ANALYSIS OF DOSE-RESPONSE DATA IN THE PRESENCE OF EXTRA-BINOMIAL VARIATION

Dennis D. Boos

Department of Statistics, North Carolina State University
Raleigh, N.C. 27695-8203

and

Division of Biometry and Risk Assessment
National Institute of Environmental Health Sciences
Research Triangle Park, NC 27709

Institute of Statistics Mimeo Series No. 1984

October 1990

Analysis of Dose-Response Data in the Presence of Extra-Binomial Variation

by

Dennis D. Boos

SUMMARY

Binary dose-response data often exhibit extra-binomial variation when the responses arise naturally in groups or "litters." This paper investigates the use of generalized Wald and score statistics for robustifying the standard inference methods based on the binomial likelihood. Special attention is given to the probit analysis of a parallel assay of the teratogenic effects on mice of several dioxins.

Keywords: Probit analysis; Logistic regression; Teratology; Empirical variance; Score tests; Goodness-of-fit.

1 Introduction

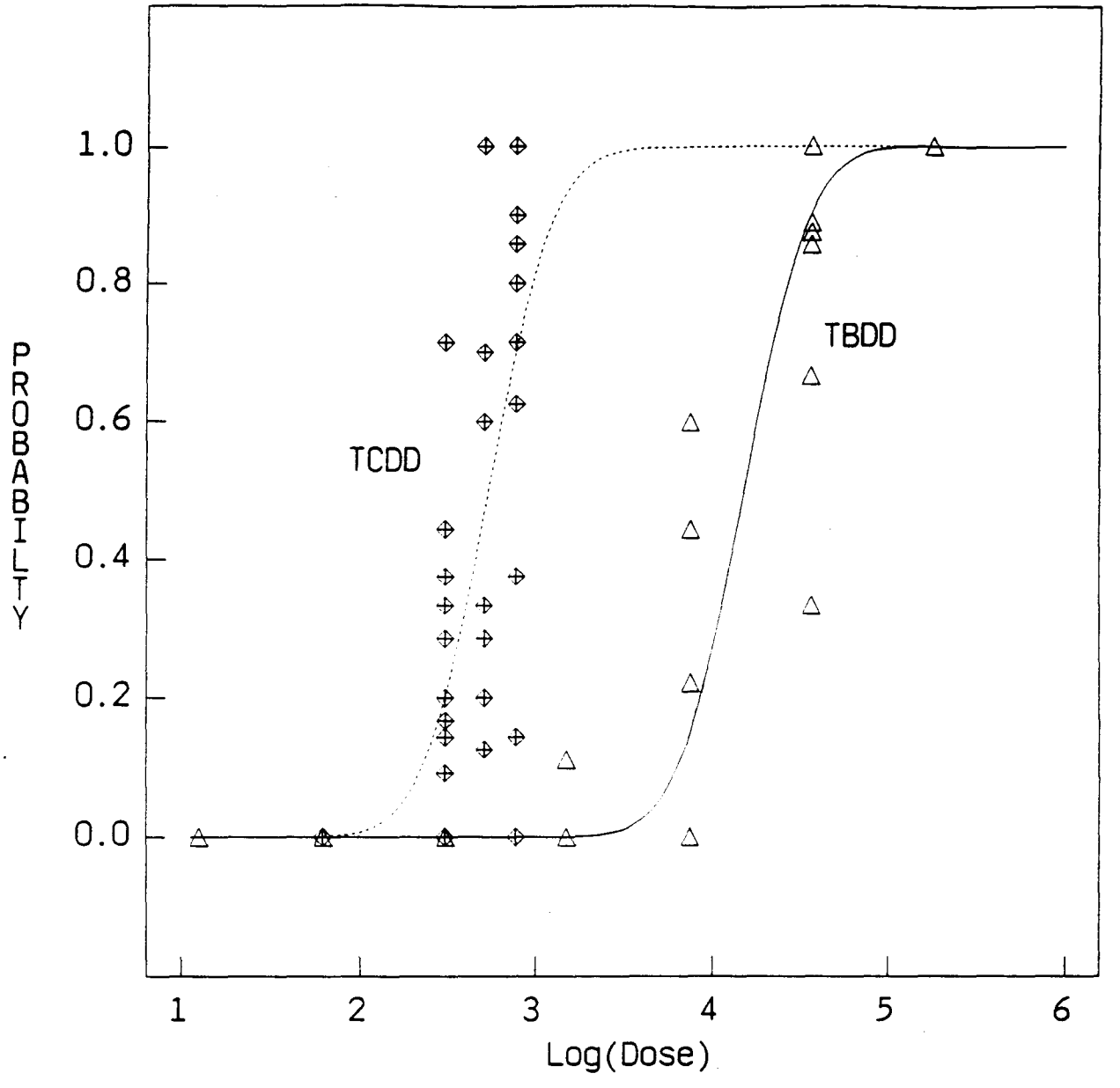
Logistic regression and probit analysis are often used in dose-response modeling of binary responses where binomial likelihoods form the basis for a well-developed theory of estimation and inference (e.g., Cox, 1989, Hosmer and Lemeshow, 1989, and Santner and Duffy, 1989). In certain situations, however, the binary responses arise naturally in groups or "litters" and a binomial likelihood description of the data is not correct due to induced correlations within litters. If Y is the number of "successes" in a litter of size n with $\mathrm{E}(Y|n) = np$, then typically $\mathrm{Var}(Y|n) > np(1-p)$, and the data are said to have extra-binomial variation (see Haseman and Kupper, 1979).

