THE UNIVERSITY OF MANITOBA

A Retrospective Analysis of Survival Data of Snapping Turtle Embryos using Generalized Linear Models

BY

AMELIA TERESA SHEOCHARAN

A PRACTICUM SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF STATISTICS

June 2005

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION

"A Retrospective Analysis of Survival Data of Snapping Turtle Embryos using Generalized Linear Models"

BY

Amelia Teresa Sheocharan

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment of the requirement of the degree

Of

MASTER OF SCIENCE

Abstract

The effect of temperature, moisture and site on the proportion of turtle eggs that survived after hatching was analyzed. Preliminary main effect plots showed that temperature and moisture had no effect on the survival of a turtle egg. On the other hand site showed some differences. Next, the interaction effects were then examined using interaction plots. Here it was found that site by temperature showed some slight differences.

The data were analyzed using a newly developed technique called Generalized Linear Models. After the analysis was completed it was found that indeed site had an effect on the survival rate of turtle eggs.

Acknowledgements

First and foremost I would like to thank my supervisor Dr. K. Mount for taking me on as one of his students, and for all his help, support and patience. I would also like to thank my committee, Dr. B. Macpherson and Mr. W. Falk for taking time out of their busy schedule to proof read the practicum and offer their suggestions. I would also like to thank the Department of Statistics, Dr. S. Cheng for giving me the many opportunities while completing my project.

Finally, I would like to thank all my friends and family for their continued love and support.

Table of Content

List of Figures

List of Tables

Introduction

In 1988 Michele Bobyn, a student at the University of Guelph, conducted an experiment to help her determine the incubation conditions needed to influence embryonic survival in snapping turtles. The research in this practicum re-evaluates the same study but uses modern methods to analyze the 1988 data.

The objective of this practicum is to determine the effect of temperature, moisture and site on the embryonic survival of a snapping turtle egg. The analysis used is a newly developed method called generalized linear model.

A generalized linear mixed model will be fit and used to help determine what conditions are needed to have the highest survival rate of a turtle egg and estimate the probability reflected in those conditions.

In chapter 1 we focus on the factors our researcher is considering and make some initial main effect and interaction evaluations.

Chapter 2 examines the temperature effects and introduces new concepts needed to evaluate the data.

Chapter 3 builds on the temperature model and examines the added effects in the model when the factor moisture is introduced.

Chapter 4 introduces the last effect site and uses all our data to fit a model that best reflects the data. The model is then used to make predictions for the turtles using best linear unbiased techniques.

Finally in chapter 5 we summarize our conclusions and give final remarks.

Chapter 1: Overview

1.1 THE OBJECTIVE

The following information was provided through a case study report written by the experimenter.

The experiment was conducted in June 1988 by Michele Bobyn at the University of Guelph under the supervision of Dr. R.J. Brooks.

The objective was to:

'determine the effect of temperature, moisture and site on the embryonic survival of a snapping turtle egg and to determine how embryonic survival varies from clutch to clutch.'[3]

1.2 THE EXPERIMENT

Snapping turtle eggs were collected from four Ontario nesting sites:

