Nonparametric simulation extrapolation for measurement-error models

Dylan SPICKER\(^1\*\), Michael P. WALLACE\(^2\), and Grace Y. YI\(^3\)

\(^1\)Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada
\(^2\)Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada
\(^3\)Department of Statistical and Actuarial Sciences, Department of Computer Science, University of Western Ontario, London, Ontario, Canada

Key words and phrases: Heavy-tailed errors; kernel density estimation; measurement error; non-normal errors; remeasurement procedures; simulation extrapolation.

Abstract: The presence of measurement error is a widespread issue, which, when ignored, can render the results of an analysis unreliable. Numerous corrections for the effects of measurement error have been proposed and studied, often under the assumption of a normally distributed, additive measurement-error model. In many situations, observed data are nonsymmetric, heavy-tailed, or otherwise highly non-normal. In these settings, correction techniques relying on the assumption of normality are undesirable. We propose an extension of simulation extrapolation that is nonparametric in the sense that no specific distributional assumptions are required on the error terms. The technique can be implemented when either validation data or replicate measurements are available, and is designed to be immediately accessible to those familiar with simulation extrapolation.

1. INTRODUCTION

Measurement error, where a variate of interest is not accurately observed, is a pervasive issue that can undermine the validity of an analysis. Numerous methods exist that correct for the effects of measurement error. These techniques commonly assume an additive model with

*Corresponding author: dylan.spicker@mcgill.ca

© 2023 The Authors. The Canadian Journal of Statistics/La revue canadienne de statistique
published by Wiley Periodicals LLC on behalf of Statistical Society of Canada / Société statistique du Canada.
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
distributed errors. This assumption, though appealing in its simplicity, is often unreasonable in real-world applications. See, for instance, National Research Council (1986), Nusser et al. (1996), Bollinger (1998), Purdom & Holmes (2005), McKenzie et al. (2008), Xu, Kim & Li (2017), and Rajan & Desai (2018). When the assumptions made are difficult to test, such as those regarding the distribution of error terms, concerns regarding the validity of correction procedures are amplified.

The accommodation of non-normal errors is therefore an important area of study. Nonparametric and semiparametric methods, which do not impose strict assumptions on the distribution of error terms, are important in providing flexible ways to correct for the effects of measurement error; see, for instance, Li & Vuong (1998), Gorfine, Hsu & Prentice (2004), Carroll et al. (2006), Schennach & Hu (2013), Xu, Kim & Li (2017), and Yi (2017). In addition to these nonparametric and semiparametric methods, several parametric techniques have been developed to account for non-normal errors (Augustin, 2004; Koul & Song, 2014).

In order to facilitate error correction in the presence of non-normal errors, we present an extension of the commonly used simulation extrapolation (SIMEX) method (Cook & Stefanski, 1994). Our extension does not assume that errors are normally distributed. Instead, we propose a nonparametric SIMEX procedure that consistently corrects for the effects of additive measurement error regardless of the distribution of the error. We refer to the standard SIMEX procedure as P-SIMEX, or the parametric SIMEX, to distinguish it from our nonparametric procedure, which we refer to as NP-SIMEX. It has been shown that P-SIMEX can be resilient to deviations from the assumption of error normality in some settings (Cook & Stefanski, 1994). Other authors have shown that when errors are non-normal, the bias resulting from P-SIMEX correction can be substantial (Yi & He, 2012; Koul & Song, 2014).

P-SIMEX is a three-step procedure consisting of a simulation step, an estimation step, and an extrapolation step. The simulation step of P-SIMEX has been described as a remeasurement method (Novick & Stefanski, 2002), emphasizing P-SIMEX’s similarities to bootstrap procedures. Just as bootstrap procedures can be made nonparametric by resampling from an empirical distribution, P-SIMEX can be made nonparametric by remeasuring using the empirical error distribution. Doing so allows assumptions regarding the distribution of the error terms to be relaxed. We make explicit this nonparametric remeasurement procedure. This allows NP-SIMEX to accommodate a wide range of error models without making specific distributional assumptions, which distinguishes it from parametric extensions of SIMEX to Laplace errors (Koul & Song, 2014).

2. BACKGROUND

2.1. Notation and Available Data

Suppose that we have a sample indexed by \( i \in \{1, \ldots, n\} \). We take \( Y_i \) to represent an outcome of interest, \( X_i \) to be an explanatory variable subject to measurement error, and \( Z_i \) to be a vector of explanatory variables measured without error. The explanatory variable \( X_i \) is taken to be univariate for ease of notation. We are concerned with estimating some parameter \( \theta \) that relates the distribution of \( Y_i \) to \( X_i \) and \( Z_i \). Instead of observing \( X_i \), we observe \( X_i^* = X_i + U_i \), where \( U_i \) is the error term, which we assume to be independent of \( Y_i, X_i, \) and \( Z_i \). We assume that, in the error-free setting, we have an estimator \( \hat{\theta}(Y_i, X_i, Z_i)_{i=1}^n \) that is consistent for \( \theta \).

Generally, measurement-error-correction techniques rely on auxiliary data to infer information about the errors. Validation data, which can either be internal or external, involve the observation of the true \( X_i \) alongside the proxy measurement \( X_i^* \) for some set of individuals. Internal validation data arise in the situation where a subset of the sample of interest has this true measurement taken. That is, \( \{Y_i, X_i, X_i^*, Z_i\} \) is observed for some subset of the \( n \) total observations, while only \( \{Y_i, X_i^*, Z_i\} \) is observed for the remainder of the sample. External
validation data arise in the setting where, inside the sample of interest, only \( \{ Y_i, X_i^*, Z_i \} \) is measured, but pairs \( \{ X_i, X_i^* \} \) are observed in a separate dataset. In order to make use of external validation data, we must make assumptions regarding the transportability of the error models. We assume that the error process for the external sample is equivalent to the error process within the sample of interest. This assumption allows us to use information from the external sample.

When validation samples are not available, we may instead use replicate measurements under some assumptions about the model. Here we observe repeated proxy measurements \( \{ Y_i, X_{i1}^*, \ldots, X_{ik}^*, Z_i \} \) for all \( i \), where, for \( j \in \{ 1, \ldots, \kappa \} \),

\[
X_{ij}^* = X_i + U_{ij},
\]

and \( \kappa \) is the number of replicates. It is further assumed that \( U_{ij1}, \ldots, U_{ijn} \) are independent and identically distributed according to \( U_i \), and each \( U_{ij} \) is independent of \( \{ X_i, Z_i, Y_i \} \). These auxiliary data can be used to estimate the variance of \( U_i \), denoted \( \sigma^2_U \).

### 2.2. Parametric Simulation Extrapolation

We begin by briefly presenting P-SIMEX. For full details of this technique, see Cook & Stefanski (1994) and Carroll et al. (1996). In P-SIMEX, we assume that the error term \( U_i \) is normally distributed with mean zero and constant variance. We also assume that auxiliary data exist, which allow for an estimate \( \hat{\sigma}^2_U \) of \( \sigma^2_U \). Once \( \hat{\sigma}^2_U \) has been estimated, P-SIMEX fixes a non-negative, real-valued grid \( \Lambda_p \) with \( M \) elements, where \( M \) is an analyst-specified positive integer. For each \( \lambda_p \in \Lambda_p \) and each \( i \in \{ 1, \ldots, n \} \), we define the quantity \( X_{bi}^*(\lambda_p) = \overline{X}_i^* + (\lambda_p^{1/2} \hat{\sigma}_U) \nu_{bi} \), where \( \nu_{bi} \sim \mathcal{N}(0, 1) \) is independent of all other terms generated by the analyst, and \( \overline{X}_i^* \) is the sample average of the \( X_{i}^* \)’s. This is considered to be a remeasured dataset, which is then used to compute \( \hat{\theta}_b(\{ Y_i, X_{bi}^*(\lambda_p), Z_i \}_{i=1}^n) \) by replacing \( X_i \) with \( X_{bi}^*(\lambda_p) \) in the estimator obtained from the standard estimation method in error-free contexts. This process is repeated for \( b \in \{ 1, \ldots, B \} \), with \( B \) specified by the analyst, to produce a set of estimates which are then averaged as

\[
\hat{\theta}(\{ Y_i, X_i^*(\lambda_p), Z_i \}_{i=1}^n) = B^{-1} \sum_{b=1}^B \hat{\theta}_b(\{ Y_i, X_{bi}^*(\lambda_p), Z_i \}_{i=1}^n).
\]

After repeating this process across the entire grid, the analyst will have a set of estimates and the corresponding \( \lambda_p \) values, represented as \( \{ (\lambda_p, \hat{\theta}(\{ Y_i, X_i^*(\lambda_p), Z_i \}_{i=1}^n) : \lambda_p \in \Lambda_p \} \}. These pairs of values can be used to fit a parametric relationship \( G(\lambda) \) via least squares estimation to give \( \hat{G}(\lambda_p) \), which describes \( \hat{\theta} \) as a function of \( \lambda_p \). Intuitively, \( X_{bi}^*(\lambda) \) is such that \( E[X_{bi}^*(\lambda)|X_i] = X_i \) and \( \text{var}(X_{bi}^*(\lambda)|X_i) = (\lambda + 1)\sigma^2_U \). With \( \lambda = -1 \), \( X_{bi}^* \) behaves as though it is \( X_i \). The P-SIMEX estimator is taken to be \( \hat{\theta}_{P-SIMEX} = \hat{G}(-1) \). Cook & Stefanski (1994) initiate this algorithm and Carroll et al. (1996) demonstrate that, under regularity conditions and the correct specification of \( G(\cdot) \), \( \hat{\theta}_{P-SIMEX} \) is a consistent estimator of \( \theta \).

