Regression Models for Recurrent Event Data: Parametric Random Effects Models with Measurement Error

Bruce W. Turnbull, Wenxin Jiang
Statistics Center, Rhodes Hall, Cornell University, Ithaca, NY 14853-3801
and
Larry C. Clark
Arizona Cancer Center, University of Arizona, Tucson AZ 85724

March 5, 1997

Abstract

Statistical methodology is presented for the statistical analysis of nonlinear measurement error models. Our approach is to provide adjustments for the usual maximum likelihood estimators, their standard errors and associated significance tests in order to account for the presence of measurement error in some of the covariates. We illustrate the technique with a mixed effects Poisson regression model for recurrent event data applied to a randomized clinical trial for the prevention of skin tumors.

1. INTRODUCTION

In this paper, we will be concerned with the effect that measurement error in covariate readings can have on the statistical analysis of recurrent endpoint data from medical trials. In such trials, subjects experience a series of recurrent events over the duration of followup. Applications include infections in AIDS patients, epileptic seizures, asthma attacks, bladder cancer and many others. Our own motivation came from the "Nutritional Prevention of Cancer" (NPC) trial, the purpose of which was to study the long-term safety and efficacy of a daily 200μg nutritional supplement of selenium (Se) for the prevention of cancer. This is a double blind, placebo controlled randomized clinical trial and has accrued approximately 1300 patients since it started in 1983. More details on the design of this trial are given by Clark *et al.*[1]. A number of endpoints were considered, but here we shall concentrate on just one — namely squamous cell carcinomas (SCC) of the skin. For each patient, the time (measured from date of randomization) of each new occurrence of a SCC was observed. At randomization a number of baseline covariates were also recorded. Of course the most important of these was the treatment assignment (Se or placebo), but others included such variables as age, clinic, gender, smoking status, previous history of skin cancer, and blood biochemical levels, in particular plasma Se status. While some of these variables are recorded accurately, others, such as plasma Se status, are subject to measurement error. The purpose of this paper is to study the effect that such error can have upon standard inferential procedures.

2. A MIXED EFFECTS POISSON REGRESSION MODEL

The negative binomial regression model has proved useful for analyzing data of this kind.[2,3,4] Suppose there are n subjects. For subject i $(1 \leq i \leq n)$, we let Y_i denote the number of occurrences of the event of interest during the followup time of length T_i, and let X_i denote the vector of covariate values (including an intercept term). We then model the responses $\{Y_i\}$ as conditionally independent and Poisson distributed

$$p(y_i \mid \theta_i) \quad \sim \quad \text{Poisson}(\theta_i T_i e^{X_i'\beta})$$

where $\theta_i \quad \sim \quad g(\theta)$ are Gamma distributed with mean 1 and variance α. (The mean can be taken to be one without loss of generality because an intercept term is included in the vector X_i.) This can be viewed as a Poisson regression model with fixed covariate effects, but with extra-Poisson variation (*i.e.* that not explained by the fixed covariates) introduced via random subject effects. Alternatively it can be viewed as an empirical Bayes model where "frailties" $\{\theta_i\}$ have a prior distribution g.

With this model we can write down the likelihood function:

$$\prod_{i=1}^{n} \int_0^\infty p(y_i \mid \theta) g(\theta) d\theta \quad \propto \quad \prod_{i=1}^{n} \frac{\Gamma(Y_i + \frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})} \frac{[\alpha \exp(X_i'\beta)]^{Y_i}}{[1 + \alpha T_i \exp(X_i'\beta)]^{Y_i + \alpha^{-1}}} \tag{1}$$

Estimates of α and of the regression coefficients β, together with standard errors can then be obtained via the method of maximum likelihood. The computation can be done using one of several software packages such as the procedure NBREG in STATA 4.0.[5] Alternatively, EVENTREG[6] provides a particularly user friendly program to perform stepwise regression with this model.

Abu-Libdeh *et al.*[4] analyzed interim data from the selenium trial described in Section 1, using the above parametric model. In fact they went further, allowing Weibull inter-event times instead of exponential times as implied by the underlying Poisson model. However, using regression diagnostics tools that they developed, they were able to conclude that the Poisson model did provide an adequate fit to their data. (They also looked at Poisson models when the events could be of different types — for example, squamous and basal cell carcinomas in the skin cancer study.)

Although Abu-Libdeh *et al.*[4] examined a number of covariates, all were of a categorical type such as treatment assignment, gender, clinic, etc., where there were presumably few inaccuracies. However, none were of the kind where there might be ample opportunity for some measurement error, such as blood biochemical levels (Se, Vitamin E etc.). The baseline plasma Se status covariate is of obvious importance for two reasons. First, a nested cohort study of association between baseline Se and disease early in the trial might indicate what treatment effect might ultimately be expected. As such, the results could provide useful information to a monitoring committee considering the termination of the trial.[7] Second, baseline Se could be an effect modifier — a nutritional supplement of Se might be of less benefit to those subjects with already adequate baseline Se status. In fact, it was the first consideration that lead to the research that we report here. A regression analysis of early interim data using the models described in this section led to a smaller magnitude for baseline Se effect on the recurrence of skin tumors than might have been expected from previous observational studies. However it is well-known[8] that, in simple linear regression models with normal errors, the naive least squares slope estimate is biased — it underestimates the magnitude of the true slope if there is measurement error present for the regressor variable. This is known as the "attenuation" phenomenon. If one acknowledges the presence of measurement error, the questions that arise naturally are:

a. What is the effect on the corresponding estimated regression coefficients?

b. What is the effect, if any, on estimated regression coefficients for other variables that are measured without error, especially treatment ?

c. What is the effect on standard errors, tests, and confidence intervals ?

