Quality and
Reliability
Engineering
International

(wileyonlinelibrary.com) DOI: 10.1002/qre.1912

Published online in Wiley Online Library

Incorporating Confidence Intervals on the Decision Threshold in Logistic Regression

Michael J. Kist and Rachel T. Silvestrini*†

This paper discusses an application of confidence intervals to the threshold decision value used in logistic regres]sion and discusses its effect on changing the quantification of false positive and false negative errors. In doing this a grey area, in which observations are not classified as success (1) or failure (0), but rather 'uncertain' is developed. The size of this grey area is related to the level of confidence chosen to create the interval around the threshold as well as the quality of logistic regression model fit. This method shows that potential errors may be mitigated. Monte Carlo simulation and an experimental design approach are used to study the relationship between a number of responses relating to classification of observations and the following factors: threshold level, confidence level, noise in the data, and number of observations collected. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: generalized linear model; predictive analytics; optimal design; Pareto frontier

1. Introduction

Predictive analytics provides several valuable models that can and should be used in applications across many fields of work. Linear and generalized linear regression models (such as logistic regression) are among the most widely used analytical models in practice. These models can be used to provide a better understanding of what can be expected from a specific test or event and therefore allow the user to make decisions based on that information. In addition to a decision aid, prediction models can greatly benefit users in terms of efficiency, such as saving time, money, and other valued resources. Medical and military applications of predictive analytics are essential for success in those domains, yet there is an ethical dilemma that follows these applications. If the analysis is poorly conducted or inaccurate, the resulting loss could be at the cost of someone's life.

In the medical and defense industries, many outcomes of tests, treatments, and missions can be summarized with binary data. The logistic regression model, using the logit or probit link function, is a commonly used mathematical model for data that is binary or dichotomous in nature. A nonlinear fit is provided for a set of known data that best describes the relationship between the response probability and one or more independent variables; this is outlined by Hosmer et al[1]. The results of the logistic regression will provide a probability that an event will occur, given a choice of settings for the independent or predictor variables. An example output may be the probability of a drug to take effect or the probability of a bullet to penetrate a plate of armor. The ability to accurately predict an outcome is invaluable.

In medicine an effective treatment could determine whether or not a patient survives an illness. Criteria such as level of risk of a medical procedure or effective dosage of medicine can be made based off of previously collected and analyzed data. For example, Sun et al.[2] observed the mortality rates of patients suffering from liver disease and applied a logistic regression. They found the 3-month mortality rate based on various factors. Initially, they utilized a scoring system called the model for end-stage liver disease (MELD) to predict mortality rates. After applying the same problems to a logistic regression they found it to be more accurate than the MELD scoring system at predictive analysis. In another medical example, DCE-MRI testing is conducted for the detection and diagnosis of breast lesions and the conclusion can be one of two outcomes, benign, or malignant. One of the methods used in an attempt of accurate predictive analysis is logistic regression. This approach is applied by Mclaren et al.[3]. In particular a comparison is being drawn between the logistic regression approach and the use of artificial neural networks (ANN). Simulations were conducted using both logistic regression and ANN to determine which method was more accurate with predictions. As a result of the testing the logistic regression model produced correct predictions in 75% of tests, and the ANN method produced correct predictions in 76% of tests. The conclusions drawn by Mclaren et al.[3] were that the ANN and logistic regression models were roughly similar in regards to predictive capability.

Industrial and Systems Engineering Department, Rochester Institute of Technology, Rochester, NY, USA
**Correspondence to: Rachel T. Silvestrini, Industrial and Systems Engineering Department, Rochester Institute of Technology, Rochester, NY, USA.*
†E-mail: rtseie@rit.edu

Similar to the seriousness of medical application, in the US Military, expensive equipment is constantly in use to either protect soldiers or confront enemy forces. If a piece of equipment malfunctions, the result could be an injured soldier or a failed mission as well as added expenses. If outcomes can be more accurately predicted, these kinds of failure can be anticipated and avoided, ensuring greater success. When the stakes are this high, the information must be presented with a level of certainty or confidence in the outcome to allow for appropriate precaution.

In statistics, a tool frequently used to report precision in a measurement is a confidence interval. The application of confidence intervals in predictive analytics can help the user establish specific levels of certainty. Confidence intervals produce a range of data surrounding the expected parameter. The true value of the parameter is expected to exist within this range with a specific degree of certainty. The specifics of confidence intervals are outlined in Sullivan[4].

Application of confidence intervals in predictive analytics, such as logistic regression, is quite beneficial. Decisions regarding the binary response based on a logistic regression model require the specification of a threshold value. When an input or predictor variable is specified, the logistic regression equation produces a probability of an event occurring under those conditions. Based on the probability and prespecified threshold a 1 or 0 will be predicted. For example, when testing bullet penetration at varying bullet velocity with a threshold at .5, if the regression model produces a .55 probability of penetration, the predicted outcome would be a 1 (a penetration) as it exceeds the threshold.

The threshold value is used to determine whether an event is expected to be a success or failure. If the prediction does not match the expected result, it is classified as one of two types of error. If an input is expected to produce a 1 and its respective prediction falls below the threshold, the result is a false positive. If an input is expected to produce a 0 and its prediction is above the threshold, the result is a false negative. False positive and false negative errors can be both fatal and very expensive in medical and military applications. In a standard logistic regression model, these errors tend to occur with higher frequency in the critical region of the regression. This critical region is typically around the area where the probability of an event is 50%.

While confidence intervals can be useful for quantification of uncertainty and regions of uncertainty, confidence intervals applied to the threshold in logistic regression have rarely been identified in the literature. Consider the previous example mentioned, if a logistic regression model produces a .55 probability of penetration for a specific input and the threshold is 0.5, it is classified as a penetration. Although, when a 95% confidence interval is applied to the 0.5 threshold probability, the threshold may span from 0.4 to 0.6 for example. Then the question stands as to whether or not a bullet with that velocity should be labeled as likely to penetrate the armor as the input probability of 0.55 falls within that threshold range. In this paper we investigate the use of confidence intervals around the threshold probability value and instead of labeling observations that fall within this region of uncertainty as a 1 or 0, we label them as being in a region of '*high uncertainty*' which we call a '*grey area*.'

This paper will discuss an application of confidence intervals to the threshold decision value used in logistic regression and discusses its effect on changing the quantification of false positive and false negative errors. In some cases this shows that potential errors are mitigated. Also, the size of the threshold interval and amount of errors can lead to insights on the amount of uncertainty in a logistic model fit. Section 2 contains a brief literature review of relevant work. Section 3 presents the methodology, outlining the process taken to collect and analyze the data presented. Section 4 provides the results, including a detailed discussion. Section 5 addresses an application of this model to a real world study involving bullet penetration armor testing and briefly discusses the use of experimental design and its influence on the region of uncertainty. Finally the paper concludes in Section 6 with a brief summary of the outcomes, conclusions drawn, and suggestion for future studies.

