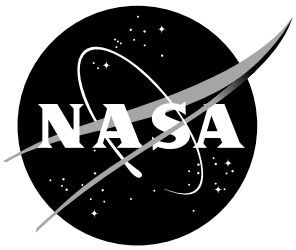
NASA/TM–2019–220427



Statistical Modeling of Quiet Sonic Boom Community Response Survey Data

Jasme Lee
National Institute of Aerospace, Hampton, Virginia

Jonathan Rathsam
NASA Langley Research Center, Hampton, Virginia

Alyson Wilson
North Carolina State University, Raleigh, North Carolina

November 2019

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 NASA Aeronautics and Space Database and its public interface, the NASA Technical Report Server, thus providing one of the largest collection 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:

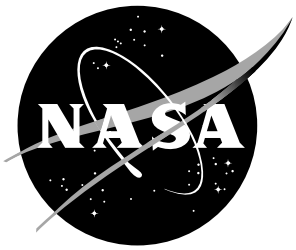
- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers, but having 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–2019–220427



Statistical Modeling of Quiet Sonic Boom Community Response Survey Data

Jasme Lee
National Institute of Aerospace, Hampton, Virginia

Jonathan Rathsam
NASA Langley Research Center, Hampton, Virginia

Alyson Wilson
North Carolina State University, Raleigh, North Carolina

National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia 23681-2199

November 2019

Acknowledgments

The authors would like to acknowledge funding from the Commercial Supersonic Technology Project for providing a Graduate Research Assistantship and post-graduate funding administered via the National Institute of Aerospace. This work was performed at the Structural Acoustics Branch at the NASA Langley Research Center.

Available from:

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

Abstract

The existing ban on commercial supersonic flight overland is largely due to the effects of loud and startling sonic booms on communities. NASA is planning a nationwide campaign of community response surveys using the experimental X-59 Quiet SuperSonic Technology (X-59 QueSST) aircraft to understand how communities perceive the sounds of quiet supersonic flight. The X-59 community response survey data will be presented to noise regulators, who are considering replacing the ban with a noise-based certification limit so quiet supersonic vehicles can fly over land. In this document, we use pilot community response survey data to explore and assess multiple approaches to statistically model the dose-response relationship between single-event sonic boom sound exposure and human annoyance. The models have two primary functions—estimating two types of quantities that support setting regulations and experimental design of future surveys. The dataset is available on the NASA Technical Reports Server as a comma separated values (.csv) file (https://ntrs.nasa.gov/search.jsp?R=20190002702).

Table of Contents

List of Tables

List of Figures

Terminology

- A **statistical model** specifies a data-generating process, which includes randomness.
- **Statistical inference** is used to estimate characteristics of a population using sample data.
- The **frequentist or classical approach** considers the observed data to be random assuming a fixed (but unknown) set of parameters that describe its data-generating mechanism (also called the sampling distribution). Uncertainty comes from sampling variability, which means that if the experiment was performed again, a different set of data would be observed, leading to different inferences about the parameters. Inference about the parameters can be expressed using confidence intervals, which are constructed to work well "on average." If the experiment were conducted many times and consequently many different samples were seen, the confidence interval constructed for each experiment would contain the population parameters (for a 95% confidence interval) 95% of the time.
- The **Bayesian approach** considers the observed data to be fixed and models the fixed (but unknown) set of parameters as random given the data. Uncertainty about the parameters is expressed using a probability distribution. Before the data are observed, this probability distribution is called the **prior distribution**; after the data are observed, the prior distribution is updated using Bayes Theorem to the posterior distribution. The **posterior distribution** represents the uncertainty about the parameters after the data are observed, and inference about the parameters is made using the posterior distribution. For example, we can calculate a 95% credible interval, where l is the lowerbound and u is the upperbound, from the posterior distribution. One common method for calculating the 95% credible interval is to take l to be the 0.025 quantile of the posterior distribution and u to be the 0.975 quantile of the posterior distribution. The credible interval gives a range of values such that the probability that the unknown parameter falls in the range is 0.95.
- **Likelihood** is a function that describes the plausibility of the parameter values given the observed data.
- The **support** of a probability distribution refers to the set of possible values of a random variable.
- **Design matrix** is a matrix of values with rows corresponding to observations, and columns corresponding to each explanatory variable in the model; note that if an intercept term is included in the model, the design matrix will include a column of 1's.
- In regression analysis, **heteroskedasticity** describes non-constant variance in the dependent variable with respect to the independent variable(s) whereas

homoskedaticity describes constant variance across observations. For example, the data are heteroskedastic if the variance of the annoyance ratings increases with respect to the level of the sonic thump. Homoskedasticity is an assumption for a linear regression model.

- A highly correlated sample contains less information than an independent sample of the same size, and the **effective sample size** is used to quantify this. The effective sample size estimates the number of independent samples that the correlated sample is equivalent to.

- Habituation, practice or fatigue are examples of **order effects**, which may cause the survey responses to depend on the order in which sonic thumps were heard in addition to other factors like the acoustic levels.

- **Cross-sectional data** are independent because each participant responds once whereas **longitudinal or panel sample data** are correlated because each participant responds multiple times. Cross-sectional data only consider single responses from one moment in time whereas longitudinal data consider multiple responses over time. For example, a survey collecting cross-sectional data will only require each survey participant to response once, whereas a survey collecting longitudinal data will request each participant to respond multiple times. Longitudinal and cross-sectional data require different modeling techniques.

- **Dose-response relationship** refers to the relationship between the exposure or dosage of a stimuli and a measured response. For example, the stimuli is the sound level of a sonic thump and the response is human annoyance.

- **Noise dose** refers to the sound level of a sonic thump.

Notation

The notation in this document follows common conventions in statistics.

- Uppercase denotes a random variable and lowercase denotes an observed instance of a random variable. For example, Y denotes the random variable for annoyance rating and y denotes the observed annoyance rating.
- The probability distribution function (PDF) for random variable Y is denoted as $f(y)$, and the cumulative distribution function (CDF) is denoted as $F(y)$.
- Expected value of random variable Y is denoted as $E(Y)$, which is an average of all possible outcomes of Y weighted by their respective probabilities.
- Tilde [$\sim$] denotes "distributed as." For example, $Y \sim N(\mu, \sigma^2)$ denotes Y is normally distributed with mean μ and variance σ^2.
- In this document, *log* refers to natural logarithm rather than logarithm of base 10.
- θ denotes a set of parameters.
- $X|Y$ denotes the random variable X conditioned on the value of the random variable Y.
- $f(\theta)$ denotes the prior distribution, $f(y|\theta)$ denotes the likelihood, and $f(\theta|y)$ denotes the posterior distribution.
- $logit(p)$ denotes the the logit function evaluated at p. The logit function is $logit(p) = log(p/(1-p))$. The inverse of the logit function is denoted as $logit^{-1}(x) = exp(x)/(1+exp(x))$.

1 Introduction

Air travel has become a very common mode of transportation. Although there have been many advances in aviation technology in recent years, such as improvement in fuel efficiency, decreases in aircraft emissions, and reductions in aircraft noise, commercial flights have not seen advances in travel speed due to the current ban on commercial supersonic flights overland. Most current commercial aircraft fly at subsonic speeds of about Mach 0.8, whereas supersonic flights at speeds over Mach 1 would reduce travel time.

The first steady, level supersonic flight occurred in 1947 when the Bell X-1 flew faster than the speed of sound. Commercial supersonic flights followed in 1976 with the introduction of the passenger jet Concorde. Flying supersonically meant a reduction in travel time, especially for long distance travel. However, the Concorde had a few problems, one of which was the creation of loud sonic booms when flying supersonically. When traveling faster than the speed of sound, the pressure waves created by the aircraft combine together to form a shock wave. When the shock wave propagates to the ground, it creates a sonic boom. Traditional sonic booms are startling because they are impulsive sounds that come without warning. The sonic boom can typically be heard up to 25 miles on either side of the flight path along the entire supersonic flight route.

Due to the loud sonic booms, the Federal Aviation Administration (FAA) banned commercial supersonic flights overland in 1973 based on aircraft and other transportation community noise surveys conducted in the 1960s. As a result, the Concorde was only allowed to fly supersonically over water. This restricted the flight routes of the Concorde. Since the ban on commercial overland supersonic flights, there has been much research on how to create quieter sonic booms (Maglieri et al., 2014). One method to achieve a quieter sonic boom is to change the shape and design of the aircraft. The National Aeronautics and Space Administration (NASA) and its partners have been pursuing this research and recently began building an experimental aircraft to demonstrate the quiet supersonic technology. NASAs X-59 Quiet SuperSonic Technology X-plane, or X-59 QueSST, is an experimental aircraft with its shape strategically designed to ensure that its sonic boom is quieter and less startling than a traditional sonic boom. Building the X-59 began in late 2018 and is expected to be complete by the early 2020's. The X-59 will not only demonstrate quiet supersonic technology but also be used to conduct surveys of community response to quiet sonic booms. These quieter sonic booms are also referred to as sonic thumps to emphasize that the noise created is softer. Traditional sonic booms sound like two loud bangs whereas a sonic thump would resemble distant thunder.

Seeing the potential to reduce air travel time without loud sonic booms, noise regulators are considering replacing the current ban with a noise-based certification standard. This standard would allow commercial supersonic aircraft to fly overland as long as the sonic booms are below a specified noise limit. NASA is planning to conduct multiple community response surveys using the X-59 to collect data on human perception of quiet sonic thumps. NASA's goal is to build a nationally representative database of community responses to quiet supersonic flight.

In preparation for the X-59 community tests (both in terms of operations and

data analysis), NASA conducted two pilot studies. We use a subset of the data from the first study, which was conducted in 2011 at Edwards Air Force Base (Page et al., 2014). The sonic thumps were simulated with an F-18 dive maneuver, and community members were asked to complete surveys about their perception of the sounds they heard. This was done through rating how annoyed or bothered they were by the events.

The goal of this document is to explore and assess multiple approaches to statistically model the dose-response relationship between sonic boom sound exposure and human annoyance. The models will have two primary functions—estimating two types of quantities that support regulations and experimental design of future surveys. This document is an extension to Jasme Lee's Master's thesis (2019).

1.1 Literature Review

There are many proposed methods in the literature for modeling the dose-response relationship between transportation noise sources and community response. Some examples of the community response are human annoyance or sleep disturbance. A dose-response relationship is of interest because regulators would often like to predict the degree of annoyance associated with certain noise exposure levels. We summarize some of the proposed methods and describe the contributions and shortcomings of each. In addition, we describe how the convention arose for using prevalence of "high annoyance" as the response. Not all proposed methods are statistical models as some are curve-fitting methods, which specify a function to fit through data points, and do not consider the data-generating process and the associated randomness. In addition, most of the methods proposed in the literature are for modeling cross-sectional data rather than longitudinal data, which are the format of our pilot study data. For cross-sectional data, each participant only responds once whereas for longitudinal data, each participant responds multiple times.

In the community noise literature, the first major attempt to model the relationship between noise exposure and annoyance dates back to the Schultz curve (Schultz, 1978). The objective of the Schultz curve was to create one comprehensive dose-response relationship for different transportation noise sources. The Schultz curve is an average of multiple 3rd degree polynomial curves, each fit to data from individual surveys from a community noise survey database. Schultz combined data from different studies with various ordinal annoyance rating scales, and various noise sources, such as aircraft, traffic and railroad noise. Schultz suggested that the average of the 3rd degree polynomial fits to each dataset was the best readily available relationship between community annoyance and transportation noise exposure. Since Schultz used a 3rd degree polynomial fit, extrapolation beyond the observed range of noise doses using the curve may result in predictions that are outside of the reasonable range of 0 to 100%.

Schultz set a precedent for using "high annoyance" as the response in modeling the community annoyance data. He believed high annoyance to be highly correlated with noise exposure and that it would reduce the influence of non-acoustical factors that may have caused lower degree of annoyance. Soon after, in 1982, the United States Environmental Protection Agency (EPA) adopted "percent highly annoyed"

as the impact criterion of noise on communities (U.S. Environmental Protection Agency, 1982). Since the individual studies that Schultz combined did not use the same ordinal response scale, Schultz determined a cutoff for highly annoyed responses for each dataset individually. This led to criticism of the Schultz curve. Later, Fields et al. (2001) recommended that the cutoff for "high annoyance" be 4 or above for a 5-point scale, and 8 or above for an 11-point scale based on their study for internationally compatible community noise surveys with appropriate language, questions, and ordinal scales. They also recommended the use of either a 5-point or 11-point scale in community noise surveys.

While the Schultz curve was an empirical curve-fit to the data, Fidell et al. (1988) proposed to use a theoretically based curve-fitting method. They proposed to fit the curve $p = e^{-A/m}$ to the observed percentages of highly annoyed responses, where m is a transformation of the noise dose based on Stevens' Power Law (Stevens, 1975). This curve takes on an S-shape and constrains the percent highly annoyed between 0 to 100%. Fidell et al. (2011) fit this psychoacoustics-based curve to a database consisting of multiple community noise surveys. Fidell et al. (2011) fit each individual dataset to the exponential relationship, $p(HA) = e^{-A/m}$, to estimate A for each community. They also coined the term "community tolerance level" or CTL as the noise dose corresponding to 50% highly annoyed. The CTL is a metric with units of dB used to describe the variation in annoyance among communities that can be attributed to non-acoustical factors. Fidell et al. (2011) estimated a CTL value for each of the communities in the survey database and fit a linear regression model with percent highly annoyed as the response, and the noise dose and CTL as the covariates. This linear regression model neither takes on an S-shape, nor restricts the response between 0 and 100%. The current limitation on using CTL to characterize a community is that the CTL can only be estimated after a community noise survey is conducted. Little is known about what the relevant non-acoustical covariates are for predicting CTL.

The Federal Interagency Committee on Noise (Federal Interagency Committee on Noise, 1992) suggests a logistic regression model instead of the Schultz curve because it is similar in fit but constrains the predictions between 0 and 100%. Logistic regression is a common model for binary data.

Miedema & Vos (1998) proposed that different curves need to be derived for different transportation noise sources instead of pooling all data. They proposed quadratic polynomial regression and a multilevel variation, which considers the different noise sources as the higher level in the model hierarchy. Both models used percent highly annoyed as the response with noise dose as the covariate. The quadratic regression model suffers from the same drawback as the Schultz curve and the linear regression model proposed by Fidell et al. (2011) in that it does not constrain the percent highly annoyed between 0 and 100%. Miedema & Vos (1998) proposed a method for combining the different ordinal scales by first scaling all responses to between 0 and 100 with equally spaced intervals. This scaling method is carried over to the multilevel interval regression model proposed by Groothuis-Oudshoorn & Miedema (2006), which also models the transportation noise source as the grouping level. One assumption necessary for the proposed model is that the ordinal levels are equally spaced. Groothuis-Oudshoorn & Miedema (2006) suggested a multilevel

model in order to combine data from multiple surveys, but a multilevel model is also a technique that can be used to model longitudinal data as described by Schäffer et al. (2017).

One of the few examples from the community noise literature that addresses modeling longitudinal data is Schäffer et al. (2017), who described two different methods—a marginal model and a multilevel model. Their main goal was to describe how to model repeated binary responses and to compare results from the two models. The difference between the two types of models is how the correlation among multiple responses from each participant is modeled. A marginal model specifies a correlation matrix for the multiple observations from each participant and assumes the same correlation matrix for each participant. A multilevel model uses participant-level parameters, assumed to come from a common distribution, to model the correlation among the multiple responses from each participant. Parameters for the population are estimated on average with a marginal model, whereas parameters for each individual are estimated with a multilevel model (Hu et al., 1998). Schäffer et al. (2017) fit both a marginal and a multilevel model to a few datasets and showed that results from the two models could be very similar or very different depending on the data. Because of this, they suggest determining the goal or research questions before deciding which model to fit. They suggest a multilevel model when the interest lies at the individual level and a marginal model when interest lies in the population on average, and not necessarily individual-specific parameters.

Wilson et al. (2017) described a multilevel model with two levels to quantify the variations in community noise surveys at individual and community levels. The individual level corresponds to variation in responses from different individuals in a community (variation within one community), and the community level corresponds to variation in responses from different communities (variation among multiple communities). The model was fit to cross-sectional data and not longitudinal data. But this can be easily extended to longitudinal data by including a third level in the model hierarchy for multiple responses per individual.

For a recent FAA study of airport noise, Miller et al. (2014) proposed analysis methods for data collected from 20 airports in the United States to establish an updated dose-response relationship between noise exposure and community response. In the proposed analysis plan, they listed the non-multilevel and multilevel variations of the logistic regression and the first-principles based exponential (Fidell et al., 2011) models for modeling the data. They concluded that the multilevel models fit the data better based on visual analysis. But it is unclear how the population representative curves were derived from the multilevel models. Their final dose-response analysis has not been published yet.

1.2 Data

The data we analyze are from a 2011 pilot study conducted at Edwards Air Force Base (Page et al., 2014). The community members were asked to respond to two sets of surveys—cumulative and single-event surveys. Cumulative surveys ask participants to rate their overall annoyance to all sonic booms in one day, and there

were typically multiple booms in a day. The single-event survey asks participants to rate their annoyance to each sonic boom occurrence. We analyze only the single-event survey data, but the methods should generalize to the cumulative survey data as well. Each observation in the data corresponds to a participant's response to the single-event survey. The data are longitudinal because each participant provides multiple responses. There were a total of 110 booms in the entire test, which spanned two weeks. When an observation is missing, it is not clear whether the participant did not hear the boom, was too busy to respond, or had dropped out of the survey.

There were multiple questions on the single-event survey, but our analysis focuses on the responses to the question: "How much did the sonic boom bother, disturb, or annoy you?" The annoyance responses or ratings are ordinal responses on a 0 to 10 scale, where 0 is not at all annoyed and 10 is extremely annoyed. The convention of using an 11-point scale and the descriptions for anchoring the scale ("not all annoyed" and "extremely") are described by Fields et al. (2001). Figure 1.1 shows the distribution of annoyance responses. There are a total of 1981 responses from a total of 49 participants. The data cleaning process and a data dictionary are given in Appendix A. Notice that most of the responses are concentrated on the low end of the scale. Because the impact criterion of noise established by the EPA (U.S. Environmental Protection Agency, 1982) is "percent highly annoyed," we dichotomize the annoyance responses to highly annoyed or not highly annoyed. The recommended cutoff for "high annoyance" on an 11-point scale is 8 or above (Fields et al., 2001). We see in Figure 1.1 that only about 6.7% of the collected responses are highly annoyed responses.

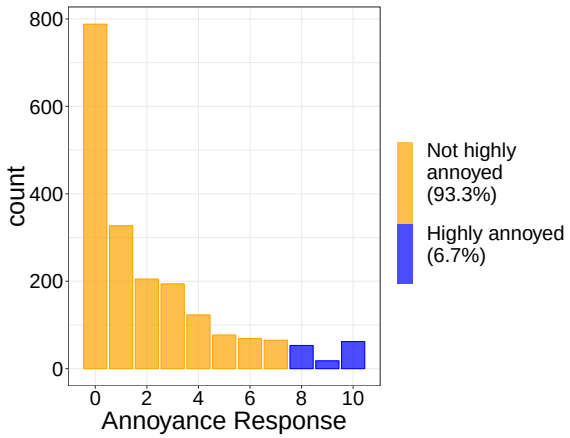


Figure 1.1: Distribution of survey responses.

Each observation has an estimated noise dose. The noise doses are estimated at participants' locations based on measurements from noise monitors at the test site. From the sonic boom waveform, we can calculate multiple metrics for the noise dose. We use the acoustic metric Stevens' Mark VII Perceived Level (PL) because the X-59 design criterion in expressed in PL. In addition, PL is a single-number dB rating that correlates well with human response to impulsive sounds (Rathsam et

al., 2018). We calculate PL from the one-third octave band spectrum of the sonic boom's pressure-time history (Shepherd & Sullivan, 1991; Stevens, 1972).

The PLs observed in the test ranged from 63 to 106 dB, and each estimated noise dose is rounded to the nearest integer because humans generally do not perceive differences less than 1 dB (Yost, 2013). Figure 1.2 shows the distribution of the observed survey responses at each PL.

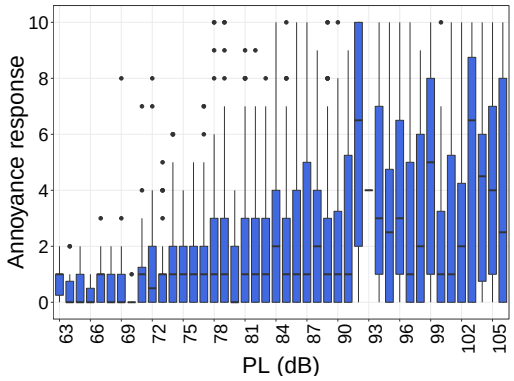


Figure 1.2: Distribution of ordinal responses at each PL.

To emphasize that there are different number of observations at each PL, Figure 1.3 shows the number of responses observed at each PL. There are very few responses at the low PLs (below 66 dB), which may be because a low level sonic boom was inaudible to some participants.

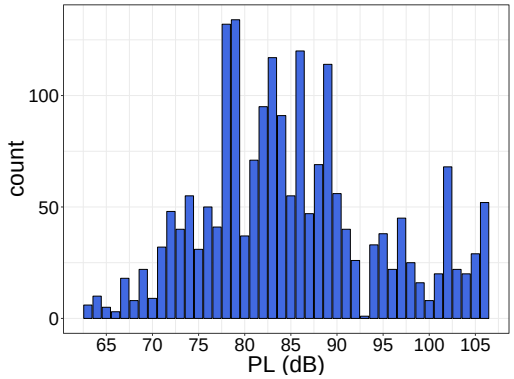


Figure 1.3: Number of observations collected at each PL.

1.3 Research Questions

Before we discuss the relevant research questions, we first consider how to model our pilot study data. Most previous surveys analyzed in the literature are cross-sectional, where each participant responds once, whereas our survey contains longitudinal data, where each participant responds multiple times. The two methods

for modeling longitudinal data described by Schäffer et al. (2017) are taken into consideration when we consider how to model the correlation among multiple responses from each participant. From the longitudinal data, we are interested in modeling average community response across the population. A multilevel model considers the mean response conditional on the individual-level parameters. We can also estimate the marginal mean response by integrating out the individual-level parameters. A marginal model, on the other hand, considers the marginal mean response and we cannot recover the individual-level parameters. We choose to use a multilevel model because for the future X-59 tests, we expect to combine data from multiple communities to estimate a dose-response curve representative of the entire U.S. population. The multilevel model structure we consider in this document models the multiple responses from individuals of one community. This can be easily generalized to multiple communities by adding an additional level in the model hierarchy.

In this document, the dose-response estimate for the pilot study data informs two research goals relevant to the planning of future X-59 community surveys. The first is to investigate the impact of a more limited noise dose range on estimating quantities for setting regulations. The second goal is to calculate sample sizes for planning the X-59 surveys.

The quantity of interest for estimation depends on how noise regulations will be set. One method is to fix the percent highly annoyed at a specific level and estimate the corresponding noise limit. The second is to fix the noise dose and estimate the percent of people highly annoyed. Both methods are plausible so we consider two types of quantities for estimation: the noise dose for a given percent highly annoyed, and the percent highly annoyed for a specified noise dose.

In the 2011 pilot study, the dose range spanned about 40 dB in Perceived Level (PL), but we expect a much smaller dose range for future community tests due to limitations of X-59. To understand the impacts of a restricted dose range, we compare the analysis results from the full dataset to the analysis results from a subset of the 2011 pilot study data with a smaller noise dose range.

Another important aspect of planning the X-59 surveys is the necessary sample size, which informs how many participants to recruit. The criterion that will guide this decision is the precision desired for estimation. We demonstrate the sample size calculations for two cases. The first case does not include missing data whereas the second case does include missing data, which is more realistic given observed rates of nonresponse. Both cases make some assumptions to simplify the calculations.

1.4 Outline

This document describes how to statistically model data from a sonic boom community response survey and their applications in planning future surveys using the 2011 pilot study single-event data. In Section 2, we describe the candidate models, all of which are fit using a Bayesian approach. We calculate the summary dose-response curves and credible intervals, and use posterior predictive checking to assess model fit. In Section 3, we first discuss how to select the best models. Then we use them to investigate the effects of a reduced noise dose range and to calculate sample sizes.

In Section 4, we present a summary of the results with suggestions for future work.

2 Candidate Models

In this section, we introduce seven candidate models and fit them to the pilot survey data using a Bayesian approach. After fitting each model, we use posterior predictive checks to assess the model fit. We also compare our models to select the best models for sample size calculation and for analyzing the effects of a reduced noise dose range.

Since most methods suggested in the community noise literature focus on the prevalence of high annoyance, we are ultimately interested in a dose-response estimate with boom exposure level as the dose and percent highly annoyed as the response. We consider three approaches to model the response data: the binary response of highly annoyed or not highly annoyed; the original ordinal survey responses; and a continuous response for proportion of highly annoyed responses. Two model classes are used to model the binary data: logistic regression as suggested by the Federal Interagency Committee on Noise (1992), and a modification of the curve-fitting method proposed by Fidell et al. (2011, 1988). Instead of using the original curve-fitting method, we fit a model similar to logistic regression with the exception of how the probability of occurrence is defined. We refer to this model as the Community Tolerance Level or CTL model. In addition to the models for binary response, we also consider ordinal regression, which models the ordinal responses. For all three model classes, we consider a non-multilevel and multilevel version to learn how the analysis results differ if we model the data without accounting for the longitudinal structure. The last model we consider is a piecewise linear model proposed in the pilot study data analysis report (Page et al., 2014). We only consider a non-multilevel specification of the piecewise linear model because of its relatively poor fit. Table 2.1 provides a summary of the seven candidate models.

Table 2.1: Summary of the seven candidate models.

Model number & name	**Model class**	**Response variable**	**Models longitudinal structure?**
1. Non-multilevel logistic regression (LR)	Logistic regression	Binary: 0/1 for not highly annoyed/highly annoyed	No
2. Multilevel logistic regression (MLR)	Logistic regression	Binary: 0/1 for not highly annoyed/highly annoyed	Yes
3. Non-multilevel CTL (CTL)	Community Tolerance Level[1]	Binary: 0/1 for not highly annoyed/highly annoyed	No
4. Multilevel CTL (MCTL)	Community Tolerance Level	Binary: 0/1 for not highly annoyed/highly annoyed	Yes
5. Non-multilevel ordinal regression (OR)	Ordinal regression	Ordinal: 0 to 10	No
6. Multilevel ordinal regression (MOR)	Ordinal regression	Ordinal: 0 to 10	Yes
7. Non-multilevel piecewise linear regression	Piecewise linear regression	Continuous: proportion of highly annoyed responses at each PL	No

We use a Bayesian approach because it naturally has a hierarchical or multilevel model structure, which is the model structure for three of our candidate models. The number of parameters greatly increases for the multilevel models, and a frequentist optimization approach (maximum likelihood) becomes more complicated. The Bayesian approach draws random samples from the posterior distribution instead of calculating the maximum. The Bayesian approach accounts for uncertainty in the parameters via the posterior distribution, with which we can estimate the probability distribution of any function of the parameters.