3. NONLINEAR MEASUREMENT ERROR MODELS

The mixed effects Poisson regression model of the previous section is a nonlinear model and now we have introduced covariate measurement error. The recent book[9] by Carroll, Ruppert and Stefanski (1995) provides a comprehensive account of current statistical methodology. There is a large literature on the subject and there are a number of different approaches. One approach considers the full likelihood, although this is feasible usually only for discrete covariates where measurement error is termed "misclassification". Examples of this approach are contained in Whittemore[10] and Gong *et. al.*[11]. Various approximate methods have been proposed appropriate in certain circumstances – *e.g.* the small error approximation (Whittemore and Keller[12]), the small incidence approximation for logistic regression (Rosner *et al.*[13]). For survival data regression, Prentice[14] considers the partial likelihood, while a "corrected" score function approach is taken by Stefanski and Carroll[15] and by Nakamura.[16] Only a few papers have been mentioned here – for a complete account, see Carroll *et al.*[9] cited above.

We now describe our general approach. We suppose that the response Y_i of the ith subject $(1 \leq i \leq n)$ has a density (or mass function if discrete) denoted by $f(y, X_i, \beta)$ conditional on the covariate vector X_i for that subject. Here β denotes the vector of all unknown parameters we wish to estimate and would include, for example, the parameter α in the negative binomial model (1) as well as the regression coefficients. We assume that the $\{Y_i\}$ are conditionally independent. Suppose however, for each i, that X_i is not measured exactly, but instead a surrogate Z_i is recorded. We will be assuming a "structural" model (Carroll et al.[9], page 6); that is, we will regard the true unobserved X_i's as independent realizations of a (vector) random variable X, for which we will be modeling the distribution. If the true covariates $\{X_i\}$ were known, the log-likelihood function would be given by:

$$\ell(b) = \sum_{i=1}^{n} \ell(Y_i, X_i; b) = \sum_{i=1}^{n} \log f(Y_i, X_i, b). \tag{2}$$

However, since the X's are not observed, this likelihood function cannot be evaluated. Instead we consider the *naive* log-likelihood function

$$\sum_{i=1}^{n} \ell(Y_i, Z_i; b)$$

where the observed surrogate variables are simply substituted for them. The *naive* maximum likelihood estimator (MLE) is then defined as:

$$\hat{b} = \arg\max \sum_{i=1}^{n} \ell(Y_i, Z_i; b).$$

(Here $\arg\max g(b)$ denotes a value of b that maximizes g.) Now typically,

$$\hat{b} \to b = b(\beta) \text{ almost surely as } n \to \infty,$$

where the limit, $b(\beta)$ say, depends on the true β, but is not equal to it. However the relationship can usually be inverted to obtain a consistent estimator for β, namely $\hat{\beta} = b^{-1}(\hat{b})$. The relationship between b and β is thus used to compute a "corrected" or "adjusted" estimate of β and it can also be used to obtain standard errors and test statistics, as will be described below.

First, however, we consider how this strategy applies in the case of simple linear regression through the origin with additive measurement error in the single predictor. The model is

$$\begin{aligned} Y &= \beta X + \epsilon \qquad (3) \\ Z &= X + U \qquad (4) \end{aligned}$$

We assume that X, U, ϵ are independent with zero means and variances $\sigma_X^2, \sigma_U^2, \sigma_\epsilon^2$, respectively. We observe n independent pairs $\{(Y_i, Z_i)\ 1 \le i \le n\}$. If we assume that the errors $\{\epsilon\}$ are normally distributed, the naive MLE is the least squares estimate that minimizes $\sum_{i=1}^n (Y_i - bZ_i)^2$. Thus $\hat{b} = \sum Y_i Z_i / \sum Z_i^2$. Using Slutsky's lemma and the law of large numbers, we see that the consistent limit of $\hat{b}$ is

$$b = b(\beta) = \frac{EYZ}{EZ^2} = \frac{EXZ}{EZ^2}\beta = \frac{EX^2}{EZ^2}\beta = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2}\beta.$$

Hence our estimate $\hat{\beta}$ of β is

$$b^{-1}(\hat{b}) = \frac{\sigma_X^2 + \sigma_U^2}{\sigma_X^2}\hat{b} = \frac{\sigma_X^2 + \sigma_U^2}{\sigma_X^2} \frac{\sum Y_i Z_i}{\sum Z_i^2}. \qquad (5)$$

We note that this estimator depends on the parameters of the distribution of X and U, which typically will be unknown. These parameters must then be estimated. Typically these estimates are based on a second data set, where pairs (X, Z) can be observed directly. This is usually termed a"validation" data set (Carroll et al.[9], page 12). It turns out that such a data set is available for the skin cancer example of Section 1; the details are given in Section 6.3. We also note here, in this simple example, the "attenuation" phenomenon that $\mid b \mid \ < \ \mid \beta \mid$ and the magnitude of the "naive" slope estimate $\mid \hat{b} \mid$ underestimates $\mid \beta \mid$. This is a common feature when measurement error is ignored in analyzing regression models.

We just presented the simplest example. The theorem that we will state in the next section applies to a wide variety of parametric models. It can be seen that a "model" consists of two parts – (a) a response model (the distribution of Y given X), and (b) a covariate error structure model (the joint distribution of Z and X). In our simple linear regression example, these are represented by (3) and (4), respectively. For response models, the methodology can be applied to many common situations, for example normal, logistic, probit, Poisson, exponential, proportional hazards and negative binomial regression models; these examples are discussed by Jiang[17]. Here we will be specifically interested in the repeated events model described in Section 2.

