Comparing logistic regression methods for completely separated and quasi-separated data

by

Michelle Botes

Submitted in partial fulfilment of the requirements for the degree
MSc: Mathematical Statistics
In the Faculty of Natural & Agricultural Sciences
University of Pretoria
Pretoria
(30 August 2013)

Contents

List of Figures x

Introduction xii

I Theory of logistic regression 1

1 Logistic Regression 2

1.1 Introduction . . . 2
1.2 Generalised linear models . . . 2
1.3 Two-way contingency table . . . 4
1.4 Binary observations . . . 4
1.5 Deriving the logit function . . . 5
1.6 Dummy variables . . . 6
1.7 Covariate Patterns . . . 7
1.8 Relationship between the probability and the odds of an event . . . 9
1.9 Estimating the parameters . . . 11
1.10 Goodness of fit . . . 14
1.10.1 Pearson's chi-square test and deviance test . . . 15
1.10.2 Hosmer-Lemeshow statistic . . . 18
1.10.3 Model fit statistics . . . 19
1.10.4 Classification tables . . . 20
1.11 Significance of coefficients . . . 22
1.11.1 Test Statistics . . . 22
1.11.2 Confidence interval . . . 24
1.12 Conclusion . . . 25

2 Complete and quasi-complete separation and overlap 26

2.1 Introduction . . . 26
2.2 Complete separation . . . 27
2.3 Quasi-complete separation . . . 28

2.4 Overlap 29

2.5 Identifying complete or quasi-complete separation 31
2.6 Conclusion 33

3 Methods used to deal with separation 34
3.1 Introduction 34
3.2 Different methods 34
3.2.1 Changing the model 35

3.2.2 Working with the likelihood function 35
3.2.3 Other methods 36
3.3 Exact logistic regression 36
3.3.1 The ML estimates 36
3.4 Firth's Model 40
3.4.1 The model 40

3.4.2 The ML estimates 43

3.4.3 Method applied to complete and quasi-complete separation 45

3.5 Hidden logistic regression 46
3.5.1 The model 47
3.5.2 The ML method 48
3.5.3 Determining the values for δ_0 and $\delta_1$be 49
3.6 Conclusion 53

II Practical Application 54

4 Overview/ Outline of Part II 55

5 Complete separation (small sample) 58
5.1 Introduction 58
5.2 Continuous covariates 59
5.2.1 HIV status example 59
5.2.2 General logistic regression model 59
5.2.3 Exact logistic regression: HIV status 61
5.2.4 Firth's Method: HIV status 63
5.2.5 Hidden logistic regression model 65
5.2.6 Conclusion: HIV status example 66
5.3 Continuous and categorical covariates 68

5.3.1 Breast cancer example 68
5.3.2 General logistic regression model 68
5.3.3 Exact logistic regression 70
5.3.4 Firth's method 72
5.3.5 Hidden logistic regression 73
5.3.6 Conclusion: Breast cancer example 74
5.4 Categorical covariates 76
5.4.1 Titanic example 76
5.4.2 General logistic regression model 76
5.4.3 Exact logistic regression 78
5.4.4 Firth's method 80
5.4.5 Hidden logistic regression 81
5.4.6 Conclusion: Titanic example 82
5.5 Conclusion 83

6 Quasi-complete separation (large sample) 85

6.1 Introduction 85
6.2 Categorical covariates 85
6.2.1 Titanic example: large sample 85
6.2.2 General logistic regression model revisited 86
6.2.3 Exact logistic regression 87
6.2.4 Firth's method 89
6.2.5 Hidden logistic regression 90
6.3 Conclusion 91

7 Summary and Conclusion 94

References 96

Appendix A: Data on HIV status patients 100

Appendix B: Data on breast cancer patients 101

Appendix C: Data on Titanic passengers: small sample 102

Appendix D: Data on Titanic passengers: large sample 103

Appendix E: SAS program code for penalised likelihood function 104

Appendix F: SAS program code for classification tables 106

Appendix G: R program for hidden logistic regression 108

Appendix H: SAS program code for Pearson chi-square, deviance and Hosmer-Lemeshow statistics 110

Declaration

I, Michelle Botes, declare that the dissertation, which I hereby submit for the degree MSc:Mathematical Statistics at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

SIGNATURE: ...

DATE: .....

Acknowledgements

I would like to thank Dr Lizelle Fletcher who under took to act as my supervisor despite her many other academic and professional commitments. Her knowledge and commitment inspired and motivated me to complete my dissertation. I would also like to thank Prof D. Meyer and Prof F.E. Steffens for providing me with the HIV data set given in Appendix A. Finally I would like to thank my mother, Christine, and my fiancé, Christoffel, for their love, support, motivation and constant patience.

Summary

In experimental analysis of data the information most often required and of interest is how the changes in one variable (independent variable) affects another variable (dependent variable). The most distinctive difference between a logistic regression model and a linear regression model is the nature of the dependent variable. A linear regression model has a continuous dependent variable whereas with the logistic regression model the response is typically binary or dichotomous.

The logistic regression function is the log odds of a success expressed linearly as a combination of all the covariates included in the model. For a simple binary model where each observation can only take one of two possible forms, one cut-off value is implemented for the probability of a success of the dependent variable. If the probability of a success for a specific observation is above the cut-off value the observation is assigned to a specific group, for example, group 1, if the probability of a success for a specific observation is below the cut-off value the observation is assigned to another group, say group 2.

When one of the independent variables can perfectly classify the observations into the respective groups of the response variable, the likelihood function has no maximum and therefore no finite value can be found for the coefficient estimates. There are three different mutually exclusive and exhaustive classes into which the data from a logistic regression can be classified: complete separation, quasi-complete separation and overlapping data. Complete and quasi-complete separation imply that only an infinite or a zero maximum likelihood estimate could be obtained for the odds ratio which rarely can be assumed to be true in practice.

Numerous methods to detect complete separation or quasi-complete separation have been developed over the years. Exact logistic regression, Firth's method and hidden logistic regression will be discussed in this dissertation followed by practical examples in part II. These methods will be compared to one another in different scenarios when two covariates are considered. A small sample where complete separation is present is investigated and compared to a large sample in which quasi-complete separation is present. In each of the sample size cases a plot of the observations and the significance of the parameters are considered to confirm whether complete or quasi-complete separation is present in the

data.

If the data is non-overlapping, exact logistic regression, Firth's method and hidden logistic regression are applied to the data set. For each of these models the significance of the parameters, the goodness of fit of the model (Pearson's chi-square, deviance and Hosmer-Lemeshow test statistic) and the classification table are considered. Overall, the best results are obtained from exact logistic regression when working with a large sample and a data set which is not sparse. Firth's method gives significant coefficient estimates for all cases and transforms the data to represent a logistic curve which gradually increases to an estimated probability from 0 to 1. Finally the hidden logistic regression model gives perfect classification for all cases, but still closely resembles a model under complete or quasi-complete separation.

Abstract

An occurrence which is sometimes observed in a model based on dichotomous dependent variables is separation in the data. Separation in the data is when one or more of the independent variables can perfectly predict some binary outcome and it primarily occurs in small samples. There are three different mutually exclusive and exhaustive classes into which the data from a logistic regression can be classified: complete separation, quasi-complete separation and overlap. Separation (either complete or quasi-complete) in the data gives rise to a number of problems since it implies infinite or zero maximum likelihood estimates which are idealistic and does not happen in practice. In this dissertation the theory behind a logistic regression model, the definition of separation and different methods to deal with separation are discussed in part I. The methods that will be focussed on are exact logistic regression, Firth's method which penalises the likelihood function and hidden logistic regression. In part II of this dissertation the three fore mentioned methods will be compared to one another. This will be done by applying each method to data sets which exhibit either complete or quasi-complete separation for different sample sizes and different covariate types.

List of Figures

- Figure 1.1: Logistic curve for values where $\beta_j > 0$ 11
- Figure 1.2: Logistic curve for values where $\beta_j < 0$ 11
- Figure 2.1: Scatter plot of x and y under complete separation 27
- Figure 2.2: Log−likelihood function as a function of β for complete separation 28
- Figure 2.3: Scatter plot of x and y under quasi-complete separation 29
- Figure 2.4: Log−likelihood function as a function of β under quasi−complete separation 30
- Figure 2.5: Scatter plot of x and y for overlapping data 31
- Figure 2.6: The log−likelihood function as a function of β for overlapping data 31
- Figure 3.1: Modified score function. 42
- Figure 3.2: The penalised log−likelihood function as a function of β for complete separation. 45
- Figure 3.3: The penalised log−likelihood function as a function of β for quasi−complete separation. 46
- Figure 3.4: Probability structure of the true status against the observable response 47
- Figure 5.1: Scatter plot of $x_{V3.74}$ vs. $x_{V1.18}$ grouped according to the observed value of y 59
- Figure 5.2: HIV example predicted probabilities for exact, Firth and hidden 67
- Figure 5.3: Scatter plot of x_{NODES} vs. x_{AGE} grouped according to the observed value of y 68
- Figure 5.4: Breast cancer example predicted probabilities for exact, Firth and hidden 75

- Figure 5.5: Scatter plot of x_{SEX} vs. x_{CLASS} according to the observed value of y 76
- Figure 5.6: Titanic example predicted probabilities for exact, Firth and hidden 83
- Figure 6.1: Large sample scatter plot of x_{SEX} vs. x_{CLASS} according to the observed value of y. 86
- Figure 6.2: Large sample Titanic example predicted probabilities for exact, Firth and hidden 92

Introduction

Regression is an essential part of any statistical data analysis used to explain the relationship between a dependent and one or more independent variables. The logistic regression model is used to predict a discrete outcome as opposed to a continuous value obtained from a linear regression model.

Before studying the theory of logistic regression it is essential to recognise that the object of this method is the same as for any other regression method: to obtain a model which explains as much of the variation in dependent variable as possible. The coefficient estimates for a regression model can be obtained in many different ways, in this dissertation the coefficient estimates for the logistic regression model will be obtained with maximum likelihood estimation.

When the outcome variable in a data set is discrete, the situation of complete or quasi-complete separation can arise within the observations. This occurs when one or more of the independent variables can perfectly predict the dependent variable. For example, consider a medical study where breast cancer for different genders is examined. It is much more likely that a female will have breast cancer than a male patient. Therefore, if a small sample is considered and only the gender of a patient is used to predict if a patient has breast cancer or not, it is very likely that all the female patients in the sample will have breast cancer and all the males in the study not. This scenario is an example of complete separation. When either complete or quasi-complete separation is present in a data set, the maximum likelihood estimates will not be obtainable due to non-convergence in the iteration process.

This dissertation comprises a study of the logistic regression model and ways to identify if complete or quasi-complete separation is present in a data. When separation has been identified, different solutions to obtain a model with convergent coefficient estimates are investigated. The models obtained under the different approaches are then compared to one another to identify which model is preferred under specific constraints from the original data set.

Part I

Theory of logistic regression

Chapter 1

Logistic Regression

1.1 Introduction

A logistic regression model is introduced in part I as a generalised linear model (GLM). This model is derived to describe the relationship between the response and the input. The most distinctive difference between a logistic regression model and a linear regression model is the response or dependent variable. A linear regression model has a continuous dependent variable whereas for the logistic regression model the response is binary or dichotomous. A great many variables in any field are dichotomous : male vs. female, guilty vs. not guilty, defective vs. non-defective just to name a few.

In Chapter 1 the derivation of the logistic regression model will be discussed. The different types of categorical observations will be investigated in Section 1.3 and 1.4 from which the logit function can be derived in Section 1.5. For categorical input values, dummy variables need to be created as shown in Section 1.6 and the definition of a sparse data set is defined in Section 1.7. Since a logistic regression model is based on two possible outcomes, the probability and the odds that links to a dichotomous variable will be investigated in Section 1.8. For any model the coefficients need to be estimated and evaluated, this will be addressed in Section 1.9. Finally different ways to test the significance of the parameters and the model will be addressed in Sections 1.10 and 1.11.

1.2 Generalised linear models

In experimental analysis of data the information most often required and of interest is how the changes in one variable (independent variable) affects another variable (dependent variable). Observations obtained from a survey, census, etc. is the dependent variable which represent the experimental or survey units example: students, companies, patients etc. (the items on which the observations were made). The different independent variables (input) considered for each observation example: age, weight, sex, etc. are also known as

the covariates. In matrix notation, let the set of observations be represented by a $n \times 1$ column vector $\mathbf{y} = [y_1, ..., y_n]^T$ and the independent variables be denoted by a $n \times (m+1)$ matrix $\mathbf{X}$. Each row in $\mathbf{X}$ represents a different observation and each column represents a covariate. Associated which each covariate is a set of coefficient or parameter values represented by a $(m+1) \times 1$ column vector $\boldsymbol{\beta} = [\beta_0, ..., \beta_m]^T$. The classical linear model can be expressed (McCullagh & Nelder 1989, p. 9) as the relationship between the independent and dependent variables by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1.1}$$

where $\boldsymbol{\varepsilon}(\boldsymbol{\beta})$ is a $n \times 1$ column vector of the residual terms.

The assumption under a linear regression model is that the observations of the dependent or response variable are independent, this assumption of independence is carried over to the wider class of generalised linear models. If a relationship between the dependent and independent variables exists a model can be fitted to predict some continuous value for the dependent variable. For a set of independent variables there is a range of possible values for the dependent variable and vice versa, this variation will occur due to measurement errors and variation between experimental units.

Over a period of time it was found that not only continuous values were desired for the response variable, but discrete values for the enumeration of probabilities were also needed. In 1952, Dyke and Patterson were of the first authors to publish a study on cross-classifying data consisting of the proportion of subjects who has a good knowledge of cancer. This analysis of counts in the form of proportions can be modelled by using a Bernoulli distribution to indicate the probability of a "success" or a "failure" of a single event; to model the number of "success" or "failures" in a fixed pool of survey units the binomial distribution will be suitable.

A model was developed, the logistic regression function, which is the log odds of a success expressed linearly as the combination of all the covariates considered. The odds of a success is bound by the log function to ensure only a small range of values for the response variable. For a simple binary model where the observation can only take one of two possible forms, one cut-off value is implemented for the probability of a success of the dependent variable. If the probability of a success for a specific observation is above the cut-off value the observation is assigned to a specific group, for example, group 1, if the probability of a success for a specific observation is below the cut-off value the observation is assigned to the other group, say group 2. Assigning each observation to a specified group ensures a discrete value for the dependent variable. Since only discrete values for the dependent variable Y are considered for the logistic regression model, the error or residual terms can rarely be used to accurately assess the fit of a logistic regression.

Keeping this in mind a number of different ways to assess the fit of the logistic regression model will be explored in Section 1.10.

1.3 Two-way contingency table

When the values obtained from the dependent and the independent variable can be categorized into a finite number of groups, the observations can be cross-classified in a contingency table. Let the independent variable X have k groups and the dependent variable Y have g groups then the k by g possible outcomes can be expressed in a table with k rows for the groups of X and g columns for the groups of Y. The cells in the table represent the kg possible outcomes.

Consider a study on $n = 20$ individuals where the dependent variable is marital status and there are 4 possible groups: single, married, divorced or widow/ed. The independent variable of interest is the number of children of each individual and is categorized in the following groups: no children - group 1, one or two children - group 2 and finally for more than two children the observation is allocated to group 3. The contingency table for variables X and Y with $k = 3$ and $g = 4$ groups respectively can be expressed in Table 1.1.

Table 1.1: Two-way contingency table

		Marital Status				
		Single	Married	Divorced	Widow/ed	**Total**
Children	Group 1	4	1	1	0	**6**
	Group 2	1	3	3	1	**8**
	Group 3	0	3	3	0	**6**
	Total	**5**	**7**	**7**	**1**	**20**

The contingency table expressed in Table 1.1 cross-classifies only two variables X and Y, this is called a two-way contingency table. A contingency table which cross-classifies three variables is called a three-way table and so forth. A two-way contingency table with k rows and g columns as given in Table 1.1 is expressed as a $k \times g$ table, or for this example, a 3×4 table.

1.4 Binary observations

When the dependent variable only has two groups i.e. $g = 2$ then Y is a binary variable. Consider a group of $n = 30$ mice on which an experiment is conducted with two different

possible treatments available. Each mouse can only receive a single treatment and the result for each mouse after a period of time will be whether it survived or not. If the group of mice is divided into $n_1 = 13$ mice receiving treatment 1 and $n_2 = 17$ mice receiving treatment 2, the 2×2 contingency table can be given in Table 1.2.

Table 1.2: Contingency table for binary observations

		Outcome		
		Survived	Died	**Total**
Treatment	Treatment 1	12	1	**13**
	Treatment 2	11	6	**17**
	Total	**23**	**7**	**30**

Since the response now only has one of two possibilities, the terms "success" and "failure" can be used to identify the two different responses. In Table 1.2 the columns are the response levels of the dependent variable Y. Let a success be the event that a mouse survived a treatment and a failure be the event that a mouse died after a treatment. From the $n = 30$ mice considered for the experiment 23 survived and 7 died, therefore the overall probability of a success is $23/30 = 76.67\%$ which will be denoted by π and the overall probability of a failure is $7/30 = 23.33\%$ which will be indicated by $(1 - \pi)$.

1.5 Deriving the logit function

When considering a binary logistic regression model, the dependent variable Y_i; $i = 1, 2, ..., n$, is defined as some categorical variable with a qualitative value of say 0 or 1. The outcome of 0 or 1 is the result of condensing a more complex input value. The value of 1 usually represents a “success” whereas 0 typically indicates a “failure”. Different probabilities are assigned to the different outcomes of Y_i by $P(Y_i = 1) = \pi_i$ and $P(Y_i = 0) = 1 - \pi_i$ for $i = 1, 2, ..., n$ where the n responses are assumed to be independent. Most common in practice is when the values of π_i is restricted to a single value. If Y has its distribution defined by the single probability π, the probability of a success and a failure can be simplified to $P(Y = 1) = \pi$ and $P(Y = 0) = 1 - \pi$ respectively.

The examples considered in chapters 1.3 and 1.4 are examples of the probability of a success depending on a single explanatory variable X. The model can be extended to a multiple regression model where the binary response variable, which can only assume categorical values, depends on m explanatory variables which may be either quantitative or qualitative.

The probability of a success can be modelled as a function of the independent variables $X_1, X_2, ..., X_m$, therefore a logistic regression function is a linear function of the observed values of X_i which can be expressed (Cox and Snell 1989, p. 26) by equation (1.2)

$$\theta_i = \text{logit}(\pi_i) = \log\left[\frac{\pi_i}{1-\pi_i}\right] = x_{i0}\beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{im}\beta_m = \sum_{j=0}^{m} x_{ij}\beta_j \quad (1.2)$$

where $i = 1, 2, ..., n$ and $j = 0, 1, 2, ..., m$. The value of $\log\left[\frac{\pi_i}{1-\pi_i}\right]$ is defined as the logit function and is the natural logarithm of the odds for event i which can also be described as the ratio between the probability of a successes (π_i) and the probability of a failure $(1 - \pi_i)$. The model represented by equation (1.2) can also be expressed by

$$\begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix} = \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1m} \\ x_{20} & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{nm} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix} \quad (1.3)$$

which is simplified to $\boldsymbol{\Theta} = \mathbf{X}\boldsymbol{\beta}$ with $\boldsymbol{\Theta} : n \times 1, \boldsymbol{\beta} : (m+1) \times 1, \mathbf{X} : n \times (m+1)$ where the i^{th} row vector of observation i is represented by $\mathbf{x}_i = [x_{i0}, x_{i1}, ..., x_{ij}, ..., x_{im}]$, the j^{th} column vector of covariate j is represented by $\mathbf{x}_j = [x_{1j}, x_{2j}, ..., x_{ij}, ..., x_{nj}]^T$ and the first column vector of $\mathbf{X}$ is represented by $\mathbf{x_0} = [1, 1, ..., 1]^T$.

1.6 Dummy variables

When a logistic regression model is fitted to a data set which contains categorical explanatory variables, it is important to create dummy variables which represent the different categories. This is done since if a categorical value is represented by a numerical value, this value only represents a category and has no numerical properties. The number of dummy variables considered for a specific covariate, is the number of categories for the specific explanatory variable (k) minus 1. Therefore for the example considered in Section 1.3, the three different groups for the number of children will be replaced by a dummy variable with two indicators say x_{d_1} and x_{d_2} where group 1 can be represented by $x_{d_1} = 0$ and $x_{d_2} = 0$, group 2 can be denoted by $x_{d_1} = 1$ and $x_{d_2} = 0$ and finally $x_{d_1} = 0$ and $x_{d_2} = 1$ to represent group 3. This coding is illustrated in Table 1.3.

If the j^{th} explanatory variable which is represented by $\mathbf{x}_j = [x_{1j}, x_{2j}, ..., x_{nj}]^T$ is a categorical variable with k_j possible levels, then $k_j - 1$ indicators will be needed. Let the indicators be represented by $x_{d_{jl}}$, then the coefficients for these indicators will be expressed by $\beta_{jl}; l = 1, 2, ..., k_j - 1$. Therefore, for a logistic regression model with m explanatory

Table 1.3: Coding for dummy variables

		Indicator	
		x_{d_1}	x_{d_2}
Children	Group 1	0	0
	Group 2	1	0
	Group 3	0	1

variables and the j^{th} variable a categorical variable, the model in equation (1.2) can be represented (Hosmer & Lemeshow 2001, p. 33) by

$$\theta_i = \beta_0 x_{i0} + x_{i1}\beta_1 + x_{i2}\beta_2 + ... + \sum_{l=1}^{k_j-1} x_{d_{jl}}\beta_j + ... + x_{im}\beta_m. \tag{1.4}$$

1.7 Covariate Patterns

Consider the logistic regression model as shown in equation (1.3) and suppose the fitted model has m covariates. A single observed set of covariates for say the i^{th} observation can be given by the i^{th} row of $\mathbf{X}$, $\mathbf{x}_i = [x_{i0}, x_{i1}, x_{i2}, ..., x_{im}]$. There are two types of covariate patterns in a data set, the first is where every observed set of covariates are distinct for all observations $i = 1, ..., n$. The second is where a covariate set is repeated for two or more different observations. Therefore if a specific number of observations all share the covariate set $\mathbf{x}_i = [x_{i0}, x_{i1}, x_{i2}, ..., x_{im}]$, let v_c denote the number of observations which share this covariate pattern in the c^{th} covariate class. Individuals sharing a specific covariate set is said to form a covariate class.

To illustrate this using a simple example consider the result of a specific course a student enrolled for (student passing the course is indicated by $y_i = 1$ or failing the course is indicated by $y_i = 0$). The results of this course are dependent on class attendance (never attended is indicated by $x_{i1} = 1$, sometimes attended by $x_{i1} = 2$ or often attended $x_{i1} = 3$) and on the semester mark obtained throughout the year (a semester mark equal to or above 50% is indicated by $x_{i2} = 1$ and a semester mark strictly below 50% is indicated by $x_{i2} = 2$). A sample of 10 students is considered and the observed values are tabulated in the first column of Table 1.4.

