**PI Name, Affiliation and Contact Address**: Elaine Symanski, PhD, Division of Epidemiology and Disease Control, University of Texas School of Public Health, 1200 Herman Pressler Dr., RAS 643, Houston, TX 77030; Phone number: 713-500-9238; Email: Elaine.symanski@uth.tmc.edu

**Institution to which the award was made**: University of Texas Health Science Center at Houston, Post Office Box 20036, Houston, Texas 77225-0036

**Project title:** Evaluation of Exposure Measurement Error

**Date:** December 19, 2007

**Co-investigators:** Wenyaw Chan, PhD
Mary Ann Smith, PhD

**Grant number:** 5 R03 OH 007834

**Starting and Ending Dates:** 9/30/2003-8/31/2006

TABLE OF CONTENTS

## List of Terms and Abbreviations

ALA – δ-aminolevulinic acid
ANOVA – Analyis of Variance
AR(1) – Autoregressive (first-order)
CDF – Cumulative Distribution Function
CS – Compound Symmetry
DMA – dimethylarsonic acid
ICC – Intra-class Correlation Coefficient
ISIC – International Standard Industrial Classification
MMA – methylarsonic acid
OLS – Ordinary Least Squares (OLS)
REML – Restricted Maximum Likelihood
TDA – 2,4-toluenediamine

## ABSTRACT

In a simple linear regression model with a latent independent random variable, the slope coefficient is smaller than it should be when a surrogate (i.e., error-prone) measure is used instead. The ratio of the slope coefficients from the regression models using the surrogate or latent variables reflects the attenuation bias. This bias can also be expressed as the ratio of the variance of the latent variable to the sum of the variances of the latent and surrogate variables. The traditional approach for estimating the attenuation bias uses this ratio and its estimated variances. Recognizing that the slope ratio cannot be estimated because the latent variable is unobservable, we developed a Bayesian-type estimator using the observations from the surrogate and outcome variables. One objective of this study was to examine the distributional behavior of the traditional and proposed estimators of attenuation bias. In the empirical study, we found that if the focus is on estimating attenuation bias, then it is preferable to use the traditional method; if the focus is on estimating the latent slope and the error variance of the simple linear regression is relatively large as compared to the variance of the latent variable, then it is preferable to use the proposed method; and if the focus is on estimating the latent slope and the error variance of the simple linear regression is relatively small as compared to the variance of the latent variable, then the traditional method may be preferable because it has a higher probability of producing estimates that are closer to the latent slope even though it has a larger mean bias. Finally, the traditional method was more likely to encounter numerical difficulties and thus produced a higher proportion of estimates that could not be obtained. In this case, the proposed method would provide a valid alternative for estimating the latent slope or the attenuation bias.

For a given set of outcome and independent surrogate measurements in a simple linear regression, we expressed the proposed latent slope and attenuation bias in an almost explicit form that can be used for approximation. These forms provide us with an alternative computational method for finding the estimates, other than the simulation method that was also presented. In addition, we found that the conditional distribution of the latent slope given the observed outcome and surrogate variables that were used for the proposed estimator, is proportional to the ratio of a noncentral chi-squared variate to a normal variate.

Due to non-dependence of the numerator and the denominator of the latent slope, the proposed estimators of the latent slope and the attenuation bias are complicated. This leads to difficulty in comparing the distribution of the attenuation bias derived from the proposed method with the traditional method known as an F or approximated F distribution depending upon whether the surrogate measurements are balanced or unbalanced. For the latent slope, the estimator using the traditional method is complicated and its distribution is not yet identified.

Further, we examined attenuation bias using a large database of multiple exposure measures collected on workers in a variety of different workplaces. Our findings indicate that there was more measurement error (attenuation bias) in biomarkers with shorter rather than longer half-lives and in exposure measurements collected via personal rather than biological sampling, although both types of monitoring data provided examples of exposure measures fraught with error. Our results also indicate substantial imprecision in the estimates of exposure measurement error, suggesting that greater attention needs to be given in the future to studies that collect sufficient data to allow for a better characterization of the attenuating effects of an error-prone exposure measure.

## HIGHLIGHTS/SIGNIFICANT FINDINGS

In a simple linear regression model with a latent independent random variable, the slope coefficient is smaller than it should be when a surrogate (i.e., error-prone) measure is used instead. The ratio of the slope coefficients from the regression models using the surrogate or latent variables reflects the attenuation bias. This bias can also be expressed as the ratio of the variance of the latent variable to the sum of the variances of the latent and surrogate variables. The traditional approach for estimating the attenuation bias uses this ratio and its estimated variances. Recognizing that the slope ratio cannot be estimated because the latent variable is unobservable, we developed a Bayesian-type estimator using the observations from the surrogate and outcome variables. Our results suggest it is preferable to use the traditional method if the focus is on estimating attenuation bias. If the focus is on estimating the latent slope and the error variance of the simple linear regression is relatively large as compared to the variance of the latent variable, then it is preferable to use the proposed method. On the other hand, if the error variance is relatively small as compared to the variance of the latent variable, then the traditional method has a higher probability of producing estimates that are closer to the latent slope but it has a larger mean bias. We also found that the traditional method was more likely to encounter numerical difficulties and thus produced a higher proportion of estimates that could not be obtained. In this case, the proposed method would provide a valid alternative for estimating the latent slope or the attenuation bias.

For a given set of outcome and independent surrogate measurements in a simple linear regression, we have expressed the proposed latent slope and attenuation bias in an almost explicit form that can be used for approximation. These forms can provide us with an alternative computational method for finding the estimates, other than the simulation method that was also presented. In addition, we found that the distributional behavior of the proposed estimator is proportional to the ratio of a noncentral chi squared variate to a normal variate.

We also examined attenuation bias using a large database of multiple exposure measures collected on workers in a variety of different workplaces and found that there was more attenuation bias in biomarkers with shorter rather than longer half-lives and in exposure measurements collected by personal rather than biological sampling. Our results also indicate substantial imprecision in the estimates of attenuation bias, suggesting that greater attention needs to be given in the future to ensure that sufficient data are collected to improve exposure assessment and diminish the attenuating effects of an error-prone exposure measure.

## TRANSLATION OF FINDINGS OUTCOMES/RELEVANCE/IMPACT

In a simple linear regression model with a latent independent random variable, the slope coefficient is smaller than it should be when a surrogate (i.e., error-prone) measure is used instead. Traditional methods focus on the study of the attenuation bias, i.e., on the ratio of the slope coefficients from the regression models using the surrogate or latent variables. This research focused on developing an estimator for the latent slope using a surrogate variable when the independent variable is unobservable. We also examined the distributional behavior of different estimators of attenuation bias. Given that exposure assessments are always fraught with

error due to intra-individual variability in exposure and errors in sampling and analysis, our results provide a framework for better understanding the nature of bias in studies of adverse health effects of occupational exposures.

We also examined attenuation bias using a large database of multiple exposure measures collected on workers in a variety of different workplaces and found that there was more attenuation bias in biomarkers with shorter rather than longer half-lives and in exposure measurements collected by personal rather than biological sampling. Finally, our results also indicated substantial imprecision in the estimates of attenuation bias, suggesting that greater attention needs to be given in the future to ensure that sufficient data are collected to improve exposure assessment and diminish the attenuating effects of an error-prone exposure measure.

## SCIENTIFIC REPORT

### Background for the Project

**Sources of exposure variability**: Two major components of exposure variability are the variation that occurs from day-to-day (intra-individual variation) and the variation that occurs among workers (inter-individual variation). For occupational exposures, intra-individual variation arises as a result of numerous factors related to the process and the work environment and to a lesser degree, to measurement errors associated with sampling and analysis (Nicas et al., 1991). Inter-individual variation in airborne contaminant levels arises as a result of differences in workers' job tasks or work practices (Rappaport, 1991). The intra-individual source of variation in biomarkers of exposure is influenced by the intra-individual variation in the external exposure. The degree to which fluctuations in external exposures are transmitted to biological levels of the contaminant or its metabolite in bodily fluids is heavily dependent on the half-life of the particular contaminant in the body (Rappaport, 1985). Fluctuations in levels of biomarkers with short half-lives predominantly reflect variation in air levels over time whereas there is physiological dampening of variability in biomarkers with longer half-lives (Droz and Wu, 1991). With respect to inter-individual variability in biomarker levels, differences in the external exposure received by workers, effects due to ethnicity, gender, and age, anthropometric and lifestyle factors, and physiologic differences in the rates of uptake, distribution, and elimination are all likely to play a role.

Intra- and inter-individual sources of variability have been characterized for a large number of airborne exposures arising in myriad industries worldwide (Kromhout et al., 1987; Heederik et al., 1991; Burdorf et al., 1994; Woskie et al., 1994; Kromhout and Heederik, 1995; Nieuwenhuijsen et al., 1995; Kumagai et al., 1996; Milton et al., 1996; Lagorio et al., 1997; Peretz et al., 1997; deCock et al., 1998; Makinen et al., 2000), as well as for biological measures of exposure (Rappaport et al., 1995; Simcox et al., 1999; Symanski et al., 2000 Symanski et al., 2001b; Hines and Deddens, 2001). In addition, two sizable databases of personal (Kromhout et al., 1993) and biological (Symanski and Greeson (2002) monitoring data provided an opportunity to compare sources of exposure variability among groups of workers in different industries. These studies have typically relied upon the one-way random-effects model to quantify the intra- and inter-individual sources of variation in exposure through estimation of the so-called within- and between-worker variance components.

**Consequences of measurement error**: Intra-individual variability in exposure can induce error and thereby can adversely affect epidemiologic studies by diminishing measures of effect and the power to detect significant effects (Thomas et al., 1993; Armstrong, 1998). The type of error that is introduced depends upon the nature of the exposure assessment, i.e., whether an individual- or group-based approach is adopted. In the former case, individual workers' exposures are estimated in which classical measurement error may be present whereas in estimating exposures for a group of workers Berkson error predominates (Armstrong, 1998).

**Linear regression and the effects of classical measurement error**: We assume that an underlying long-term exposure level is the relevant exposure index in determining the risk of an adverse health outcome. Let $Z_i$ represent an imperfect measure of true exposure for the *i-th* individual, $X_i$, so that the additive measurement error $\varepsilon_i$ is defined by:

$$Z_i = X_i + \varepsilon_i \qquad 1$$

In equation 1 it is assumed that $\varepsilon_i$ is distributed normally with mean 0 and variance $\sigma_W^2$, and that $X_i$ is distributed normally with mean and variance, $\mu$ and $\sigma_B^2$, respectively. It directly follows that $Z_i$ is distributed normally with a mean of $\mu$ and a variance of $\sigma_Z^2$ where $\sigma_Z^2 = \sigma_B^2 + \sigma_W^2$.

We assume that $Y_i$, the response or outcome of the *i-th* individual, is adequately explained by a simple linear regression model:

$$Y_i = \alpha + \beta X_i + e_i \qquad 2$$

where $\alpha$ is the intercept and $\beta$ is the slope. In model 2, we assume that $e_i$ is distributed normally with a mean of 0 and variance $\sigma_1^2$. Given the distributional assumptions that are made about the $X_i$'s in model 2 and that $X_i = x_i$, $Y_i$ is distributed normally with a mean and variance designated as $\mu_Y = \alpha + \beta x_i$ and $\sigma_Y^2$, respectively. Thus, the outcome variable ($Y_i$) depends on true exposure through the regression coefficient, $\beta$.

If all relevant explanatory variables are included in model 2 and under the standard assumptions of ordinary least squares (OLS) estimation, namely $E(e_i) = 0$ and $Cov(X_i, e_i) = 0$, we can construct consistent estimates of regression parameters. However, this is not possible when the true exposure, $X_i$, is unobservable.

When true exposure $X_i$ is unobservable, we rely on the imperfect measure $Z_i$ [see equation 1]. For individuals whose true exposure levels have been estimated by $Z_i$, the exposure–response relation, under expectation, becomes (Thomas et al., 1993):

$$
\begin{aligned}
E(Y_i|Z_i) &= E\left[E(Y_i|X_i|Z_i)\right] \\
&= E\left[\alpha+\beta X_i|Z_i\right] \\
&= \alpha+\beta\left[cZ_i+(1-c)\mu\right]
\end{aligned}
\qquad 3
$$

where $c=\dfrac{\sigma_B^2}{\sigma_B^2+\sigma_W^2}$.

Note that in equation 3 the third equality is calculated through the posterior distribution by treating the distribution of $X_i$ as a prior.

Thus,

$$
\begin{aligned}
E(Y_i|Z_i) &= \alpha+\beta cZ_i+\beta(1-c)\mu \\
&= a+bZ_i
\end{aligned}
\qquad 4
$$

where $a=\alpha+\beta(1-c)\mu_X$ and $b=\beta c$.

The ratio of the 'observed' coefficient (often referred to as the 'naïve' coefficient) to the true coefficient becomes:

$$\frac{b}{\beta}=\frac{\sigma_B^2}{\sigma_B^2+\sigma_W^2} \qquad 5$$

Thus, regressing $Y_i$ on the imperfect exposure measure $Z_i$, instead of true exposure $X_i$, yields a slope coefficient $\beta'$ that is attenuated. The factor $\sigma_B^2/\left(\sigma_B^2+\sigma_W^2\right)$ represents the degree of true attenuation of the true slope coefficient $\beta$. Using conventional notation, we define $\hat{b}$ as an estimator of $b$ and note that in the simple linear regression model with imperfectly measured exposure, $\hat{b}$ is a consistent estimator of $\beta\left[\sigma_B^2/\left(\sigma_B^2+\sigma_W^2\right)\right]$.

Similarly, if we observe $n$ $Z_{ij}$'s for each $X_i$ and compare the true exposure-response regression with the regression of $Y_i$'s on $\overline{Z}_i$, then, since $E(X_i|Z_{i1},\ldots,Z_{in})=E(X_i|\overline{Z}_i)$, the ratio of the observed coefficient to the true coefficient becomes:

$$\frac{b}{\beta}=\frac{\sigma_B^2}{\sigma_B^2+\dfrac{\sigma_W^2}{n}} \qquad 6$$

As before, regressing $Y_i$ on measured exposure $\overline{Z}_i$, instead of true exposure $X_i$, yields a slope coefficient $b$ that is attenuated. The factor $\sigma_B^2/\left(\sigma_B^2+\sigma_W^2/n\right)$ represents the degree of true attenuation of the (true) slope coefficient $\beta$.

**Corrections for measurement error**: If true exposure is unobservable, replication studies can be carried in which repeated measurements of an error-prone exposure measure are obtained on a subset of the study population. Suppose repeated observations are available on a sample of $n$ individuals. The *j-th* observation of the surrogate measure $Z_i$ on the *i-th* individual will be denoted by $Z_{ij}$. A random-effects model (Searle et al., 1992) for $Z_{ij}$ is:

$$Z_{ij} = \mu + w_i + \varepsilon_{ij} \qquad 7$$

where $w_i$ and $\varepsilon_{ij}$ are independently distributed random variables, $w_i = X_i - \mu$ represents the random intercept of model 7 and is distributed normally with a mean 0 and variance $\sigma_B^2$, and $\varepsilon_{ij}$ denotes the random deviation of the *j-th* measurement from the *i-th* individual's mean value and is distributed normally with a mean 0 and variance $\sigma_W^2$. In model 7, $E(Z_{ij}) = X_i$, $Var(Z_{ij}) = \sigma_Z^2 = \sigma_B^2 + \sigma_W^2$, and $Cov(Z_{ij}, Z_{ij'}) = \sigma_B^2$ for $j \neq j'$. The correlation between two measurements collected on the same person is $Corr(Z_{ij}, Z_{ij'}) = \left[\sigma_B^2 / \left(\sigma_B^2 + \sigma_W^2\right)\right]$ for $j \neq j'$, which is commonly referred to as the intra-class correlation coefficient. The attenuation bias is now identified with the intra-class correlation coefficient of the one-way random-effects model. An estimator of the intra-class correlation coefficient can be obtained by substituting each variance parameter with its estimated value, i.e., $\hat{\sigma}_B^2 / \left(\hat{\sigma}_B^2 + \hat{\sigma}_W^2\right)$ (Donner, 1986). While this estimator is consistent, it is not unbiased (however, the degree of bias is quite small) (Donner, 1986).

