Adjusting for Covariate Errors with Nonparametric Assessment of the True Covariate Distribution

By Donald A. Pierce

Radiation Effects Research Foundation, 5-2 Hijiyama Park, Hiroshima 732-0815, Japan

pierce@rerf.or.jp

And Albrecht M. Kellerer

Radiobiological Institute, Ludwig-Maximilians University, Schillerstrasse 42, Munich 80336, Germany

amk.sbi@lrz.uni-muenchen.de

SUMMARY

A well-known and useful method for generalised regression analysis when a linear covariate x is available only through some approximation z is to carry out more or less the usual analysis with $E(x \mid z)$ substituted for x. Sometimes, but not always, the quantity $\mathrm{var}(x \mid z)$ should be used to allow for overdispersion introduced by this substitution. These quantities involve the distribution of true covariables x, and with some exceptions this requires assessment of that distribution through the distribution of observed values z. It is often desirable to take a nonparametric approach to this, which inherently involves a deconvolution that is difficult to carry out directly. However, if covariate errors are assumed to be multiplicative and lognormal, simple but accurate approximations are available for the quantities $E(x^k \mid z),\ k = 1, 2, \cdots$. In particular, the approximations depend only on the first two derivatives of the logarithm of the density of z at the point under consideration and the coefficient of variation of $z \mid x$. The methods will thus be most useful in large-scale observational studies where the distribution of z can be assessed well enough in an essentially nonparametric manner to approximate adequately those derivatives. We consider both the classical and Berkson error models. This approach is applied to radiation dose estimates for atomic-bomb survivors.

Some Key Words: Errors in covariables, Berkson errors, Classical error model, Deconvolution of covariate errors, Generalised regression models, Regression calibration

1. INTRODUCTION

We are concerned with fairly general regression settings, such as generalised linear models or regression models for response times, that include a covariate x represented in the data only by some estimate or approximation z. We assume that covariate-estimation errors vary independently between individuals, and as usual that conditional on x the values of z are uninformative. It is well known that when the covariate x enters the model linearly it is often useful to replace x by the quantity $E(x \mid z)$ rather than z and carry out essentially, often exactly, the same analysis as if x had been observed. This is often referred to as regression calibration, although that term is used more generally. There are several alternative methods, some of which are indicated below, for dealing with covariate errors, and our aim is not to argue that regression calibration is always the most useful one among these. However, it is widely used, effective, reasonably well investigated, and has some particular advantages in applications motivating this work.

Even within regression calibration methods there are substantial variations. In particular one may sometimes estimate $E(x \mid z)$ directly by regression in a validation dataset where both x and z are available. In doing this, under the classical error model defined below and of primary interest here, one must take care that the marginal distribution of x in the validation dataset is approximately the same as in the primary dataset; see for example the admonition in Carroll et al. (1995, §1.3.3). Alternatively, if there is an assumed error model formulated in terms of the conditional distribution $p(z \mid x)$, then $E(x \mid z)$ can be computed from the joint distribution $p(x, z) = p(x)p(z \mid x)$ provided that one can make suitable inference about $p(x)$ from the observed distribution of the estimates z. This is the approach of primary interest here, where the distribution $p(x)$ is considered nonparametrically and it is assumed that $p(z \mid x)$ is lognormal with scale parameter x.

Direct nonparametric estimation of $p(x)$ from the marginal distribution of z-values in this setting, referred to as deconvolution when on some scale $z = x + e$, is

difficult although not totally impossible. A useful discussion of this and related matters is given in Carroll et al. (1995, Ch.12). In standard asymptotic formulations, explicit nonparametric estimation of $p(x)$ suffers from slow convergence, basically because convolution smooths out details of $p(x)$ that cannot be recovered. As is noted in that reference, two issues are particularly important for current needs: how best to use smoothness assumptions regarding $p(x)$, and that it is easier to estimate $E(x \mid z)$ than $p(x)$ itself. The contributions of the present paper are precisely in those directions. A recent approach of Schafer (2001) uses an EM approach to approximate directly something more general than $E(x \mid z)$ but not $p(x)$ in its entirely, namely the relevant score equation, conditional on z, for estimation of the regression parameter.

Developed here is a way of reducing deconvolution difficulties when $p(z \mid x)$ is taken as lognormal with scale parameter x, assuming only that $p(x)$ is suitably smooth. First, highly accurate approximation of $E(x \mid z)$ in terms of $p(x)$ is given by the ratio for $k = 1$ and $k = 0$ of Laplace approximations to integrals $\int x^k \; p(x)\, p(z \mid x)\, dx$, which involve only the first two derivatives of $\log p(x)$ evaluated at z. The approach also applies for $k > 1$ in the numerator, thus providing approximations to $E(x^k \mid z)$ that are useful for several purposes. A second and more crucial step involves similar Laplace approximations relating the required derivatives of $\log p(x)$ to those of the logarithm of the distribution of z. The approach is thus useful for samples large enough that these latter derivatives can be usefully estimated in an essentially nonparametric manner. The proposed approximation to $E(x \mid z)$ is an explicit function of the coefficient of variation of the distribution $p(z \mid x)$.

This approach enables tractable consideration of settings where the covariate errors are a combination of two types; namely that indicated above in line with the factorisation $p(x, z) = p(x)\, p(z \mid x)$, along with another component for which the more relevant factorisation is $p(x, z) = p(z)\, p(x \mid z)$. This distinction involves what are commonly referred to as the 'classical' and 'Berkson' covariate error models. In the classical error model z is an estimate of x in the ordinary sense, and in the Berkson model z is a different kind of approximation such as one arising from grouping x-values. Results for the combination of classical and Berkson errors are

given in the penultimate section of the paper, and until then all considerations are for only classical errors.

An alternative and popular class of approaches, where consideration of $E(x \mid z)$ does not arise, consists of adjustment of estimating equations to reduce biases in parameter estimators, which might be thought of as extending to generalised linear models the classical correction-for-attenuation methods (Fuller, 1987, p. 5). For example, this includes methods of Stefanski & Carroll (1987), Stefanski (1989) and Nakamura (1990, 1992); see also Carroll et al. (1995, Ch. 6). Generally, the adjusted estimating equations must be derived on a case-by-case basis in regard to the probability model for the primary response data. Another approach, attractive in being very direct and general although computationally intensive, is simulation extrapolation as developed by Cook & Stefanski (1994), Stefanski & Cook (1995), and described in Carroll et al. (1995, Ch. 5). None of these methods make explicit use of theoretical or estimated distributions $p(x)$ and $p(x \mid z)$, or of the marginal distribution of the observed z-values. On the contrary, some of them and much classical literature in the area strongly consider the collection of x-values as nuisance or incidental parameters, a type of modelling referred to as 'functional'.