Let q where $c = 1, 2, ...q$ denote the number of distinct covariate patterns, then if every observation has its own unique covariate pattern, $q = n$ otherwise if some observations have tied covariate patterns then $q < n$. When one or more of the covariates are continuous it is very likely that the observations will have their own unique covariate pattern. In the event that the number of tied covariate patterns are few, the data set is seen as a sparse data set (McCullagh & Nelder 1989, p. 120). Sparseness does not necessarily indicate

Table 1.4: Representing covariate patterns

Observations represented individually			Observations represented by covariate class			
i	Covariate (x_{i1}, x_{i2})	Response y_i	c	Covariate (x_{c1}, x_{c2})	Class size v_c	Response y_{c1}
1	$(2,1)$	0	1	$(1,1)$	1	0
2	$(1,1)$	0	2	$(1,2)$	1	0
3	$(2,2)$	0	3	$(2,1)$	1	0
4	$(3,1)$	0	4	$(2,2)$	3	2
5	$(1,2)$	0	5	$(3,1)$	2	1
6	$(3,1)$	1	6	$(3,2)$	2	2
7	$(3,2)$	1				
8	$(2,2)$	1				
9	$(2,2)$	1				
10	$(3,2)$	1				

that the observations does not express much about the data set or that the parameters obtained will give a poor representation of the covariates. To the contrary, as discussed in (McCullagh & Nelder 1989, p. 120) when the sample size is large, the asymptotic approximation is quite accurate.

Covariate classes make it visually easier to analyse and interpret the data set by highlighting the patterns which occur most often. Adding covariate classes does however have the disadvantage that the order of the original data set is lost and the original data set cannot be reconstructed from the covariate class summary. If the order of the data set is not important, which is most often the case when using random samples, no information will be lost.

For binary observations which are grouped into covariate classes the probability of a success for each class is expressed by $\frac{y_{c1}}{v_c}$ for $c = 1, 2, ..., q$ and $0 \leq y_{c1} \leq v_c$, where y_{c1} is the number of successes out of the v_c observations for the c^{th} covariate class. The covariate class sizes can be expressed by vector $\mathbf{v} = (v_1, v_2, ..., v_q)$. If all the observations have their own unique covariate pattern the size for all covariate classes is 1 and the covariate class vector can be expressed by $\mathbf{v} = (v_1, v_2, ..., v_n)$. The covariate classes for the student results are expressed in column 2 of Table 1.4 where $q = 6$.

Whether a data set is grouped or ungrouped according to covariate classes is important for the different goodness-of-fit tests discussed in Section 1.10. It is also essential for different methods to determine coefficient estimates for a logistic regression model which is explained in Chapter 3. One of the methods explained in Chapter 3 (exact logistic regression) takes the sum over discrete patterns of covariate values to determine the coefficients for the logistic regression model as explained in Zorn (2005). For the scenario where all the observations have their own unique covariate pattern, the exact logistic

regression model could give unreliable coefficient estimates.

1.8 Relationship between the probability and the odds of an event

Since the odds of an event is the ratio between the number of successes and failures it can also be expressed by

$$odds = \left[\frac{\pi_i}{1-\pi_i}\right] = \frac{P(Y_i = 1)}{P(Y_i = 0)} \tag{1.5}$$

and since $0 \leq \pi_i \leq 1$ the odds of the event will always be greater than or equal to 0. The relationship between the odds of an event and the probabilities π_i is illustrated in Table 1.5.

Table 1.5: Relationship between probability, odds and log(odds)

π_i	$odds = \left[\frac{\pi_i}{1-\pi_i}\right]$	$\log\left[\frac{\pi_i}{1-\pi_i}\right]$
1	Not defined	Doesn't exist
0.75	3	0.478
0.5	1	0
0.25	0.333	-0.478
0	0	Doesn't exist

Therefore a value of $0 < \pi_i < 0.5$ will lead to $odds < 1$ and $\log(odds) < 0$; $0.5 < \pi_i < 1$ will give $odds > 1$ and $\log(odds) > 0$ and finally an equal probability of $\pi_i = 0.5$ will lead to the odds of a event being 1, i.e. even odds, and the log of this event is 0.

When the value of π_i is estimated the estimated value of each binary observation y_i namely $\widehat{y_i}$ can be calculated. Since $\widehat{y_i}$ will have a continuous value, a cut-off value for $\widehat{\pi_i}$ must be specified: if the outcome is above the cut-off value the observation will be classified as $\widehat{y_i} = 1$ and for any value below this cut-off value the observation will be labelled as $\widehat{y_i} = 0$. As is inevitable in model fitting, some observations could be misclassified, i.e. $\widehat{y_i} = 1$ when $y_i = 0$ and vice versa. Unlike linear regression however, the fitting error cannot be quantified since the true value cannot be decomposed as the sum of the fitted value and an error term, therefore a classification table (discussed in Section 1.10.4) will be used to report misclassified values.

The logit model expressed in equation (1.2) can be written in terms of the row vector $\mathbf{x}_i$ and column vector $\boldsymbol{\beta}$, i.e.

$$\log\left[\frac{\pi_i}{1-\pi_i}\right] = \mathbf{x}_i\boldsymbol{\beta}. \tag{1.6}$$

Taking the antilog of equation (1.6) yields

$$\begin{aligned} & e^{\log\left[\frac{\pi_i}{1-\pi_i}\right]} = e^{\mathbf{x}_i\boldsymbol{\beta}} \\ \text{i.e.} \quad & \frac{\pi_i}{1-\pi_i} = e^{\mathbf{x}_i\boldsymbol{\beta}} \\ \text{i.e.} \quad & \pi_i = e^{\mathbf{x}_i\boldsymbol{\beta}}(1-\pi_i) \\ \text{i.e.} \quad & \pi_i + \pi_i e^{\mathbf{x}_i\boldsymbol{\beta}} = e^{\mathbf{x}_i\boldsymbol{\beta}} \end{aligned}$$

$$\therefore \quad \pi_i = \frac{e^{\mathbf{x}_i\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i\boldsymbol{\beta}}} \tag{1.7}$$

and

$$\begin{aligned} 1-\pi_i &= 1 - \frac{e^{\mathbf{x}_i\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i\boldsymbol{\beta}}} \\ &= \frac{1+e^{\mathbf{x}_i\boldsymbol{\beta}} - e^{\mathbf{x}_i\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i\boldsymbol{\beta}}} \\ &= \frac{1}{1+e^{\mathbf{x}_i\boldsymbol{\beta}}}. \end{aligned} \tag{1.8}$$

Therefore $\pi_i = P(Y_i = 1) = \frac{e^{\mathbf{x}_i\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i\boldsymbol{\beta}}}$ (Cox and Snell 1989, p. 19) or equivalently, if the equation is multiplied with $\left(\frac{e^{-\mathbf{x}_i\boldsymbol{\beta}}}{e^{-\mathbf{x}_i\boldsymbol{\beta}}}\right)$ the probability of a success is given by

$$\pi_i = P(Y_i = 1) = \frac{1}{1+e^{-\mathbf{x}_i\boldsymbol{\beta}}}. \tag{1.9}$$

To visualise how the function in equation (1.7) responds to different values of π_i one can simplify the model in equation (1.7) to include only one explanatory variable, say the j^{th} covariate, and no constant term expressed by

$$\pi_i = \frac{e^{x_{ij}\beta_j}}{1+e^{x_{ij}\beta_j}}. \tag{1.10}$$

Figures 1.1 and 1.2 (Allison 2012, p. 91), illustrate the logistic curve of the probability π_i at different values of x_{ij} when β_j is positive and when β_j is negative. It can be noted from Figure 1.1 that there is an upward trend when $\beta_j > 0$, and from Figure 1.2 that there is a downward trend when $\beta_j < 0$.

Figure 1.1: Logistic curve for values where $\beta_j > 0$

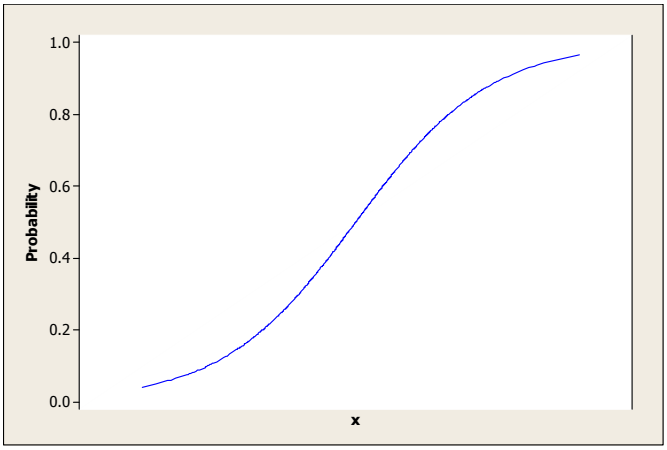
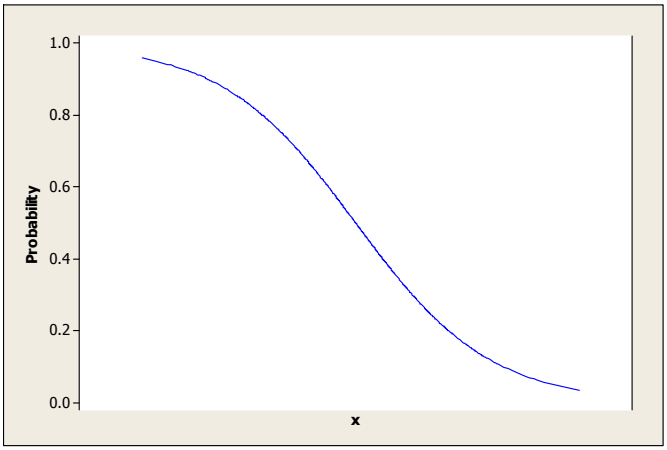


Figure 1.2: Logistic curve for values where $\beta_j < 0$



1.9 Estimating the parameters

To use the function in equation (1.2) one needs to determine the elements of the column vector $\boldsymbol{\beta}$. One of the most popular methods to estimate unknown coefficients is using maximum likelihood (ML) estimation. Since the values of $\mathbf{y} = [y_1, ..., y_n]^T$ are assumed to be independent, the likelihood function (Collett 2003, p. 67) of the n binary observations as a function of $\boldsymbol{\beta}$ is:

$$l(\boldsymbol{\beta}) = P(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n; \pi_1, ..., \pi_n) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}. \tag{1.11}$$

The log-likelihood function of $\boldsymbol{\beta} = [\beta_0, ..., \beta_m]^T$ is obtained by determining the natural logarithm of the likelihood function in equation (1.11) and by rewriting equation (1.6) as $\log(\pi_i) = \mathbf{x}_i\boldsymbol{\beta} + \log(1 - \pi_i)$, that is

$$\begin{aligned}
\log l(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \{y_i \log \pi_i + (1-y_i)\log(1-\pi_i)\} \\
&= \sum_{i=1}^{n} \{y_i(\mathbf{x}_i\boldsymbol{\beta} + \log(1-\pi_i)) + (1-y_i)\log(1-\pi_i)\} \\
&= \sum_{i=1}^{n} \{y_i\mathbf{x}_i\boldsymbol{\beta} + y_i\log(1-\pi_i) + \log(1-\pi_i) - y_i\log(1-\pi_i)\} \\
&= \sum_{i=1}^{n} \{y_i\mathbf{x}_i\boldsymbol{\beta} + \log(1-\pi_i)\} \\
&= \sum_{i=1}^{n} \left\{y_i\mathbf{x}_i\boldsymbol{\beta} - \log(1+e^{(\mathbf{x}_i\boldsymbol{\beta})})\right\}. \qquad (1.12)
\end{aligned}$$

The ML estimate for $\beta_j; j = 0, 1, ..., m$ can be obtained by getting the derivative of equation (1.12) with respect to β_j and setting this derivative equal to 0. The derivative of equation (1.12) for $j = 1, 2, ..., m$ as given in Heinze and Schemper (2002) is

$$\begin{aligned}
U(\beta_j) &= \frac{\partial \log l(\boldsymbol{\beta})}{\partial \beta_j} \\
&= \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} \frac{e^{x_{ij}\beta_j}(x_{ij})}{1+e^{x_{ij}\beta_j}} \\
&= \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} \pi_i x_{ij} \\
&= \sum_{i=1}^{n} (y_i - \pi_i) x_{ij} \overset{set}{=} 0. \qquad (1.13)
\end{aligned}$$

The derivative of equation (1.12) with respect to β_0, set equal to 0, simplifies to

$$\begin{aligned}
U(\beta_0) &= \frac{\partial \log l(\boldsymbol{\beta})}{\partial \beta_0} \\
&= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{e^{\beta_0}}{1+e^{\beta_0}} \\
&= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi_i \overset{set}{=} 0 \qquad (1.14)
\end{aligned}$$

since $x_{i0} = 1$ for $\forall i$. From equation (1.14) and Hosmer and Lemeshow (2001, p. 10) it can be noted that the solution for the sum of the observed values can be expressed as

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{\pi}_i \qquad (1.15)$$

where $\widehat{\pi}_i$ is the predicted values of π_i.

In logistic regression the equations given in (1.13) and (1.14) are nonlinear in $\beta_j; j = 0, 1, ..., m$ and thus require special methods to solve. The solution of $\boldsymbol{\beta}$ from equations (1.13) and (1.14) which is called the ML estimator ($\widehat{\boldsymbol{\beta}}$), is obtained by using iterative methods. For a simple logistic regression model, a model is saturated when the number of groups for the independent variable, k, equals the number of unknown parameters $(m + 1)$ in the model. For saturated models the equations in (1.13) and (1.14) can be solved explicitly for the ML estimator $\widehat{\boldsymbol{\beta}}$. Examples hereof include the case where the binary response variable can be found as a function with only one variable X (with two possible outcomes) such that observed values obtained can be expressed in a 2×2 contingency table as expressed in Table 1.2. In this case the observed frequencies can be used in the natural logarithm of the cross product ratio to obtain the ML estimates (Allison et al., 2004)

$$\widehat{\beta} = \log \frac{f_{11} f_{22}}{f_{12} f_{21}} \tag{1.16}$$

where f_{ij} indicates the frequency for the i^{th} row and j^{th} column in the 2×2 contingency table.

For most models, however, the model cannot be classified as saturated and therefore no explicit solution can be determined. In this scenario the ML estimates have to be obtained with numerical methods, the most often applied method being Newton-Rhapson.

Consider the i^{th} row vector $\mathbf{x}_i = (1, x_{i1}, x_{i2}, ...x_{im})$ as an $(m+1) \times 1$ column vector of the covariates for observation i i.e. $\mathbf{x}_i^T$. The first derivative of the log-likelihood with respect to $\boldsymbol{\beta}$, set equal to 0, can be found (Allison 2012, p. 44) as

$$\mathbf{U}(\boldsymbol{\beta}) = \frac{\partial \log l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{x}_i^T y_i - \sum_{i=1}^{n} \mathbf{x}_i^T (1 + e^{\beta \mathbf{x}_i^T})^{-1} \tag{1.17}$$

where $\mathbf{U}(\boldsymbol{\beta})$ is a $(m+1) \times 1$ column vector of partial derivatives $\left[\frac{\partial \log l(\boldsymbol{\beta})}{\partial \beta_0}, \frac{\partial \log l(\boldsymbol{\beta})}{\partial \beta_1}, ..., \frac{\partial \log l(\boldsymbol{\beta})}{\partial \beta_m}\right]^T$ and is also known as the gradient or score function. The second derivative with respect to $\boldsymbol{\beta}$ of the log-likelihood function set equal to 0 is indicated (Allison 2012, p. 44) by

$$\mathbf{I}(\boldsymbol{\beta}) = \frac{\partial^2 \log l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{x}_i (1 + e^{\beta \mathbf{x}_i^T})^{-1} (1 - (1 + e^{\beta \mathbf{x}_i^T})^{-1}) \tag{1.18}$$

where $\mathbf{I}(\boldsymbol{\beta})$ is a matrix of second partial derivatives and is also known as the Hessian matrix. The Hessian matrix is not only used to estimate coefficients from the well developed theory of ML estimation (Rao, 1973), but is also used to estimate the variances and covariances of the estimated coefficients.

The diagonal and off-diagonal elements of the Hessian matrix can be expressed (Hosmer & Lemeshow 2001, p. 34) by

$$\frac{\partial^2 \log l(\boldsymbol{\beta})}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i (1-\pi_i) \tag{1.19}$$

and

$$\frac{\partial^2 \log l(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi_i (1-\pi_i) \tag{1.20}$$

for $j, l = 0, 1, 2, ..., m$. The variances and covariances of the estimated coefficients are obtained from the inverse of the Hessian matrix i.e.

$$Var(\boldsymbol{\beta}) = \mathbf{I}^{-1}(\boldsymbol{\beta}). \tag{1.21}$$

The value for $Var(\boldsymbol{\beta})$ at the estimated value $\widehat{\boldsymbol{\beta}}$ will be donated by $Var(\widehat{\boldsymbol{\beta}})$, where the values in this matrix for the j^{th} coefficient estimate will be expressed by $\widehat{Var}(\widehat{\beta}_j)$ and $\widehat{Cov}(\widehat{\beta}_j, \widehat{\beta}_l)$.

The estimates for $\boldsymbol{\beta}$ is then calculated (Allison 2012, p. 44) by the Newton-Rhapson algorithm formulated by

$$\boldsymbol{\beta}^{new} = \boldsymbol{\beta}^{old} - \mathbf{I}^{-1}(\boldsymbol{\beta}^{old})\mathbf{U}(\boldsymbol{\beta}^{old}) \tag{1.22}$$

where $\mathbf{I}^{-1}(\boldsymbol{\beta})$ is the inverse of the Hessian matrix $\mathbf{I}(\boldsymbol{\beta})$. The column vector of starting values ($\boldsymbol{\beta}^{old}$) is substituted on the right hand side of equation (1.22), this will yield the new value ($\boldsymbol{\beta}^{new}$) on the left hand side of (1.22), $\boldsymbol{\beta}^{new}$ is then used on the right hand side of the equation to obtain the next new value on the left hand side. This process will be repeated until the left hand side of the equation is equal to the right hand side which is an indication that the process converged. This process usually takes fewer than 25 iterations for convergence, if the process did not converge after 25 iterations the chance that it will converge is very low. When the process does not converge, it is a strong indication that separation is present in the data, which will be discussed in Chapter 2.

1.10 Goodness of fit

Many goodness-of-fit measures exist to show how well a given logistic regression model fits the data. There is however no overall best method which can be singled out to assess the adequacy or inadequacy of a given logistic regression model. Each method has its own advantages and disadvantages depending on the sample size, type of covariates used, etc. Four different types of methods will be considered, the first method comprises of

tests which are based on covariate patterns and consist of the Pearson's chi-square test (Pearson, 1900) and the deviance test. The second method uses estimated probabilities from the assumed model which is known as Hosmer and Lemeshow's $\widehat{C}$ and $\widehat{H}$ tests. The third method is informal model fit statistics, followed by using a tabular method, the classification tables. There are a number goodness-of-fit models available in practice and the interested reader is referred to Liu (2007) and Hosmer and Lemeshow (2001).

1.10.1 Pearson's chi-square test and deviance test

In linear regression the significance of a model is found by computing the squared distance between the observed and the predicted outcome value also known as the SSE. This value indicates how close the model correctly predicted the outcome variable. If the size of the SSE is large it implies a large distance between the observed and the estimated outcome which is an indication that the model is not a good predictor. In logistic regression the same approach is followed i.e. to take the difference between the observed and the predicted outcome. According to Hosmer and Lemeshow (2001) the deviance statistic for logistic regression plays the same role as the SSE plays in linear regression.

To derive the Pearson chi-square statistic, assume n independent observations from a Bernoulli distribution where the probability of a success for a single observation is given by π_i. Then the logistic regression model can be derived as described in Section 1.5. The estimated parameters of the logistic regression model can be obtained with numerical methods as discussed in Section 1.9. Suppose a logistic regression model is derived as in equation (1.2), then if the fitted model has m covariates, each observation $y_i; i = 1, 2, ..., n$ has a single covariate set $\mathbf{x}_i = [x_{i0}, x_{i1}, ..., x_{im}]$ which is represented in a single row of $\mathbf{X}$ in (1.3). As discussed in Section 1.7 two types of covariate patterns can exist in a data set. The first is where each observation has its own unique covariate set and there are no tied covariate sets (i.e. $q = n$), this is also referred to as a sparse data set. The second type is where more than one observation share the same covariate set (i.e. $q < n$).

By considering the second type of covariate pattern ($q < n$), if v_c number of observations in the c^{th} covariate class share the same covariate set then $\sum_{c=1}^{q} v_c = n$. If n_s is the total number of successes and n_f is the total number of failures over all the covariate classes then $n = n_s + n_f$. If the total number of successes in the c^{th} covariate class is given by y_{c1} and the total number of failures is given by y_{c0} then $\sum_{c=1}^{q} y_{c1} = n_s$ and $\sum_{c=1}^{q} y_{c0} = n_f$.

The expected number of successes for the c^{th} covariate class can be given by (Hosmer and Lemeshow 2001, p. 145)

$$\widehat{y_{c1}} = v_c\widehat{\pi}_c \tag{1.23}$$

where $\widehat{\pi}_c$ is the ML estimator of π_c of the c^{th} covariate class given by $\widehat{\pi}_c = \frac{e^{\mathbf{x}_c\widehat{\boldsymbol{\beta}}}}{1+e^{\mathbf{x}_c\widehat{\boldsymbol{\beta}}}}$.

The likelihood function in equation (1.11) is that of n independent observations from a Bernoulli distribution. Now we have the case where v_c observations share the same covariate set and therefore a likelihood function can be derived in terms of v_c, y_{c1}, π_c given by (Liu 2007, p. 20)

$$l(\boldsymbol{\beta}) = \prod_{c=1}^{q} \begin{pmatrix} v_c \\ y_{c1} \end{pmatrix} \pi_c^{y_{c1}}(1-\pi_c)^{v_c - y_{c1}} \tag{1.24}$$

and the natural logarithm of the function in equation (1.24) is given by

$$\log l(\boldsymbol{\beta}) = \sum_{c=1}^{q} \left\{ \log \begin{pmatrix} v_c \\ y_{c1} \end{pmatrix} + y_{c1}\log(\pi_c) + (v_c - y_{c1})\log(1-\pi_c) \right\}. \tag{1.25}$$

For a specific covariate pattern of covariate class c the Pearson residual (Hosmer and Lemeshow 2001, p. 145) is given by

$$r(y_{c1}, \widehat{\pi}_c) = \frac{(y_{c1} - v_c\widehat{\pi}_c)}{\sqrt{v_c\widehat{\pi}_c(1-\widehat{\pi}_c)}} \tag{1.26}$$

and the Pearson chi-square test statistic is expressed by

$$\chi^2 = \sum_{c=1}^{q} r(y_{c1}, \widehat{\pi}_c)^2 \tag{1.27}$$

with $q - (m+1)$ degrees of freedom.