For example, the data in Figure 1 are from a study on the teratogenic effects of certain chemicals including 2,3,7,8-tetrabromodibenzo-p-dioxin (TBDD) in C57BL/6N mice (Birnbaum, Morrissey, and Harris, 1990). The responses are the proportions Y/n of cleft palate incidence in each litter for pregnant dams treated on gestation day 10 and examined on gestation day 18. Also plotted in Figure 1 are similar data for 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) from Birnbaum (1989). One of the goals of the statistical analysis of these data was to estimate the relative potency of TBDD to TCDD after determining if a probit model with common slope could be fit to the data. We shall see in Section 3 that these data exhibit considerable extra-binomial variation which upsets the usual binomial likelihood inference. One possible result of using standard packaged programs based on the binomial model would be that the assumption of common slope could be rejected due to underestimation of the variability. Another result would be that confidence intervals for the relative potency would be too narrow.

A number of different methods have been developed to deal with extra-binomial variation. For example, the binomial model can be expanded to a beta-binomial model and likelihood techniques used (see Haseman and Kupper, 1979, Segreti and Munson, 1981), or the mean-variance relationship may be modeled and generalized linear model methods used (see Williams, 1982), or the extra-binomial aspect can be handled by bootstrap techniques (Carr, 1989).

The approach of this paper is to use the binomial likelihood for estimation but to robustify inferences by using empirical variance estimates which do not rely on the binomial assumption. This general approach is fairly standard (e.g., Kent, 1982, White, 1982, Royall, 1986, Boos, 1990), but its use in binomial regression contexts is just now developing (e.g., Moore and Tsiatis, 1989). It is similar in spirit to common adjustments for heterogeneity or overdispersion (Finney, 1971,

Figure 1. Common Slope Probit Fits



p. 72, McCullagh and Nelder, 1989, p. 127) but is more general because it allows for different amounts of extra-variation to be estimated at each dose level. Moreover, it fits into a general theory for estimating equations as described in Boos (1990). Although the methods discussed are appropriate for the general binomial regression context, I will emphasize dose-response modeling with data having replication at each dose-treatment combination.

The paper is organized as follows. Section 2 introduces the notation and general approach and develops tests about regression parameters. Section 3 then discusses tests for extra-binomial variation and goodness-of-fit tests for adequacy of the mean specification in the presence of extra-binomial variation. The methods are illustrated throughout with the data from Figure 1.

2 The Model and Inference Method

Assume that there are k dose levels, where at the ith dose d_i we observe $\{Y_{ij},\ n_{ij},\ \text{j=1,...,}m_i\}$. In order that results will be fairly general for the binomial regression context, we let the ith dose level have a b x 1 vector of explanatory variables x_i. For dose-response models one typically has $x_i^T = (1, d_i)$ or $x_i^T = (1, d_i, d_i^2)$, and we might allow for treatment structure with additional dummy variables. The model for the mean is $\mathrm{E}(Y_{ij}|n_{ij}, x_i) = n_{ij}p_i(\beta)$, where $p_i(\beta) = F(x_i^T\beta)$ and $F(z)$ is a distribution function such as the logistic or normal and β is a vector of unknown parameters.

If the Y_{ij} are distributed as independent binomial random variables, then the log likelihood is

$$l(\beta) = c + \sum_{i=1}^{k}\sum_{j=1}^{m_i}[Y_{ij}\log\{p_i(\beta)\} + (n_{ij} - Y_{ij})\log\{1 - p_i(\beta)\}].$$

