COVARIATE MEASUREMENT ERRORS IN THE ANALYSIS OF COHORT AND CASE-CONTROL STUDIES

Ross Prentice

Fred Hutchinson Cancer Research Center and Department of Biostatistics, University of Washington

1. Introduction

This paper discusses the analysis of 'failure' time data, when predictor variables are subject to measurement error. The author's symposium presentation concentrated on a partial likelihood approach to relative risk estimation when covariates are subject to measurement error; material that mostly will appear in Prentice (1982). To avoid undue repetition the presentation here will emphasize full likelihood and marginal likelihood approaches to this problem. The accommodation of covariate measurement errors in the context of case-control sampling will also be briefly considered.

In failure time studies, as well as in many other areas of application, covariate values are subject to measurement errors. Particular applications that motivated this work include a study of the relationship between radiation exposure level and cancer mortality in atomic bomb survivors and a study of cardiovascular disease risk factors in a large cohort study. In the former study, one is interested in cancer mortality dose-response effects corresponding to individual gamma and neutron exposures. These exposure level estimates were, however, imputed from distance (from the presumed hypocenter) and shielding information obtained by interview. Such estimates may differ sub-

stantially from the 'true' exposure levels; in fact, the quality of the dosimetry data has recently been the subject of much controversy (e.g., Marshall, 1981). In the cardiovascular disease study, data on covariates such as blood pressure, serum cholesterol level and leukocyte counts were obtained in biennial clinic visits, taking place over a 20-year period. Each of these measured predictor variables is subject to considerable variation partially due to limitations of the measuring process but primarily because a large number of additional factors influence the measured values. For example, one may be interested in the relationship between some intrinsic blood pressure level and coronary heart disease incidence, but the measured blood pressure may be a rather imprecise approximation thereto, since it depends so heavily on the person's recent activities, state of relaxation, and current position in the diurnal cycle, to name a few factors. In many studies it will be possible to make some reasonable specifications of the error distributions associated with covariate measurements. Whether or not there is much basis for such specification, the sensitivity of results to various error distribution assumptions will be of interest.

2. Induced Models and Parametric Estimation

Consider a failure time random variable $T \geq 0$ and, for the moment, a fixed covariate $z = (z_1,\ldots,z_p)$. Throughout f will be used generically to denote probability, or probability density, function, so that $f(t|z)$ denotes the conditional density for T given z. Characteristics, such as relative risk parameters, used in the specification of $f(t|z)$ will usually be the primary target of estimation. Now suppose that, rather than z, one observes only the 'measured' covariate $x = (x_1,\ldots,x_q)$. Usually there will be a one-to-one correspondence between components of x and z, and $p = q$, but this is not required in the discussion that follows. In considering error distribution assumptions it is natural to think of a specification of the distribution of x given z, along with a marginal distribution for z. As will be seen below, however, it is only necessary to specify the conditional probability distribution, $f(z|x)$, for z

given x, rather than their joint distribution, in order to proceed with estimation of $f(t|z)$.

We will require a conditional independence between T and x, given z; that is,

$$f(t|z,x) = f(t|z) \quad . \tag{1}$$

This condition is a statement that the measured covariate has no prognostic value if the true covariate is known. If (1) does not hold, x is not simply an 'estimator' of z and direct modelling of $f(t|z,x)$ is indicated.

The induced probability function for T given the measured covariate x is readily derived as the expectation over the distribution of z given x of $f(t|z,x)$, which under (1) can be written

$$f(t|x) = E_x\{f(t|z)\} \quad . \tag{2}$$

If the error distribution $f(z|x)$ is completely specified, this induced model $f(t|x)$ will involve only the parameters of $f(t|z)$. It is then of interest to identify failure time and error distribution models that lead to tractable induced models for failure time, given the observable covariate. Such induced models can then be applied to failure time data in order to carry out inferences on parameters of interest.

