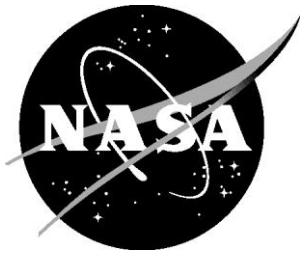
NASA/TM–20240015301



Summary and Annotated Bibliography of Measurement Error Corrections with Potential Application in Future Quesst Mission Community Noise Studies

Nathan B. Cruze
Langley Research Center, Hampton, Virginia

November 2024

NASA STI Program . . . in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA scientific and technical information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI program provides access to the NTRS Registered and its public interface, the NASA Technical Reports Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA Programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counter-part of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include organizing and publishing research results, distributing specialized research announcements and feeds, providing information desk and personal search support, and enabling data exchange services.

For more information about the NASA STI program, see the following:

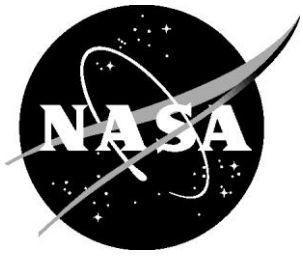
- Access the NASA STI program home page at http://www.sti.nasa.gov

- E-mail your question to help@sti.nasa.gov

- Phone the NASA STI Information Desk at 757-864-9658

- Write to:
  NASA STI Information Desk
  Mail Stop 148
  NASA Langley Research Center
  Hampton, VA 23681-2199

NASA/TM–20240015301



Summary and Annotated Bibliography of Measurement Error Corrections with Potential Application in Future Quesst Mission Community Noise Studies

Nathan B. Cruze
Langley Research Center, Hampton, Virginia

National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia 23681-2199

November 2024

Available from:

NASA STI Program / Mail Stop 148
NASA Langley Research Center
Hampton, VA 23681-2199
Fax: 757-864-6500

This page intentionally left blank.

Abstract

This document is motivated by likely needs of the Quesst mission community response tests, which will culminate in data collection and estimation of dose-response regression relationships for consideration by domestic and international aviation regulators. Furthermore, basic research questions evaluating interactions between rates of community annoyance, dose levels, and indicators of the presence of rattle, vibration, and startle hinge on hypothesis testing in the context of regression models. For a variety of reasons, noise doses may be known only imprecisely and may not reflect the actual level experienced by responding subjects. These differences between true dose and estimated dose, be they systematic or random, constitute *covariate measurement error*. Available statistics literature speaks to the impacts of measurement error on regression models, both in terms of bias in estimated coefficients and predicted values, and in terms of the loss of statistical power for hypothesis testing. Given the particulars of a categorical annoyance response variable and a continuous noise dose predictor variable subject to measurement error during testing, the emphasis of this report is on findings and methods pertinent to generalized linear (and mixed) models likely to be employed during the Quesst mission community tests. We reach the following conclusions:

1. Of four reviewed methods, structural Bayesian measurement error models and simulation extrapolation (SIMEX) may be the most readily applicable to Quesst mission community noise study objectives.
2. If warranted, a linear measurement model can help model systematic sources of measurement error that the classical measurement error does not.
3. For its ready implementation and small additional input requirements, simulation extrapolation may be ideally suited for addressing secondary research questions involving interactions between annoyance, noise dose, and other factors through hypothesis testing.
4. For their flexibility and ability to propagate uncertainty, structural Bayesian hierarchical models have great appeal for mission purposes; some care may be needed in developing appropriate probability models describing actual noise exposure during testing.

An annotated bibliography logs additional papers and resources that may be of value to analysts in other projects and disciplines.

Contents

1 Summary **1**

1.1 Why Covariate Measurement Error Matters in Regression Models 1

1.2 Presumptive Analyses to be Performed Following the Quesst Community Noise Studies 2

1.2.1 Notation and Terminology 3

1.2.2 Predicting Levels of Community Annoyance Given Dose: Obtaining a Population Average Dose-Response Curve 4

1.2.3 Assessing Interactions Between Annoyance, Dose, and Other Effects Through Hypothesis Testing 4

1.3 Approaches to Modeling Covariate Measurement Error 5

1.3.1 Functional Approaches–Regression Calibration and Simulation Extrapolation 5

1.3.2 Structural Approaches–Structural Bayesian Measurement Error Models and Data Cloning 7

1.4 Modeling Measurement Error in Past NASA Risk Reduction Study Data 8

1.4.1 Naive Models 10

1.4.2 Simulation Extrapolation (SIMEX) 12

1.4.3 Structural Bayesian Measurement Error Models 15

1.5 Findings and Recommendations in Relation to Future Community Testing with X-59 17

2 Annotated Bibliography **19**

2.1 Measurement Error Text Books and Review Articles 19

2.2 Regression Calibration 21

2.3 Simulation Extrapolation (SIMEX) 21

2.4 Bayesian Hierarchical Models 23

2.5 Data Cloning 25

2.6 Additional References 27

List of Figures

1 Demonstration of the effects of covariate measurement error 2
2 A notional SIMEX plot 6
3 Histogram of WSPR doses as measured and bar chart of ordinal perceptual responses 9
4 Histogram of QSF18 doses as measured and bar chart of ordinal perceptual responses 9
5 Notional effect of measurement error on naive analysis 10
6 Linear and Quadratic SIMEX plots for model parameters of the random-intercept logistic model for the WSPR (top row) and QSF18 (bottom row) single-event data . 13
7 Comparison of population average curves at WSPR and QSF18 test sites under naive analysis and SIMEX adjusted analysis 15
8 Example population average dose response curves derived by closed-form approximation 18

List of Tables

1 Summary of NASA study data 10
2 Naive AGHQ estimates and standard errors compared to naive Bayesian posterior means and standard deviations 12
3 Naive and SIMEX estimates from WSPR and QSF18 studies 14
4 Posterior means and standard deviations from WSPR study 16
5 Posterior means and standard deviations from QSF18 study 17

1 Summary

The NASA Quesst mission will culminate in a collection of human response data to low-noise supersonic overflights generated with multiple passes of the X-59 demonstrator aircraft. A variety of regression analyses performed using models of categorical response data regressed on a continuous noise 'dose' predictor, with levels measured in decibels, will form a foundation for regulatory decision making discussions and address related research questions.

For a variety of reasons, noise doses may only be known imprecisely. For one, recruited subjects may frequently be indoors, whereas measurements and estimates of dose are to be taken outdoors. Thus, the levels experienced by the subject may differ from best outdoor estimates due to the effects of outdoor-to-indoor transmission. Additionally, the noise doses during Quesst mission will be a synthesis of precise measurements taken from a sparse network of noise monitors fused with a physics-based model describing sonic boom propagation in order to provide estimates of outdoor dose with the greatest area coverage possible. Model uncertainty in the physics-based model and sparsity and uncertainty in the measurements at monitors themselves will result in some degree of prediction error or uncertainty at any given latitude and longitude coordinate in the test region.

Any difference between the best estimate of dose and the actual dose experienced by the recruited subject constitutes *measurement error.* The first part of this document summarizes several implications of measurement error in Section 1.1 and describes prospective analyses to be performed given data collected during the Quesst mission in Section 1.2. We identify and discuss several mitigations for covariate measurement error in Section 1.3, and we demonstrate application of two methods on past risk reduction study data in Section 1.4 with some final observations and conclusions in Section 1.5. Section 2 comprises an annotated bibliography of a variety of references that may offer some utility for addressing measurement error in the analysis of Quesst mission community test data. Furthermore, these collected references should be of broader interest to any analyst recognizing that collected predictor variables may be subject to imprecision.

1.1 Why Covariate Measurement Error Matters in Regression Models

The following reproduced example is discussed in Carroll et al. (2006), and it illustrates what the authors referred to as the 'triple whammy' of measurement error. The top panel of Figure 1 depicts a relationship between a continuous response variable Y and a continuous predictor variable X that varies about the mean function $E[Y|X] = \sin(2X)$. Now assume that in lieu of the true predictor, an error-prone version W is observed, where $W = X + U$ and the left-or-right additive perturbations, U, are independent random deviates from a normal distribution with mean zero, and standard deviation $\sigma_u = \frac{2}{3}$. The resulting scatter is shown in the lower panel of Figure 1.

Contrasting the two panels of Figure 1, measurement error poses the following issues:

1. *Features of graphical and exploratory analysis may be hidden.* In the plots below, the most obvious feature, the sinusoidal relationship, is effectively washed out in the bottom panel.

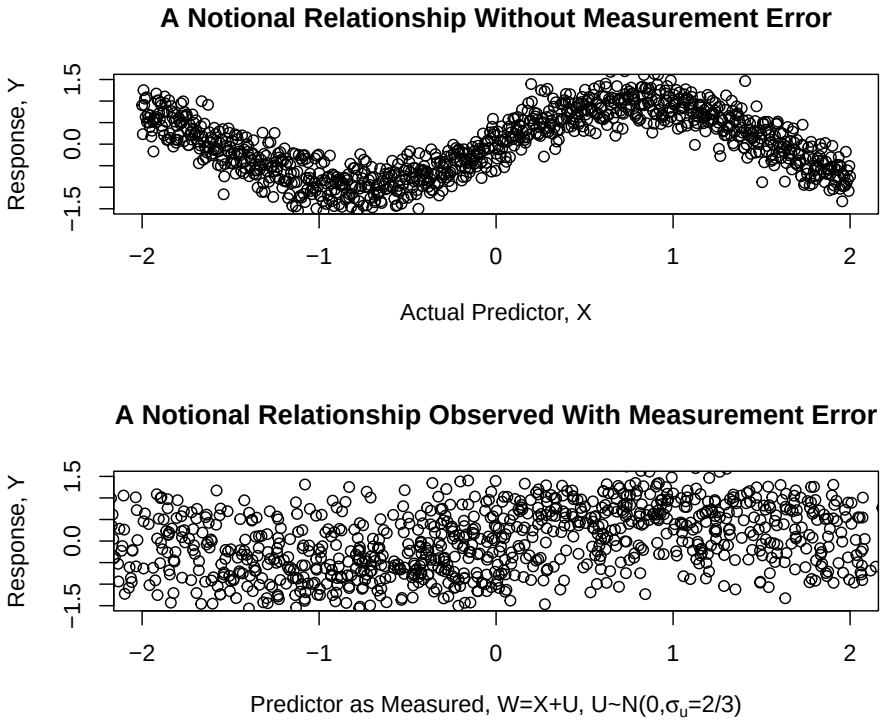
2. *Bias.* Since the sinusoidal pattern is obscured, estimates of period, amplitude, and other features may be systematically affected, or an entirely different functional relationship between response and predictor may be posited as a result.

3. *Loss of statistical power in hypothesis tests.* One might ask if the average response at a particular level of predictor (e.g., -2) is non-zero. In the top panel with pristine measurements, the answer is apparent; the average response is non-zero. In the lower panel, the error prone measurements concentrated at -2 span a much greater range along the vertical axis and cover

zero. This depicts a loss of statistical power, a decrease in the probability detecting an effect given that one is actually present.

The latter two points are of immediate concern as they relate directly to analyses following the Quesst mission community test campaigns.

Figure 1: Demonstration of the effects of covariate measurement error



1.2 Presumptive Analyses to be Performed Following the Quesst Community Noise Studies

Following procedures used in past NASA risk reduction studies, a recruited sample will be selected from communities to respond to a NASA-generated noise stimulus. In the future Quesst mission studies, the noise stimulus will be the low-noise supersonic overflight of the X-59 demonstrator. The recruited subjects will complete *repeated surveys* promptly after each supersonic overflight, and they will complete multiple end-of-day surveys in which their reflections on each flight day will be captured. These survey instruments will contain questions about the degree of annoyance a subject may experience, as indicated on a five-category ordinal scale.

As NASA seeks to quantify population-average behavior, rates of annoyance as a function of noise dose level, two common approaches arise for regressing the non-normal response variable on repeated exposures to continuous noise dose: generalized estimating equations (GEE), and generalized linear (and mixed) models (GLMM). At this time, the prevailing thinking is to obtain population average through some form of GLMM, therefore, the remainder of this review looks at measurement error in that context, using the random intercept logistic regression as a concrete example.

1.2.1 Notation and Terminology

Equation 1 represents the class of GLMM in matrix form:

$$g\left(E\left[\boldsymbol{y}|\boldsymbol{c}\right]\right) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{c}. \tag{1}$$

For a given link function $g(\cdot)$, the mean value of a response $\boldsymbol{y}$, conditioned on one or more random effects $\boldsymbol{c}$, is linearly related to a set of fixed effects $\boldsymbol{\beta}$ and the subject-specific random effects.

Defining the conditional mean response $E\left(y_{ij}|x_{ij}, c_i\right) \equiv p\left(x_{ij}, c_i\right)$ and choosing the *logit* link function, the random intercept logistic regression is specified as

$$logit\left(p\left(x_{ij}, c_i\right)\right) = \beta_0 + \beta_1 x_{ij} + c_i \text{ with } c_i \sim N\left(0, \sigma_c^2\right) \tag{2}$$

where $i \in \{1, 2, \ldots, I\}$ is an index over the number of unique participants, $j \in \{1, 2, \ldots, J_i\}$ indexes the (possibly different) number of responses provided by the i^{th} participant and x_{ij} denotes the noise exposure (dose) administered to the i^{th} participant during the j^{th} boom event. The scalar random effect c_i denotes a subject-specific intercept, assumed to be drawn from a normal distribution with mean zero and finite variance, σ_c^2. The categorical response, y_{ij}, is a binary indicator variable, taking value 1 if the participant is highly annoyed and 0 otherwise.

