Model Averaging with AIC Weights for Hypothesis Testing of Hormesis at Low Doses

Steven B. Kim¹ and Nathan Sanders¹

Abstract
For many dose–response studies, large samples are not available. Particularly, when the outcome of interest is binary rather than continuous, a large sample size is required to provide evidence for hormesis at low doses. In a small or moderate sample, we can gain statistical power by the use of a parametric model. It is an efficient approach when it is correctly specified, but it can be misleading otherwise. This research is motivated by the fact that data points at high experimental doses have too much contribution in the hypothesis testing when a parametric model is misspecified. In dose–response analyses, to account for model uncertainty and to reduce the impact of model misspecification, averaging multiple models have been widely discussed in the literature. In this article, we propose to average semiparametric models when we test for hormesis at low doses. We show the different characteristics of averaging parametric models and averaging semiparametric models by simulation. We apply the proposed method to real data, and we show that P values from averaged semiparametric models are more credible than P values from averaged parametric methods. When the true dose–response relationship does not follow a parametric assumption, the proposed method can be an alternative robust approach.

Keywords
hypothesis testing, hormesis, model mispecification, model averaging, Akaike information criterion

Introduction
In toxicology, hormesis is known to be a biphasic dose–response relationship with a stimulatory response at low doses and an inhibitory response at high doses.¹² In this article, we focus on statistical hypothesis testing for hormesis when the outcome of interest is binary. The null hypothesis is the absence of hormesis denoted by \( H_0 \), and the alternative hypothesis is the presence of hormesis denoted by \( H_1 \). For illustration, 2 models for \( H_0 \) and 2 models for \( H_1 \) are shown in Figure 1.

Mathematically speaking, we do not have a sign change in the slope of a dose–response curve at low doses when \( H_0 \) is true. The slope is entirely positive at low doses (top left in the figure), or the zero slope (ie, flat line) becomes a positive slope after passing some threshold dose point (top right in the figure). On the other hand, we have 1 sign change in the slope at low doses when \( H_1 \) is true. The starting slope is negative, and the slope becomes positive at some dose point (bottom left and bottom right in the figure). When \( H_0 \) is true, for a given significance level \( \alpha \), we want the probability of rejecting \( H_0 \) to be at \( \alpha \) or below. When \( H_1 \) is true, we want the statistical power as high as possible.

We can increase the statistical power by increasing the number of experimental doses and/or the sample size inside the hormetic range. In such a case, various statistical methods are available including polynomial, fractional polynomial modeling, splines, and nonparametric smoothing techniques.³ However, having such an ideal experimental design is not always possible in practice. Although we prefer statistical methods that require weak assumptions, we need to borrow a parametric assumption to gain statistical power in small-sample studies. Several useful parametric models used in dose–response assessments are equipped in Benchmark Dose Model software.⁴ These models can be modified to quadratic or other alternative forms in order to model a nonmonotonic
dose–response relationship. A parametric model yields a high statistical power when it is correctly specified, but it can lead to a very low statistical power when it is misspecified. The loss of statistical power due to model misspecification will be shown in this study. There are 2 main reasons. First, the observed data points cannot be adequately modeled by the parametric structure. Second, in the case of model misspecification, data points at high doses make situations even worse because they have too much contribution in the parameter estimation. Such data points are called high-leverage points.

We have 2 aims in this article. First, we explicitly address the impact of model misspecification and high-leverage points in parametric modeling when we perform hypothesis testing for hormesis at significance level \( \alpha \). Second, we propose averaging multiple semiparametric models using the weights calculated by Akaike information criterion (AIC). In both frequentist and Bayesian frameworks, model averaging methods have been widely discussed in cancer risk assessments, particularly for the estimation of benchmark dose. Model averaging allows us to reduce the impact of model misspecification and to account for model uncertainty. In this article, we present the simulation study to compare parametric methods and the proposed semiparametric method. For illustration, we apply the methods to some data discussed in Calabrese and Baldwin which seem to show evidence for hormesis at various degrees. In the application, we show that \( P \) values calculated from parametric models can be misleading, and \( P \) values calculated from the semiparametric method better match with observed dose–response trend.

**Statistical Methods**

In this section, we review a logistic regression model in “Logistic Regression Model” subsection. We base our discussion on the logistic model among many parametric models because of its popularity, and the same discussion can be carried out for another form of parameterization. We briefly review the model averaging method based on the AIC in “Model Averaging in Parametric Models (L3 and L4)” subsection. We then discuss the application of model averaging to semiparametric models in “Application of Model Averaging to Semiparametric Models” subsection. More mathematical detail is included in the Appendix.