2. Literature review

The potential for logistic regression to provide an accurate predictive model is what makes it so useful in medicine and defense applications. While it is common to observe the 50% threshold, that probability may not be appropriate in some cases, as 50/50 odds are rarely ideal. When applied to medical procedures or the durability of military equipment, criteria for success may be much higher such as 70%. Tan et al.[5] outline an application of this level of logistic regression to multiple factors that influence the likelihood of a successful coronary angioplasty. By taking the critical variables and assessing success rates in logistic regression they were able to classify patients into three groups, low, high, and intermediate probability of success. These ranges were determined by the 30% and 70% thresholds. The percentage of successful treatments in the high region was 91% with a 95% confidence interval spanning 83% to 96%. Alternatively in the low region held failures 81% of the time with a 95% confidence interval of 64% to 92%. These results show the accuracy of the model and prove that logistic regression can be a useful predictive tool for optimal patient selection.

The motivating application for this study is in bullet penetration testing of armor. Logistic Regression finds many applications in military settings. Halstead and Brown[6] outline an interesting application of logistic regression for choice analysis of enlisted army applicants. In their analysis, observations are made on army applicants in the delayed entry program or DEP. Applicants in the program will eventually decide to honor the contract and enter basic training or return to the civilian population. There are great financial losses associated with an applicant choosing not to honor the contract. As a result of the analysis by Halstead and Brown[6], the most relevant variables were identified and addressed. In this case, time in DEP, age, and education were identified in the logistic regression model as the most influential variables on whether the applicant enters basic training. This is a valuable result, as based on the outcome the United States Army Accession Command can modify their recruitment process and provide a base of applicants more likely to enter basic training after involvement in the DEP.

In another application of logistic regression, Woods et al.[7] identify signs of mental disorders in a sample of individuals in the military. Their model allowed for an improved method of psychological measurement by using logistic regression. The logistic regression was used to determine the likelihood of personality disorders among the military sample based on self and peer reflective data taken from personality assessments. In particular, they observed strong relationships among their variables to disinhibition, entitlement, exhibition, negative temperament, and workaholism.

In many applications of logistic regression, the result the user wishes to find is the input value for which the probability of success is 50%. This is often notated as the letter representing the input variable with the subscript 50. Lu et al.[8] use logistic regression in this very way to find C_{50}, or the concentration of a dose of medication for which the probability of the effect occurring is 50%. From there they wanted to confirm whether or not this regression analysis was accurate. In each simulation, the data generated was used to find the actual dosage of C_{50}, or the input dosage that corresponds to the 50% threshold. A 95% confidence interval was generated around each C_{50} value generated. This gave them a method to determine the accuracy of their model. Using those intervals for the expected C_{50} dose, they compared their results with the actual data. The results showed that, when bias was properly controlled, in 90% of cases the C_{50} value fell within the expected C_{50} 95% confidence interval. This is one successful example which displays the applicability of confidence intervals to logistic regression in regards to threshold analysis.

Adding the appropriate confidence interval to a linear regression is fairly simple, and many statistical software packages can produce the intervals, but the calculations become a bit more complex for logistic regression. Sofroniou and Hutcherson[9] outline the approach to adding a confidence interval to various regressions in a way that will account for random binomial error. The outcome is accurate confidence intervals for the expected response in logistic regression. One of the primary approaches is applying a confidence interval to the β coefficient by adding and subtracting the asymptotic standard error multiplied by the z value. The approach in this paper is slightly different in how the confidence interval is formed. In this paper we incorporate a model that takes the fit of the generated regression and adds and subtracts the standard error of that fit.

The points at which the probability of an outcome is 30%, 50%, and 70% are common examples of threshold values in logistic regression. To test the accuracy of applying confidence intervals to logistic regression threshold, Hurwitz and Remund[10] compare two related, yet independent logistic regressions. They take the same threshold value from each, find the difference, and apply a 95% confidence interval to the outcome. The result was interesting in that, by using a Monte Carlo simulation, they were able to verify, at multiple probabilities, that the percent coverage of confidence intervals generated were within 1% of the desired level of confidence. In this paper the analysis will be focused on one example of applied logistic regression, in regards to bullet penetration. The testing done by Hurwitz and Remund[10] displays strong evidence that confidence intervals are consistent at multiple thresholds. Their approach takes the difference of two independent tests at varying thresholds and finds the respective confidence intervals at those points. The testing was performed at thresholds of .1, .2, .5, .8, and .9. They analyze the conservative thresholds as well as the most extreme thresholds and found consistently accurate results.

To yet again address the wide variety of applications of this sort of statistical approach the next example shows the application of these principles in wildlife management. The main desire of many of these models is an accurate model for predicting future outcomes based on current data. While these models cannot predict events with absolute certainty, it can help the user make better decisions regarding the case they study as now they have a strong impression of what to expect. Gude et al.[11] outline three essential things that a logistic regression model must do in order to be accurate for predictive analysis. A good logistic regression model should show clear difference between the high and low probability instances, it should produce predictions that are close to the actual observed results of the population, and it should accurately account for error and uncertainty.

Where this paper will differentiate from these sources is in the interpretation of the confidence interval and how it can be used to eliminate predictive error. Specifically, confidence intervals will be applied to multiple threshold values. The threshold value is used to determine the necessary level of acceptance in predictive analysis. This was addressed previously in section 1. In the basic analysis of the bullet penetration example, if the logistic regression produces a probability of .55 and the threshold is 50% it would be classified as a 1(a penetration). The confidence interval is essential in this application as logistic regression is merely a point estimate. The true 50% threshold value could lie within a large range of percentage values depending on the size of the interval. Applying this confidence interval around the threshold creates a better-established region where truly confident predictions can be made. For example, in the same bullet penetration example, if the 50% threshold value has a 95% confidence interval, its confidence bounds could be 42% and 58%. In that case, the previous result of .55 would no longer be classified as a penetration as it is not greater than 58%. Depending on the experiment it would be classified as a grey area point. This is a valuable approach to eliminating predictive error as error occurs most frequently around the threshold.

A 50% threshold is not typically a confident prediction when it comes to situations involving someone's life. A more strict threshold such as 70% or higher allows for a larger safety net when live testing; this is because a higher threshold is known to lower the amount of false positives produced.

3. Methodology

Monte Carlo simulation studies, mimicking live test data from body armor testing, were used to investigate the relationship between threshold choices, confidence interval region, and correct/incorrect identification of binary output using a logistic regression. The programming language R was used to generate all experimental data and collect results. This section is used to discuss the approach used to study various threshold settings and the inclusion of a confidence interval for predicting the binary outcome of a response.

3.1. *Simulation*

The flow chart in Figure 1 outlines the generic process of the simulation and data collection used in this research. Initially, the user inputs specific settings that will define the logistic regression. These settings are the logistic regression coefficients, the variance associated with the random error, and the approximate range that the data spans. With these settings, simulated sets of data that would be collected in live testing are generated.

The data generated contains *n* rows consisting of the input, *x*, and a binary response. The data is generated using the following logistic regression equation.

$$P(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

After the initial data set is created a logistic regression is fit to this data. Using pre-specified threshold and confidence intervals, observations are predicted as success (1), failure (0), or unknown (grey). Figure 2 provides a visual representation and shows how each point is classified. The different classifications separate the correct points, error points, and grey points.

