A PARADOX:
MORE MEASUREMENT ERROR CAN LEAD TO MORE EFFICIENT ESTIMATES

Eugene Demidenko
Dartmouth Medical School
HB 7927, Hanover, NH 03755-3861, USA

Donna Spiegelman
Harvard School of Public Health
677 Huntington Avenue, Boston, MA 02115, USA

Published in

Cimmunications in Statistics. Theory and Methods
26(7), 1649-1675 (1997)

Key Words: Errors-in-variables; Information matrix; Berkson model; Maximum likelihood; Logistic regression.

ABSTRACT

We investigated the conditions under which the information matrix, conditional on covariates, and the unconditional version, integrated over marginal distribution of covariates, are increasing functions of the measurement error variance, and when the conditional and unconditional asymptotic variances of the maximum likelihood estimate are decreasing functions of the measurement error variance. We say that a paradox occurs when one can decrease the variance of the maximum likelihood estimate by increasing measurement error. Two covariate measurement error models were considered: the Berkson and the classical measurement error models, with continuous and binary

dependent variables. The measurement error variance was assumed known. We found that in the linear model with Berkson covariate measurement error, the paradox can occur when the model variance is known and quite large. With binary data and the Berkson covariate measurement error, the paradox is likely to occur for rare events. The paradox based upon the unconditional variance, calculated as the inverse of the unconditional information matrix, does not occur with the classical measurement error model. However, it does occur for the conditional variance, as calculated by the inverse of the observed information matrix, in some circumstances. Empirical evidence for this is illustrated using data from the Nurses' Health Study.

1. INTRODUCTION

It is commonly believed that measurement error decreases the efficiency of estimates. In particular, when measurement error is present, one expects that the asymptotic variance of the maximum likelihood estimate would increase with increasing the measurement error variance. The goal of this paper to show that this is not true for some measurement error models.

Let y be a dependent variable and x a covariate measured with error. The conditional distribution of y given x is known up to an unknown parameter vector β; i.e. $y \mid x \sim f(y|x;\beta)$. The observations on x are not available, instead there exist a surrogate variable z, which represents measured values of x with some error variance σ^2. Let us assume that the conditional distribution of x given z, $f(x|z;\theta)$ and the marginal distribution of z, $f(z;\gamma)$, are known up to an unknown parameter vector (θ,γ). A further assumption is made: $f(y \mid x,z;\beta,\theta) = f(y \mid x;\beta)$; or in other words, given x, the primary model is conditionally independent of measurement error. Under these assumptions, the distribution for available observations (y,z) can be written as

$$f(y,z;\beta,\theta,\gamma) = f(z;\gamma)f(y \mid z;\beta,\theta) = f(z;\gamma)\int f(y|x,z;\beta,\theta)f(x|z;\theta)dx$$

$$= f(z;\gamma)\int f(y|x;\beta)f(x|z;\theta)dx.$$

Therefore, the log-likelihood for available observations is

$$l(y;z;\beta;\theta;\gamma) = \log\int f(y|x;\beta)f(x|z;\theta)dx + \log f(z;\gamma)$$

where β is the parameter of interest. Two important examples of this general measurement error scheme are considered below.

1.1. The Berkson measurement error model

Here we assume that $x\,|\,z \sim N(z;\sigma^2)$; $\quad z \sim N(\mu_z;\sigma_z^2)$; where μ_z and σ_z^2 are unknown. In this paper, the error variance σ^2 is assumed to be known. Then,

$$f(y|z;\beta) = \frac{1}{\sqrt{2\pi\sigma^2}}\int f(y|x;\beta)e^{-\frac{1}{2\sigma^2}(x-z)^2}dx.$$

1.2. The classical measurement error model

In this case, the roles of variable z and x swap, or more precisely $z\,|\,x \sim N(x;\sigma^2)$; $\;x \sim N(\mu_x;\sigma_x^2)$ where μ_x and σ_x^2 are unknown. The conditional distribution $x\,|\,z \sim N(\mu_{x|z};\sigma_{x|z}^2)$ where $z \sim N(\mu_x;\sigma^2+\sigma_x^2)$ and

$$\mu_{x|z} = \mu_x + \frac{\sigma_x^2}{\sigma_x^2+\sigma^2}(z-\mu_x);\quad \sigma_{x|z}^2 = \frac{\sigma_x^2\sigma^2}{\sigma_x^2+\sigma^2} \tag{1.1}$$

Then,

$$f(y|z;\beta;\mu_{x|z|};\sigma_{x|z}^2) = \frac{1}{\sqrt{2\pi\sigma_{x|z}^2}}\int f(y|x;\beta)e^{-\frac{1}{2\sigma_{x|z}^2}(x-\mu_{x|z|})^2}dx.$$

In classical and Berkson measurement error models the observed and true variables interchange their orders: in Berkson case $x = z+\delta$, and in classical case $z = x+\delta$. The order has a critical impact on the properties of estimates. For instance, the OLS-estimator is consistent and efficient in Berkson case and inconsistent in the classical one. A detailed discussion of the two measurement models can be found in Fuller (1987). Another aspect of this difference, expressed in terms of influence of the level measurement error on the quality of estimates, will be demonstrated in the present paper.

Two cases will be considered in this paper: when y is a continuous normal variable and when y is a binary random variable.

1.3. Definitions

Let $\tau = (\beta; \theta)$ denote the combined vector of unknown parameters where β may be interpreted as a parameter of interest and θ as a nuisance parameter. Recall, we have a pair of observations: y and z: By convenience, the conditional expectation on z is taken with respect to conditional distribution of $y \mid z$; the unconditional expectation is taken with respect to the pair $(y; z)$: In this paper the term "variance" means the asymptotic variance of the MLE and consequently is equal to the inverse of the information matrix.

Four quantities associated with the efficiency of the maximum likelihood estimate, as a function of the error variance, are considered:

1. Conditional information:
$$I_\tau(\sigma^2|z) = -E\left(\frac{\partial^2 l}{\partial \tau^2}\bigg| z\right): \tag{1.2}$$
This is a commonly used information matrix for the variance assessment in practice (see item 4). It is simple and reliable because it does not require taking assumption on distribution of z. However, it may be misleading, in particular, it may lead to a paradox when in fact it does not take place for more accurate unconditional information (see the next item).

2. Unconditional information:
$$I_\tau(\sigma^2) = -E\left(\frac{\partial^2 l}{\partial \tau^2}\right) = -E_z E\left(\frac{\partial^2 l}{\partial \tau^2}\bigg| z\right): \tag{1.3}$$
It is most accurate information matrix calculation and requires assumption on distribution of z: However, one rare uses it in practice because it involves integration.

3. Conditional variance is calculated as diagonal element of the inversed estimated information matrix, namely
$$\mathrm{var}_l(\sigma^2 \mid z_1; \ldots; z_n) = -\left(\frac{1}{n}\sum_{i=1} E\left(\frac{\partial^2 l}{\partial \tau^2}\bigg| z_i\right)\right)^{-1}_{ll};$$

where τ_l is the lth element of vector parameter τ. Apparently, the inversed matrix is an estimate of the conditional information matrix (1.2). Particularly, we are interested in the variances of the elements of β. It should be noted that the conditional variance, based on the observed information matrix, can only be investigated with respect to given data $\{z_i, i = 1, \ldots, n\}$ because matrix $E\left(\frac{\partial^2 l}{\partial \tau^2}\Big| z\right)$ is singular if τ is a vector.

4. Unconditional variance is calculated as diagonal element of the inversed information matrix, namely

$$\mathrm{var}_l(\sigma^2) = (I_\tau^{-1}(\sigma^2))_{ll} \tag{1.4}$$

where $I_\tau(\sigma^2)$ is defined in (1.3).

We would expect that the information should decrease and the variance increase as the error variance σ^2 increases. Comparing two symmetric matrices A and B we write $A \leq B$ if matrix $B - A$ is a nonnegative definite matrix. The conditional information paradox takes place if (1.2) is not a decreasing function for some values of z, σ^2 and τ. Since σ^2 is scalar we can investigate paradoxical situations via matrix derivatives based on the following fact. Let $A = A(s)$ be a symmetric matrix as a function of scalar s. Then, if dA/ds is a nonnegative definite matrix for all s, then $A(s)$ is an increasing function of s, i.e., $s_1 \leq s_2$ implies $A(s_1) \leq A(s_2)$. Vice versa, if there exists s and a vector $q = q(s)$ such that $q'(dA/ds)q < 0$ then $A(s)$ is not an increasing function of s. Based on this fact, we infer that the unconditional paradox occurs when there exists a vector q, which may depend on τ, σ^2 and z, such that $q'(dI_\tau(\sigma^2 | z)/d\sigma^2)\, q < 0$. The unconditional information paradox takes place if (1.3) is not a decreasing function for some σ^2 and θ. This occurs when there exists a vector q which may be dependent on τ, σ^2, such that $q'(dI_\tau(\sigma^2)/d\sigma^2)\, q < 0$. Variance paradox takes place if (1.4) is not an increasing function for some σ^2 and τ. At last, we say that partial information paradox takes place if it takes place assuming that part of parameters is fixed, i.e., paradox takes place for a main submatrix of the information matrix. Similarly, we speak about partial variance paradox, i.e., when part of parameters are fixed. Clearly, if, for instance, the unconditional information matrix is a decreasing function of σ^2 then all main submatrices, including diagonal elements, are decreasing functions of σ^2, but not vice versa.