In order to develop mathematically convenient induced models (2), it is natural to consider normally distributed failure time and error random variables. Suppose $Y = \log T$ satisfies the normal linear regression model

$$Y = \log T = \alpha + z\beta + \sigma V \quad ,$$

where α, $\sigma > 0$ and $\beta(p\times 1)$ are real parameters and V is a standard normal random

variable. Also, suppose that the true covariate distribution, given the corresponding measurement x, is normal with mean vector $(1\times p)$ μ_x and variance matrix $\sum_x$. The induced model (2) for $Y = \log T$ given x is then readily shown to be normal with mean vector $\alpha + \mu_x\beta$ and variance $\sigma^2 + \beta' \sum_x \beta$, where β' denotes the vector transpose of β. This simple result may provide an adequate basis for exploring the implications of covariate measurement errors on regression testing and estimation, in a variety of failure time and non-failure time applications. Specifically, an iterative maximum likelihood procedure could be readily implemented for β estimation, that would not be unduly complicated by the presence of right censorship.

Other distributional assumptions may also yield explicit induced models. For example, a Weibull regression model with 'linear' hazard ratio $(1+z\beta) \geq 0$, can be written

$$\lambda(t|z) = \lambda p(\lambda t)^{p-1} (1+z\beta) \quad ,$$

where λ denotes the hazard, or instantaneous failure rate function, and λ, $\sigma > 0$ and $\beta(p\times 1)$ are parameters. A normal distribution for z given x yields, after some algebra, an induced hazard function

$$\lambda(t|x) = \lambda p(\lambda t)^{p-1} [1 + \{\mu_x - (\lambda t)^p \beta' \sum_x\} \beta] \quad .$$

In fact, some bounds on the support for z given x will be required in order that $1+z\beta \geq 0$ not be violated. A normal model for z given x, and the above induced model $\lambda(t|x)$ should, however, provide adequate approximations if β, μ_x and $\sum_x$ are such that $1+z\beta \geq 0$ with probability close to one at each x. Note that the hazard ratio corresponding to any pair of x-values is no longer constant, but rather converges monotonically to unity as $t \to \infty$. Computational methods for fitting this induced model could be derived, though the model is perhaps too complicated to expect much use.

The fitting of such models to failure time data will, as usual, require an independent censorship assumption in order that (2) be identifiable. Such an assumption can be written

$$\lambda\{t|x, \text{ no censorship in } [0,t)\} = \lambda(t|x) \quad . \tag{3}$$

In some problems an independent censorship assumption applied to t given z, rather than t given x, would be more appropriate. In such circumstances censoring will typically be mildly dependent and (2) will not strictly be identifiable. This seems unlikely to be a practical problem, however, unless covariate errors are very substantial and censorship depends heavily on z.

In order to use standard likelihood expression one will also require the independence of failure times given the corresponding measured covariate x_i, $i = 1,\ldots,n$. Such independence will follow, for example, if z_i, $i = 1,\ldots,n$ can be viewed as i.i.d. from some distribution and both t_i given z_i and x_i given z_i are independent for $i = 1,\ldots,n$.

It seems appropriate to make some comment on the specification of the probability function $f(z|x)$. For example, in order to specify the mean μ_x and variance matrix $\sum_x$ in the above normal densities, one might suppose that the basic regression vector z can be viewed as normally distributed with mean μ and variance $\sum$ and that the measured covariate x arises via $x = z+w$, where w is normal with mean zero and variance matrix C. If z and w are independent the density for z given x is then normal with mean $\mu_x = \mu + \sum\left(\sum + C\right)^{-1}(x-\mu)$ and variance matrix $\sum_x = \sum - \sum\left(\sum + C\right)^{-1}\sum$, the latter of which is independent of x. In the normal regression model described above (for $Y = \log T$) the induced regression equation in x will then have regression coefficient $\sum\left(\sum + C\right)^{-1}\beta$. Ignoring covariate measurement, errors would then give rise to coefficient estimates that are systematically too close to zero in simple linear regression and that are 'deflated' by the matrix $\sum\left(\sum + C\right)^{-1}$ in the multiple regression problem. More generally, z and w may be allowed to be correlated. A normal

