CHAPTER 11

Models with additional non-linear parameters

11.1 Introduction

The word 'non-linear' in the chapter title is used in a specialized sense because, apart from the subset of classical linear models, all generalized linear models are in a strict sense non-linear. However, their non-linearity is limited in that it enters only through the variance function and the link function, the linearity of terms contributing to the linear predictor being preserved.

So far we have assumed that the variance is a known function of the mean, except possibly for a multiplicative dispersion parameter, and that the link function is also known. In this chapter we describe a number of models in which unknown parameters enter either the variance or the link function or both. In addition we also consider the use of terms in the linear predictor of an intrinsically non-linear type. One example of such a model occurred in section 8.4.4, where

$$\log \mu = \beta_0 + \beta_1 T + \beta_{-1}/(T - \delta)$$

was used as a model for the mean. Parameters such as δ are called non-linear in this specialized sense.

Intrinsically non-linear parameters complicate the fitting algorithm either by introducing an extra level of iteration or by introducing covariates that change at each iteration. Either of these effects may render convergence of the iterative process much less certain, and may also require starting values to be given for the non-linear parameters. In addition asymptotic covariances for the linear terms may be produced by the fitting algorithm conditional on fixed values of the non-linear parameters, and so need adjustment if uncertainties in the non-linear parameters are to be allowed for.

11.2 Parameters in the variance function

In the generalized linear models described in Chapters 3–8, five distributions for error were used. Two of these, the Normal and gamma, contain dispersion parameters explicitly. The discrete distributions in their standard forms do not contain such parameters, although here quasi-likelihood arguments extended the analysis to include an unknown dispersion parameter also. Provided that the dispersion parameter is constant, its value is not required in the solution of the likelihood equations for $\hat{\boldsymbol{\beta}}$. In that sense σ^2 plays a special role and its estimation is treated separately from the regression parameters.

The negative binomial distribution provides an example of a variance function containing an unknown parameter that is not a dispersion parameter. The distribution, which is discrete, can be written in the form

$$\text{pr}(Y = y; \alpha, k) = \frac{(y + k - 1)!}{y!\,(k-1)!} \frac{\alpha^y}{(1+\alpha)^{y+k}}; \qquad y = 0, 1, 2 \ldots$$

This may be contrasted with the expression in section 6.2.3 in which a different parameterization is used. In the above parameterization, the mean and variance are given by

$$E(Y) = \mu = k\alpha,$$
$$\text{var}(Y) = k\alpha + k\alpha^2 = \mu + \mu^2/k.$$

The log likelihood can be written in the form

$$l = y \log\{\alpha/(1+\alpha)\} - k \log(1+\alpha) + (\text{function of } y, k),$$

which, for fixed k, has the form of a generalized linear model with canonical link

$$\eta = \log\left(\frac{\alpha}{1+\alpha}\right) = \log\left(\frac{\mu}{\mu + k}\right),$$

and variance function

$$V = \mu + \mu^2/k.$$

The term μ can be thought of as the Poisson variance function and μ^2/k as the extra component arising from mixing the Poisson distribution with a gamma distribution for the mean to obtain the negative binomial.

Ordinarily k is not known *a priori*, and is clearly not a dispersion parameter. Estimates of k for single samples and for several samples have been discussed by Anscombe (1949), but we require an estimator for arbitrarily structured data. The maximum-likelihood estimate requires the solution of a non-linear equation involving the digamma function. Alternative estimators are those that make the mean deviance equal to unity or the Pearson X^2 statistic equal to its expectation.

Little use seems to have been made of the negative binomial distribution in applications; in particular the use of the canonical link is problematical because it makes the linear predictor a function of a parameter of the variance function. Note that if μ varies from observation to observation the above formulation assumes that, of the two parameters α and k in the mixing distribution, only α changes with k remaining fixed. See Manton *et al.* (1981) for an analysis of survey data in which both α and k are made to depend upon the classifying covariates; such a model, though of undoubted interest, lies outside the present framework.

For another example of additional parameters in the variance function, consider data that are to be modelled with gamma errors, but which have been collected with an absolute measurement (rounding) error, rather than with the desirable proportional error. With proportional rounding error or the retention of a fixed number of digits, the error variance retains the form $V = \sigma^2\mu^2$: with absolute rounding error, or the retention of a fixed number of decimal places, the variance function takes the form $V = \tau^2+\sigma^2\mu^2$. The first term arises from the constant rounding error and the second from the assumed underlying gamma errors. The effect of this modified variance function is to reduce relatively the weight given to small observations. The quasi-likelihood model with this variance function would require the estimation of σ^2/τ^2 in the same way that k must be estimated for models using the negative binomial distribution.