Maximum likelihood might be considered as an alternative to the approaches indicated above, but our view is more that this is an ideal standard, usually very difficult to implement, to which other methods may be at least in principle compared. Likelihood methods involve averaging with respect to $p(x \mid z)$ the likelihood function if x were observed, and to implement this involves deconvolution in some sense. As noted above, Schafer (2001) considers maximum likelihood with a different approach to the deconvolution from that taken here. Our interest in regression calibration is largely as an approximation to maximum likelihood, and as noted below this approximation can in some settings be quite accurate, to the extent that $E(x \mid z)$ is determined. Since adjusted estimating equation methods do not involve $p(x)$ or $p(x \mid z)$, their merits seem to consist more of not having to assess these distributions than in being a natural approach for approximating maximum likelihood. This is not to say, however, that adjusted estimating equation methods cannot be highly efficient. These issues are discussed in Chapter 7 of Carroll et al. (1995), where it is noted that the general theoretical matters relevant to this issue are not clear.

Although it is not our aim to advocate regression calibration generally over such alternative methods, but rather to facilitate this for those who may be inclined for some reason to use it, the approach does have some particular advantages in our motivating applications: analyses of the atomic-bomb survivor data. There x is true radiation dose and fairly imprecise approximations z are provided by a general dosimetry system using survivor location and shielding. At our institutes and elsewhere a great many analyses of the same cohort data are made for different purposes and in exploratory modelling. These involve analyses for different health endpoints, employing several types of generalised regression models including survival analysis, and exploratory work involving choices and modelling of additional covariates. For this it is very convenient that once values of $E(x \mid z)$ are made available in the database, essentially the same methods for ongoing analyses can be employed as if x were observed. It would be enormously inhibiting for exploratory work, and probably seldom done even for final results, if either specialised adjusted estimating equations or computationally intensive simulation-based methods were required. Furthermore, in the final section we note a particularly important interpretation of $E(x \mid z)$ for our cohort study.

It may be helpful to clarify issues regarding use of more or less the same methods of analysis with regression calibration as if the true covariate were available. We will do this very casually, and in particular the Taylor's approximation below requires further attention or modification in a more careful treatment. It is difficult to document the history of the methodological suggestion and its development, but it has been popular in biostatistics since the work of Armstrong & Oakes (1982), Armstrong (1985) and Prentice (1982a,b); see further references in Carroll et al. (1995, Ch. 3). Ignoring covariates measured without error, and first considering x as observed, express the response data for an individual as $y = \eta(\alpha + \beta x) + e$, where conditionally on x we have $E(e) = 0$ and $\mathrm{var}(e) = \kappa(\alpha + \beta x)$. For exponential family generalised linear models, iteratively weighted least squares in this formulation is equivalent to maximum likelihood estimation. Writing $z' = E(x \mid z)$ and $y = \eta\{\alpha + \beta z' + \beta(x - z')\} + e$ suggests the approximation

$$y \mathrel{\dot{=}} \eta(\alpha + \beta z') + \dot{\eta}(\alpha + \beta z')\beta(x - z') + e = \eta(\alpha + \beta z') + u + e \ .$$

Then conditionally on z the error term for regression analysis becomes $u+e$. It is not difficult to see that $E(u+e\,|\,z)=0$ and

$$\text{var}(u+e\,|\,z)=E\{\kappa(\alpha+\beta x)\,|\,z\}+\{\dot{\eta}(\alpha+\beta z')\,\beta\}^2\,\text{var}(x\,|\,z)\ .$$

Employing exactly the usual methods means using iteratively re-weighted least squares in the model $y=\eta(\alpha+\beta z')+w$, where $w=u+e$, using variance function $\kappa(\alpha+\beta z')$. Since $E(u+e\,|\,z)=0$ this will lead to consistent estimation even though the variance function may not be appropriate. The substitution of $\kappa(\alpha+\beta z')$ for $E\{\kappa(\alpha+\beta x)\,|\,z\}$ will usually be innocuous, and the issue requiring consideration involves whether the term $\{\dot{\eta}(\alpha+\beta z')\,\beta\}^2\,\text{var}(x\,|\,z)$ representing overdispersion due to covariate error can be ignored. That the overdispersion involves $\text{var}(x\,|\,z)$ is one of the reasons why we provide approximations for $E(x^k\,|\,z)$ for $k>1$. Another reason is that when the regression model involves x^k, for $k>1$, then the regression calibration method entails substituting $E(x^k\,|\,z)$ for x^k.

When y is binomial with sample size m, the variance function becomes $k(\alpha+\beta z')+(m^2-m)\{\dot{\eta}(\alpha+\beta z')\,\beta\}^2\,\text{var}(x\,|\,z)$, so for the Bernoulli case $m=1$ there is no overdispersion. In fact, $y\,|\,z$ remains Bernoulli and iteratively re-weighted least squares is maximum likelihood. When y is Poisson, the variance function is $k(\alpha+\beta\tilde{z})+\{\dot{\eta}(\alpha+\beta\tilde{z})\,\beta\}^2\,\text{var}(x\,|\,z)$, and the overdispersion is usually negligible when $E(y)$ is small. For either the binomial case with large m or the Poisson case with large mean, the overdispersion should be taken into account. This might be done either by using the correct variance function, or by using the variance function $\kappa(\alpha+\beta\tilde{z})$ with perhaps only modest loss of efficiency, along with a robust method of estimating standard errors not relying on the incorrectly assumed variance function. For the analysis of survival data with response time T with hazard linear in x, Prentice (1982a) showed that using the standard analysis with substitution of z' for x provides a close approximation to maximum likelihood estimation provided that $E(x\,|\,z,T>t)/E(x\,|\,z)$ is near unity; that is, in comparison to z, the response time T carries relatively little information regarding x. Prentice (1982b) has shown that similar considerations hold for case-control studies.

We first present the main results for the classical error model, where the factorisation $p(x,z) = p(x)\,p(z \mid x)$ is most natural, and then apply these to our motivating example involving radiation doses for atomic-bomb survivors. Following that we consider the extension when Berkson errors, where the natural factorisation is $p(x,z) = p(z)\,p(x \mid z)$, are involved as well.

