Model Choice Considerations for Fitting Concentration-response Curves

by
William Nathaniel Cumberland

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington
2012

Committee:
Youyi Fong (Chair)
Ying Huang
Xuesong Yu

Program Authorized to Offer Degree:
School of Public Health - Biostatistics

University of Washington

Abstract

Model Choice Considerations for Fitting Concentration-response Curves

William N. Cumberland

Chair of the Supervisory Committee:
Affiliate Assistant Professor Youyi Fong
Department of Biostatistics

Many bioassays use inverse prediction, sometimes referred to as calibration, in order to estimate an unknown concentration from a machine response, using a training set. This thesis attempts to evaluate the effect model selection has upon the inverse prediction, specifically with regard to the four and five parameter log logistic models. It is known that the bias-variance trade off relationship is at the center of this decision. We conducted two simulation studies and one cross validation study to evaluate the effects of asymmetry, denoted by f, and the effects of noise, denoted by σ, on the model's accuracy. Our criterion measures were a global MSE measure referred to as S^1 and an absolute relative bias measure referred to as S^2.

Our initial simulations, which involved a direct comparison of the two fitted curves with respect to the true curve, found that the five parameter performed better for both scores (S^1 and S^2) when dealing with asymmetric data, as anticipated. Our second set of simulations, which evaluated our fitted curves' accuracy in predicting future simulated samples, found an unexpected trend for the S^2 scores. The four parameter outperforms the five parameter with regard to the absolute relative bias for a majority of the positive asymmetric ($f > 1$) cases. Our cross validation study, using actual lab data, was conducted to further investigate this finding. The cross validation study confirmed that for highly estimated asymmetric levels ($f > 3.5$), a majority of the cases reported that the four parameter provided a better prediction. From these last two studies, we could conclude that the four parameter provides a better fit for future predictions than the five parameter for almost all levels of asymmetry (excluding $f < 0.8$).

There are some possible explanations for these findings. During our relative bias measurements, we had to omit certain points due to extrapolation into unrealistic concentrations. We tried various exclusion criteria but the pattern remained relatively the same. Another possible explanation may be that the fitting algorithm actually passes the asymmetry to the other parameters at higher levels of asymmetry, allowing the four parameter to undermine the f parameter's intention at the higher levels. In the end, while the four parameter appears to perform better for most of the positive asymmetric regions, a further study of this incident is needed to clarify exactly what leads to this result.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my family for their encouragement over the past few years. Their constant support over the past few years has been invaluable. I look forward to seeing my family much more often back in California, including the two new dogs.

I would also like to thank my advisor Scott Emerson for his help in tackling the program. No matter how difficult or overwhelming the work appeared, he always provided advice on how to approach it.

I would like to thank Youyi Fong for being my advisor, for always being helpful and patient. No matter how confusing something was, he has always been willing to explain in detail. I would like to also thank Youyi for his constant input during the lengthy thesis process.

I would like to thank Xuesong Yu for helping with the Thesis review process, providing valuable feedback.

I would also like to thank Ying Huang for joining the committee on short notice and for generously accepting.

Lastly, I would like to thank the whole Biostatistics department for their company over the past three years. To those still on the PhD track, I wish you the best of luck!

1 Introduction

Every bioassay relies upon a well defined model in order to properly associate an outcome with the sample's concentration. We will refer to our machine measured outcome as Y and our sample concentration as X during the course of this thesis. A number of studies have used models for predicting unknown single outcomes of Y, such as blood pressure for a potential patient, with known concentrations of drugs and/or other predictors. This paper study instead deals with inverse prediction, which is sometimes referred to as calibration, where we predict an unknown X, estimating concentrations, through the use of training datasets. Our study aims at differentiating the advantages of two different model assumptions formed from the four and five parameter log logistic functions. We will also expand and apply our findings to a recent example involving HIV vaccine treatments wherein we attempt to measure unknown cytokine concentrations.

This thesis is mostly a simulation study that compares the four and five parameter models by looking at their relative performances with various ranges of asymmetry and present noise in the reported machine outputs. In order to construct these models, we will need to make a few assumptions about the general design used, as well as provide some background on designs and how they can factor into the study. Generally, the labs performing these assays already have some of this information, such as reasonable concentration ranges, machine noise levels, and slight asymmetry tendencies from prior experiments or background. There are three sources of variation involved with our estimator for the unknown concentrations. The first is machine noise, the variance of Y, which has a direct impact upon the variance of our estimator of interest. We are interested in the effects of this noise, which we denote with ε and is distributed as $N(0,\sigma^2)$. It is also plausible that the amount of variation or noise present can depend on X. For example, machine measurements may be pinpoint accurate when the concentration is either incredibly low or high, and slightly more volatile when there is a moderate concentration present. The second source is model choice, which is the added variance induced by over-parametrizations of the truth. This thesis is focused primarily on the variance and bias trade-offs between the four and five parameter models. The final source is the design choice, where the allocation of our sample affects the accuracy of the inverse prediction. For this thesis, we maintain the same design throughout using the standard design in order to avoid further complications involving this factor. We are primarily interested in the criterion scores for varying levels of σ and f for the two models of interest.

When we perform these assays, we usually collect our unknown samples through a form of "baking" or scanning, acquiring our machine measured responses. However, this cooking process has several drawbacks. The first problem is that each experiment will have different results due to uncontrollable experimental conditions. This means the results in experimental 1 will not be

similar to experiment 2's results since the underlying parameter estimates (for the model in mind) have changed. The second problem is associated with the cost of these bioassays. It is usually unreasonable to bake multiple samples across multiple plates to address these experimental issues. Thus we need to include what is known as a design sample on each plate, which serves as a unique data set for that experiment and can provide the unique parameter estimates for our model. However, this brings us to our third and final problem, which is the limited space available on each plate. Usually, these plates have 96 wells, which need to accommodate all of the unknown samples of interest and the design sample. Usually, bioassays will have 2-4 replicates per distinct X for the design, which can take around 20 to 48 of the well positions. To make matters more complex, designs can be modified in three areas: the sample allocation to each X (can be different for each), number of distinct X, and the locations for each distinct X. We will therefore focus on the smallest design standard, which consists of 20 points, allocated evenly across 10 distinct points on the log scale of X, when comparing our two model assumptions. In the end, our aim is to solely address the machine noise variation, given our experimental choices.

The overarching problem is that the optimal model and design choice changes with respect to time. Gottschalk and Dunn [2005] hints that improvements in technology will continue to reduce the noise in the machine measurements, revealing subtle response patterns that were previously masked by noise. In this paper, we are interested in the increasing asymmetry observed in recent dose response studies. As a result, new models are designed to accommodate these discoveries. However, there are no established guidelines on how to use these findings on a regular basis. Instead, there is confusion regarding which method is the best. As a result, potentially better methods, such as the five parameter model, are ignored in favor of previously established protocols. Papers, which provide a general approach under specific assumptions that may or may not occur, have not provided a road map or a guideline offering different solutions depending on the scenario.

In Gottschalk and Dunn [2005], they stressed the use of the five parameter log logistic in the presence of asymmetric data on the premise that machine error is or will eventually be negligible. As a result, the bias will far outweigh the variance component when measuring the MSE under the four parameter model. When one switches from the four to the five parameter model, we should expect the bias to decrease substantially and the variance to potentially increase only slightly. However, their paper only focused on the amount of error (bias) dropped when switching from the four to the five parameter for various levels of asymmetry. They never considered the small amount of variation present could increase and potentially make the switch from the four to five parameter worthless for various levels of asymmetry. It was never clear how small an error would be needed to achieve the simple automatic five parameter dominance assumption. With the current state of technology and the likely future, it may be reasonable to use the four parameter in certain cases for a while longer.

In François et al. [2004], they investigated whether there existed any alternative designs which could outperform the standard uniform allocation design. They discovered that there existed new designs which achieved better specific criterion scores, which meant these designs should achieve a lower variance. However, this paper limited itself to the four parameter model and never investigated whether the same designs were viable for the five parameter model. In addition, the criterion scores also focused only on the variance and not the MSE as a whole. It never included the effect of bias with the new designs. It was possible that the new designs are susceptible to higher bias. In the end, there may exist some design which performs better overall for both the five parameter and four parameter models, allowing labs to freely select the proper model if there is strong asymmetry present.

To address these concerns, we had developed a new criterion score which is related to the MSE, called the S^1 Criterion. It is the integral of the squared difference between the underlying truth and the fitted curve of interest. The criterion allows us to effectively measure how well a model assumption performs in lowering the MSE when compared to an alternative assumption. The lower the criterion score, the better. We also have a second criterion,called S^2, which is the absolute relative bias. This criterion is the absolute difference between our fitted and truth, relative to the truth. Using these criterions, we evaluated performances by taking the differences between the four and five parameter models' criterion scores for various levels of asymmetry and error levels. The final goal was to determine the regions of asymmetry and error which favored the four and five parameter assumptions. We anticipated the variance component of the MSE to increase with higher errors. Further, we expected the bias to increase for the four parameter assumption for increasing levels of asymmetry. Our end goal was to determine where the four parameter loses its advantage due to increasing levels of asymmetry alongside various levels of error. We also included sample tests for the criterions to evaluate how well they perform in a realistic scenario along with a cross validation study of a recent HIV vaccine trial.

2 Four and Five Parameter Log Logistic Model Selection

2.1 The Basic Five and Four Parameter Models

All dose responses are assumed to follow a monotonic relationship where increasing concentrations yield greater responses, resulting in an S Shaped curve (or a reverse S in some cases where higher concentrations are associated with decreased responses) with an upper and lower asymptote. The four parameter log logistic function was designed to fit closely to these dose responses. However, recent improvements have shown there may exist some asymmetry in the responses, requiring an additional parameter to accommodate. To adjust for this asymmetry, the five parameter function was introduced,

$$\begin{aligned} f_{5pl}\left(t; c, d, b, e, f\right) &= c + \frac{d-c}{\left\{1+\left(\frac{x}{e}\right)^{b}\right\}^{f}} \\ &= c + \frac{d-c}{\left\{1+\exp\left(b\left(t-\log e\right)\right)\right\}^{f}}, \end{aligned} \tag{1}$$

where x is the standard concentration, $t\left(x\right) = \log x$ is the log concentration, and $y = f_{5pl}\left(x\right)$ is the machine measured output. The parameter c represents the lower asymptote while d represents the upper asymptote. The f parameter represents the asymmetry in the five parameter model. The four parameter model is a special case when $f = 1$, as seen in Figure 1. The parameter e is considered the inflection point only when $f = 1$ while b is considered a general term.

Even though we are simply deciding between two models, the choice is not simple. If one underfits a model, even if the underfit is just a lower parameter variant of the truth, the bias will increase. Likewise, If one over fits a model, they will increase the variance. With either of the two improper model selections, the Mean Squared Error (MSE) increases. If we were to pick a good model, be it the true model or a close yet different model, with no over or under fitting, then we would expect to see little to no bias and the desired minimal variance. The problem is that there does not exist a single perfect model that can be used for every bioassay. That is an unrealistic ideal and the four and five parameter models were developed to be a general approach. The decision between the four and five parameter models boils down to a bias and variance trade-off relationship for supposedly lightly asymmetric data. For heavily asymmetric data, we might intuitively select the five parameter without taking into consideration the four parameter. In the end, we are aiming to find the best model assumption which has the lowest MSE and absolute relative bias for various levels of asymmetry and error.

