On the Simultaneous Effects of Model Misspecification and Errors-in-Variables

Paul Gustafson

Department of Statistics
University of British Columbia
333-6356 Agricultural Road
Vancouver, B.C.
Canada V6T 1Z2

gustaf@stat.ubc.ca

May 21, 2002

On the Simultaneous Effects of Model Misspecification and Errors-in-Variables

May 21, 2002

Abstract

Misspecified models and noisy covariate measurements are two common sources of bias in statistical inferences. While there is considerable literature on the consequences of each problem in isolation, this article investigates the effect of both problems in tandem. In the context of linear models, the large-sample error in estimating the regression function is partitioned into two terms, one resulting from model misspecification and the other from covariate imprecision. While trivial to establish, this decomposition proves interesting. Specifically, it reveals tradeoffs between the deleterious effects of model misspecification and covariate imprecision in a number of scenarios. A finite-sample version of the decomposition is also presented. This permits study of the relative impacts of model misspecification, covariate imprecision, and sampling variability, with reference to the detectability of the model misspecification via diagnostic plots.

Keywords: bias; errors-in-variables; measurement error; misclassification; model misspecification.

1 Introduction

In keeping with adages such as "all models are wrong but some are useful" (Box, 1979), it is recognized that biases induced by model misspecification are ubiquitous in statistical practice. The hope is that screening out models which are clearly incorrect will render these biases small relative to other uncertainties present in a statistical analysis. There is a sizeable literature on the effects of model misspecification, or "wrong-model analysis" in various scenarios. Some key references include Kent (1982), Gould and Lawless (1988), Ramsey (1969), and White (1981, 1982).

While less ubiquitous than model misspecification, the noisy or imprecise measurement of covariates is another common difficulty in statistical analysis. This problem is often referred to under the rubric of *errors-in-variables*, or more specifically *measurement error* in the case of continuous covariates and *misclassification* in the case of categorical covariates. In either situation, the effect of undetected or ignored imprecision in a covariate is typically a bias towards zero in the corresponding regression coefficient. Errors-in-variables are particularly common in biostatistics and epidemiology, where many putative risk factors can only be measured roughly at the level of the individual. Some general errors-in-variables references include Carroll, Ruppert, and Stefanski (1995), Thomas, Stram, and Dwyer (1993), and Willett (1989).

While interesting research questions still abound, the effects of model misspecification and covariate imprecision are reasonably well understood. For the most part, however, the two topics have been investigated in isolation from one another. The effect of model misspecification is typically studied in the context of precise covariate measurements, while the effect of errors-in-variables is typically studied in the context of a correctly specified model. In contrast, this article examines the simultaneous effects of model misspecification and covariate imprecision.

There are both before-the-fact and after-the-fact rationales for the present investigation. *A priori*, the ubiquity of model misspecification suggests it may often be an issue when covariate imprecision is manifested. Thus it seems sensible to consider the joint effect of the two problems. *A posteriori*, our findings point toward an intriguing relationship between model misspecification and covariate imprecision. In examples we see tradeoffs in the form of an inverse relationship between the error arising from model misspecification and the error arising from covariate imprecision. Moreover, we garner some sense of which problem is more damaging in a particular situation.

The article is organized as follows. The main ideas are developed in the next section, where the large-sample error in estimating a regression function is partitioned into a component due to model misspecification and a component due to covariate imprecision. Sections 3 through 5 then detail the application of this decomposition in different scenarios. A finite-sample version of the decomposition is presented in Section 6, along with simulations designed to illustrate the practical ramifications of simultaneous model misspecification and covariate imprecision. A short discussion ensues in Section 7. Some of the details needed in the examples of Sections 3, 4, and 5 are relegated to an Appendix.

2 Quantifying Errors

Let Y be the response variable, and let $V = (V_1, \ldots, V_k)'$ be the collection of potential predictor variables. Say that the *actual* relationship between Y and V is governed by $S = \{S_1(V), \ldots, S_q(V)\}'$ via

$$E(Y|V) \quad = \quad \alpha_1 S_1 + \ldots \alpha_q S_q, \tag{1}$$

but model misspecification arises because the analyst regards $T = \{T_1(V), \ldots, T_p(V)\}'$ as the relevant predictors. That is, $E(Y|V)$ is *incorrectly* assumed to be a linear function of the components of T. Moreover, say that the measurement of T is error-prone. Thus while the analyst is focussed on the relationship between Y and T, the available data are measurements of Y and U, where $U = \{U_1, \ldots, U_p\}'$ is a noisy surrogate for T. Note that S, T, and U are viewed as random rather than fixed, and throughout the article the existence of joint second moments for (Y, S, T, U) is assumed.

To quantify the effects of model misspecification and covariate imprecision, let β be the large-sample limiting coefficients for least-squares regression of Y on T, and let γ be the large-sample limiting coefficients for least-squares regression of Y on U. If both the model misspecification and the measurement error are undetected or ignored, then with enough data the analyst is incorrectly led to believe that

$$E(Y|V) \quad = \quad \gamma_1 T_1 + \ldots \gamma_p T_p. \tag{2}$$

Thus in estimating the regression function the difference between (2) and (1) is the net large-sample error that results from both model misspecification and covariate imprecision. This is formalized by defining the average squared error (ASE) as

$$ASE \quad = \quad E\left\{(\alpha' S - \gamma' T)^2\right\}, \tag{3}$$

where the expectation is with respect to the underlying distribution of V. That is, we are averaging the squared error in estimating the regression function with respect to the distribution of the covariates.

We can also define average squared errors for model misspecification and covariate imprecision separately. The error due to misspecification (hence the subscript M) is captured by

$$ASE_M \quad = \quad E\left\{(\alpha' S - \beta' T)^2\right\}, \tag{4}$$

the average squared difference between the actual regression function $E(Y|V) = \alpha' S$ and the large-sample estimate $\beta' T$ which results from using the misspecified model and regressing Y on precise measurements of T. Similarly, the error due to covariate imprecision (hence the subscript I) is captured by

$$ASE_I \;=\; E\left\{(\beta' T - \gamma' T)^2\right\}, \tag{5}$$

the average squared difference between the estimated regression function based on the misspecified model with precise measurements and that based on the misspecified model with imprecise measurements. We emphasize that the covariate imprecision is undetected or ignored, so the analyst believes to be regressing Y on T while actually regressing Y on U. Consequently, the large-sample estimated regression function is $\gamma' T$ rather than $\gamma' U$.