The deviance residual for the c^{th} covariate class is given by (Hosmer and Lemeshow 2001, p. 146)

$$d(y_{c1}, \widehat{\pi}_c) = \pm \left\{ 2 \left[y_{c1}\log\left(\frac{y_{c1}}{v_c\widehat{\pi}_c}\right) + (v_c - y_{c1})\log\left(\frac{(v_c - y_{c1})}{v_c(1-\widehat{\pi}_c)}\right) \right] \right\}^{1/2}. \tag{1.28}$$

If the c^{th} covariate class has no successes ($y_{c1} = 0$) it results in $\left(\frac{y_{c1}}{v_c\widehat{\pi}_c}\right) = 0$ then the deviance residual can be calculated by

$$d(y_{c1}, \widehat{\pi}_c) = -\sqrt{2v_c\,|\log(1-\widehat{\pi}_c)|}. \tag{1.29}$$

Similarly, if only successes are observed in the c^{th} covariate class ($y_{c1} = v_c$) it leads to $\left(\frac{(v_c - y_{c1})}{v_c(1-\widehat{\pi}_c)}\right) = 0$ and the deviance residual can be calculated by

$$d(y_{c1}, \widehat{\pi}_c) = \sqrt{2v_c\,|\log(\widehat{\pi}_c)|}. \tag{1.30}$$

From equations (1.28), (1.29) and (1.30) the deviance statistic can be calculated by

$$D = \sum_{c=1}^{q} d(y_{c1}, \widehat{\pi}_c)^2. \tag{1.31}$$

If the model is correct, the deviance test statistic is approximately a chi-square distribution with $q - (m + 1)$ degrees of freedom.

When the first covariate type ($q = n$) is present in a data set, i.e. $v_c = 1$ for $\forall c$, Pearson's statistic reduces to (by using equation (1.15)) as shown in McCullagh and Nelder (1989, p. 121)

$$\begin{aligned}
\chi^2 &= \sum_{c=1}^{q} \frac{(y_{c1} - \widehat{\pi}_c)^2}{\widehat{\pi}_c(1 - \widehat{\pi}_c)} \\
&= \sum_{i=1}^{n} \frac{(y_i - \widehat{\pi}_i)^2}{\widehat{\pi}_i(1 - \widehat{\pi}_i)} \\
&= \sum_{i=1}^{n} \frac{(y_i - 2y_i\widehat{\pi}_i + \widehat{\pi}_i{}^2)}{\widehat{\pi}_i(1 - \widehat{\pi}_i)} \\
&= \sum_{i=1}^{n} \frac{(y_i - 2y_i^2 + y_i^2)}{y_i(1 - y_i)} \\
&= \sum_{i=1}^{n} \frac{(y_i - y_i^2)}{y_i(1 - y_i)} \\
&= \sum_{i=1}^{n} \frac{y_i(1 - y_i)}{y_i(1 - y_i)} \\
&= n.
\end{aligned} \tag{1.32}$$

From equation (1.32) one can note that the Pearson chi-square statistic is reduced to the sample size which is not a very useful test for goodness-of-fit.

Similarly when the first covariate type is present in a data set, the deviance residual from equation (1.28) is reduced to

$$d(y_i, \widehat{\pi}_i) = \pm \left\{ 2 \left[y_i \log\left(\frac{y_i}{\widehat{\pi}_i}\right) + (1 - y_i) \log\left(\frac{(1 - y_i)}{(1 - \widehat{\pi}_i)}\right) \right] \right\}^{1/2} \tag{1.33}$$

since $v_c = 1$, $y_{c1} = y_i$ and all classes are of size 1. The only two possible values for y_i is 0 or 1 which implies that the values for $(1 - y_i)\ln(1 - y_i)$ and $y_i \ln y_i$ will be reduced to 0 for either case. By taking this into consideration, combined with the result from equation (1.15), the deviance test statistic in equation (1.31) can be rewritten, as shown by (Liu 2007, p. 22) , to be

$$
\begin{aligned}
D &= \sum_{c=1}^{q} d(y_{c1}, \widehat{\pi}_c)^2 \\
&= \sum_{i=1}^{n} d(y_i, \widehat{\pi}_i)^2 \qquad (1.34) \\
&= \sum_{i=1}^{n} \left(\left\{ 2 \left[y_i \log\left(\frac{y_i}{\widehat{\pi}_i}\right) + (1-y_i) \log\left(\frac{(1-y_i)}{(1-\widehat{\pi}_i)}\right) \right] \right\}^{1/2} \right)^2 \\
&= 2 \sum_{i=1}^{n} \left[y_i \log\left(\frac{y_i}{\widehat{\pi}_i}\right) + (1-y_i) \log\left(\frac{(1-y_i)}{(1-\widehat{\pi}_i)}\right) \right] \\
&= 2 \sum_{i=1}^{n} \left[y_i \log y_i - y_i \log \widehat{\pi}_i + (1-y_i) \log(1-y_i) - (1-y_i) \log(1-\widehat{\pi}_i) \right] \\
&= 2 \sum_{i=1}^{n} \left[-y_i \log \widehat{\pi}_i - (1-y_i) \log(1-\widehat{\pi}_i) \right] \\
&= -2 \sum_{i=1}^{n} \left[\widehat{\pi}_i \log \widehat{\pi}_i + (1-\widehat{\pi}_i) \log(1-\widehat{\pi}_i) \right]. \qquad (1.35)
\end{aligned}
$$

From equation (1.35) it is seen that the deviance only makes use of the estimated probabilities $\widehat{\pi}_i$ and does not take into consideration the agreement between the observed binary values and their corresponding fitted probabilities (Collett 2003, p. 68). The test statistics represented in equations (1.32) and (1.35) only occurs when sparse data is observed ($q = n$) therefore extreme caution should be applied when interpreting the Pearson chi-square or deviance test statistic when the first covariate type is present in the data set.

When there is a significant difference between the Pearson chi-square and deviance test statistic it could imply that the sampling distribution of the Pearson chi-square and the deviance test statistic are not correctly approximated by the Chi-square distribution with $q - (m + 1)$ degrees of freedom; this could be the result of a sparse or a small data set.

1.10.2 Hosmer-Lemeshow statistic

One advantage of using the Hosmer-Lemeshow statistic as a measure of goodness-of-fit is that this test does not require that a covariate set $\mathbf{x}_i = [x_{i0}, x_{i1}, ..., x_{im}]$ needs to be repeated over more than one observation. The Hosmer-Lemeshow test can be used for the case where each observation has its own unique covariate set ($q = n$), unlike for the Pearson chi-square and deviance test. Therefore the Hosmer-Lemeshow tests can be used on a data set with both continuous and categorical independent variables.

There are two different types of Hosmer-Lemeshow tests available (Liu 2007, p. 24), the one is based on predetermined cut-off values of the estimated probability of a success and is indicated by $\widehat{H}$. To calculate $\widehat{H}$, ten groups are formed by setting the upper interval for the fitted probabilities for each group to $0.1, 0.2, ..., 1$. This may however lead to groups with small and/or unequal sizes which is why this particular method is not often used in practice and therefore will not be discussed any further.

The second statistic, $\widehat{C}$, is more prevalent and will be discussed. This test is based on the percentiles of the estimated probabilities. The first step to determine $\widehat{C}$ is to calculate the estimated probabilities under the assumed logistic regression model derived for the specific data set (as derived in equation (1.7)) of all the observations. The observations are then sorted into ascending order according to the corresponding fitted probabilities. From this ordered list the observations are grouped. The groups can be selected manually such that the groups have equal sizes.

Suppose there is a total of a groups where each group is of size v_a. In each of these groups the observed values for the dependent variable $y_i; i = 1, 2, ..., a$, are added to obtain the observed number of successes, o_i in that group. Similarly, the fitted probabilities in each group are added to obtain the estimated expected number of successes, e_i. In each group the average success probability is the expected number of successes divided by the total number of observations in that group, i.e. $\widehat{\pi}_i = \frac{e_i}{v_a}$.Using this information the Hosmer-Lemeshow test statistic is given by (Hosmer and Lemeshow 2001, p. 148)

$$\widehat{C} = \sum_{i=1}^{a} \frac{(o_i - v_a\widehat{\pi}_i)^2}{v_a\widehat{\pi}_i(1 - \widehat{\pi}_i)}. \tag{1.36}$$

This statistic is an approximate chi-square distribution with $(a - 2)$ degrees of freedom when the fitted model is appropriate. This leads to a formal goodness-of-fit hypothesis test where each observation has its own unique covariate set. This value should be interpreted with caution since the value is greatly influenced by the total number of observations, the number of observations within each group and how the values are split into different groups. Care should especially be taken when interpreting this value when the number of covariate patterns are less than the number of observations ($q < n$), as shown in Bertolini et al. (2000). Any conclusion based on this statistic should only be taken as a guideline on assessing the goodness-of-fit of a logistic regression model.

1.10.3 Model fit statistics

Penalized fit statistics, as discussed in (Allison 2012, p. 22) can be used to informally compare models with different number of covariates against each other; these values can

however not be used to construct a formal hypothesis testing such as the Pearson chi-square, deviance test statistic and Hosmer-Lemeshow statistic.

The most fundamental test of the model fit statistics is the maximised value of the log likelihood function as defined in equation (1.11) multiplied by -2. The value of

$$-2LogL \tag{1.37}$$

is greatly dependent on the number of observations and the number of covariates considered in the data set. A high value for $-2LogL$ is usually a indication of a badly fitted model, but since this value is affected by the data set that is used, it is advised to only use this value as a comparative measure for different models fitted on the same data set. If a data set has more covariates, the value of $-2LogL$ tends to decrease, if the number of observations in the data set increases the value of $-2LogL$ tends to be inflated.

Since the $-2LogL$ value decreases (showing a better fit) as the number of covariates increases, another measure needs to be considered which penalises models with more covariates. The Akaike's information criterion (AIC) places such a penalty on models since it is calculated by

$$AIC = -2LogL + 2k \tag{1.38}$$

where k is the number of covariates plus the intercept term, therefore $k = m + 1$. For each additional covariate introduced in a model the value of $-2LogL$ is thus penalised (increased) by a factor of 2.

Schwarz criterion (SC) is a statistic that even more severely penalises $-2LogL$ for each additional covariate added and is given by

$$SC = -2LogL + k \log n \tag{1.39}$$

Each additional covariate in the model is now penalised by a factor of $k \log n$ where $k = m+1$ and $\log n$ is the natural logarithm of the number of observations used in the model. For example if a model is built based on a sample of 50 observations, each additional covariate introduced in the model will be penalised by a factor of $\log(50) = 3.912$.

1.10.4 Classification tables

The methods and tests mentioned up to this stage involve formal hypothesis testing and informal model fit statistics to assess whether the logistic regression model fits the data adequately. These methods however do not allow a visual representation of the

actual results. When a logistic regression model has been obtained for a specific data set the predicted values can be calculated and compared with the observed values. A classification table (Hosmer and Lemeshow 2001, p. 156) can then be compiled by cross-classifying the observed outcome with the predicted values obtained from the logistic regression model.

A classification table allows one to evaluate to which extent the logistic regression model correctly predicts the group membership of an observation, i.e. whether the model is appropriate and fits the data well. Consider the outcomes (win or a loss) of a team for 10 matches. The predicted outcome for the team depends on the number of hours they have practised, the team's morale, the proportion of wins the coach has and the location of the match. A logistic regression model can be constructed to predict a win, dependent on the above mentioned independent variables. Fictitious predicted and actual outcomes of the 10 games (win=1 and loss=0) for the specific team are summarised in Table 1.6.

Table 1.6: Predicted and actual outcome for team

Match	1	2	3	4	5	6	7	8	9	10
Predicted outcome	1	1	1	1	1	0	0	0	1	0
Actual outcome	1	0	1	1	0	0	1	0	1	0

The predicted values in Table 1.6 are obtained by specifying a cut-off predicted probability value to classify the predicted outcome as either a loss or a win. If the predicted probability is below this cut-off predicted probability value the predicted outcome will be classified as a 0. If the predicted probability exceeds the cut-off value the derived outcome will be equal to 1. The most often used cut-off value is 0.5. By cross-classifying the actual and predicted values on a cut-off value of 0.5 in Table 1.6 the classification table in Table 1.7 is obtained.

Overall the percentage which was correctly classified by the model is 70%. For the individual outcomes the win category is predicted most accurately (80%) whereas the prediction for a loss is not as accurately predicted (60%).

This method allows us to examine a model from the predictive accuracy perspective. With this method it is important to remember that the predictions were obtained from the same data used to fit the model and could therefore give very positive results. For

Table 1.7: Classification table for team results

		Predicted		
		Win=1	Loss=0	Percentage Correct
Observed	Win=1	4	1	80%
	Loss=0	2	3	60%
Overall %				70%

example if a correctly classified percentage of wins is 70%, it suggests a good fit on the face of it. If however it is much more likely that a team will win than lose, 70% correctly classified as a win may be a bad prediction. It should be borne in mind that classification is susceptible to the relative sizes of the two groups and always promotes classification into the group with the larger size.

Classification is not just dependent on the sample size but also on the predictive outcome that was obtained from the model. To illustrate this, consider (Hosmer and Lemeshow, 2001) a data set with 100 patients where the predictive probability to have a disease is $\widehat{\pi} = 0.51$ for all 100 patients. If the cut-off value is 0.5 and considering that the model is appropriate would imply that 51 patients have the disease and 49 does not. Then 51 would fall above the cut-off value and would be correctly classified, whereas 49 out of the 100 would be misclassified. Therefore classification tables should only be utilised as an illustrative measure of the predictive outcome. It is good practice to keep a separate validation sample which can be used to evaluate the predictive accuracy of the model.

1.11 Significance of coefficients

1.11.1 Test Statistics

When considering which of the covariates to include in a model, it is important to consider the following question: does the model that include this specific covariate predict the outcome better than a model that does not include this specific covariate? To answer this question one can compare the output of a model that includes this particular covariate with one that does not. From this (in addition to a few tests mentioned below) one can decide which covariates to include in the model. It is important to remember in the situation when a covariate is categorical and dummy variables were used to represent this covariate, that these dummy variables form one group. Therefore if the categorical covariate is excluded from the model, all the dummy variables which represent this covariate must be excluded. Section 1.11 just focuses on the significance of the estimated coefficients where the significance of a model as a whole (goodness-of-fit) is discussed in Section 1.10.

One way to test the significance of a coefficient representing covariate j is to compute the difference of the deviance (as defined in equation (1.31)) when the covariate is not included in the model and when the covariate is included in the model, shown by (Hosmer and Lemeshow 2001, p. 14)

$$G = D(\text{model without the covariate}) - D(\text{model with the covariate}). \tag{1.40}$$

From equation (1.40) the likelihood ratio statistic can be explained. For any specific data set, the derived ML estimates can be used to set up the current model. The maximised likelihood under this current model can be denoted by $\widehat{L_c}$. The current model can however not be used on its own since it is dependent on the number of observations and covariates used and therefore needs to be compared to a baseline model. The baseline model typically used is one that perfectly fits the data by building a model for which the fitted values match the actual values. This baseline model is a saturated model. The maximised likelihood of this saturated model is denoted by $\widehat{L_s}$. From this the likelihood ratio (LR) is given by (Collett 2003, p. 66)

$$LR = -2\log\frac{\widehat{L_c}}{\widehat{L_s}}. \tag{1.41}$$

Under the null hypothesis that β_j is equal to 0, LR follows a chi-square distribution with degrees of freedom equal to the difference in the number of parameters estimated by the two models. For this test a large sample size n is required.

Before excluding any of the coefficients, the univariate Wald statistic also needs to be considered, it is given by (Hosmer and Lemeshow 2001, p. 37)

$$WALD_j = \frac{\widehat{\beta}_j}{\widehat{SE}(\widehat{\beta}_j)} \tag{1.42}$$

where the standard error of the estimated coefficients $\beta_j; j = 0, 1, ..., m$ is given by

$$\widehat{SE}(\widehat{\beta}_j) = \left[\widehat{Var}(\widehat{\beta}_j)\right]^{1/2}. \tag{1.43}$$

From equation (1.21), the estimated variance of the vector $\widehat{\boldsymbol{\beta}}$ is given by

$$\widehat{Var}(\widehat{\boldsymbol{\beta}}) = \widehat{\mathbf{I}}^{-1}(\widehat{\boldsymbol{\beta}}). \tag{1.44}$$

A formulation of the Hessian matrix (expressed by equations (1.18), (1.19) and (1.20)) can be given by $\widehat{\mathbf{I}}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}^T\mathbf{W}(\widehat{\boldsymbol{\beta}})\mathbf{X}$ where $\mathbf{X}$ is the design matrix and $\mathbf{W}(\widehat{\boldsymbol{\beta}})$ is a $n \times n$ matrix where the diagonal elements are given by $\widehat{\pi}_i(1-\widehat{\pi}_i)$ i.e.

$$\mathbf{W}(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} \widehat{\pi}_1(1-\widehat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \widehat{\pi}_2(1-\widehat{\pi}_2) & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & \widehat{\pi}_n(1-\widehat{\pi}_n) \end{bmatrix}. \tag{1.45}$$

Similarly, the matrix $\mathbf{W}(\boldsymbol{\beta})$ is a $n \times n$ matrix where the diagonal elements are given by $\pi_i(1-\pi_i)$ i.e.

$$\mathbf{W}(\boldsymbol{\beta}) = \begin{bmatrix} \pi_1(1-\pi_1) & 0 & \cdots & 0 \\ 0 & \pi_2(1-\pi_2) & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & \pi_n(1-\pi_n) \end{bmatrix}. \tag{1.46}$$

Under the null hypothesis that a specific coefficient is not significant, the Wald statistic will follow a standard normal distribution.

For a logistic regression model with more than one covariate the Wald statistic can be calculated by (Hosmer and Lemeshow 2001, p. 39)

$$\begin{aligned} WALD &= \widehat{\boldsymbol{\beta}}^T \left[\widehat{Var}(\widehat{\boldsymbol{\beta}})\right]^{-1} \widehat{\boldsymbol{\beta}} \\ &= \widehat{\boldsymbol{\beta}}^T (\mathbf{X}^T \mathbf{W}(\widehat{\boldsymbol{\beta}})\mathbf{X})\widehat{\boldsymbol{\beta}}. \end{aligned} \tag{1.47}$$

The Wald statistic given by equation (1.47) follows a chi-square distribution with $m+1$ degrees of freedom under the null hypothesis that each of the $m+1$ coefficients is equal to 0.

The LR statistic and the Wald statistic can provide guidance as to which covariates significantly contribute to predicting the outcome. The score test can also be used to analyse the significance of the estimated parameters in the model and the interested reader is referred to Hosmer and Lemeshow (2001, p.152). One should however not entirely base the decisions on these tests, an overall assessment of the entire model and the effect of each of the covariates should also be considered.

1.11.2 Confidence interval

For any estimate in statistics an interval in which this estimate falls can be calculated. For logistic regression the endpoints for a $100(1-\alpha)\%$ confidence interval for the coefficient estimate of the j^{th} covariate is given by (Hosmer and Lemeshow 2001, p. 18)

$$\widehat{\beta}_j \pm z_{1-\frac{\alpha}{2}} \widehat{SE}(\widehat{\beta}_j)$$

where $z_{1-\frac{\alpha}{2}}$ is the upper $100(1-\alpha/2)\%$ point from the standard normal distribution and $\widehat{SE}(\widehat{\beta}_j)$ is as defined in equation (1.43).

1.12 Conclusion

Logistic regression predicts the outcome of a dichotomous variable based on a set of covariates. When deriving a logistic regression model it is very important to investigate the type of covariates in your data set. If the covariates are categorical then dummy variables should be introduced into the model, in which case the number of covariate classes present in the data set is more likely to be less than the sample size ($q < n$). For this situation the Pearson chi-square and the deviance test can be used to test the goodness-of-fit of the model. When the covariates used are continuous no dummy variables are required in the logistic regression model and the data set is more likely to be sparse. In this case the Pearson chi-square and deviance test should be applied with caution and it is advisable to rather use the Hosmer-Lemeshow test to evaluate the model's goodness-of-fit.

Chapter 2

Complete and quasi-complete separation and overlap

2.1 Introduction

When one of the independent variables, X, can perfectly classify the observations into the respective groups of the response variable, the likelihood function has no maximum and therefore no finite value can be found for the estimates of $\boldsymbol{\beta}$. If the maximum of the likelihood function does not exist, it follows that the ML estimates also do not exist. This problem is known as monotone likelihood. Three different mutually exclusive and exhaustive classes into which the data from a logistic regression can be classified exists (Albert and Anderson, 1984): complete separation, quasi-complete separation and overlap. Complete and quasi-complete separation imply that only an infinite or a zero ML estimate could be obtained for the odds ratio which rarely can be assumed to be true in practice. Although perfect prediction is aimed for in practice, if the sample size is small and perfect prediction occurs, it is probably as a result of random variation and not that of a true infinite or zero odds ratio.

To illustrate complete separation, quasi-complete separation and overlapping data, practical examples are considered as illustrated in Allison et al. (2004). For the examples considered in Section 2.1, 2.2 and 2.3 let $x_i \leq 0$ be indicated by $x_i = 0$ and $x_i > 0$ be indicated by $x_i = 1$. In Chapter 2 one independent variable will be used to predict the outcome of the dependent variable, the coefficient representing this one independent variable is indicated by β.

2.2 Complete separation

From Albert and Anderson (1984), complete separation occurs when there exists a $(m+1)$ vector of coefficients $\boldsymbol{\beta}$ such that when $\mathbf{x}_i\boldsymbol{\beta} < 0$ the outcome is $y_i = 0$ and when $\mathbf{x}_i\boldsymbol{\beta} > 0$ the outcome is $y_i = 1$. Whenever a linear function of the independent variable X_i can perfectly predict the response variable, complete separation occurs.

Table 2.1 contains observations from a completely separated model where there is only one explanatory variable X and the response variable Y takes on the value $y_i = 0$ whenever $x_i < 0$ and $y_i = 1$ whenever $x_i \geq 0$ for $i = 1, 2, ..., 10$.

Table 2.1: Example of complete separation

i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x_i	-5	-4	-3	-2	-1	1	2	3	2	5
y_i	0	0	0	0	0	1	1	1	1	1

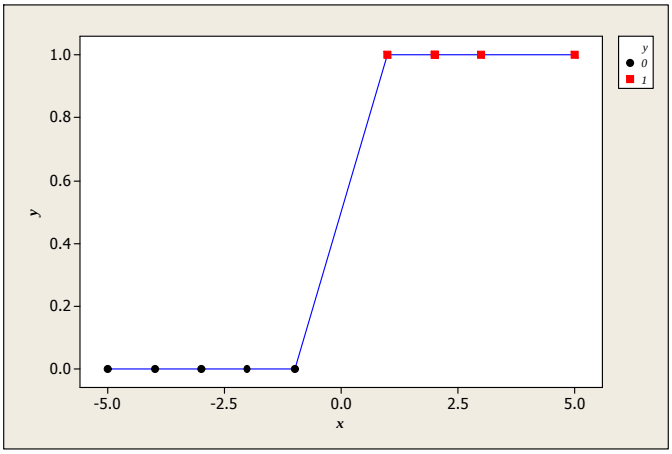
The observations in Table 2.1 can be summarised by a 2×2 contingency table.