The following notation is used in the section. Let \( x_j \geq 0 \) denote the \( j \)th fixed experimental dose for \( j = 1, \ldots, J \), where \( J \) is the total number of experimental doses. Without loss of generality, let \( x_1 = 0 \) be the control dose and \( x_1 < x_2 < \ldots < x_J \). Let \( Y_j \) denote the binary random variable, where \( Y_j = 1 \) if the \( j \)th experimental unit treated at dose \( x_j \) shows a toxic outcome and \( Y_j = 0 \) otherwise. Let \( \pi_j \) denote the unknown probability of observing \( Y_j = 1 \) (ie, the probability of observing a toxic outcome at dose \( x_j \)). Let \( n_j \) denote the number of experimental units observed at dose \( x_j \), and let \( N = \sum_j n_j \) denote the total sample size for the experiment.

**Logistic Regression Model**

To model a potential “J-shaped” dose–response relationship, a logistic regression model with the quadratic term

\[
\log\left(\frac{\pi_j}{1 - \pi_j}\right) = \beta_0 + \beta_1 x_j + \beta_2 x_j^2
\]

is briefly discussed by May and Bigelow. They considered the log dose, but it is not an important distinction in our discussion. Throughout the article, the logistic regression model is referred to as the L3 model, where L3 stands for the logistic regression model with 3 parameters \( \beta_0, \beta_1, \) and \( \beta_2 \). The quadratic term allows a hormetic dose–response curve at low doses, and similar parameterizations are discussed by Bogen and Kim et al using different link functions. In the quadratic form, hypothesis testing for hormesis is simply formulated as \( H_0: \beta_1 \geq 0 \) versus \( H_1: \beta_1 < 0 \). In a large sample, we can make inference for \( \beta_1 \) based on the maximum likelihood estimator for \( \beta_1 \), but we usually do not observe such a large sample size particularly inside the hormetic range in practice. In this article, we base our inference on bootstrapping to approximate the sampling distribution of the maximum likelihood estimator for \( \beta_1 \). Our focus is on the consequence of wrong implementation of this model.

The L3 model is an efficient statistical strategy when it is correctly specified. On the other hand, when the model is misspecified, it can lead us to a misleading result even in a large sample. Under model misspecification, observations made at
Model Averaging in Parametric Models (L3 and L4)

The advantage of the L3 model is efficiency (high statistical power) when the true hormetic zone is symmetric and the true dose–response relationship follows the quadratic form. The disadvantage is a loss of statistical power when the model is misspecified. It is also sensitive to high-leverage points. The impact of model misspecification and high-leverage points can be increased particularly when the experiment is poorly designed. On the other hand, the advantage of the L4 model is flexibility, and it can maintain relatively high statistical power when the model is misspecified. It is also sensitive to high-leverage points. The disadvantage is a loss of statistical power when the model is misspecified. When our goal is parameter estimation. However, our goal is hypothesis testing not parameter estimation. Though \( \beta_1 \) in L3 and \( \beta_1 \) in L4 have different meanings, both \( \beta_1 < 0 \) in L3 and \( \beta_1 > 0 \) in L4 have the same meaning (presence of hormesis) in the context of our hypothesis testing. Under each model using bootstrapping, we may seek evidence for \( \beta_1 < 0 \) through the bootstrap distribution of estimated \( \beta_1 \). Therefore, when a mixture bootstrap distribution of estimated \( \beta_1 \) is mostly negative (regardless of magnitude), it may be regarded as evidence for hormesis.

Application of Model Averaging to Semiparametric Models

Recall that we let \( \pi_j \) denote the probability of a toxic outcome at dose \( x_j \). In general, we need evidence for \( \pi_1 > \pi_2 \) to reject \( H_0 \)
in favor of \( H_1 \). In this section, we introduce a proposed model averaging method which does not require a link function between the probability of a toxic outcome and the dose.

Let \( M_1 \) denote the saturated model with \( J \) free parameters, \( \pi_1, \pi_2, \ldots, \pi_J \), so each \( \pi_j \) is estimated by the observed proportion at dose \( x_j \). Let \( M_2 \) denote a model with \( J - 1 \) free parameters by the condition \( \pi_2 = \pi_3 = \ldots = \pi_J \), so \( \pi_2 \) is estimated by the observed proportion at the 2 doses \( x_2 \) and \( x_3 \). Let \( M_3 \) denote a model with \( J - 2 \) free parameters by the condition \( \pi_2 = \pi_3 = \pi_4 \), so \( \pi_2 \) is estimated by the observed proportion at the 3 doses \( x_2, x_3, \) and \( x_4 \). Using general notation, let \( M_k \) denote the model with \( J - k + 1 \) free parameters such that \( \pi_2 = \ldots = \pi_k+1 \). We consider up to \( M_{J-2} \) and obtain the AIC-weights \( w_1, \ldots, w_{J-2} \) by maximizing the log-likelihood function (see Appendix for detail). When we have more experimental units inside the hor- metric range, the model averaging will become more robust regardless of the values of \( x_1, \ldots, x_J \) because the model structure depends on the order \( x_1, \ldots, x_J \) and estimated \( \pi_1, \ldots, \pi_J \). It does not assume a particular shape of dose–response curve.