The definitions (3), (4) and (5) lead quite naturally to the following theorem. Part (i) simply states that the overall average squared error (3) decomposes as the sum of misspecification term (4) and the measurement error term (5). While this may seem somewhat predictable, it appears that such a decomposition has not been developed in the model misspecification literature or the errors-in-variables literature. Parts (ii) and (iii) of the theorem give expressions for the component terms (4) and (5) respectively.

Theorem 1 *Assume that $E(SS')$, $E(TT')$ and $E(UU')$ are all of full rank. Then*

(i) $ASE = ASE_M + ASE_I$.

(ii) $ASE_M = \alpha' A_M \alpha$, where

$$A_M \;=\; E(SS') - E(ST')E(TT')^{-1}E(TS').$$

(iii) $ASE_I = \alpha' A_I \alpha$, where $A_I = B' E(TT') B$, with

$$B \;=\; E(TT')^{-1}E(TS') - E(UU')^{-1}E(US').$$

Proof. To establish (i) it suffices to show that

$$E\left\{(\alpha' S - \beta' T)T'(\beta - \gamma)\right\} \;=\; 0. \tag{6}$$

But note that standard results for misspecified models (see, for instance, White, 1982) give $\beta = \text{argmin}_\xi E\{(Y - \xi' T)^2\}$ and hence

$$\begin{aligned} E(TT')\beta &= E(TY) \\ &= E\{TE(Y|V)\} \\ &= E(TS')\alpha. \end{aligned} \tag{7}$$

Therefore the left-hand side of (6) can be expressed as

$$\alpha' E\left[\left\{S - E(ST')E(TT')^{-1}T\right\} T'\right](\beta - \gamma) \;\; = \;\; 0,$$

giving the desired result. The expression in (ii) follows straightforwardly from (7), while the analogous definition for γ, namely

$$E(UU')\gamma \;\; = \;\; E(US')\alpha,$$

leads immediately to (iii). □

If some of the regressors are correctly specified (i.e. components of T which appear in S) then intuitively one expects these components, or more specifically the magnitudes of their coefficients in (1), will not contribute to the error arising from model misspecification. Furthermore, if such components are measured precisely, then one expects they will not contribute to the error arising from covariate imprecision. These intuitions are verified in the following easily-established lemma.

Lemma 1

(i) If $T_j = S_i$ for some i and j, then the i-th row and column of A_M are zero. That is, ASE_M does not depend on α_i.

(ii) If $U_j = S_i$ for some i and j, then the i-th row and column of A_M and A_I are zero. That is, ASE_M and ASE_I (and hence ASE) do not depend on α_i.

Proof. Assume the condition in (i) holds, and without loss of generality say $i = j = 1$. Thus $T_1 = S_1$. Since A_M is symmetric and non-negative definite, it suffices to show that its $(1,1)$ entry is zero to establish (i). Evaluating ASE_M when $\alpha = (1,0,\ldots,0)'$ gives $(A_M)_{11} = \min_\xi E\{(S_1 - \xi' T)^2\} = 0$ as desired. Similarly, say $U_1 = S_1$. Then evaluating ASE when $\alpha = (1,0,\ldots,0)'$ gives $A_{11} = \text{argmin}_\xi E\{(S_1 - \xi' U)^2\} = 0$. Since A_M and A_I are symmetric and non-negative definite, (ii) follows immediately.□

3 Example: Missed Curvature or Interaction

Say an analyst is interested in the relationship between a response variable Y and predictors X and W. Without loss of generality assume the predictors have been scaled so that

$E(X) = E(W) = 0$ and $Var(X) = Var(W) = 1$, and let $\rho = Corr(X, W)$. The analyst views $T = (1, X, W)'$ as the predictors of interest. However, X is subject to nondifferential additive measurement error, so that the actual regressors used are $U = (1, X^*, W)'$, where X^* and (W, Y) are conditionally independent, with $E(X^*|X) = X$ and $Var(X^*|X) = \tau^2$. Note that since X is standardized, τ^2 can be regarded as the measurement error variance as a fraction of the predictor's variance.

Now say the real relationship between Y and (X, W) is governed by

$$\begin{aligned} E(Y|X) &= \alpha' S \\ &= \alpha_1 + \alpha_2 X + \alpha_3 W + \alpha_4 X^2, \end{aligned} \tag{8}$$

so that the analyst is 'missing' the curved effect of X. Lemma 1 indicates that ASE_M can depend only on α_4, while ASE_I can depend only on α_2 and α_4. Indeed, straightforward calculation using part (ii) of Theorem 1 yields expressions for A_M and A_I which are given in the Appendix. It is particularly instructive to consider the situation where W depends on X in a linear manner; that is, $E(W|X) = \rho X$. The expressions then simplify to

$$ASE_M = (m_4 - 1 - m_3^2)\alpha_4^2, \tag{9}$$

$$ASE_I = c(\tau, \rho)(\alpha_2 + m_3\alpha_4)^2, \tag{10}$$

where $m_i = E(X^i)$, and

$$c(\tau, \rho) = \frac{(1 - \rho^2)\tau^4}{(1 + \tau^2 - \rho^2)^2}. \tag{11}$$

To put (9) and (10) in context, let $\lambda = \mathrm{Var}(\alpha_2 X + \alpha_4 X^2) = \alpha_2^2 + (m_4 - 1)\alpha_4^2 + 2m_3\alpha_2\alpha_4$ represent the total "signal" due to X, and let $\omega = \mathrm{Var}(\alpha_4 X^2)/\lambda = (m_4 - 1)\alpha_4^2/\lambda$ be the fraction of the total signal which is due to the quadratic term. Then we have