2. MAIN RESULTS

Results in this section are based on the factorisation $p(x,z) = p(x)\,p(z \mid x)$ most useful in the classical error model. As noted, we will assume that covariate estimation errors are multiplicative and $p(z \mid x)$ is lognormal with scale parameter x and constant coefficient of variation. Henceforth, we clarify notation by usually expressing the distributions of x and z as $p_X(\bullet)$ and $p_Z(\bullet)$, and so forth, unless the meaning is otherwise clear. We first provide an approximation to $E(x \mid z)$ as a functional of $p_X(\bullet)$, and then extend this to a functional of $p_Z(\bullet)$.

The argument and results are more direct on the log scale, and we write $x^* = \log(x)$ and $z^* = \log(z)$. Under our model we then have that $z^* = x^* + e^*$, where the covariate error e^* is Gaussian and independent of x^*. We first assume that $E(e^*) = 0$, so that $E(z^* \mid x^*) = x^*$, and then provide modifications for the case that $E(z \mid x) = x$. For either case write $\sigma^2 = \mathrm{var}(z^* \mid x^*)$, noting that the square root of this is the approximate coefficient of variation of $z \mid x$. Of course assessment of σ is an important and difficult matter, but one considered outside of the scope of this paper.

Our first main result is the following approximation to $E(x^k \mid z)$ as a functional of $p_X(\bullet)$, or more directly of $p_{X^*}(\bullet)$. This is given in terms of correction factors $C_k(z)$ defined and approximated as

$$E(x^k \mid z) \equiv \{C_k(z)\,z\}^k \quad \text{with} \quad C_k(z) \doteq \omega^{\{k+2d_1(z)\}/\{1-\sigma^2 d_2(z)\}}. \tag{1}$$

Here $\omega = \exp(\sigma^2/2)$ and $d_1(z)$ and $d_2(z)$ are the first two derivatives of $\log p_{X^*}(\bullet)$ with respect to x^*, evaluated at $z^* = \log(z)$. The result (1) is only valid when $1 - \sigma^2 d_2(z) > 0$. The formal derivation is very straightforward, following directly

from replacing $\log p_{X^*}(\bullet)$ by a second-order expansion at the point z^* in each of the integrals $\int \exp(kx^*)\,p(x^*)\,p(z^* \mid x^*)dx^*$ and $\int p(x^*)\,p(z^* \mid x^*)dx^*$, whose ratio is $E(x^k \mid z)$. This means that the relationship in (1) is exact when $p_X(\bullet)$ is lognormal, but our interest is in a general $p_X(\bullet)$. These approximations to the two integrals are Laplace-type (Barndorff-Nielsen & Cox, 1989), where the asymptotic index is σ^{-2}. The key to impressive accuracy, in this and many other settings, is that the underlying second-order expansion is made locally to each value of z and need only be reliable for x-values where $p(z \mid x)$ is substantial. Moreover, the approximation to the ratio is substantially better than that to either term since moderate errors tend to cancel (Tierney & Kadane, 1986). More formally, the approach requires that $d_2(z)$ be continuous, and then it is well known that for such Laplace approximations the error for each integral is $O(\sigma^{-2})$ and that for the ratio of the two integrals is $O(\sigma^{-4})$.

We now turn to approximations as functionals of $p_Z(\bullet)$, the distribution of observed quantities z. Further Laplace-type approximation explained below relates the derivatives $d_j(z)$ arising in (1) to the derivatives $\tilde{d}_j(z)$ of $\log p_{Z^*}(\bullet)$, with respect to z^* and evaluated at $z^* = \log(z)$. This relationship is given by

$$d_j(z) \triangleq \tilde{d}_j(z)/\{1+\sigma^2\tilde{d}_2(z)\}, \quad j=1,2\ , \tag{2}$$

and substituting these into (1) yields our second main result,

$$E(x^k \mid z) \equiv \{\tilde{C}_k(z)\,z\}^k \quad \text{with} \quad \tilde{C}_k(z) \triangleq \omega^{k\{1+\sigma^2\tilde{d}_2(z)\}+2\tilde{d}_1(z)}\ . \tag{3}$$

Implementing (3) involves estimating the derivatives required there from the empirical distribution of the observed covariate estimates z. Although the local nature of the underlying approximations might suggest using nonparametric local estimates of the derivatives $\tilde{d}_1(z)$ and $\tilde{d}_2(z)$, this should probably be avoided in most applications unless very substantial smoothing is employed. This is because the true derivatives being estimated will typically be quite smooth functions of z, and it will be unattractive at best if the correction factors $\tilde{C}_k(z)$ are not correspondingly smooth. Therefore the intended use of these methods, which should usually present no

serious difficulty, is that one fit globally to the empirical distribution of $\log p_{Z^*}(\bullet)$ a curve suitable for analytical differentiation, and with continuous second derivative. For this, taking $f\{\log p_{Z^*}(\bullet)\}$ as a third- or fourth-degree polynomial in $g(z^*)$, for some suitable functions $f(\bullet)$ and $g(\bullet)$ arrived at in an exploratory manner, should usually be adequate. Splines might be used, but with some attention to adequate smoothing and continuous second derivative.

Relationships (2) can be formally derived as follows. We have that $z^* = x^* + e^*$, where e^* is Gaussian and independent of x^*. Equations (2) would arise directly if x^* were Gaussian, in terms of moments of x^* and z^*. Furthermore, along lines similar to the development of (1), local approximations show that (2) holds more generally. In particular, differentiating the convolution equation yields that

$$p_{Z^*}(z^*)\tilde{d}_1(z^*) = \int p_{X^*}(z^* - e^*)d_1(z^* - e^*)\,p(e^*)de^*$$
$$p_{Z^*}(z^*)\left\{\tilde{d}_1(z^*)^2 + \tilde{d}_2(z^*)\right\} = \int p_{X^*}(z^* - e^*)\left\{d_1(z^* - e^*)^2 + d_2(z^* - e^*)\right\}p(e^*)de^* \ .$$

Of course these functional equations relating the derivatives have solution (2) when $p_{X^*}(\bullet)$ is Gaussian. Use of local second-order approximations of $\log p_{X^*}(z^* - e^*)$ at each z^* then provides (2) as an approximate solution to the functional equations. This is a Laplace-type approximation, but less accurate than (1) because the aim is not evaluation of the integrals but solution of the functional equations. Nonparametric deconvolution is a notoriously difficult problem, and the evaluation in the following section of the adequacy of approximation (3) is especially important.

There is interest in conditions such that $E(x \mid z) < z$, since one might expect this to be true for large values of z when the distributions $p_X(\bullet)$ or $p_Z(\bullet)$ are highly skewed to the right. Clearly, in the case of approximations (1) and (3) that condition holds in terms of $p_X(\bullet)$ when $d_1(z) < -1/2$, or in terms of $p_Z(\bullet)$ when $\tilde{d}_1(z) < -\{1 + \sigma^2\tilde{d}_2(z)\}/2$.