Once each point is classified it is recorded and the number of points in each of the eight regions is calculated. The entire process in Figure 1 is repeated 50 times for a specific set of inputs. Based on initial tests, a replication value of 50 was shown to provide

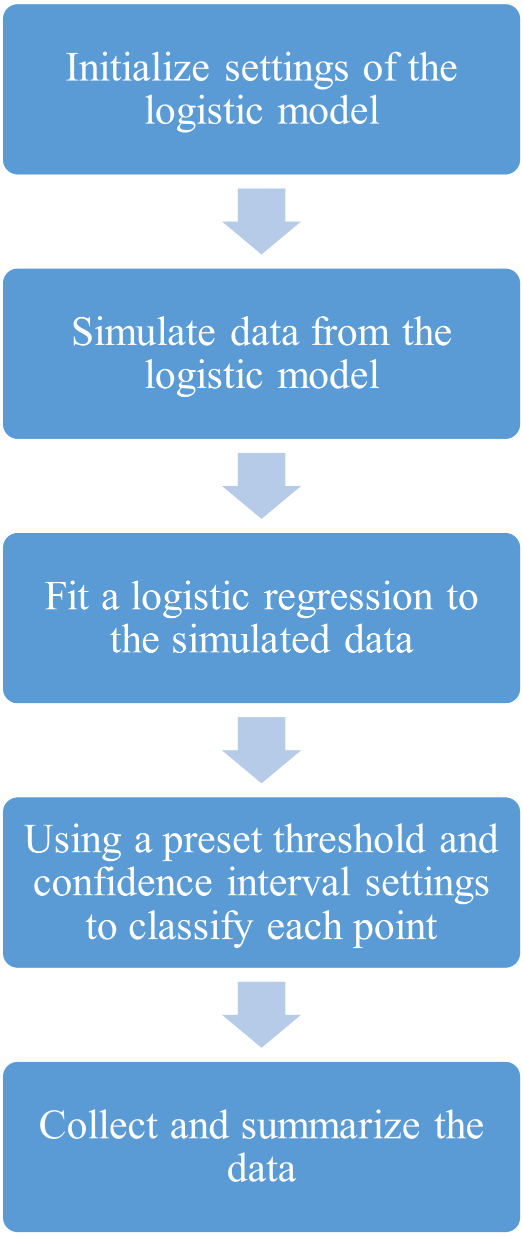
Initialize settings of the logistic model
Simulate data from the logistic model
Fit a logistic regression to the simulated data
Using a preset threshold and confidence interval settings to classify each point
Collect and summarize the data

Figure 1. Simulation and data collection flow chart

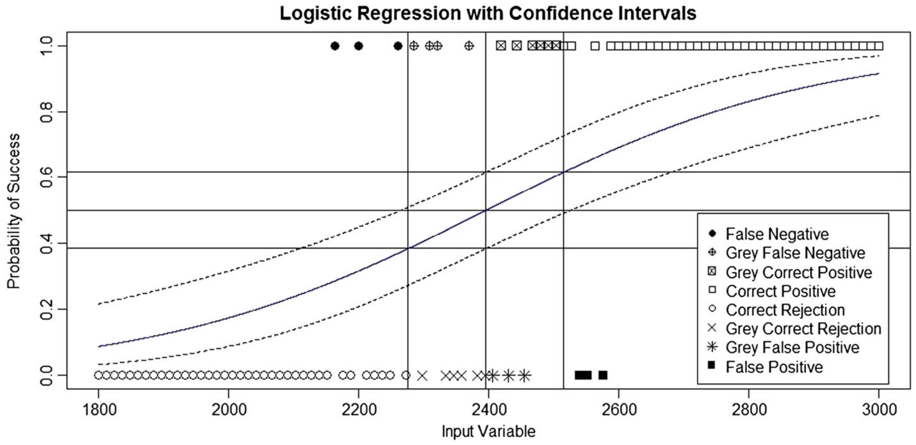

Figure 2. Diagram showing the eight regions of points

asymptotically stable results across a variety of models and chosen levels of error. After those 50 replications the data is averaged and summarized.

To observe the effects of the confidence interval on the amount of error, we extract the data from Figure 2 with and without the confidence intervals around the threshold, these results are presented numerically in Table I. In the standard simulation, the result is 87% of the points were predicted correctly and 13% were error points. When the confidence interval is applied, a grey region is formed. The grey region represents points of uncertainty where we cannot determine with enough certainty whether or not they will be greater than the threshold. In this case, the grey region is formed using a 90% confidence interval. Using this region to omit the points in the grey region results in 75% are correct and 6% represent error. There is a 7% reduction in incorrectly classified observations, but this is at the expense of 12% correctly classified observations.

3.2. *Experimental design*

The process outlined in Figure 1 is repeated 50 times per run; this is considered a single trial. The model we developed can run through the replicated trial in a few seconds. In a single test there are four main variables of interest that are presented in Table II: threshold choice, sigma, sample size, and alpha.

An experimental design approach was used to study the influence of each of these factors on the 13 responses of interest presented in Table III. Generating a full factorial design in the variables shown in Table II results in a total of 81 trials.

Table IV shows five of the 81 trials. For example, trial 2 simulates a sample set of data with a size of 20, alpha is set at .05 indicating a 95% confidence interval around the threshold, squared root of random error variance (labeled as sigma) of 1, and a .5 threshold.

3.3. *Specifics of how the model works*

The motivating application for this study is body armor testing. Using approximate data from live tests, relatively realistic logistic regression coefficients were chosen, the actual data from experimental tests is classified, and units of measurement will not be discussed. The equation shown below is the regression equation used in this model to generate the data, including the specific coefficients.

Table I. Results of simulation from Figure 2

	% correct	% error	% grey
No confidence interval	0.87	0.13	
Confidence interval	0.75	0.06	0.19
Difference	.12	.07	

Table II. Variables of interest in experimental study

Factor	Levels		
Threshold	.3	.5	.7
Sigma	1	1.5	2
Sample size	20	50	100
Alpha	.05	.1	.2

Table III. Responses collected through the experimental design

Response
% correct rejection
% correct acceptance
% correct
% false positive
% false negative
% error
% grey correct rejection
% grey correct acceptance
% grey correct
% grey false positive
% grey false negative
% grey error
% grey

Table IV. Set of trials

Trial	Sample size	Alpha	Sigma	Threshold
1	20	0.05	1	0.3
2	20	0.05	1	0.5
3	20	0.05	1	0.7
4	20	0.05	1.5	0.3
5	20	0.05	1.5	0.5

$$P(x) = \frac{1}{1 + e^{-(-11+.0046x)}}$$

Threshold, sample size, and alpha values are briefly discussed. The threshold value (0.3, 0.5, or 0.7) represents the cut off decision point in order to classify an observation as a 1 or 0. The sample size refers to the sample size that is intended to be used (thus the number of data points that will be simulated) in the actual live body armor testing experiment. The input alpha, α, with values between 0.05 and 0.2, are used to determine the confidence interval level that will be applied around the threshold cut off. For example, for a threshold choice of 0.5, a 95% confidence interval ($\alpha = 0.05$) around the predicted probability may result in a window of 0.3 to 0.7, whereas an 80% confidence interval ($\alpha = 0.2$) will result in a much narrower window, say 0.45 to 0.55. The confidence interval around the threshold determines the grey area, and only bullets with velocities above the threshold upper limit will be classified as a 1 (penetration) and below the lower limit threshold will be classified as 0.