If the measurement error is not normally distributed or if \( G(\cdot) \) is not correctly specified, this method will produce only approximately consistent estimators with nonzero asymptotic bias. Koul & Song (2014) provide a framework that can be adapted to different parametric error distributions by changing the distribution of \( \nu_{bi} \) and allowing \( \lambda_p \) to impact parameters other than the variance. Their method provides a flexible way to accommodate errors when there is a scientific rationale for specifying the distribution. Our proposal circumvents this requirement when no such rationale exists.
3. NONPARAMETRIC SIMULATION EXTRAPOLATION

3.1. Estimation Procedure

The key property of P-SIMEX is that the characteristic function of $U_i + [\lambda \sigma_U^2]^{1/2} \nu_{bi}$ tends to 1 as $\lambda$ tends to $-1$ (Koul & Song, 2014). We can exploit this property nonparametrically. Suppose that we are able to form the set $U^*$, which contains all observed error terms for all individuals. Sampling $U^*$ from $U$ is then sampling from the empirical error distribution. Denote the characteristic function and the distribution function of a random variable $X$ by $\phi_X$ and $F_X$, respectively. Denote empirical distributions and characteristic functions by $\hat{F}_X$ and $\hat{\phi}_X$, respectively. That is, $\hat{F}_X(x) = n^{-1}[I(X_1 \leq x) + \cdots + I(X_n \leq x)]$, where $I(\cdot)$ is the indicator function.

Note that $U^*$ drawn from $U^*$ has the distribution function $\hat{F}_U$ and the characteristic function $\hat{\phi}_U(t)$. Denote the cardinality of a set $S$ as $|S|$ and take $|U^*| = m$. Then as $m \to \infty$, $\hat{\phi}_U(t) \to \phi_U(t)$ for any $t \in \mathbb{R}$. Sampling $U_1^*, U_2^*, \ldots, U^*_\lambda$ from $U^*$, where $\lambda$ is a positive integer, the characteristic function of $U + U_1^* + \cdots + U^*_\lambda$ converges pointwise to $\phi_U$ as $m \to \infty$ since, for independent random variables $X$ and $W$, the sum $X + W$ has the characteristic function $\phi_X \phi_W$. This function converges to 1 as we extrapolate $\lambda$ to $-1$. This suggests the proposed NP-SIMEX procedure below.

1. Form the set $U^*$.
2. Specify a fixed grid $\Lambda$ of non-negative integers.
3. For each $\lambda \in \Lambda$, $b \in \{1, \ldots, B\}$, and every $i$, form

$$\tilde{X}_b^*(\lambda) = X_i^* + \sum_{j=1}^{\lambda} U_{bi,j}^*,$$

where the $U_{bi,j}^*$s are sampled independently with replacement from $U^*$.
4. Using $\tilde{X}_b^*(\lambda)$, compute $\hat{\theta}_b(\{Y_i, X_b^*(\lambda), Z_i\})$ for $b \in \{1, \ldots, B\}$. Then compute

$$\hat{\theta}(\{Y_i, \tilde{X}_b^*(\lambda), Z_i\}_{i=1}^n) = B^{-1} \sum_{b=1}^B \hat{\theta}_b(\{Y_i, \tilde{X}_b^*(\lambda), Z_i\}).$$

5. Fit a parametric regression model to $\{(\lambda, \hat{\theta}(\{Y_i, \tilde{X}_b^*(\lambda), Z_i\}_{i=1}^n)) : \lambda \in \Lambda\}$ and then extrapolate to $\lambda = -1$.

This procedure relies on being able to form the set $U^*$. The method for doing this depends on the auxiliary data that are available.

3.2. Forming the Empirical Error Distribution

If we observe an internal validation sample, such that for $i \in \{1, \ldots, n_1\}$ we observe $\{Y_i, X_i, X_i^*, Z_i\}$ and for $i = n_1 + 1, \ldots, n$ we observe $\{Y_i, X_i^*, Z_i\}$, then we have implicitly observed a subsample including $U_i$. For $i \in \{1, \ldots, n_1\}$, we can define $U_i = X_i^* - X_i$. Supposing then that the additive measurement error model is correct, we can form $U^*$ as $\{U_i\}_{i=1}^{n_1}$. If we have an external validation sample, such that we have $\{Y_i, X_i^*, Z_i\}$ observed for $i \in \{1, \ldots, n\}$ and $\{X_i, X_i^*\}$ observed in a separate dataset for $i \in \{1, \ldots, n_1\}$, we can perform the same process and form $U^*$ as the set of $X_i^* - X_i$ for $i \in \{1, \ldots, n_1\}$. We require the same assumptions as with an internal validation sample, as well as the transportability assumption for external validation data. In either case, no further restrictions are required on the distribution of $U$.
With replicate measurements, we form \( \mathcal{U} \) by restricting the error distributions that we consider to only those that are symmetric around a known constant. Given that observed errors often follow heavy-tailed \( t \)-distributions, this assumption may be defensible (Bailey, 2017; Rajan & Desai, 2018). We take the known constant to be 0 without loss of generality. Assume that for all \( i \in \{1, \ldots, n\} \), we observe \( \{Y_i, X^*_i, \ldots, X^{n_i}_i, Z_i\} \), where \( X^{n_i}_i \) are as in model (1). Take \( \kappa = 2 \) as an example. Then consider \( \tilde{X}_i = 0.5 \left( X^*_i + X^{n_i}_i \right) \). By model (1), \( \tilde{X} = X_i + 0.5 \left( U_i + U_{i2} \right) \), which can be viewed as an error-prone measurement of \( X_i \). If we define \( \bar{U}_i = 0.5 \left( X^{n_i}_i - X^{*}_i \right) \), then by symmetry we have that \( \bar{U}_i = 0.5 \left( U_i - U_{i2} \right) \). For the consistency of \( \bar{U}_i \), we observe \( \bar{U}_i = X_i + \tilde{U}_i = X_i + u_i + \bar{U}_i \). As a result, we can form \( \bar{U} = \{\bar{U}_1, \ldots, \bar{U}_n\} \), and conduct our analysis using the mean response \( \bar{X}_i \) in place of \( X_i \), and sampling empirical errors from \( \bar{U} \). This procedure also can be applied when \( \kappa \neq 2 \) (see the details in the Appendix).

3.3. Illustration and Theoretical Justification

We first illustrate our procedure with a simplified example. Suppose that we wish to estimate the fourth moment of \( X \) (assuming its existence) using \( \hat{\theta}(X) = n^{-1}(X^4 + \cdots + X^4_n) \). In place of \( X_i \), we observe \( X^*_i = X_i + U_i \), where the \( U_i \) are symmetrically distributed about 0 and independent of \( X_i \) and each other. An application of the weak law of large numbers demonstrates that as \( n \to \infty \), \( \hat{\theta}(\{X^*_i\}_{i=1}^n) \xrightarrow{p} \mu_4 + 6\mu_2\sigma_U^2 + E[U^4] \), where \( \xrightarrow{p} \) refers to convergence in probability, and \( \mu_4 = E[X^4] \) for \( j \in \{2, 4\} \). This is generally biased for \( \mu_4 \) and renders the naive estimator inconsistent.

To apply P-SIMEX, consider \( \hat{\theta}(\{X^*_i(\lambda_p)\}_{i=1}^n) \), which has a limit in probability, as \( n \to \infty \), of \( \mu_4 + 6\mu_2(1 + \lambda_p)\sigma_U^2 + E[U^4] \), with \( \lambda_p \sim \mathcal{N}(0, \lambda_p\sigma_U^2) \). This can be expanded to \( \mu_4 + 6\mu_2(1 + \lambda_p)\sigma_U^2 + 3\lambda_p\sigma_U^4 + 6\lambda_p\sigma_U^4 + E[U^4] \). Since \( E[U^4] \) is functionally independent of \( \lambda_p \), we can take \( G(\lambda) = a + b\lambda + c\lambda^2 \) to be the extrapolant. Then, \( G(-1) = \mu_4 + E[U^4] - 3\sigma_U^4 \). When \( U_{ij} \sim \mathcal{N}(0, \sigma_U^2) \), then \( E[U^4] = 3\sigma_U^4 \) and P-SIMEX consistently corrects for the effects of measurement error. If instead we take \( U_{ij} \sim \mathcal{T}_5 \), then \( E[U^4] = 25 \). Combined with \( \sigma_U^2 = 25/3 \), P-SIMEX leaves a residual asymptotic bias of 50/3.

Consider applying NP-SIMEX with \( \hat{\theta}(\{X_i + U_i(\lambda)\}_{i=1}^n) \), where \( U_i(\lambda) \xrightarrow{d} U_0 + \cdots + U_4 \). Using the same argument as above gives \( \mu_4 + 6\mu_2(1 + \lambda)\sigma_U^2 + (\lambda + 1)E[U^4] + 3(\lambda + 1)\lambda\sigma_U^4 \). This produces an extrapolant that is exactly quadratic, giving \( G'(\lambda) = a' + b'\lambda + c'\lambda^2 \). This leads to the conclusion that \( G'(-1) = \mu_4 \) regardless of the value of \( E[U^4] \). The analytic tractability of this example allows it to serve as an illustration of the conditions required for consistency.