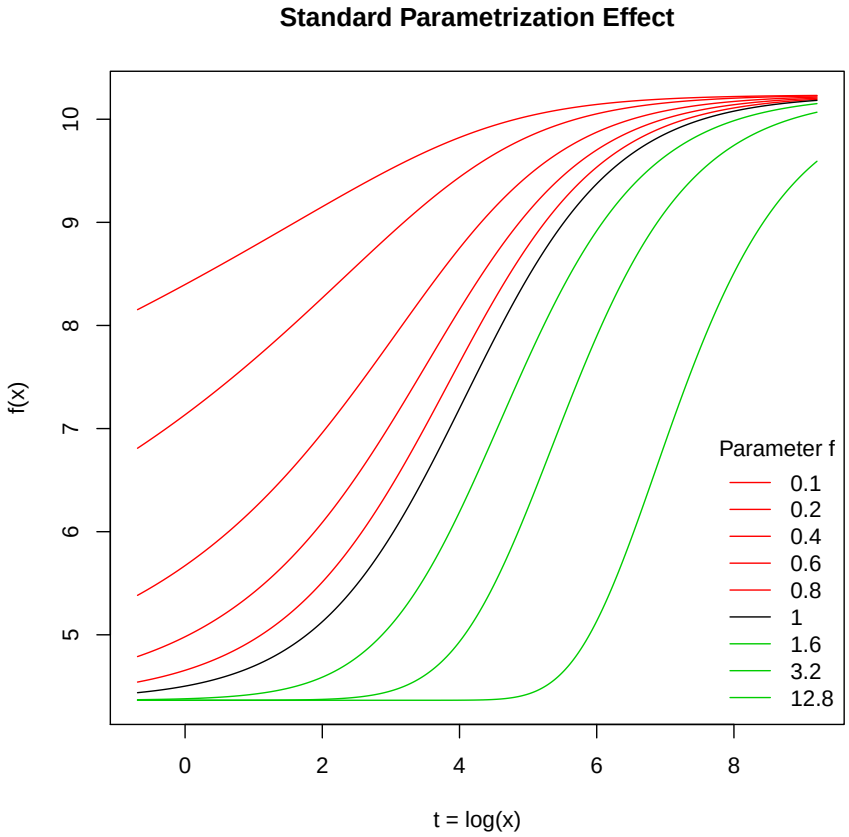

Figure 1: The effects of various levels of asymmetry with the standard five parameterization form, with parameters: c = 4.37, d = 10.24, e = 59.07, b = -0.92.

2.2 History of Four and Five Parameter Models

For a while, the four parameter model was considered the best choice. The standard process involved collecting the output measurements, such as luminescence ratings, from our unknown samples and designs. Using our design data set, we would fit the curve and then use the curve to find the unknown concentrations via inverse prediction. The four parameter model has been used for a long time to acquire meaningful results and was the standard choice. As time progressed, our tools and computing power continued to improve as well, reducing the machine noise when reporting Y.

However, these improvements should not be confused with improvements across experiments. The experimental conditions between experiments still vary

too much, returning different parameter estimates each time. With the recent improvements, we are getting to the point where deciding between the four and five parameter in certain situations can yield better results. In addition, this evidence brings up multiple questions: "Was the four parameter model an under fit of the truth? Is the five parameter closer to the truth or simply an overfit? Is the truth something entirely different?". However, the five parameter's adoption has been slow, partly on concerns of a potential increase in the variance for symmetric data. So which model should people use by default? The answer is: "There is no definitive default at this time". With the current state of technology lowering the machine noise, the bias component of the MSE is no longer insignificant in contrast to the variance component (Gottschalk and Dunn [2005]) We are at the point where the four and five parameter model selection can actually have an impact on the MSE depending on certain factors such as the asymmetry present. While the final model decision is left to each lab to decide, this paper will cover the potential pitfalls of each assumption, as well as attempt to provide guidelines on model selection for the present and future.

2.2.1 Re-parametrization of the Five Parameter Model

While the standard five parameter model is used in most of the fitting programs, it actually does not provide a direct visual of the asymmetry levels alongside the other variables. The variables e and b do not have a clear interpretation when f is not equal to 1, thus whenever the asymmetry shifts, so does their exact interpretation and roles with respect to the asymmetry. Another parametrization of the five and four parameter log logistic, the GH parametrization, allows for a more meaningful interpretation of the variables

$$f_{5pl}\left(t; c, d, g, h, f\right) = c + \frac{d-c}{\left\{1 + \left(\frac{1}{f}\right)\exp\{-\left(\frac{h}{d-c}\right)\left(1+\frac{1}{f}\right)^{f+1}\left(t-g\right)\}\right\}^{f}}$$

$$g(e, b, f) = \log(e) - \frac{1}{b}(\log(f))$$

$$h(c, d, b, f) = (-b)\frac{(d-c)}{(1+\frac{1}{f})^{(f+1)}}$$

where g is the inflection point and h is the slope at the inflection point for all values of f (Fong et al. [2012] and references therein).To effectively observe the differences in asymmetry, we must also adjust the parameters e and b in the standard model, resulting in different curves as seen in Figure 2.

2.3 Model Choice Problem

When we are deciding between the two log logistic models, where the only difference is the number of parameters, we are basically weighing the bias-variance

trade off in an attempt to minimize the MSE. Our preliminary simulations found for a majority of our cases, the four parameter performed poorly, specifically with highly asymmetric simulations as anticipated. As we initially thought, the four parameter was likely to suffer from bias issues when dealing with asymmetric data. However, these simulations had no direct control or direction with regard to f or σ, and were simulated using various sets of parameters. Therefore any direct conclusion about the bias/variance trade off optimality involving varying degrees of asymmetry required a more strenuous study.

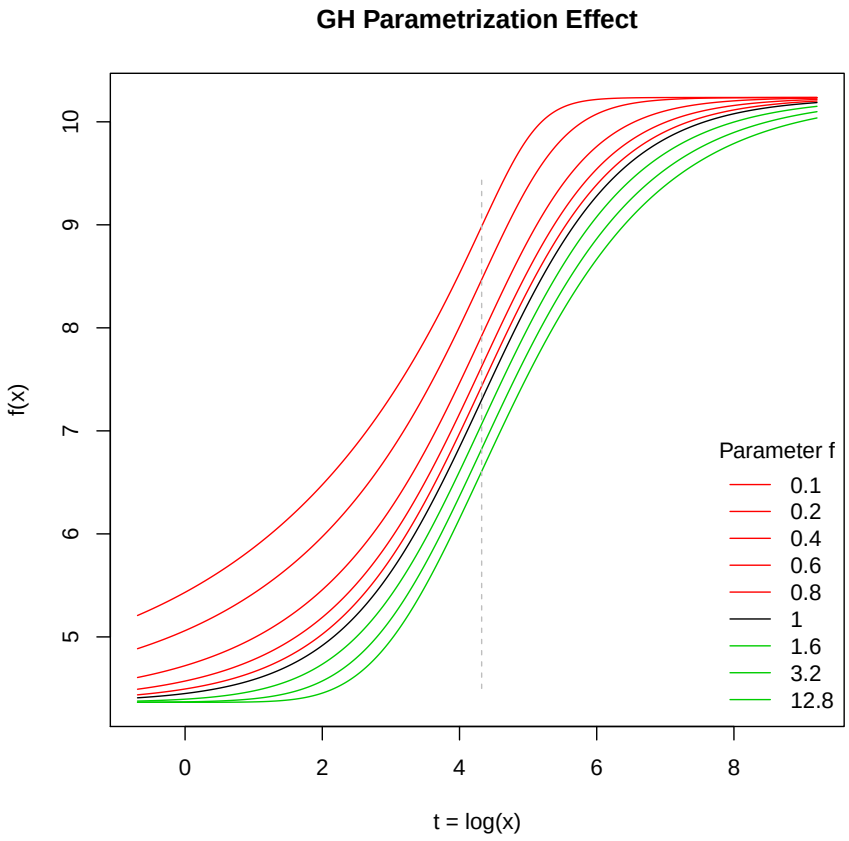

Figure 2: The effects of various levels of asymmetry with the "GH" five parameterization form with parameters: c = 4.37, d = 10.24, g = 4.32, h = 1.43. The grey dotted line represents the inflection point.

As Gottschalk and Dunn [2005] mentioned, they stated that the five parameter will eventually be the best alternative due to the constant bias outweighing the ever decreasing variance. However, it was never clear for which levels of

asymmetry it was better to assume the four parameter model over the five parameter model, given certain error structure levels. Thus we are interested at what levels of asymmetry and present error do you want to decide between either a four or five parameter model assumption when fitting.

2.4 Simulations

We conducted a simulation experiment to compare the four and five parameter model fitting assumptions by calculating the S^1 and S^2 Criterion Scores, which are considered a global measure of MSE and absolute relative bias, for various levels of asymmetry and error present in the data. We randomly generated the data with the following form,

$$Y = f_{5pl}^{GH}(t; c, d, g, h, f) + \varepsilon \tag{2}$$

where Y is the machine measured log fluorescence intensity response value, t is the $\log(x)$ where x is the concentration, $f_{5pl}^{GH}(t; c, d, g, h, f)$ is the true curve using the GH parametrization (Figure 2), and ε is the error term, which is drawn from a $N(0, \sigma^2)$ where σ ranges from 0.04 to 0.20. Our simulations assume that the machine noise was constant with respect to X, while it is entirely plausible that in some actual assays the noise may vary with respect to the concentration. The only parameters we varied between each set of simulations were f and σ using the GH parametrization. The other parameters c, d, g,and h were left unchanged.

Both the standard and GH parametrization use and interpret c and d in the same way - as asymptotes for the curves. The parameters e and b from the standard parametrization do change when we change f since we are using the GH parametrization for the truth. They change with respect to f so that the inflection point, as well as the slope, remains the same for the varying levels of f. If we had used the standard parametrization with varying f, our inflection point and the slope at the inflection point would have changed with f. The end effect would lead to some undesirable and unrealistic curves, as seen when comparing Figure 2 to Figure 1.

For our simulations, our realistic parameter values come from a HIV vaccine trial which examined blood samples drawn from patients who were recently administered a vaccine. These blood samples contained unknown cytokine concentrations, which were created in response to the vaccine. The accompanying design consisted of cytokine concentrations of $10,000$ with additional 3 fold dilutions down to a concentration of about 0.5, with two samples per dilution level. From this study, they fitted a realistic design curve with the following estimated GH parameters: $c = 4.37$, $d = 10.24$, $g = 4.32$, $h = 1.43$,and $f = 1.25$. These parameter estimates served as our realistic truth, where only f would change. To get a relative idea of how much f should change in order to ensure we simulated only realistic curves, we looked into additional assays using a

five parameter fit, and found f ranging mostly from $f = 0.1$ to $f = 12.8$. The GH parametrization continued to provide realistic curves for various levels of asymmetry (f) as long as the other parameters remained the same. We used the GH parametrization when simulating the machine response for its ability to control and we used the standard parametrization for the fitting process since it was the standard in the DRC fitting program.

We used the same design when comparing the four and five parameter models, which had a size of 20 with 10 distinct $X\prime s$, each with 2 samples per distinct X. On the natural log scale, the design concentrations were evenly allocated across -0.68 to 9.21 (on the standard scale, the concentration range went from 0.5 to $10,000$, using 3 fold dilutions: $10,000$, $3,333$, $1,111$, etc.). With a design X and its corresponding Y, we then fitted the standard four and five parameter models with the DRC package by Ritz and Streibig [2005] and the Ruminex package by Fong [2012] for R.

2.5 Direct Criteria Evaluation

Our first set of simulations involved comparing the two fitted curves of interest with respect to the assumed truth curve by calculating two different criterion scores. The first is the integral of the difference between the true and fitted curves squared, which we shall call S^1 and it serves as a global measure of MSE.

$$S^1 = \int_{\min(t)}^{\max(t)} (\widehat{f}(t) - f_0(t))^2 dt \tag{3}$$

where t is the log concentration (recall that x is the standard concentration), f_0 is the truth using the correct parameters to form the curve, and $\widehat{f}$ is the fitted curve, using the parameter estimates from the simulated data using the DRC package(Ritz and Streibig [2005]). The integral used for S^1 is taken from the minimum to the maximum of $\log(x) = t$,our design of interest. From the previous section, they were -0.677 and 9.210 respectively, from the logs of concentrations 0.5 and $10,000$.