I. North Madawaska drainage system of Algonquin Park (AP)

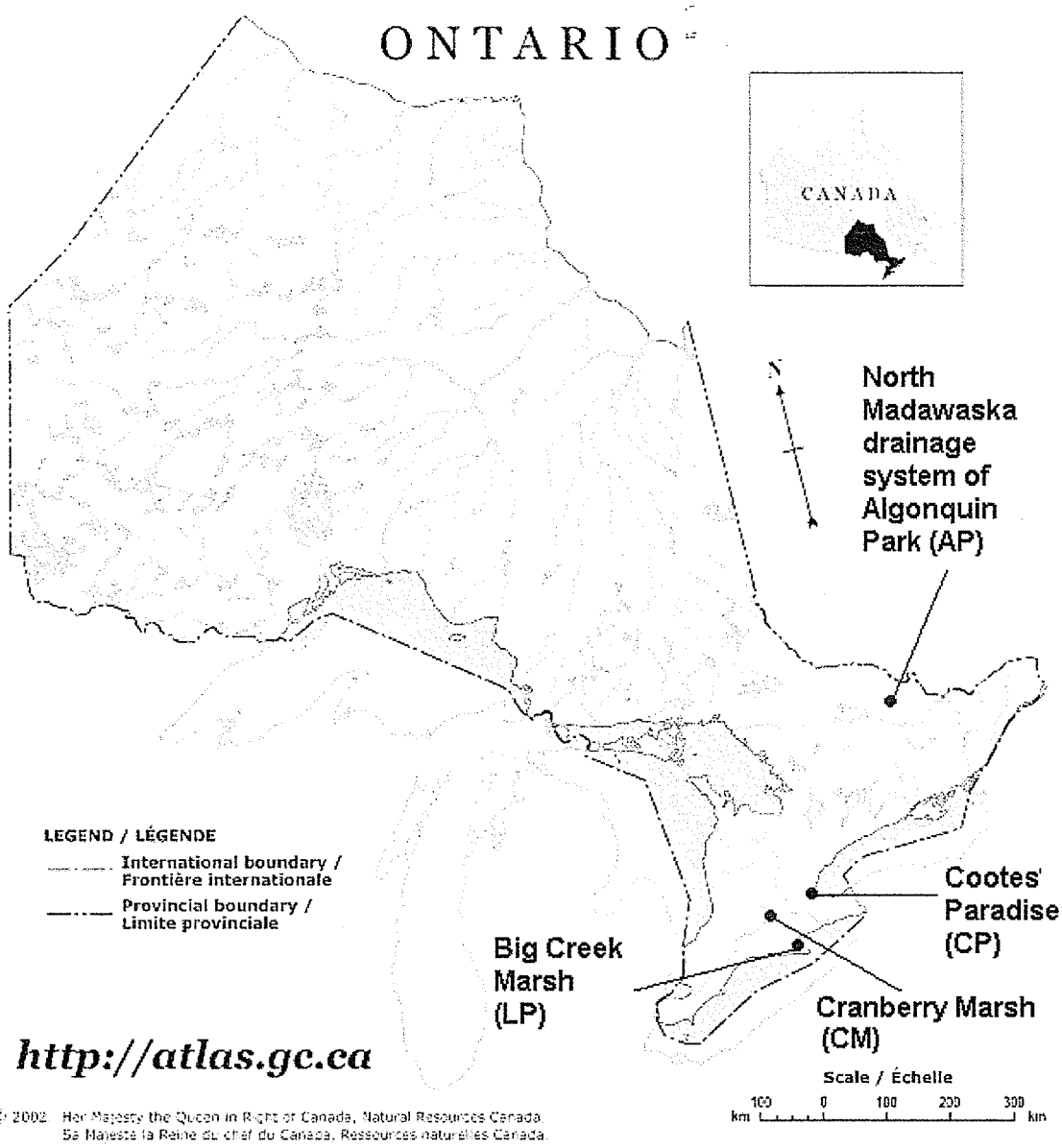
II. Cootes' Paradise near Hamilton (CP)

III. Big Creek Marsh near Long Point (LP)

IV. Cranberry Marsh near Ajax (CM).

The sites are shown on figure 1.1.

Figure 1.1:


ONTARIO
CANADA
N
North Madawaska drainage system of Algonquin Park (AP)
LEGEND / LÉGENDE
International boundary / Frontière internationale
Provincial boundary / Limite provinciale
Cootes' Paradise (CP)
Big Creek Marsh (LP)
Cranberry Marsh (CM)
http://atlas.gc.ca
Scale / Échelle
100 0 100 200 300
km
km
© 2002 Her Majesty the Queen in Right of Canada, Natural Resources Canada.
Sa Majesté la Reine du chef du Canada, Ressources naturelles Canada.

Each egg was labelled with:

- its location (AP, CP, LP, CM)
- clutch identification
- egg number (#1 = last laid egg).

For transport the eggs were arranged in a single layer plastic shoe box and covered with a mixture of vermiculite (*a type of mineral*) and water. The shoe box was covered with aluminium foil to prevent dehydration. All eggs were maintained at a constant 20°C before being placed into an incubator.

For the experiment there were six incubators used, two set at each of three temperatures: 21°C, 25°C and 29.5°C. A total of 720 eggs were collected. The following table shows how many eggs were collected from each site.

Table 1.1:

Site	***No. of clutches***	***No. of eggs***
AP	6	212
CP	5	183
LP	5	154
CM	5	171
TOTAL	21	720

The eggs were then randomly assigned to a tray. Each tray had a total of 40 eggs with a maximum of 2 eggs from each clutch at each site.

The trays were then randomly assigned to one of two moistures: dry or wet.

A total of three trays were placed in each incubator. The combinations of moistures were either two wet and one dry or two dry and one wet tray per incubator.

When the first sign of hatching occurred in an incubator all the eggs were removed at that temperature and placed in a glass jar.

Less then a week into the start of the incubation period one of the incubators at 29.5^0C malfunctioned and 120 eggs were lost. This left five incubators for the study.

The response variable, survival, was measured two months after the egg hatched. Survival was scored a 1 or 0 depending on whether the turtle lived or died.

1.3 DESIGN OF THE EXPERIMENT

The design of the experiment followed a split- split plot layout. At the whole plot level the factor is temperature and the experimental units are the incubators. Two incubators were randomly assigned to each temperature level.

At the split plot level the factor is moisture and the experimental units are the trays within incubators. At this point the trays were assigned to one of two moisture levels; wet or dry.