Note that rounding from Z to Y has the effect of increasing the variance. In fact the rounding error, although numerically a

deterministic function of Z, is essentially statistically independent of Z, and not of Y. See Exercise 11.1.

11.3 Parameters in the link function

While link functions in generalized linear models are usually assumed known, it may be useful on occasion to assume that the link comes from a class of functions, members of the class being indexed by one or more unknown parameters. The goodness of fit expressed as a function of these parameters can then be inspected to see what range of parameter values is consistent with the data. If a particular value is of interest we can perform a goodness-of-link test (Pregibon, 1980) by comparing the deviance for that value with the deviance for the best-fitting value, or by using a score test.

11.3.1 *One link parameter*

A commonly considered class of link functions is that given by the power function, either in the form

$$\eta = \begin{cases} \mu^\lambda & \text{for } \lambda \neq 0, \\ \log \mu & \text{for } \lambda = 0, \end{cases}$$

or, in the form having continuity at $\lambda = 0$,

$$\eta = \frac{\mu^\lambda - 1}{\lambda}.$$

Exploration of this class of functions, used as transformations of the data rather than of the fitted values, was considered by Box and Cox (1964). For any fixed value of λ the model can be fitted with that power link function, and the deviance obtained in the usual way. When this is done for a range of λ-values, the deviances may be plotted against λ to display the range of λ-values that are most consistent with the observed data (and the model formula used).

If we wish to optimize over λ we can adopt the linearizing strategy proposed by Pregibon (1980), whereby we expand the link

function in a Taylor series about a fixed λ_0 and take only the linear term. Thus for the power family we have

$$\begin{aligned} g(\mu;\lambda) = \mu^{\lambda} &\simeq g(\mu;\lambda_0) + (\lambda - \lambda_0) g'_{\lambda}(\mu;\lambda_0) \\ &= \mu^{\lambda_0} + (\lambda - \lambda_0)\mu^{\lambda_0} \log \mu, \end{aligned} \tag{11.1}$$

so that we can approximate the correct link function $\eta = \mu^{\lambda}$ by

$$\begin{aligned} \eta_0 = \mu^{\lambda_0} &= \mu^{\lambda} - (\lambda - \lambda_0)\mu^{\lambda_0} \log \mu \\ &= \sum \beta_j x_j - (\lambda - \lambda_0)\mu^{\lambda_0} \log \mu. \end{aligned}$$

Given a first estimate λ_0 of λ, with corresponding fitted values $\hat{\mu}_0$, we then extend the linear predictor to include the covariate $-\hat{\mu}_0^{\lambda_0} \log \hat{\mu}_0$. Its parameter estimate gives the first-order adjustment to λ_0. The reduction in deviance from its inclusion gives a test of whether λ_0 is an acceptable value for λ. To obtain the maximum-likelihood estimate for λ we repeat the above process forming a new adjusted value for λ at each stage. Convergence is not guaranteed, however, and requires that λ_0, our starting value, be sufficiently close to $\hat{\lambda}$ for the linear expansion (11.1) to be adequate. To obtain convergence, an inner iteration, whereby the extra covariate's values are refined for fixed λ_0, may be needed. Pregibon (1980) comments 'that the method is likely to be most useful for determining if a reasonable fit can be improved, rather than for the somewhat more optimistic goal of correcting a hopeless situation'.

Figure 11.1 shows the effect with the car-insurance data of changing the link by varying λ in the power family $\eta = \mu^{\lambda}$. The linear predictor contains the main effects only and the variance function is taken as $V(\mu) \propto \mu^2$. The minimum deviance of 124.51 occurs near $\lambda = -1$, corresponding to the reciprocal link originally chosen for the analysis, though the 95% limits

$$\left\{\lambda : \text{dev}(\lambda) - 124.51 < \frac{124.51}{108} \times 3.93\right\}$$

show an appreciable range of compatible values for λ, including zero, corresponding to the log link.