To describe the part of the model that relates to the covariate error structure, we first need to partition the true covariate vector as $X' = (1, \tilde{X}', W')$. Here the components represent the intercept term, those covariates $(\tilde{X})$ measured with error, and those covariates (W) measured without error, which are of dimensions $1, p, q$, respectively, say. We correspondingly partition the observed surrogate variable vector as $Z' = (1, \tilde{Z}', W')$. A covariate error structure model is one which specifies the joint distribution of X and Z. In our application in Section 6, we shall use a simple normal additive error model (NADD), where $\tilde{Z} = \tilde{X} + U$, $\tilde{X} \sim N(0, \Sigma_{\tilde{X}})$, $U \sim N(0, \Sigma_U)$

and are independent. It is worth noting that this implies the conditional distribution:

$$\tilde{X} \mid Z \quad \sim \quad N(\Lambda'\tilde{Z},\, \Sigma) \quad \text{where } \Lambda = \Sigma_{\tilde{Z}}^{-1}\Sigma_{\tilde{X}}, \quad \Sigma = \Sigma_U \Sigma_{\tilde{Z}}^{-1}\Sigma_{\tilde{X}}, \quad \Sigma_{\tilde{Z}} = \Sigma_{\tilde{X}} + \Sigma_U. \tag{6}$$

We will call the matrix Λ the "attenuation" matrix and it plays an important role. With little added effort, it is possible to apply the techniques we describe to more general models (CN) for the joint distribution of X and Z, in which the conditional distribution of $\tilde{X}$ given Z is normal with mean linear in Z and constant variance:

$$\tilde{X} \mid Z \quad \sim \quad N(C'Z,\, \Sigma) \quad \text{where} \quad C'Z = C_0' + \Lambda'\tilde{Z} + C_W'W$$

for general vector C_0 and general matrices C, Σ, C_W, and Λ. A special case of this model (CN) but which is more general than (NADD), is the model (CN1) in which the distribution of $\tilde{X}$ does not depend on W, the covariates measured without error. ($\tilde{X}$ of course still depends on the surrogates $\tilde{Z}$). This assumption might be valid if, for example, W is treatment assignment in a randomized trial. It might not be valid, however, if W is gender and $\tilde{X}$ is a blood biochemical level, for example. In model (CN1), $C_W = 0$ and without loss of generality we can re-center $\tilde{X}$, $\tilde{Z}$ to have mean zero, so that $C_0 = 0$ and $E(\tilde{X} \mid \tilde{Z}) = \Lambda'\tilde{Z}$, which is of the same form as in (6) for the additive model (NADD). The hierarchy is (NADD) $\subseteq$ (CN1) $\subseteq$ (CN). Another model, not in this hierarchy, but to which the methodology can apply is the multiplicative model[18], in which $Z = X \odot U$ with X and U independent, $U > 0$, $E[U] = 1$. We shall only be concerned with the NADD model here; for application of the more general models, see Jiang.[17]

4. A KEY THEOREM

Under mild regularity conditions listed in the Appendix, the following results hold:

I. $\hat{b} \xrightarrow{\text{a.s.}} b(\beta)$ and is asymptotically normally distributed as $n \to \infty$.

II. $\hat{\beta}_n = b^{-1}(\hat{b}_n) \xrightarrow{\text{a.s.}} \beta$ and $\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{\mathcal{D}} N(0, D'I^{-1}VI^{-1}D)$ with $I = -\mathrm{E}\nabla^2\ell(Y,Z,b)\Big|_{b(\beta)}$, the information matrix, $V = \mathrm{E}\nabla\ell\nabla\ell'\Big|_{b(\beta)}$, where $\nabla\ell$ is the score vector, and $D = \nabla b^{-1}\Big|_{b(\beta)}$, the gradient of b^{-1}.

III. $b = b(\beta)$ satisfies the estimating equation $\nabla \mathrm{E}_\beta \ell(Y, Z; b) = 0$.

The proof of this theorem will appear elsewhere.[17] However, essentially this follows from the results described in White[19] on "misspecified likelihoods," of which the naive likelihood is an example.

The strategy then is to obtain $b(\beta)$ from (I) or (III). If we find that b^{-1} is unique and invertible (as it was in our simple linear regression example of the previous section), we can "correct" or "adjust" the naive MLE by using $\hat{\beta} = b^{-1}(\hat{b})$.

The second part (II) of the above theorem shows how standard errors and test statistics can be constructed. The "naive" asymptotic variance (matrix) of $\hat{b}$ that ignores the presence of measurement error is $(nI)^{-1}$. The correct (robust) asymptotic variance of $\hat{b}$ is $\frac{1}{n}I^{-1}VI^{-1}$, the so-called "sandwich formula" — Huber[20], Carroll *et al.*[9] page 263. The asymptotic variance of the adjusted estimate $\hat{\beta}$ is thus given by $\frac{1}{n}D'I^{-1}VI^{-1}D$. If the expectations in I and V are not available then quantities based on sample averages can be used in the usual way, *e.g.* for I use

$-n^{-1}\sum \nabla^2 \ell(Y_i, Z_i, b)\Big|_{b(\hat{\beta})}$. To test the significance of a particular regressor variable, the jth say, using the Wald method, the naive test statistic that ignores the presence of measurement error would be the Z-value:

$$\mathcal{Z}_N = \frac{\hat{b}_j}{\sqrt{\mathrm{Avar}_N \hat{b}_j}}$$

where $\mathrm{Avar}_N(\hat{b}_j)$ is the jth diagonal element of $(nI)^{-1}$. The correct Z-statistic is:

$$\mathcal{Z}_{\mathrm{Adj.}} = \frac{\hat{\beta}_j}{\sqrt{\mathrm{Avar}\hat{\beta}_j}}.$$

where Avar $\hat{\beta}_j$ is the jth diagonal element of $\frac{1}{n}D'I^{-1}VI^{-1}D$.

5. APPLICATION OF THE THEOREM

We now apply the theorem to the mixed effects Poisson regression model and the likelihood (1) of Section 2. Suppose we consider the CN1 covariate error model described in Section 3, which includes the normal additive error model (NADD).

