STATISTICAL INFERENCE IN LOGISTIC REGRESSION MODELS

WU SWEE LENG

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

INSTITUTE OF MATHEMATICAL SCIENCES
FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2011
UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: WU SWEE LENG  (I/C No: 610101-08-5460)
Registration/Matric No: SHAQ00021
Name of Degree: Doctor of Philosophy
STATISTICAL INFERENCE IN LOGISTIC REGRESSION MODELS
Field of Study: GENERALIZED LINEAR MODELS

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this Work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya (“UM”), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Wu Swee Leng
(signed)
Date: 31 May 2011
(Candidate’s signature)

Subscribed and solemnly declared before

Pooi Ah Hin
(signed)
Date: 31 May 2011
(Witness’s signature)
Name: Pooi Ah Hin
Designation: Professor
ABSTRAK

Model regresi logistik merupakan sejenis model linear teritlak yang telah pun digunakan secara luas untuk memodelkan data binari. Tesis ini cuba membekalkan analisa yang lebih halus untuk model regresi logistik mudah.

Taburan bagi statistik ujian untuk menguji kekurangan penyuaian model dan keertian parameter lereng \( \beta_1 \) diperoleh dengan lebih tepat supaya dapat memaparkan ciri-ciri yang bertertib lebih tinggi seperti kepencongan dan kurtosis taburan.

Selang keyakinan berasaskan ujian hipotesis diterbitkan untuk parameter lereng. Dalam proses mencari selang keyakinannya, taburan bagi anggaran kebolehjadian maksimum \( \hat{\beta}_1 \) digunakan apabila vektor parameter \( (\beta_0, \beta_1) \) dalam model memenuhi syarat-syarat di bawah hipotesis nol dan adalah paling dekat kepada anggaran kebolehjadian maksimum \( (\hat{\beta}_0, \hat{\beta}_1) \). Kajian simulasi yang telah dijalankan menunjukkan bahawa selang keyakinan bagi \( \beta_1 \) yang berasaskan kaedah ujian hipotesis memberi prestasi yang lebih memuaskan, baik dari segi kebarangkalian laporan mahupun ukur panjang selang yang dijangkakan, apabila dibandingkan dengan selang keyakinan yang diperoleh daripada varians asimptot.

Suatu selang keyakinan berasaskan ujian hipotesis kemudian diterbitkan untuk kebarangkalian kejayaan apabila vektor pembolehubah penjelasan mengambil nilai \( x^* \) pada masa yang akan datang. Akhirnya selang keyakinan bagi kebarangkalian kejayaan ini diperbesarkan untuk membentuk suatu selang ramalan bagi kadar kejayaan apabila sebilangan percubaan dilakukan di bawah syarat-syarat yang ditetapkan oleh \( x^* \). Kajian simulasi yang dikendalikan menunjukkan selang ramalan yang dihasilkan mempunyai purata kebarangkalian laporan yang agak hampir dengan nilai sasaran.
ABSTRACT

Logistic regression model is a type of generalized linear model which has been used widely for modelling binary data. The present thesis aims to provide more refined analyses of the simple logistic regression model.

The distributions of the test statistics for testing the lack of fit of the model and the significance of the slope parameter $\beta_1$ are found more accurately so as to capture the higher order characteristics such as skewness and kurtosis of the distribution.

A confidence interval based on hypothesis testing is derived for the slope parameter. In the process of finding the confidence interval, we make use of the distribution of the maximum likelihood estimate $\hat{\beta}_1$ when the parameter vector $(\beta_0, \beta_1)$ in the model satisfies the conditions under the null hypothesis and is the nearest to the maximum likelihood estimate $(\hat{\beta}_0, \hat{\beta}_1)$. The simulation studies which have been performed show that the confidence interval for $\beta_1$ based on the method of hypothesis testing performs better both in terms of coverage probability and expected length, as compared to the confidence interval derived from the asymptotic variance.

A confidence interval based on hypothesis testing is next derived for the probability of success when the vector of explanatory variables assumes the value $x^*$ in the future. Finally the confidence interval for the probability of success is enlarged to form a prediction interval for the future observed proportion of successes when a number of trials are performed under the conditions specified by $x^*$. The simulation studies which have been performed show that the resulting prediction interval has an average coverage probability which is fairly close to the target value.
ACKNOWLEDGEMENTS

I wish to convey my deepest gratitude to my supervisor, Professor Dr Pooi Ah Hin for his constant guidance, invaluable advice and effective motivation given throughout the course of my study. He has been extremely patient with me in handling the much technical work involved in the project. I am indeed indebted to him for without his encouragement and supervision, I may not have made this writing of thesis materialized.