$$ASE_M = \lambda\gamma\omega, \tag{12}$$

$$ASE_I = c(\tau, \rho)\lambda(1 - \gamma\omega), \tag{13}$$

where $\gamma = 1 - \{m_3^2/(m_4 - 1)\}$. Note that $\gamma \in (0, 1]$ by the Cauchy-Schwartz inequality, and that $\gamma = 1$ when X has a symmetric distribution. Clearly (12) and (13) reveal a tradeoff. As the real relationship becomes more curved (ω increases while λ is fixed), the error due to misspecification increases, but the error due to covariate imprecision decreases. Straightforward analysis shows that $c(\tau, \rho) < 1$, thus the misspecification term always dominates in the sense that ASE increases with ω for fixed λ. However $c(\tau,\rho)$ can be appreciable when τ is relatively large. Therefore as the underlying relationship becomes more curved, increased error due to misspecification is partially offset by decreased error due to covariate imprecision.

We also consider what happens when the analyst misses an interaction between the two predictors. That is, the true relationship (8) is replaced with

$$\begin{aligned} E(Y|X) &= \alpha' S \\ &= \alpha_1 + \alpha_2 X + \alpha_3 W + \alpha_4 XW. \end{aligned}$$

Again, resultant expressions for A_M and A_I are given in the Appendix. In this scenario instructive expressions result when the dependence of W on X is assumed to be both linear and homoscedastic; that is, $E(W|X) = \rho X$ and $Var(W|X) = 1 - \rho^2$. Under these conditions it can be shown that (12) and (13) hold again, with $\lambda = Var(\alpha_2 X + \alpha_4 XW)$, $\omega = Var(\alpha_4 XW)/\lambda$, $c(\tau,\rho)$ defined previously as per (11), and γ now given by

$$\gamma = 1 - \frac{\rho^2 m_3^2}{1 - \rho^2 + \rho^2(m_4 - 1)}.$$

As previously, the Cauchy-Schwartz inequality implies that $\gamma \in (0, 1]$, while $\gamma = 1$ if X has a symmetric distribution. Thus the effect of simultaneous model misspecification and covariate imprecision in the missed interaction scenario is very similar to that arising in the missed curvature scenario.

4 Example: Covariate Dichotomization

The next example expands upon qualitative findings in Gustafson and Le (2001). For the sake of tractability we assume the two predictors (X, W) have a bivariate normal distribution, an assumption not required in the previous section. As previously we assume without loss of generality that both X and W have been scaled to have mean zero and variance one, and let $\rho = \text{Corr}(X, W)$. Imagine the analyst is contemplating two choices of regressors, namely $T^{(C)} = (1, X, W)'$ and $T^{(D)} = (1, I\{X > 0\}, W)'$. In particular, he is considering dichotomization of X by comparison to its mean of zero. The practice of creating categorical covariates from continuous covariates is relatively common, especially in biostatistics and epidemiology. Also, say that X is subject to nondifferential normal measurement error. Thus the actual regressors are either $U^{(C)} = (1, X^*, W)$ or $U^{(D)} = (1, I\{X^* > 0\}, W)$, where $X^*|X, W, Y \sim N(X, \tau^2)$.

In this scenario it proves fruitful to consider what happens when the true relationship between the response and predictors falls in between the possibilities considered by the analyst. Specifically, say that

$$\begin{aligned} E(Y|X,W) &= \alpha^T S \\ &= \alpha_1 + \alpha_2 \left\{ (1-\omega)X + \omega\sqrt{2\pi}\left(I\{X>0\} - \frac{1}{2} \right) \right\} + \alpha_2 W, \end{aligned} \tag{14}$$

for some $\omega \in [0,1]$. Thus as ω increases $T^{(D)}$ becomes a more appropriate choice of regressors while $T^{(C)}$ becomes a less appropriate choice. As an aside, the centering and scaling by $\sqrt{2\pi}$ of the indicator function in (14) appears for a technical reason. In particular, for any value of ω it leads to $\beta = \alpha$ when considering $T^{(C)}$ as the predictors. Thus ω is more readily interpreted as the weight given to the dichotomous component in (14).

Entries for the matrices needed to compute A_M and A_I via Theorem 1 are given in the Appendix, for both $T^{(C)}$ and $T^{(D)}$. In light of Lemma 1, ASE_M and ASE_I can depend only on α_2 in either case. Figure 1 gives plots of ASE_M, ASE_I and ASE as a function of w, where the vertical scale is based on fixing $\alpha_2 = 1$. Both choices of T and different values of τ and ρ are considered.

The behaviour of the model misspecification term ASE_M in Figure 1 is entirely predictable. If $T^{(C)}$ is used, ASE_M increases with ω as the true relationship moves away from the postulated model. If $T^{(D)}$ is used, ASE_M decreases with w as the true relationship moves toward the postulated form. However, the behaviour of the covariate imprecision term ASE_I is more curious. If $T^{(C)}$ is used, ASE_I does not depend on ω. That is, the deleterious effect of measurement error is the same regardless of the form of the underlying relationship, amplifying findings in Gustafson and Le (2001). On the other hand, if $T^{(D)}$ is used then ASE_I increases with w. Thus the decomposition of ASE into ASE_M and ASE_I quantifies a tradeoff noted by Gustafson and Le concerning the dichotomization of a continuous predictor. As the model fit improves, the damaging effect of measurement error worsens. This phenomenon is particularly acute in the $\tau = 0.75$ scenarios shown in Figure 1. Here ASE actually increases with ω for larger values of ω, so that the overall error *decreases* as the true relationship moves further away from the postulated model!