Table 2.2: Two-way table for complete separation

		y	
		0	**1**
x	**0**	5	0
	1	0	5

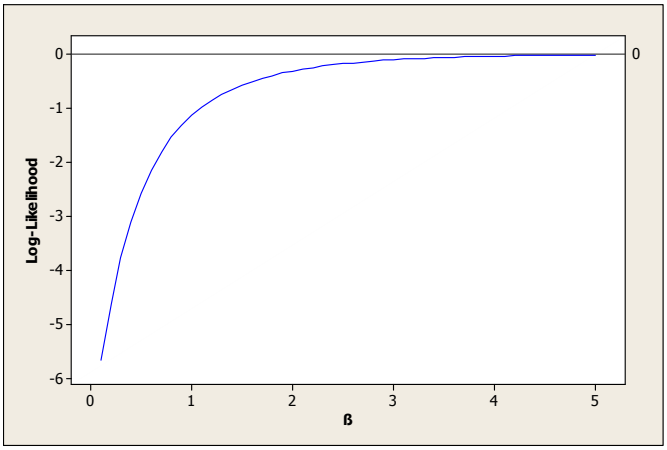
Whenever the two off-diagonal cells in a 2×2 contingency table has frequencies of 0, it is an indication of complete separation. To illustrate complete separation in X and Y consider the scatterplot of the observed values in Table 2.1 given in Figure 2.1. From Figure 2.1 it is noted that there is a horizontal jump from $y_i = 0$ to $y_i = 1$.

Figure 2.1: Scatter plot of x and y under complete separation



By substituting the values of the contingency table into equation (1.16) the ML estimator is $\widehat{\beta} = \log \frac{(5)(5)}{(0)(0)}$, which does not exist. As the value of the estimator β increases, the log likelihood does not reach a maximum value. Even though the log likelihood is bounded by the value of 0, no significant values for β can be estimated, this scenario is illustrated by Figure 2.2 (Allison et al., 2004). Since the likelihood function is flat the diagonal elements in the variance matrix of the coefficient in equation (1.44) will be infinite in size, which leads to an infinite standard error of the coefficient representing the covariate.

Figure 2.2: The log-likelihood function as a function of β for complete separation



2.3 Quasi-complete separation

One also gets the situation, as explained by Albert and Anderson (1984), where there exists some $(m+1)$ vector of coefficients $\boldsymbol{\beta}$ such that $y_i = 0$ when $\mathbf{x}_i\boldsymbol{\beta} \leq 0$ and $y_i = 1$ when $\mathbf{x}_i\boldsymbol{\beta} \geq 0$ and for at least one category of the outcome variable the equality holds. This is known as quasi-complete separation.

An example of quasi-complete separation (Allison et al., 2004) is represented in Table 2.3, where there is once again only one explanatory variable X and the response variable Y assumes the value of $y_i = 0$ whenever $x_i < 0$ and $y_i = 1$ whenever $x_i > 0$. There is however one value for X, $x_i = 0$, for which both $y_i = 0$ and $y_i = 1$ is observed.

Table 2.3: Example of quasi-complete separation

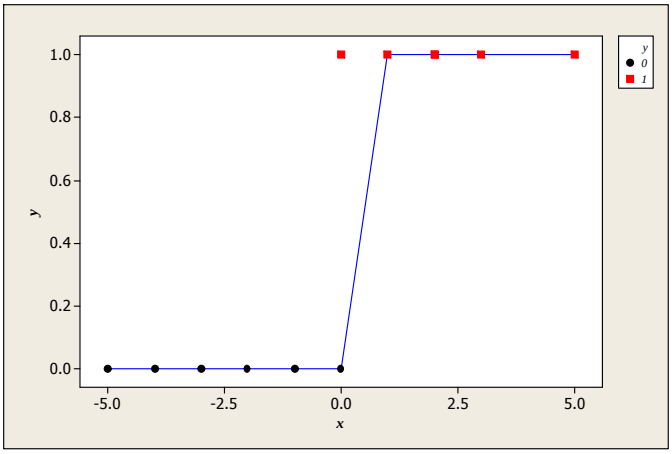
i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
x_i	-5	-4	-3	-2	-1	0	0	1	2	3	2	5
y_i	0	0	0	0	0	0	1	1	1	1	1	1

Table 2.4: Two-way table for quasi-complete separation

		y	
		0	**1**
x	**0**	6	1
	1	0	5

By summarising the values in Table 2.3 in a 2×2 contingency table, Table 2.4 is obtained. If either one of the off-diagonal cells in a 2×2 contingency table contains a value of 0 it is an indication of quasi-complete separation. The observations considered in Table 2.3 can be illustrated by a scatter plot shown in Figure 2.3. For the case of quasi-complete separation there is at least one overlapping observation when $x_i = 0$ as seen in Figure 2.3.

Figure 2.3: Scatter plot of x and y under quasi-complete separation

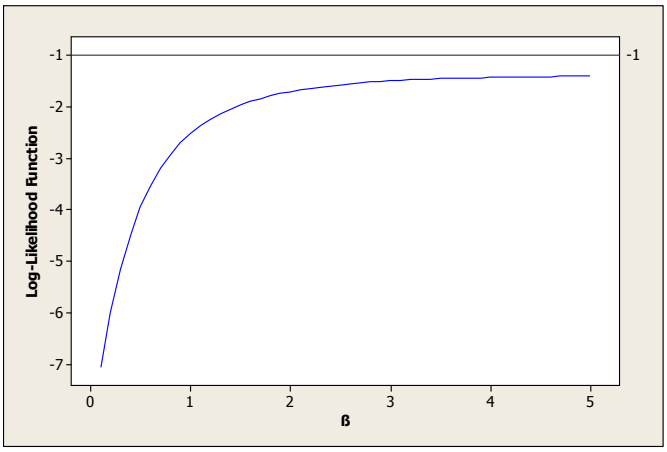


By substituting the values of the contingency table into equation (1.16) the ML estimator is found by $\widehat{\beta} = \log \frac{(6)(5)}{(0)(1)}$, which does not exist. As the value of the estimator β increases, the log likelihood does not reach a maximum value. Even though the log likelihood is bound by some value smaller than 0, in this case bounded by -1 as seen in Figure 2.4 , no finite value for β can be estimated. For quasi-complete separation the standard error in equation (1.43) will be infinite in size. This situation is far more common in practice than complete separation.

2.4 Overlap

When neither complete nor quasi-complete separation has occurred in the data it can be assumed that an unique finite solution for the ML estimates exist, this is known as overlap (Silvapulle, 1981).

Figure 2.4: The log-likelihood function as a function of β under quasi-complete separation



To determine the estimates of $\boldsymbol{\beta}$ by using the Newton-Rhapson algorithm (1.22) the inverse of the Hessian matrix (1.18) must exist, therefore a simple criterion to determine if the data overlaps is to test if the Hessian matrix is positive definite. This will occur when the dependent variable Y cannot be perfectly predicted by the independent variable X. Consider the following data set with $n = 10$.

Table 2.5: Example of overlapping data

i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x_i	-5	2	-3	-2	-1	0	0	1	2	3
y_i	0	1	1	1	0	0	1	0	1	1

As observed in Table 2.5 multiple values (0 and 1) exist for y_i if the corresponding value of x_i is above or below any specific value. This data set can be summarised in a 2×2 contingency given in Table 2.6 and illustrated by a scatter plot given in Figure 2.5. From Figure 2.5 it can be observed that the observations fluctuate between $y_i = 0$ and $y_i = 1$, there is no single horizontal jump for a certain value of x_i.

Table 2.6: Two-way table for overlapping data

		y	
		0	**1**
x	**0**	3	3
	1	1	3

From equation (1.16) the ML estimator can be obtained to be $\widehat{\beta} = \log \frac{(3)(3)}{(3)(1)} = 1.0986$ as indicated in Figure 2.6.

Figure 2.5: Scatter plot of x and y for overlapping data

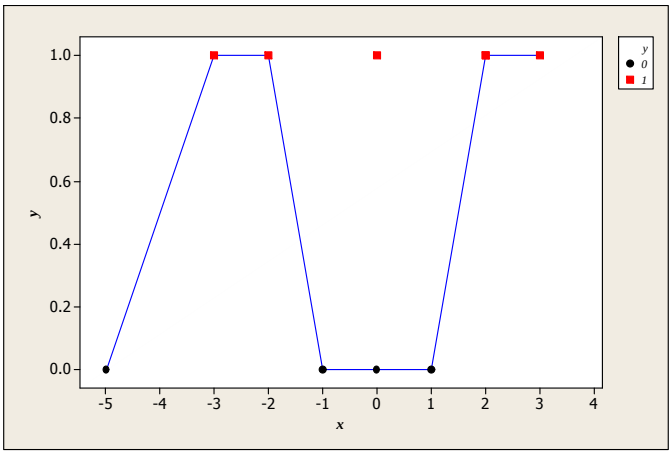
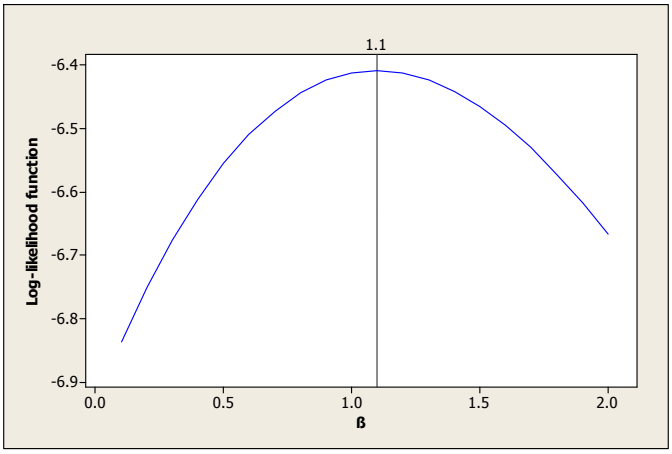


Figure 2.6: The log-likelihood function as a function of β for overlapping data



2.5 Identifying complete or quasi-complete separation

The process of identifying whether a data set exhibits complete or quasi-complete separation is very important to be able to report significant estimators for the logistic regression model. Different statistical software programs have been tested with the data set represented in Tables 2.1 and 2.3 by Allison et al. (2004) to asses which of the programs give applicable warnings of possible separation in the model. Only MINITAB, SAS GENMOD, SAS LOGISTIC, SAS CATMOD and SPSS will be considered to compare with results for 2013 since these are the software packages readily available at the University of Pretoria. Table 2.7 indicates the warning which is given by the program, whether the applicable program gives false convergence, reports unreliable estimates and finally if the program reports LR statistics.

The study done in Allison et al. (2004) is based on software that were available in 2002.

Table 2.7: Computer packages under complete and quasi-complete separation

Results obtained by Allison et al. (2004)								
Program	Warning Messages		False Conv.		Report Est.		LR Stats	
	Comp.	Quasi.	Comp.	Quasi.	Comp.	Quasi.	Comp.	Quasi.
MINITAB					*	*		
SAS GENMOD		A	*		*	*	*	*
SAS LOGISTIC	C	C			*	*		
SAS CATMOD	A	A	*	*	*	*		
SPSS	C			*		*		
Results obtained for 2013								
Program	Warning Messages		False Conv.		Report Est.		LR Stats	
	Comp.	Quasi.	Comp.	Quasi.	Comp.	Quasi.	Comp.	Quasi.
MINITAB	D, E, H	D, E, H			*	*		
SAS GENMOD		J	*	*	*	*	*	*
SAS LOGISTIC	K, F, E	L, F, E			*	*		
SAS CATMOD	G	G	*	*	*	*		
SPSS	H,I	H,I			*	*		

A: Ambiguous Warning
C: Clear Warning
D: Algorithm did not converge
E: Results may be unreliable
F: MLE does not exist
G: Parameter estimates are infinite
H: Max number of iterations reached
I: No solution could be found
J: Hessian matrix not positive definite
K: Complete Separation
L: Quasi-complete Separation

Comparing the results in Table 2.7 from 2002 to 2013, one observes that some of the programs have improved whilst some stayed the same.

One of the computer packages that improved a great deal is MINITAB, even though it still reports estimates: from no warning in 2002 to giving a warning that even if the maximum number of iterations have been reached, no convergence was reached and the results obtained could be unreliable in 2013.

The SAS GENMOD procedure still reports false converge, values for the estimates and LR statistics. The warning message in 2002 was seen as ambiguous, the message as seen in 2013 reports that the Hessian matrix is not positive definite. This warning only occurs when quasi-complete separation is present in the data set; for complete separation no warning is given. If an user has a good understanding of how the coefficients are estimated, he or she would know that if the Hessian is not positive definite it will not be possible to obtain reliable estimates.

A clear warning is still given by SAS LOGISTIC in 2013; it reports estimates but gives a clear warning that the estimates are infinite and that the ML estimates does not exist. The SAS LOGISTIC procedure is the only program from the list in Table 2.7 that gives a warning "Complete separation" or "Quasi-complete separation" if either one of those scenarios occur within the data set.

The third SAS procedure to be considered is the SAS CATMOD procedure, this procedure still gives false convergence and reports estimates, but from an ambiguous warning in 2002 it now reports that the parameter estimates are infinite, which gives the user of the program a good indication to apply caution when interpreting the estimated parameters.

The final program considered is SPSS. In 2002 this program gave a clear warning but only for complete separation and it reported false convergence and estimates for quasi-complete separation. In 2013 it still reports estimates but now gives a warning that no solution could be found even though the maximum number of iterations have been reached for both complete and quasi-complete separation.

2.6 Conclusion

When constructing a logistic regression model it is not always the case that the coefficient estimates will exist or, if the coefficient estimates do exist, will be reliable. Therefore it is essential to test whether complete or quasi-complete separation is present in a data set when the dependent variable is dichotomous, especially when the sample size is small. When complete or quasi-complete separation is present in the data set it is imperative not to continue with the general approach to computing a logistic regression model, but to follow a different approach. A few of these different approaches to deal with complete or quasi-complete separation are mentioned and investigated in Chapter 3.

Chapter 3

Methods used to deal with separation

3.1 Introduction

The problem of complete separation and quasi-complete separation in binomial data modelling was first documented by Day and Kerridge (1967). As shown in Silvapulle(1981), when the data overlaps in a logistic regression model the ML estimates exist and are unique, if separation in the data occurs the ML estimates does not exist. A formal criterion to distinguish between complete separation and quasi-complete separation was proposed by Albert and Anderson (1984) as discussed in Chapter 2. It is of utmost importance to know if either complete separation and quasi-complete separation is present within the data set in order to interpret the estimated regression coefficients correctly.

Numerous methods to detect complete separation or quasi-complete separation have been developed over the years. A linear-programming method was proposed by Albert and Anderson (1984) which indicates when the data does not overlap. This linear programming method was extended to a mixed linear programming algorithm by Santer and Duffy (1986) which determines for each data set whether there is complete separation, quasi-complete separation or if the data overlap. Christmann and Rousseeuw (2001) developed a method not just to determine whether the data overlaps, but also to determine to which degree the data overlap in the data set.

3.2 Different methods

3.2.1 Changing the model

As shown in Section 2.5, to identify whether the ML estimates exist is easily managed with most statistical software packages. The next step is to obtain a solution when dealing with non-overlapping data. One often used method in practice is to omit some of the covariates which is the cause of complete separation or quasi-complete separation. This method is however not recommended, since by omitting the covariates that hold a strong relationship to the occurrence of interest is similar to deliberately introducing specification bias, which in turn leads to biased parameter estimates.

Another approach which changes the original covariates included in the model is by adding artificial data across the various patterns to "fill the gaps" which was caused by separation. Various techniques are available to impute observations. This method was proposed by Clogg *et al.* (1991) and as with omitting covariates, changes the covariates that hold a strong relationship with the outcome. This method is not recommended as shown in Heinze and Schemper (2002).

If the problematic covariate is a nominal variable with more than 2 categories, for example single, married and divorced, the categories can be grouped together. For example single and divorced can be considered as one group and married can be seen as the other group. This method usually solves quasi-complete separation but does not solve complete separation. It should also be borne in mind that if the groups are merged and the observed values are more condensed, it will not be possible to allocate the observations to the original groups, i.e. to restore the data set back to its original dimension.

3.2.2 Working with the likelihood function

Methods based on the likelihood function are preferred to methods which change the original data set. One option is to use exact logistic regression (Section 3.3) as proposed by Cox and Snell (1989) but is computing-intensive. Another solution to dealing with non-overlapping data is by penalising the likelihood function with Jeffreys prior as proposed by Firth (1993). Firth's method is explained in Section 3.4 and has many advantages as shown in Heinze (2006). Firth's method of penalising the likelihood function was extended by Heinze and Schemper (2002) to calculate the confidence intervals based on this method instead of using Wald intervals which are unreliable when dealing with non-overlapping data. If multicollinearity is present in the data set, a second penalising term can be introduced as shown in Gao and Chen (2007). In 2003, Rousseeuw and Christmann introduced the concept of hidden logistic regression discussed in Section 3.5. This involves obtaining a maximum estimated likelihood estimator which always exists and is robust to separation and outliers.

Since altering the original data set is not recommended and has been shown to give undesirable results in Heinze and Schemper (2002), this method will not be discussed further. Exact logistic regression, Firth's method and hidden logistic regression will be extended on in Section 3.3 to 3.5. All three methods will be applied to practical examples in part II where they will be compared to each other in different scenarios.

3.2.3 Other methods

Alternative methods which merit mentioning but will not be elaborated on in this dissertation includes an approach using Markov chain Monte Carlo methods in which one would work in a Bayesian paradigm. Caution is advised when using these methods since uninformative priors can lead to no convergence of the coefficient estimates and the use of informative priors may result to misleading results as discussed in Abrahantes and Aerts (2012) and Allison et al. (2004).

When dealing with separation one should also be aware of other properties within a data set which could lead to misleading results. Multicollinearity can be problematic if not dealt with correctly. Gao and Chen (2007) discuss this phenomenon and introduce a second penalising term to address this problem. Another statistical problem which can be observed in practice is outliers. An outlier-robust method is discussed in Rousseeuw and Christmann (2003). These are just a few different methods to deal with certain aspects which can be observed within a data set and one should always exercise caution when using different methods and interpreting results.

3.3 Exact logistic regression

Exact logistic regression was first proposed by Cox and Snell (1989). This method is also known as exact conditional inference. As stated by King and Ryan (2002), the general idea of exact logistic regression is "to base inferences on exact permutational distributions of the sufficient statistics that correspond to the regression parameters of interest, conditional on fixing the sufficient statistic of the remaining parameters at their observed values".

3.3.1 The ML estimates

To obtain the conditional likelihood function to estimate a specific parameter, consider first the likelihood function given in equation (1.11) and the logit function defined in

equation (1.2); then the unconditional likelihood function is given by (Collett 2003, p. 308)

$$\begin{aligned} l(\boldsymbol{\beta}) &= \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \\ &= \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{-y_i}(1-\pi_i)^{1} \\ &= \prod_{i=1}^{n} \left(\frac{\pi_i}{1-\pi_i}\right)^{y_i}(1-\pi_i) \\ &= \prod_{i=1}^{n} e^{\mathbf{x}_i\boldsymbol{\beta}y_i}(1+e^{\mathbf{x}_i\boldsymbol{\beta}})^{-1} \\ &\quad \prod_{i=1}^{n} e^{\mathbf{x}_i\boldsymbol{\beta}y_i}(1+e^{\theta_i})^{-1} \\ &= \frac{e^{\sum_{i=1}^{n}\sum_{j=0}^{m}\beta_j x_{ij} y_i}}{\prod_{i=1}^{n}(1+e^{\theta_i})} \end{aligned} \tag{3.1}$$

Let $t_j = \sum_{i=1}^{n} x_{ij}y_i$; $j = 0, 1, ..., m$ therefore for $y_i = 1$, t_j is the sum of the values of the j^{th} explanatory variable and expression (3.1) can be rewritten as

$$l(\boldsymbol{\beta}) = \frac{e^{\sum_{j=0}^{m}\beta_j t_j}}{\prod_{i=1}^{n}(1+e^{\theta_i})}. \tag{3.2}$$

The t_j values contain all the information about the $\beta_j^{'}s$ for the binary observations, therefore the $t_j^{'}s$ are the sufficient statistics for the coefficients. A full discussion on sufficient statistics and its role in conditional inference is available in Cox and Hinkly (1979).

Suppose we are only interested in one specific coefficient, say β_m, in which case the rest of the parameters $\beta_0, ..., \beta_{m-1}$ are regarded as nuisance parameters and can be expressed by the vector $\boldsymbol{\beta}_1 = [\beta_0, ..., \beta_{m-1}]$. The column vector of coefficients can then be composed as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \beta_m)^T$. The conditional likelihood function of β_m is found by using the unconditional likelihood function $l(\boldsymbol{\beta})$ and eliminating the effect of $\boldsymbol{\beta}_1$.

If $t_0, t_1, ..., t_m$ are the observed sufficient statistics from the random variables $T_0, T_1, ..., T_m$, the conditional likelihood (as discussed in Collett (2003)) for β_m given $T_0 = t_0, ..., T_m = t_m$ is given by

$$\begin{aligned} l_c(\beta_m) &= P(T_m = t_m | T_0 = t_0, T_1 = t_1, ..., T_{m-1} = t_{m-1}) \\ &= \frac{P(T_0 = t_0, T_1 = t_1, ..., T_m = t_m)}{P(T_0 = t_0, T_1 = t_1, ..., T_{m-1} = t_{m-1})}. \end{aligned} \tag{3.3}$$

The numerator of equation (3.3), $P(T_0 = t_0, T_1 = t_1, ..., T_m = t_m)$, is simply the probability of the observed data which is the sum of the values of $l(\boldsymbol{\beta})$ over all possible sets of binary data that lead to $t_0, t_1, ..., t_m$. Therefore the probability of the observed data can be calculated by (Collett 2003, p. 309)

$$P(T_0 = t_0, T_1 = t_1, ..., T_m = t_m) = \frac{c(t_0, ..., t_m) e^{\sum_{j=0}^{m} \beta_j t_j}}{\prod_{i=1}^{n} (1 + e^{\theta_i})} \tag{3.4}$$

where the number of distinct binary sequences that give the specified values of $t_0, t_1, ..., t_m$ can be denoted by $c(t_0, ..., t_m)$. The joint distribution of $P(T_0 = t_0, T_1 = t_1, ..., T_{m-1} = t_{m-1})$ can be found from equation (3.4) to be

$$P(T_0 = t_0, T_1 = t_1, ..., T_{m-1} = t_{m-1}) = \frac{\sum_u c(t_0, ..., t_{m-1}, u) e^{\beta_m u + \sum_{j=0}^{m-1} \beta_j t_j}}{\prod_{i=1}^{n} (1 + e^{\theta_i})}. \tag{3.5}$$

where the summation term in the numerator is over all values for u for which $c(t_0, ..., t_{m-1}, u) \geq 1$. Therefore the conditional likelihood function of β_m from equation (3.3), (3.4) and (3.5) can be expressed as (Collett 2003, p. 309)

$$\begin{aligned} l_c(\beta_m) &= \frac{c(t_0, ..., t_m) e^{\sum_{j=0}^{m} \beta_j t_j}}{\sum_u c(t_0, ..., t_{m-1}, u) e^{\beta_m u + \sum_{j=0}^{m-1} \beta_j t_j}} \\ &= \frac{c(t_0, ..., t_m) e^{\beta_m t_m + \sum_{j=0}^{m-1} \beta_j t_j}}{\sum_u c(t_0, ..., t_{m-1}, u) e^{\beta_m u + \sum_{j=0}^{m-1} \beta_j t_j}} \\ &= \frac{c(t_0, ..., t_m) e^{\beta_m t_m}}{\sum_u c(t_0, ..., t_{m-1}, u) e^{\beta_m u}}. \end{aligned} \tag{3.6}$$

From equation (3.6) it can be noted that the function does not depend on $\beta_1, ..., \beta_{m-1}$, therefore if this function is maximised an exact parameter estimate for β_m, also known as the conditional ML estimate will be obtained.