The input sigma, σ, is used to represent the standard deviation of the noise associated with the experiment. Note, there are several ways of generating simulated data from a binomial experiment. The binomial distribution and associated link function can be used to generate binary responses in a region of interest. Alternatively, the linear portion of the model may include an error term. We simulated data using both methods. We found that simulating the error within the link function provided us with greater flexibility to simulate sets of data that would be more or less subject to error, and at the same time allowing us to control this error in our experimentation. We were curious to understand the influence that variation in the data set had on the amount of correct and incorrect decisions made. Based on this assumption that the simulated data should be asymptotically

$$\mathbf{x}'\mathbf{b} \sim N\left[\mathbf{x}'\mathbf{b}\ , \mathbf{x}'(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{x}\right]$$

we chose a range of values representing the error variance (sigma squared) at the centroid of the data.

A single experiment, simulated 50 times, generates a collection of percentage values associated with each of the responses shown in Table V. Note that typically when users categorizing predictions from a logistic model, the approach shown in Table V is taken. The threshold divides the random data, and if it is above the threshold input value it is predicted to be a success (1), and if it is below the threshold input it is predicted to be a failure (0). Four types of classifications can be distinguished: correct positive, correct negative, false negative, and false positive. Here we only show the correct decisions as 'correct.'

In the case of our experimentation and interest, the incorporation of a confidence interval around the threshold further segments the data into eight regions as opposed to four. This breakdown is shown in Table VI. The grey region, or the area within the confidence interval around the threshold, is a new subset of the data in which the true threshold value may fall within to some degree of certainty. Note that the amount of observations that fall within the grey region can act as a surrogate for estimating the amount of noise or uncertainty in the model fit. Note however, that the size of the grey area is intrinsically related to model uncertainty, but is also directly proportional to the chosen confidence level.

Table V. Prediction classification

		Predicted response	
		1	0
Actual response	1	Correct	False negative
	0	False positive	Correct

Table VI. Prediction classification with grey area

		Predicted response			
		Outside of confidence interval		Within confidence interval	
		1	0	1	0
Actual response	1	Correct positive	False negative	Grey correct positive	Grey false negative
	0	False positive	Correct negative	Grey false positive	Grey correct negative

4. Results

In this section we divide the results into three subsections. The first subsection provides results from regression analysis used to compare the variables (Table II) and their influence on classification rate for each of the 13 responses (Table III). Based on the regression analysis, the individual effects that are shown to be significant among a large quantity of models are analyzed graphically and their affects are discussed. Finally, Pareto optimal fronts are generated to observe the ideal criteria to perform these simulations. Throughout this section, the results are referred to as percentage values but are displayed on a scale of 0 to 1.

4.1. Regression analysis

Using the experimental design (discussed in section 3) and the data collected, regression models were fit to each of the 13 responses. The general form of the regression models fit to these responses can be described by the following equation. Only significant terms (at a 95% confidence level) were retained in the model.

$$y_c = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i<j}^{k}\sum^{k} \beta_{ij} x_i x_j + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \varepsilon$$

The subscript *c* represents the responses shown in Table III. A regression model was fit to each one of these 13 responses; however we will only be addressing a few of the more interesting results. The coefficients of each of the significant terms in the regression equations were recorded as well as the goodness of the fit as judged by R^2 and R^2_{adj}. The average R^2 among the 13 models was 0.949, and the average R^2_{adj} was 0.943. Table VII shows the number of occasions in which each term was significant out of a maximum of 13. Sample size^2, the interaction between sample size and sigma, and threshold^2 appeared in all 13 models. The terms sample size, threshold, and sigma were also frequently significant.

Below is a single fitted model.

$$\begin{aligned}\hat{y}_{percent\ error} = & -.0004 + .0015x_{SampleSize} + 0.0593x_{Sigma} - .275x_{Threshold} \\ & - \left(1.14{*}10^{-5}\right)x^2_{SampleSize} + \left(4.88{*}10^{-4}\right)x_{SampleSize}x_{Sigma} \\ & + .145x_{Alpha}x_{Sigma} - .012x^2_{Sigma} + .281x^2_{Threshold} - .0008x_{SampleSize}x_{Alpha}.\end{aligned}$$

The main effects of the model shown for percentage of error are provided in the profile plot in Figure 3 below.

The interactions are interesting. The strongest interaction in the model is the interaction between alpha and sigma and is shown graphical in Figure 4. As the amount of noise associated with the linear portion of the logistic model increases, the percentage of errors (total false positive and false negative errors) increases. However, the rate of errors made is larger and at an increased rate when the confidence interval around the threshold value is narrower (alpha = 0.2).

4.2. Discussion of individual effects

4.2.1. Sample size. The sample size of the simulations run had a clear influence on the percentage of points that fell in each specific region. In practice, sample size is the number of live tests run and may or may not be controlled given budget restrictions. The results regarding the influence of sample size are shown in Figures 5 through 7. As the sample size grew larger, the percentage of error

Table VII. Number of instances in which the characteristic is significant

Model terms	# models where term is significant
Sample size	11
Alpha	6
Sigma	10
Threshold	11
Sample size^2	13
Sample size × alpha	8
Sample size × sigma	13
Sample size × threshold	6
Alpha^2	6
Alpha sigma	9
Alpha × threshold	6
Sigma^2	5
Sigma threshold	8
Threshold^2	13

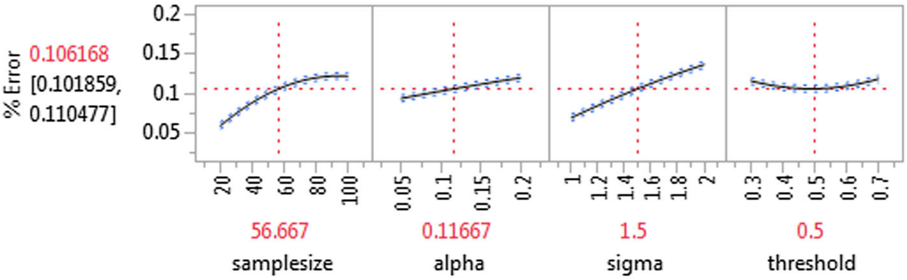

Figure 3. Profile plot for main effects in the percent error model

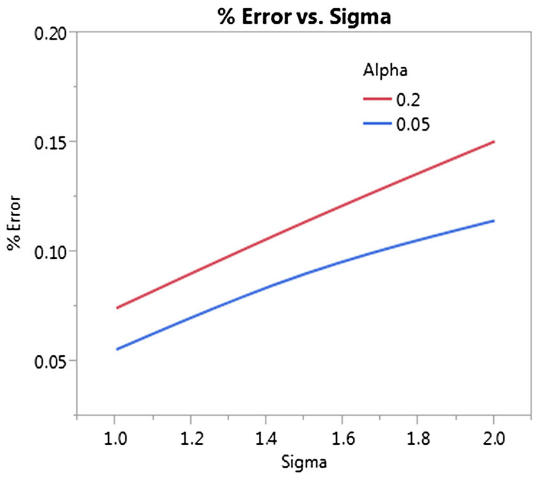

Figure 4. Percent error vs. sigma, segmented by alpha

points and correct points both increased. Having less points creates a weaker logistic regression fit, and this results in the confidence interval being proportionally larger when sample size is lower. This indicates that a larger sample size is ideal for the most accurate distribution of points; however it may be unrealistic to arbitrarily increase sample size in practice.

