Title
Transform-Both-Sides Approach for Overdispersed Binomial Data When N is Unobserved

Authors
Dong K. Kim
Jeremy M. G. Taylor

Publication Date
2011-10-24

Transform-Both-Sides Approach for Overdispersed Binomial Data When N is Unobserved

Dong K. Kim and Jeremy M. G. Taylor*

A common complication in analyzing binomial data is overdispersion, where the observed variation exceeds that predicted from the binomial distribution with parameters N and P. We consider the situation where N is not observed and variable. To estimate the regression parameters associated with covariates, we apply the transform-both-sides method. Based on the first-order asymptotic variance stabilizing transformation, we develop the arcsine transformation family indexed by a single parameter. This family includes the square root and the arcsine transformations as special cases. Asymptotic properties of the transformation methods are obtained. Simulation study indicates that the arcsine transformation family is more efficient than the square root and arcsine transformations when there is moderate overdispersion. These approaches are applied to a data set from radiobiology.

KEY WORDS: Arcsine transformation family; Dose–response model; Extrabinomial variation; Variance stabilizing transformation.

1. INTRODUCTION

1.1 Overdispersion Due to Variability of N

For binomial outcomes, y_i, $i = 1, 2, \ldots, n$ with parameters N_i and P_i, the observed variation of the response frequently exceeds the nominal variance determined by N_i and P_i. In epidemiologic studies or in certain toxicological experiments with laboratory animals, overdispersion occurs due to variability of P_i, which varies from unit to unit of the experiment (Haseman and Kupper 1979; Williams 1982).

Overdispersion due to variability of N_i can arise in certain dose–response models when the binomial count, N_i, is not known exactly (Kim 1991). This overdispersion problem can occur in various situations; for example:

1. In dose-response models, we observe the number of cells that are alive after a certain dose, but we are not exactly sure of the total number of cells before the dose is given. The response is the number of cells that remain after the treated dose.
2. We count the number of people who have a certain disease, and want to compare the disease rate between geographic regions. But we do not know exactly the total number of people who are in these geographic regions, and need to estimate this number from other sources.

The following is a data set that exhibits overdispersion due to variability of N_i in a dose–response model. Figure 1 shows the data from a specific experiment undertaken at the University of California, Los Angeles, in which surviving jejunal crypts in mice are counted after a single dose of radiation. A jejunal crypt can be thought of as a compartment that contains stem cells in a certain region of the intestine. These stem cells are ultimately responsible for maintaining the function of the intestine. We are interested in modeling the effect of dose on the response. Each point represents one animal, and we observe the number of jejunal crypts present in a cross-sectional slice following a specific dose of gamma rays. For each animal, we assume that the number of crypts follows a binomial distribution with parameters N_i and P_i, where N_i is the total number of crypts before the dose is given and P_i is a parameter related to the dose level. One problem in this data is that we are not exactly sure of the total number of crypts N_i, although N_i is believed to be approximately 160. The reason that N_i is unobserved is because it is necessary to sacrifice the animal to observe the number of jejunal crypts present.

1.2 Notation and Objective

Assume a column vector of binomial observations $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ and an $n \times p$ matrix $\mathbf{X}$ with the values of p covariates $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ as its $p \times 1$ row vector. The independent binomial observations y_i, $i = 1, 2, \ldots, n$, have parameters N_i and P_i, with P_i fixed at θ_i and N_i unknown and variable. Assume that $E(y_i|N_i) = N_i\theta_i$ and $\text{var}(y_i|N_i) = N_i\theta_i(1 - \theta_i)$, where $\theta_i = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta})$, g is a known link function, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ are unknown parameters associated with the covariates.

To estimate $\boldsymbol{\beta}$ from this overdispersed binomial data, it is necessary to assume something about the distribution of N_i. Our estimation procedures are based on the assumptions that $E(N_i) = m_i$ (m_i: known) and $\text{var}(N_i) = m_i\gamma$ (γ: unknown). The latter assumption is motivated by its common usage for count data, convenience, and the fact that for the special case $\gamma = 1$, the resulting distribution for y_i has the same first two moments as a Poisson. The aim of this article is to apply the transform-both-sides (TBS) method to estimate regression parameter $\boldsymbol{\beta}$ and obtain the properties of the estimates of $\boldsymbol{\beta}$.

In Section 2, we present the basic idea of the TBS method. In Section 3, we derive a family of transformations and describe the maximum likelihood estimation procedure for obtaining estimates of $\boldsymbol{\beta}$. The family of transformations that we call the arcsine transformation family is based on the first-order asymptotic variance stabilizing transformation. Also, we apply the TBS method using known transformations such as the square root and the arcsine transformations, which are special cases of the arcsine transformation family. We present asymptotic results concerning the estimates of

* Dong K. Kim is an Assistant Professor, Department of Mathematics, Statistics and Computer Science, and Program in Surgical Oncology, University of Illinois, Chicago, IL 60607-7045. Jeremy M. G. Taylor is an Associate Professor, Department of Biostatistics, University of California, Los Angeles, CA 90024. This work was partially supported by National Institutes of Health Grants R29-CA45216 and AI29196.

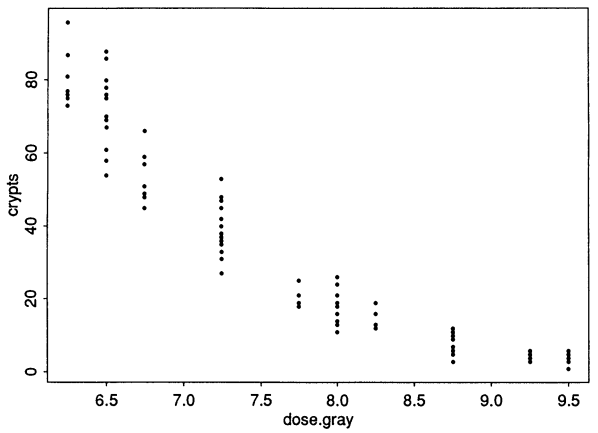

Figure 1. Surviving Jejunal Crypts as a Function of Dose. Each point indicates one animal. The number of crypts present in a cross-sectional slice is observed at the specific dose of gamma rays.

β in Section 4. We use robust procedures (Hernandez and Johnson 1980; Huber 1981; Liang and Zeger 1986) to estimate the asymptotic covariance matrix of the estimates. We describe a simulation study in Section 5 and analyze the jejunum crypts data set in Section 6. Finally, in Section 7 we present some conclusions and discussion.

2. TRANSFORM-BOTH-SIDES METHOD

Transformations for regression models have been used in several ways. We might assume that some transformed form of the response variable satisfies a normal theory linear model. Let $h(\mathbf{y}, \lambda)$ be a family of transformations of $\mathbf{y}$ indexed by λ. Box and Cox (1964) suggested the family of power transformations of the response with the aim of achieving a simple additive or linear model, homoscedastistic error, and normally distributed errors. Box and Cox used the model $h(\mathbf{y}, \lambda) = \mathbf{X}^T\beta + \varepsilon$, where

$$h(\mathbf{y}, \lambda) = \frac{\mathbf{y}^\lambda - 1}{\lambda}, \qquad \lambda \neq 0 \ (=\ln(\mathbf{y}), \lambda = 0).$$

Here ε_i are iid with $E(\varepsilon_i) = 0$ and $\text{var}(\varepsilon_i) = \sigma^2$ and distribution function F.

Carroll and Ruppert (1984) suggested the TBS method when a model $f(\mathbf{X}, \beta)$ has already been chosen empirically or from theoretical consideration to fit $\mathbf{y}$. The TBS method assumes that after applying the same power transformation to $\mathbf{y}$ and $f(\mathbf{X}, \beta)$, the residuals are normally distributed with constant variance; that is, $h(\mathbf{y}, \lambda) = h(f(\mathbf{X}, \beta), \lambda) + \varepsilon$, where ε_i, $i = 1, 2, \ldots, n$ are iid normal with mean 0 and variance σ^2. The main use of the TBS method is to achieve homogeneity of variance on the transformed scale.

3. ARCSINE TRANSFORMATION FAMILY

3.1 Variance Stabilizing Transformation

When N_i is unknown and P_i is fixed at θ_i, the mean function is $E(y_i) = E(E(y_i|N_i)) = E(N_i\theta_i) = E(N_i)\theta_i$ and the variance function is $\text{var}(y_i) = E(\text{var}(y_i|N_i)) + \text{var}(E(y_i|N_i)) = E(N_i)\theta_i(1 - \theta_i) + \text{var}(N_i)\theta_i^2$. When we add the assumptions on N_i that $E(N_i) = m_i$ (known) and $\text{var}(N_i) = m_i\gamma$ (γ: unknown), then the mean and variance functions are

$$E(y_i) = \mu_i = m_i\theta_i \tag{1}$$

and

$$\text{var}(y_i) = V_i = V(\mu_i, \gamma) = m_i\theta_i(1 - \theta_i) + m_i\gamma\theta_i^2 = \mu_i\left(1 + (\gamma - 1)\frac{\mu_i}{m_i}\right). \tag{2}$$

Some interpretations of $V(\mu_i, \gamma)$ for various values of γ are given in Table 1. We have a binomial model without extra variation when $\gamma = 0$; when $\gamma = 1$, the variance is the same as that of a Poisson model without extra variation. When $\gamma > 1$, the observations have the same variance as overdispersed Poisson data (Breslow 1984, 1990) whose distribution is negative binomial. An interesting region is $0 < \gamma < 1$. In this region, we have an overdispersed binomial model, which has different variance from the beta binomial model. In the beta binomial model, N_i is fixed and the first two moments of P_i are determined from the beta distribution (Williams 1982). Our model can also be interpreted as an underdispersed Poisson model in this region. When $\gamma < 0$, the variance in Table 1 represents an underdispersed binomial model. Although the motivation for the form of $V(\mu_i, \gamma)$ arose from the assumption that $\text{var}(N_i) = \gamma m_i$, which permits only nonnegative values of γ, negative values of γ are permitted in the family defined by $V(\mu_i, \gamma)$, provided that $V(\mu_i, \gamma)$ is positive. Brooks, James, and Grey (1991) considered sub-binomial variation compared to binomial variation in the study of sex combinations in litters of pigs. The form of the variance function in Table 1 differs from that of Brooks et al. (1991), who used a beta binomial model assuming a negative variance for P_i.