[1]Original curve-fitting method is proposed by Fidell et al. (2011, 1988). We modify the method to a statistical model similar to logistic regression. See Section 2.4

2.1 Computation

For each model, we first specify the prior distribution for each model parameter. A prior distribution is the probability distribution of the parameter value based on previous knowledge, before observing the current data. Next, we summarize the observed data via the likelihood function. Finally, we specify the posterior distribution via Bayes Rule by multiplying the prior distribution by the likelihood function and normalizing so the distribution integrates to one. Generally speaking, the prior distribution, the likelihood function, and the posterior distribution are all multidimensional. Summary statistics of the posterior distribution, such as the expected value or the variance, require integration over all dimensions. To simplify, we approximate the integrals via Monte Carlo integration of a random sample drawn from the posterior distribution. Individual draws from this random sample are called posterior draws. There are multiple methods to draw the random sample. We use Markov Chain Monte Carlo (MCMC) sampling for most models because it is a general purpose algorithm that works well across a range of statistical analyses. For the special case when the posterior distribution has only one dimension, we use a grid approximation method (see Section 2.4).

2.1.1 MCMC Sampling and Diagnostics

We use Just Another Gibbs Sampler (JAGS) (Plummer, 2003) to do the MCMC sampling to fit most of the candidate models (all but Model 3, see Section 2.4). We call JAGS from R (R Core Team, 2018) using the package `rjags` (Plummer, 2018). We are using JAGS 4.3.0 and R 3.5.0.

When fitting the candidate models, the number of posterior draws are chosen to ensure that each model parameter has an effective sample size of at least 1000. We check the traceplots and Gelman-Rubin plots (Gelman & Rubin, 1992) for indication of convergence and autocorrelation plots for signs of mixing issues. For each model, we run two MCMC chains. Appendix C provides more details and some examples on checking diagnostic plots.

2.1.2 Posterior Predictive Checking

After fitting each model, we use posterior predictive checks to assess the model fit. The idea is to check whether the data replicated using the model are similar to the observed data. We focus our posterior predictive checking on discrepancy statistics, which are one-number summaries that capture features of the data that we are interested in. The procedures for posterior predictive checking are:

1. Select a set of discrepancy statistics to calculate. For example, the number of highly annoyed responses is one applicable discrepancy statistic in this context.
2. Calculate the set of discrepancy statistics using the observed data. For example, find the number of highly annoyed responses in the observed data.
3. For each posterior draw, generate responses (i.e., 0 or 1 for a binary response or 0-10 for an ordinal response) using parameters from that draw, keeping the design matrix the same.

4. Calculate the set of discrepancy statistics using the generated responses.
5. Repeat Steps 2-3 for each draw.
6. Compare the discrepancy statistic calculated from observed data to the distribution of discrepancy statistics calculated from replicated data.

We compare the observed to replicated discrepancy statistics by plotting a histogram of the discrepancy statistics calculated from the replicated data along with a vertical line for the observed discrepancy statistic. We conclude there is lack of fit if the observed discrepancy statistic falls outside of the middle 95% probability region of the histogram because the fitted model is not generating datasets that replicate the observed data for the feature described by the discrepancy statistic.

2.1.3 Deviance Information Criterion

To compare the relative fit of our models, we calculate the deviance information criterion (DIC). We can only compare the DIC of models that are fit to the same data. For example, ordinal and binary data are different data. The model with the best relative fit has the lowest DIC. In order to calculate DIC, we first define deviance and the posterior mean deviance. Deviance is defined as

$$D(y, \theta) = -2log(f(y|\theta)) \tag{1}$$

where y is the observed data, θ is the vector of parameters in the model, and $f(y|\theta)$ is the likelihood. The posterior mean deviance is defined as

$$D_{avg}(y) = E_{\theta|y}[D(y, \theta)|y] \tag{2}$$

where $E_{\theta|y}[D(y, \theta)|y]$ is the expected value of $D(y, \theta)$ with respect to the posterior distribution. We can estimate the posterior mean deviance as

$$\hat{D}_{avg}(y) \approx \frac{1}{L} \sum_{l=1}^{L} -2log(f(y|\theta^l)) \tag{3}$$

where L is the total number of posterior draws from our MCMC algorithm, and θ^l is the l-th realization of the parameters (in other words, the parameters at the l-th posterior draw). DIC is defined as

$$DIC = \hat{D}_{avg}(y) + p_D \tag{4}$$

where $p_D = \hat{D}_{avg}(y) - D(y, \hat{\theta})$. We specify $\hat{\theta}$ as the posterior mean of the parameters.

2.2 Model 1: Non-Multilevel Logistic Regression

In the community noise literature, logistic regression is a common model because historically the data are split into highly annoyed and not highly annoyed. Although our data were collected longitudinally, we do not account for multiple responses from each participant when we use a non-multilevel logistic regression. In Section 2.3, we use a multilevel model to include participant-level parameters to account for correlation in the responses from each participant.

2.2.1 Data

To fit the logistic regression model, we dichotomize the collected survey responses to highly annoyed or not. The survey responses ranged from 0 to 10. For this 11-point scale, a response is considered as highly annoyed if it is greater than or equal to 8 (Fields et al., 2001). We coded 0 for "not highly annoyed" and 1 for "highly annoyed." The explanatory variable is noise dose given in PL. The binary response is denoted as H and the original ordinal response is denoted as Y.

2.2.2 Model

$$\begin{aligned} H_i|p_i \sim & \quad Bernoulli(p_i) \\ p_i|\beta_0, \beta_1 = & \quad logit^{-1}(\beta_0 + \beta_1 PL_i) \\ \beta_0 \sim & \quad N(0, 100) \\ \beta_1 \sim & \quad N(0, 100) \end{aligned} \tag{5}$$

- Parameters to estimate: β_0, β_1
- Independent variable: noise dose in PL (dB)
- Dependent variable: binary response $H_i = \begin{cases} 1 & \text{if } Y_i = 8, 9, \text{ or } 10 \\ 0 & \text{if } Y_i \leq 7 \end{cases}$

Note that the two model parameters are assigned noninformative prior distributions because we have no information about the model parameters a priori, so the prior distributions should contribute little information relative to the data.

2.2.3 Assumptions

We do not account for the longitudinal structure of the data, and assume that all responses are independent. We model the function for probability of high annoyance, p_i, to only depend on PL. We also assume that PL is known precisely, and do not account for order effects.

2.2.4 Fitting the Model

We draw 1000 burn-in samples and 200,000 additional samples. We center the PL values by subtracting the mean to help reduce the correlation between the two beta parameters. The traceplots and Gelman-Rubin plots both indicate convergence. The autocorrelation plots show no evidence of lack of mixing.

2.2.5 Results

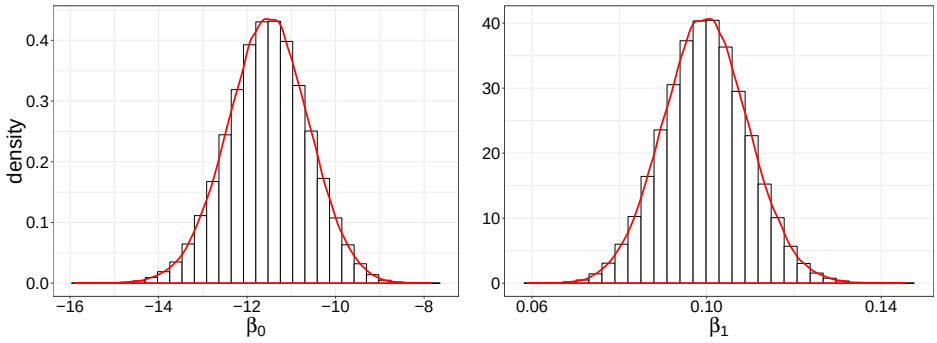
The summary statistics for the parameters are shown in Table 2.2 and the marginal posterior distributions of the two beta parameters are shown in Figure 2.1.[2] The

[2]The marginal posterior distribution of β_0 is the joint posterior distribution of β_0 and β_1 integrated over the parameter space of β_1. In other words, the marginal posterior distribution of β_0 is not a function of β_1.

precision in the summary statistics is determined by the Monte Carlo standard error (MCSE), which we estimated using the time-series standard error (time-series SE).[3]

Table 2.2: Summary statistics of the non-multilevel logistic regression parameters.

	Mean	SD	0.025 quant.	0.25 quant.	Median	0.75 quant.	0.975 quant.
β_0	-11.53	0.92	-13.36	-12.14	-11.52	-10.91	-9.76
β_1	0.0998	0.0098	0.0808	0.0932	0.0997	0.1063	0.1192



(a) (b)

Figure 2.1: Marginal posterior distributions of the non-multilevel logistic regression parameters (a) β_0 and (b) β_1.

For each candidate model, we compute a summary dose-response curve to describe the relationship between the noise doses and percent of highly annoyed responses. For the non-multilevel logistic regression model, this is done by calculating the pointwise posterior means of p_i defined in Eq. 5 for a grid of PL values. The grid is a sequence of 1000 PL values equally spaced within the observed PL range of 63 to 106 dB. In order to calculate the posterior mean of p_i or the predicted probability for each of the 1000 PL values, we first calculate the p_i for each of the 200,000 posterior draws. For each posterior draw, calculate p_i using the values of the β_0 and β_1 parameters from that draw. This results in a distribution of predicted probabilities at each of the 1000 PL values. Then we take the mean of the predicted probabilities across all draws to estimate the posterior means. In other words, at each PL value, we take the mean of the distribution of predicted probabilities, and

[3]Because we estimate the posterior summaries using a random sample from the posterior distribution, the posterior summary estimates calculated from a different random sample would be slightly different. This sampling variability is quantified using the time-series SE via a 95% interval around the posterior mean, or $\pm(2*$time-series SE$)$. For example, the posterior summaries for β_0 in Table 2.2 are reported to 2 decimal places because $2*$time-series SE $= 0.001$ for β_0. The time-series SE is $s/\sqrt{ESS}$ where s is the sample standard deviation and ESS is the effective sample size, or the number of independent samples that the correlated posterior samples is equivalent to. It is calculated as $ESS = L/[1 + 2\sum_{h=1}^{\infty} \rho(h)]$, where L is the number of posterior samples and $\rho(h)$ is autocorrelation at lag h. The time-series SE is typically reported in the R summary output of the posterior draws. Note that the MCSE is not equivalent to the posterior standard deviation, which quantifies the variability of the posterior distribution.

connect them for an estimate of the summary dose-response curve. We find the 95% credible intervals by finding the sample 0.025 and 0.975 quantiles.

For non-multilevel logistic regression, the predicted probability is the expected value of the response, $E(H_i) = logit^{-1}(\beta_0 + \beta_1 * PL_i)$. Figure 2.2 shows the summary dose-response curve and the credible intervals compared to the observed proportions of highly annoyed responses at each PL. The shading of the points indicates the sample size of observed responses at each PL (see Figure 1.3).

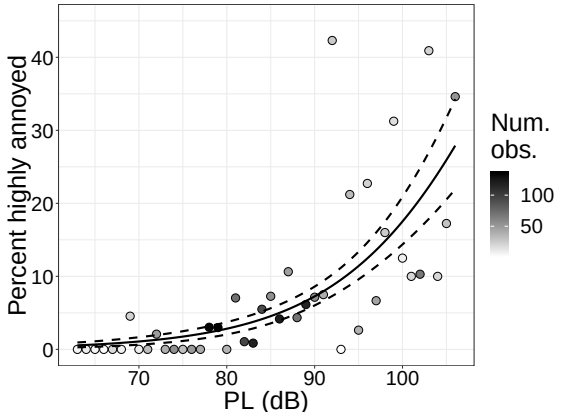
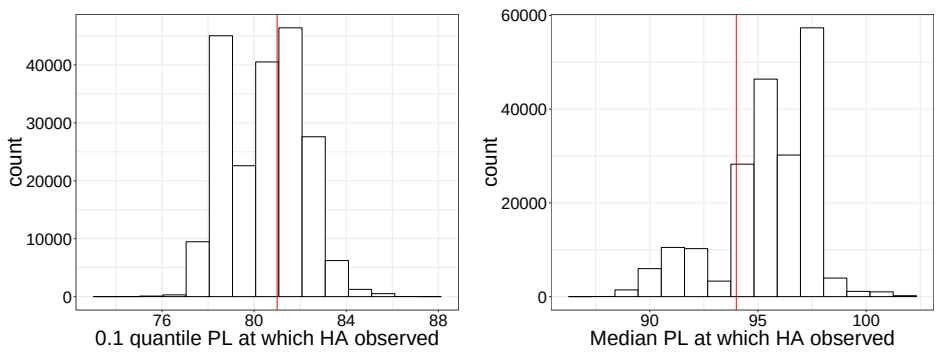


Figure 2.2: Non-multilevel logistic regression summary dose-response curve estimate and 95% credible intervals.

2.2.6 Model Assessment

To assess the fit of the non-multilevel logistic regression model, we use posterior predictive checks as described in Section 2.1. We check four discrepancy statistics for non-multilevel logistic regression—deviance, total proportion of highly annoyed responses, the 0.1 quantile PL at which highly annoyed responses occur, and median PL at which highly annoyed responses occur. We use the 0.1 quantile and median to summarize the PL distribution of highly annoyed responses.

Figure 2.3 shows the 0.1 quantile and median PL posterior predictive checks, while the other checks are plotted in Appendix B. Each histogram shows the value of the discrepancy statistic calculated from replicated data with the red vertical line indicating the statistic calculated from the observed data. We do not see lack of fit based on these checks because each of the observed statistics fall within the middle 95% probability region of their respective histograms.



(a) (b)

Figure 2.3: Posterior predictive checks for non-multilevel logistic regression for (a) the 0.1 quantile PL, and (b) median PL at which highly annoyed responses occur; each histogram indicates discrepancy statistics calculated from replicated data at each posterior draw and red indicates the observed discrepancy statistics.

2.3 Model 2: Multilevel Logistic Regression

2.3.1 Data

We use a multilevel model to account for the correlation among the responses from the same individual. We model each individual with a random intercept β_{0i}; these intercepts are assumed to come from a common distribution. To estimate the individual-level parameters, we keep track of the participants associated with each response. The data are the binary responses for highly annoyed or not, PL, and the subject IDs. A survey response of 8, 9 or 10 is coded as "1" for highly annoyed, and a survey response between 0 and 7 is coded as "0" for not highly annoyed.

2.3.2 Model

Let $i \in 1, ..., S$ be the set of indices indicating the participant and $j \in 1, ..., n_i$ be the set of indices indicating the observation for participant i, where n_i indicates the total number of responses from subject i.

$$\begin{aligned} H_{ij}|p_{ij} \sim & \quad Bernoulli(p_{ij}), \\ p_{ij}|\beta_{0i}, \beta_1 = & \quad logit^{-1}(\beta_{0i} + \beta_1 PL_{ij}) \\ \beta_{0i}|\beta_0, \sigma^2 \sim & \quad N(\beta_0, \sigma^2) \\ \beta_0 \sim & \quad N(0, 100) \\ \beta_1 \sim & \quad N(0, 100) \\ \sigma^2 \sim & \quad InvGamma(0.01, 0.01) \end{aligned} \tag{6}$$

- Parameters to estimate: $\beta_0, \beta_1, \sigma^2, \beta_{0i}\ \forall i$
- Independent variable: noise dose in PL (dB)

- Dependent variable: binary response $H_{ij} = \begin{cases} 1 & \text{if } Y_{ij} = 8, 9, \text{ or } 10 \\ 0 & \text{if } Y_{ij} \leq 7 \end{cases}$

The model parameters β_0, β_1 and σ^2 are assigned noninformative prior distributions.

2.3.3 Assumptions

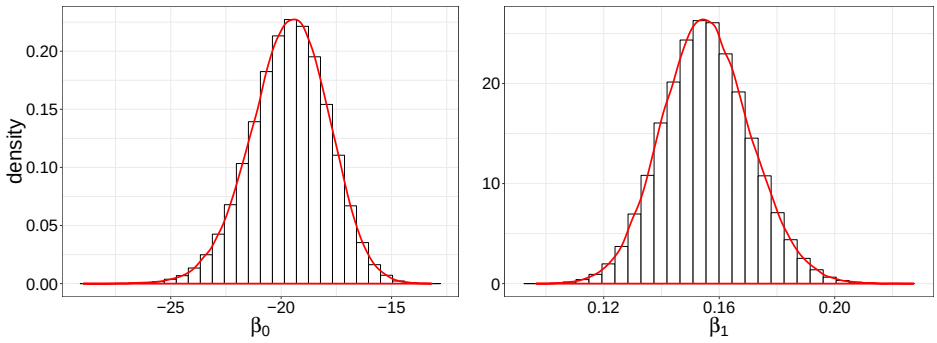
With the multilevel model, we model all individuals with a different intercept, β_{0i}, but the same slope, β_1. We assume all β_{0i} come from a common distribution, which is $N(\beta_0, \sigma^2)$. The probability of high annoyance depends on both PL and the participant. We also assume that PL is known precisely, and do not account for order effects.

2.3.4 Fitting the Model

We use 1000 burn-in samples and 80,000 additional samples. We also center the PL values by subtracting the mean from all observed values when fitting the multilevel logistic regression model. The traceplot and Gelman-Rubin plots both indicate convergence, and the autocorrelation plots do not show signs of mixing issues.

2.3.5 Results

The summary statistics for β_0, β_1 are shown in Table 2.3 and their marginal posterior distributions are shown in Figure 2.4. The precision in the summary statistics is determined by the time-series standard error.



(a) (b)

Figure 2.4: Marginal posterior distributions of the multilevel logistic regression parameters (a) β_0 and (b) β_1.

Table 2.3: Summary statistics of the multilevel logistic regression β_0 and β_1 parameters.

	Mean	SD	0.025 quant.	0.25 quant.	Median	0.75 quant.	0.975 quant.
β_0	-19.6	1.8	-23.2	-20.8	-19.6	-18.4	-16.4
β_1	0.156	0.015	0.127	0.145	0.155	0.166	0.186

For the multilevel models, we first calculate each individual's dose-response curve using procedures similar to those described for the non-multilevel case with individual-level parameters (see Section 2.2.5). Then we take an average of the individual curves to obtain a population representative summary dose-response curve at each posterior draw. So for each posterior draw, we calculate $p_{ij} = logit^{-1}(\beta_{0i} + \beta_1 PL_{ij})$ for the 1000 PL values for each of the 49 participants, and average the 49 individual curves. This averaged curve is one instance of the summary dose-response curve. When this is calculated for each posterior draw, we have a distribution of summary dose-response curves. Then we take the pointwise posterior means among the averaged curves as the mean estimate of the summary dose-response curves. We use the sample 0.025 and 0.975 quantiles to find the 95% credible intervals. The summary dose-response curve for the multilevel logistic regression model is shown in Figure 2.5.

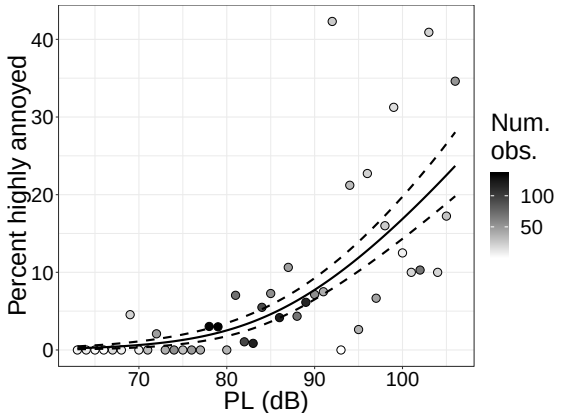


Figure 2.5: Multilevel logistic regression summary dose-response curve estimate and 95% credible intervals.

2.3.6 Model Assessment

For posterior predictive checking, we checked the discrepancy statistics: deviance, total proportion of highly annoyed responses, the 0.1 quantile, and median PL at which highly annoyed responses occurred, the mean number of highly annoyed responses per participant, the standard deviation of highly annoyed responses per participant, the maximum number of highly annoyed responses per participant, and the total number of participants who responded highly annoyed. We used more discrepancy statistics for the multilevel specification than for non-multilevel because

there is additional information at the participant level, and we can check model fit for additional features in the data. The posterior predictive checks do not show lack of fit for the listed features. Figure 2.6 shows the histograms for the 0.1 quantile PL, median PL, and the total number of participants who responded highly annoyed. The histograms for the remaining discrepancy statistics are in Appendix B.

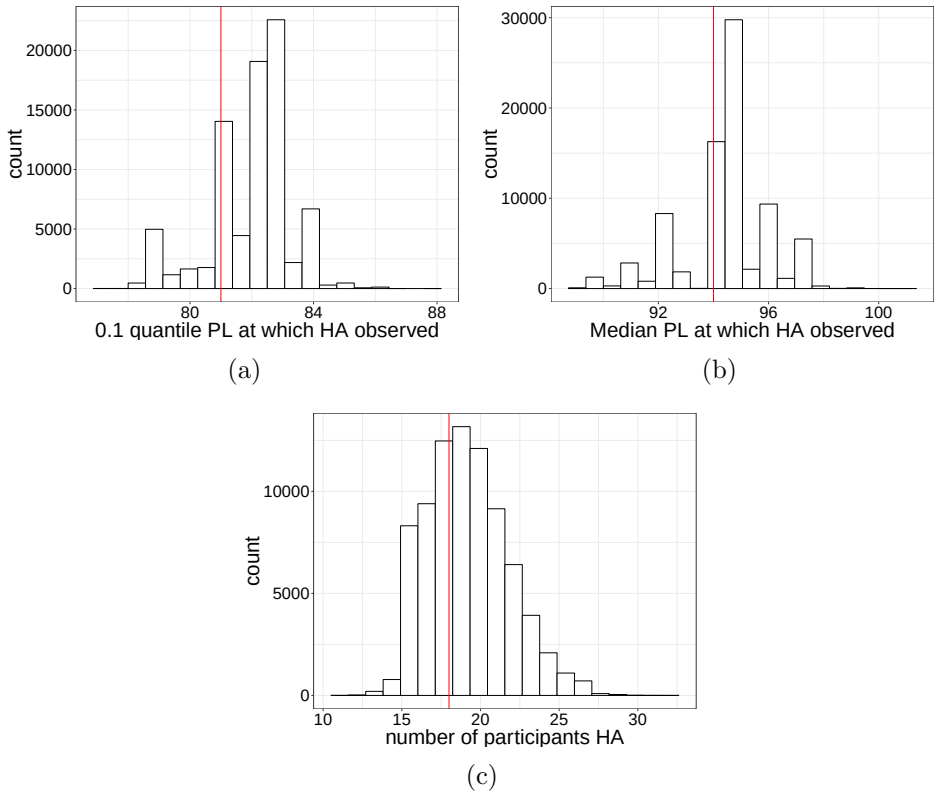


Figure 2.6: Posterior predictive checks for multilevel logistic regression for (a) the 0.1 quantile PL and (b) median PL at which highly annoyed responses occur, and (c) the number of participants highly annoyed at least once; each histogram indicates discrepancy statistics calculated from replicated data at each posterior draw and red indicates the observed discrepancy statistics.

2.4 Model 3: Non-Multilevel CTL

The original approach specified in Fidell et al. (2011, 1988) is not a statistical model, but a least squares curve-fitting method. They propose to fit the curve $p_i = e^{-C/m_i}$, where $m_i = (10^{PL_i/10})^{0.3}$, to observed proportions of highly annoyed responses versus noise dose by minimizing the squared difference between the observed data and the curve. The parameter to estimate is C. The transformation of the noise dose to m is based on Stevens' power law that relates loudness to measured sound pressure levels (Stevens, 1975). To calculate the observed proportions, we find the proportion of survey responses that are 8, 9 or 10 at each PL. When applied to our data, there are only 44 points that are used for this method. This curve-fitting

method gives equal weighting for all observed points, but our data have a different number of observations at each PL (see Figure 1.3).

We instead model the binary data for highly annoyed or not highly annoyed using a statistical model. The model is similar to logistic regression, but we specify the probability of occurrence to be $p_i = e^{-C/m_i}$ instead of $logit^{-1}(\beta_0 + \beta_1 PL_i)$. We refer to this model as the CTL model.

As with logistic regression, model 3 does not model the longitudinal structure in the data, hence the non-multilevel CTL model does not account for correlation among multiple responses from each participant. Model 4 introduces a multilevel structure to the CTL model that includes participant-level parameters to account for the correlation among multiple responses from each participant.

2.4.1 Data

The data for CTL are also binary data, so the same data are used for logistic regression and CTL. The survey responses of 8, 9 or 10 are coded as "1" for highly annoyed, and survey responses between 0 and 7 are coded as "0" for not highly annoyed. The explanatory variable is noise dose given in PL.

2.4.2 Model

$$\begin{aligned} H_i|p_i \sim &\quad Bernoulli(p_i) \\ p_i|C = &\quad e^{-C/m_i} \\ C \sim &\quad Unif(0, 10000) \end{aligned} \tag{7}$$

where m is a transformation of PL, $m_i = (10^{PL_i/10})^{0.3}$.

- Independent variable: m_{ij}, which is a transformed noise dose in PL (dB)
- Dependent variable: binary response $H_i = \begin{cases} 1 & \text{if } Y_i = 8, 9 \text{ or } 10 \\ 0 & \text{if } Y_i \leq 7 \end{cases}$
- Parameter to estimate: C

2.4.3 Assumptions

For the non-multilevel CTL model, we do not model the longitudinal structure and assume all responses are independent. The function for probability of high annoyance only depends on PL. We also assume that PL is known precisely, and do not account for order effects.

