BAYESIAN METHODS FOR INTERPRETING RISK OF INVASIVE SPECIES OCCURRENCE

A Thesis

Presented in Fulfillment of the Requirements for the

Degree of Master of Science

with a

Major in Statistical Science

in the

College of Graduate Studies

University of Idaho

by

Boya Liu

June 2014

Major Professor- Christopher J. Williams, Ph. D.

AUTHORIZATION TO SUBMIT THESIS

This thesis of Boya Liu, submitted for the degree of Master of Science with a major in Statistical Science and titled “BAYESIAN METHODS FOR INTERPRETING RISK OF INVASIVE SPECIES OCCURRENCE,” has been reviewed in final form. Permission, as indicated by the signatures and dates given below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor: ____________________________ Date ______________________
Christopher J. Williams, Ph. D.

Committee
Members: ____________________________ Date ______________________
Christine M. Moffitt, Ph. D.

____________________________ Date ______________________
Stephen Lee, Ph. D.

Department
Administrator: ____________________________ Date ______________________
Christopher J. Williams, Ph. D.

Discipline’s
College Dean: ____________________________ Date ______________________
Paul Joyce, Ph. D.

Final Approval and Acceptance

Dean of the College
of Graduate Studies: ____________________________ Date ______________________
Jie Chen, Ph. D.

ABSTRACT

Logistic regression is often used to predict the probability of an event, but it assumes perfect test sensitivity and specificity. However, most tests are not perfectly sensitive and specific. The prediction interval is an estimate of the interval in which the future observations will fall, and it can be applied to study the impact of imperfect test.

Logistic regression is applied to the study of invasive species, such as New Zealand mudsnails, which will potentially harm biodiversity and affect biotic homogenization.

In this study, we investigate how risk prediction in logistic regression on New Zealand mudsnail is affected by imperfect tests, using a Bayesian approach. The results show that the changes of mean sensitivity clearly affect the prediction interval width at a low temperature, but the effects of changes of mean specificity and the weights of prior distributions of sensitivity and specificity were less clear at both temperatures examined.

ACKNOWLEDGEMENTS

Under the guidance of my major professor, Christopher Williams, I got the opportunity to improve my research and my statistical background through my years in the graduate school. He takes care of his students, and his rigorous attitude toward teaching and research are the virtues I will strive to emulate in the rest of my life.

I appreciate the help and suggestions from my committee members, Christine Moffitt and Stephen Lee, who are very helpful in my thesis research. I would also like to thank people who discussed critical points and provided important suggestions to my research with me.

Without the support of my parents, I could not complete my graduate study in the United States. Their encouragement gives me mettle to overcome the difficulties I faced. Their support, no matter in financial, or in my study, will always be there with me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

1.1 Logistic Regression

The general linear model works well for many continuous response variables. However, it is not appropriate for a categorical response variable such as a binary response variable. With respect to the categorical response variable, logistic regression is a good option. Formally, the logistic regression model is

$$L = \ln(o) = \log \frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \qquad (1)$$

Where $\ln(o)$ is binary and represent the event of interest (response), coded as 0/1 for failure/ success, and

p is the proportion of successes,

o is the odds of the event,

L is the ln (odds of event),

x_i are the independent variables,

β_0 is the intercept and β_i is the slope coefficient (i.e., the expected change in $\ln(o)$ relative to one unit change in x_i).

The predicted probability of the logistic regression for a given x value is

$$\mathrm{p}(\mathrm{x}) = \frac{\mathrm{e}^{\hat{\beta}_0 + \hat{\beta}_1 \mathrm{x}_1 + ... + \hat{\beta}_\mathrm{k}}}{1 + \mathrm{e}^{\hat{\beta}_0 + \hat{\beta}_1 \mathrm{x}_1 + ... + \hat{\beta}_\mathrm{k}}} \qquad (2)$$

In our study, we are interested in using the prediction interval around $p(x)$.

Logistic regression is often used to predict the probability of an event, but typically assumes a test of perfect sensitivity and specificity. However, the data often come from different sources or may be collected by using different sampling tools or methods for measurement. In addition, the data may come from diagnostic tests that are assumed to be perfect (Williams and Moffitt, 2010). In fact, measurements in studies often suffer from both imperfect sensitivity and specificity. In this study, we are going to use the Bayesian approach to study the impacts of imperfect tests which can be shown by examining prediction interval width with different Bayesian prior distributions.

1.2 Background of New Zealand Mudsnail

The pioneering topic of recent global ecology—biological invasions occur when organisms come to live in the wrong place, including invasions by animals, plants, fungi, and bacteria (Simberloff and Rejmanek, 2011). In the early 20th century, people introduced non-native species to the western societies as resources. Today, some of the exotic species are assets for their aesthetic properties or economic value. However, as the introduction of non-native species increased in the 20th century, biologists had evidence to show the threat of invasive species on native species and ecosystems and human well-being. When it comes to the impacts of invasive species, scientists detected that it is uncertain and often delayed. For example, among known aquatic species introduced to six European countries, at least 69% have ecological impacts (Simberloff, Martin and el., 2011). However, these percentages are underestimated since some impacts are slight or inaccessible and can only be measured after intensive study. In addition, there were unsuspected effects of invasive species on communities and ecosystems. For example, certain invasive plants can transform ecosystems both above and below ground. To take actions to minimize the negative impacts of invasive species, there are a range of actions we can take, such as prevention, eradication, and long-term management (Simberloff, Martin, 2011).

The New Zealand mudsnail, *Potamopyrgus antipodarum*, is a very small freshwater snail with a gill and an operculum, an aquatic gastropod mollusk in the family Hydrobiidae, it has a dextral with right-handed coiling, and elongated shell with 7-8 whorls separated by deep grooves (USGS, 2012). Then, subsequent investigations have documented a rapid spread of it to the Madison River near the boundary of Yellowstone National Park, Firehole and Iower

Gibbon rivers. One reason for considering the New Zealand mudsnail to be a nuisance is that the snail has the ability to reproduce quickly and mass in high densities. When they become as dense as one- half million per meter square, they will impact the food chain of native trout and alter the physical characteristics of streams, where the West is known for great trout fishing. Another reason is its ability to survive in variable conditions. The mudsnails are able to survive in desiccation, a variety of temperature regimes, and are small enough that many types of water users could inadvertently be the mechanism for interbasin transfer of this nuisance species. In addition, the snail’s asexual reproduction causes great concern.

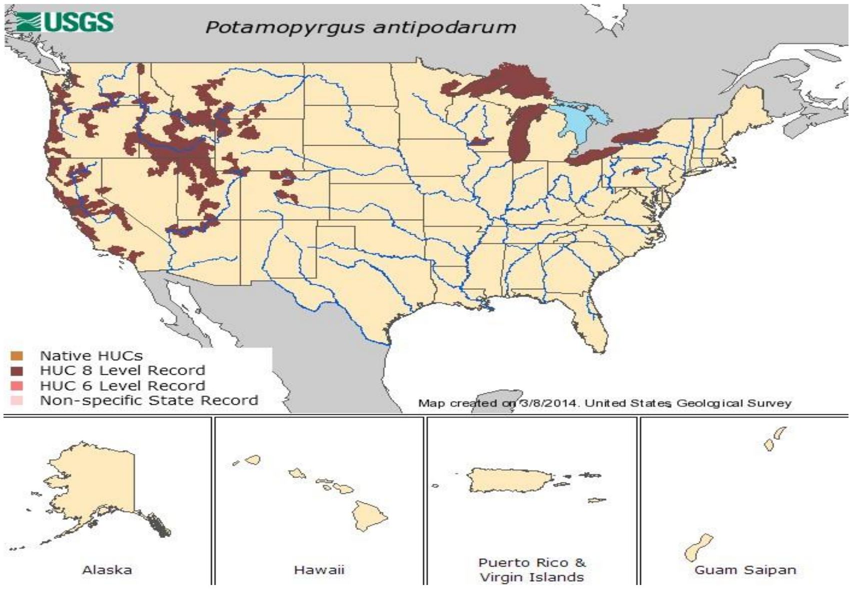

Figure 1. Potamopyrgus antipodarum HUC Distribution Map

Biological invasion, such as invasion of New Zealand mudsnail affects rivers and streams in the United States, and has negative impact on bio-systems, which will potentially harm biodiversity and simultaneously raise concern about biotic homogenization (McKinney and Lockwood, 1999). Invasive species may cause a decrease in ecological quality because of the changes in biological, chemical and physical properties of ecosystems (Elliot, 2003). In addition, biological invasion not only causes the decrease in ecological quality, but also leads to problems like economic loss, facilities harm and resource unbalance (Mann and Hanna, 2010).

Therefore, the evaluation and prevention of the effects of biological invasion has become a priority for researchers. In recent years, several papers have developed statistical methods to improve the understanding of specific biological invasions, such as aquatic biological invasion (Moffitt et al. 2012). In this study, researchers found that near and just below freezing water temperatures in localized reaches of the watershed were related to reduced populations or lack of detection. And distributions observed in winter were associated with regions of groundwater releases, or downstream of impoundments in the watershed. They speculate that the population has remained restricted because of thermal conditions. However, these relationships can be changed with watershed alterations or global climate change.

1.3 Importance of the Problem

Although the existence of imperfect sensitivity and specificity are widely known, the impacts of imperfect sensitivity and specificity on event prediction have rarely been studied. In addition, biological invasion may cause problems like unbalanced biodiversity, disordered ecosystems, economic loss, and facility harm. Therefore, it is necessary to study the influence of biological invasion when the sensitivity and specificity of the diagnostic tests used are imperfect. To study the influence of imperfect tests, we vary parameter prior distributions and examine those changes to prediction interval width using a Bayesian approach.

1.4 Objectives

The objectives of this paper are to: 1) describe statistical methods to estimate the probability of detecting biological invasion with imperfect diagnostic tests using covariate information. 2) under two scenarios, study how the prediction interval for event probabilities change when prior information changes.

1.5 Organization of the thesis

The remainder of this thesis is organized as follows. Chapter two provides a review of past efforts and application of Bayesian approach with imperfect tests. Chapter three provides the model and the methodology of Bayesian approach. Chapter four describes the data and source and the statistical model, which is used to implement the data analysis. Chapter five presents the results derived from the analysis. Chapter six provides concluding remarks.

CHAPTER 2. LITERATURE REVIEW

2.1 Bayesian Approach and Imperfect Tests

Sensitivity and specificity are two statistical measures of binary classification tests. Sensitivity measures the proportion of positives which are correctly identified; specificity identifies the proportion of negatives which are correctly measured. A perfect predictor is perfectly sensitive and specific (Ott and Longnecker, 2001). However, most tests are not perfectly sensitive and specific. In addition, problems like limited numbers of observation and different experimental procedures with imperfect sensitivity and specificity lead to further imprecision of studies (Williams and Moffitt, 2010).

A standard assumption of logistic regression is that the outcomes are measured properly, and the assumption of a perfect diagnostic tests is correct, so that sensitivity and specificity equal one. But outcomes from laboratory or field studies, such as invasive species, are usually imperfect.

To identify the impacts of imperfect sensitivity and specificity, some statistical methods and studies had been developed to analyze the imperfect tests for specific topics like wildlife disease prevalence. Williams and Moffitt (2010) conducted a study on estimating disease prevalence from imperfect diagnostic tests with data from different pool sizes by using a Bayesian approach to examine the posterior distribution of prevalence, sensitivity, and specificity. The results show that the estimates adjust for imperfect tests, which is more efficient than estimates assuming perfect tests.

The Bayesian approach depends on a subjective definition of probability compared to the approach relying on the frequency definition of probability.

$$f(\theta|x) \propto f(\theta)f(x|\theta) \tag{3}$$

The left-hand side of (3) is the posterior density $f(\theta|x)$, whereas on the right-hand side is the product of prior density $f(\theta)$ and the probability distribution for the data $f(x|\theta)$, conditional on the parameter θ.