Taking partial derivatives, the maximum likelihood estimator $\hat{\beta}$ solves

$$S(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{k}\sum_{j=1}^{m_i}\{Y_{ij} - n_{ij}p_i(\beta)\}\ \{\frac{p_i'(\beta)}{p_i(\beta)(1 - p_i(\beta))}\}x_i = 0,$$

where $p_i'(\beta) = dF(z)/dz|_{z=x_i^T\beta}$. The negative of the Hessian of $l(\beta)$ is

$$I_Y = -\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T} = \sum_{i=1}^{k}\sum_{j=1}^{m_i} n_{ij}\left[\frac{\{p_i'(\beta)\}^2}{p_i(\beta)\{1-p_i(\beta)\}}\right] x_i x_i^T$$

$$+\sum_{i=1}^{k}\sum_{j=1}^{m_i}\{Y_{ij}-n_{ij}p_i(\beta)\}\left[\frac{p_i''(\beta)}{p_i(\beta)\{1-p_i(\beta)\}} + \frac{\{p_i'(\beta)\}^2\{1-2p_i(\beta)\}}{[p_i(\beta)\{1-p_i(\beta)\}]^2}\right] x_i x_i^T,$$

where $p_i''(\beta) = d^2F(z)/dz^2|_{z=x_i^T\beta}$.

When evaluated at $\hat{\beta}$, this latter quantity is the observed information and will be denoted by $\hat{I}_Y$. The first piece of I_Y is the Fisher information matrix I_f . If the mean model is correct, then the second part of $\hat{I}_Y$ is negligible and $\hat{I}_Y$ is approximately equal to the Fisher information evaluated at $\hat{\beta}$, denoted by $\hat{I}_f$. When $F(\cdot)$ is the logistic distribution function, then $I_Y = I_f = \sum_{i=1}^{k}\sum_{j=1}^{m_i} n_{ij}p_i(\beta)(1-p_i(\beta))x_i x_i^T$.

If the mean is correctly specified, then $\hat{\beta} \xrightarrow{P} \beta$ in large samples and $\hat{\beta}-\beta$ is approximately normal(0,V), where V may be estimated by

$$\hat{V} = \hat{I}_f^{-1}\ \hat{D}_Y\ \hat{I}_f^{-1}$$

and

$$\hat{D}_Y = \sum_{i=1}^{k}\sum_{j=1}^{m_i}\{Y_{ij}-n_{ij}p_i(\hat{\beta})\}^2\ x_i x_i^T.$$

(The required regularity conditions are similar to those for the binomially distributed case which may be found in Fahrmeir and Kaufman, 1985.) In practice I suggest multiplying $\hat{D}_Y$ by N/(N-p) where N is the total sample size and p is the number of parameters estimated, in analogy with least squares regression.

Straightforward inference about β can be made using $\hat{V}$ and Wald type tests even though the data are not binomial. For example, if $\beta^T = (\beta_1^T, \beta_2^T)$ and $H_0 : \beta_2 = 0$ is of interest, then the Wald type test statistic is

$$T_{GW} = \hat{\beta}_2^T \hat{V}_{22}^{-1} \hat{\beta}_2, \tag{1}$$

where $\hat{V}_{22}$ is from the appropriate partition of $\hat{V}$.

Score and likelihood ratio tests are often preferred to Wald tests because of parameter invariance and Type I error considerations. Since likelihood ratio tests do not generalize easily to handle misspecification (see Kent, 1982), I will focus on score test generalizations.

Rao's (1948) score statistic for $H_0 : \beta_2 = 0$ is given by

$$T_S = S(\tilde{\beta})^T \tilde{I}_f^{-1} \; S(\tilde{\beta}) = S_2(\tilde{\beta})^T (\tilde{I}_{f22} - \tilde{I}_{f21}\tilde{I}_{f11}^{-1}\tilde{I}_{f12})^{-1} S_2(\tilde{\beta}),$$

where $\tilde{\beta}$ is the maximum likelihood estimator of β under H_0, and $\tilde{I}_f$ is the Fisher information evaluated at $\tilde{\beta}$ and partitioned as

$$\tilde{I}_f = \begin{pmatrix} \tilde{I}_{f11} & \tilde{I}_{f12} \\ \tilde{I}_{f21} & \tilde{I}_{f22} \end{pmatrix}.$$

If the binomial likelihood is not correct, then T_S could have inflated Type I errors under H_0.

A generalization of T_S which allows for misspecification of the likelihood is given by

$$T_{GS} = S_2(\tilde{\beta})^T \tilde{V}_{S_2}^{-1} S_2(\tilde{\beta}),$$

where

$$\tilde{V}_{S_2} = \tilde{D}_{Y22} \; - \tilde{I}_{f21} \, \tilde{I}_{f11}^{-1} \, \tilde{D}_{Y21}^T \; - \tilde{D}_{Y21} \, \tilde{I}_{f11}^{-1} \, \tilde{I}_{f21}^T \; - \tilde{I}_{f21} \, \tilde{I}_{f11}^{-1} \, \tilde{D}_{Y11} \, \tilde{I}_{f11}^{-1} \, \tilde{I}_{f21}^T \, .$$