**Estimators of attenuation bias**: Replication studies provide estimates of the variance components, $\sigma_B^2 + \sigma_W^2$, which can be used in turn to estimate the intra-class correlation coefficient by $\hat{\sigma}_B^2 / \left(\hat{\sigma}_B^2 + \hat{\sigma}_W^2\right)$. The estimated intra-class correlation can then be applied as a correction factor to adjust regression results that examine exposure-related effects based upon single measurements to estimate each study participant's exposure. Previous studies that have quantified measurement error in occupational exposure measurements have all reported point estimates of the attenuation bias without regard to sample size or other factors that might affect the reliability of such estimates.

As indicated in equation 5 above, the ratio of the 'observed' to the true slope coefficient reflects the attenuation bias, which is equal to the ratio of the variance of true exposure to the sum of the variances of true exposure and the imperfect measure of exposure. Thus, one expression for the true attenuation bias relies on a ratio measure of slope coefficients and another on a ratio of variances. In terms of expectation, they are equal to each other.

Given the different expressions for attenuation bias, it follows that different forms of the estimators can be constructed, namely, as a ratio of two estimated slope coefficients or as a ratio of estimated variances. In the former case, the estimator is a ratio of normally distributed variates, and in the latter case, the estimator is a ratio of approximately chi-squared distributed variates. This implies that the two estimators are not equal in distribution. The sampling

distribution of the estimator of the ratio of the variances of true to measured exposure (i.e., the intra-class correlation coefficient) has been derived in closed form for balanced data (Donner and Koval, 1983). In the case of unbalanced data, an approximate distribution for the estimator of the intra-class correlation coefficient has been proposed (Donner, 1986) and is the same as that of a transformation of an F variate. However, there have been no investigations, to our knowledge, that have examined the estimator of attenuation bias as a ratio of estimated slope coefficients.

## Specific Aims

**Specific Aim #1** – Evaluate the distributional behavior of the estimators of attenuation bias using empirical methods. In a simulation study, investigate the properties of the estimators for distinct combinations of sample size and of the parameters in the regression and measurement error models. Compare the impact of different parameter values, as well as different sample sizes under both a balanced and unbalanced design structure, on the behavior of the estimators of attenuation bias.

**Specific Aim #2** – Investigate the characteristics of the estimators of attenuation bias by analytically examining their distributional behavior. The comparison will focus on the distribution of the ratio of the variance of the true exposure to the variance of the error-prone exposure measure and the conditional distribution of the ratio of the estimated observed slope and the estimated true slope, given error-prone exposure and measured outcome. The moments of the distributions of both estimators will be computed and compared to each other and to the empirical results obtained in Aim #1.

**Specific Aim #3** – Estimate the degree of measurement error in multiple exposure measures collected simultaneously in a worker population. For each exposure measure, evaluate likely effects of measurement error on regression results. Make comparisons between exposure measures taking into account kinetic considerations that may affect or contribute to variability in levels of biomarker of exposure.

## Specific Aim #1

*Methods*

As indicated in equation 2 above, we denote the simple linear regression equation with unobservable predictor $X_i$ and outcome variable $Y_i$ as $Y_i = \alpha + \beta X_i + e_i, i = 1, 2, \ldots, n$, where $e_i \sim N\left(0, \sigma_1^2\right)$, and the unobservable $X_i$ is assumed to follow the $N\left(\mu, \sigma_x^2\right)$ distribution. Let $Z_i$ be a surrogate variable used to assess $X_i$, $i = 1, 2, \ldots, n$. Therefore, the linear regression equation between the outcome $Y_i$ and surrogate predictor $Z_i$, can be written as

$$Y_i = a + bZ_i + d_i, i = 1, 2, \ldots, n \qquad 8$$

where $d_i \sim N\left(0, \sigma_2^2\right)$. It is common to assume that, for any $i$, as indicated in equation 1 above, that $Z_i = X_i + \varepsilon_i$ where $\varepsilon_i \sim N\left(0, \sigma_W^2\right)$, $i = 1, 2, \ldots, n$.

Assume that $X_i, Z_i$ and $Y_i$, $i = 1, 2, \ldots, n$ are jointly normally distributed. The aforementioned assumptions imply that the random vector $\left(X_i, Z_i, Y_i\right)^T$ have a trivariate normal distribution with mean vector $E\left[\left(X_i, Z_i, Y_i\right)^T\right] = \left(\mu, \mu, \alpha + \beta\mu\right)^T = \left(\mu, \mu, a + b\mu\right)^T$ and variance-covariance matrix

$$\begin{pmatrix} \sigma_B^2 & \sigma_B^2 & b\sigma_B^2 \\ \sigma_B^2 & \sigma_B^2 + \sigma_W^2 & b\left(\sigma_B^2 + \sigma_W^2\right) \\ b\sigma_B^2 & b\left(\sigma_B^2 + \sigma_W^2\right) & b^2\left(\sigma_B^2 + \sigma_W^2\right) + \sigma_2^2 \end{pmatrix}.$$

The attenuation of the regression slope can be expressed as $b/\beta$ or $\sigma_B^2/\left(\sigma_B^2 + \sigma_W^2\right)$, hereafter referred to as the slope or variance ratio, respectively. The traditional method of estimating attenuation is known to be $\hat{\sigma}_B^2/\left(\hat{\sigma}_B^2 + \hat{\sigma}_W^2\right)$. Using this method, one can recover the latent slope $\hat{\beta}$ as $\dfrac{\hat{b}}{\hat{\sigma}_B^2/\left(\hat{\sigma}_B^2 + \hat{\sigma}_W^2\right)}$. In this project, we propose to use $E\left(\hat{\beta}\middle|\begin{pmatrix}\mathbf{Z}\\ \mathbf{Y}\end{pmatrix} = \begin{pmatrix}\mathbf{z}\\ \mathbf{y}\end{pmatrix}\right)$ as the estimated latent slope and $\dfrac{\hat{b}}{E\left(\hat{\beta}\middle|\begin{pmatrix}\mathbf{Z}\\ \mathbf{Y}\end{pmatrix} = \begin{pmatrix}\mathbf{z}\\ \mathbf{y}\end{pmatrix}\right)}$ as another method of estimating attenuation.

In this simulation study, we first generate the random variate $X_i$ for each of 10 subjects using $N\left(\mu, \sigma_B^2\right)$ distribution. The surrogate variable $Z_i$ for each subject was then generated from $x_i$ using equation 1. Since multiple observations of $Z_i$ are needed for estimating $\sigma_B^2$ and $\sigma_W^2$, four replicated $Z_i$ values were generated for each subject. The outcome variable $Y_i$ was then generated by equation 8.

For the relevant parameters involved in generating the data, we chose convenient values that can best represent different scenarios. They are $\mu = 3$, $\sigma_B^2 = 1$, $\sigma_W^2 = 0.1, 1, 4,$ and $10$, $\alpha = 1$, $\beta = 0.5$ and $\sigma_1^2 = 0.1, 1,$ and $10$. In addition, at the beginning of the simulation, we eliminate 500 random numbers as the burn-in period. For each parameter combination, there were 1000 simulation runs performed. In each run, the value for $\dfrac{\hat{b}}{\hat{\sigma}_B^2/\left(\hat{\sigma}_B^2 + \hat{\sigma}_W^2\right)}$ is calculated. In order to calculate the conditional slope, 1000 more "$X_i$" for each run, are generated based upon the

conditional distribution of $X\left|\begin{pmatrix}\mathbf{Z}\\ \mathbf{Y}\end{pmatrix}=\begin{pmatrix}\overline{\mathbf{z}}\\ \mathbf{y}\end{pmatrix}\right.$, where each $\overline{z}_i$ in $\overline{\mathbf{z}}$ is the average of 4 replicated $Z_i$ generated in the simulation run. These generated $X_i$s, along with their corresponding $z_i, y_i$ are used for calculating $E\left(\hat{\beta}\left|\begin{pmatrix}\mathbf{Z}\\ \mathbf{Y}\end{pmatrix}=\begin{pmatrix}\mathbf{z}\\ \mathbf{y}\end{pmatrix}\right.\right)$. The estimates of recovered beta from each method, $\dfrac{\hat{b}}{\hat{\sigma}_B^2/\left(\hat{\sigma}_B^2+\hat{\sigma}_W^2\right)}$ and $E\left(\hat{\beta}\left|\begin{pmatrix}\mathbf{Z}\\ \mathbf{Y}\end{pmatrix}=\begin{pmatrix}\mathbf{z}\\ \mathbf{y}\end{pmatrix}\right.\right)$, are compared with the true $\beta$. In addition, the difference between each estimate of the recovered beta and the true $\beta$ is computed, along with the proportion of runs in which the proposed method $E\left(\hat{\beta}\left|\begin{pmatrix}\mathbf{Z}\\ \mathbf{Y}\end{pmatrix}=\begin{pmatrix}\mathbf{z}\\ \mathbf{y}\end{pmatrix}\right.\right)$ is closer to $\beta$ as compared to the traditional method. Note that the $\hat{\beta}$ can only be obtained through simulation since $X_i$ is an unobservable predictor of $Y_i$.

*Results and Discussion*

Table 1 presents of proportion of runs in which the proposed method is better than the traditional method in estimating the recovered beta and attenuation (1000 runs), along with the number of estimates that could not be computed for reasons such as limitations in the numerical procedures implemented by the software. In these comparisons, only simulation runs that have estimates from both methods available are counted. Thus, the number of simulations contributing data varied slightly. In general, the proposed method has a higher proportion of recovered slopes that is closer to the "real" latent slope than that using the traditional method.

However, when $\sigma_1^2$ is very small, the traditional method performs slightly better. For estimating the attenuation the traditional method is clearly better. However, there were more estimates that could not be computed using the traditional method than the proposed method, especially when the surrogate variable had a higher within-subject variance than between-subject variance. This is partially due to the implementation of the mixed model procedures.

Table 2 presents the comparisons of the mean estimates of the recovered beta and attenuation from each method. Again, the proposed method is much better than the traditional method in estimating the recovered better, especially when $\sigma_W^2$ is not smaller than $\sigma_B^2$. When the latter is much smaller than the former and $\sigma_W^2$ is moderate, the traditional method performs poorly in estimating the recovered latent slope. When $\sigma_W^2$ is smaller than $\sigma_B^2$, $\hat{\beta}$ calculated from the proposed method is almost identical to the true latent slope. On the other hand, for estimating the attenuation, the traditional method is better than the proposed method in all scenarios.

Table 1: Proportion of runs in which the proposed method is closer to $\beta$ as compared to the traditional method in estimating the recovered beta and the attenuation bias using the following parameters, $\beta = 0.5$ *and* $\sigma_\beta^2 = 1$, along with the noted values for $\sigma_W^2$ and $\sigma_\iota^2$

| $\sigma_W^2$ | $\sigma_\iota^2$ | Recovered beta | Attenuation | No. of missing runs in the simulation |
|---|---|---|---|---|
| 0.1 | 0.1 | 0.496 (496/1000) | 0.283 (283/1000) | 0 |
| 0.1 | 1 | 0.575 (575/1000) | 0.278 (278/1000) | 0 |
| 0.1 | 10 | 0.802 (802/1000) | 0.280 (280/1000) | 0 |
| 1 | 0.1 | 0.482 (477/989) | 0.315 (312/989) | Proposed = 6, traditional = 5 |
| 1 | 1 | 0.593 (590/995) | 0.290 (289/995) | traditional = 5 |
| 1 | 10 | 0.836 (832/995) | 0.283 (282/995) | traditional = 5 |
| 4 | 0.1 | 0.415 (348/838) | 0.297 (249/838) | Proposed = 27, traditional = 135 |
| 4 | 1 | 0.594 (511/861) | 0.254 (219/861) | Proposed = 139, traditional = 135 |
| 4 | 10 | 0.837 (724/865) | 0.095 (82/865) | Proposed = 0 for $E(\hat{\beta} \mid \bar{z}, y)$ and 1 for $\hat{b} / E(\hat{\beta} \mid \bar{z}, y)$, traditional = 135 |
| 10 | 0.1 | 0.381 (243/637) | 0.223 (142/637) | Proposed = 44, traditional = 319 |
| 10 | 1 | 0.568 (381/671) | 0.207 (139/671) | Proposed = 10, traditional = 319 |
| 10 | 10 | 0.815 (549/674) | 0.200 (135/674) | Proposed = 7, traditional = 319 |

Table 2: Mean of recovered beta with parameters $\beta = 0.5$ *and* $\sigma_B^2 = 1$

| | | Unobservable slope[1] | Recovered beta | | True attenuation | Estimated Attenuation | |
|---|---|---|---|---|---|---|---|
| $\sigma_W^2$ | $\sigma_1^2$ | $\hat{\beta}$ | Proposed[2] (number available) | Traditional[3] (number available) | $\frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2 / k}$ | Proposed[4] (number available) | Traditional[5] (number available) |
| 0.1 | 0.1 | 0.503 | 0.499 (1000) | 0.507 (1000) | | 0.987 (1000) | |
| 0.1 | 1 | 0.508 | 0.503 (1000) | 0.512 (1000) | 0.976 | 0.987 (1000) | 0.970 (1000) |
| 0.1 | 10 | 0.526 | 0.519 (1000) | 0.528 (1000) | | 0.986 (1000) | |
| 1 | 0.1 | 0.503 | 0.471 (989) | 1.873 (989) | | 0.877 (994) | |
| 1 | 1 | 0.508 | 0.481 (995) | 3.137 (995) | 0.800 | 0.889 (1000) | 0.756 (995) |
| 1 | 10 | 0.526 | 0.539 (995) | 7.156 (995) | | 0.890 (1000) | |
| 4 | 0.1 | 0.503 | 0.402 (838) | 1.070 (838) | | 0.619 (973) | |
| 4 | 1 | 0.508 | 0.396 (861) | 1.265 (861) | 0.5 | 0.745 (861) | 0.486 (865) |
| 4 | 10 | 0.526 | 0.392 (865) | 1.900 (865) | | 0.535 (999) | |
| 10 | 0.1 | 0.503 | 0.329 (637) | 1.419 (637) | | 0.433 (956) | |
| 10 | 1 | 0.508 | 0.341 (671) | 1.108 (671) | 0.286 | 0.430 (990) | 0.376 (681) |
| 10 | 10 | 0.526 | 0.288 (674) | 0.299 (674) | | 0.566 (993) | |

[1]Estimated slope from the regression of $Y_i$ on $X_i$ (where $X_i$ is a latent variable that is only 'observable' in simulation).