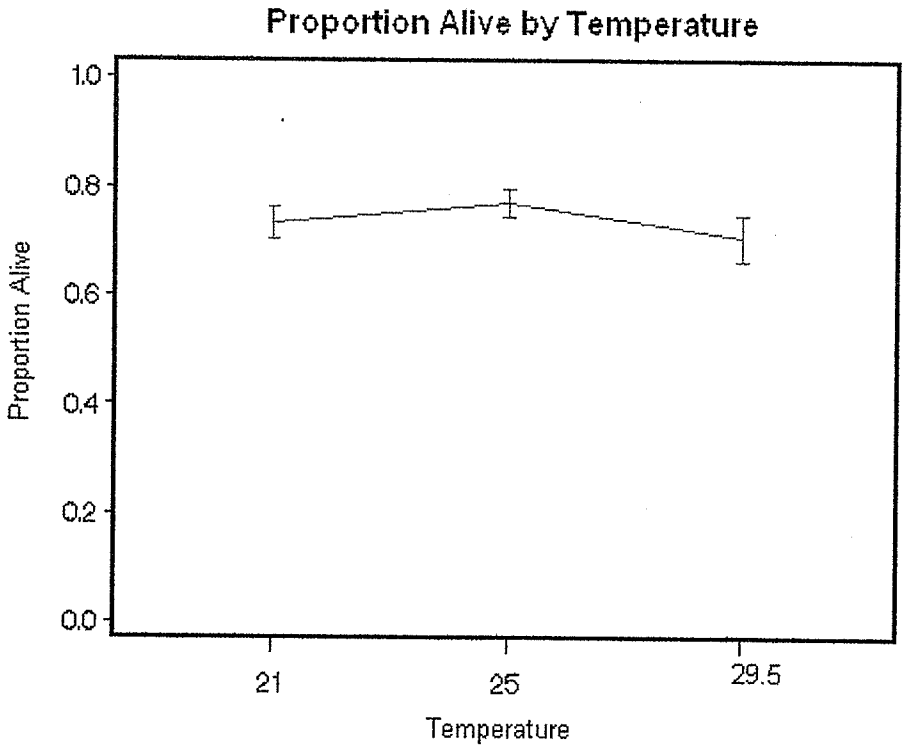
At the split- split plot level the factor is site and the experimental units are the eggs. For the experiment approximately 600 eggs were used.

1.4 INITIAL PLOTS FOR THE MAIN EFFECTS

Before any formal data analysis is done, we are going to look at main effect and interaction plots with plus and minus one standard error bands for each data point.

The first factor, temperature, has 3 levels: 21^0C, 25^0C, and 29.5^0C.

Figure 1.2:

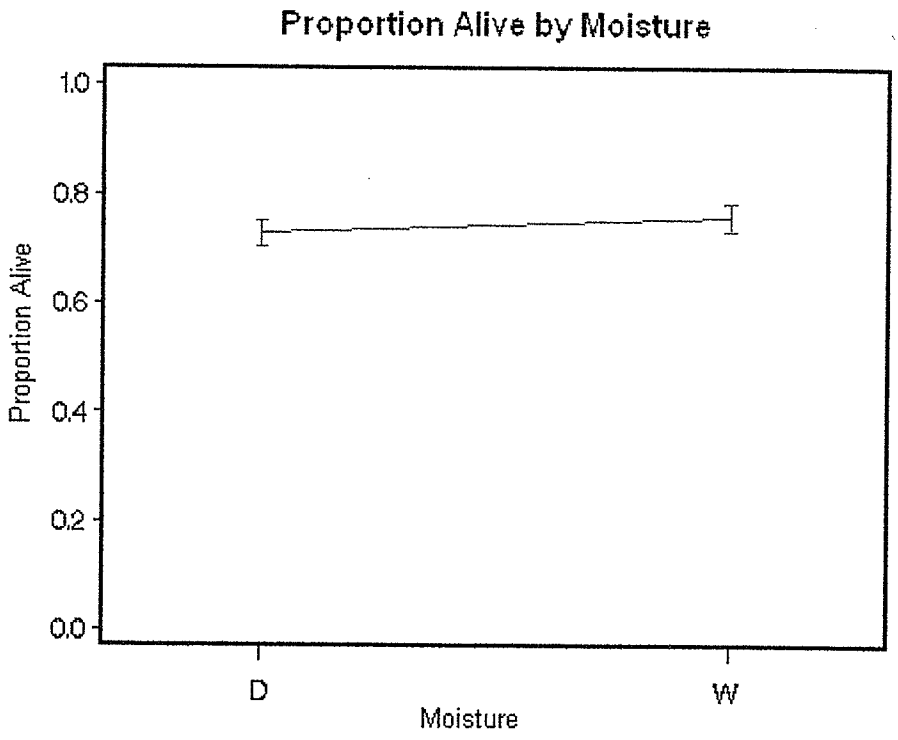


As we can see there is very little change in the proportion of turtles that survived as the temperature increased from 21^0C to 29.5^0C. This tells us that the main effect of temperature has very little effect on the survival of a turtle egg. We notice that at temperature 29.5^0C there is a larger standard error; that is because at 29.5^0C one of the incubators malfunctioned and 120 eggs were lost. At each

21^0C and 25^0C we are looking at the number of eggs that survived out of 240, whereas at 29.5^0C we are looking at the number of eggs that survived out of 120.

