AN ABSTRACT FOR THE THESIS OF

Lisa M. Ganio-Gibbons for the degree of Doctor of Philosophy in Statistics presented on August 18, 1989.

Title: Diagnostic Tools for Overdispersion in Generalized Linear Models

Abstract approved: Redacted for Privacy
Daniel W. Schafer

Data in the form of counts or proportions often exhibit more variability than that predicted by a Poisson or binomial distribution. Many different models have been proposed to account for extra-Poisson or extra-binomial variation. A simple model includes a single heterogeneity factor (dispersion parameter) in the variance. Other models that allow the dispersion parameter to vary between groups or according to a continuous covariate also exist but require a more complicated analysis. This thesis is concerned with (1) understanding the consequences of using an oversimplified model for overdispersion, (2) presenting diagnostic tools for detecting the dependence of overdispersion on covariates in regression settings for counts and proportions and (3) presenting diagnostic tools for distinguishing between some commonly used models for overdispersed data.

The double exponential family of distributions is used as a foundation for this work. A double binomial or double Poisson

density is constructed from a binomial or Poisson density and an additional dispersion parameter. This provides a completely parametric framework for modeling overdispersed counts and proportions.

The first issue above is addressed by exploring the properties of maximum likelihood estimates obtained from incorrectly specified likelihoods. The diagnostic tools are based on a score test in the double exponential family. An attractive feature of this test is that it can be computed from the components of the deviance in the standard generalized linear model fit. A graphical display is suggested by the score test. For the normal linear model, which is a special case of the double exponential family, the diagnostics reduce to those for heteroscedasticity presented by Cook and Weisberg (1983).

Diagnostic Tools for Overdispersion in Generalized Linear Models

by

Lisa M. Ganio-Gibbons

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Completed August 18, 1989

Commencement June 1990

APPROVED:

Redacted for Privacy

Professor of Statistics in charge of major

Redacted for Privacy

Head of Department of Statistics

Redacted for Privacy

Dean of Graduate School

Date thesis is presented August 18, 1989

Typed by Lisa Ganio-Gibbons for Lisa M. Ganio-Gibbons

ACKNOWLEDGEMENTS

I began this project with a great deal of enthusiasm and as it draws to a close I'm beginning to think that it was fun. But there were many hours of quiet desperation between then and now and it was during that time that I came to appreciate the help and support that has brought this endeavor to its end.

I extend my deepest appreciation to Dan Schafer, whose timely guidance, encouragement and prodding has made this work an enjoyable exercise in thought and accomplishment. His ability to apply humor and careful thinking at the same time is beyond compare. My thanks too, to the rest of my committee, Don Pierce, Cliff Pereira, Fred Ramsey and Bob Burton; and to the rest of the faculty and staff of the Statistics Department, especially Gen and Ron, who provided help with endless patience. And thanks to my parents, who were my first and best teachers.

I would like to thank Roger and Maryam, for sharing the day to day routine of study and work, and for being the best of office-mates and friends. Special thanks are due to Jeannie, Joyce and Jane, who have helped me more than they know. And above all, thanks to John, who kept my feet firmly planted on the good earth and my heart in my home.

And finally, thanks to to Gnat and Derby, who were always there.

This research was supported by NSF grant DMS 87 02111.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

DIAGNOSTIC TOOLS FOR OVERDISPERSION IN GENERALIZED LINEAR MODELS

Chapter 1

THE DEPENDENCE OF OVERDISPERSION ON COVARIATES

1.1 OVERDISPERSION

Data in the form of counts or proportions are often analyzed as observations from a Poisson or binomial distribution. Counted data, however, often exhibit greater variability than that predicted by these parametric models. This extra variability has been called extra-Poisson or extra-binomial variation or, more generally, overdispersion. This thesis will discuss the analysis of overdispersed counted data with regression models when the overdispersion may depend on covariates or factors. In particular it will provide a practical diagnostic tool for determining whether it is necessary to model this structure.

There are potentially many models for overdispersion and it is desirable to evaluate the appropriateness of each. A simple model (Finney, 1971; Wedderburn, 1974) accounts for overdispersion by a constant heterogeneity factor. Nelder and Pregibon (1987) noted that,

> "Wedderburn's original quasilikelihood model assumed that the dispersion parameter φ is constant for all observations. In certain applications it may be desirable to check this assumption, or perhaps model φ as a function of known covariates."

Thus there is a need for an easy diagnostic method for identifying patterns in the extra variability. Fitting a model where overdispersion is accounted for by a constant heterogeneity factor is relatively simple. It would be nice to know whether the data supports this model or whether it indicates that a more sophisticated model is needed.

Various models have been proposed for the probabilistic mechanisms that produce overdispersed data. For example, important covariates left out of the regression model, measurement errors in covariates, inter-subject variability and mixture models can be responsible for variation that is greater than expected. For proportions, non-independence of Bernoulli trials can also lead to extra binomial variation.

In many cases overdispersion may be directly related to factors or to continuous covariates. For example, a treatment may affect the variability of the responses as well as the mean. If a regression model is fit to a function of the mean and an important covariate is omitted, then it is possible that overdispersion is associated with the omitted covariate. If a measuring process improves over time, then overdispersion may be related to time. If a covariate in the regression model contains measurement error then the variability of the response given the measured variable will depend on a term which

is proportional to the variance of the covariate given its measurement.

In addition, the variance model with a simple heterogeneity factor may be an oversimplification of the variability in responses, even though overdispersion doesn't depend on covariates. For example, if Y is a count with mean μ, two common models for extra-Poisson variation are, $\mathrm{Var}(Y) = \sigma^2\mu$ and $\mathrm{Var}(Y) = \mu(1 + \sigma^2\mu)$. It will be shown in Chapter 3 that the latter model can be studied using a constructed covariate. It may be desired to include an assessment of the appropriateness of each of these models in the statistical analysis.

In all of these situations, the extra variation may be modeled using known covariates or factors. This thesis is concerned with simple methods for using the data to indicate whether or not this is necessary. The following examples illustrate situations in which it is desired to investigate the dependence of overdispersion on covariates as part of the statistical analysis.

1.2 INTRODUCTION TO THE EXAMPLES

The data described in Examples 1 through 5 are provided in the appendix.

1.2.1. Fish Toxicology Data

An experiment was conducted by researchers in the Environmental Health Sciences Center at Oregon State University to investigate the

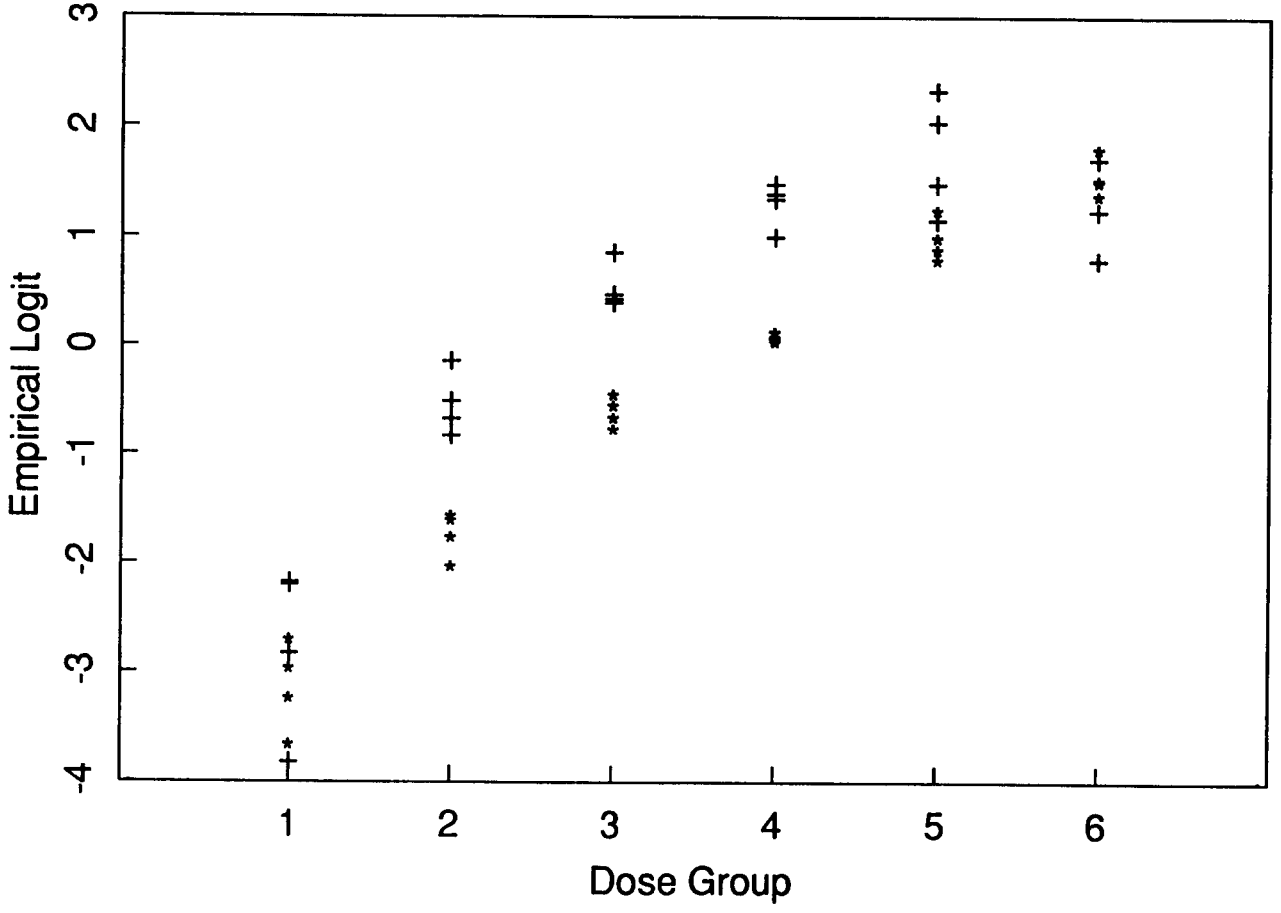
carcinogenic effects of aflatoxin, a toxic by-product produced by a mold which infects cottonseed meal, peanuts and grains. Tanks of rainbow trout embryos were exposed to either aflatoxin B1 or a related compound, aflatoxicol, at one of six doses for one hour. The fish were allowed to grow for one year and then the number of fish developing liver cancer in each tank was recorded. The entire experiment was replicated four times. The statistical analysis involves fitting a logistic regression model to determine if there is a difference in the dose response relationships for the two carcinogens. Figure 1.1 is a plot of the empirical logit versus dose level for each treatment.

Researchers involved in this experiment know that the metabolic pathway from aflatoxicol to liver cancer is much longer than the pathway from aflatoxin B1 to cancer. Thus they expect to see more variation in the outcomes for fish treated with a given dose of aflatoxicol than for an equivalent dose of aflatoxin B1. This suggests that overdispersion may depend on the treatment group. In addition, a lack of independence in the outcome for each fish due to such things as competition for food may also lead to overdispersion. It would be useful to apply the diagnostic tools described in this thesis to investigate the presumed dependence of overdispersion on treatment group.

1.2.2. Fish Vaccination Data

An experiment was conducted by researchers in the Department of Microbiology at Oregon State University. The proportion of fish dying

Figure 1.1 Fish Toxicology Data



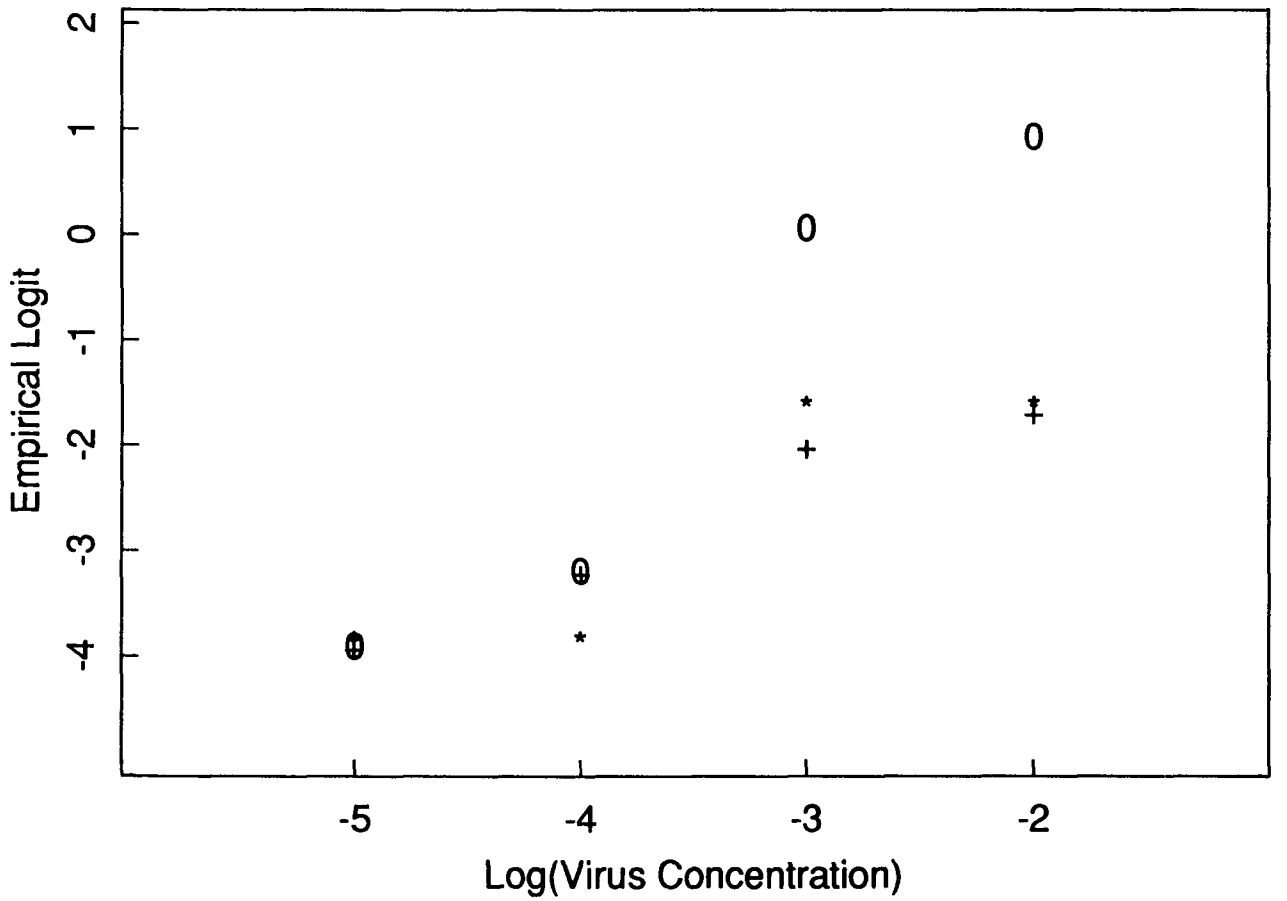
Aflatoxin B1 = (*) Aflatoxicol = (+)

due to viral infection in several treatment groups was used to study the effectiveness of an antiviral vaccine. Interest lies in comparing the expensive inoculation vaccine treatment with the inexpensive immersion vaccine.

Tanks of fish were given one of three vaccination treatments and after a period of 30-35 days were exposed to one of 5 dilutions of virus. The number of fish dying due to viral infection in each tank was recorded. The vaccination treatments were (1) no vaccine, (2) vaccine was dissolved in the water into which the fish were immersed and (3) vaccine was applied by injection to each fish. This set-up was replicated at four locations and each replication was labeled an experiment. The statistical analysis involves fitting logistic regression models to determine the relative risks of death for the inoculated and immersed treatment groups. The data from the first of the four experiments are plotted in Figure 1.2

There are two reasons to suspect overdispersion in this data set. First, as in Example 1 above, the tank effects may be thought to induce correlations between outcomes for individual fish, resulting in overdispersion. Second, the exact dosage of vaccine is known for the control and the inoculated treatment groups. However, the exact dosage is not known for the immersed fish. The absorption rate of the vaccine may depend on fish size, overall fish health or on other unidentified factors. In any case, there is some measurement error associated with the amount of vaccine received by each fish in this treatment group which would result in a higher degree of overdispersion in this treatment group than in the others.

Figure 1.2 Fish Vaccination Data



As a preliminary step in the analysis of this data it would be helpful to know whether the degree of overdispersion does, in fact, differ for the different treatment groups.

1.2.3 Salmonella Data

Simpson and Margolin (1986) reported the results of an experiment in which plates containing *Salmonella* bacteria were exposed to various doses of Acid Red 114 and the number of revertant colonies in each plate was observed. The researchers were interested in the pattern of response and the tendency of the treatment to be toxic at high dose levels. The logarithm of the count for each plate is plotted in Figure 1.3 versus the logarithm of dose.

For this problem it may be of interest to determine which models for overdispersion are appropriate. If $E(Y_i) = \mu_i$, then some possible models discussed throughout this thesis for the variance of Y_i are:

$$\text{Var}(Y_i) = \mu_i \sigma^2 \qquad \text{and}$$

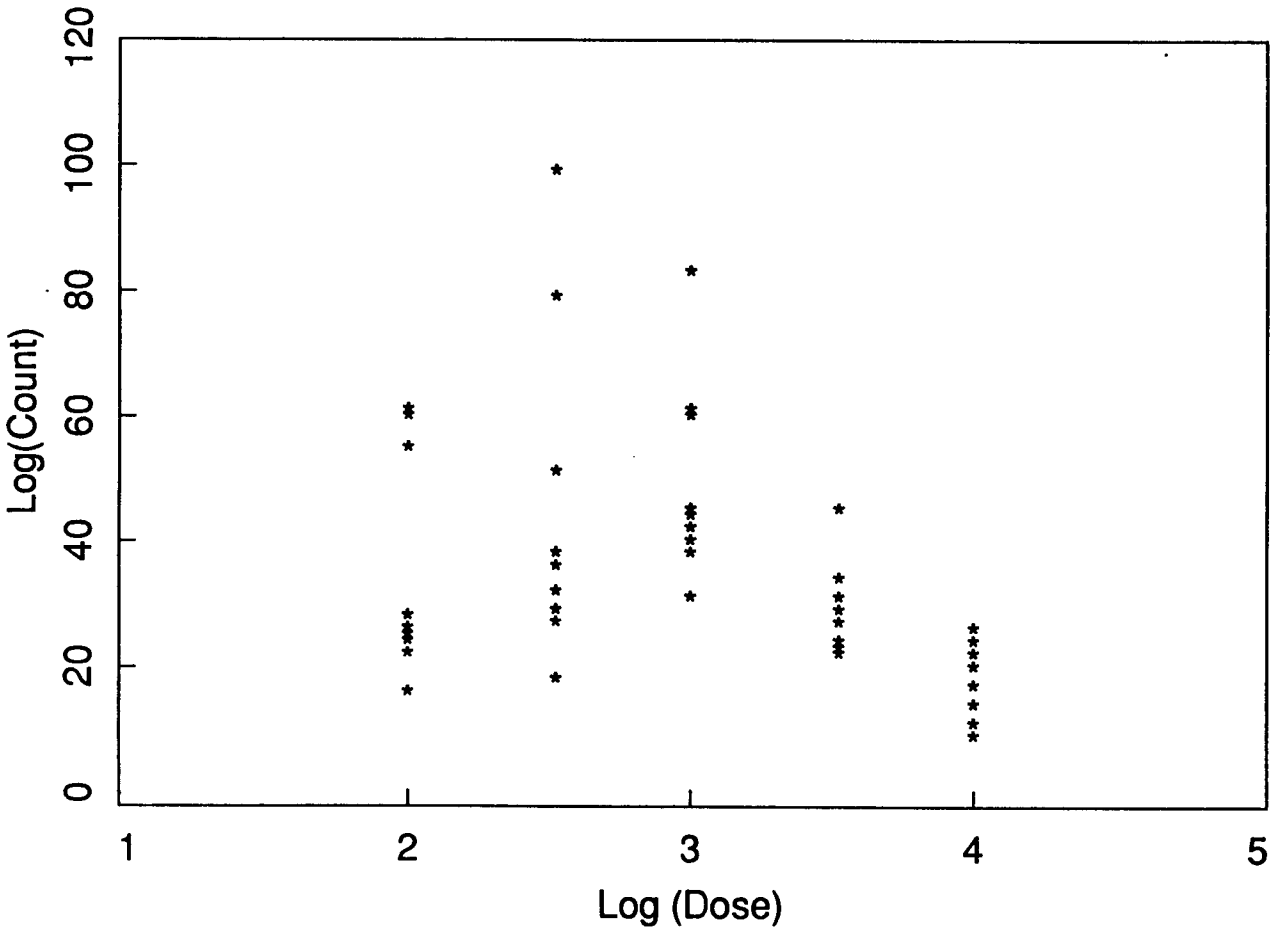
$$\text{Var}(Y_i) = \mu_i [1 + \alpha_0 \mu_i] \ .$$

The first model is relatively simple; the second model will allow extra-Poisson variation to change with the mean. A useful analysis would include an evaluation of the appropriateness of each model.

1.2.4 Chromosome Aberration Data

Blood samples from 649 survivors of the atomic bombing of Hiroshima were collected after the bomb blast. Thirty to one hundred circulating lymphocytes were examined and the number of lymphocytes with chromosome aberrations was observed (Otake and Prentice, 1984).

Figure 1.3 Salmonella Bacteria Data



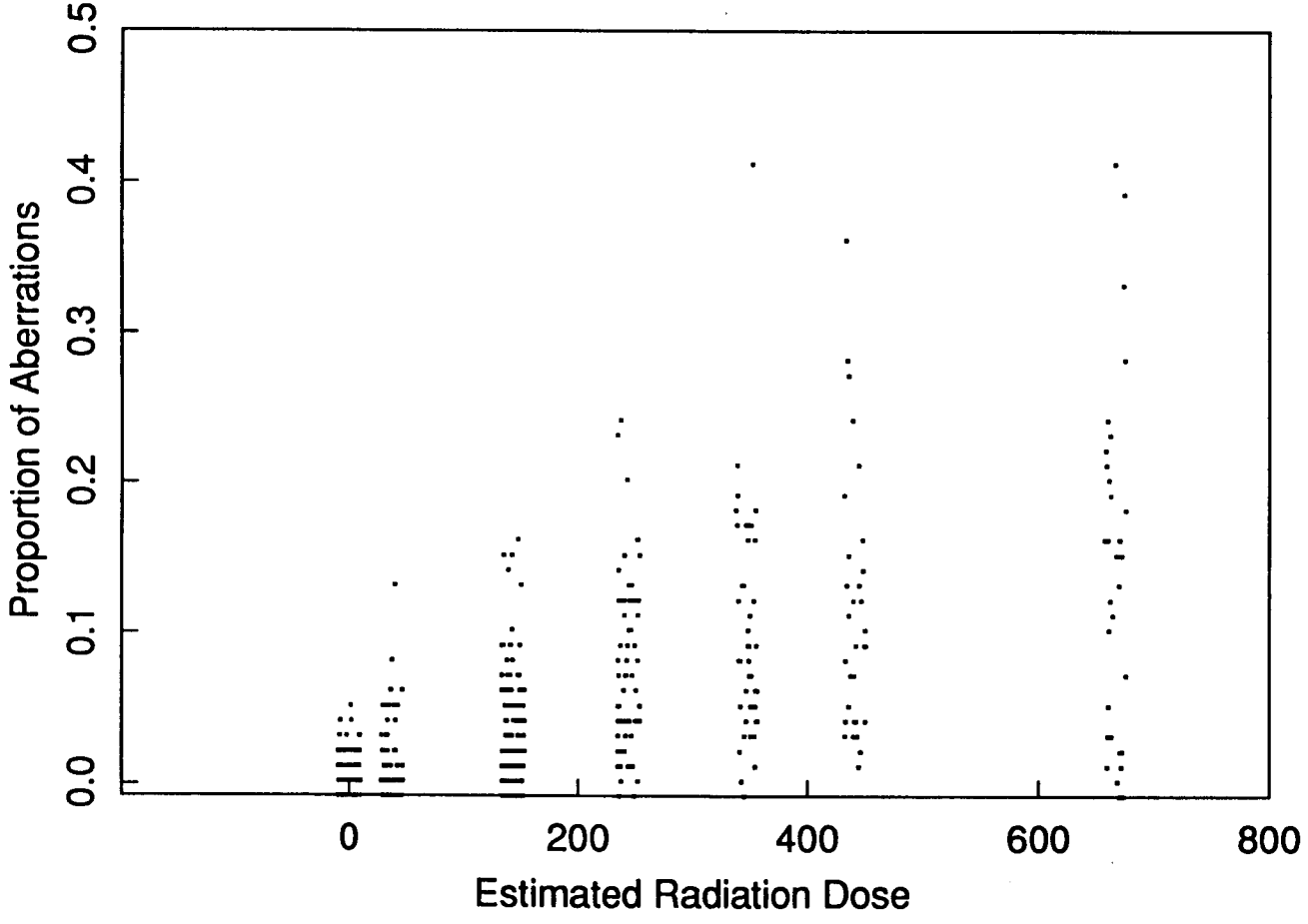
An estimate of the amount of gamma and neutron radiation received was also estimated using information provided by the survivor on their location and shielding at the time of the blast. It is of interest to know how the aberration rate depends on the total dose of radiation received. Based on biological consideration, the statistical analysis involves fitting a linear regression model to the proportion of chromosome aberrations. The data are plotted in Figure 1.4. A small random uniform number was added to each grouped average exposure to display the concentration of points in each dose category.

