Study about Probit and Logit model to analysis Dose Response data.

Project report of the independent study submitted as a partial fulfilment for the Bachelor of Science (Special) Degree

By

W.J.Jannidi
SC / 2010 / 7623

Supervisor:
Prof. L. A. L. W. Jayasekara

Department of Mathematics
University of Ruhuna
Matara.

January 2015

Declaration

This thesis is the retribution of the independent study carried out by W.J.Jannidi for B.Sc.Special Degree in Mathematics Part (II) (2014) under the supervision and guidance of Prof. L. A. L. W. Jayasekara, Department of Mathematics, University of Ruhuna, Sri Lanka.

....

Date W.J.Jannidi

I certify that this statement is correct.

....

Date Prof. L. A. L. W. Jayasekara

Acknowledgment

I would like to offer my gratitude to all those who gave me the strength to complete this project.

First and foremost, I would like to express my greatest and heartful grace to Prof:L.A.L.W.Jayasekara, Department of Mathematics, University of Ruhuna for supervising me. Furthermore I would like to thank Prof: L.A.L.W.Jayasekara for the perfect knowledge about Statistics, clear instructions and great encouragement gave me. He continually and convincingly conveyed a spirit of adventure in regard to the research. Without his guidance and persistent help this dissertation would not have been possible.

I also offer my special gratitude to Prof :(Mrs)W.T.S.D.Premachandra, Department of Zoology, University of Ruhuna for giving me the data set to this research. I like to thank Mr.M.P.A.Wijayasiri; Head of the Department of Mathematics, Dr. N. Yapage and all the members of the academic staff for their precious advices and support.

I would also like to thank to all Mr/Miss demonstrators and all non academic members for their powerful helps.

A special thanks to my family. Words cannot express how grateful I am to my golden father, mother and my siblings for all of the sacrifices that you have made on my behalf. Your prayer for me was what sustained me thus far.

Finally, I would like to acknowledge to all of my friends who supported me to succeed my goal. A very special thanks goes to lord Ganesha for everlasting benisons.

Abstract

Dose-response data are widely used in many fields of research including toxicology,medicine, laboratory animals research, economics, sociology, and genetics. Probit and Logit models are usually used to analyze these data. This article discusses a study about the statistical background of the Probit and Logit model and provides examples for both models and these examples have been analyzed using R-statistical software. Finally, an application of dose-response data which related to the laboratory experiment is introduced by this text. The experiment has been done for find a best botanical to reduce root-knot nematodes. Probit and Logit model of experimental data, interpretation of the models, are contain in this thesis.Also, using statistical models and some other measurements of toxicology, different botanicals are compaired to find out the best one.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This introductive chapter will draw the idea and the context where this research has been developed and it will give an idea of the structure of the whole thesis.

1.1 Background of the study

The purpose of this thesis is study about statistical models for analyze dose response data and apply these methods in a real world application.

The outcome od dose-response experiments is the effect or stimulus on an arganism in response to a dose administered.In this context dose can refer to any biological, chemical, radioactive stimulus, or to any other tangible stimuli that ca be graduated. Major application areas of dose response data include agriculture, biology, chemisty, medicine, pharmacology and toxicology[12].

In statistic, logit and probit models are used to analysis dose-response data and carry out a regression model for these data.Statistical background of logit and probit models has been introduced in this thesis.

LC50 value is a very special role in dose response data.It is useful to compare different chemicals, drugs etc. LC50 values of different chemicals can obtain from both logit and probit models of each chemical.A important and significant study of LC50 value has been comprised in this dissertation. The laboratory experiment of dose-response data contains in this thesis and build a logit model, a probit model and obtain LC50 values for this experimental data as a example of real world application of dose-response data.

1.2 Objectives

1.2.1 Major Object

Study Logit and Probit statistical methods to analysis dose-response data

1.2.2 Specific objects

- Apply ststistical methods of dose-response data to the experimental data.

Under this experiment

- Find LC50 values of each botanicals.
- Build logit and probit models.
- Identify the best effective botanical.

1.3 Chapter Organization

Chapter 1 of this study introduced the background of the study and described the objectives of the study.

Chapter 2 presents a review of dose-response data and describes the relationship between dose and response.

Chapter 3 presents the statistcal methods to analysis dese-response data.Logit and probit models included in this chapter.

Chapter 4 contains definition, applications and importance of LC50 value and methodology of LC50 value.

Chapter 5 contains real world application of dose-response data.

Chapter 6 offers a summary and discussion of the researchers findings, implications for practice, and recommendations for future research.

Chapter 2

Dose-response data

2.1 Definition

A quantity of a medicine or drug taken or recommended to be taken at a particular time is called a dose and any action or change of condition evoked by a stimulus is named as a response.

2.2 Dose-response relationship

The dose-response relationship describes the change in effect on an organism caused by differing levels of doses to a stressor (usually a chemical) after a certain exposure time. This may apply to individuals or to populations. Generally dose-response relationship depend on the exposure time and exposure route.

2.3 Dose-response curve

A dose-response curve is a simple X-Y graph relating the magnitude of a stressor (e.g. concentration of a pollutant, amount of a drug, temperature, intensity of radiation) to the response of the receptor (e.g. organism under study). The response may be a physiological or biochemical response, or even death (mortality), and thus can be counts (or proportion, e.g. mortality rate), ordered descriptive categories (e.g. severity of a lesion), or continuous measurements (e.g. blood pressure). A number of effects (or endpoints) can be studied, often at different organizational levels. Also the dose-response curve is a valuable tool to understand the levels at which substances begin to exert adverse effects and the degree of harm expected at various levels.

The measured dose is generally plotted on the X axis and the response is plotted on the Y axis. In some cases, it is the logarithm of the dose that is plotted on the X axis, and in such cases the curve is typically sigmoidal, with the steepest portion in the middle. Biologically based models using dose are preferred over the use of log(dose) because the latter can visually imply a threshold dose when in fact there is none. The first point along the graph where a response above zero (or above the control response) is reached is usually referred to as a threshold-dose. For most beneficial or recreational drugs, the desired effects are found at doses slightly greater than the threshold-dose. At higher doses, undesired side effects appear and grow stronger as the dose increases.

2.3.1 Shape of the curve

A standard dose-response curve is defined by four parameters.

- The baseline response(Bottom)
- The maximum response(Top)
- The slope, and
- The drug concentration that provokes a response halfway between baseline and maximum(LC50)

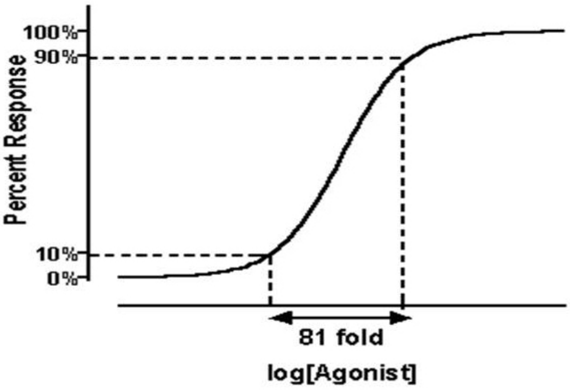


Figure 2.1: Dose-response curve(1).

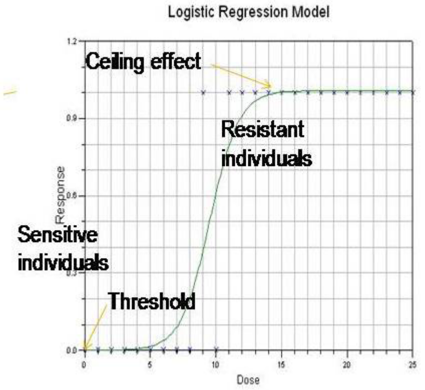


Figure 2.2: Dose-response curve(2).

Ceiling effect: no difference once all individuals are affected.

Threshold: A dose below which there are no adverse effects from exposure to chemicals.

2.3.2 Graded dose-response curve

Graded dose-response curves are constructed for response that are measured on continuous scale(efficacy). Relationship of dose and response can be plotted on curve (x-axis=dose, y-axis=response).When dose on arithmetic scale, curve is hyperbolic (not linear relationship) and dose on log scale, curve is sigmoid-shaped (semi log dose-response curve). e.g. heart rate Graded dose response means slight increase of drug brings small increase in response.

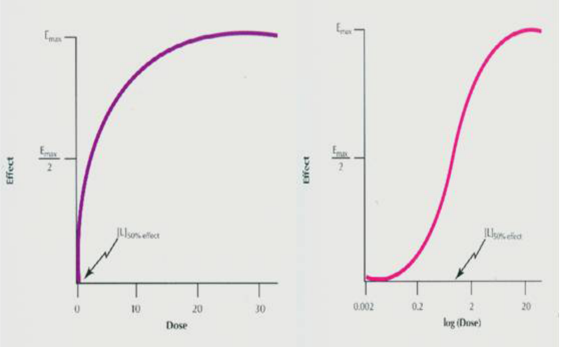


Figure 2.3: Graded dose-response curve.

2.3.3 Quantal dose-response curve

Response is an either/or event (potency) in quantal dose-response curve. Results can be plotted as Log dose-percentage curve and Gaussian distribution curve is obtained by keeping log doses on horizontal-axis and percentage of response on vertical-axis.

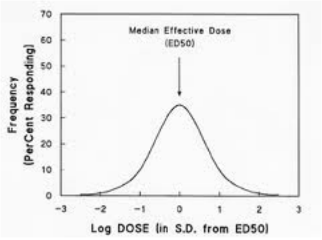


Figure 2.4: Quantal dose-response curve (1).

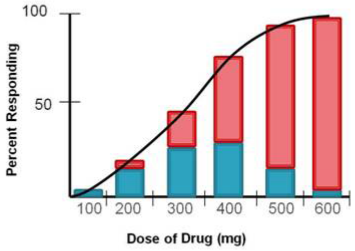


Figure 2.5: Quantal dose-response curve (2)

2.4 Dose-response curve information

Four important values can be identified from dose-response curves.

- Potency
- Efficacy
- Slope
- Variability

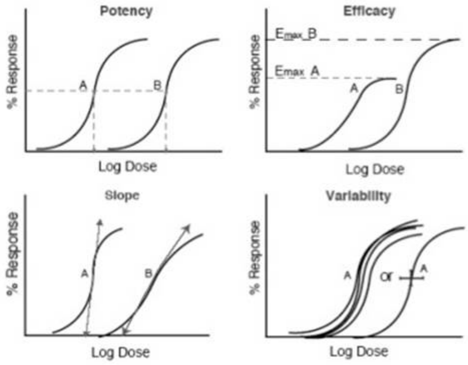


Figure 2.6: Information of dose-response curve

- Potency refers to the amount of drug necessary to produce a certain effect. A drug which produces a certain effect at 5mg dosage is ten times more potent than a drug which produces the same effect at 50mg dosage.
- Efficacy refers to the maximal response that can be obtained by a particular drug.
- Slope is effect of incremental increase in dose.
- Variability is reproductivity of data different for different organism.

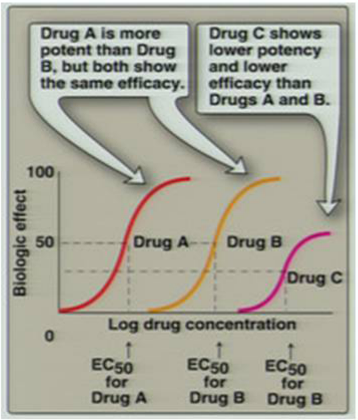


Figure 2.7: Potency and Efficacy of drugs.

2.5 NOAEL and LOAEL

NOAEL: No Observed Adverse Effect Level(highest data point at which there was not an observed toxic).

LOAEL: Low Observed Adverse Effect Level(lowest data point at which there was an observed toxic).

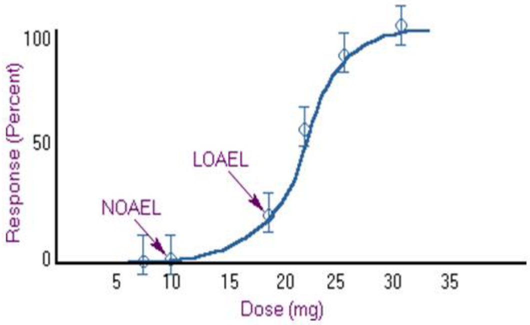


Figure 2.8: NOAEL and LOAEL

Chapter 3

Statistical methods for analysis the dose-response data

3.1 Probit model

The idea of probit analysis was originally published in Science by Chester Ittner Bliss in 1934. He worked as an entomologist for the Connecticut agricultural experiment station and was primarily concerned with finding an effective pesticide to control insects that fed on grape leaves. By plotting the response of the insects to various concentrations of pesticides, he could visually see that each pesticide affected the insects at different concentrations, i.e. one was more effective than the other. However, he didn't have a statistically sound method to compare this difference. The most logical approach would be to fit a regression of the response versus the concentration, or dose and compare between the different pesticides. Yet, the relationship of response to dose was sigmoid in nature and at the time regression was only used on linear data. Therefore, Bliss developed the idea of transforming the sigmoid dose-response curve to a straight line. In 1952, a professor of statistics at the University of Edinburgh by the name of David Finney took Bliss' idea and wrote a book called Probit Analysis. Today, probit analysis is still the preferred statistical method in understanding dose-response relationships.

Probit analyze is used to analysis many kinds of dose-response or binomial response experiments in a variety of fields and commonly used in toxicology to determine the relative toxicity of pesticides to living organisms. This is done by testing the response of an organism under various concentrations of each of the pesticide in question and then comparing the concentrations at which one encounters a response. The response is always binomial (e.g. death/no death) and the relationship between the response and the various concentrations is always sigmoid. Probit analysis acts as a transformation from sigmoid to linear and then runs a regression on the relationship.

3.1.1 Probit model in binary outcome data

Probit regression, also called a probit model, is used to model binary(dichotomous) outcome variables. In probit model, the inverse standard normal distribution of the probability is modeled as a linear combination of the predictors.

i.e.