In this example, we treated NP-SIMEX as though the estimator was computed based on random observations that are distributed as \( X + U_0 + \cdots + U_4 \). In practice, we will compute the estimator based on random quantities distributed as \( X + U + U^*_1 + \cdots + U^*_n \), where the \( U^*_i \) are sampled from \( \bar{U} \). We argued above that as \( |U| \to \infty \), this quantity will have characteristic function \( \varphi_X \varphi_U^{*+1} \) and as such will behave as though it were distributed as \( X + U_0 + \cdots + U_4 \). In general, this requires smoothness assumptions on the estimator. Take \( F_\lambda \) to be the distribution function for the sum of \( X \) and \( \lambda + 1 \) copies of \( U \), which is given by the convolution \( F_X * F_U^{\lambda+1} \), and assume that \( m = n \). We establish the following asymptotic results.

**Theorem 1.** Suppose that both \( X \) and \( U \) are absolutely continuous with respect to the Lebesgue measure, that the estimator from the error-free context can be expressed as a functional \( T(\cdot) \) over the distributions, and that \( T(F_\lambda) = G(\lambda) \) for all \( \lambda \geq -1 \), where \( G(\lambda) \) has a known parametric
form. If $T$ is continuous or bounded with respect to $\| \cdot \|_\infty$, then $\hat{\theta}_{\text{NP-SIMEX}}$ is consistent for $\theta$ as $n \to \infty$.

**Theorem 2.** Suppose that both $X$ and $U$ are absolutely continuous with respect to the Lebesgue measure, that the estimator from the error-free context can be expressed as a functional $T(\cdot)$ over the distributions, and that $T(F_\lambda) = \mathcal{G}(\lambda)$ for all $\lambda \geq -1$, where $\mathcal{G}(\lambda)$ has a known parametric form. If $T$ is Fréchet differentiable with respect to $\| \cdot \|_\infty$, then $\sqrt{n}(\hat{\theta}_{\text{NP-SIMEX}} - \theta)$ has an asymptotic normal distribution with mean 0, as $n \to \infty$.

These consistency and asymptotic distribution results are established through nonstandard, technical asymptotic theory. This makes the conditions for the application of these results difficult to assess in practice. These results rely on treating the estimator $\hat{\theta}$ from the error-free context as a functional over distributions and assessing the continuity or differentiability of these functionals. We establish sufficient conditions for consistency that are comparatively straightforward to check, though the condition of Fréchet differentiability of the statistical functional, which is sufficient for asymptotic normality, is fairly strong and difficult to translate into the standard language of estimators. Shao (1993) demonstrates that large classes of commonly used estimators satisfy this condition (for instance, differentiable functions of the mean, large classes of M-estimators, and the Cramér–von Mises test statistic). Other authors have noted that there are many statistical functionals of interest for which Fréchet differentiability with respect to $\| \cdot \|_\infty$ is not satisfied, but where consistency results can still be obtained (Fernholz, 1983).

We expect that, by using more sophisticated arguments, asymptotic normality could be obtained under weaker forms of differentiability. However, it is worth considering the other strong assumption that is being made: that the extrapolant is known and correctly specified. This is an assumption shared by other SIMEX estimators (Cook & Stefanski, 1994; Koul & Song, 2014) but it is quite strong nonetheless. In practice, this assumption is the reason why SIMEX is often treated as an approximately consistent technique for the correction of the effects of measurement error. The use of a suitable extrapolant can reduce (even if it does not entirely eliminate) bias (Cook & Stefanski, 1994; Carroll et al., 2006).

Variance estimation can be conducted through a bootstrap procedure. In certain settings, this may be undesirable because of a need for nested resampling procedures. For P-SIMEX, two additional variance estimation techniques have been proposed: one using a modified SIMEX procedure (Stefanski & Cook, 1995), and the other using the asymptotic distribution (Carroll et al., 1996). Theorem 2 allows for the use of sandwich estimation techniques to establish an estimate of the asymptotic variance. Full details are provided in Carroll et al. (1996). To use sandwich estimation techniques, we require an estimate of the covariance of the stacked influence curves of the functional representation of the estimator. In settings where these representations are common, this result can be useful. For other situations, where both bootstrapping and sandwich estimation techniques are not viable, the techniques proposed by Stefanski & Cook (1995) can be adopted for NP-SIMEX, as outlined in the Appendix.

### 3.4. Extensions of the Core Procedure

The presentation of NP-SIMEX assumed that, when using replicate data to form $U^*$, these replicates were independent and identically distributed. This assumption is stronger than is necessary. First, define a contrast vector $a$, with elements $a_1, \ldots, a_k$, such that $0 = a_1 + \cdots + a_k$ and $1 = |a_1| + \cdots + |a_k|$. Suppose that $U_{ij}$ are symmetric and independent but not necessarily...
identically distributed. For each $j$, $a_j U_{ij} \overset{d}{=} |a_j| U_{ij}$. Then, to form elements of the set $U^*$, we take

$$U^*_i = \sum_{j=1}^{\kappa} a_j X^*_i = \sum_{j=1}^{\kappa} a_j U_{ij} = \sum_{j=1}^{\kappa} |a_j| U_{ij}.$$ 

Here, $U^*_i$ represents a realization from the empirical error distribution for

$$X^*_i = \sum_{j=1}^{\kappa} |a_j| X^*_i = X_i + \sum_{j=1}^{\kappa} |a_j| U_{ij}.$$ 

We can slightly relax the assumption of symmetric errors with replicate data. Assuming that $U_{ij}$ is symmetric, for any $a_j$, we have $a_j U_{ij} \overset{d}{=} |a_j| U_{ij}$. When $a_j \geq 0$, then $a_j U_{ij} = |a_j| U_{ij}$. Supposing that $U_{ij}$ does not follow a symmetric distribution, then the previous argument still holds if $a_j \geq 0$ for all $j$. As a result, so long as at least one of the repeated measurements has a symmetric error distribution, NP-SIMEX can proceed by defining the contrast vector $a$ as before, with the restriction that each nonsymmetric entry has a positive value.

Finally, we have made the common assumption that $U_{ij} \perp X_i$, for all $i$ and $j$. It may be the case that errors depend on the true, underlying value, which would render the presented argument invalid for NP-SIMEX. If the $U_{ij}s$ are dependent on $X_i$, the errors associated with individual $i$ will be drawn from a different distribution than those from individual $i' \neq i$. Conceptually, we can replace the empirical distribution with a distribution estimated using kernel density estimation (KDE). Many techniques have been proposed for estimating a conditional density based on kernel methods (Hall, Racine & Li, 2004). With an estimated conditional KDE, denoted $\hat{f}_{U|X}$, it is possible to sample directly from this conditional distribution (see for instance, Shalizi (2022, Section 14.7)). The details of this technique are expanded upon in the Appendix.

While this procedure conceptually works, the difficulty is that we cannot directly condition on $X_i$ outside of the validation sample. Instead, we need a method for drawing from the correct error distribution given only $X_i^*$. A possible technique is to repeat this procedure using the validation sample to estimate $\hat{f}_{X|X^*}(x|x^*)$ and then draw $\tilde{X}_i$ for each individual in the sample via this KDE. This could then be used as the value of $X_i$ to condition on. An alternative approach is to draw error realizations directly from the distribution of $\hat{f}_{U|X^*}(u|x^*)$. This procedure can be directly applied over the complete sample. Despite the easier application, this procedure only approximately corrects for dependence in the errors, even in the limit, as generally conditioning on $X^*$ induces dependence between $X$ and $U$ even if none previously existed.

A discussion of using KDEs in the event that $U_i$ and $X_i$ are dependent may also prompt consideration of using KDEs under the assumption that $U_i \perp X_i$. Instead of forming $U^*$ directly, we can estimate $\hat{f}_U(u)$ and sample from this KDE. This procedure is the smoothed bootstrap (de Angelis & Young, 1992). In certain settings, smoothing can improve the performance of estimators, particularly with small sample sizes (Efron, 1982). This smoothing could be applied, under the independence assumption, with either validation or repeated measurements. We refer to this as smoothed NP-SIMEX.

### 4. SIMULATION STUDIES

In this section we present four simulation studies investigating the behaviour of the estimator in several scenarios. Additional simulation results are provided in the Appendix. The first simulation contrasts P-SIMEX and NP-SIMEX in logistic regression. We take $n = 5000$, $B = 100$, and a $\Lambda$-grid size of 10. We generate a true, unobserved variable $X$ from a $\mathcal{N}(1, 2)$ distribution...
and consider an outcome \( Y \) such that \( P(Y = 1|X) = H(1 - X) \), where \( H(\cdot) \) is the inverse-logit function. In place of \( X \), we generate two replicated responses for each individual, \( X_1 \) and \( X_2 \), which are given by \( X_j = X + U_j \), for \( j \in \{1, 2\} \) where \( U_j \) follows a \( t \)-distribution with degrees of freedom in \( \{3, 4, 5, 10, 30\} \) and is independent of all other variables. Both P-SIMEX and NP-SIMEX are implemented using the nonlinear extrapolant \( G(\lambda) = a + b/(c + \lambda) \). We compute 95% confidence intervals using a bias-adjusted bootstrap procedure with 500 bootstrap replicates. These simulations are repeated 200 times. Results are shown in Table 1.

Across all \( t \)-distributions, NP-SIMEX dramatically outperforms P-SIMEX in MSE and bias. The computed coverage probabilities are also substantially improved, though there is evidence of undercoverage, particularly for low degrees of freedom. While none of these differences is significant at the 95% level, these anticonservative results warrant caution and careful application of the bootstrap procedure, specifically when the error distribution is likely to be particularly heavy-tailed. The results suggest that bootstrapping may be feasible for quantifying uncertainty in NP-SIMEX when computation is not a problem.