I would also like to express my heartfelt thanks to all my fellow colleagues and friends in the Institute of Mathematical Sciences of University of Malaya for their support, friendship and assistance rendered. Special thanks must go to Dr Ng Kok Haur for providing ideas in computational work and offering an experienced ear to my doubts and frustration encountered during the development stage of my project. I am also grateful to Ms Ng Lee Leng for her excellent typing of the thesis.

I shall not forget to show my high appreciation to all my friends who in one way or another have given me the motivation and inspiration to strive on to accomplish my mission. Sincere thanks are due to Nicole Lee for her warm friendship and kindness.

Lastly, I shall thank all the members of my family, especially my thoughtful husband Koong W.K., two loving children ZheXian and YuQian; and my brothers and sisters, especially SweeTing, for their immense patience and understanding. To them I dedicate this thesis.
TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRAK</td>
<td>ii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>iii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>iv</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>v</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>ix</td>
</tr>
</tbody>
</table>

CHAPTER 1       INTRODUCTION

1.1 Generalized Linear Models  1
1.2 Logistic Regression Models  2
1.3 Extension Of The Logistic Regression Models  5
1.4 Outline Of The Thesis  5

CHAPTER 2       STATISTICAL INFERENCE IN LOGISTIC REGRESSION MODELS

2.1 Logistic Regression Models  7
2.2 Maximum Likelihood Estimate $\hat{\beta}$ In Simple Logistic Regression Models  8
2.3 Approximating $\hat{\beta}$ Using Taylor’s Series Expansion  9
2.4 Assessing The Adequacy Of Logistic Regression Models Using Deviance  13
2.5 Distributions Of Deviance And Wald Statistics Under Null Hypothesis 18
2.6 Modified Chi-square Distribution 19
2.7 Numerical Examples 20

CHAPTER 3 CONFIDENCE INTERVALS FOR THE SLOPE PARAMETER
3.1 Introduction 30
3.2 Method Based On Hypothesis Testing 31
3.3 Performance Study On Confidence Intervals For Slope Parameter Based On The Method Of Hypothesis Testing 32

CHAPTER 4 PREDICTION INTERVALS IN SIMPLE LOGISTIC REGRESSION MODEL
4.1 Introduction 38
4.2 Confidence Interval For The Probability Of Success 38
4.3 Prediction Interval For The Observed Proportion Of Success In The Future 41
4.4 Simulation Study On The Expected Coverage Probability Of The Prediction Interval 43

CHAPTER 5 CONCLUDING REMARKS 45

REFERENCES 47
LIST OF FIGURES

Figure 2.1 Cumulative probability distributions of deviance based on numerical integration and simulation procedure, for \( n=15, \beta_0^{(0)} = -0.8, \beta_1^{(0)} = -0.01 \). The fitted modified chi-square distribution for the deviance has parameters \( \hat{\mu} = -3.84856051, \hat{\sigma} = 0.951463042 \), and \( \hat{\lambda} = (0.156932, 0.782083, -0.025838)^T \). 23

Figure 2.2 Cumulative probability distributions of deviance based on numerical integration and simulation procedure, for \( n=25, \beta_0^{(0)} = -0.1, \beta_1^{(0)} = 0.065 \). The fitted modified chi-square distribution for deviance has parameters \( \hat{\mu} = -5.3212953, \hat{\sigma} = 0.480868054 \), and \( \hat{\lambda} = (0.0578233, 0.465036, -1.33361)^T \). 24

Figure 2.3 Cumulative probability distributions of \( \lambda(\beta | \beta_0) \) based on numerical integration and simulation procedure, for \( n=15, \beta_0^{(0)} = 0.05 \). The fitted modified chi-square distribution for \( \lambda(\beta | \beta_0) \) has parameters \( \hat{\mu} = 0.808713597, \hat{\sigma} = 0.691722826 \), and \( \hat{\lambda} = (0.7583, 0.147704, -0.999968)^T \). 25

Figure 2.4 Cumulative probability distributions of \( \lambda(\beta | \beta_0) \) based on numerical integration and simulation procedure, for \( n=20, \beta_0^{(0)} = -0.10 \). The fitted modified chi-square distribution for \( \lambda(\beta | \beta_0) \) has parameters \( \hat{\mu} = 0.787405746, \hat{\sigma} = 0.69174975, \) and \( \hat{\lambda} = (0.7583, 0.147704, -0.999968)^T \). 26

Figure 2.5 Cumulative probability distributions of Wald statistic \( W \) based on numerical integration and simulation procedure, for \( n=15, \beta_0^{(0)} = -0.2, \beta_1^{(0)} = 0.1 \). The fitted modified chi-square
distribution for $W$ has parameters $\hat{\mu} = -0.874006573$, $\hat{\sigma} = 0.638669563$, and $\hat{\lambda} = (1.12019, 0.00548593, 30.3013)^T$.  