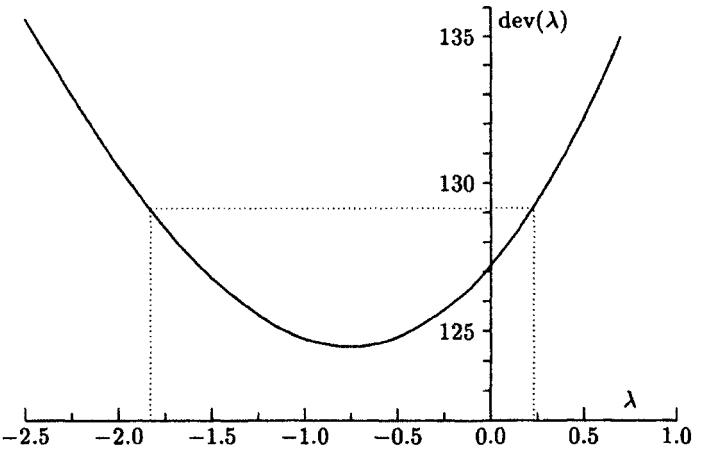

Fig. 11.1 *Car insurance data: deviance for varying λ in the link function $\eta = \mu^\lambda$, with nominal 95% confidence limits. Linear predictor includes main effects, and variance function $\propto \mu^2$.*

11.3.2 *More than one link parameter*

The above method extends in principle to more than one parameter in the link function. For each parameter λ, we add an extra covariate

$$-\left(\frac{\partial g}{\partial \lambda}\right)_{\lambda=\lambda_0}$$

to the model matrix and its parameter estimate gives the first-order adjustment to the starting value λ_0. Pregibon (1980) discusses two examples with two parameters; the first is given by

$$g(\mu;\alpha,\lambda) = \{(\mu+\alpha)^\lambda - 1\}/\lambda,$$

i.e. a shifted form of the power family, indexed by λ, but with an added unknown origin α. Note that

$$g(\mu;1,1) = \mu,$$

so that the identity link is a member of the family.

The second of Pregibon's examples is useful for models based on tolerance distributions, such as probit analysis. The generalized link function is given by

$$g(\mu;\lambda,\delta) = \frac{\pi^{\lambda-\delta}-1}{\lambda-\delta} - \frac{(1-\pi)^{\lambda+\delta}-1}{\lambda+\delta}$$

when π is the proportion responding, i.e. μ/m. The family contains the logit link as the limiting form

$$\lim_{\lambda,\delta\to 0} g(\mu;\lambda,\delta).$$

The one-parameter link family for binomial data,

$$g(\mu;\lambda) = \log\left[\{(1/(1-\pi))^{\lambda} - 1\}/\lambda\right],$$

contains both the logistic ($\lambda = 1$) and complementary log-log link ($\lambda \to 0$) as special cases. This family may be used to assess the adequacy of an assumed linear logistic model against alternatives in the direction of the complementary log-log link.

11.3.3 *Transformation of data vs transformation of fitted values*

Transformations of fitted values through link functions (whether with or without unknown parameters) must be distinguished from transformations of the data values. The latter are fully discussed in Box and Cox (1964), who deal in particular with the power family. In using a function $g(y)$, rather than y, in the analysis we usually seek a transformation that yields simultaneously additive effects in the systematic part of the model and constancy of variance for the random part. Such a search may be successful; see, for example, the data set on survival times, given by Box and Cox, where a reciprocal transformation of the data allowed a linear regression model to be applied. However, there is no guarantee that both properties will result from the same transformation. Thus Nelder and Wedderburn (1972) in their reanalysis of the tuberculin-test data of Fisher (1949) (see the example in section 6.3.1) show that while a square-root transformation produces desirable error properties, a log transformation is required for additivity of effects. It is an advantage of generalized linear models over data transformation methods that the transformation to produce additivity can be made through the link function quite independently of any transformation of the data to produce approximate Normality or constancy of variance. Indeed the latter is itself often rendered unnecessary by the possibility of using a variance function other than a constant. Thus with the Fisher data mentioned above analysis of

$$Y \text{ with variance function } V \propto \mu$$

and of

$$Y^{1/2} \text{ with variance function } V = \text{constant},$$

using a log link for each, produce effectively identical results (Baker and Nelder, 1978, Appendix D).