The expected value of a binary random variable is a well-defined probability or proportion. Whereas an unconstrained linear probability model may produce predicted probabilities outside the interval $(0, 1)$, the inverse logit maps all real values that may arise in the linear predictor space into this interval. For a specific participant, the probability that he or she is highly annoyed given dose and individual intercept is stated in Equation 3:

$$p\left(x_{ij}, c_i\right) = logit^{-1}\left(\beta_0 + \beta_1 x_{ij} + c_i\right) = \frac{1}{1 + exp\left(-\left[\beta_0 + \beta_1 x_{ij} + c_i\right]\right)}. \tag{3}$$

Let w_{ij} denote an error-prone, observed measurement that corresponds to the true but unobserved x_{ij}. The measurement error problem arises in a *naive* analysis, i.e., modeling $p(w_{ij}, c_i)$ instead when $p(x_{ij}, c_i)$ was intended. Understanding the impact of measurement error necessitates understanding the relationship between actual x_{ij} and the observed w_{ij}. Keogh et al. (2020) provides overview of several statistical measurement models that describe plausible relationships between the observed and underlying latent doses.

1. The Berkson error model: The Berkson error model is often associated with rounding, binning, or assigning identical measurements, e.g., an available average value, to all members of the same subgroup. It is represented as

   $$X = W + U \tag{4}$$

   where the true value X can be thought of as arising from measured values W and an independent, mean-zero error U. During a community noise test, a hypothetical example of Berkson error might entail assigning a reading from the nearest noise monitor to all individuals in the same apartment complex or neighboring houses.

2. Classical measurement error model: In contrast with the Berkson error above, the classical measurement error model treats manifest measurements W as arising from an underlying true level X which is independent of the mean-zero error term U:

   $$W = X + U \tag{5}$$

This is a commonly assumed measurement model for continuous covariates. In the Quesst community noise studies, estimates of outdoor noise levels may be obtained with some geographic specificity. That is, estimates may be had at indicated latitude and longitude coordinates. The fact that the estimates are predictions based on a fusion of a network of noise monitors and a physics-based model may mean the that best estimate W is subject to prediction error.

3. The linear measurement error model: The linear measurement error model allows the flexibility of addressing both systematic discrepancy and random errors.

$$W = \alpha_0 + \alpha_1 X + U \tag{6}$$

Keogh et al. (2020) note that the coefficient α_0 speaks to location bias, that is, biases that don't depend on the magnitudes of X, and that α_1 addresses scale biases which do. Note that the classical measurement error model can be viewed as a special case of Equation 6 in which $\alpha_0 = 0$ and $\alpha_1 = 1$. Hypothetical examples in community noise testing could include changes in levels due to outdoor-to-indoor transmission, which could be modeled appropriately with the α_0 coefficient. Additionally, if measurement error changes with the magnitude of the noise dose, these could be modeled with suitable values of α_1.

1.2.2 Predicting Levels of Community Annoyance Given Dose: Obtaining a Population Average Dose-Response Curve

The means of obtaining marginal (population average) predictions from conditional (subject-specific) models involves numerical integration as discussed in Pavlou et al. (2015), Hedeker et al. (2018), or relevant closed-form approximation as in Wakefield (2013) for Bayesian models. In the context of the random-intercept logistic regression model, marginalization is an integral with respect to the distribution of random intercepts:

$$\widehat{p}\,(x) = \int_{-\infty}^{\infty} \frac{1}{1 + \exp\left(-\left[\hat{\beta}_0 + \hat{\beta}_1 x + c\right]\right)} \phi\left(c|\widehat{\sigma}_c^2\right) dc \tag{7}$$

where $\phi\left(c|\widehat{\sigma}_c^2\right)$ denotes a mean zero normal distribution with estimated variance $\widehat{\sigma}_c^2$. The result is a function describing the probability of high annoyance as a function of dose, and it is not dependent on the intercept of any specific subject. However, in the naive analysis, estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are attenuated, or biased toward zero. Consequently, when there is appreciable measurement error in dose, marginal predictions of annoyance as a function of dose are also biased. Stefanski & Carroll (1985) note that there is potential to overstate annoyance at low ends of the dose range and underestimate annoyance at high ends of the noise dose range. As one of the critical objects to be delivered to regulators at the conclusion of the Quesst community noise studies, predictions of annoyance as a function of dose can potentially be improved by adopting some choice of measurement error model.

1.2.3 Assessing Interactions Between Annoyance, Dose, and Other Effects Through Hypothesis Testing

In addition to assessing its relationship with noise dose, additional research objectives call for investigating whether annoyance is related to other factors including the presence of rattle (denoted r_{ij}), vibration (v_{ij}), or startle (s_{ij}). These variables will be binary indicators determined by subject

response to survey questions about detecting rattle and vibration, or experiencing startle. Following (Fidell et al., 2020, Sec. 6.4) and (Opsomer et al., 2024, Ch. 11), the logistic model in Equation 8 and any submodel thereof can be fit to the collected data:

$$logit\left(p\left(x_{ij}, r_{ij}, s_{ij}, v_{ij}\right)\right) = \beta_0 + \beta_1 x_{ij} + \beta_2 r_{ij} + \beta_3 s_{ij} + \beta_4 v_{ij}. \tag{8}$$

Estimated coefficients for each predictor variable can be interpreted as the effect (on the log odds ratio scale) on the probability of annoyance. Traditional hypothesis tests can be used to assess whether the effect is significantly different from zero.

When a naive analysis is performed, i.e., when as-measured doses w_{ij} are used in lieu of the true dose level x_{ij} in Equation 8 above, the hypothesis tests may lack sufficient power, and the effects may be deemed statistically insignificant, when in fact, they are. Gustafson (2004) notes that in regression models with multiple predictor variables, the issues of bias in coefficients and statistical power are not limited to the variable measured with error, i.e., dose. Thus, mitigations for anticipated measurement error in dose can be used to improve complementary research objectives, even if the survey respondent knows unambiguously that they were startled, or detected rattle or vibration during an overflight.

1.3 Approaches to Modeling Covariate Measurement Error

The literature speaks to two broad classes of measurement error approaches. In *functional measurement error modeling*, the distribution of the true predictor variable is not modeled parametrically. That is, no explicit assumptions about the distribution of true dose needs to be made. In contrast, *structural measurement error modeling* entails specific parametric assumptions about the distribution to the true predictor. We sketch a few popular methods of each type, with an emphasis on the utility of simulation extrapolation and Bayesian hierarchical models for the purposes of mitigating measurement error in the Quesst mission community noise studies.

1.3.1 Functional Approaches–Regression Calibration and Simulation Extrapolation

Regression Calibration: Regression calibration is discussed by Carroll & Stefanski (1990) and even earlier in the case of the Cox proportional hazards model in Prentice (1982). Regression calibration is a popular method for its relative simplicity and broad applicability. It can be broken into a three step process. Letting Z denote additional explanatory variables, X denote the true value of a covariate, and W its error-prone surrogate, fit the calibration model $E\left[X|W, Z\right]$. Second, use the fitted model to predict $\widehat{X}$ for all cases, and substitute these into the analysis model in lieu of the unobservable X. The third step is to adjust standard errors or confidence intervals to reflect the combined uncertainties of the calibration step and the outcome model.

Note that the description above posited that some data on the true covariate X exists to begin with, at least for a subset of the data. This could be had if the study is designed to collect some internal validation data, or if some other unbiased instrument for X exists. For the study of a first-of-kind noise source, the shaped sonic boom of the X-59, there will be little available 'gold standard' data on which to build a calibration model. In addition, Carroll et al. (2006) caution on the naive use of regression calibration for the class of GLMM subject to measurement error, as it correctly specifies fixed-effects structure, but does not correctly specify the random effects structure. Consequently, biases in the variance component of the naive models may not be corrected by regression calibration. Given that the minimum requirements to fit a calibration model may not be available or equally applicable across distinct test sites and that the variance

components may not be sufficiently adjusted by regression calibration, it is not considered a viable method for correcting measurement error in the Quesst community noise studies.

Simulation Extrapolation (SIMEX): Simulation extrapolation was initially developed by Cook & Stefanski (1994). The intuition of the method is that subjecting naive models to further error induces a pattern that can be learned experimentally. After a suitable pattern is learned, it can be used to project back to the case where there is no measurement error and model point estimates are ostensibly free of bias.

The algorithm begins by fitting the naive model and noting the resulting point estimates for all model parameters; for the random intercept logistic model these are $\widehat{\Theta}_{naive} \equiv \left(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}_c^2\right)$. Provided a suitable estimate of the measurement error variance $\widehat{\sigma}_u^2$, measurement error is inflated by a factor λ along a discrete set of values determined by the practitioner, often $\{0.5, 1.0, 1.5, 2.0\}$ in practice. For each value λ, increasing amounts of measurement error $(1+\lambda)\,\widehat{\sigma}_u^2$ are injected into the error-prone covariate and the naive analysis is repeated a large number of times, denoted B. An estimate of center for each model parameter, typically a mean or a median, is computed across the B data sets for each model parameter. These simulated means are matched to the corresponding λ coordinate as illustrated in Figure 2. Finally, the SIMEX estimator, $\widehat{\Theta}_{SIMEX}$ is obtained by choosing a functional form for an extrapolant function related to the observed trend, and extrapolating to the case of $\lambda = -1$, thus 'canceling out' the measurement error. In principle, many functional forms could be chosen, but generally a small number of data points will be available. Statistical software often focuses on three extrapolant functions: linear, quadratic, and rational linear functions. The quadratic extrapolant function has been observed to perform well in a variety of settings, and Carroll et al. (2006) note that it typically results in conservative corrections for attenuation bias.

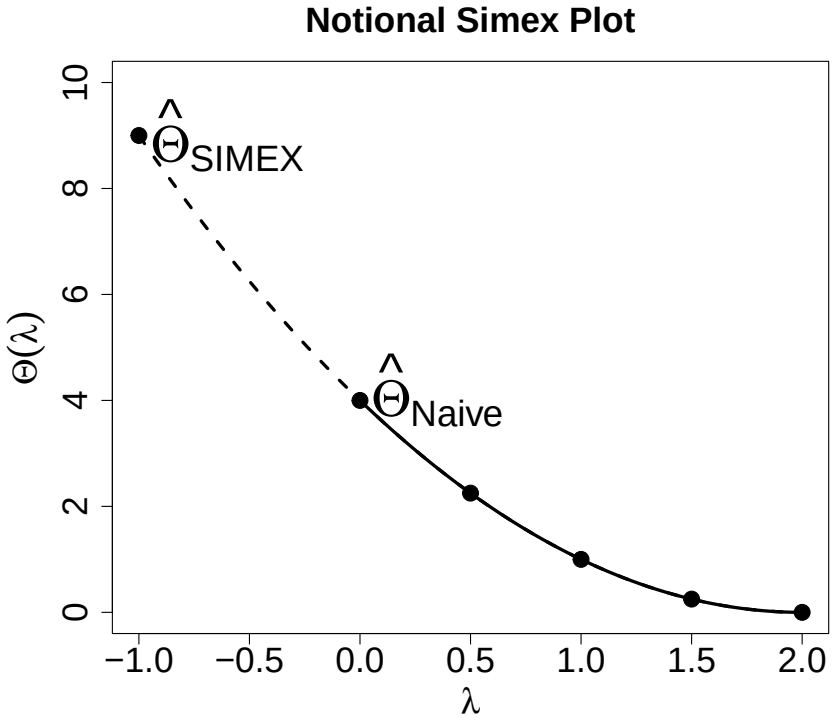


Figure 2: A notional SIMEX plot

1.3.2 Structural Approaches–Structural Bayesian Measurement Error Models and Data Cloning

Bayesian measurement error models: Richardson & Gilks (1993) and Gustafson (2004) describe a 'formula' for constructing Bayesian hierarchical models for measurement error adjustment in terms of a product of three probability models: a response model, a measurement model, and an exposure model. For the random-intercept logistic model, these are represented by Equation 9, Equation 10, Equation 11, respectively:

$$
\begin{aligned}
&f\left(\boldsymbol{x}, \beta_0, \beta_1, \sigma_c | \boldsymbol{w}, \boldsymbol{y}\right) \propto \\
&\prod_{i=1}^{I} \prod_{j=1}^{J_i} p\left(x_{ij}, c_i\right)^{y_{ij}} \left[1 - p\left(x_{ij}, c_i\right)\right]^{1-y_{ij}} \phi\left(c_i | \sigma_c\right) && (9) \\
&\times f\left(w_{ij} | x_{ij}\right) && (10) \\
&\times f\left(x_{ij}\right) && (11) \\
&\times f\left(\beta_0\right) f\left(\beta_1\right) f\left(\sigma_c\right). && (12)
\end{aligned}
$$

Noting the use of the true but unobserved dose variable x_{ij} in the response model (Equation 9), the relationship between the as-measured surrogate doses w_{ij} and actual doses x_{ij} is described by a measurement model (or combination of measurement models) like those discussed in Section 1.2.1. As written, Equation 10 is consistent with a classical measurement error assumption, that each w_{ij} arises from a corresponding unobservable x_{ij}. The part that makes the model a structural model is Equation 11, an assumed parametric form for the distribution of underlying true doses; some authors refer to this as a prior distribution. In Section 1.4.3, we identify assumptions about this distribution as a significant source of sensitivity when modeling past NASA risk reduction study data. The complete model specification requires a joint prior distribution on unknown model parameters, often taken as the product a priori independent probability distributions for each model parameter. These are denoted by the generic distributions for the intercept β_0, slope β_1, and standard deviation describing the spread of random intercepts σ_c in Equation 12. A variety of computing paradigms, often based on Monte Carlo simulations, enable access to samples from the full posterior distribution from which point estimates and corresponding estimates of uncertainty can be obtained.

Data cloning: The method of data cloning was developed by Lele et al. (2007, 2010) as a means to obtain maximum likelihood estimates using the machinery of Bayesian computation. The theory uses asymptotic results relating the Bayesian model obtained by replicating the data many times, say k. In essence, the data are modeled as though they come from k independent repetitions of the experiment, and each of those k experiments just happen to have identical results. As the number of clones k goes to infinity, the means of the posterior distributions converge to the maximum likelihood estimator, and a simple matrix estimator converges to the asymptotic variance-covariance matrix for these estimators. One appealing feature is that the obtained maximum likelihood estimates do not actually depend on the choice of assumed prior distributions on model parameters, which is sometimes held out as a source of subjectivity in the construction of Bayesian models. Torabi (2013) went on to apply this method for GLMM measurement error models.