**Figure 2.6** Cumulative probability distributions of Wald statistic $W$ based on numerical integration and simulation procedure, for $n=20$, $\beta^{(0)} = -0.8$, $\beta^{(0)} = 0.1$. The fitted modified chi-square distribution for $W$ has parameters $\hat{\mu} = 1.121162623$, $\hat{\sigma} = 0.81927581$, and $\hat{\lambda} = (1.20413, -0.161111, -0.614506)^T$.  

**Figure 3.1** Graphical presentation of $N=100$ simulated confidence intervals based on HT and WA, ranked according to the value of $\hat{\beta}_1$, when $n=15$, $\beta_0 = -0.2$, $\beta_1 = 0.1$, $x_i = i$, for $i=1,2,\ldots, n$, and $\alpha=0.05$.  

**Figure 3.2** Graphical presentation of $N=100$ simulated confidence intervals based on HT and WA, ranked according to the value of $\hat{\beta}_1$, when $n=25$, $\beta_0 = -1.0$, $\beta_1 = 0.05$, $x_i = i$, for $i=1,2,\ldots, n$; and $\alpha=0.05$.  

34  

35
Table 2.1  First four moments of $v_i$ and $v_i^*$ and their approximated QN-distributions. [$n=15$, $\beta^{(0)} = (-0.8, -0.01)^T$, $x_i = i$, for $i = 1, 2, ..., n.$]

Table 2.2  First four moments of deviance based on numerical integration and the fitted modified chi-square distribution. [$n=15$, $\beta^{(0)} = (-0.8, -0.01)^T$, $x_i = i$ for $i = 1, 2, ..., n.$]

Table 3.1  Estimated coverage probability and expected length of $N = 1000$ simulated confidence intervals for $\beta_1$ based on Wald Approach(WA), Bootstrap(BS) method and Hypothesis Testing(HT) method ($x_i = i$, for $i = 1, 2, ..., n$, $\alpha=0.05$.)

Table 3.2  Estimated coverage probability and the expected length of $N = 1000$ simulated confidence intervals for $\beta_1$ based on Wald Approach(WA) and Hypothesis Testing(HT) method when $\beta_1 = 0$. ($x_i = i$, for $i = 1, 2, ..., n$, $\alpha = 0.05$.)

Table 4.1  Estimated coverage probability and expected length of $N = 1000$ simulated confidence intervals for $p(x^*) = 1/(1 + \exp(-x^{**T}\beta))$ based on bootstrap(BS) and Hypothesis testing(HT) method. ($x_i = i$, for $i = 1, 2, ..., n$, $x^{**} = (1, 10.2)^T$, $\alpha = 0.05.$)

Table 4.2  Expected coverage probability of the prediction interval for the observed proportion $p^{(m)}(x^*)$. ($x_i = i$, for $i = 1, 2, ..., n$, $\alpha=0.05$, $M_1=1000$, $M_2=1000$)
CHAPTER 1

Introduction

1.1 Generalized Linear Models

Let $y_1, y_2, \ldots, y_n$ be a set of $n$ independent observations of which the $i$-th observation $y_i$ was taken under the conditions specified by the value $x_i = (1, x_{i1}, x_{i2}, \ldots, x_{ik})^T$ of a vector of $(k + 1)$ explanatory variables. Assume that $y_i$ has probability density function (or mass function in discrete case) from the exponential family

$$ f(y_i; \phi, \theta) = \exp \left\{ r(\phi)[y_i - g(\theta)] + h(\phi, y_i) \right\} $$