For each \( M_k \), we can obtain the bootstrap distribution of estimated \( \pi_1 - \pi_2 \), then we test for hypothesis testing \( H_0^*: \pi_1 - \pi_2 \leq 0 \) versus \( H_1^*: \pi_1 - \pi_2 > 0 \) based on the mixture of the \( J - 2 \) bootstrap distributions weighted by the AIC weights. At significance level \( \alpha \), we reject \( H_0 \) in favor of \( H_1 \) when the \( \alpha \)th quantile of the mixture distribution exceeds 0. Throughout the discussion, this model averaging method is referred to as MA\(_{\text{SP}}\) (model averaging with semiparametric models).

\section*{Results}

In this section, we compare the operating characteristics of the 4 aforementioned models: L3, L4, MA\(_p\), and MA\(_{\text{SP}}\). We consider 16 simulation scenarios with 5 scenarios under \( H_0 \) and 11 scenarios under \( H_1 \) (see Simulation Design). We summarize the simulation results by the probability of rejecting \( H_0 \) under each scenario and for each model (see Simulation Result). Then, we apply the 4 models to the data discussed in Calabrese and Baldwin\cite{14} which provided some degree of evidence for hormesis at low doses (see Application).

\section*{Simulation Design}

To control noise in the simulation, for all 16 scenarios, we assumed \( J = 6 \) experimental doses geometrically spaced as \( x_1 = 0, x_2 = 0.0625, x_3 = 0.125, x_4 = 0.25, x_5 = 0.5, \) and \( x_6 = 1, \) and we assumed \( n_j = 50 \) for each dose group so that \( N = 300 \) is the total sample size. For the case of \( H_0 \), we generated data under the logistic models L\(_3\) or L\(_4\) (scenarios 1-5). For the case of \( H_1 \), we generated data under L\(_3\) (scenarios 6 and 7), L\(_4\) (scenarios 8-13) and neither (scenarios 14-16).

The scenarios under the parametric structures are shown in Figure 3 (scenarios 1-13). For these parametric scenarios, the parameter values are presented in Table 1. For scenarios 14 to 16, we broke the parametric structures, so both L\(_3\) and L\(_4\) are misspecified models in the 3 scenarios. We made pointwise assumptions \( \pi_1 = 0.2, \pi_2 = 0.09, \pi_3 = 0.1, \pi_4 = 0.2, \pi_5 = 0.4, \) and \( \pi_6 = 0.6 \) in scenario 14; \( 0.2, 0.09, 0.07, 0.2, 0.4, \) and \( 0.6 \) in scenario 15, respectively; and \( 0.2, 0.07, 0.08, 0.2, 0.4, \) and \( 0.6 \) in scenario 16, respectively. Each scenario was simulated 1000 times, and 2000 bootstrap samples were used per simulated sample. We fixed the significance level at \( \alpha = 0.05 \), and the probability of rejecting \( H_0 \) was recorded for each method under each scenario.

\section*{Simulation Result}

Table 2 provides the simulation results. When \( H_0 \) was true in scenarios 1 to 5, the L\(_3\) model violated \( \alpha = 0.05 \) in some scenarios at mild degree, and the L\(_4\) model violated \( \alpha = 0.05 \) at serious degree in scenario 2. The estimated type I error probability was .082 which is difficult to believe that it just happened by chance for 1000 replications of the scenario. In each scenario, the model averaging method MA\(_{\text{SP}}\) rejected \( H_0 \) with a probability between the resulting probabilities in L\(_3\) and L\(_4\) with an anticipated result. The model averaging method MA\(_{\text{SP}}\) obeyed \( \alpha = 0.05 \) in the 5 null scenarios.

In scenarios 6 and 7, when \( H_1 \) was true under the L\(_3\) model, the L\(_4\) model led to slightly lower statistical powers than the L\(_3\) model due to overparameterization. The MA\(_{\text{SP}}\) model yielded statistical power between the results in M\(_3\) and M\(_4\) as anticipated. On the other hand, the MA\(_{\text{SP}}\) yielded substantially lower statistical powers in the 2 scenarios. When the true dose–response curve is generated under the simple L\(_3\) model, the L\(_3\) model outperformed which is not surprising.

In scenarios 8 to 13, when \( H_1 \) was true under the L\(_4\) model, the L\(_3\) model could not tolerate model misspecification by showing substantially lower statistical powers due to the inflexibility. The L\(_4\) model mostly showed outperformance because the scenarios belong to its own parameterization. In these 6 scenarios, the MA\(_{\text{SP}}\) model was consistently more powerful than the MA\(_p\) model despite the contribution of the true L\(_4\) model to MA\(_p\). In scenarios 9 to 11, the MA\(_{\text{SP}}\) model showed comparable results to the results from the true L\(_4\) model.

We now turn our focus on scenarios 14, 15, and 16. Recall that these 3 scenarios did not belong to any of L\(_3\), L\(_4\), MA\(_p\), and MA\(_{\text{SP}}\). The M\(_3\) model could not reject \( H_0 \) even once, the L\(_4\) model showed statistical powers of .180, .200, and .268, respectively, and the MA\(_p\) model showed statistical powers of .107, .144, and .200, respectively. On the other hand, the MA\(_{\text{SP}}\) showed statistical powers of .438, .580, and .666, respectively.