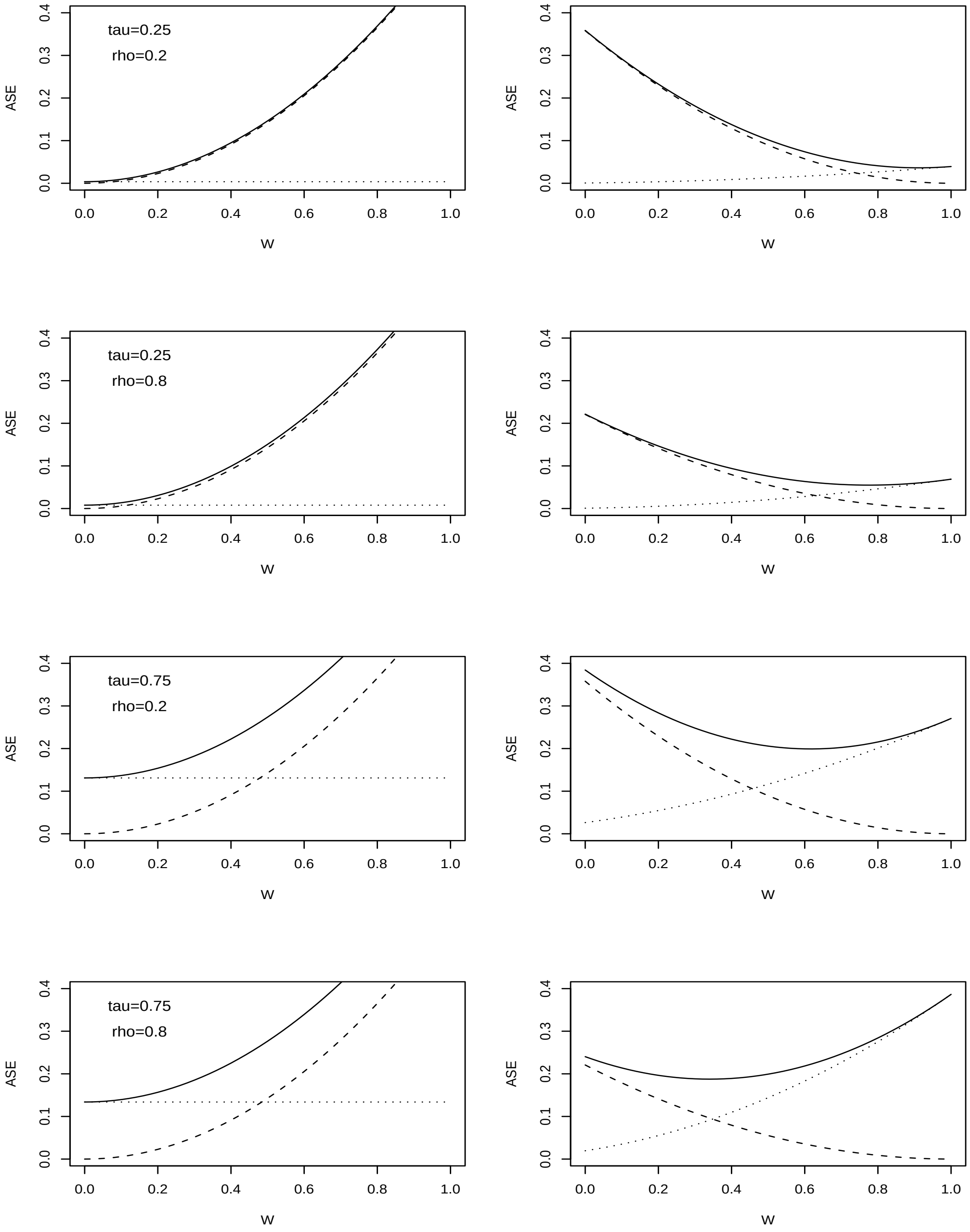

Figure 1: ASE (solid curve), ASE_M (dashed curve) and ASE_I (dotted curve) as a function of ω which indexes the true relationship (14). The four pairs of plots correspond to combinations of $\tau = 0.25$ or $\tau = 0.75$ and $\rho = 0.2$ or $\rho = 0.8$. For each pair, the left plot corresponds to $T^{(C)} = (1, X, W)'$ as the postulated predictors, while the right plot corresponds to $T^{(D)} = (1, I\{X > 0\}, W)'$ as the postulated predictors. The vertical scaling of the plots is based on $\alpha_1 = 1$.

5 Example: How Many Categories?

Often continuous predictors are categorized to avoid specifying a grossly incorrect form for the regression function. In fact with plentiful data it may be reasonable to use many categories, in order to better approximate the actual regression function. Here we examine the joint effects of model misspecification and covariate imprecision as the number of categories increases.

Let X, W and X^* have the same jointly normal distribution as in the previous section. Say that the investigator chooses to categorize the predictor X into r equi-probable categories. That is, the postulated regressors are taken to be

$$T^{(r)} = \left(1, I\left\{\frac{1}{r} \leq \Phi(X) < \frac{2}{r}\right\}, \ldots, I\left\{\frac{r-1}{r} \leq \Phi(X) < 1\right\}, W\right),$$

where $\Phi(\cdot)$ is the standard normal distribution function. The measurement error, however, implies that the actual regressors are

$$U^{(r)} = \left(1, I\left\{\frac{1}{r} \leq \Phi(X^*) < \frac{2}{r}\right\}, \ldots, I\left\{\frac{r-1}{r} \leq \Phi(X^*) < 1\right\}, W\right).$$

In this context we investigate what happens when the real relationship between Y and (X, W) is simply linear. That is,

$$\begin{aligned} E(Y|X) &= \alpha' S \\ &= \alpha_1 + \alpha_2 X + \alpha_3 W. \end{aligned}$$

Setting aside concerns about measurement error, it is clear that a better-fitting model will result as the number of categories r increases. To investigate further, note from Lemma 1 that ASE_M and ASE_I can depend only on α_2. The various quantities required to compute these terms via parts (ii) and (iii) of Theorem 1 are given in the Appendix. Figure 2 plots ASE, ASE_M, and ASE_I as functions of r, for combinations of $\tau = 0.25$ or $\tau = 0.75$, and $\rho = 0.2$ or $\rho = 0.8$. Note that the vertical scaling of the plots is based on fixing $\alpha_2 = 1$.