Furthermore, it will often be the case that $d_2(z)$ and $\tilde{d}_2(z)$ vary modestly over the range of z, in which case the ratios $C_2(z)/C_1(z)$ and $\tilde{C}_2(z)/\tilde{C}_1(z)$ are fairly constant. This would mean that $E(x^2 \mid z)$ is approximately a constant multiple of

$E(x \mid z)^2$, and thus that $\mathrm{var}(x \mid z)$ is approximately proportional to $E(x \mid z)^2$. The former is useful when testing linearity in x of the regression model by considering a quadratic alternative, and the latter for considering the effect of overdispersion introduced by the distribution of $x - E(x \mid z)$.

Finally, in the case that $E(z \mid x) = x$, rather than $E(z^* \mid x) = x^*$ as assumed above, the changes are as follows. The z-variate, z_{alt} say, unbiased on the linear scale is ω^{-1} times that unbiased on the log scale, so one could simply rescale z_{alt} and apply the above methods. We note, however, that in the case of approximation (3) the result is that $\tilde{C}_k^{\mathrm{alt}}(z_{\mathrm{alt}}) = \omega \tilde{C}_k(z)$, and for approximation (1) it is $C_k^{\mathrm{alt}}(z_{\mathrm{alt}}) = C_{k+1}(z)$. In the case of approximation (3), this involves some change in (2), namely that the expression for $j = 1$ becomes $d_1(z) = \{\tilde{d}_1(z) - \sigma^2 \tilde{d}_2(z)\} / \{1 + \sigma^2 \tilde{d}_2(z)\}$, whereas that for $j = 2$ remains the same.

3. EXAMPLE

The example involves rather imprecisely estimated radiation doses for atomic-bomb survivors followed up by the Radiation Effects Research Foundation in Hiroshima and Nagasaki, which are used in the study of radiation-related cancer and other diseases. The dosimetry system providing estimates z is documented in Roesch (1987). We note that it consists of having assessed each survivor's location and shielding through interviews, and then applying physical calculations of radiation transport through air and shielding materials.

The value of the proposed methods becomes clearer in relation to what has been used in the past for these data. Pierce et al. (1990), also reported in Pierce et al. (1992), estimated the distributions of true radiation doses for Hiroshima and Nagasaki as Weibull distributions, although restricted to the dose range used below, with cumulative distribution functions $F_X(x) = 1 - \exp(-\theta_1 x^{\theta_2})$; $\theta_1 = \{2.84, 2.33\}$, $\theta_2 = 0.5$, where the two values of θ_1 are for Hiroshima and Nagasaki, respectively. The adjustment factors $C_1(z)$ and $C_2(z)$ were then computed by numerical integration, and approximated in terms of second-degree polynomials in $\log(z)$ for routine use. We have found that the error in approximation (1) for this specific Weibull setting is totally negligible, but there was really no need for this when assuming that the above

estimates of $p_X(\bullet)$ are adequate. The unattractive aspects of the methodology were involved in the deconvolution step used in arriving at the above estimates of $F_X(x)$. First $p_Z(\bullet)$ was fitted by a Weibull distribution for each city, using the restricted dose range indicated below. Then, for selected values of $\sigma = SD\{\log(z) \mid x\}$, trial and error was used to find Weibull distributions for $p_X(\bullet)$ that correspond approximately to the estimated Weibull distributions for $p_Z(\bullet)$. The criterion for this correspondence was rather arbitrary, but the weakest link was assuming Weibull forms, particularly for $p_X(\bullet)$. This method has been used extensively at the Radiation Effects Research Foundation and elsewhere for a decade, assuming 35% coefficient of variation classical errors, that is $\sigma = 0.35$. Comparison of previous results to those obtained here is indicated below.

We illustrate the new methods here for Nagasaki, where the log empirical density $\log p_Z(\bullet)$ departs more than for Hiroshima from a second-degree polynomial, for which the approximation (1) is exact. The dose range used here is 0.10 – 6 Gy (Gray) of primary interest for the adjustments in question, which for Nagasaki involves about 4000 survivors, with dose quartiles 0.20, 0.45 and 0.85 Gy. About 85% of Nagasaki survivors in the cohort have doses less than 0.10, where radiation risks are very small, and including these low-dose survivors would greatly and unnecessarily complicate considerations here. As a perspective on radiation dose levels, with whole-body exposures a dose of 1 Gy would require serious medical attention, a dose of 3 Gy is roughly the LD_{50}, and annual occupational limits for radiation workers are around 0.02–0.05 Gy. Roughly speaking, cancer rates for a survivor are increased by about 50% per Gy for all remaining lifetime (Pierce et al., 1996).

Figure 1 shows the log-log empirical density of dose estimates z for Nagasaki, along with the fit of a third-degree polynomial on the indicated dose range, namely $\text{const} - 0.543\,z^* - 0.503(z^*)^2 - 0.110(z^*)^3$. Our initial aim was to use the exploratory method suggested in the previous section, but this worked out much more simply than expected since the transformations $f(\bullet)$ and $g(\bullet)$ suggested there were not necessary in this case.

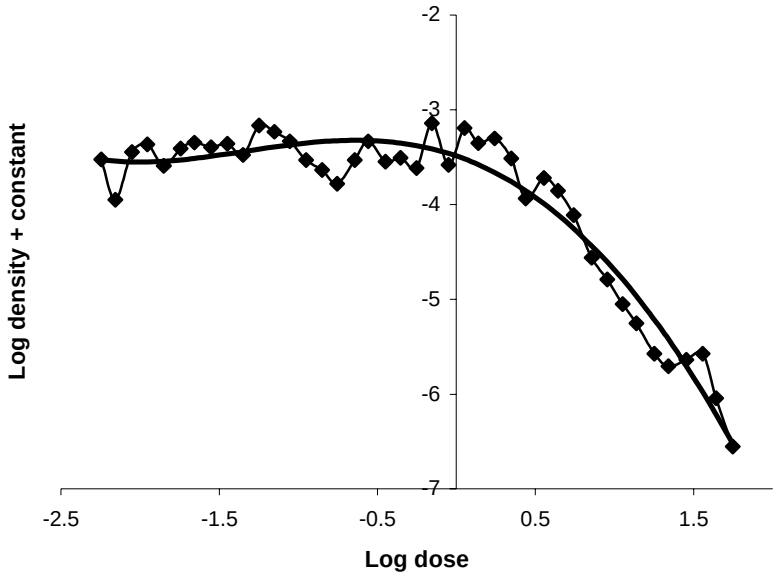

Fig. 1. Empirical density and cubic polynomial fit, log-log scale, for Nagasaki and dose range 0.10 – 6 Gy.