Although T_{GS} has a somewhat complex appearance, it may be derived from simple Taylor expansions of $S(\beta)$ (see Breslow, 1990, p. 567, and Boos, 1990, for details). Kent (1982, p. 23) has given an alternate computational form for V_{S_2},

$$\tilde{V}_{S_2} = (\tilde{I}_{f22} \; - \tilde{I}_{f21} \, \tilde{I}_{f11}^{-1} \, \tilde{I}_{f12} \,) \tilde{V}_{22} (\tilde{I}_{f22} \; - \tilde{I}_{f21} \, \tilde{I}_{f11}^{-1} \, \tilde{I}_{f12} \,).$$

Now I want to illustrate the use of $\hat{V}$ and T_{GS} in a modification of the usual probit analysis of a parallel assay. The data are shown in Figure 1 and listed in Table 1. As mentioned in the Introduction, one purpose of the assay was to estimate the relative potency of TBDD to that of TCDD for the incidence of cleft palate. A standard binomial analysis would ignore the individual Y_{ij} values and fit the model to the totals $Y_{i.} = \sum_{j=1}^{m_i} Y_{ij}$ for each dose level. We will use these same parameter estimates, but in addition we use the individual Y_{ij} values from Table

Table 1: Incidence of Cleft Palate, from Birnbaum, Morrissey, and Harris (1990) and Birnbaum (1989).

TBDD

Dose	Y_{ij}	n_{ij}	*Dose*	Y_{ij}	n_{ij}	*Dose*	Y_{ij}	n_{ij}	*Dose*	Y_{ij}	n_{ij}
3	0	7	6	0	8	24	0	9	96	6	7
3	0	11	6	0	9	24	0	9	96	7	7
3	0	10	6	0	9	24	0	7	96	3	3
3	0	9	6	0	10	24	0	9	96	9	9
3	0	10	6	0	8	24	0	5	96	10	10
3	0	8	6	0	8	24	0	9	96	2	3
3	0	7	6	0	10	24	0	8	96	7	8
3	0	10	12	0	3	24	1	9	96	1	3
3	0	9	12	0	9	24	0	11	96	9	9
3	0	10	12	0	7	24	0	6	96	8	9
3	0	2	12	0	8	24	0	9	96	8	8
3	0	9	12	0	9	24	0	9	192	6	6
3	0	9	12	0	5	24	0	8	192	9	9
3	0	10	12	0	6	24	0	6	192	4	4
3	0	9	12	0	8	24	0	9	192	6	6
6	0	11	12	0	8	48	3	5	192	7	7
6	0	6	12	0	9	48	2	9	192	10	10
6	0	3	12	0	10	48	0	8	192	7	7
6	0	7	12	0	6	48	0	8	192	5	5
6	0	3	12	0	8	48	0	10	192	9	9
6	0	9	12	0	9	48	0	5	192	4	4
6	0	10	12	0	10	48	0	8	192	7	7
6	0	9	12	0	9	48	0	3	192	8	8
6	0	3	12	0	7	48	4	9	192	9	9
6	0	9	12	0	8	48	0	9	192	10	10
6	0	11	12	0	11	48	0	8			

TCDD

Dose	Y_{ij}	n_{ij}	*Dose*	Y_{ij}	n_{ij}	*Dose*	Y_{ij}	n_{ij}	*Dose*	Y_{ij}	n_{ij}
6	0	7	12	2	7	12	1	7	18	9	10
6	0	8	12	4	9	12	0	6	18	5	7
6	0	9	12	2	10	15	7	10	18	10	10
6	0	10	12	0	7	15	3	5	18	1	7
6	0	5	12	0	10	15	3	9	18	6	7
6	0	9	12	3	8	15	6	6	18	0	10
6	0	8	12	5	7	15	2	7	18	3	8
6	0	9	12	3	8	15	7	10	18	8	10
6	0	10	12	1	6	15	1	8	18	5	8
6	0	10	12	1	11	15	2	10	18	11	11
6	0	10	12	1	3	15	5	5			

1 to robustly estimate variances. Analysis of the actual extra-binomial variation is deferred until the next section.

For a parallel assay one hopes that a common slope and two intercepts will provide an adequate fit to the data. In the first part of Table 2 are given p-values for the Wald and score tests for the common slope hypothesis and also for the linear versus quadratic hypothesis. The Wald tests require that the larger 4-parameter models be fitted, whereas the score tests only require that the smaller 3-parameter models be fitted. I used SAS NLIN to get the parameter estimates and then SAS IML for the matrix manipulations.

The entry labeled "SAS" is from PROC PROBIT in SAS 6.03. The Wald p-values for this entry are based on an estimated covariance matrix adjusted for heterogeneity as explained in Finney (1971, p. 72). All tests here suggest that neither separate slopes nor a quadratic term help improve the fit very much.