For a two-sided hypothesis test the null and alternative hypothesis is given by $H_0 : \beta_m = 0$ and $H_1 : \beta_m \neq 0$ respectively. The appropriate p-value for this hypothesis test, according to Collett (2003, p. 312), can be obtained by adding the probabilities of all the values of the sufficient statistics of coefficient β_m which have a probability smaller than or equal to that of the observed value. This is similar to a two-sided p-value for Fisher's exact test. This test is generally known as the conditional exact test and a discussion on this can be found in Collett (2003). A full discussion on exact logistic regression can be found in Cox and Snell (1989), Mehta and Patel (1995), Collett (2003) and Hosmer and Lemeshow (2001).

For exact logistic regression the estimates of $\beta_0, ..., \beta_m$ are obtainable even if empty cells are observed in the 2×2 contingency table expressed in Table 1.2. Even though it is possible to determine parameter estimates in the presence of complete or quasi-complete separation, the drawback of this method is that it can get computationally difficult as the number of covariates and samples increase. As stated by Hosmer and Lemeshow (2001, p. 337), when using exact logistic regression, caution should be taken when analysing the Pearson chi-square and deviance test statistics since these tests are based on a large sample assumption and this is not usually the case when using exact methods. It is recommended that one should rather use visual methods like a classification table to investigate the agreement between the observed and the predicted value.

Another disadvantage of exact logistic regression method is that the sufficient statistics for the different parameter estimates need to be summed over discrete patterns of covariate values. As stated by Zorn (2005), exact logistic regression will lead to unreliable estimates when using a combination of both continuous and categorical covariates. Therefore if exact logistic regression is applied to a data set which contains both categorical and continuous covariates, caution should be applied when interpreting the coefficient estimates.

3.4 Firth's Model

Another way to approach the problem of separation is to reduce the bias that is found in the ML estimates. The ML estimates are unbiased with asymptotic variance equal to the inverse of the Fisher information matrix which is given by $\mathbf{I}(\boldsymbol{\beta}) = \phi^{-1}\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta})\mathbf{X}$. The dispersion parameter is given by $\phi = 1$ (for a binomial distribution), $\mathbf{X}$ is the model matrix and $\mathbf{W}$ is a $n \times n$ matrix where $\mathbf{W} = diag(\pi_i(1-\pi_i))$ as shown in equation(1.46). For a large sample size McCullagh and Nelder(1989, p. 119) showed that

$$E(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) = O(n^{-1}) \tag{3.7}$$

and expressed in terms of the covariate matrix

$$cov(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta})\mathbf{X})^{-1}\left\{1 + O(n^{-1})\right\}. \tag{3.8}$$

Then, by Firth (1993), the asymptotic bias of a single ML estimate $\widehat{\beta}$ of parameter β for an m dimensional model can be written as

$$b(\beta) = \frac{b_1(\beta)}{n} + \frac{b_2(\beta)}{n^2} + ...\tag{3.9}$$

The aim of Firth's method is to reduce the bias of the parameter estimates, specifically by removing the $O(n^{-1})$ term. Two methods already exist where the term $\frac{b_1(\beta)}{n}$ is removed from the asymptotic bias, namely the jackknife method (Quenouille, 1949, 1956) and by simply substituting $\widehat{\beta}$ for the unknown β in $\frac{b_1(\beta)}{n}$. The bias corrective parameter estimate, $\widehat{\beta}_{BC}$, is then given by

$$\widehat{\beta}_{BC} = \widehat{\beta} - \frac{b_1(\widehat{\beta})}{n}. \tag{3.10}$$

Firth's method is different from the procedures mentioned above in that the parameter is not corrected after it is estimated, but a corrective procedure is applied to the ML estimate (score function) before the parameter estimate is calculated.

3.4.1 The model

To illustrate this method consider the modelling of binomial observations where each observation y_i have a true probability of success equal to π_i i.e. $\frac{y_i}{n_i}$. The binomial likelihood function can be expressed by (Collett 2003, p. 66)

$$\begin{aligned} p(y_i|\pi_i) &= \binom{n_i}{y_i} \pi_i^{y_i}(1-\pi_i)^{n_i-y_i} \\ &= \binom{n_i}{y_i} \pi_i^{y_i}(1-\pi_i)^{-y_i}(1-\pi_i)^{n_i} \\ &= \binom{n_i}{y_i} \left(\frac{\pi_i}{1-\pi_i}\right)^{y_i} (1-\pi_i)^{n_i}. \end{aligned} \tag{3.11}$$

Taking the exponential of the log of this expression becomes

$$\begin{aligned} & \exp\left\{\log\left(\binom{n_i}{y_i}\left(\frac{\pi_i}{1-\pi_i}\right)^{y_i}(1-\pi_i)^{n_i}\right)\right\} \\ = & \exp\left[\log\binom{n_i}{y_i} + y_i\log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i\log(1-\pi_i)\right] \\ = & \exp\left[\frac{y_i\log\left(\frac{\pi_i}{1-\pi_i}\right) - [-n_i\log(1-\pi_i)]}{1} + \log\binom{n_i}{y_i}\right]. \end{aligned} \tag{3.12}$$

The form of equation (3.12) is the same as that of n independent observations $y_1, ..., y_n$ with an exponential density expressed by (McCullagh and Nelder 1989, p. 28)

$$p(y_i|\theta_i, \phi) = \exp\left[\frac{(y_i\theta_i - d(\theta_i))}{a_i(\phi)} + c(y_i, \phi)\right], i = 1, 2, ..., n. \tag{3.13}$$

Comparing equation (3.13) to equation (3.12) the functions in equation (3.12) can be identified as $\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$ (i.e. $\theta_i = \mathbf{x}_i\boldsymbol{\beta}$ from equation(1.2)), $d(\theta_i) = -n_i\log(1-\pi_i)$, $a_i(\phi) = 1$, $c(y_i, \phi) = \log\binom{n_i}{y_i}$ where $\log\left(\frac{\pi_i}{1-\pi_i}\right) = logit(\pi_i)$ is the canonical link function for the binomial distribution and β_j is the canonical parameter. The logit function is the canonical link function (McCullagh and Nelder 1989, p. 32) since it is also a function of $\mu_i = E(Y_i) = n_i\pi_i$ which can be seen by

$$logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \log\left(\frac{n_i\pi_i}{n_i(1-\pi_i)}\right) = \log\left(\frac{\mu_i}{n_i-\mu_i}\right) = \eta_i(\mu_i). \tag{3.14}$$

Since $\eta(\mu_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{j=0}^{m} x_{ij}\beta_j$ then $\eta(\mu_i) = \theta_i$ as shown in equation (1.2) and $t_j = \sum_{i=1}^{n} x_{ij}y_i$ is the sufficient statistic for β_j where $j = 0, 1, ..., m$.

For a single parameter β_j the ML estimate is derived as a solution of the derivative of the log likelihood $l(\boldsymbol{\beta})$ or score function set equal to 0, i.e.

$$\frac{\partial \log l(\boldsymbol{\beta})}{\partial \beta_j} = U(\beta_j) = 0. \tag{3.15}$$

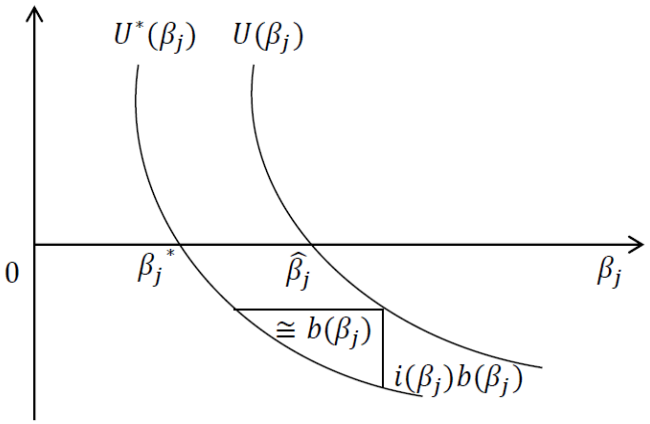
From equation (1.12) and the definition of sufficient statistic t_j, the log likelihood function for a single coefficient β_j can be expressed by $\log l(\beta_j) = \sum_{i=1}^{n} \left\{ y_i x_{ij} \beta_j - \log(1 + e^{(x_{ij}\beta_j)}) \right\} = t_j \beta_j - K(\beta_j)$ where $K(\beta_j) = \sum_{i=1}^{n} \log(1 + e^{(x_{ij}\beta_j)})$. The score function (as obtained in equation(1.13)) can then be expressed by

$$U(\beta_j) = \frac{\partial \log l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} (x_{ij} y_i - x_{ij} \pi_i) = t_j - K'(\beta_j) \tag{3.16}$$

which implies that the sufficient statistic t_j only affects the location of $U(\beta_j)$ and not the gradient.

It can be shown (Firth, 1993) that the bias of the estimate $\widehat{\beta}_j$ comes from the unbiasedness of the score function, $E\left[U(\beta_j)\right] = 0$ at the correct value of β_j and due to the bend in the score function shown in Figure 3.1, $U''(\beta_j) = K'''(\beta_j) \neq 0$.

Figure 3.1: Modified score function



If the score function $U(\beta_j)$ was a linear function of β_j then $E\left[\widehat{\beta}_j\right] = \beta_j$, but since this is clearly not the case as shown in Figure 3.1 a bias is induced in $\widehat{\beta}_j$. The idea behind Firth's model is to implement a small bias in the score function to reduce the bias in $\widehat{\beta}_j$. A suitable modification to $U(\beta_j)$ is given by

$$U^*(\beta_j) = U(\beta_j) - i(\beta_j) b(\beta_j). \tag{3.17}$$

This originates from the triangle geometry shown in Figure (3.1). If the estimator $\widehat{\beta}_j$ has a positive bias of $b(\beta_j)$, the score function can be shifted downward by degree $i(\beta_j)b(\beta_j)$ where the gradient of $U(\beta_j)$ is given by $U'(\beta_j) = -i(\beta_j)$. The modified estimate β_j^* can then be calculated by setting the modified score function equal to 0, i.e. $U^*(\beta_j) = 0$.

3.4.2 The ML estimates

To obtain the modified score function explained above, one can penalise the log likelihood function given in equation(1.11) by Jeffreys (1946) invariant prior (for exponential family models). The Jeffreys invariant prior density is given by $|\mathbf{I}(\boldsymbol{\beta})|^{1/2} = \left|\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta})\mathbf{X}\right|^{1/2}$ where the vector of unknown parameters are given by $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_m)^T$ and the Fisher information matrix is given by $\mathbf{I}(\boldsymbol{\beta}) = \phi^{-1}\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta})\mathbf{X}; \phi = 1$. The penalised likelihood function for Firth's model is thus (Firth, 1992b)

$$l^*(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) \times |\mathbf{I}(\boldsymbol{\beta})|^{(1/2)}. \tag{3.18}$$

Taking the natural logarithm of equation (3.18) yields

$$\log l^*(\boldsymbol{\beta}) = \log l(\boldsymbol{\beta}) + (1/2)\log|\mathbf{I}(\boldsymbol{\beta})|. \tag{3.19}$$

To find the maximum, the partial derivative of β_j $(j = 0, 1, ..., m)$ in equation (3.19) is taken and set equal to 0 (Firth, 1992a)

$$\begin{aligned} U^*(\beta_j) &= U(\beta_j) + \left(\frac{1}{2}\right)\frac{\partial}{\partial\beta_j}\log\left|\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta})\mathbf{X}\right| \\ &= U(\beta_j) + (1/2)trace\left\{\left(\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta})\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{W}(\beta_j)\mathbf{X}\right)\right\}; j = 0, 1, ..., m \end{aligned} \tag{3.20}$$

where $U^*(\beta_j) = \frac{\partial \log l^*(\boldsymbol{\beta})}{\partial\beta_j}$, $U(\beta_j) = \frac{\partial \log l(\boldsymbol{\beta})}{\partial\beta_j}$ as given in equation (1.13) and $\mathbf{W}(\beta_j) = \frac{\partial\mathbf{W}(\boldsymbol{\beta})}{\partial\beta_j}$.

The value of $\mathbf{W}(\beta_j)$ is calculated by

$$
\begin{aligned}
& \mathbf{W}(\beta_j) \\
= \; & diag\left(\frac{\partial(\pi_i(1-\pi_i))}{\partial\beta_j}\right) \qquad (3.21) \\
= \; & diag\left(\frac{\partial\{(1+e^{x_{ij}\beta_j})^{-1}(1+e^{-x_{ij}\beta_j})^{-1}\}}{\partial\beta_j}\right) \\
= \; & diag\left\{-(1+e^{x_{ij}\beta_j})^{-2}\left(x_{ij}e^{x_{ij}\beta_j}\right)(1+e^{-x_{ij}\beta_j})^{-1}+(1+e^{-x_{ij}\beta_j})^{-2}\left(x_{ij}e^{-x_{ij}\beta_j}\right)(1+e^{x_{ij}\beta_j})^{-1}\right\} \\
= \; & diag\left\{x_{ij}\left[\frac{-e^{x_{ij}\beta_j}}{(1+e^{x_{ij}\beta_j})^2(1+e^{-x_{ij}\beta_j})}+\frac{e^{-x_{ij}\beta_j}}{(1+e^{x_{ij}\beta_j})(1+e^{-x_{ij}\beta_j})^2}\right]\right\} \\
= \; & diag\left\{x_{ij}\left[\frac{-e^{x_{ij}\beta_j}(1+e^{-x_{ij}\beta_j})+\left(e^{-x_{ij}\beta_j}\right)(1+e^{x_{ij}\beta_j})}{(1+e^{x_{ij}\beta_j})^2(1+e^{-x_{ij}\beta_j})^2}\right]\right\} \\
= \; & diag\left\{x_{ij}\left[\frac{-e^{x_{ij}\beta_j}-1+e^{-x_{ij}\beta_j}+1)}{(1+e^{x_{ij}\beta_j})^2(1+e^{-x_{ij}\beta_j})^2}\right]\right\} \\
= \; & diag\left\{x_{ij}\left[\frac{-(1+e^{x_{ij}\beta_j})+(1+e^{-x_{ij}\beta_j})}{(1+e^{x_{ij}\beta_j})^2(1+e^{-x_{ij}\beta_j})^2}\right]\right\} \\
= \; & diag\left\{x_{ij}\left[\frac{-(1+e^{x_{ij}\beta_j})}{(1+e^{x_{ij}\beta_j})^2(1+e^{-x_{ij}\beta_j})^2}+\frac{(1+e^{-x_{ij}\beta_j})}{(1+e^{x_{ij}\beta_j})^2(1+e^{-x_{ij}\beta_j})^2}\right]\right\} \\
= \; & diag\left\{x_{ij}\left[\frac{-1}{(1+e^{x_{ij}\beta_j})(1+e^{-x_{ij}\beta_j})^2}+\frac{1}{(1+e^{x_{ij}\beta_j})^2(1+e^{-x_{ij}\beta_j})}\right]\right\} \\
= \; & diag\left\{x_{ij}\left[-(\pi_i)^2(1-\pi_i)+(\pi_i)(1-\pi_i)^2\right]\right\} \\
= \; & diag\left\{x_{ij}(\pi_i)(1-\pi_i)\left[-\pi_i+1-\pi_i\right]\right\} \\
= \; & diag\left\{x_{ij}(\pi_i)(1-\pi_i)\left[1-2\pi_i\right]\right\}. \qquad (3.22)
\end{aligned}
$$

Let $\mathbf{V}(\beta_j)$ be an $n \times n$ diagonal matrix were the i^{th} diagonal element is represented by $x_{ij}(1-2\pi_i); i = 1, 2, ..., n$, i.e. $\mathbf{V}(\beta_j) = diag\left(x_{ij}(1-2\pi_i)\right)$, then $\mathbf{W}(\beta_j) = \mathbf{W}(\boldsymbol{\beta})\mathbf{V}(\beta_j)$ from equation (1.46) and (Firth, 1992(a)).

Therefore equation (3.20) can be simplified to

$$
\begin{aligned}
U^*(\beta_j) &= U(\beta_j) + (1/2)trace\left\{\left(\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta})\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{W}(\beta_j)\mathbf{X}\right)\right\} \\
&= U(\beta_j) + (1/2)trace\left\{\left(\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta})\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta})\mathbf{V}(\beta_j)\mathbf{X}\right)\right\} \\
&= U(\beta_j) + (1/2)trace\left\{\left[\mathbf{W}(\boldsymbol{\beta})\mathbf{X}\left(\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta})\mathbf{X}\right)^{-1}\mathbf{X}^T\right]\mathbf{V}(\beta_j)\right\} \\
&= U(\beta_j) + \frac{1}{2}\sum_{i=1}^{n} h_i x_{ij}(1-2\pi_i) \qquad (3.23)
\end{aligned}
$$

where $j = 0, 1, ..., m$ and h_i is the i^{th} diagonal element of the hat matrix $\mathbf{H} = \mathbf{W}(\boldsymbol{\beta})\mathbf{X}\left(\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta})\mathbf{X}\right)^{-1}\mathbf{X}^T$. This can be done since $tr(\mathbf{AB}) = tr(\mathbf{BA})$ and the trace of a matrix is the sum of its diagonal elements .

Substituting the value of $U(\beta_j)$ given in equation (1.13) in equation (3.23) the modified score function for $j = 0, 1, ..., m$ is (Firth, 1992a)

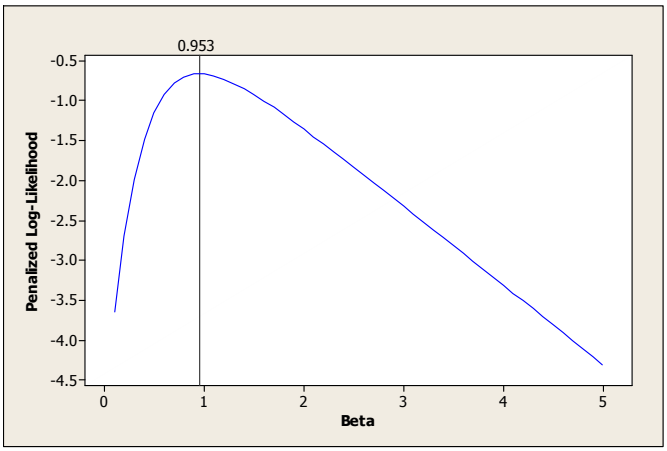
$$\begin{aligned} U^*(\beta_j) &= U(\beta_j) + \frac{1}{2}\sum_{i=1}^{n} h_i x_{ij}(1 - 2\pi_i) \\ &= \sum_{i=1}^{n}(y_i - \pi_i)x_{ij} + \frac{1}{2}\sum_{i=1}^{n} h_i x_{ij}(1 - 2\pi_i) \\ &= \sum_{i=1}^{n}\left(y_i + \left(\frac{h_i}{2}\right) - h_i\pi_i - \pi_i\right)x_{ij} \overset{set}{=} 0. \end{aligned} \tag{3.24}$$

If the value of h_i is known then the value of equation (3.24) can be obtained by using standard statistical or mathematical software. If the value of h_i is unknown (which is more likely the case since it is a function of the unknown parameter β_j) then this value should be calculated by an iterative procedure as explained in Firth (1992a).

3.4.3 Method applied to complete and quasi-complete separation

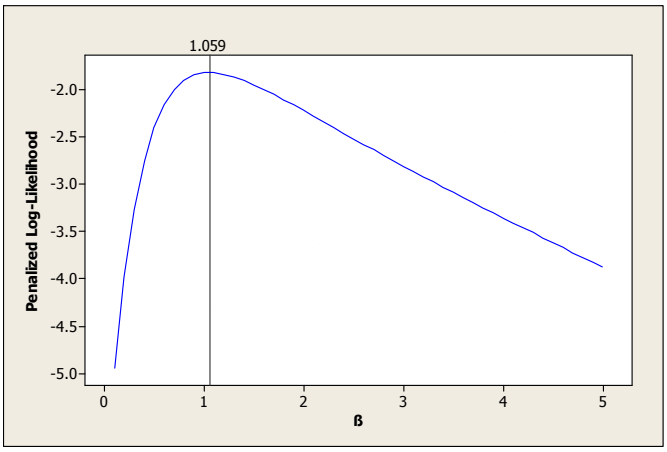
To illustrate the effect of the penalised likelihood function as described above we will revisit the completely separated data considered in Section 2.1 and the quasi-complete separated data in Section 2.2. In both these instances the value for the log-likelihood and coefficient estimate increased to positive infinity and did not converge to one value; by applying Firth's model to the completely separated values the following figure can be obtained (SAS program available in Appendix E)

Figure 3.2: The penalized log-likelihood function as a function of β for complete separation



From Figure 3.2 it is noted that the estimate of β now converges to a value of 0.953. For the quasi-complete separation case the log-likelihood as a function of the estimate can be illustrated by Figure 3.3.

Figure 3.3: The penalized log-likelihood function as a function of β for quasi separation



Firth's approach corresponds to adding 0.5 in each cell of a 2×2 contingency table which gives a solution to the empty cell problem. Another way to view this is "splitting each original observation i into two new observations having response values Y_i and $1 - Y_i$ with iteratively updated weights $1 + h_i/2$ and $h_i/2$, respectively" as stated by Heinze and Schemper (2002). By following this approach one can ensure that all the parameter estimates for the covariates exist whether the data exhibits complete separation or quasi-complete separation. An important property of Firth's method is that it yields consistent parameter estimates which can be confirmed by the fact that given an overlapping data set where the ML estimates exists, the estimates obtained under Firth's method will converge to the ML estimates for a logistic regression model, as the sample size increases.

3.5 Hidden logistic regression

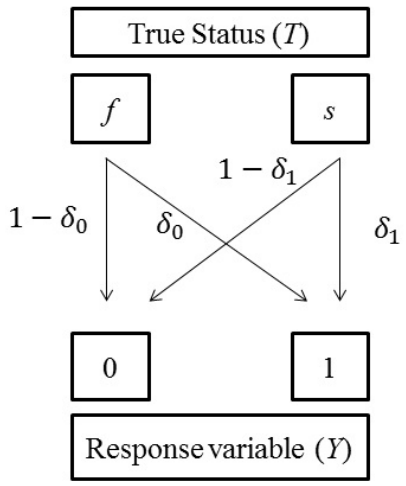
The last model to be discussed is the hidden logistic regression model. This has been proposed by Ekholm and Palmgren (1982) and by Copas (1988) to introduce the idea of a hidden layer in a neural network. The same idea was further developed by Rousseeuw and Christmann (2003) but with a slightly different approach. For this method it is assumed that the derived logistic regression model always has an intercept term. The main idea behind this model from Ekholm and Palmgren (1982), Copas (1988) and Rousseeuw and Christmann (2003) is that the true responses are unobservable. The response values that are needed to fit the model are therefore not the true values but are strongly related to the unobservable true responses. As illustration, consider a medical test which indicates

whether a patient has a specific disease or not. Even if this test shows that a patient does have the disease, this test is not 100% without fault and there could be a small chance that the patient should have been classified in the other group.