Paradoxes based on quantities (1.2,1.3,1.4) are not equivalent: for instance, the unconditional variance paradox can take place and the unconditional information paradox may not, because if diagonal elements of matrix $B - A$ are positive it does not imply matrix $B - A$ is nonnegative definite.

Remarks:

1. If matrix $dI_\tau(\sigma^2)/d\sigma^2$ is nonnegative definite, then $\mathrm{var}_l(\sigma^2)$ is a decreasing function of σ^2 for all l and paradox occurs. If the conditional information paradox takes place for a specific z^* then it is likely that the conditional variance paradox occurs, if data $\{z_i; i = 1;...;n\}$ are concentrated around z^*. In fact, let us assume that for some z^* and τ matrix $I_\tau(\sigma^2|z)$ is an increasing function of σ^2; i.e. matrix $dI_\tau(\sigma^2|z^*)/d\sigma^2$ is positive definite. Then
$$\frac{d\mathrm{cov}(\sigma^2 \mid z_1;...;z_n)}{d\sigma^2} = \left(\frac{1}{n}\sum_{i=1} E\left(\frac{\partial^2 l}{\partial\tau^2}\Bigg| z_i\right)\right)^{-1}$$
$$\times \left(\frac{1}{n}\sum_{i=1}\frac{dE\left(\frac{\partial^2 l}{\partial\tau^2}\Big| z_i\right)}{d\sigma^2}\right)\left(\frac{1}{n}\sum_{i=1} E\left(\frac{\partial^2 l}{\partial\tau^2}\Bigg| z_i\right)\right)^{-1}:$$
Since z_i are close to z^* we can expect that the matrix in the middle is positive definite, and consequently the derivative of the covariance matrix is negative definite — the conditional variance paradox occurs.

2. If the conditional information paradox does not hold, then the unconditional information paradox and the asymptotic variance paradox do not hold.

3. If the conditional information paradox occurs in the Berkson case, then it is likely that the unconditional information paradox takes place. More precisely, if for some $\tau;\sigma^2;z_*$ the derivative matrix $dI_\tau(\sigma^2 \mid z = z_*)/d\sigma^2$

is not nonnegative definite, or in other words, there exists a vector q that is not dependent on z_* such that $q'(dI_\tau(\sigma^2 \mid z = z_*)/d\sigma^2)\, q < 0$; then for $\mu_z = z_*$ and $\sigma_z^2 \to 0$, we have

$$q' \left(dI_\tau(\sigma^2)/d\sigma^2 \right) q = \lim_{\sigma_z^2 \to 0} q' \left(dI_\tau(\sigma^2 \mid z = z_*)/d\sigma^2 \right) q < 0$$

and the unconditional information paradox occurs. Also, if conditionally on z, matrix $dI_\tau(\sigma^2 \mid z = z_*)/d\sigma^2$ is negative definite, then the variance paradox takes place.

4. One should distinguish between the conditional and unconditional information. Even when the conditional information paradox occurs but the unconditional one does not, it is possible that the conditional information matrix $\sum I_\tau(\sigma^2 | z_i)$ is an increasing and $\mathrm{cov}(\hat{\beta}) = (\sum I_\tau(\sigma^2 | z_i))^{-1}$ a decreasing function of σ^2. The latter is the conditional covariance matrix of the maximum likelihood estimate.

The goal of the present article is to find conditions when the conditional or unconditional information/variance paradox occurs for different measurement error models. The linear model is investigated in sections 3 and 4, the binary model - in sections 4 and 5.

2. THE LINEAR MODEL WITH BERKSON MEASUREMENT ERROR

We start with the linear model

$$y_i = \beta_0 + \beta x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2) \tag{2.1}$$

where y_i is the dependent variable, β_0 and $\beta \neq 0$ are the parameters of interest, x_i is the unobserved true covariate, $i = 1, \ldots, n$; σ_ε^2 is the model variance. The Berkson model for measurement error is described in Section 1.1. The variance of measurement errors, σ^2 is assumed to be known and $(y_i, x_i, z_i, \varepsilon_i, \delta_i)$ are i.i.d., where $\delta_i = x_i - z_i$ is the measurement error. The conditional distribution of the observed data is

$$y_i \mid z_i \sim N(\beta_0 + \beta z_i; \sigma_\delta^2 + \sigma^2\beta^2). \qquad (2.2)$$

Our aim is to investigate the behavior of the asymptotic variance of the maximum likelihood estimate of β as function of σ^2. Since the information matrix is block-diagonal for $(\beta_0; \beta; \sigma_\delta^2)$ and $(\mu_z; \sigma_z^2)$ we can assume μ_z and σ_z^2 given. We show that the knowledge of $\beta_0; \beta; \sigma_\delta^2$ is the key.

2.1. The unconditional variance paradox does not occur when all parameters are unknown

The conditional information matrix for $\theta = (\beta_0; \beta; \sigma_\delta^2)$ in (2.2) can be written as

$$I_\theta(\sigma^2 \mid z) = K \begin{bmatrix} 1 & z & 0 \\ z & 2\sigma^4 K\beta^2 + z^2 & \sigma^2 K\beta \\ 0 & \sigma^2 K\beta & \frac{1}{2}K \end{bmatrix}$$

where $K = (\sigma_\delta^2 + \sigma^2\beta^2)^{-1}$. Note that the conditional information paradox for the (2,2)th element of this matrix occurs when

$$\frac{d(K(2\sigma^4 K\beta^2 + z^2))}{d\sigma^2} = -\left((\beta^2 z^2 - 4\sigma_\delta^2)\sigma^2 + \sigma_\delta^2 z^2\right)\beta^2 K^3 > 0; \qquad (2.3)$$

which particularly takes place if $|\beta z| < 2\sigma_\delta$. Following our definition in the previous section, we say that in this case the partial conditional information paradox for β takes place if β_0 and σ_δ^2 are known. Similarly, the partial unconditional information paradox for this element can occur when $|\beta|\sqrt{\sigma_z^2 + \mu_z^2} < 2\sigma_\delta$; because the derivative of the (2,2)th element of matrix

$$I_\theta(\sigma^2) = K \begin{bmatrix} 1 & \mu_z & 0 \\ \mu_z & 2\sigma^4 K\beta^2 + \mu_z^2 + \sigma_z^2 & \sigma^2 K\beta \\ 0 & \sigma^2 K\beta & \frac{1}{2}K \end{bmatrix}: \qquad (2.4)$$

is positive when this condition is met.

However, the unconditional variance paradox does not occur, because the unconditional variances of all parameters are increasing functions of σ^2:

$$\begin{aligned} \mathrm{var}_{\beta_0}(\sigma^2) &= (\sigma_\delta^2 + \sigma^2\beta^2)(1 + \mu_z^2/\sigma_z^2); \quad \mathrm{var}_\beta(\sigma^2) = (\sigma_\delta^2 + \sigma^2\beta^2)/\sigma_z^2 \\ \mathrm{var}_{\sigma_\delta^2}(\sigma^2) &= 4\beta^2\sigma_z^{-2}\sigma^4(\sigma_\delta^2 + \sigma^2\beta^2) + 2(\sigma_\delta^2 + \sigma^2\beta^2)^2. \end{aligned}$$

2.2. The unconditional variance paradox when parameter σ_ε^2 is known

The situation changes if the parameter σ_ε^2 is known. For example, it could be the case when the only errors in the true model (2.1) are associated with measurement errors with a known degree of inaccuracy reflected by σ_ε^2. This kind of model can be found in technical applications. Accordingly our definition, now we speak about partial paradox. The maximum likelihood estimate for β does not coincide with the ordinary least squares estimator anymore and is the solution to the following cubic equation:

$$-n\sigma^2\beta(\sigma_\varepsilon^2+\sigma^2\beta^2)+\sigma^2\sum(y_i^o-\beta z_i^o)^2\beta+(\sigma_\varepsilon^2+\sigma^2\beta^2)\sum(y_i^o-\beta z_i^o)z_i^o=0,$$

where $y_i^o = y_i - \bar{y}$, $z_i^o = z_i - \bar{z}$.

Now the unconditional information matrix has the form:

$$I_\tau(\sigma^2)=K\begin{bmatrix}1 & \mu_z\\ \mu_z & 2\sigma^4K\beta^2+\mu_z^2+\sigma_z^2\end{bmatrix}$$

leading to

$$\mathrm{var}_\beta(\sigma^2)=\frac{(\sigma_\varepsilon^2+\sigma^2\beta^2)^2}{2\sigma^4\beta^2+\sigma_z^2(\sigma_\varepsilon^2+\sigma^2\beta^2)}. \tag{2.5}$$

The formula for the variance of β_0 is given in Section 2.1. As is shown below, the asymptotic variance (2.5) is not always an increasing function of σ^2. Here are some properties of (2.5) as the function of σ^2

1. $\mathrm{var}_\beta(\sigma^2=0)=\sigma_\varepsilon^2/\sigma_z^2$ corresponds to the classical regression model with the variance of the OLS-estimator.