The second simulation investigates the impact of sample size on estimate variability. We use the example from Section 3.3, which involves estimating the fourth moment of \( X \) from a \( \mathcal{N}(5,4) \) distribution. We take two error-prone measurements, both subject to additive error from a \( t_5 \) distribution. The errors are independent of each other and of \( X \). We vary the sample size from 100 to 100,000, and replicate each setting 1000 times. We take \( M = 10 \) and \( B = 500 \). Results are shown in Table 2.

Predictably, the naive method performs unsatisfactorily and demonstrates the utility of both P-SIMEX and NP-SIMEX in reducing the impact of measurement error. While MSE is quite large for small \( n \) regardless of the method, this is also true for the true estimator, which sees only an 8.3% and 11.3% increase in the relative MSEs over the truth for NP-SIMEX and P-SIMEX, respectively (when \( n = 100 \)). While raw MSE decreases for both correction procedures as \( n \) increases, the relative MSE increases for both. However, NP-SIMEX remains more or less comparable to the truth for all values of \( n \), while for larger values of \( n \), P-SIMEX performs substantially worse. P-SIMEX and the naive estimator are left with substantial bias even for large sample sizes, whereas NP-SIMEX appears to mostly eliminate it. There is no substantial difference between the smoothed and unsmoothed estimators.

Our third simulation considers the use of validation data in place of replicate measurements. We generate the true variate \( X \) from a gamma distribution with shape parameter 1 and scale parameter 2 so that \( E(X) = 2 \) and \( \text{var}(X) = 4 \). We then contaminate \( X \) with an additive error term \( U_i \) that has mean 0 and follows a Laplace distribution. We consider several values for

| Table 1: Mean squared error (MSE), mean bias, and coverage probability from 200 replicate simulations estimating the slope parameter in a logistic regression model where the variate has \( t \)-distributed error with varying degrees of freedom (DFs). |
|---|---|---|---|---|---|---|
| DFs | P-SIMEX | NP-SIMEX |
| MSE | Mean bias | Coverage | MSE | Mean bias | Coverage |
| 3 | 0.014 | -0.108 | 0.200 | 0.002 | -0.005 | 0.920 |
| 4 | 0.014 | -0.111 | 0.165 | 0.001 | 0.002 | 0.930 |
| 5 | 0.015 | -0.116 | 0.110 | 0.001 | 0.005 | 0.930 |
| 10 | 0.016 | -0.119 | 0.095 | 0.001 | 0.002 | 0.940 |
| 30 | 0.016 | -0.120 | 0.100 | 0.001 | 0.001 | 0.950 |

Note: Coverage probability is computed using a bias-corrected bootstrap with 500 bootstrap resamples.

The Canadian Journal of Statistics / La revue canadienne de statistique

DOI: 10.1002/cjs.11777
Table 2: Relative MSE (and mean bias) from 1000 replicated simulations estimating the fourth moment of a contaminated random variable for different sample sizes (n).

| n       | Naive      | P-SIMEX    | NP-SIMEX   | Smoothed NP | Truth         |
|---------|------------|------------|------------|-------------|---------------|
| 100     | 2.279 (148.305) | 1.682 (9.600) | 1.273 (−12.185) | 1.299 (−23.709) | 34647.104 (−6.013) |
| 500     | 5.048 (151.618) | 1.771 (34.205) | 1.386 (3.955) | 1.431 (−2.972) | 6660.699 (2.029)  |
| 1000    | 8.275 (148.976) | 1.759 (35.874) | 1.383 (−1.329) | 1.390 (−6.045) | 3350.250 (0.672)  |
| 5000    | 34.404 (147.637) | 3.561 (36.599) | 1.456 (−1.495) | 1.467 (−4.127) | 0664.606 (−1.443) |
| 10,000  | 70.330 (149.476) | 5.259 (35.557) | 1.364 (0.751) | 1.349 (−1.344) | 324.617 (0.483)   |
| 20,000  | 133.467 (149.920) | 8.622 (34.759) | 1.464 (0.068) | 1.484 (−1.383) | 170.587 (0.134)   |
| 50,000  | 350.646 (149.509) | 21.772 (35.821) | 1.446 (0.491) | 1.540 (−0.053) | 64.119 (0.320)    |
| 100,000 | 683.371 (149.037) | 39.205 (35.045) | 1.547 (−0.164) | 1.564 (−0.442) | 32.579 (0.032)    |

Note: The main values are the MSE divided by the MSE computed using the error-free variable (Truth) for the same sample size; bracketed values are the mean biases.

the measurement error variance by taking the ratio \( \sigma_U/\sigma_X \) from \{0.1, 0.5, 1, 2\}. The sample size is selected from \{1000, 10,000, 100,000\}, and we assume that an internal validation sample is available comprised of \{5%, 10%, 50%\} of the total sample. All results use the nonlinear extrapolant for both P-SIMEX and NP-SIMEX. The MSEs for the naive, P-SIMEX, and NP-SIMEX estimators across all scenarios are presented in Tables 3 and 4.

In Table 3, we see that NP-SIMEX outperforms both P-SIMEX and naive estimation, particularly when the ratio of variances grows. With a small validation sample, and with small measurement error, we see that P-SIMEX performs at the same level as NP-SIMEX. However, as the estimators stabilize, by increasing either \( n \) or the proportion of validation samples, NP-SIMEX correction substantially outperforms both the other methods. This table excludes results with a ratio of standard deviations equal to 1 when the validation sample size is 100 or fewer and the results where the ratio of standard deviations was 2 for validation samples up to 1000 individuals. The results of these scenarios are provided in Table 4.

The results in Table 4 demonstrate the instability of the nonparametric procedure for small sample sizes and when the error is sufficiently large. Note that, as would be expected, the naive estimators are not impacted by the size of the validation sample, and the impact on P-SIMEX is fairly small. The results show that if the validation sample is too small or the estimated variation is too large, nonparametric techniques are not appropriate. This emphasizes the importance of considering the fact that, when using validation data, convergence of the correction happens in \( n_1 \) rather than \( n \). Fortunately, while these results demonstrate clear instability for small sample sizes, the breakdown in performance is easy to see. The techniques, when unstable, will often result in estimates that are unreasonable from a subject matter perspective. We stress careful application of these techniques in settings where sample sizes may lead to instability.

In the final simulation, we demonstrate the viability of the NP-SIMEX strategy using KDE. We take \( X \sim \mathcal{N}(1, 4) \) and \( U|X \sim \mathcal{N}(\rho(X - 1), 1) \), where \( \rho \) is a parameter selected from \{0, 0.5, 1, 2\}, and a binary outcome \( Y \) with \( P(Y = 1|X) = H(1 - X) \). The sample size is \( n = 1000 \) with a 20% validation sample. The simulations are repeated 500 times. Within this context, we compare four estimation strategies: first, a standard application of SIMEX; second, a version of NP-SIMEX where we first sample \( X \) from \( X|X^* \) and then from \( U|X \); third, a version of NP-SIMEX where we sample directly from \( U|X^* \); and finally, sampling \( X|X^* \) directly and averaging over many iterations. Results are presented in Table 5.

DOI: 10.1002/cjs.11777
TABLE 3: MSE from 1000 replicated simulations estimating the slope parameter in a logistic regression model for different sample sizes (n), validation sample size percentages (%), and ratios of standard deviations (\( \sigma_U/\sigma_X \)).

| \( \sigma_U/\sigma_X \) = 0.1 | \( \sigma_U/\sigma_X \) = 0.5 | \( \sigma_U/\sigma_X \) = 1 | \( \sigma_U/\sigma_X \) = 2 |
|---|---|---|---|
| % | Naive | P | NP | Naive | P | NP | Naive | P | NP |
| n = 1000 | | | | | | | | | |
| 5 | 0.011 | 0.111 | 0.111 | 0.225 | 0.114 | 0.124 | – | – | – |
| 10 | 0.011 | 0.111 | 0.112 | 0.220 | 0.108 | 0.039 | – | – | – |
| 50 | 0.012 | 0.112 | 0.113 | 0.221 | 0.108 | 0.024 | 0.726 | 0.264 | 0.086 | – | – | – |
| n = 10,000 | | | | | | | | | |
| 5 | 0.002 | 0.002 | 0.001 | 0.219 | 0.103 | 0.009 | 0.726 | 0.259 | 0.071 | – | – | – |
| 10 | 0.002 | 0.002 | 0.001 | 0.220 | 0.103 | 0.007 | 0.726 | 0.257 | 0.048 | – | – | – |
| 50 | 0.002 | 0.002 | 0.001 | 0.219 | 0.102 | 0.006 | 0.724 | 0.254 | 0.035 | 10.230 | 0.856 | 0.092 |
| n = 100,000 | | | | | | | | | |
| 5 | 0.001 | 0.001 | 0.000 | 0.219 | 0.102 | 0.004 | 0.725 | 0.254 | 0.033 | 10.230 | 0.858 | 0.100 |
| 10 | 0.001 | 0.001 | 0.000 | 0.219 | 0.102 | 0.004 | 0.725 | 0.254 | 0.032 | 10.230 | 0.857 | 0.076 |
| 50 | 0.001 | 0.001 | 0.000 | 0.219 | 0.102 | 0.004 | 0.725 | 0.254 | 0.032 | 10.230 | 0.857 | 0.068 |

Note: The results compare the naive estimators, those from P-SIMEX (P), and those from NP-SIMEX (NP). This table contains results with a validation sample that is sufficiently large relative to the measurement error variance.