An important result of this analysis is shown in Figure 7. This figure shows the relationship between sample size and the percentage of grey correct and error points. As the sample sizes change from 20 to 100, the average percent of grey correct points drops from approximately 17% to approximately 7%. Alternatively, the percent of grey error points drops from approximately 11% to approximately 6%. This is very important to recognize, as the goal is to minimize grey correct points and maximize grey error points in order to reduce error. As sample size increases both percentages decrease, but the percent of grey correct decrease in greater numbers which is very beneficial.

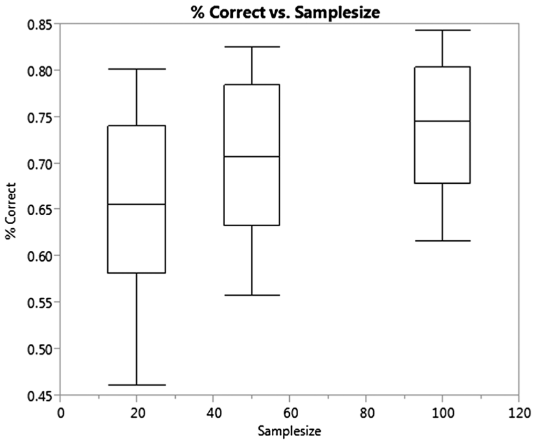

Figure 5. Percent correct vs. sample size

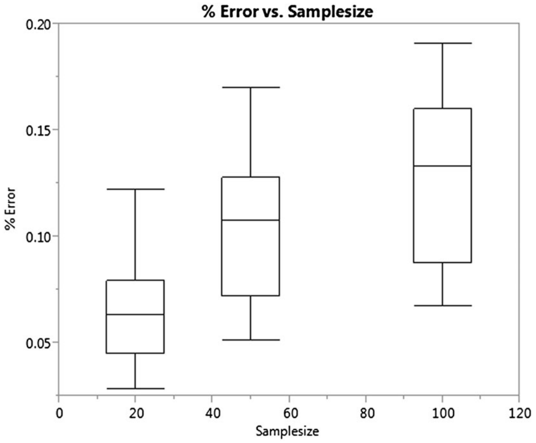

Figure 6. Percent error vs. sample size

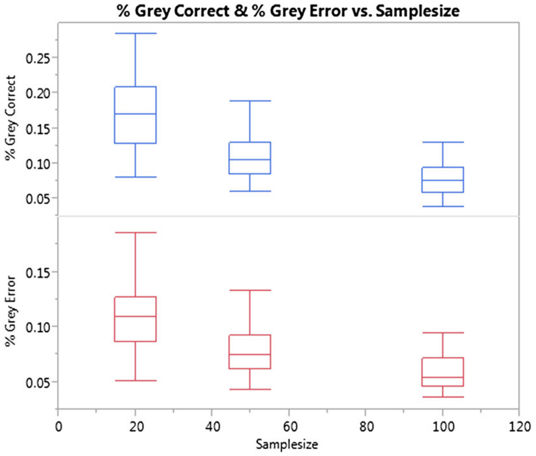

Figure 7. Percent grey correct and percent grey error vs. sample size

4.2.2. Sigma. Changing the value of sigma causes varying levels of random error or noise added to the regression. As the sigma value increases so does the range of the noise. The levels used in this model were 1, 1.5, and 2. These sigma values reflect realistic testing conditions.

The value of sigma had a significant impact with regards to the percent of error and percent of correct points. Because sigma increases the range of random error it is seen in Figure 8 that, as sigma grows, the variance in both the percentage of error and correct points increases. The problem this poses is less predictability. If there is a great degree of random error in an experiment, predictive

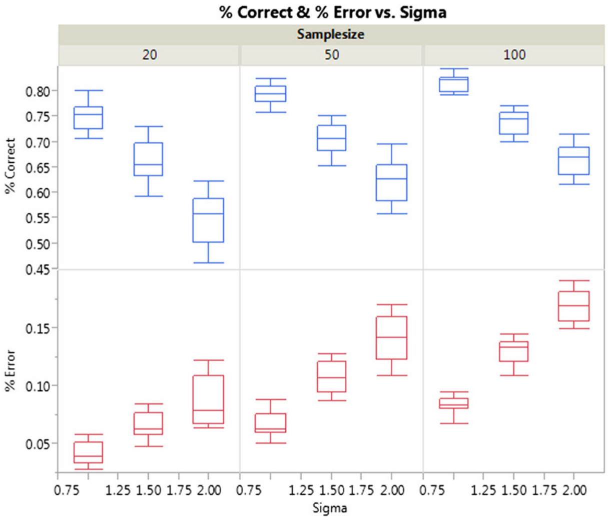

Figure 8. Percent correct and percent error vs. sigma grouped by sample size

analysis is very limited. It can also be seen that, as sigma increases, the amount of error points increases, and the amount of correct points decrease. This is an expected result.

Sigma has an effect on the type of points that are located in the grey region formed by the confidence interval and threshold. Figure 9 illustrates this relationship. As sigma increases, the amount of points in the grey region will increase. It is not beneficial to have too many points in the grey or uncertainty region. Because sigma cannot be controlled, as it is the natural degree of random error, it is important to control the experiment in way to compensate for the error. For example, keeping the sample sizes higher allows there to be much less variance in the grey region. From this it can be determined that if the grey area is too large the model being tested is very weak. For example, Figure 9 illustrates that at sample size of 20 and a sigma level of 2, upwards of 35% of all points are in the grey region. This would be a clear indicator that the model in question either needs to be reworked, its parameters need to be modified or the data needs to be recollected. This result also speaks to the uncertainty in the data. Fewer samples and higher noise increase the amount of uncertainty in the model, which is reflected in the percent of observations in the grey region.

4.2.3. *Alpha*.

The variable alpha influences the size of the confidence interval that will surround the regression. The smaller the alpha is the larger the interval will be. Once these varying alpha's were plotted with the responses several things can be determined. Figure 10 illustrates that when alpha was increased the amount of points in the grey area decreased. This is because the grey area is based on various intersections with the confidence interval. So as the region encompassed by the confidence interval increases the region the grey area is in will also grow. Because the grey area is used to eliminate the false positives and false negatives it comes to no surprise that if the alpha increases, the amount of error points will increase, but the grey area will decrease.