The take-home message is clear. The parametric L\(_3\) and L\(_4\) models sometimes showed outperformance within their own parameterizations (statistical power of .984 and .998 from L\(_3\) in scenarios 6 and 7, respectively; statistical power of .570, .474, .706, .525, and .768 from L\(_4\) in scenarios 8, 9, 10, 12, and 13, respectively; see Table 2), but they performed poorly when the truth was not under the model (statistical power of 0 from L\(_3\) in scenarios 14, 15, and 16; statistical power of .180, .200, and .268 from L\(_4\) in scenarios 14, 15, and 16 which are less than one half when compared to .438, .580, and .666 from MA\(_{\text{SP}}\) in the respective scenarios; see Table 2). This is a
general phenomenon in statistics. On the other hand, the \( \text{MASP} \) model was relatively less sensitive, and the model averaging among the semiparametric model (\( \text{MASP} \)) showed greater statistical power than the model averaging among the parametric models (\( \text{MAP} \)) in the scenarios except when \( L_3 \) was the true model (see scenarios 8-16 in Table 2).

**Application**

In this section, we apply the 4 models (\( L_3, L_4, \text{MAP}, \) and \( \text{MASP} \)) to some of binary data discussed in Calabrese and Baldwin\(^{14} \) and compare \( P \) values from the 4 models.

Calabrese and Baldwin\(^{14} \) discussed the effects of saccharin on hyperplasia of the urinary bladder. The 6 experimental doses were 0, .01, .1, 1, 5, and 7.5 (\% in diets). The respective observed proportions of tumor incidence were 10/73 (14 \%), 6/71 (8 \%), 4/81 (5 \%), 4/76 (5 \%), 6/64 (9 \%), and 19/62 (31 \%) for male rats (see the top left panel in Figure 4). When we implemented the \( L_3, L_4, \text{MAP}, \) and \( \text{MASP} \) models, the respective \( P \) values were .042, .027, .034, and .059, respectively. This is one of the cases when the parametric models adequately described the dose–response relationships, and the \( \text{MASP} \) model could not achieve the significance level \( \alpha = .05 \). At the same experimental doses, the respective observed

---

**Table 1. Parameter Values for Scenarios 1 to 13 Under the Logistic Model.**

| Scenario | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 | 13 |
|----------|----|----|----|----|----|----|----|----|----|----|----|----|----|
| \( \beta_0 \) | –1.3 | –1.3 | –1.3 | –1.3 | –2.3 | –1.39 | –1.39 | –1.39 | –1.39 | –1.39 | –1.39 | \( \text{MASP} \) |
| \( \beta_1 \) | 0  | 0  | 0  | 0  | 0  | –7.33 | –10.02 | –5.18 | –4.62 | –7.09 | –6.31 | –7.33 | –10.02 |
| \( \beta_2 \) | 0  | 1  | 3  | 1  | 3  | 14.66 | 20.04 | 7.33 | 5.82 | 10.02 | 7.95 | 14.66 | 20.04 |
| \( \beta_3 \) | 0  | 1  | 1  | .5 | .5 | 1  | 1  | .5 | .33 | .33 | .5 | .5 | .5 |
proportions of incidence were 3/34 (9%), 15/34 (44%), 8/45 (17.8%), and 10/76 (13%) respectively (see the bottom left panel in Figure 4). As shown in the figure (the top right panel), the data points in the observed range are not symmetric, so the L3 model was not able to model the observed nonmonotonic dose–response relationship adequately.

Calabrese and Baldwin14 also discussed the effect of 3-Methylcholanthrene on pulmonary tumors in female rats. The 9 experimental doses were 0, .005, .015, .046, .137, .4, 1.2, 3.7, and 11.1 (µg). The observed proportions of tumor incidence were 15/34 (44%), 1/18 (6%), 5/19 (26%), 7/18 (39%), 6/20 (30%), 12/24 (50%), 8/11 (73%), 10/10 (100%), and 11/11 (100%), respectively (see the bottom left panel in Figure 4). Compared to the previous data set (effect of saccharin on urinary bladder), it has a smaller total sample size but a larger number of experimental doses. The L3, L4, MAP, and MASP models yielded P values of .869, .070, .224, and .003, respectively. As shown in Figure 4 (the bottom left panel), the L3 model was not flexible enough to describe the observed data, and the L4 model was quite flexible to follow the observed nonmonotonic trend though it did not achieve statistical significance. Under the MASP model, the data served as significance evidence for hormesis with a P value .003 < α = .05.