In each case illustrated in Figure 2, ASE_M decreases with r, to reflect the improved fit that results from using more categories. This improvement, however, is partly mitigated by ASE_I which increases with r in all cases. The net effect is that while the overall error ASE decreases with the number of categories, the rate of decrease is less when τ is larger. For instance, in the $\tau = 0.75$ and $\rho = 0.8$ scenario there appears to be little benefit in increasing r beyond about four. Thus another tradeoff is identified: creating more categories improves the fit of the model but this is partially offset by an increase in the error due to covariate imprecision.

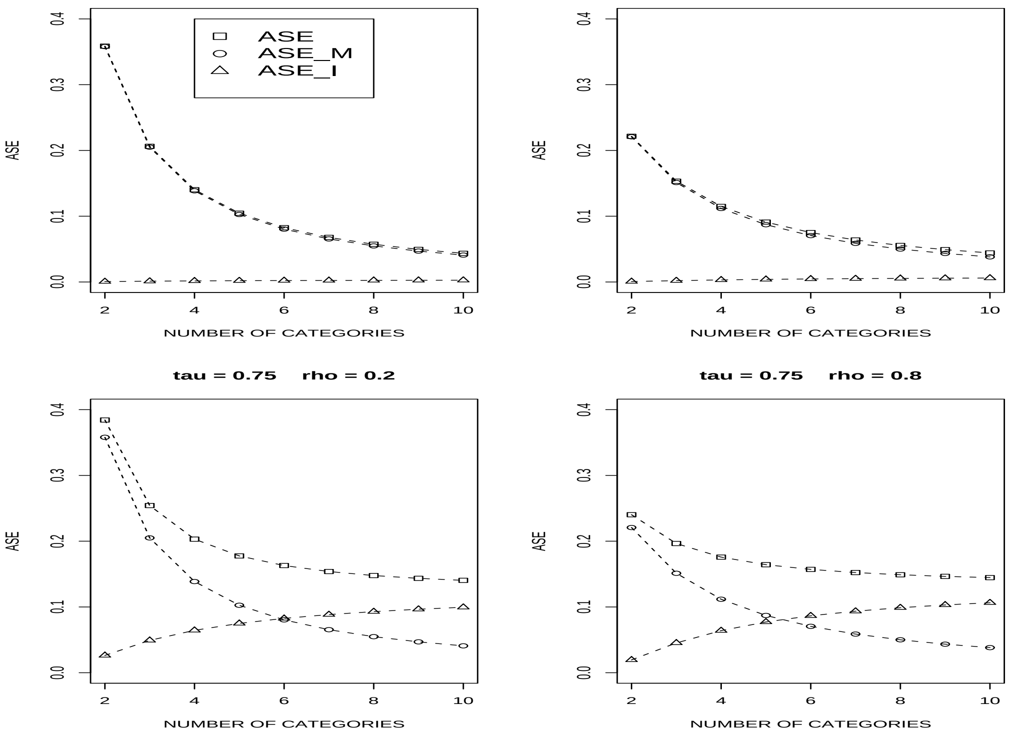

Figure 2: ASE, ASE_M and ASE_I as functions of r, the number of categories used for categorization. The four plots correspond to combinations of $\tau = 0.25$ or $\tau = 0.75$ and $\rho = 0.2$ or $\rho = 0.8$. The underlying relationship is such that $E(Y|X, W)$ is linear in X and W. The vertical scaling of the plots is based on $\alpha_2 = 1$.

Note also that in comparing the $\rho = 0.2$ and $\rho = 0.8$ scenarios when $\tau = 0.75$, ASE is smaller when ρ is larger, at least when the number of categories is small. Moreover, this difference is primarily attributable to ASE_M rather than ASE_I. That is, a larger correlation between X and W is helpful in obtaining a better fit, as the actual linear effect of X can be better reflected by the postulated linear effect of W in tandem with the postulated categorized effect of X.

6 A Finite-Sample Decomposition

It is possible to construct a decomposition analogous to Theorem 1 that describes finite samples and incorporates variability as well as bias. First we establish notation for the finite-sample case. Let the $n \times q$ matrix D_s and the $n \times p$ matrices D_t and D_u be design matrices such that the i-th rows of these matrices constitute a random draw from the joint distribution of (S, T, U), with independence across rows. Let D denote the three matrices collectively, and say the vector of responses is distributed as $Y|D \sim N(D_s\alpha, \sigma^2 I_n)$.

The decomposition operates on fitted values and their expectations. Let

$$\nu_1(Y;D) \quad = \quad D_s(D_s'D_s)^{-1}D_s'Y \tag{15}$$

be the fitted values arising from a correct regression of Y on S, while

$$\nu_2(Y;D) \quad = \quad D_t(D_t'D_t)^{-1}D_t'Y \tag{16}$$

are the fitted values from an incorrect regression of Y on T. Furthermore, if the actual measured regressors are U but the resulting coefficients are regarded as describing the relationship between Y and T, then the envisioned fitted values are

$$\nu_3(Y;D) \quad = \quad D_t(D_u'D_u)^{-1}D_u'Y. \tag{17}$$

For each of (15) through (17), let $\nu_i(D) = E\{\nu_i(Y;D)|D\}$ be the expectation of the fitted values given the design matrices. That is, $\nu_1(D) = D_s\alpha$, $\nu_2(D) = D_t(D_t'D_t)^{-1}D_t'D_s\alpha$, and $\nu_3(D) = D_t(D_u'D_u)^{-1}{D_u}'D_s\alpha$.

