelihood

To estimate the elements of the parameter vector $\boldsymbol{\beta}$ in (18.3.2), which becomes (18.3.7) for the binomial example, we use the method of maximum likelihood. This requires regarding each of (18.3.1), (18.3.4), and similar appropriate formulas for other distributions, as a function of the parameters in $\boldsymbol{\beta}$, given a specific set of Y_i; we then call such a quantity the *likelihood function* and maximize it with respect to the elements of $\boldsymbol{\beta}$. More specifically, we usually maximize the *log likelihood function* obtained by taking the natural logarithm of formulas like (18.3.4); this leads to the same estimate $\mathbf{b}$ of $\boldsymbol{\beta}$.

The actual maximization procedure, while mathematically interesting, is tedious to work through. [See details in, for example, Dobson (1990).] It results in some normal equations of the general form

$$\mathbf{X}'\mathbf{W}\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{W}\mathbf{z}, \tag{18.3.9}$$

which, we can quickly appreciate, see Eq. (9.2.9), are the normal equations for a generalized least squares model fit. There is an additional catch here, however, in that $\mathbf{W}$ and $\mathbf{z}$ involve $\mathbf{b}$ also. Thus the solution of (18.3.9) usually requires an iterative calculation. An initial value for $\mathbf{b}$, say, $\mathbf{b}_0$, is inserted into $\mathbf{W}$ and $\mathbf{z}$, and (18.3.9) is solved for $\mathbf{b}$ to give solution $\mathbf{b}_1$, say. Then $\mathbf{b}_1$ replaces $\mathbf{b}_0$ and a new solution $\mathbf{b}_2$ is calculated, and so on, until convergence is attained. Standard programs exist for doing these *iteratively reweighted least squares* calculations. They typically permit specification of the distribution (e.g., binomial, Poisson), the link function [e.g., $\ln\{p/(1 - p)\}$ for the binomial, $\ln \lambda$ or λ for the Poisson with mean and variance λ], and the $\mathbf{x}_i$ values, whereupon the solution $\hat{\boldsymbol{\beta}}$ is given. Perhaps the GLIM program (produced by Numerical Algorithms Group, Oxford, U.K.) is the best known program in some parts of the world, but versions are also available in BMDP, GENSTAT, SAS, SPSS, SPLUS, and so on.

Deviance

Comparisons between nested models in ordinary least squares are usually made via extra sums of squares F-tests, which essentially compare residual sums of squares. In

the case of generalized linear models, the residual sum of squares role is played by the deviance

$$-2\{l(\hat{\boldsymbol{\beta}}) - l(\hat{\boldsymbol{\theta}})\}, \tag{18.3.10}$$

where $l(\cdot)$ is the log likelihood function, $l(\hat{\boldsymbol{\beta}})$ is its value for the maximum likelihood value $\hat{\boldsymbol{\beta}}$, and $l(\hat{\boldsymbol{\theta}})$ is its value when $\hat{\theta}_i = Y_i$ is used, that is, when each θ_i in (18.3.1) is estimated by its corresponding Y_i value. Thus we can rewrite (18.3.10) as

$$\text{Deviance} = -2\{l(\hat{\boldsymbol{\beta}}) - l(\mathbf{Y})\}. \tag{18.3.11}$$

The deviance is asymptotically distributed as a χ^2 variable with $(n - p)$ degrees of freedom when the model is correct, and so large deviance values are taken to be indicative of a bad fit for the $\mathbf{x}'\boldsymbol{\beta}$ model. We compare the deviance value with a selected percentage point of the appropriate χ^2 distribution whose df correspond to the number of parameters in $\boldsymbol{\beta}$. Differences between deviances can also be used in a way parallel to the use of extra sums of squares.

It should be noted that the χ^2 approximation can be inaccurate for small samples, for example, if the Y_i are Poisson variables with means close to 1, or binomial. See McCullagh and Nelder (1989, p. 121).

18.4. PERFORMING THE CALCULATIONS: AN EXAMPLE

Table 18.1 shows a set of data on reported occurrences of a communicable disease in two areas of the country at ten 2 month intervals, 2, 4, . . . , 20. There are 20 data points, so that $i = 1, 2, \ldots, 20$ below.

We assume that the occurrences Y_i are Poisson variables with $E(Y_i) = \mu_i$ and that

$$g(\mu_i) = \ln \mu_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \tag{18.4.1}$$

is the model under consideration. We take $X_{1i} = \ln(2i)$ corresponding to ln(month indicator) and $X_{2i} = 0$ for the observations from area A and $X_{2i} = 1$ for observations in area B. Other choices of X_{1i} and X_{2i} are possible. Also, another link function could be used; the one chosen, $g(\mu_i) = \ln \mu_i$, is the "natural" link, corresponding to the "natural parameter."