The probit model is represented as prob(y=1|x)=Φ(Xβ),where Φ indicates the cumulative standard normal probability distribution function.

$$F(\mathrm{X}\beta)=\Phi(\mathrm{X}\beta)=\int_{-\infty}^{X\beta}\Phi(\mathrm{z})\mathrm{dz}$$

where

$$\Phi(z) = \frac{1}{\sqrt{2\Pi}}e^{\frac{-z^2}{2}}$$

Let

$$X\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

Consider

$$\mathrm{Pr(y=1|x)}=\Phi(\mathrm{X}\beta)=\Phi(\beta_0 + \beta_1 x_1)$$

$$\Phi(\beta_0 + \beta_1 x_1) = \int_{-\infty}^{\beta_0+\beta_1 x_1} \frac{1}{\sqrt{2\Pi}}e^{\frac{-t^2}{2}}dt = \Pi$$

$$\beta_0 + \beta_1 x_1 = \Phi^{-1}(\Pi)$$

3.1.2 Likelihood contribution of probit model

Likelihood contribution for single observation

$$\Phi(x_i\beta) if y_i = 1 \text{ and } 1 - \Phi(x_i\beta) if y_i = 0$$

we can write this as

$$[\Phi(x_i\beta)]^{y_i}[1 - \Phi(x_i\beta)]^{1-y_i}$$

Log-likelihood function for sample of n observations

$$lnL(\beta) = \sum_{i=1}^{n}[y_i ln\Phi(x_i\beta) + (1 - y_i)ln(1 - \Phi(x_i\beta))]$$

$$\frac{\partial lnL(\beta)}{\partial\beta} = \sum_{i=1}^{n} \frac{dlnL(\Phi(x_i\beta))}{d\Phi(x_i\beta)} \frac{\partial\Phi(x_i\beta)}{\partial\beta}$$

$$\frac{\partial lnL(\beta)}{\partial\beta} = \sum_{i=1}^{n}\{\frac{y_i}{\Phi(x_i\beta)} - \frac{1-y_i}{1-\Phi(x_i\beta)}\}\Phi(x_i\beta)x_i{}'$$

$$\frac{\partial lnL(\beta)}{\partial\beta} = \sum_{i=1}^{n}\{\frac{y_i-\Phi(x_i\beta)}{\Phi(x_i\beta)(1-\Phi(x_i\beta))}\}\Phi(x_i\beta)x_i{}'$$

where $\Phi(z)$is the standard normal density function.

$$\frac{\partial^2 lnL(\beta)}{\partial\beta\partial\beta'} = -\sum_{i=1}^{n}\{\frac{\Phi(x_i\beta)^2}{\Phi(x_i\beta)(1-\Phi(x_i\beta))}\}x_i{}'x_i$$

MLE $\widehat{\beta}$is approximately normal distributed with

$$\widehat{\beta} \sim N(\beta, [E(\frac{\partial^2 lnL(\beta)}{\partial\beta\partial\beta'})]^{-1})$$

3.1.3 Interpretation of coefficients

$$\frac{\partial p}{\partial x_i} = F^{'}(x\beta)\beta_i$$

$$= \Phi(x\beta)\beta_i$$

Value of $x\beta$ is taken to be Z-value of a normal distribution and higher value of $x\beta$ means that event is more likely to happen.

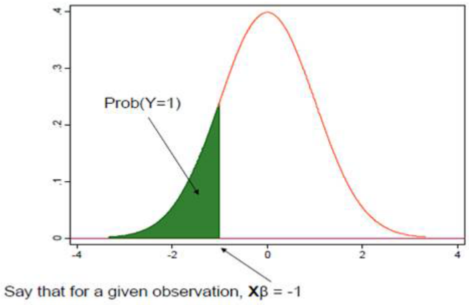


Figure 3.1: Density function of standard normal distribution-I

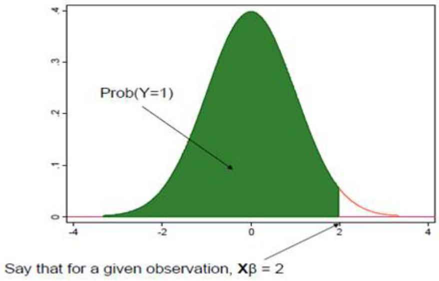


Figure 3.2: Density function of standard normal distribution-II

Density function of standard normal distribution -II A one unit change in x_i leads to a β_i change in the z score of y.

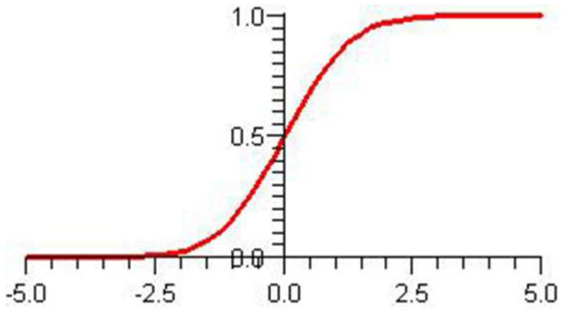


Figure 3.3: Conditional density function of standard normal distribution

Probit models represent the binomial probability pr(y=1) and pr(y=0) in terms of the standard normal conditional density function $\Phi(x\beta)$.

$$pr(y = 1) = \Phi(x\beta)$$

$$pr(y = 0) = 1 - \Phi(x\beta)$$

3.1.4 Marginal effects

Marginal effects reflect the change in the probability of y=1 given a one unit change in an independent variable x.

Two type of marginal effects.

1. Marginal index effects
2. Marginal probability effects

Marginal index effects

Marginal index effects are the partial effects of each explanatory variable on the probit index function $x_i\beta$.

Case 1: x_i is a continuous

Marginal index effect of variable $x_i = \frac{\partial E(y_i|x_i)}{\partial x_i}$

$$= \frac{\partial x_i\beta}{\partial x_i}$$

$$= \beta_i$$

Case 2: x_i is a binary variable

Marginal index effect of variable x_i = value of $x\beta$ when $x_i = 1$ and other regressor are fixed. or

value of $x\beta$ when $x_i = 0$ and other regressor equal fixed.

Marginal probability effects

Marginal probability effects are partial effects of each independent variables on the probability that the observed dependent variable $y_i = 1$.

Case 1: x is continuous

$Marginalprobabilityeffectofx = \frac{\partial pr(y=1)}{\partial x} = \frac{\partial \Phi(x\beta)}{\partial x}$

Using chain rule

$$= \frac{d\Phi(x\beta)}{d(x\beta)}\frac{\partial x\beta}{\partial x}$$

$$= \Phi(x\beta)\frac{\partial x\beta}{\partial x}$$

Case 2: x is binary

Marginal probability effect of x is difference between value of $\Phi(x\beta)$ when $x = 1$ and other regressor equal fixed and value of $\Phi(x\beta)$ when $x = 0$ and other regressor equal the same fixed.

i.e.

$\Phi(x_1\beta) - \Phi(x_0\beta)$

Compare the MIE and MPE of a continuous explanatory variable x_i

$$\text{MIE } x_i = \frac{\partial x\beta}{\partial x_i}$$

$$\text{MPE } x_i = \frac{\partial \Phi(x\beta)}{\partial x_i} = \Phi(x\beta)\frac{\partial x\beta}{\partial x_i}$$

Relationship: for a continuous explanatory variable x_i, the MPE is proportional to the MIE of x_i where the factor of proportionality is the standard normal p.d.f. of $X\beta$

MPE of $x_i = \Phi(x\beta)$MIE of x_i

Example

Let $X\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 {x_2}^2 + \beta_4 x_3 + \beta_5 D + \beta_6 D x_3 + e$

Where $x_1, x_2, x_3$are continuous and D is a binary variable.

D=1 when observation exhibits some attribute

D=0 otherwise

MIE of $x_1 = \frac{\partial (x\beta)}{\partial x_1} = \beta_1$

MPE of $x_1 = \Phi(X\beta)\beta_1$

MIE of $x_2 = \frac{\partial(x\beta)}{\partial x_2} = \beta_2 + 2\beta_3 x_2$

MPE of $x_2 = \Phi(X\beta)\frac{\partial(x\beta)}{\partial x_2} = \Phi(X\beta)(\beta_2 + 2\beta_3 x_2)$

MIE of $x_3 = \frac{\partial(x\beta)}{\partial x_3} = \beta_4 + \beta_6 D$

$= \beta_4 + \beta_6$ when D=1

$= \beta_4$ when D=0

MPE of $x_3 = \Phi(X\beta)\frac{\partial(X\beta)}{\partial x_3} = \Phi(X\beta)(\beta_4 + \beta_6 D)$

$= \Phi(X\beta)(\beta_4 + \beta_6)$ if D=1

$= \Phi(X\beta)\beta_4$ if D=0

MIE of D

when D=1

$X\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 {x_2}^2 + \beta_4 x_3 + \beta_5 + \beta_6 x_3$

when D=0

$X\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 {x_2}^2 + \beta_4 x_3$

Index function difference=$(X\beta)_{D=1} - (X\beta)_{D=0} = \beta_5 + \beta_6 x_3$

MIE of D=$\beta_5 + \beta_6 x_3$

MPE of D=$\Phi(X\beta)_{D=1} - \Phi(X\beta)_{D=0}$

3.1.5 Goodness of fit test

Goodness of fit may be judged by McFaddens Pseudo R^2. It is a measure for proximity of the model to the observed data.

Comparison of the estimated model with a model which only contains a constant as independent variable.

- $ln\widehat{L}(M_{full})$:Likelihood of model of interest
- $ln\widehat{L}(M_{intercept})$:Likelihood with all coefficients except that of the intercept restricted to zero
- It always holds that $ln\widehat{L}(M_{full}) \geqslant ln\widehat{L}(M_{intercept})$

The Pseudo R^2 is defined as:

$$PseudoR^2 = R^2_{McF} = 1 - \frac{ln\widehat{L}(M_{full})}{ln\widehat{L}(M_{intercept})}$$

Similar to the R^2 of the linear regression model, it holds that $0 \leq R^2_{McF} \leq 1$.

An increasing Pseudo R^2 may indicate a better fit of the model, whereas no simple interpretation like for the R^2 of the linear regression model is possible.

3.1.6 Examples

Ex:1

Analysis of the effect of a new teaching method in economic sciences

- Grade: Dependent variable. Indicates whether a student improved his grades after the new teaching method PSI had been introduced (0 = no, 1 = yes).

- PSI: Indicates if a student attended courses that used the new method (0 = no, 1 = yes).
- GPA: Average grade of the student
- TUCE: Score of an intermediate test which shows previous knowledge of a topic.

Table 3.1: Source: Spector, L. and M. Mazzeo, Probit Analysis and Economic Education. In:Journal of Economic Education, 11, 1980, pp.37-44

GPA	TUCE	PSI	Grade	GPA	TUCE	PSI	Grade
2.66	20	0	0	2.89	22	0	0
3.28	24	0	0	2.92	12	0	0
4	21	0	1	2.86	17	0	0
2.76	17	0	0	2.87	21	0	0
3.03	25	0	0	3.92	29	0	1
2.63	20	0	0	3.32	23	0	0
3.57	23	0	0	3.26	25	0	1
3.53	26	0	0	2.74	19	0	0
2.75	25	0	0	2.83	19	0	0
3.12	23	1	0	3.16	25	1	1
2.06	22	1	0	3.62	28	1	1
2.89	14	1	0	3.51	26	1	0
3.54	24	1	1	2.83	27	1	1
3.39	17	1	1	2.67	24	1	0
3.65	21	1	1	4	23	1	1
3.1	21	1	0	2.39	19	1	1

Probit regression model is:

$p(Grade = 1) = \Phi(\beta_0 + \beta_1 GPA + \beta_2 TUCE + \beta_3 PSI)$.

Also we can find,

$p(Grade = 0) = 1 - \Phi(\beta_0 + \beta_1 GPA + \beta_2 TUCE + \beta_3 PSI)$

Table 3.2: Probit regression model Ex:1

```
call:
glm(Grade~GPA+TUCE+PSI,binomial(link=”probit”))

Deviance Residuals:
 Min                                 1Q        Median    3Q        Max
 -1.9392                             -0.6508   -0.2229   0.5934    2.0451

 Coefficients:
                                     Estimate  Std. Error  z value  Pr(> |z|)
 (Intercept)                         -7.45231  2.57152     -2.898   0.00376 **
 GPA                                 1.62581   0.68973     2.357    0.01841 *
 TUCE                                0.05173   0.08119     0.637    0.52406
 PSI                                 1.42633   0.58695     2.430    0.01510 *

 Null deviance: 41.183 on 31 degrees of freedom
 Residual deviance: 25.638on 28 degrees of freedom
 AIC: 33.638

 The log likelihood function is -12.818803 (df=4)
```

Table 3.3: Average marginal effects of the model and their interpretation.

Variable	Estimated marginal effect	Interpretation
GPA	0.36078701	If the average grade of a student goes up by an infinitesimal amount, the probability for the variable grade taking the value one rises by 36.07٪.
TUCE	0.01147916	Analog to GPA,with an increase of 1.1٪.
PSI	0.31651968	If the dummy variable changes from zero to one, the probability for the variable grade taking the value one rises by 31.65٪.

Ex:2

The table(3.4) contains a set of well-known data from Bliss (1935) showing the results of experiments in which beetles were exposed to different concentrations of carbon disulphide. The data file shows the dose, the number of beetles exposed, and the number of beetles killed.

Table 3.4: Experimental data

Dose	Exposed	Killed	Alive
1.6907	59	6	53
1.7242	60	13	47
1.7552	62	18	44
1.7842	56	28	38
1.8113	63	52	11
1.8369	59	53	6
1.861	62	61	1
1.8839	60	60	0

For this data, the dependent variable Y is the proportion of exposed beetles at each dose that died, calculated by Y = Exposed / Alive. There is a single predictor variable X = Dose. There are a total of n = 481 subjects.