distribution for z given x is readily derived from any joint normal distribution for z and w. In the very special case, sometimes referred to as the Berkson model, z and w have a joint normal distribution as above except that the covariance of z and w is -C. It follows that $\mu_x = x$ and $\sum_x = C$. It is worth noting again that specification of the joint distribution of z and x is unnecessary, since only the distribution of z given x appears in (2). Joint normal distributions, of the type just described, may then be used as a guide toward the specification of μ_x and $\sum_x$ in a normal model for z given x, but do not need to be explicitly assumed. The reader is referred to the review paper, Cochran (1968), for further comments on error distribution specification and on the effects of measurement errors in the ordinary regression model.

3. Cox Model Estimation with Covariate Errors

The models described above, particularly the induced log-normal model for T given x, provide the basis for a parametric approach to accommodating measurement errors in failure time analyses. The partially parametric regression model of Cox (1972) is an attractive alternative to failure time analyses. Desirable features include the ability to interpret the regression parameter in terms of relative risk, substantial model flexibility, and the availability of many important generalizations, as summarized in Kalbfleisch and Prentice (1980). In its most general form the method gives a computationally feasible method of exploring the dependence of the relative risk function on covariates and follow-up time, without placing any model restrictions, except the presumed parametric form for the relative risk function.

The special case of the Cox model in which the relative risk is independent of t can be written

$$\lambda(t|z) = \lambda_0(t)\, g(z\beta) \quad , \tag{4}$$

where $\lambda_0(\cdot) \geq 0$ is unrestricted, $g(\cdot) \geq 0$ is a specified function standardized so

that $g(0) = 1$ and $\beta(p\times 1)$ is a regression parameter to be estimated. Note that $g(z\beta) = \lambda(t|z)/\lambda(t|z=0)$ is the risk associated with regression vector z, relative to that at $z = 0$. Usually the relative risk function has been defined by $g(u) = \exp(u)$, though other forms such as $g(u) = 1 + u$ have also sometimes been used.

Various approaches have been considered for the estimation of β in (4); most notably, the partial likelihood approach of Cox (1972,1975). Kalbfleisch and Prentice (1973) utilized a marginal likelihood approach that was based on the distribution of failure time ranks.

In the presence of covariate errors the model induced from (4) via (2) has the rather complicated corresponding hazard function

$$(5) \qquad \lambda(t|x) =$$

$$\lambda_0(t)\left[\int g(z\beta)\, \exp\{-g(z\beta) \int_0^t \lambda_0(u)du\}\, f(z|x)dx \Big/ \int \exp\{-g(z\beta) \int_0^t \lambda_0(u)du\}\, f(z|x)dz\right],$$

where the integrals (or sums) are over the range of z, given x. In the special case $g(z\beta) = 1+z\beta$, and z given x normal with mean μ_x and variance matrix $\sum_x$, (5) simplifies to

$$\lambda(t|x) = \lambda_0(t)\left[1 + \{\mu_x - \int_0^t \lambda_0(u)du\ \beta'\ \textstyle\sum_x\}\beta\right] \quad ,$$

generalizing the Weibull regression result given above. In spite of the complexity of (5), the induced class of models retains the property of functional invariance under monotone-increasing differentiable transformations on t. One can show, as in Kalbfleisch and Prentice (1973), that the distribution of the failure time ranks does not involve the baseline hazard function $\lambda_0(\cdot)$. In fact, the failure time rank vector is marginally sufficient for β in the sense described by Kalbfleisch and Prentice. Assuming the expectation operators and order statistic integrals in the generalized rank vector probability can be