2. $\lim_{\sigma^2\to\infty}\mathrm{var}_\beta(\sigma^2)=.5\beta^2$, is not equal to zero, as might be expected since the noise goes to infinity.

3. The sign of the derivative of $\mathrm{var}_\beta(\sigma^2)$ with respect to σ^2 coincides with the sign of the expression $(\beta^2\sigma_z^2-4\sigma_\varepsilon^2)\sigma^2+\sigma_z^2\sigma_\varepsilon^2$. In the neighborhood of $\sigma^2=0$, the function $\mathrm{var}_\beta(\sigma^2)$ is always increasing because $\mathrm{var}'_\beta(\sigma^2=0)>0$.

4. The partial variance paradox occurs if and only if

$$t = \frac{j\,¯\,j}{¾_2 = ¾_z} < 2: \qquad (2.6)$$

If (2.6) holds then for some $¾^2$ we have $\mathrm{var}^0_{¯}(¾^2) < 0$.

5. If the model variance, $¾^2_2$ is large enough, i.e. if (2.6) holds, then $\mathrm{var}_{¯}(¾^2)$ has a maximum at $¾^2 = ¾^2_z ¾^2_2 = (4¾^2_2 \; ¡ \; ¯^2 ¾^2_z)$:

The paradox is illustrated in Figure 1, where $¾^2_z = 1$; $¾^2_2 = 1$. If $¯ = 3$; then $t = 3$ and the variance of $¯$ is an increasing function of $¾^2$: If $¯ = 1$; then $t = 1$ and the unconditional variance paradox occurs.

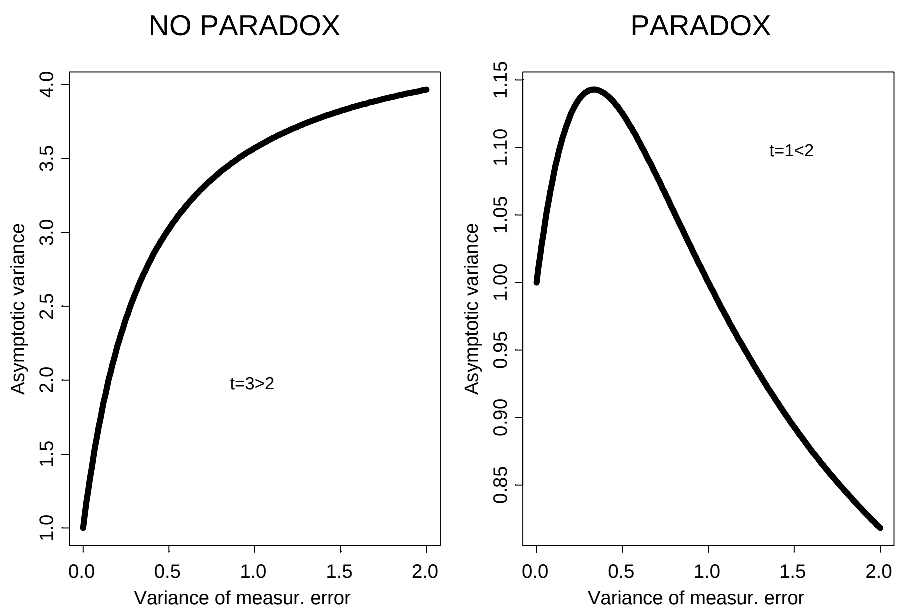

FIG.1 Linear Berkson model.

We suggest the following explanation of the discovered phenomena. First, let us consider the linear regression with known model variance $¾^2$ and without measurement error with covariate z having a small range of values, i.e. $¾^2_z$ is small. Then, the MLE (which in this case coincides with the OLS-estimate) is not stable in sense that its variance is large. Moreover, it goes to in¯nity when

σ_z^2 tends to zero. Now, let us introduce measurement error in observations of z. Although this measurement error increases the overall variance of the dependent variable, $\sigma_\varepsilon^2 + \sigma^2\beta^2$, it also increases the range for observations on z (now x). The result of two opposite effects decreases the asymptotic variance of the MLE in Berkson model if condition (2.6) is fulfilled.

3. THE CLASSICAL MEASUREMENT MODEL

The primary model is (2.1) and the measurement error model is as given in Section 1.2. The key difference between the classical and the Berkson measurement error models is that now the observed covariate, z_i, depends on the model error ε_i. As before, the index i will be omitted. We start with the conditional information. The conditional distribution $y \mid z$ is normal, and $N(\beta_0 + \beta\mu_{x|z}, \sigma_\varepsilon^2 + \beta^2\sigma_{x|z}^2)$ where $\mu_{x|z}$ and $\sigma_{x|z}^2$ are defined in (1.1). Thus, the conditional log-likelihood is

$$l_{y|z} = -\frac{1}{2}\log(\sigma_\varepsilon^2 + \beta^2\sigma_{x|z}^2) - \frac{1}{2}(\sigma_\varepsilon^2 + \beta^2\sigma_{x|z}^2)^{-1}(y - \beta_0 - \beta\mu_{x|z})^2.$$

We consider only the simplest case when all parameters are known but β. Then the conditional information for β is

$$I(\sigma^2 \mid z) = E\left(\frac{dl_{y|z}}{d\beta}\right)^2 = (\sigma_\varepsilon^2 + \beta^2\sigma_{x|z}^2)^{-1}\mu_{x|z}^2 + 2\beta^2\sigma_{x|z}^4(\sigma_\varepsilon^2 + \beta^2\sigma_{x|z}^2)^{-2}.$$

Clearly,

$$I(0 \mid z) = \sigma_\varepsilon^2 z^2; \quad I(\infty \mid z) = (\sigma_\varepsilon^2 + \beta^2\sigma_x^2)^{-1}\mu_x^2 + 2\beta^2\sigma_x^4(\sigma_\varepsilon^2 + \beta^2\sigma_x^2)^{-2}$$

and for certain values of $(\sigma_\varepsilon^2, z, \sigma_x^2, \mu_x)$ we may have $I(0 \mid z) < I(\infty \mid z)$. Therefore, the conditional information paradox can occur. Consequently, if a large enough proportion of a sample satisfies this condition, the conditional variance paradox will occur.

To consider the unconditional information paradox, we represent the model in the following way

$$u_i = \begin{bmatrix} y_i \\ z_i \end{bmatrix} \sim N(\mu, V)$$

where

$$\mu = \begin{bmatrix} \beta_0 + \beta\mu_x \\ \mu_x \end{bmatrix}; \quad V = \begin{bmatrix} \beta^2\sigma_x^2 + \sigma_\varepsilon^2 & \beta\sigma_x^2 \\ \beta\sigma_x^2 & \sigma_x^2 + \sigma^2 \end{bmatrix}$$

with the log-likelihood function for the ith observation

$$l_i = -\frac{1}{2}\log|V| - \frac{1}{2}(u_i - \mu)' V^{-1}(u_i - \mu). \tag{3.1}$$

As is shown in the Appendix, the 5×5 information matrix for $(\beta, \beta_0, \mu_x, \sigma_x^2, \sigma_\varepsilon^2)$ is

$$D^{-1}\begin{bmatrix} 2\sigma_x^2\beta^2 D^{-1}\sigma^4 + \mu_x^2\sigma^2 + \tau & & & & \\ \mu_x\omega & \omega & & & \\ \mu_x\beta\sigma^2 & \beta\sigma^2 & \theta & & \\ D^{-1}\sigma^2\sigma_x^2\beta\theta & 0 & 0 & .5D^{-1}\theta^2 & \\ D^{-1}\sigma^2\sigma_x^2\beta\omega & 0 & 0 & D^{-1}.5\beta^2\sigma^4 & .5D^{-1}\omega^2 \end{bmatrix} \tag{3.2}$$

where

$$D = \beta^2\sigma^2\sigma_x^2 + \sigma_\varepsilon^2(\sigma_x^2 + \sigma^2); \quad \tau = \sigma_x^2(\mu_x^2 + \sigma_x^2); \quad \theta = \beta^2\sigma^2 + \sigma_\varepsilon^2; \quad \omega = \sigma_x^2 + \sigma^2.$$

The derivative of this matrix with respect to σ^2 is a positive semidefinite matrix. It follows that all submatrices of the information matrix are positive semidefinite. Consequently, the unconditional information paradox does not take place for this model regardless of what parameters are known or unknown. The diagonal elements of the inverse information matrix (3.2), i.e. the variances of parameters, are as follows:

$$\begin{aligned} \mathrm{var}_{ML}(\sqrt{n}\hat{\beta}) &= (2\beta^2\sigma^4 + \sigma_y^2\sigma^2 + \sigma_\varepsilon^2\sigma_x^2)\sigma_x^{-4}; \quad \mathrm{var}_{ML}(\sqrt{n}\hat{\mu}_x) = \sigma_x^2 + \sigma^2; \\ \mathrm{var}_{ML}(\sqrt{n}\hat{\sigma}_x^2) &= 2(\sigma_x^2 + \sigma^2)^2; \quad \mathrm{var}_{ML}(\sqrt{n}\hat{\sigma}_\varepsilon^2) = 2(\sigma_\varepsilon^2 + \beta^2\sigma^2)^2; \\ \mathrm{var}_{ML}(\sqrt{n}\hat{\beta}_0) &= \left[2\mu_x^2\beta^2\sigma^4 + (\sigma_x^2 + \mu_x^2)\beta^2\sigma_x^2\sigma^2 + (\sigma_x^2 + \mu_x^2)\sigma_\varepsilon^2\sigma_x^2\right]\sigma_x^{-4}. \end{aligned}$$

The first formula was given by Fuller (1987, p.16). All variances are increasing functions of σ^2 - no paradox occurs.