In order to demonstrate the accuracy of the Laplace approximations in (1), we first pretend that the curve in Fig. 1 estimates the distribution $p_X(\bullet)$ of true doses. Figure 2 shows, for $\sigma = 0.35$ and $\sigma = 0.7$, correction factors $C_1(x)$ and $C_2(x)$ computed both by numerical integration using the cubic polynomial and by approximation (1). Henceforth we make plots on the range 0.25–6 Gy to avoid, in integrations, excessive extrapolation to doses below the 0.10 used in Fig. 1. The adjustment factors, both exact and approximate, appear to be unstable for doses less than 0.25. For the application we have little interest in this, particularly in view of the extrapolative uncertainties involved, since risk estimation relative errors at very low doses dominate effects of dose adjustment.

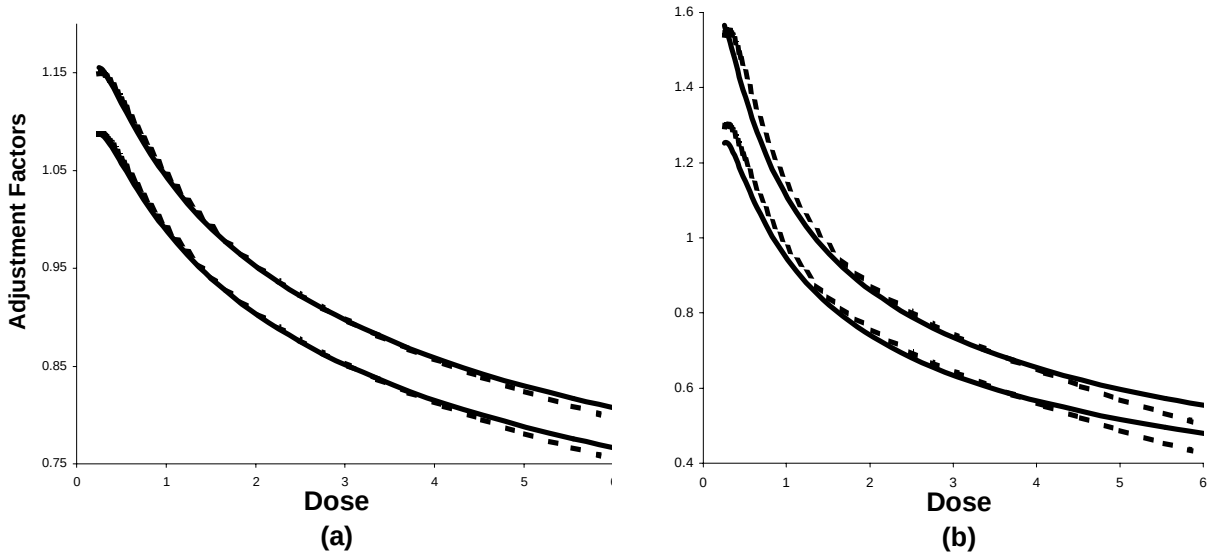

Fig. 2. Comparison, for (a) $\sigma = 0.35$ and (b) $\sigma = 0.70$, of approximation (1) (dashed lines) to exact results (solid lines) based on numerical integration, taking the curve in Fig. 1 as estimating $p_X(\bullet)$ rather than $p_Z(\bullet)$. The lower pair of lines pertains to $C_1(x)$ and the upper pair to $C_2(x)$.

In Fig. 3 we show for $\sigma = 0.35$, the choice ordinarily used for our setting, final results of using $\tilde{C}_1(z)$ and $\tilde{C}_2(z)$ given in (3). These are compared to $C_1(z)$ and $C_2(z)$ given by (1) when taking $d_j(z) = \tilde{d}_j(z)$, to demonstrate the effect of the deconvolution to estimate $p_X(\bullet)$ from $p_Z(\bullet)$. We note that these results differ substantially from what was obtained using the previous method indicated at the outset of this section. For example, whereas $\tilde{C}_1(4) = 0.78$ in Fig. 3, the value obtained before was $C_1(4) = 0.87$. We believe that this reflects the inadequacies of the previous methods, and that the current assessment is essentially correct, given the assumptions. However, we note that such a discrepancy is not really large in relation to uncertainties of modelling covariate errors. Beyond the dependence on the assumed value of σ, we note that assuming $E(z \mid x) = x$ rather than $E(z^* \mid x) = x^*$, as we have done here, would result in $\tilde{C}_1(4) = 0.83$. Such variations, however, do not have a large effect on eventual estimation of quantities such as radiation-related cancer risk, and the primary value of methods under consideration is to show that covariate errors roughly as considered here actually have very modest effect on final regression analyses of that nature.

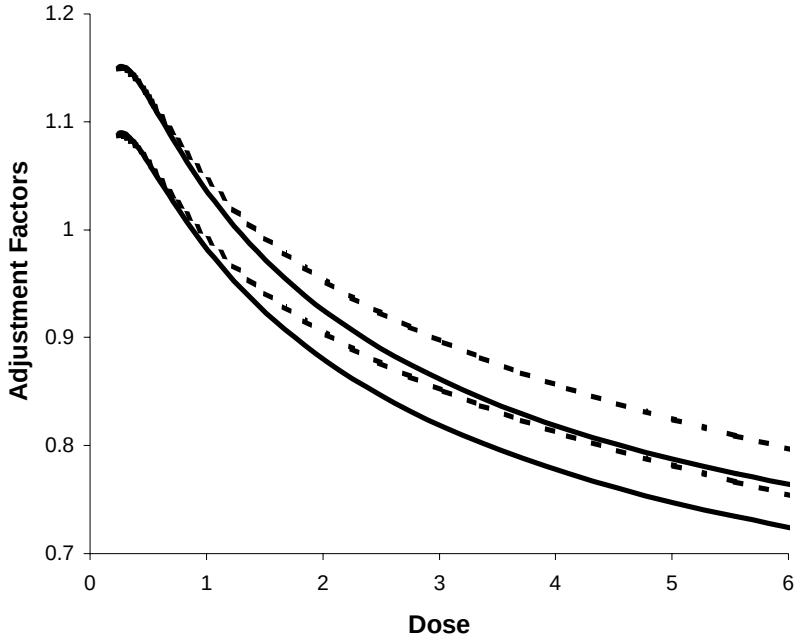

Fig. 3. Final results, with $\sigma = 0.35$, for $\tilde{C}_1(z)$ and $\tilde{C}_2(z)$, the solid curves of each pair, where the lower pair contains $\tilde{C}_1(z)$. Dashed curves in each pair are $C_1(z)$ and $C_2(z)$ when one takes the derivatives of $p_X(\bullet)$ as those of the observed $p_Z(\bullet)$, omitting the deconvolution step.