2.4.4 Fitting the Model

Since Eq. 7 is a one parameter model, MCMC sampling is not necessary. Instead, we use a grid approximation method to sample from the posterior distribution of C. The general procedure to approximate the posterior distribution is outlined in Section 2.5 of Albert (2007). Below is an outline for applying this method to fit the

non-multilevel CTL model:

1. Find a vector of values of C such that the support is not 0 for the posterior distribution
2. Calculate the likelihood and prior at the vector of C values found in Step 1
3. Multiply the likelihood and prior, and normalize by the sum of the products to create weights
4. Take a random sample of C values with replacement using the weights to simulate a sample from the posterior distribution

2.4.5 Results

Figure 2.7 shows an approximation to the posterior distribution of C, which is the histogram of the sample of size 100,000 from Step 4. The summary statistics for the parameter C are shown in Table 2.4.

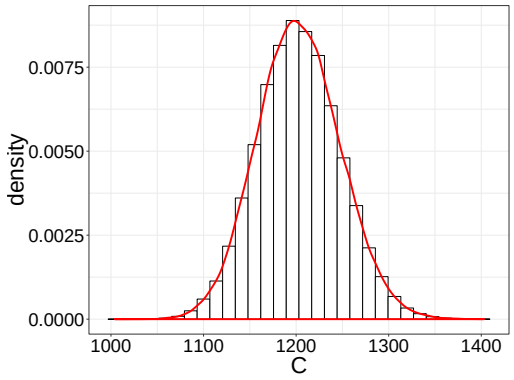


Figure 2.7: Posterior distribution of the non-multilevel CTL parameter C.

Table 2.4: Summary statistics of the non-multilevel CTL parameter.

	Mean	SD	0.025 quant.	0.25 quant.	Median	0.75 quant.	0.975 quant.
C	1203	45	1118	1172	1202	1233	1293

We compute the summary dose-response curve for the non-multilevel CTL model following the procedures in Section 2.2.5. The probability of high annoyance is changed to $p_i = e^{C/m_i}$ instead of $p_i = logit^{-1}(\beta_0 + \beta_1 PL_i)$. Figure 2.8 shows the summary dose-response estimate for non-multilevel CTL. The model fits the low PL well but seems to overpredict at high PL.

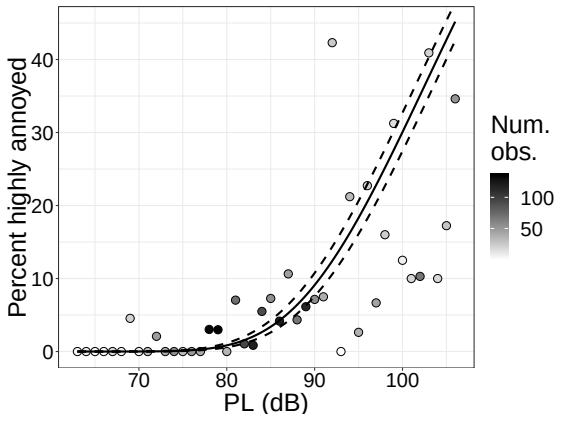
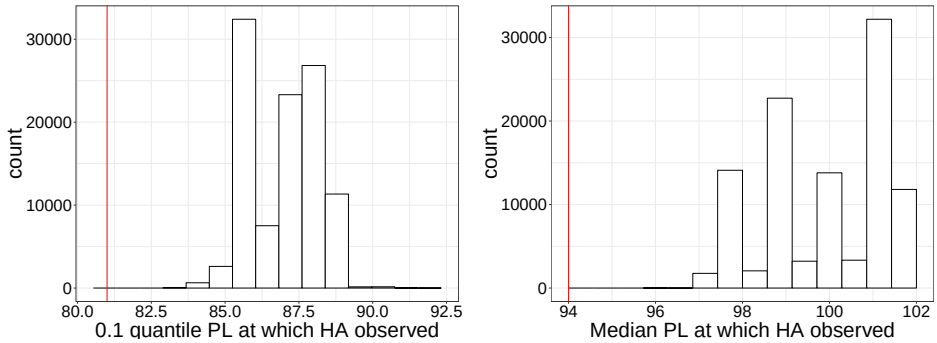


Figure 2.8: Non-multilevel CTL summary dose-response curve estimate and 95% credible intervals.

2.4.6 Model Assessment

For non-multilevel CTL, we check the same four discrepancy statistics that we use for non-multilevel logistic regression—deviance, total proportion of highly annoyed responses, the 0.1 quantile and median PL at which highly annoyed responses occur. Figure 2.9 shows the 0.1 quantile and median PL posterior predictive checks. The remaining two checks are in Appendix B. We see that the PL distribution of highly annoyed responses replicated from the model is shifted to the right relative to the observed. These two plots indicate that some observed highly annoyed responses occur at lower PL than the non-multilevel CTL model predicts. This is consistent with the model overpredicting at high PL observed in Figure 2.8.



(a) (b)

Figure 2.9: Posterior predictive checks for non-multilevel CTL for (a) the 0.1 quantile PL, and (b) median PL at which highly annoyed responses occur; each histogram indicates discrepancy statistics calculated from replicated data at each posterior draw and red indicates the observed discrepancy statistics.

2.5 Model 4: Multilevel CTL

We model the longitudinal structure of the data with a multilevel specification of CTL by including individual-level parameters to account for the correlation among multiple responses from each participant.

2.5.1 Data

The data for the multilevel CTL model are the same as for multilevel logistic regression. We use the binary data for highly annoyed or not highly annoyed, the PL, and the subject IDs. The binary data are coded from the survey responses: a survey responses of 8, 9 or 10 is coded as "1" for highly annoyed, and a survey response between 0 and 7 is coded as "0" for not highly annoyed.

2.5.2 Model

$$\begin{aligned} H_{ij}|p_{ij} \sim &\quad Bernoulli(p_{ij}), \\ p_{ij}|C_i = &\quad e^{-C_i/m_{ij}} \\ C_i|C_0, \sigma^2 \sim &\quad N(C_0, \sigma^2) \\ C_0 \sim &\quad Gamma(24, 0.005) \\ \sigma \sim &\quad N(900, 100^2) \end{aligned} \tag{8}$$

where $m_{ij} = (10^{PL_{ij}/10})^{0.3}$.

- Independent variable: m_{ij}, which is a transformed noise dose in PL (dB)
- Dependent variable: binary response $H_{ij} = \begin{cases} 1 & \text{if } Y_{ij} = 8, 9 \text{ or } 10 \\ 0 & \text{if } Y_{ij} \leq 7 \end{cases}$
- Parameters to estimate: C_i, C_0, σ

2.5.3 Specifying Informative Priors

For this model, we use informative priors to help with convergence. Strictly for model fitting, there are many values of C_i that are plausible. However, there are some restrictions to a reasonable range of values that are physically plausible. For example, the probability of high annoyance will be low at PL levels near ambient and reach 1 at a very high PL. The prior distributions are picked conservatively because we only need to approximate roughly the appropriate range of C_i values. This will lead to a narrower range of possible values rather than all positive real values. We check that the chosen prior distributions are reasonable using prior predictive checks.

The prior distributions are chosen to bound C_i between 70 to 10,000 based on the following two criteria, which are educated guesses based on discussions with a NASA subject matter expert. The prior for C_0 is set to the middle of the 70 to 10,000 range with some variation, while the prior for σ is chosen using prior predictive checks.

1. A gunshot recorded at the listener's ears is about 135 dB in PL (Doebler & Rathsam, 2019). At a much higher PL such as 200 dB, we think it is reasonable to assume that almost anyone would be highly annoyed, so the probability of high annoyance is approximately equal to 1. As C_i decreases, the curve shifts to the left and probability of high annoyance at 200 dB gets closer to 1. From trial-and-error and examining plots of the resulting curves, we found an upper bound of 10,000 for C_i. In other words, the highest value of C_i that would result in probability of high annoyance approximately equal to 1 at 200 dB is approximately $C_i = 10,000$.

2. Car door slams from a neighboring house are approximately 60 dB PL (Doebler & Rathsam, 2019) so we consider sound levels 10 dB lower than that to be at approximate ambient levels. If a sound level is at ambient level, there is a very small probability of high annoyance. At 50 dB, we think it is reasonable to assume that the highest probability of high annoyance is 0.1. As C_i increases, the curve shifts to the right and probability of high annoyance at 50 dB decreases. Again from trial-and-error, we found a lower bound of 70 for C_i. In other words, the lowest value of C_i that would result in at most 10% highly annoyed at 50 dB is approximately $C_i = 70$.

The steps for prior predictive checks are:

1. Draw from joint prior of C_0 and σ and generate one value of C_i from the (C_0, σ) pair

2. Repeat Step 1 10,000 times

3. Plot histogram of the C_i values to check if they fall in a reasonable range of values for C_i. If they do not, change prior and repeat steps 1-3 until the generated histogram of C_i looks reasonable

Figure 2.10 shows 10,000 randomly drawn C_i from the joint prior of C_0 and σ specified in Eq. 8. We see the C_i are restricted to approximately the 70 to 10,000 range.

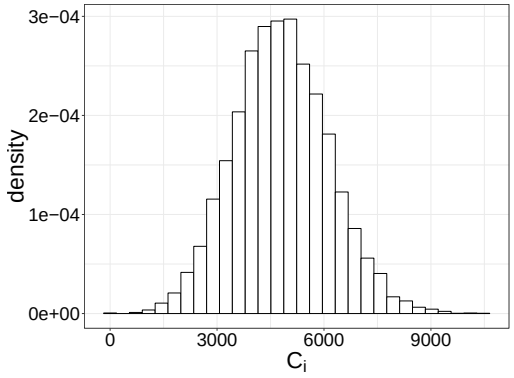


Figure 2.10: C_i generated from informative priors specified for multilevel CTL.

2.5.4 Assumptions

For the multilevel model, we include individual-level parameters, C_i, and assume that they come from a common distribution: $N(C_0, \sigma^2)$. The individual-level parameters account for correlation among the multiple responses from each participant. The function for probability of high annoyance depends on both PL and the individual. We also assume that PL is known precisely, and do not account for order effects.

2.5.5 Fitting the Model

We use JAGS to fit the multilevel CTL model, and run 5000 burn-in samples with an additional 80,000 iterations. The traceplots and Gelman-Rubin plots all indicate convergence. The autocorrelation plots do not indicate problems with mixing.

2.5.6 Results

Table 2.5 shows the summary statistics for C_0, and Figure 2.11 shows the marginal posterior distribution. The precision in the summary statistics is determined by the time-series standard error.

Table 2.5: Summary statistics of the MCTL parameter C_0.

	Mean	SD	0.025 quant.	0.25 quant.	Median	0.75 quant.	0.975 quant.
C_0	2600	200	2200	2500	2600	2700	3100

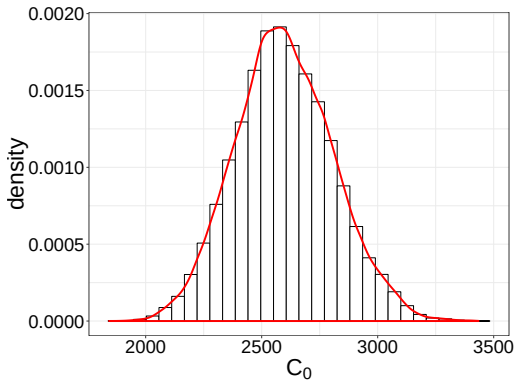


Figure 2.11: Marginal posterior distribution of the multilevel CTL parameter C_0.

After fitting the model, we calculate a summary dose-response curve, which is an average of the individual curves. The procedures are outlined in Section 2.3.5, and the function for calculating individual probability of high annoyance is $p_{ij} = \exp(-C_i/m_{ij})$. The summary dose-response curve for multilevel CTL is shown in Figure 2.12. We see the multilevel CTL summary dose-response curve fits the data much better than the non-multilevel curve.

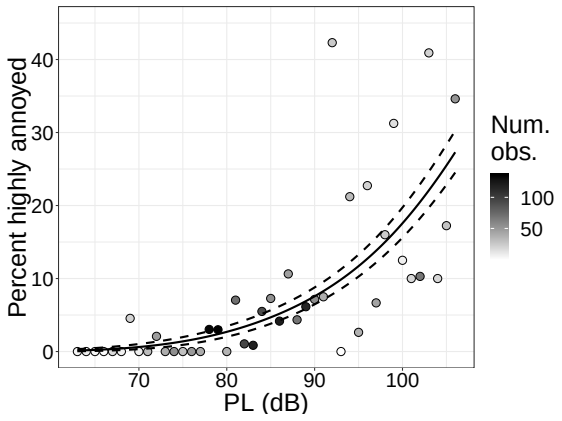


Figure 2.12: Multilevel CTL summary dose-response curve estimate and 95% credible intervals.

2.5.7 Model Assessment

To assess the model fit, we check the following discrepancy statistics—deviance, total proportion of highly annoyed responses, the 0.1 quantile and median PL at which highly annoyed responses occurred, the mean number of highly annoyed responses per participant, the standard deviation of highly annoyed responses per participant, the maximum number of highly annoyed responses per participant, and the total number of participants who responded highly annoyed at least once. Figure 2.13 shows the posterior predictive checks for the 0.1 quantile PL, median PL, and the number of participants highly annoyed at least once. The remaining histograms are plotted in Appendix B. Only the posterior predictive check for the number of participants highly annoyed at least once indicate lack of fit. Figure 2.13c indicates that the model predicts too many participants highly annoyed at least once compared to the observed number.

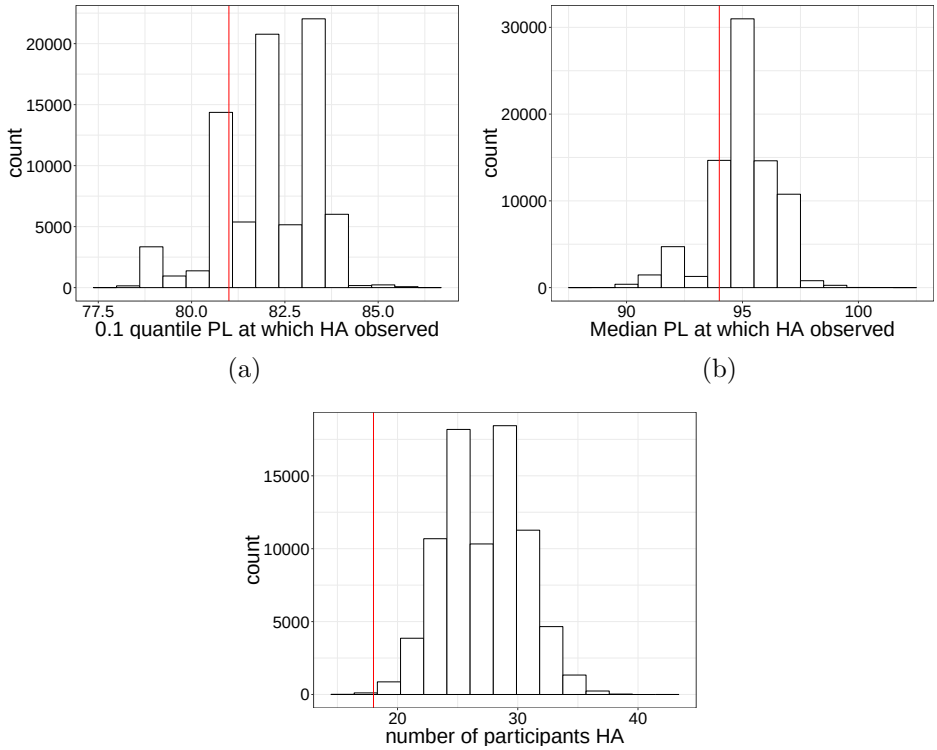


Figure 2.13: Posterior predictive checks for multilevel CTL for (a) the 0.1 quantile PL and (b) median PL at which highly annoyed responses occur, and (c) number of participants highly annoyed at least once; each histogram indicates discrepancy statistics calculated from replicated data at each posterior draw and red indicates the observed discrepancy statistics.

2.5.8 Model Comparison

We calculate the DIC as discussed in Section 2.1 to compare the four models discussed thus far. The DIC for the four models are shown in Table 2.6, ordered by lowest DIC. Lower DIC indicates better relative fit. The best fit to the data based on DIC is multilevel logistic regression. The relative fit of both multilevel models are better than their non-multilevel counterparts.

Table 2.6: Comparison of DIC for the four binary models.

Model	DIC
Multilevel logistic regression	510.5
Multilevel CTL	534.4
Non-multilevel logistic regresion	863.7
Non-multilevel CTL	925.07

From posterior predictive checking, we see that the multilevel CTL model fails to

capture one important feature of the data that multilevel logistic regression is able to capture—the number of participants highly annoyed at least once. Combining the results from posterior predictive checks and DIC, multilevel logistic regression fits the data better than multilevel CTL. Figure 2.14 compares the summary dose-response curves from the two multilevel models. The two multilevel summary does-response curves overlap until about 100 dB where the multilevel CTL model starts to estimate slightly higher than the multilevel logistic regression model. This may be related to how the multilevel CTL model predicts a higher number than the multilevel logistic regression model for the number of participants highly annoyed at least once.

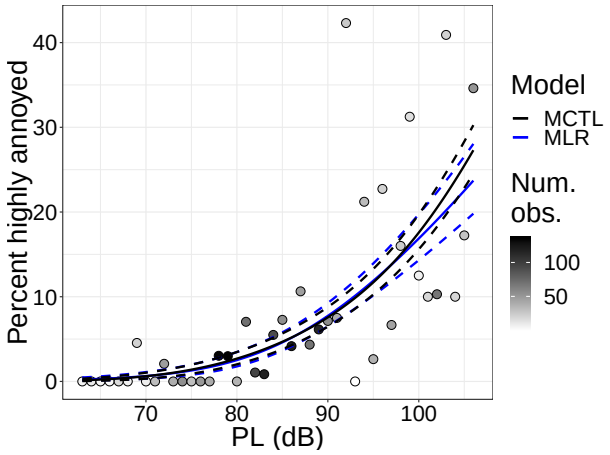


Figure 2.14: Comparison of multilevel logistic regression and multilevel CTL summary dose-response curves.

2.6 Model 5: Non-Multilevel Ordinal Regression

A different analysis approach is to model the ordinal responses directly in order to fully make use of the data available. First, we consider a non-multilevel specification, which does not account for the longitudinal structure of the data and assume all responses are independent.

2.6.1 Data

Let Y be the ordinal annoyance responses. This is the survey response to the question about how much the sonic boom annoyed, bothered or disturbed the participant. The response scale is from 0 to 10.

2.6.2 Model

There are two common ways for parameterizing the ordinal regression model—a cumulative link model (Agresti, 2003) or a latent variable model (Long, 1997). We use the latent variable model by assuming that an underlying continuous variable Y^* with range $(-\infty, \infty)$ can be mapped back to the ordinal response Y.

The relationship between the ordinal response and the latent variable is analogous to a letter grade that is derived from a continuous numerical score. A numerical score can be mapped to a letter grade, but a letter grade can only be mapped to a range of possible values, not an exact numerical score. For ordinal regression, we assume the thresholds for distinguishing the intervals of numerical scores corresponding to the letter grades are unknown and need to be estimated from the data. If we assume that the thresholds are fixed and known, then we can use interval regression instead, which is proposed by Groothuis-Oudshoorn & Miedema (2006).

Mathematically, the relationship between the latent variable Y^* and ordinal variable Y is:

$$Y_i = k \text{ if } \gamma_k < Y_i^* \leq \gamma_{k+1} \text{ for } k = 0, ..., 10$$

where k is the ordinal response, γ_k corresponds to the lower threshold, and γ_{k+1} corresponds to the upper threshold on the latent variable scale. The first and last thresholds are $\gamma_0 = -\infty$ and $\gamma_{11} = \infty$. Note that the thresholds or gamma parameters are strictly increasing, so $\gamma_1 < \gamma_2 < ... < \gamma_{10}$.

The model specifies that the latent variable is linearly related to PL: $Y_i^* \sim N(\beta_0 + \beta_1 PL_i, \sigma^2)$. There are two ways that Long (1997) suggests to make the model identifiable, and we use the first method:

1. Assume variance $\sigma^2 = 1$, and $\gamma_1 = 0$, or
2. Assume variance $\sigma^2 = 1$ and $\beta_0 = 1$

So, the model is

$$\begin{aligned}
Y_i|\pi_i \sim \quad & Multinomial(1, \boldsymbol{\pi_i}) = Categorical(\boldsymbol{\pi_i}) \\
\boldsymbol{\pi_i}|\boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \boldsymbol{\gamma_2}, ..., \boldsymbol{\gamma_{10}} = \quad & \begin{bmatrix} \Phi(0 - \beta_0 - \beta_1 PL_i) \\ ... \\ \Phi(\gamma_{10} - \beta_0 - \beta_1 PL_i) - \Phi(\gamma_9 - \beta_0 - \beta_1 PL_i) \\ 1 - \Phi(\gamma_{10} - \beta_0 - \beta_1 PL_i) \end{bmatrix} \\
\beta_0 \sim \quad & N(0, 100) \\
\beta_1 \sim \quad & N(0, 100) \\
\gamma_k \sim \quad & N(0, 10) \text{ for } k = 2, ..., 10
\end{aligned} \tag{9}$$

- Independent variable: noise dose in PL (dB)
- Dependent variable: annoyance responses (ordinal)
- Parameters to estimate: $\beta_0, \beta_1, \gamma_2, ..., \gamma_{10}$

Figure 2.15 shows the mapping from the ordinal variable to the latent variable, and the relationship among the two variables and the covariate. This is a modified version of Figure 23.6 from Kruschke (2014). Each level for the ordinal variable Y is mapped to an interval of values for the latent variable Y^*. We model Y^* to have a linear relationship with PL, and the line shows the expected value of Y^*,

$E(Y^*)$. Once we have specified $E(Y^*)$, we have fully specified the model for Y^*, which is a normal distribution with mean $E(Y^*)$ and variance 1, as shown vertically in Figure 2.15. The normal distribution is necessary for deriving the probabilities associated with each ordinal level in $\pi_\mathbf{i}$ in Eq. 9. Note that $\pi_\mathbf{i}$ is an 11 element vector with each element corresponding to the probability of each ordinal level. Note that when the value of PL changes, $E(Y^*)$ also changes. When $E(Y^*)$ changes, the normal distribution will shift either up or down, thus causing the probabilities associated with each ordinal level to change. Notice that the gamma parameters are not assumed to be equally spaced and are estimated based on the data.

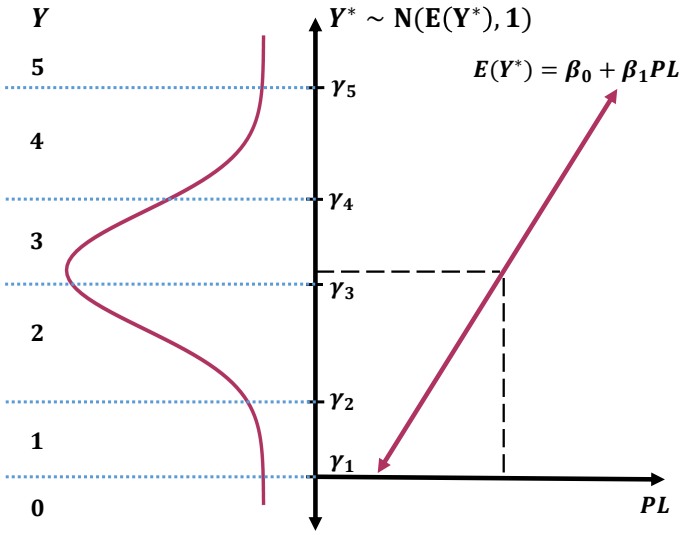


Figure 2.15: Relationship among the ordinal variable (Y), latent variable (Y^*) and covariate (PL).

2.6.3 Assumptions

We assume that all observations are independent for the non-multilevel ordinal regression model. So the vector of probabilities, π_i, only varies with PL. We also assume that the latent variable is normally distributed and has a linear relationship with PL. We also assume that PL is known precisely, and do not account for order effects.

2.6.4 Fitting the Model

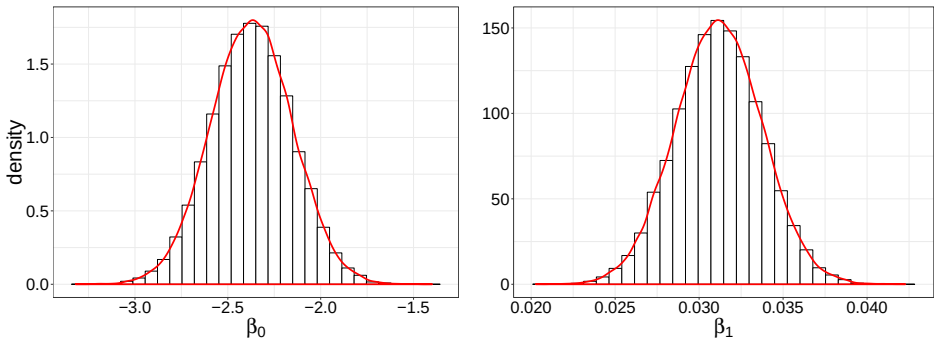
We draw 1000 burn-in samples and 40,000 additional samples. We also center the PL value when fitting the non-multilevel ordinal regression model. Both traceplots and Gelman-Rubin plots indicate convergence. The autocorrelations for the γ parameters drop slowly because they are highly correlated, but the autocorrelation plots do not indicate problems with mixing.

2.6.5 Results

Table 2.7 shows the table of summary statistics for the parameters, and Figure 2.16 shows the marginal posterior distributions of β_0 and β_1. The precision in the summary statistics is determined by the time-series standard error.

Table 2.7: Summary statistics of the ordinal regression parameters β_0, β_1 and all γ_k.

	Mean	SD	0.025 quant.	0.25 quant.	Median	0.75 quant.	0.975 quant.
β_0	-2.37	0.22	-2.80	-2.52	-2.37	-2.22	-1.94
β_1	0.0312	0.0026	0.0262	0.0294	0.0312	0.0329	0.0362
γ_2	0.426	0.022	0.384	0.412	0.426	0.441	0.470
γ_3	0.71	0.03	0.66	0.69	0.71	0.73	0.76
γ_4	1.02	0.03	0.95	1.00	1.02	1.04	1.08
γ_5	1.26	0.04	1.18	1.23	1.26	1.28	1.33
γ_6	1.44	0.04	1.36	1.41	1.44	1.46	1.52
γ_7	1.63	0.04	1.55	1.60	1.63	1.66	1.72
γ_8	1.87	0.05	1.77	1.84	1.87	1.90	1.97
γ_9	2.14	0.06	2.03	2.10	2.14	2.18	2.26
γ_{10}	2.28	0.06	2.15	2.23	2.27	2.32	2.40



(a) (b)

Figure 2.16: Marginal posterior distributions of the non-multilevel ordinal regression parameters (a) β_0 and (b) β_1.