11.4 Non-linear parameters in the covariates

As remarked in section 3.3.1 a function of x, such as e^{kx}, is an acceptable covariate in a linear predictor, provided that k is known; we simply use the values of e^{kx} in place of x in the model matrix. However, if k is to be estimated from the data, then non-linearity arises. Box and Tidwell (1962) describe a fitting technique by linearization, which follows closely that for non-linear parameters in the link function described above. If $g(x;\theta)$ is the covariate to be used, with θ unknown, we expand about an initial value θ_0 to give the linear approximation

$$g(x;\theta) \simeq g(x;\theta_0) + (\theta - \theta_0)[\partial g/\partial\theta]_{\theta=\theta_0}.$$

Thus if a non-linear term in the linear predictor is given by

$$\beta g(x;\theta),$$

we replace it by two linear terms

$$\beta u + \gamma v,$$

where

$$u = g(x;\theta_0), \qquad v = [\partial g/\partial\theta]_{\theta=\theta_0} \quad \text{and} \quad \gamma = \beta(\theta - \theta_0).$$

An extra level of iteration is again required, and after fitting a model including u and v as covariates we obtain

$$\theta_1 = \theta_0 + \hat{\gamma}/\hat{\beta}$$

as the improved estimate, and iterate. Convergence is not guaranteed for starting values arbitrarily far from the solution. If the process does converge then the presence of the extra term γv ensures that the asymptotic covariances produced for the remaining

parameters are correctly adjusted for the fitting of θ. If we wish to obtain the asymptotic variance of $\hat{\theta}$ directly, we need a final iteration with $\hat{\beta}v$ in the place of v; the components of $(\mathbf{X}^T\mathbf{W}\mathbf{X})^-$ corresponding to that covariate then give the approximate variance of $\hat{\theta}$ and its covariances with the other parameters.

While this technique is undoubtedly useful, and indeed probably under-used, it is usually unwise to try to include more than a very few non-linear parameters in this way, especially when the other covariates are themselves appreciably correlated in the data set. It will usually be found that estimates of the non-linear parameters have large sampling errors, and are highly correlated with the linear parameters and perhaps with each other. This is especially likely to be so in models where the systematic part consists of sums of exponentials of the form

$$\beta_0 + \beta_1 e^{k_1 x_1} + \beta_2 e^{k_2 x_2},$$

with the ks in the exponents requiring to be estimated as well as the βs.

One example where non-linear parameters arise in a fairly natural way concerns models for the joint action of a mixture of drugs (see the example in section 11.5.3). Here, apart from a single parameter that enters the model non-linearly, the model is of the generalized linear type with covariate $\log(x_1 + \theta x_2)$, where x_1 and x_2 are the amounts of the two drugs in the mixture. One method of analysis is to use the linearizing technique described above. Alternatively, following Darby and Ellis (1976), we may maximize the likelihood for various values of θ and plot the residual sum of squares $\mathrm{RSS}(\theta)$ against θ, thereby obtaining a profile deviance curve similar to that shown in Fig. 11.1. The minimum, usually unique, gives $\hat{\theta}$ and the residual mean deviance $s^2 = D(\hat{\theta})/(n-p)$, where p is the total number of parameters, including θ. Approximate confidence limits for θ can be found from

$$\{\theta : \mathrm{RSS}(\theta) - \mathrm{RSS}(\hat{\theta}) < s^2 F^*_{1,n-p,\alpha}\},$$

where $F^*_{1,n-p,\alpha}$ is the upper $100(1-\alpha)$ percentage point of the F-distribution on 1 and $n-p$ degrees of freedom.

Unfortunately this method of analysis does not allow the covariances of the parameter estimates, allowing for the uncertainty in θ, to be calculated easily. If these are required, the linearization technique should be used (see example in section 11.5.3).

11.5 Examples

11.5.1 *The effects of fertilizers on coastal Bermuda grass*

Welch *et al.* (1963) published the results of a 4^3 factorial experiment with the three major plant nutrients, nitrogen (N), phosphorus (P) and potassium (K), on the yield of coastal Bermuda grass. The experiment was performed to produce a response surface for the effects of the three nutrients, so that an optimal dressing could be predicted. The four levels for the three factors (all in lb/acre) were:

Levels	1	2	3	4
N	0	100	200	400
P	0	22	44	88
K	0	42	84	168

The grass was cut about every five weeks and oven-dried. The yields (in tons/acre) averaged over the three years 1955–57 and three replicates are shown in Table 11.1 with the factor levels coded 0, 1, 2, 3. Inspection of the data shows a hyperbolic type of response to the nutrients but the yield for the (0,0,0) plot shows that it will be necessary to allow for the nutrients already present in the soil if inverse polynomials (Section 7.3.3) are to be used to describe the response surface. We consider, therefore, an inverse linear response surface with