The final fitted model is then $\mathrm{E}(Y_{ij}|n_{ij}, d_i) = n_{ij}\Phi(-14.26 + 3.39d_i)$ for TBDD and $\mathrm{E}(Y_{ij}|n_{ij}, d_i) = n_{ij}\Phi(-9.26+3.39d_i)$ for TCDD. The standard errors for these parameter estimates are given in the middle part of Table 2. Notice that the standard errors from the information matrix $\hat{I}_f^{-1}$ are approximately 55% of those based on $\hat{V} = \hat{I}_f^{-1}\hat{D}\hat{I}_f^{-1}$. The standard errors from SAS PROC PROBIT using the heterogeneity correction factor are between those from $\hat{I}_f^{-1}$ and $\hat{V}$.

The last part of Table 2 gives the median effective dose (ED50) for TBDD, 65.2=exp{14.16/3.39}, and for TCDD, 15.4=exp{9.26/3.39}, and the relative potency, 4.25=exp{(14.16–9.26)/3.39}. The confidence intervals are constructed using Fieller's Theorem as outlined in Finney (1971, p.78). Note that the relative potency confidence interval (3.61,4.81) based on $\hat{V}$ is considerably larger than the interval (3.89,4.64) based on $\hat{I}_f^{-1}$.

3 Model Adequacy

Two Pearson chi-squared statistics for testing adequacy of the mean model $\mathrm{E}(Y_{ij}|n_{ij}, \boldsymbol{x}_i) = n_{ij}p_i(\beta) = n_{ij}F(\boldsymbol{x}_i^T\beta)$ with the binomial likelihood structure are

$$\chi_a^2 = \sum_{i=1}^{k} \frac{[Y_{i\cdot} - n_{i\cdot}p_i(\hat{\beta})]^2}{n_{i\cdot}p_i(\hat{\beta})(1 - p_i(\hat{\beta}))}, \tag{2}$$

Table 2: Probit Analysis of TBDD and TCDD Data

1) P-Values for 3-Parameter vs. 4-Parameter Models

	Wald			Score	
Hypothesis	T_W	T_{GW}	SAS	T_S	T_{GS}
Common Slope	.72	.84	.81	.72	.84
Quadratic Term	.83	.89	.90	.84	.89

2) Final Model Estimates and Standard Errors

		Standard Errors		
Parameter	Estimate	$\hat{I}^{-1}$	$\hat{I}^{-1}\hat{D}\hat{I}^{-1}$	SAS
Intercept(TBDD)	−14.16	1.14	2.06	1.74
Intercept(TCDD)	−9.26	.74	1.26	1.13
Common Slope	3.39	.27	.48	.42

3) Median Effective Dose and Relative Potency Estimates

		95% Confidence Limits					
		$\hat{I}^{-1}$		$\hat{I}^{-1}\hat{D}\hat{I}^{-1}$		SAS	
Parameter	Estimate	L	R	L	R	L	R
ED50(TBDD)	65.2	60.5	70.4	58.1	72.0	56.8	75.2
ED50(TCDD)	15.4	14.7	16.1	14.1	17.0	14.2	16.8
Rel. Potency	4.25	3.89	4.64	3.61	4.81		

Note: The T_{GW} test statistic is defined in (1) and T_W has the same form but with $\hat{I}_{f22}$ in place of $\hat{V}_{22}$. SAS results are from SAS PROC PROBIT. $\hat{D}$ has been multiplied by N/(N–p)=149/145 or 149/146, and the p-values in Part 1) use t percentiles with N–p=145 or 146 degrees of freedom.

which is based on totals $Y_{i.} = \sum_{j=1}^{m_i} Y_{ij}$ and $n_{i.} = \sum_{j=1}^{m_i} n_{ij}$, and

$$\chi_b^2 = \sum_{i=1}^{k} \sum_{j=1}^{m_i} \frac{[Y_{ij} - n_{ij} p_i(\hat{\beta})]^2}{n_{ij} p_i(\hat{\beta})(1 - p_i(\hat{\beta}))},$$

which is based on the individual litter values (Y_{ij}, n_{ij}).

Either of these statistics will be sensitive to both mean specification and to extra-binomial variation. I think it is helpful, however, to separate the distributional part from the mean specification since typically the mean specification is the focus of the investigation. Although Section 2 shows how to study the mean part without requiring the binomial assumptions to hold, it can be of interest to check the binomial assumption and perhaps quantify the extent of the extra-binomial variation. I will first illustrate the latter with the example data and then discuss methods for assessing the mean fit.