Table 3.5: Probit regression model Ex:2

```
call:
glm(y ~ dose,binomial(link="probit"))
Deviance Residuals:
Min                       1Q        Median    3Q        Max
-1.5714                   -0.4703   0.7501    1.0632    1.3449
Coefficients:
                          Estimate  Std. Error  z value  Pr(> |z|)
(Intercept)               -34.935   2.648       -13.19   <2e-16 ***
dose                      19.728    1.487       13.27    <2e-16 ***

Null deviance: 284.20 on 7 degrees of freedom
Residual deviance:10.12 on 6 degrees of freedom
AIC: 40.318
```

The fitted model for the sample data is

$$p(killed) = \Phi(-34.935 + 19.728Dose)$$

$pseudoR^2 = 1 - \frac{Residualdeviance}{Nulldeviance} = 1 - \frac{10.12}{284.20} = 0.96439$

The regression explains about 96.44٪of the deviance of a model without Dose. The P-value for Dose is very small, indicating that it is a statistically significant predictor for the proportion of beetles Killed.

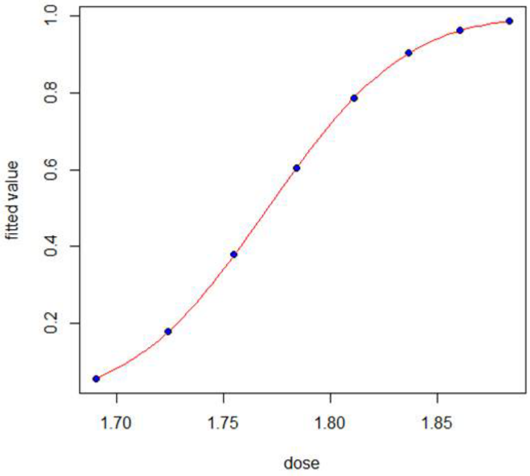


Figure 3.4: Plot of Fitted Model

3.2 Logit model

Logistic model is appropriate when the response takes one of only two possible values representing success and failure, or more generally the presence or absence of an attribute of interest.Logistic regression sometimes called the logistic model or logit model, analyzes the relationship between multiple independent variables and a categorical dependent variable, and estimates the probability of occurrence of an event by fitting data to a logistic curve. There are two models of logistic regression, binary logistic regression and multinomial logistic regression. Binary logistic regression is typically used when the dependent variable is dichotomous and the independent variables are either continuous or categorical. When the dependent variable is not dichotomous and is comprised of more than two categories, a multinomial logistic regression can be employed.

Example

The following figure shows 30 day mortality in a sample of septic patients as a function of their baseline APACHE II Score. Patients are coded as 1 or 0 depending on whether they are dead or alive in 30 days, respectively.

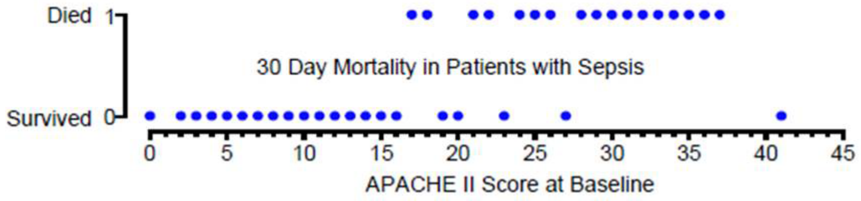


Figure 3.5: APACHE II Score and Mortality

Note that linear regression would not work well here since it could produce probabilities less than zero or greater than one.

We use the quantity $\Pi(x) = E(Y|x)$ to represent the conditional mean of Y given x when the logistic distribution is used. The specific form of the logistic regression model we use is:

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

A transformation of $\Pi(x)$ that is central to our study of logistic regression is the logit transformation. This transformation is defined, in terms of $\Pi(x)$, as:

$$g(x) = ln[\frac{\Pi(x)}{1-\Pi(x)}]$$

$$= \beta_0 + \beta_1 x$$

The importance of this transformation is that g(x) has many of the desirable properties of a linear regression model. The logit, g(x), is linear in its parameters, may be continuous, and may range from$-\infty to + \infty$, depending on the range of x.

Logistic regression equation describes a family of sigmoidal curves, three examples of which are given below.

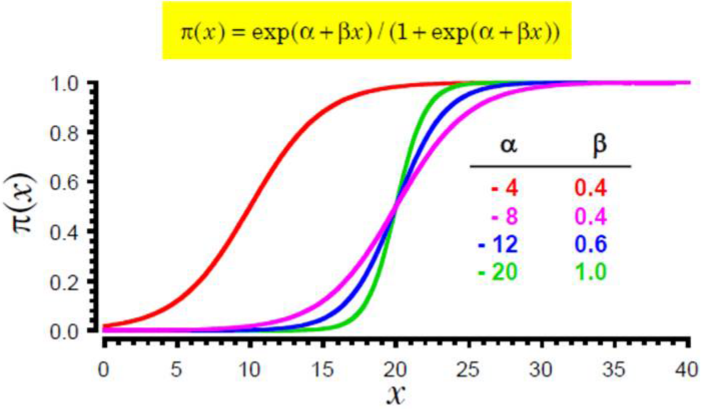


Figure 3.6: Parameter values and the shape of the regression curve

Consider

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

For now assume that $\beta_1 > 0$

For negative values of x,$e^{\beta_0+\beta_1 x} \rightarrow 0 as x \rightarrow -\infty$

and hence $\Pi(x) \rightarrow 0/(1+0) = 0$

For very large value of x, $e^{\beta_0+\beta_1 x} \rightarrow \infty$ and hence $\Pi(x) \rightarrow \infty/(1+\infty) = 1$

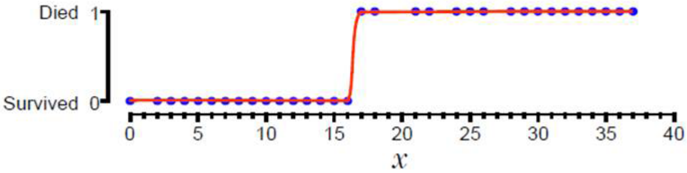
When $x = -\beta_0/\beta_1, \beta_0 + \beta_1 x = 0$ and hence $\Pi(x) = 1/(1+1) = 0.5$

The slope of $\Pi(x)$ when $\Pi(x) = 0.5$ is β_1/β_0

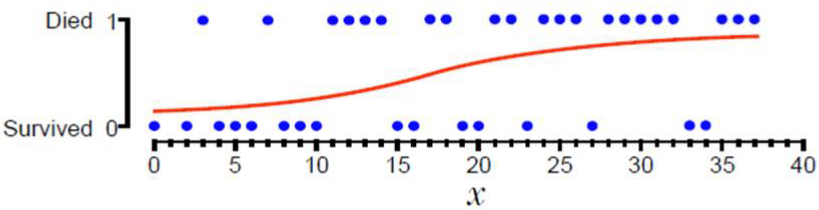
Thus β_1 controls how fast $\Pi(x)$ rises from 0 to 1.

For given β_1, β_0controls were the 50%survival point is located.

Data that has a sharp survival cut off point between patients who live or die should have a large value of β_1.



Data with a lengthy transition from survival to death should have a low value of β_1.



- **The Logistic Curve**

Logistic regression is a method for fitting a regression curve, $y = f(x)$, when y consists of binary coded (0, 1- -failure, success) data. When the response is a binary (dichotomous) variable and x is numerical, logistic regression fits a logistic curve to the relationship between x and y. Logistic curve is an S-shaped or sigmoid curve, often used to model population growth (Eberhardt and Breiwick, 2012). A logistic curve starts with slow, linear growth, followed by exponential growth, which then slows again to a stable rate.

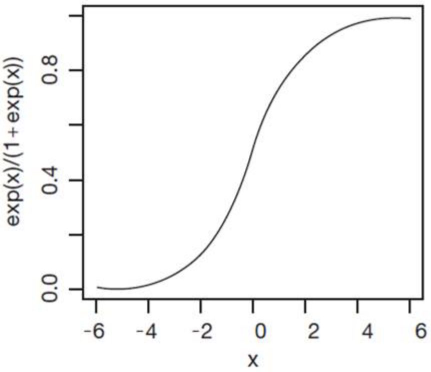


Figure 3.7: Logistic Curve

3.2.1 Fitting the logistic regression model

Consider a sample of n independent observations of the pair (x_i, y_i), i=1,2,.....n where y_i denotes the value of a dichotomous outcome variable and x_i is the value of the independent variable for the ith subject. Furthermore, assume that the outcome variable has been coded as 0 or 1, representing the absence or the presence of the characteristic, respectively. This coding for a dichotomous outcome is used throughout the text. Fitting the logistic regression model in logistic equation to a set of data requires that we estimate the values of β_0 and β_1, the unknown parameters.

The least squares method is used for estimate unknown parameters in linear regression.In that method we choose those values of β_0 and β_1 that minimize the sum-of-squared deviations of the observed values of Y from the predicted values based on the model. Under the usual assumptions for linear regression the method of least squares yields estimators with a number of desirable statistical properties.But, when the method of least squares is applied to a model with a dichotomous outcome, the estimators no longer have these same properties.

When the error are normally distributed, the general method of estimation that leads to the least squares function under the linear regression model is called maximum likelihood.This method provides the foundation to estimation with the logistic regression model.In a general sense, the method of maximum likelihood yields values for the unknown parameters that maximize the probability of obtaining the observed set of data.We must first construct a function, called the likelihood function to apply this method.

Likelihood function expresses the probability of the observed data as a function of the unknown parameters. The maximum likelihood estimators of the parameters are the values that maximize this function. Thus, the resulting estimators are those that agree most closely with the observed data.

$$Y = \begin{cases} 1 & \text{characteristic is presence} \\ 0 & \text{characteristic is absence} \end{cases}$$

$Pr(Y = 1|x) = \Pi(x)$ and $Pr(Y = 0|x) = 1 - \Pi(x)$ where,

$\Pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$ for arbitrary value of $\beta = (\beta_0, \beta_1)$.

Thus, for those pairs (x_i, y_i), where $y_i = 1$, the contribution to the likelihood function is $\Pi(x_i)$, and for those pairs where $y_i = 0$, the contribution to the likelihood function is $1 - \Pi(x_i)$, where the quantity $\Pi(x_i)$ denotes the value of $\Pi(x)$ computed at x_i.

The likelihood function for the pair (x_i, y_i) is

$$\Pi(x_i)^{y_i}[1 - \Pi(x_i)]^{1-y_i} \tag{3.1}$$

As the observations are assumed to be independent, the likelihood function is obtained as the product of the terms given in equation (3.1) as follows:

$$l(\beta) = \prod_{i=1}^{n} \Pi(x_i)^{y_i}[1 - \Pi(x_i)]^{1-y_i} \tag{3.2}$$

The loglikelihood is defined as:

$$L(\beta) = ln[l(\beta)] = \sum_{i=1}^{n} \{y_i ln[\Pi(x_i)] + (1 - y_i) ln[1 - \Pi(x_i)]\} \tag{3.3}$$

To find the value of β that maximizes $L(\beta)$ we differentiate $L(\beta)$ with respect to β_0 and β_1 and set the resulting expressions equal to zero. These equations, known as the likelihood equations, are

$$\sum [y_i - \Pi(x_i)] = 0 \tag{3.4}$$

and

$$\sum x_i[y_i - \Pi(x_i)] = 0 \tag{3.5}$$

In linear regression, the likelihood equations, obtained by differentiating the sum-of-squared deviations function with respect to β are linear in the unknown parameters and thus are easily solved. For logistic regression the expressions in equations (3.4) and (3.5) are nonlinear in β_0 and β_1, and thus require special methods for their solution. These methods are iterative in nature and have been programmed into logistic regression software.

The value of β given by the solution to equations (3.4) and (3.5) is called the maximum likelihood estimate and is denoted as $\widehat{\beta}$.

The sum of the observed values of y is equal to the sum of the predicted (expected) values. That is,

$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{\Pi}(x_i)$

3.2.2 Testing for the significance of the coefficients

After estimating the coefficients, our first look at the fitted model commonly concerns an assessment of the significance of the variables in the model. This usually involves formulation and testing of a statistical hypothesis to determine whether the independent variables in the model are significantly related to the outcome variable.

We can use different methods for test the significance of the coefficients.

1. **Likelihood ratio test**

   In logistic regression, comparison of observed to predicted values is based on the log-likelihood function defined in equation (3.3). To better understand this comparison, it is helpful conceptually to think of an observed value of the response variable as also being a predicted value resulting from a saturated model. A saturated model is one that contains as many parameters as there are data points.

   The comparison of observed to predicted values using the likelihood function is based on the following expression:

$$D = -2ln[\frac{(likelihood \quad of \quad the \quad fitted \quad model)}{(likelihood \quad of \quad the \quad saturated \quad model)}] \tag{3.6}$$

   Otherwise, $D = -2ln[Likelihood \quad Ratio]$

   Using equation (3.3), equation (3.6) becomes

$$D = -2\sum_{i=1}^{n}[y_i ln(\frac{\widehat{\pi_i}}{y_i}) + (1-y_i)ln(\frac{1-\widehat{\pi_i}}{1-y_i})] \tag{3.7}$$

where $\widehat{\pi_i} = \widehat{\pi}(x_i)$.

The statistic, D, in equation (3.7) is called the deviance, and for logistic regression, it plays the same role that the residual sum-of-squares plays in linear regression.

In particular, to assess the significance of an independent variable we compare the value of D with and without the independent variable in the equation. The change in D due to the inclusion of the independent variable in the model is:

G = D(model without the variable) - D(model with the variable)

This statistic, G, plays the same role in logistic regression that the numerator of the partial F-test does in linear regression. Because the likelihood of the saturated model is always common to both values of D being differenced, G can be expressed as,

$$G = -2ln[\frac{(likelihood \quad without \quad the \quad variable)}{(likelihood \quad with \quad the \quad variable)}] \tag{3.8}$$

2. **Wald test**

With nonnull standard error SE of $\widehat{\beta}$, the test statistic

$z = \frac{\widehat{\beta}-\beta_0}{SE}$

has an approximate statndard normal distribution when $\beta = \beta_0$. Once refers z to the standard normal table to obtain one or two sided p-values. Equivalently, for the two sided alternative, z^2 has a chi-squared null distribution with 1 degree of freedom(df); the p-value is then the right-tailed chi-squared probability above the observed value. This type of statistic, using the nonnull standard error, is called a Wald statistic(Wald 1943).

