WEIBULL MODEL FOR DOSE RESPONSE DATA
and Akaike Information Criterion for model selection.

By

Petronilla Mumbua Stephen

A project submitted for the partial fulfillment of the degree of
Master of Science in Mathematical Statistics.

Kenyatta University.

DECLARATION.

This is my original work and has not been presented for a degree award in any other university.

Sign ____________

PETRONILLA MUMBUA STEPHEN.

This work has been presented with my approval as the University Supervisor.

Sign ____________

DR. ROMANUS ODHIAMBO OTIENO

Mathematics department

Jomo Kenyatta University.
ABSTRACT.

Statistical linear models are used to study dose response models in the bioassay. These have given rise to many statistical problems since the dose response data do not follow linear model. This has led to the use of non-linear models such as probit and Logit. The Logit model has been widely used to analyze the data. Several non-linear models have also been proposed which can be treated in a fashion similar to the parametric logistic model.

In this project, we review the parametric logistic model and study the analytical method used in its analysis thoroughly. We study in detail the Weibull dose-response model, following the same method of logistic model. We are able to show that their structures are similar.

We study the Akaike Information Criterion for model selection and use it to select a better model between logistic and Weibull models.
DEDICATION.

To my father, Cyprian Stephen Munyao (deceased) and mother, Kasendi and my siblings, Itumbi, Munyiva, Nthale, and Mbini.
ACKNOWLEDGEMENTS.

My sincere thanks go first to my siblings and my parents for their financial support they chipped in since my study was self sponsored.

Secondly I owe alot to my Supervisor Dr Romanus Odhiambo for his professional guidance especially when everything was strange to me he was always patient and available.

I acknowledge the encouragement and professional guidance by Dr Wafula, Dr Odongo, Dr Njenga and all other lecturers in mathematics department, Kenyatta University, who have seen me through my course successfully.

I can’t forget my colleagues in the struggle, Mulati, Biwott, Kisaka, Otwombe and Monica (deceased) for their cooperation and support, Thankyou.

Special thanks goes to Munyiva (sister), Arimi (Mathematics Department, Computer), Tony (AVU) and Alati (Cyber Café) whose technical support enabled me to complete my work in time.

I also express my sincere thanks to all persons who in one way or other contributed to this study.

Finally I give thanks to the Almighty God for the gift of life for each one of us with which we serve Him through serving one another.

MAY GOD BLESS US ALL, AMEN.
CONTENTS.

Abstract                      i
Dedication                    ii
Acknowledgment               iii
Table of contents             iv
Prologue                     vii

CHAPTER ONE.

LOGISTIC MODEL OVERVIEW.

1.1 Introduction.             1
1.1.1 Regression Analysis.    1
1.1.2 Non-linear regression.  2
1.1.3 Some methods of fitting curves. 3
1.1.4 The general regression model. 4
1.2. The logistic model development. 5
1.3. Estimation of parameters. 7
1.4. Maximum likelihood estimation for logistic regression.
    (General formation of the iterative procedure). 8
1.5. The Logit model assumptions. 10
1.6. Statistical problems.     11
CHAPTER TWO
THE WEIBULL MODEL USING LOGIT MODEL.

2.1 Introduction. 19
2.2 General review of the Weibull distribution. 19
2.3 Relationship between Weibull and exponential distribution. 21
2.4 Maximum likelihood estimation for Weibull distribution (General formulation of the iterative procedure). 22
2.5 Estimation of parameters. 24
2.6 The method of scoring for the Weibull model. 27
2.7 The connection with the iterated weighted regression (Generalized model). 32

CHAPTER THREE
AKAIKE INFORMATION CRITERION.

3.1 Introduction. 34
3.2 K-L distance definition 35
3.3 The K-L estimator. 36
3.4 K-L distance and AIC relationship. 37
3.5 The AIC definition. 39
3.6 An application of AIC. 41
3.7 Akaike Predictive Expected log-likelihood. 42
3.8 A second order AIC important refinement. 45
3.9 Useful Preliminaries. 46
3.10 A general derivation of AIC. 53

CHAPTER FOUR.

AIC ON LOGISTIC AND WEIBULL MODEL.

4.1 Introduction. 60
4.2 Application of AIC on Weibull and logistic model. 65
4.3 Estimation of parameters. 70
4.4 Results. 75

CHAPTER FIVE

5.1 Conclusion. 79
5.2 Areas for further research. 80

REFERENCE. 81
Dose response model has been used extensively in bioassay to describe the probability of a biological response as a function of dose or drug. Many statistical methods have been proposed to quantify the dose effects. Linear statistical models have been proposed. Besides we have probit model due to Bliss (1984) and Gaddum (1948) which have been extensively used to analyze data. Logit model reviewed by Ashton (1972) is more widely used as well.

In many practical situations, it is likely that the response depends on several characteristics of the study subject. This causes us to use distributions that try to express this more accurately. Wijesinha and Piantadosi (1995) suggested some parametric dose-response models that can be treated in a fashion similar to the parametric logistic model. In the project we will be studying one of the model’s (Weibull) and find out which among the two is a better model for dose response problems.

Over a century ago measures were derived for assessing the ‘distance’ between two models or probability distributions. Most relevant here is Boltzmann’s (1877) concept of generalized entropy in physics and thermodynamics, Shannon (1948) employed entropy in his famous treatise on communication theory. Kullback and Leibler (1951) derived an
information measure that happened to be the negative of Boltzmann’s entropy now referred to as Kullback-Leibler (K-L) distance. The motivation for Kullback and Leibeler’s work was to provide a vigorous definition of ‘information’ in relation to Fisher’s ‘sufficient statistics’. The K-L distance has also been called the K-L discrepancy, divergence, information, and number.

The Kullback-Leibler distance can be conceptualized as a directed ‘distance’ between two models, say f and g Kullback (1959). Strictly speaking, this is a measure of ‘discrepancy’ it is not a simple distance because the measure from g to f is not the same as the distance from f to g – it is a directed, or oriented, distance. The K-L distance is perhaps the most fundamental of all information measures in the sense of being derived from minimal assumptions and its additivity property. It is the fundamental quantity in science and information theory (Akaike 1983) and is the logical basis for model selection as defined by Akaike. We shall thus use the Akaike Information Criterion as a basis for determining the best model.

In chapter one we review the parametric logistic model. This is necessary for the following chapter. By understanding the equations we are able to develop the dose response model for Weibull. Chapter two opens up with a brief introduction of Weibull model. By using the concepts gained in chapter
one we are able to show that Weibull model behaves in a similar way like logistic dose response model. Now the question remains which is a better model. To answer this we used Akaike’s Information Criterion, which we review in chapter three. Besides the review of the general Akaike’s Information Criterion in this chapter (three) we review particularly a general derivation of Akaike’s Information Criterion for exponential model since both our two models are exponential. Finally in chapter four we estimate the unknown constants by use of iterations in both the Weibull model and the logistic model. By use of data we are able to draw conclusion on which model is better.
CHAPTER ONE

1. LOGISTIC MODEL OVERVIEW.

1.1. Introduction.

1.1.1. Regression analysis.

Analysis of data is done first by applying straight line models because they are simple to deal with. This involves regression analysis of variables (dependent and independent) where the relationship is expressed in form of equation as

\[ y = b_0 + b_1 x + \varepsilon_i \]

where \( y \) is the dependent variable and \( x_i, (i = 1,2,\ldots,n) \) is independent variable or multiple linear regression

\[ y = b_0 + \sum_{i=1}^{n} b_i x_i + \varepsilon_i \]

\( b_0, b_1, \ldots, b_n \) are regression co-efficients and \( \varepsilon_j, (j = 1,2,3,\ldots,k) \) are random errors.

Once a regression model has been established, we need to estimate the parameters. For linear models and any other model which can be linearized the method of ordinary least square is the best for estimating the parameters.

But model like
\[ y = \frac{c}{1 + \lambda^{\alpha + \beta x}} + \varepsilon_i \]

cannot be linearized. Here an iterative algorithm can be used where an initial approximation is successively improved until desired accuracy is obtained. After estimating the parameters and substituting them on the model, one is always tempted to test whether the model is adequate or not. Formulation of different types of hypothesis follows.

### 1.1.2 Non-Linear regression

When the ordinary least square method is used to estimate parameters in models that are not linear, obviously problems of precision, efficiency, consistency etc arise, hence the need for a method of estimating parameters for non linear model. These models are referred to as discrete choice models or qualitative response models, or categorical models or quantal models, Kimenta (1986). Some problem associated with linear probability model emanate from the fact that \( y_i \) is interpreted as probability of the event occurring hence giving the range as

\[ 0 \leq y_i = P_i \leq 1 \]

To solve this heteroskedastic problem we use weighted square method and restricting the parameter to lie in the range 0 and 1.
1.1.3 Some methods of fitting curves

(i) Explanatory band regression.
This involves dividing a scatter plot into a series of vertical bands then a
central point is picked or selected within each band. It has major draw backs
such as being affected by outliers, Skewness, etc. This problems can be
overcome by choosing a good power transformation on $x$ and $y$

(ii) Curvilinear regression.
The model
\[ y = \alpha ( \exp(\beta x_i)) \epsilon_i \]
can be linearized for the sake of simplified analysis by taking logs on both
sides resulting into
\[ \log_y y_i = \log \alpha + \beta x_i + \log \epsilon_i \]
Hence one can use ordinary least squares for estimating the parameters.

(iii) Non-Linear regression.
This method has its own criterion which are obeyed by models like

(a) Non-linear exponential models
\[ y_i = \alpha (\exp(- \beta x_i)) + \epsilon_i \cdot \]

(b) Negative exponential curves.
\[ y_i = \alpha (1 - \exp(- \beta x_i)) + \epsilon_i \]
(c) The two term exponential curves

\[ y_i = \frac{\theta_1}{\theta_1 - \theta_2} [\exp(-\theta_2 x_i) - \exp(-\theta_1 x_i)] + \epsilon_i \]

where \( \theta_1 \) and \( \theta_2 \) are parameters.

(d) Growth curves.

The growth rate curve forms an up-to-right curve with an asymptotic finish producing a sigmoid growth curve. The best known sigmoid model is the logistic curve given by

\[ y_i = \frac{\alpha}{1 + \gamma \exp(-\beta x_i)} + \epsilon_i \]

1.1.4 The general logistic regression model

The general logistic regression model has an equation of the form

\[ Y_i = \frac{\alpha}{1 + \beta e^{-\alpha x_i}} + \epsilon_i \quad \text{.................................1.1.1} \]

where \( Y_i \) is the dependent variable.

\( x_i \) is the independent variable.

\( \alpha \) and \( \beta \) are constant parameters which are estimated from the data \( x_i \) and \( \epsilon_i \) is the random error.
This equation produces S-shaped or sigmoid curves with rate of change of $Y_i$ with respect to $x_i$, which increases at first and decreases as $x_i$ increases.

When $x_i = 0$ we have the

$$Y_i = \frac{\alpha}{1+\gamma} \quad \text{..................................................1.1.2}$$

Equation 1.1.2 gives the $Y_i$ intercept.

Put $\alpha = 1$ and $\beta = 1$ in equation 1.1.1 then we obtain a special case of the above equation reflecting that $Y_i$ is a response, which can only take values between 0 and 1. It is 1 when there is response and 0 otherwise.

Hence the resulting equation is given as

$$Y_i = \frac{1}{1 + e^{-(\beta x_i)}} + \epsilon_i \quad \text{..................................................1.1.3}$$

Equation 1.1.3 is referred to as the logistic regression model.

1.2 The logistic model (development).

The method of ordinary least square cannot be used with this model since it will lead to predicted values that are less than zero or greater than one which however is not the case with probability of response that is 0 and 1.

Therefore we will use the logit model.

Let $P$ be the probability of success then the ratio
\( \frac{p}{1-p} \) is referred to the odds ratio, where \((1-p)\) is the probability of failure. This can be estimated as [that is the probability of success]

\[
\hat{p} = \frac{r_i}{n_i} \text{ for } i = 1, 2, ..., \Delta
\]

where \(n_i\) are subjects in a group, \(r_i\) are those who respond, and \(\Delta\) the number of doses in the experiments.

The logit model is based on the loge of this odds ratio that is

\[
\hat{L}_j = \ln \left( \frac{\hat{p}}{1-\hat{p}} \right) = \beta_0 + \beta_i x_{ij} + \varepsilon_j \text{ where } i = 1, 2, ..., k \text{ and } j = 1, 2, ..., n
\]

Now \( \hat{L}_j = \ln \left( \frac{\hat{p}}{1-\hat{p}} \right) \):

\[
\Rightarrow e^{\hat{L}_j} = \frac{\hat{p}}{1-\hat{p}}
\]

\[
\Rightarrow (1-\hat{p}) e^{\hat{L}_j} = \hat{p}
\]

\[
\hat{p}(1 + e^{\hat{L}_j}) = e^{\hat{L}_j}
\]

\[
\hat{p} = \frac{e^{\hat{L}_j}}{1 + e^{\hat{L}_j}}
\]

\[
\hat{p} = \frac{1}{1 + \frac{1}{e^{\hat{L}_j}}}
\]

\[
\hat{p} = \frac{1}{1 + e^{-\hat{L}_j}}
\]
From the equation
\[ L = \ln\left(\frac{P}{1-P}\right) \]

Put \( P = 0 \) then \( L = \ln\frac{0}{1} = 0 \Rightarrow L = -\infty \)

Put \( P = 1 \Rightarrow L = \ln\frac{1}{0} = \infty \Rightarrow L = +\infty \)

Thus \( L \) ranges from \( -\infty \) to \( +\infty \) which helps us to overcome the problem of getting responses \( Y_j \)s which go beyond the \([0,1]\) boundary in the model.

1.3 Estimation of parameters

Since the logit model follows a binomial distribution then we will use the method of maximum likelihood estimation in estimating the parameters.