Notice that the variance function involves the unknown parameter γ, and hence the variance stabilizing transformation family for this overdispersed binomial data will be indexed by the parameter γ. The variance stabilizing transformation seeks a function $h(y_i, \gamma)$ such that $\text{var}\ h(y_i, \gamma) = (h'(\mu_i, \gamma))^2 V(\mu_i, \gamma)$ is a constant, which implies

$$h(y_i, \gamma) = \int_0^{y_i} \frac{1}{\sqrt{V(\mu_i, \gamma)}}\, d\mu_i. \tag{3}$$

Proposition 1. Arcsine Transformation Family for Overdispersed Binomial Data. When y_i, conditional on N_i, is a binomial count and unconditionally has the variance

Table 1. Classification of the Overdispersed Binomial Model

γ	$V(\mu_i, \gamma)$	Model
$\gamma < 0$	$\mu_i(1 + (\gamma - 1)\mu_i/m_i)$	Underdispersed binomial model
$\gamma = 0$	$\mu_i(m_i - \mu_i)/m_i$	Binomial model
$0 < \gamma < 1$	$\mu_i(1 - (1 - \gamma)\mu_i/m_i)$	Overdispersed binomial model Underdispersed Poisson model
$\gamma = 1$	μ_i	Poisson model
$\gamma > 1$	$\mu_i(1 + (\gamma - 1)\mu_i/m_i)$	Overdispersed Poisson model

function shown in Table 1, then h is a first-order asymptotic variance stabilizing transformation where

$$h(y_i, \gamma) = \sqrt{\frac{m_i}{1-\gamma}} \sin^{-1} \sqrt{\frac{(1-\gamma)y_i}{m_i}}, \qquad 1 - \frac{m_i}{y_i} < \gamma < 1, \quad \forall\, i$$

$$h(y_i, \gamma) = \sqrt{\frac{m_i}{\gamma - 1}} \left(\ln\left(\sqrt{\frac{(\gamma-1)y_i}{m_i}} + \sqrt{\frac{(\gamma-1)y_i}{m_i} + 1} \right) \right), \qquad \gamma > 1,$$

and

$$h(y_i, \gamma) = \sqrt{y_i}, \qquad \gamma = 1.$$

Furthermore,

$$\lim_{\gamma \to 1^+} h(y_i, \gamma) = \lim_{\gamma \to 1^-} h(y_i, \gamma) = h(y_i, 1).$$

Proof. It is easy to verify that $h(y_i, \gamma)$ satisfies Equation (3). For the proof of the limit, we use the Taylor series expansion about $\gamma = 1$; that is,

$$h(y_i, \gamma) = \sqrt{y_i} \times \left(1 + \frac{1}{6} \frac{(1-\gamma)y_i}{m_i} + \frac{3}{40} \frac{(1-\gamma)^2 y_i^2}{m_i^2} + \ldots \right).$$

Thus the limiting transformation as γ tends to 1 is the square root.

In this article, because we assume that the first two moments of the response are known, we use the TBS method by applying the same transformation to the response and to the mean function. When we use the TBS method with the arcsine transformation family, we assume that $h(y_i, \lambda) = h(\mu_i, \lambda) + \varepsilon_i$, $i = 1, 2, \ldots, n$, where ε_i are iid with $E(\varepsilon_i) = 0$ and $\text{var}(\varepsilon_i) = \sigma^2$ with distribution F. Furthermore, $\sigma \simeq \frac{1}{2}$ when the observation y_i has the first two moments (1) and (2). This is because $\sigma^2 = \text{var}(\varepsilon_i) = (h'(\mu_i, \gamma))^2 \text{var}(y_i, \gamma) \simeq \frac{1}{4}$.

Anscombe (1948) derived the square root, arcsine, and hyperbolic arcsine transformations as variance stabilizing transformations for Poisson, binomial, and negative binomial distributions. The arcsine transformation family that we derived here is a continuous function of a single scalar γ and includes the aforementioned three transformations as special cases. Although it is not immediately obvious, it can be shown that $h(y_i, \gamma)$ for $\gamma > 1$ can be written as

$$h(y_i, \gamma) = \sqrt{\frac{m_i}{\gamma - 1}} \sinh^{-1} \sqrt{\frac{(\gamma-1)y_i}{m_i}},$$

which has the same form as that of Beall (1942).

3.2 Estimation

The estimation procedure for the parameters β and γ is the same as that of the Box–Cox transformation family. We use maximum likelihood assuming that the transformed data is normal. Let the Jacobian of the transformation $y_i \to h(y_i, \gamma)$ be $J_i(\gamma)$; that is,

$$J_i(\gamma) = \frac{\partial h(y_i, \gamma)}{\partial y_i}.$$

It is simple to show that

$$\frac{\partial h(y_i, \gamma)}{\partial y_i} = \frac{1}{2\sqrt{y_i}} \frac{\sqrt{m_i}}{\sqrt{m_i - (1-\gamma)y_i}}$$

for all values of γ. Under the normality assumption, the log-likelihood of $y_1, y_2, \ldots, y_n$ is given by

$$L(\beta, \gamma, \sigma) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \sum_{i=1}^{n} \frac{(h(y_i, \gamma) - h(\mu_i, \gamma))^2}{2\sigma^2} + \sum_{i=1}^{n} \log J_i(\gamma). \quad (4)$$

For fixed β and γ, the log-likelihood is maximized in σ by

$$(\sigma^2)^* = \frac{\sum_{i=1}^{n} (h(y_i, \gamma) - h(\mu_i, \gamma))^2}{n},$$

and the maximum likelihood estimates of β and γ maximize

$$L(\beta, \gamma) = L(\beta, \gamma, \sigma^*) = -\frac{n}{2} \ln\left(\frac{\sigma^*}{J^0(\gamma)} \right)^2 - \frac{n}{2},$$

where

$$J^0(\gamma) = \left(\prod_{i=1}^{n} J_i(\gamma) \right)^{1/n}.$$

Although there are several possible algorithms for finding the maximum likelihood estimates, including the Newton–Raphson algorithm, we apply the block relaxation algorithm (De Leeuw 1993). The block relaxation algorithm is an iterative scheme to find the maximum over a set of parameters. In this algorithm, we group the parameters in several blocks and maximize the likelihood function over parameters in a specific block, keeping the parameters in other blocks fixed. We iterate this scheme over all blocks until convergence. To find the maximum likelihood estimate for the TBS method using the arcsine transformation family, the block relaxation algorithm has the following iterations:

1. Set $k \leftarrow 1$; start with an initial value, $\gamma^{(k)}$, for the arcsine transformation family parameter.
2. Obtain $\beta^{(k)}$ that maximizes $L(\beta, \gamma^{(k)})$, keeping γ fixed at $\gamma^{(k)}$.
3. Obtain $\gamma^{(k+1)}$ that maximizes $L(\beta^{(k)}, \gamma)$, keeping β fixed at $\beta^{(k)}$.
4. Set $k \leftarrow k + 1$ and go to Step 2. Continue until convergence.

In Step 2, we use the Gauss–Newton nonlinear least squares method as described in Section 3.3. In Step 3, because the log-likelihood function is continuous in γ, this function may be easily maximized by standard optimization techniques. We applied a Newton-Raphson algorithm with a step-halving procedure for Step 3. We found that the block relaxation algorithm, which finds estimates of β and γ sepa-

rately, was more numerically stable than a Newton–Raphson algorithm, which finds the estimates of β and γ jointly.

3.3 Special Cases: Square Root and Arcsine Transformations

When $\gamma = 1$, the variance of $\mathbf{y}$ has the same form as Poisson data; thus we can use the square root transformation to stabilize the variance. After transforming the data, we have

$$\sqrt{y_i} = \sqrt{\mu_i} + \varepsilon_i, \qquad i = 1, 2, \ldots, n,$$

where $\mu_i = m_i\theta_i$, $E(\varepsilon_i) = 0$, and $\operatorname{var}(\varepsilon_i) = \sigma^2$. This is a nonlinear regression problem, so we can use the Gauss–Newton nonlinear least squares algorithm to estimate β.

When $\gamma = 0$, the variance of $\mathbf{y}$ is the same as that of binomial data. Thus the arcsine transformation can be used; that is,

$$\sqrt{m_i}\sin^{-1}\sqrt{\frac{y_i}{m_i}} = \sqrt{m_i}\sin^{-1}\sqrt{\frac{\mu_i}{m_i}} + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where $E(\varepsilon_i) = 0$ and $\operatorname{var}(\varepsilon_i) = \sigma^2$. Similarly, we can use Gauss–Newton nonlinear least squares to estimate β.