interchanged, the marginal likelihood for β in (5) can be written

$$L(\beta|X) = E_X L(\beta|Z) \quad , \tag{6}$$

where X has been written for the set of measured covariate vectors over the sample; that is, $X = \{x_1,\ldots,x_n\}$, Z has been written for $\{z_1,\ldots,z_n\}$, the expectation is over the distribution of $Z = \{z_1,\ldots,z_n\}$ given $X = \{x_1,\ldots,x_n\}$ and $L(\beta|Z)$ is the marginal likelihood that would arise if the true covariate vectors Z, rather than only X, were observed. Specifically,

$$L(\beta|Z) = \prod_{i=1}^{k} \left[\prod_{\ell \varepsilon F(t_i)} g(z_\ell \beta) \Big/ \left\{ \sum_{\ell \varepsilon R(t_i)} g(z_\ell \beta) \right\}^{m_i} \right] \quad , \tag{7}$$

where $t_1,\ldots,t_k$ represent the distinct (uncensored) failure times in the sample, $F(t_i)$ is the set of $m_i \geq 1$ study subjects that fail at t_i and $R(t_i)$ is the risk set just prior to time t_i. Note that the denominator of (7) involves an approximation (Breslow, 1974) to accommodate any tied failure times. The score statistic from (6) is

$$\partial \log L(\beta|X)/\partial\beta = L(\beta|X)^{-1} E_X\{L(\beta|Z)\ \partial \log L(\beta|Z)/\partial\beta\} \quad , \tag{8}$$

a weighted average of the score statistics corresponding to possible values of Z, given X. Similarly, the observed information matrix can be written

$$-\partial^2 \log L(\beta|X)/\partial\beta^2 = \tag{9}$$

$$L(\beta|X)^{-1} E_X[L(\beta|Z)\ \{-\partial^2 \log L(\beta|Z)/\partial\beta^2 - \partial \log L(\beta|Z)/\partial\beta'\ \partial \log L(\beta|Z)/\partial\beta]$$

$$+ \{\partial \log L(\beta|X)/\partial\beta'\}\{\partial \log L(\beta|X)/\partial\beta\} \quad .$$

At $\beta = 0$, $L(\beta|Z)$ is independent of Z and (8) simplifies to

$$(10) \qquad g'(0) \sum_{i=1}^{k} \left\{ \sum_{\ell \in F(t_i)} E_{x_\ell}(z_\ell) - m_i n_i^{-1} \sum_{\ell \in R(t_i)} E_{x_\ell}(z_\ell) \right\} ,$$

where n_i is the number of subjects in $R(t_i)$. It follows that it is only necessary to specify the expectations of each z-value given the corresponding x-value in order to carry out a score test for $\beta = 0$, a point that was made somewhat more generally, in the context of partial likelihood, in Prentice (1982).

In order to use (6) for general inference on the regression parameter β one needs to contend with a complicated expectation. The possibility of developing useful analytic expressions for (6) seems remote, even if mathematically convenient choices for g and $f(z|x)$ are entertained. An approximate estimation procedure, based on Monte Carlo sampling is suggested by (8) and (9). In particular, suppose that sets of regression vectors $Z_1,\ldots,Z_s$ are sampled from the joint distributions of Z given X. The score statistic (8) is then estimated by

$$\tilde{v} = \sum_{j=1}^{s} L(\beta|Z_j)\, \partial \log L(\beta|Z_j)/\partial\beta \Big/ \sum_{j=1}^{s} L(\beta|Z_j)$$

while the corresponding observed information matrix is estimated by

$$\sum_{j=1}^{s} L(\beta|Z_j) \{-\partial^2 \log L(\beta|Z_j) - \partial \log L(\beta|Z_j)/\partial\beta'\, \partial \log L(\beta|Z_j)/\partial\beta\} \Big/ \sum_{j=1}^{s} L(\beta|Z_j) + \tilde{v}'\tilde{v} .$$

Existing computer software could then be readily adapted to carry out a Newton-Raphson maximization for β. This idea amounts simply to approximating (6) by

$$s^{-1} \sum_{j=1}^{s} L(\beta|Z_j) \quad .$$

As such, the approximation can be made to be as close as desired by making s large. This is a computation-intensive approach to regression estimation. It is, however, very flexible in terms of both the relative risk function, g , and error distribution $f(z|x)$. In fact, it is not even necessary that z-values on distinct study subjects be independent, given the corresponding measured x-values. It is hoped to pursue this idea in more detail elsewhere.