TABLE 4: MSE from 1000 replicated simulations estimating the slope parameter in a logistic regression model for different sample sizes (n), validation sample size percentages (%), and ratios of standard deviations (\( \sigma_U/\sigma_X \)).

| \( \sigma_U/\sigma_X \) = 1 | \( \sigma_U/\sigma_X \) = 2 |
|---|---|
| n (%) | Naive | P | NP | Naive | P | NP |
| 1000 (5) | 0.728 | 0.295 | 9917.339 | 1.230 | 0.872 | 52.925 |
| 1000 (10) | 0.724 | 0.273 | 100.555 | 1.228 | 0.860 | 190.287 |
| 1000 (50) | – | – | – | 1.231 | 0.863 | 190.660 |
| 10,000 (5) | – | – | – | 1.230 | 0.859 | 124.587 |
| 10,000 (10) | – | – | – | 1.231 | 0.859 | 58.118 |

Note: The results compare the naive estimators, those from P-SIMEX (P), and those from NP-SIMEX (NP). This table contains results with a validation sample that is small relative to the measurement error variance.

When there is independence between \( U \) and \( X \), the standard SIMEX estimators perform well. As this dependence strengthens, the corrections are less able to address measurement error. Using \( U|X \) for the KDE NP-SIMEX provides comparable performance across all the scenarios tested and is always among the best techniques. Using \( U|X^* \) for the KDE NP-SIMEX performs comparatively poorly when there is independence, but with dependent errors this approach sees a marked improvement and performs better than any other technique. The averaging of samples...
Table 5: MSE of logistic regression slope parameter estimates from 500 simulation runs where the strength of the simulated dependence between the true variate \((X)\) and the error term \((U)\) is mediated by \(\rho\).

| \(\rho\) | SIMEX | NP-SIMEX \((U|X)\) | NP-SIMEX \((U|X^\ast)\) | Direct from \(X|X^\ast\) |
|---------|-------|----------------|---------------------|-----------------|
| 0       | 0.012 | 0.014          | 0.053               | 0.183           |
| 0.5     | 0.067 | 0.010          | 0.012               | 0.070           |
| 1       | 0.115 | 0.013          | 0.009               | 0.034           |
| 2       | 0.222 | 0.018          | 0.009               | 0.011           |

directly from \(X|X^\ast\) tends to perform fairly well in settings with a strong dependence, although it seems unlikely to be preferable to NP-SIMEX using \(U|X^\ast\) or \(U|X\).

5. DATA ANALYSIS

We now consider an analysis of the Korean Longitudinal Study of Aging (KLoSA), following that by Xu, Kim & Li (2017). The KLoSA considers South Korean citizens aged 45 and over in a longitudinal study looking to determine health effects of aging. Our analysis considers data on \(n = 9842\) individuals, with an internal validation sample of size \(n_1 = 505\). We are interested in estimating how an individual’s body mass index (BMI) impacts their propensity towards being hypertensive. In the main study, BMI is estimated through self-reported weights and heights, while the validation sample includes clinical measurements, taken to be the truth, alongside the self-reported values. We also observe each individual’s age, which we consider to be error-free.

An analysis of the validation sample demonstrates that the errors are non-normal (Figure 1), have a negative skew, and have an excess kurtosis of 2.05. This suggests that the standard P-SIMEX procedure might not be appropriate. We estimate the ratio \(\sigma_U/\sigma_X\) using the 505 validation sample observations as 0.898.

We analyze these data by fitting a logit-link logistic regression model, which includes the main effects of BMI and age. That is, we assume that

\[
\logit(E[Y_i|X_i, Z_i]) = \beta_0 + \beta_1 \text{BMI}_i + \beta_2 \text{Age}_i.
\]

We generate bootstrap estimates of standard error with 1000 replicates and compare both NP-SIMEX and P-SIMEX with a nonlinear extrapolant. We also consider an uncorrected analysis. The results are summarized in Table 6.

The three methods tend to agree on the estimate and standard error for \(\beta_2\). For both \(\beta_0\) and \(\beta_1\), we see that NP-SIMEX yields estimates and standard errors that are larger in magnitude than the naive estimator but smaller than the P-SIMEX estimator. All three techniques suggest a positive effect of BMI on hypertension, although the level of significance of this effect varies dramatically: 0.054, 0.117, and 0.133 for the naive, P-SIMEX, and NP-SIMEX estimators, respectively.

One concern with this analysis of the KLoSA is that there is evidence that the observed errors are not independent of the true values. In Figure 2, the plot of the error terms versus the true values illustrates the degree of dependence present in these data. This relationship corresponds to a correlation of approximately \(-0.464\). This pattern has also been observed in past research (Villanueva, 2001).
Figure 1: A normal Q–Q plot for the observed errors in the Korean Longitudinal Study of Aging.

Table 6: Point estimates and bootstrap standard error (SE) estimates for the logistic regression parameters when estimating the propensity of hypertension for the intercept ($\beta_0$), self-reported BMI ($\beta_1$), and age ($\beta_2$).

| Method    | $\beta_0$ Estimate | $\beta_0$ SE  | $\beta_1$ Estimate | $\beta_1$ SE  | $\beta_2$ Estimate | $\beta_2$ SE  |
|-----------|--------------------|---------------|--------------------|---------------|--------------------|---------------|
| Naive     | $-5.023$           | $0.439$       | $0.030$            | $0.016$       | $0.053$            | $0.002$       |
| P-SIMEX   | $-5.512$           | $0.833$       | $0.049$            | $0.031$       | $0.054$            | $0.003$       |
| NP-SIMEX  | $-5.061$           | $0.673$       | $0.039$            | $0.026$       | $0.054$            | $0.002$       |

Note: The estimates are based on 1000 bootstrap replicates and compare the naive method, P-SIMEX correction, and NP-SIMEX correction.

We consider conducting the same analysis using the two proposed KDE NP-SIMEX estimation techniques, based on both sampling first from $X|X^*$ and then $U|X$, and on sampling directly from $U|X^*$. The estimated coefficients and bootstrap standard errors are included in Table 7. For each of the analyses, we consider using both the quadratic and nonlinear extrapolants. The slope coefficient for age ($\beta_2$) generally was not estimable with the nonlinear extrapolant and so these results are not reported.

The resulting estimates for $\beta_0$ and $\beta_2$ do not differ substantially from the nonconditional results. The signs for these coefficients and their approximate magnitudes are comparable to the previously estimated values. The largest difference is in the estimates for $\beta_1$ when the nonlinear extrapolant was used. These results suggest that the magnitude of the effect size was severely underestimated. In comparing the use of the nonlinear extrapolant with either of the conditional distributions to the previous analyses assuming independence in the errors, we find that the
FIGURE 2: Estimated errors (U) versus the true underlying BMI for individuals within the KLoSA validation sample. The included line is a LOESS curve that is included to clearly delineate the degree of dependence observed within these data.

TABLE 7: Point estimates and bootstrap standard error (SE) estimates for the logistic regression parameters when estimating the propensity of hypertension for the intercept (β₀), self-reported BMI (β₁), and age (β₂).

| Method | Extrapolant | β₀     | SE   | β₁     | SE   | β₂     | SE   |
|--------|-------------|--------|------|--------|------|--------|------|
| U|X| Quadratic | −5.531 | 0.472 | 0.049 | 0.017 | 0.054 | 0.002 |
| U|X| Nonlinear | −4.740 | 0.669 | 0.063 | 0.028 | −      | −    |
| U|X∗| Quadratic | −5.515 | 0.473 | 0.049 | 0.017 | 0.054 | 0.002 |
| U|X∗| Nonlinear | −4.936 | 0.679 | 0.061 | 0.038 | −      | −    |

Note: The estimates are based on 500 bootstrap replicates and compare the conditional NP-SIMEX method using U|X and the conditional NP-SIMEX method using U|X∗, both with either a quadratic or nonlinear extrapolant.

previous estimates had magnitudes that were between 0.5 and 0.8 times the estimated magnitude using the conditional distribution. The P-values for a test of significance using U|X and U|X∗ were, respectively, 0.022 and 0.105.

Given the clear dependence observed between the errors and the true BMI in these data, and in past literature, we advise taking the conditional analyses as more reliable for estimators of the truth than the unconditional analyses presented originally. Agreement with respect to the intercept and age coefficient gives confidence in these estimates.
6. DISCUSSION
Measurement error is a ubiquitous issue that impacts the validity of statistical inference. Methods for correcting for measurement error often rely on assumptions regarding the distribution of the error terms. The SIMEX procedure is a commonly used correction procedure that can correct for the effects of measurement error in a wide variety of models (Carroll et al., 2006; Yi, 2017). Where the corrections are not consistent, SIMEX can reduce the bias present in a naive analysis. While P-SIMEX is easy to implement, it relies on an assumption that the errors are normally distributed (Cook & Stefanski, 1994; Stefanski & Cook, 1995; Carroll et al., 1996). We provide a nonparametric SIMEX procedure that relaxes this assumption and allows for any error distribution if validation data are available or allows for any symmetric error distribution when replicate measurements are available. The implementation of our procedure is similar to that of P-SIMEX, which allows for analysts familiar with P-SIMEX to adopt this nonparametric extension where appropriate. Our simulation results demonstrate that when there are sufficient data and the models used are conducive to SIMEX, the proposed procedure effectively corrects for the effects of measurement error across a wide variety of measurement-error settings. These results are supported by large-sample, theoretical justifications.

Our results complement those of Koul & Song (2014), who accommodate non-normal errors through parametric extensions to SIMEX. These parametric extensions carry the standard benefits of parametric methods. We would expect them to be more efficient, particularly for small sample sizes, when the assumed error distribution is correct. In contrast, NP-SIMEX provides a flexible way to accommodate errors without the need to specify a particular distribution for the error terms. This is an important extension in settings that might not have strong domain-specific justifications for particular distributional assumptions.