We have evaluated by simulation the use of approximation (3), compared to exact results based on the true distribution $p_X(\bullet)$, for settings similar to this example. For the sample size arising in the example, discussed further below, Table 1 summarises the errors in estimating $C_1(z)$, relative to the reduction $1-C_1(z)$, at $z = 4$ Gy, for the three true dose distributions shown in Fig. 4 and two values of σ. The intermediate distribution there corresponds essentially to the fitted curve shown in Fig. 1, rounding the polynomial coefficients there to $\text{const} - 0.50\,x^* - 0.50\,(x^*)^2 - 0.10\,(x^*)^3$. The other two distributions correspond to replacing the coefficient $b_3 = -0.10$ by $b_3 = 0$ and $b_3 = -0.20$. In Table 1, notation such as $> -20\%$ means at least 20% error in the negative direction. Thus, for example, with $b_3 = -0.10$ as in the example and for $\sigma = 0.35$, there were 6.9% + 1.4% = 8.3% of trials with relative error greater than 20%. As discussed below, most of the error seen in Table 1 is not due to approximation (3) but to errors in estimation of the true $p_Z(\bullet)$.

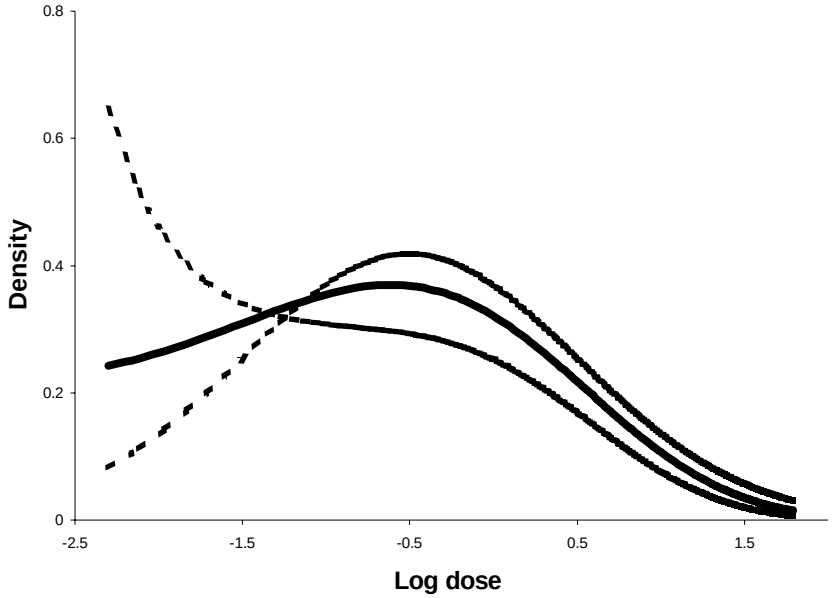

Fig.4. Assumed densities of true dose for simulation. Solid line corresponds to Fig. 1 and the dashed lines to $b_3 = 0$ and -0.20, with the former giving the more Gaussian-like curve.

Table 1. Simulation distributions, with 3,000 trials, of relative error of approximation (3) at 4 Gy

	Relative Error							
b_3	> –30%	> – 20%	> –30%	–10% to 10%	>10%	>20%	>30%	mean
				$\sigma = 0.35$				
0	0.012	0.071	0.248	0.584	0.168	0.036	0.007	0.016
-0.10	0.017	0.069	0.265	0.624	0.110	0.014	0.001	0.032
-0.20	0.040	0.135	0.380	0.545	0.075	0.005	0.000	0.069
				$\sigma = 0.50$				
0	0.010	0.071	0.258	0.573	0.169	0.041	0.007	0.017
-0.10	0.014	0.068	0.259	0.634	0.108	0.014	0.001	0.033
-0.20	0.052	0.189	0.469	0.486	0.045	0.005	0.000	0.096

Before giving some further details of the simulation, we discuss important matters of perspective that will arise more generally. For the application we are mainly interested in the value of $C_1(z)$ in the range 2–4 Gy. Although there are about 4,000 Nagasaki survivors with dose estimates in the range 0.1 to 6 Gy, only about 400 of these have dose estimates of at least 2 Gy. In the simulations we maintain for the other dose distriubutions the 4,000 individuals in the range 0.1 to 6 Gy. Most of the error seen in Table 1 comes not from the Laplace-type approximations leading to (1) and (2), but from estimating the derivatives $\tilde{d}_j(z)$. Evidence for that derives from the fact that for $b_3 = 0$ the approximations (1) and (2) are exact, and the errors for the three values of b_3 are similar. The error in estimating the $\tilde{d}_j(z)$ depends on the method used for this estimation, described below.

If instead of 4,000 there were only 2,000 survivors in the full dose range there would be so few at above 2 Gy that smoothness assumptions stronger than those used here would be required. If there were 10,000 survivors in the full dose range, errors such as those reported in Table 1 would be much smaller. For example, the 8.3% chance, referred to above, of errors of at least 20% for $\sigma = 0.35$ is then reduced to 0.8%, and essentially the same reduction occurs for $\sigma = 0.50$. Those 10-fold reductions are for $b_3 = -0.10$, and for $b_3 = -0.20$ the corresponding reductions are by factors of about 6 for $\sigma = 0.35$, and about 3 for $\sigma = 0.50$. This indicates, as claimed, that fairly generally the errors do not result mainly from approximations (1) and (2) but from the inherent difficulty of estimation of the $\tilde{d}_j(z)$ in the right-hand tail of the distribution.

In the simulation we take the approach indicated in Fig. 1 of fitting a cubic polynomial to the log-log density of z for use in (2) and (1). Use of a quartic polynomial gives very similar results. This is not strictly speaking a nonparametric approach, but indicates the general nature of our intended application. It would become somewhat more nonparametric if as suggested earlier one finds by exploration suitable nonlinear transformations of both the log density and log doses for use of the polynomial-based estimation of required derivatives. In the simulation unweighted regression of empirical densities in 40 equal width bins was used. It is important to generate true doses over a wider range that that used for estimated doses,

so that the convolution remains in effect near the endpoints. Sampling was continued until there was the desired number of estimated doses in the range 0.1–6 Gy, this being 4,000 for Table 1.