The measured radiation dose is known to contain substantial measurement error and the standard deviation of the measured radiation dose is thought to be proportional to the true radiation dose. As discussed previously, if Y is the proportion of chromosome aberrations, X is the true radiation received and Z is the measured radiation, then with a simple multiplicative model for measurement error, the variance of Y given Z would be the binomial variance plus a term which is quadratic in Z. The diagnostic tools presented in this thesis may be used at an early stage of the analysis to check on this presumed form for the overdispersion.

1.2.5 Rotenone Data

In an experiment to assess the insecticidal properties of rotenone and degulin, two compounds obtained from the roots of the plant genus *Derris*, batches of the Chrysanthemum Aphid, *Macrosiphoniella sanborni* were exposed to either rotenone or degulin at varying doses or to a 1:4 mixture of the two toxins, and the

Figure 1.4 Chromosome Aberration Data



mortality of each batch was noted. It is of interest to know whether an additive model is adequate to describe the toxic effects of rotenone and degulin or whether interaction terms are needed to represent non-parallel probit regression lines. The statistical analysis involves fitting a probit regression model to the data with and without the appropriate interaction terms. Figure 1.5 is a plot of the empirical probits versus the logarithm of dose. This data was initially reported by Martin (1942) and Finney (1971) fit a probit regression model to the data.

Since it can be difficult to distinguish between overdispersion and interaction, it is important that the overdispersion be modeled as adequately as possible. So it is worthwhile to compare the relative validity of competing models for overdispersion. If Y_i is the proportion of dead insects in a batch of size m_i and $E(Y_i) = \mu_i$, two models for the variance of Y are given by,

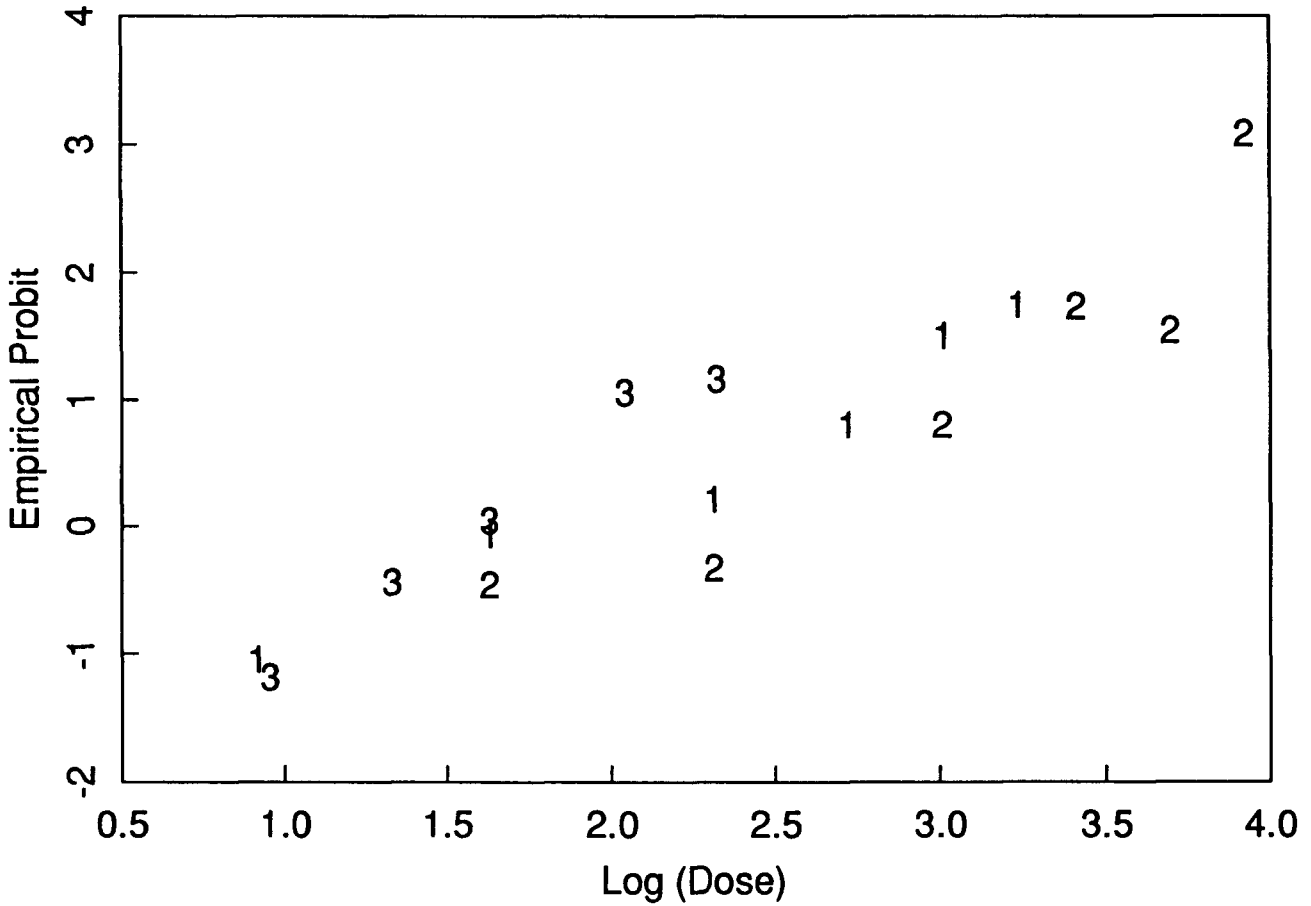
$$\text{Var}(Y_i) = \mu_i(1-\mu_i)(1/m)\ \sigma^2 \quad \text{and}$$

$$\text{Var}(Y_i) = \mu_i(1-\mu_i)(1/m)\ [1 + \sigma_0\mu_i(1-\mu_i)] \ .$$

These models are are analogous to the models for counts given in Example 3.

In this example, the treatment is applied to an entire batch of insects and not to individual aphids, and so there are potential differences in the doses actually received by individual insects. If it is suspected that the standard deviation of the measurement error

Figure 1.5 Rotenone Data



Rotenone=1 Degulin=2 Rotenone+Degulin=3

is proportional to the logarithm of the dose, than another potential model for the variance of Y_i is,

$$\text{Var}(Y_i) = \mu_i(1-\mu_i)(1/m)\ [\alpha_0 + \alpha_1 \log(\text{dose})] \quad .$$

It would be helpful to include a comparison of the validity of these three competing models in the statistical analysis.

1.3 DESCRIPTION AND SUMMARY OF THE THESIS

The examples of Section 1.2 have shown that there are often good reasons for suspecting that the amount of overdispersion can depend on a factor or covariate.

Many different models for overdispersed data exist and some are described in Chapter 2. Some account for the overdispersion through a heterogeneity factor which is constant for all observations. For example, a model for overdispersed counts that is often fit by quasilikelihood methods (see Section 2.2.3) may use $\text{Var}(Y_i) = \sigma^2\mu_i$, where $E(Y_i) = \mu_i$ and σ^2 is constant for all observations. Computer packages such as GLIM (Baker and Nelder, 1978) are available to fit this type of model. This has contributed to the growing recognition and statistical treatment of overdispersed data.

Models that will allow overdispersion to depend on a covariate or vary from group to group, such as the regression models used with extended quasi-likelihood methods (Nelder and Pregibon, 1987) or double exponential families (Efron 1986), also exist (see Section

2.2.4). However, these models may be quite sensitive to outliers, and in addition, the computer analysis using these models requires additional programming by the researcher.

The consequences of using the wrong assumptions about overdispersion will be examined in Chapter 3. Two questions that will be addressed are, (1) What can go wrong when overdispersion is ignored altogether? and (2) What can go wrong if the dependence overdispersion on a factor or covariate is ignored?

A score test for double exponential families and a simple diagnostic plot that can be used to detect the dependence of overdispersion on covariates will be presented in Chapter 4. These depend in a simple way on statistics and residuals routinely available from a standard fit to a generalized linear model. The diagnostic tools will be applied to the examples of Section 1.2.

Chapter 2

A REVIEW OF OVERDISPERSION MODELS FOR COUNTS AND PROPORTIONS

2.1 GENERALIZED LINEAR MODELS

Probit, logistic and log-linear regression models, along with others relating a response variable from a one parameter exponential family to covariates were collected into a general structure, termed a generalized linear model by Nelder and Wedderburn (1972). If Y is a response variable from a one parameter exponential family with $E(Y) = \mu$, then a generalized linear model will describe a function of μ as a linear combination of coefficients. These models have been made popular by McCullagh and Nelder (1983) and by the widespread use of statistical computer packages such as GLIM (Baker and Nelder 1978). The general definition of a generalized linear model is given below and the special cases of logit regression and log-linear models are given as examples.

Suppose $Y_1,\ldots,Y_n$ are independent random variables with density functions given by,

$$f(y_i;\theta_i,\varphi) = \exp\{[y_i\theta_i - b(\theta_i)]/a_i(\varphi) + c(y_i,\varphi)\},$$

for some specific functions $a_i(\varphi)$, $b(\theta)$ and $c(y_i,\varphi)$. Then, $E(Y_i) = \mu_i = b'(\theta_i)$, $Var(Y_i) = b''(\theta_i)a_i(\varphi)$ and $b''(\theta_i) = V(\mu_i)$ is called the variance function. φ is called the dispersion parameter and when φ is known, $f(y_i;\theta_i,\varphi)$ is a one parameter

exponential family. The function $a_i(\phi)$ often has the form $a_i(\phi) = \phi/w_i$ where w_i is called the prior weight.

Let $\eta_i = h(\mu_i) = \underline{x}_i'\underline{\beta}$ where $h(\mu_i)$ is called the link function, $\underline{x}_i$ is a (px1) vector of explanatory variables and $\underline{\beta}$ is a (px1) vector of unknown parameters. Often η is selected so that $\eta_i = \theta_i$, and the corresponding $h(\mu)$ is called the canonical link, which has desirable statistical properties. There is often however, no *a priori* reason for using the canonical link from a data analytic viewpoint; its use may be simply a mathematical convenience and other links can be used if desired.

Maximum likelihood estimates of the β_j's can be found using Fisher's scoring method which can be carried out by iteratively weighted least squares using the working dependent variable,

$$z^t = \hat{\eta}^t + (y-\hat{\mu}^t)\left[\frac{\partial \eta}{\partial \mu}\right]\Bigg|_{\hat{\mu}^t}$$

where $\hat{\mu}^t$ and $\hat{\eta}^t$ are the estimates of μ and η after (t) iterations. The weight after (t) iterations is defined to be,

$$W_t^{-1} = [b''(\hat{\theta}^t)]\left[\frac{\partial \eta}{\partial \mu}\right]^2\Bigg|_{\hat{\mu}^t}$$

where $\hat{\theta}^t$ is the estimate of θ after (t) iterations. See McCullagh and Nelder (1983) for a full description.

An important quantity in the study of generalized linear models is the deviance function. This statistic is a generalization of the

residual sum of squares from ordinary regression models. It measures the discrepancy between the completely saturated model and the model in question. If interest lies in testing H_o: $\eta_i = \underline{x}_i'\underline{\beta}$ against H_a: $\eta_i = \gamma_i$, then the deviance statistic is the likelihood ratio test statistic for testing this hypothesis. It is given by,

$$D(y;\hat{\mu})/\varphi = \Sigma_i d(y_i;\hat{\mu})/\varphi = 2\Sigma_i [y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]/a(\varphi)$$

where $\tilde{\theta}_i$ is the estimate under H_a and $\hat{\theta}_i$ is the estimate under H_o. The deviance components, $d(y_i,\hat{\mu}_i)$, are important tools in the residual analysis of a generalized linear model and will play an important part in the diagnostic presented in Chapter 4. Asymptotically, as $m_i \to \infty$ for proportions, or for counts, as $\mu_i \to \infty$, for each i, $D(y;\hat{\mu})/\varphi \sim \chi^2_{n-p}$ and, when φ is known and the μ_i are large, the goodness of fit of the model can be evaluated by comparing the deviance to the chi-square distribution with n-p degrees of freedom.

Example 1. Logit Regression. The binomial logit regression model can be put into the generalized linear model framework. Suppose that $Y \sim \text{binomial}(m,\mu)/m$ with $\text{logit}(\mu) = \underline{x}'\underline{\beta}$. Then, the generalized linear model parameters are,

$$\theta = \text{logit}(\mu) = \ln[\mu/(1-\mu)] \quad \text{and} \quad \eta = \ln[\mu/(1-\mu)].$$

Also, $a(\varphi) = \varphi/m = m^{-1}$ and $V(\mu) = \mu(1-\mu)$ and the deviance for one observation is given by,

$$d(y;\hat{\mu}) = \{y \ln[y/\mu] + (m-y)\ln[(m-y)/(m-\hat{\mu})]\}.$$

Example 2. Poisson Log-linear Models. Log-linear models can also be cast as generalized linear models. If $Y \sim P(\mu)$ and $\ln(\mu) = \underline{x}'\underline{\beta}$ then,

$$\theta = \ln(\mu), \quad \eta = \ln(\mu), \quad a(\varphi) = \varphi = 1 \quad \text{and} \quad V(\mu) = \mu.$$

The deviance is given by,

$$d(y;\hat{\mu}) = [y \ln(y/\hat{\mu}) + (y-\hat{\mu})].$$

2.2 REVIEW OF EXISTING MODELS FOR OVERDISPERSION

The presence of overdispersed data has been recognized for a long time. Greenwood and Yule (1920) use a Poisson distribution with a gamma mixing distribution to obtain the negative binomial distribution for overdispersed counts. Cochran (1943) proposed a weighted estimator of μ to account for extra-binomial variation in fractions and percentages. Skellam (1948) introduced the beta-binomial model as a parametric model for overdispersed proportions. Since then many different types of models have been proposed to help explain observed variability that is larger or smaller than that predicted by a particular distribution.

There are two general groups of models that incorporate overdispersion. The first group contains models with explicit likelihood functions. Overdispersion is modeled either through mixing distributions or, in the case of proportions, by assuming the existence of correlations between Bernoulli trials. The second group of models is based only on assumptions about the first and second moments of the response variables.

I have distinguished between models and methods in this chapter. Models are the assumed relationships between E(Y) and Var(Y) and include likelihood functions. Examples are Models I, II and III of Section 2.2.3. Methods are the processes by which estimates of the parameters are obtained from the models. Methods discussed in Section 2.2 include maximum likelihood, maximum quasi-likelihood, maximum extended quasi-likelihood and iteratively weighted least squares.

2.2.1 Likelihood Based Models for Overdispersed Proportions

2.2.1.1 BETA-BINOMIAL MODEL

One of the oldest models for extra binomial variation is the beta-binomial model. Given observations, Y, such that $mY|P \sim \text{Binomial}(m,P)$ and $P \sim \text{Beta}(\gamma,\delta)$, the unconditional distribution of mY is described by the probability mass function

$$f_Y(y) = \binom{m}{y} \left[\frac{\Gamma(\gamma+y)\ \Gamma(\delta+m-y)}{\Gamma(\gamma+\delta+m)}\right] \left[\frac{\Gamma(\gamma+\delta)}{\Gamma(\gamma)\Gamma(\delta)}\right]$$

for $y = 0,1,\ldots,m$, where $\gamma > 0$, $\delta > 0$ and $\Gamma(\cdot)$ is the gamma function.

Define $\mu = \left[\frac{\gamma}{\gamma+\delta}\right]$, $\sigma^2 = \left[\frac{1}{\gamma+\delta+1}\right]$ and $V(\mu) = \mu(1-\mu)$. Then

$$E(Y) = \left[\frac{\gamma}{\gamma+\delta}\right] = \mu \quad \text{and}$$

$$\text{Var}(Y) = (1/m)\mu(1-\mu) + (1/m)(m-1)\sigma^2\mu(1-\mu)$$

$$= (1/m)V(\mu)[1 + \sigma^2(m-1)]$$

As either γ or δ approach infinity (which means the variance of the beta distribution goes to zero), σ^2 approaches zero and the variance of Y approaches binomial variance. But as γ and δ simultaneously approach zero, σ^2 gets large and the variance of Y is dominated by the extra-binomial variability.

To incorporate a regression model into this framework, let $\text{logit}(\mu_i) = \underline{x}_i'\underline{\beta}$ where $\underline{x}_i$ is a (px1) vector of explanatory variables and $\underline{\beta}$ is a (px1) vector of parameters. Estimates of the β_j's can be obtained by maximum likelihood.

The beta-binomial model was introduced by Skellam (1948) and since then it has been used in many different applications. It has been applied to point quadrat data by Kemp and Kemp (1956), to consumer purchasing behavior by Chatfield and Goodhart (1970) and to household incidence of disease by Griffiths (1973). Williams (1975) applied the beta-binomial model to toxicology data involving litters

of mice and Aeschbacher (1977) applied it to dominant lethal tests in mice. The beta-binomial model has been favored because the flexibility of the beta distribution allows for a wide range of shapes and because the density exists in closed form so that maximum likelihood estimates are relatively easy to obtain.

If $Y = \Sigma_i W_i$ where W_i is a Bernoulli random variable with parameter μ, then the beta-binomial model implies that all the correlations between W_i's are positive. Prentice (1986) extended the beta-binomial model to allow for negative correlations under certain conditions. Altham (1978) gave a model that allowed for positive and negative correlation between observations but this model is difficult for researchers to interpret.

2.2.1.2 CORRELATED BERNOULLI MODEL

As an alternative to the beta-binomial model, Kupper and Haseman (1978) developed a simple model for either positively or negatively correlated Bernoulli trials.

Let $Y = \sum_{i=1}^{m} Y_i$ where $Y_i \sim$ Bernoulli(μ).

If $Y_1,\ldots,Y_m$ are independent, the probability mass function for Y is given by

$$P(y) = \binom{m}{y} \mu^y (1-\mu)^{m-y} \quad \text{for} \quad y = 0,\ldots,m$$

However, if correlation exists between Y_i and Y_j,, P(y), must be multiplied by a factor to adjust for this dependence. This factor is

a function of second order, third order, on up to m^{th} order correlations where Kupper and Haseman define the p^{th} order correlation to be $E\left\{\prod_{i=1}^{m}\left[\frac{y_i-\mu_i}{\mu_i(1-\mu_i)}\right]\right\}$. The factor is complicated but an approximation can be obtained by ignoring all second order and higher corellations.

If $\rho = Cov(Y_i, Y_j,)$ and all second order and higher correlations are taken to be zero, then the approximation to the correct probability mass function is given by,

$$P_2(y) = \binom{m}{y} \mu^y (1-\mu)^{m-y}\left\{1 + \frac{\rho}{2\mu^2(1-\mu)^2}\left[(y-m\mu)^2 + y(2\mu-1) - m\mu^2\right]\right\}.$$

$P_2(y)$ is a valid probability mass function if and only if:

$$\frac{-2}{m(m-1)} \min\left[\frac{\mu}{1-\mu}, \frac{(1-\mu)}{\mu}\right] \leq \frac{\rho}{\mu(1-\mu)} \leq \frac{2\mu(1-\mu)}{(m-1)\mu(1-\mu)+.25-\gamma_o}$$

where $\gamma_o = \min\{[y - (m-1)\mu - .5]^{1/2}\}$. Kupper and Haseman (1978) give a table of permissible ranges of ρ for various choices of m and μ.

Better approximations to the true probability mass function can be obtained by including higher order correlations but Kupper and Haseman reported that $P_2(y)$ performed adequately for most of the applications they studied. Estimates of ρ and μ can be otained from maximum likelihood methods.

For the data sets given in Haseman and Soares (1976), Kupper and Haseman (1978) show that the use of this model improves the fit relative to the binomial model with independent Bernoulli trials.

2.2.1.3 CORRELATED PROBIT MODEL

Ochi and Prentice (1984) presented the following correlated probit regression model. Let $w_i = (w_{i1},\ldots w_{im_i})$ be normally distributed variates with common mean θ, variance σ^2 and correlation ρ.

Then let $Y_i = \Sigma_j Y_{ij}$ where $Y_{ij} = \begin{cases} 1 & \text{if } w_{ij} > 0 \\ 0 & \text{if } w_{ij} \leq 0 \end{cases}$.

If $\rho = 0$, then this gives the standard probit model with $Y_i \sim \text{binomial}(m_i,\mu_i)$, $E(Y) = m_i\mu_i$, $\text{Var}(Y) = m_i\mu_i(1-\mu_i)$ and $\Phi^{-1}(\mu_i) = \theta/\sigma = \underline{x}_i'\underline{\beta}$ where $\Phi(\cdot)$ is the standard normal distribution function. The parameter vector $\underline{\beta}$ maybe estimated by maximum likelihood methods.

If ρ is not zero then the correlated probability mass function for Y_i is,

$$P(y_i) = \binom{m}{y_i} \int_A f_m(w_i,0,1,\rho)\,dw$$

where $A = \left\{ v_i : w_i > -\theta/\sigma \text{ when } i \leq y; \text{ or } v_i \leq -\theta/\sigma \text{ when } i > y \right\}$

For this probability mass function,

$$E(Y) = m\mu \qquad \text{and} \qquad \text{Var}(Y) = m\mu(1-\mu)[1 + (m-1)\delta]$$

where $\mu = \Phi(\theta/\sigma)$, $\Phi^{-1}(\mu) = x_i'\beta$ and δ, the correlation parameter for Y_i, is given by,

$$\delta = \left\{ \int_{-\infty}^{0} \int_{-\infty}^{0} f(w,0,1,\rho)dw - \mu^2 \right\} \left\{ \mu(1-\mu) \right\}^{-1} .$$

Ochi and Prentice report that δ is fairly stable for μ close to 0.5. The estimates of the β_j's in this model may also be obtained using maximum likelihood methods.

Regression models for ρ may also be used and negative as well as positive correlations may be incorporated into this model. Although the full likelihood consists of the product of the $P(y_i)$'s, it is computationally difficult to maximize.

2.2.1.4 LOGIT-NORMAL MODEL

Pierce and Sands (1975) proposed the following logit regression model with fixed and random effects on the logit scale. Suppose that, $mY|P \sim$ Binomial (m,μ), and $\text{logit}(P) = \underline{x}'\underline{\beta} + u$ where $u \sim N(0,\sigma^2)$. Then $E(Y|u) = mP_u$ and $\text{Var}(Y|u) = mP_u(1-P_u)$ where $P_u = \exp(\underline{x}'\underline{\beta} + u)/[1+\exp(\underline{x}'\underline{\beta} + u)]$.

Expanding $E(Y|u)$ about $u = E(u) = 0$ and calculating the unconditional expectation and variance of Y gives,

$$E(Y) = P_0 + o_p(\sigma)$$

$$Var(Y) = (1/m)P_0(1-P_0)[1 + \sigma^2(m - 1)P_0(1-P_0)] + o_p(\sigma)$$

in an asymptotic sequence where $\sigma \to 0$ and where

$P_0 = \exp(\underline{x}'\underline{\beta})/[1+\exp(\underline{x}'\underline{\beta})]$.

The unconditional variance of Y can be thought of as binomial variance plus an extra-binomial term that depends on σ^2. This model can be more directly interpreted than other models since the random variation occurs on the same scale as the covariates and the model allows for the incorporation of complicated randomization schemes. However, the likelihood is computationally difficult to maximize.

2.2.2. Likelihood Based Models for Extra-Poisson Variation

2.2.2.1 NEGATIVE BINOMIAL MODEL

For overdispersed count data, suppose $Y|U \sim$ Poisson(U) and $U \sim$ Gamma(γ, δ). Then the unconditional distribution of Y is

$$f_Y(y) = \left[\frac{\Gamma(\gamma+y)}{\Gamma(\gamma)\Gamma(y+1)}\right]\left[\frac{\delta}{\delta+1}\right]^y \left[\frac{1}{\delta+1}\right]^\gamma \quad \text{for } y = 0,1,2,\ldots$$

which is the negative binomial probability mass function.

If the probability mass function above is reparameterized so that $\mu = \gamma\delta$ and $\sigma^2 = \delta$, then unconditionally

$$E(Y) = E[E(Y|U)] = E(U) = \gamma\delta = \mu ,$$

$$\begin{aligned} Var(Y) &= E[Var(Y|U)] + Var[E(Y|U)] \\ &= E(U) + Var(U) \\ &= \gamma\delta + \gamma\delta^2 = \gamma\delta\ (1 + \delta) \\ &= \mu(1 + \sigma^2). \end{aligned}$$

This is analogous to the mean/variance relationship that arose from the beta-binomial model and the correlated probit and Bernoulli models.

Alternatively, if the model is parameterized so that $\gamma\delta = \mu$ and $1/\gamma = \sigma^2$ then

$$E(Y) = E[E(Y|U)] = E(U) = \gamma\delta = \mu ,$$

$$\begin{aligned} Var(Y) &= E[Var(Y|U)] + Var[E(Y|U)] \\ &= E(U) + Var(U) \\ &= \gamma\delta + \gamma\delta^2 = \gamma\delta\ (1 + \gamma\delta/\gamma) \\ &= \mu\ (1 + \mu\sigma^2) . \end{aligned}$$

This mean/variance relationship has a similar structure to that which arises by modeling the random effects on the same scale as the covariates for the mean in the logit normal model. Maximum

likelihood techniques are used to find estimates of μ and σ^2 for either parameterization.