Recall that the true covariate vector is written as $X' = (1, \tilde{X}', W')$, and $Z' = (1, \tilde{Z}', W')$ is observed. The component vector W denotes the covariates measured without error. The parameters of the model (1) are α and $\beta' = (\beta_0, \tilde{\beta}', \gamma')$, which is the $1 + p + q$ dimensional vector of regression coefficients partitioned in the same way as X'. (Strictly speaking, in the notation of Section 4, we should include α as part of the vector of parameters β, but it is convenient to use separate notation for the mixing parameter α.) Similarly define $b' = (b_0, \tilde{b}', g')$. The naive ML estimates of parameters are denoted by $\hat{a}$ and $\hat{b}' = (\hat{b}_0, \hat{\tilde{b}}', \hat{g}')$, which can be obtained from available software as discussed in Section 2.

The estimating equation in part (III) of the theorem in Section 4 is applied to the likelihood (1). The naive log likelihood function is the sum of n i.i.d. copies of

$$l = \sum_{j=0}^{(Y-1)^+} log(1 + aj) + Y log\mu - (Y + a^{-1})log(1 + a\mu) \tag{7}$$

where $\mu = Texp(Z'b)$, and y^+ denotes $\max(y, 0)$. Here we use (a, b) in place of (α, β), in order to emphasize that they are not true underlying parameters, but only the arguments of the naive log likelihood function.

Differentiating the expectation of (7) with respect to the κ-th component of b $(0 \leq \kappa \leq p + q)$, we obtain

$$\partial_\kappa El = E\{\frac{(Y-\mu)Z_\kappa}{1 + a\mu}\} = E\{\frac{Z_\kappa T(exp(X'\beta) - exp(Z'b))}{1 + aTexp(Z'b)}\} = 0. \tag{8}$$

First note that the CN1 model implies that

$$\begin{aligned} E[\exp(X'\beta) \mid Z] &= \exp\{E(X \mid Z)'\beta + \frac{1}{2}\beta'\Sigma_{X|Z}\beta\} \\ &= \exp\{\beta_0 + \tilde{Z}'\Lambda\tilde{\beta} + W'\gamma + \frac{1}{2}\tilde{\beta}'\Sigma\tilde{\beta}\} \end{aligned} \tag{9}$$

since $\tilde{X} \mid Z \sim N(\Lambda'\tilde{Z}, \Sigma)$ and the conditional covariance matrix $\Sigma_{X|Z} = \text{diag}(0, \Sigma, 0)$.

By conditioning first on Z and T, taking expectations and using (9), we get

$$E\{\frac{Z_\kappa T(\exp(\beta_0 + \frac{1}{2}\tilde{\beta}'\Sigma\tilde{\beta} + \tilde{Z}'\Lambda\tilde{\beta} + W'\gamma) - \exp(b_0 + \tilde{Z}'\tilde{b} + W'g))}{1 + aTexp(Z'b)}\} = 0. \tag{10}$$

Here we have tacitly made the natural assumption that, given Z, the followup time T is independent of X. An obvious solution for (10) is:

$$b_0 = \beta_0 + \frac{1}{2}\tilde{\beta}'\Sigma\tilde{\beta}, \qquad \tilde{b} = \Lambda\tilde{\beta} \qquad g = \gamma. \tag{11}$$

This solution is unique, since the expectation of (7), El, is a globally concave function of b. This can be seen by noting that

$$\partial_\kappa \partial_\lambda El = -E\frac{\mu Z_\kappa Z_\lambda}{(1 + a\mu)^2} \cdot (1 + aY) \tag{12}$$

which is negative semi-definite; strictly negative definite if the components of Z are non-degenerate, in the sense that no component of Z can be expressed as a linear combination of the other components with probability one.

The equations in (11) give the asymptotic limit $b(\beta)$ of the naive MLE $\hat{b}$. Inverting (11) gives the adjusted estimates:

$$\hat{\beta}_0 = \hat{b}_0 - \frac{1}{2}(\Lambda^{-1}\hat{\tilde{b}})'\Sigma(\Lambda^{-1}\hat{\tilde{b}}), \qquad \hat{\tilde{\beta}} = \Lambda^{-1}\hat{\tilde{b}}, \qquad \hat{\gamma} = \hat{g} \tag{13}$$

The adjusted estimate $\hat{\alpha}$ is of less interest but may be obtained by solving numerically the derivative $\partial_a El = 0$ analogous to (8).

We now turn our attention to the standard errors of our adjusted estimate $\hat{\beta}$ which come from application of part (II) of the theorem to the naive log-likelihood (7). The information matrix I as defined there is block diagonal, which can be written as $I = \text{diag}(I_{aa}, I_{bb})$, say, where the components of I_{bb} are given by (12). Define V_{bb} be the $(1 + p + q) \times (1 + p + q)$ submatrix of the matrix V, defined in part (II) of the theorem, with elements:

$$E\partial_\kappa l \partial_\lambda l \;\; = \;\; E[(\frac{Y - \mu}{1 + a\mu})^2 Z_\kappa Z_\lambda],$$

obtained from (7). Finally define the submatrix of the derivative matrix:

$$D_{bb} = \frac{\partial\beta}{\partial b} = \begin{pmatrix} 1 & 0 & 0 \\ -(\Lambda^{-1})'\Sigma\Lambda^{-1}\tilde{b} & \Lambda^{-1} & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Since (13) does not involve a and I is block diagonal, we can compute the asymptotic variances of $\hat{\beta}$ by:

$$n\text{Avar}\hat{\beta} = D'_{bb} I_{bb}^{-1} V_{bb} I_{bb}^{-1} D_{bb} \tag{14}$$

Here all quantities are evaluated at the asymptotic limit of the naive MLE as given by (I) of the theorem. Since (14) depends on unknown quantities we replace b and a by $\hat{b}$ and $\hat{a}$, respectively, and replace expectations by the sample averages.