If \( x_j \) is the \( j^{th} \) combination of \( x \) values from the logit model then the conditional probability is given by

\[ P_j = \frac{1}{1+e^{-L_j}} \]

where \( L_j = \beta_0 + \sum_{i=1}^{n} \beta_i x_{ij} \) and the \( j^{th} \) case contribute

\[ P_j \text{ if } y_j = 1 \text{ and } 1-P_j \text{ if } y_j = 0 \]
The contributions to the likelihood function is the product of the individual contribution.

\[ L = \prod_{j=1}^{n} P_{j}^{y_j} (1 - P_j)^{(1 - y_j)} \]

maximizing the log likelihood function we are able to get the max

\( (\beta_1, \beta_2, \ldots, \beta_k) \)

thus

\[ \ln L = \ln \prod_{j=1}^{n} P_{j}^{y_j} (1 - P_j)^{(1 - y_j)} \]

\[ l = \ln \prod_{j=1}^{n} P_{j}^{y_j} (1 - P_j)^{(1 - y_j)} \]

\[ l = \sum_{j=1}^{n} (y_j \ln P_j + (1 - y_j) \ln (1 - P_j)) \] 1.3.1

The expression 1.3.1 is differentiated once with respect to each parameter and then the derivatives are set to zero to obtain normal equations

\[ \frac{\partial l}{\partial \beta'} = 0 \]

which are solved by use of iterative procedure.

1.4 Maximum likelihood estimation for logistic regression [general formulation of the iterative procedure].

Let us consider a certain test done on a group of \( n \) subjects, where \( i \) denote the number of different test and \( r \), be the number of subjects who respond
then we will have a binomial distribution given by the likelihood function

\[ L = \prod_{i=1}^{k} \binom{n_i}{r_i} P_i^{r_i} (1 - P_i)^{n_i - r_i} \] for \( i = 1, 2, \ldots, k \) and

\( P_i \) is the probability of response for test \( i \) and is given as

\[ P_i = \frac{1}{1 + e^{-(\alpha + \beta x_i)}} \]

thus the log-likelihood function is

\[ l = \Phi + \sum_{i=1}^{k} r_i \ln P_i + \sum_{i=1}^{k} (n_i - r_i) \ln (1 - P_i) \]

where \( \Phi = \ln \prod_{i=1}^{k} \binom{n_i}{r_i} \)

substituting the value of \( P_i \) in equation 1.4.1 we get

\[ l = \Phi + \left( -\sum_{i=1}^{k} r_i \ln (1 + e^{-(\alpha + \beta x_i)}) \right) - \sum_{i=1}^{k} (n_i - r_i) \ln (1 + e^{-(\alpha + \beta x_i)}) - \sum_{i=1}^{k} (n_i - r_i) (\alpha + \beta x_i) \]

\[ l = \Phi - \sum_{i=1}^{k} n_i \ln (1 + e^{-(\alpha + \beta x_i)}) - \sum_{i=1}^{k} (n_i - r_i) \ln (1 + e^{-(\alpha + \beta x_i)}) - \sum_{i=1}^{k} (n_i - r_i) (\alpha + \beta x_i) \]

\[ l = \Phi - \sum_{i=1}^{k} n_i \ln (1 + e^{-(\alpha + \beta x_i)}) - \sum_{i=1}^{k} (n_i - r_i) (\alpha + \beta x_i) \]...1.4.2

Let \( \beta' \) be a vector representing \( (\alpha, \beta) \) and since the logit on a vector is a vector that is

\( \text{Logit } (P) = X' \beta \)

where \( X' = (X_1, X_2, \ldots, X_k) \)
and \( k \) is the number of combination of explanatory variables that are employed in the exp. \( x \) then we can generalize equation 1.4.2 as

\[
l = \Phi - \sum_{i=1}^{k} n_i \ln \left(1 + e^{-(x_i' \beta)}\right) - \sum_{i=1}^{k} (n_i - r_i)(x_i' \beta) \]

1.4.3

where \( x_i \) denote the \( i^{th} \) row matrix \( x \)

hence from equation 1.4.3 we get

\[
\frac{\partial l}{\partial \beta_i} = \sum_{i=1}^{k} x_i (r_i - n_i P_i) \]

........................................... 1.4.4.

\[
\frac{\partial l}{\partial \beta} = \sum x_i (1 - P_i) P_i \]

These equations (1.4.4) allow us to write the updated estimates using the Newton- Raphson as

\[
\beta^{t+1} = \beta^t + (x'vx)^{-1} x's \]

This is similar to standard linear regression method where

\[
v = \text{diag}(n_i P_i (1 - P_i)) \quad \text{and} \quad s = r_i - n_i P_i \]

These two \((v, and, s)\) are evaluated at the \( i^{th} \) iterate \( \beta^t \).

1.5 The logit model assumptions.

The logit model regression has the following assumptions
(a). Model should be specified correctly

(b). $X$ variable must be accurately measured and no additional less significant variable are included or omitted.

(c). The cases are independent

(d). None of the $X$ variables is a linear combination of the other for if multicolinearity would bring about imprecise estimates. Just like in ordinary least square method, employing nested models can do hypothesis test in logit regression.

1.6 Statistical problems.

Several statistical problems arise with logit regression which include

(a). Multicollinearity among the $X$ variables.

(b). High discrimination or our ability to detect the response $Y$.

(c). Curvilliniarity, we assume that the logit is linearly related to $X$-variables.

(d). Influential cases like the outliers should be dealt with through transformation

(e). For skewed $X$-variables, symmetrizing transformation can be applied before the logit model.
The method of scoring for the logit model.

The log-likelihood is given as

\[ l = \Phi + \sum_{i=1}^{k} r_i \ln P_j + \sum_{i=1}^{k} (n_i - r_i) \ln (1 - P_i) \]

where \( \Phi = \ln \prod_{i=1}^{k} \left( \frac{n_i}{r_i} \right) \)

for the logit model

\[ P_j = \frac{1}{1 + e^{-(\alpha + \beta r_j)}} \]

Let \( \lambda \) denote the vector \( (\alpha, \beta) \) then

\[ \frac{\partial l}{\partial \lambda} = \sum_{i=1}^{k} \frac{r_i \partial P_i}{P_i} - \sum_{i=1}^{k} \left( \frac{n_i - r_i}{(1 - P_i)} \right) \]

\[ = \sum_{i=1}^{k} \frac{\partial P_i n_i (p_i - P_i)}{\partial \lambda P_i (1 - P_i)} \]

where \( p_i = \frac{r_i}{n_i} \)

then we will have
\[ \frac{\partial l}{\partial \alpha} = \sum_{i=1}^{k} \frac{n_i(p_i - P_i)}{P_i(1 - P_i)} f(d_i) \]

\[ \frac{\partial l}{\partial \beta} = \sum_{i=1}^{k} n_i d_i (p_i - P_i) f(d_i) \]

\[ \frac{\partial l}{\partial \alpha} = \sum_{i=1}^{k} n_i (p_i - P_i) \]

\[ \frac{\partial l}{\partial \beta} = \sum_{i=1}^{k} n_i d_i (p_i - P_i) \]

To obtain solutions for \( \alpha \) and \( \beta \) we need to find \( \frac{\partial l}{\partial \alpha} \) and \( \frac{\partial l}{\partial \beta} \text{ and equate to zero } \)

but \( f(d_i) = P(d)(1 - P(d)) \) (note that \( d_i \) can be used interchangeably with \( x_i \))

for all the logit case.

Therefore \( \frac{\partial l}{\partial \alpha} \) and \( \frac{\partial l}{\partial \beta} \) become (from equations 1.7.1)

Now for a general model

\[ \frac{\partial^2 l}{\partial \lambda^2} = \sum_{i=1}^{k} \frac{\partial^2 P}{\partial \lambda^2} n_i \left( \frac{\partial P_i}{\partial \lambda} \right) \] + Terms with factors of \( (p_i - P_i) \)

Furthermore \( E(p_i) = P_i \) and since \( p_i = \frac{r_i}{n_i} \) which results from the binomial distribution \( B(n_i, P_i) \) with mean \( n_i P_i \)

hence

\[ E \frac{\partial^2 l}{\partial \lambda^2} = -\sum_{i=1}^{k} \frac{n_i}{P_i(1 - P_i)} \left( \frac{\partial P_i}{\partial \lambda} \right)^2 \]

---------- 1.7.3
similarly,

\[
E\left( \frac{\partial^2 l}{\partial \alpha \partial \beta} \right) = - \sum_{i=1}^{k} n_i \left( \frac{\partial P_i}{\partial \alpha} \frac{\partial P_i}{\partial \beta} \right) \frac{1}{P_i(1 - P_i)} \tag{1.7.4}
\]

To obtain solution to the equations,

\[
\frac{\partial l}{\partial \alpha} = \frac{\partial l}{\partial \beta} = 0
\]

we need to iterate using the method of scoring by Morgan (1992) which gives

\[
\begin{pmatrix} \alpha^{t+1} \\ \beta^{t+1} \end{pmatrix} = \begin{pmatrix} \alpha^t \\ \beta^t \end{pmatrix} + A^{-1} \begin{pmatrix} \frac{\partial l}{\partial \alpha} \\ \frac{\partial l}{\partial \beta} \end{pmatrix} \tag{1.7.5} \text{ for } t \geq 0
\]

where

\[
A = -E \begin{bmatrix}
\frac{\partial^2 l}{\partial \alpha^2} & \frac{\partial^2 l}{\partial \alpha \partial \beta} \\
\frac{\partial^2 l}{\partial \alpha \partial \beta} & \frac{\partial^2 l}{\partial \beta^2}
\end{bmatrix}
\]

1.8 The relation of logistic model and iterated weighted regression; generalized model.

From the equations 1.7.2, 1.7.3 and 1.7.4.

\[
A = \begin{bmatrix}
\sum_{i=1}^{k} n_i P_i(1 - P_i) & \sum_{i=1}^{k} n_i P_i(1 - P_i) d_i \\
\sum_{i=1}^{k} n_i d_i P_i(1 - P_i) & \sum_{i=1}^{k} n_i d_i^2 P_i(1 - P_i)
\end{bmatrix}
\]
thus 1.7.5 can be rewritten as

$$A\left(\alpha^{(t+1)}\right) = A\left(\alpha^t\right) + \left(\frac{\partial l}{\partial \alpha}\right)\left(\frac{\partial l}{\partial \beta}\right)$$

for \( t \geq 0 \)

$$\Rightarrow \left[ \sum_{i=1}^{k} n_i P_i (1 - P_i) \frac{\partial}{\partial \alpha} + \sum_{i=1}^{k} n_i d_i P_i (1 - P_i) \frac{\partial}{\partial \beta} \right] \left(\alpha^{(t+1)}\right) = \left[ \sum_{i=1}^{k} n_i P_i (1 - P_i) \frac{\partial}{\partial \alpha} + \sum_{i=1}^{k} n_i d_i P_i (1 - P_i) \frac{\partial}{\partial \beta} \right] \left(\alpha^t\right) + \left(\frac{\partial l}{\partial \alpha}\right)$$

Now let \( w_i = P_i (1 - P_i) \) then the above equation becomes

$$\left[ \sum_{i=1}^{k} n_i w_i \frac{\partial}{\partial \alpha} + \sum_{i=1}^{k} n_i d_i w_i \frac{\partial}{\partial \beta} \right] \left(\alpha^{(t+1)}\right) = \left[ \sum_{i=1}^{k} n_i w_i \frac{\partial}{\partial \alpha} + \sum_{i=1}^{k} n_i d_i w_i \frac{\partial}{\partial \beta} \right] \left(\alpha^t\right) + \left(\frac{\sum_{i=1}^{k} n_i (P_i - P_i)}{\sum_{i=1}^{k} n_i d_i (P_i - P_i)}\right)$$

$$= \left[ \sum_{i=1}^{k} n_i w_i \alpha' + \sum_{i=1}^{k} n_i d_i w_i \beta' \right] + \left(\frac{\sum_{i=1}^{k} n_i (P_i - P_i)}{\sum_{i=1}^{k} n_i d_i (P_i - P_i)}\right)$$

$$= \left[ \sum_{i=1}^{k} n_i w_i \alpha' + \sum_{i=1}^{k} n_i d_i w_i \beta' + \sum_{i=1}^{k} n_i (P_i - P_i) \right]$$

$$\sum_{i=1}^{k} n_i d_i w_i \alpha' + \sum_{i=1}^{k} n_i d_i^2 w_i \beta' + \sum_{i=1}^{k} n_i d_i (P_i - P_i^2)$$
where

\[ y'_i = \alpha'^i + d_i \beta'^i + \frac{p_i - P'_i}{w'_i} \]

\( 1 \leq i \leq k \)

\[ \Rightarrow \alpha'^{i+1} = \left( \frac{\sum_{i=1}^{k} n_i w'_i y'_i}{\sum_{i=1}^{k} n_i w'_i d'_i} - \frac{\sum_{i=1}^{k} n_i w'_i d'_i y'_i}{\sum_{i=1}^{k} n_i w'_i} \right) \left( \frac{\sum_{i=1}^{k} n_i w'_i}{\sum_{i=1}^{k} n_i w'_i d'_i} - \frac{\sum_{i=1}^{k} n_i w'_i}{\sum_{i=1}^{k} n_i w'_i d'_i} \right)^2 \]

and

\[ \beta'^{i+1} = \left( \frac{\sum_{i=1}^{k} n_i w'_i y'_i}{\sum_{i=1}^{k} n_i w'_i d'_i} - \frac{\sum_{i=1}^{k} n_i w'_i d'_i y'_i}{\sum_{i=1}^{k} n_i w'_i} \right) \left( \frac{\sum_{i=1}^{k} n_i w'_i}{\sum_{i=1}^{k} n_i w'_i d'_i} - \frac{\sum_{i=1}^{k} n_i w'_i}{\sum_{i=1}^{k} n_i w'_i d'_i} \right)^2 \]

We note that \( y_i \) are called the working logits. This method of scoring can be performed using a suitable computer program that allows for weighted regression, such the weights can be changed accordingly. Hence it follows that after the \( t^{th} \) iteration, we have estimates \( \alpha'^t \) and \( \beta'^t \) which provides us with latest estimates of the logits as

\[ L'_j = \alpha'^t + \beta'^t d_j \]

\( 1 \leq i \leq k \)
and the working logits are given by

\[ y'_i = \left( \alpha' + d_i \beta' + \left( \frac{p_i - P_i'}{w'_i} \right) \right) \]

where

\[ p_i = \frac{r}{n_i} \]

\[ P_i' = \frac{1}{1 + e^{-\beta'}} \]

\[ w'_i = P_i' (1 - P_i') \]

From the \( t \)th iteration we proceed to obtain the \((t + 1)\)th iteration. This process is repeated until the accuracy of solving \( \frac{\partial l}{\partial \alpha} = \frac{\partial l}{\partial \beta} = 0 \) is achieved.

1.9 Summary.

Dose response models have been used extensively in bioassay to describe the probability of a biological response as a function of the dose of some drug or agent. A number of statistical methods have been proposed to quantify dose effects. These include:

(i) Probit method [Bliss (1934) and Gaddum (1948) reviewed later by Finney (1971)].