While this is an interesting methodological accomplishment, some practical matters are worth considering as they relate to utility for future Quesst community noise studies. First, while data cloning is said to be invariant to the choice of prior distributions on model parameters, the method is firmly a structural measurement error approach. Like the structural Bayesian measurement error model it is built on, there remains some possible sensitivity due to (mis)specification of the exposure

model. Second, the matter of practical computing remains. Where the larger of two NASA risk reduction studies was on the order of 5,000 data rows, those coming from each community test under Quesst mission could be on the order of *80,000 data rows*, assuming current targeted values of 80 supersonic passes over a test duration and assuming working targets of 1,000 recruited subjects at each test site. In order for the posterior means from the Bayesian model to converge to the maximum likelihood estimators, a large number of clones may be necessary, and the task becomes a more intensive task of computing a Bayesian model on data sets of size $k \times 80,000$. Data cloning will not remove the need for a choice of exposure model, and it comes with greater computing cost, so that it may not be a practical method for application in the Quesst mission. However, since it uses the framework of Bayesian computation, some of the references dedicated to reducing the computational burden of data cloning as noted in the annotated bibliography in Section 2.5 may still expedite fitting structural Bayesian measurement error models on data of the size anticipated during the Quesst community noise studies.

1.4 Modeling Measurement Error in Past NASA Risk Reduction Study Data

NASA conducted two past risk reduction studies, generating data with some shared or anticipated features of the upcoming X-59 community studies. The first, called Waveforms and Sonic Boom Perception and Response (WSPR), was conducted in the vicinity of Edwards Air Force Base in southern California in 2011. A second, larger study called Quiet Supersonic Flights 2018 (QSF18) was conducted near Galveston, Texas. For procedural details on these studies, see Page et al. (2014), Page et al. (2020a), Page et al. (2020b).

In brief, both studies used research F-18 aircraft to generate a series of low-amplitude sonic booms delivered across multiple flight days and approximating some features and levels of the anticipated X-59 acoustic profile. Recruited subjects promptly completed single-event surveys after each supersonic dive maneuver, as well as daily summary surveys capturing attitudes toward cumulative noise exposures after each day of testing. The single-event data sets with as-measured doses form the basis for the measurement error analysis in this report. Histograms of as-measured doses for both studies are depicted in the left-hand panels of Figure 3 and Figure 4, respectively. Note that the WSPR data set has a larger dose range and maximum level than the QSF18 data, given deliberate testing of traditional sonic booms, and even the capture of adventitious sonic booms happening near the Edwards Air Force Base area at the time. In short, the WSPR dose data were generated by a *mixture* of 84 planned low-noise supersonic dive maneuvers, 5 planned traditional sonic booms, and 21 unplanned traditional sonic sonic booms, for a total of 110 single noise events. During QSF18, NASA completed 52 low-noise supersonic dive maneuvers. In addition, estimates of noise dose uncertainty, assumed constant across the test areas and expressed as a standard deviation, were obtained by procedures adopted in Page et al. (2014) and Page et al. (2020a).

The bar charts in the right hand panels of Figure 3 and Figure 4 capture the multiple ordinal responses provided by recruited subjects in the WSPR and QSF18 studies, respectively. Between WSPR and QSF18, NASA experimented with two different recognized socio-acoustic response scales, an ll-point ordinal scale, and a 5 category verbal scale. Total numbers of responses collected in each bin are annotated in the figures. A commonly adopted convention dichotomizes responses into 'highly annoyed' (responses of 8 or greater on the 11-point scale, and 'very' or 'extremely annoyed' on the verbal scale) and 'not highly annoyed' (otherwise). Despite the different choices of response scale, the collected data both point to the rarity of the 'highly annoyed' outcome, happening in less than 7% of observed cases in WSPR, and in less than 1% of observed cases in QSF18. Important summary data including sample size, total responses, total highly annoyed responses, and summary statistics for doses as measured are gathered in Table 1.

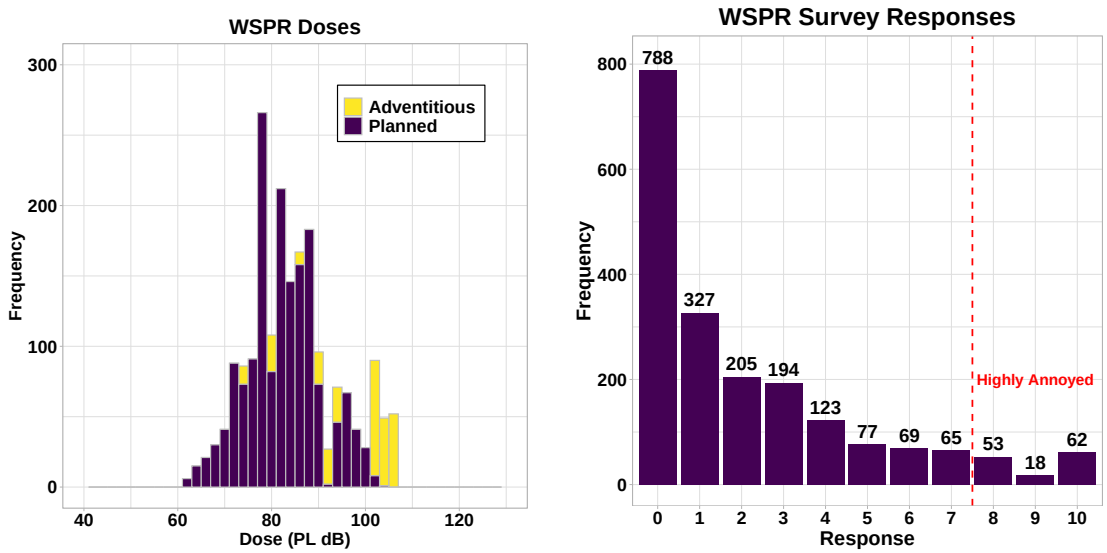


Figure 3: Histogram of WSPR doses as measured and bar chart of ordinal perceptual responses

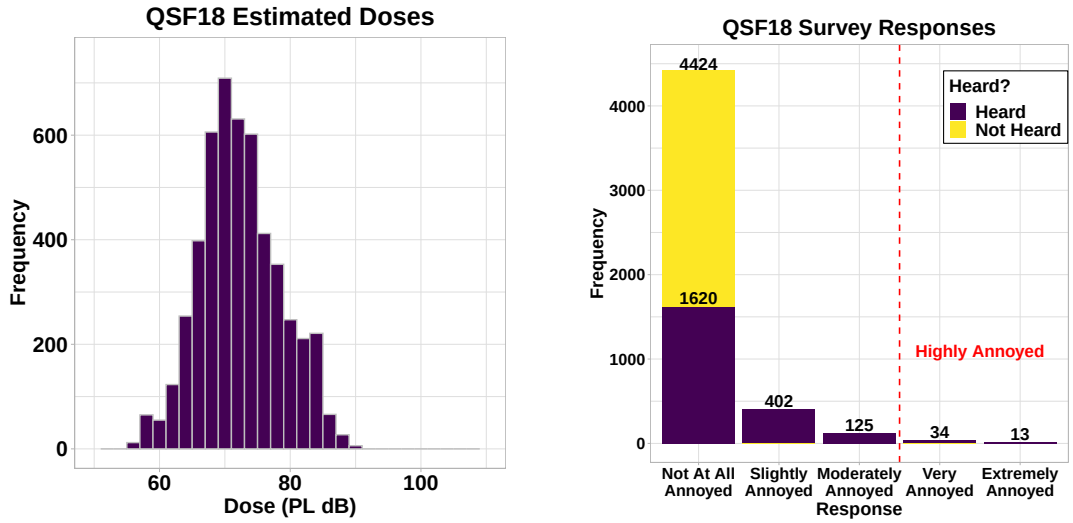


Figure 4: Histogram of QSF18 doses as measured and bar chart of ordinal perceptual responses

Table 1: Summary of NASA study data

Study	**WSPR**	**QSF18**
Number of subjects, I	49	371
Number of supersonic events	110	52
Total responses	1,981	4,998
Total highly annoyed, $\sum_i \sum_j y_{ij}$	133	47
Range of as-measured doses, w_{ij}, in PL dB	63 to 106	56 to 90
Deviation, $\hat{\sigma}_u$, in PL dB	3.7	4.9

1.4.1 Naive Models

Following the discussion in Section 1.2.2, bias is a key issue with naive regression analysis when one or more predictor variables are subject to measurement error. In the context of the presumptive dose-response analysis, noise doses may be subject to error as described by one (or more) of the measurement models described in Section 1.3. In the case of non-differential, classical measurement error, the effect on coefficients of the subject specific model is that of *attenuation*, or a bias toward zero. Biases in estimated coefficients of the subject-specific model then propagate into the population-average dose response model, which depends on integration with respect to the distribution of random effects, taking estimated coefficients as given in the integration. The effect on the predicted mean response is overprediction of annoyance at the low end of the dose range and underprediction of annoyance at the high end of the dose range Stefanski & Carroll (1985); this effect is illustrated in the notional plot in Figure 5.

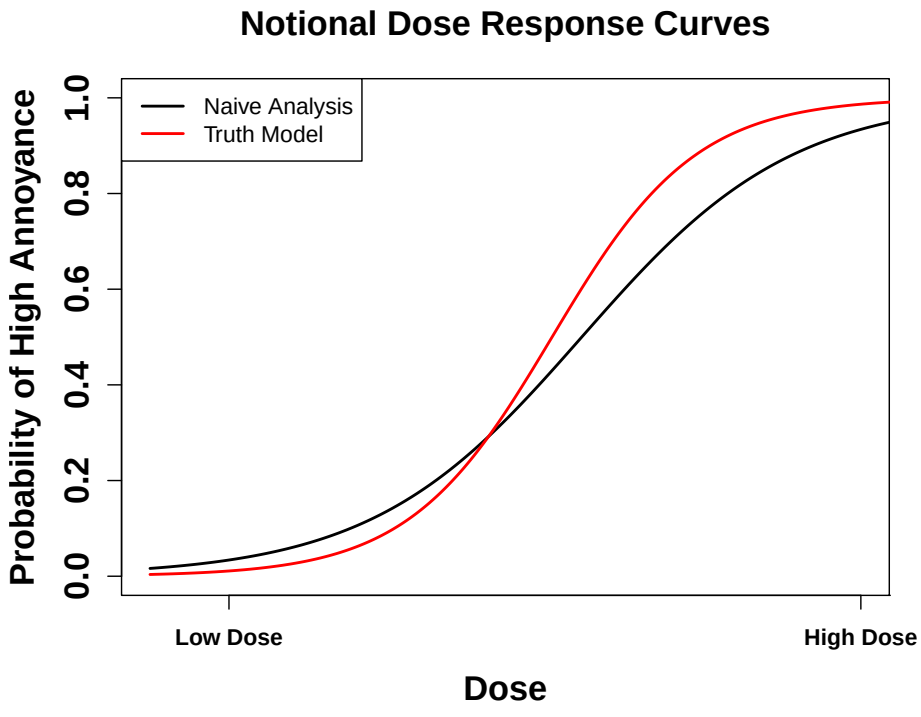


Figure 5: Notional effect of measurement error on naive analysis

In the naive analysis, the observed covariate, as-measured dose in decibels, w_{ij}, is included in the subject-specific model. The naive, subject-specific model can be fit in a Frequentist paradigm using maximum likelihood estimation, or in a Bayesian paradigm, specifying a choice of prior distributions on unknown model quantities, e.g., parameters β_0, β_1,σ_c^2, in the random intercept logistic model.

Maximum likelihood via Adaptive Gauss-Hermite Quadrature: The maximum likelihood approach proceeds on the basis of optimization, finding values of parameters that simultaneously maximize the marginal likelihood function, that is, obtain $\arg\max_{\beta_0,\beta_1,\sigma_c^2} L\left(\beta_0,\beta_1,\sigma_u^2|\boldsymbol{w},\boldsymbol{y}\right)$ where

$$\begin{aligned} L\left(\beta_0,\beta_1,\sigma_u^2|\boldsymbol{w},\boldsymbol{y}\right) &= \prod_{i=1}^{I} L_i\left(\beta_0,\beta_1,\sigma_c^2|\boldsymbol{w}_i,\boldsymbol{y}_i\right) \\ &= \prod_{i=1}^{I}\int_{-\infty}^{\infty}\prod_{j=1}^{J_i} p\left(w_{ij},c_i\right)^{y_{ij}}\left[1-p\left(w_{ij},c_i\right)\right]^{1-y_{ij}}\phi\left(c_i|\sigma_c^2\right)dc_i. \end{aligned} \tag{13}$$

In the context of naive analysis of QSF18 data, Cruze et al. (2022) discussed the availability of a variety of computational maximum likelihood estimators, each designed to approximate the integral or the integrand in Equation 13. Based on the findings in that report, a 25-node Adaptive Gauss-Hermite Quadrature (AGHQ) estimator is used in the present analysis. This is important for two reasons. First, the estimator is believed to offer sufficient accuracy, so that any biases in estimated coefficients of a naive model can be attributed wholly to measurement error and not to numerical instability of the estimator. Second, the application of Simulation Extrapolation in latter sections depends on repeated application of a choice of naive estimator, which we make explicit.