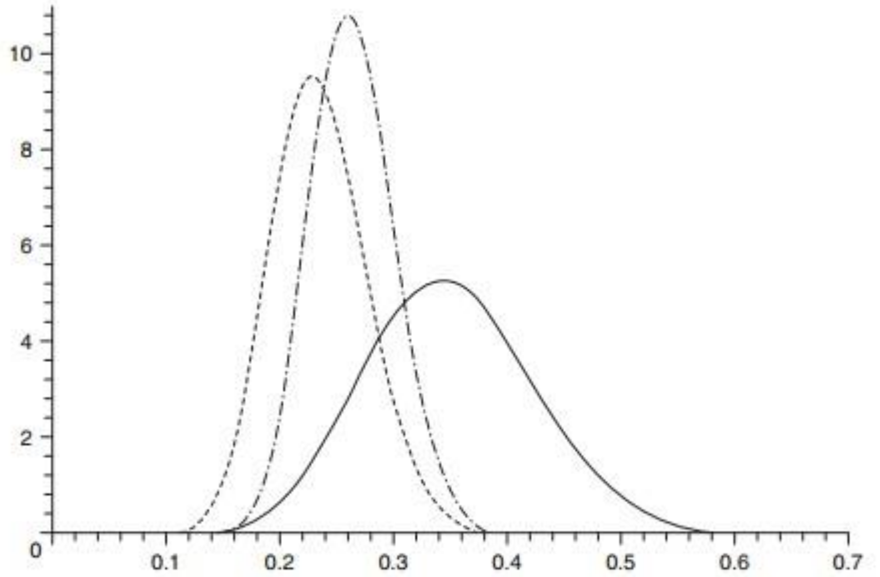

Figure 2. Explanation of Bayesian Approach

Figure 2 illustrates the use of Bayesian approach. The solid line is the prior distribution, the dashed line is the likelihood, and the posterior distribution is the dot-dashed line.

General Ideas of Bayesian Inference are: the posterior distribution combines the information in both the prior distribution and the likelihood; it is a compromise between the values that are supported by the prior and by the data separately; the posterior is generally more similar to, and is centered nearer to the center of, the stronger information source (Hagan, 1996).

Past studies tried to use external estimates of sensitivity and specificity into the likelihood for logistic regression. However, the limitation of this method is that it treated sensitivity and specificity as fixed. In order to fix this limitation, Tu, et al (1999) developed a Bayesian model to allow the uncertainty of test parameters. Focusing on application of simple Bayesian methods for binomial regression analysis of risk factor studies with imperfect outcome measurement, McInturff et al (2004) used WinBUGS and conditional means priors to allow for inclusion of prior data and expert opinion in the estimation of odds ratios, probabilities, risk ratios, risk differences, and diagnostic test sensitivity and specificity. The simple method of obtaining Bayes factors for link selection is presented. And the regression coefficient estimates are shown to change noticeably when prior and imperfect sensitivity and specificity are incorporated into the model. Even though studies and methods on imperfect test sensitivity and specificity have been devised, the impacts of imperfect testing on monitoring events have rarely been studied.

2.2 Invasive Species and Its Impact

Shimberloff (1996) gives the geography of invasive species in his study, to show that not all states are affected equally. Among the most affected states are Hawaii and Florida.

The impacts of invasive species include ecosystem modification, resource competition, aggression and its analogs, predation, herbivory (Simberloff 2010). He pointed out that the best thing to do with the invasive species is to keep them out to establish their own population, and another way is find and eradicate them to keep their population at low levels.

Meinesz (2003) pointed out that introduced species have invaded natural habitats to harm one or more native species, not to say the economic consequences of varying degree, including loss of recreation and tourism, such as invasions threaten biodiversity in those habitats. At first, the invasive species will maintain itself in a limited range of habitats without spreading and without upsetting the equilibrium of the ecosystem. At the second level, they spread to the detriment of one or a few natives, and it thus threatens native biodiversity. At the third level, the invasive species turns dominant and alters the entire ecosystem. At the fourth level, the invasive species affect several ecosystems, thus threatening an even larger swath of biodiversity.

Britton and Pegg (2010) pointed out that surveillance programs are developed to minimize the opportunities to develop from initial introductions via early detection. These methods are dependent on surveillance methods being able to detect species at low levels of abundance to avoid false- negative recordings through imperfect detection. The data indicates that small pest fishes such as *P. parva* may be inclined to imperfect detection when at low densities and this is consistent with other invasive species. In addition, Britton and Pegg indicated the importance of designing surveillance programs with known statistical power to control conservation resource expenditure and optimize management outcomes.

2.3 New Zealand Mudsnails

When it comes to the effects of New Zealand mudsnails, Evans and Buffalo (2011) found that the invasive New Zealand mudsnail has largely been studied as a consumer of periphyton, algae and diatoms. Kolosovich (2012) studied the short-time survival and potential grazing effects of the New Zealand mudsnail in the Truckee River and Lake Tahoe. As a result, snail survivorship in the Truckee River ranged from 50-85 percent across treatments and snail survivorship ranged from 5-40 percent in the Lake Tahoe. They suggested that the Truckee River is more vulnerable to establishment by New Zealand mudsnails than Lake Tahoe.

Bennett (2011) found that the snail densities fluctuated between very low in the winter to relatively high during spring and summer months. In addition, field and laboratory experiments allowed them to examine the effects of abiotic factors on New Zealand mudsnail and the impacts of New Zealand mudsnail on stream algal assemblages, invertebrates, and Western Toad tadpoles. They found that the New Zealand mudsnail reduces the growth and survival of native taxa by changing the abundance and composition of their shared food resource, benthic algae.

Moffitt and James (2011) investigated the distribution and seasonal dynamics of the snail population in the Silver Creek watershed in Idaho for two years, and they found that density of New Zealand mudsnail was highest during summer months, but the distribution was patchy. Furthermore, they found that near-to-below freezing winter water temperatures in the watershed increased the mortality and extirpation of colony.

Moffitt (2012) studied the invasion of New Zealand mudsnails, in the Silver Creek watershed in Idaho, related to the water temperature.   The results show that thermal conditions restricted the population.

Stockton and Moffitt(2011) pointed out that New Zealand mudsnails will flourish within and surrounding aquatic facilities due to their constant temperature and flow and enhanced nutritional resources. In addition, various ways can be used in and surrounding facilities to determine the risk of invasion, including hydrocyclone filtration, mixed-cell rearing units, depuration strategies, and barriers.

CHAPTER 3. METHODOLOGY

Here we introduce the Bayesian formulation of the logistic regression model, and discuss how we studied the width of certain prediction intervals as the prior information for sensitivity and specificity are varied.

3.1 The Logistic Regression Model

Applying the method from McInturff et al (2004), we use a dichotomous test to detect the risk of invasive species D. Let the variable z to indicate the truth, where $z = 1$ indicates D is present. Furthermore, let variable y denotes the test results, where $y = 1$ indicates that D is detected and $y = 0$ otherwise. They define sensitivity and specificity as $\eta = \Pr(y = 1|z = 1)$ and $\lambda = \Pr\left(y = 0|z = 0\right)$respectively. For the individual with covariate information x, a binomial regression model will be applied to model the probability that an individual risk, $p_{x_j}(\beta) \equiv \Pr\left(z = 1|x_j\right)$. We know,

$$p_{x_j}(\beta) = \frac{\exp(x_j^{'}\beta)}{(1+\exp(x_j^{'}\beta))} \tag{4}$$

where β is a vector of regression coefficients in the logistic regression. To be more general, $g(\pi_x) = x'\beta$ or $\pi_x = g^{-1}(x'\beta)$ for the monotone link function $g(\cdot)$ in the generalized linear model (GLM).

Thus, a positive test result for an individual with covariate information x has the probability

$$q_x = \Pr\left(y=1|x\right) = p_x(\beta)\eta + (1-p_x(\beta))(1-\lambda) \tag{5}$$

From the data, we observe (y_j, x_j) with $j = 1,2, \dots, n$, where y_j is the diagnostic test outcome and x_j denotes row vector of covariate information for the jth individual in independent sample of size n. To simplify, we assume $y_j|x_j \sim Bernoulli(q_j)$ with the probability of success from (4) and $p_{x_j}(\beta) \equiv \Pr\left(z_j = 1|x_j\right)$. For convenience, let X denote the usual $n \times k$ regression covariate matrix, and $Y = \{y_j; j = 1,2,\dots,n\}'$. Then the likelihood function is the product of probabilities given in (1),

$$\mathrm{L}\left(y,x|p_{x_j}(\beta),\eta,\lambda\right)= \prod_{j=1}^{n}\left[p_{x_j}(\beta)\eta+\left(1\text{-}p_{x_j}(\beta)\right)(1\text{-}\lambda)\right]^{y_j}\left[p_{x_j}(\beta)(1\text{-}\eta)+\left(1\text{-}p_{x_j}(\beta)\right)\lambda\right]^{1\text{-}y_j} \tag{6}$$

In our study, we assumed independent beta priors for sensitivity and specificity. Therefore, the joint posterior is

$$p(\beta,\eta,\lambda|x,y)\propto \mathrm{L}(x,y|\beta,\eta,\lambda)p(\beta)p(\eta)p(\lambda) \tag{7}$$

Where $p(\beta)$, $p(\eta)$, $p(\lambda)$ are the prior densities for parameters β, η, λ.

It is possible to use Gibbs sampling to sample from it. WinBUGS was designed to do Gibbs sampling. Spiegelhalter and Thomas (2003) gave more details about WinBUGS in the WinBUGS User Manual.

Gibbs sampling entails Monte Carlo sampling from each variable which has been modelled, conditional on all other variables and the data in some order. After a ‘burn-in’ phase, the samples will generally be from the joint posterior distribution.

Convergence diagnostics are available in WinBUGS. It is standard to simply monitor and to look for white noise and no trend in those plots as evidence of convergence.

To better understand the impacts from the imperfect tests, we studied prediction interval widths in this study. To study the prediction interval width of our model with changes of the beta distribution priors for η and λ, an efficient way to vary the range of priors on the prediction interval is the Central Composite Design.

3.2 Central Composite Design (CCD)

In a given model with “n” parameters, there are three types of points in the CCD:

1. Axial points: the design will have 2*n axial points.
2. Factorial or Cube points: which contain 2^n cube points from a full factorial design.

3. Center points: there are usually multiple points, and two central points are applied in our study.

The pictorial representation of the central composite design is shown in Figure 3 (Kuehl, 1999).

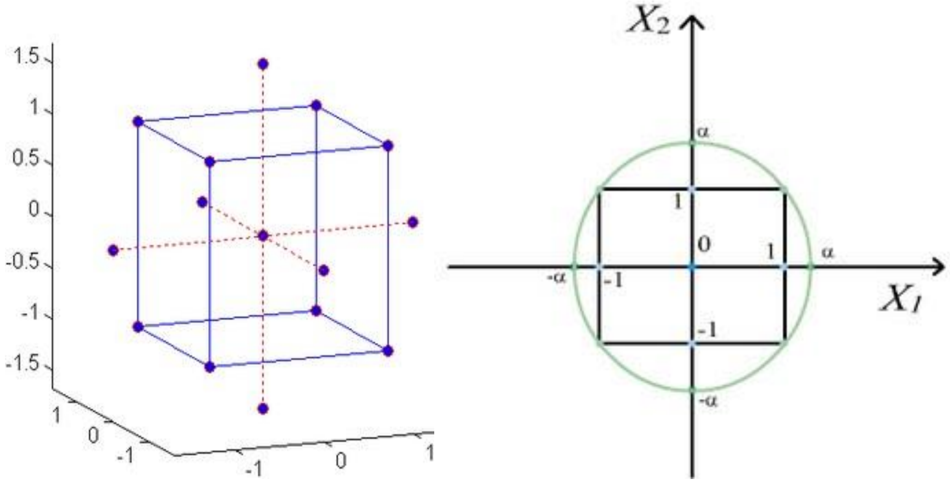

Figure 3. Central Composite Design for Two Factors and Three Factors

In our model, we have four parameters, the mean of the prior distribution of sensitivity, the mean of the prior distribution of specificity, the weight of the prior for sensitivity, and the weight of the prior for specificity. In our study, n=4, thus we have 2*n=2*4= 8 axial points, $2^n = 2^4 = 16$ factorial points, and two center points.

Regarding to our data set, the initial prior distributions, and weight for sensitivity and specificity are shown in Table 1 below:

	Factorial	Center	Axial
Mean Se	0.85±0.07	0.85	0.85±0.14
Mean Sp	0.85±0.07	0.85	0.85±0.14
Se Weight	70±30	70	70±60
Sp Weight	70±30	70	70±60

Table 1. Initial Priors of Data Set

Given the weights and the initial priors listed above, we obtain the values of parameters of sensitivity and specificity, and plug them into the likelihood function of

WinBUGS to get the prediction interval width. We can get α and β by the following two equations: mean value of se/sp= $\frac{\alpha}{\alpha+\beta}$, and weight of se/sp= $\alpha+\beta$.

Given the data from a central composite design, a second order regression model can be built to understand the effect of model parameters. A response surface of a regression model helps us understand the model.