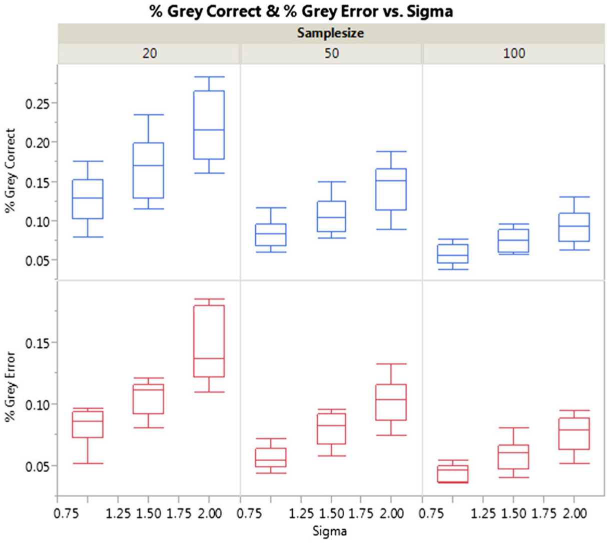

Figure 9. Percent correct and percent error vs. sigma grouped by sample size

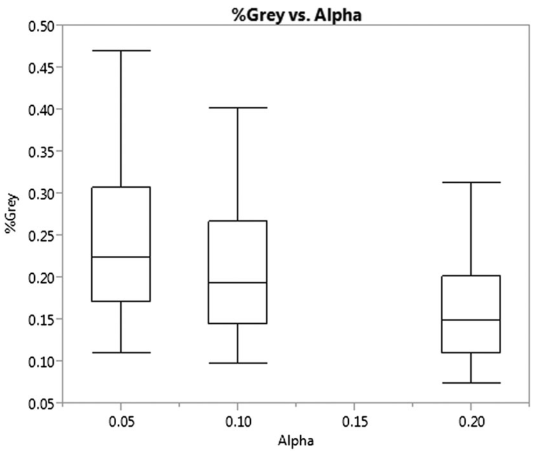

Figure 10. Percent grey vs. alpha

The tradeoff with decreasing the sizes of the confidence interval is that there is now less certainty in the analysis. Another issue is that both the percentage of correct and the percentage of error points will increase. Increasing the percent correct is beneficial but increasing error is harmful. As a result is important to analyze the rate at which they change. Figures 11 and 12 illustrate that as alpha increases from .05 to .2, the percent correct increases by roughly 5% and the percent error also increases by about 5%. The ideal alpha is dependent on the goals of the experiment. If the goal is to maximize percent correct but reduce some error, a higher alpha would be ideal. If the goal is to eliminate as much error as possible while trying to maintain some correct points, then a lower alpha is best.

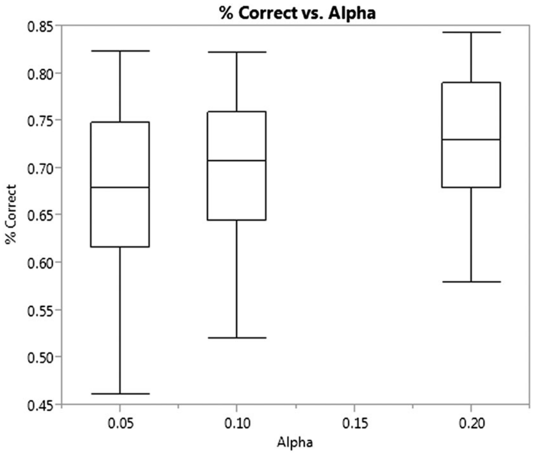

Figure 11. Percent correct vs. alpha

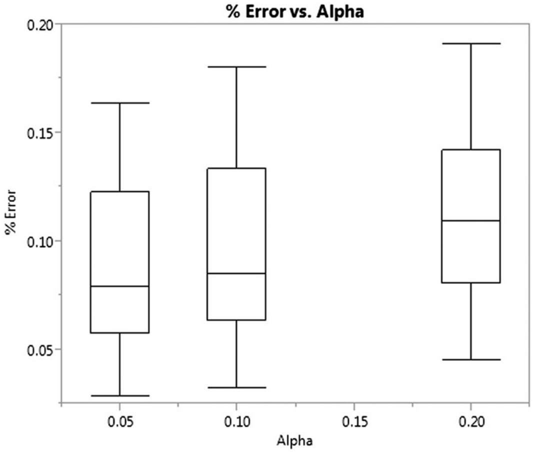

Figure 12. Percent error vs. alpha

4.2.4. Threshold. The threshold value determined where the center of the grey area will be located. It is a cutoff probability point used to determine bullet penetration (1) or no bullet penetration (0). Because the distribution of points is usually consistently spread along the x axis, there is no correlation between where the threshold value is and how many points lie in the grey region. The clear pattern that appears based on the threshold is the false negative and false positive points. When the threshold is moved closer to either side of the regression either the false positives or false negatives will increase and the other will decrease. This is understandable because in logistic regression the extreme points that are toward the edges of the plot tend to be more consistent and overlap far less than at the center. This causes the lopsided type of error. This relationship is clearly shown in Figure 13. The plots on the left side of Figure 13 show a smoothed line through the mean, whereas the right plots present the box plots, which include an aspect of variability.

Another key relationship is the total amount of error at each threshold. Figure 14 shows that on average the threshold of .5 will allow the most errors to be removed. This is because the greatest amount of uncertainty occurs where the probability is 50%. Centering the confidence interval at that area will allow the grey area to cover the region with the highest density of error points.

Dividing other areas of analysis by threshold reveals some interesting relationships. For example, in the case of the bullet penetration testing, false negative errors are the most detrimental. Figures 15 and 16 illustrate the various relationships of false negative errors with respect to threshold, sample size, and alpha. In all cases the lowest levels of false negative errors occur at the threshold of .3, and they also have the lowest rates of change with respect to other variables like alpha and sample size. The inverse could be said regarding false positive errors should that be the more critical type of error.

4.3. Pareto optimal fronts

In this analysis, two sets of Pareto optimal fronts were generated to examine the ideal setting given a certain amount of random error. The plots were created by eliminating any non-optimal points. In this experiment the goal is to eliminate errors. In order to

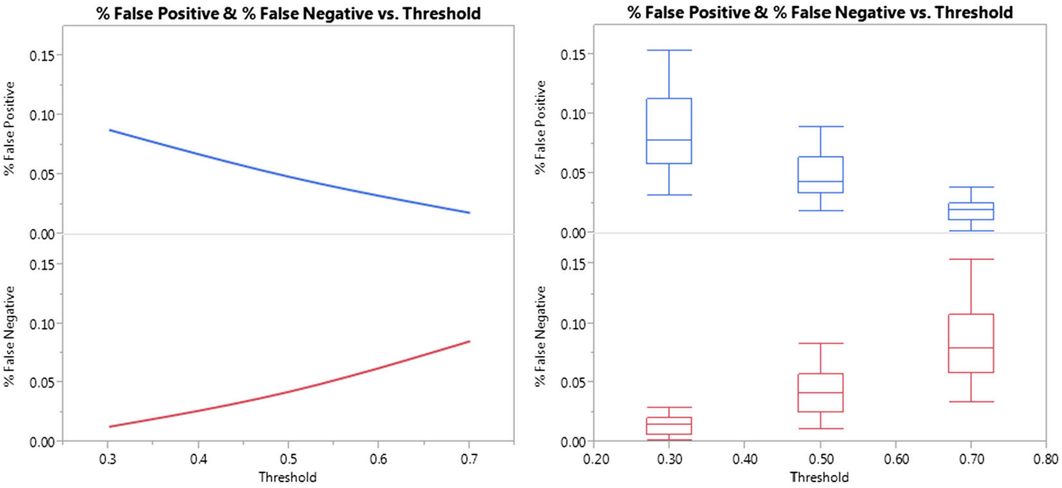

Figure 13. Relationship of false positive and false negative errors vs. threshold