The binomial likelihood can be tested without modeling the mean by simply replacing $p_i(\hat{\beta})$ by $\overline{Y_{i.}} = Y_{i.}/n_{i.}$ in χ_b^2 yielding

$$\chi_c^2 = \sum_{i=1}^{k} \sum_{j=1}^{m_i} \frac{[Y_{ij} - n_{ij}\overline{Y_{i.}}\,]^2}{n_{ij}\overline{Y_{i.}}\,(1 - \overline{Y_{i.}}\,)}. \tag{3}$$

See Tarone (1979) for other appropriate test statistics. Table 3 lists the components of χ_c^2 by dose for TBDD and TCDD for all cases where $\overline{Y_{i.}} > 0$. For TBDD in Dose=24, there was just one incidence of cleft palate in all litters and thus little distributional information. For the other doses listed and for the overall χ_c^2 there is strong evidence of extra-binomial variation.

One way to quantify this extra-binomial variation is to compute the "heterogeneity factor" $\chi 2/(df - 1)$ for each dose level. Another approach is to use the ratio of the empirical estimate of $\mathrm{Var}(\overline{Y_{i.}}\,|n_{i1}, \ldots, n_{im_i})$ given by

$$(\frac{m_i}{m_i - 1})\frac{1}{n_{i.}^2} \sum_{j=1}^{m_i} [Y_{ij} - n_{ij}\overline{Y_{i.}}\,]^2 \tag{4}$$

to that under the binomial assumption, $\overline{Y_{i.}}\,(1 - \overline{Y_{i.}}\,)/n_{i.}$. The resulting ratio listed in Table 3 is

$$\text{Var Ratio} = (\frac{m_i}{m_i - 1}) \sum_{j=1}^{m_i} \frac{[Y_{ij} - n_{ij}\overline{Y_{i.}}\,]^2}{n_{i.}\overline{Y_{i.}}\,(1 - \overline{Y_{i.}}\,)}. \tag{5}$$

Table 3: Extra-Binomial Variation in the TBDD and TCDD Treatment Groups

Treatment	Dose	χ^2	df	$\chi^2/(df-1)$	Var Ratio
TBDD	24	14.88	17	0.93	1.00
TBDD	48	31.06	11	3.11	2.99
TBDD	96	21.62	11	2.16	1.29
		67.56	39		
TCDD	12	30.72	15	2.19	2.22
TCDD	15	19.69	8	2.81	2.77
TCDD	18	57.50	12	5.22	5.56
		107.91	35		

Note: The "χ^2" entries are the components of χ^2_c in (3), and the "Var Ratio" entries are defined in (5).

The factor $m_i/(m_i - 1)$ has been added so that if the n_{ij} have a common value, then (3) is an unbiased estimate of $\mathrm{Var}(\overline{Y_{i.}} \mid n_{i1}, \ldots, n_{im_i})$. In that case the Var Ratio is just the component of χ^2_c for that dose divided by $m_i - 1$ and thus equal to the heterogeneity factor for that dose. The differences between $\chi^2/(df - 1)$ and the Var Ratio are small except for TBDD at Dose=96 where there are three n_{ij} equal to 3 and the rest lie between 7 and 10. Dose 18 for TCDD seems to have considerably more overdispersion than the other dose-treatment combinations. Note also that (4) is the unstructured version of $\hat{V} = \hat{I}_f^{-1}\hat{D}_Y\hat{I}_f^{-1}$ given in Section 2 and closely related to the jackknife variance estimator of Gladen (1979).

Now we turn to checking the adequacy of the mean specification $\mathrm{E}(Y_{ij}|n_{ij}, \boldsymbol{x}_i) = n_{ij}F(\boldsymbol{x}_i^T\beta)$. In a standard binomial analysis, one would typically use χ^2_a in (2) to assess adequacy of the mean. Here we want to develop tests for mean adequacy without assuming that the data are binomial.

The form of the generalized score statistic T_{GS} given in Section 2 for testing $H_0 : \beta_2 = 0$ is not appropriate here since we want to test $H_0 : \mathrm{E}(Y_{ij}|n_{ij}, \boldsymbol{x}_i) = n_{ij}F(\boldsymbol{x}_i^T\beta)$. Instead we let $\theta^T = (\theta_1, \ldots, \theta_k)$, where $\theta_i = \mathrm{E}(Y_{ij}/n_{ij}|n_{ij}, \boldsymbol{x}_i)$ is the unmodeled true mean parameter for the ith dose. The null hypothesis is then $H_0 : \theta = g(\beta)$, where $g(\beta)^T = (F(\boldsymbol{x}_1^T\beta), \ldots, F(\boldsymbol{x}_k^T\beta))$. In this formulation the generalized score statistic is