(ii) Logit method [Ashton (1972)]
Some parametric dose response models that can be treated in a fashion similar to the logistic model include, Wijesinha and Piantadosi (1995),

(a) Logistic (α fixed) \[ p(d) = \frac{1}{1 + e^{-\beta(d-\alpha)}} \]

(b) Weibull \[ p(d) = 1 - e^{-\beta d^a} \]

(c) Weibull (α fixed) \[ p(d) = 1 - e^{-\beta d^a} \]

(d) Weibull \[ p(d) = 1 - e^{-\mu - \beta d^a} \]

(e) Single hit \[ p(d) = 1 - e^{-\beta d} \]

(f) Single hit \[ p(d) = 1 - e^{-\alpha - \beta d} \]

(g) Two stage \[ p(d) = 1 - e^{-\alpha - \beta d - \mu d^2} \]

(h) Two stage \[ p(d) = 1 - e^{-\alpha - \beta d - \mu d^2 - \delta d^3} \]

(i) Sub-maximal response \[ p(d) = \frac{1 - e^{-\beta d}}{1 + \alpha} \]

These models have not been studied thoroughly as logistic. This project uses Weibull model to study dose-response dynamics.
CHAPTER TWO

2. WEIBULL DISTRIBUTION.

2.1 Introduction.

Important examples of non-negative random variables occurring in applications are lifetimes, waiting times, learning times, duration of epidemics and traveling times. Nontemporal examples of non-negative random variables include material integration, particle dimensions, radioactive intensities, rainfall amounts and costs of industrial accidents. Although exponential or gamma distribution provide reasonable fits to the frequency distribution of some of these random variables in some cases the fit is not as close as desired and in other cases the fit is unsatisfactory. Hence other classes of distributions have been introduced to explain the variability of some of these phenomena. One such family is the Weibull distribution named after the Swedish physicist (Hank (1994)).

2.2 General review of the Weibull model.

The density $f(d)$ is given by

$$f(d) = \alpha \beta d^{\alpha-1} \exp(-\beta d^\alpha), \quad d > 0$$
With the shape parameter \( \alpha > 0 \) and scale parameter \( \beta > 0 \). The cumulative distribution function \( G(d) \) equals

\[
G(d) = 1 - \exp(-\beta d^\alpha), d \geq 0.
\]

The mean and the squared coefficient of variation of Weibull density are given by \( E(x) = \frac{1}{\beta^{\frac{1}{\alpha}}} \Gamma\left(1 + \frac{1}{\alpha}\right) \) and \( C_x^2 = \frac{\Gamma\left(1 + \frac{2}{\alpha}\right)}{\left(\Gamma\left(1 + \frac{1}{\alpha}\right)\right)^2} - 1 \).

A unique Weibull distribution can be fitted to each positive random variable with given first two moments. Thus a nonlinear equation in \( \alpha \) must be solved numerically. The Weibull density is always unimodal with a maximum at

\[
d = \beta^{-\frac{1}{\alpha}} \left(1 - \frac{1}{\alpha}\right)^{\frac{1}{\alpha}} \text{ if } C_x^2 < 1 (\alpha > 1)
\]

and

\[
= 0 \text{ if } C_x^2 \geq 1 (\alpha \leq 0)
\]

hence the failure rate function is increasing from 0 to infinity if \( C_x < 1 \) and is decreasing from infinity to zero if \( C_x^2 > 1 \) the gamma and Weibull densities are similar in shape and when \( C_x^2 < 1 \) the lognormal density takes on shapes similar to those of Weibull and gamma.

The gamma and Weibull densities have their maximum value at \( t = 0 \) so that most outcomes will be small and very large outcomes occur occasionally.
On the other hand lognormal distributions will typically produce fewer small outcomes when $t \to 0$ this explains the popular use of lognormal distribution in actual studies. The difference between the densities is their tail behavior where the lognormal has the longest tail followed by gamma then Weibull.

2.3 Relationship between Weibull and Exponential distribution.

The Weibull distribution arises in the theory of extreme values and in life testing. The survivor function.

$$F(t) = \exp(-\beta t^\alpha), \alpha > 0$$

reduces to exponential distribution when $K = 1$ (number of parameters). Every random variable with a Weibull distribution is a power law transformation of an exponentially distributed random variable. A consequence of this correspondence is that if $K$ is known, estimation problems for the Weibull distribution reduce to that of exponential. However if $K$ is not known the solution of the maximum likelihood estimation cannot be obtained in closed form. Estimators of $\beta$ and $\alpha$ have been given by also Cox [1990].
2.4 Maximum likelihood estimation for Weibull distribution. [General formulation of the iterative procedure].

Let us consider a certain test done on a group of \( n_i \) subjects where \( i \) denote the number of different test and \( r_i \) be the number of subjects who respond.

Then will have the likelihood function given as

\[
L = \prod_{i}^{k} \left( \begin{array}{c} n_i \\ r_i \end{array} \right) p_i^n (1 - p_i)^{n_i - r_i} \quad \text{for} \quad i = 1, 2, \ldots, k.
\]

\( p_i \) is the probability of response, for test \( i \) and is given as

\[
p_i = 1 - e^{-\beta x^a}.
\]

Taking the log of the likelihood we have

\[
L = \Phi + \sum_{i}^{k} r_i \ln P_i + \sum (n_i - r_i) \ln (1 - P_i) \quad \text{2.4.1.}
\]

where \( \Phi = \ln \prod_{i}^{k} \left( \begin{array}{c} n_i \\ r_i \end{array} \right) \)

on expanding 2.4.1. By substituting \( P_i \) we obtain

\[
L = \Omega + \sum_{i=1}^{k} r_i \ln (1 - e^{-\beta x^a}) + \sum (n_i - r_i) \ln \left( 1 - (1 - e^{-\beta x^a}) \right)
\]

which simplifies to

\[
L = \Omega + \sum_{i=1}^{k} r_i \ln (1 - e^{-\beta x^a}) - \sum (n_i - r_i) (\beta x^a)
\]
Let us represent \((\alpha, \beta)\) with a vector \(\beta'\) such that \(\beta' = (\alpha, \beta)\) and assume that the Weibull on a vector is also a vector that is Weibull, \(W(P) = x'\beta\)

where \(P' = (P_1, P_2, \ldots, P_k)\)

and \(k\), is the total number of different combinations of explanatory variables that are employed in the experiments. \(X\) is a \(k \times (n + 1)\) design matrix \((x_i)\) the \(i^{th}\) row of which contains value of the explanatory variables for \(P_i\). We can thus generalize the log likelihood function to

\[
L = \Omega + \sum_i^k r_i \ln(1 - e^{-x'P_i}) - \sum_i^k (n_i - r_i) x_i'\beta
\]

\(X_i\) is the \(i^{th}\) row of the matrix \(X\). If we differentiate the above equation with respect to \(\beta\) we get

\[
\sum_i^k \left( \frac{r_i - n_i P_i}{P_i} \right) x_i y_i 
\]

where

\[P_i = 1 - e^{-x_i'\beta}\]

and

\[
\frac{\partial^2 L}{\partial \beta_i \beta_k} = -\sum_i^k r_i (1 - P_i) \frac{x_i x_j}{P_i^2}
\]

The above equation generalize the results by noting that

\[
\frac{\partial P_i}{\partial \beta_j} = \sum_i^k x_i (1 - P_i)
\]
The equations formed by equating equation 2.4.2. to zero can be written in vector notation

\[ X'S = 0 \]

where

\[ S = \frac{r_i - n_i P_i}{P_i} \]

These equations allows us to write the updated estimates using Newton-Raphson method as

\[ \beta'^{t+1} = \beta' + (x'vxtx')^{-1}x's \]

where \( v \) is a diagonal matrix such that \( v = diag(1 - \hat{p}_i) \).

### 2.5 Estimation of Parameters.

In most analysis the outcome variable follows a normal distribution, but there are cases where the outcome variable follows a binomial rather than normal distribution. The Weibull model follows a binomial distribution. Here the maximum likelihood estimation method is preferred to the ordinary least square method in estimation of parameter.

If \( x_i \) is the \( j^{th} \) combination of \( x \) values from the Weibull model, the conditional probability.

\[ E(y \mid x = x) \]
is given by
\[ P_i = 1 - e^{-\beta_i} \]

The contribution to the likelihood function is
\[ P_i^{y_i} (1 - P_i)^{-y_i} \]

If all the cases are independent with no auto-correlation, then the likelihood function is the product of the individual contribution
\[ L = \prod_{i} (P_i^{y_i} (1 - P_i))^{-y_i} \]

We look for the parameter \( \beta' = (\beta_1, \beta_2, \ldots, \beta_s) \), which gives the highest possible value of the likelihood function. This is equivalent to maximizing the log-likelihood function
\[ \ln L = \left( \prod_{i} P_i^{y_i} (1 - P_i)^{-y_i} \right) \]

This equation is differentiated once with respect to each parameter and set to zero that is
\[ \frac{\partial l}{\partial \beta'} = 0 \]

The set of equations obtained are nonlinear and hence cannot be solved using ordinary least square method. The best method is to use iteration procedure in which case a computer find successive approximation for the parameter from initial guess. The iteration stops when the relative changes
on each co-efficient goes below a given value for example 0.00001. For the
above to hold we need to make the following assumptions.

(i) Model should be specified correctly. The conditional probability is a
Weibull function of $x$ variables and hence, the combination of these
two help us in the use of the likelihood function.

(ii) No additional less significant variable are included or omitted from
the model and the $x$ variables must be accurately measured.

(iii) The cases are independent.

(iv) The $x$ variables are independent of each other.

These assumptions make the maximum likelihood estimate unbiased and
efficient.

Problems that arise.

(i) Multicollinearity among the $x$-variables. This leads to biased
estimates of the parameters.

(ii) The ability to detect the response $y$ is very low and can cause costly
results.

(iii) For skewed $x$-variables, symmetrizing transformation can be carried
out before the Weibull analysis is carried out.
Bearing in mind what we have discussed let us go deeper and find out how the Weibull model behaves when the method applied to logistic model is applied to it too.

### 2.6 The method of scoring for the Weibull model.

The likelihood function is given as

\[ L = \prod_{i=1}^{k} \left( \frac{n_i}{r_i} \right) P_i^{n_i} (1 - P_i)^{r_i - n_i} \]

and the log is thus given by

\[ l = \Omega + \sum_{i=1}^{k} r_i \ln P_i + \sum_{i=1}^{k} (n_i - r_i) \ln (1 - P_i) \]

For the Weibull model

\[ P_i = 1 - e^{-\beta t_i^a} \]

then the log-likelihood simplifies to

\[ l = \Omega + \sum_{i=1}^{k} r_i \ln \left( 1 - e^{-\beta t_i^a} \right) + \sum_{i=1}^{k} (n_i - r_i) \ln \left( 1 - \left( 1 - e^{-\beta t_i^a} \right) \right) \]

\[ = \Omega + \sum_{i=1}^{k} r_i \ln \left( 1 - e^{-\beta t_i^a} \right) - \sum_{i=1}^{k} (n_i - r_i) \beta d_i^a \]

now differentiating with respect to \( \alpha \) will give

\[ \frac{\partial l}{\partial \alpha} = \sum_{i=1}^{k} r_i \beta e^{-\beta t_i^a} \frac{\partial d_i^a}{\partial \alpha} - \sum_{i=1}^{k} (n_i - r_i) \beta \frac{\partial d_i^a}{\partial \alpha} \]
\[
\begin{align*}
= & \sum_{i=1}^{k} \frac{r_i \beta e^{-\beta d_i^2} d_i^2 \ln d_i}{1 - e^{-\beta d_i^2}} - \sum_{i=1}^{k} (n_i - r_i) \beta d_i^2 \ln d_i \\
= & \sum_{i=1}^{k} \frac{r_i \beta (1 - P_i) d_i^2 \ln d_i}{P_i} - \sum_{i=1}^{k} (n_i - r_i) \beta d_i^2 \ln d_i \\
= & \sum_{i=1}^{k} \left( \frac{r_i - n_i P_i}{P_i} \right) \beta d_i^2 \ln d_i \\
= & \sum_{i=1}^{k} n_i \left( \frac{P_i}{P_i} - 1 \right) \beta d_i^2 \ln d_i
\end{align*}
\]

note that \( P_i \) is different from \( p_i \)

where \( p_i = \frac{r_i}{n_i} \)

and

\[
\frac{\partial l}{\partial \beta} = \sum_{i=1}^{k} \frac{r_i d_i e^{-\beta d_i^2}}{1 - e^{-\beta d_i^2}} - \sum_{i=1}^{k} (n_i - r_i) d_i^2 \\
= \sum_{i=1}^{k} \frac{r_i d_i^2 (1 - P_i)}{P_i} - \sum_{i=1}^{k} (n_i - r_i) d_i^2 \\
= \sum_{i=1}^{k} \left( \frac{r_i - n_i P_i}{P_i} \right) d_i^2 \\
= \sum_{i=1}^{k} n_i \left( \frac{P_i}{P_i} - 1 \right) d_i^2
\]

where \( p_i = \frac{r_i}{n_i} \)
and \( p_i \neq P_i \).