Our second is referred to as the absolute relative bias, represented by S^2,

$$S^2 = \int_{\min(y_0)}^{\max(y_0)} \frac{|e^{\widehat{f}^{-1}(y)} - e^{f_0^{-1}(y)}|}{e^{f_0^{-1}(y)}} dy \tag{4}$$

where S^2 is the criterion score for the absolute relative bias and y_0 represents the range of y for the true curve given our defined truth parameters and design range. The relative bias in this case uses the standard concentration values instead of the log concentration, hence the exponentiation involved. For acquiring S^2, we integrated from the minimum to the maximum response values of the true curve with regard to our concentration range X. There is no single minimum or

maximum for y_0 since it varies with regard to f. Using the original parameters, with $f = 1.25$, we had a minimum of 4.39 and a maximum 10.17 (for $f = 1$; it was 4.41 and 10.19). However, sometimes the fitted curves will never reach the minimum or maximum value of y_0 (e.g. $\min(y_0) < \widehat{f}(\min(t))$). As a result, if we attempted to directly integrate, we would have some cases where $\widehat{f}^{-1}(y) = \pm\infty$, throwing the results. Since we have no interest in extrapolating into regions of meaningless negative concentrations or highly saturated concentrations, we replaced $\widehat{f}^{-1}(y)$ with $(\min(t), \max(t))$ respectively whenever $\widehat{f}^{-1}(y)$ returned an extrapolated result.

With these outlines, we performed 1000 simulations for each combination of f and σ, recording the S^1 and S^2 criterion scores for the four and five parameter fitting assumptions. We calculated the average for each criterion score with respect to every combination of f and σ for the five and four parameter models, and then took the differences between the four and five parameter models for each f and σ:

$$W(f, \sigma) = \frac{\sum_{i=1}^{1000} S_i^{5pl}(f, \sigma)}{1000} - \frac{\sum_{i=1}^{1000} S_i^{4pl}(f, \sigma)}{1000} \tag{5}$$

where W is the criterion score mean difference between the two model assumptions. The ideal scenario is when the fitted curve and truth are identical, yielding a criterion score of 0 for both the S^1 and S^2. Thus, the lower the criterion score, the better the fit is. With that in mind, a positive score for W is associated with the four parameter model having a better fit while a negative score is associated with the five parameter model having a better fit. However, it is possible that for some combinations of f and σ some criterions will favor the five parameter while at the same time other criterions will favor the four parameter model.

$\sigma \backslash f$	0.10	0.20	0.40	0.60	0.80	1.00	1.60	3.20	12.80
0.04	-0.173	-0.108	-0.037	-0.010	-0.001	0.001	-0.006	-0.028	-0.060
0.06	-0.172	-0.107	-0.036	-0.010	-0.000	0.002	-0.005	-0.028	-0.059
0.08	-0.171	-0.106	-0.035	-0.008	0.001	0.003	-0.004	-0.027	-0.059
0.10	-0.170	-0.104	-0.033	-0.007	0.002	0.004	-0.003	-0.026	-0.058
0.12	-0.169	-0.102	-0.031	-0.005	0.004	0.006	-0.001	-0.025	-0.057
0.14	-0.167	-0.099	-0.029	-0.002	0.007	0.008	0.001	-0.024	-0.056
0.16	-0.166	-0.097	-0.026	0.000	0.009	0.011	0.002	-0.022	-0.055
0.18	-0.164	-0.094	-0.023	0.003	0.012	0.013	0.004	-0.021	-0.053
0.20	-0.162	-0.091	-0.020	0.006	0.015	0.016	0.007	-0.019	-0.051

Table 1: Simulation results for the direct curve integration evaluation comparing the mean of the S1 Five parameter simulation results to the mean of the S1 Four parameter results. (Global MSE evaluation)

Table 1 presents the simulation results for the global MSE. As we expected when dealing with increasing σ, the four parameter model assumption gains an advantage when dealing with an increasing variance at a set level of asymmetry (source of bias). This bias effect increases the further we stray away from $f = 1$ However, when $f = 1$, the four parameter model is the ideal model, since it is the theoretical truth and thus the bias element shrinks towards nothing. Hence, the four parameter will always have a better mean S^1 score for this segment. Up to this point, the simulations have followed everything we expected in theory.

Looking further into the simulations, we noticed that the four parameter assumption still performs better than the five parameter assumption for slight levels of asymmetry. The amount of bias introduced under the four parameter model is minimal for slight levels of asymmetry (about $0.80 < f < 1.25$). When we take into consideration the relative bias simulation results in Table 2, we see a similar pattern favoring the four parameter model when there is only a slight asymmetry present, reinforcing our early conclusions about the bias component being minimal for slight asymmetry. For moderate levels of asymmetry (around $f = 1.6$, potentially up to 2.0), if there is a decent amount of error ($\sigma \geq 0.14$), the four parameter model can still outperform the five parameter model. The scale of improvement is something else entirely; where $f = 1.60$ and $\sigma = 0.12$ yields a tiny difference of 0.005, while for $f = 1.60$ and $\sigma = 0.06$, that difference is instead 0.064 in favor of the five parameter. The potential gains under the four parameter are moderately small. At the same time, the potential losses scale far worse for the four parameter under a poor assumption (i.e. very asymmetric data and/or low σ). It appears that the losses are almost 10 times worse than the potential gains. What we can conclude so far from our simulations is that, if we know the general range for either f or σ somehow, we can almost safely pick when to use and when not to use the four parameter model. If we blindly pick the four parameter, it is theoretically risky and can lead to far worse results.

It is interesting to note that both the global MSE and relative bias tables share a similar cone shaped pattern, as seen in Figure 3, which represents the region where the four parameter model assumption is better. For increasing σ the cone widens, but not uniformly with respect to decreasing and increasing levels of f. A negative shift toward zero is much stronger in introducing bias than is an equal positive shift from $f = 1$.

2.6 Testing Criteria Evaluation

Simulations involving testing data allow us to check the accuracy of the inverse prediction since it emulates an actual bioassay process. It also allows us to observe the effect the number of replicates has, unlike the prior simulation. The main distinction is that the prior simulations evaluated the scores with respect to

$\sigma \backslash f$	0.10	0.20	0.40	0.60	0.80	1.00	1.60	3.20	12.80
0.04	-0.899	-0.703	-0.369	-0.154	-0.025	0.015	-0.099	-0.283	-0.425
0.06	-0.790	-0.623	-0.308	-0.109	-0.005	0.022	-0.064	-0.235	-0.383
0.08	-0.680	-0.543	-0.254	-0.074	0.010	0.029	-0.040	-0.196	-0.343
0.10	-0.571	-0.467	-0.206	-0.047	0.022	0.036	-0.021	-0.164	-0.307
0.12	-0.468	-0.396	-0.164	-0.025	0.032	0.043	-0.007	-0.138	-0.275
0.14	-0.369	-0.329	-0.126	-0.006	0.042	0.051	0.005	-0.117	-0.247
0.16	-0.281	-0.270	-0.093	0.011	0.051	0.057	0.015	-0.098	-0.222
0.18	-0.199	-0.215	-0.064	0.026	0.060	0.064	0.023	-0.083	-0.201
0.20	-0.123	-0.166	-0.037	0.040	0.068	0.071	0.031	-0.069	-0.181

Table 2: Simulation results for the direct curve integration evaluation comparing the mean of the S2 Five parameter simulation results to the mean of the S2 Four parameter results. (Absolute Relative Bias Measure)

two curves, the fitted and the truth, while these new testing simulations instead evaluate how well the fitted curve performs in predicting future concentrations. In better detail, we have generated a large dataset containing both our training sample, which is used for fitting the four and five parameter curves, and a testing sample, which assumes the role of our unknown concentrations of interest. While we are evaluating how well our training set helps us in predicting our testing set, we still have and use the true concentration values to calculate the absolute relative bias. This simulation therefore recreates the bioassay dish experiment with the two sample groups: a standard sample, representing the design, and an unknown sample, representing our concentration of interest that we wish to inversely predict.

As with most bioassays, we never know the truth, even when there are plenty of prior experiments dealing with the same bioassay. The purpose of these simulations was to evaluate how well each model performed for individual averaged responses, emulating their potential bioassay performance. If the fitted curve is close to our simulated unknowns, we can then conclude that the model assumptions are good.

On a more technical level, the testing simulations allow us to evaluate how practical the models are for bioassays. It allows us to see the fitting algorithms performance with regard to each assumption. For example, one assumption may yield consistent fits, which in turn serves as a good source of prediction for the unknown samples. Another assumption may yield inconsistent fits, where the curves vary significantly due to the slight noise. As a result, these curves would be poor in predicting future samples drawn from the same population. However, there are more possibilities when it comes to this form of testing. It is entirely possible that an incorrect model assumption can yield consistent results while the correct model assumption may return inconsistent results due to the fitting

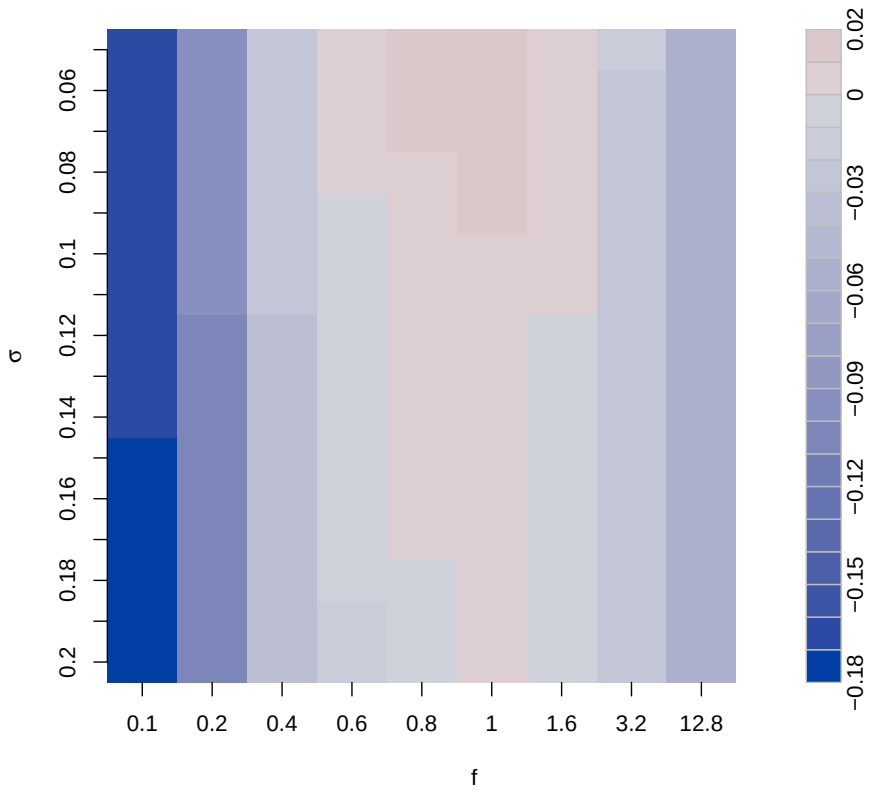

Figure 3: Color plot of Figure 1: S1-Simulation results for relative MSE - emphasizing the regions of f and sigma

process. In the end, it could be determined by the fitting algorithms accuracy for each assumption, not whether the models are the correct choice.

For our testing simulation, we used 1000 simulations with 1,2,3 and 20 replicates per distinct X of interest for the unknown sample. For simplicity and to ensure we drew our sample from the same population, we decided to generate the unknown samples using the same distinct $X's$ from the design set. For our S^1, validation test, it would be average of:

$$S_j^1 = (\widehat{f}(t_j) - \overline{y_j})^2 \tag{6}$$

where j is an identification associated with the concentration dilution of interest, $\overline{y_j}$ was the average of the responses associated with a supposed log concentration

t_j(if only 1 replicate: $y_j = \overline{y_j}$) , while $\widehat{f}(t_j)$ was the fitted curve's response (for either the four or five parameter) for the specified log concentrate.