$$1/\mu = \eta = \beta_0 + \beta_1 u_1 + \beta_2 u_2 + \beta_3 u_3,$$

where $u_i = 1/(x_i + \alpha_i)$, $i = 1, 2, 3$. Here $x_i (i = 1, 2, 3)$ are the applied amounts of N, P and K respectively, while α_i are the (unknown) amounts in the soil. If we assume for the moment that the proportional standard deviation of the yield is constant, we have a model with gamma errors and the reciprocal (canonical) link, but with three non-linear parameters α_1, α_2 and α_3 to be estimated. The linearizing technique of section 11.4 leads to our adding the extra covariates $v_i = \partial u_i/\partial \alpha_i = -u_i^2$ to the model, giving the corrections

$$\delta a_i = c_i/b_i; \qquad i = 1, 2, 3,$$

Table 11.1 *Yields of coastal Bermuda grass as affected by* N, P *and* K

Nitrogen (N)	*Phosphorus* (P)	*Potassium* (K) 0	1	2	3
0	0	1.98	2.13	2.19	1.97
0	1	2.38	2.24	2.10	2.60
0	2	2.18	2.56	2.22	2.47
0	3	2.22	2.47	2.94	2.48
1	0	3.88	3.91	3.66	4.07
1	1	4.35	4.59	4.47	4.55
1	2	4.14	4.36	4.55	4.35
1	3	4.26	4.72	4.83	4.85
2	0	4.40	4.91	5.10	5.23
2	1	5.01	5.64	5.68	5.60
2	2	4.77	5.69	5.80	6.07
2	3	5.17	5.45	5.85	6.43
3	0	4.43	5.31	5.15	5.87
3	1	4.95	6.27	6.49	6.54
3	2	5.22	6.27	6.35	6.72
3	3	5.66	6.24	7.11	7.32

Data from Welch *et al.* (1963).

to the current estimates of α_i in each iteration, where a_i is the estimate of α_i, c_i is the coefficient of v_i and b_i that of u_i. Starting values are required for a_i and these can be obtained by taking reciprocals of the data, forming N, P and K margins and plotting these against u_i for various trial values of α_i. The following suggested values are obtained

$$a_1 = 40, \qquad a_2 = 22, \qquad a_3 = 32.$$

Six iterations refine these to

$$a_1 = 44.60, \qquad a_2 = 15.56, \qquad a_3 = 32.39,$$

with a final deviance of 0.1965 on 57d.f., corresponding to a percentage standard deviation per observation of 5.9. The X^2 statistic is 0.1986, trivially different from the deviance, as is typical when the coefficient of variation is small. A final iteration with v_i replaced by $b_i v_i$ enables us to obtain the asymptotic standard errors

of the a_i directly. The parameter estimates with their standard errors, are given by:

b_0	0.09746 ± 0.00963
b_1	13.5 ± 1.350
b_2	0.7007 ± 0.457
b_3	1.336 ± 0.956
a_1	44.6 ± 4.18
a_2	15.6 ± 8.44
a_3	32.4 ± 19.1

These agree closely with those given by Nelder (1966) who used an approximate non-iterative method.

The correlation matrix shows high correlations between the a_i and b_i. These are respectively

$$0.9702, \qquad 0.9850, \qquad 0.9849,$$

and reflect the fact that if the a_i are taken as known the standard errors of the b_i are reduced by factors of from 4 to 6. Note too the large standard errors for a_2 and a_3, which do not exclude impossible negative values; the assumptions of (asymptotic) Normality must be treated cautiously here.

The inverse linearity of the response may be tested by including the inverse quadratic terms $(x_i + \alpha_i)$ in the model. This gives a deviance of 0.1938 with 54d.f., a negligible reduction. The Pearson residuals $(y - \hat{\mu})/\hat{\mu}$ show one possible outlier, the yield 2.94 for levels $(0, 3, 2)$. The fitted value is 2.43, so that the possibility of a transposition at some stage from 2.49 to 2.94 might be investigated. Omission of this point does not change the fit greatly, the largest effect being on b_2. A plot of the residuals against fitted values does not contradict the assumption of gamma errors.

As shown in Nelder (1966), the quadratic polynomial with 10 parameters fits less well than the inverse linear surface with unknown origins, which has seven parameters. The latter is also additive for the three nutrients whereas the quadratic polynomial requires all the two-factor interaction terms for an adequate fit.