Working conditionally given the design matrices, the overall average squared error in estimating $E(Y|D) = \nu_1(D) = D_s\alpha$ by the fitted values $\nu_3(Y;D)$ can be defined as

$$ASE^* \quad = \quad n^{-1}E\left\{ \|\nu_1(D) - \nu_3(Y;D)\|^2 \middle| \, D \right\}.$$

The error due to model misspecification alone is defined as

$$\begin{aligned} ASE_M^* \quad &= \quad n^{-1}\|\nu_1(D) - \nu_2(D)\|^2 \\ &= \quad n^{-1}\|\{I_n - D_t(D_t'D_t)^{-1}D_t'\}D_s\alpha\|^2 \\ &= \quad n^{-1}\,\alpha'D_s'\{I_n - D_t(D_t'D_t)^{-1}D_t'\}D_s\alpha, \end{aligned}$$

while similarly the error due to covariate imprecision is

$$\begin{aligned} ASE_I^* \quad &= \quad n^{-1}\|\nu_2(D) - \nu_3(D)\|^2 \\ &= \quad n^{-1}\|D_t\{(D_t'D_t)^{-1}D_t' - (D_u'D_u)^{-1}D_u'\}D_s\alpha\|^2. \end{aligned}$$

Finally, the error due to sampling variation is naturally defined as

$$\begin{aligned} ASE_V^* \quad &= \quad n^{-1}E\left\{ \|\nu_3(D) - \nu_3(Y;D)\|^2 \middle| \, M \right\} \\ &= \quad n^{-1}\sigma^2 \,\mathrm{tr}\{D_t(D_u'D_u)^{-1}D_t'\}. \end{aligned}$$

In light of Theorem 1, the analogous result for the finite-sample case comes as no surprise. In particular,

$$ASE^* \quad = \quad ASE^*_M + ASE^*_I + ASE^*_V. \tag{18}$$

It is straightforward to establish (18). The two cross-terms involving $v_3(D) - v_3(Y;D)$ are clearly zero, while $\{v_1(D) - v_2(D)\}'\{v_2(D) - v_3(D)\} = 0$ follows from the fact that the second term is in the column space of D_t while the first term is in the orthocomplement of the column space of D_t. We also note that taking expectations of all terms in (18) with respect to M yields an unconditional version of the decomposition.

We apply the decomposition in the missed curvature scenario of Section 3, taking (X, W) to have a bivariate normal distribution with standardized marginals and $\rho = Corr(X, W) = 0.75$. Also we set $\sigma^2 = Var(Y|X) = (0.2)^2$, and $n = 100$. The coefficients $\alpha_1 = 0$, $\alpha_2 = 0.5$ and $\alpha_3 = 0.25$ are fixed in (8), while various values of α_4 and τ, representing various degrees of missed curvature and measurement error, are considered. Note that $\alpha_4 = 0$ and $\tau = 0$ yield a 'strong-signal' scenario, as the ratio of $VarE(Y|X, W)$ to $VarY$ is 0.93. Using simulated data, Table 1 presents the decomposition (18) for each combination of $\tau = 0$, $\tau = 0.1$, $\tau = 0.25$, $\tau = 0.5$ and $\alpha_4 = 0$, $\alpha_4 = 0.005$, $\alpha_4 = 0.029$, $\alpha_4 = 0.085$. Given that X is standardized, the values of $\tau = SD(X^*|X)$ range from no measurement error in X to 50% measurement error in X. The four values of α_4 are chosen to match the four values of τ by setting $ASE_I = ASE_M$ via the expressions given in Section 3. Note that to minimize simulation variability the samples of (Y, D_s, D_t, D_u) for the sixteen scenarios in Table 1 are all constructed by suitable transformation of the same underlying Monte Carlo samples.

To provide some context for Table 1, residual plots based on the data from each scenario are also given. In particular, a curved effect for a predictor is commonly detected from a plot of residuals versus that predictor. Since we are studying the effect of using a noisy surrogate X^* as a regressor in lieu of X, we plot residuals from the fit of Y to $U = (1, X^*, W)'$ against X^*. These appear in Figure 3.

Examining Table 1 and Figure 3 in tandem gives some insight to the potential damage induced by simultaneous model misspecification and covariate imprecision in practice. For instance, consider the $\alpha_4 = 0.029$, $\tau = 0.25$ scenario. Even though there is considerable measurement error, the misspecification, while not detectable from the residual plot, makes the largest contribution to the overall error ASE^*. As a more extreme example, consider the $\alpha_4 = 0.085$ scenarios. In the absence of measurement error ($\tau = 0$), such a substantial model misspec-

α_4	$\tau = 0$	$\tau = 0.1$	$\tau = 0.25$	$\tau = 0.5$
0.000	0.0012 (0.00 0.00 1.00)	0.0013 (0.00 0.06 0.94)	0.0028 (0.00 0.59 0.51)	0.0138 (0.00 0.92 0.08)
0.005	0.0013 (0.06 0.00 0.94)	0.0013 (0.06 0.05 0.89)	0.0028 (0.03 0.56 0.41)	0.0138 (0.01 0.92 0.07)
0.029	0.0036 (0.67 0.00 0.33)	0.0037 (0.66 0.02 0.32)	0.0050 (0.47 0.30 0.23)	0.0158 (0.15 0.78 0.07)
0.085	0.0215 (0.94 0.00 0.06)	0.0216 (0.94 0.00 0.06)	0.0227 (0.89 0.06 0.05)	0.0328 (0.62 0.35 0.03)

Table 1: The decomposition (18) for simulated data. For each combination of $\tau = SD(X^*|X)$ and the missed curvature coefficient α_4, the upper entry is the overall error ASE^*. The lower entry comprises the constituent terms (ASE^*_M ASE^*_I ASE^*_V), expressed as proportions of ASE^*.