The Gauss–Newton algorithm is an iterative scheme that minimizes

$$\sum_{i=1}^{n} (h(y_i) - h(\mu_i))^2.$$

Each iteration is based on the linearization of $h(\mu_i)$ and given by

$$\beta^{(s+1)} = \beta^{(s)} + \left(\sum_{i=1}^{n} \mathbf{D}_i^T (h'(\mu_i))^2 \mathbf{D}_i\right)^{-1} \times \sum_{i=1}^{n} \mathbf{D}_i^T h'(\mu_i)((h(y_i) - h(\mu_i)),$$

where $h'(\mu_i) = \partial h(\mu_i)/\partial\mu_i$ and $\mathbf{D}_i = \partial\mu_i/\partial\beta$.

4. ASYMPTOTIC RESULTS

4.1 Asymptotic Bias and Asymptotic Normality

We derived the arcsine transformation family based on the first-order variance stabilizing transformation and used a normality assumption to estimate parameters (β, γ, σ). Suppose that the observation, y_i, comes from the probability density function $w(\ )$. The exact form of the distribution $w(\ )$ may be unknown; however, the first two moments of y_i are given by (1) and (2). From the log-likelihood (4), we define the score functions for (β, γ, σ) by

$$U_1(\beta, \gamma, \sigma) = \frac{\partial L(\beta, \gamma, \sigma)}{\partial\beta}, \qquad U_2(\beta, \gamma, \sigma) = \frac{\partial L(\beta, \gamma, \sigma)}{\partial\gamma},$$

and

$$U_3(\beta, \gamma, \sigma) = \frac{\partial L(\beta, \gamma, \sigma)}{\partial\sigma}.$$

Let $(\beta^*, \gamma^*, \sigma^*)$ be the estimate obtained by setting these score functions equal to 0 and let $(\beta^{00}, \gamma^{00}, \sigma^{00})$ be a solution of $E_w(U_i(\beta, \gamma, \sigma)) = 0$, $i = 1, 2, 3$. Under certain regularity conditions, we have $(\beta^*, \gamma^*, \sigma^*) \to (\beta^{00}, \gamma^{00}, \sigma^{00})$ almost surely (Hernandez and Johnson 1980). Also, let (β^0, γ^0) be the true value of the parameters under the assumption that the observations y_i come from the underlying distribution $w(\ )$ with the first two moments given by (1) and (2). How close $(\beta^{00}, \gamma^{00})$ is to (β^0, γ^0) depends on how close the transformed distribution is to a normal distribution with homogeneous variance.

Because the data after transformation by a member of the arcsine transformation family are not exactly normal with constant variance, we expect β^* to be an inconsistent estimate of the true value β^0. But we can show that the magnitude of the bias is small under certain additional assumptions. Heuristic arguments given in the Appendix show that the limiting term of the large-sample bias of β^* can be expressed as

$$|E_w(\beta^*) - \beta^0| \le O(\mu_H^{-1}),$$

where μ_H is a weighted harmonic average of μ_i^0.

This result says that when the μ_i^0's are large enough, so that $1/\mu_H$ is small, the bias of β^* is small and thus can essentially be ignored. This result is intuitively sensible, because if the μ_i's are all small, then it is unreasonable to assume that transformed binomial responses are normal, whereas if the μ_i's are large, then it is reasonable to assume that we can transform the discrete y_i's to achieve approximate normality.

When the normality assumption is satisfied, the Fisher information gives the asymptotic covariance matrix of the estimates. In the TBS method, it is assumed that the same value of γ achieves both homogeneity of variance and normality of the transformed data; in real applications, it is unlikely that both requirements can be achieved. So to account for the failure of these assumptions, we use robust procedures (Hernandez and Johnson 1980; Huber 1981; Liang and Zeger 1986) to obtain the covariance matrix of the estimates.

Theorem 2. As $n \to \infty$, we have the following.

a. The asymptotic covariance matrix of $n^{1/2}(\beta^* - \beta^{00})$ is

$$\lim_{n\to\infty} n4\left(\sum_{i=1}^{n} (\mathbf{D}_i^{00})^T (V_i^{00})^{-1} \mathbf{D}_i^{00}\right)^{-1} \times \left(\sum_{i=1}^{n} (\mathbf{D}_i^{00})^T (V_i^{00})^{-1} M_{AF}^2 \mathbf{D}_i^{00}\right) \times \left(\sum_{i=1}^{n} (\mathbf{D}_i^{00})^T (V_i^{00})^{-1} \mathbf{D}_i^{00}\right)^{-1}, \tag{5}$$

where $\mathbf{D}_i^{00} = \partial\mu_i/\beta^{00}$, $V_i^{00} = V(\mu_i^{00}, \gamma^{00})$, and $M_{AF}^2 = E_w(h_{AF}(y_i, \gamma^{00}) - h_{AF}(\mu_i^{00}, \gamma^{00}))^2$. Here $h_{AF}(\ )$ is the arcsine transformation family. The dependence of M_{AF}^2 on i is deleted for convenience of notation.

b. When the normality assumption is satisfied, β^* and γ^* in the arcsine transformation family are asymptotically uncorrelated.

Proof. See the Appendix.

We can evaluate these expressions at the estimates to estimate the covariance matrix and hence construct confidence

intervals or standard errors for β^{00}. Thus we obtain approximate inference for β^0. The covariance matrix in Theorem 2 is robust, because the first and last terms of the expression come from the choice of the transformation method and the second term comes from the data. So the second term protects against choosing an inappropriate transformation that fails to satisfy the normality assumption. When the transformed data are normal with constant variance, $M_{AF}^2 = \sigma^2$ and the asymptotic covariance matrix of $n^{1/2}(\beta^* - \beta^{00})$ reduces to

$$\lim_{n\to\infty} n4\sigma^2\left(\sum_{i=1}^{n} (\mathbf{D}_i^{00})^T (V_i^{00})^{-1}\mathbf{D}_i^{00}\right)^{-1}, \qquad (6)$$

which is the same as the inverse of the Fisher information matrix from the normal density.

From the asymptotic result, we can estimate the covariance matrix of β^* as if we know the value of γ. This result agrees with the approximate result of the TBS method using the Box–Cox transformation family; the limiting distribution of β^* is the same whether or not the transformation is known when the sample size is large and $\sigma \to 0$ (Carroll and Ruppert 1984).

To test a hypothesis for H_0: $\gamma = \gamma^0$, we apply the likelihood ratio test. Within iterations of the block relaxation algorithm, we can easily extract the likelihood ratio test statistics. Let $(\beta^*, \gamma^*, \sigma^*)$ be the maximum likelihood estimate under the full model and let $(\beta^*(\gamma^0), \gamma^0, \sigma^*(\gamma^0))$ be the maximum likelihood estimate under the null hypothesis. The likelihood ratio test for the null hypothesis H_0 is defined by $r = -2(L(\beta^*(\gamma^0), \gamma^0, \sigma^*(\gamma^0)) - L(\beta^*, \gamma^*, \sigma^*))$, which has an approximate asymptotic χ^2 distribution with 1 degree of freedom. The profile likelihood confidence interval for γ can be easily obtained directly by inverting the likelihood ratio test. The $100(1-\alpha)\%$ likelihood-based confidence interval for γ is

$$\left(\gamma^0 : L(\beta^*(\gamma^0), \gamma^0, \sigma^*(\gamma^0)) - L(\beta^*, \gamma^*, \sigma^*) > -\frac{1}{2}\chi^2_{\alpha,1}\right).$$

Instead of estimating γ from the data, one might use a known transformation, such as the square root or the arcsine transformation, to estimate β. Let β^{**} denote the estimate from the Gauss–Newton nonlinear least squares algorithm. It can be shown, using arguments similar to those in the Appendix, that as $n \to \infty$,

$$|E_w(\beta^{**}) - \beta^0| \le O(\mu_H^{-1}) + O\left(\frac{|\gamma^0 - \gamma^+|}{\mu_H}\right),$$

where γ^+ is the assumed value of γ; that is, $\gamma^+ = 0$ for the arcsine transformation and $\gamma^+ = 1$ for the square root transformation.

This result shows that when μ_H is large and the assumed value, γ^+, is close to the true value, γ^0, the bias tends to be small; however, when γ^0 is far from γ^+, the bias will increase.

Similar to the result in the Theorem 2, we can robustly estimate the covariance matrix in cases of the square root and arcsine transformations.

Corollary 3. Let β^{000} be a limit of β^{**}. The asymptotic covariance matrix of $n^{1/2}(\beta^{**} - \beta^{000})$ is

$$\lim_{n\to\infty} n4\left(\sum_{i=1}^{n} (\mathbf{D}_i^{000})^T V_i^{000(+)^{-1}} \mathbf{D}_i^{000}\right)^{-1} \times \left(\sum_{i=1}^{n} (\mathbf{D}_i^{000})^T V_i^{000(+)^{-1}} M_{(+)}^2 \mathbf{D}_i^{000}\right) \times \left(\sum_{i=1}^{n} (\mathbf{D}_i^{000})^T V_i^{000(+)^{-1}} \mathbf{D}_i^{000}\right)^{-1},$$

where $V_i^{000(+)} = \mu_i^{000}$ and $M_{(+)}^2 = M_{ST}^2 = E_w(h_{ST}(y_i) - h_{ST}(\mu_i^{000}))^2$ for the square root transformation and $V_i^{000(+)} = \mu_i^{000}(m_i - \mu_i^{000})/m_i$ and $M_{(+)}^2 = M_{AT}^2 = E_w(h_{AT}(y_i) - h_{AT}(\mu_i^{000}))^2$ for the arcsine transformation. Here $h_{ST}(\ )$ and $h_{AT}(\ )$ are the square root and the arcsine transformations and $\mu_i^{000} = m_i g^{-1}(x_i^T \beta^{000})$. Also, the dependence of M_{ST}^2 and M_{AT}^2 on i is suppressed for ease of notation.