Greenwood and Yule (1920) introduced this model and Skellam (1948) showed how it is obtained as the limiting form of the beta-binomial distribution. McCaughran and Arnold (1976) presented this model for use in the study of embryonic deaths in mice and Moore (1985) gives a full description of this model and its score functions. Collings and Margolin (1985) proposed tests for extra Poisson variation based on the negative binomial distribution when (1) the mean was constant, (2) the mean depended on a single covariate and the regression line passed through the origin and (3) the mean took on a fixed number of values according to a one-way layout. Dean and Lawless (1989) extended these tests to include arbitrary Poisson regression models.

2.2.3 Models Based Only on First and Second Moment Assumptions

2.2.3.1 HISTORICAL RESULTS FOR I.I.D. SETTINGS

Cochran (1943) recognized extra-binomial variation in data that was reported as fractions and percentages. He proposed

$$\text{Var}(Y) = \frac{\mu(1-\mu)}{m} + \sigma^2$$

as a model for the variance of the proportion Y, where σ^2 is the extraneous variance. He noted that while it is not always clear what assumptions can be made about σ^2, there is probably not one set of assumptions for all data sets.

Given observed percentages, Y_i, based on m_i trials, with $E(Y_i) = \mu$, for $i = 1,\ldots,n$, Cochran proposed a weighted estimator of μ,

$$\hat{\mu} = \sum_{i=1}^{n} \frac{w_i y_i}{\Sigma w_i}$$

where w_i could be either (1) the binomial sample sizes m_i, (2) 1, or (3) m_i for the lower third of the ordered m_i and $\min(m_i)$ for the upper two thirds or the ordered m_i, where the minimum is over the upper two thirds of the binomial samples. The efficiencies of these weighting schemes depend on the proportions of binomial and extraneous variation that are present.

Kleinman (1973) extended Cochran's ideas by estimating σ^2. He supposed that given a percentage Y based on m trials, $mY|P \sim$ binomial(m,P), $E(P) = \mu$ and $Var(P) = \sigma^2\mu(1-\mu)(1/m)$. It can be noted that these assumptions are analogous to those made in deriving the beta binomial distribution. However, the form of the unconditional distribution is not specified here. Unconditionally, $E(Y_i) = \mu$ and $Var(Y_i) = \mu(1-\mu)(1/m_i)[1+\sigma^2(m_i-1)]$.

Defining $\hat{\mu} = \sum_i \frac{w_i y_i}{\Sigma w_i}$ and $S = \sum_i w_i (y_i-\hat{\mu})^2$, Kleinman set $\hat{\mu}$ and S equal to their expected values and solved for $\hat{\sigma}^2$. Then he proposed the following scheme for estimating w_i.

1. Letting $w_{i1} = 1$ or m_i, use the equations for $\hat{\mu}$ and S to solve for $\hat{\sigma}^2$.
2. Let $w_{i2} = m_i/[1+\hat{\sigma}^2(m_i-1)]$ and evaluate $\hat{\mu}$ with $w_i = w_{i2}$.

Finney (1971), working in the context of probit analysis, also recognized the presence of overdispersion. If Y is an observed

proportion and the mean of Y is estimated correctly, then, in the absence of overdispersion, the generalized Pearson chi-squared statistic $\hat{\chi}^2$ should, on the average, equal its degrees of freedom, (f). Finney suggested that a significantly large value of $\hat{\chi}^2$ would imply that the variance of Y was inflated by a heterogeneity factor, h, and h could be estimated by $\hat{\chi}^2/f$. Finney noted that it is difficult to distinguish between overdispersion and inadequacies in the model for the mean.

2.2.3.2 REGRESSION SETTINGS AND QUASI-LIKELIHOOD MODELS

A general class of regression models in which the variance of the response variable is proportional to a function of its mean was described by Wedderburn (1974). Given $Y_1,\ldots,Y_n$ independent observations such that $E(Y_i) = \mu_i$, $\eta_i = h(\mu_i) = \underline{x}_i'\underline{\beta}$ and $Var(Y_i) = \sigma^2 V(\mu_i)$, where $\underline{x}_i$ is a (px1) vector of covariates, the quasi-likelihood function $\Sigma_i Q(y_i,\mu_i)$ is defined by:

$$\Sigma_i \frac{\partial}{\partial \mu_i} Q(y_i,\mu_i) = \Sigma_i \frac{y_i-\mu_i}{V(\mu_i)} .$$

The maximum quasi-likelihood estimate of the vector $\underline{\beta}$ is the $\hat{\underline{\beta}}$ such that $\Sigma_i Q(y_i,\hat{\mu}_i) \geq \Sigma_i Q(y_i,\mu_i)$ for all μ where $g(\mu) = \underline{x}_i'\underline{\beta}$ and $g(\hat{\mu}) = \underline{x}_i'\hat{\underline{\beta}}$.

Wedderburn showed that $\Sigma_i Q(y_i,\mu_i)$ has many properties similar to log-likelihoods. He showed that a quasi-likelihood function is identical to a likelihood function if and only if the distribution of Y is from the exponential family. Thus if the mean/variance

relationship is known and is the same as a known exponential family (up to a multiplicative constant in the variance) then the maximum quasi-likelihood estimates of μ are identical to the maximum likelihood estimates. Thus quasi-likelihood estimation generalizes maximum likelihood estimation of generalized linear models in the same way that least squares estimation generalized maximum likelihood estimation for normal theory regression.

The dispersion parameter, σ^2 is estimated separately. The suggested estimator of σ^2,

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i-\mu_i)^2}{V(\mu_i)} = \hat{\chi}^2/(n-p)$$

is the generalized Pearson chi-squared statistic divided by its degrees of freedom, as in the method suggest by Finney (see Section 2.2.3). Since σ^2 is estimated separately from the β_j's, quasi-likelihood models can be fit in the framework of generalized linear models using Fisher's scoring method (McCullagh and Nelder, 1983). When the assumed mean/variance relationship is the same as that of a one parameter exponential family except for the multiplicative constant σ^2, the maximum quasi-likelihood estimates of the β_j's are identical to those obtained by maximum likelihood under the corresponding one parameter exponential family model but the estimated standard errors of the estimates must be scaled by $\hat{\sigma}$.

McCullagh (1983) studied the asymptotic properties of quasi-likelihood estimators and showed that among all estimators of $\underline{\beta}$

for which the influence function is linear, quasi-likelihood estimates have minimum asymptotic variance. Firth (1987) investigated the efficiency of maximum quasi-likelihood estimators in the presence of overdispersion with respect to maximum likelihood estimators and found that for mixing distributions with regular cumulant behavior maximum quasi-likelihood estimators are greater than 90% efficient if σ^2 is less than 1.3.

Williams (1982) summarized logistic regression models for proportions given below and provided macros to fit these models in GLIM (Baker and Nelder, 1978).

Model I. $E(Y) = \mu \quad \text{logit}(\mu) = \underline{x}'\underline{\beta} \quad \text{Var}(Y) = \mu(1-\mu)/m$

Model II. $E(Y) = \mu \quad \text{logit}(\mu) = \underline{x}'\underline{\beta}$

$$\text{Var}(Y) = \mu(1-\mu)/m\,[1 + \sigma^2(m-1)]$$

Model III. $E(Y) = \mu \quad \text{logit}(\mu) = \underline{x}'\underline{\beta}$

$$\text{Var}(Y) = \mu(1-\mu)(1/m)[1+\sigma^2(m-1)\mu(1-\mu)]$$

Model I is the binomial model without overdispersion. Model II contains the same mean/variance relationship that arose using the beta-binomial distribution or the correlated Bernoulli or the correlated probit distribution. Model III contains approximately the same mean/variance relationship as the logit normal distribution discussed in Section 2.2.1. It arises by modeling random variation

on the same scale as covariates. It will be described further in the next section.

Similarly, three models for overdispersed counts can be provided.

Model I'. $E(Y) = \mu \quad \log(\mu) = \underline{x}'\underline{\beta} \quad Var(Y) = \mu$

Model II'. $E(Y) = \mu \quad \log(\mu) = \underline{x}'\underline{\beta} \quad Var(Y) = \mu\,[1 + \sigma^2]$

Model III'. $E(Y) = \mu \quad \log(\mu) = \underline{x}'\underline{\beta} \quad Var(Y) = \mu[1+\sigma^2\mu]$

Model I' is the Poisson model without overdispersion. Model II' contains the same mean/variance relationship as the negative binomial model and the correlated models of Section 2.2.1. Similar to the model for proportions, Model III' arises by modeling random variation on the same scale as the covariates. Model II' and Model III' can be obtained from the negative binomial model for counts as discussed in Section 2.2.2.

Breslow (1984) discussed model III' for overdispersed Poisson data. Given an observed count, d, with fixed denominator m, and an unknown rate parameter λ assume $d|\lambda \sim$ Poisson $(m\lambda)$ and $\log(\lambda) = \underline{x}'\underline{\beta} + u$ where $E(u) = 0$ and $Var(u) = \sigma^2$. This is analogous to Williams' model III in that variability occurs on the same scale as the covariates.

Then, if d is large, log(d/m) has an approximate normal distribution with mean $\underline{x}'\underline{\beta}$ and variance $[\sigma^2 + E(d)^{-1}]$. Estimates of

the β_j's can be obtained using weighted least squares with weights $(\sigma^2 + d^{-1})$.

For small d, Breslow suggested expanding $E(d|\lambda m)$ about $u = E(u) = 0$ to obtain,

$$E(d) = \exp\{\log(m) + \log[E(\lambda)]\} = \exp\{\log(m) + \underline{x}'\underline{\beta}\}$$
$$= \mu$$
$$Var(d) = E[Var(d|m,\lambda)] + Var[E(d|m,\lambda)]$$
$$\approx \mu(1 + \sigma^2\mu).$$

This model can be fit using quasi-likelihood methods and the estimated prior weight, $(1 + \hat{\sigma}^2\hat{\mu}_i)^{-1}$, where $\hat{\sigma}^2$ is the generalized Pearson chi-squared statistic divided by its degrees of freedom.

Moore (1987) presented an extension of the quasi-likelihood method for modeling the variance of overdispersed proportions. Under his model, $E(Y) = \mu$, $h(\mu) = \underline{x}'\underline{\beta}$ and

$$Var(Y) = (1/m)\mu(1-\mu)[1 + \sigma^2\mu^{\xi}(1-\mu)^{\xi}]$$

where ξ is an additional parameter. The case of $\xi = 2$ corresponds approximately to Williams' model III. Moore suggests that an appropriate value of ξ may be chosen by examining residuals or by minimizing

$$Q(\xi) = \Sigma_i \{\hat{e}_i^2 - \hat{\mu}_i(1-\hat{\mu}_i)(1/m_i) - \hat{\sigma}^2\hat{\mu}_i^{\xi}(1-\hat{\mu}_i^{\xi})\}^2$$

where $\hat{e}_i^2 = (y_i - \hat{\mu}_i)$ and $\hat{\mu}_i$ and $\hat{\sigma}^2$ are obtained using the quasi-likelihood method.

2.2.3.3 REGRESSION SETTINGS FOR MODELS III AND III'

In addition to the logit normal model discussed in 2.2.1, Pierce and Sands (1975) presented Model III above, in the context of logit regression. They modeled random effects as additive on the same scale as the covariates without specifying a distribution for u.

In a derivation similar to the one in 2.2.1, suppose that, $Y|P_u \sim$ Binomial $(m,P_u)/m$, and $\text{logit}(P_u) = \underline{x}'\underline{\beta} + u$ where $E(u) = 0$ and $\text{Var}(u) = \sigma^2$. Expanding $E(y|u)$ about $u = E(u) = 0$ and calculating the unconditional variance of Y gives,

$$(1/m)E(Y) = P_0 + o_p(\sigma) ,$$

$$\text{Var}(Y) = (1/m)P_0(1-P_0)[1 + \sigma^2(m - 1)P_0(1-P_0)] + o_p(\sigma)$$

in an asymptotic sequence where $\sigma \to 0$ and where $P_0 = \exp(\underline{x}'\underline{\beta})/[1+\exp(\underline{x}'\underline{\beta})]$. This model can incorporate variability due to omitted variables, random effects and complicated randomization schemes.

In the absence of overdispersion, where $u_i = 0$ for all i, this is the usual logit regression model and $\hat{\underline{\beta}}$, the maximum likelihood estimator of $\underline{\beta}$ can be found using iteratively weighted least squares.

For $u \sim N(0,\sigma^2)$, Pierce and Sands show how to find the maximum likelihood estimates using numerical quadrature. On the basis of simplicity, they suggested estimating σ^2 with,

$$\hat{\sigma}^2 = \left[\left(\frac{\Sigma_i (y_i-\hat{y})^2}{n-m}\right) - 1\right]\left[\frac{\Sigma_i m_i \hat{P}_i (1-\hat{P}_i)}{\Sigma_i m_i}\right]^{-1}.$$

where logit $(P_i) = \underline{x}'\underline{\beta}$. This estimator is inadmissible, although Pierce and Sands say that it fails only slightly to be admissible for the examples they have studied. Unlike the maximum likelihood estimate, this unbiased estimator can be explicitly evaluated once P_i is estimated and its variance function is similar to that of maximum likelihood estimators for moderate σ^2.

2.2.4 Models that Incorporate Covariates Into the Variance

2.2.4.1 EXTENDED QUASI-LIKELIHOOD

Quasi-likelihood methods provide for a dispersion parameter that is constant for all observations in a dataset and the model given by Moore (1987) allows for a dispersion parameter that depends on the mean and one additional parameter. Based on the discussion of Chapter 1, it may be reasonable to model the dispersion parameter as a function of known covariates which may or may not include the mean. Nelder and Pregibon (1987) introduced extended quasi-likelihood functions in order to allow for comparisons of link functions, linear

predictors and variance functions between competing models as well as regression models for the dispersion parameter.

Suppose $Y_1,\ldots,Y_n$ are independent observations with $E(Y_i) = \mu_i$, $Var(Y_i) = \sigma^2_i V(\mu_i)$ and deviance components $d(Y_i;\mu_i)$. The extended quasi-likelihood function is defined to be

$$Q^+(y,\underline{\mu},\underline{\sigma}^2) = -1/2\ \Sigma_i\ [\log\{2\pi\sigma^2_i V(y_i)\} - d(y_i;\mu_i)/\sigma^2_i]$$

The maximum quasi-likelihood estimates of μ are the estimates which maximize Q^+. If $\sigma^2_i = \sigma^2$ for all i, the estimate of σ^2 obtained by maximizing Q^+ is $\hat{\sigma}^2 = \Sigma_i d(y_i;\hat{\mu}_i)/n$. Notice that the extended quasi-likelihood function depends only on first and second moment assumptions.

Now suppose $\eta_i = h(\mu_i) = \underline{x}_i'\underline{\beta}$ and $w(\sigma^2_i) = \underline{z}_i'\underline{\alpha}$ where $\underline{x}_i$ is a (px1) vector of covariates for the mean, $\underline{\beta}$ is a (px1) vector of unknown parameters for the mean, $\underline{z}_i$ is a (qx1) vector of covariates for the dispersion parameter, $\underline{\alpha}$ is a (qx1) vector of unknown paramters for the dispersion parameter, $h(\cdot)$ is the link function for the mean and $w(\cdot)$ is the link function for the dispersion parameter. Nelder and Pregibon suggest the following scheme for finding the estimates that maximize Q^+.

1. Hold σ^2_i fixed at $\hat{\sigma}^2_{i0}$. Using $\hat{\sigma}^2_{i0}$ as a prior weight and fitting the generalized linear model with $E(Y_i) = \mu_i$, $Var(Y_i) = \hat{\sigma}^2_{i0} V(\mu_i)$, $h(\mu_i) = \underline{x}_i'\underline{\beta}$, obtain the updated estimates of μ_i, $\hat{\mu}_{i1}$,.

2. Holding μ_i fixed at $\hat{\mu}_{i1}$, σ_i^2 is estimated by fitting the generalized linear model with $d(y_i;\hat{\mu}_{i1})$ as the dependent variable with $E(d(y_i;\hat{\mu}_{i1}) = \sigma_i^2$, $w(\sigma_i^2) = \underline{z}_i'\underline{\alpha}$ and $V(\sigma_i^2)=(\sigma_i^2)^2$.

3. Hold σ_i^2 fixed at the new estimate, $\hat{\sigma}_{i1}^2$, obtained in step 2, and refit the generalized linear model in step 1 with $\hat{\sigma}_{i_0}^2$ replaced with $\hat{\sigma}_{i1}^2$.

Iterate between steps 2 and 3 until convergence. The standard errors obtained at each step for each set of parameter estimates are conditional on the values of the other set being equal to their estimates. This procedure is relatively simple to program using existing computer software.

For a single observation, y, if an extended quasi-likelihood model is chosen with the same variance function as the inverse Gaussian or normal distributions, then Q^+ is the log-likelihood function for that exponential family. For the gamma distribution, Q^+ differs from the log-likelihood by a factor that depends on σ^2. For the negative binomial, Poisson or binomial distributions, Q^+ can be obtained from the log-likelihood by replacing k! with $(2\pi k)^{1/2}k^k e^{-k}$ (Nelder and Pregibon, 1987).

By multiplying $\exp(Q^+)$ by a normalizing factor $c(\mu,\sigma^2)$, a distribution can be formed. However Nelder and Pregibon argue that since $c(\mu,\sigma^2)$ contains little information about μ or σ^2, very little is lost in maximizing the unnormalized extended quasi-likelihood. This is similar to the situation that occurs in the double exponential families presented below.

2.2.4.2 DOUBLE EXPONENTIAL FAMILIES

Double exponential families were presented by Efron (1986) in order to allow for the estimation of a dispersion parameter independent of the mean which could depend on covariates. In Efron's words, double exponential families are a way to take the "quasi" out of quasi-likelihood. Let Y have a one parameter exponential family distribution with density $f(y;\mu,)$ given by,

$$f(y;\mu) = \exp\left\{[y\theta - b(\theta)]m + c(y)\right\}$$

where $\theta = \theta(\mu)$ is the canonical parameter, $E(Y) = \mu$ and $Var(Y) = V(\mu)$ and m is a known constant. The double exponential family density is defined to be,

$$g(y;\mu,m) = c(\mu,\varphi,m)\ \varphi^{1/2}\ \left\{f_Y(y;\mu)\right\}^{\varphi}\ \left\{f_Y(y;y)\right\}^{1-\varphi}.$$

Efron (1986) shows that under the double exponential family, $E(Y) \approx \mu$, $Var(Y) \approx V(\mu)m/\varphi$ and $c(\mu,\varphi,m) \approx 1$. In addition, with μ and m fixed, g(y) is an exponential family with parameter φ. Notice that if a quasi-likelihood model was used to describe this mean/variance relationship, the quasi-likelihood dispersion parameter, σ^2, corresponds to φ^{-1}, where φ is the double exponential family dispersion parameter.

The deviance for one observation, y, is defined to be

$$D(y;\hat{\mu}) = 2\{\log [f(y;y)] - \log [f(\hat{\mu};y)] \}$$

where $\hat{\mu}$ is the maximum likelihood estimate of μ under the model of interest and $D(y,\hat{\mu}) \sim \chi^2_1$ as $\mu \longrightarrow \infty$, or as $m \longrightarrow \infty$ for proportions.

Also, $I_{\mu,\varphi}$, the expected Fisher Information matrix for μ and φ (m fixed) is approximately

$$I_{\mu,\varphi} = \begin{bmatrix} m\varphi/V(\mu) & 0 \\ 0 & (2\varphi^2)^{-1} \end{bmatrix}.$$

Regression models for both μ and φ can be incorporated into double exponential families. Let $\eta_i = h(\mu_i) = \underline{x}_i'\underline{\beta}$ and $\xi = w(\varphi_i) = \underline{z}_i'\underline{\alpha}$ where $h(\cdot)$ and $w(\cdot)$ represent link functions for the mean and dispersion parameter respectively. The expected Fisher Information matrix for $\underline{\alpha}$ and $\underline{\beta}$ is given by

$$I_{\underline{\alpha},\underline{\beta}} = \begin{bmatrix} \underline{X}'\underline{W}\underline{X} & 0 \\ 0 & 1/2\ \underline{Z}'\underline{V}\underline{Z} \end{bmatrix} = \begin{bmatrix} \underline{I}_{\underline{\beta}} & 0 \\ 0 & \underline{I}_{\underline{\alpha}} \end{bmatrix} \quad \text{where,}$$

$\underline{X}$ is the (nxp) matrix with row $\underline{x}_i'$,

$\underline{Z}$ is the (nxq) matrix with row $\underline{z}_i'$,

$$\underline{W} = \text{diag}\left[\left(\frac{b''(\theta_i)m_i}{\varphi_i}\right)^{-1}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2\right], \quad \text{and}$$

$$\underline{V} = \text{diag}[\varphi_i^{-2}\{\partial\varphi_i/\partial(\underline{z}_i'\underline{\alpha})\}] \ .$$

Fisher's scoring method can be used to find the maximum likelihood estimates of the α_j's and β_j's. If $\underline{\alpha}^t$ and $\underline{\beta}^t$ are the vectors of estimates after the (t) iteration, then the improved estimates, $\underline{\alpha}^{t+1}$ and $\underline{\beta}^{t+1}$, are given by,

$$\underline{\alpha}^{t+1} = \underline{\alpha}^t + \underline{\mathcal{I}}_{\underline{\alpha}}^{-1}(\underline{\partial l/\partial \alpha}^t)$$
$$\underline{\beta}^{t+1} = \underline{\beta}^t + \underline{\mathcal{I}}_{\underline{\beta}}^{-1}(\underline{\partial l/\partial \beta}^t)$$

where $\underline{\partial l/\partial \alpha}^t$, $\underline{\partial l/\partial \beta}^t$ are the score vectors evaluated at $\underline{\alpha}^t$ and $\underline{\beta}^t$.

Double exponential families are similar to extended quasi-likelihood models presented by Nelder and Pregibon (1987). It can be shown that the double exponential family log-likelihood is equal to the extended quasi-likelihood except for an additive term that does not depend on the parameters. Thus, estimates which maximize the extended quasi-likelihood are identical to estimates which maximize the double exponential family likelihood. Double exponential families are also similar to West's (1985) scaled exponential family, and to Jorgensen's (1987) exponential dispersion model.

2.3 DISCUSSION OF EXISTING MODELS AND METHODS

The models given in Section 2.2 have been derived in an attempt to find an understandable way to explain extra-Poisson or

extra-binomial variability. Some likelihood-based models have added parameters to account for correlations that induce overdispersion; other have incorporated mixing distributions to model variability in parameters. The models based on first and second moment assumptions such as quasi-likelihood models, have traded some efficiency for the desirable property of robustness. In addition, the availability of GLIM (Baker and Nelder, 1978) has made quasi-likelihood models very easy to use in practice and estimates are available with a minimum of time invested in programming.

Before discussing the models presented in 2.2, however, two other methods, a jackknife estimator and transformations, for handling overdispersed data are presented below.

2.3.1 Other Methods

Gladen (1979) proposed a jackknife estimator for μ, the true proportion of affected fetuses out of m_i fetuses in a litter in teratological experiments. If M is $\Sigma_i m_i$ and r_i is the proportion of affected fetuses in the i^{th} litter where $i = 1,\ldots,n$ and $\hat{p}$ is the estimate of μ and if $y_i = m_i(r_i - \hat{p})(M - m_i)^{-1}$, then Gladen estimated the variance of $\hat{\mu}$ with

$$\mathrm{Var}(\hat{\mu}) = M^{-1}\,(M - 1)\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Gladen reported that the jackknife estimates of μ are almost fully efficient with respect to the maximum likelihood estimates under various models.

Another method of analysis for data which exhibit extra variation which has not been discussed previously involves the use of transformations. When it is desired to use analysis of variance or regression techniques with proportions or counts, transformations such as the Freeman-Tukey binomial or Freeman-Tukey Poisson transformation are often used to stabilize variance and/or transform the data to approximate normality. See Mosteller and Youtz (1961) for a description of the transformations. Because the transformations are relatively easy to apply and ANOVA and regression techniques are well known, such analyses are often carried out in practice. But such an analysis assumes constant variance and using transformations in a situation where overdispersion varies from group to group may result in incorrect inference.

2.3.2 Comparison of Models

Some results are available on the efficiency of maximum likelihood estimates and maximum quasi-likelihood estimates. Firth (1987) studied the asymptotic relative efficiency of quasi-likelihood estimates when the mean/variance relationship arose from overdispersion relative to an exponential family. If $\mathrm{Var}(Y) = V(\mu)$ under the exponential family and $\mathrm{Var}(Y) = \sigma^2 V(\mu)$ under the corresponding quasi-likelihood model, he found that for any mixing distribution with regular cumulant behavior, maximum quasi-likelihood estimation has efficiency greater than 90% if $\sigma^2 < 1.3$.

Lawless (1987) studied the robustness of the maximum likelihood estimator, $\hat{\beta}$, obtained from the negative binomial distribution (see

2.2.2) with a constant dispersion parameter when the negative binomial assumption was wrong. For the cases, $\mu_i = \mu$ for all i, and $\mu_i = \exp[\beta_0 + \beta_1 x_i]$ where one third of the x's are 0, one third are 1 and one third are -1, he compared the covariance matrix given under the negative binomial distribution assumption with the true covariance matrix given by White (1982). He found that the incorrect maximum likelihood procedure slightly underestimated the true variance in large samples.