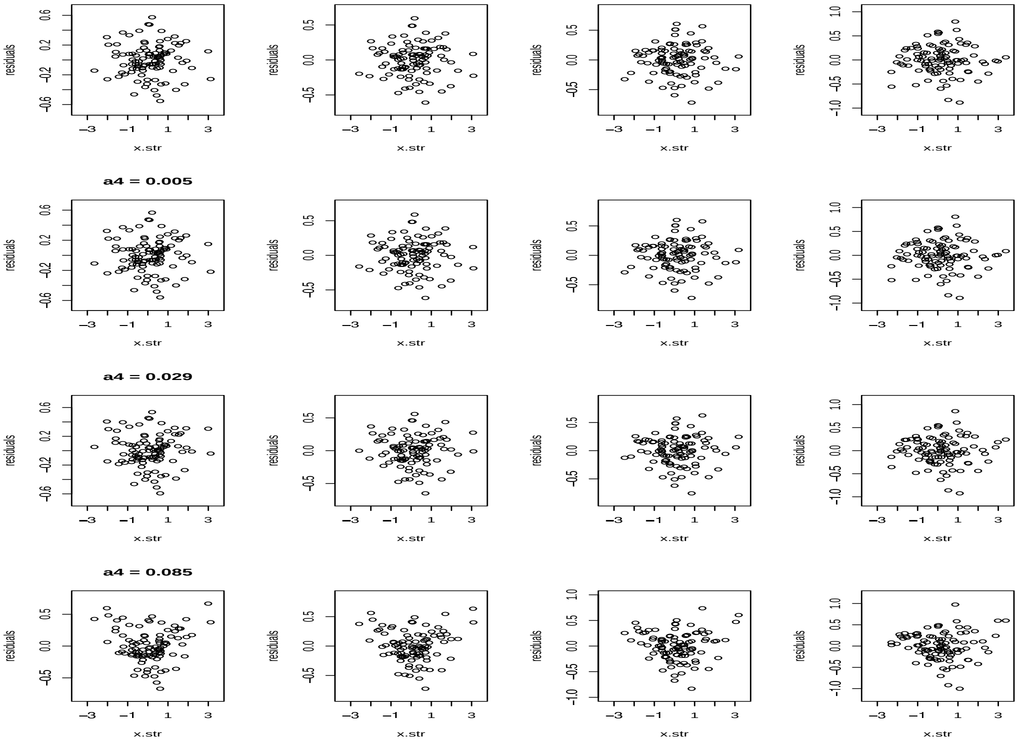

Figure 3: Residual plots in the scenarios described by Table 1. In each case the residuals from the fit of Y to $(1, X^*, W)'$ are plotted against X^*. The row and column layout matches that of Table 1.

ification is clearly indicated by the residual plot. As one might expect, however, the curved pattern in the residuals becomes less pronounced as τ increases. When there is very substantial measurement error ($\tau = 0.5$), the residual plot would likely be viewed as 'passable' by most. However, the model misspecification still contributes about twice as much to the overall error ASE^* as does this sizeable measurement error. Thus measurement error can deliver a deleterious 'double-whammy'. As well as inducing bias in estimated coefficients, it can mask a poorly specified model which might easily be screened-out by diagnostic procedures in the absence of measurement error. Indeed, the present results indicate that the latter problem can be a larger source of error than the former in some situations.

7 Discussion

Consider the two choices of S in Section 3, the two choices of T in Section 4, and Section 5. An inverse relationship between ASE_M and ASE_I arises in four of these five scenarios. For Section 3 and the dichotomized choice of T in Section 4, ASE_M and ASE_I are inversely related as the true underlying relationship between the response and possible covariates is varied. In section 5 the inverse relationship obtains when the analyst's choice of predictors is varied. The exceptional case is the continuous choice of T in Section 4, where in fact ASE_I does not change as the underlying relationship is varied. We do not have an overarching explanation for why an inverse relationship might be typical. The findings sound a cautionary note nonetheless, as they suggest a "no-free-lunch" principle regarding bias due to model misspecification and bias due to covariate imprecision.

The findings in Section 6 speak to the practical ramifications of simultaneous model misspecification and covariate imprecision. They show that the bias induced by a substantial amount of measurement error can be relatively small compared to the bias induced by model misspecification, *without that misspecification being readily detected!* In fact, the very presence of the measurement error can make the detection of misspecification more difficult. Thus methods to correct for measurement error may not be very helpful without evidence that the model linking the response variable to the unobserved precise predictor is appropriate, and such evidence may be hard to obtain. On the one hand this speaks to the desirability of nonparametric regression methods which account for measurement error. On the other hand, this is known to be a difficult problem related to deconvolution (Fan and Truong 1993, Carroll, Maca and Ruppert 1999).

Although it is not illuminating in the present context, a more refined decomposition of ASE than that of Theorem 1 is possible in some situations. In particular, if the chosen predictors T are all functions of the actual predictors S, then the error arising in estimating the regression function $E(Y|S)$ by $\beta' T$ can be viewed as arising from two sources. The first is the difference between the 'full' regression function $E(Y|S) = \alpha' S$ and the 'reduced' regression function $E(Y|T) = \alpha' E(S|T)$, while the second is the error in estimating the reduced regression function by $\beta' T$. This leads easily to $ASE_M = ASE_{M1} + ASE_{M2}$, where

$$ASE_{M1} \quad = \quad E([\alpha'\{S - E(S|T)\}]^2)$$

reflects the loss of information that results from viewing T rather than S as predictors, and

$$ASE_{M2} \quad = \quad E[\{\alpha' E(S|T) - \beta' T\}^2]$$

reflects the large-sample error in estimating $E(Y|T) = \alpha' E(S|T)$ by a linear function of T. Of course if each component of $E(S|T)$ happens to be linear in T, then $ASE_{M2} = 0$, and ASE_M is entirely due to the loss of information in reducing to T from S.

Appendix

In the missed curvature scenario of Section 3, the only non-zero entry of A_M is

$$(A_M)_{44} \quad = \quad m_4 - 1 - \frac{m_3^2 + m_{2,1}^2 + 2\rho m_3 m_{2,1}}{1 - \rho^2}$$

where $m_i = E(X^i)$ and $m_{i,j} = E(X^i W^j)$. On the other hand, A_I is given as

$$A_I \quad = \quad \frac{\tau^4}{(1-\rho^2)(1+\tau^2-\rho^2)^2} \left(\begin{array}{cccc} 0 & 0 & 0 & 0 \\ & (1-\rho^2)^2 & 0 & (1-\rho^2)(m_3 - \rho m_{2,1}) \\ & & 0 & 0 \\ & & & (m_3 - \rho m_{2,1})^2 \end{array} \right).$$