TheWald test for logistic regression model is,

$$W_j = \frac{\widehat{\beta_j}}{SE(\widehat{\beta_j})} \tag{3.9}$$

3. **Score test**

The third method uses the score statistic, due to R.A.Fisher and C.R.Rao. The score test is based on the slope and expected curvature of the log-likelihood function $L(\beta)$ at the null value β_0.It utilizes the size of the score function

$u(\beta) = \partial L(\beta)/\partial\beta$, evaluated at β_0.

The value $u(\beta_0)$ tends to be lager in absoute value when $\widehat{\beta}$ is farther from β_0. Denote $-E[\partial^2 L(\beta)/\partial\beta^2]$ evaluated at β_0 by $\iota(\beta_0)$. The score statistic is the ratio of $u(\beta_0)$ to its null SE, which is $[\iota(\beta_0)]^{1/2}$. This has an approximate standard normal null disrtibution. The chi-squared from of the score statistic is

$$\frac{[u(\beta_0)]^2}{\iota(\beta_0)} = \frac{[\partial L(\beta)/\partial\beta_0]^2}{-E[\partial^2 L(\beta)/\partial\beta_0^2]} \tag{3.10}$$

where the partial derivative notation reflects derivatives with respect to β that are evaluated at β_0.

A comparison of the three tests

The following figure is a generic plot of a log-likelihood $L(\beta)$ for the univariate case. It illustrates the three tests of $H_0 : \beta = 0$. The wald test uses the behavior of $L(\beta)$ at the ML estimate $\widehat{\beta}$, having chi-squared form $(\widehat{\beta}/SE)^2$. The SE of $\widehat{\beta}$ depends on the curvature of $L(\beta)$ at $\widehat{\beta}$.

The score test is based on the slope and curvature of $L(\beta)$ at $\beta = 0$.

The likelihood-ratio test combines information about $L(\beta)$ at both $\widehat{\beta}$ and $\beta = 0$.

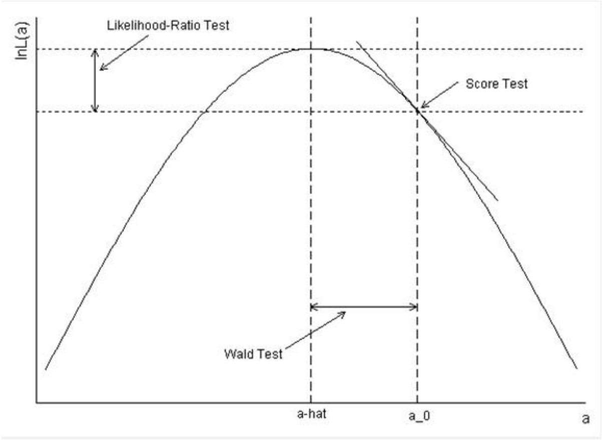


Figure 3.8: Log-likelihood function and information used in three test

3.2.3 Confidence interval estimation

The basis for construction of the interval estimators is the same statistical theory we used to formulate the tests for significance of the model. In particular, the confidence interval estimators for the slope and intercept are, most often, based on their respective Wald tests and are sometimes referred to as Wald-based confidence intervals. The endpoints of a $100(1-\alpha)$%confidence interval for the slope coefficient are

$$\widehat{\beta_1} \pm z_{1-\alpha/2}\widehat{SE}(\widehat{\beta_1}) \tag{3.11}$$

and for the intercept they are

$$\widehat{\beta_0} \pm z_{1-\alpha/2}\widehat{SE}(\widehat{\beta_0}) \tag{3.12}$$

where $z_{1-\alpha/2}$ is the upper $100(1-\alpha/2)$%point from the standard normal distribution and $\widehat{SE}$ denotes a model-based estimator of the standard error of the respective parameter estimator.

3.2.4 Multiple logistic regression model

Consider a collection of p independent variables denoted by the vector $x\prime = (x_1, x_2, ...x_p)$. For the moment we assume that each of these variables is at least interval scaled. Let the conditional probability that the outcome is present be denoted by $Pr(Y = 1|x) = \pi(x)$. The logit of the multiple logistic regression model is given by the equation

$$g(x) = ln(\frac{\pi(x)}{1 - \pi(x)}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \tag{3.13}$$

where, for the multiple logistic regression model,

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \tag{3.14}$$

Assume that we have a sample of n independent observations (x_i, y_i) , i=1,2...n.

As in the univariable case,

The likelihood equations that result may be expressed as follows:

$$\textstyle\sum_{i=1}^{n}[y_i - \pi(x_i)] = 0$$

and

$$\textstyle\sum_{i=1}^{n} x_{ij}[y_i - \pi(x_i)] = 0$$

for j=1,2,.....p

Partial derivatives of the log-likelihood function have the following general form

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i (1 - \pi_i) \tag{3.15}$$

and

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi_i (1 - \pi_i) \tag{3.16}$$

for j,l=0,1,2....p where π_i denotes $\pi(x_i)$

The estimated standard errors of the estimated coefficients, which we denote as

$$\widehat{SE}(\widehat{\beta_j}) = [\widehat{Var}(\widehat{\beta_j})]^{1/2} \tag{3.17}$$

3.2.5 Interpretation of the fitted logistic regression model

The interpretation of any fitted model requires that we be able to draw practical inferences from the estimated coefficients in the model. In most instances, the intercept coefficient is of little interest. The estimated coefficients for the independent variables represent the slope (i.e., rate of change) of a function of the dependent variable per unit of change in the independent variable. Thus, interpretation involves two issues: determining the functional relationship between the dependent variable and the independent variable, and appropriately defining the unit of change for the independent variable.

The first step is to determine what function of the dependent variable yields a linear function of the independent variables. This is called the link function [McCullagh and Nelder (1989), or Dobson (2002)]. In the case of a linear regression model, the link function is the identity function as the dependent variable, by definition, is linear in the parameters.

In the logistic regression model the link function is the logit transformation

$g(x) = ln\{\pi(x)/[1 - \pi(x)]\} = \beta_0 + \beta_1 x.$

In the logistic regression model, the slope coefficient is the change in the logit corresponding to a change of one unit in the independent variable. That is,

$$\beta_1 = g(x + 1) - g(x)$$

3.2.6 Dichotomous independent variable

We assume that the independent variable, x, is coded as either 0 or 1. The difference in the logit for a subject with x = 1 and x = 0 is

$$g(1) - g(0) = (\beta_0 + \beta_1 1) - (\beta_0 + \beta_1 0) = \beta_1$$

To provide a more meaningful interpretation,need to introduce the odds ratio as a measure of association.

3.2.7 Polychotomous independent variable

The independent variable which has k>2 categories with distinct values is called a polychotomous independent variable.For example, variables that denote the county of residence within a state, the clinic used for primary health care within a city, or race. Each of these variables has a fixed number of discrete values and the scale of measurement is nominal. Therefore, we must form a set of design variables to represent the categories of the variable. We can use Reference cell coding method or Deviation from means coding method for creating design variables for polychotomous independent variables.

Ex: Consider, Risk of disease categorized as Less,Same, and More.

- Reference cell coding method

RateRisk(Code)	D1	D2
Less(1)	0	0
Same(2)	1	0
More(3)	0	1

- Deviation from means coding method

RateRisk(Code)	D1	D2
Less(1)	-1	-1
Same(2)	1	0
More(3)	0	1

3.2.8 Contingency table

The joint distribution between two categorical variables determines their relationship. This distribution also determines the marginal and conditional distribution.

Let X and Y denote two categorical response variables, X with I categories and Y with J categories. Classification of subjects on both variables have IJ possible combinations. The responses (X,Y) of a subject chosen randomly from some population have a probability distribution. A rectangular table having I rows for categories of X and J columns for categories of Y displays this distribution. The cells of the table represent the IJ possible outcomes. When the cells contain frequency counts of outcomes for a sample, the table is called a contingency table, a term introduced by Karl Pearson (1904).

Let π_{ij} denote the probability that (X,Y) occurs in the cell in row i and column j. The marginal distributions are the row and column totals that result from summing the joint probabilities.

Table 3.6: Frequency counts for I ×J table

		Y					
		1	2	..	j	..	J
	1	$n_1 1$	$n_1 2$	..	$n_1 j$	..	$n_1 J$
	2	$n_2 1$	$n_2 2$	..	$n_2 j$	..	$n_2 J$
X		..			..		
	i	$n_i 1$	$n_i 2$	..	$n_i j$	..	$n_i J$
		..			..		
	I	$n_I 1$	$n_I 2$	..	$n_I j$	..	$n_I J$

Table 3.7: Cell probabilities for I ×J table

		Y						
		1	2	..	j	..	J	
	1	$\pi_1 1$	$\pi_1 2$	..	$\pi_1 j$	..	$\pi_1 J$	$\pi_1 +$
	2	$\pi_2 1$	$\pi_2 2$	..	$\pi_2 j$	..	$\pi_2 J$	$\pi_2 +$
X		..			..			
	i	$\pi_i 1$	$\pi_i 2$	..	$\pi_i j$	..	$\pi_i J$	$\pi_i +$
		..			..			
	I	$\pi_I 1$	$\pi_I 2$	..	$\pi_I j$	..	$\pi_I J$	$\pi_I +$
		$\pi_+ 1$	$\pi_+ 2$	..	$\pi_+ j$	..	$\pi_+ J$	

3.2.9 Odds Ratio

Odds

For a probability π of success, the odds are defined to be $\Omega = \pi/(1 - \pi)$.

The odds are nonnegative, with $\Omega > 1$ when a success is more likely than a failure.

Table 3.8: Values of the logistic regression model when the independent variable is dichotomous

	Independent Variable (x)	
Outcome Variable (y)	$x = 1$	$x = 0$
y=1	$\pi(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\pi(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
y=0	$1 - \pi(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$	$1 - \pi(0) = \frac{1}{1+e^{\beta_0}}$
Total	1.0	1.0

The possible values of the logistic probabilities from a model containing a single dichotomous covariate coded 0 and 1 are displayed in the 2 ×2 table,shown in Table 3.4. The odds of the outcome being present among individuals with x = 1 is $\pi(1)/[1 - \pi(1)]$. Similarly, the odds of the outcome being present among individuals with x = 0 is $\pi(0)/[1 - \pi(0)]$. The odds ratio, denoted OR (θ), is the ratio of the odds for x = 1 to the odds for x = 0, and is given by the equation

$$OR = \frac{\pi(1)/[1 - \pi(1)]}{\pi(0)/[1 - \pi(0)]} \tag{3.18}$$

Substituting the expressions for the logistic regression model probabilities in Table 3.4 into equation (3.18) we obtain

$$OR = \frac{(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}})/(\frac{1}{1+e^{\beta_0+\beta_1}})}{(\frac{e^{\beta_0}}{1+e^{\beta_0}})/(\frac{1}{1+e^{\beta_0}})} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{(\beta_0+\beta_1)-\beta_0} = e^{\beta_1} \tag{3.19}$$

Hence, for a logistic regression model with a dichotomous independent variable coded 0 and 1, the relationship between the odds ratio and the regression coefficient is

$OR = e^{\beta_1}$

The odds ratio is widely used as a measure of association as it approximates how much more likely or unlikely (in terms of odds) it is for the outcome to be present among those subjects with x = 1 as compared to those subjects with x = 0.

- **95%Confidence interval of odds ratio**

Odds ratio with 95%confidence interval (CI) can be used to assess the contribution of individual predictors (Katz, 1999). It is important to note however, that unlike the p value, the 95%CI does not report a measures statistical significance. It is used as a proxy for the presence of statistical significance if it does not overlap the null value (e.g. OR=1). The 95%CI is used to estimate the precision of the OR. A large CI indicates a low level of precision of the OR, whereas a small CI indicates a higher precision of the OR. An approximate confidence interval for the population log odds ratio is

95%CI for $ln(OR) = ln(OR) \pm 1.96SE[ln(OR)]$

where ln(OR) is the sample log odds ratio, and SE[ln(OR)] is the standard error of the log odds ratio(Morris and Gardner, 1988). Taking the antilog, we get the 95%confidence interval for the odds ratio:

95%CI for $OR = e^{lnOR \pm 1.96SE(lnOR)}$

3.2.10 Relative risk

In certain settings, the odds ratio can approximate another measure of association called the relative risk, which is the ratio of the two outcome probabilities,

$RR = \frac{\pi(1)}{\pi(0)}$

It follows from equation (3.18) that the odds ratio approximates the relative risk if $[1 - \pi(0)]/[1 - \pi(1)] \approx 1$.This holds when $\pi(x)$ is small for both x = 0 and x = 1, often referred to in medical/epidemiological research as the rare disease assumption.

- **Confounding variable**

Frequently whenwe are studying the relationship between the status of some disease and the status of some risk factor, we are aware of another variable that may be associated either with the disease, with the risk factor, or with both in such a way that the true relationship between the disease status and the risk factor is masked. Such a variable is called a confounding variable.

3.2.11 The Mantel-Haenszel statistic

Mantel-Haenszel procedure allows us to test the null hypothesis that there is no association between status with respect to disease and risk factor status. Initially used only with data from retrospective studies, the Mantel-Haenszel procedure is also appropriate for use with data from prospective studies as discussed by Mantel(23). In the application of the Mantel-Haenszel procedure, case and control subjects are assigned to strata corresponding to different values of the confounding variable. The data are then analyzed within individual strata as well as across all strata.

The data under analysis are from a retrospective or a prospective study with case and noncase subjects classified according to whether they have or do not have the suspected risk factor. The confounding variable is categorical, with the different categories defining the strata. If the confounding variable is continuous it must be categorized.

Table 3.9: Subjects in the ith stratum of a confounding variable classified according to status relative to a risk factor and whether they are cases or controls

	Sample		
Risk factor	Cases	Controls	Total
Presrnt	a_i	b_i	$a_i + b_i$
Absent	c_i	d_i	$c_i + d_i$
Total	$a_i + c_i$	$b_i + d_i$	n_i

Application of the Mantel-Haenszel procedure consists of the following steps.