11.5.2 *Assay of an insecticide with a synergist*

The data for this example, shown in Table 11.2, are taken from an unpublished paper by Morse, McKinlay and Spurr on the estimation of lowest-cost mixtures of insecticides and synergists. They relate to assays on a grasshopper *Melanopus sanguinipes* (F.) with the insecticide carbofuran and the synergist piperonyl butoxide (PB), which enhances the toxicity of the insecticide. The first model to be tried is of a type suggested by Hewlett (1969) and having the form of a logit link and binomial error with 'linear' predictor given by

$$\eta = \alpha + \beta_1 x_1 + \frac{\beta_2 x_2}{\delta + x_2},$$

where x_1 is the log dose of insecticide and x_2 is the dose of the synergist PB. The effect of the synergist is thus modelled as affecting the intercept by adding a hyperbolic term tending to β_2 for large x_2. The slope of β_1 is assumed unaffected by the amount of PB. If δ were known we could set $u = x_2/(\delta + x_2)$ and a generalized linear model would result. To estimate δ we set up u for some starting value of δ and include the derivative $\partial u/\partial\delta = -u^2/x_2$ as a further covariate. Starting with $\delta = 1$ the standard process converges in four iterations to $\hat{\delta} = 1.763$ and a deviance of 53.34 with 11 d.f. The fit is poor with a deviance nearly five times the base level of 11.

Inspection of the residuals shows that the major part of the discrepancy comes from the low doses of insecticide where the fitted kills are all considerably greater than those measured. The alternative links, probit and complementary log-log, give very similar results, suggesting that the log dose is not a satisfactory scale for the insecticide. The low kills for the low doses suggest that there may be a threshold value for the insecticide, and we can test this by putting a second non-linear parameter θ in the model to represent the threshold. The model now takes the form

$$\eta = \alpha + \beta_1 \log(z - \theta) + \beta_2 x_2/(\delta + x_2),$$

where z is the dose of insecticide. Given current values, θ_0 and δ_0, of θ and δ, the linearized form is given by

$$\eta = \alpha + \beta_1 \log(z - \theta_0) - \gamma_1 \frac{1}{z - \theta_0} + \beta_2 \frac{x_2}{\delta_0 + x_2} - \gamma_2 \frac{x_2}{(\delta_0 + x_2)^2}.$$

Table 11.2 *Data from assay on insecticide and synergist*

Number killed, y	*Sample size, m*	*Dose of insecticide*	*Dose of synergist*
7	100	4	0
59	200	5	0
115	300	8	0
149	300	10	0
178	300	15	0
229	300	20	0
5	100	2	3.9
43	100	5	3.9
76	100	10	3.9
4	100	2	19.5
57	100	5	19.5
83	100	10	19.5
6	100	2	39.0
57	100	5	39.0
84	100	10	39.0

Data courtesy of Drs Morse, McKinlay and Spurr of Agriculture Canada.

With starting values $\delta_0 = 1.76$ from the first model and $\theta_0 = 1.5$ the estimation process again converges quickly to give estimates $\hat{\theta} = 1.67$, $\hat{\delta} = 2.06$, with a deviance of 18.70 with 10d.f., clearly a great improvement on the first model. A test of variation in the slope β_1 with level of x_2 now gives no significant reduction in the deviance, whereas with the first model the deviance was nearly halved by allowing the slope to vary. A final iteration multiplying the two derivative covariates by b_1 and b_2 and using the mean deviance as a heterogeneity factor gives the estimates, standard errors and correlations shown in Table 11.3.

Table 11.3 *Results of the analysis of the insecticide-synergist assay*

Parameter	*Estimate*	*SE*	*Correlations*			
α	−2.896	0.340				
β_1	1.345	0.143	−0.97			
θ	1.674	0.154	0.78	−0.77		
β_2	1.708	0.241	−0.31	0.26	−0.03	
δ	2.061	1.49	0.09	−0.07	0.06	0.61

Note that the two non-linear parameters are estimated almost

independently (correlation −0.07), and that δ is ill-determined. In particular δ must be positive, so that the standard error is suspect; a plot of the deviance against fixed values of δ near 2 shows a curve that is non-quadratic. The use of a square-root transformation for δ is a great improvement and indicates that confidence limits for δ should be calculated on the square-root scale. The fitting of a separate intercept for each level of synergist in place of the term in $x_2/(\delta+x_2)$ gives a trivial reduction in deviance, indicating that the hyperbolic form is satisfactory. There remain two large residuals, for units 1 and 2, of opposite sign, whose removal from the fit reduces the deviance from 18.70 to 5.69; however, there seems to be no good reason to reject them. The relevant values are:

Unit	y	m	$\hat{\mu}$	r_P
1	7	100	14.67	−2.17
2	59	200	43.51	2.66

This analysis indicates that the effect of doubling the dose of insecticide is to increase the odds of a kill by an estimated factor of $2^{\beta_1} = 2.54$. Since there is apparently no interaction between the insecticide dose and the synergist dose, this odds factor of 2.54 applies at any fixed dose of the synergist. The synergist exhibits decreasing returns of scale, a large dose increasing the odds of a kill by the factor $\exp(\hat{\beta}_2) = 5.52$. A moderate dose of 19.5 units increases the odds of a kill by an estimated $\exp(1.54) = 4.69$. These factors apply at any fixed dose of insecticide.