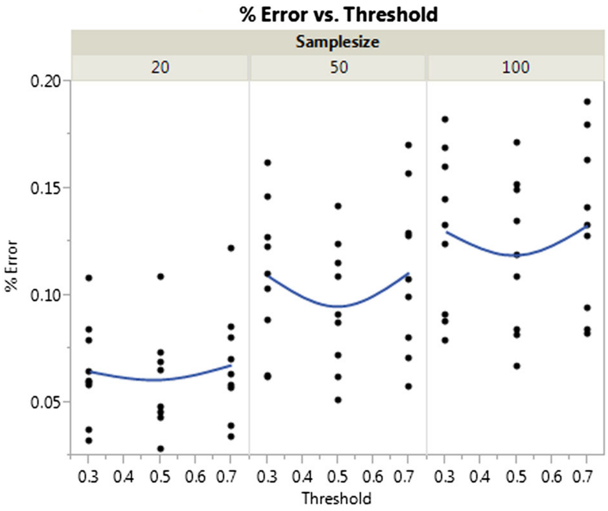

Figure 14. Percent error vs. threshold

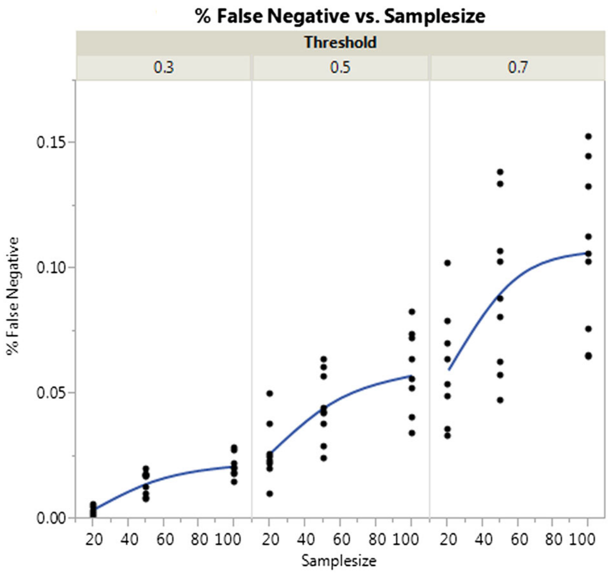

Figure 15. Percent False negative vs. sample size

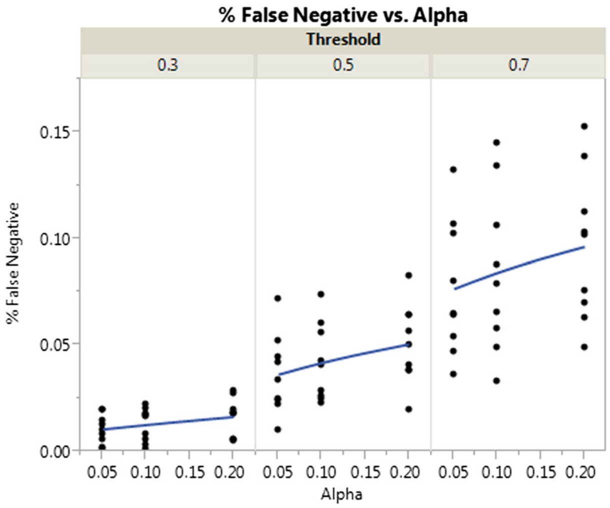

Figure 16. Percent negative vs. alpha

remove those errors points are classified in the grey region and are removed. Several of these grey points are predicted correctly and removing them is not beneficial. In order to determine optimal criteria these two are compared against each other. The best case scenario is to have both low error and low grey correct points. Figure 17 is used to determine the ideal sample size to test, and Figure 18 is used to determine the ideal confidence interval size. Figure 17 illustrates that as sample size increases, grey correct decreases, yet error increases. For the least grey correct, larger sample size is ideal, and for the least error smaller sample size is more appropriate. Keeping the sample size around 50 is clearly the best case to eliminate portions of each type of occurrence, especially as error increases.

Figure 18 shows the relationship of sigma and alpha. It shows that in cases of low error, there is little change among the various confidence intervals in terms of the optimal amount of error and the optimal amount of grey correct points. The trends show that as alpha increases, percent error increases as well and percent grey correct will decrease. Overall the changes are very minimal. As a result, in these cases it is better to use a smaller alpha as it has higher confidence. Alternatively, in cases where standard error is high, using a larger alpha would help reduce more error and grey correct points at the expense of some confidence in classification of points.

4.4. Predictive results

All of these experiments were run again under slightly different conditions. Rather than comparing the set to its own regression, a second set of data generated from the same initial settings was compared against the regression, and the results were nearly identical across the board. There were no clear differences in the results produced. This verifies the model as a useful predictive tool.

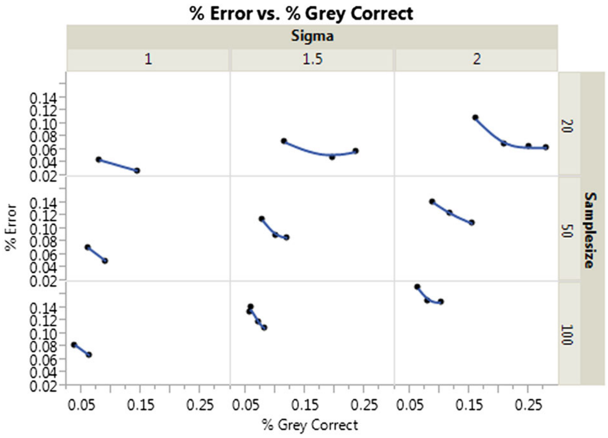

Figure 17. Pareto optimal front of percent error vs. percent grey correct subset by sigma and sample size

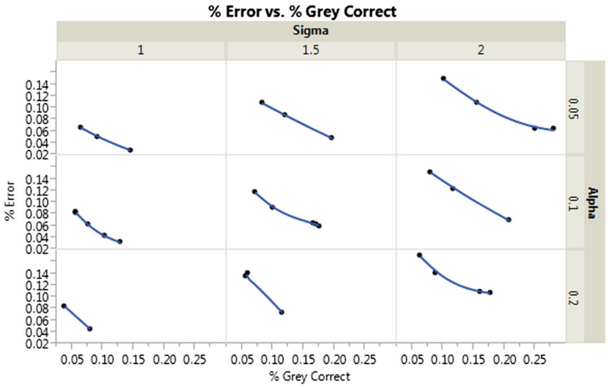

Figure 18. Pareto optimal front of percent error vs. percent grey correct subset by sigma and alpha

5. Applications

This approach is a great method to reduce predictive error, and as a result it has a wide variety of applications. Specifically, this type of simulation can truly benefit areas in which live testing is risky or expensive, so long as accurate data already exists. The value of this model is that it recognizes the potential error in the logistic regression analysis. The issue with the point estimate of a logistic regression is that in very sensitive cases, a few percent of leeway on a threshold value may result in false predictions such as a misdiagnosis or a judgment call that costs the life of a soldier. By expanding the threshold from a point estimate to a confidence region, the analysis becomes discerning.