While nonparametric corrections for the effects of measurement error are important given the prevalence of non-normal error distributions, there are drawbacks to such an approach. Primarily, nonparametric methods tend to require substantially more data to behave in a stable manner, and our correction is no exception.

All the code used to perform the simulations and analyses in this article are available publicly on GitHub at https://github.com/DylanSpicker/np-simex.

ACKNOWLEDGEMENTS
This work was partially supported by funding from the Natural Sciences and Engineering Council of Canada (NSERC). Yi is a Canada Research Chair in Data Science (Tier 1). Her research was undertaken in part thanks to funding from the Canada Research Chairs Program. The authors thank the anonymous reviewers for their helpful comments.

REFERENCES
Augustin, T. (2004). An exact corrected log-likelihood function for Cox’s proportional hazards model under measurement error and some extensions. Scandinavian Journal of Statistics, 31(1), 43–50.
Bailey, D. C. (2017). Not normal: The uncertainties of scientific measurements. Royal Society Open Science, 4(1), 160600.
Bollinger, C. R. (1998). Measurement error in the current population survey: A nonparametric look. Journal of Labor Economics, 16(3), 576–594.
Carroll, R., Ruppert, D., Stefanski, L., & Crainiceanu, C. (2006). Measurement Error in Nonlinear Models: A Modern Perspective, II, Chapman and Hall/CRC, Boca Raton, FL.
Carroll, R. J., Küchenhoff, H., Lombard, F., & Stefanski, L. A. (1996). Asymptotics for the SIMEX estimator in nonlinear measurement error models. Journal of the American Statistical Association, 91(433), 242–250.
Cook, J. R. & Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428), 1314–1328.

de Angelis, D. & Young, G. A. (1992). Smoothing the bootstrap. *International Statistical Review*, 60(1), 45–56.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*, Society for Industrial and Applied Mathematics, Philadelphia. ISBN: 978-0-89871-179-0

Fernholz, L. T. (1983). *von Mises Calculus for Statistical Functionals*, Springer, New York.

Gorfine, M., Hsu, L., & Prentice, R. L. (2004). Nonparametric correction for covariate measurement error in a stratified Cox model. *Biostatistics*, 5(1), 75–87.

Hall, P., Racine, J., & Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468), 1015–1026.

Koul, H. L. & Song, W. (2014). Simulation extrapolation estimation in parametric models with Laplace measurement error. *Electronic Journal of Statistics*, 8(2), 1973–1995.

Li, T. & Vuong, Q. (1998). Nonparametric estimation of the measurement error model using multiple indicators. *Journal of Multivariate Analysis*, 65(2), 139–165.

McKenzie, H. W., Jerde, C. L., Visscher, D. R., Merrill, E. H., & Lewis, M. A. (2008). Inferring linear feature use in the presence of GPS measurement error. *Environmental and Ecological Statistics*, 16(4), 531–546.

National Research Council (1986). *Nutrient Adequacy*, National Academies Press, Washington, DC. DOI: 10.17226/618

Novick, S. J. & Stefanski, L. A. (2002). Corrected score estimation via complex variable simulation extrapolation. *Journal of the American Statistical Association*, 97(458), 472–481.

Nusser, S. M., Carriquiry, A. L., Dodd, K. W., & Fuller, W. A. (1996). A semiparametric transformation approach to estimating usual daily intake distributions. *Journal of the American Statistical Association*, 91(436), 1440–1449.

Purdom, E. & Holmes, S. P. (2005). Error distribution for gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 16.

Rajan, A. & Desai, S. (2018). Non-Gaussian error distributions of galactic rotation speed measurements. *The European Physical Journal Plus*, 133(3), 107.

Schennach, S. M. & Hu, Y. (2013). Nonparametric identification and semiparametric estimation of classical measurement error models without side information. *Journal of the American Statistical Association*, 108(501), 177–186.

Shalizi, C. R. (2022). *Advanced Data Analysis from an Elementary Point of View*, Cambridge University Press, Cambridge.

Shao, J. (1993). Differentiability of statistical functionals and consistency of the jackknife. *The Annals of Statistics*, 21(1), 61–75.

Shorack, G. R. (1979). The weighted empirical process of row independent random variables with arbitrary distribution functions. *Statistica Neerlandica*, 33(4), 169–189.

Stefanski, L. A. & Cook, J. R. (1995). Simulation-extrapolation: The measurement error jackknife. *Journal of the American Statistical Association*, 90(432), 1247–1256.

Villanueva, E. V. (2001). The validity of self-reported weight in US adults: A population based cross-sectional study. *BMC Public Health. Public Health*, 1(1), 11.

Xu, Y., Kim, J. K., & Li, Y. (2017). Semiparametric estimation for measurement error models with validation data. *The Canadian Journal of Statistics*, 45(2), 185–201.

Yi, G. Y. (2017). *Statistical Analysis with Measurement Error or Misclassification*, Springer, New York.

Yi, G. Y. & He, W. (2012). Bias analysis and the simulation-extrapolation method for survival data with covariate measurement error under parametric proportional odds models. *Biometrical Journal*, 54(3), 343–360.

DOI: 10.1002/cjs.11777
APPENDIX

Empirical Error Distribution with $\kappa \neq 2$ Replicates

In the main article, we presented a technique for forming $U^*$ when there are $\kappa = 2$ replicates. If $\kappa \neq 2$, a similar process can be followed. Define a $\kappa$-dimensional contrast $(a_1, \ldots, a_\kappa)$ such that $0 = a_1 + \cdots + a_\kappa$ and $1 = |a_1| + \cdots + |a_\kappa|$. Consider the sums

$$\sum_{j=1}^\kappa a_j X_{i,j}^* \quad \text{and} \quad \sum_{j=1}^\kappa |a_j| X_{i,j}^*.$$ 

Since $a_1 + \cdots + a_\kappa = 0$, the former satisfies

$$\sum_{j=1}^\kappa a_j U_{i,j} = \sum_{j=1}^\kappa |a_j| U_{i,j},$$

where the distributional equality holds since the $U_{i,j}$ are symmetric by assumption. For the latter sum, we note that since $|a_1| + \cdots + |a_\kappa| = 1$, the sum simplifies to $X_i^* + |a_1| U_{i1} + \cdots + |a_\kappa| U_{i\kappa}$. As a result, we can take $X_i^* = |a_1| X_{i1}^* + \cdots + |a_\kappa| X_{i\kappa}^*$ and $U_i = a_1 X_{i1}^* + \cdots + a_\kappa X_{i\kappa}^*$, and apply the same argument as above. In the case of $\kappa = 2$, we have used the contrast $(1/2, -1/2)^T$. This is naturally extended to

$$\left(\begin{array}{c} 1/\kappa, \ldots, 1/\kappa, -1/\kappa, \ldots, -1/\kappa \\ \kappa/2 \text{ terms} & \kappa/2 \text{ terms} \end{array}\right)^T,$$

when $\kappa$ is even. When $\kappa$ is odd, we use

$$\left(\begin{array}{c} 1/(\kappa + 1), \ldots, 1/(\kappa + 1), -1/(\kappa - 1), \ldots, -1/(\kappa - 1) \\ (\kappa+1)/2 \text{ terms} & (\kappa-1)/2 \text{ terms} \end{array}\right)^T.$$

Proof of Theorems

In order to prove Theorems 1 and 2, we first present several lemmas. Recall that $\hat{F}_{n,m,\lambda}$ represents the empirical cumulative distribution function (CDF) of $X_i^*(\lambda)$, $F_{m,\lambda} = F_X * F_U * \hat{F}_U^\lambda$ when $|U| = m$, and $F_j = F_X * F_{U_j}^{\lambda+1}$. Moreover, we take $m = n$.

Lemma A1. Suppose that $X$ is a random variable that is absolutely continuous with respect to the Lebesgue measure, with density $f$ and distribution function $F$. Take $G$ to be a distribution function, and $G_n$ to be a sequence of distribution functions. Then

1. $\|F * G_n - F * G\|_\infty = \|F * (G_n - G)\|_\infty \leq \|G_n - G\|_\infty$ and
2. if $\|G_n - G\| \xrightarrow{a.s.} 0$ as $n \to \infty$, then $\|F * (G_n - G)\| \xrightarrow{a.s.} 0$ as $n \to \infty$.

Proof of Lemma A1. Note that this is essentially a specific case of Young’s Convolution Inequality. Also note that the second claim follows immediately from the first, so we prove the inequality. Then

$$\|F * (G_n - G)\|_\infty = \sup_{x \in \mathbb{R}} \left| \int_{-\infty}^{\infty} f(t) \left[ G_n(x-t) - G(x-t) \right] \, dt \right|$$

DOI: 10.1002/cjs.11777
\[
\leq \sup_{x \in \mathbb{R}} \left[ \int_{-\infty}^{\infty} f(t) \left| G_n(x - t) - G(x - t) \right| dt \right] \leq \|f\|_p \|G_n - G\|_q,
\]
where \(1/p + 1/q = 1\) for \(p, q \in [1, \infty]\). The second inequality is simply Hölder’s Inequality. Take \(q = \infty\) so that \(p = 1\).