Figure 2.17 plots the values of each sampled gamma vector against its mean to keep the gamma parameters at each draw plotted together. Since the gamma parameters are highly correlated, it may be misleading to examine the marginal posterior distributions of each gamma parameter individually. The dashed lines indicate the posterior means for each gamma parameter, and the yellow points indicate one of the sampled gamma vectors. We see in Figure 2.17 that the gamma parameters are not estimated to be equally spaced. For example, the distance between the first two gamma parameters (the two leftmost yellow points) is not the same as the distance between the last two gamma parameters (the two rightmost yellow points). This indicates that the ordinal regression model may be preferred

over the interval regression model proposed by Groothuis-Oudshoorn & Miedema (2006), which assumes the gamma parameters are equally spaced. We also see that there is higher variability in the distribution of the gamma parameters at the higher end of the response scale because there are more responses on the lower end of the 11-point scale. Thus, there is less uncertainty in the posterior distributions of the gamma parameters corresponding to the lower end of the response scale.

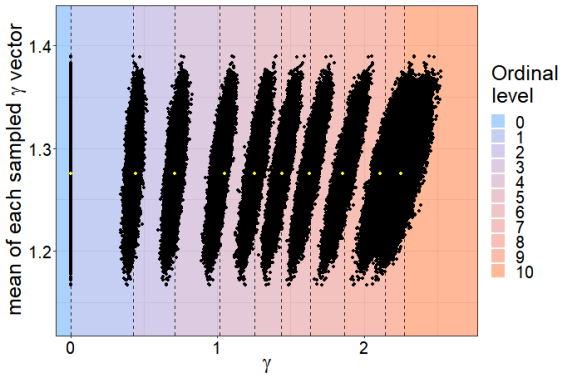


Figure 2.17: Posterior distribution of non-multilevel ordinal regression γ parameters; dashed lines indicate posterior means of each gamma parameter, and yellow points indicate the gamma parameters for one random posterior draw.

Since the dose-response relationship is between noise exposure and percent highly annoyed, we need to calculate the probability of high annoyance (and multiply by 100) from the ordinal regression model. It is straightforward for logistic regression and CTL because we modeled the binary variable for highly annoyed or not, and specified the functional form of the probability of occurrence. For ordinal regression models, there are two ways to calculate the function for the probability of high annoyance, both yielding the same function. The two methods rely on using either the ordinal scale or the latent variable scale. First, if we consider the ordinal scale, we know the probability for each ordinal level based on π_i from Eq. 9. The probability of high annoyance is $p_i = P(Y \geq 8) = P(Y = 8) + P(Y = 9) + P(Y = 10)$. If we let $z_i = \beta_0 + \beta_1 * PL_i$, then

$$\begin{aligned} P(Y \geq 8) &= P(Y = 8) + P(Y = 9) + P(Y = 10) \\ &= (\Phi(\gamma_9 - z_i) - \Phi(\gamma_8 - z_i)) + (\Phi(\gamma_{10} - z_i) - \Phi(\gamma_9 - z_i)) + (1 - \Phi(\gamma_{10} - z_i)) \\ &= 1 - \Phi(\gamma_8 - z_i) \\ p_i &= 1 - \Phi(\gamma_8 - \beta_0 - \beta_1 * PL_i) \end{aligned} \tag{10}$$

We can alternatively consider the latent variable, Y^*, for calculating the probability of high annoyance like Groothuis-Oudshoorn & Miedema (2006). The idea is that instead of identifying all responses greater than or equal to 8 on the ordinal scale as highly annoyed, we identify these responses from the latent variable scale. In other words, we need to identify all responses greater than or equal to the gamma parameter corresponding to 8 on the latent variable scale as highly annoyed. For

the 11-point scale, γ_8 corresponds to the high annoyance cutoff. So the probability of high annoyance is $p_i = P(Y^* \geq \gamma_8)$. Note that $p_i = P(Y^* \geq \gamma_8) = P(Y \geq 8)$. Because we made distributional assumptions about Y^* (more specifically, $Y^* \sim N(\beta_0 + \beta_1 PL_i, 1)$), we can easily find $P(Y^* \geq \gamma_8)$.

$$\begin{aligned} P(Y^* \geq \gamma_8) = & \quad 1 - P(Y^* < \gamma_8) \\ p_i = & \quad 1 - \Phi(\gamma_8 - \beta_0 - \beta_1 * PL_i) \end{aligned} \tag{11}$$

Both methods result in the same function for calculating probability of high annoyance.

We calculate the summary dose-response curve by following the procedures outlined in Section 2.2.5 using Eq. 11 for p_i. The summary dose-response curve estimate and 95% credible intervals are shown in Figure 2.18. The estimates are higher than observed proportions at low PL, which indicates potential lack of fit.

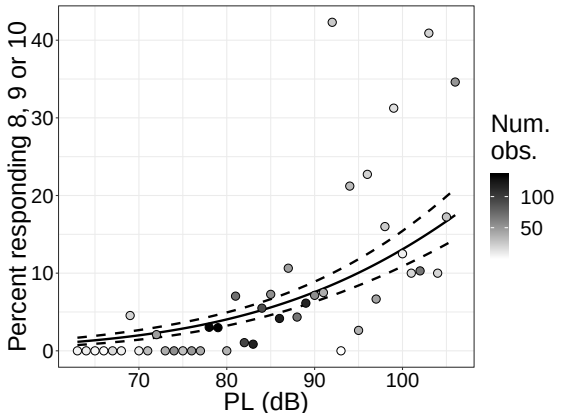
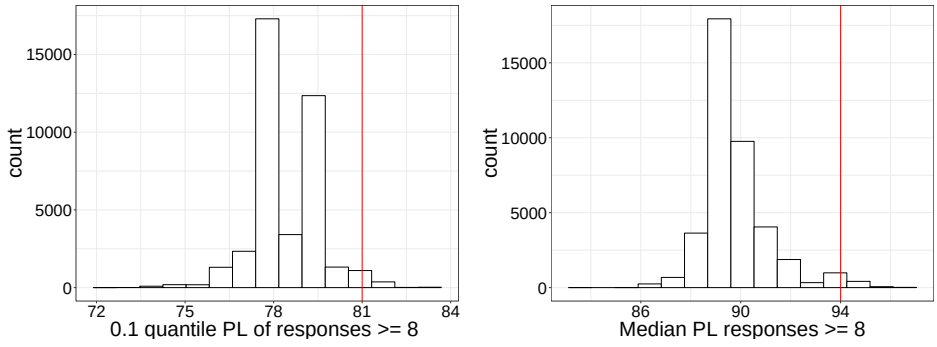


Figure 2.18: Non-multilevel ordinal regression summary dose-response curve estimate and 95% credible intervals.

2.6.6 Model Assessment

To assess model fit, we check the same four discrepancy statistics that we did for other non-multilevel models—deviance, total proportion of highly annoyed responses, the 0.1 quantile and median PL at which highly annoyed responses occur. Figure 2.19 shows the posterior predictive checks for the 0.1 quantile and median PL. Note that for ordinal regression, we consider the responses that are 8, 9 or 10 for highly annoyed responses. We see that the distribution of PL at which highly annoyed responses occur for the ordinal regression model is shifted to the left (lower end) of the PL range relative to the observed value. The empirical quantiles for both observed statistics are 0.99, indicating lack of fit for these two data features. The histograms for the other two discrepancy statistics are in Appendix B, and neither show lack of fit.



(a) (b)

Figure 2.19: Posterior predictive checks for non-multilevel ordinal regression for (a) the 0.1 quantile PL and (b) median PL at which responses of 8, 9 or 10 occur; each histogram indicates discrepancy statistics calculated from replicated data at each posterior draw and red indicates the observed discrepancy statistic.

2.7 Model 6: Multilevel Ordinal Regression

We consider a multilevel ordinal regression model that accounts for the longitudinal structure of the data. To model the correlation among multiple responses from the same participants, we introduce individual-level parameters β_{0i}, which all come from a common distribution, $N(\beta_0, \sigma_0^2)$.

2.7.1 Data

The data for the multilevel ordinal regression model are the same as the non-multilevel model with the addition of subject IDs for keeping track of the participants. The response is the ordinal survey response ranging from 0 to 10.

2.7.2 Model

We will once again assume a latent variable Y^* with range $(-\infty, \infty)$ that can be mapped back to the ordinal response Y. The mathematical relationship between the latent variable Y^* and ordinal response Y is:

$$Y_{ij} = k \text{ if } \gamma_k < Y_{ij}^* \leq \gamma_{k+1} \text{ for } k = 0, ..., 10$$

where k is the ordinal response, γ_k corresponds to the lower (random) threshold and γ_{k+1} corresponds to the upper (random) threshold on the latent variable scale. The first and last thresholds remain as $\gamma_0 = -\infty$ and $\gamma_{11} = \infty$. Note that the gamma parameters are strictly increasing, so $\gamma_1 < \gamma_2 < ... < \gamma_{10}$. We will use the same method for identifying the model as with non-multilevel ordinal regression by assuming the variance of the latent variable $\sigma^2 = 1$ and the second threshold $\gamma_1 = 0$ (note that the indexing for the thresholds start at 0).

So, the model is

$$
\begin{aligned}
Y_{ij}|\pi_{ij} \sim & \quad Multinomial(1, \pi_{\mathbf{ij}}) \\
\pi_{\mathbf{ij}}|\beta_{\mathbf{0i}}, \beta_{\mathbf{1}}, \gamma_{\mathbf{2}}, ..., \gamma_{\mathbf{10}} = & \begin{bmatrix} \Phi(0 - \beta_{0i} - \beta_1 PL_{ij}) \\ ... \\ \Phi(\gamma_{10} - \beta_{0i} - \beta_1 PL_{ij}) - \Phi(\gamma_9 - \beta_{0i} - \beta_1 PL_{ij}) \\ 1 - \Phi(\gamma_{10} - \beta_{0i} - \beta_1 PL_{ij}) \end{bmatrix} \\
\beta_{0i}|\beta_0, \sigma_0^2 \sim & \quad N(\beta_0, \sigma_0^2) \\
\beta_0 \sim & \quad N(0, 100) \\
\beta_1 \sim & \quad N(0, 100) \\
\gamma_k \sim & \quad N(0, 10) \text{ for } k = 2, ..., 10 \\
\sigma_0^2 \sim & \quad InvGamma(0.01, 0.01).
\end{aligned} \tag{12}
$$

The probabilities are derived from the model for the latent variable Y^*,

$$
\begin{aligned}
Y_{ij}^*|\beta_{0i}, \beta_1 \sim & \quad N(\beta_{0i} + \beta_1 PL_{ij}, 1) \\
\beta_{0i}|\beta_0, \sigma_0^2 \sim & \quad N(\beta_0, \sigma_0^2)
\end{aligned} \tag{13}
$$

- Independent variable: noise dose in PL (dB)
- Dependent variable: annoyance responses (ordinal)
- Parameters to estimate: $\beta_0, \beta_1, \sigma_0^2, \beta_{0i}\ \forall i$, and γ_k for $k = 2, 3, .., 10$

2.7.3 Assumptions

We model the correlation among multiple responses from the same participant using individual-level parameters β_{0i}. The probabilities for each ordinal level, π_{ij}, are dependent on the participant and PL. As with the non-multilevel model, we assume that the latent variable is normally distributed. We also assume that the individual-level parameters β_{0i} come from a common normal distribution with mean β_0 and variance σ_0^2, that is $\beta_{0i} \sim N(\beta_0, \sigma_0^2)$. We also assume that PL is known precisely, and do not account for order effects.

2.7.4 Fitting the Model

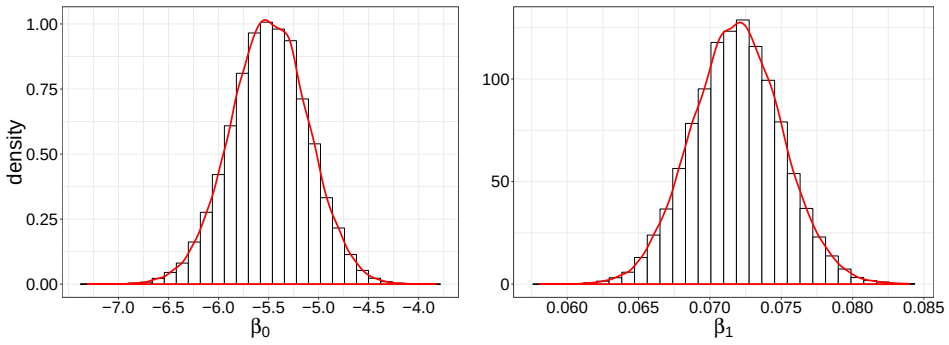
We draw 1000 burn-in samples and 40,000 additional samples. We also center the PL value when fitting the multilevel ordinal regression model. The traceplots and Gelman-Rubin plots both indicate convergence, and the autocorrelation plots do not show any problems with mixing.

2.7.5 Results

Table 2.8 shows summary statistics for β_0, β_1 and the γ parameters. The precision in the summary statistics is determined by the time-series standard error. The marginal posterior distributions of β_0 and β_1 are shown in Figure 2.20. Figure 2.21 shows each sampled vector of gamma parameters plotted against its mean. As with the non-multilevel ordinal regression results, the intervals are not estimated to be equally spaced, and there is more variation in the gamma values associated with the higher end of the response scale.

Table 2.8: Summary statistics of the multilevel ordinal regression parameters β_0, β_1 and all γ_k.

	Mean	SD	0.025 quant.	0.25 quant.	Median	0.75 quant.	0.975 quant.
β_0	-5.49	0.39	-6.26	-5.75	-5.49	-5.23	-4.74
β_1	0.072	0.003	0.066	0.070	0.072	0.074	0.078
γ_2	1.03	0.05	0.93	0.99	1.03	1.06	1.13
γ_3	1.68	0.06	1.56	1.64	1.68	1.72	1.80
γ_4	2.32	0.07	2.18	2.27	2.32	2.36	2.46
γ_5	2.77	0.07	2.62	2.72	2.77	2.82	2.92
γ_6	3.09	0.08	2.94	3.04	3.09	3.14	3.24
γ_7	3.42	0.08	3.26	3.37	3.42	3.48	3.59
γ_8	3.81	0.09	3.63	3.75	3.82	3.88	3.99
γ_9	4.25	0.10	4.05	4.18	4.25	4.32	4.45
γ_{10}	4.45	0.11	4.24	4.38	4.45	4.53	4.67



(a) (b)

Figure 2.20: Marginal posterior distributions of the multilevel ordinal regression parameters (a) β_0 and (b) β_1.

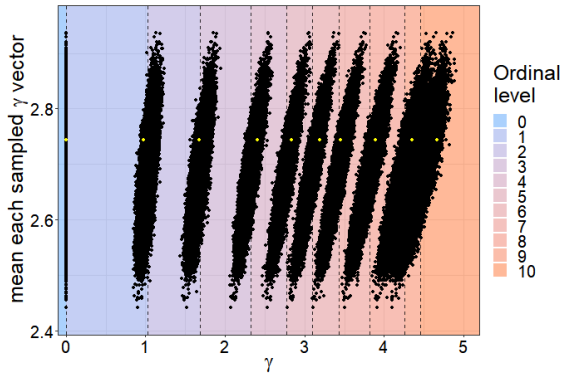


Figure 2.21: Posterior distribution of the multilevel ordinal regression γ parameters; dashed lines indicate posterior means of each gamma parameter, and yellow points indicate the gamma parameters for one random posterior draw.

We calculate a summary dose-response curve for multilevel ordinal regression following procedures outlined in Section 2.3.5, with $p_{ij} = 1 - \Phi(\gamma_8 - \beta_{0i} - \beta_1 * PL_{ij})$. Notice that we replaced β_0 in Eq. 11 with β_{0i} to calculate each individual's dose-response curve at each posterior draw for the grid of PL values. The summary dose-response curve for the multilevel ordinal regression model is shown in Figure 2.22.

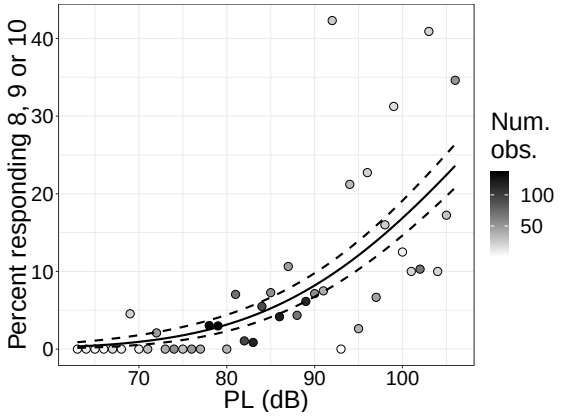


Figure 2.22: Multilevel ordinal regression summary dose-response curve estimate and 95% credible intervals.

2.7.6 Model Assessment

To assess model fit, we check the same discrepancy statistics as for other multilevel models—deviance, total proportion of highly annoyed responses, the 0.1 quantile and median PL at which highly annoyed responses occur, the mean number of highly annoyed responses per participants, the standard deviation of number of highly annoyed responses per participants, the maximum number of highly annoyed responses per participants and the total number of highly annoyed participants. We

again count any replicated responses that are 8, 9 or 10 as a highly annoyed response. Figure 2.23 shows the posterior predictive checks for the 0.1 quantile PL, median PL and the number of participants highly annoyed at least once. The remaining checks are in Appendix B. None of the checks indicate lack of fit. We also do not see the lack of fit in the PL distribution for highly annoyed responses as we saw with the non-multilevel specification. Note that the empirical quantile for the observed statistic in Figure 2.23c is 0.17.

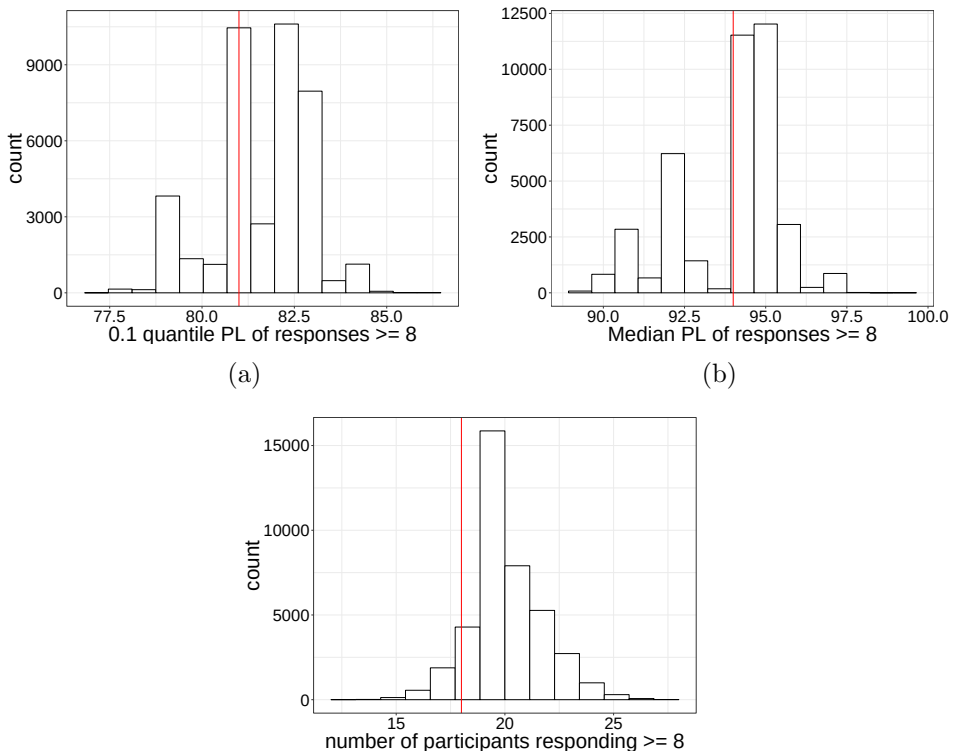


Figure 2.23: Posterior predictive checks for multilevel ordinal regression for (a) the 0.1 quantile PL and (b) median PL at which responses of 8, 9 or 10 occur, and (c) number of participants responding 8, 9 or 10 at least once; each histogram indicates discrepancy statistics calculated from replicated data at each posterior draw and red indicates the observed discrepancy statistics.

2.7.7 Model Comparison

The non-multilevel and multilevel ordinal regression models are compared using the DIC metric. Although the multilevel model structure introduces many more parameters, it fits the data better based on DIC and the posterior predictive checking results. The DIC for the non-multilevel ordinal regression model is 7412.6, and for the multilevel ordinal regression model is 5108.82. Recall that lower DIC indicates better relative fit to the data, hence the DIC suggests that the multilevel ordinal regression model fits the data better. From posterior predictive checking, we also saw

that the non-multilevel ordinal regression model does not fit well to the data for the two discrepancy statistics: the 0.1 quantile and median PL at which highly annoyed responses occur. On the other hand, the multilevel ordinal regression model does not show lack of fit in these two statistics. Figure 2.24 compares the two summary dose-response curves for the two ordinal regression models. The multilevel ordinal regression summary dose-response curve seems to follow the gradual increase in observed proportions of highly annoyed responses more closely than the curve from non-multilevel ordinal regression.

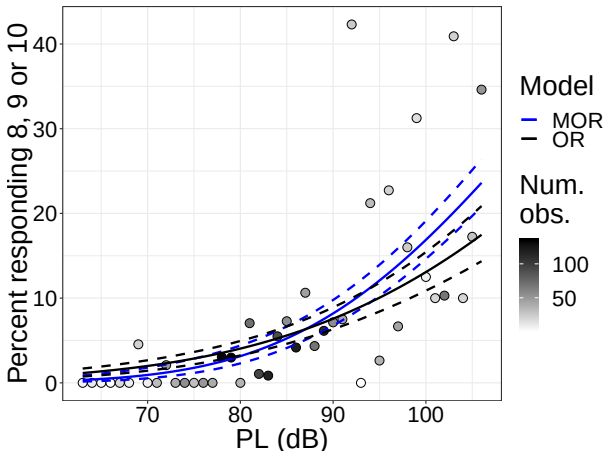


Figure 2.24: Comparison of non-multilevel and multilevel ordinal regression dose-response curves.

For all three model classes (logistic regression, CTL, and ordinal regression), the multilevel versions fit the data better than the non-multilevel versions. Figure 2.25 compares the three multilevel summary dose-response curves. The three curves are similar at low PL values, indicating the percent highly annoyed values predicted from the three models are similar. At high PL values, however, we see the multilevel CTL model estimates of percent highly annoyed are the highest among the three models. Meanwhile, the multilevel logistic regression and multilevel ordinal regression model estimates of percent highly annoyed remain similar. We cannot compare all three models using DIC because they are not all fit to the same data. The multilevel logistic regression and multilevel CTL models can be compared using DIC since both are fit to binary data, but the multilevel ordinal regression model is fit to ordinal data. Recall from Table 2.6 that the multilevel logistic regression model fit the data better than the multilevel CTL model.

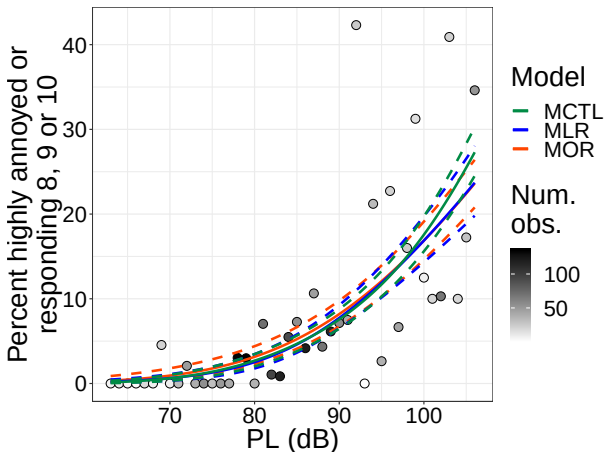


Figure 2.25: Comparison of the multilevel logistic regression, multilevel CTL and multilevel ordinal regression summary dose-response curves.

2.8 Model 7: Piecewise Linear Regression

For piecewise linear regression, we only consider the non-multilevel model due to its poor fit to the data and a few issues with the model. The model does not account for the longitudinal structure of the data.

2.8.1 Data

The piecewise linear regression model is proposed in the pilot study analysis report for cumulative exposure analysis (Page et al., 2014). We consider this model for the single-event data. This model is fit to proportions of highly annoyed responses as the response and PL as the covariate. At each PL, we calculate the proportion of survey responses that are 8, 9 or 10.

2.8.2 Model

The piecewise linear model is used when we think that the linear relationship between the response and the covariates changes for different regions of the covariates' range. The PL value at which the two linear models intersect is called the knot. Specifically in our case, we assume that the proportion of highly annoyed responses can be modeled by one linear model below the knot, and by another linear model above the knot. We expect the first segment to be flat (slope close to zero), and the second segment to have a larger slope than the first segment. This should look similar to a hockey stick. If we denote the proportion of highly annoyed responses P_i, then the model is

$$
\begin{aligned}
P_i|\mu_i, \sigma^2 \sim & \quad N(\mu_i, \sigma^2) \\
\mu_i|\beta_0, \beta_1, \beta_2 = & \quad \begin{cases} \beta_0 + \beta_1 * (PL_i - C) & \text{if } PL_i < C \\ \beta_0 + \beta_2 * (PL_i - C) & \text{if } PL_i \geq C \end{cases} \\
\beta_0 \sim & \quad N(0, 1) \\
\beta_1 \sim & \quad N(0, 0.0055^2) \\
\beta_2 \sim & \quad N(0, 0.0055^2) \\
\sigma^2 \sim & \quad InvGamma(0.01, 0.01) \\
C \sim & \quad N(85, 8.33^2)
\end{aligned} \tag{14}
$$

- Independent variable: noise dose in PL (dB)
- Dependent variable: proportion of highly annoyed responses
- Parameters to estimate: $C, \beta_0, \beta_1, \beta_2, \sigma^2$

This parameterization is borrowed from an example of piecewise linear regression in the OpenBUGS User Manual Examples Volume II (Spiegelhalter et al., 2014).[4] This parameterization models two parameters for the slopes of the two linear segments: β_1 and β_2. The knot is specified at $PL = C$, where the value of μ_i is β_0. So, the knot is at (C, β_0).

2.8.3 Specifying Informative Priors

We considered the following when specifying the informative priors.