The response surface lets the researcher visually study the response in a region of interesting factor levels and to estimate the effect of the treatment factors on the response variable. The response surfaces can also be applied to investigate the interaction of factors on the response variable, for example, the interplay between sensitivity and specificity prior distributions. The response can be represented graphically, either in the three-dimensional space or as contour plots that help visualize the shape of the response surface. Contours are curves of constant response drawn in the x, y plane keeping all other variables fixed. Each contour corresponds to a particular height of the response surface. A central composite design (CCD) is the most commonly used response surface methodology (RSM) design, and RSM is often used for the study of quadratic effects of factors.

3.3 Data Source

The data in our study contains 400 simulated observations, and we are interested in the low temperature (45 degrees Fahrenheit) and high temperature (75 degrees Fahrenheit) since these two degrees are the two extremes of the living temperature.

CHAPTER 4. DATA ANALYSIS AND RESULTS

In this study, we used the logistic regression model in OpenBUGS. Details of our OpenBUGS program code are provided in the appendix.

To better compare and recognize the risk of an event with imperfect test measurements, the first step of our study was to run the logistic regression to obtain the maximum likelihood estimates model for our simulated invasive species data. Also, OpenBUGS provided estimates for logistic model by assuming perfect tests. To capture the impacts of imperfect sensitivity and specificity, models with imperfect tests will be run.

To show the logistic relationship between temperature and the existence of New Zealand mudsnails, we use Figure 4. From the graph, we can find that as the temperature increases, the existence of New Zealand mudsnails decreases.

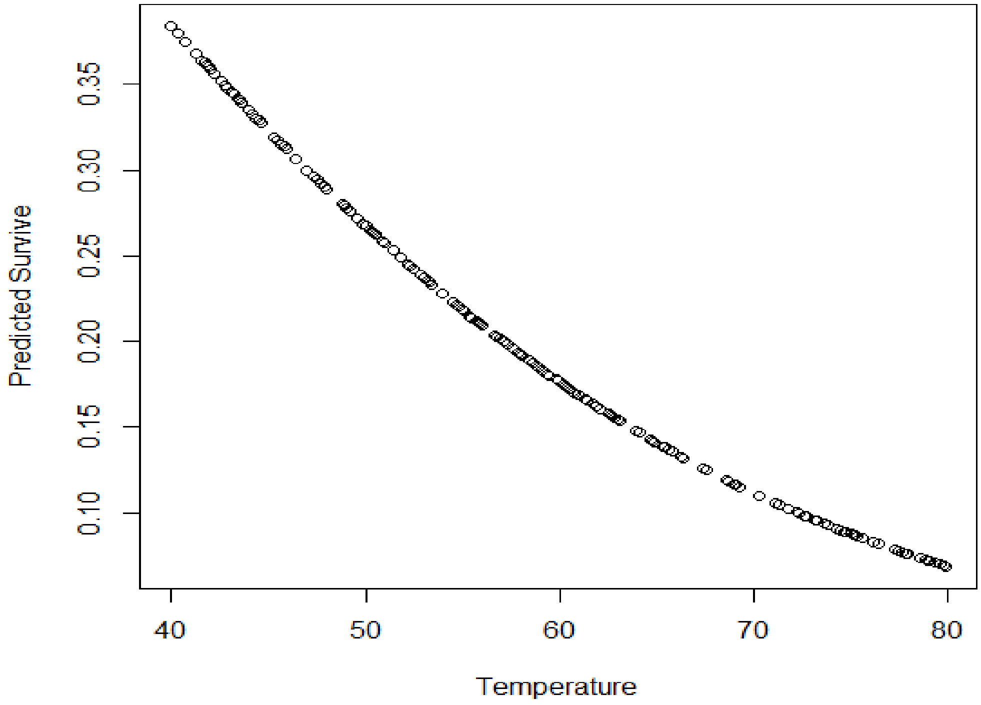

Figure 4. Logistic Regression for Simulated New Zealand Mudsnails Data

Based on the logistic regression from R, we have $\log\frac{p(x_i)}{1-p(x_i)}=1.6671-0.0535x_i$, from which we have $p(45)=0.32$, and $p(75)=0.088$.

The fitted logistic model based on the result from OpenBugs is

$$\log\frac{p(x_i)}{1-p(x_i)}=-1.371-0.05331\left(x_i-\bar{x}\right) \quad (8)$$

Based on the equation, we have $p(45)=0.99$, and $p(75)=0.12$.

After we extended Table 1 to 26 combinations of parameters of sensitivity and specificity, we used these values with OpenBugs to get distributions of predicted values under low (45 degrees Fahrenheit) and high (75 degrees Fahrenheit) temperatures, and we selected the 2.5% and 97.5% percentiles. The difference between the 2.5% and 97.5% percentiles gives the value of the prediction interval width under the low or the high temperature degrees. The parameters, and the corresponding prediction interval width are listed in the Table 2 below.

To understand the impacts of changes on the priors on prediction interval width, we fitted a best multiple linear regression model based on the values of mean sensitivity, mean specificity, their weights, and the prediction interval widths shown in Table 2. The general regression models under 45 and 75 degrees Fahrenheit were

$$\begin{aligned}\text{width} = {} & b_0+b_1\text{mean\_se}+b_2\text{mean\_sp}+b_3\text{se\_weight}+b_4\text{sp\_weight}\\ & +b_5\text{meanse}^2+b_6\text{meansp}^2+b_7\text{seweight}^2+b_8\text{spweight}^2\\ & +b_9\text{mean\_se}*\text{mean\_sp}+b_{10}\text{mean\_se}*\text{seweight}\\ & +b_{11}\text{mean\_se}*\text{spweight}+b_{12}\text{mean\_sp}*\text{seweight}\\ & +b_{13}\text{mean\_sp}*\text{spweight}+b_{14}\text{seweight}*\text{spweight}+e\end{aligned} \quad (9)$$

In our study, we used best subset model selection to find the best fitted model. We selected the best multiple linear regression by SAS. The results are as follows:

Under 45 degrees Fahrenheit, the best model is

$$\begin{aligned}w_{45}={} & -28.34+23.3\text{mean\_se}+48.97\text{mean\_sp}+0.005086\text{se\_weight}\\ & -15.77\text{meanse}^2-28.88\text{meansp}^2-0.000043\text{seweight}^2\end{aligned}$$

$$\textit{with}\ R^2=0.9216. \qquad (10)$$

Under 75 degrees Fahrenheit, we did a Box-Cox transformation to the prediction interval width with a natural logarithm, and then the model is

$$\begin{aligned} \mathrm{lnw}_{75} =& -231.7+360\mathrm{mean\_se}+109.3\mathrm{mean\_sp}+0.3649\mathrm{se\_weight} \\ & -0.2078\mathrm{sp\_weight}-116.8\mathrm{meanse}^2-14.84\mathrm{meansp}^2 \\ & -0.0009443\mathrm{seweight}^2-0.01442\mathrm{meanse}*\mathrm{meansp} \\ & -0.02404\mathrm{meanse}*\mathrm{se\_weight}+0.2343\mathrm{meansp}*\mathrm{sp\_weight} \end{aligned}$$

$$\textit{with}\ R^2=0.9396. \qquad (11)$$

Mean se	Mean sp	Se weight	Sp weight	Width 45	Width75
0.85	0.85	70	70	0.99470	0.00690
0.85	0.85	70	70	0.99470	0.00690
0.78	0.78	40	40	1.00000	0.00004
0.78	0.78	100	40	1.00000	0.00029
0.78	0.78	40	100	1.00000	0.00000
0.78	0.78	100	100	1.00000	0.00006
0.78	0.92	40	40	1.00000	0.00531
0.78	0.92	100	40	1.00000	0.01431
0.78	0.92	40	100	0.99974	0.00747
0.78	0.92	100	100	0.99978	0.03081
0.92	0.78	40	40	0.42700	0.00534
0.92	0.78	40	100	0.38040	0.00184
0.92	0.78	100	40	0.40450	0.00469
0.92	0.78	100	100	0.36350	0.00135
0.92	0.92	40	40	0.44220	0.05653
0.92	0.92	100	40	0.44717	0.04924
0.92	0.92	40	100	0.45231	0.05161
0.92	0.92	100	100	0.40147	0.06312
0.99	0.85	70	70	0.39790	0.01794
0.71	0.85	70	70	1.00000	0.00002
0.85	0.99	70	70	0.34660	0.19958
0.85	0.71	70	70	0.53730	0.00038
0.85	0.85	10	70	1.00000	0.00001
0.85	0.85	130	70	0.70640	0.00587
0.85	0.85	70	10	0.98700	0.01077
0.85	0.85	70	130	0.99810	0.00683

Table 2. Combinations of Sensitivity, Specificity, Prediction Interval and Their Weights

From the above table, we see that at 45 degrees the prediction interval width decreased significantly when the mean value of sensitivity increases from 0.85 to 0.99, where the prediction interval width decreases to 0.01794. Furthermore, the prediction interval width under 45 degrees Fahrenheit with the changes in the mean value of specificity fluctuated, so there is less evidence to show that the priors of specificity will affect the prediction interval. In addition, we cannot find similar changes with respect to the increase in the mean value of specificity. The increases in weights for sensitivity did not reduce the prediction interval width under 45 degrees Fahrenheit, and the weights of specificity were not included in our model based on (6). Thus, changes in the priors of sensitivity help to decrease the prediction interval width, but the effects of mean specificity and the weight of sensitivity was not clear.

In addition, under 75 degrees Fahrenheit, it is hard to detect the influence from either sensitivity or the specificity or their weights.

To visually examine how the changes of sensitivity and specificity affect the prediction width, we use values from Table 1 to graph the mean value of sensitivity, specificity or the weights with respect to the widths. We also used contour plots. The two variables (X and Y) in the contour plots are for mean value of sensitivity and specificity. The third variable (Z) is the predicted value in our model with changes of sensitivity and specificity. And the contour plot helps to answer the following question: how does the variable Z change as the function of the X and Y. It helps us to understand how our predicted interval width changes as the changes of mean sensitivity and mean specificity.

4.1 Model for x =45 Degrees Fahrenheit

At 45 degrees Fahrenheit, we examined how prediction interval width changes with mean values of sensitivity or specificity. And from (10), we know the significant terms are mean sensitivity, mean specificity, their weights, the second order terms of sensitivity, specificity, and the weight of sensitivity. Figure 5 shows the changes of prediction interval width under the changes of mean value of sensitivity.

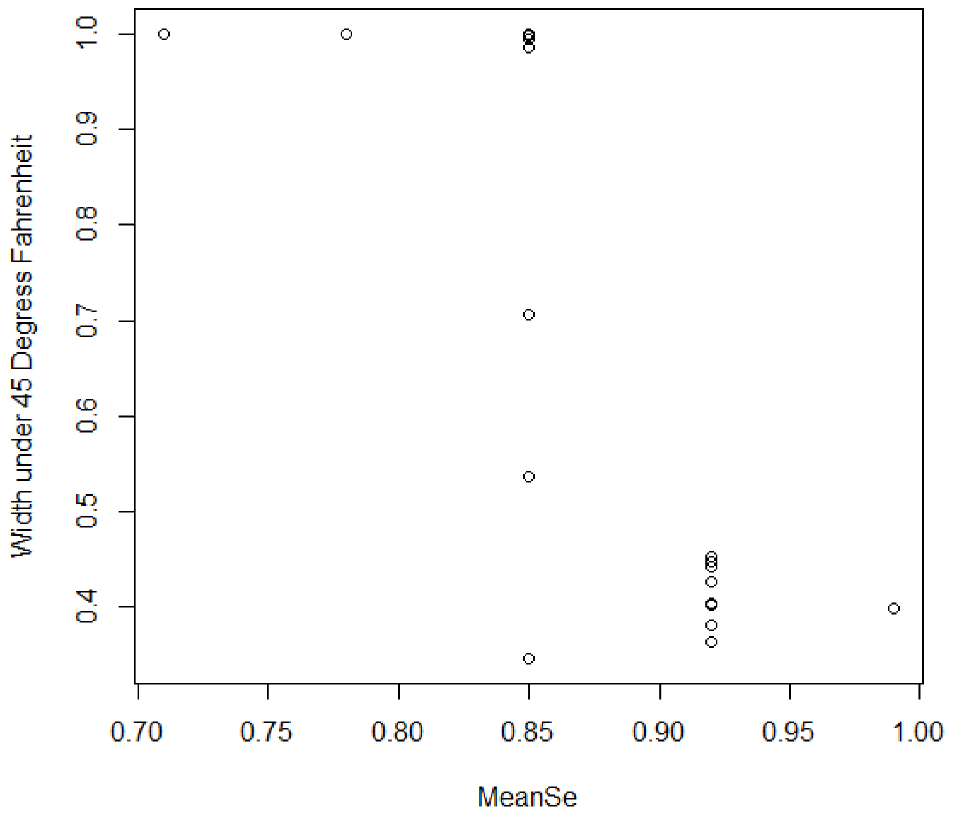

Figure 5. Plot of Prediction Interval Width under the Change of Mean value of sensitivity at 45 Fahrenheit

From the above graph, we can see that the prediction interval width decreases as the mean value of sensitivity increases to 1, which means the prediction width tends to be smaller as the mean value of sensitivity converges to 1, the perfect sensitivity.