TABLE 18.1. Reported Occurrences of a Communicable Disease in Two Areas of the Country

	Occurrences in		
	Area A	Area B	Month, e^{X_1}
	8	14	2
	8	19	4
	10	16	6
	11	21	8
	14	23	10
	17	27	12
	13	28	14
	15	29	16
	17	33	18
	15	31	20
Dummy X_2	0	1	

To fit (18.4.1) in GLIM, one needes these instructions:

```
glim
  ?   $units 20$            (the ? prompt appears automatically. The instruction
                            indicated 20 observations are coming)
  ?   $data y x w $         (Names the variables as y, x, and w.)
      $read                 (No $ closure here.)
      $REA?  8 0.693 0      ($REA? appears automatically on each line. Data are
                            entered line by line)
      $REA?  8 1.386 0
      $REA? 10 1.792 0
      $REA? 11 2.079 0
      $REA? 14 2.303 0
      $REA? 17 2.485 0
      $REA? 13 2.639 0
      $REA? 15 2.773 0
      $REA? 17 2.890 0
      $REA? 15 2.996 0
      $REA? 14 0.693 1
      $REA? 19 1.386 1
      $REA? 16 1.792 1
      $REA? 21 2.079 1
      $REA? 23 2.303 1
      $REA? 27 2.485 1
      $REA? 28 2.639 1
      $REA? 29 2.773 1
      $REA? 33 2.890 1
      $REA? 31 2.996 1      (Last data group. The ? prompt now reappears.)
  ?   $yvar y $             (Declares y as the response.)
  ?   $error poisson $      (Declares the error distribution.)
  ?   $link L $             (Can be omitted for the natural link, here g(μ_i) =
                            ln μ_i)[1]
  ?   $fit x + w $          (Fit additive terms in x and w.)
scaled deviance = 3.1258 at cycle 3 (Printed out.)
d.f. = 17
  ?   $display e $          (Gives printout as follows.)
          estimate s.e.      parameter
  1       1.684    0.2202        1 (Estimate of β0)
  2       0.3784   0.08524       X (Estimate of β1)
  3       0.6328   0.1094        W (Estimate of β2)
scale parameter taken as 1.000
                                (By retyping the $link, $fit, and $display e
                                instructions, other fits can be made using the
                                same data set.)
```

TABLE 18.2. Some Reasonable Choices of Link Functions

Symbol to be Entered	Name of Link	Form of $g(u)$	Used for
C	Complementary log-log	$\ln\{-\ln(1-u)\}$	Binomial
E	Exponential	e^u	
G	Logit	$\ln\{u/(1-u)\}$	Binomial
I	Identity	u	Normal
L	Logarithm	$\ln u$	Poisson, Gamma
P	Probit	$\Phi^{-1}\{(u-\mu)/\sigma\}$	Binomial
R	Reciprocal	u^{-1}	Gamma, Poisson
S	Square root	$u^{1/2}$	Poisson

[1]Available link choices are given in Table 18.2. It is not always clear in practice what link should be employed and data are often analyzed by comparing several alternative choices.

The fitted model is

$$\widehat{\ln Y} = 1.684 + 0.3784X_1 + 0.6328X_2 \tag{18.4.2}$$

Compared with $\chi^2(3, 0.95) = 7.81$, the scaled deviance of 3.13 is nonsignificant, indicating a reasonable fit.

18.5. FURTHER READING

There are several excellent books in this area. Dobson (1990) provides a good basic introduction. McCullagh and Nelder (1989) is the technical bible. Aitken et al. (1989) and Healy (1988) provide good practical examples. Francis, Green, and Payne (1993) is about the GLIM program. The paper by Hilbe (1994) reviews seven software packages. These are the places to start, the individual selection depending on your objective. Other useful books are Collett (1991); Cox and Snell (1989); Fahrmeir and Tutz (1994); Freund and Littell (1991); Green and Silverman (1994); Hosmer and Lemeshow (1989); and Littell, Freund, and Spector (1991). For comments on interpretation of fitted models, see the note by Chappell (1994).

EXERCISE FOR CHAPTER 18

A. Suppose we have n observations of variables $X_1, X_2, \ldots, X_k, Y$, where the X's are predictors and Y is a response variable. If the Y's are binomial ratios $Y_i = r_i/m$, say, where m is constant, what types of analyses are feasible?