Proof. See the Appendix.

The covariance matrix from the foregoing results are robust, because the second term is evaluated from the data. When $M_{ST}^2 = \sigma^2$ and $M_{AT}^2 = \sigma^2$ in each transformation, the covariance matrix reduces to the standard asymptotic form obtained from a Gauss–Newton nonlinear least squares approach (Jennrich 1969). Also, the first term is obtained from the assumed variance function that each transformation is attempting to make homogeneous.

4.2 Numerical Evaluation of the Asymptotic Relative Efficiency

In this section we evaluate the asymptotic relative efficiency of the arcsine transformation family to the square root and the arcsine transformations in a specific dose–response model. In the presence of overdispersion, there is a possible loss of efficiency in using a model that does not allow overdispersion, although the efficiency loss may be not so serious in moderate overdispersion (Cox 1983). In this section we quantify the loss of efficiency of the square root and arcsine transformations compared to the arcsine transformation family when the binomial data are overdispersed.

Table 2 shows the true model used in the study. We mimic a dose–response model in an animal experiment. We use a logit link function and one covariate in the linear predictor. We assume that the mean of N_i is 160 and consider a range of values for γ ($-.5 \le \gamma \le 3.0$).

To evaluate the asymptotic relative efficiency, we approximate M_{AF}^2, M_{ST}^2, and M_{AT}^2 as follows:

Table 2. True Model

Model:	$\log(\theta/(1-\theta)) = A_0 + A_1 d$
	d = dose level
	θ = probability of response with d chosen such that
	Design 1: $\theta = \{.05, .20, .50, .80, .95\}$, $0 \le \gamma \le 3$
	Design 2: $\theta = \{.05, .25, .45, .65, .75\}$, $-.3 \le \gamma \le 3$
	Design 3: $\theta = \{.05, .15, .25, .45, .65\}$, $-.5 \le \gamma \le 3$
	A_0 and A_1 = parameters ($A_0 = 4.0$ and $A_1 = -1.0$)
	The distribution of y_i has two moments μ_i and V_i defined by (1) and (2)

$$M_{AF}^2 \simeq \left(\frac{\partial h_{AF}(y_i;\gamma)}{\partial y_i}\bigg|_{\mu_i}\right)^2 E(y_i-\mu_i)^2 = \left(\frac{1}{2\sqrt{V_i}}\right)^2 V_i = \frac{1}{4},$$

$$M_{ST}^2 \simeq \left(\frac{\partial h_{ST}(y_i)}{\partial y_i}\bigg|_{\mu_i}\right)^2 E(y_i-\mu_i)^2 = \left(\frac{1}{2\sqrt{\mu_i}}\right)^2 V_i = \frac{1}{4}\frac{V_i}{\mu_i},$$

and

$$M_{AT}^2 \simeq \left(\frac{\partial h_{AT}(y_i)}{\partial y_i}\bigg|_{\mu_i}\right)^2 E(y_i-\mu_i)^2$$

$$= \frac{1}{4}\left(\frac{\mu_i(m_i-\mu_i)}{m_i}\right)^{-1} V_i.$$

Figure 2 shows the asymptotic relative efficiency (ARE) of the arcsine transformation family to the square root and arcsine transformations. A.1 and B.1 are the ARE's for the intercept (A_0), and A.2 and B.2 are the ARE's for the slope (A_1). In A.1 and A.2, we can see that the ARE is 1 when $\gamma = 1$. This is because the arcsine transformation family corresponds to the square root transformation when $\gamma = 1$ and there is no inflation of the variance associated with estimating γ (Theorem 2). When γ is greater than or less than 1, we observe a loss of efficiency if the square root transformation is used. Especially when $\gamma < 0$, which means that when the data come from the underdispersed binomial model, the loss of efficiency from the square root transformation is quite large. Also, when $\gamma = 0$ for which the arcsine transformation family reduces to the arcsine transformation, the ARE is 1 (B.1 and B.2). When $\gamma > 0$ or $\gamma < 0$, we observe a loss of efficiency if the arcsine transformation is used instead of the arcsine transformation family. Design 1, Design 2, and Design 3 indicate three different sets of doses with corresponding different sets of response probabilities, as shown in Table 2. From these three designs, we can see that the more widely the probability of response is distributed, the greater the efficiency gain that can be achieved by using the arcsine transformation family.

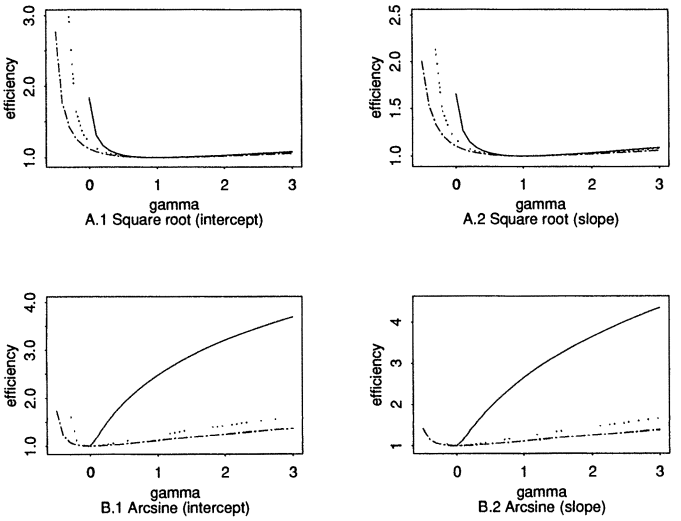

Figure 2. Asymptotic Relative Efficiencies of the Arcsine Transformation Family. Each line indicates the theoretical evaluation of the ARE in Design 1 (——), Design 2 (· · · ·), and Design 3 (– ·–). The ARE to the square root is 1 when $\gamma = 1$, and the ARE to the arcsine is 1 when $\gamma = 0$.

5. SIMULATION STUDY

A Monte Carlo simulation was performed to evaluate the bias, variability, and coverage properties of the parameter estimates. In this study we mimic the dose–response model in the animal study. The simulation study was undertaken on an IBM PC 286 using the GAUSS programming language. The data generating scheme for the responses y_i with overdispersion due to variability of N_i is as follows:

Scheme 1 ($\gamma \geq 0$)

1. Generate N_i from $N(m, m\gamma)$ and round to the nearest integer, where $m = 160$ and $\gamma = \{.1, .3, .5, .7, 0, 1.0, 3.0, 5.0, 7.0, 9.0\}$.

2. Given N_i, generate y_i from a binomial distribution with parameters N_i and θ_i, where θ_i is determined by a true model corresponding to the response probabilities in Design 1 ($0 \leq \gamma \leq 1$) and Design 2 ($\gamma \geq 1$).

Scheme 2 ($\gamma \leq 0$)

1. Calculate $\mu_i = m\theta_i$ and $V_i = V(\mu_i, \theta)$, where $m = 160$, $\gamma = \{-.5, -.4, -.3, -.2, -.1, 0\}$ and θ_i is from Design 3.

2. Generate y_i from $N(\mu_i, V_i)$ and round to the nearest integer.

We compare the arcsine transformation family to the square root and arcsine transformations. We generate 50 binomial responses at each dose level corresponding to θ, where θ is shown in Table 2, so we have 250 binomial responses in each data set. For each given value of γ, we generate 500 data sets to compare the transformation methods. We use the same linear logistic models as described in Table 2.

Figure 3 shows the estimated bias and variance of the transformation methods when $0 \leq \gamma \leq 1$ (Design 1). A.1 and A.2 show the biases for the intercept and the slope. We can see that the biases are not serious for the arcsine transformation family. The arcsine transformation gives systematically larger bias than the square root transformation and the arcsine transformation family, except when $\gamma = 0$. We observe that the bias from the square root transformation is close to the bias from the arcsine transformation family when $0 \leq \gamma \leq 1$. B.1 and B.2 show the variance of the estimate from the transformation methods for the intercept and the slope. We have divided the variances of each method by that of the arcsine transformation family. The arcsine transformation gives systematically larger variance than the other transformations, except when $\gamma = 0$. Also, the square root transformation gives larger variance than the arcsine transformation family when γ is far from 1 and relative variance is close to 1 when γ is close to 1.

Figure 4 shows the bias and variance of the estimates when $\gamma \geq 1$ for Design 2. When $\gamma \geq 1$, the arcsine transformation has systematically larger variance than the arcsine transformation family. The loss of efficiency of the square root transformation is slightly increased when γ is far from 1.