Next, we looked at the factor moisture.

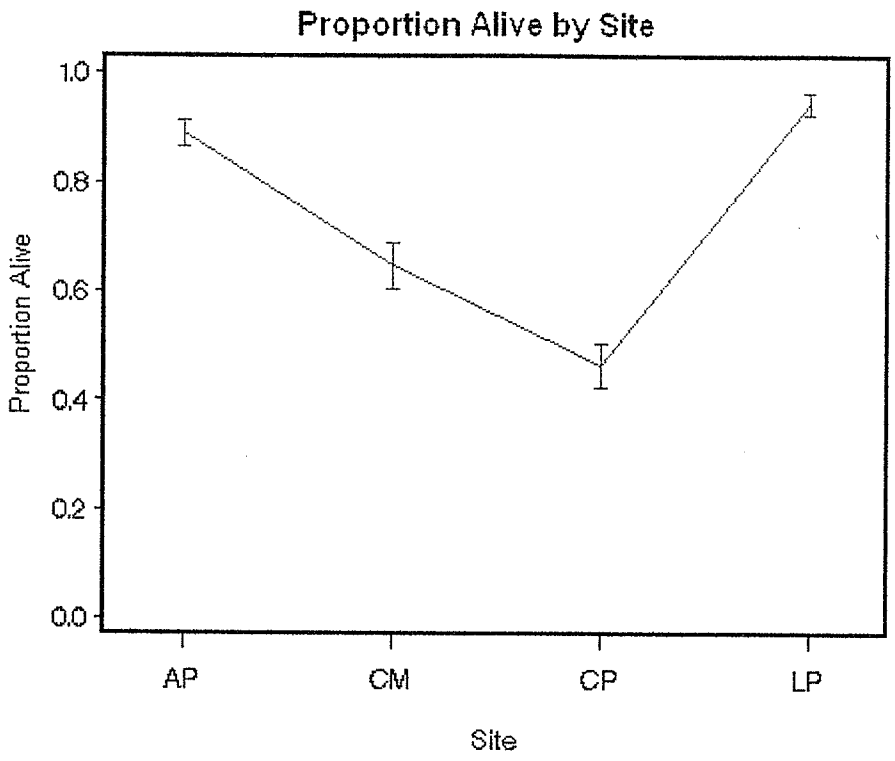
Figure 1.3:



As we can see moisture shows very little difference between wet and dry, which tells us that the factor moisture has no notable effect on the survival of a turtle.

The next factor is site.

Figure 1.4:



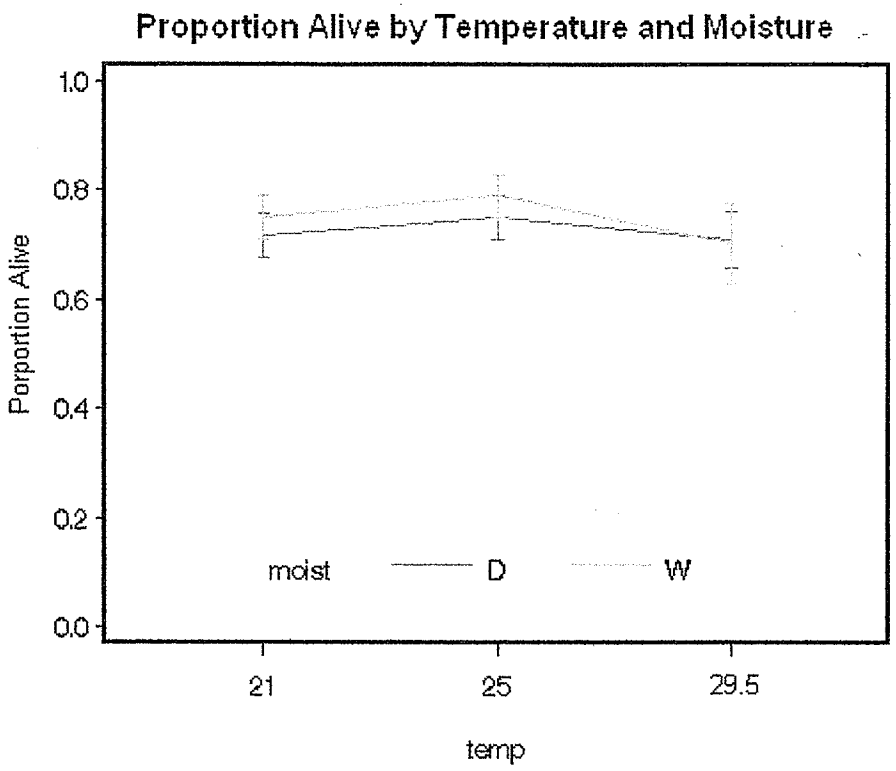
We notice there is a considerable difference as to where the turtle egg came from and whether it survived. It seems Big Creek Marsh (LP) had the highest mean survival rate and Cootes' Paradise (CP) the lowest.