As a final example, the same paper14 discussed the effect of cadmium chloride on testicular tumors. From the 7 experimental doses 0, 1, 2.5, 5, 10, 20, and 40 (µmol/kg), the observed proportions of incidence were 8/45 (17.8%), 1/30 (3.3%), 3/29 (10.3%), 3/30 (10.0%), 4/30 (13.3%), 21/29 (72.4%), and 24/29 (82.8%), respectively (see the bottom right panel in Figure 4). Despite the strong hormetic trend, P values from L3, L4, MAP, and MASP were .99, .64, .65, and .06, respectively, and they seem to be influenced by the 2 data points at the high doses 20 and 40 µmol/kg. The P values from the parametric methods (L3, L4, and MAP) do not seem credible, and the P value from averaging the semiparametric models seems more credible based on the observed trend before modeling. In the figure (the bottom right panel), the L3 model was not able to model the asymmetric hormetic trend, and the flexibility of the L4 model was used to chase the 2 data points at the high experimental doses rather than the data points at low doses. An interesting issue with the L4 model is discussed in the following section.

**Discussion**

The focus of this article is not to argue the existence of hormesis for a particular carcinogen. Our focus is a valid hypothesis testing for a hormetic effect at low doses. We presented the different characteristics of averaging parametric models and averaging semiparametric models. In conclusion, when the true hormetic relationship is under the simple L3 model, the parametric approach MAP (and individual L3 and L4) outperformed the semiparametric approach MASP, which is not surprising. On the other hand, when we compare the 2 averaging methods, MASP outperformed MAP when the true hormetic relationship is nonparametric and even when the truth is generated under the parametric L4 model. It is also shown that the parametric approaches cannot tolerate model misspecification for the hypothesis testing. To this end, when the truth does not follow a parametric form, MASP can be useful for more robust and higher statistical power.

In large sample studies with many experimental doses in a hormetic range, a nonparametric method can be a more reasonable approach because it can disconnect information between doses inside a hormetic range and higher doses. May and Bigelow7 discussed the practical challenges due to insufficient sample sizes and a lack of experimental doses (missing a potential hormetic range). In small sample studies, borrowing mathematical structure (ie, parametric models) to gain efficiency seems inevitable. As discussed in “Results,” however, it sometimes gives us a misleading result not simply due to a lack of evidence but due to model misspecification and high-leverage data points (recall the low statistical power from L3, L4, and MA P in scenarios 14-16 in Table 2). Motivated by this fact, we considered averaging semiparametric models (MA SP) with AIC weights to test for a hormetic effect at low doses. When we implemented real data, calculated P values from MASP made more sense than P values from L3, L4, and MAP particularly in the last applied example in “Application” (cadmium chloride on testicular tumors).

In the simulation study, we showed pros and cons of the parametric methods (L3, L4, and MAP) and of the semiparametric method (MA SP). Under the correctly specified parametric form, the statistical power from L3 was highest among the 4 methods as shown in scenarios 6 and 7 in Table 2.
2, and MAP outperformed MASP in the 2 scenarios generated by L3. Under the wrong parametric form, we lost statistical power substantially as shown in scenarios 14 to 16 (zero statistical power from L3 in Table 2). We observed that MASP outperformed MSP in the 3 scenarios generated outside of L3 or L4. Even when the true scenario was generated by L4, MASP could yield higher statistical power than MAP as shown in scenarios 8 to 13 (see Table 2). In practice, we cannot guarantee that observed outcomes (which are based on the unknown true dose–response relationship) will follow a parametric model as seen in the cadmium chloride data, and it is not under our control. Instead, the proposed model averaging method with semiparametric models does not require the parametric form, and we learned that it is relatively less sensitive to the shape of observed dose–response trend.

For the L4 model, we refitted the cadmium chloride data with the restriction $0 < \beta_3 < 1$. The restricted parameterization yielded a $P$ value lower than $\alpha = .05$ by adequately modeling the asymmetric hormetic zone. However, when we ran a simulation study under the null scenarios, it violated the significance level $\alpha = .05$ seriously. In other words, the restricted parameterization tended to favor $H_1$ too often when $H_0$ is true. Additionally, we implemented the nonparametric regression method discussed in Hall and Heckman in the simulation study. The advantage of a nonparametric method is insensitivity to data points at high doses, but it yielded lower statistical powers than MASP in all of the alternative scenarios (generally below .2). We also studied other methods among others. We modified the test statistic in Baraud et al., which is more appropriate for the binomial data (ie, difference in the 2 probabilities, comparing the control group and the lowest nonzero dose group). It tended to be too conservative, and it did not seem appropriate for a sparse experimental doses. Similarly, we tested an umbrella shape at low doses (decreasing then increasing) using the additive constrained regression. Due to the small number of dose groups, the model fitted the data exactly, and it inflated the type I error rate in null scenarios. A nonparametric method is often useful for a large sample study, but 50 experimental units per dose group with 6 experimental units do not seem sufficiently large particularly for binary outcomes.