4. BERKSON ERRORS

The classical error model is appropriate when z is in the usual sense an estimate of x, such as that on the original or logarithmic scale $z = x + e$ with $\mathrm{cov}(e, x) = 0$. The Berkson error model arises when z is some other type of approximation, such as would arise from grouping x-values, in which case it is better to think of $x = z + e$ where $\mathrm{cov}(e, z) = 0$, along with the factorisation $p(x, z) = p(z)\,p(x \mid z)$. In the classical case z is more variable than x, and in the Berkson case it is less variable.

Although we chose not to complicate the above example with such matters, and could not deal with them using previous methods, for the atomic-bomb survivor setting the overall error is a composite of the two types. Classical error arises mainly from estimation of survivor location and shielding, whereas Berkson error arises from using the location and shielding information only to some approximation, which would be necessary even if this information were known without error.

A formulation for our lognormal setting allowing for both classical and Berkson errors can be expressed in terms of a latent variable $u = \exp(u^*)$ as

$$u^* = x^* + e_C^* \tag{4}$$

$$u^* = z^* + e_B^* \tag{5}$$

where e_C^* is classical with variance σ_C^2 and independent of x, e_B^* is Berkson with variance σ_B^2 and independent of (z, e_C^*). That is, u can be thought of as an unobserved 'estimate' of x as in (4), whereas because of some smoothing or grouping what is available is only z as provided by (5). A slightly different formulation was utilized by Tosteson & Tsiatis (1988) and Reeves et al. (1998), equivalent to the above at least when all the variables are normally distributed.

The methods of this paper can be readily employed for this more general setting as follows. Suppose first, artificially, that the required derivatives of $p_X(\bullet)$ were known. Even though u is not observed, approximation (1) can be applied to equation

(4) to obtain $E(x^k \mid u) \triangleq \{C_k(u)u\}^k$. Then u can be eliminated from this result by taking the expectation corresponding to equation (5), conditionally on z. Usually, the functions $\{C_k(u)u\}^k$, for $k = 1, 2$ are close enough to linear in u to allow using for this expectation simply $\{C_k(z)z\}^k$.

To carry this out in practice requires us, as before, to relate the derivatives of $p_{X^*}(\bullet)$ to those of $p_{Z^*}(\bullet)$ in the more general setting. We show at the end of this section that the generalization of the previous relationship (2) that corresponds to the model given by (4) and (5) is

$$d_j(z) \triangleq \tilde{d}_j(z) / \{1 + (\sigma_C^2 - \sigma_B^2)\tilde{d}_2(z)\}. \tag{6}$$

Thus, given estimates of the derivatives $\tilde{d}_j(z)$, which do not depend on the error model, one uses (6) to estimate the derivatives $d_j(z)$ to be used in approximation (1), taking $\sigma^2 = \sigma_C^2$, for obtaining the intermediate result $E(x^k \mid u) \triangleq \{C_k(u)u\}^k$. Then equation (5) involving σ_B^2 is employed to compute the expectation $E[\{C_k(u)u\}^k \mid z] \triangleq E(x^k \mid z)$. As noted above, one may usually take $E[\{C_k(u)u\}^k \mid z] \triangleq \{C_k(z)z\}^k$ with negligible error. If necessary, this error may be reduced by approximating $\{C_k(u)u\}^k$ as log-quadratic in $u^* - z^*$, so that the expectation under (5) can be evaluated exactly when e_B^* is normally distributed.

For our example, consider the two cases where we assume that $\sigma_B = \sigma_c = 0.35$ on the one hand, or $\sigma_c = 0.35$, $\sigma_B = 0$ on the other. In the first case we see from (6) that the net effect of the deconvolution is nil, so that we apply approximation (1) with $d_j(z) = \tilde{d}_j(z)$. Intuitive support for this is provided by recalling that classical errors alone result in z being more variable than x, and Berkson errors alone result in z being less variable than x, so when both types of error are equally present these effects cancel. Since we can use the approximation $E\{C_1(u)u \mid z\} \triangleq C_1(z)z$, we have already shown the comparison of interest in the bottom two curves in Fig.3, where the dashed one was given there simply to illustrate the effect of omitting the deconvolution.

We were surprised to find that adding the Berkson error results in adjustment factors nearer to unity, meaning that adding these errors actually decreases the bias in the naïve analysis when simply replacing x by z. On reflection, it seems this may be related to results following up on a suggestion by Wald, that under the classical error model the bias may be reduced by grouping the data; see for example Cheng & Van Ness (1999, Ch. 4).

To derive (6), denote by $\hat{d}_j(\bullet)$ the derivatives of $p_{U^*}(\bullet)$ analogous to those already considered. Applying the previous result (2) to equation (4) yields

$$d_j(u) \doteq \hat{d}_j(u)/\{1+\sigma_C^2\,\hat{d}_2(u)\}, \quad j=1,2\ ,$$

and application of (2) to equation (5) yields

$$\tilde{d}_j(u) \doteq \hat{d}_j(u)/\{1+\sigma_B^2\,\hat{d}_2(u)\}, \quad j=1,2\ .$$

Solving these four equations provides the result (6).

5. DISCUSSION

In §1 we noted that the quantities $E(x\,|\,z)$ have a particularly important interpretation for the atomic-bomb survivor data. The dosimetry system does not take into account information regarding dose provided by the fact of survival of the bombs, but only survivor location and shielding. To a substantial extent the neglected information is reflected in the distribution $p_X(\bullet)$, this being very highly skewed with relatively few survivors at high doses. If the survival of a dose x were proportional to $p(x)$, then $E(x\,|\,z)$ would take into account the information provided by survival. What is actually desired, though, is an appropriate dose estimate not specifically for survivors but for those belonging to the cohort study, and $E(x\,|\,z)$ may be considered the appropriate dose estimate for them. These issues are discussed more thoroughly in Pierce et al. (1990) and Pierce et al. (1992). In our view, similar considerations are involved in any observational study, where $E(x\,|\,z)$ reflects aspects of selection of the study population that are normally not considered in arriving at the estimates z.