Furthermore, we cannot detect the same pattern when the independent variables are mean value of specificity or the weight of the sensitivity prior distribution. Figure 6 shows how the width changes with mean specificity.

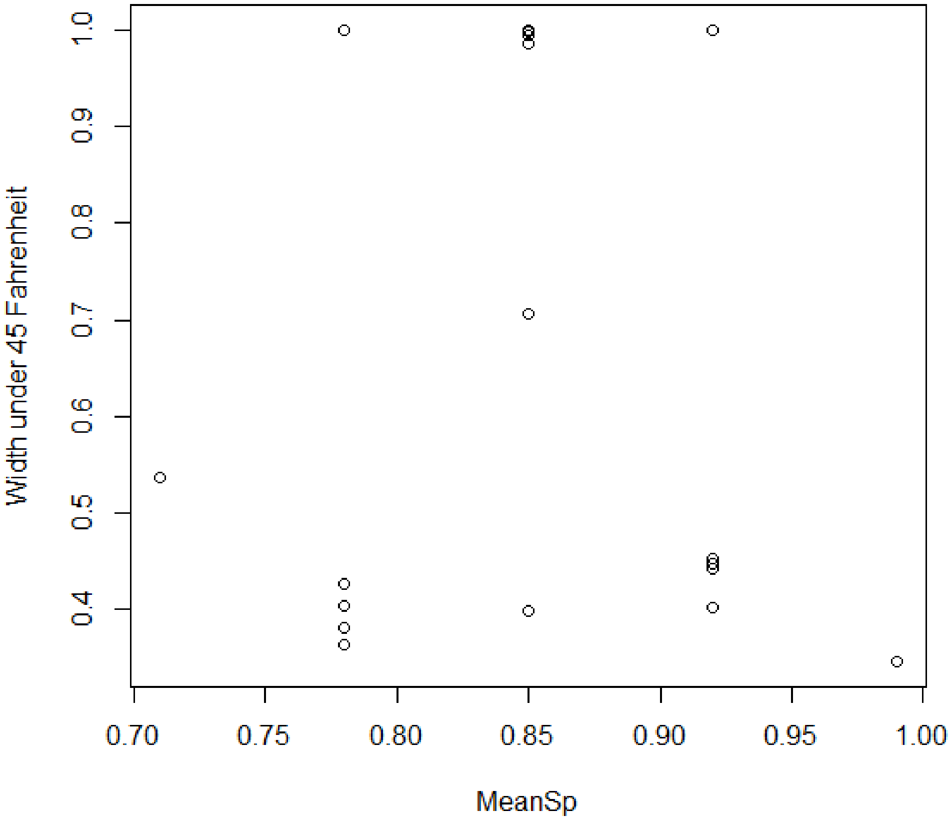

Figure 6. Plot of Width under the Changes of Mean Value of Specificity at 45 Fahrenheit

In the above plot, when the mean value of specificity ranges from 0.75 to 0.95, we cannot find a decreasing trend for the width. In addition, when mean value of specificity is low as 0.7, the width is not high, but between 0.75 and 0.95, the width is often higher. These results show there is less clear trend of the width of prediction under the changes in mean specificity. It may be caused by the interaction of independent variables since the fitted model in (10) has second order terms.

In addition, a similar pattern is seen for the width as the sensitivity weight increases in Figure 7 below, which may also be affected by the interaction between the independent variables.

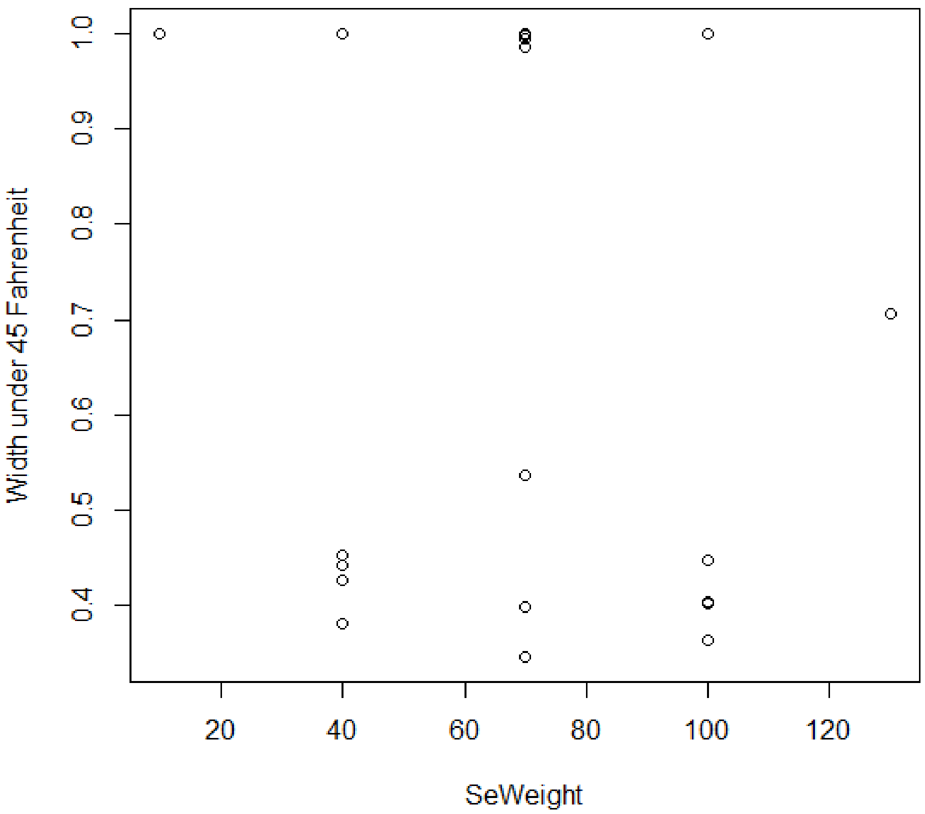

Figure 7. Plot of Width under the Changes of Specificity Weight at 45 Fahrenheit

To better understand the effects of sensitivity and specificity, we used contour plots in our study with combinations of sensitivity, specificity, and their weights as shown in Table 2.

Figure 8 shows how the width changes with mean sensitivity and mean specificity when the sensitivity weight equals 40.

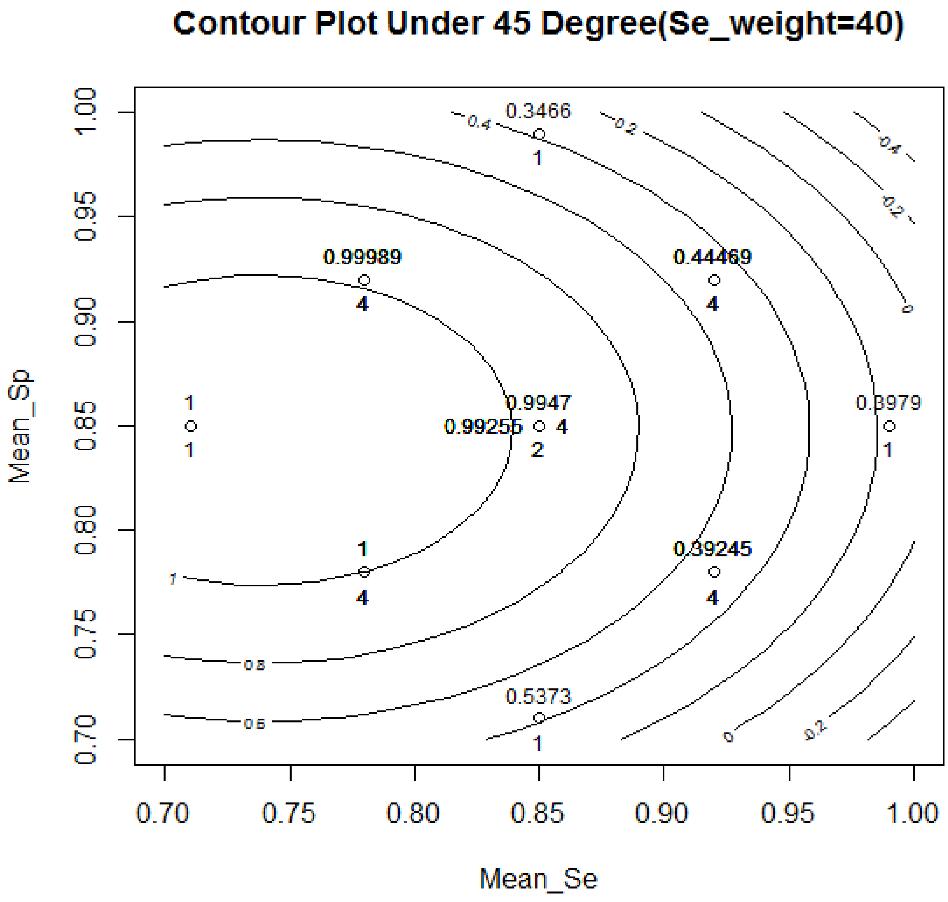

Figure 8. Contour Plot at 45 Degree with Sensitivity Weight =40

We see that it is a mound shaped distribution. The increase in the mean value of sensitivity decreases the prediction interval width under the setting where sensitivity weight equals 40. Thus, we see how the sensitivity has an impact on the prediction width.

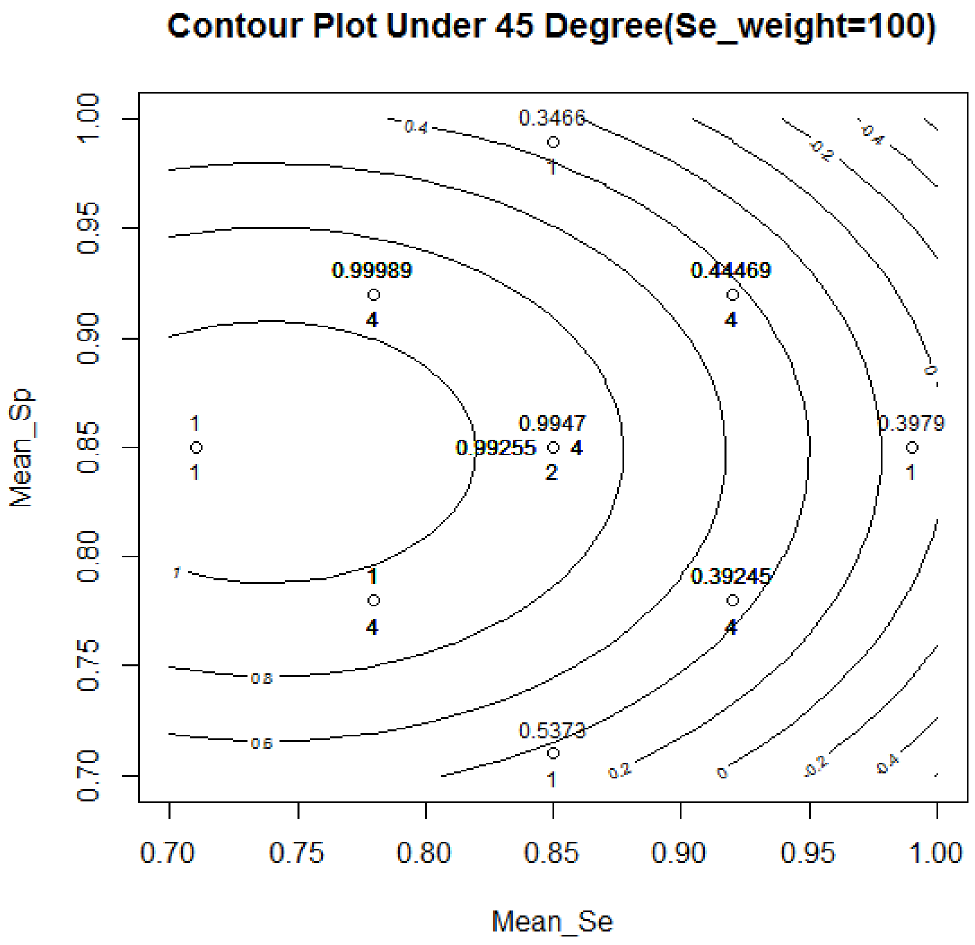

Figure 9. Contour Plot at 45 Degree with Sensitivity Weight =100

In Figure 9, we see that the increases in the mean value of sensitivity will also decrease the prediction interval width under the condition that sensitivity weight equals 100. For example, when the sensitivity increases from 0.78 to 0.92, the prediction width decreases from 1 to 0.38. Thus, when specific sensitivity weight equals 100, we can see the influence of mean sensitivity to the prediction width.