From the above results

\[
\frac{\partial^2 I}{\partial \beta^2} = \frac{\partial}{\partial \beta} \left( \sum_{i=1}^{k} \frac{r_i - n_i P_i}{P_i} d_i^\alpha \right)
\]

\[
= \frac{\partial}{\partial \beta} \sum_{i=1}^{k} \left( \frac{r_i}{1 - e^{-\beta d_i^\alpha}} - n_i \right) d_i^\alpha
\]

\[
= -\sum_{i=1}^{k} r_i d_i^\alpha e^{-\beta d_i^\alpha} \left( 1 - e^{-\beta d_i^\alpha} \right)
\]

\[
= \sum_{i=1}^{k} n_i P_i d_i^\alpha (1 - P_i)
\]

likewise,

\[
\frac{\partial^2 I}{\partial \alpha^2} = \frac{\partial}{\partial \alpha} \left( \sum_{i=1}^{k} \frac{r_i - n_i}{P_i} d_i^\alpha \beta \ln d_i \right)
\]

\[
= \sum_{i=1}^{k} r_i d_i^\alpha \beta \ln d_i \frac{\partial}{\partial \alpha} \left( 1 - e^{-\beta d_i^\alpha} \right)^{-1}
\]

\[
= -\sum_{i=1}^{k} r_i \beta^2 (d_i^\alpha (\ln d_i))^2 \frac{e^{-\beta d_i^\alpha}}{1 - e^{-\beta d_i^\alpha}}
\]

\[
= \sum_{i=1}^{k} n_i P_i (1 - P_i) \beta^2 (d_i^\alpha (\ln d_i))^2
\]

\( r_i \) results from the binomial distribution \( B(n_i, P_i) \) with mean \( n_i P \) and

\[ E(p_i) = P_i \]

hence
\[-E\left(\frac{\partial^2 I}{\partial \lambda^2}\right) = \sum_{i=1}^{k} \frac{n_i \left( \frac{\partial P_i}{\partial \lambda} \right)^2}{P_i(1-P_i)}\]

where \(\lambda\) denotes the vector \(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}\)

similarly,

\[-E\left(\frac{\partial^2 I}{\partial \alpha \partial \beta}\right) = \sum_{i=1}^{k} \frac{n_i \left( \frac{\partial P_i}{\partial \alpha} \right) \left( \frac{\partial P_i}{\partial \beta} \right)}{P_i(1-P_i)}\]

To obtain solutions to the equations

\[
\frac{\partial l}{\partial \alpha} = \frac{\partial l}{\partial \beta} = 0
\]

we use the method of scoring Morgan [1992], which provides the iterations

\[
\begin{pmatrix} \alpha^{t+1} \\ \beta^{t+1} \end{pmatrix} = \begin{pmatrix} \alpha^t \\ \beta^t \end{pmatrix} + A^{-1} \begin{pmatrix} \frac{\partial l}{\partial \alpha} \\ \frac{\partial l}{\partial \beta} \end{pmatrix} \quad \text{............................2.5.1.}
\]

for \(t \geq 0\)

where \(A\) is expected Fisher information matrix given by

\[
A = -E \begin{bmatrix}
\frac{\partial^2 l}{\partial \alpha^2} & \frac{\partial^2 l}{\partial \alpha \partial \beta} \\
\frac{\partial^2 l}{\partial \alpha \partial \beta} & \frac{\partial^2 l}{\partial \beta^2}
\end{bmatrix}
\]

but
\[ E \left( \frac{\partial^2 l}{\partial \alpha^2} \right) = \sum_{i=1}^{k} n_i \left( \frac{\partial P_i}{\partial \alpha} \right)^2 \frac{1}{P_i(1-P_i)} \]

\[ = \sum_{i=1}^{k} n_i \left( \frac{\beta d_i^\alpha e^{-\beta d_i^\alpha \ln d_i}}{P_i(1-P_i)} \right)^2 \]

\[ = \sum_{i=1}^{k} n_i \frac{(1-P_i)}{P_i} \left( \beta d_i^\alpha \ln d_i \right)^2 \]

similarly

\[ -E \left( \frac{\partial^2 l}{\partial \beta^2} \right) = \sum_{i=1}^{k} n_i \left( \frac{\partial P_i}{\partial \beta} \right)^2 \frac{1}{P_i(1-P_i)} \]

\[ = \sum_{i=1}^{k} n_i \left( \frac{d_i^\alpha (1-P_i)}{P_i(1-P_i)} \right)^2 \]

\[ = \sum_{i=1}^{k} n_i \frac{(1-P_i)d_i^{2\alpha}}{P_i} \]

hence

\[ -E \left( \frac{\partial^2 l}{\partial \alpha \partial \beta} \right) = \sum_{i=1}^{k} n_i \left( \frac{\partial P_i}{\partial \alpha} \right) \left( \frac{\partial P_i}{\partial \beta} \right) \frac{1}{P_i(1-P_i)} \]

\[ = \sum_{i=1}^{k} n_i \frac{(1-P_i)\beta d_i^{2\alpha} \ln d_i}{P_i} \]

where
2.7 The connection with the iterated weighted regression.

From the above section we can substitute the derivations to obtain.

\[
A = \begin{bmatrix}
\sum_{i=1}^{k} n_i \frac{1-P_i}{P_i} (\beta d_i^{\alpha} \ln d_i)^2 & \sum_{i=1}^{k} n_i \frac{1-P_i}{P_i} \beta d_i^{2\alpha} \ln d_i \\
\sum_{i=1}^{k} n_i \frac{1-P_i}{P_i} \beta d_i^{2\alpha} \ln d_i & \sum_{i=1}^{k} n_i \frac{1-P_i}{P_i} d_i^{2\alpha}
\end{bmatrix}
\]

\[
= \begin{bmatrix}
\sum_{i=1}^{k} n_i \beta (\ln d_i)^3 w_i & \sum_{i=1}^{k} n_i \beta (\ln d_i) w_i \\
\sum_{i=1}^{k} n_i \beta (\ln d_i) w_i & \sum_{i=1}^{k} n_i w_i
\end{bmatrix}
\]

where \( w_i = \frac{(1-P_i)d_i^{2\alpha}}{P_i} \).

hence equation 2.5.1. above is equivalent to

\[
A_{\alpha^{t+1}} = A_{\beta^{t}} + \left( \frac{\partial f}{\partial \alpha} \right)_{\beta^{t}}
\]

substituting the value for \( \frac{\partial f}{\partial \alpha} \) and \( \frac{\partial f}{\partial \beta} \) we have

\[
A_{\alpha^{t+1}} = \left( \sum_{i=1}^{k} n_i \left( \beta w_i^{\alpha'} (\ln d_i) + \beta w_i^{\alpha'} \ln d_i + \frac{P_i-P_i^t}{P_i} \beta d_i^{\alpha'} \ln d_i \right) \right)
\]

\[
= \left( \sum_{i=1}^{k} n_i \beta (\ln d_i) w_i^t y_i^t \right)
\]

where
\[ y'_i = \alpha' \beta \ln d_i + \beta' + \frac{P_i - P_i'}{P_i' w_i} d_i' \]

To obtain the \((t + 1)\)th iterates that is, \((\alpha^{t+1}, \beta^{t+1})\) From the above expression, we just obtain the inverse of \(A\) and multiply on both sides of equation 2.5.2.

to obtain \(A^{-1} = \frac{1}{D} \begin{bmatrix} \sum_{i=1}^{k} n_i w_i & -\sum_{i=1}^{k} n_i \beta w_i \ln d_i \\ -\sum_{i=1}^{k} n_i w_i \ln d_i & \sum_{i=1}^{k} n_i w_i \beta^2 (\ln d_i)^2 \end{bmatrix} \]

where \(D = \) determinant, then

\[
\alpha^{t+1} = \frac{\left( \sum_{i=1}^{k} n_i w_i \beta (\ln d_i) w'_i y'_i \right) - \left( \sum_{i=1}^{k} n_i w_i \beta (\ln d_i) \right) \left( \sum_{i=1}^{k} n_i w'_i y'_i \right)}{\left( \sum_{i=1}^{k} n_i w_i \beta^2 (\ln d_i)^2 \right) \left( \sum_{i=1}^{k} n_i w_i \right) - \left( \sum_{i=1}^{k} n_i w_i \beta (\ln d_i) \right)^2}
\]

\[
\beta^{t+1} = \frac{\left( \sum_{i=1}^{k} n_i w_i \beta^2 (\ln d_i)^2 \right) \left( \sum_{i=1}^{k} n_i w'_i y'_i \right) - \left( \sum_{i=1}^{k} n_i w'_i y'_i \beta (\ln d_i) \right) \left( \sum_{i=1}^{k} n_i w_i \beta (\ln d_i) \right)}{\left( \sum_{i=1}^{k} n_i w_i \beta^2 (\ln d_i)^2 \right) \left( \sum_{i=1}^{k} n_i w_i \right) - \left( \sum_{i=1}^{k} n_i w_i \beta (\ln d_i) \right)^2}
\]

We have seen that when the method of scoring is applied to both models, Weibull and Logistic, we end up with a structure, which looks like that of logistic model. Now the question is which is the optimal model. We shall use the idea of Akaike Information Criterion to answer this, which we review in chapter three.
CHAPTER THREE

AKAIKE'S INFORMATION CRITERION.

3.1 Introduction.

In this section we present a heuristic introduction to the derivation of AIC. A more technical account is given later. Akaike’s [1973] seminal paper proposed the use of the Kullback–Liebler [K-L] distance as a fundamental basis for model selection.

However K-L distance cannot be computed without full knowledge of the true model \( f(x) \), the form of the proposed model, \( g(x | \theta) \), and a defined parameter space \( \Theta \) (that is the set of all possible values of \( \theta \) ). The specific value of the parameter \( \theta \) needed in the model can be computed given \( f(x) \) and the functional form of the model.

There has to be a specific value of \( \theta \) defining the model \( g(x | \theta) \) as one member of a family of models indexed by \( \theta \in \Theta \) which is unique in minimizing K-L distance

\[ I(f;g) , \ g = g(x | \theta) \]

and depends on the truth \( f \) the model \( g \) through its structure and sample space.
3.2. Kullback-Leibler distance definition.

From the above it follows that there is a true value of $\theta$ underlying maximum likelihood estimation, but that value of $\theta$ depends on the assumed model and can vary by model. For the true $\theta$ under $g$, K-L is minimized. This property of the model $g(x | \theta)$ as the minimizer of K-L, over all $\theta \in \Phi$ is an important feature involved in derivation of AIC.

In data analysis the model parameter must be estimated, and there is usually substantial uncertainty in this estimation in that model based on estimated parameters $\hat{\theta}$ not $\theta$. This difference causes us to change the model selection to that of minimizing expected estimated K-L distance rather than minimizing the known K-L distance

$$\hat{I}(f;g) = \int f(x) \log \left( \frac{f(x)}{g(x | \hat{\theta}(y))} \right) dx \quad \text{..........3.2.1.}$$

where it is important to note that the variable of integration $x$ does not represent the data but the data is represented by $y$. In fact proper thought of $x$ has been integrated out of $I(f,g)$ and what remains is really a function of $\theta$ given the structural model $g$ and influence of the truth $f$.

hence $x$ disappears in $I(f,g)$ and $\hat{I}(f,g)$. Using K-L distance does not necessarily need data but is necessary for computation of $y$. 

3.3 The Kullback-Leibler estimator.

In context of repeated sampling properties of an inference procedure we expect our estimated K-L to have on average the positive value

\[ E_g(\hat{I}(f,g)) = \int f(y) \left( \int f(x) \log \left( \frac{f(x)}{g(x | \theta(y))} \right) dx \right) dy \] .......................... 3.3.1.

Since for the model structure \( g \), minimizes K-L distance then

\[ E_g(\hat{I}(f,g)) > I(f,g) \]

Now from equation 3.2.1

\[ \hat{I}(f,g) = \int f(x) \log \left( \frac{f(x)}{g(x | \theta(y))} \right) dx \]

\[ = \int f(x) \log f(x) dx - \int f(x) \log \left( g(x | \theta(y)) \right) dx \]

\[ = constant - Ex \left( \log \left( g(x | \theta(y)) \right) \right) \] .......................... 3.3.2.

And from 3.3.1. We have

\[ E(\hat{I}(f,g)) = \int f(y) \left( \int f(x) \log \left( \frac{f(x)}{g(x | \theta(y))} \right) dx \right) dy \]
\[
\int f(y) \left( \int f(x) \log f(x) \, dx \right) \, dy - \int f(y) \left( \int f(x) \log g(x \mid \hat{\theta}(y)) \, dx \right) \, dy
\]

\[
= \left( \int f(y) \, dy \right) \left( \int f(x) \log f(x) \, dx \right) - \int f(y) \left( \int f(x) \log g(x \mid \hat{\theta}(y)) \, dx \right) \, dy
\]

\[
= \text{const} - E_y E_x \left( \log g(x \mid \hat{\theta}(y)) \right). \quad \text{..................3.3.3.}
\]

### 3.4 The relationship between Kullback-Leibler and Akaike Information criterion.

If we were to base K-L model selection on 3.3.2, we would have to compute

\[
E_x \left( \log g(x \mid \hat{\theta}(y)) \right)
\]

which is impossible thus we are forced to use the 3.3.3. above, where the necessary basis for applied K-L information theoretic model selection is the quantity

\[
E_y E_x \left( \log g(x \mid \hat{\theta}(y)) \right) \quad \text{..................3.4.1}
\]

which is related to the K-L distance by

\[
E_y E_x \left( \log g(x \mid \hat{\theta}(y)) \right) = \text{const} - E_{\hat{\theta}} \left( \hat{\imath}(f, g) \right)
\]

thus \(E_y E_x \left( \log g(x \mid \hat{\theta}(y)) \right)\) is the relative expected K-L distance, which depends on truth \(f\), model structure and true parameter \(\theta\) given \(f \text{ and } g\) and can be estimated based on data and model, Burnham et al (1994)
Consequently, we can determine a method to select the model \( g_i \), that on average minimizes, over the set of models, \( g_i \rightarrow \cdots \), a very relevant expected K-L distance. The model \( g_i(x | \theta) \) that maximizes 3.4.1 is the K-L best model when \( \theta \) must be estimated, and maximum likelihood theory is used to get \( \hat{\theta} \).

In applications we fit each model to the data (getting \( \hat{\theta} \) by model) and then get estimated values of our criterion 3.4.1. For large sample sizes the maximum of both

\[
E_x E_{x'} (\log(g(x | \hat{\theta}(y)))
\]

and

\[
E_{x'} (\log(g(x | \hat{\theta}(y)))
\]

have a ratio almost 1, so AIC will then essentially be selecting on average the K-L best model from the set of models.
3.5 The Akaike Information Criterion definition.

Akaike (1973, 1985, 1994) first showed that the critical issue for getting an applied K-L model selection method was to estimate

$$E_x E_y \{ \log \{ g(x | \hat{\theta}(y)) \} \} .$$

This expression involves the log of the probability of the model for the data. That is the conceptual

$$\log \{ g(x | \hat{\theta}(y)) \}$$

bears a strong resemblance to the actual log-likelihood

$$\log \{ L(\hat{\theta}(y) | y) \} = \log \{ g(y | \hat{\theta}(y)) \}$$

Moreover, because we have directly available to estimate this target model selection criterion is our log likelihood (for each model) the temptation is to just estimate

$$E_x E_y \{ \log g(x | \hat{\theta}(y)) \}$$

by the maximized

$$\log \{ L(\hat{\theta}(y) | y) \} = \log \{ g(y | \hat{\theta}(y)) \}$$

for each model.