3.5.1 The model

According to the approach taken by Rousseeuw and Christmann (2003) one assumes that the true status (T) of the patient has one of two possible outcomes, (s) which indicates a success e.g. the patient does have the disease and (f) which indicates a failure i.e. the patient does not have the specified disease. If the true status is a success then the probability of observing a 1 for the response variable is $P(Y = 1|T = s) = \delta_1$, if the true status is a success but the patient is classified into group 0, then the misclassification probability is given by $P(Y = 0|T = s) = 1 - \delta_1$. If the true status of the patient is that he or she does not have the disease then the probability that the patient is correctly classified to group 0 is $P(Y = 0|T = f) = 1 - \delta_0$ and the probability that the patient is classified into group 1 is $P(Y = 1|T = f) = \delta_0$. This probability structure is represented in Figure 3.4 (Rousseeuw and Christmann, 2003).

Figure 3.4: Probability structure of the true status against the observable response



The values of $1-\delta_0$ and δ_1 indicate the probability of observing the true response for $T = f$ and $T = s$ respectively. The probability of observing the true response is assumed to be higher than observing the incorrect true value. Therefore the probabilities of correctly classifying a observation can be assumed to be greater than 0.5 i.e. $0 < \delta_0 < 0.5 < \delta_1 < 1$. If for example a patient does have the disease and the probability of classifying the patient correctly is 95% i.e. $P(Y = 1|T = s) = \delta_1 = 0.95$ then the probability of misclassifying the patient as not having the disease is $P(Y = 0|T = s) = 1 - \delta_1 = 0.05$.

3.5.2 The ML method

Since the true response is the unobservable value that needs to be estimated, an ML estimate needs to be calculated for T. For this case as illustrated by Rousseeuw and Christmann (2003) in Figure 3.4, the estimate for T only has one of two possibilities (since it is a binary variable) and for each outcome there is a probability that it is correctly classified or not. If the true response is a failure (f) then the likelihood of observing $Y = 0$ is greater than observing $Y = 1$, likewise if the true response is a success (s) the likelihood of observing $Y = 1$ is greater than observing $Y = 0$. From this argument the ML estimator for T, given the response $(Y = y)$, can be expressed by

$$\begin{aligned} \widehat{T}(Y &= 0) = f \\ \widehat{T}(Y &= 1) = s. \end{aligned} \tag{3.25}$$

If the observed value is assigned to group 1 i.e. $Y = 1$ then one of two possibilities could have occurred, the true response is a success and was correctly classified or the true response is a failure and was misclassified. Therefore the probability of an observed value being placed into group 1 conditional on the ML estimator for T can be given by

$$P(Y = 1|\widehat{T}) = \begin{matrix} \delta_0 & \text{if} & y = 0 \\ \delta_1 & \text{if} & y = 1 \end{matrix} \tag{3.26}$$

where y is the observed value of Y. To express equation (3.26) in a single line equation, let $\widetilde{Y}$ denote equation (3.26) then

$$\widetilde{Y} = \delta_0 + (\delta_1 - \delta_0)Y = (1 - Y)\delta_0 + Y\delta_1. \tag{3.27}$$

The value for $\widetilde{Y}$ is a weighted average of δ_0 and δ_1 with weights of $(1 - Y)$ and Y respectively.

Since logistic regression is based on n observations from a Bernoulli distribution, the observed value for Y for the i^{th} sample can be expressed by

$$\widetilde{y_i} = (1 - y_i)\delta_0 + y_i\delta_1 \tag{3.28}$$

where $\widetilde{y_i}$ is the pseudo-observation for each individual. This pseudo-observations is a deterministic result of the observed value y_i.

The resulting estimated likelihood function of the pseudo-observations $\widetilde{y_i}$ is given by

$$l(\boldsymbol{\beta}|\widetilde{y_1}, ..., \widetilde{y_n}) = \prod_{i=1}^{n} \pi_i^{\widetilde{y_i}}(1 - \pi_i)^{1-\widetilde{y_i}}. \tag{3.29}$$

Since the true likelihood function depends on the unobservable values $t_1, ..., t_n$ the likelihood function in equation (3.29) is only an estimation. If $\delta_0 = 0$ and $\delta_1 = 1$ the true likelihood is known, therefore, since $\widetilde{y}_i$ is just an estimation of this value the value for $\widetilde{y}_i$ will strictly lie between 0 and 1. To obtain the ML estimator for the model one can take the natural logarithm of equation (3.29) given by

$$\log l(\boldsymbol{\beta}|\widetilde{y_1}, ..., \widetilde{y_n}) = \sum_{i=1}^{n} (\widetilde{y_i} \log \pi_i + (1 - \widetilde{y_i}) \log(1 - \pi_i)) \tag{3.30}$$

and set the derivative of equation (3.30) with respect to $\beta_j; j = 1, ..., m$ equal to 0 i.e.

$$\begin{aligned} \frac{\log l(\boldsymbol{\beta}|\widetilde{y_1}, ..., \widetilde{y_n})}{\partial \beta_j} &= \sum_{i=1}^{n} \widetilde{y_i} x_{ij} - \sum_{i=1}^{n} \frac{e^{x_{ij}\beta_j}(x_{ij})}{1 + e^{x_{ij}\beta_j}} \\ &= \sum_{i=1}^{n} \widetilde{y_i} x_{ij} - \sum_{i=1}^{n} \pi_i x_{ij} \\ &= \sum_{i=1}^{n} (\widetilde{y_i} - \pi_i) x_{ij} \overset{set}{=} 0. \end{aligned} \tag{3.31}$$

When only the partial derivative with respect to β_0 is considered the derivative of the log-likelihood function set equal to 0 is given by

$$\begin{aligned} \frac{\log l(\boldsymbol{\beta}|\widetilde{y_1}, ..., \widetilde{y_n})}{\partial \beta_0} &= \sum_{i=1}^{n} \widetilde{y_i} - \sum_{i=1}^{n} \frac{e^{\beta_0}}{1 + e^{\beta_0}} \\ &= \sum_{i=1}^{n} \widetilde{y_i} - \sum_{i=1}^{n} \pi_i \overset{set}{=} 0. \end{aligned} \tag{3.32}$$

This leads to the result that the sum of the pseudo-observations is equivalent to the sum of the estimated probabilities expressed by

$$\sum_{i=1}^{n} \widetilde{y_i} = \sum_{i=1}^{n} \widehat{\pi}_i. \tag{3.33}$$

Equations (3.29), (3.30), (3.31) and (3.33) are all equivalent to equation (1.11), (1.12), (1.13) and (1.15) respectively where $y_i = \widetilde{y_i}$.

3.5.3 Determining the values for δ_0 and $\delta_1$be

Since the true values for a type I and type II error are rarely known in practice, in most cases the values for δ_0 and δ_1 must be calculated from theoretical values. One approach for different values of δ_0 and δ_1 is explained by Copas (1988) and is based on a fixed

value. Since δ_0 is the probability of an observation being misclassified into group 1 when the actual outcome is group 0, it should be a small value. δ_1 is the probability of an observation being correctly classified into group 1 and therefore this is expected to be a high value. From this information let

$$\begin{aligned} \delta_0 &= \gamma \\ \delta_1 &= 1-\gamma \end{aligned} \tag{3.34}$$

with $\gamma \geq 0$ and the value for γ typically ranging from 0.01 to 0.02. As shown in Copas (1988) desirable results are obtained when $\gamma = 0.01$ whereas a value of $\gamma = 0.05$ is a too high. This method is preferable when simplicity is required.

The method developed by Copas (1988) is most effective when the data set is symmetric, i.e. when the dependent variable has an equal number of 0 and 1's. When an asymmetric data set is dealt with, the approach discussed by Rousseeuw and Christmann (2003) in Section 3.5.3 can be considered. For further discussion of the method done by Rousseeuw and Christmann (2003) assume π_i is restricted to one value for all $i = 1, 2, ..., n$ namely π. It is more appropriate to consider the marginal distribution of y_i from which the estimate, $\widehat{\pi}$ can be obtained by using the number of 0's and 1's observed.

To obtain the estimated probability of observing a success, one can argue that $\widehat{\pi}$ is the average of the pseudo observations in equation (3.33)

$$\begin{aligned} \sum_{i=1}^{n} \widehat{\pi}_i &= \sum_{i=1}^{n} \widetilde{y}_i \\ \text{i.e.} \quad \sum_{i=1}^{n} \widehat{\pi} &= \sum_{i=1}^{n} \widetilde{y}_i \\ \text{i.e.} \quad n\widehat{\pi} &= \sum_{i=1}^{n} \widetilde{y}_i \\ \text{i.e.} \quad \widehat{\pi} &= \frac{1}{n} \sum_{i=1}^{n} \widetilde{y}_i. \end{aligned} \tag{3.35}$$

By combining the results of equations (3.28), (3.33) and (3.35) the estimated probability can be expressed (Rousseeuw and Christmann, 2003) in terms of δ_0 and δ_1.

$$\begin{aligned} \widehat{\pi} &= \frac{1}{n}\sum_{i=1}^{n}\widetilde{y}_i = (1-\widehat{\pi})\delta_0 + \widehat{\pi}\delta_1 \\ \text{i.e.} \quad \widehat{\pi} - \widehat{\pi}\delta_1 &= \delta_0 - \widehat{\pi}\delta_0 \\ \text{i.e.} \quad \frac{1-\delta_1}{\delta_1 - \widehat{\pi}} &= \frac{\delta_0}{\widehat{\pi} - \delta_0} \end{aligned} \tag{3.36}$$

The two ratios given in equation (3.36) are both equal to a small positive number, say δ, since to it reasonable to assume

$$\delta_0 < \widehat{\pi} < \delta_1. \tag{3.37}$$

Therefore from equation (3.36) the value for δ_0 and δ_1 can be expressed in terms of $\widehat{\pi}$ and δ by

$$\begin{aligned} \delta &= \frac{1-\delta_1}{\delta_1 - \widehat{\pi}} \\ \delta(\delta_1 - \widehat{\pi}) &= 1 - \delta_1 \\ \delta\delta_1 + \delta_1 &= 1 + \delta\widehat{\pi} \\ \delta_1(\delta + 1) &= 1 + \delta\widehat{\pi} \\ \delta_1 &= \frac{1 + \delta\widehat{\pi}}{\delta + 1} \end{aligned} \tag{3.38}$$

and

$$\begin{aligned} \delta &= \frac{\delta_0}{\widehat{\pi} - \delta_0} \\ \delta(\widehat{\pi} - \delta_0) &= \delta_0 \\ \delta\widehat{\pi} + \delta\delta_0 &= \delta_0 \\ \delta\delta_0 - \delta_0 &= -\delta\widehat{\pi} \\ \delta_0 &= \frac{-\delta\widehat{\pi}}{\delta - 1} \\ \delta_0 &= \frac{\delta\widehat{\pi}}{\delta + 1}. \end{aligned} \tag{3.39}$$

Since for this model the extreme values of 0 and 1 are not applicable to the estimated probabilities, the value for $\widehat{\pi}$ cannot simply be the average of the observations, $\overline{\pi} = \frac{1}{n}\sum_{i=1}^{n} y_i$. There has to be a constraint to ensure a value which is strictly less than 1 and strictly more than 0. One possible bound on $\widehat{\pi}$ is given by

$$\widehat{\pi} = \max(\delta, \min(1-\delta, \overline{\pi})) \tag{3.40}$$

which ensures $0 < \widehat{\pi} < 1$.

The ratios in equations (3.38) and (3.39) correspond with the constraint given in equation (3.37) and the fact that $0 < \widehat{\pi} < 1$. This can be verified by

$$\delta_1 = \frac{1+\delta\widehat{\pi}}{\delta+1} > \frac{\widehat{\pi}+\delta\widehat{\pi}}{\delta+1} = \frac{\widehat{\pi}(1+\delta)}{\delta+1} = \widehat{\pi} \tag{3.41}$$

and

$$\delta_0 = \frac{\delta\widehat{\pi}}{\delta+1} < \frac{\widehat{\pi}+\delta\widehat{\pi}}{\delta+1} = \frac{\widehat{\pi}(1+\delta)}{\delta+1} = \widehat{\pi}. \tag{3.42}$$

The two misclassification probabilities of this model is $P(Y=0|T=s) = 1-\delta_1$ and $P(Y=1|T=f) = \delta_0$. These two values will always be smaller than the small positive number δ as shown by

$$\delta_0 = \frac{\delta\widehat{\pi}}{\delta+1} < \frac{\delta}{\delta+1} < \delta \tag{3.43}$$

and

$$1-\delta_1 = \frac{1+\delta-1-\delta\widehat{\pi}}{\delta+1} = \frac{(1-\widehat{\pi})\delta}{1+\delta} < \frac{\delta}{1+\delta} < \delta. \tag{3.44}$$

The commonly used value for δ is 0.01 as advocated by Rousseeuw and Christmann (2003). Then if a balanced data set is observed the estimated probability is given by $\widehat{\pi} = 0.5$ which implies $\delta_0 = 1-\delta_1$ from equations (3.38) and (3.39). This yields the same results as for a symmetric data set discussed by Copas(1988) and shown by equation (3.34).

If one considers a asymmetric data set with say 20 observations and 19 observed $y_i = 1$, with $\delta = 0.01$ the estimated probability from equation (3.40) is given by

$$\begin{aligned}
\widehat{\pi} &= \max\left(0.01, \min\left(1-0.01, \frac{19}{20}\right)\right) \\
&= \max(0.01, \min(0.99, 0.95)) \\
&= \max(0.01, 0.95) \\
&= 0.95.
\end{aligned}$$

Therefore the probability of a misclassification, calculated by equation (3.39), is

$$\delta_0 = \frac{\delta\widehat{\pi}}{\delta+1} = \frac{0.01 \times 0.95}{1.01} = 0.0094$$

and the probability of an observation to be correctly classified into group 1, from equation (3.38), is

$$\delta_1 = \frac{1 + \delta\widehat{\pi}}{\delta + 1} = \frac{1 + (0.01 \times 0.95)}{1.01} = 0.9995.$$

For the asymmetric case this approach gives less biased estimates than for the case where δ_0 and δ_1 are set to a fixed value.

This method has many advantages; it is robust against separation and against outliers. Since the issue of outliers will not be examined further the interested reader is referred to Rousseeuw and Christmann (2003) where a full discussion with examples and simulations are given.

3.6 Conclusion

Many different solutions for complete and quasi-complete separation exist in practice. Each method has its own benefits and disadvantages, which should be investigated before applying each method. For the methods which were considered exact logistic regression performs well with categorical covariates but caution should be applied when using this method to a sparse data set. Firth's method and the hidden logistic regression method can be applied to a data set which contains both continuous and categorical covariates. Each method behaves differently with different covariate types and sample sizes; this will be investigated using practical applications in part II.

Part II

Practical Application

Chapter 4

Overview/ Outline of Part II

In part I the theory of logistic regression is summarised and the problem of complete and quasi-complete separation was identified in Chapter 2 of part I. Three different methods to use when either complete or quasi-complete separation is present were discussed in Chapter 3 of part I. The three different methods mentioned can however not always be applied to any given data set to predict a binary outcome. Keeping this in mind Chapter 5 will investigate a small sample of twenty observations which exhibits complete separation. To study different types of independent variables the first example will include only continuous covariates, the second both continuous and categorical covariates and the final example of Chapter 5 will investigate the case where only categorical covariates are used to predict a binary outcome. In Chapter 6 a large sample with only categorical covariates will be investigated.

For each of the different combinations of covariate types mentioned, the general logistic regression model will be investigated to confirm if complete separation is present. From SAS, PROC LOGISTIC, it is confirmed that all the examples of Chapter 5 exhibit complete separation and in Chapter 6 exhibit quasi-complete separation, but for the case where the computer packages mentioned in Section 2.5 are not available a few exploratory methods can be considered. First a scatter plot of the categorical covariates (which is obtained with MINITAB) will be introduced from which one can visually confirm if there is a split between the groups or not. Secondly by considering numerical analysis it can be examined whether the coefficient estimates converges and if the standard error of the coefficient estimates tends to infinity as the number of iterations increase or not, obtained by both SAS and MINITAB. Once the coefficient estimates are obtained it can be confirmed with the Wald and LR test statistic (obtained from SAS) if $\widehat{\boldsymbol{\beta}}$ is significantly different from 0 or not.

All the covariates which are considered for each model are significant in predicting the dichotomous outcome of that example. All interactions between the covariates were also originally included in the model. The final models given in equations (5.1), (5.2) and

(5.3) are obtained from only choosing the significant covariates to predict the outcome according to forward stepwise logistic regression in SPSS.

For each of the different independent variable types mentioned (continuous and categorical) three methods will be applied, namely; exact logistic regression, Firth's method and finally hidden logistic regression. Once parameter estimates are obtained for each of these methods one needs to test if the coefficient estimates are significant or not. For exact logistic regression and Firth's method the coefficient estimates with the appropriate significance level are calculated with SAS. The two-sided conditional exact test is used to test the significance of the coefficients obtained under exact logistic regression whereas Wald's test statistic is used for the coefficients obtained under Firth's method. For hidden logistic regression the R program code given in Appendix G is applied to obtain the coefficient estimates and the appropriate significance levels since there is no procedure available in SAS as of yet to calculate these values. The Wald test statistic is also used to calculate the significance of the coefficient estimates.

After testing the significance of the individual coefficient estimates, the overall validity of the model needs to be examined. Some of the goodness-of-fit measures mentioned in Section 1.10 will be considered namely the Pearson chi-square, deviance, Hosmer-Lemeshow tests and finally the classification table. For the classification tables, the rows represent the observed values and the columns represent the predicted value for the specified model. All the examples will make use of only two covariates to compare the different approaches on the same number of observations. Therefore none of the model fit statistics discussed in Section 1.10.3 will be necessary, as these tests are used to compare models with different number of covariates to each other. To calculate the formal goodness-of-fit test statistics a SAS, PROC IML code is available in Appendix F and H for the classification tables and Pearson chi-square, deviance and Hosmer-Lemeshow tests respectively. This has been done since the mentioned goodness-of-fit test statistics are not available for exact or hidden logistic regression in SAS. From the goodness-of-fit statistics it can be concluded if the method considered is a good choice for that specific combination of covariates.

For the small sample case one should keep in mind that the Wald, LR, Pearson chi-square and deviance test statistics all depend on the assumption of an approximate chi-square distribution and could therefore give misleading results. When the statistics give conflicting results it could be a sign that the large-sample approximation is unreliable. Also, the Pearson chi-square and deviance test statistic are most reliable when considering a data set with only categorical observations. The numerical values for the Pearson chi-square and the deviance test statistic should be close to each other in value as stated in Collett (2003, p. 87): "Large differences between the two statistics can be taken as an

indication that the chi-squared approximation to the distribution of the deviance or the X^2-statistic is not adequate."

When considering the Hosmer-Lemeshow test, as explained by Xie et al. (2008), it has been shown that when the number of groups are few this test statistic almost always shows that the model is adequate and therefore when possible only consider the test statistic for a large number of groups.

Chapter 5

Complete separation (small sample)

5.1 Introduction

Complete and quasi-complete separation is most likely to occur in small data sets which typically occur in medical sciences as discussed in Heinze (2006), in economic indices as shown in Beggs et al. (1981) and even in marketing as indicated in Chapman (1984). The most common occurrence of complete or quasi-complete separation is found in biostatistics (medical sciences) where the application of a clinical trial usually demands a dichotomous outcome (if a patient survived or died, if a treatment worked or did not work, etc.) and since most medical test are typically expensive, a clinical trial will most likely consist of a small sample.

To compare the different approaches mentioned in Chapter 3, three different data sets will be considered. The three data sets will be identical in size ($n = 20$) and all of them will be balanced, i.e. $y_i = 1$ for 10 observations and $y_i = 0$ for the remaining 10 observations. Each of the three examples exhibit complete separation and will have two covariates to predict a dichotomous outcome.

The first clinical trial from Philippeos et. al (2009) is used to predict whether a patient is HIV positive or HIV negative based on two continuous covariates (NMR-based metabonomics).

The second example from Haberman (1973) is based on a sample which consists of a combination of a categorical covariate (age group of a patient) and a continuous covariate (number of positive axillary nodes detected) to predict if a patient survived or died within 5 years of receiving a breast cancer operation .

Finally the third example is a sample acquired from Kaggle (2012); this data set predicts whether a passenger who was Section the RMS Titanic survived or died based on two categorical covariates, the class of the passenger (first class, second class or third class) and the sex of the passenger (female or male).

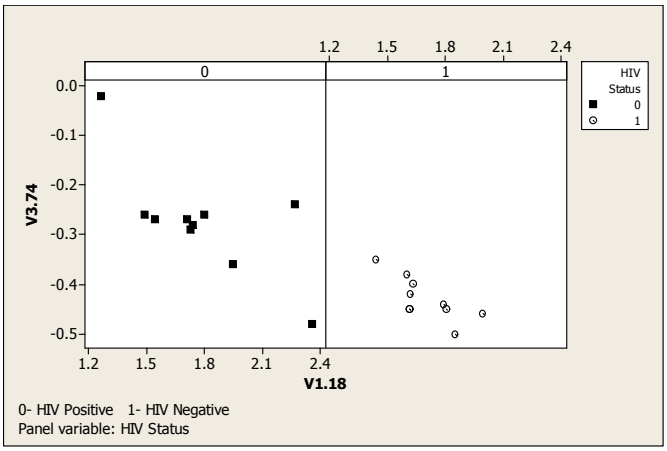
5.2 Continuous covariates

5.2.1 HIV status example

Consider the data set: HIV Status (Appendix A) which was supplied by Prof. D. Meyer. This data set consists of high-resolution proton NMR (^{1}H-NMR) spectroscopic profiling of biological fluids; these metabonomics have contributed to the identification and study of human disease. In Philippeos, Steffens and Meyer (2009) these metabonomics are specifically applied to identify whether a patient is HIV negative or HIV positive.

This experiment involves 20 patients of whom 10 is HIV negative (HIV$^-, y_i = 1$) and 10 of the patients are HIV positive and currently on anti-retroviral therapy (HIV$^+, y_i = 0$). From the initial data set, 88 covariates where obtained from the NMR instrument from which only 2 were significant, $x_{V3.74}$ and $x_{V1.18}$. Both the covariates $x_{V3.74}$ and $x_{V1.18}$ are continuous and complete separation is found within this data set which can also be confirmed by the scatterplot in Figure 5.1. $x_{V3.74}$ and $x_{V1.18}$ are shown in Figure 5.1 and the groups are indicated for $y_i = 1$ and $y_i = 0$.

Figure 5.1: Scatter plot of $x_{V3.74}$ vs. $x_{V1.18}$ grouped according to the observed value of y.



Since this data set consists of only continuous covariates, it is a sparse data set which is confirmed by the fact that there are only two observations that share the same covariate pattern therefore there are 19 unique covariate patterns i.e. $q = 19$.

5.2.2 General logistic regression model

Since this data set consists of continuous covariates, no dummy variables are required to formulate a logistic regression model. The equation for a logistic regression model applied to the HIV status data set is given by equation (5.1).

$$\text{logit}(\pi_i) = \beta_0 + \beta_{v3.74} x_{V3.74} + \beta_{v1.18} x_{V1.18} \tag{5.1}$$

The effect of the interaction between $x_{V3.74}$ and $x_{V1.18}$ is not significant and therefore not included in the model. From the SAS, PROC LOGISTIC, procedure the coefficient estimates for equation (5.1) after 8 iterations are found as shown by the first column of Table 5.1.