[2] $E\left(\hat{\beta} \middle| \hat{z}, y\right)$ [3] $\hat{b} \Big/ \left( \frac{\hat{\sigma}_B^2}{\hat{\sigma}_B^2 + \hat{\sigma}_W^2 / 4} \right)$ [4] $\frac{\hat{b}}{\hat{\beta}}$ [5] $\frac{\hat{\sigma}_B^2}{\hat{\sigma}_B^2 + \hat{\sigma}_W^2 / 4}$

The empirical cumulative distribution functions (CDFs) of the recovered latent slope using the proposed and the traditional methods in the simulation analysis (1000 runs) are shown in Appendix A. Each CDF is also compared with the "real" latent slopes, which are only available in simulation. In instances when $\sigma_W^2$ is small compared to $\sigma_B^2$, both methods result in almost identical distributions. When $\sigma_W^2$ is the same as $\sigma_B^2$, the distribution of the recovered beta using the proposed method is very close to that of the "real" latent slope. For most cases, the traditional method results in a distribution of the recovered slope that is shifted to the right. When $\sigma_W^2$ is larger than $\sigma_B^2$, the proposed method is likely to generate an outlier as it can be seen

from the CDF graphs. Note that, for a given data set, the proposed method can produce a distribution of the estimated latent slope, while the traditional method can only provide a single estimate. When researchers are using the proposed method, they have options to choose any summary statistic (such as mean, trimmed mean or median) from the generated estimates to estimate the recovered beta, depending on their outlying scenarios or their research needs.

The empirical cumulative distribution functions of the estimated attenuation using the proposed and the traditional methods from the 1000 simulated runs of different scenarios are presented in Appendix B. Clearly, from its algebraic formula, the traditional method does not depend on $\sigma_1^2$, as shown in equation 6 or described in the first paragraph of the section entitled, 'Estimators of attenuation bias'. These figures also show that the distributions of the estimated attenuation using the proposed method is not as sensitive to changes in the magnitude of $\sigma_1^2$, especially when $\sigma_W^2$ is relatively small compared to $\sigma_B^2$.

*Conclusions*

Our results from the empirical study allows us to evaluate the distributional behavior of different estimators of the recovered latent slope and attenuation bias and to make recommendations when using a surrogate variable to estimate the latent slope of a simple linear regression:

1. If focus is on estimating the attenuation bias, then it may be preferable to use the traditional method;
2. If focus is on estimating the latent slope and the error variance of the simple linear regression is relatively large as compared to the variance of the latent variable, then use the proposed method;
3. If the focus is on estimating the latent slope and the error variance of the simple linear regression is relatively small as compared to the variance of the latent variable, then the traditional method has a higher probability of producing estimates that are closer to the latent slope but it has a larger mean bias; and,
4. If traditional method fails to produce an estimate, the proposed method provides a valid alternative for estimating the recovered beta or for estimating the attenuation bias. Note that failure to produce an estimate is more likely to occur in the traditional method than in the proposed method.

**Specific Aim #2**

*Methods*

The commonly used estimator $\hat{\sigma}_B^2/\left(\hat{\sigma}_B^2+\hat{\sigma}_W^2\right)$ has been shown to have a chi-squared distribution when the surrogate variable $Z_i$ is repeatedly observed $k$ times for each subject. When each subject has a different number of repeated measurements on $Z_i$, $\hat{\sigma}_B^2/\left(\hat{\sigma}_B^2+\hat{\sigma}_W^2\right)$ has an approximate chi-squared distribution (see Searle et al., (1992)) Although the different

functional forms $\sigma_B^2/(\sigma_B^2+\sigma_W^2)$ and $b/\beta$ are identical in expressing the slope attenuation, their corresponding point estimates use different principles of estimation. In the former instance, the estimate $\hat{\sigma}_B^2/(\hat{\sigma}_B^2+\hat{\sigma}_W^2)$ can be computed directly using either the maximum likelihood or restricted maximum likelihood estimation method (particularly for unbalanced realizations of the surrogate variable) if the invariance property applies. In contrast, the estimate $\hat{b}/\hat{\beta}$ cannot be obtained directly, since $X_i$'s are not observable. However, $E\left(\hat{b}/\hat{\beta}\middle| Z_i's \text{ and } Y_i's\right)$ can be interpreted in a Bayesian framework and can serve as the estimator for $b/\beta$.

This project examined the conditional distributions of the estimated attenuation $\hat{b}/\hat{\beta}$ and the estimated latent slope $\hat{\beta}$, given $(\mathbf{Z},\mathbf{Y})^T=(\mathbf{z},\mathbf{y})^T$, where $\mathbf{z}=(z_1,z_2,\ldots,z_n)^T$ and $\mathbf{y}=(y_1,y_2,\ldots,y_n)^T$. In this aim, we first studied the conditional distribution of $\hat{\beta}\left|(\mathbf{Z},\mathbf{Y})^T=(\mathbf{z},\mathbf{y})^T\right.$. Using the formula for the estimator of the regression slope (see for example, Neter et al, 1996), one can derive $\hat{\beta}=\sum_{i=1}^{n}(X_i-\bar{X})y_i\Big/\sum_{i=1}^{n}(X_i-\bar{X})^2$. We also explored the conditional distributions of $\sum_{j=1}^{n}(X_j-\bar{X})^2\left|(\mathbf{Z},\mathbf{Y})^T=(\mathbf{z},\mathbf{y})^T\right.$ and $\sum_{j=1}^{n}(X_j-\bar{X})y_j\left|(\mathbf{Z},\mathbf{Y})^T=(\mathbf{z},\mathbf{y})^T\right.$. In addition, the properties of the estimated latent slope $E\left(\hat{\beta}\middle|(\mathbf{Z},\mathbf{Y})^T=(\mathbf{z},\mathbf{y})^T\right)$ and the attenuation bias $E\left(\frac{\hat{b}}{\hat{\beta}}\middle|(\mathbf{Z},\mathbf{Y})^T=(\mathbf{z},\mathbf{y})^T\right)$ were investigated.

From the property of the conditional distribution in a multivariate normal random vector (see for example, Anderson (2003)), we know that $X_i\left|(Z_i,Y_i)^T=(z_i,y_i)^T\right.,i=1,2,\ldots,n$ has a normal distribution with mean $\mu_i$ and variance $\sigma_i^2$, where $\mu_i=\mu+\frac{\sigma_B^2}{\sigma_B^2+\sigma_W^2}(z_i-\mu)$ and $\sigma_i^2=\frac{\sigma_B^2\sigma_W^2}{\sigma_B^2+\sigma_W^2}\equiv\tau^2$.

## 1. Conditional Distribution of $\hat{\beta}\left|(\mathbf{Z},\mathbf{Y})^T=(\mathbf{z},\mathbf{y})^T\right.$

To prove the properties in this section, we need the following lemmas proved by Searle (1971).

***Theorem A***: Let $X\sim N(\boldsymbol{\mu},V)$, then $X'AX\sim\chi^{2'}\left(r,\frac{1}{2}\boldsymbol{\mu}'A\boldsymbol{\mu}\right)$ if and only if $AV$ is idempotent with rank $r$.

***Theorem B:*** Let $X \sim N(\mu, V)$, then $X'AX$ **and** $BX$ are independent if and only if BVA=0.

**Proposition 1**: The conditional distribution of $\sum_{j=1}^{n}(X_j - \bar{X})y_j \Big| (\mathbf{Z},\mathbf{Y})^T = (\mathbf{z},\mathbf{y})^T$ is a normal distribution with mean $\bar{\mu} = \frac{\sum_{i=1}^{n}\mu_i}{n} = \mu + \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}(\bar{z} - \mu)$ and variance $\sum_{j=1}^{n}(y_j - \bar{y})^2\tau^2$.

*Proof.* By recognizing $\sum_{j=1}^{n}(X_j - \bar{X})y_j = \sum_{j=1}^{n}(y_j - \bar{y})X_j$ as a linear combination of independent normal variates $X_i \Big| (Z_i, Y_i)^T = (z_i, y_i)^T$, its mean can be calculated as $\sum_{i=1}^{n}(\mu_i y_i - n\bar{\mu}\bar{y}) = \sum_{i=1}^{n}(\mu_i - \bar{\mu})(y_i - \bar{y})$ and its variance is $\sum_{j=1}^{n}(y_j - \bar{y})^2\tau^2$.

**Proposition 2**: The conditional distribution of $\frac{1}{\tau^2}\sum_{j=1}^{n}(X_j - \bar{X})^2 \Big| (\mathbf{Z},\mathbf{Y})^T = (\mathbf{z},\mathbf{y})^T$ is a non-central chi-squared distribution with degrees of freedom $n-1$ and noncentrality $\lambda = \frac{\sum_{i=1}^{n}(\mu_i - \bar{\mu})^2}{2\tau^2}$.

*Proof.* We first rewrite the quadratic form $\frac{1}{\tau^2}\sum_{j=1}^{n}(X_j - \bar{X})^2$ as $X'AX$, where $A = \frac{1}{\tau^2}\left(I_n - \frac{1}{n}J_n\right)$, $I_n$ is the $n \times n$ identity matrix and $J_n$ is the $n \times n$ matrix with all elements equal to 1. By recognizing $(X_1, X_2, \cdots, X_n)^T \Big| (\mathbf{Z},\mathbf{Y})^T = (\mathbf{z},\mathbf{y})^T$ as a multivariate normal vector $N(\boldsymbol{\mu}, V)$, where $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_n)^T$ and $V = \tau^2 I_n$, and verifying that $AV = I_n - \frac{1}{n}J_n$ is idempotent and has a rank equal to $n-1$, we can show that the conditional distribution of $\frac{1}{\tau^2}\sum_{j=1}^{n}(X_j - \bar{X})^2 \Big| (\mathbf{Z},\mathbf{Y})^T = (\mathbf{z},\mathbf{y})^T$ is a non-central chi-squared distribution with degrees of

freedom $n-1$. Its noncentrality can be calculated as $\lambda = \frac{1}{2}\boldsymbol{\mu}' A \boldsymbol{\mu} = \frac{\sum_{i=1}^{n}(\mu_i - \bar{\mu})^2}{2\tau^2}$, where

$$\bar{\mu} = \frac{\sum_{i=1}^{n}\mu_i}{n} = \mu + \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}(\bar{z} - \mu).$$

**Proposition 3**: (i) $E\left(\sum_{j=1}^{n}(X_j - \bar{X})^2 \middle\| (\mathbf{Z},\mathbf{Y})^T = (\mathbf{z},\mathbf{y})^T\right) = \tau^2(n-1) + \frac{\sum_{i=1}^{n}(\mu_i - \bar{\mu})^2}{2}$ and

(ii) $Var\left(\sum_{j=1}^{n}(X_j - \bar{X})^2 \middle\| (\mathbf{Z},\mathbf{Y})^T = (\mathbf{z},\mathbf{y})^T\right) = 2\tau^2\left(\tau^2(n-1) + \sum_{i=1}^{n}(\mu_i - \bar{\mu})^2\right)$.

Proposition 3 can be proved by directly applying the mean and the variance of a non-central chi-squared distribution. (see, for example, Johnson and Kotz (1970)) .

**Proposition 4**: $\sum_{j=1}^{n}(X_j - \bar{X})y_j \middle| (\mathbf{Z},\mathbf{Y})^T = (\mathbf{z},\mathbf{y})^T$ and $\sum_{j=1}^{n}(X_j - \bar{X})^2 \middle| (\mathbf{Z},\mathbf{Y})^T = (\mathbf{z},\mathbf{y})^T$ are not conditionally independent, given $(\mathbf{Z},\mathbf{Y})^T = (\mathbf{z},\mathbf{y})^T$.

*Proof.* By rewriting $\sum_{j=1}^{n}(X_j - \bar{X})y_j$ as $\sum_{j=1}^{n}(y_j - \bar{y})X_j (= B\mathbf{X})$, where $B = (y_1 - \bar{y}, y_2 - \bar{y}, \ldots, y_n - \bar{y})$ and $\frac{1}{\tau^2}\sum_{j=1}^{n}(X_j - \bar{X})^2$ as $\boldsymbol{X'AX}$ , where A is the same as what was presented in Proposition 2, we can apply Theorem B after verifying $BVA = B \neq \mathbf{0}$.

Note that the non-independence of the numerator and denominator of $\hat{\beta}$ complicates the conditional distribution $\hat{\beta} \middle| (\mathbf{Z},\mathbf{Y})^T = (\mathbf{z},\mathbf{y})^T$, which cannot be analytically derived.

## 2. Approximate forms of the latent slope and the attenuation bias

This subsection will present an approximate form that can be used for calculating $E\left(\hat{\beta} \middle| (\mathbf{Z},\mathbf{Y})^T = (\mathbf{z},\mathbf{y})^T\right)$ and $E\left(\frac{\hat{b}}{\hat{\beta}} \middle| (\mathbf{Z},\mathbf{Y})^T = (\mathbf{z},\mathbf{y})^T\right)$ without using simulation as was presented in Aim 1.

Let $U=\frac{\sum_{j=1}^{n}(X_j-\bar{X})^2}{\tau^2}$ and $W=\sum_{j=1}^{n}(X_j-\bar{X})y_j$, we first state the following propositions.

**Proposition 5:** The conditional mean and the conditional variance of $\frac{1}{U}$ can be expressed as

$$E\left(\frac{1}{U}\middle|(\mathbf{Z},\mathbf{Y})^T=(\mathbf{z},\mathbf{y})^T\right)=\sum_{i=1}^{\infty}\frac{e^{-\frac{\lambda}{2}}\left(\frac{\lambda}{2}\right)^i}{i!\left(n-1+2(i-1)\right)}\frac{1}{\tau^2} \text{ and}$$

$$Var\left(\frac{1}{U}\middle|(\mathbf{Z},\mathbf{Y})^T=(\mathbf{z},\mathbf{y})^T\right)=\left(\sum_{i=1}^{\infty}\frac{e^{-\frac{\lambda}{2}}\left(\frac{\lambda}{2}\right)^i}{i!\left(n-1+2i-2\right)\left(n-1+2i-4\right)}\right)-\left(\sum_{i=1}^{\infty}\frac{e^{-\frac{\lambda}{2}}\left(\frac{\lambda}{2}\right)^i}{i!\left(n-1+2(i-1)\right)}\right)^2.$$

*Proof.* Since $\frac{1}{\tau^2}\sum_{j=1}^{n}(X_j-\bar{X})^2\left|(\mathbf{Z},\mathbf{Y})^T=(\mathbf{z},\mathbf{y})^T\right.$ has a non-central chi-squared distribution with degrees of freedom equal to $n-1$, we can use the definition of non-central chi-squared distribution (see, for example, p132, Johnson and Kotz) to calculate

$$E\left(\frac{1}{U}\middle|(\mathbf{Z},\mathbf{Y})^T=(\mathbf{z},\mathbf{y})^T\right)=\sum_{i=1}^{\infty}\frac{e^{-\frac{\lambda}{2}}\left(\frac{\lambda}{2}\right)^i}{i!}\int_0^{\infty}u^{-1}f_{\chi_{n-1+2i}}(u)du=\sum_{i=1}^{\infty}\frac{e^{-\frac{\lambda}{2}}\left(\frac{\lambda}{2}\right)^i}{i!\left(n-1+2(i-1)\right)}\frac{1}{\tau^2}, \text{ where}$$

$f_{\chi_{n-1+2i}}(x)$ is the probability density function of a chi-squared distribution with degrees of freedom $n-1+2i$ and the integral inside the first summand can be obtained by some algebraic manipulations

$$\int_0^{\infty}u^{-1}f_{\chi_{n-1+2i}}(u)du=\int_0^{\infty}\frac{u^{\frac{n-1+2i-2}{2}-1}e^{-\frac{u}{2}}}{2\cdot 2^{\frac{n-1+2i-2}{2}}\cdot\left(\frac{n-1+2i-2}{2}\right)\cdot\Gamma\left(\frac{n-1+2i-2}{2}\right)}du=\frac{1}{n-1+2i-2}.$$

Similarly, we can calculate $E\left(\frac{1}{U^2}\middle|(\mathbf{Z},\mathbf{Y})^T=(\mathbf{z},\mathbf{y})^T\right)=\sum_{i=1}^{\infty}\frac{e^{-\frac{\lambda}{2}}\left(\frac{\lambda}{2}\right)^i}{i!\left(n-1+2i-2\right)\left(n-1+2i-4\right)}$.