**Figure 4.** Fitted dose–response curves under the L3 and L4 models using the maximum likelihood estimates.
To further study L3, L4, MA\textsubscript{P}, and MA\textsubscript{SP} for increased sample sizes (fixing the experimental doses), we manipulated the cadmium chloride data by doubling the sample size while maintaining the observed proportions, that is 16/90 (17.8\%), 2/60 (3.3\%), 6/58 (10.3\%), 6/60 (10.0\%), 8/60 (13.3\%), 42/58 (72.4\%), and 48/58 (82.8\%). Recall the \(P\) values were .99, .64, .65, and .06 for L3, L4, MA\textsubscript{P}, and MA\textsubscript{SP}, respectively, before the manipulation (see Application). After the manipulation, the \(P\) values changed to .99, .727, .727, and .006 for L3, L4, MA\textsubscript{P}, and MA\textsubscript{SP}, respectively. When we multiplied the sample size by 10, that is 80/450 (17.8\%), 10/300 (3.3\%), 30/290 (10.3\%), 30/300 (10.0\%), 40/300 (13.3\%), 210/290 (72.4\%), and 240/290 (82.8\%), the \(P\) values were close to 1. .943, .943, and close to 0 for L3, L4, MA\textsubscript{P}, and MA\textsubscript{SP}, respectively. Then, we tested an additional simulation scenario under the assumptions (1) \(\pi_1 = .178\), \(\pi_2 = .033\), \(\pi_3 = .103\), \(\pi_4 = .100\), \(\pi_5 = .133\), \(\pi_6 = .724\), and \(\pi_7 = .828\) and (2) \(n_1 = 450\), \(n_2 = 300\), \(n_3 = 290\), \(n_4 = 300\), \(n_5 = 300\), \(n_6 = 290\), and \(n_7 = 290\). The resulting statistical powers were near .005, .005, and near 1 for L3, L4, MA\textsubscript{P}, and MA\textsubscript{SP}, respectively. The results illustrate the impact of model misspecification even with large data under the parametric methods when we test for hormesis.

The Generic Hockey Stick model proposed by Bogen\textsuperscript{6} is a parametric model with the enhanced polynomial flexibility and the link function used in the linearized multistage model.\textsuperscript{23} In sparse data, it may have the advantage of utilizing 3 parameters to allow greater flexibility than the L3 models considered in this study. However, the Fisher expected information tells us that the parameter estimation still heavily depends on high data points when we use a polynomial predictor. In practice, high-response data are removed when (1) they are irrelevant to low-dose inference and (2) they severely deviate from an assumed parametric model. We concern about too few data points after removing the data points, and it may increase the type I error rate under the null scenario. It is our future study.

Based on the simulation study, we have thought that averaging L4 and MA\textsubscript{SP} can balance the sensitivity (when the observed data points can be approximated by the assumed parametric structure) and the robustness (when they do not follow the assumed parametric structure). It is our current research direction.

Appendix

Model Averaging of L3 and L4

Let \(y_{ij} = 1\) if we observe the \(i\)th experimental unit treated by dose \(x_i\) showed a toxic outcome (and \(y_{ij} = 0\) otherwise) for \(i = 1, \ldots, n\) and \(j = 1, \ldots, J\) (see Model Averaging in Parametric Models [L3 and L4] subsection). Let \(\pi_j\) denote the probability of observing a toxic event at dose \(x_j\) which is unknown for \(j = 1, \ldots, J\). Given the data, the log-likelihood function is

\[
I = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left( y_{ij} \log(\pi_j) + (1 - y_{ij}) \log(1 - \pi_j) \right)
\]

The L3 model assumes

\[
\pi_j = \frac{e^{\beta_0 + \beta_1 x_j + \beta_2 x_j^2}}{1 + e^{\beta_0 + \beta_1 x_j + \beta_2 x_j^2}}
\]

so \(I\) is a function of \(\beta_0, \beta_1, \) and \(\beta_2,\) denoted by \(I_{L3}(\beta_0, \beta_1, \beta_2)\). The L4 model assumes

\[
\pi_j = \frac{e^{\beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \beta_3 x_j^3}}{1 + e^{\beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \beta_3 x_j^3}}
\]

so \(I\) is a function of \(\beta_0, \beta_1, \beta_2, \) and \(\beta_3,\) denoted by \(I_{L4}(\beta_0, \beta_1, \beta_2, \beta_3)\). Let \(\tilde{I}_{L3}\) denote the maximized \(I_{L3}(\beta_0, \beta_1, \beta_2)\) with respect to \((\beta_0, \beta_1, \beta_2)\), and let \(\tilde{I}_{L4}\) denote the maximized \(I_{L4}(\beta_0, \beta_1, \beta_2, \beta_3)\) with respect to \((\beta_0, \beta_1, \beta_2, \beta_3)\). The AIC of the L3 model is defined as \(\text{AIC}_{L3} = -2\tilde{I}_{L3} + 2p = -2\tilde{I}_{L3} + 6\) because the model has \(p = 3\) parameters. The AIC of the L4 model is defined as \(\text{AIC}_{L4} = -2\tilde{I}_{L4} + 2p = -2\tilde{I}_{L3} + 8\) because it has \(p = 4\) parameters. Then the AIC weight of the L3 model and of the L4 model is

\[
W_{L3} = \frac{e^{-0.5\text{AIC}_{L3}}}{e^{-0.5\text{AIC}_{L3}} + e^{-0.5\text{AIC}_{L4}}}, \quad W_{L4} = \frac{e^{-0.5\text{AIC}_{L4}}}{e^{-0.5\text{AIC}_{L3}} + e^{-0.5\text{AIC}_{L4}}},
\]

respectively.