$$T_{GS} = S(\tilde{\theta}^T)[\tilde{D}_Y^{-1} - \tilde{D}_Y^{-1}\,\tilde{I}_Y\,\tilde{G}(\tilde{G}^T\tilde{I}_Y\,\tilde{D}_Y^{-1}\,\tilde{I}_Y\,\tilde{G})^{-1}\tilde{G}\tilde{I}_Y\,\tilde{D}_Y^{-1}\,]S(\tilde{\theta}), \tag{6}$$

where now

$$S(\tilde{\theta})_i = \frac{[Y_{i.} - n_{i.}p_i(\tilde{\beta})]}{p_i(\tilde{\beta})(1 - p_i(\tilde{\beta}))}, i = 1, \ldots, k,$$

$$\tilde{I}_Y = Diag\left[n_{i.}\left\{\frac{\overline{Y_{i.}}}{\{p_i(\tilde{\beta})\}^2} + \frac{1 - \overline{Y_{i.}}}{\{1 - p_i(\tilde{\beta})\}^2}\right\}, i = 1, \ldots, k\right],$$

$$\tilde{D}_Y = Diag\left[\sum_{j=1}^{m_i}\frac{\{Y_{ij} - n_{ij}p_i(\tilde{\beta})\}^2}{[p_i(\tilde{\beta})\{1 - p_i(\tilde{\beta})\}]^2}, i = 1, \ldots, k\right],$$

and $\tilde{G} = \partial g(\beta)/\partial\beta^T = \tilde{F}'X$ with $X = [\boldsymbol{x}_1|\boldsymbol{x}_2|\cdots|\boldsymbol{x}_k]$ and

$$\tilde{F}' = Diag\left[\frac{dF(z)}{dz}\Big|_{z=\boldsymbol{x}_i^T\tilde{\beta}}, i = 1, \ldots, k\right].$$

The general form for T_{GS} in this context is derived in Boos (1990). $S(\theta)$, I_Y, and D_Y are different from those given at the beginning of Section 2 because the "full model" here assumes that Y_{ij} is distributed as binomial(n_{ij}, θ_i) and the "reduced model" has Y_{ij} distributed as binomial$(n_{ij}, F(x_i^T\beta))$. Under $H_0 : \theta_i = F(x_i^T\beta), i = 1, \ldots, k$ where β is b x 1, T_{GS} converges to a chi-squared random variable with k–b degrees of freedom.

Rewrite (6) as $T_{GS} = T_{GS1} - T_{GS2}$, where $T_{GS1} = S(\tilde{\theta})^T \tilde{D}_Y^{-1} S(\tilde{\theta})$. If the mean specification and binomial likelihood are both correct, then T_{GS1} is asymptotically equivalent to the Pearson statistic χ_a^2 in (1), and T_{GS2} converges to zero in probability as the cell sizes m_i tend to infinity. As in Section 2 we may replace $\tilde{I}_Y$ by the Fisher information

$$\tilde{I}_f = Diag[n_{i.}/p_i(\hat{\beta})\{1 - p_i(\hat{\beta})\}, i = 1, \ldots, k]$$

since both $\overline{Y_{i.}}$ and $p_i(\hat{\beta})$ converge to $p_i(\beta)$ under H_0. If suitable replication is not available at each dose level, then some grouping and a generalization of Tsiatis (1980) may be used.

For the 3-parameter probit fit given in the middle of Table 3, Table 4 lists the components of χ_a^2 and $T_{GS1} = S(\tilde{\theta})^T \tilde{D}_Y^{-1} S(\tilde{\theta})$ for the 6 dose-treatment combinations where at least one cleft palate appeared. Since $T_{GS} = 1.25$ is not near significance when compared to a χ_3^2 distribution, the probit fit seems adequate. (If $\tilde{I}_f$ is used in place of $\tilde{I}_Y$, then $T_{GS2} = .48$ and $T_{GS} = 1.12$.) Notice, however, that $\chi_a^2 = 19.36$ is highly significant with a very large component in the first row of Table 4. This is a case where $\overline{Y_{i.}} = 1/142 = .007$ and $p_i(\hat{\beta}) = .00035$. If that single cleft palate had not occurred, then $\chi_a^2 = 0.32$ and $T_{GS} = 0.10$ yielding much more similar results for the two statistics. It is interesting that the likelihood ratio statistic $G^2 = 5.67$ (not shown) does not seem to have the same sensitivity to that one cell as does its analogue $\chi_a^2 = 19.36$. SAS PROC PROBIT prints out G^2 but keys on χ_a^2 and uses the heterogeneity factor $19.4/8 = 2.43$ to multiply I^{-1} for inference purposes as seen in Table 2 parts 2) and 3). In a sense, the inference based on the heterogeneity factor brings the analysis closer to my analysis based on empirical variances but seemingly for the wrong reasons (i.e., mean lack-of-fit in that one cell).