A nonparametric maximum likelihood approach to estimation in (4) would lead in the presence of measurement errors to a likelihood function for β that can again be written

$$\tilde{L}(\beta|X) = E_X \tilde{L}(\beta|Z) \quad ,$$

where $\tilde{L}(\beta|Z)$ is the likelihood function, given the 'true' covariate values Z, after maximizing out the baseline hazard function. The approximate likelihood of Breslow (1974) would lead once again to (6), in the presence of covariate errors.

Prentice (1982) considered a partial likelihood approach to this problem. A partial likelihood function for β, given the measured covariate values, X , in the sample can be written

$$(11) \qquad \prod_{i=1}^{k} \left[\prod_{\ell \in F(t_i)} E_{(t_i, x_\ell)} g(z_\ell \beta) \Big/ \left\{ \sum_{\ell \in R(t_i)} E_{(t_i, x_\ell)} g(z_\ell \beta) \right\}^{m_i} \right] \quad ,$$

where a tied failure time approximation has again been made and the expectations in the i^{th} term of the product are conditional on both $T \geq t_i$ and the measured covariate values. The partial likelihood approach accommodates time-dependent covariates as may be defined to test or relax the proportional

hazards assumption in (4) or may be utilized to relate failure rate to some stochastic covariate process. The result (10) could equally well be derived from (11). Unfortunately, however, (11) does not provide an adequate answer to more general testing and estimation problems in many applications, since the expectations in (11) typically involve the baseline incidence function $\lambda_0(\cdot)$, as is evident from (5) upon noting that

$$\lambda(t|x) = \lambda_0(t)\, E_{(t,x)}\, g(z\beta) \quad .$$

The application that motivated Prentice (1982) was such that the dependence of $E_{(t,x)}\, g(z\beta)$ on the condition $T \geq t$, and hence the dependence of the expectation on $\lambda_0(\cdot)$, could be ignored. If such dependence cannot be ignored, it would be useful to consider iterative estimation procedures in which a trial value of β is used to produce an empirical estimate of the cumulative hazard function $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ that appears in (5), which in turn, is used to obtain an updated β-value on the basis of (11). Even in the simple special case $g(u) = 1+u$ with normally distributed covariate errors, a nonparametric maximum likelihood approach to estimating $\Lambda_0(t)$, at a specified β, is complicated. On the other hand, unless covariate errors are quite substantial, it would presumably be accurate enough to obtain an empirical estimate of $\Lambda_0(t)$, ignoring covariate measurement errors, and subsequently use this esimate in the partial likelihood function (11). Such a usage would be quite routine, for example, in the circumstances mentioned above in which z, given x, is normally distributed and the relative risk function g is of a linear form. Numerical evaluation of this proposal would be worthwhile.

4. Covariate Errors in Case-Control Studies

Suppose now that a Cox-type model (4) holds for the incidence (or mortality) rate for a disease. A case-control study involves selecting both diseased (cases) and disease-free (control) subjects and sampling their

corresponding covariate data z. Often z will include summarizations of certain exposure histories along with personal characteristics. In the presence of covariate measurement errors one will sample the measured covariate x, rather than z.