Figure 5 shows the bias and the variance of the estimates when the data generating model is an underdispersed binomial model. We can see that the loss of efficiency from

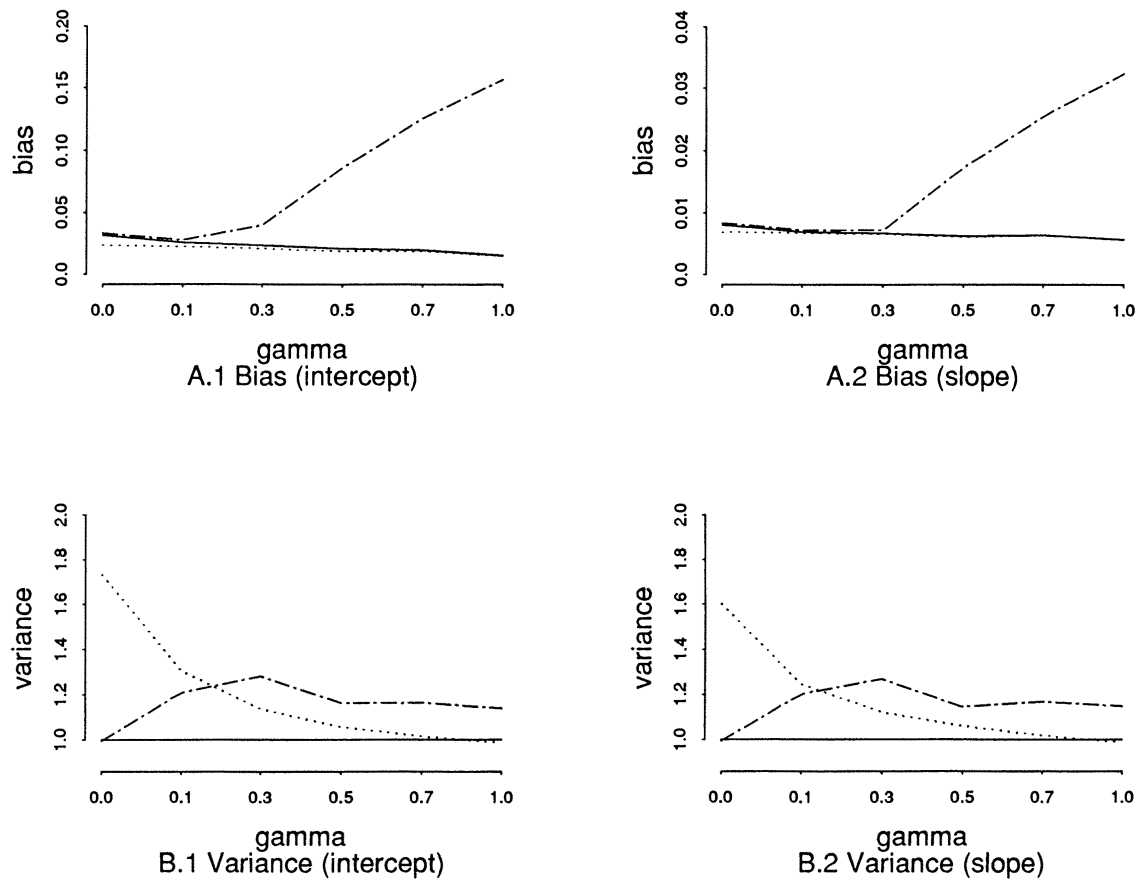

Figure 3. Estimated Biases and Variances of the Transformation Methods (Design 1). Bias is calculated by the absolute value of the difference between the average of the estimates and the true value. Variance indicates the ratio of Monte Carlo variance from each transformation divided by the Monte Carlo variance from the arcsine transformation family. When $\gamma \simeq 0$, the arcsine transformation (- · -) behaves similarly to the arcsine transformation family (——), and when $\gamma \simeq 1$, the square root (· · · ·) performs similarly to the arcsine transformation family.

the square root transformation is largest in this region. The arcsine transformation has larger variance than the arcsine transformation family when γ is far from 0.

In these simulation experiments, we observed that the arcsine transformation performed well near $\gamma = 0$ and the square root transformation performed well near $\gamma = 1$. These results indicate that there is no cost associated with estimating γ as predicted in Theorem 2, because γ^* and β^* are asymptotically uncorrelated when the normality assumption is satisfied. Moreover, we can see that the Monte Carlo asymptotic relative efficiencies in Design 1, Design 2, and Design 3 (Figs. 3, 4, and 5) are quite close to the theoretical asymptotic relative efficiencies (Fig. 2).

Figure 6 shows the coverage rates of the three transformation methods. Coverage rates are averaged over the intercept and the slope, because the two results are almost identical. The confidence intervals are obtained using the robust asymptotic variances as described in Equation (5). In the asymptotic covariance matrix, we estimate M_{AF}^2, M_{ST}^2, and M_{AT}^2 from the data without assuming that each transformation method produces homogeneous errors. So we use the estimate

$$(M_{AF}^2)^* = (h_{AF}(y_i, \gamma^*) - h_{AF}(\mu_i^*, \gamma^*))^2 \quad \text{for each } i.$$

Estimates of M_{ST}^2 and M_{AT}^2 are obtained in a similar way.

From the simulation, we observe that the arcsine transformation family gives quite good coverage rates, with a range of 86.0–96.0% coverage at the 95% nominal rate. The coverage rates from the square root transformation method are also very close to the nominal rate. On the other hand, the arcsine transformation has poor coverage rates when γ is far from 0. We observe a 56.1% coverage at the 95% nominal rate when $\gamma = 1.0$ in Design 1. One of the reason for the poor coverage rate of the arcsine transformation is that when γ is far from 0, the bias of the arcsine transformation is quite large in Figures 3, 4, and 5, so even the "robust" method will not give correct coverage properties.

We compared the coverage rates of the robust confidence intervals to those of the nonrobust confidence intervals obtained from the standard Fisher information matrix (6). We found very little difference between two coverage rates for three designs considered, suggesting that using the robust variance estimate is frequently not necessary.

Table 3 shows the average and standard deviations of the estimates of the parameter γ from the arcsine transformation family. We can see that the estimates are close to the true value of γ, but that the biases and the standard deviations increase as γ increases. This indicates that we can get more precise estimates of γ when $\gamma \leq 1$, especially for underdispersed binomial data ($\gamma < 0$). For overdispersed Poisson

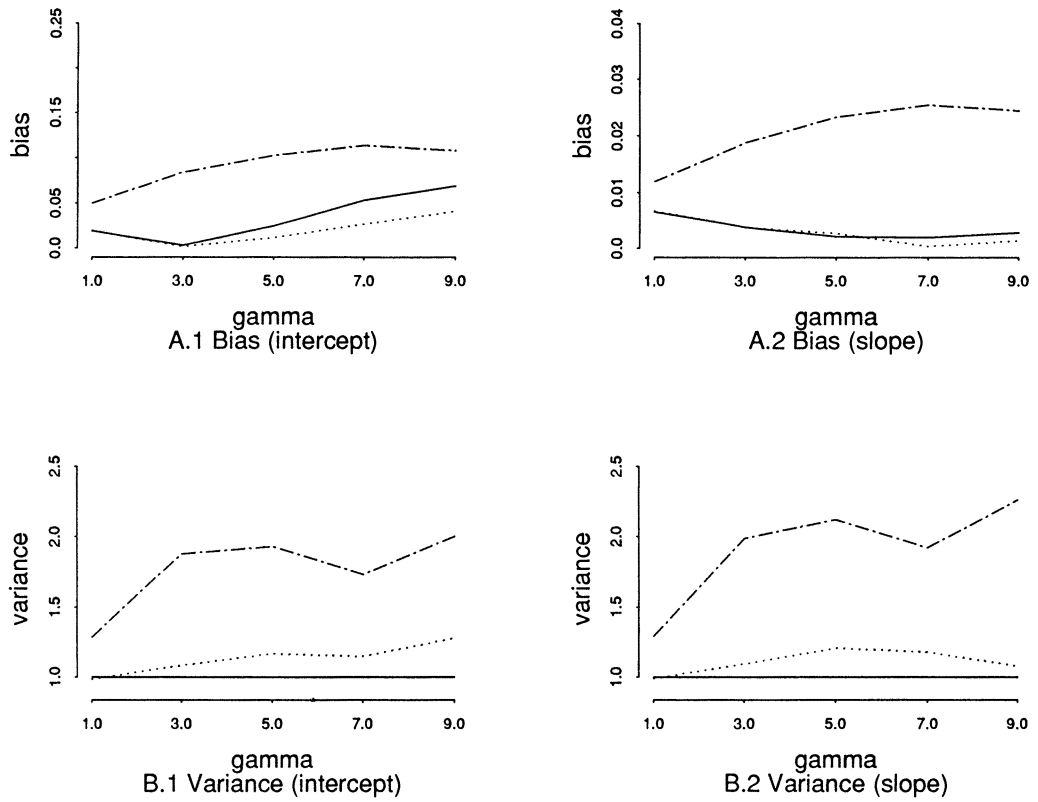

Figure 4. Estimated Biases and Variances of the Transformation Methods (Design 2). When $\gamma > 1$, efficiency losses relative to the square root (· · · ·) and arcsine transformations (– · –) are observed.

data ($\gamma > 1$), the estimates of γ have large standard deviations, so the estimates may be less precise. Figure 6B shows the coverage rates of the parameter γ from the arcsine transformation family for Design 1, calculated by the likelihood-based approach. We observe rates of 91.6–94.8% at the 95% nominal level. For the other two designs, the coverage rates were similar, except when γ was very large or much less than 0.

6. APPLICATION TO RADIATION BIOLOGY DATA

Surviving jejunal crypts in mice following a specific dose of gamma rays are studied at the department of Radiation Oncology, University of California, Los Angeles. We have 126 observations in this data set, as shown in Figure 1.

As a first approximation, it is reasonable to believe that the response (i.e., the number of crypts) is binomial with parameters N_i and θ_i, but N_i is unknown. The overdispersion problem in these data is due to variability of the unobserved total counts N_i, as described in Section 1.1.

We apply the TBS method using the three different approaches we have described. Because we have prior knowledge that the total count before the dose is given, N_i, is distributed around 160, we use this knowledge to analyze the data. Thus we use $E(N_i) = 160$ for the TBS method.

Table 4 shows the estimates of the regression parameter β and the additional parameter γ. The first line is a result from a standard logistic regression analysis in which N_i is assumed to be known and fixed at 160, and the following lines are results from the TBS method. We can see that the point estimates of the regression parameter are quite similar to those from the TBS method, but the standard errors of the logistic regression model are slightly smaller than those given by the TBS method.