For our S^2 test, which evaluates the absolute relative bias at selected points, we calculated the average of:

$$S_j^2 = \frac{|e^{\widehat{f}^{-1}(\overline{y_j})} - x_j|}{x_j} \tag{7}$$

where x_j was the concentration associated with $\overline{y_j}$, and $\widehat{f}^{-1}(\overline{y_j})$ was the fitted curve's x for the specified $\overline{y_j}$. There are a few problems with the aforementioned measurements, similar to the problems for the direct testing with the absolute relative bias. Like before, there are times when $\overline{y_j} < min(\widehat{f}(x))$ or $\overline{y_j} > max(\widehat{f}(x))$, resulting in $\widehat{f}^{-1}(\overline{y_j}) = \infty$. However, we can not set $\widehat{f}^{-1}(\overline{y_j}) = (\min(x), \max(x))$ for the sampling tests, since it could result in $S_i^2 = 0$, signifying an optimal measurement when in fact it was off. Since we are measuring the average of our criterion scores and that these samples were arbitrarily selected, we could remove these points. The end result is that the remaining points will have a heavier weight with regard to the simulation, and may in turn potentially bias the results in favor of one of the assumptions.

These tables present the various differences between the four and five parameter assumptions when evaluating the absolute relative bias for various levels of f and σ. While some results may mirror the pattern in the previous simulations for the criterions involving the truth, there are some incidents where the validation results do not reflect the previous simulations.

In Table 3, we began to see something totally different in contrast to tables 1 and 2. Instead of higher σ being associated with an improved four parameter advantage across the board, it instead followed a more haphazard pattern. For example, when $f < 1$, it appears the five parameter model gains for moderate σ(appears to the best at around 0.06 to 0.14), which goes against the concept of the five parameter advantage weakening with regard to increasing σ (even when the five is still better overall). While the five parameter still has the strong advantage in the negative asymmetry region ($f \leq 0.60$), the four parameter fit appears to be a better model for future cases when dealing with most positive asymmetric regions (Note the $f = 12.80$ case). This trend appears to get stronger with increasing f and is not simply unique to $f = 12.8$. These strange regions appear to debunk our original cone region where we anticipated the four parameter to outperform. However, there two things to take into consideration for these odd patterns. The first is that the simulation currently uses 1 replicate per distinct sample - a large number of bioassays use at least 2 replicates per unknown sample. Thus this odd case is not a cause for alarm yet. In addition, we are including all potential estimated points that can be measured, whether

$\sigma \backslash f$	0.10	0.20	0.40	0.60	0.80	1.00	1.60	3.20	12.80
0.04	-0.283	-0.163	-0.069	-0.023	-0.000	0.001	-0.035	-0.067	0.020
0.06	-0.334	-0.164	-0.054	-0.018	-0.002	0.001	-0.011	0.003	0.104
0.08	-0.311	-0.163	-0.055	-0.015	-0.002	0.001	0.007	0.049	0.163
0.10	-0.299	-0.167	-0.062	-0.022	-0.002	0.003	0.021	0.087	0.208
0.12	-0.267	-0.170	-0.066	-0.018	-0.010	0.004	0.033	0.115	0.244
0.14	-0.275	-0.185	-0.058	-0.027	-0.014	0.000	0.040	0.133	0.268
0.16	-0.244	-0.186	-0.062	-0.036	-0.005	0.003	0.048	0.147	0.283
0.18	-0.234	-0.177	-0.067	-0.033	-0.016	0.009	0.056	0.157	0.296
0.20	-0.229	-0.182	-0.073	-0.032	-0.010	0.008	0.064	0.164	0.303
Number of Missing									
0.04	1.241	0.751	0.234	0.057	-0.012	0.008	0.203	0.422	0.620
0.06	1.153	0.699	0.211	0.057	0.017	0.021	0.162	0.389	0.631
0.08	0.959	0.592	0.197	0.035	0.010	0.024	0.135	0.359	0.517
0.10	0.827	0.514	0.168	0.012	0.010	0.032	0.114	0.284	0.433
0.12	0.687	0.451	0.149	0.007	0.019	0.019	0.091	0.259	0.360
0.14	0.592	0.385	0.096	0.025	0.015	0.039	0.100	0.224	0.315
0.16	0.487	0.337	0.071	0.018	0.000	0.034	0.103	0.196	0.283
0.18	0.413	0.301	0.054	0.036	0.020	0.034	0.081	0.168	0.249
0.20	0.366	0.245	0.064	0.015	0.011	0.017	0.048	0.135	0.206

Table 3: Absolute Relative Bias comparison with only 1 replicate per unknown sample with no LODi restrictions. The bottom half represents the average missing observations due to exclusion for the five parameter model minus the average missing observations for the four parameter model. A higher score represents more missing cases for the five parameter model.

or not they are well within reasonable detection ranges, which we will cover later on.

The lower half of Table 3 also presents the overall differences between the two model assumptions' missing averaged observations. From the count section, the five parameter appears to have a larger number of observations omitted from its analyses for both regions of asymmetry (negative and positive). Upon closer inspection, the five parameter appears to miss more observations depending on the value of f while the four parameter does not appear to miss any with regard to changes in f. However, both tend to miss observations as σ increases. The reason why the five parameter appears to improve with increasing σ is due to the fact that the five parameter can't seem to lose much more observations with increasing σ when inside the asymmetric regions. This may explain why the five parameter appears to be gaining an advantage for increasing σ, since the four parameter's valid estimate counts seems unaffected by the asymmetry and only σ. It is plausible that the strange trend may be attributed to the odd samples located at the lower and higher concentrations, which in turn may be controlled with additional replicates. In the end, if we increase our replicates towards

infinity, we should see something similar to the direct simulation results. Given we are dealing with only 1 replicate, we can only conclude so far that 1 replicate for each distinct sample is unstable. To determine if the trend persists, we need to look into the higher replicate cases.

$\sigma \backslash f$	0.10	0.20	0.40	0.60	0.80	1.00	1.60	3.20	12.80
0.04	-0.264	-0.163	-0.077	-0.021	0.003	0.003	-0.048	-0.112	-0.025
0.06	-0.287	-0.167	-0.058	-0.017	0.003	0.003	-0.030	-0.048	0.048
0.08	-0.306	-0.165	-0.045	-0.015	-0.003	0.002	-0.012	-0.002	0.107
0.10	-0.299	-0.157	-0.052	-0.012	-0.002	0.001	0.004	0.035	0.154
0.12	-0.297	-0.158	-0.051	-0.008	0.000	0.004	0.012	0.067	0.189
0.14	-0.274	-0.157	-0.056	-0.008	0.000	0.005	0.025	0.092	0.216
0.16	-0.261	-0.156	-0.045	-0.011	-0.003	0.008	0.034	0.107	0.239
0.18	-0.248	-0.168	-0.050	-0.017	-0.002	0.004	0.043	0.121	0.255
0.20	-0.230	-0.169	-0.050	-0.024	0.001	0.009	0.049	0.136	0.267
Number of Missing									
0.04	1.168	0.727	0.240	0.024	-0.020	0.017	0.190	0.384	0.507
0.06	1.191	0.762	0.224	0.101	0.033	0.037	0.238	0.453	0.628
0.08	1.119	0.700	0.215	0.068	0.040	0.052	0.211	0.406	0.629
0.10	0.991	0.619	0.222	0.065	0.007	0.053	0.155	0.364	0.542
0.12	0.862	0.543	0.173	0.028	-0.001	0.037	0.158	0.306	0.436
0.14	0.737	0.476	0.149	0.008	0.018	0.036	0.132	0.286	0.385
0.16	0.627	0.437	0.125	0.011	0.004	0.038	0.122	0.277	0.343
0.18	0.564	0.393	0.119	0.022	0.014	0.060	0.103	0.243	0.316
0.20	0.497	0.341	0.119	0.035	0.025	0.044	0.104	0.215	0.284

Table 4: Absolute Relative Bias comparison with 2 replicates per unknown sample with no LODi restrictions. The bottom half represents the average missing observations due to exclusion for the five parameter model minus the average missing observations for the four parameter model. A higher score represents more missing cases for the five parameter model.

Table 4 is the same as Table 3, except it presents the case when we have 2 replicates averaged instead of a single replicate. Strangely, the model extreme cases were barely affected. While the peculiar trends for $f < 1$ have become more stable (five parameter loses strength generally for increasing σ instead of sporadically, little spikes still) and appear to approach the general trends as we anticipated. The positive asymmetry region is more complicated. The strange feature is the growing parabolic shaped region which favors the five parameter assumption within the positive asymmetric defined region. Extra simulations, with replicate counts of 3 (Table 6) and 20 (Table 7), continued to expand the five parameter parabolic region. In theory, if we had an infinite number of replicates, the average of the replicates responses would yield the truth and thus

achieve the desired truth we had observed in the prior simulations. Based on what we are seeing, it makes some sense that the parabolic region will continue to stretch ever onward in such a way that we would eventually see the truth. However, even with the additional replicates, the haphazard pattern for the negative asymmetric region did not disappear entirely, even with the 20 replicate case. In the end, it appears the parabolic shape will continue to grow for finite samples while the negative asymptotic range would eventually stabilize with respect to σ for only a portion of $f < 1$ cases. At this point, the five parameter seems to favor the $f \leq 1$ region while the four parameter favors $f \succeq 1$.

$\sigma \backslash f$	0.10	0.20	0.40	0.60	0.80	1.00	1.60	3.20	12.80
0.04	-0.184	-0.178	-0.133	-0.101	-0.046	-0.054	-0.112	-0.110	-0.221
0.06	-0.188	-0.150	-0.110	-0.141	-0.127	-0.122	-0.131	-0.130	-0.225
0.08	-0.157	-0.126	-0.121	-0.172	-0.198	-0.196	-0.161	-0.168	-0.233
0.10	-0.136	-0.125	-0.182	-0.195	-0.215	-0.234	-0.182	-0.209	-0.223
0.12	-0.141	-0.137	-0.206	-0.201	-0.229	-0.232	-0.232	-0.229	-0.223
0.14	-0.126	-0.163	-0.240	-0.210	-0.249	-0.246	-0.237	-0.258	-0.224
0.16	-0.122	-0.197	-0.252	-0.230	-0.237	-0.261	-0.259	-0.315	-0.260
0.18	-0.131	-0.215	-0.248	-0.260	-0.249	-0.262	-0.269	-0.309	-0.270
0.20	-0.138	-0.221	-0.257	-0.295	-0.260	-0.269	-0.294	-0.321	-0.297
Five Parameter Above, Four Parameter Below: Missing 20 Rep vs 2 Rep.									
0.04	-0.003	-0.004	-0.019	-0.049	0.018	-0.037	-0.019	0.027	0.004
0.06	-0.061	-0.041	-0.084	-0.034	-0.012	-0.090	-0.100	-0.009	-0.010
0.08	-0.166	-0.135	-0.135	-0.144	-0.128	-0.169	-0.176	-0.157	-0.090
0.10	-0.296	-0.232	-0.179	-0.174	-0.216	-0.238	-0.269	-0.247	-0.234
0.12	-0.391	-0.315	-0.258	-0.216	-0.240	-0.267	-0.289	-0.316	-0.361
0.14	-0.458	-0.363	-0.275	-0.216	-0.251	-0.269	-0.312	-0.353	-0.404
0.16	-0.476	-0.341	-0.292	-0.236	-0.261	-0.266	-0.318	-0.383	-0.418
0.18	-0.475	-0.351	-0.287	-0.252	-0.241	-0.242	-0.325	-0.399	-0.414
0.20	-0.480	-0.378	-0.283	-0.263	-0.254	-0.268	-0.320	-0.405	-0.425

Table 5: The top table represents the difference in missing cases between the 20 replicate case and the 2 replicate case of the five parameter assumption. The bottom table represents the difference in missing cases between the 20 replicate case and the 2 replicate case of the four parameter assumption.