Therefore,

$$Var\left(\frac{1}{U}\middle|(\mathbf{Z},\mathbf{Y})^T=(\mathbf{z},\mathbf{y})^T\right)=\left(\sum_{i=1}^{\infty}\frac{e^{-\frac{\lambda}{2}}\left(\frac{\lambda}{2}\right)^i}{i!(n-1+2i-2)(n-1+2i-4)}\right)-\left(\sum_{i=1}^{\infty}\frac{e^{-\frac{\lambda}{2}}\left(\frac{\lambda}{2}\right)^i}{i!(n-1+2(i-1))}\right)^2.$$

**Proposition 6:** The conditional mean and the conditional variance of $W/\tau^2$ can be expressed as

$$E\left(W/\tau^2\middle|(\mathbf{Z},\mathbf{Y})^T=(\mathbf{z},\mathbf{y})^T\right)=\frac{\sum_{i=1}^{n}\mu_i y_i-n\bar{\mu}\bar{y}}{\tau^2}\text{ and }Var\left(W/\tau^2\middle|(\mathbf{Z},\mathbf{Y})^T=(\mathbf{z},\mathbf{y})^T\right)=\frac{\sum_{j=1}^{n}(y_j-\bar{y})^2}{\tau^2}.$$

Note that Proposition 6 is a direct consequence of Proposition 1.

Observe that

$$\begin{aligned}E\left(\hat{\beta}\middle|(\mathbf{Z},\mathbf{Y})^T=(\mathbf{z},\mathbf{y})^T\right)&=E\left(\frac{W/\tau^2}{U}\middle|(\mathbf{Z},\mathbf{Y})^T=(\mathbf{z},\mathbf{y})^T\right)\\&=\left[E\left(W/\tau^2\middle|(\mathbf{Z},\mathbf{Y})^T=(\mathbf{z},\mathbf{y})^T\right)E\left(\frac{1}{U}\middle|(\mathbf{Z},\mathbf{Y})^T=(\mathbf{z},\mathbf{y})^T\right)\right]\\&\quad+\rho\left[Var\left(W/\tau^2\middle|(\mathbf{Z},\mathbf{Y})^T=(\mathbf{z},\mathbf{y})^T\right)Var\left(\frac{1}{U}\middle|(\mathbf{Z},\mathbf{Y})^T=(\mathbf{z},\mathbf{y})^T\right)\right]^{1/2}\end{aligned}\qquad 9$$

where $\rho$ represents the conditional correlation between $\frac{1}{U}$ and $W$.

Since all terms in equation 9, except $\rho$ can be expressed in an explicit form of y's and z's, as shown in Propositions 5 and 6, we can obtain an algebraic form for $E\left(\hat{\beta}\middle|(\mathbf{Z},\mathbf{Y})^T=(\mathbf{z},\mathbf{y})^T\right)$ in terms of $\rho$ which is bounded by 1, although this is a more complicated function of y's and z's.

**Proposition 7:** The conditional mean of $\tau^2/W$ can be expressed as

$$E\left(\tau^2/W\middle|(\mathbf{Z},\mathbf{Y})^T=(\mathbf{z},\mathbf{y})^T\right)=\tau^2\left(\phi/\theta\right)\sum_{k=0}^{\infty}\left[-\left(\phi/\theta\right)\right]^{2k}\frac{(2k)!}{2^k k!}$$

where $\theta=\sum_{i=1}^{n}\mu_i y_i-n\bar{\mu}\bar{y}$ and variance $\phi^2=\tau^2\sum_{j=1}^{n}(y_j-\bar{y})^2$.

*Proof.*

$$E\left(\tau^2/W \middle| (\mathbf{Z},\mathbf{Y})^T = (\mathbf{z},\mathbf{y})^T\right) = \tau^2 \int_{-\infty}^{\infty} \frac{1}{w}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{w-\theta}{\phi}\right)^2\right)dw$$

$$= \tau^2 \int_{-\infty}^{\infty} \frac{1}{y+\theta/\phi}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)dy = \tau^2\left(\phi/\theta\right)\int_{-\infty}^{\infty}\sum_{n=0}^{\infty}\left(-\left(\phi/\theta\right)y\right)^n \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)dy$$

$$= \tau^2\left(\phi/\theta\right)\sum_{n=0}^{\infty}\left(-\left(\phi/\theta\right)\right)^n \int_{-\infty}^{\infty} y^n \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)dy$$

$$= \tau^2\left(\phi/\theta\right)\sum_{k=0}^{\infty}\left(-\left(\phi/\theta\right)\right)^{2k}\frac{(2k)!}{2^k k!}.$$

Here the first equality is the result of the definition, the second equality is a routine change-of-variable technique in integral, the third equality uses the power series, the fourth equality assumes that the integrals and the infinite series are convergent and that the conditions of Fubini's theorem are satisfied, and the last equality uses the result of central moments from the standard normal distribution. Note that if conditions of Fubini's theorem and transformation of power series are not satisfied, other methods of approximation of the above integral, such as Monte Carlo integration would be necessary.

Observe that

$$E\left(\frac{\hat{b}}{\hat{\beta}}\middle|(\mathbf{Z},\mathbf{Y})^T = (\mathbf{z},\mathbf{y})^T\right) = \hat{b}E\left(\frac{U}{W/\tau^2}\middle|(\mathbf{Z},\mathbf{Y})^T = (\mathbf{z},\mathbf{y})^T\right)$$

$$= \left[E\left(\tau^2/W \middle| (\mathbf{Z},\mathbf{Y})^T = (\mathbf{z},\mathbf{y})^T\right)E\left(U\middle|(\mathbf{Z},\mathbf{Y})^T = (\mathbf{z},\mathbf{y})^T\right)\right] \qquad 10$$

$$+\gamma\left[Var\left(\tau^2/W \middle| (\mathbf{Z},\mathbf{Y})^T = (\mathbf{z},\mathbf{y})^T\right)Var\left(U\middle|(\mathbf{Z},\mathbf{Y})^T = (\mathbf{z},\mathbf{y})^T\right)\right]^{1/2}$$

where $\gamma$ represents the conditional correlation between $U$ and $\tau^2/W$.

Since all terms in equation 10, except $\gamma$ can be expressed in an explicit form of y's and z's, as shown in Propositions 3 and 7, we can obtain an algebraic form for $E\left(\frac{\hat{b}}{\hat{\beta}}\middle|(\mathbf{Z},\mathbf{Y})^T = (\mathbf{z},\mathbf{y})^T\right)$ in terms of $\gamma$ which is bounded by 1, although this is a more complicated function of y's and z's.

Given that there is not an explicit form for calculating the $Var\left(\tau^2/W \middle| (\mathbf{Z},\mathbf{Y})^T = (\mathbf{z},\mathbf{y})^T\right)$, we would apply the Monte Carlo integration to estimate

$E\left(1/W^2 \middle| (\mathbf{Z},\mathbf{Y})^T = (\mathbf{z},\mathbf{y})^T\right) = \int_{-\infty}^{\infty} \frac{1}{w^2}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{w-\theta}{\phi}\right)^2\right]dw$ and use the result from Proposition 7 to obtain the conditional variance.

*Results and Discussion*

For a given set of $\mathbf{z} = (z_1, z_2, \ldots, z_n)^T$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$, we have expressed the proposed latent slope and attenuation in an almost explicit form that can be used for approximation. These forms can provide us with an alternative computational method for finding the estimates, other than the simulation method presented in Aim1. In addition, from Propositions 1-2, we found that the distributional behavior of $\hat{\beta}\left|(\mathbf{Z},\mathbf{Y})^T = (\mathbf{z},\mathbf{y})^T\right.$ follows the $c\frac{noncentral\ \chi^2}{normal}$ form.

*Conclusions*

The analytic form of the proposed latent slope and attenuation is complicated. It is difficult to compare the distribution of the attenuation bias derived from this method with the traditional method known as an F or approximated F distribution depending upon whether the surrogate measurements are balanced or unbalanced. For the latent slope, the estimator using the traditional method $\left(\frac{\hat{b}}{\hat{\sigma}_B^2 \Big/ \left(\hat{\sigma}_B^2 + \hat{\sigma}_W^2\right)}\right)$ is complicated and its distribution is not yet identified.

**Specific Aim #3**

Material related to Specific Aim #3 was abstracted from Symanski et al., 2007.

*Methods*

The database contributing measurements to this study has been described elsewhere (Symanski and Greeson, 2002). In total, the database contains 121 sets of data, comprised of 6174 biological and personal exposure measurements collected on 577 workers from 55 groups. In addition to the exposure measurements, details about the industry, the airborne contaminant, the sampling protocol, and the worker were abstracted from the original studies. For each biomarker of exposure, information about the half-life of the contaminant or its metabolite was also recorded. For inclusion in the current study, data were eliminated on the basis of the following criteria: 1) a single biological monitoring method was used to evaluate exposure, 2) biological measurements were collected prior to or at the start of the work-shift (if comparisons were to be made to personal monitoring data), 3) fewer than 5 workers had been monitored, 4) reported use of respirators was inconsistent among workers with both personal and biological

monitoring data, or 5) workers came from more than one workplace (drycleaners). To enhance comparability among multiple measures of exposures collected on groups of workers, data were further examined to insure that the measurements were collected concurrently on the same workers on common days of sampling. In instances where data were missing on some workers (i.e., measurements were available on one exposure measure but not for another) or where the periods of sampling for different exposure measures did not exactly coincide, monitoring data on specific workers or individual measurements were omitted to obtain sets of data that were balanced across exposure measures on the basis of both the numbers of workers sampled and days monitored. Following these exclusions, the data were re-examined to exclude groups with fewer than five workers or less than 10 measurements.

To assess the attenuation bias for each measure of exposure, a one-way random-effects model was applied following natural logarithmic transformation of the personal or biological measurements of exposure. Under the assumption that the variance and covariance functions do not depend on the time of measurement, we considered two covariance structures for the log-transformed measurements collected on the same individual: 1) a compound symmetric structure and 2) an exponential correlation structure. Under compound symmetry (CS), it is assumed that the covariance between measurements collected on the same worker is the same irrespective of the time interval separating them $\left[Cov\left(lnZ_{ij}, lnZ_{ij'}\right) = \sigma_B^2 \; for \; j \neq j'\right]$. It directly follows that the correlation between measurements on the same person can be expressed as $\left[Corr\left(lnZ_{ij}, lnZ_{ij'}\right) = \sigma_B^2 / \left(\sigma_B^2 + \sigma_W^2\right) \; for \; j \neq j'\right]$, which is also known as the intra-class correlation coefficient (ICC). We expected in some cases that the correlation between repeated measurements would decrease as a function of the time interval between them, as might occur for a biomarker with a long half-life in which a series of measurements collected on the same worker were collected over a relatively short period. Thus, an exponential (EXP) correlation structure (in which the correlation is dependent upon the time interval between measurements) was also applied. Here, the covariance function is $Cov\left(lnZ_{ij}, lnZ_{ij'}\right) = \sigma_B^2 + \rho^{\tau_{jj'}} \sigma_W^2$ for $j \neq j'$ and where $\rho^{\tau_{jj'}}$ is the correlation between measurements separated by the interval $\tau_{jj'}$ $\left[\tau_{jj'} = \left|t_j - t_{j'}\right|\right]$. Under this structure, the correlation decays toward zero as the time separation between measurements increases (which represents a generalization of the first-order autotoregressive [AR(1)] structure for equally spaced data). Under both correlation structures, the within-worker ($\sigma_W^2$) and between-worker ($\sigma_B^2$) variances were estimated using the method of restricted maximum likelihood (REML) implemented with the PROC MIXED procedure from the SAS System Software (Version 9.1, Cary, NC, USA).

Since in a simple linear regression it is the relative magnitude of the variances that provides an indication of the bias in the slope coefficient due to measurement error in the exposure measure, estimated values of the intra-class correlation coefficient under compound symmetry were computed as $\left[\hat{\sigma}_B^2 / \left(\hat{\sigma}_B^2 + \hat{\sigma}_W^2\right)\right]$. Exact 95% confidence limits on the intra-class correlation coefficient were also obtained according to the methods outlined by Searle et al. (1992), which rely on functions of the mean squares associated with the sources of variation in the one-way random effects model. Under assumptions of normality, the REML and ANOVA

estimators for the variance components are identical in the balanced case *except* when the ANOVA estimators yield a negative estimate (since REML estimates are always non-negative). In this instance, the ANOVA estimators of the variances were obtained to allow for the interval estimation of the intra-class correlation coefficients. Negative point or interval estimates were set to 0. Estimated variances using the ANOVA method of estimation were obtained with PROC NESTED in SAS.

Finally, in expanding upon the work of Lin et al. (2005), we derived an expression for the intra-class correlation coefficients for serially correlated, irregularly spaced measurements as follows:

$$\frac{\hat{\sigma}_B^2}{\hat{\sigma}_B^2 + \frac{\hat{\sigma}_W^2}{n}\left(1 + \frac{2}{n}\sum_{j=2}^{n}\sum_{j'=1}^{j-1}\hat{\rho}^{t_{jj'}}\right)}$$

This equation was applied to all data sets with evidence of serial correlation ($p < 0.05$) under the model with an exponential correlation structure.

*Results and Discussion*

After applying the criteria for inclusion in the study, data were available on 32 groups of workers with multiple measures of exposure representing a total of 77 datasets – each group represents workers from a common workplace and each data set compiles monitoring data for a single exposure measure. Fourteen groups of workers were sampled via both personal and biological monitoring and 18 groups were assessed with multiple biological measures of exposure. Table 3 provides a listing of groups of workers on whom both personal and biological monitoring had been conducted with information on the industry, the exposure measure, the duration of the monitoring period, the time of sampling, and numbers of workers and measurements. Groups 3 and 7 had personal samples on two related exposures (alkyl lead and inorganic lead and PAHs and benzo(a)pyrene, respectively), and five groups (groups 1, 2, 3, 10, and 13) had been monitored using two biological measures of exposure. Table 4 provides a breakdown of groups on whom only biological monitoring had been conducted. Among these workers, groups 15, 16, 24, 25, and 27 had been evaluated with three or more biomarkers. Measurements on groups 24, 25, and 27 were collected in the same workplace as groups 3, 10, and 13, respectively (Table 3), but not on the same set of workers or over the same sampling intervals. Characteristics of the air-biomarker and multiple biomarker data sets are summarized in Table 5, which provides a profile of the two databases on the basis of type of industry and agent, residence time of the biomarkers, duration of monitoring, and sample size.

For the groups of workers on whom both personal and biological monitoring had been conducted, 15 of the 35 data sets (43%) were characterized by greater levels of between-worker variability than within-worker variability when the point estimates of the variances were compared. The point and interval estimates of the intra-class correlation coefficients (ICC) (based upon the model with a compound symmetry covariance structure) appear in Table 6. For these data sets, the estimated ICC ranged from 0 to 0.993 with a median value of 0.474. In stratifying the data by type of exposure measure, the ICC median value dropped from 0.627 for the biological measures (n=19 data sets) to 0.263 for the airborne measures (n=16). On a group

by group basis, the estimated ICC values were larger for the biological monitoring data as compared to the personal sampling data for 8 out of the 14 groups (57%).