4. GENERALIZED LINEAR MODEL FOR BINARY DATA BERKSON CASE

Let us consider a generalized linear model for binary data with errors-in-variables as follows

$$\Pr(D_i = 0|x_i) = \mu(\beta_0 + \beta x_i); \quad x_i \mid z_i \sim N(z_i; \sigma^2); \quad z_i \sim N(\mu_z; \sigma_z^2); \quad (4.1)$$

where $\mu(\cdot)$ is a link function, and $x_i - z_i$ is the measurement error. As before, we assume that σ^2 is known, and our aim is to investigate the dependence of the variance of the maximum likelihood estimate for β on σ^2. We assume that $\mu' < 0$ and $s \cdot \mu'(s) \to 0$ if $s \to \infty$. Also, as in the linear model, it is assumed that $\Pr(D_i = 0; \beta|x_i; z_i) = \Pr(D_i = 0; \beta|x_i)$. As easy to see, the conditional probability with respect to z_i is:

$$\begin{aligned}
P_i(\sigma^2) &= \Pr(D = 0; \sigma^2 \mid z_i) = E_{x|z}(\mu(\beta_0 + \beta x)) \\
&= (2\pi\sigma^2)^{-1/2} \int \mu(\beta_0 + \beta s) \exp\left(-\frac{1}{2\sigma^2}(s - z_i)^2\right) ds \\
&= E_U \mu(\beta_0 + \beta z_i + \sigma\beta U); \qquad U \sim N(0; 1).
\end{aligned}$$

The maximum likelihood estimate of $\beta = (\beta_0; \beta)$ based on observations $D_1; \ldots; D_n$ maximizes the log-likelihood function

$$l(\beta_0; \beta; \sigma^2) = \sum_{D_i=0} \log P_i(\sigma^2) + \sum_{D_i=1} \log(1 - P_i(\sigma^2)). \quad (4.2)$$

4.1. The conditional information

The conditional information for $\hat{\beta}$ is equal to

$$\begin{aligned}
I &= I(\beta) = \left(\frac{1}{P(0; \beta)} + \frac{1}{1 - P(0; \beta)}\right) \left(\frac{\partial P}{\partial \beta}\right) \left(\frac{\partial P}{\partial \beta}\right)' \\
&= \frac{1}{P(0; \beta)(1 - P(0; \beta))} \left(\frac{\partial P}{\partial \beta}\right) \left(\frac{\partial P}{\partial \beta}\right)'. \quad (4.3)
\end{aligned}$$

For computational convinience, we consider properties of the information and the variance of the maximum likelihood estimate as a function of σ, instead of

$\sigma^2.$ Thus, the conditional information (4.3) for the model (4.1) becomes

$$I_{\beta_0,\beta}(\sigma \mid z) = A(\sigma)\begin{bmatrix} E^2\mu'(\sigma) & E\mu'(\sigma)\cdot E[(z+\sigma U)\mu'(\sigma)] \\ E\mu'(\sigma)\cdot E[(z+\sigma U)\mu'(\sigma)] & E^2[(z+\sigma U)\mu'(\sigma)] \end{bmatrix} \tag{4.4}$$

where $P(\sigma) = E_U\mu(\beta_0+\beta z+\sigma\beta U)$; $\quad A(\sigma) = P^{-1}(\sigma)(1-P(\sigma))^{-1}$ and $\mu'(\sigma) = \mu'(\beta_0+\beta z+\sigma\beta U).$ The $(2,2)$th element of this matrix

$$I_\beta(\sigma \mid z) = \frac{E^2[(z+\sigma U)\mu'(\beta_0+\beta z+\beta\sigma U)]}{E[\mu(\beta_0+\beta z+\beta\sigma U)](1-E[\mu(\beta_0+\beta z+\beta\sigma U)])} \tag{4.5}$$

is the conditional information for $\hat{\beta}.$ To show that the conditional information paradox can take place here, it is sufficient to show that $I_\beta(\sigma \mid z)$ is not a decreasing function of σ for some z, β_0, β and $\sigma \geq 0.$ Thus, partial paradox is considered since β_0 is assumed known. Recall, that in this case the asymptotic variance for the MLE of β is the reciprocal of (4.5).

4.2. Information properties

We analyze the dependence of I_β given by (4.5) on σ for fixed values of β_0, β and $z.$ As shown in the Appendix,

$$I'_\beta(\sigma = 0 \mid z) = 0 \quad \text{and} \quad \lim_{\sigma\to\infty} I_\beta(\sigma \mid z) = 0. \tag{4.6}$$

The last equation means that the variance of the MLE goes to infinity when $\sigma \to \infty.$ This is in contrast to linear case (Section 2.2), when the limit of the variance was $.5\beta^{-2} > 0$. The conditions for the occurrence of the paradox are described in the following theorem.

Theorem 4.1. The conditional information paradox occurs in the neighborhood of $\sigma = 0$ if

$$2\mu''\mu(1-\mu)(z\beta) + (\mu'''\mu(1-\mu) - \mu'\mu''(.5-\mu))(z\beta)^2 < 0 \tag{4.7}$$

where μ and all its derivatives are calculated at $\beta_0 + z\beta.$

The proof of this theorem is given in the Appendix.

As follows from the Theorem, the conditional information paradox takes place when the second derivative of the information with respect to σ at zero is positive. We will illustrate this with the logistic regression model, where

$$\mu(s) = \frac{1}{1+e^s}, \quad \mu'(s) = -\frac{e^s}{(1+e^s)^2},$$
$$\mu''(s) = \frac{e^s(e^s-1)}{(1+e^s)^3}, \quad \mu'''(s) = \frac{e^s[3-(e^s-2)^2]}{(1+e^s)^4}.$$

Assume that $\beta_0 = 0.$ Then the inequality (4.7) can be expressed in terms of probability $p = e^s/(1+e^s) = \Pr(D=1).$ From the Theorem, the information for $\hat{\beta}$ is an increasing function of σ for all $p < p_*,$ where p_* is the solution to the nonlinear equation

$$4(2p-1) - (8p^2 - 8p + 1)\log\frac{p}{1-p} = 0 \tag{4.8}$$

which has the root $p_* = 0.0129.$ Thus, when $\beta_0 = 0$, as follows from Theorem 4.1, if the probability of the occurrence of the event is less than 0.0129 the paradox occurs for small measurement error. Therefore, we conclude that the paradox likely occurs for rare events. Note, since the information (4.5) is invariant to the choice of the link function, μ or $1-\mu,$ the equation (4.8) is symmetric for $p.$ Consequently, "the occurrence of the event" may mean $D_i = 0$ or $D_i = 1.$

4.3. Generalization to nonnormal Berkson errors

Here we will show that the paradox occurs for rare events with nonnormal Berkson measurement errors. Let λ be a density of a random variable with mean 0 and variance $1.$ The model (4.1) will be written as

$$\Pr(D=0;\beta|x) = \mu(\beta_0 + \beta x), \quad \frac{x-z}{\sigma} \sim \lambda. \tag{4.9}$$

Denoting $U = \sigma^{-1}(x-z)$ the model (4.9) can be rewritten in terms of the standardized random variable $U,$

$$\Pr(D=0;\beta|U) = \mu(\beta_0 + \beta z + \beta\sigma U), \quad U \sim \lambda,$$

with the marginal probability

$$\Pr(D = 0, \beta, \sigma^2) = E_U \mu(\beta_0 + \beta z + \beta\sigma U), \qquad U \sim \lambda.$$

As is shown in the Appendix, the properties (4.6) and the result of the previous theorem are true for the nonnormal case.