Looking at the number of missing differences for the 2,3, and 20 replicate cases, we only see a very subtle change in the difference of number of missing cases. It appears that with more replicates, the difference in missing between the four and five parameter assumptions does not appear to change that much with regard to the heavy asymmetric regions. On closer inspection as see in 5, the additional replicates are reducing the number of missing almost uniformly with respect to the five and four parameter models individually (not the difference

between four and five parameter models). More replicates appear to help miss less for the five parameter for low σ more and the four parameter more for high σ. The five parameter misses fewer than the four as σ increases, but the five is reports worse overall because it was missing a lot more to begin with at $\sigma = 0.4$. We figured a model assumption may be including/excluding a number of extreme points, since the absolute relative bias results appear to approach the truth in a slow and strange manner. However, the missing cases, if they were somehow included, could work in both ways - if the excluded points were too extreme, they would strengthen the four parameters advantage. If they were normal, they could give the five parameter a stronger advantage. To better understand, it would be easier to look at individual cases to observe the trend physically.

$\sigma \backslash f$	0.10	0.20	0.40	0.60	0.80	1.00	1.60	3.20	12.80
0.04	-0.253	-0.165	-0.082	-0.024	0.001	0.003	-0.053	-0.133	-0.054
0.06	-0.267	-0.159	-0.066	-0.019	0.004	0.003	-0.035	-0.073	0.009
0.08	-0.297	-0.154	-0.055	-0.017	0.000	0.002	-0.020	-0.029	0.060
0.10	-0.309	-0.160	-0.045	-0.014	0.001	0.001	-0.004	0.007	0.104
0.12	-0.307	-0.158	-0.042	-0.014	-0.001	0.004	0.010	0.038	0.141
0.14	-0.286	-0.149	-0.050	-0.013	0.001	0.007	0.019	0.065	0.170
0.16	-0.282	-0.159	-0.058	-0.012	-0.000	0.006	0.029	0.084	0.194
0.18	-0.263	-0.162	-0.056	-0.008	0.000	0.009	0.036	0.100	0.212
0.20	-0.250	-0.165	-0.049	-0.013	0.000	0.008	0.041	0.112	0.228
Number of Missing									
0.04	1.088	0.640	0.190	-0.005	-0.048	-0.017	0.158	0.364	0.490
0.06	1.129	0.700	0.239	0.059	-0.022	-0.005	0.235	0.451	0.605
0.08	1.116	0.693	0.218	0.068	-0.009	0.019	0.226	0.426	0.615
0.10	1.033	0.659	0.189	0.039	-0.021	0.025	0.176	0.400	0.565
0.12	0.910	0.560	0.173	0.025	-0.009	0.018	0.132	0.359	0.495
0.14	0.795	0.494	0.154	0.005	-0.013	0.009	0.120	0.309	0.421
0.16	0.708	0.435	0.125	0.006	-0.017	0.018	0.099	0.283	0.362
0.18	0.624	0.364	0.097	-0.002	-0.017	0.011	0.095	0.234	0.321
0.20	0.542	0.312	0.107	0.007	0.007	0.004	0.103	0.209	0.279

Table 6: Absolute Relative Bias comparison with 3 replicates per unknown sample with no LODi restrictions. The bottom half represents the average missing observations due to exclusion for the five parameter model minus the average missing observations for the four parameter model. A higher score represents more missing cases for the five parameter model.

Therefore we started to look into individual simulation cases and noticed a high number of them favored the four parameter (not simply a few extreme cases favoring the four parameter model heavily). It would seem that for $f = 14.2$, $\sigma = 0.15$, and with 2 replicates, 70.3% of the simulations favored the four parameter

in contrast to the five parameter. We looked into individual simulations where the four parameter was the favored case and noticed a few interesting trends, as seen in Figure 4. The first major trend is that the four parameter was rising sooner than the five parameter fit, allowing the four parameter to better predict the extreme samples that were located on the lower asymptote ranges. Another interesting fact was that, even though the five parameter was closer to the truth, the four parameter was also relatively close to the truth for a majority of the region. The last interesting trend we observed was the crossing behavior of the four parameter over the five parameter. The four parameter fit would cross the five parameter at the asymptotes, allowing the four parameter curve to be closer to the replicates for several extreme concentrations. In a humorous way involving cars, it would be equivalent to somehow swerving at the ideal location to hit two targets on opposite sides of the road. These points, which were also consistent, appeared to make the four parameter model better. It appears that the four parameter model may actually be a better predictor for future samples that are asymmetric than the five parameter model when looking at the whole concentration range.

$\sigma \backslash f$	0.10	0.20	0.40	0.60	0.80	1.00	1.60	3.20	12.80
0.04	-0.261	-0.178	-0.096	-0.039	-0.004	0.005	-0.072	-0.190	-0.106
0.06	-0.258	-0.167	-0.081	-0.025	0.004	0.006	-0.054	-0.139	-0.073
0.08	-0.263	-0.163	-0.069	-0.019	0.006	0.007	-0.041	-0.096	-0.035
0.10	-0.283	-0.159	-0.059	-0.016	0.003	0.005	-0.030	-0.063	-0.001
0.12	-0.297	-0.165	-0.056	-0.015	0.003	0.004	-0.019	-0.036	0.027
0.14	-0.312	-0.161	-0.047	-0.010	0.004	0.005	-0.010	-0.014	0.055
0.16	-0.304	-0.153	-0.048	-0.007	0.006	0.009	-0.000	0.008	0.078
0.18	-0.308	-0.145	-0.046	-0.006	0.010	0.011	0.007	0.025	0.099
0.20	-0.299	-0.142	-0.045	-0.005	0.008	0.013	0.015	0.041	0.119
Number of Missing									
0.04	0.977	0.553	0.126	-0.028	-0.084	0.000	0.097	0.246	0.291
0.06	1.065	0.653	0.198	-0.006	-0.082	0.005	0.207	0.331	0.413
0.08	1.125	0.710	0.229	0.040	-0.030	0.025	0.225	0.399	0.489
0.10	1.148	0.725	0.219	0.044	0.008	0.057	0.242	0.403	0.554
0.12	1.115	0.721	0.224	0.043	0.010	0.072	0.213	0.394	0.584
0.14	1.068	0.677	0.184	0.014	0.020	0.059	0.206	0.378	0.552
0.16	0.980	0.582	0.165	0.017	0.028	0.045	0.182	0.347	0.504
0.18	0.907	0.530	0.158	0.015	0.006	0.039	0.157	0.334	0.460
0.20	0.837	0.496	0.145	0.004	0.018	0.041	0.127	0.302	0.417

Table 7: Absolute Relative Bias comparison with 20 replicates per unknown sample with no LODi restrictions. The bottom half represents the average missing observations due to exclusion for the five parameter model minus the average missing observations for the four parameter model. A higher score represents more missing cases for the five parameter model.

Thus, we thought if we were able to control this issue in a more reasonable manner, we might be able to remove what appeared to be a bias in favor of the four parameter model. It is possible that simply removing extrapolations without a clear criteria may in fact unfairly favor one of the model choices. We then began to look into new procedures involving Limits of Differentiation, sometimes referred to as LODi. This should not be confused with the Limits of Detection, which are ranges of Y which the machines can efficiently measure and report from, with each machine having a different but fixed range. Yet, our prior results could be viewed as a weak application of LoD, where we removed points that exceeded any reasonable LoD range a machine may have. Generally, it is a bad idea to try to estimate a concentration from a "LoD" result and should be dropped instead

As for the LODi, it represents a range of X where we would expect to find meaningful estimates. The range can only be calculated through an extensive process, using a grid search across the entire range of X. Its exclusion criteria revolves around calculating $\widehat{f}^{-1}(y)$ and constructing a confidence interval. If our confidence interval contains either the minimum or maximum concentration (0.5 and $10,000$ respectively), we consider that estimate to lie outside of the LODi range. LODi is thus just a more stable process for acquiring only reasonable concentration estimates for both models (Defawe et al. [2012]). It also addresses our concerns about a few unstable points determining the whole simulation result.

For both the four and five parameter models, our $\widehat{f}^{-1}(y)$ and confidence intervals will be different. Since we are interested in exclusively selecting either the four or five parameter model, we decided to first compare the models by calculating each one's scores with respect to its own unique LODi region. The LODi regions for both the four and five parameter models do not entirely cover each other in any way. If this form of removal does not bring us closer to what we originally anticipated, we would then proceed in using only estimates that were in both LODi regions. For this "shared" LODi region, both the four and five parameter models will have a missing difference of 0.

As long as both parameter models utilized the LODi rules, we assumed it would be a more fair process for both. Table 8 presents the absolute relative bias for the 2 replicate standard case while Table 9 presents the 20 replicate case. However, even this does not assume the pattern we had anticipated. There still exists some aspects of the peculiar trends for increasing σ within the negative asymmetric region ($f = 0.6, 0.8$). It still has the parabolic region for the five parameter model within the positive asymmetric region, just stronger and expanded more than Non-LODi equivalent. With additional replicates (referencing 20), we noted that the odd pattern within the negative asymmetric region was nearly gone. However, the parabolic shape for the five parameter assumption within the positive region still existed. It would seem that our standard

LODi approach simply has the effect equivalent to increasing the replicates for a non-LODi.

$\sigma \backslash f$	0.10	0.20	0.40	0.60	0.80	1.00	1.60	3.20	12.80
0.04	-0.230	-0.164	-0.080	-0.029	-0.003	0.002	-0.042	-0.094	-0.011
0.06	-0.189	-0.142	-0.063	-0.016	0.002	0.003	-0.018	-0.032	0.064
0.08	-0.162	-0.128	-0.054	-0.013	0.002	0.004	-0.006	0.011	0.122
0.10	-0.140	-0.116	-0.045	-0.014	0.001	0.004	0.006	0.046	0.166
0.12	-0.121	-0.106	-0.041	-0.016	0.001	0.004	0.016	0.073	0.205
0.14	-0.107	-0.099	-0.037	-0.014	-0.003	0.008	0.029	0.104	0.240
0.16	-0.097	-0.088	-0.034	-0.012	-0.002	0.011	0.040	0.122	0.262
0.18	-0.091	-0.080	-0.032	-0.013	0.001	0.014	0.049	0.142	0.281
0.20	-0.088	-0.074	-0.031	-0.011	0.003	0.015	0.056	0.154	0.297
Number of Missing									
0.04	0.885	0.740	0.362	0.177	0.050	0.009	0.229	0.402	0.533
0.06	0.544	0.399	0.094	0.001	-0.033	-0.015	0.046	0.124	0.205
0.08	0.319	0.157	-0.008	-0.037	-0.002	-0.007	-0.025	0.009	-0.005
0.10	0.100	0.015	-0.079	-0.052	-0.041	-0.027	-0.013	-0.053	-0.074
0.12	-0.043	-0.075	-0.071	-0.039	-0.035	-0.027	-0.021	-0.033	-0.128
0.14	-0.137	-0.110	-0.063	-0.030	-0.025	-0.044	-0.028	-0.087	-0.143
0.16	-0.190	-0.118	-0.070	-0.033	-0.039	-0.027	-0.031	-0.078	-0.133
0.18	-0.219	-0.131	-0.060	-0.041	-0.023	-0.028	-0.050	-0.068	-0.121
0.20	-0.218	-0.122	-0.055	-0.045	-0.016	-0.026	-0.045	-0.074	-0.117

Table 8: Absolute relative bias comparing the four and five parameter assumptions with 2 replicates and LODi exclusions. The bottom half represents the average missing observations due to exclusion for the five parameter model minus the average missing observations for the four parameter model. A higher score represents more missing cases for the five parameter model.