Similar results can be derived in the same fashion for other types of response models, (still assuming the CN1 covariate error model). The appropriate log-likelihood is simply substituted for (7). For example, for (*i*) the normal response model $Y \sim N(X'\beta, \sigma_\epsilon^2)$, (*ii*) the Poisson regression model with log link and possible exposure time offset $Y \sim \mathrm{P}(Te^{X'\beta})$, and (*iii*) the proportional hazards rate model where Y has hazard rate $\lambda_0(t)e^{X'\beta}$ for parametric function λ_0, the following effects are typically observed[17]:

A. $\hat{\tilde{\beta}} = \Lambda^{-1}\hat{\tilde{b}}$ and we observe the attenuation effect, namely the estimates of regression coefficients for variables measured with error are adjusted upward in magnitude.

B. $\hat{\gamma} = \hat{g}$, coefficient estimates of variables measured without error are unchanged.

C. Upon adjustment, the significance of regression coefficients is diminished, *i.e.* Z-values become smaller in magnitude. (For the normal regression model, Z-values are unchanged.) This may seem paradoxical at first in light of the fact that the magnitudes of adjusted regression coefficients are typically larger; however the diminished significance reflects the fact that the presence of measurement error has introduced more uncertainty into the problem.

We say these effects are "typical" because the theorem concerns asymptotic results. The relationships (A — C) may not hold exactly in finite samples. It is interesting to note that somewhat different phenomena are observed in the logistic regression model where $Y \sim \mathrm{Bin}[1,\ (1+\exp(-X'\beta))^{-1})]$. Then there can be an attenuation effect even on $\hat{\gamma}$, coefficient estimates of variables measured without error. Also the relationship between $\hat{\tilde{\beta}}$ and $\hat{\tilde{b}}$ is no longer linear, with $\hat{\tilde{\beta}} = \Lambda^{-1}\hat{\tilde{b}}$ holding only approximately for small $\mid \tilde{\beta} \mid$ — *cf* Rosner *et al.*[13] Space does not permit a full discussion of the details of the different regression models — this will appear elsewhere.[17] We now apply the results derived above for the mixed effects Poisson regression model to the skin cancer data and show how incorporation of measurement error can affect the results.

6. ILLUSTRATION WITH RECURRENT EVENT DATA

6.1. The model

For illustration we consider a simplified version of the model for the skin cancer endpoint data for the clinical trial described in Section 1. For a given patient, we let Y denote the number of squamous cell carcinomas experienced during followup time T. For simplicity of exposition we consider just two covariates. Covariate X is the patient's long run average baseline log(Se) plasma level. However X is measured with error and only Z is observed – the reading taken at randomization. The second covariate is W which takes on values 0 or 1, according to which of the two treatments was randomly assigned (the coding is blinded). We assume that W is known without error. (A full analysis might include other covariates and treatment by covariate interactions.) The followup time T is assumed independent of both covariates. Recall that our response model was a mixed effects Poisson regression model with offset. That is

$$Y \;\sim\; \mathrm{P}(\theta T e^{\beta_0+\beta_1 X+\gamma W}) \quad \text{with } \theta \sim \Gamma(\text{mean} = 1, \text{ var } = \alpha).$$

We adopt the additive measurement error model (NADD), which implies that:

$$Z = X + U \qquad X\ ,\ U \text{ indep.} \quad X \sim N(\mu_X, \sigma_X^2) \qquad U \sim N(0, \sigma_U^2)$$

In the notation of Section 5, $p = q = 1$ and we are more simply using X to denote $\tilde{X}$, β_1 for $\tilde{\beta}$ and b_1 for $\tilde{b}$. The matrices $\Sigma_{\tilde{X}}$ and Σ_U in the NADD model (6) are now replaced by scalars σ_X^2 and σ_U^2, respectively. The attenuation matrix Λ is the scalar $\sigma_X^2/(\sigma_X^2+\sigma_U^2)$.

6.2. Adjusted MLE's

We denote the naive ML estimates of parameters $\alpha, \beta_0, \beta_1, \gamma$ by $\hat{a}, \hat{b}_0, \hat{b}_1, \hat{g}$, respectively. These can be obtained from available software as discussed in Section 2. We are most interested in the estimates of the covariate effects. Equation (13) leads us to the adjusted estimates for parameters β_1 and γ with the simple and familiar form:

$$\hat{\beta}_1 = \frac{\sigma_X^2+\sigma_U^2}{\sigma_X^2}\,\hat{b}_1 \quad \text{and} \quad \hat{\gamma} = \hat{g}.$$

Because of the simple relationship between (β_1, γ) and (b_1, g), the derivative matrix in (14) simplifies. Thus the asymptotic variance of $\hat{\beta}_1$ is given by the center element of the 3×3 matrix $\frac{1}{n}\Lambda^{-2}(I_{bb}^{-1}V_{bb}I_{bb}^{-1})$ and that of $\hat{\gamma}$ is the bottom right entry of $\frac{1}{n}I_{bb}^{-1}V_{bb}I_{bb}^{-1}$. As before, we substitute MLE's for the unknown quantities in these expressions and use sample averages in place of expectations.

Unfortunately the estimate and asymptotic variance of $\hat{\beta}_1$ depend on σ_X and σ_U which are unknown. These cannot be estimated from the pairs (Y, Z). However, for the NPC clinical trial described in Section 1, a validation data set was available which could provide these estimates, and also be used as a check on the appropriateness of the additive covariate error model assumption.