1. From k strata corresponding to the k categories of the confounding variable.
2. For each stratum compute the expected frequency e_i.

   $e_i = \frac{(a_i+b_i)(a_i+c_i)}{n_i}$

3. For each stratum compute

   $v_i = \frac{(a_i+b_i)(c_i+d_i)(a_i+c_i)(b_i+d_i)}{n_i^2(n_i-1)}$

4. Compute the Mantel-Haenszel test statistic, as follows:

   $\chi^2_{MH} = \frac{(\sum_{i=1}^k a_i - \sum_{i=1}^k e_i)^2}{\sum_{i=1}^k v_i}$

5. Reject the null hypothesis of no association between disease status and suspected risk factor in the population if the computed value of χ^2_{MH} is equal to or greater than the critical value of the test statistic, which is the tabulated chi-square value for 1 degree of freedom and the chosen level of significance.

   We may compute the Mantel-Haenszel estimator of the common odds ratio,$\widehat{OR_{MH}}$, as follows:

   $\widehat{OR_{MH}} = \frac{\sum_{i=1}^k (a_i d_i/n_i)}{\sum_{i=1}^k (b_i c_i/n_i)}$

3.2.12 Examples

Ex:1

- Rizatriptan for Migraine

Response - Complete Pain Relief at 2 hours (Yes/No)

Predictor - Dose (mg): Placebo (0,2.5,5,10)

Dose	Patients	Relieved	%Relieved
0	67	2	3.0
2.5	75	7	9.3
5	130	29	22.3
10	145	40	27.6

Table 3.10: Logostic regression model

```
call:
glm(formula=y   dose , family=binomial(link="logit"))

Deviance Residuals:
               1          2           3         4
               1.6275     0.5096      -1.8960   0.6946

 Coefficients:
               Estimate   Std. Error  z value   Pr(> |z|)
 (Intercept)   -2.48981   0.28475     8.744     <2e-16 ***
 dose          0.16526    0.03712     -4.452    8.51e-06 ***
```

Wald test statistic of the model is:

```
Wald test
            X2    df  p(>X2)
Intercept   76.5  1   0
dose        19.8  1   8.5e-06
```

Coefficient of dose=$\widehat{\beta_1} = 0.16526$ and standard error of $\widehat{\beta_1} = 0.03712$.

95%CI: for $\widehat{\beta_1}$: $0.1651.960.037 = (0.0925, 0.2375)$

95%CI: for population odds ratio: $(e^{0.0925}, e^{0.2375}) = (1.10, 1.27)$

Conclude positive association between dose and probability of complete relief.

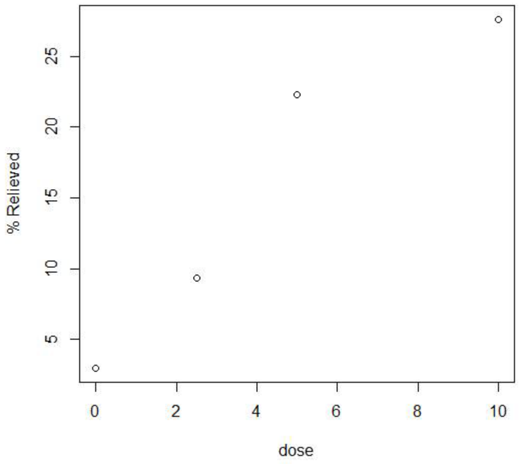


Figure 3.9: Dose vs. Percentage of Reilived

Ex:2

The data set is the Tax Increment Financing(TIF)data from http://www.stat.sc.edu

Table 3.11: Data set

y	income	y	income	y	income	y	income	y	income
0	9.2	0	12.9	0	9.2	1	9.6	0	12.5
0	9.3	1	10.1	0	9.4	1	10.3	1	13.5
0	9.5	1	10.9	0	9.5	1	10.9	1	13.2
0	9.5	1	11.1	0	9.6	1	11.1	0	12.3
0	9.7	1	11.5	0	9.7	1	11.1	1	13.1
0	9.8	1	11.8	0	9.8	1	11.9	0	12.1
0	9.9	1	12.1	0	10.5	1	12.2	1	12.9
0	10.9	1	12.6	0	10.5	1	12.5	0	11.7
0	11	1	12.6	0	11.2	1	12.6	1	12.9
0	11.2	1	12.9	0	11.5	1	12.9	0	11.8

Table 3.12: Logistic regression model Ex:2

```
call:
glm(y ~income,family=binomial(link=”logit”))
Deviance Residuals:
Min                                         1Q         Median      3Q        Max
-1.8781                                     -0.8021    -0.4736     0.8097    1.9461
Coefficients:
                                            Estimate   Std. Error  z value   Pr(> |z|)
(Intercept)                                 -11.3487   3.3511      -3.387    0.000708 ***
income                                      1.0019     0.2954      3.392     0.000695 ***

Null deviance:69.235 on 49 degrees of freedom
Residual deviance:53.666 on 48 degrees of freedom
AIC:57.666
```

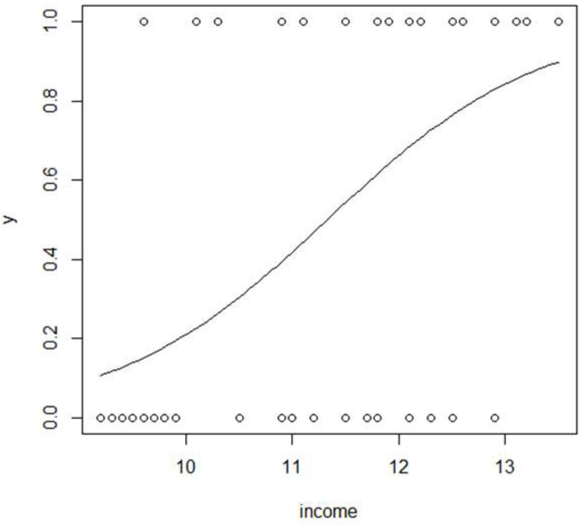


Figure 3.10: A plot of the data with the estimated logistic curve

Inference in logistic regression

- Getting the Likelihood Ratio(LR) test statistic and P-value

From Table(3.12), Likelihood ratio test statistic is 15.569. LR test p-value is 7.955886e-05. Null hypothesis of LR test is $H_0 : \beta_1 = 0$. Since the p-value is very small (<0.0001), we reject H_0 that is β_1 is not zero and conclude that income has a significant effect on the probability a city uses TIF.

Table 3.13: Analysis of Deviance Table

	Df	Deviance	Resid. Df	Resid. Dev
NULL			49	69.235
income	1	15.569	48	53.666

- Odds ratio and Confidence interval of odds ratio

Estimated odds ratio is 2.723401. An approximate 95%CI for the odds ratio is (1.526, 4.858).

Chapter 4

LC_{50} Value

4.1 Definition

LC stands for "Lethal Concentration". LC values usually refer to the concentration of a chemical in air but in environmental studies it can also mean the concentration of a chemical in water.

According to the Organization for Economic Cooperation and Development (OECD) guidelines for the testing of chemicals, a traditional experiment involves groups of animals exposed to a concentration (or series of concentrations) for a set period of time (usually 4 hours). The animals are clinically observed for up to 14 days.

The concentration of the chemical in air that kills 50%of the test animals during the observation period is the LC_{50} value. Other durations of exposure (versus the traditional 4 hours) may apply depending on specific laws.

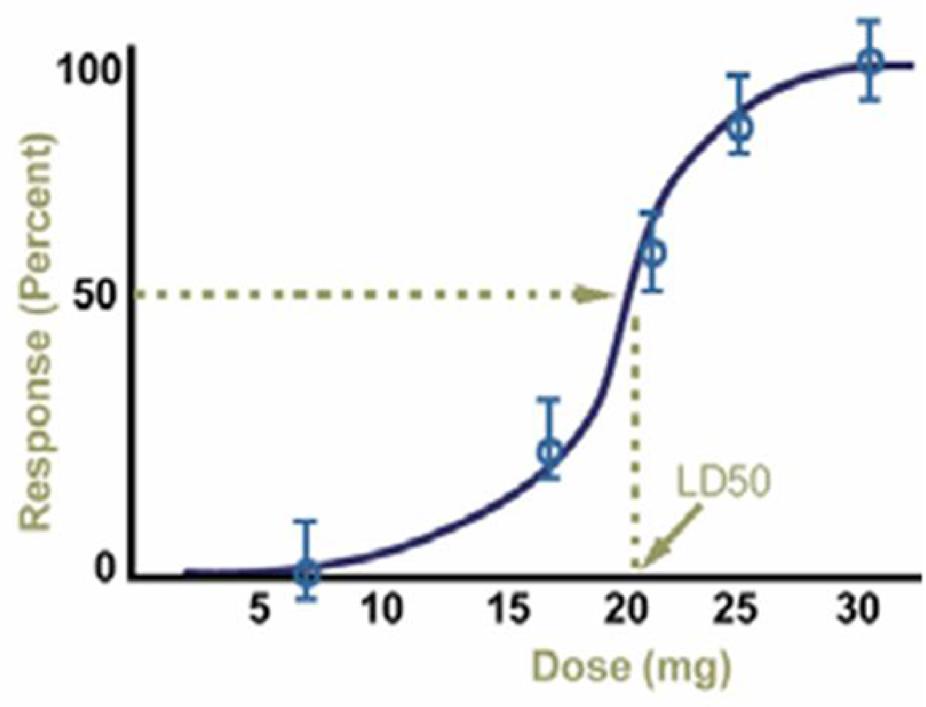


Figure 4.1: LC_{50} value from dose-response curve

4.2 Significance of the study

Chemicals can have a wide range of effects on our health. Depending on how the chemical will be used, many kinds of toxicity tests may be required.

Since different chemicals cause different toxic effects, comparing the toxicity of one with another is hard. We could measure the amount of a chemical that causes kidney damage, for example, but not all chemicals will damage the kidney. We could say that nerve damage is observed when 10 grams of chemical A is administered, and kidney damage is observed when 10 grams of chemical B is administered. However, this information does not tell us if A or B is more toxic because we do not know which damage is more critical or harmful.

Therefore, to compare the toxic potency or intensity of different chemicals, researchers must measure the same effect. One way is to carry out lethality testing (the LC50 tests) by measuring how much of a chemical is required to cause death. This type of test is also referred to as a "quantal" test because it is measures an effect that "occurs" or "does not occur".

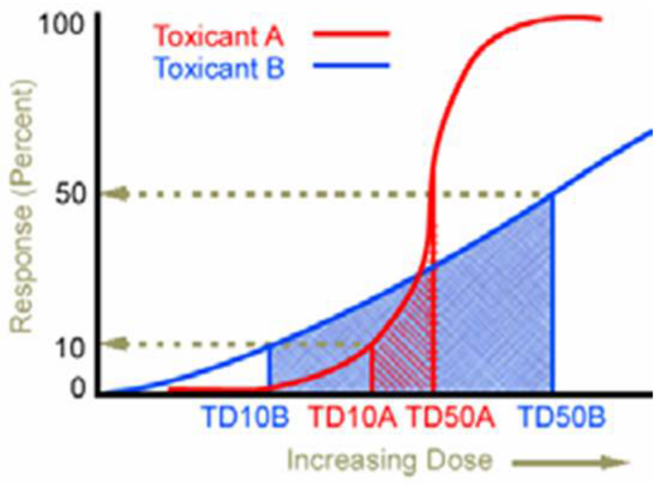


Figure 4.2: LC_{50} value of A and B

4.3 Becoming of the LC_{50}?

In 1927, J.W. Trevan attempted to find a way to estimate the relative poisoning potency of drugs and medicines used at that time. He developed the LC_{50} test because the use of death as a "target" allows for comparisons between chemicals that poison the body in very different ways. Since Trevan's early work, other scientists have developed different approaches for more direct, faster methods of obtaining the LC_{50}.

4.3.1 Acute toxicity

Acute toxicity is the ability of a chemical to cause ill effects relatively soon after one oral administration or a 4-hour exposure to a chemical in air. "Relatively soon" is usually defined as a period of minutes, hours (up to 24) or days (up to about 2 weeks) but rarely longer. So, LC_{50} value is considered as a measure of acute toxicity.

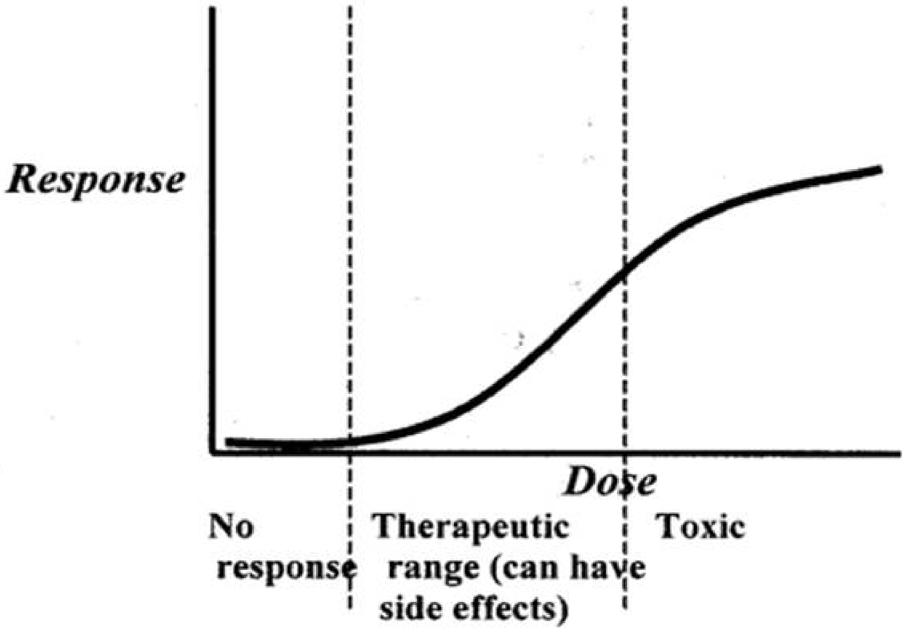


Figure 4.3: Toxic effects on Dose-response curve

4.4 Test of LC_{50}

In nearly all cases,LC_{50} tests are performed using a pure form of the chemical. Mixtures are rarely studied.