4.4. The unconditional variance paradox

From Section 1.3, we can conclude that the unconditional information paradox for Berkson normal and non-normal errors holds too, because in the previous notation $q = (0, 1)'$, that corresponds the case when the intercept term β_0 is known. The variance paradox takes place if $dI_{\beta_0,\beta}(\sigma \mid z)/d\sigma < 0.$

The unconditional information matrix for the model (4.1) is

$$I(\sigma) = \begin{bmatrix} I_{11}(\sigma) & I_{12}(\sigma) \\ I_{12}(\sigma) & I_{22}(\sigma) \end{bmatrix}$$

where

$$\begin{aligned}
I_{11}(\sigma) &= E_z \frac{E^2 \mu'(\beta_0 + \beta z + \beta\sigma U)}{E[\mu(\beta_0 + \beta z + \beta\sigma U)](1 - E[\mu(\beta_0 + \beta z + \beta\sigma U)])}, \\
I_{12}(\sigma) &= E_z \frac{E[(z + \sigma U)\mu'(\beta_0 + \beta z + \beta\sigma U)] \cdot E\mu'(\beta_0 + \beta z + \beta\sigma U)}{E[\mu(\beta_0 + \beta z + \beta\sigma U)](1 - E[\mu(\beta_0 + \beta z + \beta\sigma U)])}, \\
I_{22}(\sigma) &= E_z \frac{E^2[(z + \sigma U)\mu'(\beta_0 + \beta z + \beta\sigma U)]}{E[\mu(\beta_0 + \beta z + \beta\sigma U)](1 - E[\mu(\beta_0 + \beta z + \beta\sigma U)])}
\end{aligned} \tag{4.10}$$

and $z \sim N(\mu_z, \sigma_z^2).$ Then, the asymptotic variance for the maximum likelihood estimate of β is the (2,2)th element of the inversed information matrix, as usual. If the conditional information paradox occurs when $z = \mu_z$ for given β_0, then it is likely to occur for the variance $I_{22}^{-1}(\sigma)$ assuming β_0 is known and the variance of z is small enough. For example, the variance paradox does not occur when $\beta_0 = 1, \beta = -0.1, \mu_z = 1, \sigma_z^2 = 1$, where $\Pr(D = 1) = 0.711.$ When $\beta_0 = -3$ and all other parameters are the same as previously, $\Pr(D = 1) = 0.043$, then the paradox occurs. For σ up to 7, the asymptotic variance of the maximum likelihood estimate decreases (FIG. 2). The optimal σ here is about 5.

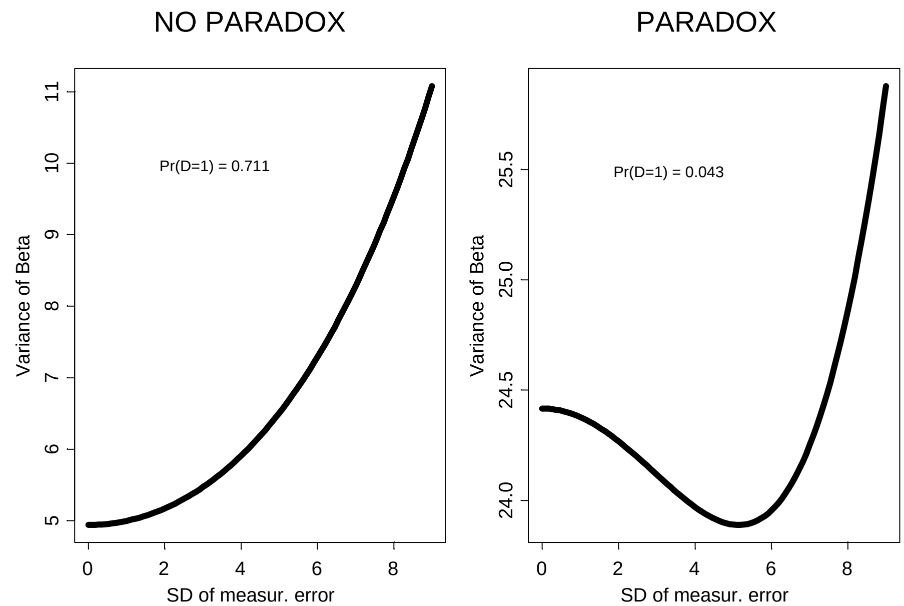

FIG. 2. Logistic regression with the Berkson covariate measurement error. Paradox likely to ocur for rare events.

4.5. Example

The occurrence of the conditional variance paradox is illustrated using the Nurses' Health Study, a prospective cohort of 89,538 white married females, 601 of whom developed breast cancer between 1980 and 1984 (Willett et al., 1987). The 4 year cumulative incidence of breast cancer was a rare event with probability 0.0067 over the four year study. Let us assume the standard logistic regression model for the variables measured without error

$$\mathrm{logit}P(D_i = 0) = \beta_0 + \gamma \mathrm{Age}_i + \beta_1 \mathrm{Alcohol}_i + \beta_2 \mathrm{Fat}_i + \beta_3 \mathrm{Calories}_i; \quad i = 1, \ldots, n.$$

which is an example of a more general model of form logit $P(D_i = 0) = \gamma' w_i + \beta' x_i$ where $w_i = (1, \mathrm{Age}_i)'$ is measured without error, and we assume $x_i = z_i + \epsilon_i$, where $z_i = (\mathrm{Alcohol(g/day)}_i, \mathrm{Fat(g/day)}_i, \mathrm{Calories(Kcal/day)}_i)'$.

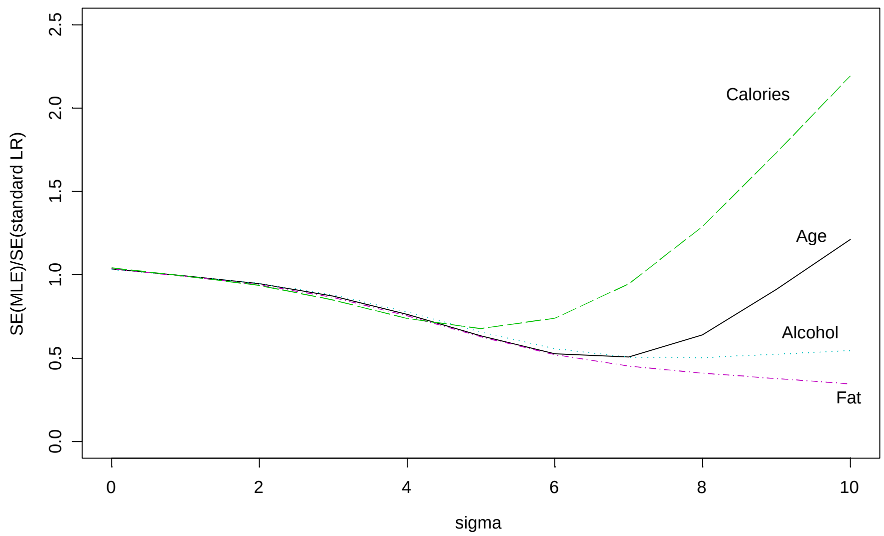

FIG. 3. Relative standard errors of the maximum likelihood estimate vs. standard error of standard logistic regression as functions of the error variance in the Nurses' Health Study.

It is assumed that the standard error of the measurement error ϵ_i is proportional to its covariate's mean, i.e. $\epsilon_i \sim N(0; \sigma^2 \mathrm{diag}(\bar{z}^2))$; where $\bar{z} = (6.626; 27.792; 1.579)'$ is the vector of empirical means. Then,

$$\Pr(D_i = 0 \mid z_i) = E_{t_i}[1 + \exp(\gamma' w_i + t_i)]^{-1}; \quad t_i \sim N(\beta' z_i; \sigma^2(\beta' \bar{z})^2). \quad (4.11)$$

To calculate the information matrix we can assume that the z_i are normally distributed. After estimating its mean and covariance matrix, similarly to (4.10), we can calculate the unconditional information through a multidimensional integral. Alternatively, the conditional information matrix can be calculated avoiding the integration problem mentioned above. Let $I_i(\sigma \mid z_i)$ denote the conditional information matrix for model (4.11). Then

$$\mathrm{cov}(\sigma) = \left(\frac{1}{n} \sum_{i=1} I_i(\sigma \mid z_i) \right)^{-1} \quad (4.12)$$

is the 5×5 dimensional asymptotic covariance matrix of β based on the conditional information matrix. FIG. 3 displays the standard errors of the maximum likelihood estimates of the coefficients $\gamma, \beta_1, \beta_2, \beta_3$ calculated by (4.12) as functions of σ, compared with the standards errors of the maximum likelihood estimate when $\sigma = 0$ (standard logistic regression with no measurement error). An algorithm suggested by Crouch & Spiegelman (1990) was used to calculate the integral involved in (4.11). A paradoxical situation takes place: the quality of the estimates is improved with increased measurement error. The picture is different for different variables. When $\sigma \simeq 7$, the standard errors for all estimates is improved by 20% approximately.