Kupper *et al.* (1986), Williams (1988) and Pack (1986) studied bias and hypothesis testing for beta-binomial models fit using maximum likelihood. Kupper *et al.* (1986) used the beta-binomial model to fit dose response regressions to the proportion of affected fetuses in teratology experiments. They found that the maximum likelihood estimator obtained from this model become biased when it is assumed that the intra-litter correlations are homogeneous. They used a simulation study where the number of affected fetuses, Y_{ij}, i=1,2,3 and j = 1,...,n_i, had a beta-binomial distribution with $E(Y_{ij}) = m_{ij}\mu_i$, and $Var(Y_{ij}) = m_{ij}\mu_i(1-\mu_i)[1 + \sigma_i^2(m_i-1)]$ where σ_i^2 is the intra-litter correlation for the ith group. The regression model was taken to be

$$\log[\mu_i/(1-\mu_i)] = \beta_0 + \beta_1 \ln d_i$$

for ln d_i = 1,2,3. They found that if it was assumed that $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$ then $\hat{\beta}_1$ is negatively biased if $\sigma_1^2 < \sigma_2^2 < \sigma_3^2$ and positively biased if $\sigma_1^2 > \sigma_2^2 > \sigma_3^2$.

Williams (1988), investigating these results, suggested that for a single group with mean $\mu < 0.5$, the maximum likelihood estimator, $\hat{\mu}$, will be negatively biased if it is assumed that σ^2 is smaller than the true σ_i^2 and positively biased if it is assumed that σ^2 is larger than the true σ_i^2. The results would be reversed for $\mu > 0.5$. Williams showed that $\hat{\mu}$ is approximately unbiased when it is assumed that σ^2 is equal to zero or when it is equal to its true value. He suggested regressing σ_i^2 on d_i or using Moore's (1987) method to reduce the number of parameters in the model. He also notes that the bias could be eliminated by using the more robust quasi-likelihood model.

Pack (1986) studied power and type 1 error rates for likelihood ratio tests under the beta-binomial model for the hypotheses,

H1	$\mu_1 = \mu_2 \quad \sigma_1^2 = \sigma_2^2$	v.s.	H1	$\mu_1 \neq \mu_2 \quad \sigma_1^2 \neq \sigma_2^2$
H2	$\mu_1 = \mu_2 \mid \sigma_1^2 = \sigma_2^2$	v.s.	H2	$\mu_1 \neq \mu_2 \mid \sigma_1^2 \neq \sigma_2^2$
H5	$\mu_1 = \mu_2$	v.s.	H5	$\mu_1 \neq \mu_2$

He found that, in general, the likelihood ratio test of H5 had acceptable error rates for all (μ, σ^2) combinations considered and it was the most powerful in a broad range of situations. He compared the above test with the Student's T-test on the Freeman-Tukey transformed data and Kleinman's (1973) weighted estimator and found that for small differences in means (about 0.02) none of the tests had a clear advantage. He also gives a good summary of prior studies of control versus treatment comparisons for reproductive studies.

Results from Table 2 of Pack (1986) showed that if $\mu_1 < \mu_2$ and $\sigma_1^2 < \sigma^2_2$ and the null hypothesis $\mu_1 = \mu_2$ is tested then the likelihood ratio test assuming $\sigma_1^2 \neq \sigma_2^2$ is more powerful than the likelihood ratio test assuming $\sigma_1^2 = \sigma_2^2$. Williams (1988) notes that this power difference can be partially attributed to the underestimation of $\mu_1 - \mu_2$ when σ_1^2 and σ_2^2 are incorrectly assumed to be equal.

These results suggest that using maximum likelihood estimates from incorrectly specified distributions can lead to problems. The loss in efficiency of maximum quasi-likelihood estimation does not seem to be a large problem. However, the results of Kupper *et al.* (1986) and Williams (1988) suggest that making the incorrect assumption of homogeneous correlations can in the case of the beta-binomial model lead to incorrect inference about the μ_i's.

Moment methods such as quasi-likelihood provide estimates with minimal assumptions about the distribution and hence are robust alternatives to the use of specific likelihood functions. The loss in efficiency due to using maximum quasi-likelihood methods over maximum likelihood seems to be small and well worth the gain in robustness especially since it is often difficult to specify the exact form of the distribution with much confidence. The robustness of quasi-likelihood models and the ease with which they can be fit make them attractive models to use in practice.

Once it has been decided to use a method based on first and second moment assumptions it becomes necessary to choose the mean/variance relationship. The relationships considered here are

given by Models I, II and III and I',II' and III' of Section 2.2.3. Model I and I' can be obtained when extra variation is not present and should be used in the absence of overdispersion. However, as McCullagh and Nelder (p. 127, 1983) note, it is wise to assume that overdispersion is present unless there is strong evidence to the contrary. Models II and II' can be obtained by hypothesizing the existence of a constant correlation between observations. Models III and III' came about by modeling random variation on the same scale as the covariates for the mean. In designed experiments or observational studies variability due to random effects are often thought to be additive on the same scale as the covariates. Model III or III' is appropriate in this case. In the discussion to Diaconis and Efron (1985), Pierce notes that when the binomial sample sizes are very different model II or II' may not be a reasonable model to use. He suggested model III or III' as an alternative.

Williams (1982) notes that the effective difference between model II and model III is a factor of $\mu_i(1-\mu_i)$ in the weight term and this factor is relatively constant for μ_i between 0.2 and 0.8. He suggests that it will only be possible or important to distinguish between model II and model III if there are a substantial number of observations for which μ_i is close to zero or one.

If all the m_i are approximately equal, which is often the case in data analysis, than model II or II' can be reparameterized as

$$\mathrm{Var}(Y_i) = V(\mu)\ \varphi^{-1},$$

where $\varphi^{-1} = [1 + \sigma^2(m-1)]$ and such a model can be fit using quasi-likelihood methods. A generalization of Model II is,

Model IV. $E(Y_i) = \mu_i \quad h(\mu_i) = \mathbf{x}_i'\beta \quad Var(Y_i) = V(\mu_i)\varphi(\underline{z}_i'\underline{\alpha})^{-1}$

where φ^{-1} is allowed to vary between observations according to the covariate vector $\underline{z}_i$.

Chapter 3 will investigate the consequences of using the model with $Var(Y) = V(\mu)\varphi^{-1}$ when model IV is more appropriate. It will also be shown how model III or III' can be written as model IV. Hence the consequences of using the model with $Var(Y) = V(\mu)\varphi^{-1}$ when model III or III' is more appropriate can also be investigated.

2.3.3. Covariates in the Variance Function

In the previous sections, the variance of Y has been modeled generally as $V(\mu)\varphi_i^{-1}$ where $V(\mu)$ is called the variance function and φ_i^{-1} has been written as σ^2, $[1 + \sigma^2(m-1)]$ or $[1 + \sigma^2(m-1)V(\mu)]$. In all of the models discussed so far the variance function has been assumed known and it depends on the covariates $\underline{x}$ only through $h(\mu) = \underline{x}'\underline{\beta}$.

One class of models which has not been discussed contains models of the form,

Model V. $E(Y) = \mu \quad h(\mu) = \underline{x}'\underline{\alpha} \quad Var(Y) = \sigma^2 V(\underline{z}, \underline{\alpha}, \theta)$

where the variance function $V(\underline{z}, \underline{\alpha}, \theta)$ depends on the mean μ, not only through the vector $\underline{\alpha}$ but it also depends on the unknown parameter θ

and the known vector of covariates $\underline{z}$ (which may or may not include $\underline{x}$). Models of this type and methods of estimation are discussed by Davidian and Carroll (1987).

In many applications, especially in the area of quality control, interest lies not only in the mean response, but also in patterns of variability and the factors which affect variability. For such applications, model V is very useful. However, for the problems discussed in this thesis, interest lies mainly with inference about the mean. Accounting for variability and overdispersion are important but approximations to the exact form are acceptable for this type of problem.

For example, Efron (1986) using a binomial double exponential family distribution to model the proportion of subjects testing positive for toxoplasmosis, as a function of rainfall, used a quadratic function of the binomial sample sizes in the regression model for the dispersion parameter. Such a model does not give clear insights into the patterns of variability but it serves as a good approximation for estimating standard errors.

In the rest of the thesis, models for the variance of Y having the form, $V(\mu)\varphi_i^{-1}$, will be investigated. Various forms for φ_i^{-1}, corresponding to models I, II and III discussed earlier will be used. Chapter 3 will explore the consequences of using simple forms for φ_i^{-1} when in fact more complicated forms are appropriate. Diagnostic tools to help decide when φ_i^{-1} does not have a simple form will be presented in Chapter 4 and the tools will be applied to the examples given in Chapter 1.

Chapter 3

CONSEQUENCES OF USING INCORRECT ASSUMPTIONS ABOUT OVERDISPERSION

Many models and methods exist that allow a researcher to incorporate overdispersion into an analysis. In some cases the design of the experiment or prior knowledge held by the researcher dictate which model for overdispersion is chosen. In other cases the researcher settles on a method out of convenience or because there is no evidence to support a particular model.

It is possible that the chosen model is not the most appropriate one and that the assumptions upon which the chosen model is based do not apply. It is of interest to know the extent to which inference based, incorrect assumptions affect the estimated coefficients and the associated tests.

In this chapter the following four models for overdispersion will be discussed. Suppose that $E(Y) = \mu$ and $h(\mu) = \underline{x}'\underline{\beta}$ define the regression model for the mean and consider the models below for the variance of Y.

Model 0. $\quad Var(Y) = V(\mu)$

Model 1. $\quad Var(Y) = \varphi(\alpha_0)^{-1}V(\mu)$

Model 2. $\quad Var(Y) = \varphi(\alpha_0 + \underline{z}_i'\underline{\alpha}_1)^{-1}V(\mu)$

Model 3. $\quad Var(Y) = V(\mu)(1/m)[1 + \sigma^2 k(m)V(\mu)]$

$V(\mu)$ is the variance function, $\underline{z}_i$ is a (qx1) vector of covariates for the dispersion parameter φ, $\underline{\alpha}_1$ is a (qx1) vector of unknown parameters, α_0 is an unknown scalar parameter and k(m) is a function

of the known m's. For counts, k(m) is equal to 1 and k(m) is equal to (m-1) for proportions. As discussed previously, model (0) can be used for the variance of Y under the one parameter exponential family model. Model (1) was used by Finney (1971) and has been made popular by quasi-likelihood estimation methods. Model (2), used by Nelder and Pregibon (1987) and also by Efron (1986), allows the dispersion parameter to depend on covariates or factors and can be fit, for example, using either an extended quasi-likelihood model or a double exponential family distribution. Model (3), discussed by Pierce and Sands (1975), Williams (1982) and Breslow (1984), arises in modeling extra variation on the same scale as the covariates for the mean. The main focus of this chapter will be to investigate some consequences of using model (1) when the variation of Y is more appropriately described by either model (2) or model (3).

3.1 Consequences of Ignoring Overdispersion

In their introduction to the analysis of count data using log-linear models, McCullagh and Nelder (p. 127, 1983), note that often counts do not occur according to the Poisson model of randomness, but occur in clusters or batches. They suggest,

> "Unless there is strong evidence to the contrary we avoid the assumption of Poisson variation and assume only that
>
> $$\mathrm{Var}(Y_i) = \sigma^2 E(Y_i)$$
>
> where σ^2, the dispersion parameter is assumed constant over the data."

The estimated β_j's obtained from fitting model (0) are identical to the estimates obtained from model (1) using quasi-likelihood methods and accounting for a constant dispersion parameter (Wedderburn, 1974). Ignoring the presence of overdispersion, however, can result in underestimation of standard errors and incorrect inference about parameters. Consequently the presence of ignored extra variability may also prompt the researcher to include unnecessary interaction terms or extra explanatory variables in the regression.

Consider Example 5 from Chapter 1 where batches of aphids are exposed to insecticides at different dose levels. Suppose that the regression model for the mean is given by

$$\text{probit}(\mu_i) = \beta_0 + \beta_1 \log(\text{conc}) + \beta_2 \text{class2} + \beta_3 \text{class3} + \beta_4 I_2 + \beta_5 I_3$$

where log(conc) is the logarithm of the toxin concentration used, (class 2) is an indicator for the degulin group, (class 3) is an indicator for the rotenone+degulin group and I_2 and I_3 are the [log(conc) x class2] and [log(conc) x class3] interactions respectively. The maximum likelihood estimates of the coefficients and the associated standard errors under model (0) and the maximum quasi-likelihood estimates and their estimated standard errors under model (1) are given in Table 3.1 below. The estimated coefficients are identical under both models but the standard errors are larger under model (1). In addition, the deviance statistic no longer has an asymptotic chi-squared distribution so that the deviance goodness

of fit test, mentioned in 2.1, is not valid. Notice that while there is still evidence for the significance of the interaction terms in the second fit, it is less conclusive.

Table 3.1. Rotenone Data. Estimates and Standard Errors Under Model (0) and Model (1).

Coefficient	Model (0) Estimate	S.E.	Model (1) Estimate	S.E.
β_0	-2.8870	0.3510	-2.8870	0.4505
log(conc)	1.8300	0.2087	1.8300	0.2678
class2	0.2201	0.4944	0.2201	0.6344
class3	0.8565	0.4431	0.8565	0.5687
I_2	-0.6465	0.2434	-0.6465	0.3124
I_3	-0.7371	0.2391	-0.7371	0.3068

Cox (1983) showed that the maximum likelihood estimates obtained from model (0) retained high efficiency with respect to the maximum quasi-likelihood estimates obtained under model (1) in the presence of small amounts of overdispersion, so that for $\varphi(\alpha_0)^{-1} < 1.2$, for example, the loss of efficiency due to using model (0) is minimal.

3.2 CONSEQUENCES OF INCORRECTLY ASSUMING A SIMPLE HETEROGENEITY MODEL

3.2.1 Normal Theory Regression

To study the consequences of estimating regression coefficients in a model for E(Y) and assuming model (1) for overdispersion when model (2) is correct, we can begin with the special case of normal linear models. Some consequences of using model (1), when model (2) is appropriate, for normally distributed data have been studied. In

this case, model (1) corresponds to the usual assumption of constant variance for linear models and model (2) corresponds to a particular form of heterogeneity where the variance depends on a known covariate.

First for the case of two independent normal samples and the test of equal means, consider the model,

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \text{where } x_i = \begin{cases} 0 & i = 1,\ldots m \\ 1 & i = m+1,\ldots,m+n \end{cases}$$

$$E(\epsilon_i) = 0 \quad \text{and} \quad Var(\epsilon_i) = \alpha_0 + \alpha_1 x_i$$

and the test of the hypothesis that $\beta_1 = 0$ when it is incorrectly assumed α_1 is zero. As discussed in Section 3.1, the estimate of β_1 in this example remains unbiased but its standard error is underestimated. The resulting Student's T-test will not necessarily have a Type 1 error rate equal to the nominal level. For example, Wetherill (1981) showed that the probability of exceeding the nominal 5% limit is equal to 5% if n = m or if $\alpha_1 = 0$. However, for $n \neq m$ and $\alpha_1 \neq 0$, the probability differs from 0.05 and the difference can be substantial. For example, if $n/m = 2$ and $\alpha_0/(\alpha_0 + \alpha_1) = 0.2$ then the Type 1 error probability is 0.15 and if $\alpha_0/(\alpha_0 + \alpha_1) = 2.0$ the probability is equal to 0.029.

Next, consider the following regression model:

$$Y_i = \underline{x}_i'\underline{\beta} + \epsilon_i \qquad E(\epsilon_i) = 0 \qquad Var(\epsilon_i) = \sigma^2 w_i \qquad \text{for } i = 1,\ldots n$$

where $\underline{x}_i$ is a (px1) vector of known covariates and w_i is a known scalar. It is well known (Draper and Smith, 1981) that the

unweighted least squares estimator of $\underline{\beta}$ is inefficient and a more efficient estimator is obtain by weighted least squares.

If ordinary least squares is used to estimate $\underline{\beta}$ in the model above, then the covariance matrix for the estimated $\underline{\beta}$, $\hat{\underline{\beta}}_{ls}$, is ,

$$\text{Cov } (\hat{\underline{\beta}}_{ls}) = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{W}\underline{X} \ (\underline{X}'\underline{X})^{-1}\sigma^2$$

and the covariance matrix of the estimated $\underline{\beta}$ under a weighted least squares analysis, $\hat{\underline{\beta}}_{wls}$ is given by ,

$$\text{Cov } (\hat{\underline{\beta}}_{wls}) = (\underline{X}\ \underline{W}^{-1}\underline{X})^{-1}\sigma^2$$

where $\underline{X}$ is the (nxp) matrix with i^{th} row $\underline{x}_i'$ and $\underline{W}$ is the diagonal matrix with ith diagonal entry w_i. The standard errors of the $\hat{\underline{\beta}}_{ls}$ obtained from Cov $(\hat{\underline{\beta}}_{ls})$ are larger than the standard errors of the $\hat{\underline{\beta}}_{wls}$.

3.2.2 Generalized Linear Models

3.2.2.1 GENERAL RESULTS

In the context of generalized linear models, it is desired to assess the consequences of using model (1) when model (2) is correct. Some of the results of White (1982) on the asymptotic properties of maximum likelihood estimators from misspecified likelihoods will be used to assess the consequences of model misspecification for the special case of generalized linear models with the canonical link. For this application the asymptotic results do not require that the

full distribution be known, only the first two moments must be specified. For this reason, the following discussion can be based on the double exponential family distribution (Efron, 1986) without loss of generality, at least in an asymptotic sense. In the absence of overdispersion the double exponential family is a one parameter exponential family and maximum likelihood estimates obtained from it are equivalent to maximum likelihood estimates obtained from the one parameter exponential family. When model (1) is used, the maximum likelihood estimates obtained from the double exponential distribution are the same as the maximum quasilikelihood estimates. So double exponential families provide a useful framework into which models (0), (1) and (2) can be placed. It will be shown in Section 3.3 that model (3) can be approximately described by model (2) and placed in the double exponential family framework as well.

Given $Y_1,\ldots,Y_n$, independent observations from a double exponential distribution, as described in Section 2.2.4, with true density $\Sigma_i g[y_i;\mu_i,\varphi(\alpha_0 + \underline{z}_i'\underline{\alpha}_1)]$ and canonical parameter, θ_i, suppose that,

$$E(Y_i) = \mu_i, \quad \theta_i = h(\mu_i) = \underline{x}_i'\underline{\beta},$$

and

$$\mathrm{Var}(Y_i) = (1/m_i)V(\mu_i)\varphi_i^{-1} \qquad \text{where } \varphi_i = \varphi(\alpha_0 + \underline{z}_i'\underline{\alpha}_1),$$

for some positive function $\varphi(\cdot)$. If $\underline{\alpha}_1 = \underline{0}$ then φ_i is constant so

that the mean/variance structure corresponds to model (1); otherwise it corresponds to model (2). The question of interest is how the use of the incorrect assumption of a constant dispersion parameter (that is, falsely assuming $\underline{\alpha}_1 = \underline{0}$), affects asymptotic standard errors, confidence intervals and relative efficiency of the estimated β_j's.

Suppose model (1) is assumed by a researcher to be correct and the likelihood is taken to be $\Sigma_i g[\mu_i, \varphi(\alpha_0); y_i]$, when in reality model (2) is correct and the true likelihood is $\Sigma_i g[y_i; \mu_i, \varphi(\alpha_0 + \underline{z}_i'\underline{\alpha}_1)]$. In addition, suppose $(\tilde{\underline{\beta}}, \tilde{\alpha}_0)$ maximizes $\Sigma_i g[\mu_i, \varphi(\alpha_0); y_i]$, the misspecified likelihood, and let $(\underline{\beta}^0, \alpha_0^0, \alpha_1^0, \ldots, \alpha_q^0) = (\underline{\beta}^0, \alpha_0^0, \underline{\alpha}_1^0)$ be the true values of the parameters under model (2). Then, under suitable regularity conditions (White 1982), $\tilde{\underline{\beta}}$ converges in probability to $\underline{\beta}^0$ and $\varphi(\tilde{\alpha}_0)$ converges, approximately, in probability to

$$\varphi(\alpha_0^*) = n \left\{ \sum\nolimits_i \varphi(\alpha_0 + \underline{z}_i'\underline{\alpha}_1)^{-1} \right\}^{-1} .$$

In addition, $\sqrt{n}\,[(\tilde{\underline{\beta}}', \tilde{\alpha}_0) - (\underline{\beta}^{0\prime}, \alpha_0^*)]$ converges in law to a multivariate normal distribution with mean 0 and asymptotic covariance matrix, $(\underline{A}^{-1}\underline{B}\,\underline{A}^{-1})$, where,

$$\underline{A} = \lim_{n\to\infty} n^{-1} \sum\nolimits_{i=1}^{n} m_i V(\mu_i^0)\, \underline{x}_i \underline{x}_i'$$

$$\underline{B} = \lim_{n\to\infty} n^{-1} \sum\nolimits_{i=1}^{n} m_i V(\mu_i^0)\, [\varphi(\alpha_0^0 + \underline{z}_i'\underline{\alpha}^0)]^{-1}\, \underline{x}_i \underline{x}_i'$$

and $h(\mu_i^0) = \underline{x}_i'\underline{\beta}^0$.

Under the misspecified model (1), the researcher would think that the asymptotic covariance of $\tilde{\underline{\beta}}$ was,

$$\text{Cov}(\tilde{\underline{\beta}}) = \varphi(\alpha_o)^{-1} \left[\sum m_i V(\mu_i) \underline{x}_i \underline{x}_i' \right]^{-1} . \qquad (4)$$

The incorrectly estimated asymptotic covariance matrix, $\widetilde{\text{Cov}}(\tilde{\underline{\beta}})$, would be obtained from (4) with μ_i replaced with $\tilde{\mu}_i$ and α_o replace with $\tilde{\alpha}_o$.

However, the previous results show that in the presence of model misspecification the covariance matrix of the β_j's is not given by (4) but by $(n^{-1}\underline{A}^{-1}\underline{B}\,\underline{A}^{-1})$. Thus the standard errors of the β_j's are incorrectly estimated. The question of how the use of the wrong covariance matrix due to model misspecification affects inference and efficiency of the β_j's is addressed in the next two sections.

In the first section, coverage probabilities of asymptotic confidence intervals will be used to study the effect of model misspecification on inference. Using results derived above, approximate coverage probabilities for asymptotic nominal 95% confidence intervals for the β_j's can be obtained. These approximate coverage probabilities are explicitly evaluated for some simple cases. In the second section the asymptotic relative efficiency of the β_j's is evaluated for these simple cases.

3.2.2.2 COVERAGE PROBABILITIES

To explore the effect of the incorrect standard errors on inference, consider the true asymptotic coverage probabilities of 95% confidence intervals based on the incorrect standard errors. Under

the misspecified model (1) and a double exponential family distribution with a variance structure as in model (1), the maximum likelihood estimate of $\varphi(\alpha_o)$ is, $\tilde{\varphi}(\tilde{\alpha}_o) = n\,[\Sigma_i d(y_i, \tilde{\mu}_i)]^{-1}$ where the $d(y_i;\mu_i)$'s, the deviance components are defined in Section 2.1. Using this and the previous results on the true asymptotic distribution of $\tilde{\beta}$ and $\tilde{\alpha}_0$, it can be shown that the incorrectly estimated asymptotic covariance matrix of the β_j's, $\widetilde{Cov}(\tilde{\beta})$, converges in probability to approximately, $[\varphi(\alpha_0^*)^{-1}\ (n\ \underline{A})^{-1}]$.

When model (2) is correct, the true asymptotic coverage probabilities of nominal 95% confidence intervals for β_j, based on the incorrect standard errors from model (1), can be approximated by

$$1 - 2\ Pr\left[\ Z > 1.96\ \left\{Var(\tilde{\beta}_j)^*/\ Var(\tilde{\beta}_j)\right\}^{1/2}\ \right] \qquad (5)$$

where $Var(\tilde{\beta}_j)^*$ is the j^{th} diagonal entry of $[\varphi(\alpha_0^*)^{-1}\ (n\ \underline{A})^{-1}]$ and $Var(\tilde{\beta}_j)$ is the j^{th} diagonal entry of $(n^{-1}\ \underline{A}^{-1}\underline{B}\ \underline{A}^{-1})$.