The differences between the standard errors are not large, which indicates that the overdispersion problem is not serious in this data set. The estimate of the additional parameter, γ, which allows for overdispersion in the arcsine transformation family, is 2.41.

Figure 7 shows the residual plots from the TBS method. A.1, A.2, and A.3 show the residual plots versus the fitted value from the arcsine transformation, the square root transformation, and the arcsine transformation family. We can see that the arcsine transformation does not produce homogeneous errors (A.1), whereas the square root transformation (A.2) and the arcsine transformation family (A.3) appear quite homogeneous. We calculated the Spearman rank correlation between the absolute residuals and the fitted values to access the homogeneity of variance of the residuals (Carroll and Ruppert 1988, p. 147). The correlation from the arcsine transformation is .179 (p = .045), whereas the correlation from the square root transformation (.062, p = .487) and the arcsine transformation family (−.015, p = .868) are smaller. B.1, B.2, and B.3 of Figure 7 show the Q-Q plots of the residual from the transformation methods. We can see that the Q-Q plot of the arcsine transformation is quite curved and that of the square root transformation is nearly linear. The Q-Q plot of the arcsine transformation family is very nearly linear, indicating that the residuals are

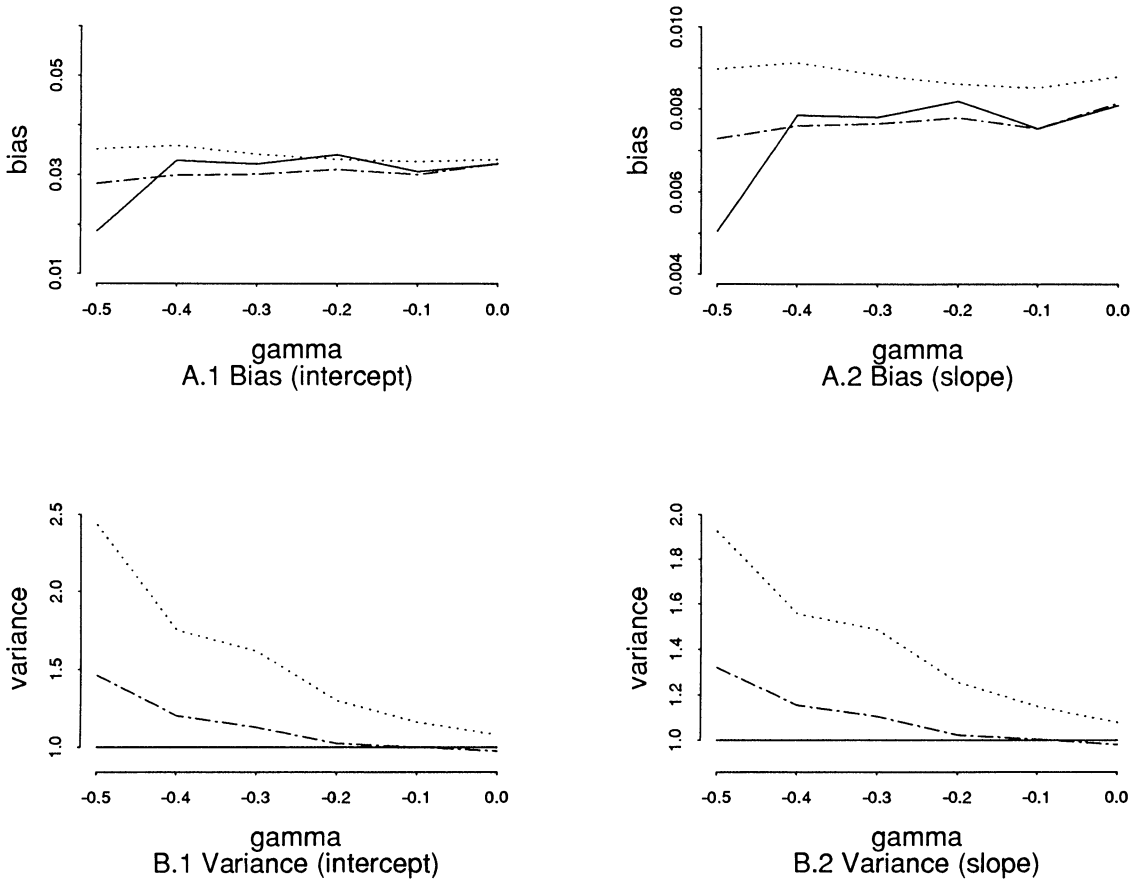

Figure 5. Estimated Biases and Variances of the Transformation Methods (Design 3). When $\gamma < 0$, efficiency losses relative to the square root (· · · ·) are large. Also, the arcsine (– · –) loses efficiency except when γ is close to 0.

approximately normally distributed. The residual plot and the Q-Q plot of the square root transformation are close to those of the arcsine transformation family because $\gamma^* = 2.41$ and the confidence interval, (.56, 6.73), includes the square root transformation ($\gamma = 1$). We can see that the profile confidence interval for γ is asymmetric around the point estimate.

7. CONCLUSION

In this article we focus on the overdispersion problem in binomial data due to variability of N_i, where N_i is an unobserved random variable. When overdispersion exists in binomial data, the parameter estimates from standard methods of analysis that ignore overdispersion will tend not to be seriously biased, but the standard errors will generally be too small.

We apply the TBS method to solve this overdispersion problem. We develop the arcsine transformation family, indexed by a single parameter, which contains the square root and the arcsine transformation as special cases. The arcsine transformation family is asymptotically equivalent to the fixed arcsine and the square root transformations when γ is equal to 0 or to 1, and is more efficient than those fixed transformations in other cases. Moreover, the regression parameter β and additional parameter γ are asymptotically uncorrelated for the TBS method using the arcsine transformation family when the transformed data are assumed to follow a normal distribution. The simulation study and the application from radiobiology support these results.

For data with considerable overdispersion ($\gamma > 1$), we found that the square root transformation performs almost as well as the arcsine transformation family. So in this setting, the square root TBS approach is a simple and attractive method. But for approximately binomial data ($\gamma \simeq 0$), using the square root transformation is not as efficient as using the arcsine transformation family. The arcsine transformation performs substantially worse than the arcsine transformation family, unless γ is very close to 0. We do not recommend using the arcsine transformation if overdispersion is suspected.

An attractive feature of the TBS method is that it is computationally fairly simple. When γ is known, we can use Gauss–Newton nonlinear least squares to estimate β. When γ is unknown, we need an optimization routine to estimate the parameters β and γ, which maximize the likelihood. For a numerical technique, we applied the block relaxation algorithm. This algorithm separates β and γ and gives us a computationally stable method for obtaining the maximum likelihood estimates. Another numerical technique is an application of the Gauss–Newton algorithm by adapting "pseudo-observations" (Carroll and Ruppert 1988).

In this article we have focused on applying the arcsine transformation family in the setting where we have binomial data but N_i is unobserved. But the arcsine transformation

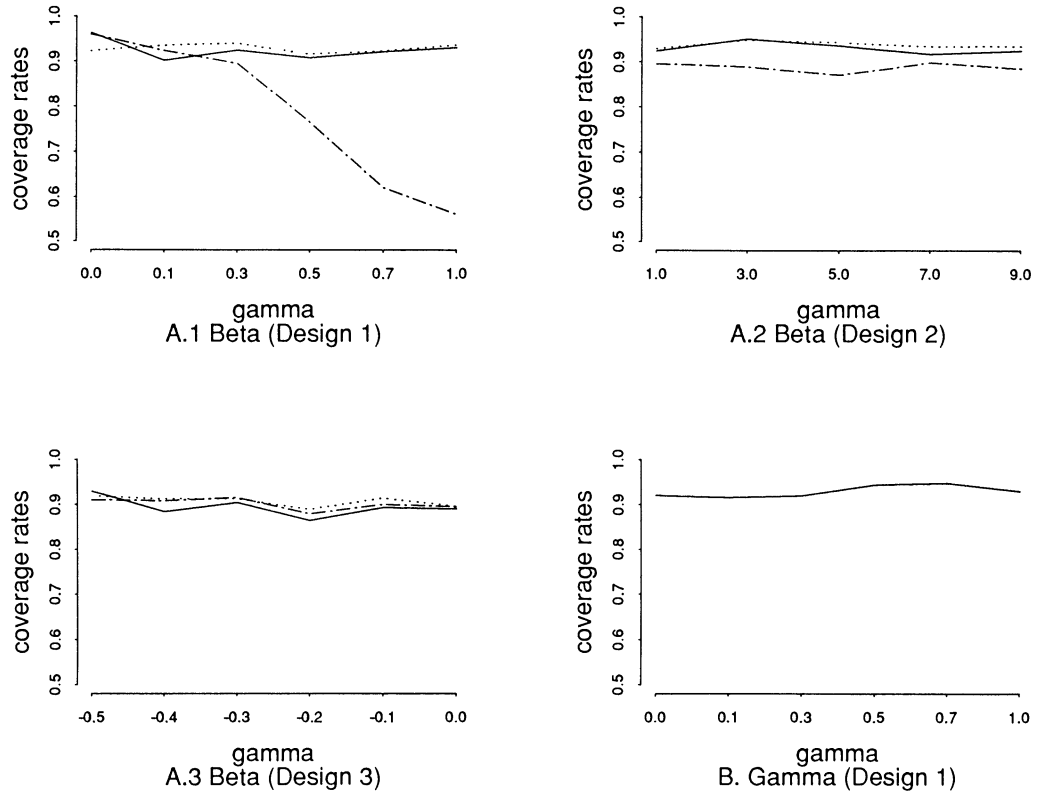

Figure 6. Coverage Rates of Beta and Gamma From the Transformation Methods; arcsine family (——), square root (· · · ·), arcsine (- · -). Coverage rates of beta are based on the robust method. Coverage rates are close to the nominal rate, except for the arcsine transformation in Design 1. Coverage rates of gamma are based on the likelihood-based approach.

family may have broader applicability to the situation where we observe overdispersed binomial data with known N_i or overdispersed or underdispersed Poisson count data.