However Akaike (1973) showed that the maximized log likelihood is

(i) Biased upward as an estimator of the model selection target criterion

3.3.4
(ii) Under certain conditions this bias is approximately equal to $K$, the number of estimable parameters in the approximate model.

Thus an approximate unbiased estimator of

$$T = E_xE_y \left( \log \left( g \left( x \mid \hat{\theta}(y) \right) \right) \right)$$

for large samples and "good" models is

$$\hat{T} = \log \left( L \left( \hat{\theta} \mid y \right) \right) - K \quad \ldots \quad 3.5.1$$

where the notation $\hat{\theta}(y) = \hat{\theta}$

This result is equivalently to

$$\log L \left( \hat{\theta} \mid y \right) - K = \text{cons} \tan t - \hat{E}_\theta \left( \ell(f, g) \right)$$

or

$$- \log L \left( \hat{\theta} \mid y \right) + K = \text{estimated relative expected K-L distance.}$$

Akaike [1973] then defined an information criterion (AIC) by multiplying equation 3.5.1 with $-2$ to get

$$\text{AIC} = -2 \log L \left( \hat{\theta} \mid y \right) + 2k$$

It is important to note that AIC has a strong theoretical underpinning, based on information theory and Kullback Liebler information within a realistic data analysis philosophy that no model is true; rather, truth as $f$ is far more complex than any model used. Thus one should select the model that yields the smallest value of AIC because this model is estimated to be "closest" to
the unknown reality that generated the data from among the candidate models considered.

3.6 An application of AIC.

Let candidate models $g_1, g_2, g_3$, and $g_4$ have AIC values of 3400, 3560, 3380 and 3415 respectively. Then one would select model $g_3$ as the best single model as the basis for inference because $g_3$ has the smallest value. Since these values are relative (additive) scale, one could subtract 3000 to achieve 400, 560, 380 and 415. This does not change the ranks of the models nor pairwise differences in the AIC values. Thus the absolute size of the AIC value is not important but the difference between these values. Hence we recommend computing the AIC differences (rather than the actual AIC values)

$$D_i = AIC_i - \min AIC = E_\theta(\hat{I}(f, g)) - \min E_\theta(\hat{I}(f, g))$$

over all candidate models in the set which estimate the relative expected K-L differences between $f$ and $g_i(x | \theta)$.

These $D_i$ values are easy to interpret and allow quick comparison and ranking of candidates models and are also useful in computing Akaike
weights. The larger \( D \) is, the less plausible is the fitted model \( g_r(x | \hat{\theta}) \) as being the K-L best model for sample such as the data above has.

Thus when \( D \) is

\[ \leq 2 \; ; \text{then the models have substantial support and should be considered in making inference.} \]

\[ 4 \leq D \leq 7 \; ; \text{the models have considerable less support} \]

\[ > 10 \; ; \text{the models have essentially no support, and might be omitted from further consideration, or at least the models fail to explain some substantial explainable variation in the data.} \]

In the above example model \( g_3 \) has \( D_3 = 3380 - 3380 = 0 \) whilst for model \( g_1, g_2, \) and \( g_4 \) we have \( D \) values of 20, 180 and 35 respectively. This rescaling to a minimum relative AIC makes comparison between the best model and another easy. Hence in this example model \( g_3 \) is best fitted.

Now let us look at the derivation of AIC in details.

3.7 Akaike predictive expected log-likelihood.

This section provides further insights into Akaike's (1973) result that an approximately unbiased estimator of the relative K-L distance for large samples and good models is \( \log \log (\hat{\theta} | data) - k \). The argument that we should select a model to minimize K-L distance leads us to focus on the quantity
\[ E_x(\log(g(x | \theta))) \]

Next we introduce \( \hat{\theta}(y) \) into this function. Conceptually we can do so, but still

\[ E_x \log(g(x | \hat{\theta}(y))) \]

cannot be computed although it can be approximated. A second order Taylor's series expansion of

\[ \log(g(x | \hat{\theta}(y))) \]

about \( \theta \) leads to an interesting quadratic approximation. If the model is close to the truth then

\[ E_x(\log(g(x | \hat{\theta}(y)))) \approx E_x(\log L(\hat{\theta}|data)) - \frac{1}{2} k - \frac{1}{2} (\hat{\theta} - \theta) \Sigma^{-1}(\hat{\theta} - \theta) \ldots \ldots 3.7.1. \]

The general result of the derivation is given in detail later in the chapter.

The \( \log(L(\hat{\theta} | data)) \) is unbiased estimator of \( E_\hat{\theta}(\log L(\hat{\theta} | data)) \) where \( k \) is known.

There are several ways to compute a good estimate of the unknown variance-covariance matrix, \( \Sigma \)

hence we get

\[ \hat{E}(\log(g(x | \hat{\theta}(y)))) = \log(L(\hat{\theta} | data)) - \frac{1}{2} k - \frac{1}{2} (\hat{\theta} - \theta) \Sigma^{-1}(\hat{\theta} - \theta) \]

taking expectation on both sides of equation 3.7.1, we have

\[ E_y E_x(\log(g(x | \hat{\theta}(y)))) = E_\hat{\theta}(\log(L(\hat{\theta} | data))) - k \]
(The data are \( y \), and expectation with respect to \( \hat{\theta} \), \( E_{\hat{\theta}} \) is the same as \( E_{y} \))

Akaike, in defining and deriving AIC, sometimes uses a particular predictive expectation for the log-likelihood function of the data and unknown parameters. This has advantages and properties that are still not well recognized in the literature. Deeper insight into the derivation of AIC are given in Akaike (1973, 1981, 1992), Bozdogan (1981), Sakamoto (1991), de Leeuw (1992), and later on in this chapter.

This alternative path AIC again shows us that we have to adopt our criterion for selecting the best fitted model, the maximization of

\[ E_{y}E_{x}(\log(g(x | \hat{\theta}(y)))) \]

the expected relative K-L distance.

Hence we are really implementing formulae 3.3.3. as our model selection criterion since we have to estimate \( \theta \). With this realization we need to consider what \( \log(g(x | \hat{\theta}(y))) \) might mean; in fact, which Akaike called predictive likelihood. Deriving AIC is sometimes given in terms of a satisfied expectation based on a different, independent sample, \( x \), as well as data, \( y \). Then one considers the conceptual predictive model \( g(x | \hat{\theta}(y)) \) used for making prediction about a new sample \( x \) that could be collected, based on actual data \( y \). The predictive log likelihood is given as
with expected value
\[ E_y E_x \left( \log \left( \frac{g(x | \hat{\theta}(y))}{g(x | \hat{\theta}(y))} \right) \right) \]

which is useful as a criterion for model selection. This expectation over a second conceptual independent 'data set' provides AIC with a cross validation property. Symbolically Akaike's expected predictive log-likelihood could be given as
\[ E_p \left( \log \left( \frac{L(\hat{\theta})}{L(\hat{\theta})} \right) \right) = E_y E_x \left( \log \left( \frac{L(\hat{\theta} | y)}{L(\hat{\theta} | x)} \right) \right) \]

where
\[ E_p \] denotes Akaike's predictive expectation.

3.8. A second-order AIC important refinement

Akaike derived an estimator of the K-L information quantity. However, AIC may perform poorly if there are too many parameters in relation to the size of the sample (Sugiura 1978, Sakamoto et al. 1986) Sugiura (1978) derived a second order variant of AIC that he called c-AIC. Hurvich and Tsai (1989) further studied this small sample (second order) bias adjustment, which lead to a criterion called \( AIC_c \), which is given as
\[ AIC_c = -2 \log(L(\hat{\theta})) + 2k \left( \frac{n}{n-k-1} \right) \]

where the penalty term is multiplied by the correction factor \( \left( \frac{n}{n-k-1} \right) \)

\( AIC_c \) can also be written as
\[ AIC_c = -2 \log(L(\hat{\theta})) + 2k + \frac{2k(k+1)}{n-k-1} \]
\[ = AIC + \frac{2k(k+1)}{n-k-1} \text{ where } n \text{ is sample size.} \]

If \( n \) is large with respect to \( k \) then the second order correction is negligible and AIC works well. While \( AIC_c \) is derived under Gaussian assumptions for linear models (fixed effects) Burnham et al. (1994) found this second order approximation to the K-L distance to be useful in product multinomial models. Generally \( AIC_c \) is suitable when \( \frac{n}{k} \) is small (say \( \leq 40 \)).

3.9 Useful preliminaries.

The sole purpose for this section is to provide a summary of the basic notation, concepts and mathematical background needed to understand the derivation of AIC.
In model selection criterion, it is clear that AIC is for a model with $K$ estimated parameters, $\hat{\theta}$ being the MLE of those parameters, computed from the data $x$ under assumed model $g(x \mid \theta)$.

$$= -2\log L(\hat{\theta}) + 2K$$

Without loss of generality we take the likelihood of $\theta, as, L(\theta \mid x) = g(x \mid \theta)$ by simply interpreting $g$ as a function of $\theta$ given $x$. A second dual usage of notation for the random variable $x$ arises. Sometimes as the data (as a random variable) and times denotes the variable of integration with respect to $f(x)$. Always the data arise from $f()$ but not from $g(\mid \theta)$. AIC has been motivated justified and derived in a variety of ways for example Akaike 1973, Sawa 1978, Sugiura 1978, Chow 1981, Stone 1982, Shibata 1989, Bozdogan 1987, but these derivations are often cryptic and thus difficult to follow however we shall give a general derivation.

The most general approach of deriving AIC was the Taylor series expansion to second order. If $h(\theta)$ is real valued function on $K$ dimensions then the Taylor series expansion about some value $\theta_0$ near $\theta$ is given as

$$h(\theta) = h(\theta_0) + \left( \frac{\partial h(\theta)}{\partial \theta} \right)(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^T\left( \frac{\partial^2 h(\theta_0)}{\partial \theta^2} \right)(\theta - \theta_0) + \text{Re} \ldots \ldots 3.9.1.$$
where \( \theta \) and \( \theta_0 \) are two different points in the space over which \( h() \) is defined and \( \text{Re} \) is the remainder term for quadratic Taylor series expansion which has an approximation of order \( O\left(\|\theta - \theta_0\|^3\right) \)

for any vector argument \( Z - W \)

\[
\|Z - W\| = \sqrt{\sum_{i=1}^{K} (Z_i - W_i)^2}
\]

which denotes the Euclidian distance between the two points in the \( K \)-dimensional space. The 1st differential term on the right side of equation 3.9.1. \( \frac{\partial h(\theta)}{\partial \theta} \) is a \( K \times 1 \) column vector that is.

\[
\begin{bmatrix}
\frac{\partial h(\theta)}{\partial \theta_1} \\
\frac{\partial h(\theta)}{\partial \theta_2} \\
\vdots \\
\frac{\partial h(\theta)}{\partial \theta_K}
\end{bmatrix}
\]

and for the second differential term we have

\[
\left( \frac{\partial^2 h(\theta)}{\partial \theta^2} \right) = \left( \frac{\partial^2 h(\theta)}{\partial \theta_i \partial \theta_j} \right)_{\theta = \theta_0}
\]

\( i = 1, 2, \ldots, k \) and \( j = 1, 2, \ldots, k \).

which denotes the \( K \times K \) matrix of second mixed partial derivative of \( h(\theta) \) with respect to \( \theta_1, \ldots, \theta_k \) evaluated at \( \theta = \theta_0 \). This matrix is called the
Hessian of $h(\theta)$. Expression 3.9.1. above when terminated at the quadratic term then we get an approximation of $h(\theta)$.

For cases of interest $h(.)$ will be a log-likelihood based on a probability distribution which gives rise to $E(\hat{\theta}) = \theta_0$ of $O\left(\frac{1}{n}\right)$

For large samples $n \to \infty, \hat{\theta} \to \theta_0$ with probability 1 and Taylor expression 3.9.1. above is quite good hence simplifies to

$$h(\hat{\theta}) = h(\theta_0) + \frac{\partial h(\theta_0)}{\partial \theta}(\hat{\theta} - \theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)'\left(\frac{\partial^2 h(\theta_0)}{\partial \theta^2}\right)(\hat{\theta} - \theta_0) + O\left(\frac{1}{n}\right)$$

In the context of parametric maximum likelihood estimator the standard approach is to assume that one specific member of a family of models generates the data. By assumption, truth corresponds to one specific (but unknown) value of $\theta$ which is denoted by $\theta_0$ is our target model under likelihood inference.

Approached theoretically ignoring issues of data and estimation, the best approximating model $g$ in a class of models considered, under the (compelling) K-L information measure, is simply the model that produces the minimum K-L over $\Phi$ (the whole space). Hence we look for a unique value of $\theta \in \Phi$ which we will denoted by $\theta_0$ that provides the K-L best approximating model $\theta_0$ is the solution to the optimization problem
Clearly \( g(x \mid \theta_0) \) is the best model here where \( \theta_0 \) satisfies the vector equations

\[
\frac{\partial}{\partial \theta} \int f(x) \log \left( \frac{f(x)}{g(x \mid \theta_0)} \right) dx = 0
\]

\[= \frac{\partial}{\partial \theta} \int f(x) \log f(x) dx - \frac{\partial}{\partial \theta} \int f(x) \log g(x \mid \theta) dx = 0\]

Since in the 1st term there is no \( \theta \) in \( f(.) \) then it simplifies to 0 whilst the second is

\[
\int f(x) \left( \frac{\partial}{\partial \theta} \log g(x \mid \theta) \right) dx = E_f \left( \frac{\partial}{\partial \theta} \log g(x \mid \theta_0) \right)_{\theta = \theta_0} = 0
\]

\[\Rightarrow E_f \left( \frac{\partial}{\partial \theta} \log g(x \mid \theta_0) \right) = 0\]

But we know if \( x_i \) is and an iid of sample size \( n \) then probability distribution function \( f(x) = \prod_{i=1}^{n} f(x_i) \) and we consider the maximum likelihood estimate under model \( g(x \mid \theta) = \prod_{i=1}^{n} g(x_i \mid \theta_0) \) then for every \( n \) we have \( K \) likelihood equations (expressed as mean wlg)

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log \left( g(x_i \mid \hat{\theta}_0) \right) = 0 \quad \text{If } h(.) \text{ is the log likelihood}
\]

\[
\log(g(x \mid \theta)) \text{, then we have 3.9.2 as}
\]
\[
\left( \frac{\partial^2 \log(g(x | \theta))}{\partial \theta_i \partial \theta_j} \right)_{\theta = \theta_0} \text{ Which is related to the Fisher information matrix}
\]

\[
I(\theta_0) = E_q \left( - \frac{\partial^2 \log(g(x | \theta))}{\partial \theta_i \partial \theta_j} \right)_{\theta = \theta_0} \quad \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots 3.9.3
\]

If \( g(.) \) is the true model form that is \( f \equiv g \) then the sampling variance-covariance matrix \( \sum \) of the maximum likelihood estimator is (for large sample) \( \sum = (I(\theta_0))^{-1} \) that is

\[
\sum = E (\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)' \text{ is } (I(\theta_0))^{-1}.
\]

In fact in deriving AIC we have expectations with respect to \( f \) thus

\[
I(\theta_0) = E_f \left( - \frac{\partial^2 \log(g(x | \theta))}{\partial \theta_i \partial \theta_j} \right) \quad \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots 3.9.4
\]

In cases where \( f = g \) then we have \( I(\theta_0) = I(\theta_0) \)

Additional information here is the empirical but unknown matrix

\[
\hat{I}(\theta_0) = \left( - \frac{\partial^2 \log(g(x | \theta))}{\partial \theta_i \partial \theta_j} \right)_{\theta = \theta_0}
\]

For simpler notation we shall use \( I(\theta_0) = E_f \left( - \frac{\partial^2 \log g(x | \theta_0)}{\partial \theta^2} \right) \) which means exactly the same as 3.9.4 and hence simpler notation for the \( \hat{I}(\theta_0) \) is

\[
\hat{I}(\theta_0) = - \frac{\partial^2 \log g(x | \theta_0)}{\partial \theta^2}. \text{ It is obvious that } E_f \hat{I}(\theta_0) = I(\theta_0)
\]

and as \( n \to \infty \)
\[ \hat{i}(\theta) = I(\theta) + \text{Re} \]

where \( \text{Re} \) is \( O\left(\frac{1}{n}\right) \)

There are two ways of computing Fisher information matrix \( I(\theta) \) of equation 3.9.3 when \( f = g \).