Table 7 summarizes the results for the 18 groups of workers whose exposures were evaluated with two or more biomarkers (42 sets of data). Information about the half-life of the measured compound (short=less than 7 days, intermediate=7 days to 4 weeks, and long=greater than 4 weeks) is also given. Overall, 55% (n=23) of the data sets were characterized by greater levels of between-worker variability than within-worker variability. Estimates of the intra-class correlation coefficients ranged from 0 to 0.998 with a median value of 0.526. On the basis of residence time in the body, only five groups (groups 18, 23, 24, 27, and 29) were measured with biomarkers of different half-lives ('short', 'intermediate' or 'long'). For group 18, the estimated ICC values for the 'short-lived' biomarker as compared to the 'intermediate-lived' biomarker were nearly the same (0.570 and 0.509 for urinary tetachlorophenol and pentachorophenol, respectively). Likewise, results on group 24 for urinary lead ('intermediate') and urinary ALA ('long') were similar (0.637 and 0.579). Biomarkers with longer residence times were characterized by larger ICC values for group 23 (0.168 and 0.457 for urinary lead ('intermediate') and urinary ALA ('long'), respectively), group 24 (0.456 and 0.637 for blood lead ('intermediate') and urinary lead ('long'), respectively) and group 29 (0.330 and 0.591 for blood mercury ('intermediate') and urinary mercury ('long'), respectively).

In evaluating the results from the random-effects model with an exponential correlation error structure, significant serial correlation ($p < 0.05$) was detected in data collected on tobacco farmers (groups 30, 31, and 32; both measures), chloralkali processing plant workers (group 14; blood Hg), sawmill workers (group 18; both measures), forestry workers (group 22; blood arsenic) and alkyl lead manufacturing workers (group 24; urinary lead). For these groups, the estimated values of the intra-class correlation coefficients under the one-way random effects model with either a CS or exponential correlation error structure, computed on the basis of the number of measurements collected on each worker (equations 4 and 5, respectively), are shown in Table 8. Generally, the intra-class correlation coefficients appear to be over-estimated when serial correlation was present but ignored in the analyses.

Table 3. Groups of Workers whose Exposures were Assessed by Air and Biological Monitoring (n=14 groups)[1]

| Group | N | k | Duration of Sampling | Industry | Type | Time | Units |
|---|---|---|---|---|---|---|---|
| 1 | 27 | 9 | 1 week | Hospital sterilizer unit | Environmental ethylene oxide | Postshift | $mg/m^3$ |
| | | | | | Blood ethylene oxide | Postshift | µg/L |
| | | | | | Alveolar ethylene oxide | Postshift | $mg/m^3$ |
| 2 | 25 | 5 | 1 week | Dry alkaline battery manufacturing | Inorganic mercury | 8AM-2PM | $µg/m^3$ |
| | | | | | Blood mercury | 2PM | µg/100ml |
| | | | | | Urinary mercury | 2PM | µg/g creatinine |
| 3 | 80 | 5 | 1 month | Alkyl lead manufacturing | Alkyl lead | During shift | $µg/m^3$ |
| | | | | | Inorganic lead | During shift | $µg/m^3$ |
| | | | | | Urinary lead | During shift | µg/L |
| | | | | | Urinary ALA | During shift | mg/100ml |
| 4 | 20 | 5 | 1 week | Plastic boat manufacturing | Styrene | During shift | ppm |
| | | | | | Alveolar styrene | Postshift | ppm |
| 5 | 15 | 5 | 1 week | Carbon black manufacturing | Dust | During shift | $mg/m^3$ |
| | | | | | Urinary 1-hydroxypyrene | Postshift | µmol/mol creatinine |
| 6 | 40 | 20 | 1 year | Lead-acid battery manufacturing | Lead | During shift | $µg/m^3$ |
| | | | | | Blood lead | During shift | µg/100ml |
| 7 | 20 | 10 | 1 week | Coke manufacturing | Benzo[a]pyrene | During shift | $µg/m^3$ |
| | | | | | Total PAHs | During shift | $µg/m^3$ |
| | | | | | Urinary 1-hydroxypyrene | Postshift | ng/ml |
| 8 | 16 | 8 | 1 week | Coke manufacturing | Benzo[a]pyrene | During shift | $µg/m^3$ |
| | | | | | Urinary 1-hydroxypyrene | Postshift | ng/ml |
| 9 | 25 | 5 | 1 week | Pulp and paper industry machine manufacturing | Water-soluble chromium | During shift | $mg/m^3$ |
| | | | | | Urinary chromium | 4:00PM | µg/g creatinine |
| 10 | 16 | 8 | 10 months | Plastic boat manufacturing | Styrene | During shift | ppm |
| | | | | | Urinary mandelic acid | Postshift | mg/g creatinine |
| | | | | | Urinary phenylglyoxylic acid | Postshift | mg/g creatinine |
| 11 | 10 | 5 | 1 week | Viscose rayon plant | Carbon disulfide | During shift | ppm |
| | | | | | Urinary TTCA | Postshift | mg/g creatinine |
| 12 | 18 | 6 | 1 week | Creosote impregnation plant | Pyrene | During shift | $µg/m^3$ |
| | | | | | Urinary 1-hydroxypyrene | Endshift | µmol/mol creatinine |
| 13 | 12 | 6 | 7 months | Plastics manufacturing | Styrene | During shift | $mg/m^3$ |
| | | | | | Blood styrene | Midshift | mg/L |
| | | | | | Urinary mandelic acid | Postshift | mmol/mol creatinine |
| 14 | 63 | 7 | 1 month | Chloralkali processing plant | Inorganic mercury | During shift | $µg/m^3$ |
| | | | | | Blood mercury | Postshift | nmol/L |

[1]Abbreviations: N = no. of measurements; k = no. of workers; ALA=δ -aminolevulinic acid, PAHs=polycyclic aromatic hydrocarbons, TTCA=2-thiothiazolidine-4-carboxylic acid

Table 4. Groups of Workers whose Exposures were Assessed by Multiple Biomarkers (n=18 groups)[1]

| Group | N | k | Duration of Sampling | Industry | Type | Time | Units |
|---|---|---|---|---|---|---|---|
| 15 | 22 | 11 | 1 week | Copper smelting | Urinary inorganic arsenic | Postshift | µg/L |
| | | | | | Urinary methylarsonic acid | Postshift | µg/L |
| | | | | | Urinary dimethylarsinic acid | Postshift | µg/L |
| | | | | | Urinary trimethylarsenic compounds | Postshift | µg/L |
| 16 | 30 | 10 | 1 week | Rotogravure printing | Alveolar air toluene | Postshift | µg/L |
| | | | | | Blood toluene | Postshift | mg/kg |
| | | | | | Urinary hippuric acid | Postshift | g/L |
| 17 | 26 | 13 | 2 years | Cadmium pigment manufacturing | Urinary cadmium | | µg/L |
| | | | | | Blood cadmium | | µg/100ml |
| 18 | 124 | 44 | 5 months | Sawmilling | Urinary tetrachlorophenol | | ppb |
| | | | | | Urinary pentachlorophenol | | ppb |
| 19 | 12 | 6 | 1 week | Plastic boat manufacturing | Urinary mandelic acid + phenylglyoxylic acid | 4:00PM | mmol/h |
| | | | | | Urinary mandelic acid + phenylglyoxylic acid | 9:00PM | mmol/h |
| 20 | 66 | 12 | 5 years | Chromium plating | Urinary chromium | During shift | µg/g creatinine |
| | | | | | Sister Chromatid Exchanges/cell (mean) | During shift | |
| 21 | 12 | 6 | 1 week | Polyurethane manufacturing | Plasma 2,4-TDA | 3:30PM | ng/ml |
| | | | | | Plasma 2,6-TDA | 3:30PM | ng/ml |
| 22 | 40 | 5 | 2 months | Forestry | Blood arsenic | Postshift | ppm |
| | | | | | Urinary arsenic | 24 hours | µg/24hrs |
| 23 | 30 | 5 | 1 month | Alkyl lead manufacturing | Urinary lead | During shift | mg/L |
| | | | | | Urinary ALA | During shift | mg/100ml |
| 24 | 25 | 5 | 1 month | Alkyl lead manufacturing | Blood lead | During shift | µg/100g |
| | | | | | Urinary lead | During shift | µg/L |
| | | | | | Urinary ALA | During shift | mg/100ml |
| 25 | 22 | 11 | 2 years | Plastic boat manufacturing | Blood styrene | Postshift | mg/L |
| | | | | | Urinary mandelic acid | Postshift | mg/g creatinine |
| | | | | | Urinary phenylglyoxylic acid | Postshift | mg/g creatinine |
| 26 | 10 | 5 | 1 week | Smelting | Urinary inorganic As + MMA + DMA | Preshift | µg/g creatinine |
| | | | | | Urinary inorganic As + MMA + DMA | Preshift | µg/L |
| 27 | 14 | 7 | 2 months | Plastics manufacturing | Blood styrene | Midshift | mg/L |
| | | | | | Urinary mandelic acid | Postshift | mmol/mol creatinine |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | 06-styrene-guanine lymphocyte adducts | Midshift | adducts/10^8 dNp |
| 28 | 32 | 8 | 1 week | Nickel refining | Urinary nickel | Postshift | µmol/L |
| | | | | | Urinary nickel | Postshift | mmol/mol creatinine |
| 29 | 18 | 9 | 6 months | Chemical manufacturing | Blood mercury | | µg/L |
| | | | | | Urinary mercury | | µg/g creatinine |
| 30 | 106 | 22 | 2 months | Tobacco farming | Plasma cholinesterase | PM | µmols/ml/min |
| | | | | | Red blood cell cholinesterase | PM | µmols/ml/min |
| 31 | 55 | 11 | 2 months | Tobacco farming | Plasma cholinesterase | PM | µmols/ml/min |
| | | | | | Red blood cell cholinesterase | PM | µmols/ml/min |
| 32 | 62 | 14 | 2 months | Tobacco farming | Plasma cholinesterase | PM | µmols/ml/min |
| | | | | | Red blood cell cholinesterase | PM | µmols/ml/min |

[1]Abbreviations used: N = No. of measurements; k = No. of workers; ALA=δ-aminolevulinic acid; TDA=2,4-toluenediamine, MMA=methylarsonic acid; DMA=dimethylarsonic acid

Table 5. Characteristics of the Air-Biological Exposure Measures and Multiple Biomarkers Databases

| | Air-Biomarker database | Multiple Biomarker database |
|---|---|---|
| No. of groups | 14 | 18 |
| No. of data sets | 35 | 42 |
| Industrial Sector[1] | | |
| Agriculture, hunting and forestry | -- | 1 group |
| Mining and quarrying | -- | 3 groups |
| Manufacturing | 13 groups | 14 groups |
| Health and social work | 1 group | -- |
| Type of contaminant (exposure) | | |
| Gas/vapor | 5 groups | 6 groups |
| Aerosol | 6 groups | 12 groups |
| Combination | 3 groups | -- |
| No. of workers | | |
| Total | 104 | 204 |
| Range (per group) | 5-20 | 5-44 |
| Average (across groups) | 7 | 11 |
| Median (across groups) | 6 | 10 |
| No. of measurements | | |
| Total | 1034 | 1547 |
| Range (per data set) | 10-80 | 10-124 |
| Average (across data sets) | 30 | 37 |
| Median (across data sets) | 20 | 26 |
| Sampling interval | | |
| 1 week or less | 9 groups | 6 groups |
| 1 month-3 months | 2 groups | 7 groups |
| 5 months to 1 year | 3 groups | 2 groups |
| > 1 year | -- | 3 groups |
| Residence time of biomarker | | |
| Short (< 7 days) | 12 data sets | 21 data sets |
| Intermediate (1 to 4 weeks) | 5 data sets | 11 data sets |
| Long (> 4 weeks) | 2 data sets | 9 data sets |

[1]Based on the International Standard Industrial Classification (ISIC) coding (United Nations, 1990)

Table 6. Point and Interval Estimates of the Intra-Class Correlation Coefficients (ICC) for Workers' Exposures Assessed in Parallel by Air and Biological Monitoring (Based upon the Model with Compound Symmetry)

| Group | Type | ICC[1] | 95 % CI |
|---|---|---|---|
| 1 | Environmental ethylene oxide | 0.000 | -[2] |
| | Blood ethylene oxide | 0.000 | (0.000, 0.273) |
| | Alveolar ethylene oxide | 0.000 | -[2] |
| 2 | Inorganic mercury | 0.757 | (0.426, 0.966) |
| | Blood mercury | 0.798 | (0.494, 0.972) |
| | Urinary mercury | 0.985 | (0.947, 0.998) |
| 3 | Alkyl lead | 0.204 | (0.043, 0.722) |
| | Inorganic lead | 0.000 | (0.000, 0.294) |
| | Urinary lead | 0.634 | (0.353, 0.937) |
| | Urinary ALA | 0.649 | (0.368, 0.941) |
| 4 | Styrene | 0.000 | (0.000, 0.619) |
| | Alveolar styrene | 0.000 | (0.000, 0.628) |
| 5 | Dust | 0.228 | (0.000, 0.839) |
| | Urinary 1-hydroxypyrene | 0.738 | (0.271, 0.965) |
| 6 | Lead | 0.993 | (0.984, 0.997) |
| | Blood lead | 0.943 | (0.864, 0.977) |
| 7 | Benzo[a]pyrene | 0.000 | (0.000, 0.167) |
| | Total PAHs | 0.000 | (0.000, 0.143) |
| | Urinary 1-hydroxypyrene | 0.399 | (0.000, 0.805) |
| 8 | Benzo[a]pyrene | 0.495 | (0.000, 0.871) |
| | Urinary 1-hydroxypyrene | 0.636 | (0.000, 0.913) |
| 9 | Water-soluble chromium | 0.658 | (0.288, 0.947) |
| | Urinary chromium | 0.866 | (0.630, 0.983) |
| 10 | Styrene | 0.297 | (0.000, 0.801) |
| | Urinary mandelic acid | 0.267 | (0.000, 0.789) |
| | Urinary phenylglyoxylic acid | 0.083 | (0.000, 0.705) |
| 11 | Carbon disulfide | 0.196 | (0.000, 0.866) |
| | Urinary TTCA | 0.627 | (0.000, 0.952) |
| 12 | Pyrene | 0.512 | (0.021, 0.897) |
| | Urinary 1-hydroxypyrene | 0.497 | (0.007, 0.892) |
| 13 | Styrene | 0.677 | (0.000, 0.946) |
| | Blood styrene | 0.339 | (0.000, 0.868) |
| | Urinary mandelic acid | 0.474 | (0.000, 0.903) |
| 14 | Inorganic mercury | 0.397 | (0.152, 0.788) |
| | Blood mercury | 0.860 | (0.693, 0.969) |

[1]ICC calculated as $\hat{\sigma}_B^2 / \left( \hat{\sigma}_B^2 + \hat{\sigma}_W^2 \right)$

[2]Lower and upper confidence limits were negative.