- The prior on the knot C: the prior on C restricts the support to be in the observed PL range of about 60 to 110 dB and centered at the mean of the range. The standard deviation of 8.33 is chosen so that $\pm 3\sigma$ will cover the desired dB range
- The prior on the intercept β_0: at the knot, (C, β_0), the proportion should be restricted between 0 and 1. A reasonable prior on β_0 is normal with mean at 0 and variance of 1. We choose variance of 1 to restrict the prior on β_0 to be closer to the mean. For large variance, such as 100, there are many values for β_0 that would not be within the 0 to 1 range
- The priors on the slopes β_1, β_2: as described in Section 2.5, we make an educated guess that almost 100% of the population would be highly annoyed at 200 dB. Based on this information, we choose the priors on the slopes such that the largest possible slope values are included. If we consider the range for the knot, the steepest value for the slope would be if the knot was at 110 dB with the proportion of highly annoyed responses at 0. If that is the case, the slope between (110, 0) and (200, 1) is 0.011. This is an extreme case and the

[4] The example is titled "Stagnant: a change point problem."

maximum value for the slope given that the proportion of highly annoyed responses at 200 dB is 1 and the knot is between 60 to 110 dB. If the proportion of highly annoyed responses at the knot was greater than 0 or the knot was at a lower PL, then the slope will be smaller. So, for both slope parameters, we choose normal priors centered at 0 with $\pm 2\sigma = 0.011$. This results in standard deviation of 0.0055 or variance of 3.025e-5. If we draw a histogram of 10,000 random draws from $N(0, 0.0055^2)$, the support is nonzero for ± 0.02.

2.8.4 Assumptions

For the piecewise linear regression model specified, we assume all observations are independent and do not account for the multiple responses from each participant. In addition, we assume that there is only one knot and there is a linear relationship between proportion of highly annoyed responses and PL both below and above the knot. If we do not believe a linear relationship is reasonable, we could use higher order polynomials. We also assume that PL is known precisely, and do not account for order effects.

Because we model the proportion of highly annoyed responses as normally distributed, one obvious problem with this model is that the response is not bounded between 0 and 1. So extrapolation can lead to some values of proportion highly annoyed that do not make sense. Also, the model does not account for different sample sizes at each PL and gives equal weighting to all observed proportions.

2.8.5 Fitting the Model

We draw 1000 burn-in samples and 50,000 additional samples. The traceplots and Gelman-Rubin plots indicate convergence. The autocorrelation plots indicate that C and β_0 are highly autocorrelated but do not show signs of mixing issues.

2.8.6 Results

Table 2.9 shows the summary statistics for the parameters, and the precision in the summary statistics is determined by the time-series standard error. Figure 2.26 shows the marginal posterior distributions of $\beta_0, \beta_1, \beta_2$ and C.

Table 2.9: Summary statistics of the piecewise linear regression parameters $\beta_0, \beta_1, \beta_2$ and C.

	Mean	SD	0.025 quant.	0.25 quant.	Median	0.75 quant.	0.975 quant.
β_0	82	8	68	76	81	86	100
β_1	0.002	0.003	-0.006	-0.000	0.002	0.004	0.007
β_2	0.0073	0.0025	0.0020	0.0061	0.0074	0.0087	0.0117
C	0.04	0.05	-0.04	0.00	0.03	0.07	0.17

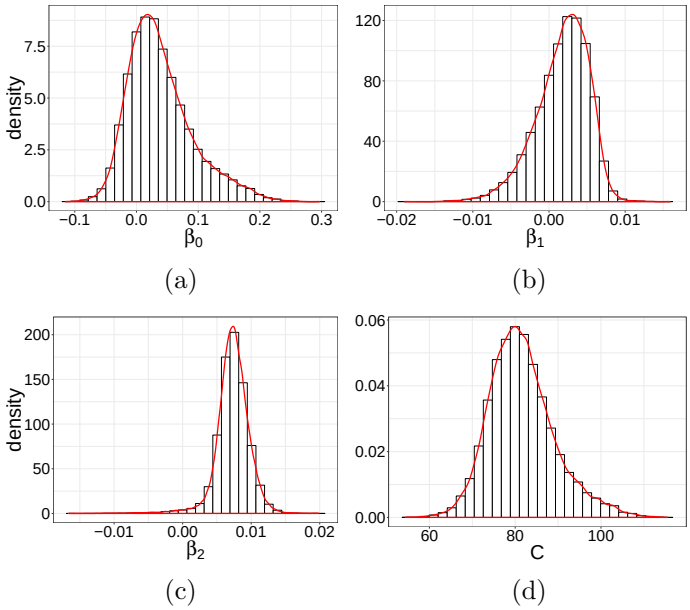


Figure 2.26: Marginal posterior distributions of the piecewise linear regression parameters (a) β_0, (b) β_1, (c) β_2 and (d) C.

The summary dose-response curve is calculated in a similar manner as for other non-multilevel models. We follow the procedures outlined in Section 2.2.5 and $p_i = E(P_i) = \beta_0 + \beta_1 PL_i$. Figure 2.27 shows the summary dose-response curve estimate and credible intervals. We see that the credible intervals reach into the negative values for the proportion of highly annoyed responses. We also see that the first segment to the left of the knot is not flat, which is different from what we hypothesized. This may be because we do not restrict the proportions to be greater than 0. Note that the summary dose-response curve does not exhibit the "hockey stick" shape that we expect for a piecewise linear regression model because taking pointwise posterior means as the curve estimate results in smoothing out the "hockey sticks" calculated from each posterior draw.

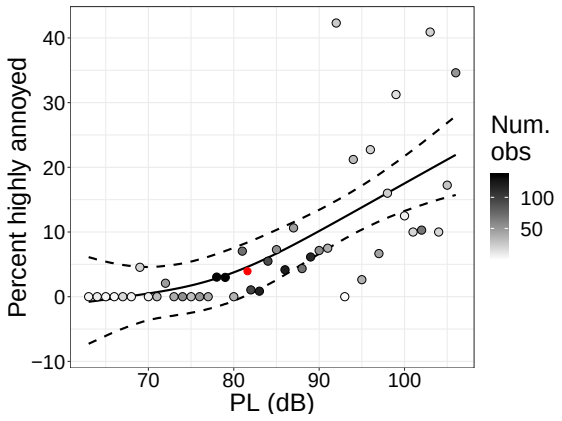
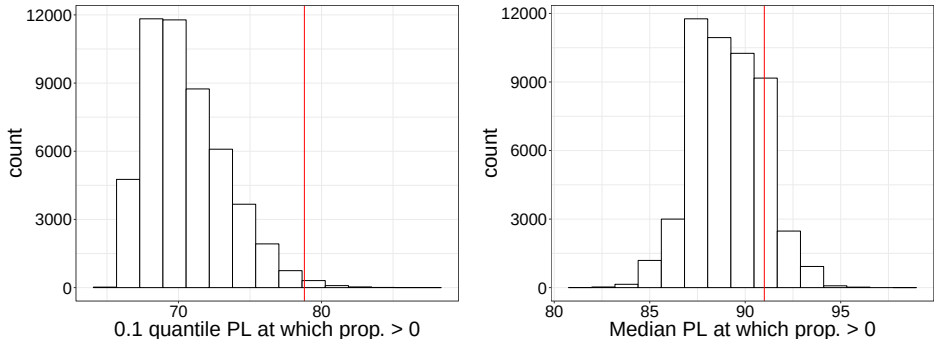


Figure 2.27: Piecewise linear regression summary dose-response curve estimate and 95% credible intervals; knot is estimated as (posterior mean of C, posterior mean of β_0) and indicated by red.

2.8.7 Model Assessment

We check model fit using the same four discrepancy statistics that we did for the other non-multilevel models—deviance, the 0.1 quantile and median PL at which highly annoyed responses occur, and the total proportion of highly annoyed responses. For the total proportion of highly annoyed responses, we multiply the replicated proportions at each PL by the sample sizes, then sum the products, and divide by the total number of observations. Figure 2.28 shows the posterior predictive checks for the 0.1 quantile and median PL. For these two posterior predictive checks, we consider at each PL whether the proportion of highly annoyed responses is greater than 0. If it is greater than 0, then there is at least 1 highly annoyed response. Otherwise, there are no highly annoyed responses at that PL. This method only considers the 44 PL values. The remaining two checks are in Appendix B. We only see lack of fit in the 0.1 quantile PL of highly annoyed responses, but not for the other three discrepancy statistics. This indicates that the model starts predicting highly annoyed responses at a lower PL than observed.



(a) (b)

Figure 2.28: Posterior predictive checks for piecewise linear regression for (a) the 0.1 quantile PL and (b) median PL at which proportions of highly annoyed responses are greater than 0; each histogram indicates discrepancy statistics calculated from replicated data at each posterior draw and red indicates the observed discrepancy statistics.

To illustrate the drawbacks of the model, we also compare the proportions of highly annoyed responses replicated from the model to the observed proportions across PL. We first choose 20 random posterior draws from which to replicate data and to calculate the replicated proportions. We then compare to the observed proportions. This is similar to checking multiple discrepancy statistics simultaneously; but instead, we only compare the observed values to replicated values from 20 random posterior draws.

Figure 2.29 compares the replicated and observed proportions. The observed data are represented by the green diamonds, and the replicated proportions are represented by the black points. The size of the black points indicate how many of the 20 draws have the same number of highly annoyed responses replicated. Note that the sample size at each PL is not displayed.

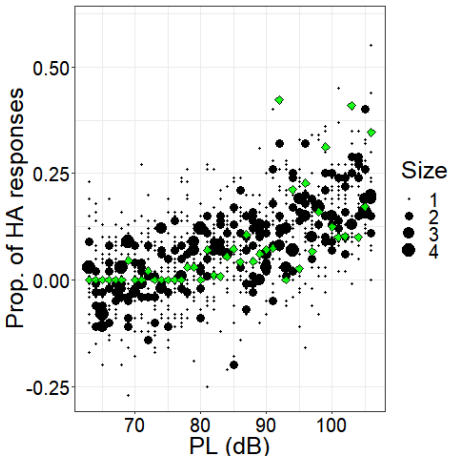


Figure 2.29: Comparison of replicated proportions of highly annoyed responses from piecewise linear regression model (black points) to observed proportions (green diamonds) at each PL; the replicated proportions are rounded to hundredths place to allow grouping.

From Figure 2.29, we see two issues with this model and make some suggestions for fixing them but we will not pursue this model further. First, some of the replicated proportions of highly annoyed responses are negative. We suggest to enforce a constraint on the proportion of highly annoyed values to be greater than or equal to 0. We can choose, for example, the truncated normal or gamma distribution instead of normal distribution. Secondly, the specified model assumes homoskedasticity (equal variance in proportion of highly annoyed responses across PL), but the data are heteroskedastic (unequal variance in proportion of highly annoyed responses across PL) as seen by the variation in the black points as we increase PL. The first suggestion is to try a data transformation. The second suggestion is to weight the observations based on estimated variance. In particular, the weights will be the inverse of the estimated variance.

2.8.8 Sensitivity Analysis

Because we specified informative priors in the model, we do a sensitivity analysis for the piecewise linear regression to check how changing the priors affects the posterior distributions and the results. The first analysis inflates the standard deviation of the β_1 prior by 3, and the second inflates the standard deviation of β_2 prior by 3. These are both less restrictive priors, so we expect to see some posterior draws farther out in the tails of the posterior distribution.

First, we change the prior for β_1 from $\beta_1 \sim N(0, 0.0055^2)$ to $\beta_1 \sim N(0, 0.0165^2)$. Figure 2.30 shows the approximate marginal distributions of the $C, \beta_0, \beta_1,$ and β_2 parameters for the original model in pink, and for the model with the inflated β_1 prior in blue. The posterior distributions do not have substantial differences from changing the prior for β_1.

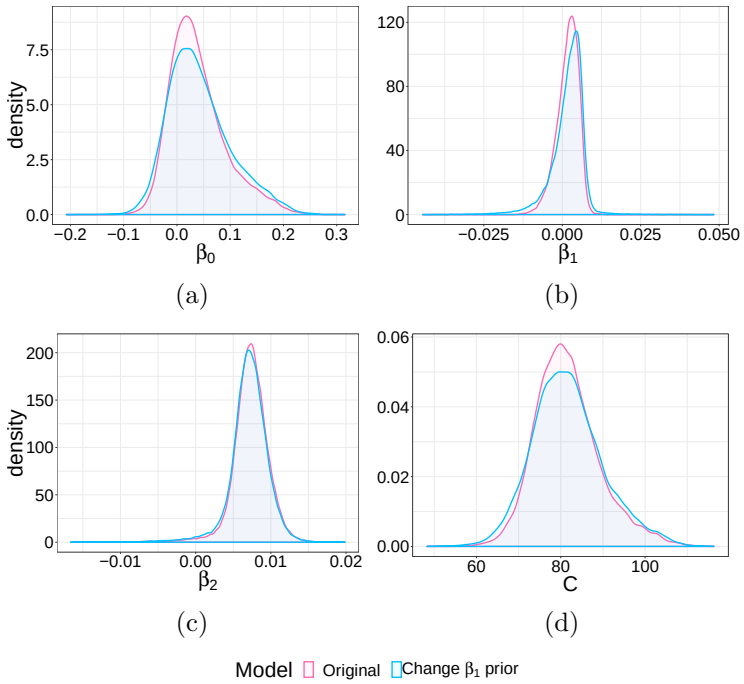


Figure 2.30: Comparison of marginal posterior distributions of the piecewise linear regression parameters (a) β_0, (b) β_1, (c) β_2 and (d) C from the original and inflated β_1 prior specifications.

The small changes in the posterior distributions do not substantially change the summary dose-response curve estimate, DIC, or posterior predictive check results. The DIC for the original model is -91.5, and for the model with an inflated β_1 prior is -91.03. Figure 2.31 shows that the summary dose-response curves are very similar. The knots are indicated by the red square and triangle, and are estimated to be at similar points. Figure 2.32 compares the posterior predictive checks for the original model and the model with inflated β_1 prior. The two checks are for the 0.1 quantile and median PL of highly annoyed responses. We see that the model fit results are the same after inflating the β_1 prior, with lack of fit for the 0.1 quantile PL of highly annoyed responses but not for median PL.

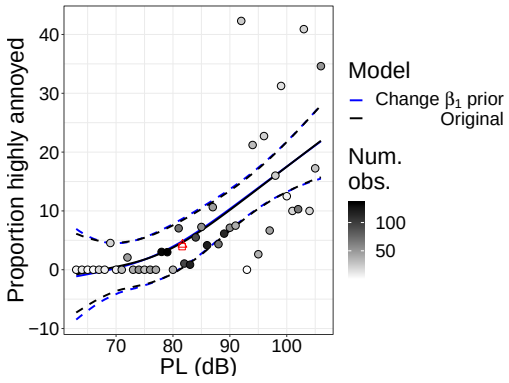
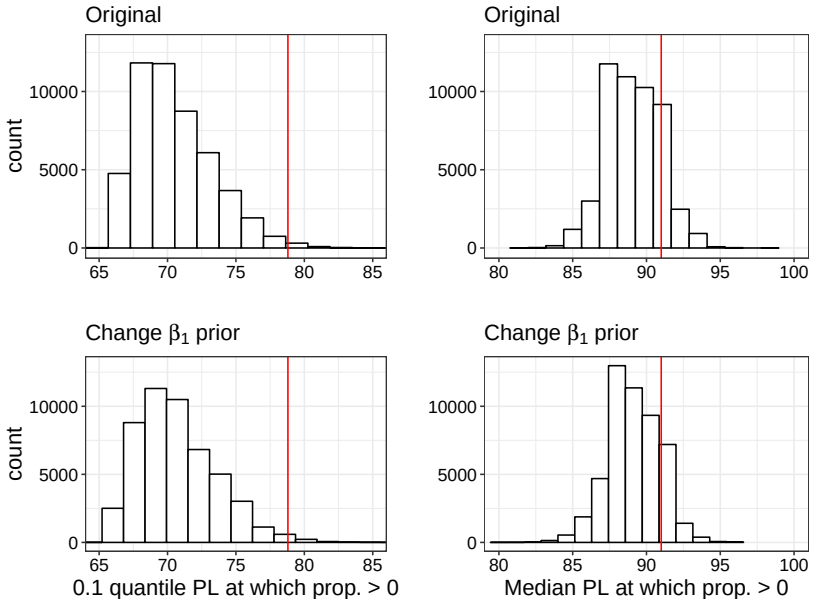


Figure 2.31: Comparison of piecewise linear regression summary dose-response curve estimates after inflating β_1 prior; knots indicated by red, square for original model and triangle for inflated β_1 prior



(a) (b)

Figure 2.32: Comparison of posterior predictive checks for the original piecewise linear regression model (top row) and the model with inflated β_1 prior (bottom row) for (a) the 0.1 quantile PL and (b) the median PL at which the proportion of highly annoyed responses exceeds 0; each histogram indicates discrepancy statistics calculated from replicated data at each posterior draw and red indicates the observed discrepancy statistics.

Next, we change the prior for β_2 from $\beta_2 \sim N(0, 0.0055^2)$ to $\beta_2 \sim N(0, 0.0165^2)$ (note that the variance of the β_1 prior is not inflated). Figure 2.33 compares the approximate marginal posterior distributions of the C, β_0, β_1, and β_2 parameters be-

fore and after inflating the variance of the β_2 prior. We again see only slight changes in the posterior distributions. The estimated summary dose-response curve, DIC, and posterior predictive check results do not substantially change, so we do not show the plots again. The DIC for the model after changing the β_2 prior specification is -91.29, and for the original model specification is -91.55. Because of the problems identified with the non-multilevel specification, we do not pursue a multilevel version of the piecewise linear model.

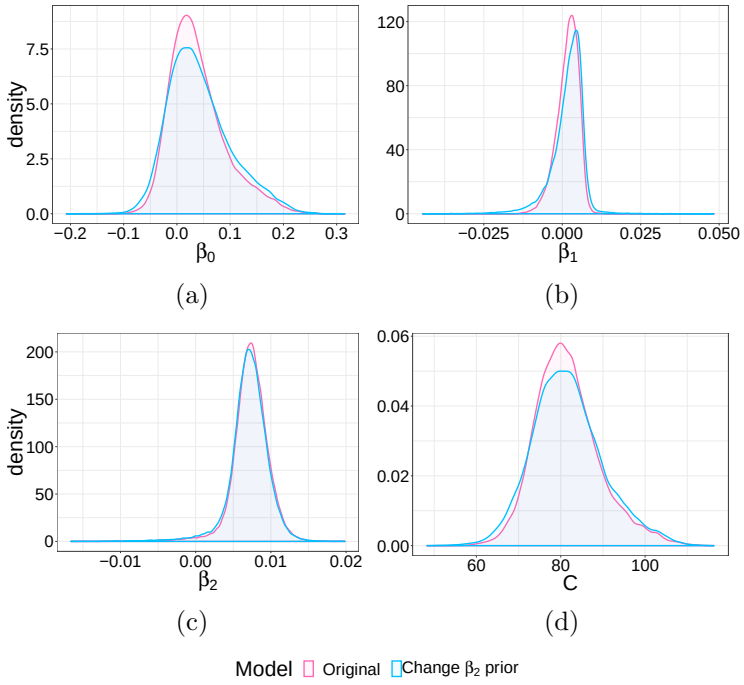


Figure 2.33: Comparison of marginal posterior distributions of the piecewise linear regression parameters (a) β_0, (b) β_1, (c) β_2 and (d) C from the original and changed β_2 prior specifications.

After fitting all the candidate models, we need to select the best models. We will use the subset of best models to make recommendations relevant to sample size, and to investigate the effects of a reduced noise dose range in future studies.

3 Results

3.1 Selection of Models

After fitting all candidate models and assessing model fit, we select the best models based on posterior predictive checking and DIC to investigate the effects of a reduced noise dose range and to calculate sample sizes for subsequent experiments. For models that are fit to the same data (for example, ordinal responses and binary responses are not the same data), we can use DIC for comparing the relative fit of the models. Based on the process described below, we arrive at the multilevel logistic regression and multilevel ordinal regression models as our two best models.

We eliminate piecewise linear regression because it does not model the heteroskedasticity in the data nor does it bound the proportion of highly annoyed responses to the reasonable range of 0 to 1. The posterior predictive checks also indicate lack of fit for the 0.1 quantile PL at which highly annoyed responses occur.

Next we compare the non-multilevel and multilevel versions for the other three model classes: logistic regression, CTL and ordinal regression. For the logistic regression models, the posterior predictive checks for neither model shows lack of fit. Based on DIC, the multilevel model (DIC = 510.5) fits the data better than the non-multilevel model (DIC = 863.77). Recall that lower DIC indicates better relative fit.

For the CTL models, we also see that the multilevel specification fits better than the non-multilevel model based on DIC (multilevel CTL DIC = 534.4 and non-multilevel CTL DIC = 925.07). The posterior predictive checks for non-multilevel CTL indicate that the model predicts high annoyance at a higher PL than observed. The posterior predictive checks for multilevel CTL do not show lack of fit for these aspects, but indicate lack of fit for the number of participants highly annoyed at least once. Based on DIC and the posterior predictive check for the number of participants highly annoyed at least once, we conclude that the multilevel logistic regression model fits the data better than the multilevel CTL model.

The last class of models that we consider is ordinal regression. As with other classes of models, the relative fit of the multilevel version (DIC = 5108.82) is better than the non-multilevel model (DIC = 7412.6). From the posterior predictive checks on the 0.1 quantile and median of the PL distribution for highly annoyed responses (see Figure 2.19), we see that the non-multilevel ordinal regression model predicts highly annoyed responses starting at a lower PL than observed. None of the posterior predictive checks for the multilevel ordinal regression indicate lack of fit.

We cannot compare DIC for the multilevel logistic and multilevel ordinal regression models because they are fit to different data. However, we can compare the fits based on posterior predictive checks, neither of which indicate substantial lack of fit for either model.

To summarize the model selection process, we first eliminate the piecewise linear regression model due to its poor fit to the data. Then we compare the multilevel and non-multilevel models within each model class (logistic regression, CTL and ordinal regression). From the comparisons, we find that all multilevel specifications fit better than their non-multilevel counterparts based on DIC. From the three

multilevel models, we downselect to multilevel logistic regression and multilevel ordinal regression as the best models because the multilevel CTL model overpredicts the number of participants highly annoyed at least once. Figure 3.1 compares the summary dose-response curves of the two models, which overlap for most of the dose range in the plot.

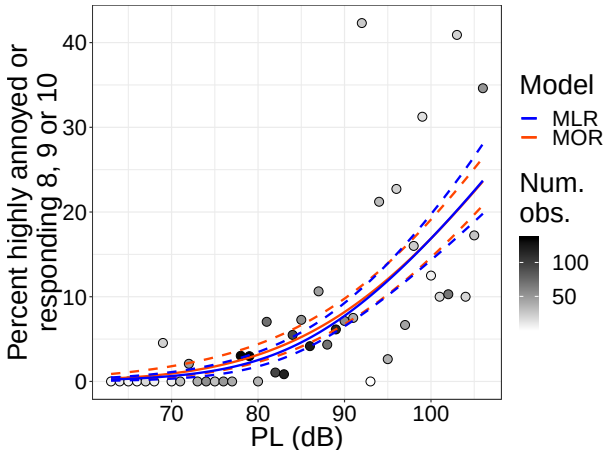


Figure 3.1: Comparison of multilevel logistic and multilevel ordinal regression summary curves.

3.2 Reduced vs. Full Range Analysis

Since there are limitations to the noise doses that can be achieved by X-59, the range of noise doses for the X-59 tests may be smaller than that of the pilot study. To understand the impacts of testing in a limited dose range, we compare the estimates from fits to the full dataset and a subset with a reduced dose range from the 2011 pilot study. The full range analysis consists of all the responses from the study, whereas the reduced range analysis considers responses between 70-80 dB only. The selection of the 10 dB range is based on initial estimates around the design guideline for the X-59 noise dose, which is 75 dB. The 10 dB range should be changed based on the actual limitations of the X-59 when that information is available. Ideally, the range should include the presumed noise limit in order to avoid extrapolation.

To compare the analysis using the full and reduced range data, we refit the two best models (the multilevel logistic and ordinal regression models) to the reduced range data. Figure 3.2 shows the distribution of the ordinal responses within the reduced range only. Figure 1.2 shows the distribution of the ordinal responses at each PL. There are only 9 out of 609 responses (1.5%) that are considered highly annoyed in the reduced range data, and 133 out of 1981 (6.7%) responses that are considered highly annoyed in the full range data. Notice that the reduced range also has a smaller sample size. So for this analysis, both the range and the sample size vary.

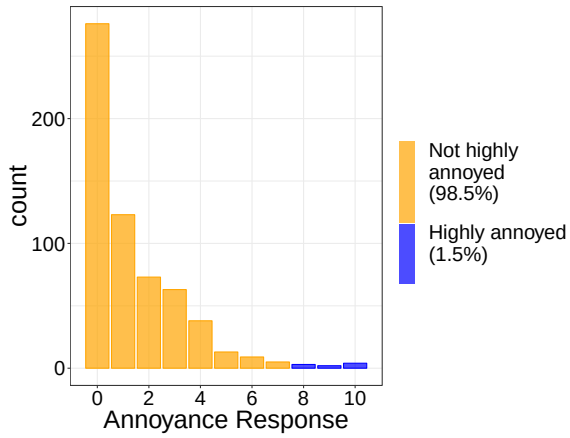


Figure 3.2: Distribution of survey responses within 70 to 80 dB.

After refitting the models to the reduced range data, we estimate six quantities of interest and compare the credible intervals for those quantities. The quantities of interest are based on the presumed uses for the summary dose-response curve. One use would be to estimate the percent highly annoyed for a fixed noise exposure. This approach would be used if the noise regulation were based on public perception of a particular noise dose, such as the target noise dose of the X-59. An alternate approach would be to fix the percent highly annoyed at some value and estimate the associated noise dose. A hypothetical percent highly annoyed value would be 12.3%, which corresponds to the limit of 65 dB in day-night average sound level (DNL) reaffirmed by Federal Interagency Committee on Noise (1992) as the threshold of significant cumulative noise exposure around airports. Since either approach is plausible, both are presented here. We select the following six quantities of interest, for which we will calculate point estimates and credible intervals:

- percent of highly annoyed responses at PL of 70, 80 and 82 dB
- PL for 1, 1.5 and 15 % highly annoyed

For both sets of quantities, we specifically choose two values that do not require extrapolation for all four models, and one value that does. The reduced range fits include data from 70 to 80 dB, so extrapolation is required only for estimating percent highly annoyed at 82 dB. Extrapolation is also required when estimating PL at 15% highly annoyed.