Bayesian analysis with assumed Cauchy family priors: In contrast to maximum likelihood estimation, which is based on optimization involving a difficult integral, Bayesian computation relies on *sampling*. In the Bayesian paradigm, additional probability distributions known as prior distributions are assigned to the unknown model parameters β_0, β_1,σ_c^2. A fully joint relationship between parameters and data can be expressed as a product of the full likelihood function (Equation 14) and priors (Equation 15), and the joint posterior distribution is proportional to this product, up to a constant determined by the marginal distribution of the data:

$$f\left(\beta_0,\beta_1,\sigma_c|\boldsymbol{w},\boldsymbol{y}\right) \propto$$

$$\prod_{i=1}^{I}\prod_{j=1}^{J_i} p\left(w_{ij},c_i\right)^{y_{ij}}\left[1-p\left(w_{ij},c_i\right)\right]^{1-y_{ij}}\phi\left(c_i|\sigma_c\right) \tag{14}$$

$$\times\, f\left(\beta_0\right) f\left(\beta_1\right) f\left(\sigma_c\right). \tag{15}$$

Any of a wide variety of prior distributions may be assumed. Commonly, noninformative prior distributions are assumed. For the naive analysis we assume popular choices of a priori independent, noninformative Cauchy family priors discussed in Gelman et al. (2008), Polson & Scott (2012). Namely, the assumed prior distributions on regression coefficients and the *standard deviation* of random intercepts are, respectively, $\beta_0 \sim t(0,10,1)$, $\beta_1 \sim t(0,2.5,1)$, and $\sigma_c \sim t^{+}(0,1,1)$, where the triplet of arguments refer to location, scale, and degrees of freedom parameters of these distributions. This assumption is important in as much as it produces naive estimates comparable to the AGHQ estimator. Further, by retaining these choices in the latter application of Bayesian measurement error models, we can attribute the observed changes in estimated coefficients to differences in structural assumptions on the distribution of true dose.

The estimated coefficients of the naive, subject-specific models are presented in Table 2, and it represents a template for results presented in subsequent sections. First, within each data set, the maximum likelihood estimates and standard errors obtained from the AGHQ estimator and the posterior means and posterior standard deviations from the Bayesian models are approximately equal. Thus, any change or improvement, whether obtained from SIMEX or Bayesian approaches, will be measured from the same starting values. The coefficients presented in Table 2 are features of the subject-specific model. For each subject, i, dose that elicits high annoyance with probability p is given by

$$d_p = \left[\log\left(p/\left(1-p\right)\right) - \left(\widehat{\beta}_0 + \widehat{c}_i\right)\right]/\widehat{\beta}_1. \tag{16}$$

By symmetry arguments, when $\widehat{c}_i = 0$, $\widehat{d}_{50} = \frac{-\widehat{\beta}_0}{\widehat{\beta}_1}$ is also a point on the *population average dose response curve*. Thus, a point estimate of the noise dose level that would annoy 50% of the population can be obtained by simple arithmetic given estimated coefficients from the subject-specific model. The naive estimates imply that the two populations at Edwards Air Force Base (WSPR) and Galveston (QSF18) respond similarly, and that doses of approximately 124 to 125 dB would cause half of each population to respond as highly annoyed.

Table 2: Naive AGHQ estimates and standard errors compared to naive Bayesian posterior means and standard deviations

Data	Estimator	$\widehat{\beta}_0$ (SE)	$\widehat{\beta}_1$ (SE)	$\widehat{\sigma}_c$ (SE)	$\widehat{d}_{50}$ (dB)
WSPR	AGHQ	-19.29 (1.70)	0.154 (0.015)	3.34 (—)	125.3
WSPR	Bayes	-19.22 (1.69)	0.154 (0.015)	3.43 (0.72)	125.0
QSF18	AGHQ	-18.67 (2.40)	0.151 (0.028)	2.49 (—)	123.6
QSF18	Bayes	-18.62 (2.40)	0.150 (0.029)	2.50 (0.50)	124.1

1.4.2 Simulation Extrapolation (SIMEX)

The SIMEX procedure, a functional measurement error approach, was described in Section 1.3.1. Among decisions to be made by the practitioner, SIMEX relies on a sequence of measurement error variance inflation factors (λ_m), a number of pseudoreplicates (B), and a choice of extrapolant functions used to 'remove' the effects of measurement error. As applied to the WSPR and QSF18 single-event data sets, we assume $\lambda_m \in \{0.5, 1.0, 1.5, 2.0\}$, $B = 250$, and experiment with two choices of extrapolant fuction: linear and quadratic functions.

Figure 6 contains SIMEX plots for the parameters of the subject-specific models applied to both the WSPR data (top row) and the QSF18 data (bottom row). The naive AGHQ point estimates shown earlier in Table 2 appear at the $\lambda = 0$ coordinates in each panel. The simulation phase of the SIMEX procedure uses a large number of pseudoreplicates subject to increasing levels of measurement error. The points in each panel with positive λ coordinates represent the arithmetic averages of naive point estimates obtained from each of the 250 pseudoreplicate datasets.

A pattern is learned experimentally, and through a choice of extrapolant function, the pattern is projected back to the case of 'no measurement error' at the $\lambda = -1$ coordinate. The act of plotting should help guide the choice of extrapolant function. For each data set, the linear trend (SIMEX-L, black line) and quadratic trend (SIMEX-Q, red line) are fit to the same collection of five points. Given the assumptions surrounding the measurement error (nondifferential, classical, constant variance), both choices would seem to adjust estimates in the expected direction. That is, where the effect of measurement error on a naive analysis is one of attenuation bias (estimates

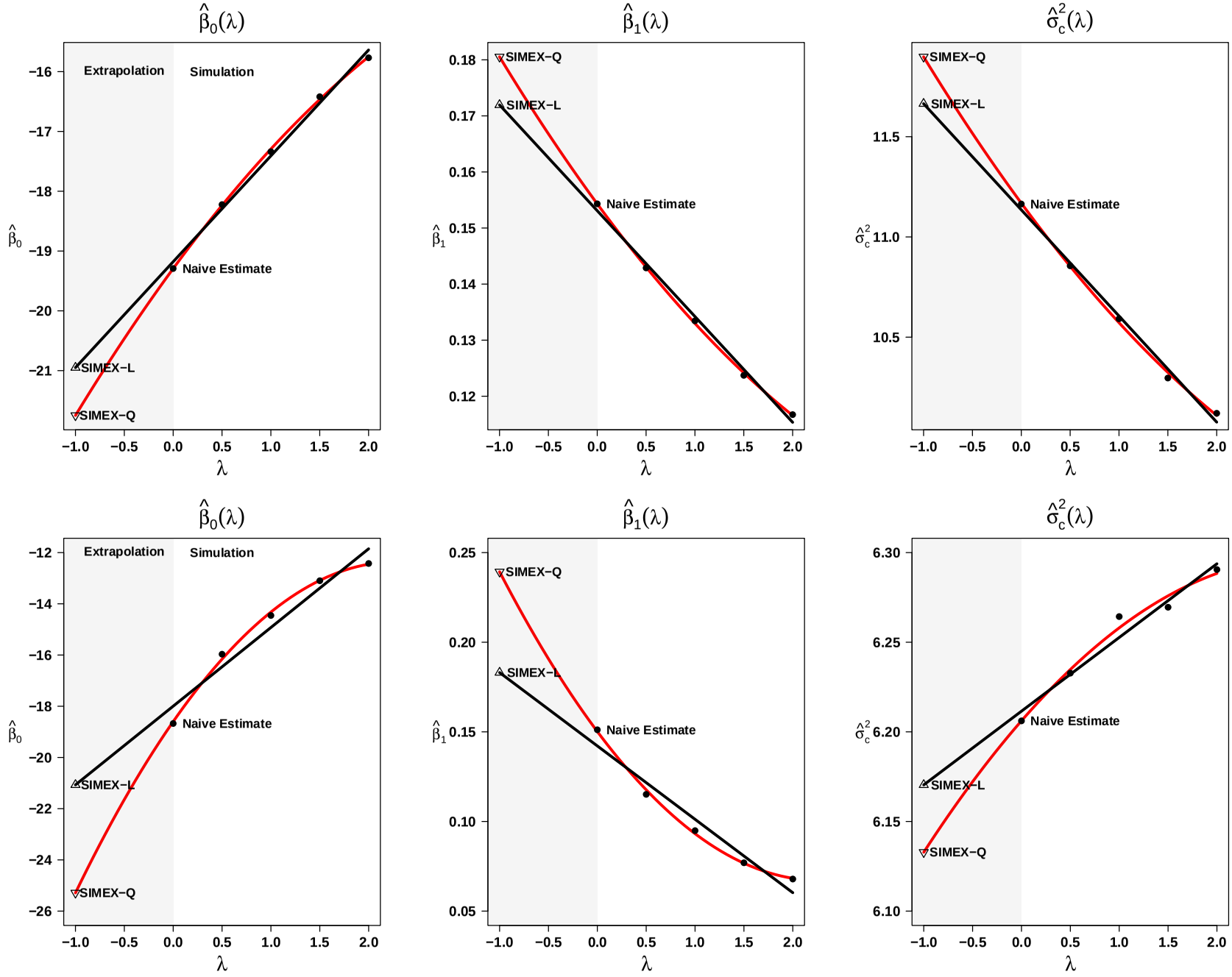


Figure 6: Linear and Quadratic SIMEX plots for model parameters of the random-intercept logistic model for the WSPR (top row) and QSF18 (bottom row) single-event data

of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ pulled closer to zero) the SIMEX-adjusted estimates are moved further away from zero. In the WSPR data set, the linear and quadratic extrapolant relationships shows only modest departures from one another, with the quadratic extrapolant producing larger adjustments in intercept and slope coefficients. By contrast, the QSF18 data shows a much stronger quadratic fit, and the resulting SIMEX-L and SIMEX-Q estimates of intercept and slope are further apart from one another.

Table 3 contains SIMEX point estimates and standard errors of coefficients in the subject-specific models, as well as estimates of $\widehat{d}_{50}$ for both single-event data sets. Again, the intercept and slope estimates are adjusted away from zero. The standard errors were computed by jackknife variance routine described in (Carroll et al., 2006, Appendix B.4.1); note that the standard errors also increase relative to the standard errors of the naive models. Where naive analysis suggests that 50% of both populations would become highly annoyed around 124-125 dB, the SIMEX-Q estimates suggest that the two communities differ in their respective tolerances, with the Galveston community approaching 50% annoyance at 105 dB. Population average dose response curves, produced using integration a la Pavlou et al., are shown in Figure 7. Both panels point to the likely downward bias of the naive population average curves at the highest end of the dose range. After accounting for measurement error at both test sites, what seemed to be a similar response across both populations is shown to differ. Even within the range of doses observed at Galveston ($\leq$90 dB), the probability of annoyance becomes greater than that of the Edwards Air Force Base community. Comparing the SIMEX-Q curves across both panels, the difference in predicted response becomes even more apparent, perhaps highlighting habituation of a community at an Air Force base versus a population seldom exposed to sonic boom noise of any kind.

Table 3: Naive and SIMEX estimates from WSPR and QSF18 studies

Data	Estimator	$\widehat{\beta}_0$ (SE)	$\widehat{\beta}_1$ (SE)	$\widehat{\sigma}_c$	$\widehat{d}_{50}$ (dB)
WSPR	Naive (AGHQ)	-19.29 (1.70)	0.154 (0.015)	3.34	125.3
WSPR	SIMEX-L	-20.95 (1.80)	0.172 (0.016)	3.42	121.8
WSPR	SIMEX-Q	-21.75 (1.86)	0.181 (0.017)	3.45	120.2
QSF18	Naive (AGHQ)	-18.67 (2.40)	0.151 (0.028)	2.49	123.6
QSF18	SIMEX-L	-21.07 (2.63)	0.183 (0.031)	2.48	115.1
QSF18	SIMEX-Q	-25.29 (3.02)	0.239 (0.037)	2.48	105.8

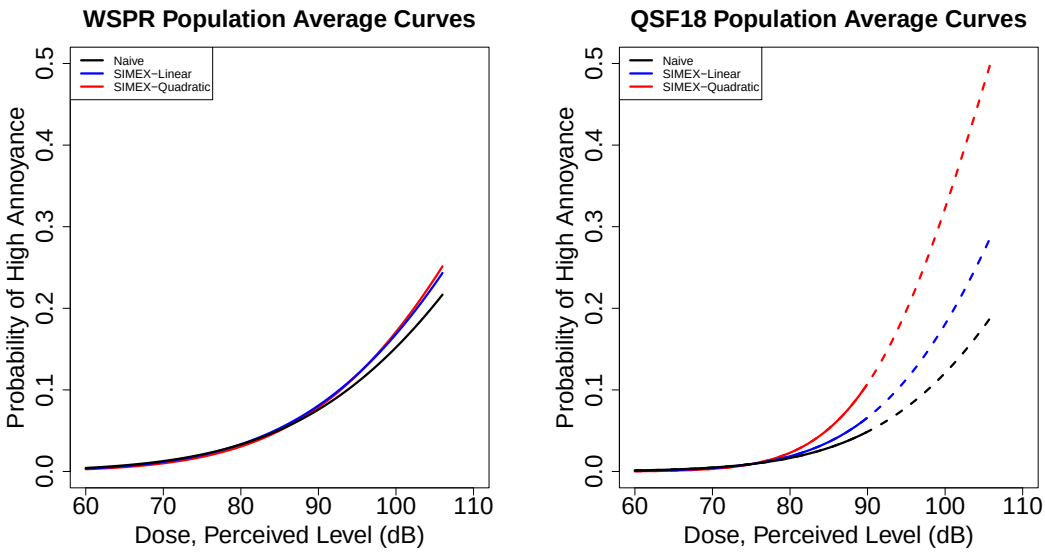


Figure 7: Comparison of population average curves at WSPR and QSF18 test sites under naive analysis and SIMEX adjusted analysis

1.4.3 Structural Bayesian Measurement Error Models

Structural Bayesian measurement error models were discussed in Section 1.3.2. In particular, the structural measurement error model requires an explicit assumption about the so-called exposure model, in this context, a probability distribution describing the nature of 'true dose' in the community noise studies. Retaining assumptions of classical measurement error ($w_{ij}|x_{ij} \sim N(x_{ij}, \widehat{\sigma}_u)$) and the same assumed Cauchy family priors from Section 1.4.1 ($\beta_0 \sim t(0, 10, 1)$, $\beta_1 \sim t(0, 2.5, 1)$, $\sigma_c \sim t^+(0, 1, 1)$), we focus on varying the exposure model, keeping other model features fixed. A variety of assumptions and insights may be used to develop an exposure model, and we note this as an apparent source of sensitivity as applied to past risk reduction study data.

From Figure 3, a feature of the WSPR dose data, and indeed, the intended test design, was bimodality, or combination of low-amplitude sonic booms captured with planned and unplanned traditional sonic booms. Four choices of exposure model are considered, each completing a different Bayesian model.