1.5 PRELIMINARY INTERACTION PLOTS

Next, we consider interactions between factors.

First we look at temperature by moisture.

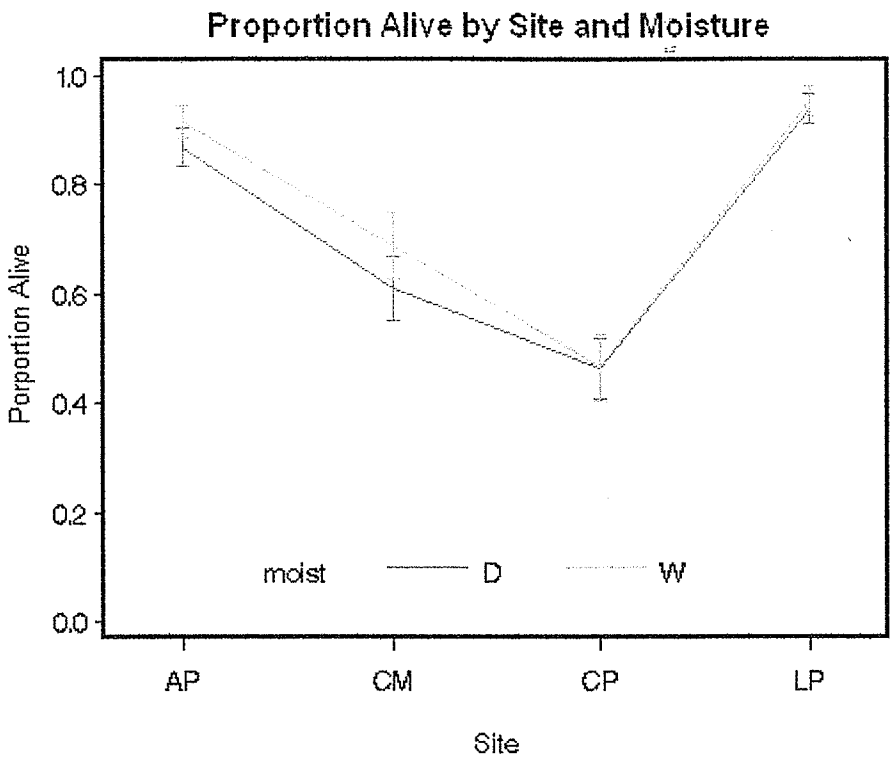
Figure 1.5:



As we can see the interaction plot of temperature by moisture shows no interaction effect.

Next, we look at site by moisture.

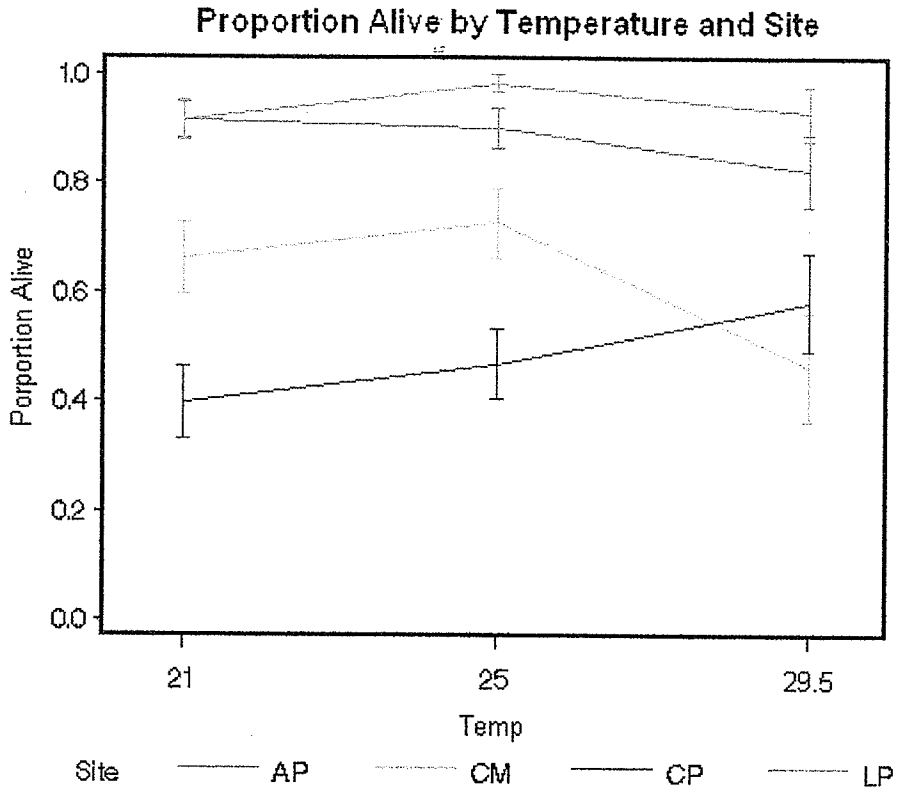
Figure 1.6:



Again we see no interaction between these two factors.