3.2.1 Computation

For fitting the two models to the reduced range data, we use the same priors described in Section 2. We check the traceplots and Gelman-Rubin plots to determine both the number of burn-ins and iterations needed for reaching convergence. We then compare the fits of the reduced and full range data for each model visually. We also assess the model fits using posterior predictive checks. We check the mean number of highly annoyed responses per participant, the standard deviation of highly

annoyed responses, and the total number of participants highly annoyed at least once in addition to the four discrepancy statistics for the non-multilevel models.

The posterior predictive checking indicate that the multilevel ordinal regression model fit to the reduced range data predict high annoyance at a lower PL than observed. This was not the case for the multilevel ordinal regression model fit to the full range data. The posterior predictive checking for the multilevel logistic regression model fit to the reduced range data do not show lack of fit.

The procedures for calculating the point estimates and credible intervals for percent highly annoyed values given PL are similar as for calculating the summary dose-response curve for the grid of dose values (see Section 2.2). Instead of calculating at a grid of 1000 PL values, we only calculate at 3 fixed PL values. We calculate at each posterior draw the probability of high annoyance by first calculating each individual's probability and averaging them. Then we calculate the posterior median[5] for the point estimate and quantiles for the credible intervals. To calculate each individual's probability, we calculate $p_{ij} = \beta_{0i} + \beta_1 * PL_{ij}$ for multilevel logistic regression and $p_{ij} = 1 - \Phi(\gamma_8 - \beta_{0i} - \beta_1 * PL_{ij})$ for multilevel ordinal regression.

To calculate the PL given percent highly annoyed, we use linear interpolation between points on the summary dose-response curve. For example, to calculate the PL at 1% highly annoyed, we iteratively search through the summary dose-response curve to find the two PL values that bound 1% and linearly interpolate between them. Since the summary curve is defined by 1000 points, the interpolation interval is 0.043 dB. This process is then repeated for every posterior draw to obtain a distribution of PL values corresponding to 1% highly annoyed. We then take the median and sample quantiles for the point-estimate and credible intervals.

3.2.2 Observations

Figures 3.3 and 3.4 show the width differences in the full range and reduced range credible intervals for the six quantities listed above. First, the width of the credible intervals are calculated for each quantity, for both the reduced and full range fits. Then, the credible interval width from the full range fit is subtracted from the credible interval width from the reduced range fit. In both Figures 3.3 and 3.4, the credible interval width differences are greater than zero for all six quantities. In other words, there is larger uncertainty in the estimates from the reduced range analysis than from the full range analysis across all four models for all six quantities. The increased uncertainty of the estimates seen in the reduced range analysis could be due to the decrease in the noise dose range, the decrease in the sample size in the reduced range, or both. To isolate the effect of the reduced range only, we suggest conducting a simulation study where the reduced and full range have the same sample sizes. Not surprisingly, extrapolating at values outside of the observed data range (i.e., calculating percent highly annoyed at 82 dB and calculating the PL at 15% highly annoyed for the reduced range) typically leads to a larger increase in the uncertainty in the model estimates. We therefore recommend including the presumed noise limits in the future X-59 tests.

[5]Note that we calculate the posterior mean instead of median for the summary dose-response curves.

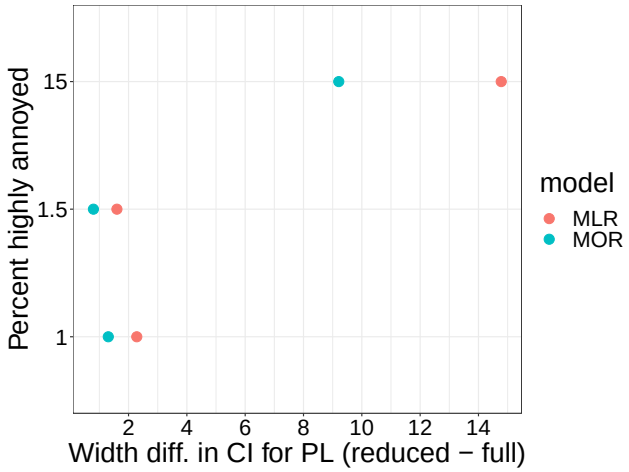


Figure 3.3: Differences in widths of the full range and reduced range credible intervals for PL (dB) given percent highly annoyed for the multilevel logistic regression and the multilevel ordinal regression models.

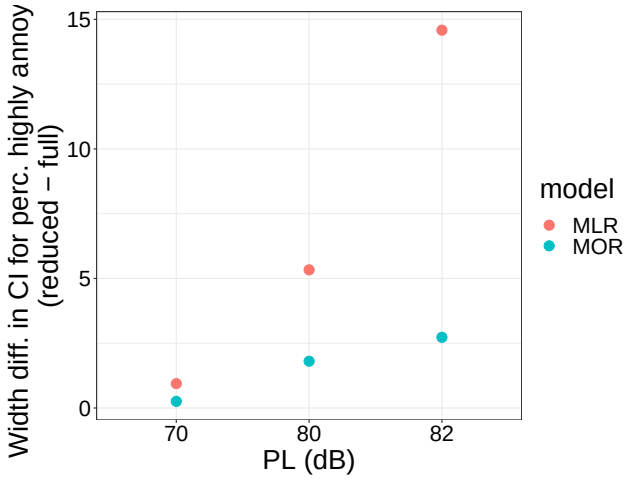


Figure 3.4: Differences in widths of the full range and reduced range credible intervals for percent highly annoyed given PL for the multilevel logistic regression and the multilevel ordinal regression models.

3.3 Sample Size Calculations

We anticipate that panel sampling will be used for the future X-59 community tests, and one important aspect for planning the tests is how many participants to recruit. To determine the sample size, we need to decide on the objective first, which can be either estimation or testing. Here we will consider estimation, which focuses on how many samples are needed to estimate a quantity of interest with a set level of

precision[6]. As previously mentioned, the design guideline for the X-59 noise dose is 75 dB PL. So, one reasonable sample size criterion is estimating the percent highly annoyed at 75 dB PL with some desired precision, which we choose to be 1%. We use this sample size criterion to demonstrate two sample size calculations, but the quantity of interest and desired precision can be changed easily. The sample size calculations are carried out using simulated data from the models fit in Section 2. We consider a simple case first without missing data, and then a second case that simulates missing responses from some survey participants.

We first introduce some notation, then describe the procedures and discuss the results. The procedures and notations follow those outlined by Wang & Gelfand (2002).

3.3.1 Notation

- θ is the vector of parameters in the model. For example, $\theta = (\beta_0, \beta_1)$ for non-multilevel logistic regression.
- $f^{(s)}(\theta)$ is the sampling prior; the prior for sampling θ. For example, we use the posterior draws from fitting the models in Section 2 as the sampling priors. Alternatively, we can approximate the posterior distribution to use as the sampling prior. For example, approximate the posterior distribution using a multivariate normal or t distribution and draw random samples from the approximated distribution instead of drawing a random sample of the posterior draws.
- $f^{(f)}(\theta)$ is the fitting prior; the prior for fitting the model once we generate data. For example, the fitting priors are the priors specified for each model in Section 2.
- $y^{(n)}$ is a vector of responses of length n. For example, a vector of 0 and 1 values for the binary models, or a vector of 0 to 10 values for the ordinal models.
- $T(y^{(n)})$ function of interest; the general form of the sample size criterion is $E[T(y^{(n)})] \leq \epsilon$ (minimum sample size to satisfy this criterion). For example, the function of interest $T(y^{(n)})$ can be the credible interval length for a specific quantity we want to estimate. Then, $E[T(y^{(n)})]$ is the expected credible interval length.
- n is sample size
- superscript $*$ will indicate generated values. For example, θ^* is set of generated parameters

[6]On the other hand, if the objective was testing, then we would focus on how many samples are needed to detect a change or difference in a quantity of interest. A common example of a testing problem is a power analysis.

3.3.2 Computation

The procedures for a very general case are described below.

1. Generate θ^* from $f^{(s)}(\theta)$
2. Generate a single set of new responses $y^{(n)*}$ from $f(y^{(n)}|\theta^*)$
3. Fit model to $y^{(n)*}$ using fitting priors $f^{(f)}(\theta)$ and calculate $T(y^{(n)*})$
4. Repeat steps 1-3 B times; B should be a large number (to have higher precision)
5. Approximate $E[T(y^{(n)})|y^{(n)} \sim f^{(s)}(y^{(n)})] \approx \frac{1}{B}\sum_{i=1}^{B} T(y^{(n)*})$
6. Repeat 1-5 for multiple values of n
7. Find minimum n that satisfies the sample size criterion

For each n, we simulate $B = 5000$ times to demonstrate how to carry out the sample size calculation. We consider the posterior distributions from fitting the model to the original data as our sampling prior. To generate new θ^*, we randomly sample one of the posterior draws. The fitting priors are the same as those specified in Section 2.

For the sample size criterion, we use the average length criterion, which means we find the minimum sample size required for the expected length of the credible intervals of percent highly annoyed at 75 dB to be less than or equal to a prespecified value, l. So $T(y^{(n)*})$, the function of interest that we calculate for each of the B simulations, is the credible interval length. Note that a larger value of B will result in a better approximation of the expected credible interval length. Our sample size criterion is then

$$E[F^{-1}_{p|y^{(n)}}(1-\alpha/2) - F^{-1}_{p|y^{(n)}}(\alpha/2)] \leq l, \tag{15}$$

where p is the probability of high annoyance at 75 dB, $F^{-1}_{p|y^{(n)}}$ is the inverse CDF of p given sample $y^{(n)}$, l is the prespecified length or desired precision, and α is the prespecified probability. Note that $F^{-1}_{p|y^{(n)}}(q)$ is the q quantile. Wang & Gelfand (2002) discuss other choices for the sample size criterion.

3.3.3 Simulation Design

For this sample size simulation, we assume that all participants receive the same noise exposure at exactly the planned levels. This is usually not the case because the noise exposure of a sonic boom varies based on many factors, such as the location of the participants. In addition, the actual precision on the estimates of the exposure levels is not yet known.

For the noise dose design, we use the design from the 2011 pilot study, which has a total of 87 booms: 29 booms at each of the 3 noise doses of 66, 83 and 90 dB PL. We first consider a simplified case where there are no missing data by making the

assumption that all participants recruited will respond to all booms throughout the test. This is likely not the case, and so we consider a second case with nonresponse. The results from the two cases are then compared.

For the second case with nonresponse, we estimate the response rates based on the current dataset. Because response rates varied with the sonic boom level, we estimate a separate response rate for low, medium and high level booms. At low levels, there was a lower response rate likely due to participants not hearing the boom. The boom levels for the observed data were categorized based on peak overpressure of the acoustic pressure-time history. Peak overpressure is the maximum pressure above the normal atmospheric pressure caused by a shock wave (Segarra et al., 2010). The peak overpressure for a low level boom in the current dataset is between 0.03 and 0.23 psf, for a medium boom is between 0.23 and 0.43 psf, and for high boom is between 0.43 and 0.63 psf (Page et al., 2014). The midpoints of the ranges correspond roughly to levels of 66, 83 and 90 dB PL. In the 2011 study, some booms measured peak overpressure greater than 0.63 psf, which were categorized as "higher" and "highest" levels and we grouped them with the high level booms.

Figure 3.5 shows the response rate distributions for low, medium, high and all booms for the current dataset. As we expected, the response rate distributions for the three level booms are not the same. We use the medians, indicated by the red lines in Figure 3.5, to estimate the response rates at each of the three levels. If we choose to be more conservative, we could, for example, choose the 0.1 quantile of the response rate distributions. When considering the simulation with nonresponse, we assign the same response rate to each participant at each of the three boom levels. This means the response rates vary by level of boom, but not by participant. The median response rates are 0.17, 0.44 and 0.45 for low, medium and high booms, respectively. When simulating the responses in Step 2 outlined in Section 3.3.2, we generate (median response rate * 29) responses for each boom level for each participant. Then, each participant is simulated to respond to 5 low, 12 medium and 12 high booms for a total of 29 responses per participant. All other steps remain the same.[7]

[7]Another method is to use the median response rate as a probability of occurrence and generate Bernoulli random variables for whether to simulate a response for each of the 87 responses for each participant. However, this method will result in slightly different number of responses per participant.

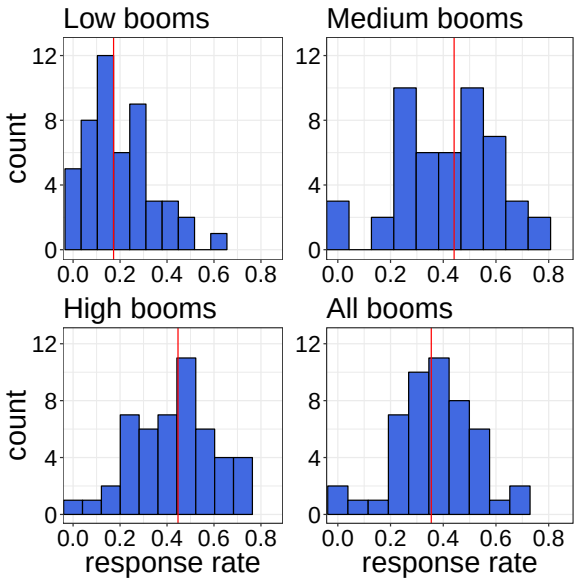


Figure 3.5: Response rates estimated from participants in pilot study separated by low, medium, high and all booms; red indicates median.

For the sample size, we consider increments of 10 between 10 and 90 participants for the no missing data case, and increments of 10 between 10 and 250 participants for the nonresponse case. We have not completed the calculations for the multilevel ordinal regression model for the second case so results are only shown for the multilevel logistic regression model. Our sample size criterion is the expected 95% credible interval length for percent highly annoyed at 75 dB PL, and our desired precision is 1% highly annoyed. Figure 3.6 shows the estimated expected credible interval lengths for the multilevel logistic regression model plotted against the sample sizes we consider. The criterion of 1% is indicated by the black horizontal dashed line. We approximate the sample sizes where each each of the two curves intersect the black dashed line.

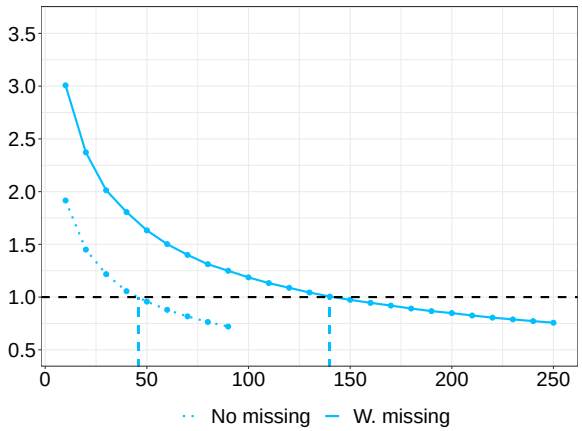


Figure 3.6: Sample size calculation without and with missing data using the multilevel logistic regression model fit to simulate data (for the with missing data case, all participants are assigned the same response rates, which are the medians of estimated response rate distributions for low, medium and high level booms); sample size criterion is expected length of the 95% credible interval of percent highly annoyed at 75 dB less than or equal to 1%.

For the case with missing data, the estimated number of participants required to reach 1% precision in the credible intervals of percent highly annoyed at 75 dB PL is 140 participants for the multilevel logistic regression model. For the case without missing data, the required sample size is 47 participants for the multilevel logistic regression model. By simulating nonresponses, we see that we need more participants to reach the same desired precision, specifically, about three times more participants. This is not too surprising because the number of responses we simulate per participant is now one third of the number for the first simulation (29 versus 87 responses). The total number of responses from 47 participants for the case without missing data is 4089 (87 booms * 47 participants), and the total number of responses from 140 participants for the case with 29 responses per participant is 4060 (29 booms * 140 participants). So, we see that the total numbers of responses required to achieve the sample size criterion are similar for the two cases, but the number of participants required greatly increases because of nonresponses.

These two simulations suggest that the number of required responses are similar when we assume no missing data and introduce nonresponse by assigning the same response rate to each individual. The number of participants increase by about 3 times when we simulate each participant to respond to one third of the 87 booms. However, this could be due to the fact that we control for each subject to have the same response rates and the same number of responses. We would need to include varying response rates per participant to simulate a more realistic scenario.

For this simulation, the sampling priors used to simulate the data are from the multilevel logistic regression model fit. More specifically, the simulated data for the multilevel logistic regression model are longitudinal binary data. To calculate the minimum sample size for the multilevel ordinal regression model and to compare the results between the two models, we would need to use the same simulated

longitudinal ordinal data. We suggest to generate data using the posterior draws of the multilevel ordinal regression model because the generated data will be similar to the observed data. Also, we can fit both the multilevel logistic regression and the multilevel ordinal regression models to the ordinal data whereas the multilevel ordinal regression model cannot be fit to binary data.

4 Discussion and Future Work

In this document, we explored:

1. how to model longitudinal data from sonic boom community response surveys using a Bayesian approach
2. how to derive a summary dose-response curve representative of the population from a multilevel model
3. how to assess and compare models, and to make recommendations for selecting best model(s) using posterior predictive checking and DIC
4. some effects of modeling single-event survey data without accounting for the longitudinal nature
5. some effects of having a smaller noise dose range in future surveys
6. the procedures for calculating sample size for subsequent surveys

There are multiple methods proposed in the community noise literature for modeling annoyance data from community noise surveys. Much of the literature is focused on cross-sectional data whereas our sonic boom community survey data are collected longitudinally. To model these types of data, we consider four model classes: logistic regression, the first-principles based Community Tolerance Level (Fidell et al., 2011), ordinal regression and piecewise linear regression. We consider modeling the binary response for highly annoyed or not as there is precedent in the community noise literature to use "percent highly annoyed" as the impact factor of noise. We consider the binary response for the non-multilevel and multilevel versions of the logistic regression and Community Tolerance Level (CTL) models. CTL specifies $p = e^{-c/m}$ where m is a transformation of PL. We also consider modeling the proportion of highly annoyed responses as a continuous response without accounting for the sample sizes for the piecewise linear regression model. Lastly, we consider modeling the ordinal responses from the annoyance surveys using ordinal regression with both non-multilevel and multilevel specifications.

When considering models that account for the multiple responses from each participant, we choose to use a multilevel model rather than a marginal model. A marginal model uses a correlation structure to model the correlation among the multiple responses, and estimates the mean responses for the population. A multilevel model uses individual-level parameters to model the correlation among the multiple responses, and estimates the mean response conditional on the individual. We prefer a multilevel model over a marginal model because it naturally fits the nested structure we expect for the X-59 data: multiple responses from individuals of multiple communities. For these data, we can calculate each participant's dose-response curve because the multilevel model enables estimation of individual-level parameters. In order to estimate a population representative summary dose-response curve for this community, we calculate a pointwise average of all participants' dose-response curves. We could also consider a marginal model since it models the population on

average. We could compare the summary dose-response curve estimates, estimates on the two types of quantities and sample size calculations from the marginal and multilevel models.

For complicated models such as the multilevel models proposed in Section 2, using a maximum likelihood approach to estimate model parameters can become a complicated optimization problem. Instead, we use Markov chain Monte Carlo (MCMC) in conjunction with Bayesian modeling to estimate model parameters and the associated uncertainties. We use the MCMC sampling software Just Another Gibbs Sampler (JAGS) to fit most of our candidate models. After fitting each model, we check the model fit using posterior predictive checking and compare models using DIC. All multilevel models fit the data better than their non-multilevel counterparts based on DIC. The multilevel logistic regression model is the best of the four binary models and the multilevel ordinal regression model is better than the non-multilevel ordinal regression model based on DIC. The two best models are the multilevel logistic regression and the multilevel ordinal regression models.

We use the multilevel logistic and the multilevel ordinal regression models to estimate quantities that support setting noise regulations to investigate the effects of a reduced noise dose range. We expect a smaller noise dose range relative to the pilot study in future X-59 surveys due to limitations of the vehicle. The two types of quantities of interest are percent highly annoyed given PL and PL given percent highly annoyed. The type of quantity to estimate depends on the method for setting noise regulations. The first method is to fix the PL and estimate what percentage of the population would be highly annoyed, while the second method is to fix the percent highly annoyed and estimate the corresponding PL.

When planning for future community surveys, we suggest including the presumed noise limit or noise doses of interest in the test design to avoid the need to extrapolate, which leads to increased uncertainty in the estimates of the quantities described. Since the reduced range resulted in both a smaller noise dose range and sample size, we suggest conducting a simulation study where the reduced and full range data have the same sample size to isolate the impact of a smaller noise dose range on the credible interval estimates.

We then demonstrate calculating sample size for future surveys. We consider two cases: the first without missing data and the second with missing data due to nonresponse. We find that the total number of responses required is similar for the two cases, but the number of participants required increases by about three times due to nonresponse. The increase by a factor of three reflects the fact that each participant is simulated to respond to only one-third of the planned booms in this missing data case.

From the downselection of candidate models, we arrived at the two best models, which are similar based on the visual analysis of their summary dose-response curves. Comparisons of model estimates, such as PL for a fixed percent highly annoyed, could shed light on how different the models estimate. In addition, a simulation study could help investigate potential biases of the two models.

In the analysis presented, we assume the dose is fixed and known precisely, but the noise doses are usually estimated from a combination of acoustic measurements and predictions. For the X-59 tests, the noise doses cannot be measured exactly

because the test area covers hundreds or thousands of square miles. There are limited resources for setting out microphones to make acoustic measurements in such a large test area. Therefore, there is currently unknown precision in the estimated noise doses for our survey participants. A possible extension to the statistical models is to model the uncertainty in the dose as well to understand how the uncertainty in the noise doses affects the summary dose-response curve estimate, the quantities of interest and the sample size calculations.

In Section 3, we consider two cases for calculating sample size. We first assume no missing data, then simulate missing data due to nonresponse. When simulating nonresponse, we assign the same nonresponse rates to all individuals, but this is likely not true. To make this simulation more realistic, we can assign a different response rate to each participant, which would mean the response rates vary by both boom level and by participant. In addition, there may be other reasons for nonresponse, such as attrition, and this is another element that can be added to a more realistic case for the sample size simulation. The missing data case assumes the data are missing at random, but this is likely not true. The patterns of missingness will require further analysis.

To compare the minimum sample sizes from the two models, we need to use the same simulated data for all models. We suggest to simulate data using the multilevel ordinal regression model because this model generates ordinal data, which is the same type of response as the observed annoyance responses, and both the multilevel logistic regression and multilevel ordinal regression models can be fit to this type of simulated data. We would want to compare results from the two models to check whether they are similar. We would not use the sample size simulation results to choose among the candidate models.

We focus on the single-event data from the pilot study for this document, but we can also model the cumulative data from the end-of-day annoyance survey in the future. For the cumulative data, each participant is asked to respond once each day of the test period. The models presented are still applicable because the participants are asked to respond to multiple end-of-day annoyance surveys as well. The cumulative data are comparable to other noise studies, such as for airport or railroad noise surveys. We can consider calculating sample size using the cumulative data, and make sample size suggestions for future surveys using both types of data.

Aside from the data used for the analysis presented, there are two other datasets available. The first is from the same 2011 pilot study but the participants on the panel are different, and the survey response scale is 5-point scale rather than 11-point. The number of participants on the panel is 48. The two 2011 pilot study datasets are not combined due to the different response scales. The second dataset is from a second pilot study conducted in 2018 in Galveston, Texas. The survey for the Galveston test also uses a 5-point scale. The number of participants recruited is about 500. In the future, we would like to fit all candidate models to the additional two datasets to compare results, especially for the choice of best models. We can also consider methods for combining the datasets, such as adding a third level in the multilevel model hierarchy. We presume the data from future X-59 community tests will be combined because the goal is to estimate a dose-response curve representative of the entire U.S. population based on the surveys from multiple communities.

References

Agresti, A. (2003). *Categorical data analysis.* Wiley.

Albert, J. (2007). *Bayesian computation with R.* Springer New York.

Doebler, W., & Rathsam, J. (2019). Stevens perceived levels of common impulsive noises, sonic booms, and sonic thumps. *The Journal of the Acoustical Society of America*, *145*(3).

Federal Interagency Committee on Noise. (1992). *Federal agency review of selected airport noise analysis issues* (Tech. Rep.). Federal Interagency Committee on Noise.

Fidell, S., Mestre, V., Schomer, P., Berry, B., Gjestland, T., Vallet, M., & Reid, T. (2011). A first-principles model for estimating the prevalence of annoyance with aircraft noise exposure. *The Journal of the Acoustical Society of America*, *130*(2), 791-806. doi: 10.1121/1.3605673

Fidell, S., Schultz, T., & Green, D. M. (1988). A theoretical interpretation of the prevalence rate of noiseinduced annoyance in residential populations. *The Journal of the Acoustical Society of America*, *84*(6), 2109-2113. doi: 10.1121/1.397056

Fields, J., De Jong, R., Gjestland, T., Flindell, I., Job, R. S., Kurra, S., ... Schumer, R. (2001). Standardized general-purpose noise reaction questions for community noise surveys: Research and a recommendation. *Journal of Sound and Vibration*, *242*(4), 641-679.

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457-511.

Groothuis-Oudshoorn, C. G. M., & Miedema, H. M. E. (2006). Multilevel grouped regression for analyzing self-reported health in relation to environmental factors: the model and its application. *Biometrical Journal*, *48*(1), 67–82. doi: 10.1002/bimj.200410172

Hu, F. B., Goldberg, J., Hedeker, D., Flay, B. R., & Pentz, M. A. (1998, 04). Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. *American Journal of Epidemiology*, *147*(7), 694-703. Retrieved from `https://dx.doi.org/10.1093/oxfordjournals.aje.a009511` doi: 10.1093/oxfordjournals.aje.a009511

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.* Elsevier Science.