The second half of Table 8 presents the number of missing observations difference when using the LODi for the 2 replicate case. As we can see, the number of missing observations is no longer lopsided to a single assumption. There is no clear pattern here that can explain the irregularities. The five parameter suffers from additional missing observations for low σ while the four parameter assumption suffers from additional missing observations for moderate levels of σ. We can therefore conclude that simply using the LODi does not yield the pattern we had previously noticed during the truth simulations.

$\sigma \backslash f$	0.10	0.20	0.40	0.60	0.80	1.00	1.60	3.20	12.80
0.04	-0.271	-0.205	-0.117	-0.048	-0.007	0.004	-0.078	-0.208	-0.132
0.06	-0.268	-0.192	-0.095	-0.038	-0.003	0.003	-0.054	-0.156	-0.096
0.08	-0.261	-0.175	-0.078	-0.026	0.001	0.006	-0.034	-0.108	-0.054
0.10	-0.241	-0.160	-0.067	-0.018	0.005	0.009	-0.019	-0.074	-0.014
0.12	-0.220	-0.147	-0.062	-0.012	0.007	0.010	-0.008	-0.041	0.020
0.14	-0.201	-0.136	-0.056	-0.010	0.009	0.012	0.001	-0.014	0.050
0.16	-0.183	-0.132	-0.050	-0.008	0.009	0.013	0.009	0.008	0.082
0.18	-0.164	-0.124	-0.046	-0.006	0.008	0.016	0.014	0.026	0.112
0.20	-0.151	-0.121	-0.040	-0.004	0.008	0.018	0.022	0.040	0.135
Number of Missing									
0.04	1.541	1.293	0.845	0.396	0.089	0.001	0.521	0.985	0.862
0.06	1.572	1.312	0.743	0.347	0.076	0.046	0.402	0.849	0.956
0.08	1.456	1.141	0.521	0.194	0.036	0.025	0.252	0.610	0.829
0.10	1.250	0.916	0.362	0.129	0.005	-0.005	0.128	0.425	0.638
0.12	1.047	0.713	0.270	0.058	-0.020	-0.007	0.081	0.304	0.446
0.14	0.858	0.539	0.204	0.028	-0.012	-0.038	0.063	0.219	0.343
0.16	0.691	0.416	0.128	0.023	0.000	-0.010	0.051	0.173	0.241
0.18	0.522	0.333	0.114	0.010	0.000	-0.012	0.043	0.133	0.165
0.20	0.392	0.257	0.091	0.000	-0.011	-0.031	0.008	0.089	0.118

Table 9: Absolute relative bias comparing the four and five parameter assumptions with 20 replicates and LODi exclusions. The bottom half represents the average missing observations due to exclusion for the five parameter model minus the average missing observations for the four parameter model. A higher score represents more missing cases for the five parameter model.

Up until this point, all of our changes have yielded only slight steps in the direction we wanted, but with no strong indications of truly approaching the truth. The four parameter model still returns a better result for the absolute relative bias for peculiar regions. An increase in the number of replicates were only partially successful in approaching the theoretical result. While the large simulations showed a slight shift toward our truth, these simulations were also impractical cases of actual bioassays. We are limited to only 96 samples per experiment. We believed there may still exists some form of bias that is returning these odd results.

We therefore began to think that a common LODi region might be the proper approach - using only the concentrations that can be accurately inverse predicted for both assumptions. For our next simulation step looking into the absolute relative bias issue, we decided to use a stricter exclusion criteria for evaluating validation relative bias. This test excluded any points which were outside of either the four parameter's or five parameter's limits of detection range. We still excluded for both tests any points which were outside of the range of the curve for either curve. In this respect, we are strictly dealing with

points that were within all acceptable ranges for both model assumptions.

$\sigma \backslash f$	0.10	0.20	0.40	0.60	0.80	1.00	1.60	3.20	12.80
0.04	-0.178	-0.130	-0.068	-0.026	-0.004	0.002	-0.014	-0.010	0.111
0.06	-0.150	-0.120	-0.054	-0.021	-0.003	0.002	0.001	0.035	0.168
0.08	-0.130	-0.113	-0.045	-0.017	-0.003	0.003	0.014	0.068	0.210
0.10	-0.116	-0.105	-0.040	-0.016	-0.003	0.004	0.024	0.095	0.241
0.12	-0.104	-0.096	-0.037	-0.015	-0.003	0.006	0.033	0.116	0.266
0.14	-0.097	-0.088	-0.035	-0.015	-0.003	0.008	0.042	0.134	0.283
0.16	-0.092	-0.081	-0.035	-0.015	-0.003	0.010	0.050	0.146	0.295
0.18	-0.089	-0.076	-0.035	-0.016	-0.002	0.011	0.056	0.155	0.302
0.20	-0.088	-0.071	-0.035	-0.016	-0.001	0.013	0.061	0.161	0.308

Table 10: Comparison of the five and four parameter absolute relative bias using 2 replicates with only a common LODi region

Unfortunately, as we expected Table 10 shows that this new common LODi region is actually worse than the original LODi plan when addressing the parabolic problem. It did, however, remove the odd patterns for $f < 1$ when increasing σ. Even with additional replicates the pattern remained the same. Thus, we can conclude that the four parameter assumption, for realistic sample replicates sizes of 2, is the best choice for a realistic scenario. Even though it is not the truth, it is the only one returning realistic relative estimated concentrations. As we mentioned before, it was entirely possible that an incorrect model may actually be better in predicting future outcomes over the actual true model.

As for our tests involving S^1, we acquired the same pattern observed during our direct simulations for S^1, as seen in Table 11. Changing the number of replicates to either 1 or 3 had little to no impact upon the overall pattern. This seems to make our prior results with the absolute relative bias a little confusing. If we take the Global MSE and Absolute relative bias measures together in their current state, it seems to be implying that the variance may be impacted by high levels of f for some odd reason. Unlike the tests for the absolute relative bias measures, we have no missing observations for this test as it is not susceptible to undefined or infinite values.

$\sigma \backslash f$	0.10	0.20	0.40	0.60	0.80	1.00	1.60	3.20	12.80
0.04	-0.018	-0.011	-0.004	-0.001	-0.000	0.000	-0.001	-0.003	-0.006
0.06	-0.018	-0.011	-0.004	-0.001	-0.000	0.000	-0.001	-0.003	-0.006
0.08	-0.018	-0.011	-0.004	-0.001	0.000	0.000	-0.000	-0.003	-0.006
0.10	-0.018	-0.011	-0.004	-0.001	0.000	0.000	-0.000	-0.003	-0.006
0.12	-0.018	-0.011	-0.003	-0.001	0.000	0.001	-0.000	-0.003	-0.006
0.14	-0.018	-0.011	-0.003	-0.000	0.001	0.001	0.000	-0.003	-0.006
0.16	-0.017	-0.011	-0.003	-0.000	0.001	0.001	0.000	-0.003	-0.006
0.18	-0.017	-0.010	-0.003	0.000	0.001	0.001	0.000	-0.002	-0.006
0.20	-0.017	-0.010	-0.002	0.001	0.002	0.002	0.001	-0.002	-0.006

Table 11: Testing Simulations for the S1 Criterion, utilizing 2 replicates and no exclusion criteria.

2.7 Discussion

Our entire simulation process was done in a step by step process. While there were very few difficulties with our criterion evaluation simulation phase, the same could not be said about the testing phase. All the difficulties stemmed from issues revolving around undefined regions and infinite concentration values when dealing with the relative bias. For the direct evaluation, the solution was straightforward, since it basically involved integration with regard to regions, with substitutions for segments where it was undefined. Here, substitutions were acceptable since we were integrating over a whole region and were not interested in the extrapolation into unreasonable regions, hence the reason for substitution.

However, for the testing phase, we could not insert any substitutions without ruining the interpretation. Some substitutions could have actually returned a relative bias of 0, thereby making poor estimates into ideal ones. Thus we began to look into other exclusion criteria alongside increasing the number of replicates per unknown sample. If we increased the number of replicates into the 1000's and towards infinity, we would expect the simulations to yield similar results as the direct simulations. However, this was not the case. We actually observed two regions which favored the four parameter model - the first consisting of the region around $f = 1$, and for the region around $f > 2$. We even treated each replicate as its own unique observation at one time, and only observed a stronger tendency for the four parameter's strange advantage. We began to believe that our exclusion criteria was favoring one of the models, that our exclusion criteria was faulty and biased. We looked into the limits of detection which utilized the confidence intervals for $\widehat{f}^{-1}(\overline{y_i})$. We theorized that any estimates $\widehat{f}^{-1}(\overline{y_i})$ whose confidence intervals included our minimum and maximum concentration bounds were inaccurate predictions (claiming the concentration lies between a negative and positive concentration is absurd). Thus we decided to include only $\widehat{f}^{-1}(\overline{y_i})$ that fell into this range of x via grid searching. However, the LODi exclusions did not return the regions we had anticipated. Instead, they appeared to act as

if we increased the number of replicates.

It would appear that the four parameter model, while weaker when dealing with heavily asymmetric data over realistic ranges of X, actually has a unique advantage in the positive asymmetric region. It seems the four parameter is better in reducing the absolute relative bias. This can be explained by looking at 4, where the four parameter fit rises much sooner and crosses the five parameter fit at ideal locations, thereby greatly reducing the absolute bias for the samples that reside on the asymptotes ranges. The relative bias for these fringe concentration values is a lot worse under the five parameter model, even when it appears to fit closer to the truth. In addition, the variance component of the MSE may in fact depend upon f in some strange way, which would explain the four parameter's growing advantage for increasing f. In the end, even though there are a large number of possible explanations, the four parameter assumption, while not true, is actually a good predictor for future samples drawn from the same population, possibly due to the asymptotic regions.

There are some possibilities that could explain this pattern which we did not cover. There is a possibility the four parameter's advantage, for positive asymmetric data, may actually be linked to g and h, the inflection point and its slope. We already know the design has a major impact upon inverse prediction for specific regions. For example, if we allocate all of our sample in the "middle", we would be able to acquire accurate inverse predictions for that region but poor inverse predictions for concentrations lying along the asymptotic ranges. Similarly, if we allocated all our samples to the "outside", we could acquire good predictions for the outer ranges. With that in mind, the g and h parameters have a large impact on the shape and the amount of the curve that is allocated to the inner and asymptote regions.

We have also made several assumptions when dealing with bioassays. It is entirely plausible the variance should actually depend upon X. For example, the noise present in the machine response can be negligible/stable for very small or large values of X. It is also possible that for some assays, the response will never pass a certain value, such as a negative response value. Our simulations never took into consideration any of these restrictions and may in turn generated some unreasonable values.

One final issue that we ran into was a slight discrepancy involving the fitting program. While the four parameter option always returned the same parameter estimates with the same data, it wasn't for the five parameter. It appears that permutations of a dataset will return slightly different estimates when using a five parameter fit. It was also noted that one alternative dose response curve fitting program returned warning errors about unrealistic fitting results.

What we can conclude thus far is that the four parameter model appears to be ideal for slightly asymmetric data ($0.7 < \sigma < 1.5$ is likely safe). For positive

strongly asymmetric data, the four parameter model actually performs better than the five parameter, probably due to the concentrations lying alongside the asymptotes. Since we are limited by our sample sizes, it may be better if we refer to the LODi results and consider only the estimated concentrations that are reasonable, but even the four parameter will outperform. Other factors, such as the location of the inflection point and its slope, may have a bigger impact than we originally thought. In the end,we only know that we are clearly limited by the sample size. If our designs were not strictly limited to sample sizes of 20, we would likely support the five parameter model unanimously, but given the limited size and the four parameters strange performance advantages, it may be better to assume the four parameter if there was any suspicion of positive asymmetry, even if the four parameter is clearly not the correct model.