Table 7. Point and Interval Estimates of the Intra-Class Correlation Coefficients (ICC) for Workers' Exposures Assessed in Parallel by Multiple Biological Measures of Exposure (Based upon the Model with Compound Symmetry)

| Group | Type | Half-Life[1] | ICC[2] | 95% CI |
|---|---|---|---|---|
| 15 | Urinary inorganic arsenic | short | 0.834 | (0.515, 0.952) |
| | Urinary methylarsonic acid | short | 0.481 | (0.000, 0.825) |
| | Urinary dimethylarsinic acid | short | 0.417 | (0.000, 0.798) |
| | Urinary trimethylarsenic compounds | short | 0.792 | (0.419, 0.939) |
| 16 | Alveolar air toluene | short | 0.423 | (0.041, 0.782) |
| | Blood toluene | short | 0.772 | (0.495, 0.930) |
| | Urinary hippuric acid | short | 0.542 | (0.167, 0.839) |
| 17 | Urinary cadmium | long | 0.162 | (0.000, 0.636) |
| | Blood cadmium | long | 0.389 | (0.000, 0.761) |
| 18 | Urinary tetrachlorophenol | short | 0.570 | (0.396, 0.719) |
| | Urinary pentachlorophenol | intermediate | 0.509 | (0.325, 0.674) |
| 19 | Urinary mandelic acid + phenylglyoxylic acid | short | 0.000 | (0.000, 0.690) |
| | Urinary mandelic acid + phenylglyoxylic acid | short | 0.000 | (0.000, 0.671) |
| 20 | Urinary chromium | intermediate | 0.723 | (0.515, 0.892) |
| | SCEs/cell (mean) | | 0.712 | (0.501, 0.887) |
| 21 | Plasma 2,4-TDA | intermediate | 0.942 | (0.698, 0.992) |
| | Plasma 2,6-TDA | intermediate | 0.998 | (0.987, 1.000) |
| 22 | Blood arsenic | short | 0.000 | (0.000, 0.132) |
| | Urinary arsenic | short | 0.000 | (0.000, 0.485) |
| 23 | Urinary lead | intermediate | 0.168 | (0.000, 0.749) |
| | Urinary ALA | long | 0.457 | (0.117, 0.894) |
| 24 | Blood lead | intermediate | 0.456 | (0.087, 0.897) |
| | Urinary lead | intermediate | 0.637 | (0.263, 0.943) |
| | Urinary ALA | long | 0.579 | (0.199, 0.930) |
| 25 | Blood styrene | short | 0.342 | (0.000, 0.764) |
| | Urinary mandelic acid | short | 0.268 | (0.000, 0.728) |
| | Urinary phenylglyoxylic acid | short | 0.000 | (0.000, 0.459) |
| 26 | Urinary inorganic As + MMA + DMA | short | 0.937 | (0.615, 0.993) |
| | Urinary inorganic As + MMA + DMA | short | 0.841 | (0.220, 0.982) |
| 27 | Blood styrene | short | 0.752 | (0.159, 0.951) |
| | Urinary mandelic acid | short | 0.380 | (0.000, 0.854) |
| | 06-styrene-guanine lymphocyte adducts | long | 0.678 | (0.008, 0.935) |
| 28 | Urinary nickel | short | 0.686 | (0.374, 0.913) |
| | Urinary nickel | short | 0.610 | (0.276, 0.886) |
| 29 | Blood mercury | intermediate | 0.330 | (0.000, 0.793) |
| | Urinary mercury | long | 0.591 | (0.000, 0.888) |
| 30 | Plasma cholinesterase | intermediate | 0.896 | (0.820, 0.949) |
| | RBC cholinesterase | intermediate | 0.508 | (0.315, 0.710) |
| 31 | Plasma cholinesterase | intermediate | 0.908 | (0.802, 0.970) |
| | RBC cholinesterase | intermediate | 0.399 | (0.143, 0.722) |
| 32 | Plasma cholinesterase | intermediate | 0.938 | (0.872, 0.977) |
| | RBC cholinesterase | intermediate | 0.327 | (0.090, 0.634) |

[1]short=< 7 days, intermediate=7 days to 4 weeks, long=> 4 weeks; [2]ICC calculated as $\hat{\sigma}_B^2 \Big/ \left(\hat{\sigma}_B^2 + \hat{\sigma}_W^2\right)$

Table 8. Estimated Intra-Class Correlation Coefficients Calculated under Models with an Exponential Error Structure [ICC(EXP)] and Compound Symmetric (CS) Structures [ICC(CS)] for Groups with Serially Correlated Measurements ($p < 0.05$)

| Group | Type | N[1] | $\hat{\rho}^{1\ day\ 2}$ | ICC (EXP)[3] | ICC (CS)[4] |
|---|---|---|---|---|---|
| 14 | Blood mercury | 8 | 0.842 | 0.967 | 0.980 |
| 18 | Urinary tetrachlorophenol | 4 | 0.991 | 0.782 | 0.844 |
| | Urinary pentachlorophenol | 4 | 0.983 | 0.788 | 0.807 |
| 22 | Blood arsenic | 8 | 0.916 | 0 | 0 |
| 24 | Urinary lead | 5 | 0.823 | 0.860 | 0.898 |
| 30 | Plasma cholinesterase | 5 | 0.845 | 0.971 | 0.977 |
| | RBC cholinesterase | 5 | 0.966 | 0.384 | 0.835 |
| 31 | Plasma cholinesterase | 5 | 0.992 | 0.624 | 0.980 |
| | RBC cholinesterase | 5 | 0.902 | 0.653 | 0.768 |
| 32 | Plasma cholinesterase | 5 | 0.868 | 0.982 | 0.986 |
| | RBC cholinesterase | 5 | 0.947 | 0.180 | 0.707 |

[1]Number of repeated measurements collected on each worker

[2]Estimated correlation coefficient for a 1-day interval between measurements

[3]ICC calculated as $\dfrac{\hat{\sigma}_B^2}{\hat{\sigma}_B^2 + \dfrac{\hat{\sigma}_W^2}{n}\left(1 + \dfrac{2}{n}\sum_{j=2}^{n}\sum_{j'=1}^{j-1}\hat{\rho}^{\tau_{jj'}}\right)}$

[4]ICC calculated as $\hat{\sigma}_B^2 \Big/ \left(\hat{\sigma}_B^2 + \hat{\sigma}_W^2/n\right)$

*Conclusions*

This work highlights the importance of quantifying the between- and within-worker sources of variation in exposure, which in turn provides useful information about the degree of measurement error that may be present in an individual-based exposure assessment. Our findings indicate that there was more measurement error in biomarkers with shorter rather than longer half-lives and in airborne measures than in biological measures among those groups studied, although both types of monitoring data provided examples of exposure measures fraught with error. Our results also indicate substantial imprecision in the estimates of exposure measurement error, suggesting that greater attention needs to be given in the future to studies that collect sufficient data to allow for a better characterization of the attenuating effects of an error-prone exposure measure.

## Literature Cited


Anderson TW: [2003] An introduction to multivariate statistical analysis, Wiley-Interscience.

Armstrong BG: [1998] Effect of measurement error on epidemiological studies of environmental and occupational exposures. Occup Environ Med 55:651-656.

Burdorf A, Lillienberg L, Brisman J: [1994] Characterization of exposure to inhalable flour dust in Swedish bakeries. Ann Occup Hyg 38:67-78.

de Cock J, Heederik D, Kromhout H, Boleij JSM, et al.: [1998] Exposure to captan in fruit growing. Am Ind Hyg Assoc J 59:158-165.

Donner A: [1986] A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. Int Stat Rev 54:67-82.

Donner A, Koval JJ: [1983] A note on the accuracy of Fisher's approximation to the large sample variance of an intraclass correlation. Commun Statist Simulat Computat 12:443-449.

Droz PO, Wu MM: [1991] Biological Monitoring Strategies. In: Exposure Assessment for Epidemiology and Hazard Control (eds SM Rappaport, TJ Smith), Lewis Publishing Co, pp 251-270.

Heederik D, Boleij JSM, Kromhout H, Smid T: [1991] Use and analysis of exposure monitoring data in occupational epidemiology: an example of an epidemiological study in the Dutch animal food industry. Appl Occup Environ Hyg 6:458-464.

Hines CJ, Deddens JA: [2001] Determinants of chlorpyrifos exposures and urinary 3,5,6-trichloro-2-pyridinol levels among termiticide applicators. Ann Occup Hyg 45:309-321.

Johnson NL, Kotz S: . [1970] Continuous Univariate Distributions. Houghton-Mifflin.

Kromhout H, Heederik D: [1995] Occupational epidemiology in the rubber industry: implications of exposure variability. Am J Ind Med 27:171-185.

Kromhout H, Oostendorp Y, Heederik D, Boleij JSM: [1987] Agreement between qualitative exposure estimates and quantitative exposure measurements. Am J Ind Med 12:551-562.

Kromhout H, Symanski E, Rappaport SM: [1993] A comprehensive evaluation of within- and between-worker components of occupational exposure to chemical agents. Ann Occup Hyg 37:253-270.

Kumagai S, Kusaka Y, Goto S: [1996] Cobalt exposure level and variability in the hard metal industry of Japan. Am Ind Hyg Assoc J 57:365-369.

Lagorio S, Iavarone I, Iacovella N, Proietto AR, Fuselli S, Baldassarri LT, Carere, A: [1997] Variability of benzene exposure among filling station attendants. Occup Hyg 4:15-30.

Lin YS, Kupper LL, Rappaport SM: [2005] Air samples versus biomarkers for epidemiology. Occup Environ Med 62:750-760.

Makinen M, Kangas J, Kalliokoski P: [2000] Applicability of homogeneous exposure groups for exposure assessment in the chemical industry. Int Arch Occup Environ Health 73:471-8.

Milton DK, Walters MD, Hammond K, Evans JS: [1996] Worker exposure to endotoxin, phenolic compounds, and formaldehyde in a fiberglass insulation manufacturing plant. Am Ind Hyg Assoc J 57:889-896.

Neter J, Kutner MH, Nachtsheim CJ, Wasserman W: [1996] Applied Linear Statistical Models, WCB/McGraw Hill.

Nicas M, Simmons BP, Spear RC: [1991] Environmental versus analytical variability in exposure measurements. Am Ind Hyg Assoc J 52:553-557.

Nieuwenhuijsen MJ, Lowson D, Venables KM, Newman Taylor AJ: [1995] Flour dust exposure variability in flour mills and bakeries. Ann Occup Hyg 39:299-305.

Peretz C, Goldberg P, Kahan E, Grady S, Goren A: [1997] The variability of exposure over time: a prospective longitudinal study. Ann Occup Hyg 41:485-500.

Rappaport SM: [1985] Smoothing of exposure variability at the receptor: Implications for health standards. Ann Occup Hyg 29:201-214.

Rappaport SM: [1991] Assessment of long-term exposures to toxic substances in air. Ann Occup Hyg 35:61-121.

Rappaport SM, Symanski E, Yager JW, Kupper LL: [1995] The relationship between environmental monitoring and biological markers in exposure assessment. Environ Health Perspect 103(suppl. 3):49-54.

Searle SR: [1971] Linear Models, Wiley, pp 57-59.

Searle SR, Casella G, McCulloch CE: [1992] Variance Components. John Wiley and Sons.

Simcox NJ, Camp J, Kalman D, Stebbins A, Bellamy G, Lee I-C, Fenske R: [1999]. Farmworker exposure to organophosphorus pesticide residues during apple thinning in Central Washington State. Am J Ind Hyg Assoc J 60:752-761.

Symanski E, Bergamaschi E, Mutti A: [2001b] Inter- and intra-individual sources of variation in levels of urinary styrene metabolites. Int Arch Environ Occup Health 74:336-344.

Symanski E, Greeson NMH: [2002] Assessment of Variability in Biomonitoring Data Using a Large Database of Biological Measures of Exposure. Am Ind Hyg Assoc J 63:390-401.

Symanski E, Greeson NMH, Chan W [2007]. Evaluating measurement error in estimates of worker exposure assessed in parallel by personal and biological monitoring. Am J Ind Med 50:112-121.

Symanski E, Sällsten G, Barregård L [2000]. Variability in airborne and biological measures of exposure to mercury in the chloralkali industry: implications for epidemiologic studies. Environ Health Perspect 108:569-573.

Thomas D, Stram D, Dwyer J: [1993] Exposure measurement error: Influence on exposure disease relationships and methods for correction. Annu Rev Public Health 14:69-93.

United Nations: [1990] International Standard Industrial Classification of All Economic Activities (Statistical papers series M, No 4, Rev*). New York: United Nations.

Woskie SR, Shen P, Eisen EA, Finkel MH, Smith TJ, Smith R, Wegman DH [1994] The real-time dust exposures of sodium borate workers: examination of exposure variability. Am Ind Hyg Assoc J 55:207-217.

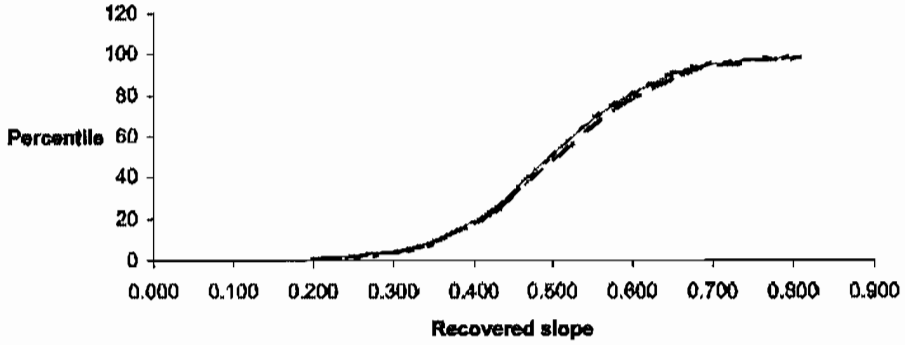
**PUBLICATIONS**

Symanski E, Greeson NMH, Chan W: [2007] Evaluating measurement error in estimates of worker exposure assessed in parallel by personal and biological monitoring. Am J Ind Med 50:112-121.