(1.1)

where $\theta_i$ is related to the mean of the distribution via the equation $E(y_i \mid x_i) = \mu_i = g'(\theta_i)$, and $\phi$ is a nuisance parameter which has a relation with the variance of the distribution given by $Var(y_i \mid x_i) = g''(\theta_i)/r(\phi)$ [see McCullagh and Nelder(1989)]. Furthermore, assume that there is a linear predictor $\eta_i = x_i^T \beta$, where $\beta$ is a vector of $(k + 1)$ unknown parameters, and a link function $s$ which provides the link

$$ \eta_i = s(\mu_i) $$

between the linear predictor and the mean of the probability distribution function of $y_i$. These assumptions outline the specifications for the generalized linear model (GLM) introduced by Nelder and Wedderburn(1972).
1.2 Logistic Regression Models

Suppose there are \( n \) independent trials and the \( i \)-th trial is carried out under the conditions specified by \( \mathbf{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{id})^T \) with an outcome \( r_i \), which is either a success \((r_i = 1)\) or a failure \((r_i = 0)\).

The logistic regression model for the \( r_i \) relates the probability of success \( p(x_i) \) in the \( i \)-th trial to the value of \( \mathbf{x}_i \) via the function

\[
p(x_i) = P(r_i = 1) = \frac{1}{1 + e^{-\mathbf{x}_i^T \beta}}, \quad i = 1, 2, \ldots, n.
\]

The above logistic regression model falls into the framework of GLM. The probability of failure (or success) in the \( i \)-th trial is given by

\[
p(x_i)^r_i[1 - p(x_i)]^{1 - r_i} = \exp[\theta_i \ln[p(x_i)] + (1 - r_i) \ln[1 - p(x_i)]], \quad r_i = 0, 1.
\]

Comparing equation Eq. (1.3) with Eq. (1.1), we see that

\[
\theta_i = \ln[p(x_i)] - \ln[1 - p(x_i)],
\]

\[
g(\theta_i) = -\ln[1 - p(x_i)] = -\ln\left[ \frac{e^{-\theta_i}}{1 + e^{-\theta_i}} \right],
\]

\[
r(\phi) = 1,
\]

\[
h(\phi, y_i) = 0
\]

and \( \mu_i = p(x_i) \), with link function given as

\[
\mathbf{x}_i^T \beta = s(\mu_i) = \ln \frac{\mu_i}{1 - \mu_i}.
\]

Logistic regression model is used in a wide variety of applications in the fields of biomedical studies, social science research, marketing, credit-scoring and genetics.

The likelihood function plays an important role in the analysis of data using logistic regression model. We may express this likelihood function as
\[ L(\boldsymbol{\beta}, \mathbf{x}_i) = \prod_{i=1}^{n} p(x_i)^{y} [1 - p(x_i)]^{1-y} . \]  

(1.4)

To find the maximum likelihood estimate \( \hat{\beta} \) of the parameter vector \( \beta \), we solve the following equations

\[ \frac{\partial \ln L(\boldsymbol{\beta}, \mathbf{x}_i)}{\partial \beta_i} = 0, \quad i = 1, 2, \ldots, n . \]  

(1.5)

For further discussion regarding maximum likelihood estimation in logistic regression models, we may refer to Albert and Anderson(1984), Berkson(1951, 1953, 1955), Cox(1958a, 1958b), Hodges(1958) and Walker and Duncan(1967).

The significance test of the hypothesis \( H_0: \beta_i = 0 \) regarding the individual parameter \( \beta_i \) is usually done using the Wald test, the likelihood ratio test or the score test. Wald test uses the maximum likelihood estimate \( \hat{\beta}_i \) and the inverse of the asymptotic covariance matrix evaluated at \( \hat{\beta} \). The likelihood ratio test uses twice the difference between the log-likelihood at \( \hat{\beta} \) and at the maximum likelihood estimate \( \tilde{\beta} \) of the model under \( H_0 \), while the score test uses the derivatives of the log-likelihood at \( \tilde{\beta} \) and the inverse of the asymptotic covariance matrix evaluated at \( \tilde{\beta} \). When the sample size is large, the acceptance region of the tests may be obtained by the standard normal distribution or the chi-square distribution with one degree of freedom and all the three tests usually give similar results. When \( |\beta| \) is relatively large, the Wald test does not show satisfactory results [see Hauck and Donner(1977), Jennings(1986)].

A confidence interval for \( \beta_i \) may be obtained as

\[ \{ \beta_i^{(0)} : H_0 \text{ that } \beta_i = \beta_i^{(0)} \text{ is accepted} \} . \]
For example, the Wald approach will yield the following approximate $(1-\alpha)100\%$ confidence interval

$$\{\beta_i^{(0)}: \hat{\beta}_i - z_{\alpha/2}(SE) \leq \beta_i^{(0)} \leq \hat{\beta}_i + z_{\alpha/2}(SE)\}$$

where $SE$ is the standard error of the estimate $\hat{\beta}_i$ of $\beta_i$ and $z_{\alpha/2}$ is the $100(1-\alpha/2)$-percentile of the standard normal distribution.