1. $Uniform(-100, 200)$: As assumed in Doebler et al. (2021), this choice of distribution spans a large, finite support with equal probability density for intervals of the same size within the support. It imparts little information about actual levels that may have been more frequently tested, and it assigns positive probability density to negative dose values.

2. $\frac{84}{110}t_1 + \frac{26}{110}t_2$ where $t_1 = t(\mu_1 = 83, \sigma_1 = 7, \nu = 4)$ and $t_2 = t(\mu_2 = 105, \sigma_2 = 6, \nu = 4)$: This mixture of t distributions uses some knowledge of test protocols, specifically the known ratios of low-amplitude sonic booms and traditional sonic booms as captured during testing. The component t_1 is centered near the empirical mean of *planned* noise doses during testing. A literature value of 105 PL dB is assumed for the mean of the component t_2 Doebler & Rathsam (2019), with assumed values for the scale parameter and degrees of freedom that govern the spread and tail-area behavior of the distribution.

3. $\frac{84}{110}N_1 + \frac{26}{110}N_2$ where $N_1 = N(\mu_1 = 83, \sigma_1 = 8)$ and $N_2 = N(\mu_2 = 97, \sigma_2 = 9)$: Like the

previous assumption, the mixing probabilities are determined by numbers of planned low-amplitude sonic boom maneuvers versus planned and adventitious sonic booms. In this case, empirical means and standard deviations from as-measured doses in the planned and adventitious booms inform the means and standard deviations of the normal mixture components.

4. $t(\mu = 85, \sigma = 10, \nu = 4)$: This choice eschews knowledge of the mixture of low-amplitude and traditional sonic booms. Instead, this unimodal distribution is centered at the empirical mean taken across all doses as-measured, and it is given an increased spread with heavy tails, so that loud events would be expected with appreciable probability. Specifically, this distribution embodies a 32% chance that a true dose exceeds 90 dB, and approximately 6% chance that a true dose exceeds 105 dB.

Table 4 shows the posterior means and posterior standard deviations for parameters obtained for each choice of exposure model. When compared to the naive model, each distinct measurement error model offers adjustment away from zero, and the absolute change in intercept and slope coefficients increases with the application of more informative exposure models. Models 3 and 4 produce results similar to the SIMEX-Q estimator, with $\widehat{d}_{50} \approx 122$ dB.

Table 4: Posterior means and standard deviations from WSPR study

Exposure Model, $f(x_{ij})$	$\widehat{\beta}_0$ (SD)	$\widehat{\beta}_1$ (SD)	$\widehat{\sigma}_c$ (SD)	$\widehat{d}_{50}$ (dB)
Naive (Bayes)	-19.22 (1.69)	0.154 (0.015)	3.43 (0.72)	125.0
1. $Uniform(-100, 200)$	-20.33 (2.02)	0.162 (0.018)	3.64 (0.79)	125.5
2. $\frac{84}{110}t_1 + \frac{26}{110}t_2$	-20.75 (2.01)	0.167 (0.018)	3.67 (0.79)	124.3
3. $\frac{84}{110}N_1 + \frac{26}{110}N_2$	-21.87 (2.14)	0.180 (0.019)	3.65 (0.79)	121.5
4. $t(\mu = 85, \sigma = 10, \nu = 4)$	-21.93 (2.15)	0.180 (0.019)	3.69 (0.79)	121.8

In contrast, Figure 3, shows that unimodality and near symmetry seemed to be a feature of the Galveston test as executed. Test procedures exclusively used the low-amplitude sonic boom maneuver, and the preponderance of events were 'Quiet', with a *maximum* undertrack loudness of 73.7 PL dB. (Page et al., 2020a, Tables 4-2 and 4-3) Below, we consider five choices of exposure model, each fully defining a distinct Bayesian measurement error model for the QSF18 single-event data. Here we experiment specifically with the spread and tail-area behaviors of the exposure model, given some notion of the center of the distribution.

1. $Uniform(-100, 200)$: As before, this choice eschews knowledge of test procedures employed, imparting little information about actual dose levels that may have been tested more frequently. It assigns a 33% chance that doses could be negative, and a 37% chance that an actual dose could exceed 90 dB.

2. $Triangular(35, 70, 105)$: Centered with a mode at 70 PL dB, this distribution has a finite support, meaning that, by assumption, actual doses could never fall below 35 PL dB or exceed 105 PL dB. It assigns approximately 9% chance that an actual dose could exceed 90 dB.

3. $N(\mu = 70, \sigma = 12)$: Similar to assumptions made by Erciulescu & Opsomer (2023), the normal distribution is also centered at 70 dB. Under this assumption, there is less than a one percent chance that true doses fall outside the range 35 to 105 dB, and about a 5% chance that an actual dose could exceed 90 dB.

4. $t(\mu = 70, \sigma = 7, \nu = 4)$: Again centered at 70 dB, this distribution assigns just 2% chance that an actual dose would exceed 90 dB, and, by symmetry, a 2% chance that actual dose could fall below 50 dB.

5. $Triangular(50, 70, 90)$: Since the triangular distribution has finite support, it assigns no probability to outcomes outside the minimum and maximum doses of 50 and 90 dB. It assigns 75% chance that actual doses fall between 60 and 80 dB.

The posterior means and standard deviations from each measurement error model are given in Table 5. Each of the five Bayesian measurement error models adjusts regression coefficients of intercept and slope further away from zero, with biggest changes seemingly related to greater concentration of probability density about the mode at 70 dB. Along with greater absolute change in estimated intercept and slope coefficients, the corresponding posterior standard deviations show a dramatic increase relative to the naive model. Model 3 produces a $\widehat{d}_{50}$ on par with the SIMEX-L estimate, whereas Models 4 and 5 produce outcomes more similar to the SIMEX-Q estimator with $\widehat{d}_{50}$ approaching 108 to 110 dB.

Table 5: Posterior means and standard deviations from QSF18 study

Exposure Model, $f(x_{ij})$	$\widehat{\beta}_0$ (SD)	$\widehat{\beta}_1$ (SD)	$\widehat{\sigma}_c$ (SD)	$\widehat{d}_{50}$ (dB)
Naive (Bayes)	-18.62 (2.40)	0.150 (0.029)	2.50 (0.50)	124.1
1. $Uniform(-100, 200)$	-20.34 (3.00)	0.165 (0.034)	2.69 (0.54)	123.3
2. $Triangular(35, 70, 105)$	-21.53 (3.35)	0.181 (0.038)	2.73 (0.58)	119.0
3. $N(\mu = 70, \sigma = 12)$	-22.29 (3.54)	0.193 (0.041)	2.65 (0.55)	115.5
4. $t(\mu = 70, \sigma = 7, \nu = 4)$	-24.07 (4.07)	0.217 (0.047)	2.76 (0.59)	110.9
5. $Triangular(50, 70, 90)$	-25.64 (4.54)	0.237 (0.053)	2.73 (0.59)	108.2

For each data set, Figure 8 shows one example of a population average curve adjusted for measurement obtained via Wakefield closed-form approximation, which can be used to generate the mean response curve as well as its credible intervals. Once again, the naive population average curves between the two test sites are more similar to one another, and the QSF18 curve shows greater adjustment after accounting for dose measurement error through the chosen exposure model. The shaded area in the right panel indicates dose values greater than 90 dB, the maximum dose measured during the QSF18 study. Note that the mean response of the measurement error model (solid red line) falls outside the upper 90% credible interval at the high end of the QSF18 dose range and beyond. This emphasizes the likely downward bias of the naive model.

1.5 Findings and Recommendations in Relation to Future Community Testing with X-59

We conclude with some considerations for the future community noise studies.

1. Of the four broad methods reviewed, simulation extrapolation and structural hierarchical Bayesian models may be the most readily applicable to Quesst mission community noise study objectives. Regression calibration entails a presumption of gold standard data on which to build the calibration model. Such data may not be obtainable given the novelty of the noise source, and regression calibration may not address biases in the variance components of a subject-specific model. Given its connection to Bayesian computing, data cloning requires the

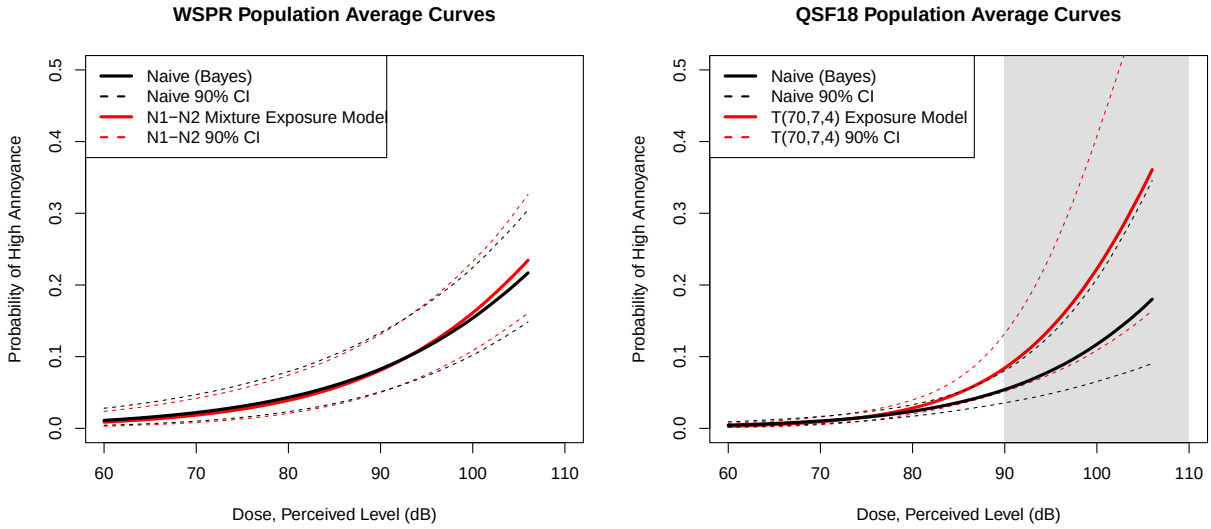


Figure 8: Example population average dose response curves derived by closed-form approximation

same structural assumption of a plausible exposure model (a potential source of sensitivity) while inviting increasing computational burden as the number of clones grows. We do note that some of the computational techniques designed to make data cloning more tractable are portable back to the Bayesian hierarchical model.

2. To date, NASA efforts have focused primarily on the classical measurement error model, but systematic sources of error may be present in the future community noise studies. One anticipated example is the difference between the best outdoor estimates and the levels experienced by indoor listeners. Additionally, scale biases could result if differences in predictive accuracy of the noise estimation procedure vary as a function with the magnitude of the noise level or lateral distance from under track. If warranted, use of the linear measurement error model encapsulates both these random and systematic components of measurement error.

3. Simulation extrapolation (SIMEX) may be ideally suited to the analysis of secondary research questions, which involves interpreting magnitudes of logistic regression coefficients as well as their associated hypothesis tests. SIMEX offers a viable means of mitigating bias and restoring statistical power to hypothesis tests. The SIMEX routine is available in an off the shelf implementation for unweighted logistic regressions in the `simex` R package (Lederer et al., 2022). Furthermore, it can be readily programmed for logistic regression analysis incorporating survey weights.

4. A benefit of the structural Bayesian hierarchical model is that it potentially corrects biases in the subject-specific model and propagates the uncertainty into the population average model. We identified the choice of the exposure model as a potential source of sensitivity in these types of models, and defining spread or tail-area behavior may be particularly difficult without auxiliary knowledge. Because this particular prior knowledge is acoustical in nature, it should be developed in consultation with acoustics subject matter experts. Already planned efforts during the second (acoustic validation) phase of the Quesst mission will be the first chance the NASA team has to observe the phenomenon of the X-59 shaped sonic boom. Data

collected during these flights could be used opportunistically to better inform the exposure models during community tests during the the execution phase of the mission.

2 Annotated Bibliography

The bibliography below was generated upon review of the extensive statistics, epidemiology, econometrics, and acoustics literatures regarding methodological advancements and applications of measurement error corrections. Given the aims of the NASA Quesst mission and the community response testing in its final phase, the focus was narrowed primarily to provide insights into methods pertinent to generalized linear mixed models, e.g., multilevel logistic regression. It is hoped, however, that the references collected, particularly the reviews and textbooks, can be of general benefit whenever potential explanatory variables can only be observed or measured imprecisely.

2.1 Measurement Error Text Books and Review Articles

1. Fuller, W. (1987). *Measurement Error Models*. John Wiley and Sons, Inc. https://doi.org/10.1002/9780470316665

   This is an early and foundational textbook on measurement error, with a major focus on linear regression models with errors in independent variables. Chapter 3 extends the topic to cases of 1) non-normal errors and unequal variances, 2) nonlinear models, and 3) measurement errors that depend on the magnitudes of the true variable.

2. Gustafson, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Chapman & Hall/CRC. https://doi.org/10.1201/9780203502761

   The text offers some in depth description of the impacts of measurement error (Chapter 2) and its cousin in the discrete variables case, called misclassification (Chapter 3). Chapter 4 is of particular interest given the emphasis on Bayesian corrections for logistic regression models, and discussions related to specifying the exposure model, a major source of sensitivity in structural measurement error models. Advanced topics related to model misspecification and computation are presented in Chapter 6 and an Appendix.

3. Carroll, R., Ruppert, D., Stefanski, L., & Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781420010138

   This text focuses on methods for *nonlinear* models, including logistic regression and other forms of generalized linear (and mixed) models. The authors were among some of the most prolific writers on measurement error methodologies, and the textbook consolidates many of their references presented and discussed later below. Particular emphasis is given to the distinction between *functional* methods (Chapters 3-6) which require no specific assumptions about the distribution of the true predictor variable and structural methods (Chapters 7-8) which do require such assumptions. Some particular discussion in Chapter 11 calls into question the utility of regression calibration for longitudinal data collections like the proposed Quesst mission community noise studies.