In the application and simulations, m_i was set at 160. One might expect the performance of the procedure to break down if m_i is small (<10) or if $E(y_i)$ $(=m_i\theta_i)$ is very small (<5) for a substantial number of the observations, because in these situations it may not be reasonable to approximate a discrete distribution by a normal.

Other approaches to analyzing data with overdispersion due to uncertainty in N_i are possible. In other work (Kim 1991), we have considered two alternative methods. The first is a quadratic estimating equations/quasi-likelihood approach based on the first two moments of y_i. The second is a finite mixture model approach, where the distribution of N_i has known range and is specified either parametrically, nonparametrically, or semiparametrically. A comparison of these methods is beyond the scope of this article.

APPENDIX: PROOFS

Sketch Proof of the Result that $|E(\beta^*) - \beta^0| \le O(\mu_H^{-1})$

We prove this when β is univariate. Let $\beta^*(\gamma^*)$ be the estimate of β when γ is estimated and let $\beta^*(\gamma^{00})$ be the estimate of β when $\gamma = \gamma^{00}$, where $\gamma^* \to \gamma^{00}$ in probability. Then we have

$$E_w(\beta^* - \beta^0) = E_w(\beta^*(\gamma^*) - \beta^0)$$
$$= E_w(\beta^*(\gamma^*) - \beta^*(\gamma^{00})) + E_w(\beta^*(\gamma^{00}) - \beta^0).$$

Because $\gamma^* \to \gamma^{00}$ and $\beta^*(\gamma)$ is a continuous function of γ, the first term converges to 0. Let $L(\beta, \gamma^{00})$ be the likelihood equation with fixed $\gamma = \gamma^{00}$. Then the score and the information functions for $\beta^*(\gamma^{00})$ are defined by

Table 3. Average and Standard Deviations of Estimates of γ From the Arcsine Transformation Family

Design 1			Design 2			Design 3		
True γ	Average	S.D.	True γ	Average	S.D.	True γ	Average	S.D.
.0	.002	.011	1.0	0.953	0.367	.0	−.037	.131
.1	.107	.038	3.0	2.888	1.157	−.1	−.126	.123
.3	.303	.085	5.0	4.744	2.099	−.2	−.209	.087
.5	.487	.127	7.0	6.298	2.999	−.3	−.297	.065
.7	.693	.198	9.0	7.596	3.928	−.4	−.392	.038
1.0	.957	.251				−.5	−.494	.014

NOTE: Number of simulations: 500.

Table 4. Estimates of the Regression Parameter β and the Additional Parameter γ

Method	β		γ (95% C.I.)
	Constant (s.e.)	Slope (s.e.)	
Logistic[a]	7.4315 (.1748)	−1.1853 (.0241)	—
Transform-both-sides			
Arcsine	7.4979 (.2010)	−1.1956 (.0267)	—
Square-root	7.4469 (.1953)	−1.1893 (.0259)	—
Arcsine family	7.4144 (.1941)	−1.1853 (.0257)	2.41 (.56, 6.73)[b]

[a] Logistic regression with N_i is fixed at $m = 160$.
[b] Profile likelihood confidence interval.

$$U_n(\beta(\gamma^{00}), \gamma^{00}) = \frac{\partial L(\beta, \gamma^{00})}{\partial \beta} = \sum_{i=1}^{n} \left(\frac{h(y_i, \gamma^{00}) - h(\mu_i, \gamma^{00})}{\sigma^2}\right)\left(\frac{\partial h(\mu_i, \gamma^{00})}{\partial \mu_i}\right)\left(\frac{\partial \mu_i}{\partial \beta}\right)$$

$$\text{and} \quad I_n(\beta(\gamma^{00})) = -\frac{\partial^2 L(\beta, \gamma^{00})}{\partial \beta^2}.$$

We assume that $U_n(\beta(\gamma^{00}), \gamma^{00})$ is continuous and that

$$i(\beta, \beta^0) = E_w(I_n(\beta(\gamma^{00}))) \to \Gamma(\beta) > 0.$$

We also assume that under certain regularity conditions, we have in the open neighborhood of β,

$$\frac{I_n(\beta(\gamma^{00}))}{n} \to \Gamma(\beta), \quad \text{ip.} \tag{A.1}$$

Here, "ip" means to converge in probability. Under the assumption of continuity of $U_n(\beta(\gamma^{00}), \gamma^{00})$, we can obtain β^+ in the interval $(\beta^*(\gamma^{00}), \beta^0)$ such that

$$U_n(\beta^*(\gamma^{00}), \gamma^{00}) = U_n(\beta^0, \gamma^{00}) - I_n(\beta^+)(\beta^*(\gamma^{00}) - \beta^0).$$

Because $\beta^*(\gamma^{00})$ is a solution of $U_n(\beta(\gamma^{00}), \gamma^{00}) = 0$, we have

$$\beta^*(\gamma^{00}) - \beta^0 = \left(\frac{I_n(\beta^+)}{n}\right)^{-1}\left(\frac{U_n(\beta^0, \gamma^{00})}{n}\right).$$

From (A.1), when n is large enough, we can obtain β^+ that satisfies

$$\frac{I_n(\beta^+)}{n} > \frac{\Gamma(\beta^+)}{4}.$$

Thus we have

$$|E_w(\beta^*(\gamma^{00})) - \beta^0| = \left|E_w\left(\left(\frac{I_n(\beta^+)}{n}\right)^{-1}\left(\frac{U_n(\beta^0, \gamma^{00})}{n}\right)\right)\right| < \frac{4}{\Gamma(\beta^+)}\left|E_w\left(\frac{U_n(\beta^0, \gamma^{00})}{n}\right)\right|.$$

Let μ_i^0 be a evaluation of μ_i at β^0. Then we have

$$\frac{U_n(\beta^0, \gamma^{00})}{n} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{h(y_i, \gamma^{00}) - h(\mu_i^0, \gamma^{00})}{\sigma^2}\right)\left(\frac{\partial h(\mu_i, \gamma^{00})}{\partial \mu_i^0}\right)\left(\frac{\partial \mu_i}{\partial \beta^0}\right).$$

From the Taylor series expansion of $h(y_i, \gamma^{00})$ at μ_i^0, we have

$$h(y_i, \gamma^{00}) = h(\mu_i^0, \gamma^{00}) + h'(\mu_i^0, \gamma^{00})(y_i - \mu_i^0) + \frac{h''(\mu_i^0, \gamma^{00})}{2}(y_i - \mu_i^0)^2 + (\text{higher-order terms}).$$

Because the first two moments of y_i are defined, we have

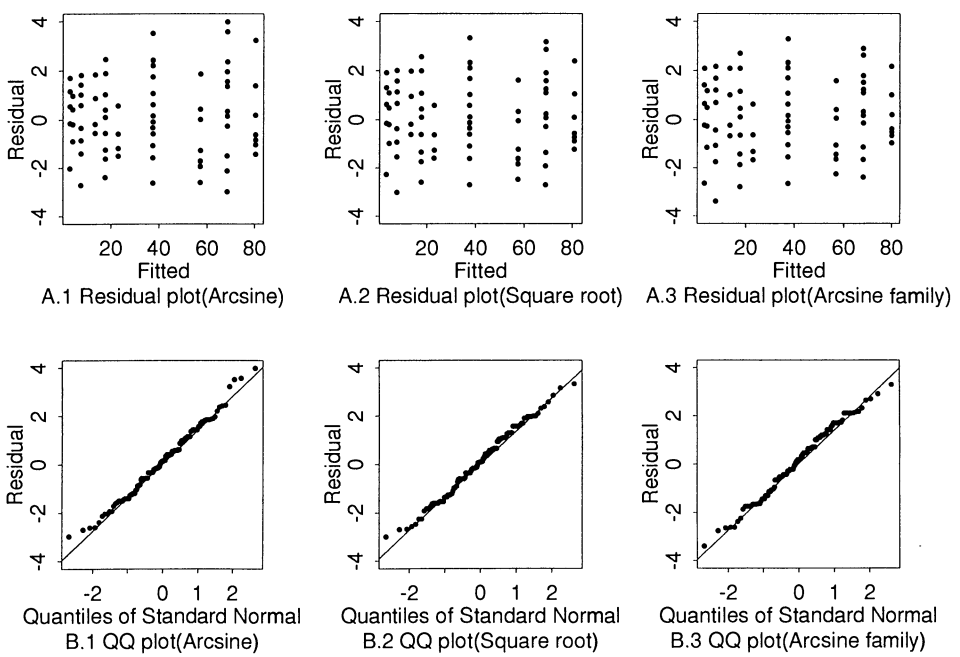

Figure 7. Residual Plots From the Transformation Methods. Standardized residuals are shown (Carroll and Ruppert 1988, p. 147).

$$E_w(y_i - \mu_i^0) = 0 \quad \text{and} \quad E_w(y_i - \mu_i^0)^2 = V(\mu_i^0, \gamma^0).$$

Also, we can verify that

$$h'(\mu_i, \gamma) = \frac{1}{2\sqrt{V_i}}$$

and

$$h''(\mu_i, \gamma) = -\frac{1}{4V_i\sqrt{V_i}}\left(1 + 2(\gamma - 1)\frac{\mu_i}{m_i}\right).$$