Since \( \int g(x \mid \theta) dx = 1 \)

then it follows that \( \int \frac{\partial g(x \mid \theta)}{\partial \theta} dx = 0 \)

which implies

\[
\frac{\partial \log(g(x \mid \theta))}{\partial \theta} = \frac{1}{g(x \mid \theta)} \left( \frac{\partial g(x \mid \theta)}{\partial \theta} \right)
\]

and hence we get

\[
\int g(x \mid \theta) \left( \frac{\partial}{\partial \theta} \log(g(x \mid \theta)) \right) dx = 0
\]

Now applying partial derivative vector we have

\[
\int g(x \mid \theta) \left( \frac{\partial}{\partial \theta} \log(g(x \mid \theta)) \right) \left( \frac{\partial}{\partial \theta} \log(g(x \mid \theta)) \right) dx + \int g(x \mid \theta) \frac{\partial^2 \log(g(x \mid \theta))}{\partial \theta^2} = 0 \quad \text{3.9.5}
\]

(0 is a \( K \times K \) matrix of zero elements)

We can rewrite the above equation 3.9.5. as

\[
E_g \left( \frac{\partial}{\partial \theta} \log(g(x \mid \theta)) \right) \left( \frac{\partial}{\partial \theta} \log(g(x \mid \theta)) \right) = E_g \left( - \frac{\partial^2 \log(g(x \mid \theta))}{\partial \theta^2} \right)
\]

\[
\Rightarrow E_g \left( \frac{\partial}{\partial \theta} \log(g(x \mid \theta)) \right) \left( \frac{\partial}{\partial \theta} \log(g(x \mid \theta)) \right) = I(\theta) = J(\theta)
\]
Where \( J(\theta) = E(\partial/\partial \theta \log(g(x | \theta))) \partial/\partial \theta \log(g(x | \theta)) \)

But \( J(\theta) = E_f(\partial/\partial \theta \log(g(x | \theta))) \partial/\partial \theta \log(g(x | \theta)) \) ..........3.9.6.

\( J(\theta) = J(\theta) \) when \( f = g \) although \( J(\theta) = I(\theta) \). There is no general equality between \( I(\theta) \) and \( J(\theta) \) when \( g \) is only and approximation of \( f \), hence when the K-L discrepancy between \( f \) and \( g \), \( I(f,g) \) is greater than 0. However we can have near equalities \( I(\theta_0) = J(\theta_0) \), \( I(\theta_0) = I(\theta_0) \) and \( J(\theta_0) = J(\theta_0) \) when \( I(f,g) = 0 \)

hence a good approximating model is used.

There is a large sample relationship among \( I(\theta_0), J(\theta_0) \) and \( \sum \) which is used more \( I(\theta_0)\sum = J(\theta_0)(I(\theta_0))^{-1} \) ..3.9.7.

3.10 A general derivation of AIC.

\( I(f,g(\theta_0)) = \int f(x) \log \left( \frac{f(x)}{g(x | \theta_0)} \right) dx \) which is independent of any data nor any value \( x \), as \( x \) has been integrated out. Given that we have data sample \( y \) from \( f(\cdot) \) then we need to find maximum likelihood estimator \( \hat{\theta} = \hat{\theta}(y) \) and compute \( I(f,g(\hat{\theta}(y))) = \int f(x) \log \left( \frac{f(x)}{g(x | \hat{\theta}(y))} \right) dx \).
There is no unique way [path] from K-L to AIC we shall thus use two approaches.

(i) Since our estimator \( \hat{\theta}(y) \) would not be equal to \( \theta_0 \) for continuous distribution and for special discrete distribution then the above equality will be with probability \(<1\). Any value of \( \hat{\theta}(y) \) other than \( \theta_0 \) will result to

\[
I(f, g(\hat{\theta}(y))) > I(f, g(\theta_0))
\]

hence our model will not be minimizing \( I(f, g(\theta_0)) \) but slightly larger value on average given by

\[
E_y[I(f, g(\hat{\theta}(y)))] > I(f, g(\theta_0))
\]

Note that

\[
E_y[I(f, g(\theta_0))] = \frac{1}{2} tr(I(\theta_0))I(\theta_0)^{-1}
\]

Now given the reality we must estimate \( \theta \) thus we adopt the criterion select the model \( g \) to minimize

\[
E_y[I(f, g(\hat{\theta}(y)))]
\]

\[
\Rightarrow E_y[I(f, g(\hat{\theta}(y)))] = \int f(x)\log(f(x))dx - E_y[\int f(x)\log g(x | \hat{\theta}(y))dx]
\]

\[
= \text{constant} - E_xE_y\log(g(x | \hat{\theta}(y)))
\]

\( E_xE_y\log(g(x | \hat{\theta}(y)) \) can be estimated thus we use AIC to give the best model for K-L.
(ii) The second, and less compelling, approach that we can take in going from K-L to AIC. We start with

\[ I(f(g(\cdot | \theta_0))) = \text{const} \, t - E_x \log(g(x | \theta_0)) \]

and see whether we can compute \( E_x(\log(g(x | \hat{\theta}(y))) \) based on Taylor series expansion. We can derive the result

\[ E_x(\log(g(x | \hat{\theta}(y)))) = E_x \log g(x | \hat{\theta}(x)) - \frac{1}{2} \text{tr}(J(\theta_0)I(\theta_0)^{-1}) - \frac{1}{2}(\hat{\theta}(y) - \theta_0)I(\theta_0)(\hat{\theta}(y) - \theta_0) \]

taking expectation on both sides we have

\[ E_x E_y \log g(x | \hat{\theta}(y)) = E_x \log g(x | \hat{\theta}(x)) - \text{tr}(J(\theta_0)I(\theta_0))^{-1} \]

Thus either of the two approaches demonstrates that we have to change our objective from model selection based on minimum K-L with known \( \theta_0 \) given \( g \), to selecting the model with estimated \( \theta \) based on minimizing an expected K-L information measure. It is still the case only a relative minimum can be found based on \( E_x E_y \log(g(x | \theta)) \) as the target objective function to be maximized. That is we want to estimate without bias the value of \( T = \int f(y) \int f(x) \log(g(x | \theta(y))) dx \, dy \) This is so because we have to estimate the parameters in model \( g \), which simplifies the expression above to

\[ T = \int \int \log(g(x | \hat{\theta})) \] where maximum likelihood estimator \( \hat{\theta} \) is based on sample \( y \) and the two expectations are for \( x \) and \( y \) (hence \( \hat{\theta} \)) both with
respect to the truth, \( f \). It is because \( T \) is also a double expectation based, conceptually, on two independent samples, that AIC-based model selection is asymptotically equivalent to cross validation; cross validation is a well-accepted basis of model selection.

Step 1 is an expression of the form 3.9.1 applied to

\[
\log(g(x|\hat{\theta})) \text{ around } \theta_0, \text{ for any given } x:
\]

\[
\log(g(x|\hat{\theta})) = \log(g(x|\theta_0)) + \left(\frac{\partial \log(g(x|\theta_0))}{\partial \theta}\right)(\hat{\theta} - \theta_0) + \frac{1}{2} (\hat{\theta} - \theta_0)\left(E_x \frac{\partial^2 \log(g(x|\theta_0))}{\partial \theta^2}\right)(\hat{\theta} - \theta_0)
\]

now taking expectation w.r.t. \( x \) we have

\[
E_x \left( \log(g(x|\hat{\theta})) \right) = E_x \left( \log(g(x|\theta_0)) \right) + \left(E_x \left( \frac{\partial \log(g(x|\theta_0))}{\partial \theta}\right) \right)(\hat{\theta} - \theta_0) + \frac{1}{2} (\hat{\theta} - \theta_0)\left(E_x \frac{\partial^2 \log(g(x|\theta_0))}{\partial \theta^2}\right)(\hat{\theta} - \theta_0)
\]

but

\[
E_x \left( \frac{\partial \log(g(x|\theta_0))}{\partial \theta} \right) = 0
\]

then

\[
E_x \log(g(x|\hat{\theta})) = E_x \log(g(x|\theta_0)) + \frac{1}{2} (\hat{\theta} - \theta_0)\left(E_x \frac{\partial^2 \log(g(x|\theta_0))}{\partial \theta^2}\right)(\hat{\theta} - \theta_0)
\]

\[
\Rightarrow E_x \left( \log(g(x|\hat{\theta})) \right) = E_x \log(g(x|\theta_0)) - \frac{1}{2} (\hat{\theta} - \theta_0) I(\theta_0)(\hat{\theta} - \theta_0)
\]

Taking expectation again with respect to \( \hat{\theta} \) we have

\[
E_{\theta} E_x \log(g(x|\hat{\theta})) = E_{\theta} E_x \log(g(x|\theta_0)) - \frac{1}{2} \text{tr} \left[I(\theta_0) \left(E_{\hat{\theta}} \left( \hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)' \right) \right] \right)
\]

\[
= E_x \log(g(x|\theta_0)) - \frac{1}{2} \text{tr} \left[I(\theta_0) E_{\hat{\theta}} \left( \hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)' \right) \right]
\]

but

\[
\]

56
\[ E_{\theta}E_{x}(\log g(x | \theta)) = T \quad \text{and} \quad E_{\theta} \left( (\hat{\theta} - \theta_o)(\hat{\theta} - \theta_o)' \right) = \sum \]

then \( T = E_{x}(\log g(x | \theta_o)) - \frac{1}{2} \text{tr}(I(\theta_o)\sum) \) \ldots 3.10.1.

Step 2 starts with the need of a relationship between \( T \) and \( E_{x}(\log g(x | \hat{\theta}(x))) \) which is expected of the actual log-likelihood at the MLE. We now do a second expansion this time of \( \log g(x | \theta_o) \) about \( \hat{\theta}(x) \) treating \( x \) as the sample data hence getting the maximum estimator of \( \theta \) for this \( x \).

Now applying Taylor’s series approximation (3.9.1) but with the roles of \( \hat{\theta} \) and \( \theta_o \) switched and \( \hat{\theta} = \hat{\theta}(x) \)

\[ \log g(x | \theta_o) = \log g(x | \theta) + \left( \frac{\partial \log g(x | \theta)}{\partial \theta^2} \right) (\theta_o - \hat{\theta}) + \frac{1}{2} (\theta_o - \hat{\theta})' \left( \frac{\partial^2 \log g(x | \theta)}{\partial \theta^2} \right) (\theta_o - \hat{\theta}) \]

The maximum likelihood estimator of \( \hat{\theta} \) is the solution of the equations

\[ \frac{\partial \log g(x | \hat{\theta})}{\partial \theta} = 0 \]

Therefore, the linear term in the above equation vanishes.

Taking the needed expectation we can write

\[ E_{x} \log g(x | \theta_o) = E_{x} \log g(x | \hat{\theta}) - \frac{1}{2} \text{tr}E_{x} \left( I(\hat{\theta})(\theta_o - \hat{\theta})(\theta_o - \hat{\theta})' \right) \]

but \( E_{x} \left( I(\hat{\theta})(\theta_o - \hat{\theta})(\theta_o - \hat{\theta})' \right) = I(\theta_o)E_{x}(\theta_o - \hat{\theta})(\theta_o - \hat{\theta})' = I(\theta_o)\sum \)

hence \( E_{x}(\log g(x | \theta_o)) = E_{x}(\log g(x | \hat{\theta}(x)) - \frac{1}{2} \text{tr}(I(\theta_o)\sum) \) \ldots 3.10.2
From 3.9.1, \( T = E_\xi \left[ \log \left( g(x \mid \hat{\theta}_0) \right) \right] - \frac{1}{2} tr I(\theta_0) \Sigma \)

And using the two equations 3.10.1 and 3.10.2,
\[
T = E_\xi \log g(x \mid \hat{\theta}(x)) - tr I(\theta_0) \Sigma
\]

But now using equation 3.9.5, the above equation is usually presented in literature as \( T \) whose unbiased estimator is
\[
\hat{T} = \log g(x \mid \hat{\theta}) - tr I(\theta_0) \Sigma
\]

The best model to use is the one with the largest value of \( \hat{T} \), because this will produce a model with the smallest estimated expected K-L distance. As a matter of fact the criterion is often stated as that of minimizing
\[
-2 \log g(x \mid \hat{\theta}) + 2 tr (J(\theta_0)(I(\theta_0))^{-1}) \text{........................................3.10.3.}
\]

If \( f \) is a subset of \( g \) that is if \( g = f \) then \( J(\theta_0) = I(\theta_0) = J(\theta_0) = I(\theta_0) = \sum^{-1} \) and hence \( tr (I(\theta_0) \Sigma) = k \text{........................................3.10.4.} \)

Which justifies AIC as a special case of equation 3.10.3.

\[
AIC = -2 \log g(x \mid \hat{\theta}) + 2k \text{........................................3.10.5.}
\]

Remarks

\( i \) It is not required that the truth \( f \) be in the set of models to which we apply AIC model selection. Many derivations are quite misleading by making assumptions that \( f = g, or, f \subset g \)
(ii) The trace term does not depend on the sample size. Rather for good models it is about equal to $K$.