3 Cross Validation Study

3.1 Cross Validation Basics

Cross validation tests were conducted in a similar way as our previous simulation testing segment. The noticeable difference between these two testing methods is that one utilizes simulations while the other uses actual assay data. The simulation approach tests whether the fitted curve was a good model for future unknown samples (mimicking samples drawn from the same population). The cross validation instead takes a whole dataset, which was originally used to fit a model, and breaks it into two smaller datasets. These partitions are called the training set and the validation set, where the training is used for fitting the model while the validation is treated as our unknown sample. There are also a number of types of cross validation. Some run through all of the iterations of interest, while others may randomly sample. Our cross validation of interest, which is also known as "Leave-One-Out" validation, breaks samples of 20 into samples of 19 and 1 respectively, and then goes through all iterations (resulting in 20 calculations per dataset). In the end, cross validation helped us evaluate the accuracy of the measure by using a realistic dataset whereas previously we had only been able to emulate realism through simulations.

There are a number of advantages Cross Validation has over our prior simulation process. Since we are using realistic data, the data is no longer generated from our standard five parameter function, which in turn may have favored the five parameter strongly (recall back to the perfect model problem, where the perfect would have no bias and the lowest variance). By using this data, we also avoided the possibility of simulating unrealistic data. It was possible, even though we used estimates from a prior assay fit, that when we varied f while holding all of the other parameters fixed, we were in turn generating unrealistic data. Lastly, this cross validation test will evaluate the four and five parameter models in general, without any specific fixed parameters. There are only two weaknesses in contrast to our simulation tests. The first involves the fact we can not accurately observe the effects of increasing replicates for each unknown sample. The second is that we do not have enough data to observe both the effects of σ and f, and since the peculiar pattern was mainly with increasing f, we chose to observe f. Our goal in the end is twofold, to confirm whether the odd patterns observed previously in simulations actually occur in reality and to evaluate the relative bias and Global-MSE criterions.

3.2 Dataset Origin and Breakdown

Our realistic dataset comes from a recent HIV vaccine trial, aimed at observing patient responses to the vaccines. Patients who were enrolled in the trial were administered the vaccine, and after a short period of time, had their blood

drawn. These blood samples would contain concentrations of cytokine, proteins the body makes in response to the vaccines. Our assay's goal was to accurately inverse predict the concentration of cytokine within these blood samples. The samples we had on hand were only the standard samples, the training data with the expected concentrations. The samples were artificially made, not drawn from a patient, and were allocated to wells on the plate using 3 fold dilutions, with 2 replicates per dilution. All machine measured responses were collected using a luminescent test.

Our cross validation tests used 754 separate datasets, each consisting of exactly 20 observations where each concentration had 2 replicates. These datasets were acquired over 26 separate assays with regard to 62 analytes of interest. For each assay, not every analyte was tested, so some assays have more analytes than others. Each combination of assay and analyte constituted a single dataset.

As for our cross validation of choice, we decided to go with a simple yet exhaustive test. We split the 20 sample size into a sample 19 and a sample of 1, using the 19 to fit the curve and the 1 for inverse predicting. We then repeated this for each individual point once, so that each point could act as the unknown concentration once.

3.3 Cross Validation Test Process

We performed the Cross Validation test with respect to the S^1 and S^2 criterions in the same way we performed the testing simulations. We had 754 individual datasets of 20. For the S^1, validation test, we calculated the average of:

$$S_j^1 = (\widehat{f_j^*}(t_j) - y_j)^2 \tag{8}$$

where y_j was the individual point response, t_j was the expected log concentration for the individual point, and $\widehat{f_j^*}$ was the fitted function without using j^{th} observation for fitting. In other words, we fit a function using all the points except for the j^{th} point, and then feed that point's values to the function and compare it..

For the S^2, which is our term for the absolute relative bias, we calculated the average of:

$$S_j^2 = \frac{|e^{\widehat{f_j^*}^{-1}(y_j)} - x_j|}{x_j} \tag{9}$$

first with the four parameter model and then with the five parameter model, where $x_j$was the expected concentration for the individually removed point and

$\widehat{f_j^*}^{-1}(y_j)$ was the fitted curve's x without the j^{th} point for the specific y_j. Similar to the two prior issues with the relative bias, we had to drop any S_j^2 where $y_j < min(\widehat{f_j^*}(x))$ or $y_j > max(\widehat{f^j}^*(x))$ for our specified range of x to avoid meaningless extrapolation and especially cases when $\widehat{f_j^*}^{-1}(y_j) = \infty$. For each dataset, we only acquire a single average for the five and four parameter models. We do not acquire multiple averages from multiple simulations or repeated samples as we did previously. We are limited to one outcome for each model for each distinct experiment. To best understand how the varying levels of f may affect our model assumption averages, we also acquired an estimate of f by fitting the five parameter model with the full 20 samples for each datasets. We then took the difference between the four and five parameter cross validation averages for each dataset.

Looking at the various estimates of f, we found that a third of the datasets have $f < 0.70$ and 90% of all datasets appear to have $f < 3.462$. There are only a very few cases that have $f > 7.51$ (99% percentile). This would imply that the odd behavior we had previously observed may be unique to unrealistic cases. However, as we shall see that is not entirely true.

Looking at the box plots for the averaged S^1 validation comparisons, we again noticed the same trend we had seen back at Table 1, where the five parameter was better for asymmetric cases and the four parameter was better for the symmetric cases. In Figure 5, we observed a parabolic relationship pivoting around $f = 1$ when looking at the S^1 criterion results, as we anticipated when σ was assumed to be somewhat stable.

As seen in Figure 6, The relative bias cross validation tests appear to support our prior odd findings. As before, it appears the five parameter gains little to no advantage for increasing positive asymmetry ($f > 1$). However, the pattern is not as extreme as the simulations because we do not have the necessary amount of datasets that stretch into the $f > 7.51$ region. However, what is interesting is that the four parameter appears to be equal or have a better advantage for samples which reported $f > 0.629$. The Five parameter's advantage appears to be slightly weaker for the negative asymmetric region. In addition, the five parameter assumption may only have a strong advantage for cases when $f < 0.629$.

When we use the LODi restrictions as in Figure 7, the five parameter immediately loses its advantage for all $f > 1$ ranges of asymmetry. When we limit

ourselves to only estimates which fall within the LODi region, the four parameter is exceptional. However, this sudden shift was unexpected in contrast to the prior simulations. Previously, we had anticipated the five parameter to gain a slightly stronger advantage when implementing the LODi, expanding the five parameter favored "parabolic" within the positive asymmetry region. Instead, the five parameter's advantage was completely negated by limiting ourselves to the LODi range. In the end, it appears the results from the cross validation phase are very similar to the results in the testing simulation phase.

3.4 Discussion

There are only a few problems with the cross validation that should be addressed. The biggest concern is that we were restricted only to the datasets we had on hand, which did not thoroughly cover all the ranges of f we had looked at during our simulation phase. While we did observe an increasing advantage for the four parameter for increasing f, we couldn't tell for certain due to the limited sample on hand. Only a very small portion was close $f = 12.8$ range we used during the simulation phase. Thus, we urge some caution still when claiming the four parameter as superior to the five parameter for most asymmetric regions.

This confirmation brings us back to our previous problem, with the odd trend still existing. It may be caused by removing all extrapolated estimates. However, with the LODi restriction it appears the relative bias is increasing with respect to f. The unexpected changes we observed, where the five parameter model no longer has any advantage when using our LODi restriction may be due to the fact that the five parameter model is not the actual truth anymore. It is possible that the truth follows some other functional form. "Additional" replicates, as seen during the simulations, were basically absent and if included would likely help us approach the truth to a degree. The LODi restrictions, as seen during the simulations, should have the same effect in bringing us closer to the truth (in this case, away from the five parameter). Following that logic, we may be shifting closer toward the truth and away from the log logistic formula we have assumed the bioassays always followed.

Since we were using actual data this time around, we were able to specifically observe the model score difference trends with respect to f. We had previously thought that the other parameters may have had a stronger impact than anticipated. With the cross validation results being similar to the simulations', we can conclude that the parameter values we had selected for c, d, g, h, e, and b had little to no impact upon these score differences.

In the end, our conclusions have not strongly changed from our prior conclusions involving the relative bias from the testing simulations. We have only

really eliminated some of the false possibilities, such as our parameter values for the simulations having a larger impact, that may have explained the odd pattern. We are still observing a four parameter advantage in positive asymmetric data that we can not seem to fully explain.

4 Conclusion

4.1 Overall

Both our testing simulation and cross validation returned similar results, reinforcing the idea that the four parameter model may in fact be better at predicting actual concentrations. We know that the four parameter model is not the true model, and the direct simulation results revealed that the five parameter model fits closer to the true curve than the four parameter does when dealing with asymmetric data. We also noticed that the mean S^1 score pattern never changed across all three tests. For every test dealing with asymmetric data, the five parameter performed better with respect to the mean S^1scores, but poorly with respect to the absolute relative bias. In the end, it may be that this trend is due to the fitting process, where we saw during the cross validation that there were only a few cases where f was above 7. Looking into some side simulations, it appears that regardless of how high f is set, f is grossly underestimated once it is above 6. as some extra tests. It is entirely possible that the asymmetry present in the data can be explained by e and b once past a certain threshold for f. Regardless, we can only make conjectures as to why the four parameter performs better for highly positive asymmetry. In short, it would appear that the four parameter is the best model in practice, regardless of intuition.

4.2 Last Thoughts

The root of problem I believe is still a missing data problem despite all the approaches we took. We even looked into the results for allowing "somewhat-reasonable" extrapolation (No infinite cases) and substitutions under various conditions, yet most of these simulations returned highly favored four parameter results for almost all f and σ, with no general pattern. The results we reported in this thesis all shared a part of the general pattern we had anticipated when looking into the bias-variance trade-off relationship, thus signifying that we have found something of interest, just that we can not fully explain it. It does feel a bit odd having the four parameter provide a better prediction when working with asymmetric data.

Overall, our original goal was to evaluate all aspects of bioassays that could improve both the four and five parameter fitting assumptions. This included the concept of alternate design allocations outside of the standard 2 replicates per dilution, as mentioned in François et al. [2004] for the four parameter model. We began to look into the accuracy differences between the four and five parameter models and we noticed the odd trend with respect to the absolute relative bias for testing data. After a few weeks of looking into it, we concluded that we should look further into the four and five parameter comparison, placing the

design choice considerations on hold, and eventually dropping it. However, there may exist alternative designs that can improve the four or five parameter fits. It is even possible that our design may in fact be favoring one of the model assumptions. Yet, this possibility was never a major concern because our study used the standard design a lot of labs use. The findings contained within this Thesis are applicable to a large number of labs out there already. Yet, I encourage anyone who is interested in these findings to look into the effects of alternative design allocations and overall sizes as another factor.

Note: There are some cases where we say $S1$ *instead of* S^1*. Both are the same. This is merely a limit in the writing format.*

5 Bibliography

References

O. Defawe, Y. Fong, E. Vasilyeva, M. Pickett, D. Carter, E. Gabriel, S. Rerks-Ngarm, S. Nitayaphan, N. Frahm, M. McElrath, et al. Optimization and qualification of a multiplex bead array to assess cytokine and chemokine production by vaccine-specific cells. *Journal of Immunological Methods*, 2012.

Y. Fong. Methods for calibration and normalization of immunoassay data using r. 2012. URL `http://labs.fhcrc.org/fong/Ruminex/index.html`.

Y. Fong, J. Wakefield, S. De Rosa, and N. Frahm. A robust bayesian random effects model for nonlinear calibration problems. *Biometrics*, 2012.

N. François, B. Govaerts, and B. Boulanger. Optimal designs for inverse prediction in univariate nonlinear calibration models. *Chemometrics and intelligent laboratory systems*, 74(2):283–292, 2004.

P. Gottschalk and J. Dunn. The five-parameter logistic: a characterization and comparison with the four-parameter logistic. *Analytical biochemistry*, 343(1): 54–65, 2005.