6.3. Validation data set

For all patients in the NPC trial, plasma Se measurements were taken serially at approximate six month intervals, and not just at baseline (randomization). Assuming stationarity, the repeated readings from each placebo patient should represent replicated measurements of X for that patient. We are including the natural temporal variation in Se plasma levels as a component of the "measurement" error. Recall that X represents the long run mean level for an individual untreated patient. Of course we do not include Se readings of treated patients since their subsequent levels would be affected by the nutritional supplements of Se they were taking. The stationarity assumption can be examined by checking that the group mean Se levels remain approximately constant over time and by using control chart techniques ($\bar{X}$-charts for individuals, moving range charts)[21] on the longitudinal series of measurements in samples of individual placebo subjects. For the NPC trial data, the stationarity assumption seemed reasonable. There were 637 placebo patients. Let $Z_{i1}, Z_{i2}, \ldots, Z_{ir_i}$ denote the replicate log(Se) readings for the ith placebo patient ($i = 1, \ldots, 637$). We obtain estimates of $\mu_X = \mu_Z$ and of $\sigma_Z^2 = \sigma_X^2 + \sigma_U^2$ from the baseline readings $\{Z_{i1}\}$:

$$\begin{aligned}\hat{\mu}_X = \hat{\mu}_Z &= \frac{1}{637}\sum_{i=1}^{637} Z_{i1}\\ \hat{\sigma}_Z^2 &= \frac{1}{636}\sum_{i=1}^{637}(Z_{i1} - \hat{\mu}_Z)^2\end{aligned}$$

An estimate of σ_U^2 can be obtained from the pooled within placebo subject variability:

$$\hat{\sigma}_U^2 = \frac{\sum_{i=1}^{637}\sum_{j=1}^{r_i}(Z_{ij} - \overline{Z}_i)^2}{\sum_{i=1}^{637}(r_i - 1)} .$$

Finally we have $\hat{\sigma}_X^2 = \hat{\sigma}_Z^2 - \hat{\sigma}_U^2$ and the attenuation factor is estimated as $\hat{\Lambda}^{-1} = \hat{\sigma}_Z^2/\hat{\sigma}_X^2$.

The validation data set can also be used to check some of the distributional assumptions on X and U in the additive covariate error model. To do this, we restricted ourselves to the 220 placebo patients for whom ten or more serial Se readings were available at the time of the analysis, *i.e.* those patients with $r_i \geq 10$. For such patients, the mean log(Se) reading should be a reasonably accurate estimate of the "true" X-value. Thus we replace X_i with $\hat{X}_i = \overline{Z}_i = \frac{1}{r_i}\sum_j Z_{ij}$. The $\{Z_i\}$ are of course the initial Se readings taken at randomization, that is $Z_i = Z_{i1}$. We can then take U_i to be $\hat{U}_i = Z_i - \hat{X}_i, (1 \leq i \leq 220)$. A small correlation between the $\{\hat{U}_i\}$ and the $\{\hat{X}_i\}$ would indicate the appropriateness of the independence assumption. (Note this is not automatically zero since $\hat{X}_i$ is based on all $Z_{ij}, 1 \leq j \leq r_i$ not just Z_{i1}). Similarly histograms and probability plots of the $\{\hat{U}_i\}$ and the $\{\hat{X}_i\}$ can indicate the appropriateness of the normality assumption.

6.4. The skin cancer data

We applied the techniques to data from $n = 1277$ patients available at the sixth interim analysis of the NPC trial described in Section 1. The final data from this trial[22] are currently under review. For this illustration, some of the treatment assignment indicators in the data set have been deliberately switched to prevent premature conclusions and speculation on treatment effect.[23,24] However the data do represent what might be expected in a typical randomized trial of this type. Using the validation data set, the probability plots and histograms as described in Sec. 6.3, showed close agreement with the normality assumption. The correlation between the $\{\hat{U}_i\}$ and $\{\hat{X}_i\}$ was computed to be -0.037. Thus the data were consistent with the normal additive error model (ADD) for the log(Se) measurements. The mean log(Se) level was $\hat{\mu}_X = \hat{\mu}_Z = 4.717$ (0.007). (Quantities in parentheses represent standard errors.) The units for Se are ng/ml. Also $\hat{\sigma}_Z^2 = 0.186^2$ (0.044^2), $\hat{\sigma}_U^2 = 0.148^2$ (0.022^2), $\hat{\sigma}_X^2 = 0.113^2$ (0.049^2). This yields an estimate of the attenuation factor $\hat{\Lambda}^{-1} = 2.71$, which is quite large and implies a rather large measurement error. (Recall we include temporal fluctuations in our definition of measurement error.)

Table 1 shows the results of a naive analysis of the squamous cell carcinoma (SCC) recurrent event data using the mixed effects Poisson regression model with covariates treatment and plasma log(Se) level at baseline. This ignores any measurement error in the log(Se) covariate. The entries in the table were obtained by using the software of Natarajan et al.[6] When the measurement error in the log(Se) covariate is taken into account using the methods of Section 6.2, the point estimate $\hat{\gamma}$ of treatment effect is unchanged although its significance is diminished because we now use the robust estimate (sandwich formula) for its variance as described in Section 4. The magnitude of the point estimate $\hat{\beta}_1$ of baseline Se effect is increased by the attenuation factor $\hat{\Lambda}^{-1} = 2.71$ (from -0.692 to -1.873), yet the Z-value decreases in magnitude from 2.07 to 1.89. The results are summarized in Table 2. Note that the results exhibit the typical features (A) – (C) listed in Section 5.

[TABLES 1 AND 2 ABOUT HERE.]