(iii) Quantities such as the log likelihood expected log likelihood and both $K-L$, $I(f_{\theta_0})$ and the expected $K-L$ increase linearly with sample size $n$.

Now since both models are exponential let us consider in particular Akaike's information criterion derivation for the exponential family of distribution us for these two models. We do this in the next chapter.
CHAPTER FOUR

4. SELECTION OF THE BETTER MODEL.

4.1 Introduction.

A generalization of normality based models is found in the exponential family of distributions. The canonical representation of an exponential family probability distribution function involves sums of functions of the sample values. It is convenient to denote these sums by $S_j$. Then will have the canonical representation as

$$g(x | \theta) = \exp \left( \sum S_j \theta_j + H(\theta) + G(S) \right)$$

$$= \exp (S'\theta + H(\theta) + G(S)) \quad \cdots \cdots \cdots \cdots 4.1.1.$$ 

Each $S_j$ is a function of the full sample $x = \mu(\theta) + \varepsilon$ and any covariates, $z$ involved in representing $\mu(\theta)$ or which we condition and $S = (S_1, \ldots, S_k)$.

Our goal is to evaluate

$$T = E_x E_y \left( \log \left( g \left( x | \hat{\theta}_x \right) \right) \right) = E_x E_y \left( S' \hat{\theta}_y + H(\hat{\theta}_y) + G(S_x) \right)$$

Here $S_x$ and $\hat{\theta}_x$ are thought of as based on independent samples $x$ and $y$.

We also simplify the notation using $\hat{\theta}_x$ rather than $\hat{\theta}_y$ hence
\[ T = E_x E_y \left( (S_x - S_y + S_y) \right) \hat{\theta}_y + H(\hat{\theta}_y) + G(S_x) \]
\[ = E_x E_y \left( (S_x - S_y) \right) \hat{\theta}_y + S_y \hat{\theta}_y + H(\hat{\theta}_y) + G(S_x) \]
\[ = E_x E_y \left( S_y \hat{\theta}_y + H(\hat{\theta}_y) + G(S_x) + (S_x - S_y) \right) \hat{\theta}_y \]
\[ = E_y \left( S_y \hat{\theta}_y + H(\hat{\theta}_y) + E_x \left( G(S_x) \right) + E_y \left( (E_x S_x - S_y) \right) \hat{\theta}_y \right) \]

The interchangeability of integration arguments now is used since both expectations are with respect to \( f \)

hence

\[ E_x \left( G(S_x) \right) = E_y \left( G(S_y) \right) \]

and for simplicity

\[ E_x (S_x) = E(S) \]

thus

\[ T = E_y \left( S_y \hat{\theta}_y + H(\hat{\theta}_y) + G(S_x) + E_y \left( (E(S) - S_y) \right) \hat{\theta}_y \right) \]
\[ = E_x \log \left( g(x | \hat{\theta}) \right) + E_y \left( (E(S) - S_y) \right) \hat{\theta}_y \]
\[ = E_x \log \left( g(x | \hat{\theta}) \right) - E_y \left( (S_y - E(S)) \right) \hat{\theta}_y \] \[ \text{......4.1.2.} \]

Equation 4.1.2. is an exact result, which shows the bias to be subtracted from \( E_x \left( \log \left( g(x | \hat{\theta}) \right) \right) \) to get \( T \):
bias = \( E_x \left[ (s - E(s)) \left( \hat{\theta} - \hat{\theta}_* \right) \right] \)

where
\[
\hat{\theta}_* = E(\hat{\theta}_y)
\]

to denote the exact expectation of the maximum likelihood estimator for the given sample size \( n \) and the model
\[ g(\theta_0) = \theta_x \] with asymptotic equality.

Now that only one "sample" is involved then

\[
\text{bias} = E \left[ (s - E(s)) \left( \hat{\theta} - \hat{\theta}_* \right) \right] \quad \text{..........................4.1.3}
\]

\[
= \text{tr} E \left[ (\hat{\theta} - \hat{\theta}_*) (s - E(s))' \right] = \text{tr} \left( \text{cov}(\hat{\theta}, s) \right)
\]

Hence for the exponential family an exact result is

\[
T = E_x \left[ \log \left( g \left( x \mid \hat{\theta} \right) \right) - \text{tr} \left( \text{cov}(\hat{\theta}, s) \right) \right] \quad \text{.................4.1.4}
\]

The \( k \times k \) matrix of covariate elements \( \text{cov}(\hat{\theta}, s) \) can be approximated by Taylor series method. Equation 4.1.4 above may not seem very useful because it seems to only apply only to the canonical form of the exponential family. This is not true hence further discussion but before that let us look at the bias term we consider the maximum likelihood estimators and the Hessian.

First
\[
\log(g(x \mid \theta)) = S'\theta + H(\theta) + G(S)
\]

\[
\Rightarrow \frac{\partial \log(g(x \mid \theta))}{\partial \theta} = S + \frac{\partial H(\theta)}{\partial \theta}
\]

\[
\Rightarrow \frac{\partial^2 \log(g(x \mid \theta))}{\partial \theta^2} = \frac{\partial^2 H(\theta)}{\partial \theta^2}
\]

and thus

\[
I(\theta_o) = E_f\left[-\frac{\partial^2 \log(g(x \mid \theta))}{\partial \theta^2}\right] = \frac{\partial H(\theta_o)}{\partial \theta^2}
\]

It follows that the maximum likelihood estimator satisfies

\[
S = -\frac{\partial H(\hat{\theta})}{\partial \theta}
\]

and \( \theta_o \) satisfies

\[
E_f(S) = \frac{\partial H(\theta_o)}{\partial \theta}
\]

Which is an exact result whereas

\[
E(\hat{\theta}) = \theta_o \quad O\left(\frac{1}{\sqrt{n}}\right)
\]

and the formula 3.9.6 becomes

\[
J(\theta) = E_f\{(S - E_f(S))(S - E_f(S))'\}
\]
which is the true variance – covariance matrix of $S$. The $n$ identically
independently distributed observations produce a set statistics $~S_i\ldots$ that sum to $S$ hence an estimator of $J(\theta_0)$ is

$$
\hat{J}(\theta_0) = \frac{n}{n-1}\left(\sum_{i=1}^{n}(S_i - \bar{S})(S_i - \bar{S})'\right)
$$

Now if we go back to the bias term, a first order Taylor series expansion
gives us

$$
\frac{\partial H(\hat{\theta}_0)}{\partial \theta} = -\frac{\partial H(\theta_0)}{\partial \theta} - \frac{\partial^2 H(\theta_0)}{\partial \theta^2}(\hat{\theta} - \theta_0)
$$

hence

$$
S = E(S) + I(\theta_0)(\hat{\theta} - \theta_0) \ldots \ldots \ldots \ldots \ldots 4.1.5.
$$

Now inserting 4.1.5 in 4.1.3, we have

$$
Bias = E\{(E(S) + I(\theta_0)(\hat{\theta} - \theta_0) - E(S))(\hat{\theta} - \theta.\} \ldots
$$

$$
= E\{(\hat{\theta} - \theta_0)'I(\theta_0)(\hat{\theta} - \theta_0)\} \ldots
$$

$$
= trE[I(\theta_0)(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)'] \ldots
$$

$$
= trI(\theta_0)E[(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)'] \ldots
$$

$$
= tr[I(\theta_0)\Sigma] \ldots
$$
Thus we have shown that in this common case of exponential family,

\[ T = E_x \left( \log \left( g \left( x \mid \hat{\Theta} \right) \right) \right) - trI(\hat{\Theta}) \sum \]

and the approximation is to \( O(1/n) \).

These results can be extended to any parameterized form of an exponential family model.

### 4.2 Application of AIC on Weibull and Logistic model.

Now since the Weibull and Logistic models are exponentials, we shall therefore use the derivation of AIC on exponential distributions. From equation 4.1.5, we have

\[ T = E_x \left( \log \left( g \left( x \mid \hat{\Theta} \right) \right) \right) - trI(\hat{\Theta}) \sum \]

but from equation 3.10.4, we have seen that

\[ trI(\hat{\Theta}) \sum = k \] then

\[ T = E_x \left( \log \left( g \left( x \mid \hat{\Theta} \right) \right) \right) - k \]

whose unbiased estimator is given by

\[ \hat{T} = \log \left( g \left( x \mid \hat{\Theta} \right) \right) - k \]

As a matter of convention the criterion is often stated as that of minimizing
which is equivalent to the expression 3.10.5.

Thus for the two models the likelihood is

\[ L(x \mid \theta) = \prod_{i=1}^{k} \binom{n_i}{r_i} P^{r_i} (1 - P)^{n_i - r_i} \]

\[ = \left( \frac{n_1}{r_1} \right) P^{r_1} (1 - P)^{n_1 - r_1} \times \left( \frac{n_2}{r_2} \right) P^{r_2} (1 - P)^{n_2 - r_2} \times \cdots \times \left( \frac{n_k}{r_k} \right) P^{r_k} (1 - P)^{n_k - r_k} \]

\[ = \prod_{i=1}^{k} \left( \frac{n_i}{r_i} \right) P^{r_i} (1 - P)^{n_i - r_i} \]

and the log-likelihood is given by

\[ \ln L(x \mid P) = \ln \prod_{i=1}^{k} \binom{n_i}{r_i} + \sum_{i=1}^{k} r_i \ln P + \sum_{i=1}^{k} (n_i - r_i) \ln (1 - P) \]

\[ \Rightarrow L(x \mid P) = g(x \mid P) = \exp \left( \ln \prod_{i=1}^{k} \binom{n_i}{r_i} + \sum_{i=1}^{k} r_i \ln P + \sum_{i=1}^{k} (n_i - r_i) \ln (1 - P) \right) \]

Now

\[ AIC = -2 \log(g(x \mid \hat{P})) + 2 \text{tr}(\hat{P}) \sum \text{ ................................. 4.2.1.} \]

\[ \hat{I}(P) = -\frac{\partial^2 \log(g(x \mid P_0))}{\partial P^2} \]
\[ \frac{\partial^2}{\partial P^2} \left( \ln \prod_{i=1}^{k} \left( \frac{n_i}{r_i} \right) + \sum_{i=1}^{k} r_i \ln P + \sum_{i=1}^{k} (n_i - r_i) \ln (1 - P) \right) \]

\[ = \frac{\partial}{\partial P} \left( \sum_{i=1}^{k} \frac{r_i}{P} - \sum_{i=1}^{k} \frac{n_i - r_i}{1 - P} \right) \]

\[ = -\sum_{i=1}^{k} \frac{r_i}{P^2} - \sum_{i=1}^{k} \frac{n_i - r_i}{(1 - P)^2} \]

\[ = \sum_{k} \frac{2r_i P - r_i - n_i P^2}{P^2(1 - P)^2} \]

and

\[ \hat{j}(P) = \left( \frac{\partial \log(g(x \mid P))}{\partial P} \right) \left( \frac{\partial \log(g(x \mid P))}{\partial P} \right)' \]

\[ \frac{\partial \log(g(x \mid P))}{\partial P} = \frac{\partial}{\partial P} \left( \ln \prod_{i=1}^{k} \left( \frac{n_i}{r_i} \right) + \sum_{i=1}^{k} r_i \ln P + \sum_{i=1}^{k} (n_i - r_i) \ln (1 - P) \right) \]

\[ = \sum_{i=1}^{k} \left( \frac{r_i - n_i}{P} \right) \frac{1}{1 - P} \]

\[ = \sum_{i=1}^{k} \frac{r_i - n_i P}{P(1 - P)} \]

Therefore

\[ \hat{j}(P) = \left( \sum_{i=1}^{k} \frac{r_i - n_i P}{P(1 - P)} \right) \left( \sum_{i=1}^{k} \frac{r_i - n_i P}{P(1 - P)} \right)' \]

\[ = \left( \sum_{i=1}^{k} \frac{r_i - n_i P}{P(1 - P)} \right)^2 \]

\[ \text{tr} \hat{I}(P) \Sigma = \text{tr} \hat{J}(P) \hat{I}(P)^{-1} \]

67
We know that

\[
\frac{\delta \log(g(x | P))}{\delta P} = \sum_{i=1}^{k} \frac{r_i - r_i}{P} - \sum_{i=1}^{k} \frac{n_i - r_i}{1 - P}
\]

\[
\Rightarrow \sum_{i=1}^{k} \frac{r_i}{P} = \sum_{i=1}^{k} \frac{n_i - r_i}{1 - P}
\]

\[
\Rightarrow \frac{r_i}{P} = \frac{n_i - r_i}{1 - P}
\]

\[
\Rightarrow r_i (1 - P) = P (n_i - r_i)
\]

\[
\Rightarrow r_i = n_i P
\]

\[
\Rightarrow \hat{p} = \frac{r_i}{n_i}
\]

Substituting in equation 4.2.2 above we have

\[
\text{tr} \left[ j(\hat{p} | \hat{p})^{-1} \right] = 0
\]

therefore 4.2.1, simplifies to

\[
AIC = -2 \log(g(x | P))
\]

\[
= -2 \left( \ln \prod_{i=1}^{k} \left( \frac{n_i}{r_i} \right) + \sum_{i=1}^{k} r \ln P + \sum_{i=1}^{k} (n_i - r_i) \ln(1 - P) \right)
\]

So if we get, for the two models the value of AIC then we can determine the better model by using empirical data.
[a] **Logistic model**

For the logistic model we have the value of $P$ as

$$P = \frac{1}{1 + e^{(\alpha + \beta x)}}$$

substituting this value in AIC we get

$$AIC_L = -2 \left[ \ln \prod_{i=1}^{k} \left( \frac{n_i}{r_i} \right) + \sum_{i=1}^{k} r_i \ln \frac{1}{1 + e^{-(\alpha + \beta x_i)}} \sum_{i=1}^{k} (n_i - r_i) \ln \left( 1 - \frac{1}{1 + e^{-(\alpha + \beta x_i)}} \right) \right]$$

let $\Omega = \ln \prod_{i=1}^{k} \left( \frac{n_i}{r_i} \right)$ then

$$AIC_L = -2 \left[ \Omega - \sum_{i=1}^{k} r_i \ln \left( 1 + e^{-(\alpha + \beta x_i)} \right) - \sum_{i=1}^{k} (n_i - r_i) \ln \left( 1 + e^{-(\alpha + \beta x_i)} \right) - \sum_{i=1}^{k} (n_i - r_i)(\alpha + \beta x_i) \right]$$

$$= -2\Omega + 2\sum_{i=1}^{k} n_i \ln \left( 1 + e^{-(\alpha + \beta x_i)} \right) + 2\sum_{i=1}^{k} (n_i - r_i)(\alpha + \beta x_i)$$

[b] **Weibull model**

For the weibull model we have $P$ as

$$P = 1 - e^{\beta x}$$

and the expression for the AIC is

$$AIC_W = -2 \left[ \Omega + \sum_{i=1}^{k} r_i \ln \left( 1 - e^{-(\beta x_i)} \right) + \sum_{i=1}^{k} (n_i - r_i) \ln \left( 1 - \left( 1 - e^{-(\beta x_i)} \right) \right) \right]$$

where $h(x)$ is the hazard function.
The AICs above are functions of $\alpha$ and $\beta$ which are unknown. Thus we need to estimate these parameters.