# APPENDIX A

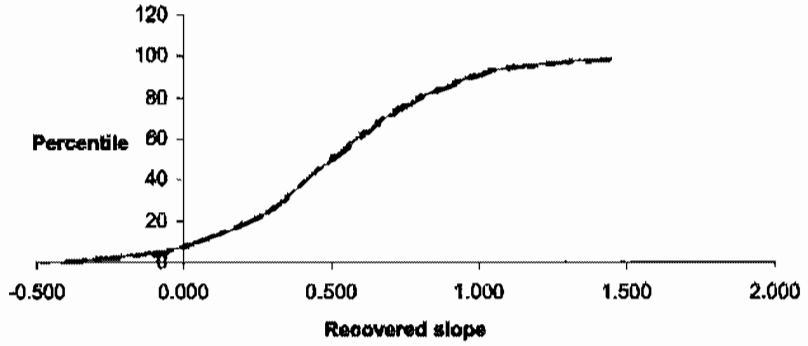
Comparison of the Empirical Distribution Functions (CDFs) of the Recovered Slope using the Proposed and the Traditional Methods

**CDF of Recovered Slope for $\sigma_W^2$ , $\sigma_1^2$ =0.1**



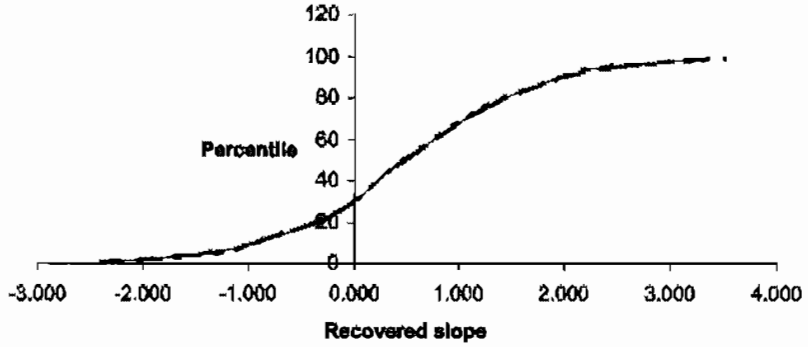

$E(\hat{\beta}|\overline{z},y)$ $\hat{b}$ $\frac{\hat{b}}{\hat{\sigma}_B^2/(\hat{\sigma}_B^2+\hat{\sigma}_W^2)}$ $\hat{\beta}$

**CDF of Recovered Slope for $\sigma_W^2$ =0.1, $\sigma_1^2$ =1**



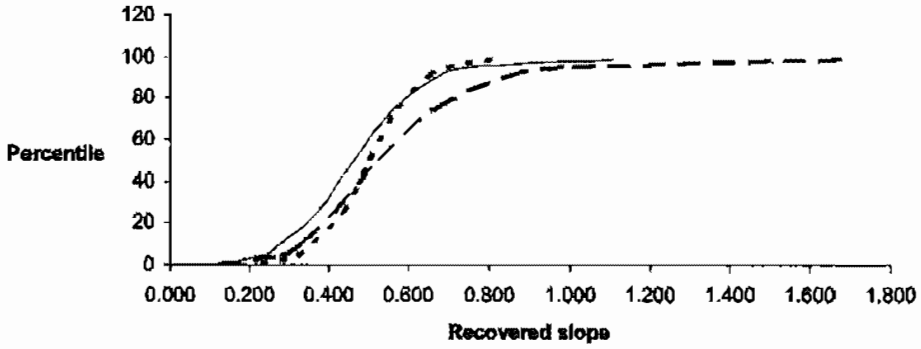

$E(\hat{\beta}|\overline{z},y)$ $\hat{b}$ $\frac{\hat{b}}{\hat{\sigma}_B^2/(\hat{\sigma}_B^2+\hat{\sigma}_W^2)}$ $\hat{\beta}$

**CDF of Recovered Slope for $\sigma_W^2$ =0.1, $\sigma_1^2$ =10**



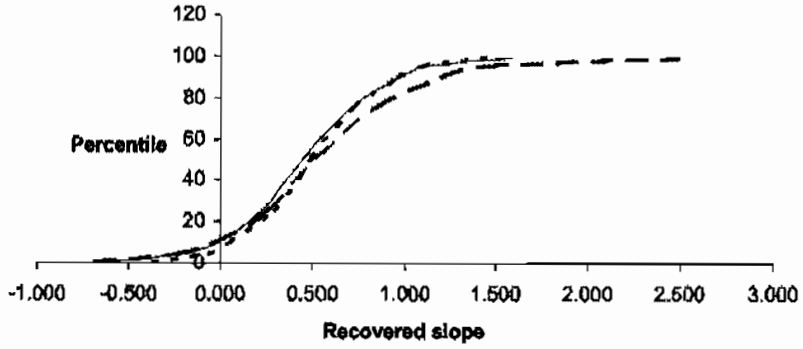

$E(\hat{\beta}|\bar{x},y)$ $\quad$ $\frac{\hat{b}}{\hat{\sigma}_B^2/(\hat{\sigma}_B^2+\hat{\sigma}_W^2)}$ $\quad$ $\hat{\beta}$

**CDF of Recovered Slope for $\sigma_W^2$ =1, $\sigma_1^2$ =0.1**



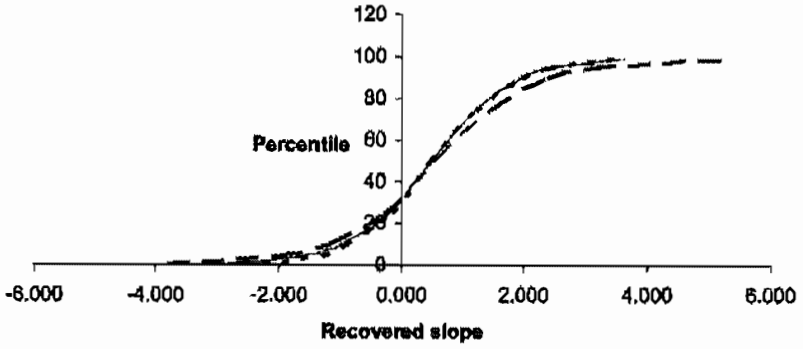

$E(\hat{\beta}|\bar{x},y)$ $\quad$ $\frac{\hat{b}}{\hat{\sigma}_B^2/(\hat{\sigma}_B^2+\hat{\sigma}_W^2)}$ $\quad$ $\hat{\beta}$

**CDF of Recovered Slope for $\sigma_w^2$ =1, $\sigma_1^2$ =1**



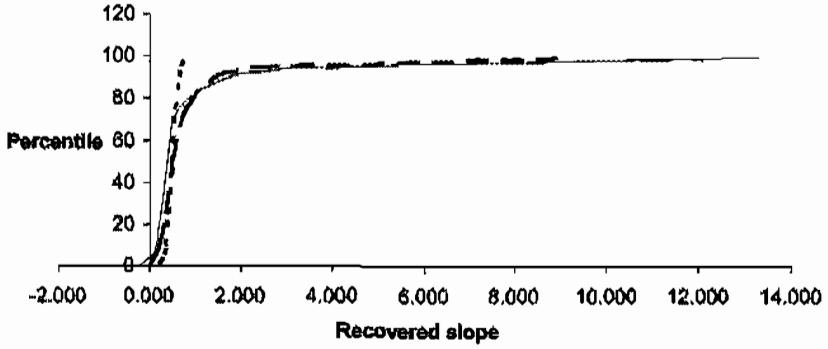

$E(\hat{\beta}|\overline{z},y)$ — — $\frac{\hat{b}}{\hat{\sigma}_B^2/(\hat{\sigma}_B^2+\hat{\sigma}_W^2)}$ - - - - $\hat{\beta}$

**CDF of Recovered Slope for $\sigma_w^2$ =1, $\sigma_1^2$ =10**



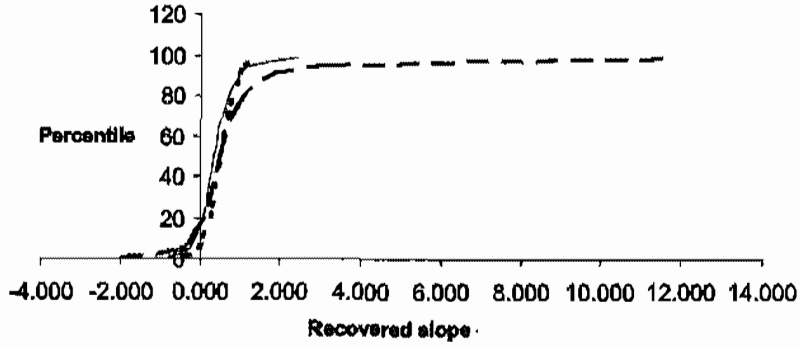

$E(\hat{\beta}|\overline{z},y)$ — — $\frac{\hat{b}}{\hat{\sigma}_B^2/(\hat{\sigma}_B^2+\hat{\sigma}_W^2)}$ - - - - $\hat{\beta}$

**CDF of Recovered Slope for $\sigma_W^2$=4, $\sigma_1^2$=0.1**



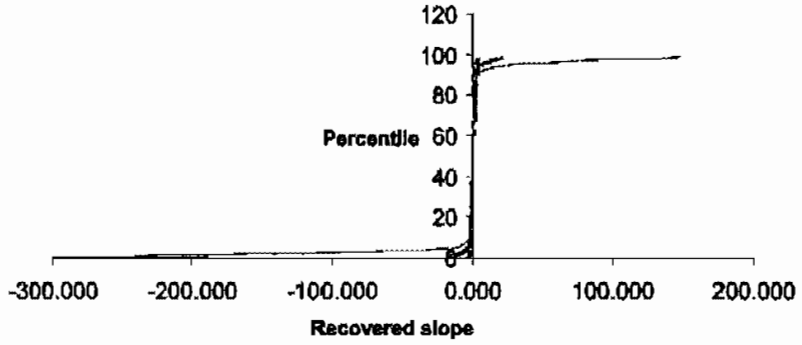

$E(\hat{\beta}|\bar{z},y)$ — $\frac{\hat{b}}{\hat{\sigma}_B^2/(\hat{\sigma}_B^2+\hat{\sigma}_W^2)}$ - - - - $\hat{\beta}$

**CDF of Recovered Slope for $\sigma_W^2$=4, $\sigma_1^2$=1**



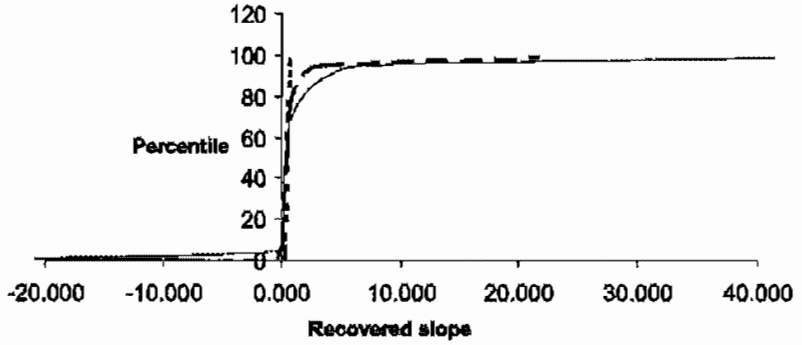

$E(\hat{\beta}|\bar{z},y)$ — $\frac{\hat{b}}{\hat{\sigma}_B^2/(\hat{\sigma}_B^2+\hat{\sigma}_W^2)}$ - - - - $\hat{\beta}$

**CDF of Recovered Slope for $\sigma_W^2$ =4, $\sigma_1^2$ =10**



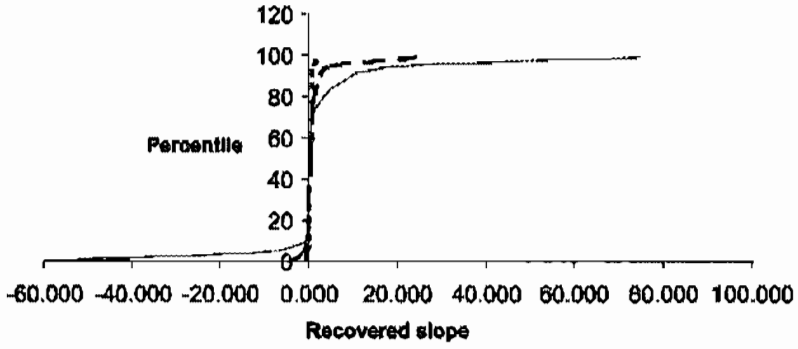

$E(\hat{\beta}|\overline{z},y)$ — $\frac{\hat{b}}{\hat{\sigma}_B^2/(\hat{\sigma}_B^2+\hat{\sigma}_W^2)}$ - - - - $\hat{\beta}$

**CDF of Recovered Slope for $\sigma_W^2$ =10, $\sigma_1^2$ =0.1**



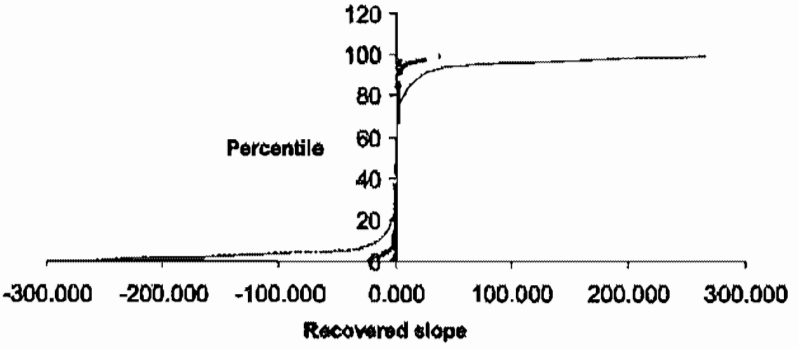

$E(\hat{\beta}|\overline{z},y)$ — $\frac{\hat{b}}{\hat{\sigma}_B^2/(\hat{\sigma}_B^2+\hat{\sigma}_W^2)}$ - - - - $\hat{\beta}$

**CDF of Recovered Slope for $\sigma_w^2$ =10, $\sigma_1^2$ =1**




$E(\hat{\beta}|\overline{z},y)$ — — $\frac{\hat{b}}{\hat{\sigma}_B^2/(\hat{\sigma}_B^2+\hat{\sigma}_w^2)}$ · · · · $\hat{\beta}$

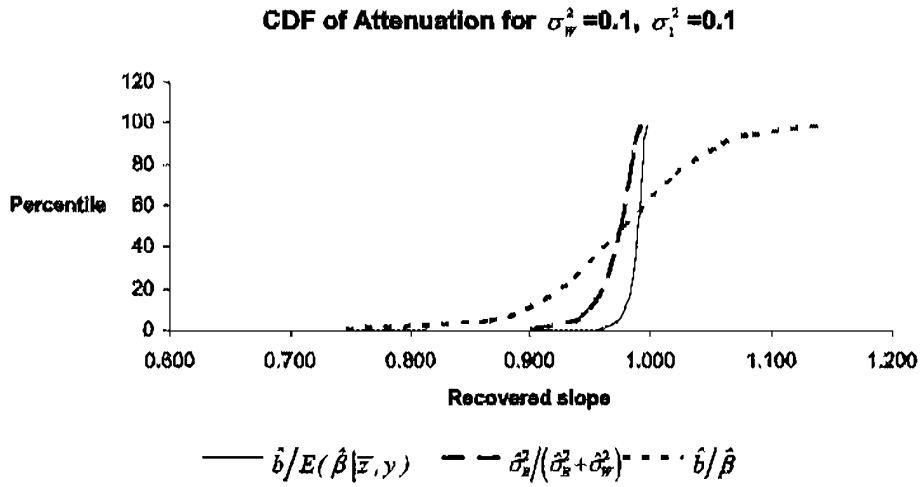
**CDF of Recovered Slope for $\sigma_w^2$ =10, $\sigma_1^2$ =10**

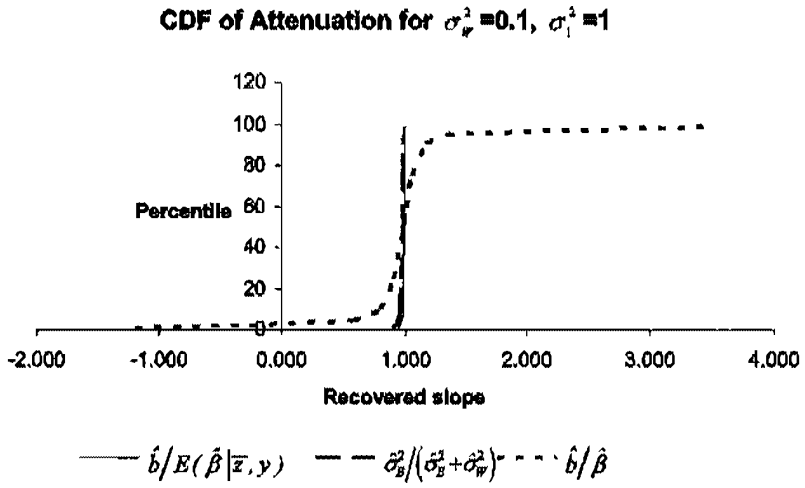



$E(\hat{\beta}|\overline{z},y)$ — — $\frac{\hat{b}}{\hat{\sigma}_B^2/(\hat{\sigma}_B^2+\hat{\sigma}_w^2)}$ · · · · $\hat{\beta}$