Lastly, I reran all the analyses for a logit model in place of the probit model and got similar results using my methods, but $\chi_a^2 = 2.66$. Thus a standard binomial analysis would not use the heterogeneity factor for inference with the logit model for these data, and confidence limits for the ED50's which are based on I_f are too narrow.

Table 4: Mean Specification Lack-of-Fit Statistics
for 3-Parameter Probit Model

Treatment	Dose	χ^2_a	T_{GS1}		
TBDD	24	18.06	0.91		
TBDD	48	1.01	0.46		
TBDD	96	0.23	0.21		
TCDD	12	0.03	0.02		
TCDD	15	0.02	0.01		
TCDD	18	0.01	0.00	$-T_{GS2}$	T_{GS}
		19.36	1.60	−0.35	1.25

Note: The "χ^2_a" entries are the components of χ^2_a in (2), and $T_{GS} = T_{GS1} - T_{GS2}$ is defined in (6).

4 Discussion

Inference for dose-response models in the presence of extra-binomial variation is easily carried out using empirical variances. No enlarged models such as the beta-binomial or the inclusion of a variance function are required. Either Wald or score tests may be used for testing nested hypotheses about the mean as shown in Section 2. For the goodness-of-fit tests in Section 3 it is useful to have replication at each dose-treatment combination as in the example data. Replication also helps in the estimates of $\hat{D}$ in Section 2 although it is not essential.

Acknowledgements

I would like to thank Joe Haseman for suggesting the problem and providing the data and Beth Gladen and Chris Portier for subsequent discussions.

REFERENCES

Birnbaum, L. S., Harris, M. W., Stocking, L. M., Clark, A. M., and Morrissey, R. E. (1989), "Retinoic acid selectively enhances teratogenesis in C57BL/6N mice," *Toxicology and Applied Pharmacology*, 98,487-500.

Birnbaum, L. S., Morrissey, R. E., and Harris, M. W. (1990), "Teratogenic effects of 2,3,7,8-tetrabromodibenzo-p-dioxin and three polybrominated dibenzofurans in C57BL/6N mice," to appear in *Toxicology and Applied Pharmacology.*

Boos, D. D. (1990), "On generalized score tests," Institute of Statistics Mimeo Series No. 1980, North Carolina State University, Raleigh, NC.

Breslow, N. (1990), "Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models," *Journal of the American Statistical Association*,85,565-571.

Carr, G. J. (1989), "Dose-response models in quantal response teratology," unpublished Ph.D. dissertation, University of North Carolina at Chapel Hill, Department of Biostatistics.

Cox, D. R. (1989), *The Analysis of Binary Data* (2nd edition), London: Chapman and Hall.

Finney, D. J. (1971), *Probit Analysis* (3rd edition), Cambridge: Cambridge University Press.

Gladen, B. (1979), "The use of the jackknife to estimate proportions from toxicological data in the presence of litter effects," *Journal of the American Statistical Association*,74,278-283.

Haseman, J. K., and Kupper, L. L. (1979), "Analysis of dichotomous response data from certain toxicological experiments," *Biometrics*,35,281-293.

Fahrmeir, L., and Kaufmann, H. (1985), "Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models," *Annals of Statistics*,13,342-368.

Hosmer, D. W., and Lemeshow, S. (1989), *Applied Logistic Regression*, New York: John wiley.

Kent, J. T. (1982), "Robust properties of likelihood ratio tests," *Biometrika*,69,19-27.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, London: Chapman and Hall.

Moore, D. F., and Tsiatis, A. (1989), "Robust estimation of the variance in moment methods for extra-binomial and extra-Poisson variation, "preprint.

Rao, C. R. (1948), "Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation," *Proceedings of the Cambridge Philosophical Society*,44,50-57.

Royall, R. M. (1986), "Model robust confidence intervals using maximum likelihood estimators," *International Statistical Review*,54,221-226.

Santner, T. J., and Duffy, D. E. (1989), *The Statistical Analysis of Discrete Data*, New York: Springer-Verlag.

Segreti, A. C., and Munson, A. E. (1981), "Estimation of the median lethal dose when responses within a litter are correlated," *Biometrics*, 37,153-156.

Tarone, R. E. (1979), "Testing the goodness of fit of the binomial distribution," *Biometrika*,66,585-590.

Tsiatis, A. A. (1980), "A note on a goodness-of-fit test for the logistic regression model," *Biometrika*,67,250-251.

White, H. (1982), "Maximum likelihood estimation of misspecified models," *Econometrica*,50, 1-26.

Williams, D. A. (1982), "Extra-binomial variation in logistic linear models," *Applied Statistics*,31,144-148.