C. Ritz and J. C. Streibig. Bioassay analysis using r. *Journal of Statistical Software*, 12, 2005. URL `http://www.bioassay.dk`.

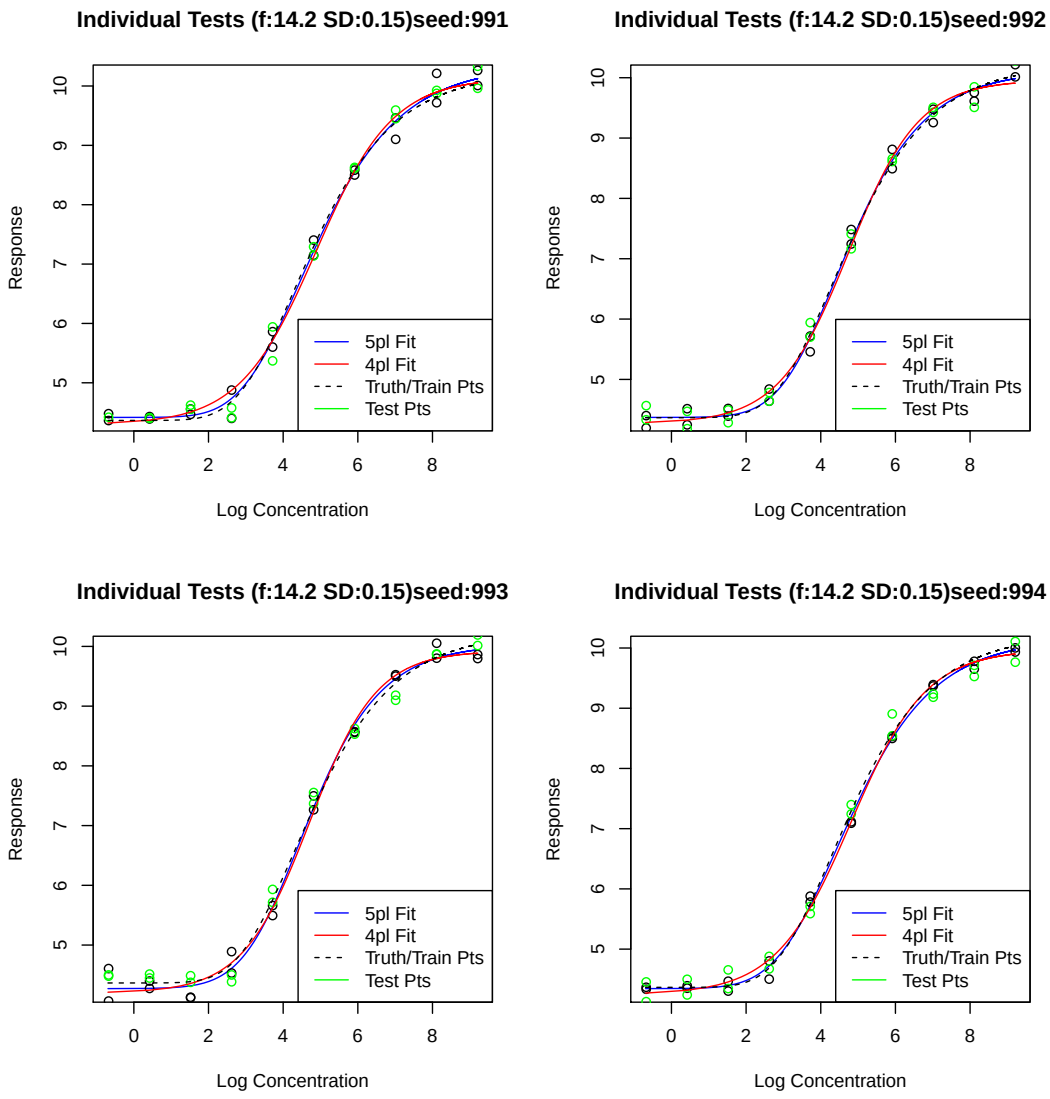

Figure 4: Four single simulations which favored the four parameter model over the five parameter model for the S2 criterion. All four were found to favor the four parameter, whether we averaged the replicates with respect to each concentration or whether we treated them as individual samples.

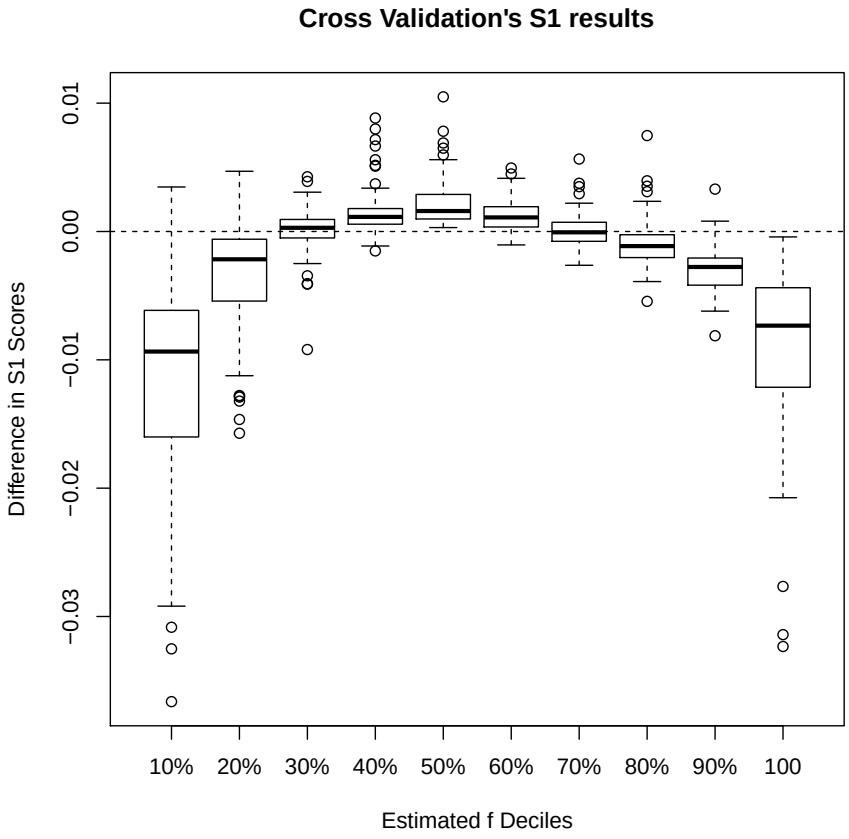

Figure 5: Box plots for the difference between the four and five parameter assumptions for the S1 Scores, grouped by relative f estimates. The estimated region boundaries are (*,0.235, 0.380, 0.629, 0.947, 1.258, 1.807, 2.257, 2.776, 3.462,*)

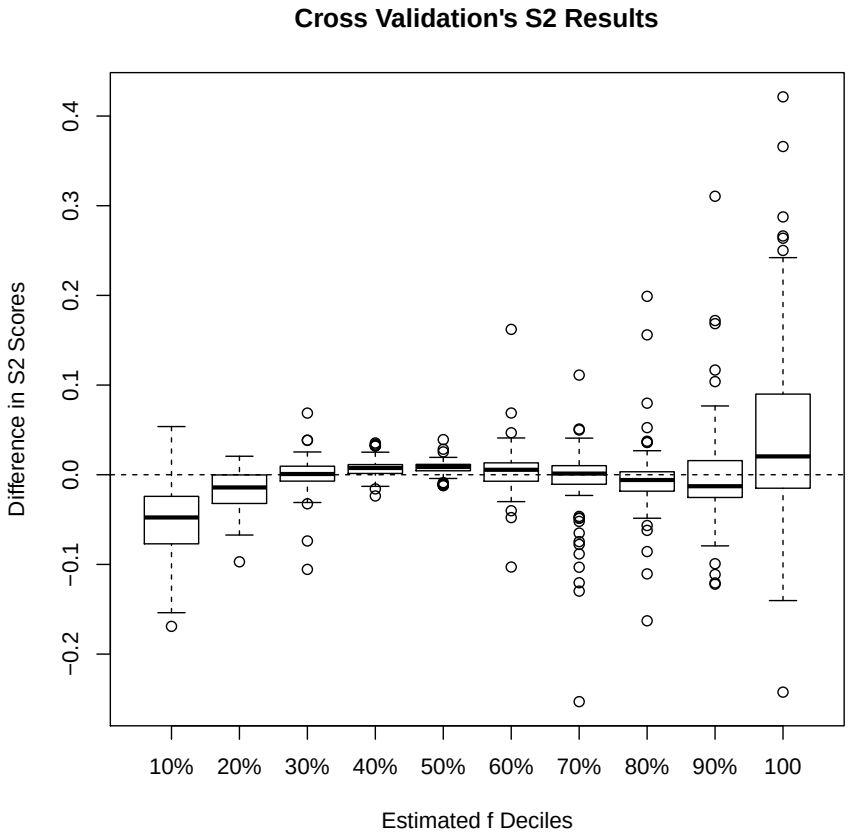

Figure 6: Box plots for the difference between the four and five parameter assumptions for the S2 scores, grouped by relative f estimates.The estimated region boundaries are (*,0.235, 0.380, 0.629, 0.947, 1.258, 1.807, 2.257, 2.776, 3.462,*)

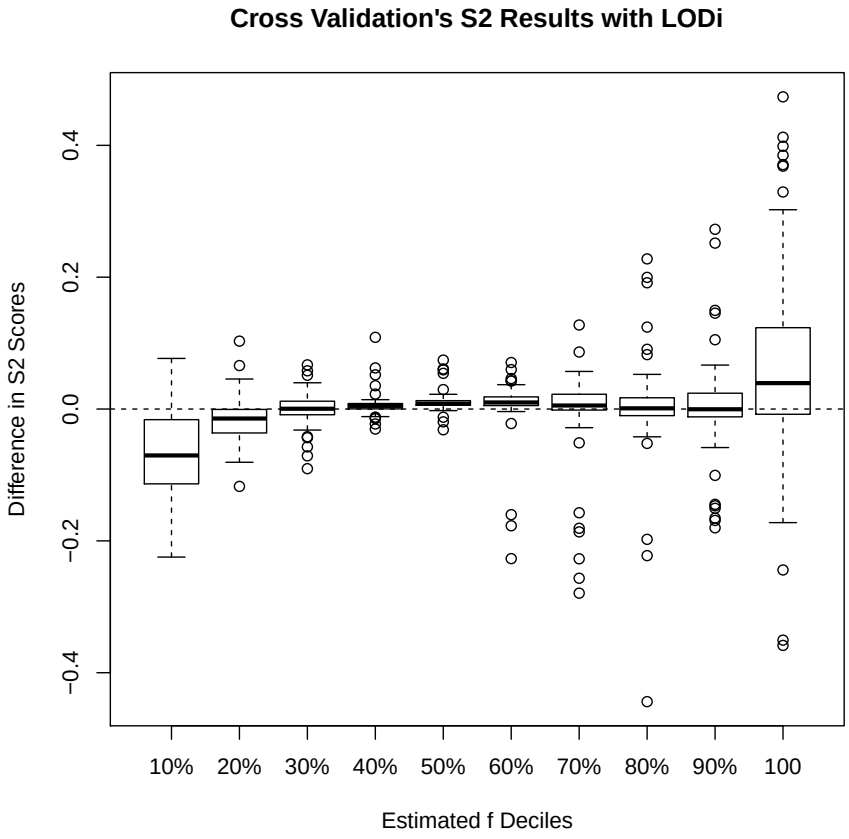

Figure 7: Box plots for the difference between the four and five parameter assumptions for the S2 Scores with LODi restrictions, grouped by relative f estimates. The estimated region boundaries are (*,0.235, 0.380, 0.629, 0.947, 1.258, 1.807, 2.257, 2.776, 3.462,*).