**Lemma 1.** Suppose that both \(X\) and \(U\) are absolutely continuous with respect to the Lebesgue measure. Then for every \(\lambda \in \Lambda\), as \(n \to \infty\),

\[
\sup_{x \in \mathbb{R}} |\hat{F}_{n,m,\lambda}(x) - F_{\lambda}(x)| = \|\hat{F}_{n,m,\lambda} - F_{\lambda}\|_{\infty} \xrightarrow{a.s.} 0.
\]

**Proof of Lemma 1.** First note that, by the triangle inequality,

\[
\|\hat{F}_{n,m,\lambda} - F_{\lambda}\|_{\infty} \leq \|\hat{F}_{n,m,\lambda} - F_{m,\lambda}\|_{\infty} + \|F_{m,\lambda} - F_{\lambda}\|_{\infty}.
\]

Appealing to Shorack (1979, Theorem 2.1), we get as \(n \to \infty\),

\[
\|\hat{F}_{n,m,\lambda} - F_{m,\lambda}\|_{\infty} \xrightarrow{a.s.} 0
\]

since \(m = n\), and for every fixed \(m\), the distribution of the realizations is given by \(F_{m,\lambda}\).

The second term converges almost surely to zero as well, which we show inductively in \(\lambda\). Take \(\lambda = 0\). Then \(F_{m,\lambda} = F_{\lambda} = F_X * F_U\), and so \(\|F_{m,\lambda} - F_{\lambda}\|_{\infty} = 0\). Consider \(\lambda = 1\). Then \(F_{m,\lambda} = F_{\lambda} = F_X * F_U * \hat{F}_U\) and \(F_{\lambda} = F_X * F_U * F_U\), and we can apply Lemma A1 with \(F = F_X * F_U, G_n = \hat{F}_U\), and \(G = F_U\). Assume that for a positive integer \(\lambda'\), this result holds for all \(\lambda = 2, 3, \ldots, \lambda' - 1\). Note that we can write \(\hat{F}_{n,m,\lambda} = \hat{F}_U * \hat{F}_{n,m,\lambda-1}\) and \(F_{\lambda} = F_{\lambda-1} * F_U\). From these identities and the triangle inequality, with \(\lambda = \lambda'\) we get that

\[
\|\hat{F}_{n,m,\lambda'} - F_{\lambda'}\|_{\infty} \leq \|\hat{F}_U * \hat{F}_{n,m,\lambda'-1} - \hat{F}_U * F_{\lambda'-1}\|_{\infty} + \|F_{\lambda'-1} * \hat{F}_U - F_{\lambda'-1} * F_U\|_{\infty}.
\]

The first term converges almost surely to 0 by the inductive hypothesis and Lemma A1. The second term converges almost surely to 0 by the standard Glivenko–Cantelli Theorem and Lemma A1.

**Lemma 2(i).** Suppose that the conditions of Lemma 1 are satisfied and that the estimator in the error-free context can be expressed as a functional \(T(\cdot)\) over distributions and is continuous with respect to \(\|\cdot\|_{\infty}\) at \(F_{\lambda}\) for all \(\lambda \in \Lambda\). Then \(\hat{\theta} = T(\hat{F}_{n,m,\lambda})\) is consistent for \(T(F_{\lambda})\) as \(n \to \infty\).

**Proof of Lemma 2(i).** By Lemma 1 we have that \(\|\hat{F}_{m,n,\lambda} - F_{\lambda}\|_{\infty} \xrightarrow{a.s.} 0\). Continuity of the functional here means that, for all \(\epsilon > 0\), there exists \(\delta > 0\) such that \(\|\hat{F}_{n,m,\lambda} - F_{\lambda}\|_{\infty} < \delta \Rightarrow |T(\hat{F}_{n,m,\lambda}) - T(F_{\lambda})| \leq \epsilon\). As a result, for \(\epsilon \geq 0\), we have

\[
P \left[ \|\hat{F}_{n,m,\lambda} - F_{\lambda}\|_{\infty} < \delta \right] \leq P \left[ |T(\hat{F}_{n,m,\lambda}) - T(F_{\lambda})| \leq \epsilon \right].
\]
and so
\[
\lim_{n \to \infty} P \left[ \| \hat{F}_{n,m,\lambda} - F_{\lambda} \|_\infty < \delta \right] \leq \lim_{n \to \infty} P \left[ | T(\hat{F}_{n,m,\lambda}) - T(F_{\lambda}) | \leq \varepsilon \right] \leq 1
\]
and
\[
1 \leq \lim_{n \to \infty} P \left[ | T(\hat{F}_{n,m,\lambda}) - T(F_{\lambda}) | \leq \varepsilon \right] \leq 1,
\]
where the last line follows since \( \| \hat{F}_{m,n,\lambda} - F_{\lambda} \|_\infty \xrightarrow{a.s.} 0 \).

**Lemma 2(ii).** Suppose that the conditions of Lemma 1 are satisfied and that the estimator from the error-free context can be expressed as a functional \( T(\cdot) \) over distributions and is bounded with respect to \( \| \cdot \|_\infty \). Then \( \hat{\theta}_{\text{NP-SIMEX}} = T(\hat{F}_{n,m,\lambda}) \) is consistent for \( T(F_{\lambda}) \) as \( n \to \infty \).

**Proof of Lemma 2(ii).** By boundedness we mean that, for a constant \( C \),
\[
| T(\hat{F}_{n,m,\lambda}) - T(F_{\lambda}) | \leq C \| \hat{F}_{n,m,\lambda} - F_{\lambda} \|_\infty.
\]
As a result, for \( \delta > 0 \),
\[
P \left[ | T(\hat{F}_{n,m,\lambda}) - T(F_{\lambda}) | < \varepsilon \right] \geq P \left[ \| \hat{F}_{n,m,\lambda} - F_{\lambda} \|_\infty < \frac{\varepsilon}{C} \right].
\]
This inequality is the same as in the proof of Lemma 2(i), but with \( \delta = \varepsilon/C \). This yields the necessary result.

**Lemma 3.** Suppose that both \( X \) and \( U \) are absolutely continuous with respect to the Lebesgue measure. Then
\[
\sqrt{n} \| \hat{F}_{m,n,\lambda} - F_{\lambda} \|_\infty = O_p(1)
\]
under certain regularity conditions (e.g., those in (Shorack, 1979, Theorem 2.2)).

**Proof of Lemma 3.** Observe that
\[
\sqrt{n} \| \hat{F}_{m,n,\lambda} - F_{\lambda} \|_\infty = \sqrt{n} \| \hat{F}_{m,n,\lambda} - F_{m,\lambda} + F_{m,\lambda} - F_{\lambda} \|_\infty
\]
\[
\leq \sqrt{n} \| \hat{F}_{m,n,\lambda} - F_{m,\lambda} \|_\infty + \sqrt{n} \| F_{m,\lambda} - F_{\lambda} \|_\infty
\]
\[
= O_p(1) + \sqrt{n} \| F_{m,\lambda} - F_{\lambda} \|_\infty.
\]
Here the first term is \( O_p(1) \) by Shorack (1979, Theorem 2.2). We now show that \( \sqrt{n} \| F_{m,\lambda} - F_{\lambda} \|_\infty = O_p(1) \).

We will show that \( \| F_{m,\lambda} - F_{\lambda} \|_\infty \leq C \| \hat{F}_U - F_U \|_\infty \) in order to obtain the necessary result. The argument is inductive in \( \lambda \). For \( \lambda = 0 \), note that \( F_{m,\lambda} = F_X \ast F_U \) and \( F_{\lambda} = F_X \ast F_U \), so the conclusion holds trivially. Assume that for \( \lambda \in \{1, \ldots, \lambda' - 1\} \), our inductive hypothesis is that
\[
\| F_{m,\lambda} - F_{\lambda} \|_\infty \leq C_\lambda \| \hat{F}_U - F_U \|_\infty.
\]
Then for \( \lambda = \lambda' \),
\[
\| F_{m,\lambda'} - F_{\lambda'} \|_\infty = \| F_X \ast F_U \ast \left( \hat{F}_U^{*,\lambda'} - F_U^{*,\lambda'} \right) \|_\infty \leq \| \hat{F}_U^{*,\lambda'} - F_U^{*,\lambda'} \|_\infty,
\]
where the inequality follows from Lemma A1. Splitting these terms then gives
\[ \| \hat{F}_U^{* \lambda} - F_U^{* \lambda} \|_\infty = \| \hat{F}_U^{(s[\lambda-t-1])} * \hat{F}_U - \hat{F}_U * \hat{F}_U^{(s[\lambda-t-1])} + \hat{F}_U * F_U^{(s[\lambda-t-1])} - F_U * F_U^{(s[\lambda-t-1])} \|_\infty \]
\[ = \| \hat{F}_U * \left( \hat{F}_U^{(s[\lambda-t-1])} - F_U^{(s[\lambda-t-1])} \right) + F_U^{(s[\lambda-t-1])} * \left( \hat{F}_U - F_U \right) \|_\infty. \]

Using this relation and the triangle inequality, we then get that
\[ \| \hat{F}_U^{* \lambda} - F_U^{* \lambda} \|_\infty \leq \| \hat{F}_U * \left( \hat{F}_U^{(s[\lambda-t-1])} - F_U^{(s[\lambda-t-1])} \right) \|_\infty + \| F_U^{(s[\lambda-t-1])} * \left( \hat{F}_U - F_U \right) \|_\infty \]
\[ \leq (C_{\lambda-t-1} + 1) \| \hat{F}_U - F_U \|_\infty, \]

where the second inequality is from Lemma A1, and the last from the inductive hypothesis. Then, since
\[ \sqrt{n} \| F_{m,\lambda} - F_{\lambda} \|_\infty \leq C_{\lambda} \sqrt{n} \| \hat{F}_U - F_U \|_\infty = O_p(1), \]
we have the desired result.

**Lemma 4.** Suppose that the conditions of Lemma 3 are satisfied and that the estimator in the error-free context can be expressed as a functional \( T(\cdot) \) over distributions and is Fréchet differentiable with respect to \( \| \cdot \|_\infty \) at \( F_\lambda \). Then
\[ \sqrt{n} \left( T(\hat{F}_{n,m,\lambda}) - T(F_\lambda) \right) \overset{d}{\to} \mathcal{N}(0, E[\psi^2_F(X)]), \]
where \( \psi_F \) is the influence function of \( T \) at \( F \).