Lastly, we look at the temperature and site interaction to see if there are any effects.

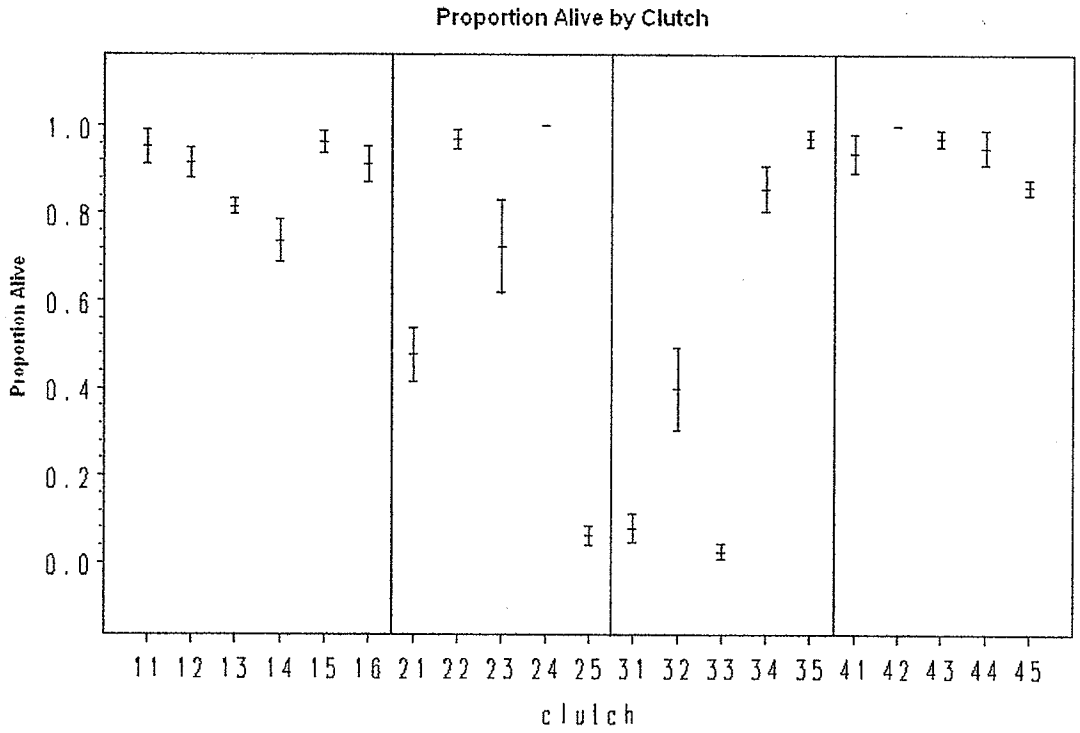
Figure 1.7:



As we can see there appear to be some differences here. As we continue we will revisit this relationship.

At this time we will look at the clutch to clutch variation. To do this we have plotted the proportion of turtle eggs that survived in each clutch from our four sites.

Figure 1.8:

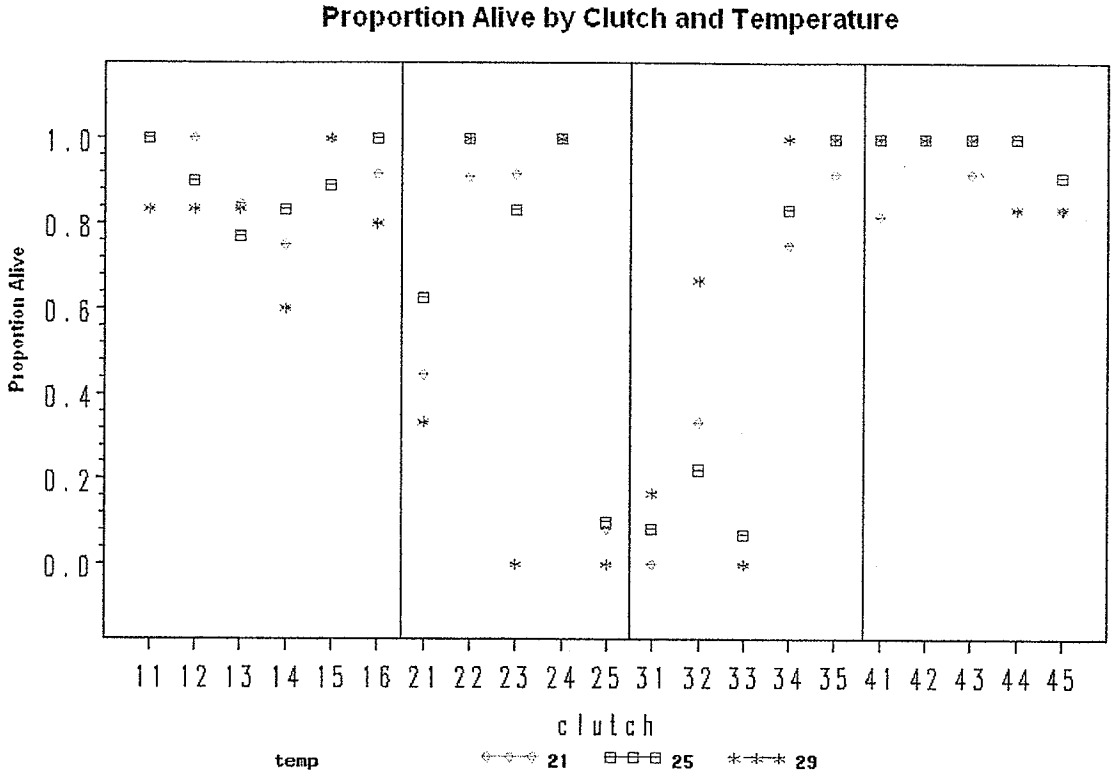


LEGEND:
Clutch #11 – 16: Site: AP
Clutch #21 – 25: Site: CM
Clutch #31 – 35: Site: CP
Clutch #41 – 45: Site: LP

From this we see there is similar variation between sites AP and LP as well as sites CM and CP. Overall, we see that sites CM and CP are more variable than sites AP and LP.

The next graph looks at the proportion of turtle eggs that survived from each clutch at each temperature level.

Figure 1.9:



LEGEND:
Clutch #11 – 16: Site: AP
Clutch #21 – 25: Site: CM
Clutch #31 – 35: Site: CP
Clutch #41 – 45: Site: LP

From this graph we do not see any patterns between clutch and temperature.

Chapter 2: Analysis of the Temperatures

2.1 INTRODUCTION

At the first level, the whole plot, one looks at the effect temperature has on the survival of a turtle egg.

If the response variable were continuous we might assume that it followed a normal distribution therefore a linear model or general linear model could be used. Because our data are binomially distributed a generalized linear model is appropriate at this stage.

A linear model has model equation $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$, where the $E(\underline{Y}) = X\underline{\beta}$. It postulates a linear relationship between a dependent or response variable $\underline{Y}$, and a linear combination of fixed predictor variables. The ε stands for the error i.e. variability in $\underline{Y}$ that cannot be accounted for by the predictors. The expected value of an error is assumed to be zero and it is independently normally distributed with constant variance i.e. $\underline{\varepsilon} \sim N(0, \sigma^2 I)$.

To estimate the unknown parameters we use the method of least squares. Because we assume the error terms are normally distributed we can carry out tests on the parameters. In addition confidence intervals for the parameters, and confidence intervals for the mean of the response variable are obtainable.

In the general linear model, $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$, $E(\underline{Y}) = X\underline{\beta}$ as above, and $\underline{\varepsilon}$ has mean zero but a more general variance covariance structure ie. $\sum$. As before, the dependent variables are expected to follow the normal distribution. To estimate the unknown parameters we now use the method of general least squares.

A generalized linear model, with form $g(\mu) = X\beta$ is a generalization of the general linear model. It was developed for data that do not follow the assumptions of a general linear model. For example, we have independent response variables $y_1, \ldots\ldots y_n$ with means $\mu_1, \ldots\ldots \mu_n$. The response variable may or may not be a continuous variable; instead it could be a count. A generalized linear model can be used in two situations:

i. for dependent variables which are discrete random variables
ii. for dependent variables, which are not linearly related to the predictor variables i.e. data that needs to be transformed so that a function of the mean of the response variable is linearly related to the predictor variables.

The generalized linear model has three aspects:

First, it extends linear models to the situation where response variables are members of the exponential family. The exponential family includes normal, binomial, Poisson, Geometric, negative binomial, exponential, gamma and inverse normal distributions. Members of the exponential family of distributions have probability density function for a response y that can be expressed in the form:

$$f_y(y;\theta,\phi)=\exp\left\{\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)\right\}$$

where:

θ is the natural location parameter. It is a function of the mean $\theta(\mu)$, where μ is the mean of the response.

ϕ is a scale or dispersion parameter.

$a(\cdot),b(\cdot)\,and\,c(\cdot)$ are specific functions.

The variance, v(μ), is a function of μ rather than θ.

Table 2.1, taken from SAS System for Mixed Models, is a table of functions for some common members of the exponential family.