4. Yi, G. Y., Delaigle, A., & Gustafson, P., editors (2021). *Handbook of Measurement Error Models*. Handbooks of Modern Statistical Methods. Chapman and Hall/CRC. https://doi.org/10.1201/9781315101279

This handbook consolidates chapters on many of the widely known methods, as well as specialty topics. Chapter 1 includes a succinct history of thought and development in the area of measurement error mitigations. Chapters 3, 4, and 5 have relevant discussions on the topic of identifiability. Of interest, there are dedicated chapters on regression calibration (Chapter 7), semiparametric methods (Chapter 9), nonparametric methods (Chapter 14), measurement error for mixed effects models (Chapter 17). Given the relative age of the other references, this reference is one of the few that even alludes to data cloning as a strategy for measurement error applications (Chapter 18, p. 394), noting it as a method to overcome some of the computational challenges, especially the impediments to working with full likelihood methods. Chapter 19 introduces spatial concepts to measurement error models. The final chapter, Chapter 24, focuses on Bayesian approaches.

5. Keogh, R. H., Shaw, P. A., Gustafson, P., Carroll, R. J., Deffner, V., Dodd, K. W., Küchenhoff, H., Tooze, J. A., Wallace, M. P., Kipnis, V., & Freedman, L. S. (2020). STRATOS Guidance Document on Measurement Error and Misclassification of Variables in Observational Epidemiology: Part 1—Basic Theory and Simple Methods of Adjustment. *Statistics in Medicine*, 39(16), 2197–2231. https://doi.org/10.1002/sim.8532

   This article is the first of a two-part review on the topic of measurement error. The first part of the review focuses on descriptions of different types of measurement error (e.g., Berkson, linear, and classical measurement error models) and distinguish between differential and non-differential errors and discussion of impacts on various types of common analyses in epidemiology. The authors provide useful discussion about developing the additional types of studies to obtain information about such errors, and ample thought is given to aspects of study design (including sample size requirements), both for the main study and in subsets of the data. Finally, regression calibration and simulation extrapolation are described as two of the simpler means of dealing with measurement error, and application of each is demonstrated on the Observing Protein and Energy (OPEN) dietary study. Reviews of available software for executing the methods are provided.

6. Shaw, P. A., Gustafson, P., Carroll, R. J., Deffner, V., Dodd, K. W., Keogh, R. H., Kipnis, V., Tooze, J. A., Wallace, M. P., Küchenhoff, H., & Freedman, L. S. (2020). STRATOS Guidance Document on Measurement Error and Misclassification of Variables in Observational Epidemiology: Part 2—More Complex Methods of Adjustment and Advanced Topics. *Statistics in Medicine*, 39(16), 2232–2263. https://doi.org/10.1002/sim.8531

   This article constitutes the second part of the two-part review, written by coauthors with interest in epidemiology. The more advanced methods alluded to include likelihood methods, Bayesian methods, moment reconstruction, moment-adjusted imputation, and multiple imputation. Lists of software are given, and code is available in the supporting information for this article. More complicated error structures are discussed, for example, the case where a predictor is subject to both Berkson and classical errors. The final contribution of this article is advice for the case when there is only external reference, partial, or no external information about the magnitudes of the measurement errors.

7. Sevilimedu, V. & Yu, L. (2022). Simulation extrapolation method for measurement error: A review. *Statistical Methods in Medical Research*, 31(8), 1617–1636. PMID: 35607297. https://doi.org/10.1177/09622802221102619

   The article is a review of two and a half decades of innovation of simulation extrapolation (SIMEX) techniques alone. Many of the innovations and modifications are discussed in section

5 and the references therein. Notably, a technique called empirical SIMEX can potentially be used when measurement error variances are unknown, or perhaps unattainable. The extension can cover the case when measurement error is unknown and even heteroskedastic. Additional SIMEX techniques exist for the case where the measurement error is systematic, i.e., has a mean other than zero.

2.2 Regression Calibration

8. Carroll, R. J. & Stefanski, L. A. (1990). Approximate Quasi-likelihood Estimation in Models with Surrogate Predictors. *Journal of the American Statistical Association*, 85(411), 652–663. https://doi.org/10.1080/01621459.1990.10474925

   This is one of the foundational papers on regression calibration, which can be reduced to a three part process of obtaining 1) a calibration model by regressing a gold standard data (validation data) on the error prone measurements and other variables, 2) replacing the unobserved (true) predictor with its estimate in the standard analysis, and then 3) adjusting standard errors to account for estimation. The authors further developed a taxonomy of likely data structures that would be necessary to identify parameters of the regression calibration models: primary data, internal validation data, internal reliability data, external validation data, and external reliability data. They note that the primary data generally do not identify all model parameters. A key takeaway is the need for additional data to perform this method.

9. Fung, K. Y. & Krewski, D. (1999). Evaluation of Regression Calibration and SIMEX Methods in Logistic Regression When One of the Predictors is Subject to Additive Measurement Error. *Journal of Epidemiology and Biostatistics*, 4(2), 64–74

   The authors developed several simulated data sets, and applied both simulation extrapolation and regression calibration. They considered cases of both classical measurement error and Berkson measurement error. The two methods were compared in terns of bias, mean squared error, and coverage of confidence intervals of the logistic regression estimates. Based on their findings in simulated data, the authors advocate for the use of regression calibration versus simulation extrapolation in all but the case of Berkson error with highly correlated predictor variables. It should be noted that they did not explore the case of logistic models incorporating random effects of any kind.

10. Horonjeff, R. D. (2023). Correcting for Bias Effects Due to Exposure Uncertainty in Community Noise Exposure-Response Analyses. *The Journal of the Acoustical Society of America*, 154(3), 1614–1627. https://doi.org/10.1121/10.0020545

    The author experimented with a variety simulated data and scenarios, varying the sound level uncertainty, the sound level range, and the distribution of sound levels over that range, e.g., the experimental design. The calibration model was able to remove biases in these scenarios, but, correctly, inflate the standard errors of model parameters.

2.3 Simulation Extrapolation (SIMEX)

11. 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. https://doi.org/10.1080/01621459.1994.10476871

    The paper that established the topic, the authors developed a simulation based means of inference for parametric measurement error models. They describe it as a 'method-of-moments'

estimation using Monte Carol estimating equations and establish equivalence to method-of-moments in the linear measurement error model case. In the logistic case, they showed that the resulting estimators were nearly asymptotically unbiased. Variance estimation and some additional theoretical justifications for the method were subsequently developed in papers noted below.

12. Stefanski, L. A. & Cook, J. R. (1995). Simulation-Extrapolation: The Measurement Error Jackknife. *Journal of the American Statistical Association*, 90(432), 1247–1256. http://www.jstor.org/stable/2291515

    The authors established a connection between SIMEX and the jackknife estimation, which is a known useful technique for reducing bias in nonlinear estimators. The major contribution of the paper was a resulting useful variance estimation procedure. Application in a logistic regression setting was demonstrated on the Framingham heart study data set.

13. 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. http://www.jstor.org/stable/2291401

    This work added to the theoretical justifications for some previously observed behaviors in SIMEX estimators. Additionally, it provided the derivation of the asymptotic distribution of SIMEX estimators, giving rise to large-sample variance estimates.

14. Wang, N., Lin, X., Gutierrez, R. G., & Carroll, R. J. (1998). Bias Analysis and SIMEX Approach in Generalized Linear Mixed Measurement Error Models. *Journal of the American Statistical Association*, 93(441), 249–261. https://doi.org/10.1080/01621459.1998.10474106

    The paper focuses on the generalized linear mixed model, and specifically, the linear mixed model; the probit, logistic, and log-linear mixed model for binary responses; and a Poisson mixed models for count data. Under assumptions of 1) additive and 2) normally distributed errors, the authors show that the naive models follow generalized mixed models with altered fixed effects and random effects structures. The assumptions provide a means of characterizing magnitudes of (asymptotic) biases. The authors discuss the applicability of regression calibration and SIMEX methods as corrections, noting inconsistency of the regression calibration approach, i.e., that the sequence of estimates produced when using increasing sample sizes *does not converge in probability* to the true value of the parameter. Regression coefficients for fixed and random effects, and especially the variance components, may still be biased when applying regression calibration to generalized linear mixed models.

15. Vaughn, A. B., Cruze, N. B., Boucher, M. A., & Doebler, W. J. (2024). Dose Error Correction Using Simulation Extrapolation for Modeling Community Noise Dose-Response Relationships. *Proceedings of Meetings on Acoustics*, 54(1), 040002. https://doi.org/10.1121/2.0001938

    The authors conducted an extensive simulation study, specifying dose ranges consistent with the nominal dose range for tests involving the X-59. They assumed logistic dose-response relationships for each of 11 populations, with increasing sensitivity to the noise stimulus. Using the rational linear extrapolant function, they showed correction of parameter estimates in the expected direction, and even that populations with low propensity to be annoyed within the observed dose range benefited from the method. The authors also recognized sensitivity

of the SIMEX approach to the estimated dose uncertainty, pointing to the need for reliable estimates of variance in its execution.

2.4 Bayesian Hierarchical Models

16. Doebler, W. J., Vaughn, A. B., Ballard, K. M., & Rathsam, J. (2021). Simulation and Application of Bayesian Dose Uncertainty Modeling for Low-Boom Community Noise Surveys. *Proceedings of Meetings on Acoustics*, volume 45. https://doi.org/10.1121/2.0001592

    This proceedings paper describes some of NASA's earliest work on measurement error. The paper discusses both Berkson and classical error structures in mixed logistic Bayesian hierarchical models. The paper analyzed both Waveforms and Sonic Boom Perception and Response (WSPR) and Quiet Supersonic Flights 2018 (QSF18) data, the limited risk reduction study available prior to NASA's Quesst mission community tests. Particular distributional assumptions were made on model parameters and a uniform prior with large support was assigned to the 'true' dose predictor, a sound level for the sonic booms as test. Under these specific assumptions, the Berkson and classical measurement error models did not seem to differ greatly from the corresponding naive models or from one another other.

17. Erciulescu, A. & Opsomer, J. (2023). Accounting for Dose Uncertainty in Dose-Response Curve Estimation Uning Hierarchical Bayes Models. *2023 Joint Statistical Meetings*. [Conference Presentation]. https://ntrs.nasa.gov/citations/20230007198

    This conference presentation summarized a portion of work performed in support of the Quesst mission planning stage, again using Bayesian hierarchical models with application to Quiet Supersonic Flights 2018 (QSF18) risk reduction study data. In particular, different distributions were assumed for the true dose predictor, ultimately resulting in larger observed changes in regression coefficients than in Doebler et al. (2021) and presumably in greater reduction of biases present in the naive model.

18. Richardson, S. & Leblond, L. (1997). Some comments on misspecification of priors in bayesian modelling of measurement error problems. *Statistics in Medicine*, 16(2), 203–213. https://doi.org/10.1002/(SICI)1097-0258(19970130)16:2<203::AID-SIM480>3.0.CO;2-T

    This paper speaks to the driving force between the observed differences reported in Doebler et al. (2021) and Erciulescu & Opsomer (2023). The idea of *model misspecification* is examined in the context of Bayesian measurement error adjustments, with a focus on the (mis)specification of the prior distribution for the 'true' predictor. (This corresponds to the distribution of true levels of noise dose in the NASA Quesst mission community testing context.) In a simulated data set, the authors deliberately specify distributions inconsistent with the known data generating process and note the changes in estimates of parameters. Structural methods, like a fully-parametric Bayesian approach, require assumptions about distribution of the true predictor, which may be difficult to formulate reasonably when information on the phenomenon is scant. Some of the concluding remarks pointed toward relaxations of the fully-parametric setup to overcome some of these sensitivities to misspecification.

19. Richardson, S. & Gilks, W. R. (1993). Conditional Independence Models for Epidemiological Studies with Covariate Measurement Error. *Statistics in Medicine*, 12(18), 1703–1722. https://doi.org/10.1002/sim.4780121806

    The paper discusses an approach, or 'recipe', for decomposing the Bayesian hierarchical measurement error model into submodels governing disease in the epidemiological context (or

any outcome in other contexts), a measurement model that describes the nature of likely mismeasurement, and the exposure model, that describes the probabilistic behavior or the 'true' predictor or covariate. In so doing, the thought process makes a variety of measurement error models approachable and accessible through dedicated sampling software. This flexibility and ability to propagate uncertainty is perhaps one of the greatest merits of the Bayesian approach. Discussion in the paper includes graphical analysis of candidate models to be fit and analysis and comparison to other method.

20. Schmid, C. H. & Rosner, B. (1993). A Bayesian Approach to Logistic Regression Models Having Measurement Error Following a Mixture Distribution. *Statistics in Medicine*, 12(12), 1141–1153. https://doi.org/10.1002/sim.4780121204

    The authors turn to Bayesian methods for describing measurement error in the context of logistic regression. *Mixtures* of distributions were implemented as a means of allowing measurement error to change form with the observed exposure, e.g., magnitude of the observed or measured exposure or other delineations. The demonstration on a Nurses Health Study data set of reported alcohol consumption exemplified the need for a mixture, as one group of subjects who truly abstained from drinking accurately reported their consumption, whereas those who did self report drinking may have incorrectly recalled their actual consumption, as determined by the difference in a maintained diet record versus a food frequency questionnaire.

21. de Castro, M., Bolfarine, H., & Galea, M. (2013). Bayesian Inference in Measurement Error Models for Replicated Data. *Environmetrics*, 24(1), 22–30. https://doi.org/10.1002/env.2179

    The importance of replication or repeated measurement in this article is that it gives rise to Bayesian measurement error approaches that do not require explicit knowledge or estimates of unknown error (co)variances. The authors focus on linear models and include homo- and heteroskedastic measurement errors variances as well as the cases of equation error and no equation error, for a total of four models under consideration.