Lee, J. (2019). *Dose-response modeling of quiet sonic boom community response survey data* (Master's thesis, North Carolina State University). Retrieved from `http://www.lib.ncsu.edu/resolver/1840.20/36354`

Long, J. (1997). *Regression models for categorical and limited dependent variables*. SAGE Publications. Retrieved from https://books.google.com/books?id=CHvSWpAyhdIC

Maglieri, D. J., Bobbitt, P. J., Plotkin, K. J., Shepherd, K. P., Coen, P. G., & Richwine, D. M. (2014). *Sonic boom: Six decades of research* (Tech. Rep. No. NASA/SP-2014-622). NASA Langley Research Center. Retrieved from https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20150006843.pdf

Miedema, H. M. E., & Vos, H. (1998). Exposure-response relationships for transportation noise. *The Journal of the Acoustical Society of America*, *104*(6), 3432-3445. Retrieved from https://doi.org/10.1121/1.423927 doi: 10.1121/1.423927

Miller, N. P., Cantor, D., Lohr, S., Jodts, E., Boene, P., Williams, D., ... Hume, K. (2014). *Research methods for understanding aircraft noise annoyances and sleep disturbance*. Washington, DC: The National Academies Press. Retrieved from https://www.nap.edu/catalog/22352/research-methods-for-understanding-aircraft-noise-annoyances-and-sleep-disturbance doi: 10.17226/22352

Page, J. A., Hodgdon, K. K., Krecker, P., Cowart, R., Hobbs, C., Wilmer, C., ... Shumway, D. L. (2014). *Waveforms and sonic boom perception and response (WSPR): Low-boom community response program pilot test design, execution, and analysis* (Tech. Rep. No. NASA/CR-2014-218180). NASA Langley Research Center. Retrieved from https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20140002785.pdf

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124).

Plummer, M. (2018). rjags: Bayesian graphical models using mcmc [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=rjags (R package version 4-8)

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org

Rathsam, J., Klos, J., Loubeau, A., Carr, D. J., & Davies, P. (2018). Effects of chair vibration on indoor annoyance ratings of sonic booms. *The Journal of the Acoustical Society of America*, *143*(1), 489-499. Retrieved from https://doi.org/10.1121/1.5019465 doi: 10.1121/1.5019465

Schäffer, B., Pieren, R., Mendolia, F., Basner, M., & Brink, M. (2017). Noise exposure-response relationships established from repeated binary observations: Modeling approaches and applications. *The Journal of the Acoustical Society of America*, *141*(5), 3175-3185. Retrieved from https://doi.org/10.1121/1.4982922 doi: 10.1121/1.4982922

Schultz, T. J. (1978). Synthesis of social surveys on noise annoyance. *The Journal of the Acoustical Society of America*, *64*(2), 377–405.

Segarra, P., Domingo, J., Lpez, L., Sanchidrin, J., & Ortega, M. (2010). Prediction of near field overpressure from quarry blasting. *Applied Acoustics*, *71*(12), 1169 - 1176. Retrieved from http://www.sciencedirect.com/science/article/pii/S0003682X10001726 doi: https://doi.org/10.1016/j.apacoust.2010.07.008

Shepherd, K. P., & Sullivan, B. M. (1991). *A loudness calculation procedure applied to shaped sonic booms* (Tech. Rep. No. NASA-TP-3134). NASA Langley Research Center. Retrieved from https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19920002547.pdf

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2014, March). OpenBUGS User Manual Examples Volume II (3.2.3 ed.) [Computer software manual]. Retrieved from http://www.openbugs.net/Examples/Volumeii.html

Stevens, S. S. (1972). Perceived level of noise by Mark VII and decibels (E). *The Journal of the Acoustical Society of America*, *51*(2B), 575-601. Retrieved from https://doi.org/10.1121/1.1912880 doi: 10.1121/1.1912880

Stevens, S. S. (1975). *Psychophysics*. Transaction Publishers.

U.S. Environmental Protection Agency. (1982). *Guidelines for noise impact analysis* (Tech. Rep. No. EPA-550/9-82-105). U.S. Environmental Protection Agency, Office of Noise Abatement and Control.

Wang, F., & Gelfand, A. E. (2002). A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statist. Sci.*, *17*(2), 193–208. Retrieved from https://doi.org/10.1214/ss/1030550861 doi: 10.1214/ss/1030550861

Wilson, D. K., Wayant, N. M., Nykaza, E. T., Pettit, C. L., & Armstrong, C. M. (2017). Multilevel modeling and regression as applied to community noise annoyance surveys. *The Journal of the Acoustical Society of America*, *141*(5), 3727-3728. Retrieved from https://doi.org/10.1121/1.4988178 doi: 10.1121/1.4988178

Yost, W. (2013). *Fundamentals of hearing: An introduction*. Brill.

Appendix A

Data Validation and Cleaning

A.1 Cleaned Dataset

The cleaned dataset is available on the NASA Technical Reports Server as a comma separated values (.csv) file (https://ntrs.nasa.gov/search.jsp?R=20190002702). The first five rows of the cleaned dataset are shown in Table A1. To avoid spaces in the .csv file column headers, shortened variable names are used: s_id, seq, PL and annoy.

Table A1: First five observations of the cleaned dataset.

Subject ID	Boom number	PL	Annoyance response
2460	1	95	5
2625	1	95	4
2725	1	95	1
2742	1	95	5
2524	1	95	6

A.2 Data Validation

Before we fit the candidate models, we use data validation to check that our dataset is the same as the one used in the pilot study data analysis report (Page et al., 2014). We recreate some of the figures and plots from the single-event analysis section. Although we find minor discrepancies, resolving them is not imperative for the purpose of this research, and so we proceed with the dataset on hand for developing the modeling framework.

To begin the process of data validation, we first compare the number of observations in our dataset against the number reported. Appendix G of the report noted that after all data cleaning, there are 2369 responses in the final dataset used for the analysis (Page et al., 2014), and we find that the dataset on hand also contains 2369 observations.

We then compare the Kendall's Tau-b correlation between annoyance rating and other attribute ratings. The values reported in the report can be found in Table 38 (Page et al., 2014). Table A2 compares the Tau-b values we calculated to those reported in Table 38 of the report. The differences in Tau-b correlations do not seem to be large. In most cases, Tau-b correlations from the report are less than the recalculated values and all differences between the reported and recalculated correlations have magnitude less than 0.01.

Table A2: Comparison of calculated to reported Kendall's Tau-b correlation.

Attribute	Tau-b (calculated)	Tau-b (report)	Difference (report - calculated)
Interference	0.7658	0.7626	−0.0032
Startle	0.6965	0.6969	0.0004
Loudness	0.5509	0.5499	−0.001
Vibration	0.4532	0.4505	−0.0027
Rattle	0.4267	0.4174	−0.0093

In addition to comparing the Kendall's Tau-b correlations, we compare the results from the reported quadratic regressions. The report presents seven quadratic polynomial fits, each with percent highly annoyed as the response variable and one of the seven acoustic metrics (ASEL, CSEL, ZSEL, PL, PNL, LLZF, LLZD) as the independent variable. There are three parts to comparing the regression results: comparison of the scatterplots, R-squared values, and regression coefficients. For the method described in the report, binning by the dose is necessary to extract proportion of highly annoyed respondents. Since a different choice of bin width will lead to different results, we use the same binning. All acoustic metrics have ten bins, but of different widths. Table A2 shows the bin widths used for each acoustic metric, which are taken from Table 39 of the report.

Table A3: Bin widths for each of the seven quadratic regressions, one for each noise metric.

Metric	Minimum value	Maximum value	Bin width
ASEL	48	95	4.7
CSEL	73	108	3.5
ZSEL	95	119	2.4
PL	63	106	4.3
PNL	69	114	4.5
LLZf	77	115	3.8
LLZd	78	115	3.7

After separating the data into bins, we first examine the individual scatterplots and compare to those from the report. The general pattern in the scatterplots from the report is evident in the replotted scatterplots. There are instances where certain points seem to not match up exactly. However, these are minor differences and since calculated values (percent highly annoyed and bins) are plotted rather than the raw data, the differences could possibly be attributed to calculation errors. Figure A1 shows both the scatterplots and the regression fits together.

Next we compare the R-squared values from our regression fits to those reported. As seen in Table A4, the differences between the R-squared from the report and the calculated values all have magnitude less than or equal to 0.03. In two of the cases, the R-squared values match exactly because the regression coefficients also match. However, keeping in mind that R-squared is not a unique value and different data could produce the same R-squared, we also compare the regression coefficients.

Table A5 compares the three beta coefficient estimates from the regression fits in the report to our recalculated estimates. The model in consideration is: $Y \sim N(\beta_0 + \beta_1 X + \beta_2 X^2, \sigma^2)$, where Y is percent highly annoyed and X is the corresponding acoustic metric for each regression.

Table A4: Comparison of R-Squared calculated to reported.

Metric	R-squared (report)	R-squared (calculated)	Difference (report - calculated)
ASEL	0.9586	0.9586	0
CSEL	0.9639	0.9522	0.0117
ZSEL	0.8632	0.8632	0
PL	0.9694	0.9623	0.0071
PNL	0.9316	0.9016	0.03
LLZf	0.8804	0.8522	0.0282
LLZd	0.8653	0.8705	−0.0052

From Table A5, we see that only two of the regression equations (for ASEL and ZSEL) have the same coefficient estimates. It is important to note that some of the regression equations presented in the report would result in percent highly annoyed values that are beyond the reasonable range of 0 to 100 for the range of observed sound levels, and thus cast some doubts on the reported curves. For example, if we substitute in values from the observed LLZf range into the equation for the LLZf acoustic metric, we obtain values that are greater than 400% highly annoyed. Figure A1 shows the results of the recalculated data points and regression fits. These plots are recreations of Figure 64 to Figure 70 from the report (Page et al., 2014).

Table A5: Comparison of beta estimates calculated to reported.

Metric	$\hat{\beta}_0$ (report)	$\hat{\beta}_0$ (calc.)	$\hat{\beta}_1$ (report)	$\hat{\beta}_1$ (calc.)	$\hat{\beta}_2$ (report)	$\hat{\beta}_2$ (calc.)
ASEL	61.7	61.71	−2.2	−2.2	0.02	0.02
CSEL	238.81	208.07	−5.89	−5.19	0.04	0.03
ZSEL	706.78	706.78	−14.18	−14.18	0.07	0.07
PL	75.64	45.45	−2.19	−1.42	0.02	0.01
PNL	127.17	124.86	−3.23	−3.16	0.02	0.02
LLZf	595.61	70.2	−3.41	−1.93	0.02	0.01
LLZd	147.28	139.81	−3.64	−3.49	0.02	0.02

When comparing the R-squared and regression coefficient estimates of the calculated and reported regression curves, there are some unresolved discrepancies. However, there are only small noticeable discrepancies when comparing the seven scatterplots of percent highly annoyed as a function of acoustic metric. For the purpose of our research, the small discrepancies observed between recalculated and reported results seem to be minor and thus, we decide it is reasonable to use the dataset on hand to develop our modeling framework.

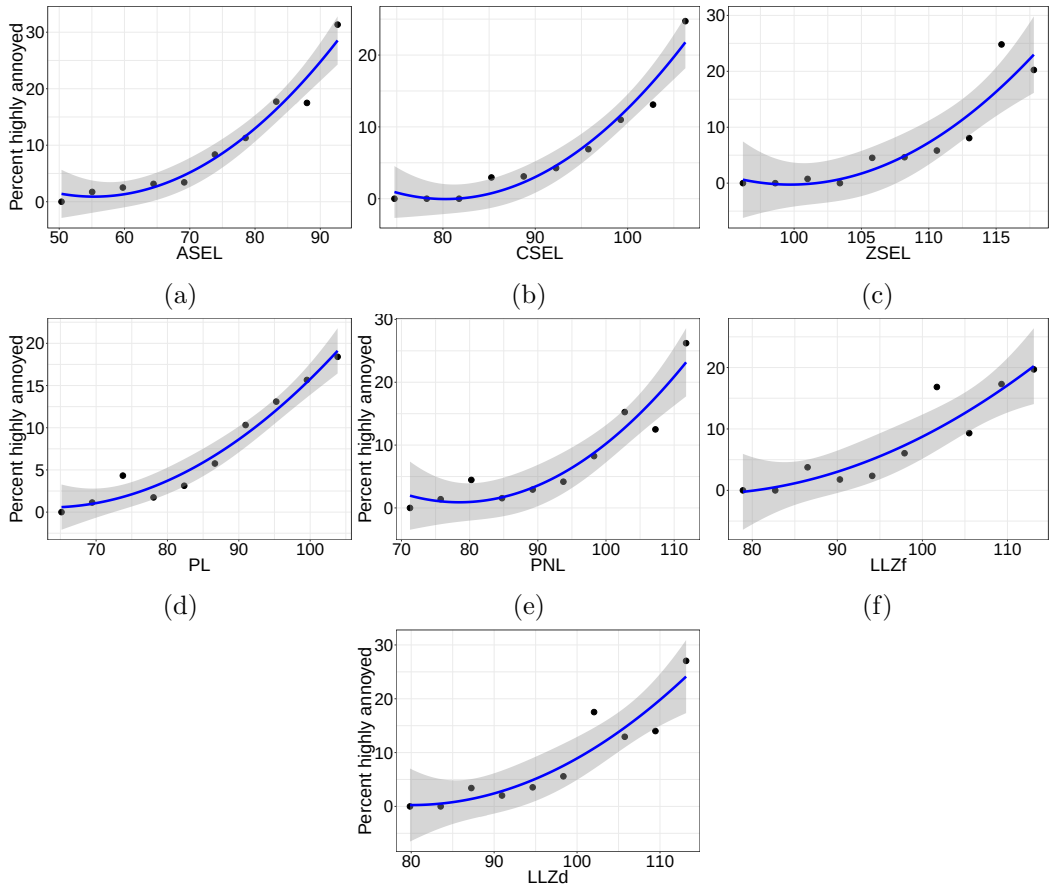


Figure A1: Quadratic regression fits to percent highly annoyed vs. (a) ASEL, (b) CSEL, (c) ZSEL, (d) PL, (e) PNL, (f) LLZf, (g) LLZd.

A.3 Data Cleaning

Before performing the data analysis, we clean the data from the original data spreadsheet containing all received responses based on a few criteria. We first take a subset of the responses that are reported indoors at-home because there was not enough information to estimate the noise exposures elsewhere. Also, since most people in the U.S. spend the majority of their time indoors, we select only responses from participants who were indoors. This reduces the number of observations from 2369 to 2060. Then we filter out all observations that have missing values for any of the variables. This step discards 27 responses, and we are left with 2033 responses. We also find that there are 47 duplicate responses in the data. Following the method reported, we discard the first of the two responses and assume that the second response is an updated response (Page et al., 2014). After discarding the 47 responses, there are 1986 responses. We also find based on the diagnostic DFFITS that there are influential points at 74 dB PL and discard all highly annoyed responses (annoy-

ance responses of 8, 9 or 10) at 74 dB PL.[A8] Since we are using this dataset to build the modeling framework rather than for substantive results, we simply take out the influential points. In addition, there are only 5 influential points so taking them out of the dataset does not reduce the sample size by much. If we are interested in using this dataset to plan for future tests, then we need to analyze with and without the influential points. After taking out the influential points, we have 1981 responses. Note that the order for cleaning and filtering data follows the order described. If we switch some procedures, we may end up with a different dataset.

A.4 Data Dictionary

The data dictionary for relevant variables is shown in Table A6. Most of the summaries are based on the data prior to cleaning, except for the sound levels, binary variable for highly annoyed or not, ordinal responses, and maximum peak overpressure.

Table A6: Data dictionary for relevant variables in the data.

Variable	**Mnemonic**	**Description**
Annoyance survey response	annoy	• Ordinal (0 to 10) • No missing values • Range observed: [0, 10] • 1st Quant.: 0 • Median: 1 • 3rd Quant.: 3
A-weighted sound exposure level	ASEL	• Originally continuous, rounded to nearest integer • Acoustic metric (in dB) • Range observed: [48, 92] • 1st Quant.: 63 • Median: 68 • Mean: 69.7 • 3rd Quant.: 74
Boom sequence or boom number	seq	• Discrete (integers) • Range possible: [1, 110] • Range observed: [1, 110]

Continued on next page

[A8] DFFITS measures the standardized difference in predicted value from the regression on all data and the regression on all data except the one data point of interest.

Table A6 Continued from previous page

Variable	Mnemonic	Description
C-weighted sound exposure level	CSEL	• Originally continuous, rounded to nearest integer • Acoustic metric (in dB) • Range observed: [74 , 108] • 1st Quant.: 89 • Median: 93 • Mean: 93.59 • 3rd Quant.: 98
Highly annoyed or not	HA	• Binary response (created in R code based on rule: `annoy` greater than or equal to 8 is HA, else not HA) • 0= not HA, 1= HA • 1848 not HA responses, 133 HA responses • Mean (prop. highly annoyed responses): 0.067
Interference rating	interfere	• Ordinal (0 to 10) • 2 missing values • Range observed: [0, 10] • 1st Quant.: 0 • Median: 0 • 3rd Quant.: 2
Kryter's perceived noise level	PNL	• Originally continuous, rounded to nearest integer • Acoustic metric (in PNdB) • Range observed: [69, 114] • 1st Quant.: 87 • Median: 93 • Mean: 93.4 • 3rd Quant.: 98
Location of participant during boom	location	• Categorical, with 4 levels • 1= indoors at home, 2= indoors elsewhere, 3= outdoors at home, 4= outdoors elsewhere • Filtered only indoors at home responses (2039 - 47 responses = 1981)

Continued on next page

Table A6 Continued from previous page

Variable	Mnemonic	Description
Loudness rating	loud	• Ordinal (0 to 10) • 1 missing value • Range observed: [0 , 10] • 1st Quant.: 2 • Median: 3 • 3rd Quant.: 5
Maximum peak over-pressure	maxpsf	• Continuous (in pounds per sq. ft.) • Range observed: [0.09, 2.24] • 1st Quant.: 0.31 • Median: 0.45 • Mean: 0.62 • 3rd Quant.: 0.74
Rattle rating	rattle	• Ordinal (0 to 10) • 5 missing values • Range observed: [0, 10] • 1st Quant.: 1 • Median: 2 • 3rd Quant.: 4
Startle rating	startle	• Ordinal (0 to 10) • 1 missing value • Range observed: [0, 10] • 1st Quant.: 0 • Median: 1 • 3rd Quant.: 3
Stevens' Mark VII Perceived level	PL	• Originally continuous, rounded to nearest integer • Acoustic metric (in dB) • Range observed: [63, 106] • 1st Quant.: 78 • Median: 84 • Mean: 85.07 • 3rd Quant.: 90
Subject ID	s_id	• Unique ID for participant • Pre-data cleaning: 52 unique IDs • Post-data cleaning: 49 unique IDs

Continued on next page

Table A6 Continued from previous page

Variable	Mnemonic	Description
Vibration rating	vibrate	• Ordinal (0 to 10) • 5 missing values • Range observed: [0, 10] • 1st Quant.: 1 • Median: 2 • 3rd Quant.: 4
Z-weighted sound exposure level (unweighted sound exposure level)	ZSEL	• Originally continuous, rounded to nearest integer • Acoustic metric (in dB) • Range observed: [95, 119] • 1st Quant.: 106 • Median: 108 • Mean: 108.5 • 3rd Quant.: 112
Zwicker's frontal incidence loudness level	LLZf	• Originally continuous, rounded to nearest integer • Acoustic metric (in phons) • Range observed: [77, 115] • 1st Quant.: 91 • Median: 95 • Mean: 96.31 • 3rd Quant.: 100
Zwicker's diffuse incidence loudness level	LLZd	• Originally continuous, rounded to nearest integer • Acoustic metric (in phons) • Range observed: [78, 115] • 1st Quant.: 91 • Median: 95 • Mean: 96.77 • 3rd Quant.: 100

Concluded

Appendix B

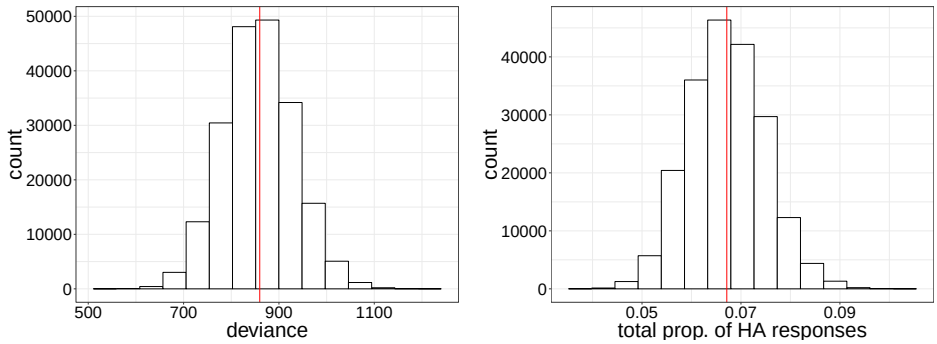
Additional Posterior Predictive Checking Plots

In Section 2, we selectively plot a few posterior predictive checks for each candidate model. The additional checks are plotted here. The procedures for calculating discrepancy statistics are in Section 3.3.2. We also check model fit across PL by comparing proportions of highly annoyed responses replicated from the model to the observed proportions. The procedures for this type of check are:

1. Randomly choose 20 posterior draws
2. Replicate responses for each of the 20 draws using the parameters at each draw
3. Calculate replicated proportions of highly annoyed responses at each PL for each of the 20 draws
4. Plot and compare the replicated proportions to observed proportions

B.1 Non-Multilevel Logistic Regression

Figure B1 shows the two additional checks for non-multilevel logistic regression: deviance and total proportion of highly annoyed responses. The two plots do not indicate lack of fit.



(a) (b)

Figure B1: Posterior predictive checks for non-multilevel logistic regression for (a) deviance and (b) total proportion of highly annoyed responses; each histogram indicates discrepancy statistics calculated from replicated data at each posterior draw and red indicates the observed discrepancy statistics.

To further investigate the model fit, we compare the proportions of highly annoyed responses from replicated responses to observed proportions across PL. The procedures for calculations are outlined in the beginning of this Appendix. Figure B2 compares the replicated proportions to the observed at each PL. The observed data

are represented by the green diamonds, and the replicated proportions are represented by the black points. The size of the black points indicates how many of the 20 replicated datasets have the same number of highly annoyed responses at that PL. Note that the sample size at each PL is not displayed.

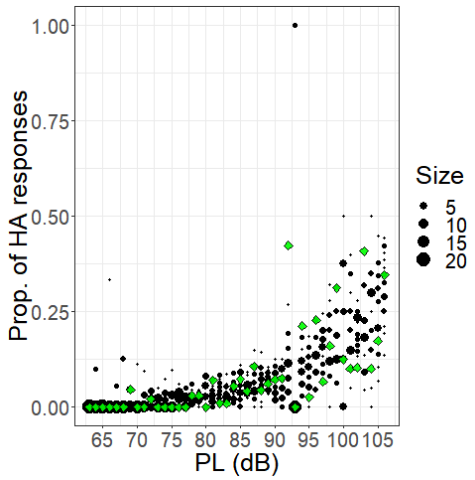


Figure B2: Comparison of replicated proportions of highly annoyed responses from non-multilevel logistic regression model (black points) to observed at each PL (green diamonds).

From Figure B2, we see that sometimes the model predicts more highly annoyed responses than observed at low PL. Note that at 93 dB, there is a black point indicating that the proportion of highly annoyed responses is 1, but it only represents one observation.

B.2 Multilevel Logistic Regression

The additional posterior predictive checks for the multilevel logistic regression model are deviance, total proportion of highly annoyed responses, the mean number of highly annoyed responses per participant, the standard deviation of number of highly annoyed responses per participant, and the maximum number of highly annoyed responses per participant. Figure B3 shows the additional checks. None of these checks indicate lack of fit.

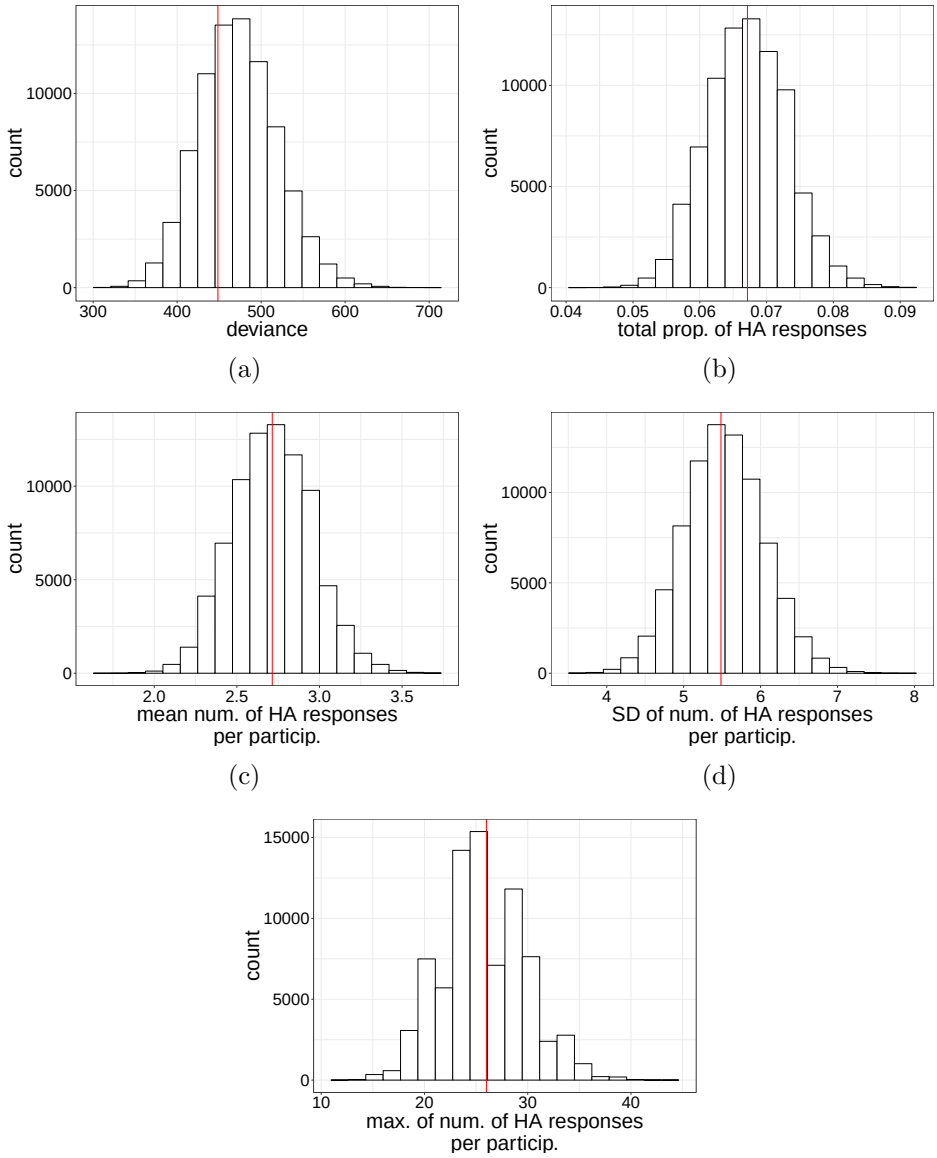


Figure B3: Posterior predictive checks for multilevel logistic regression for (a) deviance, (b) total proportion of highly annoyed responses, and (c) mean number of, (d) standard deviation of and (e) maximum number of highly annoyed responses per participant; each histogram indicates discrepancy statistics calculated from replicated data at each posterior draw and red indicates the observed discrepancy statistics.