5. GENERALIZED LINEAR MODEL FOR BINARY DATA: THE CLASSICAL MEASUREMENT ERROR MODEL

Here we assume that the observed variable is $z_i = x_i + \delta_i$, where $x_i \sim N(\mu_x, \sigma_x^2)$ and $\delta_i \sim N(0, \sigma^2)$. As before, x_i and δ_i are assumed to be independent, so $z_i \sim N(\mu_x, \sigma_x^2 + \sigma^2)$. It is assumed that σ^2 is known. To write the likelihood for observed (y_i, z_i), $i = 1, \dots, n$, the conditional distribution $x_i \mid z_i$ is needed. Using (1.1), the probability of the event $D = 0$ is

$$\Pr(D = 0 \mid z, \beta, \sigma^2) = E_U\, \mu(\beta_0 + \beta(\mu_{x|z} + \sigma_{x|z} U))$$

where U is a normalized random variable, and index i is omitted. The conditional information for β given a single observation z is

$$I(\sigma|z) = \frac{R^2(\sigma)}{P(\sigma)(1 - P(\sigma))}. \tag{5.1}$$

where

$$\begin{aligned} P(\sigma) &= E_U \left(\mu(\beta_0 + \beta(\mu_{x|z} + \sigma_{x|z} U)) \right), \\ R(\sigma) &= \frac{dP}{d\beta} = E_U \left[\left(\mu_{x|z} + \sigma_{x|z} U \right) \mu' \left(\beta_0 + \beta(\mu_{x|z} + \sigma_{x|z} U) \right) \right]. \end{aligned}$$

We wish to investigate properties of the information (5.1). We will start by finding the extreme values for this function. First, noting that

$$\mu_{x|z}(0) = z; \quad \sigma_{x|z}(0) = 0; \quad \lim_{\sigma \to \infty} \mu_{x|z}(\sigma) = \mu_x; \quad \lim_{\sigma \to \infty} \sigma_{x|z}(\sigma) = \sigma_x.$$

we obtain

$$I(0|z) = \frac{[z\mu'(\beta_0 + \beta z)]^2}{\mu(\beta_0 + \beta z)[1 - \mu(\beta_0 + \beta z)]}$$

and

$$I(\infty|z) = \lim_{\sigma \to \infty} I(\sigma|z) = \frac{E^2[(\mu_x + \sigma_x U)\mu'(\beta_0 + \beta(\mu_x + \sigma_x U))]}{E[\mu(\beta_0 + \beta(\mu_x + \sigma_x U))](1 - E[\mu(\beta_0 + \beta(\mu_x + \sigma_x U))])}.$$

Also, as shown in the Appendix,

$$I'(0|z) = 0; \tag{5.2}$$

as in the Berkson measurement error case. The conditional variance of $\hat{\beta}$ based on the conditional information will have a zero slope at $\sigma = 0$ as the function of σ, since $\mathrm{var}'(\hat{\beta}) = -(\sum I'(\sigma|z_i))(\sum I(\sigma|z_i))^{-2}$.

5.1. The conditional information paradox

The conditional paradox must take place when $I(\infty|z) > I(0|z)$. We investigate when, and if, this occurs for logistic regression. For simplicity, let $z = \mu_x$ and $\beta_0 = 0$. Then letting $H \sim N(\beta z, \sigma_x^2\beta^2)$ we can write

$$I(0|z) = \frac{[\beta z \mu'(\beta z)]^2}{\beta^2 \mu(\beta z)[1 - \mu(\beta z)]}; \quad I(\infty|z) = \frac{E^2[H\mu'(H)]}{\beta^2 E\mu(H)(1 - E\mu(H))}.$$

The difference $(I(\infty|z) - I(0|z))\beta^2$ can be expressed in terms of the probability of event $P = P(D = 1) = 1 - \mu(h)$. In FIG. 4, this difference is expressed as a function of P for different sd(H), where $\mathrm{sd}(H) = \sigma|\beta|$. When $\beta_0 = 0$, the conditional information paradox occurs for rare events: If sd(H)=1, the paradox occurs when $P < 0.0078$; for sd(H)= 0.6, it occurs when $P < 0.0108$. Thus again, we conclude that the conditional paradox is likely to occur for rare events.

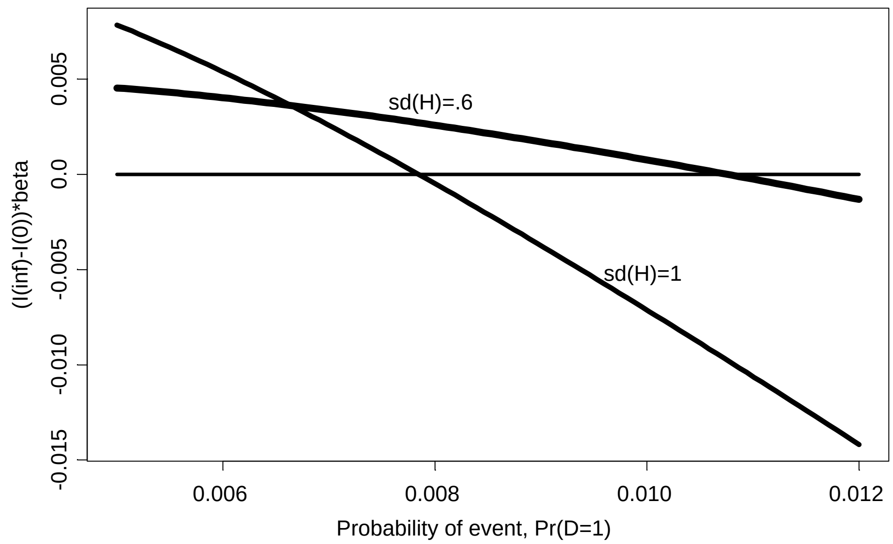

FIG. 4. Logistic regression with classical covariate measurement error. The conditional information paradox for rare events.

5.2. About the unconditional information and variance paradox

The unconditional information for β is

$$\begin{aligned} I(\sigma) &= E_{z \sim N(\mu_x; \sigma_x^2+\sigma^2)} I^{-}(\sigma|z) \\ &= (2\pi(\sigma_x^2+\sigma^2))^{-1/2} \int_{-\infty}^{+\infty} \frac{R^2(\sigma)}{P(\sigma)(1-P(\sigma))} e^{-\frac{1}{2(\sigma_x^2+\sigma^2)}(z-\mu_x)^2} dz \end{aligned} \quad (5.3)$$

We start with establishing some properties of the information $I(\sigma)$. First, let us calculate the extreme quantities

$$I(0) = E_z \frac{[z\mu'(\beta_0+\beta z)]^2}{\mu(\beta_0+\beta z)[1-\mu(\beta_0+\beta z)]}; \quad z \sim N(\mu_x; \sigma_x^2)$$

and, as shown in the Appendix,

$$I(\infty) = \frac{(E_x(x\mu'(\beta_0+\beta x)))^2}{E_x\mu(\beta_0+\beta x)(1-E_x\mu(\beta_0+\beta x))}; \quad x \sim N(\mu_x; \sigma_x^2). \quad (5.4)$$

Also, as is shown in the Appendix,

$$I(\sigma) > I(\infty); \quad I'(0) = 0; \quad (5.5)$$

i.e. at least for large σ the variance paradox does not occur. We were unable to prove that $I'(\sigma) < 0$; so the function (5.3) was investigated empirically for the logit link over a wide range of $(\beta_0, \beta, \mu_x, \sigma_x^2)$ to determine whether it is ever an increasing function of σ. In all scenarios considered, it was a decreasing function of σ and the variance of the maximum likelihood estimate was an increasing function of σ. It appears empirically that the paradox does not occur for binary models with the classical measurement.

6. DISCUSSION

The above considerations and findings are summarized in Table 1.

TABLE 1

Summary of cases considered where the paradox occurs

	Berkson meas. error model		Classical meas. error model	
	Linear	Binary	Linear	Binary
Conditional inform./ variance paradox	Yes	Yes	Yes	Yes
Unconditional information paradox	Yes	Yes	No	No*
Unconditional variance paradox	Yes, if σ_2^2 is unknown No, if σ_2^2 is known	Yes	No	No*

* − empirical evidence for the logistic regression model.

The conditional information paradox can occur in all cases considered: linear/binary and Berkson/classical measurement error model which can be interpreted as the noise to observations. As a consequence, the asymptotic variance of the maximum likelihood estimate based on the conditional information matrix may be a decreasing function of error variance (see Section 4.5). For the Berkson measurement error model, the paradox can occur both in linear and binary model under some circumstances. For instance, the unconditional variance paradox for the linear model with Berkson measurement error occurs when the model variance, σ_ϵ^2 is known and inequality (2.6) holds. It could be explained in the following way: if the model variance is known and the observed covariate without measurement error has a narrow range, then introducing measurement error decreases the power of estimation and inflates the range of observations. The combination of those two effects improves the MLE when (2.6) holds.

When inequality (4.7) holds, the unconditional variance paradox will occur in the binary model. It may be explained in the same manner, because in the binary model, the model variance is known as $\mu(1 - \mu)$. Thus, the condition of knowing the model variance is fulfilled here. One might expect the paradox in the binary model with Berkson measurement error when: (i) the incidence rate for the modeled event is small, (ii) the initial design matrix of covariates is ill-conditioned, so introducing measurement error in covariates stabilizes it. Apparently, the combination of those two circumstances led to the paradox for the Nurses' Health Study.