Type I Error Rate Under MA\textsubscript{P}

Assume the null hypothesis \(H_0\) is true (see Model Averaging in Parametric Models [L3 and L4] subsection). Let \(B\) denote the number of bootstrap samples. Let \(w\) be a number between 0 and 1 (ie, AIC weight). Let \(p_3\) denote the proportion of bootstrap estimates such that \(\beta_1 > 0\) under the L3 model. Consider a significance level \(\alpha = .05\). In this case, under the L3 model, we reject \(H_0\) in favor of \(H_1\) when \(B \times p_3 < .05 \times B\). Similarly, let \(p_4\) denote the proportion of bootstrap estimates such that \(\beta_1 < 0\) under the L4 model. Under the L4 model, we reject \(H_0\) in favor of \(H_1\) when \(B \times p_4 < .05 \times B\). Now, consider a mixture bootstrap distribution of estimated \(\beta_1\) (ie, averaging the 2 bootstrap distributions from L3 and L4). If L3 is weighted by \(w\) and L4 is weighted by \(1 - w\), we reject \(H_0\) in favor of \(H_1\) when \([B \times p_3 + (1 - w) \times p_4] < .05 \times B\) under the model averaging MA\textsubscript{P}. Let \(R_3, R_4,\) and \(R_{MA}\) denote the event that \(H_0\) is rejected under L3, L4, and MA\textsubscript{P}, respectively. We consider the 4 cases. If \(R_3 \cap R_4\) occurs, it implies \(R_{MA}\) (case 1). If \(R_3 \cap R_4^c\) occurs (ie, \(H_0\) is rejected under L3 but not under L4), it is inconclusive (case 2). If \(R_3^c \cap R_4\) occurs, it is inconclusive (case 3). If \(R_3^c \cap R_4^c\) occurs, it implies \(R_{MA}\). To this end,

\[
P(R_{MA}) \leq P(R_3 \cap R_4) + P(R_3 \cap R_4^c) + P(R_3^c \cap R_4)
\]

and
\[ P(R_{MA}) \leq P(R_3 \cap R_4) + P(R_3 \cap R_5^C) + P(R_5^C \cap R_4) \]
\[ = P(R_4) + P(R_3 \cap R_5^C) \]

In other words, the type I error rate under MAp has the upper bound
\[ P(R_{MA}) \leq \min\{P(R_3) + P(R_5^C \cap R_4), P(R_4) + P(R_3 \cap R_5^C)\} \]

If we observe an unusual original sample against \( H_0 \), both the probability of \( R_3 \) and the probability of \( R_4 \) increase, so \( P(R_3^C \cap R_4) \) and \( P(R_3 \cap R_5^C) \) are fairly small. To this end, the type I error rate under MAp cannot be too far away from a range between the type I error rate under L3 and under L4. In the simulation study (“Simulation Result" and Table 2), we observe that \( P(R_{MA}) \) is close to the minimum of \( P(R_3) \) and \( P(R_4) \) under the null scenarios (scenarios 1-5).

**Model Averaging of \( M_1, \ldots, M_{J-2} \)**

Assuming model \( M_k \) for \( k = 1, \ldots, J-2 \), the log-likelihood function is given by (see Application of Model Averaging to Semiparametric Models subsection)
\[ l_k(\pi_1, \ldots, \pi_J) = \sum_{i=1}^{l} \sum_{j=1}^{n_i} \left( y_{ij} \log(\pi_j) + (1 - y_{ij}) \log(1 - \pi_j) \right) \]
with the restriction \( \pi_2 = \ldots = \pi_{k+1} \).

- For model \( M_1 \), it is the saturated model, so the maximum \( \hat{l}_1 \) can be achieved by letting \( \pi_j = \Sigma_i y_{ij}/n_j \), the observed proportion of toxic outcomes at dose \( x_j \).
- For model \( M_2 \) with the restriction \( \pi_2 = \pi_3 \), the maximum \( \hat{l}_2 \) can be achieved by letting \( \pi_j = \Sigma_i y_{ij}/n_j \) for \( j \) not being equal to 2 or 3 and \( \pi_2 = \pi_3 = (\Sigma_i y_{ij} + \Sigma_j y_{ij})/(n_2 + n_3) \) which is the pooled estimation.
- For model \( M_3 \), with the restriction \( \pi_2 = \pi_3 = \pi_4 \), the maximum \( \hat{l}_3 \) can be achieved by letting \( \pi_j = \Sigma_i y_{ij}/n_j \) for \( j \) not being equal to 2, 3, or 4 and \( \pi_2 = \pi_3 = \pi_4 = (\Sigma_i y_{ij} + \Sigma_j y_{ij} + \Sigma_j y_{ij})/(n_2 + n_3 + n_4) \).
- This pattern continued up to \( M_{J-2} \), where \( J \) is the number of fixed experimental doses.