The specific example that influenced this paper was the case of probability of bullet penetration with respect to bullet velocity. Pinz[12] discusses the application of logistic regression to predicting probability of penetration as a function of a number of variables of interest. In our case, we are only interested in predicting probability of bullet penetration as a function of bullet velocity. Once a logistic model is fit, a number of assessments regarding the suitability of the armor for use in combat are based on the results. This situation puts soldiers at high risk. By utilizing our methodology for identifying a grey area and employing a confidence level of 90% or 95% can ensure that equipment is fit for live use with much more certainty.

Another issue to consider regarding an application of interests is which type of error is more of an issue. In the case of bullet penetration where 1 represents penetration, false negative error is of the greatest concern. In order to reduce this type of error specifically, it may be ideal to shift the threshold value from .5 to .3 or a similar value. To observe the effects of shifting thresholds, we analyzed this shift at varying confidence intervals. The tradeoff to shifting the threshold is that while one type of error is nearly eliminated the other is increased dramatically. The justification for this is that one type of error is less critical than the other. When errors are equally critical in a test, keeping the threshold centered is the best approach.

The data collected in this paper was done so with a standard set of data, spread evenly about a range of values. We refer to this as the 'base design.' It is of interest to compare the performance of this base design with an *I*-optimal design created with respect to the intended logistic regression model. When comparing the results of using a base design or optimal design approach during armor testing, eliminating false negative errors and identifying the ideal threshold value are of interest. This was done in this section, and the experimental results are shown in Tables VIII and IX. These experiments held identical descriptive criteria; a standard error of 1.5, a 90% confidence interval, a sample size of 50, and the results are a summary of 50 simulations. Where the two tests differ is

Table VIII. Base design, 90% CI, sigma = 1.5, threshold = .3, sample size of 50

		Predicted response			
		Outside of confidence interval		Within confidence interval	
		1	0	1	0
Actual response	1	.4896	.0128	.0456	.0252
	0	.0956	.1864	.0652	.0796

Table IX. *I*-optimal design, 90% CI, sigma = 1.5, threshold = .3, sample size of 50

		Predicted response			
		Outside of confidence interval		Within confidence interval	
		1	0	1	0
Actual response	1	.3544	.0252	.04	.024
	0	.0576	.3668	.068	.064

in the data that is generated to be tested. Table VIII represents the base design, where 50 points are distributed evenly across a range of values applicable to the experiment. Table IX represents an *I*-optimal design, in which five observation locations (velocities) are tested ten times each.

The two tests produced similar results; however each has their specific benefits. In the case of the base design, it has the lower percentage of false negative points by about 1.25%. As for the optimal design, it held roughly 5% more correct points. Low false negative error is the primary goal of this experiment, but keeping a high percentage of correct points is also critical. This comparison shows what can be anticipated when designing and optimal design for testing, and as expected, the results are a slight improvement.

When the two methods of design are compared with the goal of reducing all types of error and optimizing correct points the ideal threshold would be closer to .5. As a result of testing the optimal design model outperforms the base design both in terms of total percent correct and total percent error. Exploring a variety of experimental designs could provide very interesting results when compared through this model.

6. Conclusions

This paper covers the methods, approach, and results of applying confidence intervals to threshold values in logistic regression to improve the predictive ability of the basic logistic regression model. As a result of this analysis many conclusions were drawn regarding the proper methods of setting up the logistic regression model. Using the grey area classification as a region of uncertainty allows the user to minimize predictive error. Another useful interpretation of the grey area is using it to determine the quality of the model. If the grey area created from the logistic regression threshold encompasses a high percentage of predictions, the model is not strong enough for proper analysis.

There are still many potential applications of this approach to predictive analysis. The versatility of this model allows it to be applied to any set of binary data that fits a logistic regression and can simulate the experiment with varying degrees of error. We specifically analyzed its effect on bullet penetration testing; however it would be interesting to see the impact on various sets of data from a variety of domains. Also, it would be interesting to expand this study to include multiple inputs rather than a single input variable as discussed in this paper.

It would be beneficial to observe the effects of this type of model on different experimental designs as well. In this paper a comparison was drawn between the base design and the *I*-optimal design, but further work is necessary.

References

1. Hosmer DW, Lemeshow S, Sturdivant RX. *Wiley series in probability and statistics: applied logistic regression* (3rd edn). John Wiley & Sons: New York, NY, USA, 2013.
2. Sun Q, Ding J, Xu D, Chen Y, Hong L, Ye Z, Sheng J. Prediction of the prognosis of patients with acute-on-chronic hepatitis B liver failure using the model for end-stage liver disease scoring system and a novel logistic regression model. *Journal of Viral Hepatitis* 2008; **16**:464–470.
3. Mclaren C, Chen W, Nie K, Su M. Prediction of malignant breast lesions from MRI features. *Academic Radiology* 2009; **16**:842–851.
4. Sullivan L. Estimation from samples. *Circulation* 2006; **114**:445–449.

5. Tan K, Sulke N, Taub N, Watts E, Karani S, Sowton E. Determinants of success of coronary angioplasty in patients with a chronic total occlusion: a multiple logistic regression model to improve selection of patients. *Heart* 1993; **70**:126–131.
6. Halstead J, Brown D. Improving upon logistic regression to reduce army DEP loss. *IEEE SIEDS Conference* 2004; 191–201.
7. Woods C, Oltmanns T, Turkheimer E. Detection of aberrant responding on a personality scale in a military sample: an application of evaluating person fit with two-level logistic regression. *Psychological Assessment* 2008; **20**:159–168.
8. Lu W, Ramsay J, Bailey J. Reliability of pharmacodynamic analysis by logistic regression. *Anesthesiology* 2000; **99**:1255–1262.
9. Sofroniou N, Hutcheson G. Confidence intervals for the predictions of logistic regression in the presence and absence of a variance–covariance matrix. *Understanding Statistics* 2002; **1**:3–18.
10. Hurwitz A, Remund T. A large-sample confidence interval for the inverse prediction of quantile differences in logistic regression for two independent tests. *Quality Engineering* 2014; **26**:460–466.
11. Gude J, Mitchell M, Ausband D, Sime C, Bangs E. Internal validation of predictive logistic regression models for decision-making in wildlife management. *Wildlife Biology* 2009; **15**:352–369.
12. Pinz M. Data analysis of results from ballistic testing of small arms protective inserts, Master’s Thesis, Naval Postgraduate School, 2011.

Authors' biographies

Michael J Kist is a dual-degree BS/MS Industrial and Systems Engineering student at the Rochester Institute of Technology. He is a student athlete on RIT’s track and field team as a pole-vaulter. He has interest in various research topics including predictive analytics, design of experiments, and data analysis.

Dr Rachel Silvestrini is an Associate Professor of Industrial and Systems Engineering in the Kate Gleason College of Engineering at the Rochester Institute of Technology. She received her BS in Industrial Engineering from Northwestern University and her MS and PhD in Industrial Engineering from Arizona State University. Her research interests include design of experiments, response surface methods, data analysis, and simulation methodology.