We now consider the limitations of the proposed method for approximating $E(x\,|\,z)$. Approximation (1) would be usually very accurate if the true $p_X(\bullet)$ were known and the second derivative of the log density were reasonably smooth. The

more difficult issues arise with regard to the use of approximation (2), partly in assessing the accuracy of the relationship itself if the true values of the $\tilde{d}_j(z)$ employed there were known, but probably more importantly in estimating those derivatives. The simulation results in §3 assess the combination of both of these matters, providing for some separation of them since one of the choices for $p_X(\bullet)$ is normal, where the Laplace approximations are exact. Those results suggest that when the true densities are reasonably smooth, the dominant error results from estimation of the derivatives $\tilde{d}_j(z)$, which also involves smoothness assumptions and how they are used in this estimation. The difficulties are most serious when, as in our example and seems likely to be common, the main interest is in $E(x \mid z)$ for large values of z where the data for estimating the required derivatives are sparse. In practice, therefore, the most delicate matter in the proposed method seems to be to achieve suitable 'nonparametric' estimation of the derivatives $\tilde{d}_1(z), \tilde{d}_2(z)$, while imposing enough smoothness on $p_Z(\bullet)$ to cope with limitations of the data. These issues might often require more consideration than seems called for in our example, where we are quite confident in our final results.

A general aim of this paper is to encourage nonparametric use of the apparent distribution of the covariable x in computation of $E(x \mid z)$, whether by the methods of this paper or otherwise. Such a nonparametric approach is in principle rather daunting because of the deconvolution required for direct estimation of $p(x)$. Given the lognormal assumption regarding $p(z \mid x)$, the method suggested here seems likely to be reasonably accurate in relation to other modelling uncertainties provided that $p(x)$ is quite smooth and there are a few hundred observations on z in the range of primary interest for $E(x \mid z)$. When difficulties are encountered because of limitations to the extent of the data, these surely must, regardless of the specific approach, be met by stronger modelling assumptions regarding $p(x)$. It seems to us that this will often be more suitably done in terms of imposing smoothness rather than choosing some standard parametric form. Imposing this smoothness in terms of the log density derivatives considered focuses on the aspects of $p(x)$ that determine $E(x \mid z)$. If considerable smoothness in this sense is imposed, then it seems that the error in approximations (1) and (2) will be modest, and the issue becomes primarily how then

to estimate the derivatives $\tilde{d}_j(z)$. Improvements on our suggestion for this would likely be useful.

For the atomic-bomb survivor application the primary value of careful consideration of covariate errors is to see that their effect on final regressions, such as for radiation-related cancer, is actually quite modest. As reported in Pierce et al. (1990), for the magnitude of errors ordinarily considered such risk estimates are typically increased by around 10% in comparison to the naïve analysis that ignores dose errors. Detailed assumptions in error modeling appear not to alter this end result appreciably.

ACKNOWLEDGEMENT

This publication is based on research carried out partly at the Radiation Effects Research Foundation, Hiroshima and Nagasaki. This is a private nonprofit foundation funded equally by the Japanese Ministry of Health, Labor and Welfare, and the U.S. Department of Energy through the National Academy of Sciences. The work of the second author was supported by a contract with the European Union for radiation effects research. The authors thank Dan Schafer for useful discussions about modelling Berkson error.

REFERENCES

Armstrong, B. G. & Oakes, D. (1982). Effects of approximation in exposure assessments on estimates of exposure response relationships. *Scand. J. Work Envir. Health* **8**, Suppl. 1, 20-3.

Armstrong, B. (1985). Measurement error in the generalised linear model. *Commun. Statist.* B **14**, 529-44.

Barndorff-Nielsen, O. E. & Cox, D. R. (1989). *Asymptotic Techniques for Use in Statistics*. London: Chapman and Hall.

Carroll, R. J., Ruppert, D. & Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*. London: Chapman and Hall.

Cheng, C-L. & Van Ness, J. W. (1999) *Statistical Regression with Measurement Error*. London: Arnold.

Cook, J. R. & Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *J. Am. Statist. Assoc.* **89**, 1314-28.

Fuller, W. A. (1987). *Measurement Error Models*. New York: Wiley.

Nakamura, T. (1990). Corrected score function for errors-in-variables models: Methodology and applications to generalised linear models. *Biometrika* **77**, 127-37.

Nakamura, T. (1992). Proportional hazards model with covariates subject to measurement error. *Biometrics* **48**, 829-38.

Pierce, D. A., Stram, D. O. & Vaeth, M. (1990). Allowing for random errors in radiation dose estimates for the atomic bomb survivor data. *Radiat. Res.* **123**, 275-84.

Pierce, D. A., Stram, D. O., Vaeth, M. & Schafer, D. W. (1992). The errors-in-variables problem: Considerations provided by radiation dose-response analysis of the A-bomb survivor data. *J. Am. Statist. Assoc.* **87**, 351-9.

Pierce, D. A., Shimizu, Y., Preston, D. L., Vaeth, M. & Mabuchi, K. (1996). Studies of the mortality of atomic-bomb survivors. Report 12, Part 1. Cancer: 1950-1990. *Radiat. Res.* **146**, 1-27.

Prentice, R. L. (1982a). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* **69**, 331-42.

Prentice, R. L. (1982b). Covariate measurement errors in the analysis of cohort and case-control studies. In *Survival Analysis*, Ed. J. Crowley & R. A. Johnson, *IMS Lecture Notes Monograph Series, Vol. 2, pp. 137-51*. Hayward, CA: Inst. Math. Statist.

Reeves, G. K., Cox, D. R., Darby, S. C. & Whitley, E. (1998). Some aspects of measurement error in explanatory variables for continuous and binary regression models. *Statist. Med.* **17**, 2157-77.

Roesch, W. C. (Ed.) (1987). *U.S.—Japan Joint Reassessment of Atomic Bomb Radiation Dosimetry in Hiroshima and Nagasaki*. Hiroshima: Radiation Effects Research Foundation.

Schafer, D. W. (2001). Semiparametric maximum likelihood for measurement error model regression. *Biometrics* **57**, 53-61.

Stefanski, L. A. & Cook, J. R. (1995). Simulation-extrapolation: The measurement error jackknife. *J. Am. Statist. Assoc.* **90**, 1247-56.

Stefanski, L. A. (1989). Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. *Commun. Statist.* A **18**, 4335-58.

Stefanski, L. A. & Carroll, R. J. (1987). Conditional scores and optimal scores for generalised linear measurement-error models. *Biometrika* **74**, 703-16.

Tierney, L. & Kadane, J. B. (1986). Accurate approximation for posterior moments and marginal densities. *J. Am. Statist. Assoc.* **81**, 82-6.

Tosteson, T. D. & Tsiatis, A. A. (1988). The asymptotic efficiency of score tests in a generalised linear model with surrogate covariates. *Biometrika* **75**, 507-14.