11.5.3 *Mixtures of drugs*

If two drugs provoke similar responses, a mixture of both may exhibit either additive or synergistic effects. If the effect is additive one drug can be replaced by a suitable proportion of the other to give the same response. With positive synergism the joint effect is greater than the sum of the effects of the two drugs administered separately: negative synergism is the term used to describe the opposite effect. In an experiment to test for such synergism, Darby and Ellis (1976) quote the data of Table 11.4, where the response y is the conversion of (3-3H)glucose to toluene-extractable lipids in isolated rat fat cells, and the two drugs are insulin in two forms, (1) standard and (2) A1-B29 suberoyl insulin. These are given in

seven different mixtures, each at two total doses; there are four replicate readings for the 14 treatments. Darby and Ellis proposed the model

$$E(Y_{ijk}) = \alpha + \beta \log(x_{1ij} + \theta x_{2ij}) \qquad (11.2)$$

with constant-variance errors. Here i indexes the mixtures, j the total dose and k the replicates, while x_{1ij} and x_{2ij} are the amounts of the two drugs given for mixture i with dose j.

Table 11.4 *Results of insulin assay*

Mixture	*Ratio of insulin to A1-B29 suberoyl insulin*	*Total dose* (pmol l^{-1})	*Responses for four replicates*			
1	1:0	20.9	14.0	14.4	14.3	15.2
		41.9	24.6	22.4	22.4	26.7
2	1:1.85	52.9	11.7	15.0	12.9	8.3
		106	20.6	18.0	19.6	20.5
3	1:5.56	101	10.6	13.9	11.5	15.5
		202	23.4	19.6	20.0	17.8
4	1:16.7	181	13.8	12.6	12.3	14.0
		362	15.8	17.4	18.0	17.0
5	1:50.0	261	8.5	9.0	13.4	13.5
		522	20.6	17.5	17.9	16.8
6	1:150	309	12.7	9.5	12.1	8.9
		617	18.6	20.0	19.0	21.1
7	0:1	340	12.3	15.0	10.1	8.8
		681	20.9	17.1	17.2	17.4

Data from Darby and Ellis (1976).

Here θ is the non-linear parameter and we can fit the model by linearizing it, using the two covariates

$$u = \log(x_1 + \theta x_2), \qquad v = \frac{\partial u}{\partial \theta} = \frac{x_2}{x_1 + \theta x_2};$$

we fit $\alpha + \beta u + \gamma v$ for some value θ_0 and θ is then updated by $\theta_1 = \theta_0 + \gamma/\beta$. The estimate obtained after iteration is $\hat{\theta} = 0.0461 \pm 0.0036$, with a corresponding deviance (residual sum of squares) of 244.0 with 53d.f. Comparing this fit with the

replicate error of 154.8 with 42d.f., we find an F-value for residual treatment variation of $F(11, 42) = 2.20$, just beyond the 5% point. Darby and Ellis are concerned to compare this model with one in which θ is allowed to vary with the mixture, and in doing so to provide possible evidence of synergism or antagonism. Such a model, which requires a set of partial-derivative covariates, one for each mixture, reduces the deviance to 194.6 with 48d.f., giving a residual $F(6, 42) = 1.80$. While no longer significant at 5%, there is still a noticeable lack of fit, which investigation shows to be almost entirely concentrated in the first 'mixture', that for which x_2 is zero. Without this we find $\hat{\theta} = 0.0524$ and a deviance of 191.2 with 51d.f., giving a residual treatment deviance of 36.40 with 9d.f., so that the mean deviance is now close to the replicate error mean square of 3.686.