22. Bartlett, J. W. & Keogh, R. H. (2018). Bayesian Correction for Covariate Measurement Error: A Frequentist Evaluation and Comparison with Regression Calibration. *Statistical Methods in Medical Research*, 27(6), 1695–1708. PMID: 27647812. https://doi.org/10.1177/0962280216667764

    The article is somewhat tutorial in nature, as the authors advocate that Bayesian methods for dealing with covariate measurement error are well established and should be more widely adopted by practitioners. They contrast Bayesian hierarchical models specifically with regression calibration in the context of so-called replication studies, i.e., studies that have one or more **replicate** observations of the exposure for some portion of the main study sample. While the authors note the popularity of regression calibration, there are apparent drawbacks. Moderately large biases may remain in nonlinear models, even with large sample sizes. Extensions to non-trivial cases are given where the outcome is a non-linear function of the true covariate or the measurement error model is heteroskedastic. Finally, the regression calibration method does not immediately accommodate uncertainty in the parameters of the measurement model, and therefore, measures of uncertainty often stem from approximate methods. In contrast, the authors argue that the Bayesian approach naturally handles additional sources of uncertainty including measurement error, misclassification, and missing

data, and echo the arguments of Richardson & Gilks (1993) that it facilitates great flexibility to adapt to more complex modeling scenarios.

23. Muff, S., Riebler, A., Held, L., Rue, H., & Saner, P. (2015). Bayesian Analysis of Measurement Error Models Using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 64(2), 231–252. http://www.jstor.org/stable/24771892

    Computation is a recurring theme in the literature for Bayesian corrections to measurement error. In principle, the hierarchical structures outlined in Richardson & Gilks (1993) and elsewhere coupled with software to execute sampling from marginal posterior distributions makes fitting a variety of measurement error models feasible. In practice, sampling-based approaches can be rather compute intensive, and case-specific implementations can be a challenge. The primary contribution of this article is clarifying the means by which common measurement error models, e.g., classical or Berkson models, can be cast as latent Gaussian models, enabling fast approximation of the posterior marginal distributions via Integrated Nested Laplace Approximation (INLA). The focus was on application of both measurement error models in the context of generalized linear (mixed) models. Example R code is available with the online supplemental material.

2.5 Data Cloning

24. Lele, S. R., Dennis, B., & Lutscher, F. (2007). Data Cloning: Easy Maximum Likelihood Estimation for Complex Ecological Models Using Bayesian Markov Chain Monte Carlo Methods. *Ecology Letters*, 10(7), 551–563. https://doi.org/10.1111/j.1461-0248.2007.01047.x

    This paper introduced data cloning as an adaptation of a computational approach with similarities to simulated annealing algorithms. The data cloning procedure described uses the Bayesian modeling framework, but only as a means of obtaining frequentist, maximum likelihood estimates. The 'recipe' involves: 1) constructing a full Bayesian model, with fully specified, proper prior distributions, 2) substituting the likehood function for the data with a likelihood corresponding to k copies of the data (the number of clones), and 3) computing the posterior means and k times the posterior variances, which correspond to the maximum likelihood estimate and asymptotic variance, respectively. The method overcomes several of the challenges of maximum likelihood estimation for hierarchical models, and it is invariant to choices of prior distribution. Several applications to ecological data were demonstrated.

25. Ponciano, J. M., Taper, M. L., Dennis, B., & Lele, S. R. (2009). Hierarchical Models in Ecology: Confidence Intervals, Hypothesis Testing, and Model Selection Using Data Cloning. *Ecology*, 90(2), 356–362. https://doi.org/10.1890/08-0967.1

    Data cloning procedures are subject to two inferential limitations. First, the method produces so-called Wald-type confidence intervals, which may be inaccurate in small sample settings. Second, the procedure doesn't numerically evaluate the maximized likelihood function, necessary for profile-likelihood intervals, likelihood ratio hypothesis tests, and model selection. The authors develop computationally efficient methods for computing likelihood ratios using data cloning, enabling a larger set of common inferential goals.

26. Lele, S. R., Nadeem, K., & Schmuland, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association*, 105(492), 1617–1625. https://doi.org/10.1198/jasa.2010.tm09757

A follow up article to the ecology papers above, this paper focuses on some methodological developments related to diagnostics of 1) convergence of the procedure to the maximum likelihood estimates with a sufficient number of clones, and 2) estimability (identifiability) of model parameters in generalized linear mixed models. The convergence diagnostic is related to degeneracy of the posterior distribution; it involves checking that the largest eigenvalue of the posterior variance matrix is close to zero. Analytical proofs for identifiability are often difficult and rarely attempted, meaning that practitioners often carry out the analysis as though parameters are fully estimable. The graphical estimability diagnostic is important, especially for generalized mixed (and measurement error) models, as it flags the existence of unidentified model parameters.

27. Sólymos, P. (2010). `dclone`: Data Cloning in R. *The R Journal*, 2(2), 29–37. https://doi.org/10.32614/RJ-2010-011

    This article documents a specific implementation of data cloning in the R statistical programming language. The article discusses some applications of data cloning to data analyzed by a generalized linear mixed model, specifically, a Poisson regression for wildlife counts data. In addition to expediting the data cloning procedure, the article discusses the implementation of some important *diagnostics* developed and discussed in Lele et al. (2010). The functionality provided offers a convenient way for those already invested in Bayesian methods to obtain corresponding maximum likelihood estimates for complex models.

28. Torabi, M. (2012). Likelihood Inference in Generalized Linear Mixed Models with Two Components of Dispersion Using Data Cloning. *Computational Statistics & Data Analysis*, 56(12), 4259–4265. https://doi.org/10.1016/j.csda.2012.04.008

    Even in the absence of measurement error, maximum likelihood estimation of parameters of generalized linear mixed models presents a challenging numerical analysis problem, and the addition of a second random effect (variance component) amplifies the computational challenge. The author implements data cloning and compares the results to a variety of competing estimators (hierarchical Bayes, corrected penalized quasilikelihood, quasi-likelihood and method of moments) for fully-pooled and two-component mixed logistic regression model on a well-known salamander mating success data set. Data cloning, which produces the maximum likelihood estimates, produced estimates that were generally more efficient than competing estimators, and it yields a means of predicting random effects, e.g., subject-specific intercepts.

29. Torabi, M. (2013). Likelihood Inference in Generalized Linear Mixed Measurement Error Models. *Computational Statistics & Data Analysis*, 57(1), 549–557. https://doi.org/10.1016/j.csda.2012.07.018

    This may be one of the few available applications of data cloning specifically on a measurement error problem in the generalized linear mixed models setting currently available in the literature. Frequentist computation for GLMM, even in the absence of measurement error, can be quite difficult. The utility of data cloning was demonstrated on real data sets for mixed linear and mixed logistic regression models. In the linear mixed model applied to National Cancer Institute's OPEN Study, data cloning outperformed a corresponding hierarchical Bayesian measurement error model in terms of delivering greater precision; the relative efficiency (hierarchical Bayes relative to maximum likelihood via data cloning) ranged from 100% to 219%. In an analysis of fully-pooled logistic regression applied to the famed Framing-

ham Heart Study, the author compared data cloning to hierarchical Bayes models, regression calibration, and simulation extrapolation noting several advantages.

30. Baghishani, H., Rue, H., & Mohammadzadeh, M. (2012). On a Hybrid Data Cloning Method and Its Application in Generalized Linear Mixed Models. *Statistics and Computing*, 22, 597–613. https://doi.org/10.1007/s11222-011-9254-z

    In this article, the term 'hybrid' data cloning refers to the synthesis of data cloning with a computational approach known as Integrated Nested Laplace Approximation (INLA). INLA provides fast, accurate approximation of posterior distributions for a particular (broad) class of model known as the latent Gaussian model. Since data cloning requires an increasing number of clones to assure convergence, some of the appeal of the method, e.g., invariance to choice of prior distributions, may be overtaken by the increasing computational burden. The primary contribution of this article is one of computational expedience, with asymptotic theory developed for GLMMs and demonstrated on both simulated and real data.

31. Picchini, U. & Anderson, R. (2017). Approximate Maximum Likelihood Estimation Using Data-Cloning ABC. *Computational Statistics & Data Analysis*, 105, 166–183. https://doi.org/10.1016/j.csda.2016.08.006

    In a similar vein to Baghishani et al. (2012), the article investigates more computationally expedient (approximate) data cloning through approximate Bayesian computation (ABC). Approximate Bayesian computing is often used when the corresponding likelihood functions are intractable or cannot be expressed in closed form, e.g., stochastic differential equations, state-space models, and g-and-k distributions. Data cloning is combined with an ABC-MCMC sampler to execute the estimation procedure on larger intractable data sets of these types.

2.6 Additional References

32. Lee, J., Rathsam, J., & Wilson, A. (2020). Bayesian Statistical Models for Community Annoyance Survey Data. *The Journal of the Acoustical Society of America*, 147(4), 2222–2234. https://doi.org/10.1121/10.0001021

    This article has been a starting point in the naive analysis of QSF18 data, employing Bayesian hierarchical models to fit two generalized linear mixed models: the random intercept logistic regression, and the random intercept ordinal regression of QSF18 risk reduction study data.

33. Vaughn, A. B., Rathsam, J., Doebler, W. J., & Ballard, K. M. (2022). Comparison of two statistical models for low boom dose-response relationships with correlated responses. *Proceedings of Meetings on Acoustics*, volume 45. https://doi.org/10.1121/2.0001541

    This work by NASA survey team researchers performed naive analysis comparing multilevel logistic regression and generalized estimating equations to produce population average dose response curves.

34. Lau, Y. T. A. & Yan, J. (2022). Bias Analysis of Generalized Estimating Equations Under Measurement Error and Practical Bias Correction. *Stat*, 11(1), e418. https://doi.org/10.1002/sta4.418

    As one of the alternatives to GLMM, generalized estimating equations (GEE) requires specification of a working correlation structure. In the presence of covariate measurement error, the biases and mean squared errors may be greater under correct specification of the working correlation than under a working independence assumption. The authors propose a functional

bias correction approaches suitable for large samples sizes, and make further adjustments for the case of small sample sizes. The methods for these measurement error corrections are available in the `eiv` (errors-in-variables) package in the R programming language.

35. Doebler, W., Vaugh, A., Cruze, N., Ballard, K., Rathsam, J., & Parker, P. (2022). Effects of Dose Error and Sample Size on Sonic Boom Dose-Response Curves. *Journal of the Acoustical Society of America*, volume 152. Conference presentation and abstract; proceedings paper forthcoming. https://doi.org/10.1121/10.0015769

    The authors experimented with varying sample sizes and severity of of dose uncertainty in simulated data sets, noting the effects of sampling error were distinct from the loss of accuracy due to dose error (covariate measurement error), which couldn't be ameliorated with larger sample sizes. Subjecting the same simulated data sets to increasing degrees of dose uncertainty, the authors observed proportional relationships in the change. This observation is related to the choice of extrapolant function in SIMEX.

36. Doebler, W., Ballard, K., Vaughn, A., & Parker, P. (2023). Dose Error Impacts on a Collection of Realistic Dose-Response Curves Based on a NASA Sonic Boom Community Noise Survey. *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 268. https://doi.org/10.3397/IN_2023_0621

    This work examined the sensitivity of dose response curves to dose error in the context of many simulated populations with different onset and rate of annoyance. The authors noted up to a 14 dB difference in dose eliciting high rates of annoyance, attributable to measurement error, depending on the specifics of the simulated population and the severity of the dose error.

37. Stefanski, L. A. & Carroll, R. J. (1985). Covariate Measurement Error in Logistic Regression. *The Annals of Statistics*, 13(4), 1335 – 1351. https://doi.org/10.1214/aos/1176349741

    This is among the early references about the effects of covariate measurement error on logistic regression. Specifically, the paper characterizes the asymptotic bias of naive logistic regression estimators and the tendency to both overpredict response at low ends of the covariate range and underpredict predict response at the high end of the covariate range. Three alternative estimators are developed, including one estimator that makes only assumptions about moments of the distribution of measurement errors, and two alternatives based on assumptions of normally distributed measurement errors.

38. Apanasovich, T. V., Carroll, R. J., & Maity, A. (2009). SIMEX and Standard Error Estimation in Semiparametric Measurement Error Models. *Electronic Journal of Statistics*, 3(none), 318 – 348. https://doi.org/10.1214/08-EJS341

    Many of the approaches discussed proceed from modeling the mismeasured variable fully parametrically. This article provides a generalization that enables application of SIMEX that encompasses models for mismeasured variables that are fully nonparametric, fully parametric, or have some components that are modeled both parametrically and nonparametrically. A novel standard error estimator is proposed.

39. Alexeeff, S. E., Carroll, R. J., & Coull, B. (2016). Spatial Measurement Error and Correction by Spatial SIMEX in Linear Regression Models when Using Predicted Air Pollution Exposures. *Biostatistics*, 17(2), 377–389. https://doi.org/10.1093/biostatistics/kxv048

This article considers the case of linear health effects models where one of the predictor variables (air pollution) is the product of a spatial model that may be subject to estimation error as well as model misspecification. Both the these cause bias, and the latter also induces asymptotic biases on a slope coefficient. The spatial SIMEX models were developed to correct against both. The application included a study relating effects of air pollution on birth weights in Massachusetts. Given that the prevailing thinking for estimating dose involves the uses of a spatial model, the method may have potential application during the Quesst mission.

40. Baghishani, H. & Mohammadzadeh, M. (2011). A Data Cloning Algorithm for Computing Maximum Likelihood Estimates in Spatial Generalized Linear Mixed Models. *Computational Statistics & Data Analysis*, 55(4), 1748–1759. https://doi.org/10.1016/j.csda.2010.11.004

    While not a measurement error topic, per se, this reference couples the appeal of data cloning with spatial analysis of a generalized linear mixed model, which may have applicability in the X-59 community tests as the loudness generally varies spatially with lateral distance away from undertrack.