Tables 1 and 2 ignore the the uncertainty in Λ and treat it as known. From the validation data, application of the delta method leads to a standard error for $\hat{\Lambda}^{-1}$ of 0.66. The sensitivity of the qualitative conclusions can be examined by repeating the same analyses with Λ^{-1} set equal to $\hat{\Lambda}^{-1} \pm s.e.$, say. This would lead to proportional changes in the point estimate of β_1 in Table 2 yielding a range of -1.41 to -2.33. However, because we are using the robust variance estimate, the standard errors adjust by the same proportion and the Z-value in Table 2 for β_1 is unchanged. Of course the point estimates and standard errors for γ are unaffected.

7. CONCLUDING REMARKS

We have presented a unified treatment for assessing measurement error effects on MLE's, asymptotic variances and P-values in generalized linear models. We have concentrated on the mixed effects Poisson regression model for recurrent endpoint data. In fact, our approach can apply to the more general setting of misspecified models. For example, in the situation of Section 6, we might initially ignore *both* measurement error and the extra-Poisson variation. This would involve only a simple Poisson regression analysis being needed, for which there is a much larger amount of standard computer software available. The techniques of Section 4 could then be used to adjust for either measurement error or extra-Poisson variation or both. The techniques of Section 4 can also be extended to handle multi-type recurrent event data.[4] Further work including theoretical details, applications to other models and simulation results for finite sample sizes will appear in Jiang.[17]

APPENDIX: REGULARITY CONDITIONS FOR THEOREM OF SECTION 4

The following conditions are sufficient for the theorem of Section 4. Other sets of sufficient conditions are possible — see Jiang[17]. We denote the parameter space as Θ where Θ is open and $\Theta \subseteq \Re^s$. Let $\mathcal{X}$ denote the support of $\mathbf{X}$ and $\mathbf{Y}$.

(A) the support of $\exp l(\mathbf{y}, \mathbf{z}, \mathbf{b})$ does NOT depend on $\mathbf{b}$;

(B) $E_{\boldsymbol{\beta}} l(\mathbf{Y}, \mathbf{Z}, \mathbf{b})$ exists and is finite $\forall\ \mathbf{b}, \boldsymbol{\beta} \in \Theta$;

(C) $l(\mathbf{y}, \mathbf{z}, \mathbf{b}) \in \mathcal{C}^3(\Theta) \quad \forall\ \mathbf{y}, \mathbf{z} \in \mathcal{X}$.

(D) $\forall\ \mathbf{b}_0 \in \Theta$, there exists an open ball $B_\epsilon(\mathbf{b}_0)$ such that $\overline{B}_\epsilon(\mathbf{b}_0) \subset \Theta$, and

- **(i)** $\sup_{\mathbf{b} \in \overline{B}_\epsilon(\mathbf{b}_0)} |\partial_i l(\mathbf{y}, \mathbf{z}, \mathbf{b})| \leq H_i(\mathbf{y}, \mathbf{z})$,
- **(ii)** $\sup_{\mathbf{b} \in \overline{B}_\epsilon(\mathbf{b}_0)} |\partial_i \partial_j l(\mathbf{y}, \mathbf{z}, \mathbf{b})| \leq H_{ij}(\mathbf{y}, \mathbf{z})$,
- **(iii)** $\sup_{\mathbf{b} \in \overline{B}_\epsilon(\mathbf{b}_0)} |\partial_i \partial_j \partial_k l(\mathbf{y}, \mathbf{z}, \mathbf{b})| \leq H_{ijk}(\mathbf{y}, \mathbf{z})$,

where $\partial_i \equiv \frac{\partial}{\partial b_i}$, and $E_{\boldsymbol{\beta}} H_i$, $E_{\boldsymbol{\beta}} H_{ij}$, $E_{\boldsymbol{\beta}} H_{ijk} < \infty\ \forall \boldsymbol{\beta} \in \Theta$, $\forall i, j, k = 1, 2, ...s$.

(E) **(i)** $\|\partial_j \partial_i E_{\boldsymbol{\beta}} l(\mathbf{Y}, \mathbf{Z}, \mathbf{b})\|$ (which exists by (C) and (D)) is negative definite $\forall \boldsymbol{\beta}, \boldsymbol{b} \in \Theta$.

(ii) $\|\partial_j \partial_i l(\mathbf{y}, \mathbf{z}, \mathbf{b})\|$ is semi-negative definite $\forall \mathbf{b} \in \Theta$, $\mathbf{y}, \mathbf{z} \in \mathcal{X}$.

(iii) If $\exists \mathbf{b}_0 \in \Theta$ and a directional vector $\mathbf{n}_0 \in \Re^s$ such that $|\mathbf{n}_0| = 1$, and

$$\partial^2_{\mathbf{n}_0} l|_{\mathbf{b}_0} \equiv \sum_{i,j} n_{0i} n_{0j} \partial_i \partial_j l(\mathbf{y}, \mathbf{z}, \mathbf{b})|_{\mathbf{b}=\mathbf{b}_0} = 0,$$

then $\partial^2_{\mathbf{n}_0} l = 0\ \forall \mathbf{b} \in \Theta$.

(F) The equation

$$\partial_j E_{\boldsymbol{\beta}} l(\mathbf{Y}, \mathbf{Z}, \mathbf{b}) = 0, \qquad j = 1, ..., s \qquad (L)$$

has a solution (which must be unique by (E)(i)) $\mathbf{b} = \mathbf{g}(\boldsymbol{\beta})$ for all $\boldsymbol{\beta} \in \Theta$;

(G) Let $\mathbf{b}^0 = \mathbf{g}(\boldsymbol{\beta})$ be the solution of (L). For all $\boldsymbol{\beta} \in \Theta$,
$\|E_{\boldsymbol{\beta}} \partial_i l(\mathbf{Y}, \mathbf{Z}, \mathbf{b}) \partial_j l(\mathbf{Y}, \mathbf{Z}, \mathbf{b})|_{\mathbf{b}=\mathbf{b}^0}\|$ is nonsingular, and all its elements are well-defined and finite;

(H) $\mathbf{g}(\boldsymbol{\beta})$ in (F) is a $\mathcal{C}^1$-diffeomorphism of Θ onto an open set $\mathbf{g}(\Theta)$.