Table 2.1:

Table of functions for the three distributions [5]

	Poisson	Normal	Bernoulli
Mean(y)	λ	μ	π
Var(y)	λ	σ^2	$[\pi(1-\pi)]$
$\theta(\mu)$	$\log_e(\lambda)$	μ	$\log_e[\pi/(1-\pi)]$
$a(\phi)$	1	σ^2	1
$v(\mu)$	λ	1	$\pi(1-\pi)$

Secondly, as in the two previous models, it has a set of parameters $\underline{\beta}$ and explanatory variables $x^{'}=[x_1.....x_p]$

Thirdly, there is a link function g such that: $g(\mu_i) = X_i \underline{\beta}$ where $\eta_i = g(\mu_i)$ The link function is a function of the mean μ. It connects the mean of the raw data to the natural parameters to give us the basic form of the generalized linear model: $\eta = X\beta$. Various link functions can be chosen, depending on the assumed distribution of the y variable e.g. logit link: $\log\left(\frac{\pi}{1-\pi}\right)$.

In the analysis of this chapter we will use a logit link function since the data we are dealing with follow a binomial distribution.

In generalized linear models, iterative methods are needed to solve for the parameter estimates.

2.2 ITERATIVE METHOD FOR PARAMETER ESTIMATION IN GENERALIZED LINEAR MODELS.

The next two sections are a summary of the iterative method for parameter estimation from Generalized Linear and Mixed Models by Searle and McCulloch [8].

The method of maximum likelihood is the basis for parameter estimation in generalized linear models. However, the actual operation of maximum likelihood results in an algorithm based on iteratively weighted least squares.

The likelihood function is defined as the joint density function of n random variables $f(\underline{x},\theta)$. It is considered a function of θ and can be denoted as $L(\theta,\underline{x})$.

Given $\underline{x}$ we want to maximize $L(\theta,\underline{x})$ or in this case $L(\underline{y},\underline{\beta})$. We set the derivative of the log-likelihood $L(\underline{y},\underline{\beta})$ to 0 and solve for $\hat{\beta}$.

In matrix notation the maximum likelihood equations (also called the score equations) can be written as

$$X'W\Delta\underline{y} = X'W\Delta\underline{\mu}$$

where:

W = $\{_d w_i\}$ is a nxn diagonal matrix with elements $w_i = \left[v(\mu_i)g_\mu^2(\mu_i)\right]^{-1}$ and $g_\mu(\mu_i) = \frac{\partial \eta_i}{\partial \mu_i}$. $v(\mu_i)$ is the variance function for the ith observation. We note that μ_i is π_i. The weights (w_i) depend on the parameters $\underline{\mu_i}$. We can re-write w_i as $\frac{1}{\text{var}(y_i)}\left(\frac{\partial \mu_i}{\partial \eta}\right)^2$. For the turtle data, the variable y follows a binomial distribution. Thus var(y_i) equals $\left[\pi_i(1-\pi_i)\right]$. Therefore the ith diagonal element of the W matrix is $w_i = \frac{1}{\pi_i(1-\pi_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$. $\Delta = \left\{_d g_\mu(\mu_i)\right\}$ is an nxn diagonal matrix.

From this we see that W and Δ both depend on the mean $\underline{\mu_i}$. Therefore W, Δ and $\underline{\mu}$ (or $\underline{\pi}$) all involve the unknown parameter $\underline{\beta}$ through the link function. Because the maximum likelihood equation is a nonlinear function of $\underline{\beta}$ we cannot solve this equation analytically. Therefore we use the Fisher scoring method.

2.3 FISHER SCORING METHOD

Solutions for the maximum likelihood equation for $\underline{\beta}$ are performed by an iterative weighted least squares method. This can be derived as an example of the use of Fisher scoring. "Fisher scoring, the method used by SAS, is an iterative method for maximizing a likelihood and it takes on the form:

$$\theta^{m+1} = \theta^{(m)} + I\left(\theta^{(m)}\right)^{-1} \frac{\partial l}{\partial \theta}\Big|_{\theta=\theta^{(m)}}$$

where (m) represents the mth iteration, $I(\theta)$ is the information matrix and θ is the entire parameter vector" [8].

If we rewrite the previous equation in the context of our situation we have:

$$\underline{\beta}^{(m+1)} = \underline{\beta}^{(m)} + \left(X'W^{(m)}X\right)^{-1} X'W^{(m)}\Delta^{(m)}\left(\underline{y} - \underline{\mu}^{(m)}\right)$$

For this iterative procedure we need starting value estimates for the parameter $\underline{\beta}^{(0)}$. Once we have this value we can obtain starting values for μ, W, Δ (denoted by $\mu^{(0)}, W^{(0)}, \Delta^{(0)}$). These are all part of our iterative equation.

We then use these estimates to revise our equation to solve for the next parameter estimate. We linearize the model about these new values and linear least squares is applied again to find a second set of estimates. This procedure is repeated until the desired degree of convergence is obtained.