We check the model fit across PL by comparing the proportions of highly annoyed responses replicated from 20 random posterior draws against the observed proportions at each PL. Figure B4 compares the observed and replicated proportions at each PL.

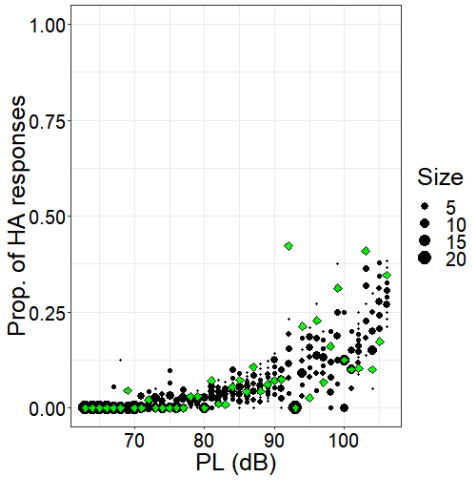


Figure B4: Comparison of replicated proportions of highly annoyed responses from multilevel logistic regression (black points) to observed proportions (green diamonds) at each PL.

We see misfit at 92 dB but this is because the observed proportion is high at 0.42 and not consistent with the gradual increasing trend of the observed proportions.

B.3 Non-Multilevel CTL

Figure B5 shows the two additional posterior predictive checks for non-multilevel CTL: deviance and total proportion of highly annoyed responses. The non-multilevel CTL model tends to predict higher proportions of highly annoyed responses than observed. The empirical quantile of the observed statistic is 0.009.

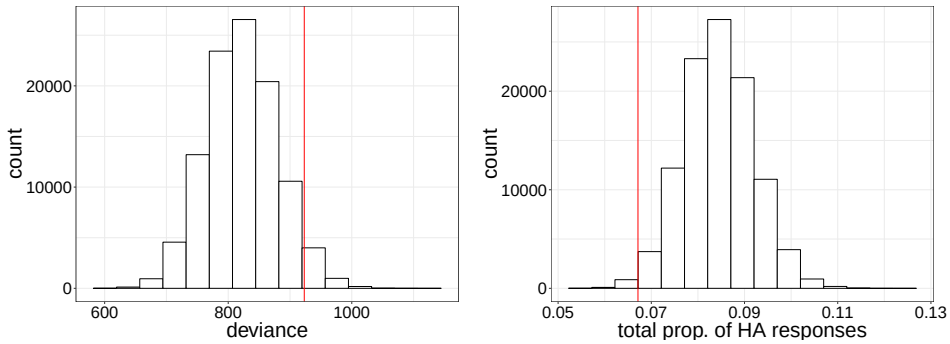


Figure B5: Posterior predictive checks for non-multilevel CTL for (a) deviance and (b) total proportion of highly annoyed responses; each histogram indicates discrepancy statistics calculated from replicated data at each posterior draw and red indicates the observed discrepancy statistics.

Figure B6 compares the proportions replicated from 20 posterior draws to the

observed proportions across PL. The non-multilevel CTL model overpredicts the proportion of highly annoyed responses.

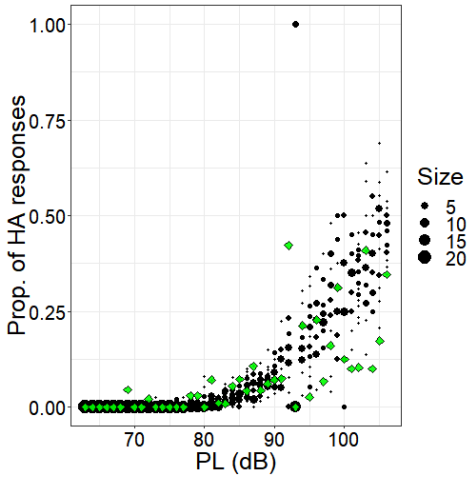


Figure B6: Comparison of replicated proportions of highly annoyed responses from non-multilevel CTL (black points) to observed proportions (green diamonds) at each PL.

B.4 Multilevel CTL

The additional posterior predictive checks for multilevel CTL are for deviance, total proportion of highly annoyed responses, mean number of highly annoyed responses per participant, standard deviation of number of highly annoyed responses per participant, and maximum number of highly annoyed responses per participant. Figure B7 shows these checks, from which we do not see lack of fit.

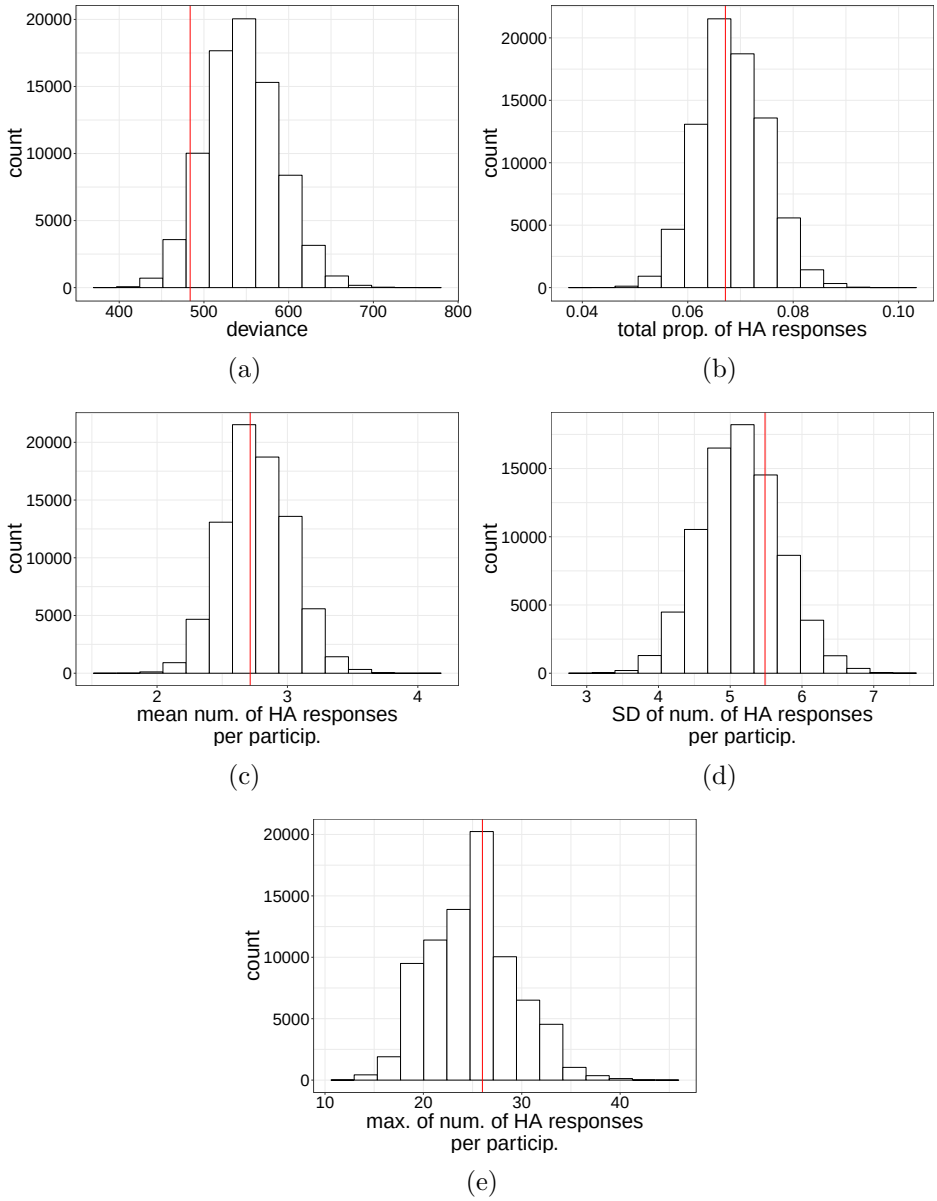


Figure B7: Posterior predictive checks for multilevel CTL for (a) deviance, (b) total proportion of highly annoyed responses, and (c) mean number of, (d) standard deviation of and (e) maximum number of highly annoyed responses per participant; each histogram indicates discrepancy statistics calculated from replicated data at each posterior draw and red indicates the observed discrepancy statistics.

We also compare the observed proportion of highly annoyed responses to replicated proportions from the model across PL in Figure B8. The multilevel CTL model replicates proportions similar to those observed across PL. We see a misfit again at 92 dB but that is due to the high observed proportion of 0.42.

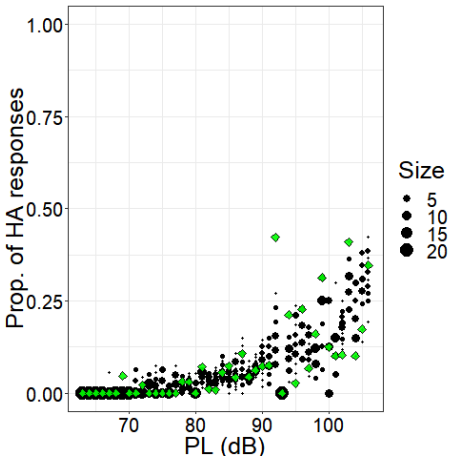
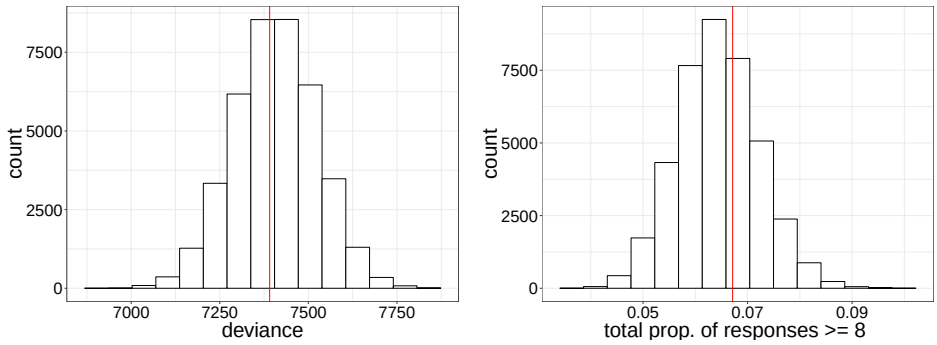


Figure B8: Comparison of replicated proportions of highly annoyed responses from multilevel CTL (black points) to observed proportions (green diamonds) at each PL.

B.5 Non-Multilevel Ordinal Regression

Figure B9 shows the two additional posterior predictive checks for non-multilevel ordinal regression: deviance and total proportion of responses that are 8, 9 or 10. We do not see lack of fit from these two checks.



(a) (b)

Figure B9: Posterior predictive checks for non-multilevel ordinal regression for (a) deviance and (b) total proportion of responses of 8, 9 or 10; each histogram indicates discrepancy statistics calculated from replicated data at each posterior draw and red indicates the observed discrepancy statistics.

Figure B10 assesses the model fit at each PL by comparing replicated proportions of highly annoyed responses to observed proportions. For the ordinal regression models, we calculate the replicated proportions by finding the proportion of responses that are 8, 9 or 10 at each PL. The model sometimes predicts higher proportions than observed at the low PL values. This is expected based on the dose-response estimate shown in Figure 2.18.

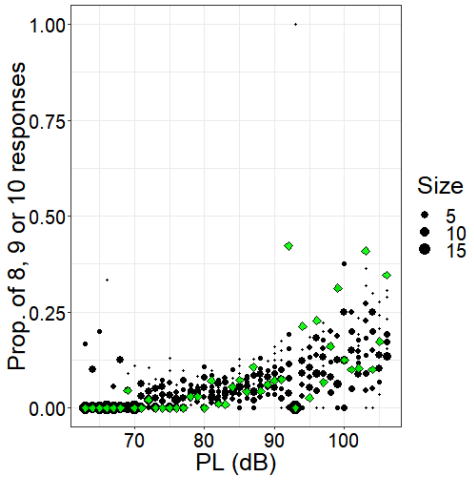


Figure B10: Comparison of replicated proportions of responses of 8, 9 or 10 from non-multilevel ordinal regression (black points) to observed proportions (green diamonds) at each PL.

B.6 Multilevel Ordinal Regression

The additional posterior predictive checks for multilevel ordinal regression are deviance, total proportion of responses of 8, 9 or 10, mean number of responses of 8, 9 or 10 per participant, standard deviation of number of responses of 8, 9 or 10 per participant, and maximum number of responses of 8, 9 or 10 per participant. Figure B11 shows the additional checks and none of these checks indicate lack of fit.

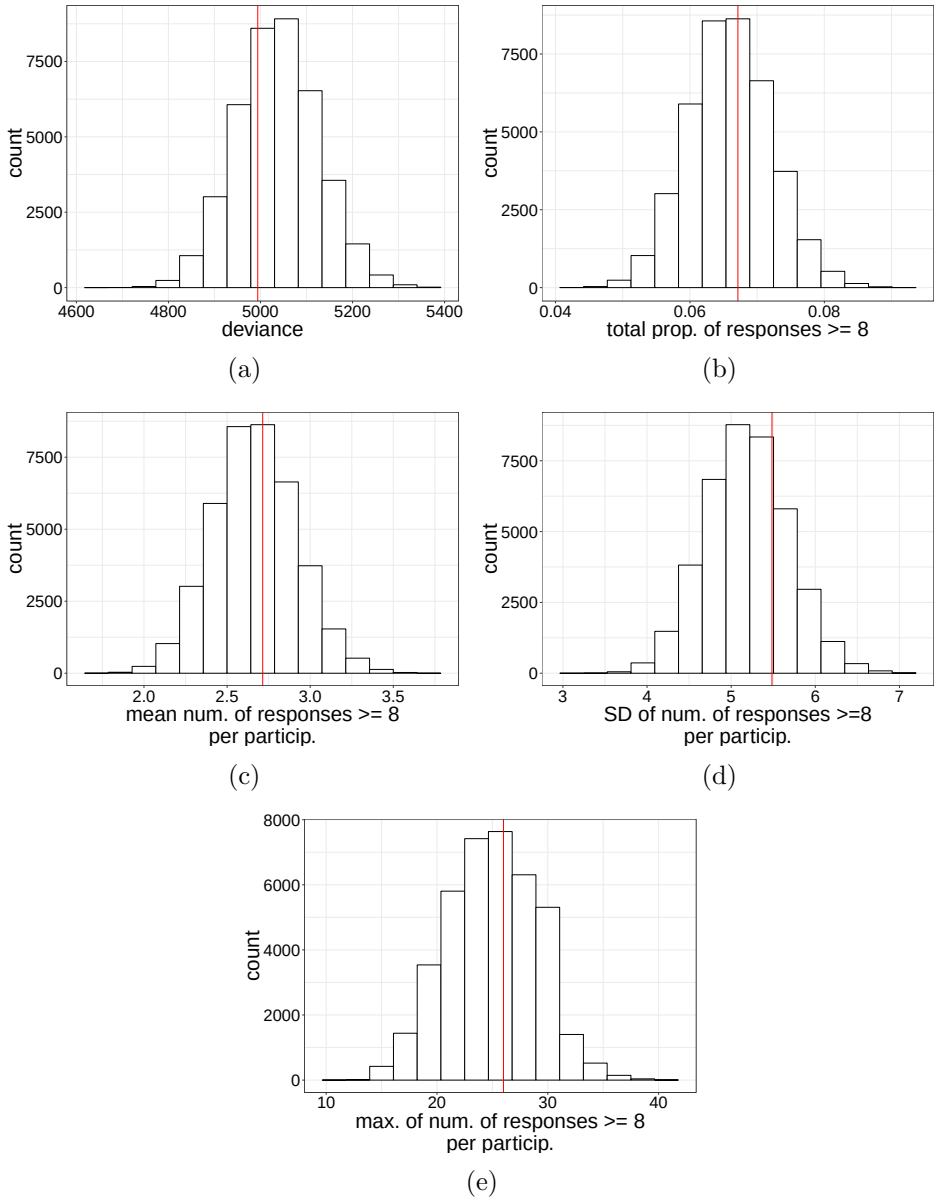


Figure B11: Posterior predictive checks for multilevel ordinal regression for (a) deviance, (b) total proportion of responses of 8, 9 or 10 , and (c) mean number of (d) standard deviation of and (e) maximum number of responses of 8, 9 or 10 per participant; each histogram indicates discrepancy statistics calculated from replicated data at each posterior draw and red indicates the observed discrepancy statistics.

Figure B12 compares the replicated proportions of highly annoyed responses to observed at each PL. The replicated proportions are the proportions of replicated responses that are 8, 9 or 10. At 93 dB, there is only 1 observation and so the black point at proportion of 1 represents 5 out of 20 randomly generated data points was generated to be 1 rather than 0.

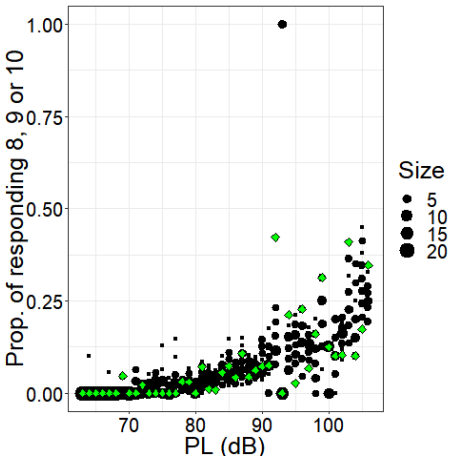
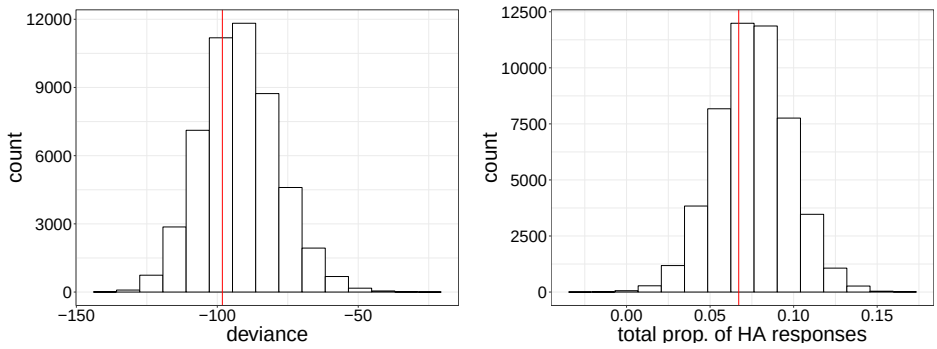


Figure B12: Comparison of replicated proportions of 8, 9 or 10 responses from multilevel ordinal regression (black points) to observed proportions (green diamond) at each PL.

B.7 Piecewise Linear Regression

Figure B13 shows the two additional posterior predictive checks for piecewise linear regression: deviance and total proportion of highly annoyed responses. Surprisingly, we do not see lack of fit in either of these aspects.



(a) (b)

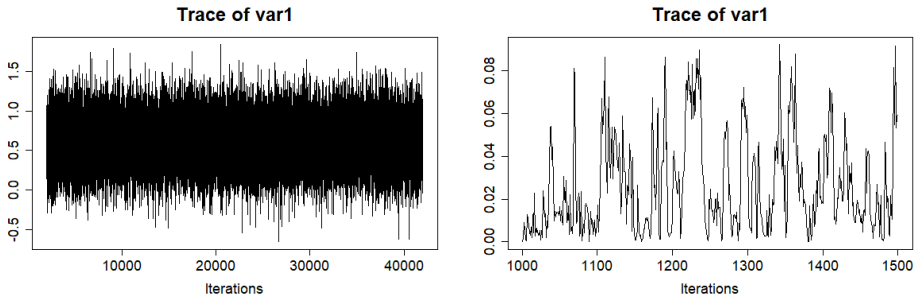
Figure B13: Posterior predictive checks for piecwise linear regression for (a) deviance and (b) total proportion of highly annoyed responses; each histogram indicates discrepancy statistics calculated from replicated data at each posterior draw and red indicates the observed discrepancy statistics.

Appendix C

MCMC Convergence Diagnostics

We use Markov Chain Monte Carlo (MCMC) sampling to fit most of the Bayesian models. JAGS (Plummer, 2003) returns a matrix of posterior draws from the MCMC sampling. The MCMC results vary because we are taking random draws from the posterior distribution. Since the results from the MCMC vary, we run multiple chains, or multiple instances of the MCMC sampling, to check whether the results are similar. Typically, we discard the first portion of an MCMC chain, known as burn-in samples, because the chain is moving from an initial value to the high-probability region of the posterior distribution. We use convergence diagnostics to help determine the number of samples to draw. Convergence indicates that we have samples from the high probability region of the posterior distribution.

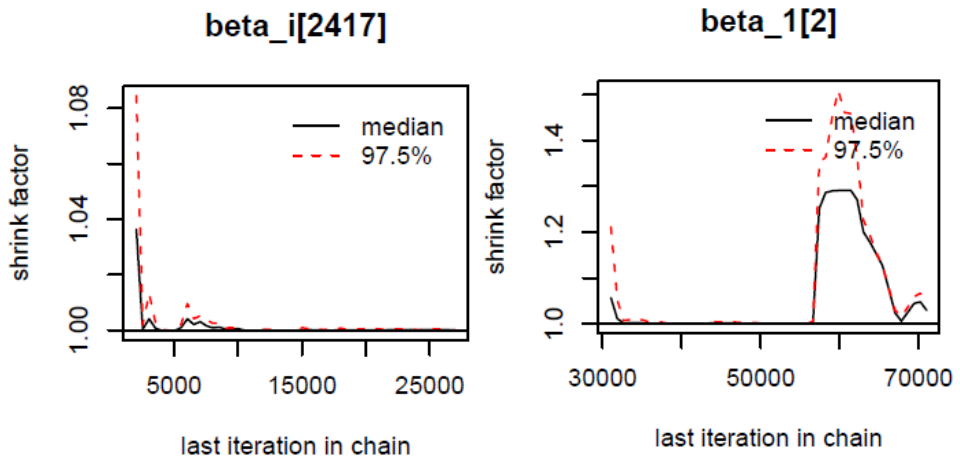
When fitting each model, we check traceplots and Gelman-Rubin plots for indication of convergence. When checking the traceplots, a random pattern indicates convergence. When we see nonrandom patterns, like cyclical patterns, then the MCMC chain has not converged yet. Figure C1 shows an example of a chain that has converged in (a), and of a chain that has not yet converged in (b).



(a) (b)

Figure C1: Example traceplots indicating MCMC (a) has converged, (b) has not converged yet.

When checking the Gelman-Rubin plots, the Gelman-Rubin statistic (labeled as shrink factor on the y-axis of the plots in Figure C2) measures the agreement among MCMC chains. Therefore, this statistic can only be calculated when we run two or more chains. The Gelman-Rubin statistic is scaled so that a value of 1 indicates perfect agreement among chains. The general rule of thumb is any value between 1 and 1.1 indicates "decent convergence" and any value greater than 1.1 indicates that the chains do not agree. Figure C2 shows an example of two chains that agree in (a), and of two chains that do not agree in (b).



(a) (b)

Figure C2: Example Gelman-Rubin plots indicating the two chains (a) agree, (b) do not agree.

We also examine autocorrelation plots to check the correlation among posterior draws. The autocorrelation is the correlation among the sequence of the posterior draws and versions of the sequence that have been shifted or "lagged" by different amounts. Figure C3 is an example of an autocorrelation plot:

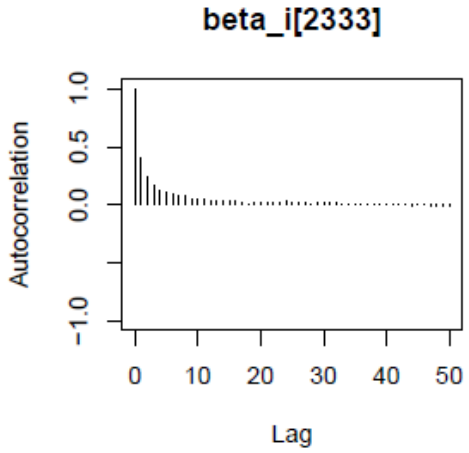


Figure C3: Example autocorrelation plot.

Ideally, the autocorrelation of the posterior samples should decrease and eventually drop to zero as lag increases because that would mean the samples are nearly independent. Highly correlated samples do not necessarily indicate problems with convergence, but may hint at slow MCMC mixing. Therefore, the autocorrelation plots alone cannot tell us whether the chain has converged or not.

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704–0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01-11-2019	Technical Memorandum	

4. TITLE AND SUBTITLE
Statistical Modeling of Quiet Sonic Boom Community Response Survey Data

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Lee, J.; Rathsam, J.; Wilson, A.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
NASA Langley Research Center
Hampton, Virginia 23681-2199

8. PERFORMING ORGANIZATION REPORT NUMBER
L–21022

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Aeronautics and Space Administration
Washington, DC 20546-0001

10. SPONSOR/MONITOR'S ACRONYM(S)
NASA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
NASA/TM–2019–220427

12. DISTRIBUTION/AVAILABILITY STATEMENT
Unclassified-Unlimited
Subject Category 71
Availability: NASA STI Program (757) 864-9658

13. SUPPLEMENTARY NOTES
An electronic version can be found at http://ntrs.nasa.gov.

14. ABSTRACT

The existing ban on commercial supersonic flight overland is largely due to the effects of loud and startling sonic booms on communities. NASA is planning a nationwide campaign of community response surveys using the experimental X-59 Quiet SuperSonic Technology (X-59 QueSST) aircraft to understand how communities perceive the sounds of quiet supersonic flight. The X-59 community response survey data will be presented to noise regulators, who are considering replacing the ban with a noise-based certification limit so quiet supersonic vehicles can fly over land. In this document, we use pilot community response survey data to explore and assess multiple approaches to statistically model the dose-response relationship between single-event sonic boom sound exposure and human annoyance. The models have two primary functions—estimating two types of quantities that support setting regulations and experimental design of future surveys. The dataset is available on the NASA Technical Reports Server as a comma separated values (.csv) file (https://ntrs.nasa.gov/search.jsp?R=20190002702).

15. SUBJECT TERMS
Dose-response, sonic boom, community response survey, statistical model

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			STI Information Desk (help@sti.nasa.gov)
U	U	U	UU	104	19b. TELEPHONE NUMBER *(Include area code)* (757) 864-9658