In the missed interaction scenario,

$$(A_M)_{44} \quad = \quad m_{2,2} - \frac{\rho^2(1-\rho^2) + m_{2,1}(m_{2,1} - \rho m_{1,2}) + m_{1,2}(m_{1,2} - \rho m_{2,1})}{1-\rho^2},$$

and

$$A_I \quad = \quad \frac{\tau^4}{(1-\rho^2)(1+\tau^2-\rho^2)^2} \left(\begin{array}{cccc} 0 & 0 & 0 & 0 \\ & (1-\rho^2)^2 & 0 & (1-\rho^2)(m_{2,1} - \rho m_{1,2}) \\ & & 0 & 0 \\ & & & (m_{2,1} - \rho m_{1,2})^2 \end{array} \right).$$

In Section 4,

$$E(SS') = \begin{pmatrix} 1 & 0 & 0 \\ & 1+(\frac{\pi}{2}-1)\omega^2 & \rho \\ & & 1 \end{pmatrix}.$$

For the continuous choice of predictors $T = T^{(C)}$ and $U = U^{(C)}$,

$$E(TT') = \begin{pmatrix} 1 & 0 & 0 \\ & 1 & \rho \\ & & 1 \end{pmatrix}, \tag{19}$$

with $E(TS') = E(TT')$ and $E(US') = E(TS')$ as well. Finally,

$$E(UU') = \begin{pmatrix} 1 & 0 & 0 \\ & 1+\tau^2 & \rho \\ & & 1 \end{pmatrix}.$$

For the dichotomized choice of predictors $T = T^{(D)}$ and $U = U^{(D)}$,

$$E(TT') = \begin{pmatrix} 1 & \frac{1}{2} & 0 \\ & \frac{1}{2} & \sqrt{\frac{1}{2\pi}}\rho \\ & & 1 \end{pmatrix},$$

$$E(TS') = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & (1-\omega)\sqrt{\frac{1}{2\pi}} + \omega\sqrt{\frac{\pi}{8}} & \sqrt{\frac{1}{2\pi}}\rho \\ 0 & \rho & 1 \end{pmatrix},$$

$$E(UU') = \begin{pmatrix} 1 & \frac{1}{2} & 0 \\ & \frac{1}{2} & \sqrt{\frac{1}{2\pi}}\frac{\rho}{\sqrt{1+\tau^2}} \\ & & 1 \end{pmatrix},$$

and

$$E(US') = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & (1-\omega)\sqrt{\frac{1}{2\pi}}\frac{1}{\sqrt{1+\tau^2}} + \omega\sqrt{2\pi}\; k(\tau,\rho) & \sqrt{\frac{1}{2\pi}}\frac{\rho}{\sqrt{1+\tau^2}} \\ 0 & \rho & 1 \end{pmatrix},$$

where

$$\begin{aligned} k(\tau,\rho) &= Pr(X^* > 0, X > 0) - Pr(X^* > 0)Pr(X > 0) \\ &= E\{\Phi(X) I_{(-\infty,0)}(X)\} \end{aligned}$$

must be evaluated numerically.

In Section 5, $E(SS')$ has the form (19), while calculation of $E(TT')$ and $E(TS')$ proceeds easily upon noting that

$$E\left(X\ I\left\{\frac{i}{r} < \Phi(X) \leq \frac{i+1}{r}\right\}\right) \quad = \quad \phi(c_i) - \phi(c_{i+1})$$

and

$$E\left(W\ I\left\{\frac{i}{r} < \Phi(X) \leq \frac{i+1}{r}\right\}\right) \quad = \quad \rho\{\phi(c_i) - \phi(c_{i+1})\},$$

where $c_i = \Phi^{-1}(i/r)$ and $\phi(\cdot)$ is the standard normal density function. Similarly, calculation of $E(UU')$ and $E(US')$ follows from

$$E\left(X\ I\left\{\frac{i}{r} < \Phi(X^*) \leq \frac{i+1}{r}\right\}\right) \quad = \quad \frac{1}{\sqrt{1+\tau^2}}\left\{\phi\left(\frac{c_i}{\sqrt{1+\tau^2}}\right) - \phi\left(\frac{c_{i+1}}{\sqrt{1+\tau^2}}\right)\right\}$$

and

$$E\left(W\ I\left\{\frac{i}{r} < \Phi(X^*) \leq \frac{i+1}{r}\right\}\right) \quad = \quad \frac{\rho}{\sqrt{1+\tau^2}}\left\{\phi\left(\frac{c_i}{\sqrt{1+\tau^2}}\right) - \phi\left(\frac{c_{i+1}}{\sqrt{1+\tau^2}}\right)\right\}.$$

References

Box, G.E.P. (1979). Robustness in the strategy of scientific model building. In *Robustness in Statistics*, eds. R.L. Launer and G.N. Wilkinson, pp. 201-236. London: Academic Press.

Carroll, R.J., Maca, J.D., and Ruppert, D. (1999). Nonparametric regression with errors in covariates. *Biometrika* **86**, 541-554.

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

Fan, J. and Truong, Y.K. (1993). Nonparametric regression with errors in variables. *Annals of Statistics* **21**, 1900-1925.

Gould, A. and Lawless, J.F. (1988). Consistency and efficiency of regression coefficient estimates in location-scale models. *Biometrika* **75**, 535- 540.

Gustafson, P. and Le, N.D. (2001). A comparison of continuous and discrete measurement error: is it wise to dichotomize imprecise covariates? Submitted. Available at
http://www.stat.ubc.ca/people/gustaf.

Kent, J.T. (1982). Robust properties of likelihood ratio tests. *Biometrika* **69**, 19-27.

Ramsey, J.B. (1969). Tests for specification errors in classical least-squares regression analysis. *Journal of the Royal Statistical Society, Series B* **31**, 350-371.

Thomas, D., Stram, D. and Dwyer, J. (1993). Exposure measurement error: influence on exposure-disease relationships and methods of correction. *Annual Review of Public Health* **14**, 69-93.

White, H. (1981). Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association* **76**, 419- 433.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1- 26.

Willett, W. (1989). An overview of issues related to the correction of non-differential exposure measurement error in epidemiologic studies. *Statistics in Medicine* **8**, 1031-1040.