Thus, when we assume that the absolute value of $\partial\mu_i/\partial\beta$ is bounded, the first-order term of $E_w(U_n(\beta^0, \gamma^{00})/n)$ disappears and the second-order term gives us

$$\left|\frac{1}{\sigma^2}\frac{1}{n}\sum_{i=1}^{n} h''(\mu_i^0, \gamma^{00})V(\mu_i^0, \gamma^0)\left(\frac{\partial h(\mu_i, \gamma^{00})}{\partial\mu_i^0}\right)\left(\frac{\partial\mu_i}{\partial\beta^0}\right)\right|$$

$$= \left|\frac{1}{16\sigma^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1 + 2(\gamma^{00} - 1)\theta_i^0}{V(\mu_i^0, \gamma^{00})\sqrt{V(\mu_i^0, \gamma^{00})}}\right)V(\mu_i^0, \gamma^0)\frac{1}{\sqrt{V(\mu_i^0, \gamma^{00})}}\frac{\partial\mu_i}{\partial\beta^0}\right| = \left|\frac{1}{n}\sum_{i=1}^{n} W_i\left(\frac{1}{\mu_i^0}\right)\right| = O(\mu_H^{-1}),$$

where

$$W_i = \frac{1}{16\sigma^2}\left(\frac{1 + 2(\gamma^{00} - 1)\theta_i^0}{1 + (\gamma^{00} - 1)\theta_i^0}\right)\left(\frac{1 + (\gamma^0 - 1)\theta_i^0}{1 + (\gamma^{00} - 1)\theta_i^0}\right)\left(\frac{\partial\mu_i}{\partial\beta^0}\right).$$

Here μ_H is a weighted harmonic average of μ_i^0.

Proof of Theorem 2

a. Let $\delta = (\gamma, \sigma^2)$. The score functions for (β, δ) are

$$U_1(\beta, \delta) = \frac{\partial L}{\partial\beta}$$

and

$$U_2(\beta, \delta) = \frac{\partial L}{\partial\delta} = \left(\frac{\partial L}{\partial\gamma}, \frac{\partial L}{\partial\sigma^2}\right)^T.$$

Using the Taylor series expansion, we have

$$\sqrt{n}\begin{pmatrix}\beta^* - \beta^{00}\\ \delta^* - \delta^{00}\end{pmatrix} = \begin{pmatrix} -\frac{1}{n}\frac{\partial U_1(\beta, \delta)}{\partial\beta^{00}} & -\frac{1}{n}\frac{\partial U_1(\beta, \delta)}{\partial\delta^{00}}\\ -\frac{1}{n}\frac{\partial U_2(\beta, \delta)}{\partial\beta^{00}} & -\frac{1}{n}\frac{\partial U_2(\beta, \delta)}{\partial\delta^{00}}\end{pmatrix}^{-1} \times \begin{pmatrix}\frac{1}{\sqrt{n}}U_1(\beta^{00}, \delta^{00})\\ \frac{1}{\sqrt{n}}U_2(\beta^{00}, \delta^{00})\end{pmatrix} + o_p(1).$$

As $n \to \infty$, $(n^{-1/2}U_1(\beta^{00}, \delta^{00}), n^{-1/2}U_2(\beta^{00}, \delta^{00})^T)$ has an asymptotic normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$, where

$$\mathbf{\Sigma} = \lim_{n\to\infty}\frac{1}{n}\begin{pmatrix}\mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12}\\ \mathbf{\Sigma}_{12}^T & \mathbf{\Sigma}_{22}\end{pmatrix},$$

whose elements are defined by

$$\mathbf{\Sigma}_{11} = \operatorname{cov}(U_1(\beta^{00}, \delta^{00})) = \sum_{i=1}^{n} E_w(U_1(\beta^{00}, \delta^{00})^T U_1(\beta^{00}, \delta^{00})),$$

$$\mathbf{\Sigma}_{12} = \operatorname{cov}(U_1(\beta^{00}, \delta^{00}), U_2(\beta^{00}, \delta^{00})) = \sum_{i=1}^{n} E_w(U_1(\beta^{00}, \delta^{00})^T U_2(\beta^{00}, \delta^{00})),$$

$$\mathbf{\Sigma}_{22} = \operatorname{cov}(U_2(\beta^{00}, \delta^{00})) = \sum_{i=1}^{n} E_w(U_2(\beta^{00}, \delta^{00})^T U_2(\beta^{00}, \delta^{00})).$$

Also, as $n \to \infty$, we have

$$-\frac{1}{n}\frac{\partial U_1(\beta, \delta)}{\partial\beta^{00}} = \frac{1}{n}\mathbf{B}_{11} + o_p(1),$$

where

$$\mathbf{B}_{11} = \frac{\sum_{i=1}^{n}(\mathbf{D}_i^{00})^T(V_i^{00})^{-1}\mathbf{D}_i^{00}}{4\sigma^2},$$

$$-\frac{1}{n}\frac{\partial U_1(\beta, \delta)}{\partial\delta^{00}} = o_p(1),$$

and

$$-\frac{1}{n}\frac{\partial U_2(\beta, \delta)}{\partial\delta^{00}} = \frac{1}{n}\mathbf{B}_{22} + o_p(1),$$

where $\mathbf{B}_{22}$ is a 2×2 matrix whose elements are the second derivatives of the score function $U_2(\beta^{00}, \delta^{00})$ with respect to δ evaluated at δ^{00}. Thus the joint asymptotic distribution of $n^{1/2}(\beta^* - \beta^{00}, \delta^* - \delta^{00})^T$ is Gaussian with mean 0 and covariance matrix

$$\lim_{n\to\infty} n\begin{pmatrix}\mathbf{B}_{11}^{-1} & 0\\ 0 & \mathbf{B}_{22}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12}\\ \mathbf{\Sigma}_{12}^T & \mathbf{\Sigma}_{22}\end{pmatrix}\begin{pmatrix}\mathbf{B}_{11}^{-1} & 0\\ 0 & \mathbf{B}_{22}^{-1}\end{pmatrix}.$$

So the covariance matrix of $n^{1/2}(\beta^* - \beta^{00})$ is

$$\lim_{n\to\infty} n(\mathbf{B}_{11}^{-1}\mathbf{\Sigma}_{11}\mathbf{B}_{11}^{-1}).$$

Here we can easily verify that

$$\mathbf{\Sigma}_{11} = \frac{(\sum_{i=1}^{n}(\mathbf{D}_i^{00})^T(V_i^{00})^{-1}E(h(y_i, \gamma^{00}) - h(\mu_i^{00}, \gamma^{00}))^2\mathbf{D}_i^{00})}{4\sigma^4}.$$

b. Under the normality assumption, we can obtain the information matrix from the second derivative of the log-likelihood (4). It is easy to prove the information matrix is block diagonal; this follows because $\mathbf{\Sigma}_{12}$ evaluated at β^{00} equals 0. (For the detailed proof, see Kim 1991.)

Proof of Corollary 3

From the Taylor series expansion of the Gauss–Newton score equation, we have

$$\sqrt{n}(\beta^{**} - \beta^{00}) = \left(-\frac{1}{n}\frac{\partial U_s(\beta)}{\partial\beta^{000}}\right)^{-1}\left(\frac{U_s(\beta^{000})}{\sqrt{n}}\right) + o_p(1).$$

Using the similar arguments as given in the proof of Theorem 2, we can verify the results.

[Received September 1992. Revised October 1993.]

REFERENCES

Anscombe, F. J. (1948), "The Transformation of Poisson, Binomial and Negative Binomial Data," *Biometrika*, 35, 246–254.

Beall, G. (1942), "The Transformation of Data From Entomological Field Experiments so That the Analysis of Variance Becomes Applicable," *Biometrika,* 32, 243–249.

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society,* Ser. B, 26, 211–246.

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

——— (1990), "Tests of Hypotheses in Overdispersed Poisson Regression and Other Quasi-Likelihood Models," *Journal of the American Statistical Association,* 85, 565–571.

Brooks, R., James, W. H., and Grey, E. (1991}, "Modeling Sub-Binomial Variation in the Frequency of Sex Combinations in Litters of Pigs," *Biometrics,* 47, 403–417.

Carroll, R. J., and Ruppert, D. (1984), "Power Transformations When Fitting Theoretical Models to Data," *Journal of the American Statistical Association,* 79, 321–328.

——— (1988), *Transformation and Weighting in Regression,* New York: Chapman and Hall.

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

De Leeuw, J. (1993), "Block Relaxation Algorithms in Statistics," unpublished manuscript, University of California, Los Angeles, Dept. of Mathematics.

Haseman, J. K., and Kupper, L. L. (1979), "Analysis of Dichotomous Response Data From Certain Toxicological Experiments," *Biometrics,* 35, 281–293.

Hernandez, F., and Johnson, R. (1980), "The Large Sample Behavior of Transformations to Normality," *Journal of the American Statistical Association,* 75, 855–861.

Huber, P. (1981), *Robust Statistics,* New York: John Wiley.

Jennrich, R. I. (1969), "Asymptotic Properties of Non-Linear Least Squares Estimators," *Annals of Mathematical Statistics,* 40, 633–643.

Kim, D. K. (1991), "Regression models for overdispersed binomial data," unpublished Ph.D. dissertation, University of California, Los Angeles, Dept. of Biostatistics.

Liang, K. Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika,* 73, 13–22.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), New York: Chapman and Hall.

Moore, D. F., and Tsiatis, A. (1991), "Robust Estimation of the Variance in Moment Methods for Extra-Binomial and Extra-Poisson Variation," *Biometrics,* 47, 383–401.

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