REFERENCES

1. Clark, L.C., Patterson, B.H., Weed, D.L. and Turnbull, B.W. 'Design issues in cancer chemoprevention trials using micronutrients: application to skin cancer', *Cancer Bulletin* **43**, 519-524 (1991).

2. Lawless, J. F. 'Regression Methods for Poisson Process Data', *The Journal of the American Statistical Association*, **82**, 808-815 (1987).

3. Thall, P.F. 'Mixed Poisson likelihood regression models for longitudinal interval count data', *Biometrics*, **44**, 197-209 (1988).

4. Abu-Libdeh, H., Turnbull, B. W., and Clark, L. C. 'Analysis of multi-type recurrent events in longitudinal studies; application to a skin cancer prevention trial', *Biometrics*, **46**, 1017-1034 (1990).

5. Computing Resource Center, *STATA, Release 4*, Santa Monica, California (1994).

6. Natarajan, R., Turnbull, B.W., Slate, E.H., Wells, M.T., Clark, L.C. and Abu-Libdeh, H. 'A computer program for the statistical analysis of repeated event data using a mixed effects regression model', *Computer Methods and Programs in Biomedicine*, **42**, 283-294 (1994).

7. Jennison, C. and Turnbull, B.W. 'Interim analyses: the repeated confidence interval approach.' (with discussion) *Journal of the Royal Statistical Society B*, **51**, 305-361 (1989).

8. Fuller, W. A. *Measurement error models*, John Wiley & Sons, New York (1987).

9. Carroll, R. J., Ruppert, D. and Stefanski, L. A. *Measurement Error in Nonlinear Models*, Chapman and Hall, New York (1995).

10. Whittemore, A. S. 'Errors-in-variables regression problems in epidemiology', *Contemporary Mathematics*, **112**, 17-31 (1990).

11. Gong, G., Whittemore, A. S., and Grosser, S. 'Censored survival data with misclassified covariates: a case study of breast-cancer mortality', *The Journal of the American Statistical Association*, **85**, 20-28 (1990).

12. Whittemore, A. S. and Keller, J. B. 'Approximations for Regression With Covariate Measurement Error', *The Journal of the American Statistical Association*, **83**, 1057-1066 (1988).

13. Rosner, B., Spiegelman, D. and Willett, W. C. 'Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error', *American Journal of Epidemiology*, **132**, 734-745 (1990).

14. Prentice, R. L. 'Covariate measurement errors and parameter estimation in a failure time regression model', *Biometrika*, **69**, 331-342 (1982).

15. Stefanski, L. A. and Carroll, D. J. 'Conditional scores and optimal scores for generalized linear measurement-error models', *Biometrika*, **74**, 703-716 (1987).

16. Nakamura, T. ‘Corrected score function for errors-in-variables models: Methodology and application to generalized linear models’, *Biometrika*, **77**, 127-137 (1990).

17. Jiang, W. *Statistical Inference with Misspecified Models: Applications to Nonlinear Regression Models with Measurement Error, Over-dispersion and Omitted Covariates,* Ph.D. dissertation, Cornell University (1996).

18. Hwang, J. T. ‘Multiplicative errors-in-variables models with applications to recent data Released by the U.S. Department of Energy’, *The Journal of the American Statistical Association*, **81**, 680-688 (1986).

19. White, H. *Estimation, Inference and Specification Analysis*, Cambridge University Press, Cambridge, England (1994).

20. Huber, P.J. ‘The behavior of maximum likelihood estimates under nonstandard conditions’, *Proceedings of the 5th Berkeley Symposium on Probability and Statistics 1*, University of California Press, 221-233 (1967).

21. Montgomery, D.C. *Introduction to Statistical Quality Control, 2nd Ed.*, Wiley, New York (1991).

22. Clark, L.C., Combs, G. F., Turnbull, B.W., Slate, E.H. . Chalker, D.K., Chow, J., Davis, L.S., Glover, R.A., Graham, G.F., Gross, E.G., Krongrad, A., Lesher, J.L., Park, H.K., Sanders, B.B., Smith, C.L., Taylor, J.R. and the Nutritional Prevention of Cancer Study Group. 'Effects of selenium supplementation for cancer prevention in patients with carcinoma of the skin: A randomized clinical trial', *The Journal of the American Medical Association.* **276**(24), 1957-1963.

23. Geller, N.L. and Pocock, S.J. 'Interim analysis in randomized clinical trials: Ramifications and guidelines for practitioners', *Biometrics*, **43**, 213-223 (1987).

24. Green, S.J., Fleming, T.R. and O'Fallon, J.R. 'Policies for study monitoring and interim reporting of results', *Journal of Clinical Oncology*, **5**, 1477-1484 (1987).

Table 1: Naive analysis of NPC trial recurrent SCC data, ignoring measurement error: Maximum likelihood estimates and Z-values

			Baseline Se	Treatment
	a	b_0	b_1	g
Estimate	2.675	1.126	**−0.692**	**0.180**
$\sqrt{\text{Avar}_{\text{N}}}$	0.299	0.089	0.337	0.125
$\mathcal{Z}_{\text{N}}$	8.99	12.58	**−2.07**	**1.44**

Table 2: Analysis of NPC trial recurrent SCC data, adjusted for measurement error: Point estimates and Z-values

			Baseline Se	Treatment
	α	β_0	β_1	γ
Estimate	2.753	1.122	**−1.873**	**0.180**
$\sqrt{\text{Avar}}$	0.135	0.092	0.992	0.130
$\mathcal{Z}$	20.45	12.14	**−1.89**	**1.39**