4.2 Model for x =75 Degrees Fahrenheit

At 75 degrees Fahrenheit from (11), we know that mean sensitivity, mean specificity, sensitivity weight, and specificity weight are all significant. In the following plots, we are going to study the relationship of these variables and the prediction interval width.

Figure 10 is a plot of width with respect to the mean sensitivity, and at the low sensitivity, such as 0.72, the width was the lowest, but when the mean value of sensitivity goes to higher values such as 0.85, the width fluctuates in a low level until the mean sensitivity is 0.95. When the mean value of sensitivity gets near 1, the prediction interval width does not change much.

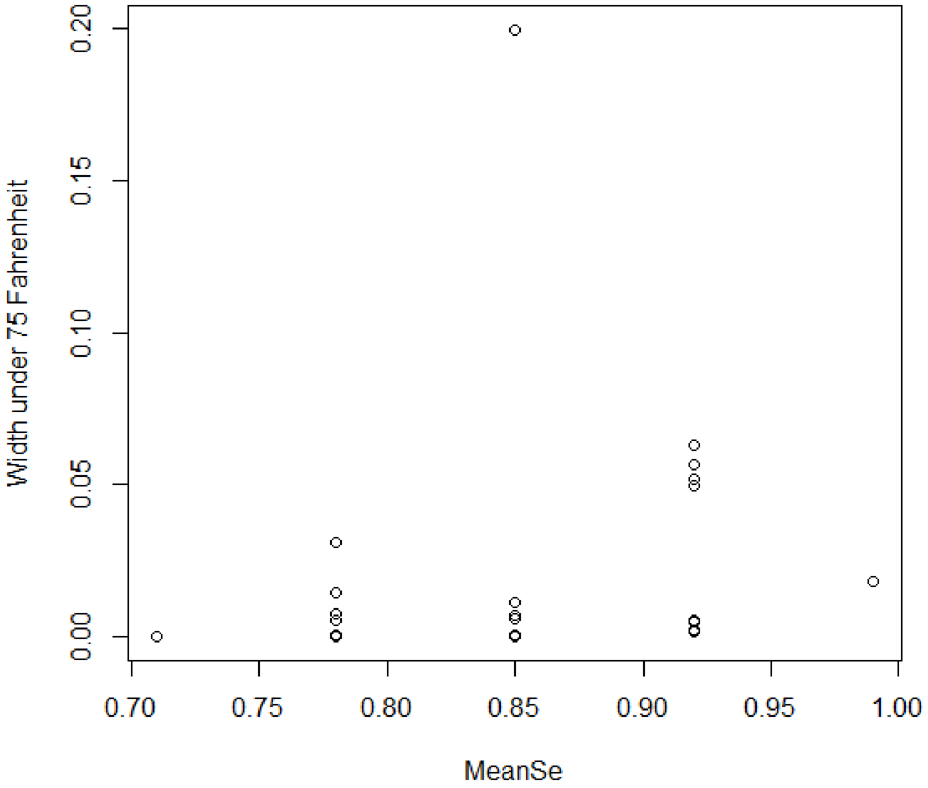

Figure 10. Plot of Prediction Interval Width under the Changes of Mean Sensitivity at 75 Fahrenheit

Figure 11 is a plot of prediction interval width with respect to the mean specificity, and at low sensitivity, such as 0.72, the width was the lowest, but the prediction width increases as specificity increases, which is different from our expectation that the width will decrease as the specificity goes to 1.

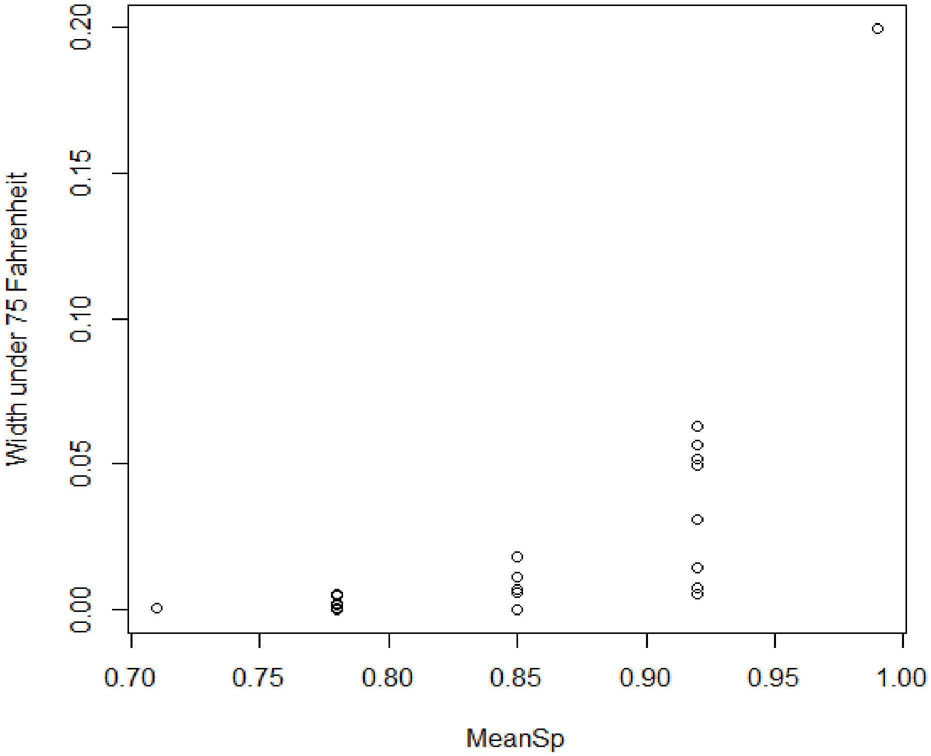

Figure 11. Plot of Prediction Interval Width under the Changes of Mean Specificity at 75 Fahrenheit

In Figures 12 and 13, the plots of prediction interval widths versus prior weights of sensitivity and specificity, and no clear trend in prediction interval width is seen as the weights of sensitivity and specificity increase. When the weights equal 40 or 100, we see the widths are all at or below 0.05, thus we cannot see obvious decreasing trend of the widths. The reason for the nondecreasing trend may also be caused by the interaction between independent variables.

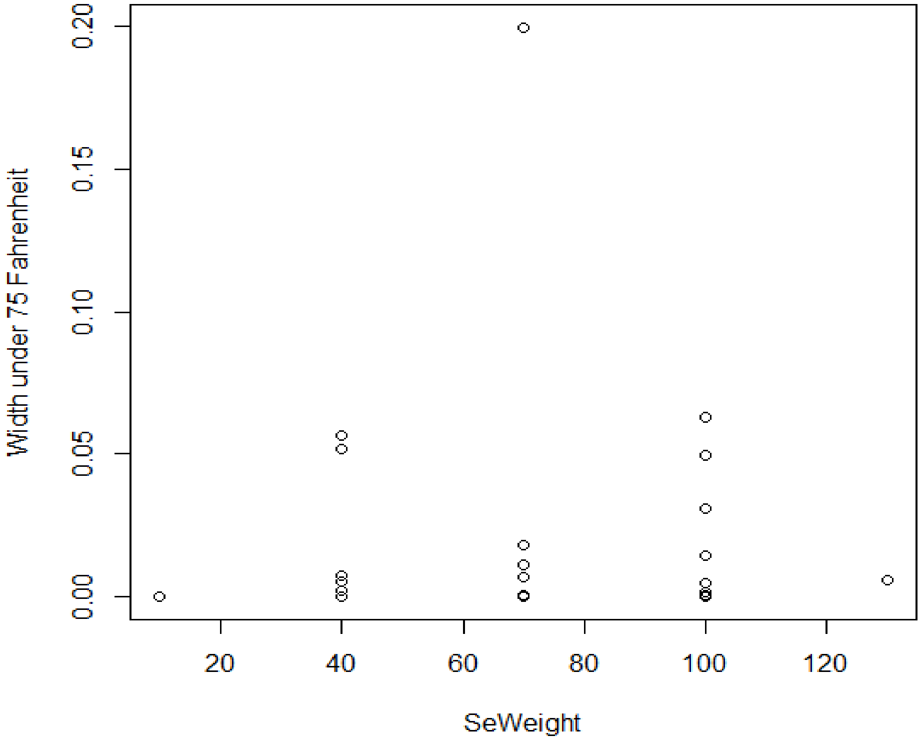

Figure 12. Plot of Prediction Interval Width under the Changes of Sensitivity Weight at 75 Fahrenheit

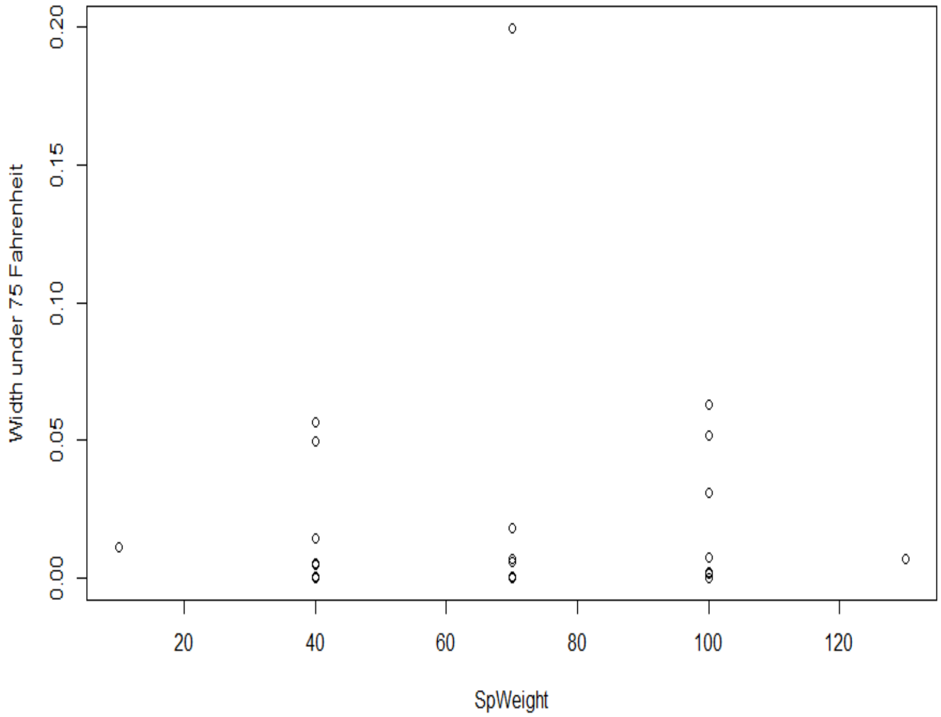

Figure 13. Plot of Prediction Interval Width under the Changes of Specificity Weight at 75 Fahrenheit

We now study contour plots at the high temperature, with different weights (40 or 100) of sensitivity and specificity :

With the weights of sensitivity and specificity are equal 40, we detect a mound shape of the width and also the curvature of the peak towards the left. For instance, when the mean value of specificity increases from 0.78 to 0.92, the width increases from 0.0004 to 0.00531. When the mean value of sensitivity increases from 0.78 to 0.92, the width increases from 0.0004 to 0.00534. In addition, when both of the sensitivity and specificity increase to 0.92, the prediction interval width also increased to 0.057, but the peak of the mound towards left as shown in Figure 14. The curvature may be caused mean sensitivity, mean specificity, and the weight of the prior distributions for sensitivity and specificity.