On this interpretation the interaction between the two drugs is expressed by saying that one unit of drug 2 is equivalent to 0.052 units of drug 1 in mixtures containing ratios by weight of 1:1.85 or more. In the absence of drug 2 the actual response to drug 1 is larger than predicted from the model (11.2), the predicted values for 'mixture' 1 (omitting it from the fit) being 12.9 and 19.8 as against 14.5 and 24.0 actually measured. Plots of residuals from the restricted model show no pattern when plotted against fitted values or mixture number, and there are no obvious outliers. The final analysis of variance is given in Table 11.5; it should be noted that this analysis differs somewhat from that of Darby and Ellis.

Table 11.5 *Analysis of variance for insulin assay*

	s.s.	*d.f.*	*m.s.*
Treatments	906.6	13	
Model (11.2)	817.4	2	408.7
Separate θs	49.4	5	9.88
Residual	39.8	6	6.63
Alternative subdivision			
Model (11.2)	817.4	2	
Removal of mixture 1	52.8	2	26.4
Residual	36.4	9	4.04
Within treatments	154.8	42	3.686

11.6 Bibliographic notes

The use of linearization methods for the optimization of non-linear functions has a long history going back to Gauss (1826), who gave a non-linear surveying problem to illustrate the technique of least squares. Its use in testing possible non-linear transformation of covariates was stressed by Box and Tidwell (1962) in the context of regression analysis.

Pregibon (1980) introduced goodness-of-link tests involving the estimation of parameters in the link function. Nelder and Pregibon (1987) described methods for the joint estimation of parameters in both link and variance functions.

11.7 Further results and exercises 11

11.1 *Rounding errors*: Let Z be a continuous random variable whose density $f(z)$ has derivatives satisfying the integrability condition

$$\int_{-\infty}^{\infty} f^{(\nu)}(z)\,dz < \infty,$$

where $\nu \geq 2$ is an even integer. Suppose that the recorded value is

$$Y = \epsilon\langle Z/\epsilon\rangle,$$

where $\langle x\rangle$ is the nearest integer to x. If $\epsilon = 10^{-d}$ then Y is Z rounded to d decimal places. We can thus write

$$Z = Y + \epsilon R,$$

where $-\frac{1}{2} < R \leq \frac{1}{2}$ is the normalized rounding error.

Using the Euler-Maclaurin summation formula (Bhattacharya and Rao, 1976, p.256; Jeffreys and Jeffreys, 1956, p.280) or otherwise, show that the joint cumulant generating function of (Z, R) satisfies

$$K_{Z,R}(\xi_1, \xi_2) = K_Z(\xi_1) + K_R(\xi_2) + O(\epsilon^\nu)$$

$$K_R(\xi) = \log\left(\frac{\sinh(\frac{1}{2}\xi)}{\frac{1}{2}\xi}\right) + O(\epsilon^\nu)$$

for small ϵ. Hence deduce that to a close approximation for small ϵ, R is uniformly distributed on $(-\frac{1}{2}, \frac{1}{2}]$, and that the cumulants of the rounded random variable Y are given by

$$\kappa_r(Y) \simeq \kappa_r(Z) \quad \text{for } r \text{ odd},$$
$$\kappa_r(Y) \simeq \kappa_r(Z) + \epsilon^r \kappa_r(R) \quad \text{for } r \text{ even}.$$

[Kolassa and McCullagh, 1987]. The curious aspect of this result is that even though R is a deterministic function of Z, the joint asymptotic distribution is such that R and Z are statistically independent to a high order of approximation, provided ν is moderately large. By contrast, R and Y are also asymptotically independent, but only to a first order of approximation. In fact $\text{cov}(R, Y) \simeq \epsilon \, \text{var}(R) = \epsilon/12$.

11.2 Deduce that if Z has the gamma distribution with index $\nu \geq 2$ and if Y is equal to Z rounded to d decimal places, then

$$\text{var}(Y) \simeq \mu_Z^2/\nu + \epsilon^2/12,$$

where $\epsilon = 10^{-d}$. What would be the effect on the variance function if Z were rounded to d significant decimal digits rather than to d decimal places?

11.3 Show that if U is uniformly distributed on $(0, 1)$ the first four cumulants are $\frac{1}{2}$, $\frac{1}{12}$, 0 and $-\frac{1}{120}$.

11.4 Use the linearization technique described in section 11.4 to fit the non-linear model (8.4) to the data in Table 8.8. Use gamma errors, log link and take weights equal to the batch sizes. Find the maximum-likelihood estimates of the four parameters, together with their asymptotic standard errors. Plot the residual deviance for fixed δ against δ in the range 50–100°C to check on the adequacy of the Normal approximation for $\hat{\delta}$.