The chemical may be given to the animals by mouth (oral); by applying on the skin (dermal); by injection at sites such as the blood veins (i.v.- intravenous), muscles (i.m. - intramuscular) or into the abdominal cavity (i.p. - intraperitoneal).

The LC_{50} value obtained at the end of the experiment is identified as the LC_{50} (oral), LC_{50} (skin), LC_{50} (i.v.), etc., as appropriate. Researchers can do the test with any animal species but they use rats or mice most often. Other species include dogs, hamsters, cats, guinea-pigs, rabbits, and monkeys. In each case, the LC_{50} value is expressed as the weight of chemical administered per kilogram body weight of the animal and it states the test animal used and route of exposure or administration; e.g., LC_{50} (oral, rat) - 5 mg/kg, LC_{50} (skin, rabbit) - 5 g/kg. So, the example "LC_{50}(oral, rat) 5 mg/kg" means that 5 milligrams of that chemical for every 1 kilogram body weight of the rat, when administered in one dose by mouth, causes the death of 50%of the test group.

If the lethal effects from breathing a compound are to be tested, the chemical (usually a gas or vapour) is first mixed in a known concentration in a special air chamber where the test animals will be placed. This concentration is usually quoted as parts per million (ppm) or milligrams per cubic metre(mg/m^3). In these experiments, the concentration that kills 50%of the animals is called an LC_{50} (Lethal Concentration 50). When an LC_{50} value is reported, it should also state the kind of test animal studied and the duration of the exposure, e.g., LC_{50} (rat) - 1000 ppm/ 4 hr or LC_{50} (mouse) - $5mg/m^3/2hr$.

4.4.1 Which LC_{50} information is the most important for occupational health and safety purposes ?

Inhalation and skin absorption are the most common routes by which workplace chemicals enter the body. Thus, the most relevant from the occupational exposure viewpoint are the inhalation and skin application tests. Despite this fact, the most frequently performed lethality study is the oral LC_{50}. This difference occurs because giving chemicals to animals by mouth is much easier and less expensive than other techniques. However, the results of oral studies are important for drugs, food poisonings, and accidental domestic poisonings. Oral occupational poisonings might occur by contamination of food or cigarettes from unwashed hands, and by accidental swallowing.

4.4.2 Compare one LC_{50} value to another

In general, the smaller the LC_{50} value, the more toxic the chemical. The opposite is also true: the larger the LC50 value, the lower the toxicity.

The LC_{50} gives a measure of the immediate or acute toxicity of a chemical in the strain, sex, and age group of a particular animal species being tested. Changing any of these variables (e.g., type animal or age) could result in finding a different LC_{50} value. The LC50test was neither designed nor intended to give information on long-term exposure effects of a chemical.

Once you have an LC_{50} value, it can be compared to other values by using a toxicity scale. It is also important to know that the actual LC_{50} value may be different for a given chemical depending on the route of exposure (e.g., oral, dermal, inhalation).Differences in the LC_{50} toxicity ratings reflect the different routes of exposure. The toxicity rating can be different for different animals.

4.5 Use of LC_{50}

The LC_{50} can be used:

- As an aid in developing emergency procedures in case of a major spill or accident.
- To help develop guidelines for the use of appropriate safety clothing and equipment. For example, if the dermal LC_{50} value for a chemical is rated as extremely toxic, it is important to protect the skin with clothing, gloves (etc.) made of the right chemical-resistant material before handling. Alternatively, if a chemical has an inhalation LC_{50} value which indicates that it is relatively harmless, respiratory protective equipment may not be necessary (as long as the oxygen concentration in the air is in the normal range - around 18٪).
- For the development of transportation regulations.
- As an aid in establishing occupational exposure limits.

The LC_{50} is only one source of toxicity information. For a more thorough picture of the immediate or acute toxicity of a chemical, additional information should be considered such as the lowest dose that causes a toxic effect (TDLO), the rate of recovery from a toxic effect, and the possibility that exposure to some mixtures may result in increasing the toxic effect of an individual chemical.

4.6 Transformation of percentage mortalities to probits

The population response may be viewed at any one test concentration over time or at any prescribed time (t) over the concentration range. In either case, a quantal response such as mortality usually follows a normal distribution. For statistical reasons we are most interested in the centre of the response curve, specially the point at which the mortality of the test population is 50%.

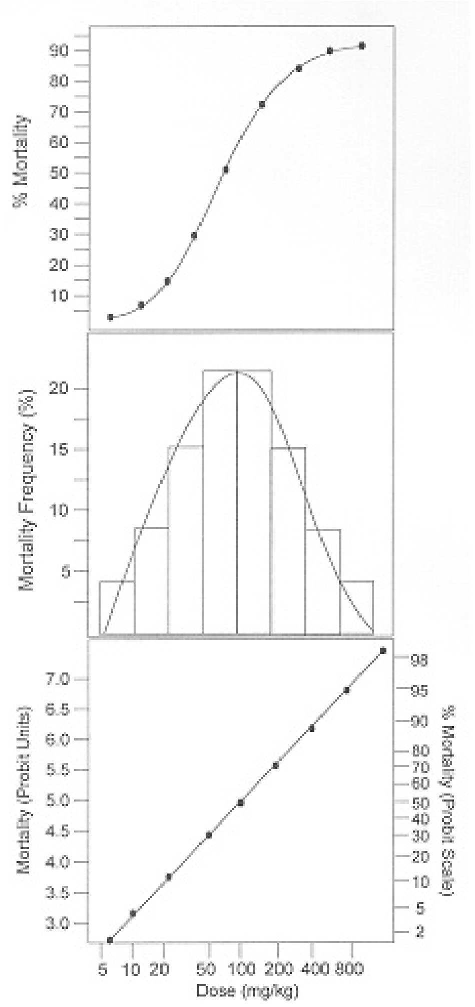


Figure 4.4: Triple graph

The middle panel is a histogram in which each bar represents the %of animals that died at each dose minus the %that died at the immediately lower dose. By definition, in the normal,the mean 1 SD represents 69.3%of the population, 2 SD is 95.5%and 3 SD is 99.7%of the population. Notice that most of the animals respond to the LD50. Those that die from the smaller doses are termed sensitive (or hypersusceptible) while those that require larger doses are resistant. The percent units can be converted to units of deviation from the mean of the normal distribution. These units, called normal equivalent deviations

%response	NED	PROBIT
0.1	-3	2
2.3	-2	3
15.9	-1	4
50.0	0	5
84.1	+1	6
97.7	+2	7
99.9	+3	8

(NEDs) simplify response calculations. The NED for 50%response is 0; -1 for 15.9%, +1 for 84.1%and so on. To avoid the use of negative numbers, the system of probability units (or PROBITS) was developed. Probits = NED + 5. The data on the bottom panel are plotted in probits. When the population is assumed to be normal, the probit transformation can be applied. The curve is much straighter, making an accurate LD50 calculation possible.

The quantile function of a distribution is the inverse of the cumulative distribution function. For a normal random variable with mean μ and variance σ^2, the quantile function is

$$F^{-1}(p) = \mu + \sigma\Phi^{-1}(p) = \mu + \sigma\sqrt{2}\quad erf^{-1}(2p-1)$$

The quantile function of the standard normal distribution is called the probit function, and can be expressed in terms of the inverse error function:

$$\Phi^{-1}(p) = \sqrt{2}\quad erf^{-1}(2p-1) \text{ where } p \in (0,1).$$

Generally, Probit value at x%mortalities=5+$\Phi^{-1}(p)$

4.7 Past studies of LC_{50}

Toxicology

LC_{50} Sediment Testing of the Insecticide Fipronil with the Non-Target Organism, Hyalella azteca (Susan Ma: May 8 2006).

Inhalation, parenteral and oral LD_{50} values of tetrahydrocannabinol in Fischer rats (Harris Rosenkrantz, Irwin A.Heyman,Monique C.Braude)

Acute and Subacute Toxicology in Evaluation of Pesticide Hazard to Avian Wildlife(Elwood F. Hill)

Fisheries Science

Determination of lethal concentration (LC_{50}) values of Cinnamomum zeylanicum hydrosol on carp fish (Kucukgul Gulec A., Altinterim B., Aksu O., July 2012)

Zoology

(LC_{50}) scores for various snakes by Stephen Spawls.

Environmental Technology

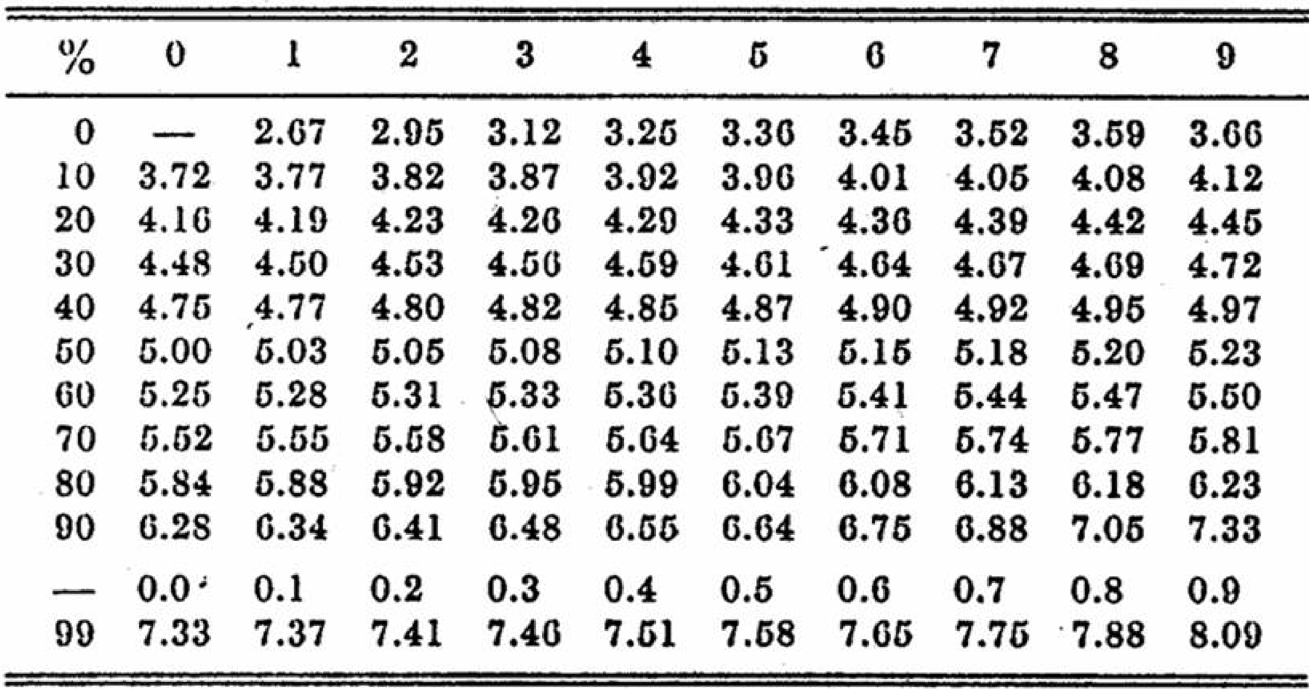
Copper (LC_{50}) to Cyprinus carpio. influence of hardness, seasonal variation, proposition of maximum acceptable toxicant concentration.(17 Dec 2008).

4.8 Examples

Ex:1

Study for estimation of LD_{50} of thymoquinone, 5 doses were given to 5 groups of rats, 10 in each group.The animals were observed for first 24th hour for any toxic symptoms.After 24 hours, the number of deceased rats was counted in each group and percentage of mortality calculated.

As the first step,Convert ⁒mortality to probits (short for probability unit) using probit table.



%	0	1	2	3	4	5	6	7	8	9
0	—	2.67	2.95	3.12	3.25	3.36	3.45	3.52	3.59	3.66
10	3.72	3.77	3.82	3.87	3.92	3.96	4.01	4.05	4.08	4.12
20	4.16	4.19	4.23	4.26	4.29	4.33	4.36	4.39	4.42	4.45
30	4.48	4.50	4.53	4.56	4.59	4.61	4.64	4.67	4.69	4.72
40	4.75	4.77	4.80	4.82	4.85	4.87	4.90	4.92	4.95	4.97
50	5.00	5.03	5.05	5.08	5.10	5.13	5.15	5.18	5.20	5.23
60	5.25	5.28	5.31	5.33	5.36	5.39	5.41	5.44	5.47	5.50
70	5.52	5.55	5.58	5.61	5.64	5.67	5.71	5.74	5.77	5.81
80	5.84	5.88	5.92	5.95	5.99	6.04	6.08	6.13	6.18	6.23
90	6.28	6.34	6.41	6.48	6.55	6.64	6.75	6.88	7.05	7.33
—	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
99	7.33	7.37	7.41	7.46	7.51	7.58	7.65	7.75	7.88	8.09

Figure 4.5: Transformation of percentages to probits

Then, take the log of concentration and plot the probits versus the log(concentration) and fit a regression line.

Dose	Log(Dose)	⁒dead	Corrected ⁒	Probits
25	1.4	0	2.5	3.04
50	1.7	40	40	4.75
75	1.88	70	70	5.52
100	2	90	90	6.28
150	2.18	100	97.5	6.96

The percentage dead for 0 and 100 are corrected before the determination of probits as under

Corrected %formula for 0 and 100 %mortality

For 0%dead = 100(0.25/n)

For 100%dead =100(n-0.25/n)

where n is the number of organism.