The lack of fit of the logistic model may be assessed by the likelihood ratio test which essentially compares the maximum of the likelihood with the maximum that would be experienced if one were to produce a perfect fit of the data. The lack of fit may also be assessed by comparing the observed counts with the fitted counts using a Pearson chi-square statistic or likelihood ratio $G^2$ statistic.

The Pearson chi-square statistic and the likelihood ratio $G^2$ statistic, which represent respectively the sum of squares of the Pearson residuals and the deviance residuals, may also be used to identify the large leverage point and to assess the influence of large leverage point on the statistical inference based on logistic regression model [see for example, Pregibon(1981) and William(1987)].

With several explanatory variables, there may be many other models which appear to fit the data reasonably well. The need to select the “best” model among these models thus arises. Goodman(1971) proposed methods similar to the forward selection and backward elimination in ordinary regression. Akaike information criterion (AIC) given as

$$\text{AIC} = -2(\text{maximum log-likelihood} - \text{number of parameters in model})$$

may also be used to select the “best” model for which the AIC gives the lowest value [see Akaike(1974)].
1.3 Extension of the Logistic Regression Models

The logistic regression model has been extended to the polychotomous logistic regression which is suitable for data involving a categorical response with more than two categories.

When the observations are of the form \( y_{it} \), \( i = 1, 2, \ldots, n; \ t = 1, 2, \ldots, T_i \), with \( y_{it} \) denoting the \( t \)-th observation in the \( i \)-th cluster, the logistic regression model has been generalized to the random effects model [see Breslow and Clayton(1993)]. In the random effects model, we add a random effect term \( z_i^T u_i \) (\( u_i \) is a vector of random effect values for cluster \( i \) and \( z_i \) is the corresponding column vector of explanatory variables) to the term \( x_i^T \beta \) to form \( \eta_i = x_i^T \beta + z_i^T u_i \) and the conditional mean \( E(y_{it} \mid u_i) \) is linked to \( \eta_i \) via the link function \( s \):

\[
s[E(y_{it} \mid u_i)] = x_i^T \beta + z_i^T u_i
\]

The generalized linear model which has the logistic regression model as a special case, has also been extended to the Bayesian model which introduces a prior distribution to the parameter vector \( \beta \). Research work on Bayesian models can be found in Albert(1988), Ibrahim and Laud(1991), Dellaportas and Smith(1993), Gelfand and Dey(1994), Bedrick, Christensen and Johnson(1996), Gosh, Natarajan, Stroud and Carlin(1998) and other references.

1.4 Outline of the Thesis

The thesis is primarily concerned with the statistical inference in simple logistic regression models. In Chapter 2, we discussed the approach on getting the approximate maximum likelihood estimate \( \hat{\beta} \) of \( \beta \) by using Taylor’s series expansion.
With that, we proceed to develop a procedure to approximate the distribution of the deviance statistics and the statistic based on approximate variance-covariance matrix for the estimate $\hat{\beta}$. We express the statistics as nonlinear functions of two uncorrelated random variables and derive approximate distribution for each statistic based on non-normal distributions of the two random variables. The resulting approximations are capable of capturing the higher order characteristics of the distribution of the statistics such as skewness and kurtosis.

In Chapter 3, we proposed a method based on hypothesis testing to construct a confidence interval for the slope parameter $\beta_1$ in the simple logistic regression models. In the process of finding the confidence interval, we have made use of the distribution of the maximum likelihood estimate $\hat{\beta}_1$ when the parameter vector $\beta$ in the model satisfies the conditions under the null hypothesis and is the nearest to the maximum likelihood estimate $\hat{\beta}$. The results from simulation studies show that the resulting confidence interval performs better than the Wald confidence interval.

In Chapter 4, we used the method based on hypothesis testing to find a confidence interval for the proportion of successes, $p(x^*)$ when the vector of explanatory variables assumes a value $x^*$ in the future. The confidence interval found is then enlarged to form a prediction interval for the future observed proportion of success when a number of trials are carried out under the conditions specified by $x^*$. The simulation studies performed show that the resulting prediction interval is satisfactory as its expected coverage probability is approximately equal to the target value.

Finally, the thesis is concluded by some concluding remarks in Chapter 5.