To be specific, consider the type of case-control study described in Prentice and Breslow (1978), in which for each case a set of time (age) matched controls are selected. The hazard function induced from (4) can, in general, be written

$$\lambda(t|x) = \lambda_0(t)\, E_{(t,x)}\, g(z\beta) \quad ; \tag{12}$$

that is, the induced relative risk of time t is the expectation of $g(z\beta)$, given the measured covariate x and given $T \geq t$. By the same argument used in Prentice and Breslow, a conditional likelihood for this relative risk function can be developed by conditioning on the set of exposure histories corresponding to each case and its matched controls. The conditional likelihood function can be written

$$L(\beta) = \prod_{i=1}^{k} E_{(t_i,x_i)}\, g(z_i\beta) \Big/ \sum_{\ell \in R_i} E_{(t_i,x_\ell)}\, g(z_\ell\beta) \quad , \tag{13}$$

where $t_1,\ldots,t_k$ denote the incidence times for the cases and R_i denotes the i^{th} case and its matched controls. As with the partial likelihood described previously, however, the induced relative risk function will depend to some extent on the baseline incidence function $\lambda_0(\cdot)$ due to the inclusion of $\{T \geq t_i\}$ in the conditioning event. If, however, the study disease is rare the distribution of z-values that correspond to a measured covariate x will be very similar among subjects without failure at some time t as was the case at $t = 0$. In this circumstance, the relative risk function will be well approximated by

(14) $$E_x\, g(z\beta)$$

and straightforward asymptotic likelihood procedures can be applied to (13) for β estimation. In the special case $g(u) = 1+u$ one then fits a relative risk model

(15) $$1 + E_x(z|x)$$

to the case-control data using (13). As a practical approach to accommodating covariate measurement errors in the estimation of exposure-response relationships Armstrong and Oakes (1982) have suggested replacing z-values by corresponding $E_x(z|x)$ values, and carrying out standard analyses. With a linear relative risk function, their proposal is supported by the development given here provided the condition $\{T \geq t\}$ can be ignored in the induced relative risk function. Estimation with a multiplicative relative risk function $g(u) = \exp(u)$ can be readily carried out with error probability functions $f(z|x)$ that have simple moment generating functions. For example, the normal probability function for z given x mentioned above, gives for (14)

$$\exp\{\mu_x\beta + \tfrac{1}{2}\beta' \textstyle\sum_x \beta\} \quad ,$$

which may be inserted into (13) for estimation of β.

Similar results could be developed for more general case-control study designs and, for example, logistic disease incidence models.

5. Concluding Remarks

Failure to acknowledge covariate measurement errors in some regression problems may lead to results that lack a useful interpretation or that are misleading. Greater effort seems warranted in respect to methods to estimate covariate error distribution properties and to utilize information on covariate

error distributions toward the estimation of key regression parameters. This paper described several approaches to the latter problem in a failure time regression context. Clearly the surface has merely been scratched on this important statistical topic.

ACKNOWLEDGEMENTS

This work was supported by grants GM-28314 and GM-24472 from the National Institutes of Health.

REFERENCES

Armstrong, B.G. and Oakes, D. (1982). Effects of approximation in exposure assessments on estimates of exposure-response relationships. To appear, Scandinavian Journal of Work, Environment and Health.

Breslow, N.E. (1974). Covariance analysis of censored survival data. Biometrics 30, 89-99.

Cochran, W.G. (1968). Errors of measurement in statistics. Technometrics 10, 637-666.

Cox, D.R. (1972). Regression models and life tables (with discussion). Journal of the Royal Statistical Society B, 34, 187-220.

Cox, D.R. (1975). Partial likelihood. Biometrika 62, 269-276.

Kalbfleisch, J.D. and Prentice, R.L. (1973). Marginal likelihoods based on Cox's regression and life model. Biometrika 60, 267-278.

Kalbfleisch, J.D. and Prentice, R.L. (1980). The Statistical Analysis of of Failure Time Data. New York, Wiley.

Marshall, E. (1981). New A-bomb studies alter radiation estimates. Science 212, 900-903.

Prentice, R.L. (1982). Covariate measurement errors and parameter estimation in Cox's failure time regression model. To appear, Biometrika.

Prentice, R.L. and Breslow, N.E. (1978). Retrospective studies and failure time models. Biometrika 65, 153-158.