4.3 Estimation of parameters.

[a] The weibull model.

The weibull distribution arises in the theory of extreme values and in life testing. The survivor function is given as

$$F(d) = \exp(-\beta d^\alpha)$$

From the parametric modeling of the survivor data we have the full likelihood from $n$ independent units as

$$L = \prod_{i=1}^{n} f(d_i; \phi) \prod_{i=1}^{n} F(d_i; \phi)$$

$$\log L = l = \sum_{i=1}^{n} \log(f(d_i; \phi)) + \sum_{i=1}^{n} \log(F(d_i; \phi)) \quad \cdots \quad 4.3.1.$$

We know that the hazard function and the survivor function are related as

$$f(d) = h(d)F(d),$$

where $h(d)$ is the hazard function.
Thus equation 4.3.1 becomes

\[ I = \sum_{n_i - r_i} \log(h(d; \phi)) + \sum_{n_i - r_i} \log(F(d; \phi)) + \sum_{n_i} \log(F(d; \phi)) \]

\[ = \sum_{n_i - r_i} \log(h(d; \phi)) + \sum_{n_i} \log(F(d; \phi)) \] 4.3.2.

we know that

\[ F(d; \phi) = \exp(-\beta d^\alpha) = \exp(-H(d; \phi)) \]

\[ \Rightarrow h(d; \phi) = \frac{\partial H(d; \phi)}{\partial d} = \alpha d_i^{\alpha - 1} \]

hence equation 4.3.2 simplifies to

\[ I = \sum_{n_i - r_i} \log(\alpha d_i^{\alpha - 1}) + \sum_{n_i} \log(\exp(-\beta d_i^\alpha)) \]

\[ = (n_i - r_i) \log \alpha + (n_i - r_i) \log \beta + (\alpha - 1) \sum_{n_i - r_i} \log(d_i) - \sum_{n_i} \beta d_i^\alpha \]

Therefore

\[ \frac{\partial I}{\partial \alpha} = \frac{n_i - r_i}{\alpha} + \sum_{n_i - r_i} \log(d_i) - \beta \sum_{n_i} d_i^\alpha \log d_i \] 4.3.3.

\[ \frac{\partial I}{\partial \beta} = \frac{n_i - r_i}{\beta} - \sum_{n_i} d_i^\alpha \] 4.3.4.

If \( \alpha \) is specified the maximum likelihood estimator of \( \beta \) can be found explicitly by solving equation 4.3.4.
\[
\frac{n_i - r_i}{\beta} = \sum_{n_i} d_i^\alpha
\]

\[\Rightarrow \hat{\beta} = \frac{n_i - r_i}{\sum d_i^\alpha} \]

substituting in equation 4.3.3. the value of \( \hat{\beta} \) we get

\[
\frac{\partial l}{\partial \alpha} = \frac{n_i - r_i}{\alpha} + \sum \log d_i - \frac{(n_i - r_i) \sum d_i^\alpha \log d_i}{\sum n_i d_i^\alpha} = 0
\]

Solving this yields the maximum likelihood estimator of \( \alpha \). An iterative process can solve this equation by Newton-Raphson formulae given as;

\[
\alpha_1 = \alpha_0 - \frac{f(\alpha_0)}{f'(\alpha_0)}
\]

\[
\alpha_2 = \alpha_1 - \frac{f(\alpha_1)}{f'(\alpha_1)}
\]

where \( \alpha_0 \) is specified.

[b]. The logistic model.

The logistic model has survivor function as

\[
F(d) = \left(1 + e^{-(\alpha + \beta d)}\right)^{-1}
\]

and the hazard function

\[
h(d) = \frac{f(d)}{F(d)}
\]

\[= \frac{\beta e^{-(\alpha + \beta d)}}{1 + e^{-(\alpha + \beta d)}}\]
Now using the parametric modeling of the survival data we have the full likelihood as

\[ L = \prod_{n_{j-1}} f(d; \phi) \prod_{n_j} F(d; \phi) \]

whose log is given by

\[ l = \sum_{n_{j-1}} \ln(f(d; \phi)) + \sum_{n_j} \ln(F(d; \phi)) \]

but

\[ f(d; \phi) = h(d; \phi) F(d; \phi) \]

then the log-likelihood becomes

\[ l = \sum_{n_{j-1}} \ln(h(d; \phi)) + \sum_{n_j} \ln(F(d; \phi)) \]

now substituting the values of \( h(d; \phi) \) and \( F(d; \phi) \) we have

\[ l = \sum_{n_{j-1}} \ln \left( \frac{\beta e^{-(\alpha + \beta d_i)}}{1 + e^{-(\alpha + \beta d_i)}} \right) + \sum_{n_j} \ln \left( 1 + e^{-(\alpha + \beta d_i)} \right)^{-1} \]

this simplifies to

\[ l = \sum_{n_{j-1}} \ln \beta - \sum_{n_{j-1}} (\alpha + \beta d_i) - \sum_{2n_i - n_j} \ln \left( 1 + e^{-(\alpha + \beta d_i)} \right) \]

differentiating partially with respect to \( \alpha \) and \( \beta \) we have

\[ \frac{\partial l}{\partial \alpha} = -(n_{j-1}) + \sum \frac{e^{-(\alpha + \beta d_i)}}{1 + e^{-(\alpha + \beta d_i)}} \]

\[ \frac{\partial l}{\partial \beta} = \sum \frac{\beta e^{-(\alpha + \beta d_i)}}{1 + e^{-(\alpha + \beta d_i)}} \]

\[ \text{4.3.5.} \]

and
\[ \frac{\partial l}{\partial \beta} = \frac{n_i - r_i}{\beta} - \sum_{n-\eta} d_i + \sum_{\eta \leq n} d_i e^{-(\alpha + \beta d_i)} \] .................4.3.6.

Now by using Newton Raphson method of iteration we can estimate the values for \(\alpha\) and \(\beta\). The Newton-Raphson method is given as

\[ \beta_1 = \beta_0 - \frac{f(\beta_0)}{f'(\beta_0)} \]

\[ \beta_2 = \beta_1 - \frac{f(\beta_1)}{f'(\beta_1)} \]

where \(\beta_0\) is specified. This is repeated until a suitable estimate value of \(\beta\) is obtained.
4.4 Results.

The data (Table 1) used in this analysis was obtained from the paper by Wijesinha and Piantadosi (1995), where $i$ denotes the number of different test, $d$, the dose, $n$, the subjects in a group and $r$, the number of the subjects who respond.

Table 1

<table>
<thead>
<tr>
<th>$d_i$</th>
<th>$r_i$</th>
<th>$n_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>23</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>25</td>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td>28</td>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td>30</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>33</td>
<td>6</td>
<td>22</td>
</tr>
<tr>
<td>35</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>40</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>45</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>50</td>
<td>12</td>
<td>19</td>
</tr>
<tr>
<td>55</td>
<td>12</td>
<td>20</td>
</tr>
<tr>
<td>58</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>60</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>63</td>
<td>16</td>
<td>20</td>
</tr>
<tr>
<td>65</td>
<td>15</td>
<td>19</td>
</tr>
<tr>
<td>d</td>
<td>r</td>
<td>n</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>23</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>25</td>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td>28</td>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td>30</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>33</td>
<td>6</td>
<td>22</td>
</tr>
<tr>
<td>35</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>40</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>45</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>50</td>
<td>12</td>
<td>19</td>
</tr>
<tr>
<td>55</td>
<td>12</td>
<td>20</td>
</tr>
<tr>
<td>58</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>60</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>63</td>
<td>16</td>
<td>20</td>
</tr>
<tr>
<td>65</td>
<td>15</td>
<td>19</td>
</tr>
</tbody>
</table>

$AIC_i = 621.9445$ is the value of AIC for logistic model.

Table 2 shows the estimates of $\alpha$ and $\beta$ for every dose since the doses are independent using the logistic model. where $\alpha_i$ is the value when $y_i = 0.00001$. This is obtained after the seventh iteration.
While $\beta$, is the corresponding value.

Table 3

<table>
<thead>
<tr>
<th>$d$</th>
<th>$r$</th>
<th>$n$</th>
<th>$\alpha$</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>20</td>
<td>512000</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>20</td>
<td>486400</td>
<td>.4999999</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>20</td>
<td>460800</td>
<td>.125</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>20</td>
<td>460800</td>
<td>9.999998E-02</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>20</td>
<td>460800</td>
<td>7.692306E-02</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>20</td>
<td>460800</td>
<td>6.666665E-02</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>20</td>
<td>460800</td>
<td>5.555553E-02</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
<td>20</td>
<td>435200</td>
<td>5.263156E-02</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
<td>20</td>
<td>435200</td>
<td>4.347824E-02</td>
</tr>
<tr>
<td>23</td>
<td>3</td>
<td>20</td>
<td>435200</td>
<td>3.999998E-02</td>
</tr>
<tr>
<td>25</td>
<td>4</td>
<td>20</td>
<td>409600</td>
<td>3.571426E-02</td>
</tr>
<tr>
<td>28</td>
<td>4</td>
<td>20</td>
<td>409600</td>
<td>3.333331E-02</td>
</tr>
<tr>
<td>30</td>
<td>5</td>
<td>20</td>
<td>384000</td>
<td>.030303</td>
</tr>
<tr>
<td>33</td>
<td>6</td>
<td>22</td>
<td>358400</td>
<td>.0285714</td>
</tr>
<tr>
<td>35</td>
<td>6</td>
<td>20</td>
<td>358400</td>
<td>.025</td>
</tr>
<tr>
<td>40</td>
<td>0</td>
<td>20</td>
<td>512000</td>
<td>2.222216E-02</td>
</tr>
<tr>
<td>45</td>
<td>10</td>
<td>20</td>
<td>256000</td>
<td>1.999989E-02</td>
</tr>
<tr>
<td>50</td>
<td>12</td>
<td>19</td>
<td>179200</td>
<td>1.818174E-02</td>
</tr>
<tr>
<td>55</td>
<td>12</td>
<td>20</td>
<td>204800</td>
<td>1.724119E-02</td>
</tr>
<tr>
<td>58</td>
<td>15</td>
<td>20</td>
<td>128000</td>
<td>1.666649E-02</td>
</tr>
<tr>
<td>60</td>
<td>15</td>
<td>20</td>
<td>128000</td>
<td>1.587277E-02</td>
</tr>
<tr>
<td>63</td>
<td>16</td>
<td>20</td>
<td>102400</td>
<td>1.538438E-02</td>
</tr>
</tbody>
</table>

$AIC_w = 831.9583$ is the AIC for the Weibull model.

Table 3 shows the estimates of $\hat{\alpha}$ and $\hat{\beta}$ using the Weibull model.
The above results gives the value of $\hat{\alpha}$ and $\hat{\beta}$ after iterations. The optimization stops after the seventh iteration (that is) when the function $f(\alpha_i)$ is approximately 0.00001 in case of the Weibull model and for logistic model when the function $f(\beta_j)$ is approximately 0.00001. The program used in the computation of these results is the Q-Basic, table 2 gives the results for the logistic model and the calculated AIC is 621.9445. Whilst table 3 shows the results under the Weibull model giving us the value for AIC as 831.9583. This implies that as having studied the Weibull model, the logistic model is still better than the Weibull model. Note that the value $-2\Omega$ was ignored during computation since both expressions contain it; and as we saw in section 3.6, the actual value of AIC does not matter but the difference between the models is enough to draw a conclusion. This means that, between the logistic and the Weibull model, the logistic model is a model of choice.
CHAPTER FIVE

5.1 Conclusion.

In chapter one we looked at the logistic dose-response model. In this chapter we were able to study the logistic model, when the logit method is applied. We also saw that we can estimate parameters by use of maximum likelihood estimation. The chapter ends with a problem from Wijesinha and Piantadosi (1995) that a few models could behave like logistic model. We choose Weibull model, from the stated models for further analysis.

Chapter 2 opened with a general knowledge of the Weibull model. We then use the same method used to study the logistic model in chapter 1, to study the Weibull model. It is observed that the results obtained have the same structures. This prompted us to wonder and ask, which is a better model. We make use of Akaike Information Criterion (AIC) to answer this question.

Chapter 3 we reviewed the Akaike Information Criterion. This is a field, which also has not been studied thoroughly. We were able to show how AIC was developed, its derivation and how it is used to select the best model.

Finally chapter four opened with the derivation of AIC for the exponential distribution which was necessary in making our decision. Later in the chapter we were able to derive the AIC for the two models and estimates of
parameters. We ended the chapter with empirical results where the logistic model emerged as a better model.

In short we managed to

(i) Study the Weibull model as a dose response model.

(ii) Study the AIC thoroughly, and were able to apply it in this project, to choose which model was better. Using the empirical results, the logistic model, was found to have a smaller AIC and therefore a better model.

5.2 Areas for further research.

The following are some of the areas that the project can be extended.

(i) In our case the analysis was done using a univariate case only that is the dose. However factors like sex, age, etc. could affect the results hence a study on the project which incorporates such covariate factors should be carried out.

(ii) We have also assumed that the response is independent of time but it is possible that the time lapse could also affect the response. Thus analysis with time need to be looked into.
REFERENCES.