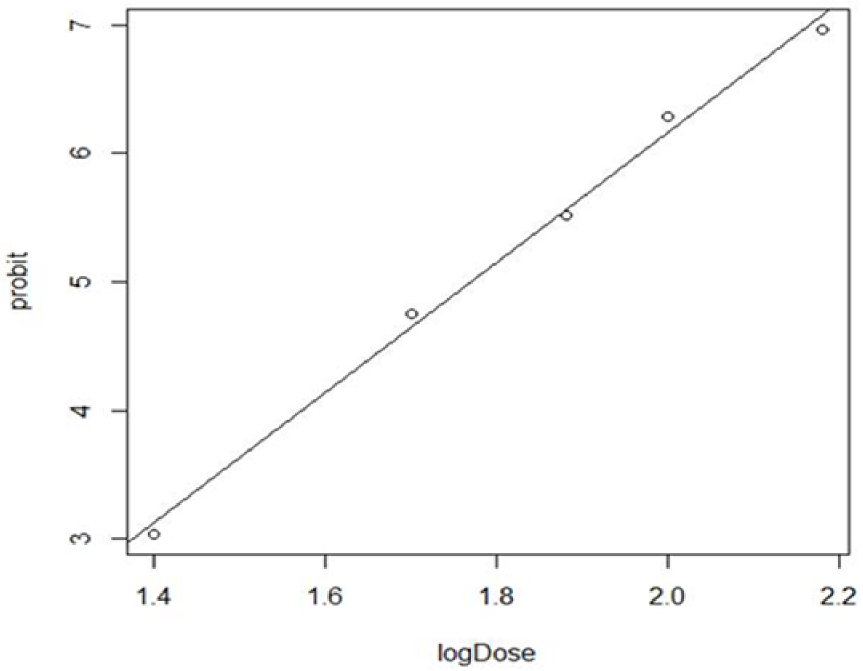


Figure 4.6: Log(Dose) vs. Probit graph

The regression equation is

probits = -3.9705 + 5.0658 logDose

To find the LC_{50} value, substitute probit=5

Then we can get,

logDose = 1.770796 and getting inverse of log, LC_{50} is 57.54mg/kg.

Ex:2

The above table shows a bunch of survival/mortality data from an experiment. Batches of snails were exposed to various high temperatures for a few hours, and recorded the number alive and dead in each batch at the end. We can find median lethal temperature as follows.

Temperature	Alive	Dead
38	20	0
40	20	0
40	20	0
40	20	0
42	19	1
42	14	6
42	15	5
44	5	15
44	3	17
44	2	18
47	0	20
47	0	20
47	0	20

Solution

Since our response variable only has two states (alive or dead), the binomial distribution is the weapon of choice.

The generalized linear model using a binomial distribution to fit a curve to temperature and alive/dead data:

Table 4.1: Binomial regression model

```
call:
glm(formula=y   temperature , family=binomial)

Deviance Residuals:
 Min                        1Q          Median      3Q         Max
 -1.30305                   -0.24134    -0.05082    0.59144    1.74125

 Coefficients:
                            Estimate    Std. Error  z value    Pr(> |z|)
 (Intercept)                69.1126     9.0186      7.663      1.81e-14 ***
 temperature                -1.6094     0.2103      -7.653     1.96e-14 ***

 Null deviance: 252.1776 on 12 degrees of freedom
 Residual deviance:8.3088 on 11 degrees of freedom
 AIC: 29.294
```

It's worth noting that the Residual Deviance here (8.309) is less than Residual Degrees of Freedom (11), meaning that our data are not overdispersed. If the Residual Deviance was something large like 20 compared to our 11 Residual Degrees of freedom, it might be worth trying a log-transformation on the treatment values. This might be the case if you were administering dosages of some drug over a very wide range of concentrations. To carry out a logistic regression on our log-transformed data, we'd simply modify the glm call slightly:

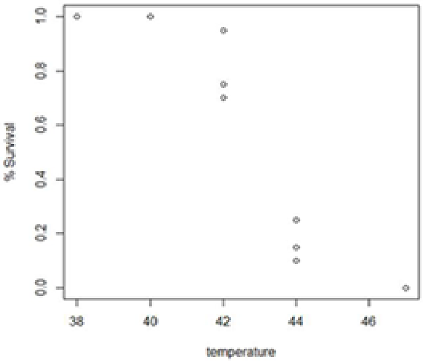


Figure 4.7: alive/dead data as %survival

Table 4.2: Logistic regression model

```
call:
glm(formula=y   log(temperature) , family=binomial)
Deviance Residuals:
 Min                                   1Q         Median     3Q        Max
 -1.28502                              -0.27691   -0.04871   0.55205   1.75589
 Coefficients:
                                       Estimate   Std. Error z value   Pr(> |z|)
 (Intercept)                           259.448    33.969     7.638     2.21e-14 ***
 log(temperature)                      -69.012    9.038      -7.635    2.25e-14 ***

 Null deviance: 252.1776 on 12 degrees of freedom
 Residual deviance:8.2032 on 11 degrees of freedom
 AIC: 29.188
```

Calculate the median lethal temperature in the experiment:

```
          Dose       SE
p = 0.5   42.94194   0.1497318
```

In this case, the value given under Dose is the median lethal dose (i.e. temperature), 42.94194. Note here that if you had to log transform your data during the glm call above, the value returned in Dose will also be log-transformed (3.7598 if we did this to the data above). You would convert it back to natural units using the exponential command:exp(3.7598) which would return 42.9 °C if we used this data set.

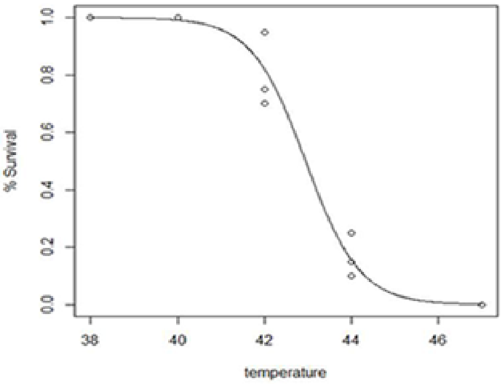


Figure 4.8: logistic curve over raw data values

Chapter 5

Application

5.1 The Experiment

Laboratory experiment was carried out to evaluate the effect of different botanicals such as Wara,Keppetiya and Maduruthala in the control of root knot nematode (M. javanica) by Pro:(Mrs)W.T.S.D.premachandra, Department Of Zoology. Approximately 50 juveniles were dispensed into petridishes containing different concentration extracts (100,80,60,40,20) of the botanicals. After 48 hours, recorded number of deaths of each petridishes.

5.2 Analysis of the experinetal data

In this experiment, outcome is a death or no death. That is, dependent variable is a binary outcome. The factors which effected to a death(independent variables) can be identify as plant type and concentration.

5.2.1 Construction of Logistic and Probit model

- The Logistic Model

Response variable is 1, when a death is occur and otherwise(i.e no death) 0. Plant type is a dummy variable. It can be design as follows:

Plant Type	D1	D2
Maduruthala	0	0
Keppetiya	1	0
Wara	0	1

Table 5.1: Logistic regression model

```
call:
glm(formula=data$ dead ˜data$Concentration + data$plant.f, family = binomial(link = ”logit”))

Deviance Residuals:
 Min                                          1Q         Median     3Q        Max
 -2.4779                                      -0.5636    -0.1515    0.5664    2.5410

 Coefficients:
                                              Estimate   Std. Error z value   Pr(> |z|)
 (Intercept)                                  -5.735211  0.131744   -43.53    <2e-16 ***
 data$Concentration                           0.063686   0.001481   42.99     <2e-16 ***
 data$plant.f2                                2.388994   0.086474   27.63     <2e-16 ***
 data$plant.f3                                2.702167   0.089134   30.32     <2e-16 ***

 Null deviance: 10374.8 on 7499 degrees of freedom
 Residual deviance:6350.3 on 7496 degrees of freedom
 AIC: 6358.3

 Note: Reference(data$plant.f1) cat should be factorize order first in R-software.
data$plant.f2 =Keppetiya ,data$plant.f3 =Wara
```

The maximum likelihood estimates of β are given from estimate values in the table. Fitted values are

$$\widehat{\pi}(x) = \frac{e^{-5.735+0.0636Con+2.3889Keppetiya+2.7021Wara}}{1+e^{-5.735+0.0636Con+2.3889Keppetiya+2.7021Wara}} \tag{5.1}$$

and the estimated logit,$\widehat{g}(x)$ is given by the equation

$$\widehat{g}(x) = -5.735 + 0.0636Con + 2.3889Keppetiya + 2.7021Wara \tag{5.2}$$

For every one unit change in concentration, the log odds of death (versus no death) increases by 0.0636.

The indicator variables for plant type have a slightly different interpretation. For example, having a death with Keppetiya plant, versus a death with a Maduruthala plant, changes the log odds of death by 2.3889.

Confidence intervals for the coefficient estimates are:

```
                     2.5 %        97.5%
(Intercept)          -5.99695190  -5.4804461
data$Concentration   0.06081724   0.0666255
data$plant.f2        2.22090055   2.5599259
data$plant.f3        2.52905116   2.8785069
```

The likelihood ratio is 4024.473, which is large enough to reject the null hypothesis of poor fit (no difference between null and full models). McFadden's pseudo R^2 is 1-(6350.3/10374.8)=0.3879111.

We can test for an overall effect of plant using the Wald test.

```
Wald test:
Chi-squared test:
X2 = 1036.2          df = 2    P(>X2) = 0.0
```

The chi-squared test statistic of 1036.2, with two degrees of freedom is associated with a p-value of 0.00 indicating that the overall effect of plant is statistically significant.

Odds ratios and their 95%CI:

Now we can say that for a one unit increase in concentration, the odds of having a death (versus not having a death) increase by a factor of 1.065.

	OR	2.5%	97.5%
(Intercept)	0.003230201	0.002486319	0.00416747
data$Concentration	1.065757750	1.062704680	1.06889510
data$plant.f2	10.902516340	9.215626290	12.93485861
data$plant.f3	14.912011210	12.541600465	17.78769387

- The Probit Model

Table 5.2: Probit regression model

```
call:
glm(formula=data$ dead ~data$Concentration + data$plant.f, family = binomial(link = "probit"))

Deviance Residuals:
 Min                                  1Q         Median     3Q       Max
 -2.5347                              -0.5739    -0.1043    0.5857   2.5829

 Coefficients:
                                      Estimate   Std. Error z value  Pr(> |z|)
 (Intercept)                          -3.2911468 0.0683004  -48.19   <2e-16 ***
 data$Concentration                   0.0371691  0.0007816  47.56    <2e-16 ***
 data$plant.f2                        1.3218435  0.0475294  27.81    <2e-16 ***
 data$plant.f3                        1.5192155  0.0486710  31.21    <2e-16 ***

 Null deviance:10374.8 on 7499 degrees of freedom
 Residual deviance:6346.2 on 7496 degrees of freedom
 AIC: 6354.2
```

The predicted probability of death is,

$\Phi(-3.2911 + 0.0372 Con + 1.3218 Keppetiya + 1.5192 Wara)$

where Φ is the cumulative distribution function of the standard normal.

The coefficient of concentration is 0.0372. This means that an increase in concentration increases the predicted probability of death. For a one unit increase in concentration, the z-score increases by 0.0372.

The constant term is -3.2911468. This means that if all of the predictors are evaluated at zero(i.e probit model of Maduruthala plant), the predicted probability of death is Φ(-3.2911468) = 0.0004988991.

Having a death of Keppettiya plant , versus a death with Mathuruthala plant (the reference group), increases the z-score by 1.3218.

95%confidence intervals for the coefficient estimates.

	2.5%	97.5%
(Intercept)	-3.42491469	-3.15969538
data$Concentration	0.03562899	0.03873375
data$plant.f2	1.23106149	1.41335706
data$plant.f3	1.42527892	1.61410480

The likelihood ratio is 4028.6, which is large enough to reject the null hypothesis of poor fit (no difference between null and full models). McFadden's pseudo R^2 is 1-(6346.2 /10374.8)=0.3883063.

We can test for an overall effect of plant using the Wald test.

```
Wald test:
Chi-squared test:
X2 = 1102.4          df = 2    P(>X2) = 0.0
```

The chi-squared test statistic of 1102.4 with two degrees of freedom is associated with a p-value of less than 0.001 indicating that the overall effect of plant is statistically significant.

Average marginal effects of the model:

(Intercept)	data$Concentration	data$plant.f2	data$plant.f3
-0.779461847	0.008802969	0.313060054	0.359804831

For each unit change in concentration, nematodes are 0.8%more likely to have death. Having a death in Keppetiya is 0.313 more likely than in Maduruthala.

5.2.2 Compararison of LC_{50} values

We can compute LC_{50} values for each botanicals using their logit and probit models. LC_{50} value means that the concentration of the botanical that kills 50%of the test animals. i.e p=0.5.

The logit model of botanical is

$\beta_0 + \beta_1 Con = ln(p/(1-p))$

when p=0.5

$\beta_0 + \beta_1 Con = ln(0.5/(1-0.5)) = ln(1)$

$\beta_0 + \beta_1 Con = 0$

and probit model is

$\beta_0^{'} + \beta_1^{'} Con = \Phi^{-1}(p)$ where Φ is the CDF of standard normal distribution.

$\beta_0^{'} + \beta_1^{'} Con = \Phi^{-1}(0.5)$

$\beta_0^{'} + \beta_1^{'} Con = 0$

Plant	LC_{50} using Logit model	LC_{50} using probit model
Maduruthala	90.1729	88.4704
Keppetiya	52.6116	52.52.9382
Wara	47.6871	47.6317

Genarally,Probit model is used to find out LC_{50} values. But we can see that there is no more different between LC_{50} values of two methods. Wara plant extract has the lowest LC_{50} value. It is enough, add $47.63mg/ml^3$ of Wara plant extract to kill 50%of the Nematode population.