These coverage probabilities will be evaluated for several examples where the $\underline{x}_i$'s are given. The $\underline{A}$ and $\underline{B}$ that appear in (5) will be replaced with their sample versions,

$$\underline{A}_n = n^{-1} \sum_{i=1}^{n} m_i V(\mu_i^0)\ \underline{x}_i \underline{x}_i' \ , \qquad (6)$$

$$\underline{B}_n = n^{-1} \sum_{i=1}^{n} m_i V(\mu_i^0)\ [\varphi(\underline{z}_i'\underline{a}^0)]^{-1}\ \underline{x}_i \underline{x}_i' \ . \qquad (7)$$

Independent Samples

As a simple example, consider two independent double-binomial samples of sizes n_1 and n_2 respectively where the j^{th} proportion in the i^{th} group is based on m_{ij} trials. Let the regression model for the mean be given by,

$$\theta_i = \text{logit}(\mu_i) = \underline{x}_i'\underline{\beta} \qquad \text{for } i = 1,2$$

where $\underline{x}_1' = (1,0)$ and $\underline{x}_2' = (1,1)$ and $\underline{\beta}' = (\beta_1, \beta_2)$, so that $\theta_1 = \beta_1$ and $\theta_2 = \beta_1 + \beta_2$. The objective here is to find approximate true coverage probabilities for the nominal 95% confidence interval for $\beta_2 = (\theta_2 - \theta_1)$ when it is incorrectly assumed that the dispersion parameter is $\varphi(\alpha_0) = \alpha_0^{-1}$ for both observations when, in fact, the true dispersion parameter is given by,

$$\varphi(\alpha_1 + \alpha_2 z_i) = (\alpha_1 + \alpha_2 x_i)^{-1}.$$

Using (5), the approximate true coverage probabilities of the nominal asymptotic 95% confidence interval for β_2 is given by,

$$1 - 2\,\Pr\left[Z > 1.96\left\{ [(1-R_2) + \delta_2 R_2] \left[\frac{1 + \lambda_2}{\delta_2 + \lambda_2}\right]\right\}^{1/2}\right]$$

where $V(\mu_i) = \mu_i(1-\mu_i)$, $R_2 = \left[\frac{n_2}{n_1 + n_2}\right]$, $\delta_2 = (\alpha_1 + \alpha_2)/\alpha_1$ and

$$\lambda_2 = \left[\frac{V(\mu_2)\ \Sigma_j m_{2j}}{V(\mu_1)\ \Sigma_j m_{1j}}\right].$$

Notice that if the assumption that $\varphi(\alpha_0) = \alpha_0^{-1}$ is true, i.e., $\alpha_2 = 0$, so that $\delta_2 = 1$, then the nominal asymptotic 95% confidence interval has the correct coverage probability. Also, if the number of observations, the total binomial sample sizes and the means are all equal in each group, the coverage probability is still correct, even though the amount of overdispersion in each group is different. Table 3.2 shows how the approximate true coverage probabilities change for various other choices of R_2, λ_2 and δ_2. Note that $m_i = \Sigma_j m_{ij}$.

Table 3.2. Approximate True Coverage Probabilities for Nominal 95% Asymptotic Confidence Intervals for $(\theta_2 - \theta_1)$.

$n_1:n_2$	$m_1V(\mu_1):m_2V(\mu_2)$	$\alpha_1:\alpha_1+\alpha_2$	coverage
1:5	5:1	5:1	1.00
1:5	5:1	2:1	.99
1:5	5:1	1:2	.88
1:5	5:1	1:5	.78
1:5	2:1	5:1	.99
1:5	2:1	2:1	.98
1:5	2:1	1:2	.90
1:5	2:1	1:5	.81
1:5	1:2	5:1	.97
1:5	1:2	2:1	.96
1:5	1:2	1:2	.93
1:5	1:2	1:5	.90
1:5	1:5	5:1	.95
1:5	1:5	2:1	.95
1:5	1:5	1:2	.95
1:5	1:5	1:5	.95
1:2	1:5	1:5	1.00
1:2	1:5	1:2	.99
1:2	1:5	2:1	.91
1:2	1:5	5:1	.88
1:2	1:2	1:5	.99
1:2	1:2	1:2	.97
1:2	1:2	2:1	.92
1:2	1:2	5:1	.88
1:2	2:1	1:5	.95
1:2	2:1	1:2	.95
1:2	2:1	2:1	.95
1:2	2:1	5:1	.95
1:2	5:1	1:5	.93
1:2	5:1	1:2	.94
1:2	5:1	2:1	.96
1:2	5:1	5:1	.98

As a second example consider 3 independent samples of proportions with the regression model for the mean given by,

$$\theta_i = \text{logit}(\mu_i) = \underline{x}_i'\underline{\beta} \quad \text{for } i = 1,2,3$$

where $\underline{x}_1' = (1,0,0)$, $\underline{x}_2' = (1,1,0)$ $\underline{x}_3' = (1,0,1)$ and $\underline{\beta}' = (\beta_1, \beta_2, \beta_3)$.

Again it is incorrectly assumed that the dispersion parameter is $\varphi(\alpha_0) = \alpha_0^{-1}$ for all observations when in reality, the true dispersion parameter is given by,

$$\varphi(\alpha_1 + \alpha_2 z_2 + \alpha_3 z_3) = (\alpha_1 + \alpha_2 z_2 + \alpha_3 z_3)^{-1}$$

where z_2 is an indicator variable for group 2 and z_3 is an indicator variable for group 3.

Then, using (5), the approximate coverage probabilities of the nominal 95% asymptotic confidence intervals for $\beta_2 = (\theta_2 - \theta_1)$ and for $\beta_3 = (\theta_3 - \theta_1)$ are given below.

$$(\theta_2 - \theta_1): \qquad 1 - 2\ \Pr\left[Z > 1.96\left\{[R_1 + R_2\delta_2 + R_3\delta_3]\left[\frac{1 + \lambda_2}{\delta_2 + \lambda_2}\right]\right\}^{1/2}\right]$$

$$(\theta_3 - \theta_1): \qquad 1 - 2\ \Pr\left[Z > 1.96\left\{[R_1 + R_2\delta_2 + R_3\delta_3]\left[\frac{1 + \lambda_3}{\delta_3 + \lambda_3}\right]\right\}^{1/2}\right]$$

where $V(\mu_i) = \mu_i(1-\mu_i)$, $R_i = n_i/(\Sigma_i n_i)$, $\lambda_k = \left[\frac{V(\mu_k)\ \Sigma_j m_{kj}}{V(\mu_1)\ \Sigma_j m_{1j}}\right]$,

$\delta_2 = \left[\frac{\alpha_1 + \alpha_2}{\alpha_1}\right]$ and $\delta_3 = \left[\frac{\alpha_1 + \alpha_3}{\alpha_1}\right]$.

If the amount of overdispersion is the same for all groups, then the coverage probability attains the nominal level, as was the case for the two sample problem. However, if the number of observations in each group, the total binomial sample sizes and the means of all the groups are equal, the coverage probability is not identically equal to 0.95 as it was for the two sample problem. The approximate true coverage probability can either be larger or smaller than the nominal probability. Table 3.3 shows how the true coverage probabilities change as R_i, λ_k and δ_j change.

Table 3.3. Approximate True Coverage Probabilities for Nominal 95% Asymptotic Confidence Intervals for the Difference in Canonical Parameters.

$n_1:n_2:n_3$	$m_1V(\mu_1):m_2V(\mu_2):m_3V(\mu_3)$	$a_1:a_1+a_2:a_1+a_3$	C2*	C3**
1:1:1	1:1:1	1:2:3	.98	.95
1:1:1	1:1:1	1:1:3	.99	.93
1:1:1	1:2:2	1:2:3	.98	.97
1:1:1	1:2:2	1:1:3	.99	.95
1:1:1	1:2:3	1:2:3	.98	.98
1:1:1	1:2:3	1:1:3	.99	.96
1:2:2	1:1:1	1:2:3	.98	.96
1:2:2	1:1:1	1:1:3	.99	.94
1:2:2	1:2:2	1:2:3	.99	.98
1:2:2	1:2:2	1:1:3	.99	.96
1:2:2	1:2:3	1:2:3	.99	.98
1:2:2	1:2:3	1:1:3	.99	.97
2:1:1	1:1:1	1:2:3	.97	.93
2:1:1	1:1:1	1:1:3	.98	.91
2:1:1	1:2:2	1:2:3	.98	.95
2:1:1	1:2:2	1:1:3	.98	.94
2:1:1	1:2:3	1:2:3	.98	.97
2:1:1	1:2:3	1:1:3	.98	.95

* C2 : coverage probability for $\theta_2 - \theta_1$.
** C3 : coverage probability for $\theta_3 - \theta_1$.

A Single Continuous Covariate

For a third example suppose that $\theta_i = h(\mu_i) = \underline{x}_i'\underline{\beta}$ where $\underline{x}_i' = [1, x_i]$, x_i is a continuous covariate, $\underline{\beta} = (\beta_0, \beta_1)$ and $i = 1,\ldots,n$. Suppose that it is incorrectly assumed that $\varphi(a) = a^{-1}$ for all observations, when in reality, the true dispersion parameter is given by $\varphi(\underline{z}_i'\underline{a})$ where $\underline{z}_i' = [1, x_i]$ and $\underline{a}' = [a_0, a_1]$. Again, using (5), the approximate coverage probability of the nominal 95% asymptotic confidence interval for β_1 is,

$$1 - 2\ \Pr\left[Z > 1.96\left\{\frac{\mathrm{Var}(\tilde{\beta}_1)^*}{\mathrm{Var}(\tilde{\beta}_1)}\right\}^{1/2}\right]$$

where $\mathrm{Var}(\tilde{\beta}_1)^*$ is approximated by the (2,2) entry of $\varphi(\alpha_0^*)^{-1}(n\,\underline{A}_n)^{-1}$ and $\mathrm{Var}(\tilde{\beta}_1)$ is approximated by the (2,2) entry of $(n^{-1}\,\underline{A}_n^{-1}\underline{B}_n\underline{A}_n^{-1})$.

Then, if $\bar{X} = (1/n)\Sigma_i x_i$,

$$\left[\frac{\mathrm{Var}(\tilde{\beta}_1)^*}{\mathrm{Var}(\tilde{\beta}_1)}\right] \approx \left\{\alpha_0 + \alpha_1\bar{X}\right\}\left\{\left[\alpha_0 + \alpha_1\left(\frac{a_1}{a_0}\right)\right] + \left[\frac{a_0a_3 - a_1a_2}{a_0a_2 - a_1^2}\right]\right\}^{-1}$$

where,
$$a_0 = \Sigma_i m_i V(\mu_i)$$
$$a_1 = \Sigma_i m_i V(\mu_i) x_i$$
$$a_2 = \Sigma_i m_i V(\mu_i) x_i^2$$
$$a_3 = \Sigma_i m_i V(\mu_i) x_i^3 .$$

The following example has been constructed in order to evaluate these expressions for a specific $\underline{X}$ matrix. This example is roughly similar to the Salmonella example of Chapter 1.

Suppose Y_i is an observed count with,

$$\log(\mu_i) = \beta_0 + \beta_1[\log(\mathrm{dose}_i)]^2$$

where $\beta_0 = 4.0$, $\beta_1 = -0.01$, $\mathrm{dose}_i = 100$ for $i = 1,\ldots,9$; $\mathrm{dose}_i = 333$ for $i = 10,\ldots,18$; $\mathrm{dose}_i = 1000$ for $i = 19,\ldots,27$;

$dose_i = 3333$ for $i = 28,\ldots,36$ and $dose_i = 10000$ for $i = 37,\ldots,44$.

Suppose that the correct model for the dispersion term is,

$$\varphi(15 - 0.15\ z_i) = (15 - 0.15\ z_i)^{-1}$$

where $z_i = [\log(dose_i)]^2$. It is incorrectly assumed that $\varphi(\alpha_0) = \alpha_0^{-1}$ for all observations. Using the formula above for $\left[\frac{Var(\tilde{\beta}_1)^*}{Var(\tilde{\beta}_1)}\right]$ with $m_i = 1$ for counts, the approximate coverage probability for the nominal 95% confidence interval for β_1 is 0.90.

3.2.2.3 ASYMPTOTIC RELATIVE EFFICIENCY

Under the misspecified model (1), where $E(Y_i) = \mu_i$, $h(\mu_i) = \underline{x}_i'\underline{\beta}$ and $Var(Y_i) = \varphi(\alpha_0)^{-1}V(\mu_i)$, the approximate efficiency attained by the estimated $\underline{\beta}$ vector, $\tilde{\underline{\beta}}$, is also of interest. Under the misspecified model (1), the correct covariance matrix of $\tilde{\underline{\beta}}$ is $(n^{-1}\ \underline{A}^{-1}\underline{B}\ \underline{A}^{-1})$ which can be estimated by $(n^{-1}\ \underline{A}_n^{-1}\ \underline{B}_n\ \underline{A}_n^{-1})$. Now if $(\hat{\underline{\beta}},\ \hat{\alpha}_0,\ \hat{\underline{\alpha}}_1)$ maximizes the true likelihood under model (2), the correct covariance matrix of $\hat{\underline{\beta}}$ is,

$$\underline{C}_n^{-1} = \left[\sum m_i V(\mu_i^0)\ [\varphi(\alpha_0 + \underline{z}_i'\underline{\alpha}_1)]\ \underline{x}_i\underline{x}_i'\right]^{-1}.$$

Then the asymptotic relative efficiency of $\tilde{\underline{\beta}}$ with respect to $\hat{\underline{\beta}}$ is approximated by,

$$A.R.E. = \left[\frac{|\underline{C}_n^{-1}|}{|n^{-1} \underline{A}_n^{-1} \underline{B}_n \underline{A}_n^{-1}|} \right]^{1/p}$$

where p is the number of elements of x_i and $\underline{A}_n$ and $\underline{B}_n$ are given by (6) and (7) respectively.

For the two and three independent sample examples given previously, the asymptotic relative efficiency is identically 1.

For the case of a single covariate, x_i, in the regression model for the mean and a single continuous covariate, z_i, in the regression model for the dispersion parameter, the asymptotic relative efficiency of $\tilde{\underline{\beta}}$ with respect to $\hat{\underline{\beta}}$ is given by,

$$A.R.E.= \left[\frac{|n\underline{A}_n|}{[\ |\underline{C}_n|\ |n\underline{B}|\ ]^{1/2}} \right]$$

where, if $h(\mu_i) = \beta_0 + \beta_1 x_i$, $\varphi(z_i'\alpha) = \varphi(\alpha_0 + \alpha_1 z_i)$ and $v_i = m_i V(\mu_i)$ then,

$$|n\underline{A}_n| = [(\Sigma_i v_i)(\Sigma_i v_i x_i^2) - (\Sigma_i v_i x_i)^2] ,$$

$$|n\underline{B}_n| = \{[\Sigma_i v_i \varphi(z_i'\alpha_i)^{-1}]\ [\Sigma_i v_i \varphi(z_i'\alpha)^{-1} x_i^2] - [\Sigma_i v_i \varphi(z_i'\alpha)^{-1} x_i]^2\}$$

and

$$|\underline{C}_n| = \{[\Sigma_i v_i \varphi(\underline{z}_i'\underline{\alpha})]\ [\Sigma_i v_i \varphi(\underline{z}_i'\underline{\alpha}) x_i^2] - [\Sigma_i v_i \varphi(\underline{z}_i'\underline{\alpha}) x_i]^2\} .$$

For the constructed example in the previous section, where $\varphi(\underline{z}_i'\underline{\alpha}) = (\underline{z}_i'\underline{\alpha})^{-1}$, $z_i = x_i$ and $v_i = \mu_i$, the asymptotic relative efficiency is approximately, 86%.

3.3 MISSPECIFICATION OF MODEL (3)

Assuming model (1) when model (3) is appropriate is another type of misspecification which might occur. Under model (3) (See Section 2.2.3.3),

$$\mathrm{Var}(Y) \approx (1/m)V(\mu)[1 + \sigma^2 k(m)V(\mu)].$$

Model (3) can be approximated by model (2) by making the following substitutions. Let $\tilde{\underline{\beta}}$ be the estimate of $\underline{\beta}$ from model (2) which can be obtained using maximum quasi-likelihood estimation methods. Let $\tilde{\mu}$ be such that $h(\tilde{\mu}) = \underline{x}_i'\tilde{\underline{\beta}}$, and let φ be defined by $\varphi(a) = a^{-1}$. Let $\tilde{z}_i = k(m)V(\tilde{\mu})$. Then,

$$\begin{aligned}\mathrm{Var}(Y) &\approx V(\mu)\ (1/m)\ [1 + \sigma^2 k(m)V(\mu)].\\ &\approx V(\mu_0)\ (1/m)\ \varphi(\alpha_0+\alpha_1\tilde{z})^{-1}\end{aligned}$$

where α_0 and α_1 replace 1 and σ^2 respectively. For proportions $\tilde{z}_i = (m-1)\tilde{\mu}_i(1-\tilde{\mu}_i)$ and for counts $\tilde{z}_i = \tilde{\mu}_i$ so that model (3) can be approximated by model (2) using the constructed covariate $\tilde{z}_i$. Thus

the results of Section 2.1 can be applied to this type of misspecification as well.

3.4 SUMMARY

For the cases of two and three independent samples discussed in Section 3.2.2, the incorrect use of model (1) when in fact model (2) is correct does not lead to a loss of efficiency. However, it does result in incorrect standard errors as evidenced by the coverage probabilities that are not equal to the nominal value. How true coverages differ from the nominal level depends on the sample sizes, on the ratios of the binomial or Poisson components of variance as well as on the relative degree of overdispersion in the samples. However, when these ratios are 2:1 or less, the coverage probabilities differ from the nominal level by 3% or less.

When both overdispersion and the mean depend on continuous covariates, coverages can differ from the nominal probability and asymptotic relative efficiencies can differ from one. However, X and Z matrices as well as parameter values are necessary to evaluate these differences. Since in practice parameter values are unknown, the extent to which mistakes may be made can be difficult to evaluate.

In the next chapter a diagnostic tool for deciding whether or not model (2) is more appropriate than model (1) will be developed. The plot is an extension of a diagnostic technique presented by Cook and Weisberg (1983) for non-constant variance in ordinary regression

and is easy to use with existing software packages for generalized linear models.

Chapter 4

A DIAGNOSTIC TOOL FOR THE DEPENDENCE OF OVERDISPERSION ON COVARIATES AND FACTORS

For overdispersed counts and proportions, the model used by Finney (1971), (see Section 2.2.3) in which the variance of the response is assumed to be a constant multiple of the binomial or Poisson variance, has proved to be quite useful. In the context of overdispersed counts, McCullagh and Nelder (p. 132, 1983) note that,

> "If the precise mechanism that produces the overdispersion or underdispersion is known (e.g. as with electronic counters), specific methods may be used. In the absence of such knowledge it is convenient to assume as an approximation that $\mathrm{Var}(Y) = \sigma^2\mu$ for some constant σ^2. This assumption can and should be checked, but even relatively substantial errors in the assumed functional form of Var(Y) generally have only a small effect on the conclusion."

It is therefore desirable to have an easy diagnostic method for deciding when it is necessary to use a more sophisticated model for the overdispersion than the one with a single heterogeneity factor.

For ordinary regression analysis the adequacy of the model and the assumptions on which the model are based can be checked using diagnostic statistics and plots. There exists a large body of literature concerning regression diagnostics and Cook and Weisberg (1982) provide a good review. To assess the assumption of constant variance in ordinary regression, a plot of residuals versus fitted values is often used. If the plot shows a megaphone shape indicating

that the residual variability increases for increasing fitted values, this is taken as evidence of variance which depends on the mean.

In generalized linear models the dependence of a dispersion parameter on known covariates can be checked by comparing models fit using extended quasi-likelihood (see Section 2.2.4). For normal, inverse Gaussian or gamma random variables, in which the dispersion parameter depends on the covariates, z_{ij}, Smyth (1989) noted that the score test statistic for testing that the coefficients of the z_{ij}'s are zero can be interpreted as one half of a regression sum of squares. This was also noted by Cook and Weisberg (1983). In addition, the method presented by Moore (1987) (see Section 2.2.3) can be used to determine whether the dispersion parameter depends on a power of the variance function. In both cases, by comparing the fit of models with and without the parameters of interest, a data analyst can decide if the dispersion parameter varies across groups or observations.

For overdispersed proportions, Williams (1982) suggests that plots of standardized residuals versus $\tilde{\mu}_i$, in which the variance of the residuals decreases markedly as $\tilde{\mu}_i$ approaches 0 or 1 may be indicative of the appropriateness of model (3).

A score test for detecting the dependence of heteroscedasticity on covariates for ordinary regression was presented by Cook and Weisberg (1983). An extension of their test and an associated diagnostic plot of overdispersion in general linear models is derived in this chapter. The results of Cook and Weisberg are given in Section 4.1. The extension to overdispersion is presented in Section

4.2. An examination of this score test and the associated graph for some simple cases are given in Section 4.3. Finally, in Section 4.4, the diagnostic is applied to the examples of Chapter 1.

4.1 A SCORE TEST FOR NON-CONSTANT VARIANCE IN ORDINARY REGRESSION

Cook and Weisberg (1983) suggested a diagnostic plot for ordinary regression based on a score test for non-constant variance. They noted that the variance may depend on known explanatory variables such as time or spatial order and developed the following model to incorporate these variables.

The Gaussian regression model can be written as

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{\epsilon} \quad \text{where} \quad \underline{\epsilon} \sim \text{MVN}(\underline{0}, \sigma^2\underline{I}) \ .$$

An alternative model that can incorporate covariates in the variance is given by,

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{\epsilon} \quad \text{where} \quad \underline{\epsilon} \sim \text{MVN}(\underline{0}, \sigma^2\underline{W})$$

where $\underline{W}$ is a diagonal matrix with i^{th} diagonal entry $w(\underline{z}_i, \underline{\alpha})$, $\underline{\alpha}$ is a (qx1) vector of unknown parameters and $\underline{z}_i$ is a (qx1) vector of known covariates for the variance and may, but need not, coincide with the variables in $\underline{X}$. The function $w(\underline{z}_i, \underline{\alpha})$ is assumed to be twice differentiable with respect to $\underline{\alpha}$ and it is assumed that there exists an $\underline{\alpha}^*$ such that $w(\underline{z}_i, \underline{\alpha}^*) = 1$ for all $\underline{z}_i$.

Cook and Weisberg (1983) suggested the general family

$$w(\underline{z}_i, \underline{\alpha}) = \exp\left[\sum_{j=1}^{q} \alpha_j (z_{ij})^{a_j} \right]$$

where $z^a = \log(z)$ for $a = 0$. This family contains two useful specific families,

$$\text{if } a_j = 1 \qquad \text{then} \quad w(\underline{z}_i, \underline{\alpha}) = \exp\left\{\underline{z}_i' \underline{\alpha}\right\}$$

$$\text{if } a_j = 0 \qquad \text{then} \quad w(\underline{z}_i, \underline{\alpha}) = \exp\left\{\Sigma_j \alpha_j \log[z_i]\right\}.$$

If it is desired to model the variance as a function of the expected response then one may use $w(\underline{z}_i, \underline{\alpha}) = w(\alpha_0 \underline{x}_i' \underline{\beta})$ where α_0 is a scalar.

Under these specific models for $w(\underline{z}_i, \underline{\alpha})$, the score test of constant variance corresponds to testing $\underline{\alpha} = \underline{\alpha}^* = \underline{0}$ and it will be shown that this test has a simple form. Define $\dot{w}(\underline{z}_i, \underline{\alpha}^*)$ to be the (qx1) vector with j^{th} entry

$$\dot{w}_j(\underline{z}_i, \underline{\alpha}^*) = \left[\frac{\partial w(\underline{z}_i, \underline{\alpha})}{\partial \alpha_j} \right]_{\underline{\alpha} = \underline{\alpha}^*}$$

and let $\underline{D}$ be the (nxq) matrix with i^{th} row $[\dot{w}(\underline{z}_i, \underline{\alpha}^*)]'$. Then the score test statistic for the hypothesis of constant variance, i.e., $\underline{\alpha} = \underline{\alpha}^*$ is given by

$$S = 1/2\ \underline{U}'\underline{D}_c(\underline{D}_c'\underline{D}_c)^{-1}\underline{D}_c'\underline{U}\ ,$$

where if $e_i = (y_i - \underline{x}_i\ \hat{\underline{\beta}})$ and $\hat{\sigma}^2 = (1/n)\ \Sigma_i e_i^2$, then $\underline{U}$ is an (nx1) vector with elements $e_i^2/\hat{\sigma}^2$ and $\underline{D}_c = [\underline{D} - \underline{1}\underline{1}'\underline{D}(1/n)]$, i.e., $\underline{D}_c$ has mean corrected columns. Under the null hypothesis, S has an asymptotic central chi-square distribution with q degrees of freedom.

When $w(\underline{z}_i, \underline{\alpha}) = \exp[\underline{z}_i'\underline{\alpha}]$ then $\dot{w}(\underline{z}_i, \underline{\alpha}^*) = \underline{z}_i$ and the ij^{th} element of $\underline{D}$ is z_{ij} under the null hypothesis. If $w(\underline{z}_i, \underline{\alpha}) = \exp[\Sigma_j \alpha_j \log(z_{ij})]$, then the ij^{th} element of $\underline{D}$ under the null hypothesis is $\log[z_{ij}]$ So that for these choices of $w(\underline{z}_i, \underline{\alpha})$, the matrix $\underline{D}$ is not difficult to compute.

S can be computed as one half of the sum of squares for the regression of $\underline{U}$ on $\underline{D}$ in the constructed model $\underline{U} = \gamma\underline{1} + \underline{D}\gamma + \underline{\epsilon}$. To obtain the statistic S, fit the regression model of interest and obtain the ordinary regression residuals, e_i. Calculate the vector $\underline{U}$ using the residuals, and the matrix $\underline{D}$. Regress the vector $\underline{U}$ on $\underline{D}$ in the model with an intercept and obtain the regression sum of squares from this constructed model. Thus the score test statistic is easy to obtain from standard regression software packages.