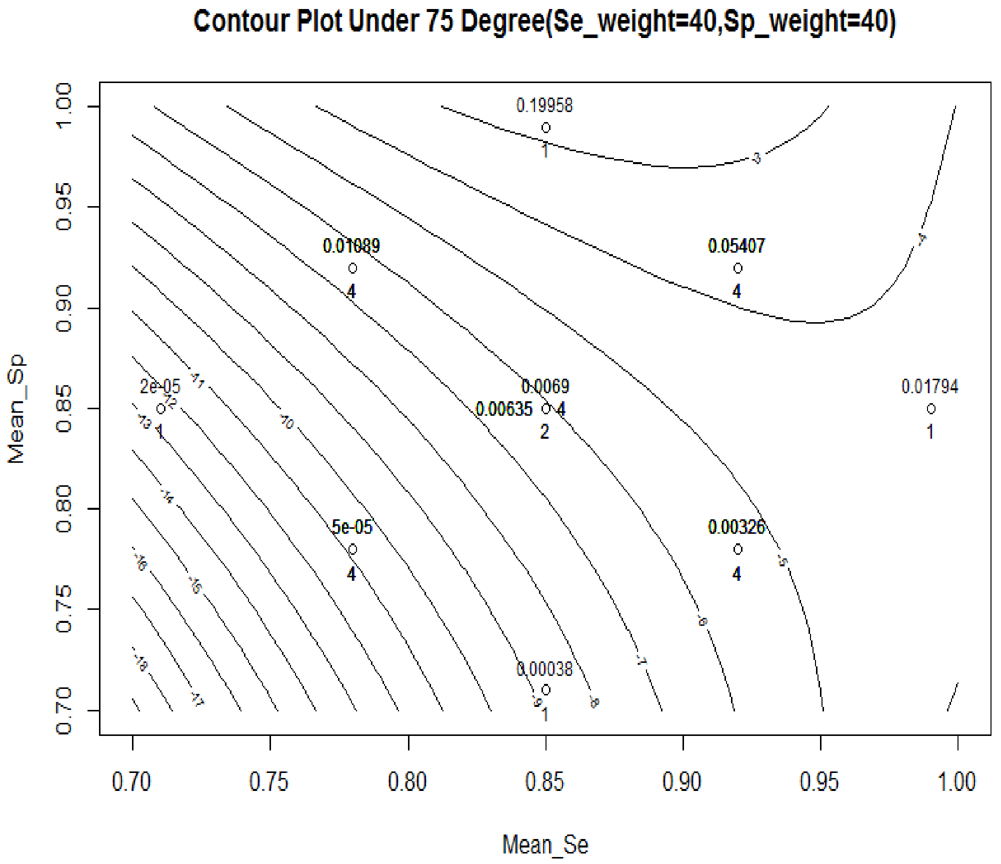

Figure 14. Contour Plot at 75 Degree with Sensitivity Weight =40, Specificity Weight=40

In Figure 15, it is seen that when the sensitivity and specificity increase, the prediction interval width increase and then decrease, and it also shows a curvature towards left at the right part of the graph of the width. For instance, when the mean value of specificity increases from 0.78 to 0.92, the width increases from 0.0000 to 0.00747. When the mean mean value of sensitivity increases from 0.78 to 0.92, the width increases from 0.0004 to 0.00184. In addition, when both of the sensitivity and specificity increased to 0.92, the prediction width also increased to 0.05661. Thus, the influence of sensitivity and specificity on the prediction interval width can be detected as shown in Figure 15.

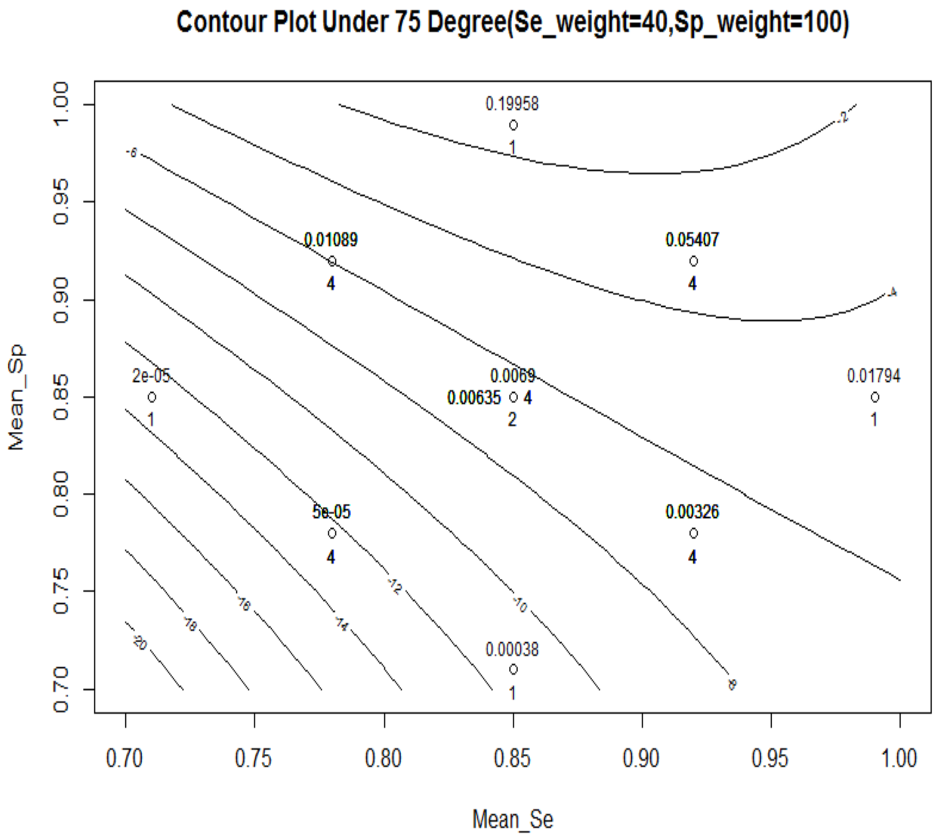

Figure 15. Contour Plot at 75 Degree with Sensitivity Weight =40, Specificity Weight=100

From Figure 16, we also see that when both of the sensitivity and specificity increase, the prediction interval width increases and then decreases, and it also shows curvature at the top of the plot. Also, the curvature may also be the result of second order terms in (11) and their interactions.

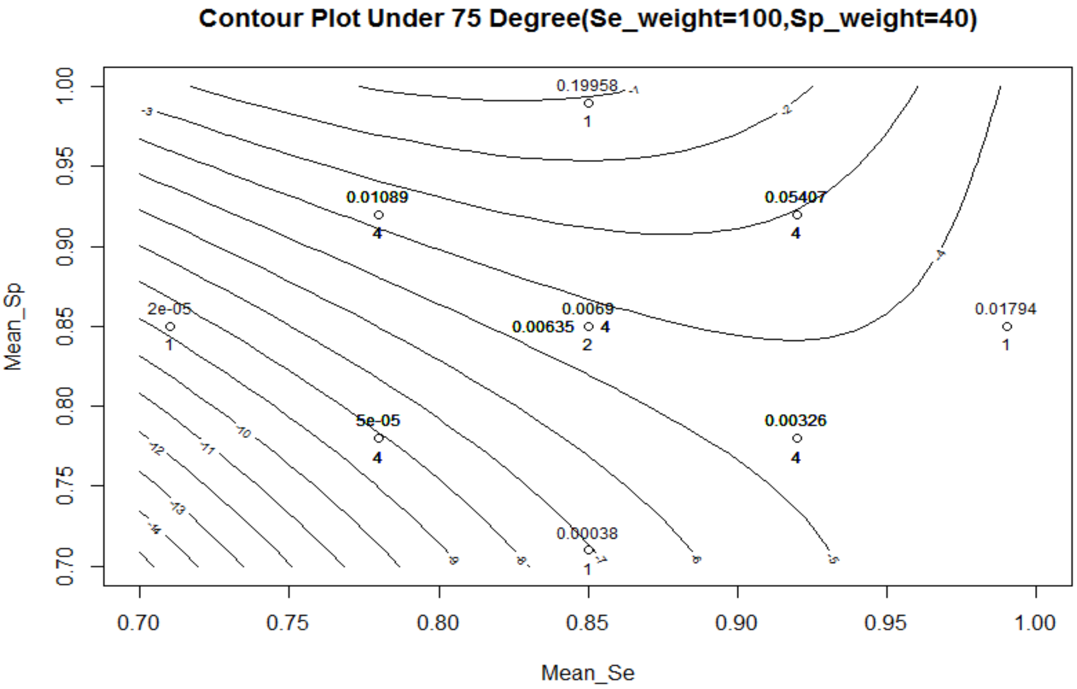

Figure 16. Contour Plot at 75 Degree with Sensitivity Weight =100, Specificity Weight=40

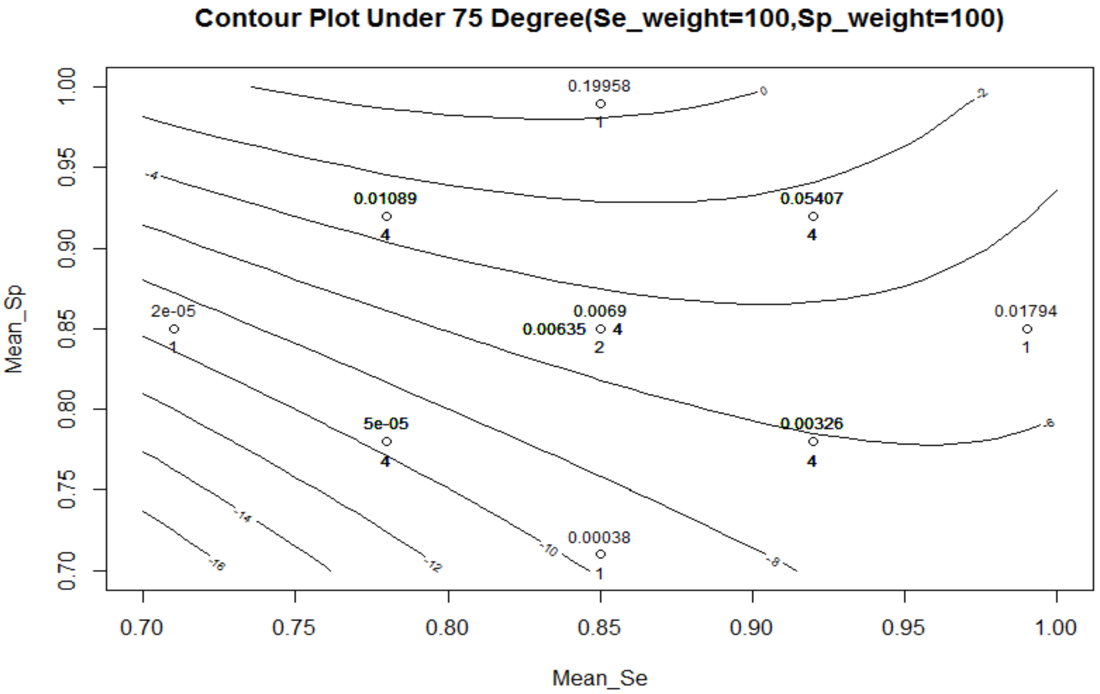

Figure 17. Contour Plot at 75 Degree with Sensitivity Weight =100, Specificity Weight=100

In Figure 17, it is seen that when the sensitivity and specificity increase, the prediction width increase and then decrease with a curvature leading to a peak at the top. For instance, when the mean value of specificity increases from 0.78 to 0.92, the width increases from 0.0006 to 0.03081. When the mean value of sensitivity increases from 0.78 to 0.92, the width increases from 0.0004 to 0.00135. In addition, when both of the sensitivity and specificity increase to 0.92, the prediction width also increases to 0.06312.

CHAPTER 5. CONCLUSIONS

Logistic regression is a common statistical method used in studying an event. The assumption of a perfect test is that sensitivity and specificity are perfect. Unfortunately, perfect sensitivity and specificity is typically unrealistic since such occurrence is rare in nature. Many reasons will cause the imperfect tests which include the imperfect sensitivity and specificity, such as difference in designing the experiment or the difference in the number of observations. To better capture the effects of imperfect tests on monitoring the invasive species, we applied a Bayesian approach, multiple linear regression, the CCD, and the contour plots in our study.

In this study, we have found that there is an association between the prevalence posterior distribution and the prior distributions of the sensitivity and specificity. From the fitted models and the plots for the prediction interval width with changes in priors, the results of our study have shown that more certainty will be detected in the prevalence prediction interval width as the mean value of sensitivity increases to 1 or 100% which means the perfect tests. However, there is no evidence to show that the same effects happened on the prediction interval width regarding to the changes in the mean value of specificity. In addition, we have also conducted the tests on changes the weights of our prior distributions. But the results showed that the bigger the weights of sensitivity and specificity, the smaller were the prevalence interval width.

To better indicate how the changes in the prior distribution on the prediction interval, we conducted a central composite design and also examined contour plots with different combinations of priors of mean sensitivity and mean specificity and their weights. From the contour plots, we showed the results from our predicted model and the plots for prediction interval width with respect to the changes in the priors. In all of these results, the increases in the mean value of sensitivity lead to the decreases in the prediction interval width.

However, changes in the mean value of specificity and weights of sensitivity and specificity do not clearly reduce the value of the prediction interval width.