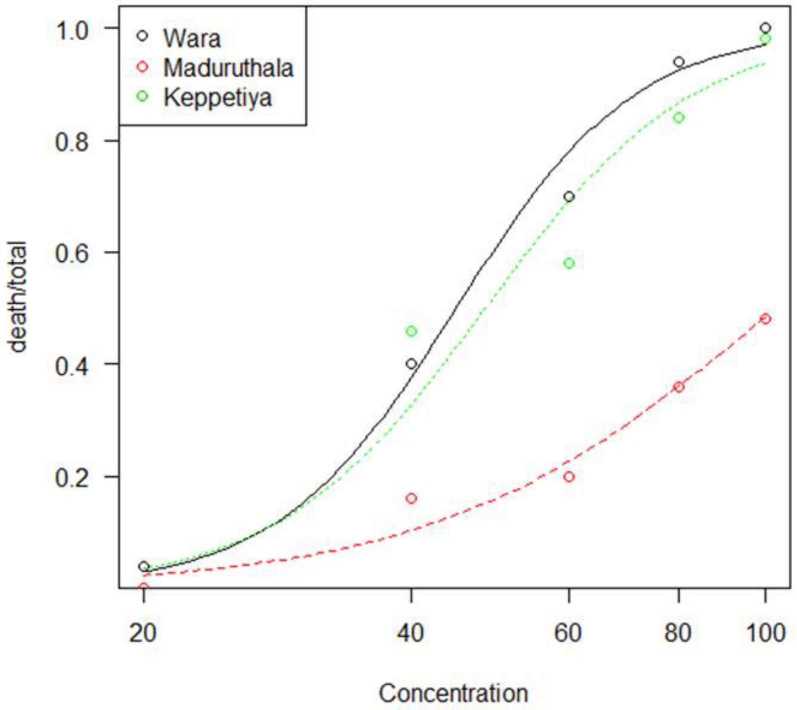


Figure 5.1: Dose-response curves of plants.

The maximal response has been obtained by Wara plant extract. That is, it has highest efficacy than others. Potency of Wara is also highest value but no more differ from Keppetiya. Maduruthala plant extract has shown lower potency and lower efficacy.

Chapter 6

Conclusions and Recommendations

6.1 Summery

In this report chapters were ordered as follows. Mainly, first chapter included the background of the study and objectives of the study. The main purpose was study Probit regression model and Logistic regression model. These two models are used to analyzed binomial response data. In real world, many problems could be found which related to binomial response from various fields such as Chemestry, Biology, ,Medicine, Economics, Sociology, pharmacology ang Zoology. So, statistical methodology of these two regression models are very important in many research to bbuild up correct conclusions.

When a patient was given any medicine, we can measure the patient state or effect of that medicine and also using different doses of medicine can get different effect. Such cases are called dose-response studies. Definition, relationship and special measures of dose-response data were included to the Chapter 2. Statistical theory of Probit regression model ang Logit regression model were presented from Chaper 3 and these theory were explained using specific examples.

Median Lethal Concentration was discussed in Chapter 4. It is a widely use measurement of dose-response data in many research, specially in Toxicology. We can compute Median Lethal Concentration using both logit and probit models but most popular method is the probit method. An application of dose-response data was included in Chapter 5. This application is a loboratory experiment of some botanicals to test effect on nematodes. Logit and Probit models of experimental data, Median Lethal Concentration of each botanicals were included in Chapter 5.

6.1.1 Discussion

This study was done for study Logistic and Probit regression model to analysis doseresponse data. A large body of literature exists on the use of logistic regression in real world applications. Comparatively less is known about a similar, but intrinsically different approach of probit regression. This article introduced probit model as another powerful and useful approach for modeling binary data. Because the function used in Probit is the inverse of the standard normal cumulative distribution function,the interpretation of the results is sometimes difficult.

I have described dose-response designs and illustrated them with simple examples. Most common outcome of a dose-response experiment in which probit analysis is used is the LC50. But in this thesis, LC50 was calculated from both Probit and Logit models.

Under the application of dose-response data, this study has shown that the leaf extracts of Maduruthala, Keppetiya andWara has some toxic effects on Root-knot Nematodes. The analysis of the results indicates that the LC50 value forWara plant is less than others. That is ,Wara plant extract has more toxic effect.

All analisis were done using R 3.0.2 statistic software.

6.2 Conclusions

The logit model is

$\widehat{g}(x) = -5.735 + 0.0636 Con + 2.3889 Keppetiya + 2.7021 Wara$

Coefficient of Wara is greater than the other plant's coefficients. Then we can say that maximum effect is caused on probability of having a death by Wara plant extract.

The probit model is

$\Phi^{-1}(p) = -3.2911 + 0.0372 Con + 1.3218 Keppetiya + 1.5192 Wara$

Also in the probit model, Wara plant extract has the greater coefficient and it affectes to increase the probablity of having a death than other two plants.

All independent variables are more significant because p-values are very small.

Coefficient of Concentration is positive in both models, that is when concentration is increase, probability of having a death is also increase. When consider the LC50 values of plants ,we can see that Wara has the least LC50 value($46.43mg/ml^3$) and conclude that it is the effective extract among these three plants. It is enough,add $46.43mg/ml^3$ of Wara plant extract to kill 50%of the population. Finally conclude that Wara plant extract is the effective botanical to use to reduce root-knot nematodes.

6.2.1 Recommendations

- Books,Articles should be published about dose-respose data analysis using R statistical software.
- More research should be carrying out to analyze dose-response data.
- Further research is needed to investigate the advantages or disadvantages in using one model over the other in dose-response data.

Appendix

R Codes

Probit model Ex:1

```
d<-read.csv(”pex.csv”,header=TRUE)
attach(d)
fit<-glm(Grade~ GPA+TUCE+PSI,binomial(link=”probit”))
summary(fit)
logLik(fit)
ProbitScalar<- mean(dnorm(predict(fit, type = ”link”)))
ProbitScalar * coef(fit)
probit0<-update(fit, formula= Grade~ 1)
McFadden<- 1-as.vector(logLik(fit)/logLik(probit0))
McFadden
```

Probit model Ex:2

```
data=data.frame(dose=c(1.6907,1.7242,1.7552,1.7842,1.8113,1.8369,1.861,1.8839), exposed= c(59,60,62,56,63,59,62,60),killed= c(6,13,18,28,52,53,61,60), alive=c(53,47,44,28,11,6,1,0))
attach(data)
y=cbind(killed,alive) or y=cbind(killed/exposed)
model= glm(y~dose,binomial(link=”probit”))
```

```
summary(model)
fitted(model)
plot(dose,fitted(model))
curve(predict(model,data.frame(dose=x),type=”resp”),add=TRUE)
points(dose,fitted(model),pch=20)
```

Logit model Ex:1

```
data= data.frame(dose= c(0,2.5,5,10),NonRelieved= c(65,68,101,105),    Relieved= c(2,7,29,40))
attach(data)
y = cbind(Relieved,NonRelieved)
model= glm(y dose,binomial(link=”logit”))
summary(model)
library(aod)
wald.test(b = coef(model), Sigma = vcov(model), Terms =1)
wald.test(b = coef(model), Sigma = vcov(model), Terms =2)
plot(dose,(Relieved / (NonRelieved+Relieved)*100), ylab = ”%Relieved”)
```

Logit model Ex:2

```
data=read.csv(”ex2.csv”,header=TRUE)
attach(data)
plot(income,y)
reg=glm(y~ income,family=binomial(link=”logit”))
summary(reg)
fitted(reg)
library(boot)
xrange=seq(from=min(income), to=max(income), length=100)
lines(xrange, inv.logit(reg$coef[1]+reg$coef[2]*xrange))
```

```
anova(reg)
LR.test.stat=anova(reg)[2,2]
LR.test.df=anova(reg)[2,1]
LR.test.Pvalue=1 - pchisq(LR.test.stat, df=LR.test.df)
est.odds.ratio=exp(summary(reg)$coef[”income”,”Estimate”])
conf.level<- 0.95
alpha<- 1 - conf.level
b1<- summary(reg)$coef[”income”,”Estimate”]
s.b1<- summary(reg)$coef[”income”,”Std. Error”]
lower<- exp(b1 - qnorm(1-alpha/2)*s.b1)
upper<- exp(b1 + qnorm(1-alpha/2)*s.b1)
print(paste(100*(1-alpha), ”percent CI for odds ratio:”, lower, upper))
```

LC_{50} value Ex:2

```
data2 = data.frame(temperature = c(38,40,40,40,42,42,42,44,44,44,47,47,47),
  alive = c(20,20,20,20,19,14,15,5,3,2,0,0,0), dead = c(0,0,0,0,1,6,5,15,17,18,20,20,20))
attach(data2)
y = cbind(alive,dead)
model.results1 = glm(y temperature,binomial)
summary(model.results1)
model.results2 = glm(y log(temperature),binomial)
summary(model.results2)
library(MASS)
dose.p(model.results1,p=0.5)
dose.p(model.results2,p=0.5)
plot(temperature,(alive / (alive + dead)), ylab = ”%Survival”)
logisticline = function(z) eta = 69.1126 + -1.6094 * z; 1 / (1 + exp(-eta))
```

```
x = seq(38,47,0.01)
lines(x,logisticline(x),new = TRUE)
```

Application

The Logistic Model

```
data<-read.csv(”RESEARCHbinary.csv”,header=TRUE)
attach(data)
data$plant.f<-factor(data$Plant)
is.factor(data$plant.f)
model= glm(data$dead ~Concentration+data$plant.f,binomial(link=”logit”))
summary(model)
confint(model)
LRtest<-model$null.deviance - model$deviance
library(aod)
wald.test(b = coef(model), Sigma = vcov(model), Terms =3:4)
exp(cbind(OR = coef(model), confint(model)))
```

The Probit Model

```
data<-read.csv(”RESEARCHbinary.csv”,header=TRUE)
attach(data)
data$plant.f<-factor(data$Plant)
is.factor(data$plant.f)
data$plant.f[1:15]
model=glm(data$dead~$Concentration+data$plant.f,binomial(link=”probit”))
summary(model)
pnorm(-3.2911468)
LRtest<-model$null.deviance-model$deviance
confint(model)
```

```
library(aod)
wald.test(b = coef(model), Sigma = vcov(model), Terms = 3:4)
ProbitScalar<- mean(dnorm(predict(model, type = ″link″)))
ProbitScalar * coef(model)
```

Compararison of LC_{50} values

```
M<-data.frame(con=c(20,40,60,80,100),death=c(0,8,10,18,24),alive=c(50,42,40,32,26))
K<-data.frame(con=c(20,40,60,80,100),death=c(0,23,29,42,49),alive=c(50,27,21,8,1))
W<-data.frame(con=c(20,40,60,80,100),death=c(2,20,35,47,50),alive=c(48,30,15,3,0))
model1<-glm(cbind(M$death,M$alive) M$con,binomial(link=″logit″))
summary(model1)
dose.p(model1,p=0.5)
model1<-glm(cbind(K$death,K$alive) K$con,binomial(link=″logit″))
summary(model1)
dose.p(model1,p=0.5)
model1<-glm(cbind(W$death,W$alive) W$con,binomial(link=″logit″))
summary(model1)
dose.p(model1,p=0.5)
model1<-glm(cbind(M$death,M$alive) M$con,binomial(link=″probit″))
summary(model1)
dose.p(model1,p=0.5)
model1<-glm(cbind(K$death,K$alive) K$con,binomial(link=″probit″))
summary(model1)
dose.p(model1,p=0.5)
model1<-glm(cbind(W$death,W$alive) W$con,binomial(link=″probit″))
summary(model1)
dose.p(model1,p=0.5)
```

Dose-response curves for different plants

```
library(drc)
con=c(20,40,60,80,100) ♯ different concentration of botanical extracts
dm<-c(0,8,10,18,24) ♯ number of deaths of Maduruthala plant extracts
dk<-c(0,23,29,42,49) ♯ number of deaths of Keppetiya plant extracts
dw<-c(2,20,35,47,50) ♯ number of deaths of Wara plant extracts
tot<-c(50,50,50,50,50)♯total number of nematodes in each sample are same
m<- drm(dm/tot con, weights=total,fct=LL.2(), type=”binomial”)
k<- drm(dk/tot con, weights=total,fct=LL.2(), type=”binomial”)
w<- drm(dw/tot con, weights=total,fct=LL.2(), type=”binomial”)
plot(w, bp=.5,conName=”control”,broken=TRUE, lty=c(1,1),xlab=”Concentration”, ylab=”death/total
plot(k, bp=.5,conName=”control”, legend=FALSE,add=TRUE,col=”green”, lty=3)
plot(m, bp=.5,conName=”control”, legend=FALSE,add=TRUE,lty=2,col=”red”)
legend(”topleft”, legend = c(”Wara”,”Maduruthala”,”Keppetiya”), col=1:3, pch=1)
```

Bibliography

[1] Mehdi Razzaghi:*Journal of Modern Applied Statistical Methods*,Bloomsburg University,May 2013, Vol. 12, No. 1, 164-169.

[2] Weng Kee Wong,Peter A. Lachenbruch:*Tutorial in Biostatistic and Designing studies for dose response*,VOL.15,343-359(1996).

[3] Susan Ma:*LC50 Sediment Testing of the Insecticide Fipronil with the Non-Target Organism*,May 8 2006.

[4] Muhammad Akram Randhawa: *http://www.ayubmed.edu.pk/JAMC/ PAST/21-3/ Randhawa*,College of Medicine, University of Dammam: 2009;21(3),

[5] K. Bondari:*Paper ST01*,University of Georgia, Tifton,GA 31793-0748.

[6] Park, Hun Myoung:*Regression models for binary dependent variables using Stata, SAS, R, LIMDEP, and SPSS*,Indiana University(2009).

[7] *Probit Analysis* By: Kim Vincent

[8] Mark Tranmer,Mark Elliot:*Binary Logistic Regression*

[9] David W. Hosmer,JR.,Stanley Lemeshow,Rodney X. Sturdivant:*Applied Logistic Regression*,Third Edition,ISBN 978-0-470-58247-3.

[10] Scott A. Czepiel:*Maximum Likelihood Estimation of Logistic Regression Models*,Theory and Implementation.

[11] Park, Hyeoun-Ae:*An Introduction to Logistic Regression*,Seoul National University,Korea,J Korean Acad Nurs Vol.43 No.2,April 2013.

[12] Christian Ritz,Jens Carl Streibig:*Dose-response modelling using R*,University of Copenhagen

[13] Raghu Prasada M.S:*Dose response curve*,Department of Pharmacology

[14] Saurabh Wani:*Dose response relationship*,Garware College

[15] Sharyn O'Halloran:*Lecture 10,Logistic regression II multinomial data.*

[16] Finney, D. J., *Ed. (1952). Probit Analysis*,Cambridge, England, Cambridge University Press.