It is interesting to observe that the unconditional paradox for the information or the asymptotic variance cannot occur with the classical measurement error either for the linear or for the binary model. For the linear model, this was proven mathematically. For the binary model, we have only empirical evidence for the logistic regression model. The mathematical proof of this question remains open.

When the measurement error variance, σ^2, is unknown and estimated based on a validation/calibration study, if the validation study is independent of the main study (external validation study) and if the main study data provide little information about $\hat{\sigma}^2_{ML}$, it is possible that the asymptotic correlations of $\hat{\sigma}_{ML}$ with other ML-estimates could be small. Then, the full information matrix will be close to block-diagonal, and the need to estimate σ^2 from validation data should not affect the occurrence of the variance paradox. However, further analysis should be done to find the specific conditions, if any, under which the paradox is true, when σ^2 is unknown and validation data are available.

What are practical implications of the paradox? First, the knowledge that the asymptotic variance of the MLE is not always an increasing function of the Berkson measurement error variance could be important when designing experiments, particularly when calculating the size of the validation study. Second, to find the variance of the MLE in classical measurement error model one has to use the inverse of the expected information matrix, not observed, because even the paradox does not hold in this case it might hold using the observed information matrix. Despite the fact that for nonlinear models the expected information may need to solve quadrature problem, like in the binary model, it is more accurate and desirable. Third, at least hypothetically, new methods of estimation can be suggested in Berkson model when the variance of the covariate is small and the model variance is known (like in binary model). However, this is the matter of future research.

7. APPENDIX. PROOFS

The information matrix for the linear classical measurement error model.

After some algebra, the log-likelihood function can be rewritten as

$$l_i = -\frac{1}{2}\log(D) - \frac{1}{2}D^{-1}T_i; \qquad i = 1;\ldots;n$$

where

$$\begin{gathered} D = \beta^2\sigma^2\sigma_x^2 + \sigma_\varepsilon^2(\sigma_x^2 + \sigma^2); \\ T_i = (\sigma_x^2 + \sigma^2)(y_i - \beta_0 - \beta\mu_x)^2 - 2\beta\sigma_x^2(y_i - \beta_0 - \beta\mu_x)(z_i - \mu_x) \\ +(\beta^2\sigma_x^2 + \sigma_\varepsilon^2)(z_i - \mu_x)^2; \quad R_i = y_i - \beta_0 - \beta\mu_x; \quad Q_i = z_i - \mu_x: \end{gathered}$$

Omitting routine calculations we obtain:

$$E\frac{\partial^2 l_i}{\partial\beta_0^2} = -D^{-1}(\sigma_x^2 + \sigma^2); \quad E\frac{\partial^2 l_i}{\partial\beta_0\partial\mu_x} = -D^{-1}\beta\sigma^2;:$$

$$\begin{aligned} E\frac{\partial^2 l_i}{\partial\mu_x^2} &= -D^{-1}(\beta^2\sigma^2 + \sigma_\varepsilon^2); \quad E\frac{\partial^2 l_i}{\partial\mu_x\partial\beta} = -D^{-1}\mu_x\beta\sigma^2; \\ E\frac{\partial^2 l_i}{\partial\beta_0\partial\beta} &= -D^{-1}\mu_x(\sigma_x^2 + \sigma^2); \quad E\frac{\partial^2 l_i}{\partial\mu_x\partial\sigma_x^2} = 0; \quad E\frac{\partial^2 l_i}{\partial\mu_x\partial\sigma_\varepsilon^2} = 0: \end{aligned}$$

The most complicated are derivatives with respect to the parameter β:

$$\frac{\partial l_i}{\partial \beta} = -\sigma^2\sigma_x^2 D^{-1}\beta + \sigma^2\sigma_x^2 D^{-2}\beta T_i + D^{-1}G_i$$

where

$$G_i = -\frac{1}{2}\frac{\partial T_i}{\partial \beta} = \mu_x(\sigma_x^2 + \sigma^2)R_i + \sigma_x^2(y_i - \beta_0 - 2\beta\mu_x)Q_i - \beta\sigma_x^2 Q_i^2.$$

To find the expected values of the second derivative, let us split the derivative $\partial l_i/\partial\beta$ into 3 terms. For the first and second term, we obtain

$$\begin{aligned} E\frac{\partial}{\partial\beta}(-\sigma^2\sigma_x^2 D^{-1}\beta) &= \sigma^2\sigma_x^2 D^{-2}\left(\sigma^2\sigma_x^2\beta^2 - \sigma_\epsilon^2(\sigma_x^2 + \sigma^2)\right); \\ E\frac{\partial}{\partial\beta}(\sigma^2\sigma_x^2 D^{-2}\beta T_i) &= \sigma^2\sigma_x^2 E\left(-4\sigma^2\sigma_x^2 D^{-3}\beta^2 T_i + D^{-2}T_i - 2D^{-2}\beta G_i\right). \end{aligned}$$

Now we notice that $ED^{-1}T_i = 2$ and $E(G_i) = -\beta\sigma_x^2\sigma^2$. Using these facts, and omitting some intermediate calculations, we find

$$E\frac{\partial^2 l_i}{\partial\beta^2} = \sigma^2\sigma_x^2 D^{-2}(D - 2\sigma^2\sigma_x^2\beta^2) - D^{-1}(\mu_x^2 + \sigma_x^2)(\sigma_x^2 + \sigma^2).$$

The first and second derivatives with respect to variances parameters σ_x^2 and σ_ϵ^2 are

$$\frac{\partial l_i}{\partial\sigma_x^2} = -\frac{1}{2}D^{-1}(\beta^2\sigma^2 + \sigma_\epsilon^2) + \frac{1}{2}D^{-2}(\beta^2\sigma^2 + \sigma_\epsilon^2)T_i - \frac{1}{2}D^{-1}(R_i^2 - 2\beta R_iQ_i + \beta^2Q_i^2)$$

and

$$E\frac{\partial^2 l_i}{\partial(\sigma_x^2)^2} = -\frac{1}{2}D^{-2}(\beta^2\sigma^2 + \sigma_\epsilon^2)^2; \quad E\frac{\partial^2 l_i}{\partial\beta\,\partial\sigma_x^2} = -\sigma^2\sigma_x^2 D^{-2}\beta(\beta^2\sigma^2 + \sigma_\epsilon^2).$$

Similarly,

$$\frac{\partial l_i}{\partial\sigma_\epsilon^2} = -\frac{1}{2}D^{-1}(\sigma_x^2 + \sigma^2) + \frac{1}{2}D^{-2}(\sigma_x^2 + \sigma^2)T_i - \frac{1}{2}D^{-1}Q_i^2$$

and

$$E\frac{\partial^2 l_i}{\partial(\sigma_\epsilon^2)^2} = -\frac{1}{2}D^{-2}(\sigma_x^2 + \sigma^2)^2; \quad E\frac{\partial^2 l_i}{\partial\sigma_\epsilon^2\,\partial\sigma_x^2} = -\frac{1}{2}D^{-2}\beta^2\sigma^4.$$

All the results are represented in the information matrix (3.2).

Proof of (4.6). The conditional information (4.5) can be written as

$$I_{\beta}(\sigma|z) = \frac{R^2}{P(1-P)}, \quad I'_{\beta}(\sigma|z) = 2\frac{R'RP(1-P) - R^2P'(.5-P)}{P^2(1-P)^2}$$

where

$$P = E\left(\Phi(\beta_0 + \beta z + \beta\sigma U)\right), \quad R = E\left((z+\sigma U)\Phi'(\beta_0 + \beta z + \beta\sigma U)\right).$$

Let us assume for a moment that $\beta > 0$, then by definition

$$\lim_{\sigma\to\infty} \Phi(\beta_0 + \beta z + \beta\sigma u) = \begin{cases} 1 \text{ for } u < 0 \\ 0 \text{ for } u > 0 \end{cases}$$

and therefore

$$\lim_{\sigma\to\infty} E[\Phi(\beta_0 + \beta z + \beta\sigma U)] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} \exp(-\frac{1}{2}u^2)du = \frac{1}{2}. \tag{7.1}$$

The same is true when $\beta < 0$. Now, letting $r = \beta_0 + \beta z$, we note that for $\beta \neq 0$

$$\lim_{\sigma\to\infty} \sigma U\Phi'(r + \beta\sigma U) = \lim_{\sigma\to\infty} \Phi'(r + \beta\sigma U) = 0$$

and

$$\begin{aligned} &\lim_{\sigma\to\infty} E[(z+\sigma U)\Phi'(r+\beta\sigma U)] \\ = &\lim_{\sigma\to\infty} zE[\Phi'(r+\beta\sigma U)] + \lim_{\sigma\to\infty} E[\sigma U\Phi'(r+\beta\sigma U)] = 0. \end{aligned}$$

which implies the first statement of (4.6) is proved. For the special case $\sigma = 0$ we have

$$\begin{aligned} P'(\sigma &= 0) = \beta E[U\Phi'(r)] = \beta\Phi'(r)E(U) = 0, \\ R'(\sigma &= 0) = E[U\Phi'(r)] + \beta zE[U\Phi''(r)] = \Phi'(r)E(U) + \beta z\Phi''(r)E(U) = 0. \end{aligned}$$

that implies the second statement of (4.6).

$$\left.\frac{dR}{d\sigma}\right|_{\sigma=0} = E[U\Phi'(r)] + \beta zE[U\Phi''(r)] = 0.$$

Proof of Theorem 4.1.