The AIC of model \( M_k \) is \( \text{AIC}_k = -2\hat{l}_k + 2p_k \), where \( p_k = J - k + 1 \) is the number of free parameters in the model. Then the AIC weight of \( M_k \) is given by
\[ w_k = \frac{e^{-0.5\text{AIC}_k}}{e^{-0.5\text{AIC}_1} + \ldots + e^{-0.5\text{AIC}_{J-2}}} \]

**Acknowledgments**

The authors appreciate the comments and suggestions from reviewers.

**Declaration of Conflicting Interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by Undergraduate Research Opportunities Center (UROC) at California State University, Monterey Bay (CSUMB). Nathan Sanders was supported by the HSI Grant (US Department of Education Hispanic-Serving Institutions Program (STEM) Program (84.031C)—Grant #P031V11021).

**References**

1. Mattson MP. Hormesis defined. *Ageing Res Rev*. 2008;7(1):1-7.
2. Calabrese EJ. Hormesis: a revolution in toxicology, risk assessment and medicine. *EMBO Rep*. 2004;5(suppl 1):S37-S40.
3. May S, Bigelow C. Modeling nonlinear dose-response relationships in epidemiologic studies: statistical approaches and practical challenges. *Dose Response*. 2005;3(4):474-490.
4. United States Environmental Protection Agency. Benchmark Dose Software (BMDS) Version BMDS: 2.6.0.1. https://www.epa.gov/bmds/download-benchmark-dose-software-bmds. 2016. Updated 2016. Accessed March 2017.
5. Hunt DL, Bowman D. A parametric model for detecting hormetic effects in developmental toxicity studies. *Risk Anal*. 2004;24(1):65-72.
6. Bogen KT. Generic hockey-stick model for estimating benchmark dose and potency: performance relative to BMDS and application to anthraquinone. *Dose Response*. 2011;9(2):182-208.
7. Kim SB, Bartell SM, Gillen DL. Estimation of a benchmark dose in the presence or absence of hormesis using posterior average. *Risk Anal*. 2015;35(3):396-408.
8. Everitt BS, Skrendal A. *The Cambridge Dictionary of Statistics*, 4th ed. New York, NY: Cambridge University Press; 2010.
9. Akaike H. A new look at the statistical model identification. *IEEE Trans Autom Control*. 1974;19(6):716-723.
10. Bailer AJ, Noble RB, Wheeler MW. Model uncertainty and risk estimation for experimental studies of quantal responses. *Risk Anal*. 2005;25(2):291-299.
11. Moon H, Kim H, Chen JJ, Kodell RL. Model averaging using the Kullback information criterion in estimating effective doses for microbial infection and illness. *Risk Anal*. 2005;25(5):1147-1159.
12. Wheeler MW, Bailer AJ. Properties of model-averaged BMDLs: a study of model averaging in dichotomous response risk estimation. *Risk Anal*. 2007;27(3):659-670.
13. Shao K, Small MJ. Potential uncertainty reduction in model-averaged benchmark dose estimates informed by an additional dose study. *Risk Anal*. 2011;31(10):1156-1175.
14. Calabrese EJ, Baldwin LA. Can the concept of hormesis be generalized to carcinogenesis? *Regul Toxicol Pharmacol*. 1998;28(3):230-241.
15. Efron B. Bootstrap methods: another look at the jackknife. *Ann Stat*. 1979;7(1):1-26.
16. Tukey JW. *Exploratory Data Analysis*. Reading, PA: Addison-Wesley; 1977.
17. Kim SB, Bartell SM, Gillen DL. Inference for the existence of hormetic dose-response relationships in toxicology studies. *Biosci Biotechnol Biochem*. 2016;71(3):523-536.
18. Kang SH, Kodell RL, Chen JJ. Incorporating model uncertainties along with data uncertainties in microbial risk assessment. *Regul Toxicol Pharmacol*. 2000;32(1):68-72.

19. Piegorsch WW, An L, Wickens AA, West RW, Pena EA, Wu W. Information-theoretic model-averaged benchmark dose analysis in environmental risk assessment. *Environmentrics*. 2013;24(3):143-157.

20. Hall P, Heckman NE. Testing for monotonicity of a regression mean by calibrating for linear functions. *Ann Stat*. 2000;28(1):20-39.

21. Baraud Y, Huet S, Laurent B. Testing convex hypotheses on the mean of a Gaussian vector: application to testing qualitative hypotheses on a regression function. *Ann Stat*. 2005;33(1):214-257.

22. Meyer MC. Semi-parametric additive constrained regression. *J Nonparametric Stat*. 2013;25(3):715-730.

23. Anderson EL, Albert RE, McGaughy R, et al. Quantitative approaches in use to assess cancer risk. *Risk Anal*. 1983;3(4):277-295.