References

Alexeeff, S. E., Carroll, R. J., & Coull, B. (2016). Spatial Measurement Error and Correction by Spatial SIMEX in Linear Regression Models when Using Predicted Air Pollution Exposures. *Biostatistics*, 17(2), 377–389. https://doi.org/10.1093/biostatistics/kxv048

Apanasovich, T. V., Carroll, R. J., & Maity, A. (2009). SIMEX and Standard Error Estimation in Semiparametric Measurement Error Models. *Electronic Journal of Statistics*, 3(none), 318 – 348. https://doi.org/10.1214/08-EJS341

Baghishani, H. & Mohammadzadeh, M. (2011). A Data Cloning Algorithm for Computing Maximum Likelihood Estimates in Spatial Generalized Linear Mixed Models. *Computational Statistics & Data Analysis*, 55(4), 1748–1759. https://doi.org/10.1016/j.csda.2010.11.004

Baghishani, H., Rue, H., & Mohammadzadeh, M. (2012). On a Hybrid Data Cloning Method and Its Application in Generalized Linear Mixed Models. *Statistics and Computing*, 22, 597–613. https://doi.org/10.1007/s11222-011-9254-z

Bartlett, J. W. & Keogh, R. H. (2018). Bayesian Correction for Covariate Measurement Error: A Frequentist Evaluation and Comparison with Regression Calibration. *Statistical Methods in Medical Research*, 27(6), 1695–1708. PMID: 27647812. https://doi.org/10.1177/0962280216667764

Carroll, R., Ruppert, D., Stefanski, L., & Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781420010138

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. http://www.jstor.org/stable/2291401

Carroll, R. J. & Stefanski, L. A. (1990). Approximate Quasi-likelihood Estimation in Models with Surrogate Predictors. *Journal of the American Statistical Association*, 85(411), 652–663. https://doi.org/10.1080/01621459.1990.10474925

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. https://doi.org/10.1080/01621459.1994.10476871

Cruze, N. B., Ballard, K. M., Vaughn, A. B., Doebler, W. J., Rathsam, J., & Parker, P. A. (2022). Comparison of Likelihood Methods for Generalized Linear Mixed Models with Application to Quiet Supersonic Flights 2018 Data. Technical Memorandum, NASA Langley Research Center. NASA/TM–20220014998. https://ntrs.nasa.gov/citations/20220014998

de Castro, M., Bolfarine, H., & Galea, M. (2013). Bayesian Inference in Measurement Error Models for Replicated Data. *Environmetrics*, 24(1), 22–30. https://doi.org/10.1002/env.2179

Doebler, W., Ballard, K., Vaughn, A., & Parker, P. (2023). Dose Error Impacts on a Collection of Realistic Dose-Response Curves Based on a NASA Sonic Boom Community Noise Survey. *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 268. https://doi.org/10.3397/IN_2023_0621

Doebler, W., Vaugh, A., Cruze, N., Ballard, K., Rathsam, J., & Parker, P. (2022). Effects of Dose Error and Sample Size on Sonic Boom Dose-Response Curves. *Journal of the Acoustical Society of America*, volume 152. Conference presentation and abstract; proceedings paper forthcoming. https://doi.org/10.1121/10.0015769

Doebler, W. J. & Rathsam, J. (2019). How Loud is X-59's Shaped Sonic Boom? *Proceedings of Meetings on Acoustics*, volume 36. https://doi.org/10.1121/2.0001265

Doebler, W. J., Vaughn, A. B., Ballard, K. M., & Rathsam, J. (2021). Simulation and Application of Bayesian Dose Uncertainty Modeling for Low-Boom Community Noise Surveys. *Proceedings of Meetings on Acoustics*, volume 45. https://doi.org/10.1121/2.0001592

Erciulescu, A. & Opsomer, J. (2023). Accounting for Dose Uncertainty in Dose-Response Curve Estimation Uning Hierarchical Bayes Models. *2023 Joint Statistical Meetings*. [Conference Presentation]. https://ntrs.nasa.gov/citations/20230007198

Fidell, S., Horonjeff, R., Tabachnick, B., & Clark, S. (2020). Independent Analyses of Galveston QSF18. Contractor Report, NASA Langley Research Center. NASA/CR–20205005471. https://ntrs.nasa.gov/citations/20205005471

Fuller, W. (1987). *Measurement Error Models*. John Wiley and Sons, Inc. https://doi.org/10.1002/9780470316665

Fung, K. Y. & Krewski, D. (1999). Evaluation of Regression Calibration and SIMEX Methods in Logistic Regression When One of the Predictors is Subject to Additive Measurement Error. *Journal of Epidemiology and Biostatistics*, 4(2), 64–74.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models. *The Annals of Applied Statistics*, 2(4), 1360 – 1383. https://doi.org/10.1214/08-AOAS191

Gustafson, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Chapman & Hall/CRC. https://doi.org/10.1201/9780203502761

Hedeker, D., du Toit, S. H. C., Demirtas, H., & Gibbons, R. D. (2018). A Note on Marginalization of Regression Parameters from Mixed Models of Binary Outcomes. *Biometrics*, 74(1), 354–361. https://doi.org/10.1111/biom.12707

Horonjeff, R. D. (2023). Correcting for Bias Effects Due to Exposure Uncertainty in Community Noise Exposure-Response Analyses. *The Journal of the Acoustical Society of America*, 154(3), 1614–1627. https://doi.org/10.1121/10.0020545

Keogh, R. H., Shaw, P. A., Gustafson, P., Carroll, R. J., Deffner, V., Dodd, K. W., Küchenhoff, H., Tooze, J. A., Wallace, M. P., Kipnis, V., & Freedman, L. S. (2020). STRATOS Guidance Document on Measurement Error and Misclassification of Variables in Observational Epidemiology: Part 1—Basic Theory and Simple Methods of Adjustment. *Statistics in Medicine*, 39(16), 2197–2231. https://doi.org/10.1002/sim.8532

Lau, Y. T. A. & Yan, J. (2022). Bias Analysis of Generalized Estimating Equations Under Measurement Error and Practical Bias Correction. *Stat*, 11(1), e418. https://doi.org/10.1002/sta4.418

Lederer, W., Seibold, H., & Küchenhoff, H. (2022). *Package 'simex': SIMEX- and MCSIMEX-Algorithm for Measurement Error Models.* https://cran.r-project.org/web/packages/simex/simex.pdf

Lee, J., Rathsam, J., & Wilson, A. (2020). Bayesian Statistical Models for Community Annoyance Survey Data. *The Journal of the Acoustical Society of America*, 147(4), 2222–2234. https://doi.org/10.1121/10.0001021

Lele, S. R., Dennis, B., & Lutscher, F. (2007). Data Cloning: Easy Maximum Likelihood Estimation for Complex Ecological Models Using Bayesian Markov Chain Monte Carlo Methods. *Ecology Letters*, 10(7), 551–563. https://doi.org/10.1111/j.1461-0248.2007.01047.x

Lele, S. R., Nadeem, K., & Schmuland, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association*, 105(492), 1617–1625. https://doi.org/10.1198/jasa.2010.tm09757

Muff, S., Riebler, A., Held, L., Rue, H., & Saner, P. (2015). Bayesian Analysis of Measurement Error Models Using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 64(2), 231–252. http://www.jstor.org/stable/24771892

Opsomer, J., Wivagg, J., Jodts, E., Ferg, R., Erciulescu, A., Monlina Manrique, I., Uppala, M., Stollery, P., Lympany, S., & Page, J. (2024). Deliverable SM-02 Survey Analysis Plan Updated Final. Technical report, HMMH Team for X-59 Community Response Testing. HMMH Report No. 312430.001.004.

Page, J. A., Hodgdon, K. K., Hunte, R. P., Davis, D. E., Gaugler, T., Downs, R., Cowart, R. A., Maglieri, D. J., Hobbs, C., Baker, G., Collmar, M., Bradley, K. A., Sonak, B., Crom, D., & Cutler, C. (2020a). Quiet Supersonic Flights 2018 (QSF18) Test: Galveston, Texas Risk Reduction for Future Community Testing with a Low-Boom Flight Demonstration Vehicle. Contractor Report, NASA Langley Research Center. NASA/CR-2020-220589/Volume I. https://ntrs.nasa.gov/citations/20200003223

Page, J. A., Hodgdon, K. K., Hunte, R. P., Davis, D. E., Gaugler, T., Downs, R., Cowart, R. A., Maglieri, D. J., Hobbs, C., Baker, G., Collmar, M., Bradley, K. A., Sonak, B., Crom, D., & Cutler,

C. (2020b). Quiet Supersonic Flights 2018 (QSF18) Test: Galveston, Texas Risk Reduction for Future Community Testing with a Low-Boom Flight Demonstration Vehicle. Contractor Report, NASA Langley Research Center. NASA/CR-2020-220589/Appendices/Volume II. https://ntrs.nasa.gov/citations/20200003224

Page, J. A., Hodgdon, K. K., Krecker, P., Cowart, R., Hobbs, C., Wilmer, C., Koening, C., Holmes, T., Gaugler, T., Shumway, D. L., Rosenberger, J. L., & Philips, D. (2014). Waveforms and Sonic Boom Perception and Response (WSPR): Low-Boom Community Response Program Pilot Test Design, Execution, and Analysis. Contractor Report, NASA Langley Research Center. NASA/CR-2014-218180. https://ntrs.nasa.gov/citations/20140002785

Pavlou, M., Ambler, G., Seaman, S., & Omar, R. Z. (2015). A Note on Obtaining Correct Marginal Predictions from a Random Intercepts Model for Binary Outcomes. *BMC Medical Research Methodology*, 15(59). https://doi.org/10.1186/s12874-015-0046-6

Picchini, U. & Anderson, R. (2017). Approximate Maximum Likelihood Estimation Using Data-Cloning ABC. *Computational Statistics & Data Analysis*, 105, 166–183. https://doi.org/10.1016/j.csda.2016.08.006

Polson, N. G. & Scott, J. G. (2012). On the Half-Cauchy Prior for a Global Scale Parameter. *Bayesian Analysis*, 7(4), 887 – 902. https://doi.org/10.1214/12-BA730

Ponciano, J. M., Taper, M. L., Dennis, B., & Lele, S. R. (2009). Hierarchical Models in Ecology: Confidence Intervals, Hypothesis Testing, and Model Selection Using Data Cloning. *Ecology*, 90(2), 356–362. https://doi.org/10.1890/08-0967.1

Prentice, R. (1982). Covariate Measurement Errors and Parameter Estimation in a Failure Time Regression Model. *Biometrika*, 69(2), 331–342. https://doi.org/10.1093/biomet/69.2.331

Richardson, S. & Gilks, W. R. (1993). Conditional Independence Models for Epidemiological Studies with Covariate Measurement Error. *Statistics in Medicine*, 12(18), 1703–1722. https://doi.org/10.1002/sim.4780121806

Richardson, S. & Leblond, L. (1997). Some comments on misspecification of priors in bayesian modelling of measurement error problems. *Statistics in Medicine*, 16(2), 203–213. https://doi.org/10.1002/(SICI)1097-0258(19970130)16:2<203::AID-SIM480>3.0.CO;2-T

Schmid, C. H. & Rosner, B. (1993). A Bayesian Approach to Logistic Regression Models Having Measurement Error Following a Mixture Distribution. *Statistics in Medicine*, 12(12), 1141–1153. https://doi.org/10.1002/sim.4780121204

Sevilimedu, V. & Yu, L. (2022). Simulation extrapolation method for measurement error: A review. *Statistical Methods in Medical Research*, 31(8), 1617–1636. PMID: 35607297. https://doi.org/10.1177/09622802221102619

Shaw, P. A., Gustafson, P., Carroll, R. J., Deffner, V., Dodd, K. W., Keogh, R. H., Kipnis, V., Tooze, J. A., Wallace, M. P., Küchenhoff, H., & Freedman, L. S. (2020). STRATOS Guidance Document on Measurement Error and Misclassification of Variables in Observational Epidemiology: Part 2—More Complex Methods of Adjustment and Advanced Topics. *Statistics in Medicine*, 39(16), 2232–2263. https://doi.org/10.1002/sim.8531

Sólymos, P. (2010). `dclone`: Data Cloning in R. *The R Journal*, 2(2), 29–37. https://doi.org/10.32614/RJ-2010-011

Stefanski, L. A. & Carroll, R. J. (1985). Covariate Measurement Error in Logistic Regression. *The Annals of Statistics*, 13(4), 1335 – 1351. https://doi.org/10.1214/aos/1176349741

Stefanski, L. A. & Cook, J. R. (1995). Simulation-Extrapolation: The Measurement Error Jackknife. *Journal of the American Statistical Association*, 90(432), 1247–1256. http://www.jstor.org/stable/2291515

Torabi, M. (2012). Likelihood Inference in Generalized Linear Mixed Models with Two Components of Dispersion Using Data Cloning. *Computational Statistics & Data Analysis*, 56(12), 4259–4265. https://doi.org/10.1016/j.csda.2012.04.008

Torabi, M. (2013). Likelihood Inference in Generalized Linear Mixed Measurement Error Models. *Computational Statistics & Data Analysis*, 57(1), 549–557. https://doi.org/10.1016/j.csda.2012.07.018

Vaughn, A. B., Cruze, N. B., Boucher, M. A., & Doebler, W. J. (2024). Dose Error Correction Using Simulation Extrapolation for Modeling Community Noise Dose-Response Relationships. *Proceedings of Meetings on Acoustics*, 54(1), 040002. https://doi.org/10.1121/2.0001938

Vaughn, A. B., Rathsam, J., Doebler, W. J., & Ballard, K. M. (2022). Comparison of two statistical models for low boom dose-response relationships with correlated responses. *Proceedings of Meetings on Acoustics*, volume 45. https://doi.org/10.1121/2.0001541

Wakefield, J. (2013). *Bayesian and Frequentist Regression Methods*. Springer Series in Statistics. Springer. https://link.springer.com/book/10.1007/978-1-4419-0925-1

Wang, N., Lin, X., Gutierrez, R. G., & Carroll, R. J. (1998). Bias Analysis and SIMEX Approach in Generalized Linear Mixed Measurement Error Models. *Journal of the American Statistical Association*, 93(441), 249–261. https://doi.org/10.1080/01621459.1998.10474106

Yi, G. Y., Delaigle, A., & Gustafson, P., editors (2021). *Handbook of Measurement Error Models*. Handbooks of Modern Statistical Methods. Chapman and Hall/CRC. https://doi.org/10.1201/9781315101279