Since testing procedures can be sensitive to the appropriateness of the normal regression model and to the presence of outliers, Cook and Weisberg (1983) suggest that a graphical procedure based on the score test offers a complementary method for distinguishing between models. The null hypothesis would tend to be rejected when the score statistic is large and this occurs when the regression sum of squares from the regression of $(e_i^2/\hat{\sigma}^2)$ on $\dot{w}_j(\underline{z}_i, \underline{\alpha}^*)$ is large. Thus

plots of $(e_i^2/\hat{\sigma}^2)$ versus $\dot{w}_j(\underline{z}_i, \underline{\alpha}^*)$ where the mean of the $(e_i^2/\hat{\sigma}^2)$ changes with $\dot{w}_j(z_i, \alpha^*)$ are evidence of non-constant variance. When only one covariate is being considered and $w(\underline{z}_i, \underline{\alpha}) = \exp[\underline{z}_i'\underline{\alpha}]$ then $\dot{w}(\underline{z}_i, \underline{\alpha}^*) = \underline{z}_i$. The plot suggested by the score test is just a plot of $e_i^2/\hat{\sigma}^2$ versus the elements of $\underline{z}_i$. If $w(\underline{z}_i, \underline{\alpha}) = \exp[\Sigma_j \alpha_j \log(z_{ij})]$, then $\dot{w}(\underline{z}_i, \underline{\alpha}^*)$ has j^{th} element $\log[z_{ij}]$ and for one covariate the plot is just a plot of $e_i^2/\hat{\sigma}^2$ versus $\log[z_i]$.

Cook and Weisberg also note that plotting $(e_i^2/\hat{\sigma}^2)$ instead of the usual residual ,e_i, places residuals with the same absolute value together. This increases the density of points in the plot which is helpful for small to moderate sample sizes. Many of these ideas extend quite naturally to generalized linear models and double exponential families if the "residuals" used are deviance residuals. This extension is discussed in detail in the next section.

4.2 A DIAGNOSTIC FOR OVERDISPERSION IN GENERALIZED LINEAR MODELS

4.2.1 Model 1 versus Model 2

For the overdispersion problem it is desirable to have a similar method for deciding whether overdispersion depends on covariates or whether it can be modeled in a simple fashion. The diagnostic should be easy to obtain and use, and it should not require fitting additional models. The diagnostic should complement the model fitting process and should not require a large investment of time.

A diagnostic procedure to explore the dependence of the dispersion parameter on covariates, analogous to the one presented by Cook and Weisberg (1983), can be derived using the double exponential family distribution. The double exponential family setting is convenient since the models discussed in this thesis for overdispersion can be fit into this framework. A double exponential family is derived from a one parameter exponential family, so overdispersed counts and proportions as well as heteroscedastic normal regression problems can be modeled. The models that have been used with quasi-likelihood and extended quasi-likelihood methods can be fit into the double exponential family framework as well. As noted in Section 2.2.4 estimates from a model obtained using extended quasi-likelihood are identical to the maximum likelihood estimates from the analogous double exponential family. A double exponential family distribution serves as a general setting for the problem of covariate dependent variation.

Suppose that $E(Y_i) = \mu_i$ and $h(\mu_i) = \underline{x}_i'\underline{\beta}$ define a regression model for the mean of Y_i. Consider the following models, presented previously in Chapter 3, for the variance of Y.

Model 1. $\quad Var(Y_i) = \varphi(\alpha_0)^{-1}V(\mu_i)$

Model 2. $\quad Var(Y_i) = \varphi(\alpha_0 + \underline{z}_i'\underline{\alpha}_1)^{-1}V(\mu_i)$

where $V(\mu_i)$ is the variance function, which refers here to the usual variance for binomial proportions or Poisson counts and $\underline{z}_i$ is a (qx1)

vector of known covariates for the dispersion parameter, φ. The parameter α_0 is an unknown scalar parameter and $\underline{\alpha}_1$ is a (qx1) vector of unknown parameters.

The quasi-likelihood method of estimation is appropriate for models in the form of model (1) where the dispersion parameter φ is constant for all observations. For model (2) the dispersion parameter can vary according to the covariate vector $\underline{z}_i$. This vector could contain indicator variables for groups, continuous covariates or combinations of these. Notice that if the vector $\underline{\alpha}_1$ is the zero vector then model (2) is the same as model (1).

A score test, similar to the one in Section 4.1 satisfies many of the requirements of a diagnostic for overdispersion. A score test of $\underline{\alpha}_1 = 0$ would correspond to testing the appropriateness of model (1) over the more general model (2). A score test would required fitting model (1) but not model (2), i.e. only the simple model needs to be fit. Finally, the form of this score test in the double exponential family setting will be shown to have a form that is easy to calculate with existing software.

To derive the score test of $\underline{\alpha}_1$ equal to zero in a double exponential family, suppose $Y_1,\ldots Y_n$, have a double exponential family distribution with density given by,

$$g(y;\mu,\varphi,m) = c(\mu,\varphi,m)\ \varphi^{1/2} \left\{f(y;\mu)\right\}^{\varphi} \left\{f(y;y)\right\}^{1-\varphi}$$

where, as described in Chapter 2, $f(y;\mu)$ is a density from a one parameter exponential family such as binomial or Poisson and

$c(\mu,\varphi,m) \approx 1$. Let regression models for the mean and dispersion parameters be given by $h(\mu_i) = \underline{x}_i'\underline{\beta}$ and $\varphi(\alpha_0 + \underline{z}_i'\underline{\alpha}_1)$, where $h(\cdot)$ is a positive link function. Assume that $\varphi(\alpha_0 + \underline{z}_i'\underline{\alpha}_1)$ is twice differentiable with respect to α. Then, under the null hypothesis, $\varphi(\alpha_0 + \underline{z}_i'\underline{\alpha}_1) = \varphi(\alpha_0)$. Let Z be the (nxq) matrix with i^{th} row $\underline{z}_i'$ and let $\underline{Z}_c = \underline{Z} - \underline{1}\underline{1}'\underline{Z}(1/n)$. Let $d(y_i,\mu_i)$ be the i^{th} deviance component. Using the approximation $E[d(y_i;\mu_i)] \approx \varphi(z_i'\alpha)^{-1}$ (Efron, 1986), the approximate score test statistic for testing the null hypothesis $\varphi(\alpha_0 + \underline{z}_i'\underline{\alpha}) = \varphi(\alpha_0)$ is given by

$$S = 1/2\ [\varphi(\tilde{\alpha}_0)]^2\ \tilde{\underline{d}}'\underline{Z}_c(\underline{Z}_c'\underline{Z}_c)^{-1}\underline{Z}_c'\tilde{\underline{d}}$$

where $\tilde{\mu}_i$ and $\varphi(\tilde{\alpha}_0)$ are the maximum likelihood estimates obtained from model (1), $\tilde{\underline{d}}$ is the (nx1) vector with i^{th} entry $d(y_i;\tilde{\mu}_i)$ and $\varphi(\tilde{\alpha}_0) = n[\Sigma_i d(y_i;\tilde{\mu}_i)]^{-1}$. A complete derivation of this statistic is given in Section 4.5.

As noted previously, when $\varphi(\alpha_0 + \underline{z}_i'\underline{\alpha}_1) = \varphi(\alpha_0)$, $\tilde{\mu}$ is the maximum quasi-likelihood estimate from model (1). The deviance components are also available from a maximum quasi-likelihood fit to this model. S is easily obtained as one half of the sum of squares of the ordinary regression of $d(y_i;\tilde{\mu}_i)\varphi(\tilde{\alpha}_0)$ on the mean corrected z's and S has an asymptotic chi-squared distribution with q degrees of freedom.

As noted in Section 4.5, the derivation of this score statistic uses the approximation, $E[d(y_i;\mu_i] \approx \varphi_i^{-1}$ (Efron, 1986), which is valid for large binomial sample sizes or for large Poisson means. However, the results of Pierce and Schafer (1986) suggest that this

approximation may be good even for small binomial sample sizes or Poisson means.

For the special case of normally distributed data, this score test reduces to the score test given by Cook and Weisberg (1983). The deviance components for the normal distribution are the squared residuals and the $\underline{Z}_c$ matrix is the $\underline{D}_c$ matrix of the previous section.

Although the score test could provide evidence to reject the null hypothesis it does not provide information on the relationships between overdispersion and the covariates. A scatter plot relating the overdispersion to covariates may be able to provide more information of this type and it can be used to see if outliers are responsible for the result of the score test. In addition, the validity of the score test statistic and its asymptotic distribution depend on large sample sizes and adequate approximations while a scatter plot relies less heavily on distributional assumptions.

The null hypothesis is rejected when the regression sum of squares is large and this suggests that plots of $[\varphi(\tilde{\alpha}_0)d(y_i;\tilde{\mu}_i)]$ versus z's could help in deciding whether or not overdispersion depends on covariates. In the case of a single covariate, a strong relationship between $[\varphi(\tilde{\alpha}_0)d(y_i;\tilde{\mu}_i)]$ and z_{ij} would mean that there is evidence for the inclusion of the covariate in the model. Since the statistics necessary for these plots are all easily computed when model 1 is fit using maximum quasilikelihood methods and existing software, the plot is a simple way of examining model assumptions.

4.2.2 Model 1 versus Model 3

If $E(Y_i) = \mu_i$, $h(\mu_i) = \underline{x}_i'\underline{\beta}$, defines a regression model for the mean μ_i, model 3 for the variance of Y_i is given by,

$$\text{Model 3.} \qquad \text{Var}(Y_i) = V(\mu_i)\ [1 + \sigma^2 k(m) V(\mu_i)]$$

where k(m) is equal to 1 for counts and to (m-1) for proportions. This model was introduced in Section 2.2.3 and it can arise when overdispersion is modeled on the same scale as the covariates for the mean. It was shown in Section 3.3 that model (3) can be written as

$$\text{Var}(Y) = \varphi(\alpha_0 + \alpha_1 \tilde{z}_i)^{-1} V(\mu_i)$$

where $\tilde{z}_i = k(m_i)V(\tilde{\mu}_i)$. So model (3) can be approximated by model (2) using the constructed covariate $\tilde{z}_i$.

The suggested diagnostic could also be applied, using the constructed covariate $\tilde{z}_i$ defined in Section 3.2, to help determine if model (3) is more appropriate than model (1).

4.3 EXAMINATION OF THE TEST FOR SPECIAL CASES

4.3.1 Two and Three Independent Samples

Suppose that Y_{ij}, $i = 1,2$ and $j = 1,\ldots,n_i$, have a double exponential family distribution with $E(Y_{ij}) = \mu_i$ and let $d_{i\cdot} = \sum_{j=1}^{n_i} d(y_{ij};\tilde{\mu}_i)$. Interest lies in testing the null

hypothesis,

$$\mathrm{Var}(Y_{ij}) = \varphi(\alpha_0)^{-1}V(\mu_i)$$

versus the alternative,

$$\mathrm{Var}(Y_{1j}) = \varphi(\alpha_0)^{-1}V(\mu_1) \quad \text{and} \quad \mathrm{Var}(Y_{2j}) = \varphi(\alpha_0 + \alpha_2)V(\mu_2).$$

The score test statistic for this hypothesis is given by

$$S = \left[\frac{n_1 + n_2}{2n_1n_2}\right]\left[\frac{n_1d_{1\cdot} - n_2d_{2\cdot}}{d_{1\cdot} + d_{2\cdot}}\right]^2.$$

When $n_1 = n_2 = n$, S can be written as,

$$S = n\left[\frac{d_{2\cdot} - d_{1\cdot}}{d_{1\cdot} + d_{2\cdot}}\right]^2.$$

This statistic is large when the difference between the weighted deviance within samples is large relative to the total deviance.

Similarly, for three independent sample of sizes n_1, n_2 and n_3, the score test statistic is given by,

$$S = \frac{1}{2}\varphi(\tilde{\alpha}_0)\left[\frac{n_1n_2(\lambda_2 - \lambda_1)^2 + n_1n_3(\lambda_3 - \lambda_1)^2 + n_2n_3(\lambda_2 - \lambda_3)^2}{d_{1\cdot} + d_{2\cdot} + d_{3\cdot}}\right]$$

where $\lambda_i = \left[\frac{d_{i\cdot}}{n_i}\right]$ and $\varphi(\tilde{\alpha}_0) = \left[\frac{n_1 + n_2 + n_3}{(d_{1\cdot} + d_{2\cdot} + d_{3\cdot})}\right]$.

If $n_1 = n_2 = n_3 = n$, then the score statistic is given by,

$$S = \left(\frac{3n}{2}\right)\left[\frac{(d_{2.} - d_{1.})^2 + (d_{3.} - d_{1.})^2 + (d_{2.} - d_{3.})^2}{(d_{1.} + d_{2.} + d_{3.})^2}\right]$$

In this case the statistic S is large when the sum of weighted pairwise differences in the total between groups is large.

For these independent sample problems, the diagnostic plot is a visual comparison of the scaled deviance components for each group. Groups for which the scaled deviance components seem to be quite a bit larger than other groups may have more variability than the other groups.

4.3.2 One Continuous Covariate

If it is thought that the dispersion parameter depends on a single, continuous covariate, z_i, for $i = 1,\ldots,n$, then the score test statistic is just the scaled regression sum of squares for the regression of the $d(y_i;\tilde{\mu}_i)$ on the z_i, i.e., S is given by,

$$S = \frac{1}{2}\left[\frac{n}{\Sigma_i d(y_i;\tilde{\mu}_i)}\right]^2\left[\frac{\Sigma_i (d_i - \bar{d}_i)(z_i - \bar{z})}{\Sigma_i (z_i - \bar{z})^2}\right]$$

where $d_i = d(y_i;\tilde{\mu}_i)$, $\bar{z} = (1/n)[\Sigma_i z_i]$ and $\bar{d} = (1/n)[\Sigma_i d_i]$.

The diagnostic plot is a scatter plot of the scaled deviance residuals versus the z_i's. The dependence of the variability on z_i is

indicated when the plot shows a dependence of the scaled deviance components on z_i.

4.3.3 Model 3

The diagnostic score test and plot can be carried out to compare the appropriateness of model (3) and model (1). This was shown in Section 4.2.2 where the z_i's, the covariate in the regression model for the variance, are $\tilde{z}_i = k(m_i)V(\tilde{\mu}_i)$. Here $\tilde{\mu}_i$ is the estimate of $E(Y_i)$ under model (1).

The procedure could be carried out as follows. First, fit model (1) and obtain the estimates of $E(Y_i)$. Next, calculate the $\tilde{z}_i$'s and $d(y_i;\tilde{\mu}_i)$'s. From these produce the scatter plot and compute the score test statistic.

The score test statistic has the same form as in the previous section,

$$S = \frac{1}{2}\left[\frac{n}{\Sigma_i d(y_i;\tilde{\mu}_i)}\right]^2\left[\frac{\Sigma_i (d_i-\bar{d}_i)(\tilde{z}_i-\bar{z})}{\Sigma_i (\tilde{z}_i-\bar{z})^2}\right],$$

where $d_i = d(y_i;\tilde{\mu}_i)$, $\bar{z} = (1/n)[\Sigma_i \tilde{z}_i]$ and $\bar{d} = (1/n)[\Sigma_i d_i]$. The plot is a scatter plot of the scaled deviance components versus the constructed covariates, $\tilde{z}_i$.

4.4 APPLICATION OF DIAGNOSTIC TOOLS TO THE EXAMPLES

4.4.1 Fish Toxicology Data

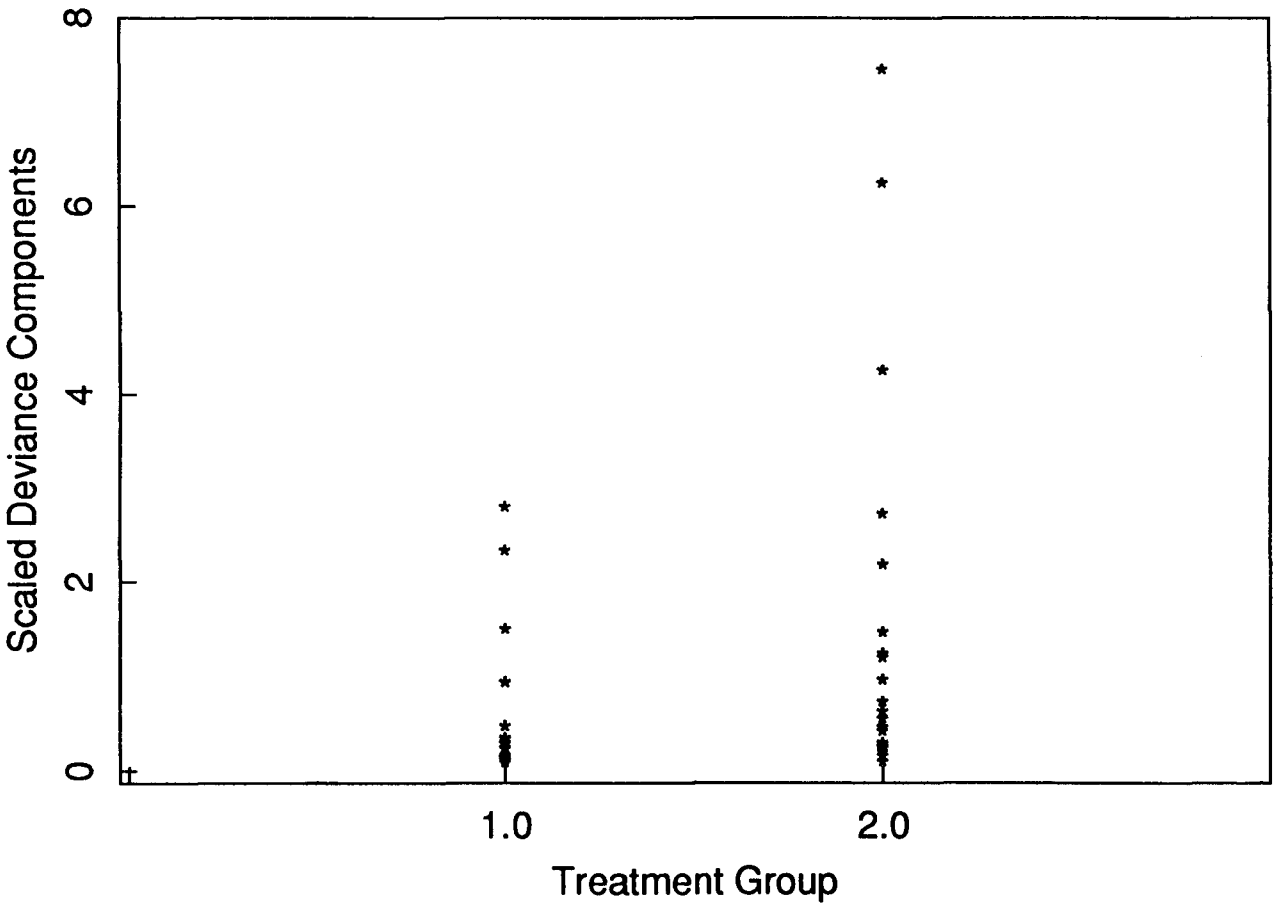
In this experiment, tanks of fish were exposed to one of six doses of either aflatoxin B1 or aflatoxicol and the proportion of fish developing liver cancer was noted. As explained in Chapter 1, it is thought that overdispersion depending on treatment might be present. In order to apply the diagnostic plot to investigate the dependence of overdispersion on treatment group, model (1) was fit to the data by maximum quasi-likelihood.

A mean was fit to each treatment/dose group and overdispersion was accounted for with a constant heterogeneity factor. From this fit, deviance components, $d(y_i, \tilde{\mu}_i)$, and the estimated heterogeneity factor, $\varphi(\tilde{\alpha}_0)$, were obtained.

Figure 4.1 is a plot of $\varphi(\tilde{\alpha}_0)d(y_i, \tilde{\mu}_i)$ versus treatment group. The plot suggests that the overdispersion varies between the aflatoxicol and the aflatoxin B1 treatment groups. The score test statistic was calculated to be S = 4.130 on one degree of freedom (p = 0.0421). Figure 4.2 is a plot of the scaled deviance components against the constructed variable, $\tilde{z}_i = (m_i-1)\tilde{\mu}_i(1-\tilde{\mu}_i)$, and there does not appear to be a strong relationship. The score test statistic for the corresponding test is S = 4.037 with one degree of freedom (p = 0.0445).

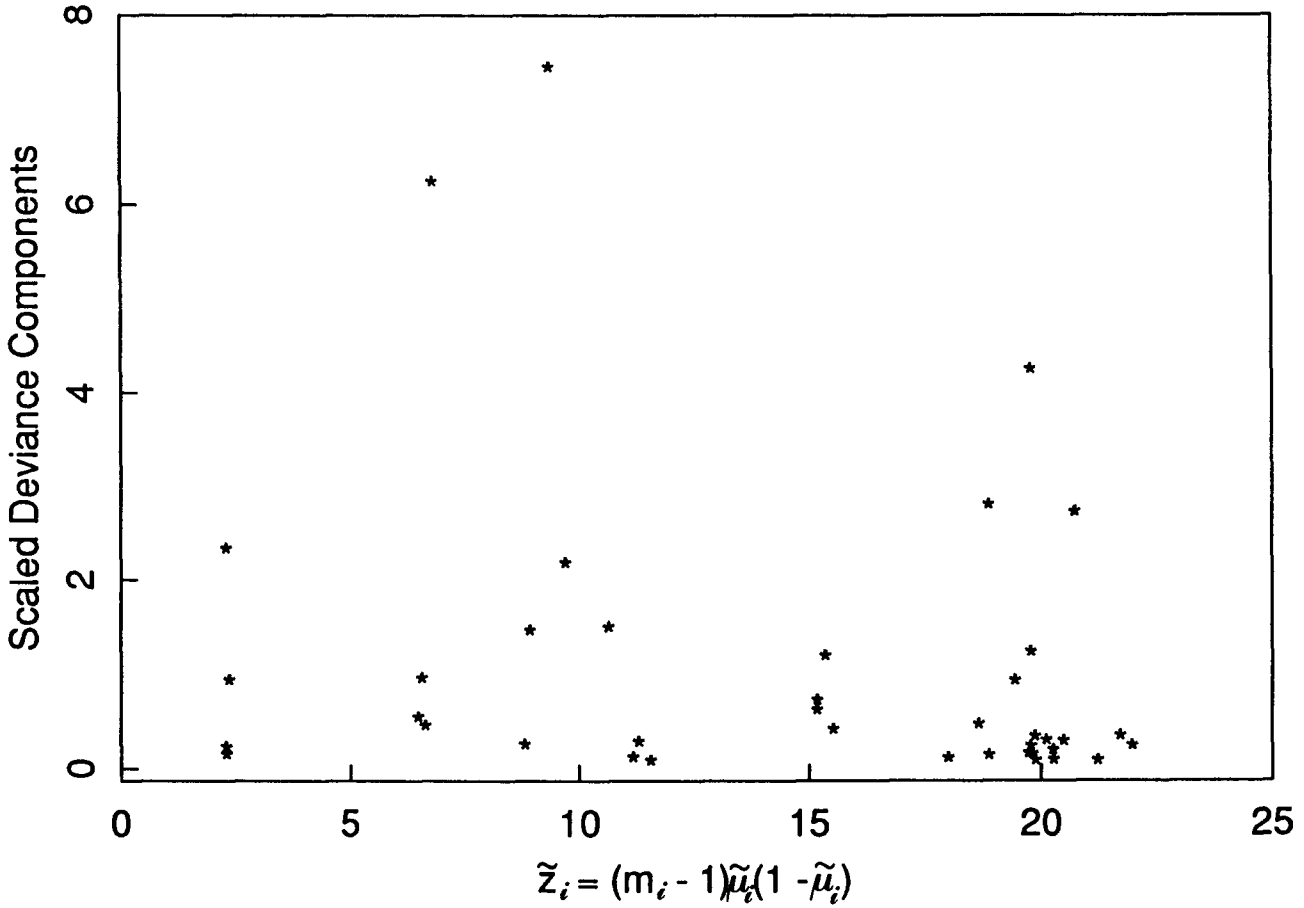
Although Figures 4.1 and 4.2 seem to indicate that the extra variation may depend on treatment group and not on $\tilde{z}_i$, the score test statistics have very similar p-values. A closer examination of the data, see Figure 1.1, shows that the difference between treatment

Figure 4.1 Fish Toxicology Data



Aflatoxin B1 = 1 Aflatoxicol = 2

Figure 4.2 Fish Toxicology Data



group means is approximately the same for all dose groups except group 6. When dose group 6 is omitted from the analysis, and the score test of the hypothesis, "overdispersion does not depend on treatment group" is conducted, S = 5.015 (p = 0.025) and the score test statistic of the hypothesis, "overdispersion does not depend on the constructed variable $\tilde{z}_i$", is S = 2.393 (p = 0.122). So the evidence that overdispersion depends on treatment group is stronger if group 6 is omitted. The researchers explained that they expected the outcomes for dose group 6 to be different than the outcomes for the other groups. A large dose of toxin was given to this group and led to the death of fish rather than to the development of cancer.

This analysis confirms what the researchers had initially expected; there may be more overdispersion in the aflatoxicol group due to the longer metabolic pathway to cancer for this compound than for aflatoxin B1.