**Proof of Lemma 4.** This follows almost immediately from Fréchet differentiability and Lemma 3. By Fréchet differentiability
\[ \sqrt{n} \left( T(\hat{F}_{n,m,\lambda}) - T(F_\lambda) \right) = \sqrt{n} \int \psi_F d \hat{F}_{n,m,\lambda} + \sqrt{n} o(\| \hat{F}_{n,m,\lambda} - F_\lambda \|_\infty) \]
\[ = \sqrt{n} \int \psi_F d \hat{F}_{n,m,\lambda} + \sqrt{n} \| \hat{F}_{n,m,\lambda} - F_\lambda \|_\infty \frac{o(\| \hat{F}_{n,m,\lambda} - F_\lambda \|_\infty)}{\| \hat{F}_{n,m,\lambda} - F_\lambda \|_\infty} \]
\[ = \sqrt{n} \int \psi_F d \hat{F}_{n,m,\lambda} + O_p(1). \]
Now, as \( n \to \infty, \)
\[ \sqrt{n} \int \psi_F d \hat{F}_{n,m,\lambda} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_F(X_{i}^\lambda) \overset{d}{\to} \mathcal{N}(0, E[\psi^2_F(X)]), \]
which, together with Slutsky’s theorem, gives the desired result.

**Proof of Theorems 1 and 2.** Proofs for Theorems 1 and 2 follow directly from the arguments in Carroll et al. (1996), Lemmas 2(i)–2(ii) (for consistency), and Lemma 4 for asymptotic normality.
In particular, note that $T(\lambda)$ can be replaced by $G(\lambda)$ in the statement of the lemmas, which gives consistency (asymptotic normality) of $T(\hat{F}_{n,m,\lambda})$ to the extrapolant. Then, by assuming that the extrapolant is correctly specified, consistency follows through extrapolation. For asymptotic normality, the argument progresses as in Carroll et al. (1996, Section 3.3).

NP-SIMEX with Kernel Density Estimation

In order to implement the KDE NP-SIMEX procedure, we need to consider the conditional density of $U_i|X_i$. We can view the sample within the validation set as $\{Y_i, X_i, U_i, Z_i\}$ for each individual (or $\{X_i, U_i\}$ in the event of an external validation set). Then we can take

$$\hat{f}_{U|X}(u|x) = \frac{\hat{f}_{U,X}(u,x)}{\hat{f}_X(x)},$$

$$\hat{f}_{U,X}(u,x) = \frac{1}{n_1h_U h_X} \sum_{i=1}^{n_1} K_U \left(\frac{u-u_i}{h_U}\right) K_X \left(\frac{x-x_i}{h_X}\right),$$

$$\hat{f}_X(x) = \frac{1}{n_1h_X} \sum_{i=1}^{n_1} K_X \left(\frac{x-x_i}{h_X}\right),$$

where $h_U$ and $h_X$ are bandwidth parameters (selected based on the observed data) and $K_U(\cdot)$ and $K_X(\cdot)$ are kernel functions (for instance, the Gaussian kernel). Estimation of the bandwidth parameters is addressed by Hall, Racine & Li (2004), who use cross validation based on integrated squared error. Once estimated, the bandwidth parameters can be used to sample from the conditional distribution given $X = x$. Specifically, to sample conditional on $X = x$, we select an individual $i \in \{1, \ldots, n_1\}$ from the validation set weighted proportional to $K_X \left(\frac{|x-x_i|}{h_X}\right)$. We then draw a realization from the $K_U$ distribution based on the kernel parameter $\hat{h}_U$ and centred at $u_i$. Using Gaussian kernels, this results in drawing a random realization from a normal distribution with mean $u_i$ and variance $\hat{h}_U^2$ (Shalizi, 2022). The analysis conducted previously can then proceed conditional on $X_i = x_i$. The convergence of this modified procedure will be in $n_1$ rather than $n$. If necessary, we can also consider conditioning on additional factors (say $Z_i$) if those are strongly informative.

For the smoothed NP-SIMEX procedure, instead of forming $U'$ directly, we can estimate $\hat{f}_U(u)$ and then sample from the KDE. To do so, with equal probability we sample an index $i \in \{1, \ldots, n_1\}$ and then draw from the distribution corresponding to $K_U$ with the bandwidth parameter $\hat{h}_U$ and centred at $u_i$.

Additional Simulations

These simulations extend the previous setting of estimating the fourth moment, this time by assuming that there are three replicated observations. The errors are taken to be from a contaminated normal distribution as in the first set of simulations, here with $\rho = 0$ and $\rho = 0.5$, and with $X$ still following a $\mathcal{N}(5,4)$ distribution. The sample size is fixed at $n = 15,000$, with $B = 500$ and $M = 10$. These simulations are replicated 1000 times. The results are repeated with two available replicates and are shown in Table A1.

We can see a reduction in MSE for both error distributions. The improvement is far more substantial when the errors are drawn from a contaminated normal distribution. In this specific context, the addition of a third replicate was more effective in decreasing the MSE than was
### TABLE A1: MSE and median based on three replicates when estimating the fourth moment (true value 1273) of a contaminated random variable with either normal or contaminated normal errors.

| Error distribution          | 2 replicates |            | 3 replicates |            |
|-----------------------------|--------------|------------|--------------|------------|
|                             | MSE          | Median     | MSE          | Median     |
| Normal                      | 279.076      | 1273.261   | 270.212      | 1272.946   |
| Contaminated normal ($\rho = 0.5$) | 4074.825     | 1269.713   | 2011.741     | 1268.731   |

*Note:* The results compare having two replicates available versus having three replicates.

![Figure A1: Estimated coverage probability based on 1000 replicated simulations for varying levels of nominal coverage using standard two-sided normal confidence intervals based on the estimated variance. The solid black line indicates a computed coverage equal to nominal coverage.](image)

**Figure A1:** Estimated coverage probability based on 1000 replicated simulations for varying levels of nominal coverage using standard two-sided normal confidence intervals based on the estimated variance. The solid black line indicates a computed coverage equal to nominal coverage.
increasing the sample size from \( n = 15,000 \) to \( n = 30,000 \) (MSE = 2131.141 when \( n = 30,000 \)). These results lend credibility to both the proposed method when including a larger number of replicates and demonstrate that the additional information is useful for improving the quality of the correction.

Additionally, we investigate the proposed variance estimation technique. Taking the same scenario as in the second simulation, with \( n \) to be one of 500, 5000, 15,000, or 50,000, we consider using a jackknife-inspired variance estimation technique (specifying a quadratic extrapolant for the variance terms). This extrapolant was chosen based on a visual inspection rather than through derived theory. The simulations are replicated 1000 times, and the results are summarized in Figure A1 and Table A2.

From these results, it is clear that this procedure tends to approximate nominal coverage adequately when the sample size is sufficiently large. When \( n = 500 \), we see fairly poor coverage results, but this improves as \( n \) increases. It is worth reiterating that these results assume a quadratic extrapolant for both variance estimation and the point estimate. While this quadratic term is theoretically justified for the point estimate, the same justification was not used for the variance terms. It has been discussed that the quadratic extrapolant tends to be conservative in the standard setting (Cook & Stefanski, 1994). While this is generally advisable for a point estimate if in doubt, it is of course less desirable when estimating the variance of an estimator. As a result, higher order extrapolants with less of a tendency to conservatively fit the data may be preferable for this purpose.

This set of simulations considers the use of validation data when the errors are nonsymmetric. In particular, we consider the true variate to be distributed according to a Gamma distribution with shape parameter 2 and scale parameter 1. The assumed errors have shape parameter 1 and scale parameter 1.5. This gives the measurement error a slightly higher variance than the variate itself, namely a standard deviation ratio of \( 3/(2\sqrt{2}) \). The sample size is taken to be \( n = 100,000 \) with a 5% validation sample, that is, \( m = 5000 \). The true model

---

**Table A2:** Actual coverage for selected nominal coverage levels for various sample sizes.

| Nominal coverage | \( n = 500 \) | \( n = 5000 \) | \( n = 15,000 \) | \( n = 50,000 \) |
|------------------|---------------|---------------|-----------------|-----------------|
| 0.900            | 0.857         | 0.898         | 0.893           | 0.903           |
| 0.950            | 0.909         | 0.954         | 0.935           | 0.946           |
| 0.990            | 0.969         | 0.995         | 0.990           | 0.990           |

**Table A3:** MSE of the estimates of the logistic regression parameters across 1000 replicated simulations for a naive fit, NP-SIMEX, and P-SIMEX.

|        | Naive | P-SIMEX | NP-SIMEX |
|--------|-------|---------|----------|
| \( \beta_0 \) | 0.7631 | 0.2866  | 0.0458   |
| \( \beta_1 \) | 0.6608 | 0.1612  | 0.0090   |
| \( \beta_2 \) | 0.0415 | 0.0110  | 0.0007   |

*Note:* Each fit is based on a validation sample of size 5000 with asymmetric errors.
for $Y_i$ is a logit-link logistic regression model with an intercept $\beta_0 = 2.5$, $\beta_1 = -1.25$ as the slope for $X_i$, and the inclusion of an independent, standard normal variate $Z_i$ with a slope of $\beta_2 = 1$.

We repeat the simulation 1000 times using $B = 200$ and $M = 10$, and consider the nonlinear extrapolant for all parameters. The results are presented in Table A3.