To better explain the results, we consider the interaction between the independent variables which may affect the effects of specificity and their weights on the prediction interval width, which also caused the curvature in the contour plot.

At 45 degrees Fahrenheit, without the limitation on the dependent variable (existence of mudsnails), the contour plots better show how the changes of sensitivity and specificity on the prediction interval widths under different widths.

However, at 75 degrees Fahrenheit, the contour plot was curved towards left at the peak since the limit on the value of dependent variable caused by the high temperature. Thus, the contour plots at 45 degrees Fahrenheit is more effective and easier to explain the changes of prediction interval widths caused by the changes in the mean sensitivity, mean specificity and their weights. However, the contour plot curvature at 75 degrees Fahrenheit leads to more difficulty in explaining the changes of mean sensitivity, mean specificity, and their weights.

REFERENCES

Bedrick, E.J., R. Christensen, and W.O. Johnson. “A New Perspective on Priors for Generalized Linear Models,” *Journal of the American Statistical Association*, 91:1450–1460, 1996.

Britton, J., J. Pegg, and R. E. Gozlan, “Quantifying imperfect detection in an invasive pest fish and the implications for conservation management”, *Biological Conservatio*n, 144:2177-2181, 2011.

Elliott, M. “Biological Pollutants and Biological Pollution– an Increasing Cause for Concern.” *Marine Pollution Bulletin*, 46: 275–280, 2003.

Evans, M. A. and V. Buffalo, New Zealand Mudsnails as Leaf Litter Decomposers. Internet site: http://www.webpages.uidaho.edu/nzms/word%20and%20pdf%20files/PDF%20Presentations/Evans_NZMS%20Conference_March%202011.pdf

Hagan, A. O. The Bayesian Approach to Statistics. Internet site: http://www.sagepub.com/upm-data/18550_Chapter6.pdf

Independent Economic Analysis Board. Economic Risk Associated with the Potential Establishment of Zebra and Quagga Mussels in the Columbia River Basin. Document IEAB 2013-2, 2013.

Kolosovich, A. S. “Short-term Survival and Potential Grazing Effects of the New Zealand Mudsnail in an Uninvaded Western Great Basin Watershed,” *Aquatic Invasions*, 7:203–209, 2012.

Kuehl, R. O. Design of Experiments: Statistical Principles of Research Design and Analysis. CA: Brooks/Cole, 1999.

Meinesz, A. The Impact of Invasive Species. Internet site: http://www.pbs.org/wgbh/nova/nature/impact-invasive-species.html

McKinney, M.L. and J.L. Lockwood. “Biotic Homogenization: a Few Winners Replacing Many Losers in the Next Mass Extinction,” *Trends in Ecology and Evolution*, 14: 450–453,1999.

McInturff, P. and W. O. Johnson. “Modelling Risk when Binary Outcomes are Subject to Error,” *Statistics in Medicine,* 23:1095-1109, 2004.

Moffitt, C. M. and C. A. James. Thermal Limits to Range Expansion of NZMS in A Highly Used Recreational Drainage in the Intermountain west. Internet site: http://www.webpages.uidaho.edu/nzms/word%20and%20pdf%20files/PDF%20Presentations/Moffitt%20Presentation.pdf

Moffitt, C and C. James, “Dynamics of Potamopyrgus antipodarum Infestations and Seasonal Water Temperatures in a Heavily Used Recreational Watershed in Intermountain North America,”*Aquatic Invasions*, Volume 7, Issue 2: 193–202, 2012.

Ott, R. L, and Longnecker, M, T. An Introduction to Statistical Methods and Data. Cengage Learning, 2001.

Simberloff, D. and M. Rejmanek. Encyclopedia of Biological Invasions, CA: University of California Press,2011.

Simberloff, D. and J.L. Martin. “Impacts of Biological Invasions:What’s What and the Way Forward,” *Trends in Ecology and Evolution*, 28(1):58-66, 2013.

Simberloff, D . Impacts of Introduced Species in the United States. Internet site: http://www.gcrio.org/CONSEQUENCES/vol2no2/article2.html

Sodhi and Ehrlich, Conservation Biology for All. Oxford: University Press, 2010.

Spiegelhalter, D. and A. Thomas. WinBUGS User Manual. Internet site： http://www.politicalbubbles.org/bayes_beach/manual14.pdf

Tu X.M., J. Kowalski, and J. Jia. “Bayesian analysis of prevalence with covariates using simulation-based techniques: applications to HIV screening,” *Statistics in Medicine*,18:3059–3073, 1999.

Williams, C.J., and Moffitt, C.M. “Estimation of fish and wildlife disease prevalence from imperfect diagnostic tests on pooled samples with varying pool sizes,” *Ecological Informatics*, 5: 273-280, 2010.

APPENDIX A

WinBUGS Code

```
model
    {
        for (i in 1:200)
           {
             y[i] ~ dbern(q[i])
             q[i] <- Se*(Pi[i]) + (1-Sp)*(1-Pi[i])
             logit(Pi[i]) <- alpha + beta*(x[i] - mean(x[]))
             phat[i] <- exp(alpha + beta*(x[i] - mean(x[])))/(1 + exp(alpha + beta*(x[i] -
mean(x[]))))
           }
        Se ~ dbeta(99, 1)
        Sp ~ dbeta(99, 1)
        alpha ~ dnorm(0, 0.0001)
        beta ~ dnorm(0, 0.0001)
phat40 <-    exp(alpha + beta*(40- mean(x[])))/(1 + exp(alpha + beta*(40- mean(x[]))))
phat45<-    exp(alpha + beta*(45 - mean(x[])))/(1 + exp(alpha + beta*(45- mean(x[]))))
phat50<-    exp(alpha + beta*(50- mean(x[])))/(1 + exp(alpha + beta*(50- mean(x[]))))
phat55 <-    exp(alpha + beta*(55 - mean(x[])))/(1 + exp(alpha + beta*(55- mean(x[]))))
phat60    <- exp(alpha + beta*(60 - mean(x[])))/(1 + exp(alpha + beta*(60- mean(x[]))))
phat65    <-    exp(alpha + beta*(65 - mean(x[])))/(1 + exp(alpha + beta*(65 - mean(x[]))))
phat70    <-    exp(alpha + beta*(70 - mean(x[])))/(1 + exp(alpha + beta*(70 - mean(x[]))))
phat75    <-    exp(alpha + beta*(75- mean(x[])))/(1 + exp(alpha + beta*(75- mean(x[]))))
phat80<-    exp(alpha + beta*(80 - mean(x[])))/(1 + exp(alpha + beta*(80- mean(x[]))))
    }

list(x = c(56.9,45.9,49.6,77.8,65.8,53.0,63.0,56.7,64.7,57.1,47.3,74.7,68.7,64.8,64.8,
58.1,54.5,78.9,42.0,54.0,52.5,69.0,55.7,52.2,49.9,50.4,44.6,40.3,73.6,42.8,
63.1,45.9,51.8,52.5,53.4,58.6,44.0,55.4,62.5,73.8,69.0,45.5,77.8,59.9,72.7,
66.4,51.0,78.6,42.2,69.1,48.0,48.9,51.8,79.1,59.4,79.3,44.5,65.3,76.2,43.1,
47.5,75.3,65.4,59.5,49.0,50.5,73.2,61.3,75.6,46.4,58.0,60.2,49.2,58.4,75.1,
60.5,48.8,50.0,77.5,60.8,66.3,56.8,53.3,57.1,60.5,43.5,43.5,71.1,57.3,79.6,
43.2,50.9,58.9,41.9,68.6,45.7,41.7,64.1,73.2,73.8,52.3,50.0,57.9,45.8,40.0,
52.8,74.5,47.0,43.6,61.4,72.2,42.9,47.9,79.8,59.8,60.5,77.7,53.2,40.7,63.9,
55.0,65.6,44.0,55.5,72.3,60.3,79.6,60.5,79.9,56.0,50.4,60.4,57.3,50.4,76.4,
74.7,41.8,44.4,55.1,68.7,75.2,65.3,43.4,47.0,71.3,50.3,42.6,62.7,53.2,41.3,
59.1,58.1,75.6,62.6,71.8,57.6,47.7,55.9,58.6,79.9,50.2,62.8,62.1,51.4,50.0,
67.6,72.7,54.7,42.2,58.0,55.8,44.5,73.3,53.3,41.6,61.9,74.8,74.3,69.3,58.8,
67.4,77.9,74.7,44.2,44.5,64.9,72.6,45.3,63.9,54.8,77.3,53.3,50.3,60.7,70.3,
60.4,79.0,75.0,61.7,61.0),

y = c(1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,0,0,0,1,0,
0,0,1,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,1,0,0,0,1,0,1,0,0,1,0,1,0,1,0,
0,1,0,1,0,0,0,0,0,0,0,1,0,0,0,0,1,1,0,0,0,1,1,0,0,1,1,0,0,0,0,0,1,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,1,0,0,1,0,1,0,1,0,1,1,0,1,0,0,1,0,0,1,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,1,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0))
# initial values
list(alpha=.0,beta=.0,Se=.9999,Sp=.9999)
```

APPENDIX B

Model Selection Code

SAS code for finding the best regression model:

```
data paper;
input mean_se mean_sp se_weight sp_weight width_45 width_75;
lnwidth_75=log(width_75);
cards;
0.85    0.85    70  70  0.9947  0.006903
0.85    0.85    70  70  0.9947  0.006903
0.78    0.78    40  40  1   0.00003626
0.78    0.78    100 40  1   0.0002881
0.78    0.78    40  100 1   0.000002375
0.78    0.78    100 100 1   0.00005969
0.78    0.92    40  40  1   0.005313
0.78    0.92    100 40  0.999999463     0.01431
0.78    0.92    40  100 0.9997375   0.007469
0.78    0.92    100 100 0.9997778   0.03081
0.92    0.78    40  40  0.427   0.005338
0.92    0.78    40  100 0.3804  0.001838
0.92    0.78    100 40  0.4045  0.004685
0.92    0.78    100 100 0.3635  0.001353
0.92    0.92    40  40  0.4422  0.05653
0.92    0.92    100 40  0.44717 0.04924
0.92    0.92    40  100 0.45231 0.05161
0.92    0.92    100 100 0.40147 0.06312
0.99    0.85    70  70  0.3979  0.01794
0.71    0.85    70  70  0.999999776     0.00002416
0.85    0.99    70  70  0.3466  0.19957807
0.85    0.71    70  70  0.5373  0.0003782
0.85    0.85    10  70  1   0.000008073
0.85    0.85    130 70  0.7064  0.005871
0.85    0.85    70  10  0.987   0.01077
0.85    0.85    70  130 0.9981  0.006834
run;

/*program of width under temp 45 degree*/
proc glm data = paper ;
model width_45 = mean_se mean_sp se_weight sp_weight mean_se*mean_se
mean_sp*mean_sp se_weight*se_weight sp_weight*sp_weight mean_se*mean_sp
mean_se*se_weight mean_se*sp_weight mean_sp*se_weight mean_sp*sp_weight
se_weight*sp_weight;
output out = rpaper p = pred r = res ;
run;
  proc plot data = rpaper vpercent = 70 ;
  plot res*pred    ;
  proc capability noprint data = rpaper lineprinter ;
  var res ;
  qqplot res /normal(mu = est sigma = est symbol='.') square    ;
  run ;
    /*adjust model 1 under temp 45 degree*/
  proc glm data=paper;
model width_45 = mean_se mean_sp se_weight sp_weight mean_se*mean_se
mean_sp*mean_sp se_weight*se_weight sp_weight*sp_weight;
output out = rpaper1 p = pred1 r = res1 ;
run;
proc plot data = rpaper1 vpercent = 70 ;
plot res1*pred1    ;
proc capability noprint data = rpaper1 lineprinter ;
var res1 ;
          qqplot res1 /normal(mu = est sigma = est symbol='.') square    ;
run ;
        /*adjust model 2 under temp 45 degree* by deleting spw*spw*/
proc glm data=paper;
model width_45 = mean_se mean_sp se_weight sp_weight mean_se*mean_se
mean_sp*mean_sp se_weight*se_weight;
output out = rpaper11 p = pred11 r = res11 ;
run;
          proc plot data = rpaper11 vpercent = 70 ;
plot res11*pred11    ;
proc capability noprint data = rpaper11 lineprinter ;
var res11 ;
qqplot res11 /normal(mu = est sigma = est symbol='.') square    ;
run ;
      /*adjust model 3 under temp 45 degree* by deleting spw*spw*/
          proc glm data=paper;
model width_45 = mean_se mean_sp se_weight mean_se*mean_se mean_sp*mean_sp
se_weight*se_weight;
output out = rpaper111 p = pred111 r = res111 ;
run;
proc plot data = rpaper111 vpercent = 70 ;
plot res111*pred111    ;
proc capability noprint data = rpaper111 lineprinter ;
var res111 ;
qqplot res111 /normal(mu = est sigma = est symbol='.') square    ;
run ;
/*program of width under 75 degree*/
proc glm data = paper ;
model width_75 = mean_se mean_sp se_weight sp_weight mean_se*mean_se
mean_sp*mean_sp se_weight*se_weight sp_weight*sp_weight mean_se*mean_sp
mean_se*se_weight mean_se*sp_weight mean_sp*se_weight mean_sp*sp_weight
se_weight*sp_weight;
output out =rpaper2 p = pred2 r = res2 ;
run;
proc plot data = rpaper2 vpercent = 70 ;
```

```
plot res2*pred2    ;
proc capability noprint data = rpaper2 lineprinter ;
var res2 ;
          qqplot res2 /normal(mu = est sigma = est symbol='.') square    ;
run ;
/*boxcox transformation of width under 75 degree*/
proc transreg data = paper ss2 plots = all;
  model BoxCox(width_75 / lambda=-.4 to .4 by 0.02)
                          = identity(mean_se mean_sp se_weight sp_weight)/ pboxcoxtable ;
      run;
/*run under log(width_75)*/
proc glm data = paper ;
model lnwidth_75 = mean_se mean_sp se_weight sp_weight mean_se*mean_se
mean_sp*mean_sp se_weight*se_weight sp_weight*sp_weight mean_se*mean_sp
mean_se*se_weight mean_se*sp_weight mean_sp*se_weight mean_sp*sp_weight
se_weight*sp_weight;
output out =rpaper3 p = pred3 r = res3 ;
run;
proc plot data = rpaper3 vpercent = 70 ;
plot res3*pred3    ;
proc capability noprint data = rpaper3 lineprinter ;
          var res3 ;
qqplot res3 /normal(mu = est sigma = est symbol='.') square    ;
run ;
    /*run under log(width_75) by deleting spw*spw, ,eanse*spw,meansp*sew, sew*spw*/
proc glm data = paper ;
model lnwidth_75 = mean_se mean_sp se_weight sp_weight mean_se*mean_se
mean_sp*mean_sp se_weight*se_weight    mean_se*mean_sp mean_se*se_weight
mean_sp*sp_weight;
          output out =rpaper33 p = pred33 r = res33 ;
run;
proc plot data = rpaper33 vpercent = 70 ;
plot res33*pred33    ;
proc capability noprint data = rpaper33 lineprinter ;
    var res33 ;
    qqplot res33 /normal(mu = est sigma = est symbol='.') square    ;
    run ;
```