Table 5.1: Logistic regression coefficient estimates for HIV status

	After 8 iterations		After 25 iterations		After 100 iterations	
Variable	$\widehat{\boldsymbol{\beta}}$	**SE of** $\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\beta}}$	**SE of** $\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\beta}}$	**SE of** $\widehat{\boldsymbol{\beta}}$
Intercept	10.6243	70.4261	45.6032	184766	96.0844	761609
$x_{V3.74}$	-160.8	272.3	-611.284	731383	-1068.22	3662281
$x_{V1.18}$	-39.7412	74.3653	-153.135	180033	-269.342	559114

From the output it is noticed that the standard error for all coefficients estimates are considerably greater than the corresponding coefficient estimates. This is the first indication that the coefficient estimates should be evaluated with care. Even though estimated values for the coefficients are obtained, there is a warning given by the program that the values haven't converged and that the validity of the model is questionable. To test the validity of the coefficient estimates after 8 iterations the LR and Wald statistic, with corresponding p-values, is given in Table 5.2 calculated by equations (1.41) and (1.47) respectively.

Table 5.2: LR and Wald statistics for HIV status

	After 8 iterations	
Variable	**Statistic**	**p-value**
LR	10.6243	< 0.0001
Wald	70.4261	0.8398

Under the null hypothesis that $\boldsymbol{\beta}$ is not significantly different from 0, the LR statistic rejects the null hypothesis and the Wald statistic does not reject the null hypothesis. The two conflicting conclusions of the LR and Wald statistics are an indication that an approximate chi-square distribution wasn't obtained. This is due to the small sample of 20 observations which is used. The maximum number of iterations were increased to 25 and 100 respectively. The results for the coefficient estimates and standard error in Table 5.1 of the coefficient estimates (obtained in MINITAB) are given columns 2 and 3 respectively. From these results it is clear that the parameter estimates increase as the log-likelihood function increases (as indicated by Figure 2.2). As the number of iterations increases, the value for $\widehat{\boldsymbol{\beta}}$ and the standard error for $\widehat{\boldsymbol{\beta}}$ increase and tend to infinity since there is no convergence. This indicates that none of the coefficient estimates found in Table 5.1 are reliable and other methods should be applied.

5.2.3 Exact logistic regression: HIV status

There is a function available in SAS, PROC LOGISTIC, which can be used to estimate the coefficients for an exact logistic regression. This function was used to estimate exact logistic regression coefficients for the HIV status model in equation (5.1). The corresponding p-values for the two sided conditional exact test are also obtained and the results are given in Table 5.3.

Table 5.3: Exact logistic regression coefficient estimates for HIV status

Variable	$\widehat{\beta}$	**p-value**
Intercept	.	.
$x_{V3.74}$	-11.9668	0.0022
$x_{V1.18}$	-1.9115	0.2857

From the values expressed in Table 5.3 it can be noted that there is no intercept value, and the effect of $x_{V3.74}$ is significant but $x_{V1.18}$ is not significant in the model. To test the overall fit of the model the Pearson chi-square, deviance (based on $q = 19$) and Hosmer-Lemeshow test under the null hypothesis that the model fits the data adequately are considered in Tables 5.4 and 5.5.

Table 5.4: Pearson and deviance test statistic for HIV under exact logistic regression

Test	**Test Statistic**	**DOF**	**p-value**	**Conclusion**
Pearson chi-square	13.56	16	0.631	The model is adequate
Deviance	17.15	16	0.376	The model is adequate

As discussed in Section 1.10.1, caution should be taken when analysing the Pearson chi-square and deviance statistic for a small sample and when working with a sparse data set. The difference between the deviance statistic value and the Pearson chi-square value should heed as a warning to the interpretation of the goodness-of-fit measures.

Since the Hosmer-Lemeshow test depends on the size of the groups chosen, groups of size $1, 2, 4$ and 5 each will be considered and are given in Table 5.5.

Table 5.5: Hosmer-Lemeshow test statistics for HIV status under exact logisitic regression

v_a	a	**Hosmer-Lemeshow test statistic**	**DOF**	**p-value**	**Conclusion**
1	20	13.65	18	0.752	The model is adequate
2	10	13.25	8	0.1	The model is adequate
4	5	9.1	3	0.03	The model is not adequate
5	4	12.26	2	0.002	The model is not adequate

Hosmer-Lemeshow's p-values for 4 and 5 groups, of size 5 and 4 respectively, are the lowest and the null hypothesis that the model is an adequate model is rejected. For 1 group of size 20 and 2 groups of size 10 each the p-values are greater than 0.05 and the null hypothesis is not rejected at a significance level of $\alpha = 0.05$. Over the different group sizes the p-values vary between 0.002 and 0.752 which is a clear indication that the group size used makes a difference with this test. From the Pearson chi-square and deviance test the model is adequate. From the Hosmer-Lemeshow statistics the indication is that the model is not adequate since the results for a larger number of groups ($v_a = 4$ and $v_a = 5$) are more reliable than for a few number of groups, and these indicate it is not a good model.

Finally consider the classification table for the exact logistic regression model with different cut-off values for the predicted probability. The cut-off values ranges from $\widehat{\pi} = 0.1$ to $\widehat{\pi} = 0.9$ in increments of 0.1. To illustrate how well the predicted model performs with the different cut-off levels the following classification tables are given (Tables 5.6 and 5.7).

Table 5.6: Classification table for HIV status under exact logisitic regression for a cut-off probability from 0.1 to 0.5.

$\widehat{\pi}$			Exact		
			Predicted values		**Percentage correct**
			HIV$^-$	**HIV$^+$**	
0.1	**Observed Values**	**HIV$^-$**	10	0	100%
		HIV$^+$	10	0	0%
	overall %				50%
0.2 to 0.4	**Observed Values**	**HIV$^-$**	10	0	100%
		HIV$^+$	8	2	20%
	overall %				60%
0.5	**Observed Values**	**HIV$^-$**	10	0	100%
		HIV$^+$	6	4	40%
	overall %				70%

Table 5.7: Classification table for HIV status under exact logisitic regression for a cut-off probability from 0.6 to 0.9.

$\widehat{\pi}$			Exact		
			Predicted values		Percentage correct
			HIV$^-$	HIV$^+$	
0.6	Observed	HIV$^-$	10	0	100%
	Values	HIV$^+$	2	8	80%
	overall %				90%
0.7	Observed	HIV$^-$	10	0	100%
	Values	HIV$^+$	1	9	90%
	overall %				95%
0.8	Observed	HIV$^-$	10	0	100%
	Values	HIV$^+$	0	10	100%
	overall %				100%
0.9	Observed	HIV$^-$	3	7	30%
	Values	HIV$^+$	0	10	100%
	overall %				65%

From Table 5.7 it can be noted that the exact logistic regression model classifies the values well for a cut-off value between 0.6 and 0.8. Since this is a balanced data set, the classification for $\widehat{\pi} = 0.5$ will be used in practice and for this cut-off value the number of observations which are correctly classified is $14/20 = 70\%$ from Table 5.6.

Overall the exact logistic regression model only has one significant parameter, is indicated as an overall good fit by Pearson chi-square and deviance tests, but not by the Hosmer-Lemeshow test. Finally this model has a relatively good classification rate.

5.2.4 Firth's Method: HIV status

Continuing with the HIV status example SAS, PROC LOGISTIC, will once again be applied with the model statement 'firth'. The coefficient estimates for the HIV status data set under Firth's model are given in Table 5.8 and the estimates converged after 5 iterations.

It can be noted from Table 5.8 that the parameter estimates for both $x_{V3.74}$ and $x_{V1.18}$ are significant at a 5% level (p-values obtained from Wald statistic discussed in Section 1.11.1). Even though the covariates used are significant it is empirical to test the overall performance of the model (given Firth's coefficient estimates). For overall goodness-of-fit

Table 5.8: Firth's method coefficient estimates for HIV status

Variable	$\widehat{\boldsymbol{\beta}}$	**p-value**
Intercept	1.9783	0.7353
$x_{V3.74}$	−42.6196	0.0147
$x_{V1.18}$	−10.0157	0.0376

the Pearson chi-square and deviance are given in Table 5.9 with the same degrees of freedom mentioned in Table 5.4.

Table 5.9: Pearson and deviance test statistic for HIV under Firth's method

Test	**Test Statistic**	**DOF**	**p-value**	**Conclusion**
Pearson chi-square	1.43	16	0.999	The model is adequate
Deviance	2.66	16	0.999	The model is adequate

From the two tests done in Table 5.9, Firth's model is an adequate model given the high p-values of the goodness-of-fit tests. Once again the value for the Pearson chi-square and the deviance statistics are slightly different which indicates that caution should be applied when interpreting these values. The Hosmer-Lemeshow statistics for the same groups sizes as considered for the exact test are given in Table 5.10.

Table 5.10: Hosmer-Lemeshow test statistics for HIV status under Firth's method

v_a	a	**Hosmer-Lemeshow test statistic**	**DOF**	**p-value**	**Conclusion**
1	20	1.435	18	0.999	The model is adequate
2	10	1.422	8	0.994	The model is adequate
4	5	0.671	3	0.88	The model is adequate
5	4	1.394	2	0.498	The model is adequate

From the results for Hosmer-Lemeshow's test for all group sizes it is clear that this model is adequate for all group sizes.

Classification of the model in equation (5.1) given die coefficient estimates in Table 5.8 for cut-off values ranging from $0.1 \leq \widehat{\pi} \leq 0.9$ in increments of 0.1 are given in Table 5.11.

Firth's model classifies the predicted outcomes perfectly for $0.3 \leq \widehat{\pi} \leq 0.8$ which includes the predicted probability for a balanced data set of $\widehat{\pi} = 0.5$.
In conclusion when using Firth's model to predict the HIV status of a patient, both covariates are significant in the model, the overall goodness-of-fit tests indicate an appropriate model and the classification table gives 100% correctly classified values for $\widehat{\pi} = 0.5$.

Table 5.11: Classification table for HIV status under Firth's method for cut-off probability from 0.1 to 0.9

$\widehat{\pi}$			Firth		
			Predicted values		Percentage correct
			HIV$^-$	HIV$^+$	
0.1	Observed Values	HIV$^-$	10	0	100%
		HIV$^+$	3	7	70%
	overall %				85%
0.2	Observed Values	HIV$^-$	10	0	100%
		HIV$^+$	1	9	90%
	overall %				95%
0.3 to 0.8	Observed Values	HIV$^-$	10	0	100%
		HIV$^+$	0	10	100%
	overall %				100%
0.9	Observed Values	HIV$^-$	8	2	80%
		HIV$^+$	0	10	100%
	overall %				90%

5.2.5 Hidden logistic regression model

The final approach, and most recently developed model, to deal with completely separated data is a hidden logistic regression model. The coefficient estimates for the HIV data set based on a hidden logistic regression model obtained with the R program in Appendix G (mentioned in the overview of part II) are given in Table 5.12. The appropriate Wald statistic p-values are also given in Table 5.12 next to each applicable coefficient estimate.

Table 5.12: Hidden logistic regression coefficient estimates for HIV status

Variable	$\widehat{\beta}$	p-value
Intercept	5.398	0.6944
$x_{V3.74}$	-82.638	0.0985
$x_{V1.18}$	-20.447	0.1351

The log-likelihood function converged after 10 iterations and finite coefficient estimates were obtained. From the values in Table 5.12 it can be noted that both the coefficient estimates for $x_{V3.74}$ and $x_{V1.18}$ are not significant at a significance level of $\alpha = 0.05$. The goodness-of-fit statistics for the hidden logistic regression model are given in Tables 5.13 and 5.14 for the same degrees of freedom considered for both exact logistic regression and Firth's model.

Table 5.13: Pearson and deviance test statistic for HIV under hidden logistic regression

Test	**Test Statistic**	**DOF**	**p-value**	**Conclusion**
Pearson chi-square	0.211	16	0.999	The model is adequate
Deviance	0.414	16	0.999	The model is adequate

Table 5.14: Hosmer-Lemeshow test statistics for HIV status under hidden logisitic regression

v_a	a	**Hosmer-Lemeshow test statistic**	**DOF**	**p-value**	**Conclusion**
1	20	0.211	18	0.999	The model is adequate
2	10	0.21	8	0.999	The model is adequate
4	5	0.059	3	0.996	The model is adequate
5	4	0.21	2	0.902	The model is adequate

All the goodness-of-fit tests done on the hidden logistic regression model indicate that it is a very good model. The classification of the predicted outcomes from the hidden logistic regression model is 100% correctly classified for all pre-defined cut-off values as shown in Table 5.15.

Table 5.15: Classification table for HIV status under hidden logistic regression for cut-off probability from 0.1 to 0.9

$\widehat{\pi}$			**Hidden logistic regression**		
			Predicted values		**Percentage correct**
			HIV$^-$	**HIV$^+$**	
0.1 to 0.9	**Observed Values**	**HIV$^-$**	10	0	100%
		HIV$^+$	0	10	100%
	overall %				100%

Overall for the hidden logistic regression model even though none of the coefficient estimates are significant, the goodness-of-fit tests and the classification table indicates it is a good model for the HIV data set.

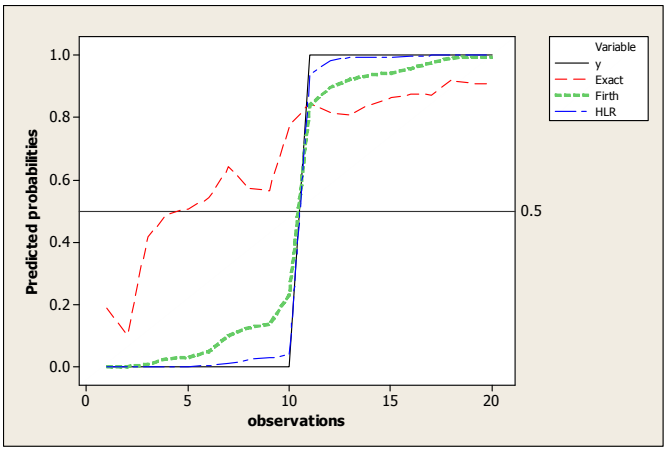
5.2.6 Conclusion: HIV status example

From the example shown above, it can be noted that even though Firth's method and the hidden logistic regression model both give 100% correctly classified values at $\widehat{\pi} = 0.5$, only the model obtained under Firth's method give significant coefficient estimates for both $x_{V3.74}$ and $x_{V1.18}$. Since the values for $\widehat{\boldsymbol{\beta}}$ for Firth's method are significant predictors

in the model it can be interpreted with more confidence than those obtained under the hidden logistic regression model.

The results obtained for the goodness-of-fit tests and classification tables can be explained by interpreting the predicted probabilities obtained under each model. The predicted probabilities for each observation for exact logistic regression, Firth's model and hidden logistic regression are plotted in Figure 5.2.

Figure 5.2: HIV example predicted probabilities for exact, Firth and hidden



The curve obtained from the predicted probabilities under the exact logistic regression model does not resemble the curve of either the original data set or that of a logistic curve (Figure 1.1). This explains the fact why the exact logistic regression model did not give significant coefficient estimates nor was it an appropriate model according to the Hosmer Lemeshow goodness-of-fit tests.

Firth's model however gives a smoother curve between $y_i = 0$ and $y_i = 1$ which enables a better prediction for $0 < y_i < 1$. The coefficient estimates for Firth's model are all significant this implies the estimated values obtained from the model can be interpreted with more confidence for all values of y_i as seen from Figure 5.2.

Finally from Figure 5.2 the good results obtained from the goodness-of-fit tests and the classification table for hidden logistic regression can be explained. The predicted probabilities obtained under the hidden logistic regression model closely resemble that of the original completely separated data set. Where the hidden logistic regression model yields predicted values close to 0 and 1, Firth's model gave a more even distribution of the predicted probabilities between 0 and 1, resembling Figure 1.1. Therefore the good results for both the goodness-of-fit tests and the classification table under a hidden logistic regression model can be explained by the fact that it closely resembles the original data set and this also explains why the coefficient estimates are not significant since this model this closely resembles a data set which is separated between $y_i = 0$ and $y_i = 1$.

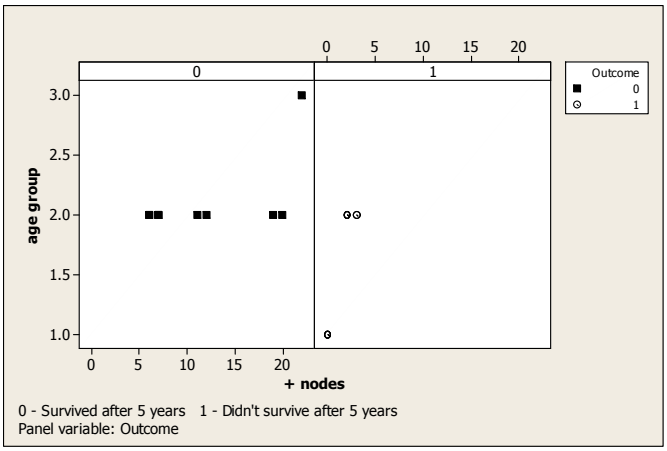
5.3 Continuous and categorical covariates

5.3.1 Breast cancer example

To illustrate a mixture of both categorical and continuous covariates, consider a study between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer, summarised by Haberman (1973). The two covariates which were significant in predicting if a patient survived or did not survive after 5 years of receiving the operation are the number of positive axillary nodes and the age group of the patient observed at the time of the operation.

A sample of 20 observations, given in Appendix B, were drawn where 10 of the patients survived after 5 years (S, $y_i = 1$) and 10 of the patients did not survive after 5 years (D, $y_i = 0$). The number of positive axillary nodes is a continuous covariate and the age group is a categorical covariate where 1 indicates a patient aged between 20 and 39, 2 indicates an age between 40 and 59 and 3 indicates a patient aged between 60 and 79. In the sample of 20 patients complete separation was observed which can be verified by a scatter plot of the age group and the number of positive nodes shown in Figure 5.3.

Figure 5.3: Scatter plot of x_{NODES} vs. x_{AGE} grouped according to the observed value of y



The number of unique covariate sets for this data set is 10, i.e. $q = 10$, which is far less than for the HIV status data set.

5.3.2 General logistic regression model

To predict the outcome of a patient after 5 years, a logistic regression model can be given by equation (5.2)

$$\text{logit}(\pi_i) = \beta_0 + \beta_{d_{11}} x_{d_{11}} + \beta_{d_{12}} x_{d_{12}} + \beta_{NODES} x_{NODES} \tag{5.2}$$

where the indicators of age are summarised in Table 5.16.

Table 5.16: Coding for age groups

		Indicator	
		$x_{d_{11}}$	$x_{d_{12}}$
Age	**Group 1 (20-39**	1	0
	Group 2 (40-59)	0	1
	Group 3 (60-79)	0	0

The interaction between the number of positive nodes detected and the age group are not significant in predicting the survival status of a patient after 5 years and is therefore not included in the model.

From SAS, PROC LOGISTIC, the coefficient estimates for equation (5.2) did not converge after 8 iterations. Since complete separation is present in the data set the log-likelihood function was unable to converge even after 100 iterations, the results of the coefficient estimates are given in Table 5.17. The standard error for each coefficient estimate obtained after 8 iterations is greater than the coefficient estimate itself, this is an indication of unreliable coefficient estimates.

Table 5.17: Logistic regression coefficient estimates for breast cancer

	After 8 iterations		After 25 iterations		After 100 iterations	
Variable	$\widehat{\boldsymbol{\beta}}$	**SE of** $\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\beta}}$	**SE of** $\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\beta}}$	**SE of** $\widehat{\boldsymbol{\beta}}$
Intercept	77.67	168.7	−289.96	457849	−251.769	652712
$x_{d_{11}}$	−68.47	175.9	263.75	481327	192.486	500000
$x_{d_{12}}$	−59.8	146.8	222.63	448157	160.788	452226
x_{NODES}	−3.95	6.1879	14.941	14654	21.3304	41523

Subsequently as the coefficient estimates did not converge, the significance of the parameters are questionable. To test the significance of vector $\widehat{\boldsymbol{\beta}}$,consider the LR and Wald statistic given in Table 5.18.

Table 5.18: LR and Wald statistics for breast cancer

	After 8 iterations	
Variable	**Test statistic value**	**p-value**
LR	27.7	< 0.0001
Wald	0.4516	0.93

Under the null hypothesis that $\boldsymbol{\beta}$ does not significantly differ from 0, for the coefficient estimates obtained after 8 iterations, it can be noted that the according to the LR test

statistic the estimates for $\boldsymbol{\beta}$ are significant. This is however contradictory to the result obtained from the Wald test statistic, where the coefficient estimates found in Table 5.17 are not significant. This is likely due to the small sample which makes the large-sample approximation unreliable. To obtain reliable coefficient estimates different methods, which are not influenced by complete separation, should be tested.

5.3.3 Exact logistic regression

When applying an exact logistic regression model to the breast cancer data one should keep the statement of Zorn(2005): "relatively sparse data and or small numbers of observations in particular patterns of categorical covariates often lead to degenerate estimates, and the inclusion of continuous covariates nearly always does so" in mind when using both categorical and continuous covariates with this method. According to SAS, PROC LOGISTIC, the coefficient estimates obtained for the exact logistic regression model, substituted in equation (5.2) are given in Table 5.19.

Table 5.19: Exact logistic regression coefficient estimates for breast cancer

Variable	$\widehat{\boldsymbol{\beta}}$	**p-value**
Intercept	.	.
$x_{d_{11}}$	.	.
$x_{d_{12}}$	.	.
x_{NODES}	-0.871	0.0004

The coefficient estimates for the intercept, $x_{d_{11}}$ and $x_{d_{12}}$ were unobtainable since the conditional distribution is degenerate. This is due to the fact that both categorical and continuous covariates are used in the model. The coefficient estimate obtained for the number of positive nodes is however a significant parameter estimate for the model.

Since not all the coefficient estimates were obtainable, the model given in equation (5.2) only includes the number of positive nodes detected to determine the patient's outcome after 5 years of surgery. The goodness-of-fit statistics for the model with only one independent variable are given in Table 5.20.

Table 5.20: Pearson and deviance test statistic for breast cancer under exact logistic regression

Test	**Test Statistic**	**DOF**	**p-value**	**Conclusion**
Pearson chi-square	54.27	6	0	The model is not adequate
Deviance	31.57	6	0	The model is not adequate

As suspected the model is not adequate to predict the outcome of a patient. The Hosmer-Lemeshow statistics, for different group sizes, are given in Table 5.21 and provides the same conclusion as given for the Pearson chi-square and deviance test.

Table 5.21: Hosmer-Lemeshow test statistics for breast cancer under exact logisitic regression

v_a	a	**Hosmer-Lemeshow test statistic**	**DOF**	**p-value**	**Conclusion**
1	20	54.27	18	0	The model is not adequate
2	10	54.27	8	0	The model is not adequate
4	5	51.18	3	0	The model is not adequate
5	4	44.57	2	0	The model is not adequate

Finally to get a tabular representation of the model obtained by exact methods, the classification table is given in Table 5.22.