Since the first derivative of the information $I_\beta(\sigma|z)$ at $\sigma = 0$ is zero, the paradox will take place iff $I''_\beta(\sigma = 0|z) > 0$. Let us do some preliminary calculations:

$$P'' = \beta^2 E[U^2\mu''(r+\beta\sigma U)];\ P''(0) = \beta^2 E[U^2\mu''(r)] = \beta^2\mu''(r)E(U^2) = \beta^2\mu''(r)$$

because $E(U^2) = 1$; where $r = \beta_0 + \beta z$. Further

$$\begin{aligned} R'' &= 2\beta E[U^2\mu''(r+\beta\sigma U)] + \beta^2 E[(z+\sigma U)U^2\mu'''(r+\beta\sigma U)]; \\ R''(0) &= 2\beta\mu''(r) + z\beta^2\mu'''(r). \end{aligned}$$

Now since $P'(0) = R'(0) = 0$, the sign of $d^2 I/d\sigma^2$ at $\sigma = 0$ coincides with the sign of

$$\begin{aligned} R''RP(1-P) - R^2P''(.5-P) &= R[R''P(1-P) - RP''(.5-P)] \\ &= z\mu'\{(2\beta\mu'' + z\beta^2\mu''')\mu(1-\mu) - z\mu'\beta^2\mu''(.5-\mu)\}, \end{aligned}$$

leading to the statement of the theorem.

Proof of (4.6) and Theorem 4.1 for the nonnormal error.

As one can see in (4.6) and (4.7), we consider extremal values of $\sigma = 0$ and $\sigma = \infty$ where only two properties of U are used: $E(U) = 0$ and $\mathrm{var}(U) = 1$. The only place where the normal law was used is in relation to formula (7.1). However, for any standardized random variable U

$$\lim_{\sigma\to\infty} E[\mu(r+\beta\sigma U)] = \begin{cases} P(U<0) = \int_{-\infty}^{0}\lambda(v)dv & \text{if } \beta > 0 \\ P(U>0) = \int_{0}^{+\infty}\lambda(v)dv & \text{if } \beta < 0 \end{cases}$$

and denominator of I_β tends to a constant while the nominator tends to 0 when $\sigma \to \infty$.

Proof of (5.2). We have

$$\frac{\partial\mu_{x|z}}{\partial\sigma} = -\frac{2\sigma_x^2\sigma}{(\sigma_x^2+\sigma^2)^2}(z-\mu_x);\quad \frac{\partial\sigma_{x|z}}{\partial\sigma} = \left(1+\frac{\sigma^2}{\sigma_x^2}\right)^{-1/2} - \sigma^2\left(1+\frac{\sigma^2}{\sigma_x^2}\right)^{-3/2};$$

that implies

$$\left.\frac{\partial\mu_{x|z}}{\partial\sigma}\right|_{\sigma=0} = 0;\quad \left.\frac{\partial\sigma_{x|z}}{\partial\sigma}\right|_{\sigma=0} = 1.$$

It is important to note that $P'(0) = 0$. Thus,

$$\frac{\partial P}{\partial \sigma} = P'(\sigma) = \beta E_U\left(\mu'(\beta_0 + \beta(\mu_{x|z} + \sigma_{x|z}U))\left(\frac{\partial \mu_{x|z}}{\partial \sigma} + \frac{\partial \sigma_{x|z}}{\partial \sigma}U\right)\right);$$

$$\left.\frac{\partial P}{\partial \sigma}\right|_{\sigma=0} = \beta E_U\,(\mu'(\beta_0 + \beta z)U)) = \beta\mu'(\beta_0 + \beta z)E(U) = 0.$$

Then, the sign of the I derivative at zero coincides with the sign of the following expression $R(0)R'(0)$. But, $R(0) = z\mu'(\beta_0 + \beta z)$ and

$$\left.\frac{\partial R}{\partial \sigma}\right|_{\sigma=0} = E_U(\mu'(\beta_0 + \beta z)U) + E_U(\mu''(\beta_0 + \beta z)zU) = 0;$$

which proves (5.2).

Proofs of (5.4) and (5.5).

By the definition,

$$I(\infty) = \lim_{\sigma\to\infty} E_{z\sim N(\mu_x;\sigma_x^2+\sigma^2)}\frac{R^2(\sigma)}{P(\sigma)(1-P(\sigma))}.$$

When $\sigma \to \infty$ from (1.1) we obtain $\mu_{x|z} \to \mu_x$ and $\sigma^2_{x|z} \to \sigma^2_x$. Thus,

$$P(\sigma) \to E_U(\mu(\beta_0 + \beta(\mu_x + \sigma_x U))) = E_x(\mu(\beta_0 + \beta x))$$

$$R(\sigma) \to E_x(x\mu'(\beta_0 + \beta x))$$

and we come to (5.4).

To prove (5.5) first, we note that the function

$$f(R;P) = \frac{R^2}{P(1-P)}; \quad 0 < P < 1$$

is convex. As follows from the convexity of this function from Jensen's inequality if R and P are random $E(f(R;P)) \geq f(E(R);E(P))$. Applying this formula to the information function (5.3), we obtain the following inequality for $I(\sigma)$:

$$E_z\frac{R^2(\sigma)}{P(\sigma)(1-P(\sigma))} \geq \frac{\left\{E_zE_U\left[\left(\mu_{x|z} + \sigma_{x|z}U\right)\mu'\left(\beta_0 + \beta(\mu_{x|z} + \sigma_{x|z}U)\right)\right]\right\}^2}{E_zE_U\left[\mu\left(\beta_0 + \beta(\mu_{x|z} + \sigma_{x|z}U)\right)\right]\left\{1 - E_zE_U\left[\mu\left(\beta_0 + \beta(\mu_{x|z} + \sigma_{x|z}U)\right)\right]\right\}}$$

However,

$$E_z E_U \left[\mu \left(\beta_0 + \beta(\mu_{x|z} + \sigma_{x|z} U) \right) \right] = E_{x \sim N(\mu_x, \sigma_x^2)} \mu(\beta_0 + \beta x)$$

and

$$E_z E_U \left[\left(\mu_{x|z} + \sigma_{x|z} U \right) \mu' \left(\beta_0 + \beta(\mu_{x|z} + \sigma_{x|z} U) \right) \right] = E_{x \sim N(\mu_x, \sigma_x^2)} (x \mu'(\beta_0 + \beta x)):$$

Since the right hand side of the last inequality is $l(\infty)$; the inequality in (5.5) is proved.

To calculate l'; we denote

$$\phi(\sigma; z) = \frac{1}{\sqrt{2\pi}} \exp \left(- \frac{1}{2(\sigma_x^2 + \sigma^2)} (z - \mu_x)^2 \right):$$

Then,

$$l(\sigma) = (\sigma_x^2 + \sigma^2)^{-1/2} \int l(\sigma|z) \phi(\sigma; z) dz$$

and

$$\begin{aligned} l' = & - \sigma(\sigma_x^2 + \sigma^2)^{-3/2} \int l(\sigma|z) \phi(\sigma; z) dz + (\sigma_x^2 + \sigma^2)^{-1/2} \int l'(\sigma|z) \phi(\sigma; z) dz \\ & - \sigma(\sigma_x^2 + \sigma^2)^{-5/2} \int (z - \mu_x)^2 l(\sigma|z) \phi(\sigma; z) dz: \end{aligned} \tag{7.2}$$

Since $l'(\sigma = 0|z) = 0$ for any z; we obtain from (7.2) $l'(0) = 0$:

ACKNOWLEDGMENTS

This work was supported by grants from National Institute of Health CA 50597 and K01 0H00106. We thank Dr. Frank Speizer, Principal Investigator of the Nurses' Health Study for his support, and Mario Casella for his thoughtful reading of the manuscript.

References

[1] Crouch, E.A.C. and Spiegelman, D. (1990). "The evaluation of integrals of the form $\int_{-\infty}^{+\infty} f(t) \exp(-t^2) dt$: application to logistic-normal models", J. Am. Statist. Assoc., 85, 464-469.

[2] Fuller, W.A. (1987). Measurement Error Models, New York: Wiley.

[3] Willett, W.C., Stampfer, M.J., Colditz, G.A., Rosner, B.A., Hennekens, C.H. and Speizer, F.E. (1987). "Dietary fat and risk of breast cancer", New England Journal of Medicine, 316, 1178-1180.