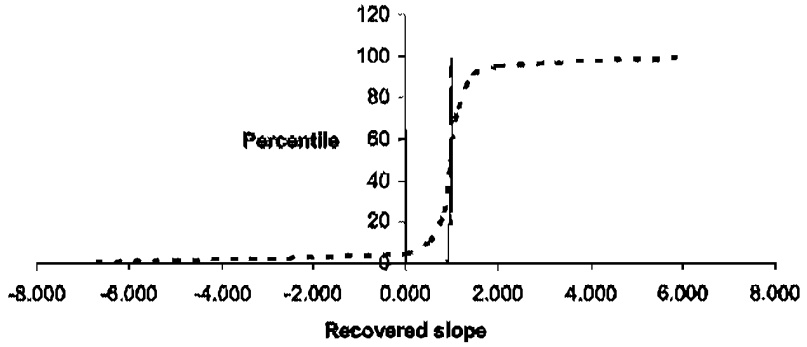
# APPENDIX B

## Comparison of the Cumulative Distribution Functions of Attenuation Using the Proposed and Traditional Methods


CDF of Attenuation for $\sigma_w^2$ =0.1, $\sigma_1^2$ =0.1
120
100
80
60
40
20
0
Percentile
0.600
0.700
0.800
0.900
1.000
1.100
1.200
Recovered slope
$\hat{b}/E(\hat{\beta}|\bar{x},y)$
$\sigma_B^2/(\sigma_B^2+\sigma_w^2)$
$\hat{b}/\hat{\beta}$

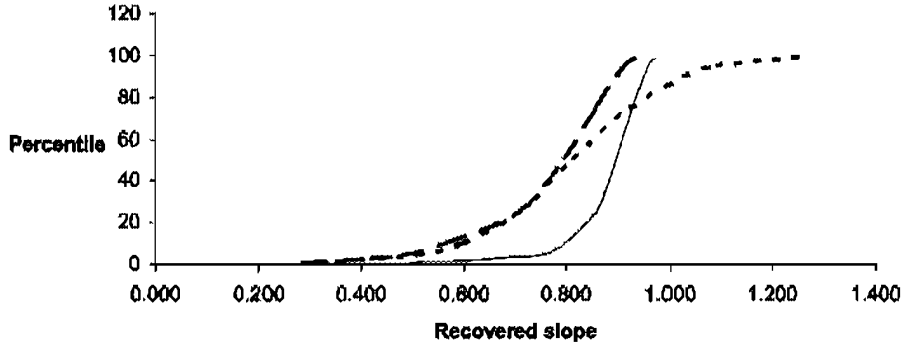

CDF of Attenuation for $\sigma_w^2$ =0.1, $\sigma_1^2$ =1
120
100
80
60
40
20
0
Percentile
-2.000
-1.000
0.000
1.000
2.000
3.000
4.000
Recovered slope
$\hat{b}/E(\hat{\beta}|\bar{x},y)$
$\sigma_B^2/(\sigma_B^2+\sigma_w^2)$
$\hat{b}/\hat{\beta}$

**CDF of Attenuation for $\sigma_w^2$ =0.1, $\sigma_1^2$ =10**

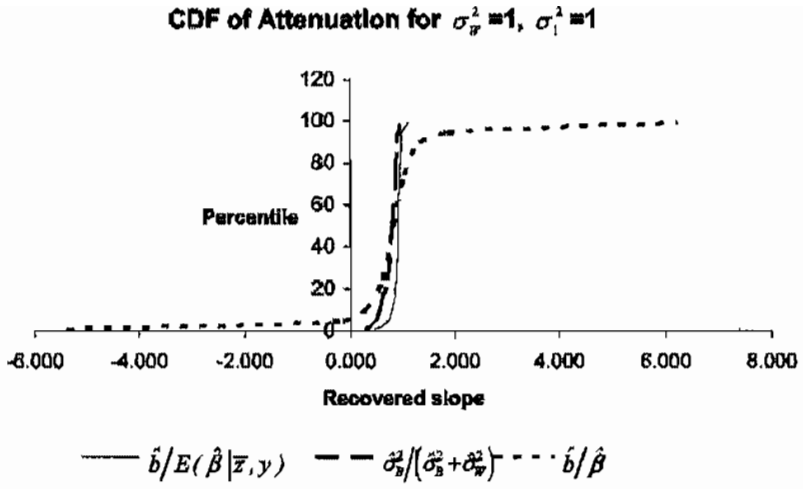


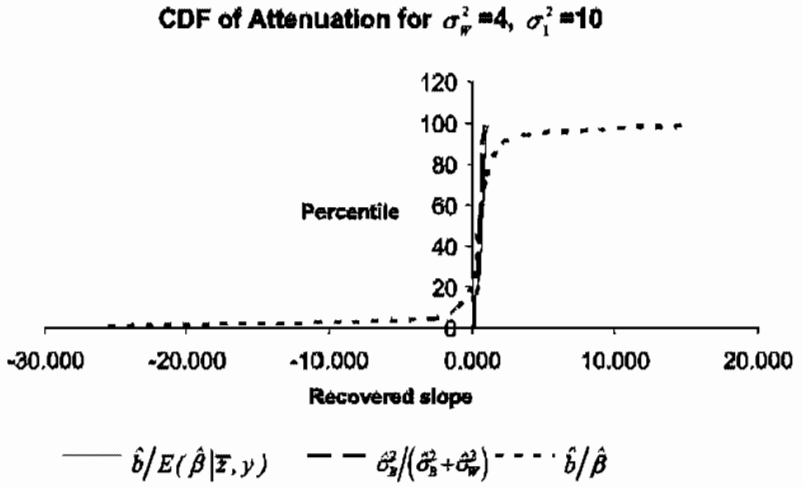

$\hat{b}/E(\hat{\beta}|\bar{z},y)$ $\sigma_\beta^2/(\sigma_\beta^2+\sigma_w^2)$ $\hat{b}/\hat{\beta}$

**CDF of Attenuation for $\sigma_w^2$ = 1 , $\sigma_1^2$ =0.1**

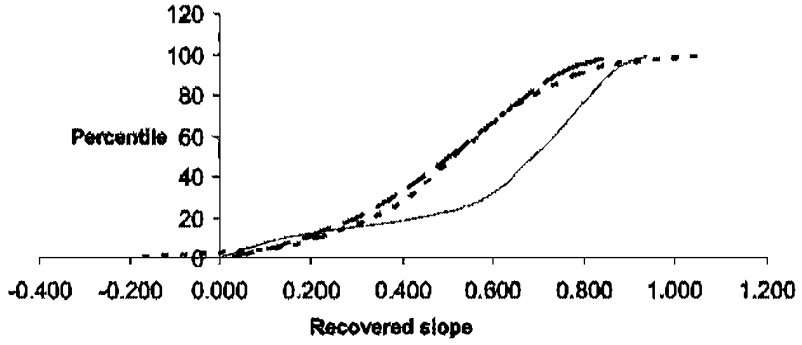



$\hat{b}/E(\hat{\beta}|\bar{z},y)$ $\sigma_\beta^2/(\sigma_\beta^2+\sigma_w^2)$ $\hat{b}/\hat{\beta}$


CDF of Attenuation for $\sigma_w^2$ =1, $\sigma_1^2$ =1
120
100
80
60
40
20
0
Percentile
-8.000
-4.000
-2.000
0.000
2.000
4.000
6.000
8.000
Recovered slope
$\hat{b}/E(\hat{\beta}|\bar{z},y)$
$\sigma_B^2/(\sigma_B^2+\sigma_W^2)$
$\hat{b}/\hat{\beta}$

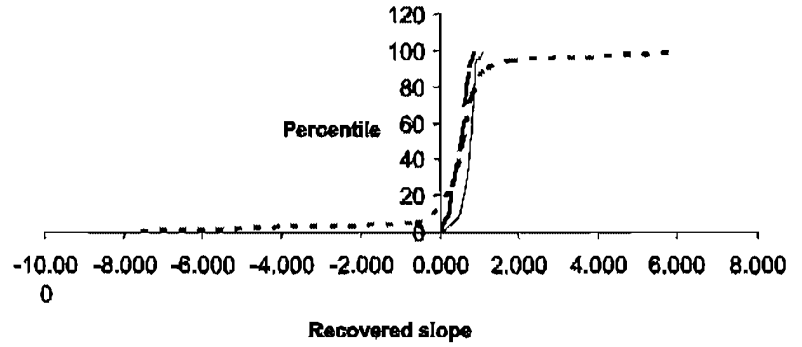

CDF of Attenuation for $\sigma_w^2$ =4, $\sigma_1^2$ =10
120
100
80
60
40
20
0
Percentile
-30.000
-20.000
-10.000
0.000
10.000
20.000
Recovered slope
$\hat{b}/E(\hat{\beta}|\bar{z},y)$
$\sigma_B^2/(\sigma_B^2+\sigma_W^2)$
$\hat{b}/\hat{\beta}$

**CDF of Attenuation for $\sigma_W^2$ =4, $\sigma_1^2$ =0.1**

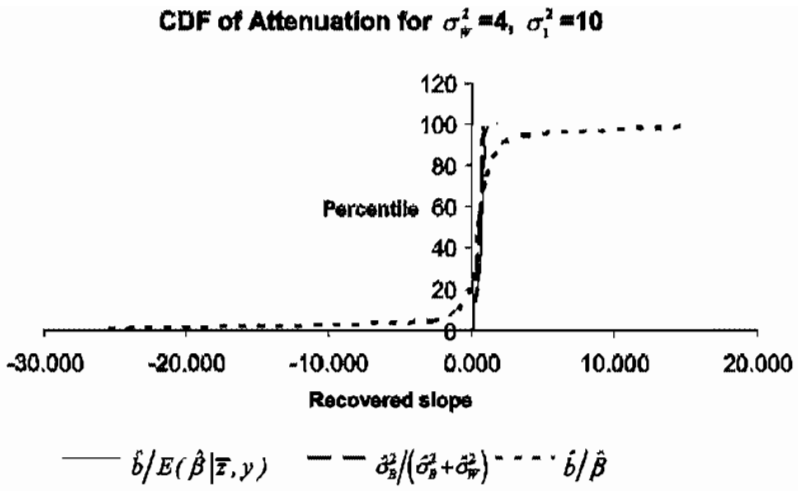


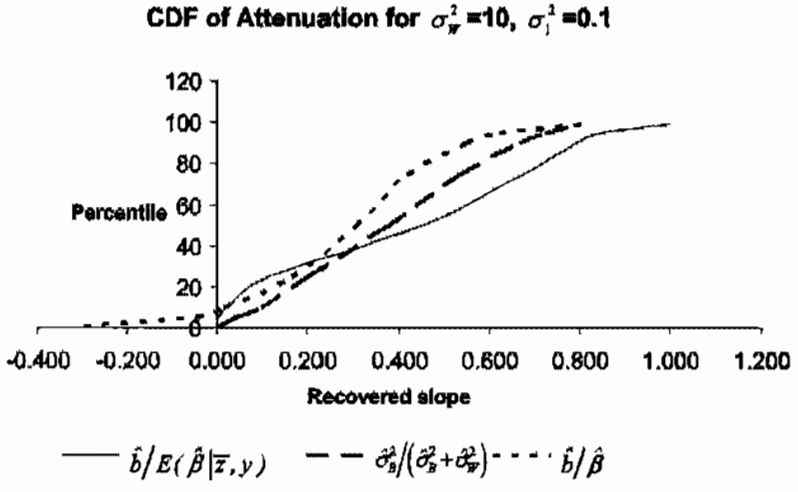

$\hat{b}/E(\hat{\beta}|\bar{z},y)$ $\hat{\sigma}_B^2/(\hat{\sigma}_B^2+\hat{\sigma}_W^2)$ $\hat{b}/\hat{\beta}$

**CDF of Attenuation for $\sigma_W^2$ =4, $\sigma_1^2$ =1**

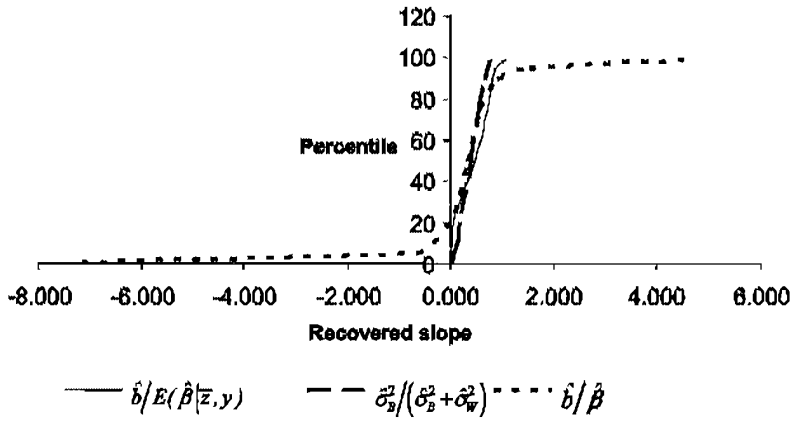



$\hat{b}/E(\hat{\beta}|\bar{z},y)$ $\hat{\sigma}_B^2/(\hat{\sigma}_B^2+\hat{\sigma}_W^2)$ $\hat{b}/\hat{\beta}$


CDF of Attenuation for $\sigma_w^2$ =4, $\sigma_1^2$ =10
120
100
80
Percentile 60
40
20
0
-30.000
-20.000
-10.000
0.000
10.000
20.000
Recovered slope
$\hat{b}/E(\hat{\beta}|\bar{x},y)$
$\hat{\sigma}_B^2/(\hat{\sigma}_B^2+\hat{\sigma}_W^2)$
$\hat{b}/\hat{\beta}$

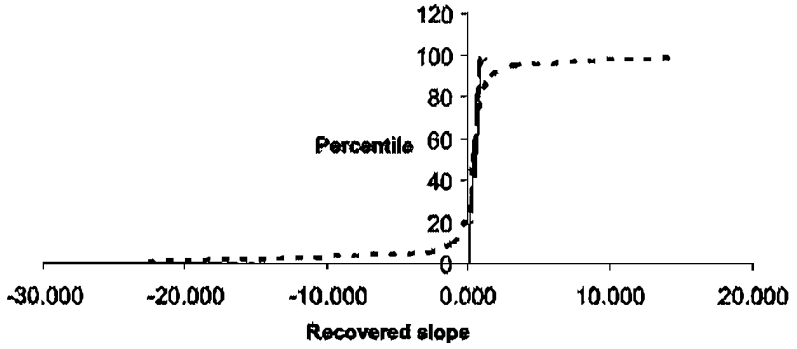

CDF of Attenuation for $\sigma_w^2$ =10, $\sigma_1^2$ =0.1
120
100
80
Percentile 60
40
20
0
-0.400
-0.200
0.000
0.200
0.400
0.600
0.800
1.000
1.200
Recovered slope
$\hat{b}/E(\hat{\beta}|\bar{x},y)$
$\hat{\sigma}_B^2/(\hat{\sigma}_B^2+\hat{\sigma}_W^2)$
$\hat{b}/\hat{\beta}$

**CDF of Attenuation for $\sigma_w^2$ =10, $\sigma_1^2$ =1**




$\hat{b}/E(\hat{\beta}|\bar{z},y)$ $\hat{\sigma}_B^2/(\hat{\sigma}_B^2+\hat{\sigma}_W^2)$ $\hat{b}/\hat{\beta}$

**CDF of Attenuation for $\sigma_w^2$ =10, $\sigma_1^2$ =10**




$\hat{b}/E(\hat{\beta}|\bar{z},y)$ $\hat{\sigma}_B^2/(\hat{\sigma}_B^2+\hat{\sigma}_W^2)$ $\hat{b}/\hat{\beta}$