4.4.2 Fish Vaccination Data

In this experiment the effectiveness of two types of fish vaccine treatments were compared. In order to compare the relative risks of death of a control group, an inoculated group and an immersed group, a logit regression model for $E(Y_i) = \mu_i$ was fit with factors representing experiment and treatment groups and virus dilution as a continuous covariate. It is suspected that overdispersion may vary between treatment groups and the diagnostic score test and plot were applied to the data to investigate this possibility.

Figure 4.3 is a plot of $\varphi(\tilde{\alpha}_0)d(y_i,\tilde{\mu}_i)$ versus the code for treatment group where $\tilde{\mu}_i$ and $\varphi(\tilde{\alpha}_0)$ are the estimates under model (1) using maximum quasi-likelihood. The plot suggests that overdispersion does vary between treatment groups. But rather than the immersed treatment group having more overdispersion as was initially proposed, it appears that the control group has more overdispersion relative to the vaccinated groups.

The score test of the null hypothesis that overdispersion does not depend on treatment group (control versus treated), yields a test statistic of S = 9.64 with one degree of freedom (p = 0.002) which also suggests that $\varphi_i = \varphi(\alpha_0 + \alpha_1 z_i)$ is to be preferred to the simple model $\varphi_i = \varphi(\alpha_0)$.

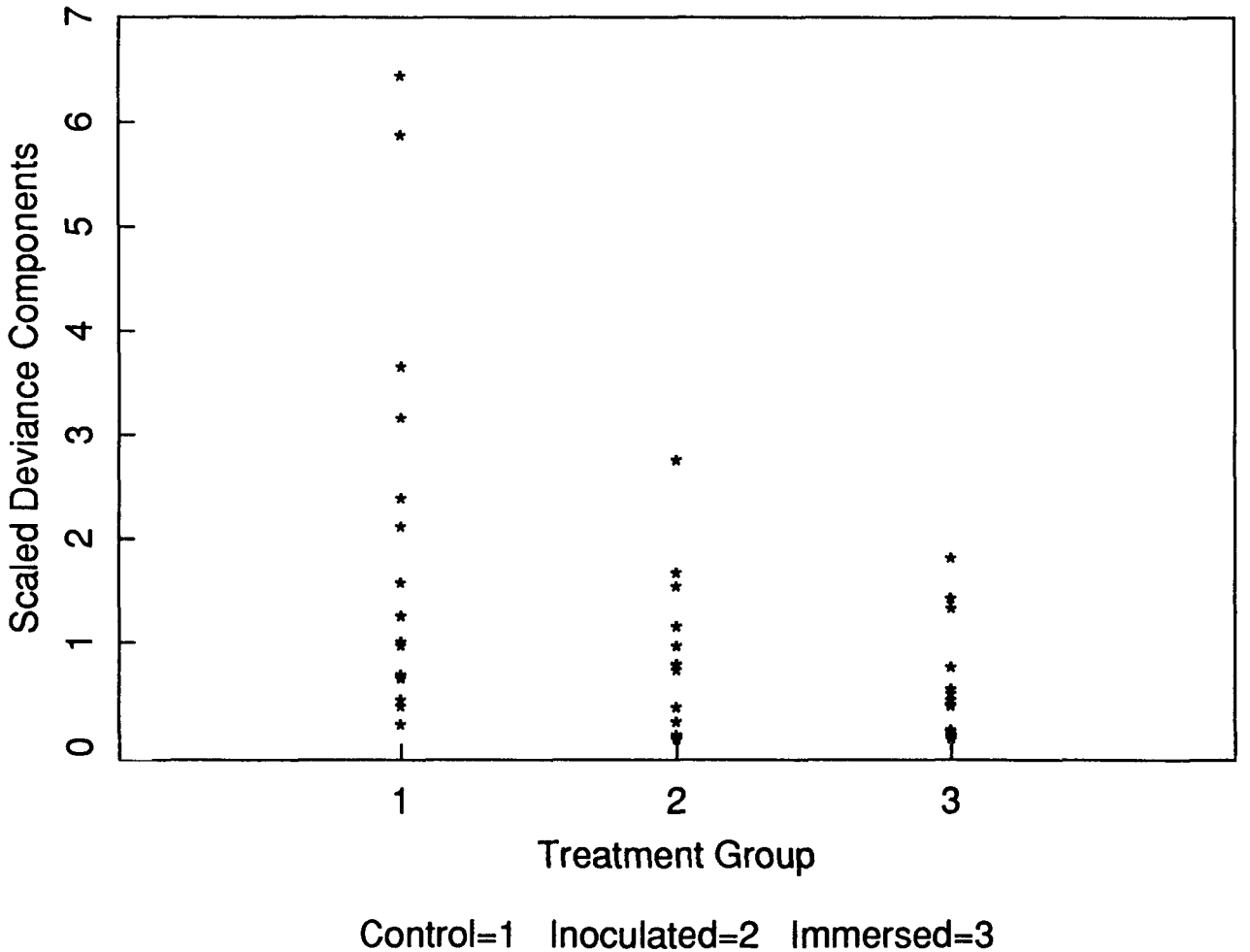
In the investigation of the appropriateness of model (3), a plot of $d(y_i;\tilde{\mu}_i)$ versus the constructed variable, $\tilde{z}_i = (m_i-1)\tilde{\mu}_i(1-\tilde{\mu}_i)$, did not show any trends and the score test statistic for the corresponding test is S = 0.20 with one degree of freedom (p = 0.66). Both the plot and the score test suggest that model (3) is not appropriate.

The diagnostic plot was able to detect a pattern in the overdispersion that was different than what had been initially suspected; it showed that groups receiving the vaccine in any form exhibited less variability than the control group.

4.4.3 Salmonella Bacteria Data

In this example, it is desired to investigate potential models for overdispersion. A log-linear model was fit to the data using log(dose) and $\log(\text{dose})^2$ as independent covariates, a factor

Figure 4.3 Fish Vaccination Data



representing replicate effects and replicate x log(dose) interactions. Several graphs were constructed and plots of deviance components versus log(dose) and $\log(\text{dose})^2$ did not reveal any obvious trends. However, Figure 4.4 shows a relationship between the deviance components and the constructed variable $\tilde{z}_i = \tilde{\mu}$. The score test statistic for the inclusion of $\tilde{z}_i$ in the dispersion parameter is S = 4.967 with one degree of freedom (p=0.026). Overdispersion doesn't appear to depend on covariates included in the model for the mean, but there is evidence that the overdispersion is not constant for all observations and that model (3) may be a more appropriate choice than either model (1) or model (2) for this data.

4.4.4 Chromosome Aberration Data

The application of the diagnostic procedures to the chromosome aberration data will be used to check on the form of the overdispersion. Under the measurement error model for this data, the variance should be quadratic in estimated radiation dose. A regression model was fit to the proportion of chromosome aberrations using total estimated radiation dose as the explanatory variable and estimates of the deviance components and a constant dispersion parameter were obtained.

Figure 4.5 is a plot of the scaled deviance components versus the squared estimated dose. The score test statistic for the inclusion of the squared estimated dose is S = 154.2 on 1 degree of freedom ($p < 0.00001$). Both the plot and the score test indicate that there is evidence to support the measurement error model.

Figure 4.4 Salmonella Bacteria Data

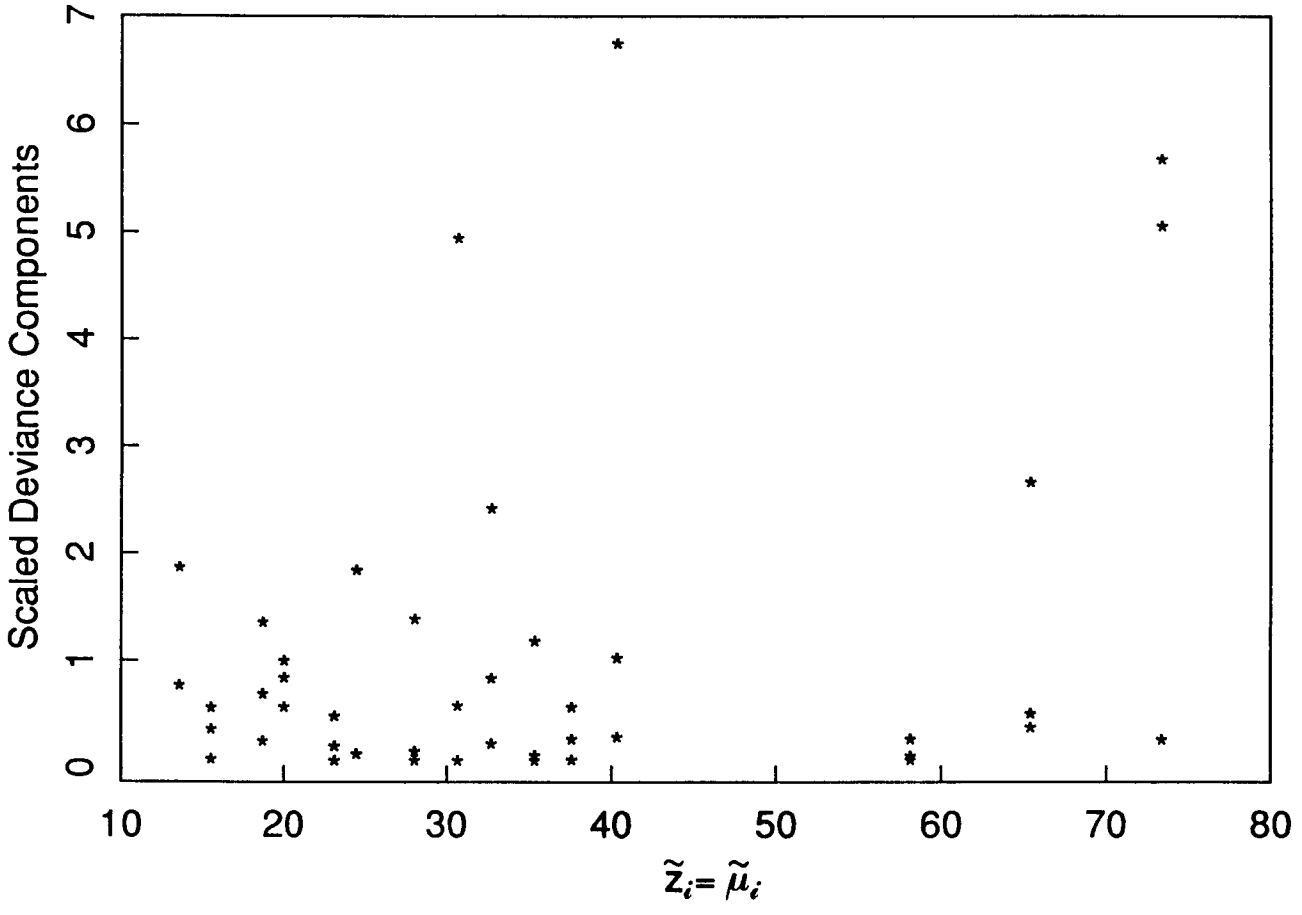
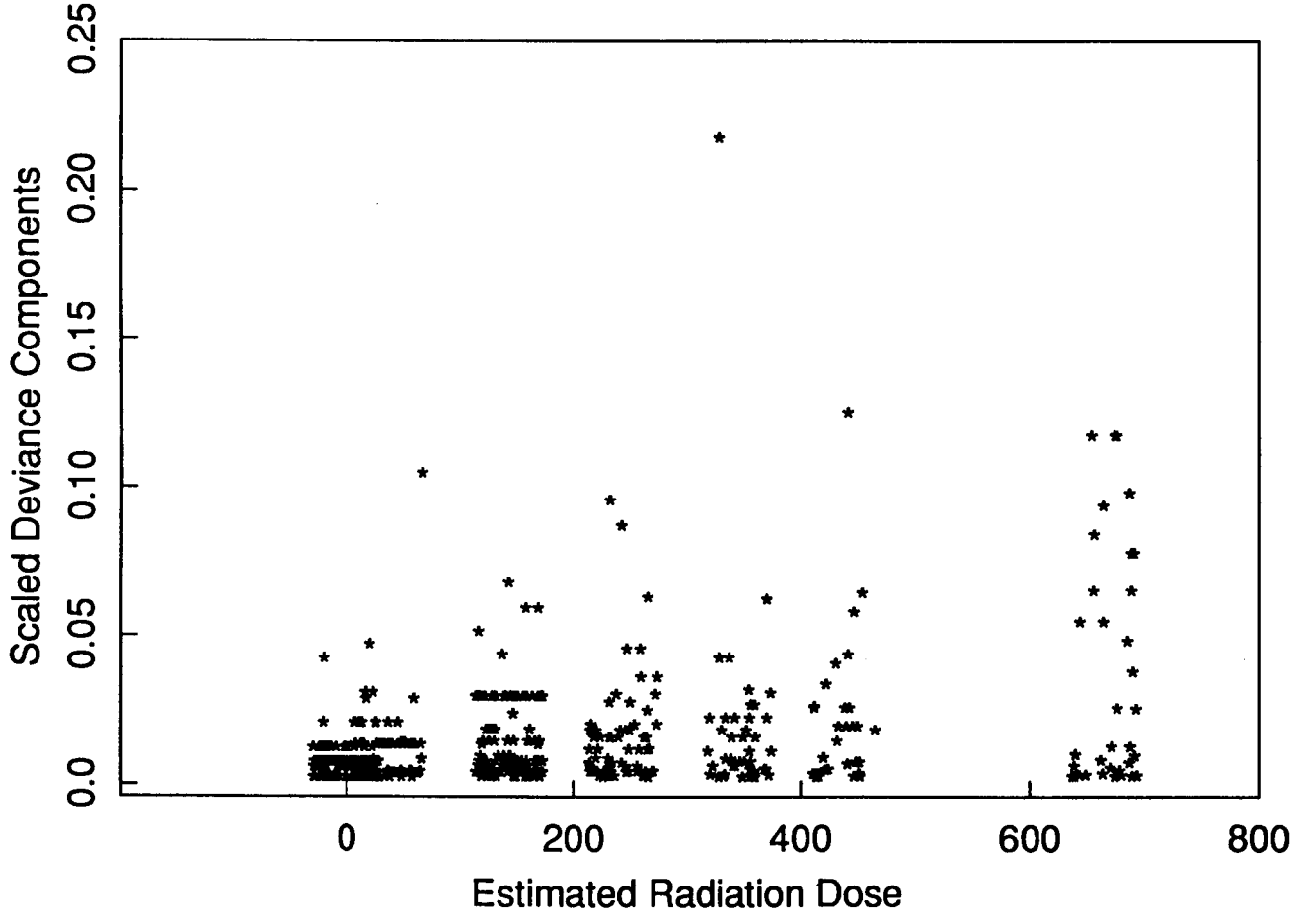


Figure 4.5 Chromosome Aberration Data



4.4.5 Rotenone Data

For this example, researchers are interested in the inclusion of interaction terms in the regression model for the mean. As discussed in Chapter 1, omitting important covariates in the regression model for the mean can make it appear as if overdispersion is present. Therefore, it is important to model the mean as completely as possible before deciding on the model for overdispersion. However, the question here is, just which terms do belong in the regression model for the mean. The procedure will be to fit a rich model for the mean, i.e., including interaction terms, then use the diagnostic tools to assess the dependence of the variability on the type of toxin and the dose level. Once a model for overdispersion has been selected, the inclusion of the interaction terms can be studied.

A probit regression model was fit to the data using [log(toxin concentration)] as a continuous covariate, factors to represent the rotenone, degulin and rotenone+degulin treatment groups and the treatment ^{x} log(concentration) interactions.

From this model deviance components and a constant dispersion parameter were estimated. Figure 4.6 and 4.7 are plots of the scaled deviance components versus log(concentration) and treatment group respectively. Although one deviance component seems larger than the others, no clear pattern emerges. The score test statistic for the dependence of overdispersion on treatment group is S = 2.9 on 2 degrees of freedom (p = 0.2357), and for the dependence of overdispersion on [log(concentration)], S = 0.4363 on 1 degree of

Figure 4.6 Rotenone Data

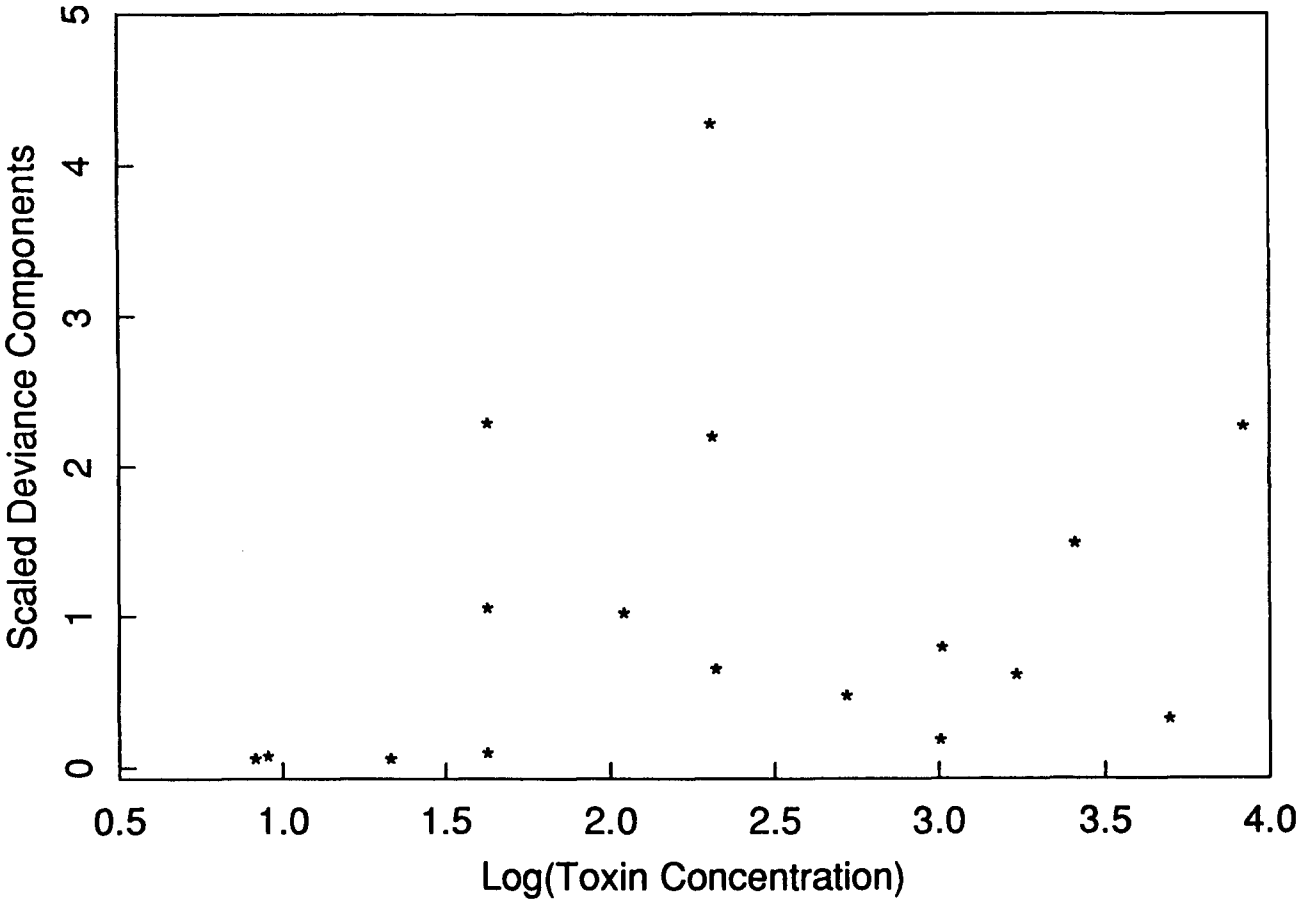
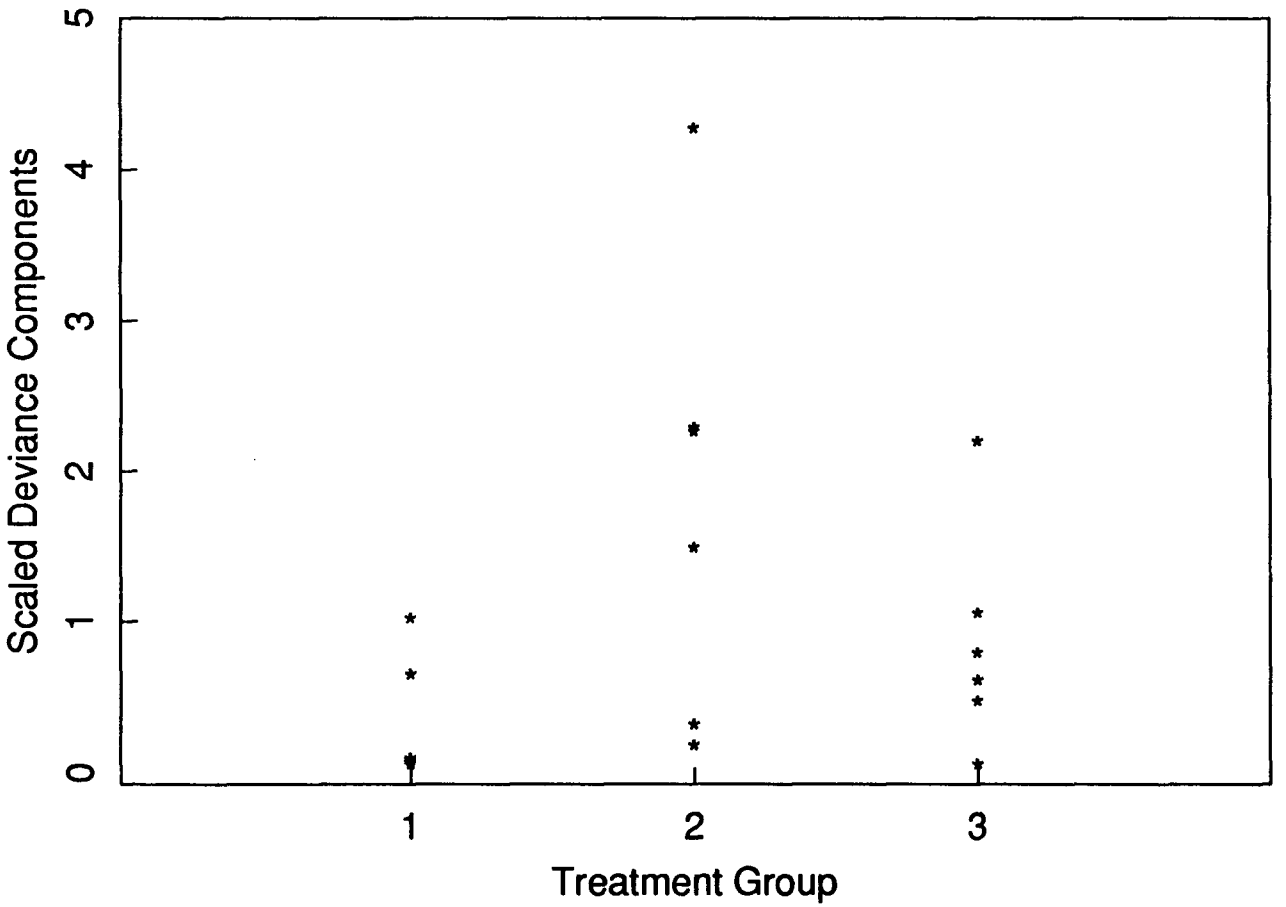


Figure 4.7 Rotenone Data



freedom (p = 0.5089). The diagnostic plot of the scaled deviance components versus the constructed variable, $\tilde{z}_i = (m_i-1)\tilde{\mu}_i(1-\tilde{\mu}_i)$, did not show any relationship either and the associated score test statistic is S = 0.5548 with one degree of freedom (p = 0.456). None of the diagnostic plots exhibited a strong pattern, suggesting that model (1) is an adequate representation of the data.

4.5 DEVELOPMENT OF THE SCORE TEST

If $f(y_i;\mu_i)$ is a one parameter exponential family density,

$$f(y_i;\mu_i) = \exp\{[y_i\theta_i - b(\theta_i)]m_i + c(y_i)\}$$

then the corresponding double exponential family distribution is given by,

$$g(y_i,\mu_i,\varphi_i) = c(\mu_i,\varphi_i)\ \varphi_i^{1/2}\ [f(y_i;\mu_i)]^{\varphi_i}\ [f(y_i;y_i)]^{1-\varphi_i}$$

(see Section 2.2.4). Efron (1986) showed that under the double exponential family distribution, $c(\mu_i,\varphi_i) \approx 1$, $E(Y_i) \approx \mu_i = b'(\theta_i)$, and $Var(Y_i) \approx V(\mu_i)/(m_i\varphi_i) = b''(\theta_i)/(m_i\varphi_i)$.

Suppose $Y_1,\ldots,Y_n$ are independent random variables from a double exponential family distribution with,

$$E(Y_i) = \mu_i = k(\eta_i) \qquad \eta_i = \underline{x}_i'\underline{\beta}$$

$$Var(Y_i) = V(\mu_i)/(m_i\varphi_i) \qquad \varphi_i = w(\gamma_i) \qquad \gamma_i = \alpha_0 + \underline{z}_i'\underline{\alpha}_1$$

where $\underline{x}_i$ is a (px1) vector of known covariates, $\underline{\beta}$ is a (px1) vector of unknown parameters, $k(\cdot)$ is a known,one to one, monotonic function, $V(\cdot)$ is a known positive function, the m_i's are known constants, $\underline{z}_{ic}$ is a (qx1) vector of known, mean corrected covariates (i.e., $\Sigma_i z_{ijc} = 0$), α_0 is a scaler parameter and $\underline{\alpha}_1$ is a (qx1) vector of unknown parameters and $w(\cdot)$ is a known positive function.