APPENDIX C

Plot Generating Code

```
#generate leverage and outlier plots
Q1<- read.table("F:/STAT Thesis/Code/leverage.txt", header=T)
Q1

meanse2 =(Q1$mean_se)^2
meansp2=(Q1$mean_sp)^2
seweight2=(Q1$se_weight)^2
meansesp=(Q1$mean_se)*(Q1$mean_sp)
meansesewei=(Q1$mean_se)*(Q1$se_weight)
meanspspwei=(Q1$mean_sp)*(Q1$sp_weight)


#fit the model under temp 45
fit_45<-lm(width_45~mean_se + mean_sp + se_weight + meanse2 + meansp2 +
seweight2,data=Q1)
summary(fit_45)


#find the studentized value
rstudent(fit_45)
studentizedres_45<-rstudent(fit_45)

#find the yhat and fitted value
ehat_45<-fit_45$residuals
yhat_45<-Q1$width_45-ehat_45
fitted(fit_45)
fit.value_45<-fitted(fit_45)

# get the hat value
hat_45<-hatvalues(fit_45)
hbar_45=mean(hat_45)
hbar_45

#plots
plot(hat_45,studentizedres_45)
abline(h=-2)
abline(h=2)
abline(v=2*hbar_45)
abline(v=3*hbar_45)
plot(fit.value_45,studentizedres_45)
abline(h=0)
```

```
qqnorm(ehat_45,ylab="Residuals",xlab="Normal Quantile")
qqline(ehat_45)

#fit the model under temp 75
fit_75<-lm(log(width_75)~mean_se + mean_sp + se_weight +sp_weight+ meanse2 +
meansp2 + seweight2+meansesp+meansesewei+meanspspwei,data=Q1)
summary(fit_75)


#find the studentized value
rstudent(fit_75)
studentizedres_75<-rstudent(fit_75)

#find the yhat and fitted value
ehat_75<-fit_75$residuals
yhat_75<-Q1$width_75-ehat_75
fitted(fit_75)
fit.value_75<-fitted(fit_75)

# get the hat value
hat_75<-hatvalues(fit_75)
hbar_75=mean(hat_75)
hbar_75

#plots
plot(hat_75,studentizedres_75)
abline(h=-2)
abline(h=2)
abline(v=2*hbar_75)
abline(v=3*hbar_75)
plot(fit.value_75,studentizedres_75)
abline(h=0)

qqnorm(ehat_75,ylab="Residuals",xlab="Normal Quantile")
qqline(ehat_75)

##Generate plots for the raw data
Q2<- read.table("F:/STAT Thesis/Data/raw data.txt", header=T)
Q2
#plots for Temp 45
```

```
plot(Q2$mean_Se, Q2$width_45,xlab="MeanSe",ylab="Width under 45 Fahrenheit")
plot(Q2$mean_Sp, Q2$width_45,xlab="MeanSp",ylab="Width under 45 Fahrenheit")
plot(Q2$Se_weight, Q2$width_45,,xlab="SeWeight",ylab="Width under 45 Fahrenheit")

#plots for Temp 75
plot(Q2$mean_Se, Q2$width_75,xlab="MeanSe",ylab="Width under 75 Fahrenheit")
plot(Q2$mean_Sp, Q2$width_75,xlab="MeanSp",ylab="Width under 75 Fahrenheit")
plot(Q2$Se_weight, Q2$width_75,xlab="SeWeight",ylab="Width under 75 Fahrenheit")
plot(Q2$Sp_weight, Q2$width_75,xlab="SpWeight",ylab="Width under 75 Fahrenheit")

#Contour plots
#under 45 degree with seweight=40
meanse_40=seq(0.70,1,0.01)
meansp_40=seq(0.70,1,0.01)

model_40=function(a,b)
    {yhat_40=(-2.834e+01)+
(2.330e+01)*a+(4.897e+01)*b+(5.086e-03)*40+(-1.577e+01)*(a^2)+
(-2.888e+01)*(b^2)+(-4.300e-05)*(40^2)}
z40=outer(meanse_40, meansp_40 ,model_40)
z40

contour(meanse_40, meansp_40, z40,
             nlevels=12, xlab = "Mean_Se", ylab = "Mean_Sp",main="Contour Plot Under
45 Degree(Se_weight=40)")
points(Q1$mean_se, Q1$mean_sp)
text(Q1$mean_se, Q1$mean_sp,Q1$median_45,cex=.75,pos=3)
text(Q1$mean_se, Q1$mean_sp,Q1$count_45,cex=.75,pos=1)



#under 45 degree with seweight=100
meanse_100=seq(0.70,1,0.01)
meansp_100=seq(0.70,1,0.01)

model_100=function(a,b)
    {yhat_100=(-2.834e+01)+
(2.330e+01)*a+(4.897e+01)*b+(5.086e-03)*100+(-1.577e+01)*(a^2)+
(-2.888e+01)*(b^2)+(-4.300e-05)*(100^2)}
z100=outer(meanse_100, meansp_100 ,model_100)
z100
```

```
contour(meanse_100, meansp_100, z100,
                    nlevels=12,xlab = "Mean_Se", ylab = "Mean_Sp",main="Contour Plot Under
45 Degree(Se_weight=100)")
points(Q1$mean_se, Q1$mean_sp)
text(Q1$mean_se, Q1$mean_sp,Q1$median_45,cex=.75,pos=3)
text(Q1$mean_se, Q1$mean_sp,Q1$count_45,cex=.75,pos=1)



#under 75 degree with seweight=40 spweight=40
meanse_4040=seq(0.70,1,0.01)
meansp_4040=seq(0.70,1,0.01)

model_4040=function(a,b)
      {yhat_4040=(-2.317e+02)+(3.600e+02)*a+(1.093e+02)*b+(3.649e-01)*40+(-2.078e-01)
*40+
(-1.168e+02)*(a^2)+(1.484e+01)*(b^2)+(-9.433e-04)*(40^2)+(-1.442e+02)*(a*b)+(-2.404e-0
1)*(a*40)+(2.343e-01)*(b*40)}
z4040=outer(meanse_4040, meansp_4040 ,model_4040)
z4040


contour(meanse_4040, meansp_4040, z4040,
                    nlevels=12,xlab = "Mean_Se", ylab = "Mean_Sp",main="Contour Plot Under
75 Degree(Se_weight=40,Sp_weight=40)")
points(Q1$mean_se, Q1$mean_sp)
text(Q1$mean_se, Q1$mean_sp,Q1$median_75,cex=.75,pos=3)
text(Q1$mean_se, Q1$mean_sp,Q1$count_75,cex=.75,pos=1)


#under 75 degree with seweight=40 spweight=100
meanse_40100=seq(0.70,1,0.01)
meansp_40100=seq(0.70,1,0.01)

model_40100=function(a,b)
      {yhat_40100=(-2.317e+02)+(3.600e+02)*a+(1.093e+02)*b+(3.649e-01)*40+(-2.078e-01
)*100+(-1.168e+02)*(a^2)+(1.484e+01)*(b^2)+(-9.433e-04)*(40^2)+(-1.442e+02)*(a*b)+(-2
.404e-01)*(a*40)+(2.343e-01)*(b*100)}
z40100=outer(meanse_40100, meansp_40100 ,model_40100)
z40100
```

```
contour(meanse_40100, meansp_40100, z40100,
                    nlevels=12,xlab = "Mean_Se", ylab = "Mean_Sp",main="Contour Plot Under
75 Degree(Se_weight=40,Sp_weight=100)")
text(Q1$mean_se, Q1$mean_sp,Q1$median_75,cex=.75,pos=3)
text(Q1$mean_se, Q1$mean_sp,Q1$count_75,cex=.75,pos=1)


#under 75 degree with seweight=100 spweight=40
meanse_10040=seq(0.70,1,0.01)
meansp_10040=seq(0.70,1,0.01)

model_10040=function(a,b)
      {yhat_10040=(-2.317e+02)+(3.600e+02)*a+(1.093e+02)*b+(3.649e-01)*100+(-2.078e-0
1)*40+(-1.168e+02)*(a^2)+(1.484e+01)*(b^2)+(-9.433e-04)*(100^2)+(-1.442e+02)*(a*b)+(-
2.404e-01)*(a*100)+(2.343e-01)*(b*40)}
z10040=outer(meanse_10040, meansp_10040 ,model_10040)
z10040

contour(meanse_10040, meansp_10040, z10040,
                    nlevels=12,xlab = "Mean_Se", ylab = "Mean_Sp",main="Contour Plot Under
75 Degree(Se_weight=100,Sp_weight=40)")
points(Q1$mean_se, Q1$mean_sp)
text(Q1$mean_se, Q1$mean_sp,Q1$median_75,cex=.75,pos=3)
text(Q1$mean_se, Q1$mean_sp,Q1$count_75,cex=.75,pos=1)


#under 75 degree with seweight=100 spweight=100
meanse_100100=seq(0.70,1,0.01)
meansp_100100=seq(0.70,1,0.01)

model_100100=function(a,b)
      {yhat_100100=(-2.317e+02)+(3.600e+02)*a+(1.093e+02)*b+(3.649e-01)*100+(-2.078e-
01)*100+(-1.168e+02)*(a^2)+(1.484e+01)*(b^2)+(-9.433e-04)*(100^2)+(-1.442e+02)*(a*b)
+(-2.404e-01)*(a*100)+(2.343e-01)*(b*100)}
z100100=outer(meanse_100100, meansp_100100 ,model_100100)
z100100

contour(meanse_100100, meansp_100100, z100100,
                    nlevels=12,xlab = "Mean_Se", ylab = "Mean_Sp",main="Contour Plot Under
75 Degree(Se_weight=100,Sp_weight=100)")
```

```
points(Q1$mean_se, Q1$mean_sp)
text(Q1$mean_se, Q1$mean_sp,Q1$median_75,cex=.75,pos=3)
text(Q1$mean_se, Q1$mean_sp,Q1$count_75,cex=.75,pos=1)
```