Then, the log-likelihood function for the vector $(\underline{\mu},\underline{\varphi})$ is given by,

$$l(\mu,\varphi,y) = \sum_{i=1}^{n} \left[(1/2)\log[\varphi_i] + \varphi_i d(y_i;\mu_i) + \log[f(y_i,y_i)] \right]$$

where $d(y_i;\mu_i)$, the deviance component is,

$$d(y_i;\mu_i) = 2[\, y_i(\tilde{\theta}_i - \theta_i) - b(\tilde{\theta}_i) + b(\theta_i)]/m_i$$

and $\tilde{\theta}_i$ is the maximum likelihood estimate of θ_i based on y_i alone.

The components of the score vector are given by,

$$S_{\underline{\beta}}(\underline{\beta}, \alpha_0, \underline{\alpha}_1) = [\partial l/\partial \underline{\beta}] = \sum_{i=1}^{n} \varphi_i \left[\frac{m_i(y_i - \mu_i)}{V(\mu_i)}\right] \dot{k}(\eta_i)\, \underline{x}_i \quad ,$$

$$S_{\alpha_0}(\underline{\beta}, \alpha_0, \underline{\alpha}_1) = [\partial l/\partial \alpha_0] = (1/2) \sum_{i=1}^{n} \left[\varphi_i^{-1} - d(y_i;\mu_i)\right] \dot{w}(\gamma_i) \quad ,$$

$$S_{\underline{\alpha}_1}(\underline{\beta}, \alpha_0, \underline{\alpha}_1) = [\partial l/\partial \underline{\alpha}_1] = (1/2) \sum_{i=1}^{n} \left[\varphi_i^{-1} - d(y_i;\mu_i)\right] \dot{w}(\gamma_i)\underline{z}_{ic} \, ,$$

where $\dot{k}(\eta_i) = \partial k(\eta_i)/\partial \eta_i$ and $\dot{w}(\gamma_i) = \partial w(\gamma_i)/\partial \gamma_i$. The information matrix is given by,

$$I(\underline{\beta}, \alpha_0, \underline{\alpha}_1) = \begin{bmatrix} \underline{I}_{\beta} & \underline{0} & \underline{0} \\ \underline{0} & I_{\alpha_0} & \underline{0} \\ \underline{0} & \underline{0} & I_{\underline{\alpha}_1} \end{bmatrix}$$

where,

$$I_{\underline{\beta}}(\underline{\beta}, \alpha_0, \underline{\alpha}_1) = -E[\partial^2 l/(\partial \underline{\beta}\ \partial \underline{\beta}')]$$

$$= \sum_{i=1}^{n} [\varphi_i m_i/V(\mu_i)]\ \dot{k}(\eta_i)^2\ \underline{x}_i\ \underline{x}_i'$$

$$I_{\alpha_0}(\underline{\beta}, \alpha_0, \underline{\alpha}_1) = -E[\partial^2 l/(\partial \alpha_0^2)]$$

$$= (1/2) \sum_{i=1}^{n} \left\{ [\varphi_i^{-1} \dot{w}(\gamma_i)]^2 - [\varphi_i^{-1} \ddot{w}(\gamma_i)] + \ddot{w}(\gamma_i) E[d(y_i; \tilde{\mu}_i)] \right\}$$

$$I_{\underline{\alpha}_1}(\underline{\beta}, \alpha_0, \underline{\alpha}_1) = -E[\partial^2 l/(\partial \underline{\alpha}_1 \partial \underline{\alpha}_1')]$$

$$= (1/2) \sum_{i=1}^{n} \left\{ [\varphi_i^{-1} \dot{w}(\gamma_i)]^2 - [\varphi_i^{-1} \ddot{w}(\gamma_i)] + \ddot{w}(\gamma_i) E[d(y_i; \tilde{\mu}_i)] \right\} \underline{z}_{ic} \underline{z}_{ic}'$$

where $\ddot{w}(\gamma_i) = \partial \dot{w}(\gamma_i)/\partial \gamma_i$.

Now, $E[d(y_i; \tilde{\mu}_i)] \approx \varphi_i^{-1}$ (Efron, 1986) when the binomial sample size is large or when the Poisson mean is large, that is, as $m_i \to \infty$. But results from Pierce and Schafer (1986) suggest that this approximation may be good even for small m_i. Using this approximation,

$$I_{\alpha_0}(\underline{\beta}, \alpha_0, \underline{\alpha}_1) = (1/2) \sum_{i=1}^{n} \left\{ \varphi_i^{-1} \dot{w}(\gamma_i) \right\}^2$$

$$I_{\underline{\alpha}_1}(\underline{\beta}, \alpha_0, \underline{\alpha}_1) = (1/2) \sum_{i=1}^{n} \left\{ \varphi_i^{-1} \dot{w}(\gamma_i) \right\}^2 \underline{z}_{ic} \underline{z}_{ic}' \quad .$$

The score test statistic for the null hypothesis $\underline{\alpha}_1 = \underline{0}$ is given by,

$$S = S_{\underline{\alpha}_1}(\tilde{\underline{\beta}}, \tilde{\alpha}_0, \underline{0})' \; [I_{\underline{\alpha}_1}(\tilde{\underline{\beta}}, \tilde{\alpha}_0, \underline{0})]^{-1} \; S_{\underline{\alpha}_1}(\tilde{\underline{\beta}}, \tilde{\alpha}_0, \underline{0})$$

where $\underline{\beta}$ and α_0 are the maximum likelihood estimates when $\underline{\alpha}_1 = \underline{0}$. Note that when $\underline{\alpha}_1 = \underline{0}$,

$$\varphi_i = w(\gamma_i) = w(\alpha_0)$$

which does not depend on i, so

$$S_{\underline{\alpha}_1}(\underline{\beta}, \alpha_0, \underline{0}) = (1/2)\dot{w}(\gamma) \left\{ \varphi_i^{-1} \sum_{i=1}^{n} \underline{z}_{ic} - \sum_{i=1}^{n} d(y_i; \mu_i)\underline{z}_{ic} \right\}$$

$$= -(1/2)\dot{w}(\gamma)\underline{Z}_c'\underline{d} \qquad \text{since} \sum_{i=1}^{n} z_{ijc} = 0,$$

where $\underline{d}$ is an (nx1) vector with i^{th} entry $d(y_i; \mu_i)$ and $\underline{Z}_c$ is an (nxp) matrix with i^{th} row $\underline{z}_{ic}'$.

Also,

$$I_{\underline{\alpha}_1}(\underline{\beta}, \alpha_0, \underline{0}) = (1/2)[\varphi^{-1}\dot{w}(\gamma)]^2 \; \underline{Z}_c'\underline{Z}_c \quad .$$

So the score test statistic is given by,

$$S = \left\{-(1/2)\dot{w}(\gamma)\underline{Z}_c'\tilde{\underline{d}}\right\}' \left\{2[\tilde{\varphi}^{-1}\dot{w}(\gamma)]^2\right\}^{-1} \left\{\underline{Z}_c'\underline{Z}_c\right\}^{-1} \left\{-(1/2)\dot{w}(\gamma)\underline{Z}_c'\tilde{\underline{d}}\right\}$$

$$= (1/2)\ \tilde{\varphi}^2\ '\tilde{\underline{d}}'\underline{Z}_c\ (\underline{Z}_c'\underline{Z}_c)^{-1}\ \underline{Z}_c'\tilde{\underline{d}}$$

where $\tilde{\mu}_i$ and $\tilde{\varphi}$ are the maximum likelihood estimates when $\underline{\alpha}_1 = \underline{0}$ and $\tilde{\underline{d}}$ has i^{th} entry $d(y_i;\tilde{\mu}_i)$.

Chapter 5

CONCLUDING REMARKS

For the analysis of overdispersed counts or proportions many different models for the variance of Y have been suggested. Finney's model, in which the variance is assumed to be a constant multiple of the binomial or Poisson variance, is useful for many applications. However, as McCullagh and Nelder (1983) and Nelder and Wedderburn (1987) point out, it is important to check the validity of this model. In the examples of Chapter 1 there are *a priori* reasons to suspect that overdispersion depends on known covariates or on some function of the mean. The diagnostic tools proposed in Chapter 4, the score test and scatter plot based on deviance components, provide an easy way to investigate the suspected dependencies with standard computer programs. The following points are to be noted in assessing the extra variability and in using the diagnostics suggested in this thesis.

1. Chapter 3 showed that the major effect of ignoring the dependence of overdispersion on covariates will be the use of incorrect standard errors of the estimated β_j's and related errors regarding the conclusions of tests. These were demonstrated through the actual (asymptotic) coverage probabilities of nominal 95% confidence intervals. For two independent samples from a double binomial population, with different amounts of overdispersion but the

same mean, the difference between the true coverage probability and the nominal value depends on (1) the ratio of the Finney heterogeneity factor in the groups, (2) the ratio of the number of observations in each sample and (3) the ratio of the total number of Bernoulli trials in each sample. When all of these ratios are between 1/2 and 2, the coverage probabilities are between 92% and 97%. For a more extreme case, consider the following example. If the first group has five times as many observations and five times as much extra variation compared to the second group, but there are five times more Bernoulli trials in the second group, the true coverage probability of the asymptotic nominal 95% confidence interval is 78%. However, such differences in sample sizes and binomial denominator sizes do not seem likely in practice.

When overdispersion depends on a continuous covariate, the difference between the true coverage probability and the nominal value depends on the values of the covariate. For the constructed example in Chapter 3, the true coverage probability was 90%, indicating that inference can be affected by ignoring the dependence of the overdispersion on the covariate.

For estimating means from independent samples from populations with different amounts of overdispersion, there is no loss of asymptotic efficiency. But for the constructed example in which overdispersion depended on on a covariate, the asymptotic relative efficiency of the vector $\underline{\beta}$ dropped to 86%. So ignoring the dependence of overdispersion on covariates may also lead to a loss of efficiency.

2. The diagnostic tools proposed in Chapter 4 are easy to obtain using standard software for generalized linear models. First, obtain the maximum likelihood fit ignoring overdispersion (or equivalently, the maximum quasi-likelihood fit) and obtain the deviance components, $d(y_i;\tilde{\mu}_i)$, and the average deviance, $\bar{D} = [n^{-1} \Sigma_i d(y_i;\tilde{\mu}_i)]$. Plots of $d_i^* = [d(y_i;\tilde{\mu}_i)/\bar{D}]$ versus the suspected covariates can then be used to investigate the dependence of extra variation on these covariates. The score statistic is one half of the regression sum of squares in the regression of d_i^* on the suspected covariates, and this may be compared to a chi-squared distribution with degrees of freedom equal to the rank of the matrix of the suspected covariates.

3. In this thesis, interest lies mainly with inference about the β_j's. Here it is important to account for overdispersion but models for the variance do not need to be realistic or "exact" to ensure good inferential techniques for the parameter $\underline{\beta}$. Often several models may be suitable. For example, if Y is a count, with $h(\mu) = \beta_0 + \beta_1 x$, the diagnostic tools might indicate that the residual variation depends on the covariate x. However, a simpler model might be one in which the variance depends on x only through the mean, such that $\text{Var}(Y) = \mu(1 + \sigma^2\mu)$. The latter would ordinarily be chosen over $\text{Var}(Y) = \mu(\alpha_0 + \alpha_1 x)$ since it involves the covariate only through the mean. The diagnostic tools can be used to decide when a simple model can serve as an adequate representation of the

data, but as in ordinary regression, there is not typically one best model.

4. It is important to use a rich model for the mean when investigating possible models for the variance, since covariates omitted from the model for E(Y) can result in overdispersion which depends on the omitted covariate (or on any variable correlated with the omitted covariate). However, in many cases it can be difficult or impossible to distinguish between interactions in the model for E(Y) and overdispersion based on the data. The choice of which to include can sometimes be settled by the prior knowledge of the researcher. In the absence of such knowledge, simple and interpretable models are preferred.

5. For large sample sizes, the score test statistic has a chi-squared distribution. But the small sample properties of the score test statistic proposed here have not yet been explored. Cook and Weisberg (1983) conducted a small simulation to explore the chi-squared approximation for the case of normally distributed data. They found that, in general, the approximation leads to a conservative test and appears to be adequate for diagnostic purposes.

6. Outliers can affect the judgment of models for the variability. Influential points make the overall residual variability appear to be larger than it really is. The diagnostic plot can help to detect outliers, but points with a large amount of

leverage will be close to the regression line and detecting them can be difficult. For this reason, it is recommended that each deviance component be replaced by the deviance component divided by $(1-h_{ii})$, where h_{ii} is the i^{th} diagonal entry of the generalized linear model equivalent of the "hat" matrix, as suggested by McCullagh and Nelder (Section 11.5.3, 1983). This improvement was also suggested by Cook and Weisberg (1983).

7. Efron, (1986, Remark 12) and Carroll and Rupert (1988) note that difficulties can arise when the regression model for the variance has been overfit. Using rich models for the variance in the double exponential family, for example, as a method of estimating β, which is robust to the form of overdispersion may lead to similar problems. For this reason, a graphical tool such as the one presented here will be useful even when a more sophisticated computer program for modeling the overdispersion is available.

BIBLIOGRAPHY

Aeschbacher, H. U., Vuataz, L., Sotek J., and Stalder R. (1977), "Use of the Beta-Binomial Distribution in Dominant-Lethal Testing for Weak Mutagenic Activity," *Mutation Research*, 44, 369-390.

Altham, P. M. E. (1978), "Two Generalizations of the Binomial Distribution," *Applied Statistics* , 27, 162-167.

Baker, R. J., and Nelder, J. A. (1978), "The GLIM System, Release 3," Numerical Algorithms Group, Oxford.

Breslow, N. E. (1984), "Poisson Variation in Log-linear Models," *Applied Statistics*, 33, 38-44.

Carroll, R. J., and Ruppert, D. (1988), *Transformation and Weighting in Regression*, Chapman and Hall, New York.

Chatfield C., and Goodhart, G. J (1970), "The Beta-Binomial Model for Consumer Purchasing Behavior," *Applied Statistics*, 19, 240-250.

Cochran, W. G. (1943), "Analysis of Variance for Percentages Based on Unequal Numbers," *Journal of the American Statistical Association*, 38, 287-301.

Collings, B. J., and Margolin, B. H. (1985), "Testing Goodness of Fit for the Poisson Assumption When Observations are Not Identically Distributed," *Journal of the American Statistical Association*, 80, 411-418.

Cook, R. D., and Weisberg S. (1982), *Residuals and Influence in Regression*, Chapman and Hall, London.

Cook, R. D., and Weisberg S. (1983), "Diagnostics for Heteroscedasticity in Regression," *Biometrika*, 70, 1-10.

Cox, D. R. (1983), "Some Remarks on Overdispersion," *Biometrika*, 70, 269-274.

Davidian, M., and Carroll, R. J.(1987), "Variance Function Estimation," *Journal of the American Statistical Association*, 82, 1079-1091.

Dean, C., and Lawless, J. F. (1989), "Tests for Detecting Overdispersion in Poisson Regression Models," *Journal of the American Statistical Association*, 84, 467-472.

Diaconis, P., and Efron, B. (1985), "Testing for Independence in a Two-Way Table: New Interpretations of the Chi Square Statistic," *The Annals of Statistics*, 13, 845-874.

Draper, N. R., and Smith, H. (1981), *Applied Regression Analysis*, Second Edition, John Wiley and Sons, New York.

Efron, B. (1986), "Double Exponential Families and Their Use in Generalized Linear Regression," *Journal of the American Statistical Association*, 81, 709-721.

Finney, D. J. (1971), *Probit Analysis*, Cambridge University Press, London, Third Edition.

Firth, D. (1987), "On the Efficiency of Quasi-Likelihood Estimation," *Biometrika*, 74, 233-245.

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.

Greenwood M., and Yule G. Y. (1920), "Inquiry Into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents," *Royal Statistical Society*, Ser. A, 83, 255-279.

Griffiths, D. A. (1973), "Maximum Likelihood Estimation for the Beta-Binomial Distribution and an Application to the Household Distribution of the Total Number of Cases of a Disease," *Biometrics*, 29, 637-648.

Haseman, J. K., and Soares, E. R. (1976), "The Distribution of Fetal Death in Control Mice and Its Implication on Statistical Tests for Dominant Lethal Effects," *Mutation Research*, 41, 277-288.

Jorgensen, B. (1987), "Exponential Dispersion Models," *Journal of the Royal Statistical Society*, Ser. B, 49, 127-162

Kemp, C. D., and Kemp, A. W. (1956), "The Analysis of Point Quadrat Data," *Australian Journal of Botany*, 4, 167-174.

Kleinman, J. C. (1973), "Proportions With Extraneous Variance: Single and Independent Samples," *Journal of the American Statistical Association*, 68, 56-64.

Kupper, L. L., and Haseman, J. K., (1978), "The Use of a Correlated Binomial Model for the Analysis of Certain Toxicological Experiments," *Biometrics*, 34, 69-76.

Kupper, L. L., Portier, C., Hogan, M. D., and Yamamoto, E. (1986), "The Impact of Litter Effects in Dose Response Modeling in Teratology," *Biometrics*, 42, 85-98.

Lawless, J. F. (1987), "Negative Binomial and Mixed Poisson Regression," *The Canadian Journal of Statistics*, 15, 209-225.

Martin, J. T. (1942), "The Problem of the Evaluation of Rotenone Containing Plants. VI The Toxicity of 1-Elliptone and of Poisons Applied Jointly, With Further Observations on the Rotenone Equivalent Method of Assessing the Toxicity of *Derris* Root," *Annals of Applied Biology*, 29, 69-81.

McCaughran, D. A., and Arnold, D. W. (1976), "Statistical Models for Numbers of Implantation Sites and Embryonic Deaths in Mice," *Toxicology and Applied Pharmacology*, 38, 325-333.

McCullagh, P. (1973), "Quasi-Likelihood Functions," *The Annals of Statistics*, 11, 59-67.

McCullagh, P., and Nelder J. A. (1983), *Generalized Linear Models*, New York and London, Chapman and Hall.

Moore, D. F. (1985), *Method of Moments Estimation for Overdispersed Counts and Proportions*, University of Washington, Seattle, Ph.D dissertation.

Moore, D. F. (1987), "Modelling the Extraneous Variance in the Presence of Extra Binomial Variation," *Applied Statistics*, 36, 8-14.

Mosteller, F., and Youtz, C. (1961), "Table of the Freeman-Tukey Transformation for the Binomial and Poisson Distribution," *Biometrika*, 48, 433-440.

Nelder, J. A., and Pregibon, D. (1987), "An Extended Quasi-likelihood Function," *Biometrika*, 74, 221-232.

Nelder, J. A., and Wedderburn, R. W. M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society*, Ser. A, 135, 370-384.

Ochi, Y., and Prentice, R. L. (1984), "Likelihood Inference in a Correlated Probit Regression Model," *Biometrika*, 71, 531-543.

Otake M., and Prentice, R. L. (1984), "The Analysis of Chromosomally Aberrant Cells Based on a Beta-Binomial Distribution," *Radiation Research*, 98, 456-470.

Pack, S. E. (1986), "Hypothesis Testing for Proportions with Overdispersion," *Biometrics*, 42, 967-972.

Pierce, D. A.,.and Sands B. R. (1975), "Extra Bernoulli Variation in Regression of Binary Data," *Technical Report #46*, Department of Statistics, Oregon State University, Corvallis, OR. 97331.

Pierce, D. A., and Schafer, D. W. (1986), "Residuals in Generalized Linear Models," *Jounal of the American Statistical Association*, 81, 977-986.

Prentice, R. L. (1986), "Correlated Binary Regression Using an Extended Beta-Binomial Distribution With Discussion of Correlation Induced by Covariate Measurement Errors," *Journal of the American Statistical Association*, 81, 321-327.

Simpson, D. G., and Margolin, B. H. (1986), "Recursive Non-Parametric Testing for Dose-Response Relationships Subject to Downturns at High Doses," *Biometrika*, 73, 589-596.

Skellam J. G. (1948), "A Probability Distribution Derived From the Binomial Distribution by Regarding the Probability of Success as Variable Between the Sets of Trials," *Journal of the Royal Statistical Society*, Ser. B, 10, 257-265.

Smyth, G. K. (1989), "Generalized Linear Models With Varying Dispersion," *Journal of the Royal Statistical Society*, Ser. B, 51, 47-60.

Wedderburn, R. W. M. (1974), "Quasi-likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method," *Biometrika*, 61, 439-447.

West, M. (1985), "Generalized Linear Models: Scale Parameters, Outlier Accommodation and Prior Distributions," *Bayesian Statistics 2*, Amsterdam, North Holland.

Wetherill, G. B. (1981), *Intermediate Statistical Methods*, London, Chapman and Hall Ltd.

White, J. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1-25.

Williams, D. A. (1975) "The Analysis of Binary Responses From Toxicological Experiments Involving Reproduction and Teratogenicity," *Biometrika*, 31, 949-952.

Williams, D. A. (1982), "Extra-binomial Variation in Logistic Linear Models," *Applied Statistics*, 31, 144-148.

Williams, D. A. (1988), "Estimation Bias Using the Beta-Binomial Distribution in Teratology", *Biometrics*, 44, 305-309.

APPENDICES

APPENDIX A

Fish Toxicology Data. Number of Fish Developing Liver Cancer.

Dose Group	Aflatoxin B1		Aflatoxicol	
	# in Tank	# with cancer	# in Tank	# with cancer
1	86	3	87	9
1	86	5	86	5
1	88	4	89	2
1	86	2	85	9
2	87	14	86	30
2	90	14	86	41
2	83	9	86	27
2	88	12	88	34
3	90	29	89	54
3	89	31	86	53
3	89	33	90	64
3	87	26	88	55
4	86	44	88	71
4	80	40	89	73
4	89	44	88	65
4	88	43	90	72
5	87	62	86	66
5	88	67	82	75
5	88	59	81	72
5	84	58	89	73
6	77	62	54	46
6	78	63	56	39
6	79	62	55	43
6	79	67	55	43

APPENDIX B

Fish Vaccination Data. Number of Fish Dying From Viral Infection.

Experiment	Dilution	Control		Immersed		Inoculated	
		# in Tank	# Dead	# in Tank	# Dead	# in Tank	# Dead
1	-2	25	18	25	4	25	4
1	-3	25	13	25	4	25	3
1	-4	25	1	25	0	25	1
1	-5	25	0	25	0	25	0
2	-2	25	21	25	17	25	8
2	-3	25	15	25	6	25	3
2	-4	25	12	25	1	25	0
2	-5	25	8	25	0	25	0
3	-2	25	25	25	18	25	14
3	-3	25	25	25	7	25	5
3	-4	25	14	25	2	25	0
3	-5	25	3	25	0	25	0
4	-3	32	32	25	23	24	18
4	-4	37	37	25	15	26	13
4	-5	30	20	25	5	22	1
4	-6	30	5	27	3	23	0

APPENDIX C

Salmonella Bacteria Data. Number of Visible Colonies After Exposure to Acid Red 114.

	Dose (μg/ml)				
Replicate	100	333	1000	3333	10000
1	60	98	60	22	23
1	59	78	82	44	21
1	54	50	59	33	25
2	15	26	39	33	10
2	25	17	44	26	8
2	24	31	30	23	
3	27	28	41	28	16
3	23	37	37	21	19
3	21	35	43	30	13

APPENDIX D

Chromosome Aberration Data; Hiroshima. Number of Cells With Aberrations Out of 100 CellS Per Subject.

Number of Aberrant Cells	Estimated Radiation Dose (rads/100)						
	0	0-.99	1-1.99	2-2.99	3-3.99	4-4.99	5+
0	139	20	23	2	1		3
1	66	23	12	2	1		1
2	35	6	20	5	1	1	2
3	17	7	23	5	1	1	2
4	3	3	6	3	3	3	2
5	2	2	12	14	3	4	
6	1	5	12	3	3	1	1
7		2	12	2	3		
8			5	4	3	2	1
9		1	2	3	3	1	
10			5	5	3	3	
11			1	2	1	1	1
12				2	1	1	1
13				7	3	2	1
14		1	1	3	2	2	1
15			1	1		1	
16			2	2		1	3
17			1	2	2	1	4
18					4		
19					2		1
20					1	1	1
21				1			1
22					1	1	1
23							1
24				1			1
25				1		1	1
26-27							
28						1	
29						1	1
30-33							
34							1
35-36							
37						1	
38-39							
40							1
41							
42					1		1

APPENDIX E

Rotenone Data. Number of Aphids Dead Due to Insecticide.

Concentration (mg/l)		# of insects used	# dead
Rotenone	Degulin		
10.2	0.0	50	44
7.7	0.0	49	42
5.1	0.0	46	24
3.8	0.0	48	16
2.6	0.0	50	6
0.0	50.5	48	48
0.0	40.4	50	47
0.0	30.3	49	47
0.0	20.2	48	38
0.0	10.1	48	18
0.0	5.1	49	16
5.1	20.3	50	48
4.0	16.3	46	43
3.0	12.2	48	38
2.0	8.1	46	27
1.0	4.1	46	22
0.5	2.0	47	7