Table 5.22: Classification table for breast cancer under exact logisitic regression for cut-off probability from 0.1 to 0.9

$\widehat{\pi}$			**Exact**		
			Predicted values		**Percentage correct**
			S	**D**	
0.1	**Observed Values**	**S**	8	2	80%
		D	0	10	100%
	overall %				90%
0.2 to 0.4	**Observed Values**	**S**	4	6	40%
		D	0	10	100%
	overall %				70%
0.5 to 0.9	**Observed Values**	**S**	0	10	0%
		D	0	10	100%
	overall %				50%

As noted from Table 5.22 for a cut-off value of $\widehat{\pi} = 0.1$ the model predicted the outcomes very well, relatively well for $0.2 \leq \widehat{\pi} \leq 0.4$ and classified all the values to be in placed in group $y_i = 0$ for a cut-off value of $0.5 \leq \widehat{\pi} \leq 0.9$. Since it is a balanced data set, when considering a cut-off value of $\widehat{\pi} = 0.5$, only 50% of the predicted values are correctly classified. Given the results from the classification table and the goodness-of-fit statistics, this model is not good in predicting the outcome of a patient 5 years after the operation. This result was anticipated when using a combination of both categorical and continuous covariates in an exact logistic regression model.

5.3.4 Firth's method

Since Firth's method is based on penalising the existing log-likelihood function, the type of covariates used should not have an effect on the significance of the model. Given the Firth specification in SAS, PROC LOGISTIC, the coefficient estimates for the covariates identified in equation (5.2) are given in Table 5.23.

Table 5.23: Firth's method coefficient estimates for breast cancer

Variable	$\widehat{\boldsymbol{\beta}}$	**p-value**
Intercept	23.2419	0.0375
$x_{d_{11}}$	−21.0352	0.0626
$x_{d_{12}}$	−18.4047	0.0420
x_{NODES}	−1.0605	0.0338

Most of the coefficient estimates given in Table 5.23 are significant at a 5% significance level. When considering $\alpha = 0.05$, $x_{d_{11}}$ is not significant but since $x_{d_{11}}$ and $x_{d_{12}}$ represents the age group together one either needs to keep or exclude both from the model. The goodness-of-fit tests for Pearson chi-square and deviance test statistics are given in Table 5.24 followed by the Hosmer-Lemeshow's test statistics in Table 5.25. All the goodness-of-fit tests indicate that Firth's model is a good model to predict the outcome of a breast cancer patient after 5 years of surgery.

Table 5.24: Pearson and deviance test statistic for breast cancer under Firth's method

Test	**Test Statistic**	**DOF**	**p-value**	**Conclusion**
Pearson chi-square	2.66	6	0.85	The model is adequate
Deviance	4.57	6	0.6	The model is adequate

Table 5.25: Hosmer-Lemeshow test statistics for breast cancer under Firth's method

v_a	a	**Hosmer-Lemeshow test statistic**	**DOF**	**p-value**	**Conclusion**
1	20	2.66	18	0.999	The model is adequate
2	10	2.5	8	0.96	The model is adequate
4	5	1.25	3	0.74	The model is adequate
5	4	2.36	2	0.31	The model is adequate

Even though Firth's model is an adequate model by the goodness-of-fit statistics, it is also necessary to compare the predicted values to the actual values for a small sample. Therefore the classification table of the model with the coefficients given in Table 5.23 is given in Table 5.26.

Table 5.26: Classification table for breast cancer under Firth's method for cut-off probability from 0.1 to 0.9

$\widehat{\pi}$			Firth		
			Predicted values		Percentage correct
			S	D	
0.1	Observed Values	S	10	0	100%
		D	3	7	70%
	overall %				85%
0.2 to 0.4	Observed Values	S	10	0	100%
		D	1	9	90%
	overall %				95%
0.5 to 0.8	Observed Values	S	10	0	100%
		D	0	10	100%
	overall %				100%
0.9	Observed Values	S	8	2	80%
		D	0	10	100%
	overall %				90%

The best classification occurs at the predicted probability cut-off value between 0.5 and 0.8. The worst classification occurs at a cut-off value of $\widehat{\pi} = 0.1$ which still predicted 85% of the observations in the correct group.

From exact logistic regression to Firth's method there has been a significant improvement in the goodness-of-fit statistics, the significance of the coefficient estimates and the classification table.

5.3.5 Hidden logistic regression

The final approach considered for the breast cancer data set is hidden logistic regression. Using the R program code given in Appendix G, the coefficient estimates given in Table 5.27 are found.

According to the Wald test statistic none of the coefficient estimates are significant at $\alpha = 0.05$. Even though none of the coefficients are significant, for pure academic interest sake, it will be used in equation (5.2) to estimate the outcome of a patient after 5 years.

One should also consider the overall fit of the model given by the Pearson chi-square, deviance and Hosmer-Lemeshow test statistics given in Tables 5.28 and 5.29. From all the given goodness-of-fit tests, the model is adequate.

Table 5.27: Hidden logistic regression coefficient estimates for breast cancer

Variable	$\widehat{\beta}$	**p-value**
Intercept	41.051	0.205
$x_{d_{11}}$	−35.748	0.281
$x_{d_{12}}$	−31.466	0.247
x_{NODES}	−2.107	0.111

Table 5.28: Pearson and deviance test statistic for breast cancer under hidden logistic regression

Test	**Test Statistic**	DOF	**p-value**	**Conclusion**
Pearson chi-square	0.23	6	0.999	The model is adequate
Deviance	0.45	6	0.998	The model is adequate

Table 5.29: Hosmer-Lemeshow test statistics for breast cancer under hidden logisitic regressiont

v_a	a	**Hosmer-Lemeshow test statistic**	**DOF**	**p-value**	**Conclusion**
1	20	0.23	18	0.999	The model is adequate
2	10	0.23	8	0.999	The model is adequate
4	5	0.061	3	0.996	The model is adequate
5	4	0.228	2	0.89	The model is adequate

The last guidance to how well the hidden logistic regression model fits the breast cancer data is the classification table for $0.1 \leq \widehat{\pi} \leq 0.9$ given in Table 5.30, which gives a 100% correct classification for all cut-off predicted probability values.

The hidden logistic regression model indicates very good prediction for the breast cancer data from the goodness-of-fit results and the classification table even though none of the coefficient estimates are significant. To investigate this final conclusion a visual comparison of the predicted probabilities for each method will be investigated in Section 5.3.6.

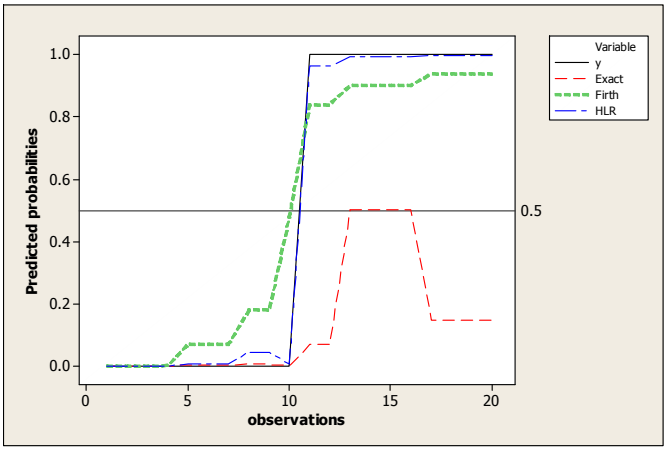
5.3.6 Conclusion: Breast cancer example

A visual comparison, in Figure 5.4, of the three methods and the actual data give insight on the results obtained for the coefficient estimates, goodness-of-fit tests and the classification tables.
Investigating the curve for the predicted probabilities obtained under exact logistic regression it can be observed that the curve does not represent the actual data or the curve in Figure 1.1. Since all the coefficient estimates were not obtainable under exact logistic regression a true representation of a logistic curve was unlikely to be observed from the

Table 5.30: Classification table for breast cancer under hidden logistic regression for cut-off probability from 0.1 to 0.9

$\widehat{\pi}$			Hidden logistic regression		
			Predicted values		Percentage correct
			S	D	
0.1 to 0.9	**Observed Values**	**S**	10	0	100%
		D	0	10	100%
	overall %				100%

Figure 5.4: Breast cancer example predicted probabilities for exact, Firth and hidden



predicted probabilities. The predicted probabilities obtained under exact logistic regression reaches a maximum of $\widehat{\pi} = 0.5$, this can explain why the classification table predicted all the values to be $y_i = 0$ for a predicted probability between 0.5 and 0.9.

Examining the predicted probabilities obtained under Firth's model there is more of a gradual increase from 0 to 1 than observed under the hidden logistic regression model. The curve obtained from the predicted probabilities under Firth's method resembles the curve obtained in Figure 1.1, which serves as explanation for the significant coefficient estimates, the results obtained from the goodness-of-fit tests and the perfect classification at $\widehat{\pi} = 0.5$.

Finally the hidden logistic regression model closely resembles the actual data same as for the HIV example. This resemblance in the predicted probabilities obtained under the hidden logistic regression model and the actual probabilities of the original data explains the 100% correct classification and the fact that the hidden logistic regression model is an adequate model. This in turn also clarifies why the coefficient estimates obtained under the hidden logistic regression model are not significant since there is still a horisontal jump in the predicted probabilities from 0 to 1.

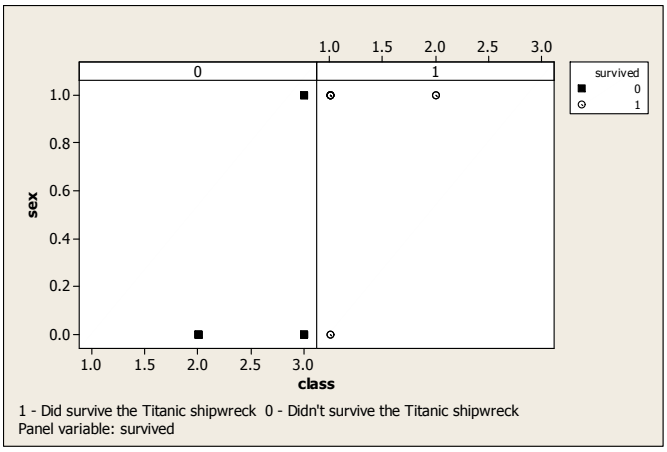
5.4 Categorical covariates

5.4.1 Titanic example

On April 15, 1912, during her maiden voyage, the RMS Titanic sank after colliding with an iceberg, killing 1502 of 2224 passengers and crew members making it one of the most infamous shipwrecks in history. The shipwreck caused a huge loss of life, since there were not enough lifeboats for the passengers and crew. Although there was some degree of luck involved in surviving the disaster, some groups of people were more likely to survive than others, such as women and the first-class passengers.

Consider a sample of 20 observations, given in Appendix C, accessed from Kaggle (2012). The two categorical covariates, sex and passenger class (first, second or third class), are used to determine if a passenger survived ($y_i = 1$) or not survived ($y_i = 0$). The sample of 20 observations exhibit complete separation as indicated by Figure 5.5.

Figure 5.5: Scatter plot of x_{SEX} vs. x_{CLASS} according to the observed value of y



Since both covariates considered are categorical the number of unique covariate sets will be less than that of the two previous examples. The number of unique covariate sets, as seen from Figure 5.5, is 6, i.e. $q = 6$. The six groups consist of first class: female and male, second class: female and male and finally third class: class female and male.

5.4.2 General logistic regression model

To predict if a passenger who was aboard the RMS Titanic survived the shipwreck, two categorical covariates are used which imply dummy variables will be required. Since sex only has two groups, where female is indicated by a 1 and male by a 0 no additional indicators will be needed for sex. There are however three different passenger classes therefore the two indicators in Table 5.31 will be used for passenger class.

Table 5.31: Coding for age groups

		Indicator	
		$x_{d_{11}}$	$x_{d_{12}}$
Class	**First class**	1	0
	Second class	0	1
	Third class	0	0

The logistic regression model to predict the odds of survival for a passenger is given by equation (5.3). In the model no interaction term between sex and class is included since it has no significant contribution in predicting the outcome of a passenger who was aboard the RMS Titanic.

$$\text{logit}(\pi_i) = \beta_0 + \beta_{d_{11}} x_{d_{11}} + \beta_{d_{12}} x_{d_{12}} + \beta_{SEX} x_{SEX} \tag{5.3}$$

From SAS, PROC LOGISTIC and MINITAB the coefficient estimates for equation (5.3) obtained for 8, 25 and 100 iterations are given in Table 5.32. Same as for the previous two results, the standard error for each coefficient estimate is much greater than the coefficient estimate itself.

Table 5.32: Logistic regression coefficient estimates for Titanic

	After 8 iterations		After 25 iterations		After 100 iterations	
Variable	$\widehat{\beta}$	**SE of** $\widehat{\beta}$	$\widehat{\beta}$	**SE of** $\widehat{\beta}$	$\widehat{\beta}$	**SE of** $\widehat{\beta}$
Intercept	-22.0708	45.0255	-72.9268	131925	-86.89	914604
$x_{d_{11}}$	29.3929	52.7734	97.1337	151752	180.483	578732
$x_{d_{12}}$	14.5855	39.7666	48.51	117531	53.8807	409067
x_{SEX}	14.3938	30.7782	48.2491	88369	59.9473	816507

From the values given in Table 5.32 it is clear that the coefficients did not converge even after 100 iterations and that the standard error of the coefficients tends to infinity as the number of iterations increase. The LR and Wald statistics for the coefficient estimates obtained after 8 iterations are given in Table 5.33 for 3 degrees of freedom. As for the two previous cases of complete separation it is observed that the conclusion given from the LR and Wald statistic is contradictory.

Table 5.33: LR and Wald statistics for Titanic

	After 8 iterations	
Variable	**Statistic**	**p-value**
LR	27.7129	< 0.0001
Wald	0.3461	0.9511

Since the general logistic regression model could not obtain finite coefficient estimates,

the exact method, Firth's method and hidden logistic regression will be applied to the observations once again.

5.4.3 Exact logistic regression

Applying exact logistic regression to the data set given in Appendix C, the coefficient estimates given in Table 5.34 are obtained.

Table 5.34: Exact logistic regression coefficient estimates for Titanic

Variable	$\widehat{\beta}$	**p-value**
Intercept	-3.1558	0.0054
$x_{d_{11}}$	4.0444	0.0025
$x_{d_{12}}$	1.4436	0.3333
x_{SEX}	2.1622	0.1333

As indicated from Table 5.34 all the coefficients were obtainable since only categorical covariates were considered for the exact method. Even though the indicator $x_{d_{12}}$ and x_{SEX} does not have significant coefficient estimates at a significance level of $\alpha = 0.05$, the overall effect of $x_{d_{11}}$ and $x_{d_{12}}$ is to be considered in the model.

By using the coefficient estimates obtained in Table 5.34 in the logistic regression model specified by equation (5.3) the Pearson chi-square, deviance and Hosmer-Lemeshow test statistics are given in Tables 5.35 and 5.36.

Table 5.35: Pearson and deviance test statistic for Titanic under exact logistic regression

Test	**Test Statistic**	**DOF**	**p-value**	**Conclusion**
Pearson chi-square	4.0144	2	0.134	The model is adequate
Deviance	0.1343	2	0.033	The model is not adequate

Table 5.36: Hosmer-Lemeshow test statistics for Titanic under exact logisitic regression

v_a	a	**Hosmer-Lemeshow test statistic**	**DOF**	**p-value**	**Conclusion**
1	20	4.0144	18	0.999	The model is adequate
2	10	4.0144	8	0.856	The model is adequate
4	5	1.949	3	0.583	The model is adequate
5	4	3.78	2	0.151	The model is adequate

The results obtained from Pearson's chi-square and the deviance test statistic are conflicting. Pearson's chi-square indicates that the model is adequate but on the other hand

the deviance test statistic indicates that the model is not adequate. When there is a difference between the values of Pearson chi-square and deviance test statistic, it is an indication that the values did not converge to an approximate chi-square distribution. This result is unexpected since theoretically the Pearson's chi-square and deviance test statistic should perform best when only categorical covariates are considered and should therefore be close in value. To investigate if this difference is truly the result of the small sample used, the same example will be revisited in Chapter 6 with a larger sample size.

As shown by the Hosmer-Lemeshow statistics in Table 5.36 the model obtained from the exact method is significant for all group sizes.

Finally the classification table, which cross-classifies the observed and the predicted values, is given in Table 5.37.

Table 5.37: Classification table for Titanic under exact logisitic regression for cut-off probability from 0.1 to 0.9

$\widehat{\pi}$			Exact		
			Predicted values		**Percentage correct**
			S	**D**	
0.1	**Observed Values**	**S**	10	0	100%
		D	6	4	40%
	overall %				70%
0.2	**Observed Values**	**S**	10	0	100%
		D	2	8	80%
	overall %				90%
0.3 to 0.6	**Observed Values**	**S**	10	0	100%
		D	0	10	100%
	overall %				100%
0.7	**Observed Values**	**S**	8	2	80%
		D	0	10	100%
	overall %				90%
0.8 to 0.9	**Observed Values**	**S**	6	4	80%
		D	0	10	100%
	overall %				80%

From the classification table given in Table 5.37 it is noted that the classification of the predicted values for exact logistic regression is 100% correctly identified for $0.3 \leq \widehat{\pi} \leq 0.6$. For the three different examples given it can be noted that only for the case where two

categorical covariates are considered did the exact logistic regression model give 100% correct classification at a cut-off value of $\widehat{\pi} = 0.5$.

5.4.4 Firth's method

By penalising the log-likelihood function of equation (5.3) the coefficient estimates given in Table 5.38 are obtained.

Table 5.38: Firth's method coefficient estimates for Titanic

Variable	$\widehat{\boldsymbol{\beta}}$	**p-value**
Intercept	-4.874	0.053
$x_{d_{11}}$	6.3635	0.028
$x_{d_{12}}$	2.8963	0.209
x_{SEX}	3.3518	0.093

The coefficient estimate for x_{SEX} is not significant at $\alpha = 0.05$ and as explained, the two indicators $x_{d_{11}}$ and $x_{d_{12}}$ cannot be evaluated separately since they form one concept (passenger class). Equation (5.3) with the coefficients given in Table 5.38 yields a model which results to a Pearson chi-square and deviance goodness-of-fit statistic given in Table 5.39.

Table 5.39: Pearson and deviance test statistic for Titanic under Firth's method

Test	**Test Statistic**	**DOF**	**p-value**	**Conclusion**
Pearson chi-square	2.025	2	0.36	The model is adequate
Deviance	3.7	2	0.16	The model is adequate

From Table 5.39 it is noticed once again (as for the exact method) that the value for the Pearson chi-square and deviance differs. Even though this difference is smaller than the difference obtained under exact logistic regression for the Titanic example, Firth's method with a large sample will also be investigated in Chapter 6. The Hosmer-Lemeshow statistics are given in Table 5.40 and indicate that the model is adequate for all group sizes.

Table 5.40: Hosmer-Lemeshow test statistics for Titanic under Firth's method

v_a	a	**Hosmer-Lemeshow test statistic**	**DOF**	**p-value**	**Conclusion**
1	20	2.025	18	0.999	The model is adequate
2	10	2.025	8	0.98	The model is adequate
4	5	1.04	3	0.79	The model is adequate
5	4	1.97	2	0.37	The model is adequate

Finally as done with all previous cases the classification table for Firth is given in Table 5.41. By using Firth's model to predict the outcome of a passenger, there is 100% correct classification of the predicted outcomes for a cut-off value between $0.2 \leq \widehat{\pi} \leq 0.7$.

Table 5.41: Classification table for Titanic under Firth's method for cut-off probability from 0.1 to 0.9

$\widehat{\pi}$			Firth		
			Predicted values		Percentage correct
			S	D	
0.1	Observed Values	S	10	0	100%
		D	6	4	40%
	overall %				75%
0.2 to 0.7	Observed Values	S	10	0	100%
		D	0	10	100%
	overall %				100%
0.8	Observed Values	S	8	2	80%
		D	0	10	100%
	overall %				100%
0.9	Observed Values	S	6	4	60%
		D	0	10	100%
	overall %				80%

5.4.5 Hidden logistic regression

The final model to be considered for the Titanic data set is the hidden logistic regression model. The coefficient estimates for this model are given in Table 5.42.

Table 5.42: Hidden logistic regression coefficient estimates for Titanic

Variable	$\widehat{\beta}$	p-value
Intercept	-11.723	0.149
$x_{d_{11}}$	15.627	0.103
$x_{d_{12}}$	7.684	0.284
x_{SEX}	7.528	0.183

Once again as with all the models found under hidden logistic regression, the coefficient estimates according to the Wald statistic are not significant. To compare the models obtained under the three different methods the coefficient estimates from Table 5.42 will be used to substitute into equation (5.3).

The goodness-of-fit statistics corresponding to the hidden logistic regression model are given in Tables 5.43 and 5.44.

Table 5.43: Pearson and deviance test statistic for Titanic under hidden logistic regression

Test	**Test Statistic**	**DOF**	**p-value**	**Conclusion**
Pearson chi-square	0.202	2	0.9	The model is adequate
Deviance	0.399	2	0.819	The model is adequate

Table 5.44: Hosmer-Lemeshow test statistics for Titanic under hidden logisitic regression

v_a	a	**Hosmer-Lemeshow test statistic**	**DOF**	**p-value**	**Conclusion**
1	20	0.202	18	0.999	The model is adequate
2	10	0.202	8	0.999	The model is adequate
4	5	0.106	3	0.991	The model is adequate
5	4	0.201	2	0.9151	The model is adequate

As for all the previous hidden logistic regression models the coefficient estimates are not significant but the model is adequate according to all goodness-of-fit tests. The classification table once again shows 100% correct classification for all cut-off values of $\widehat{\pi}$, as seen in Table 5.45.

Table 5.45: Classification table for Titanic under hidden logistic regression for cut-off probability from 0.1 to 0.9

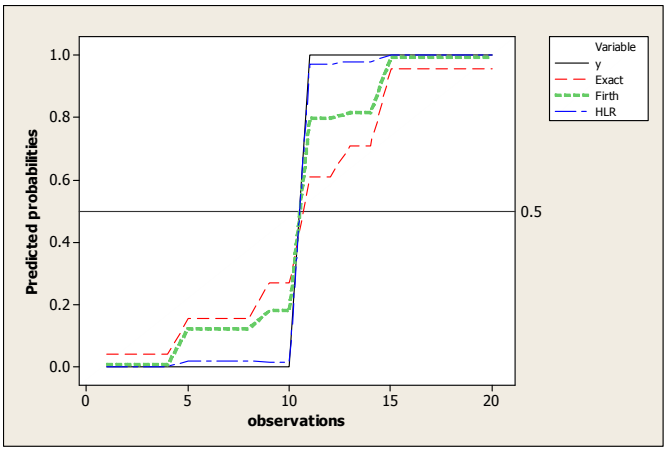
$\widehat{\boldsymbol{\pi}}$			**Hidden logistic regression**		
			Predicted values		**Percentage correct**
			S	**D**	
0.1 to 0.9	**Observed Values**	**S**	10	0	100%
		D	0	10	100%
	overall %				100%

5.4.6 Conclusion: Titanic example

When considering only categorical covariates the predicted probabilities of exact logistic regression, Firth's method and hidden logistic regression is represented in Figure 5.6.

As seen in Figure 5.6 the predicted probabilities of the exact logistic regression model more closely resembles the logistic curve in Figure 1.1 and the actual data than obtained

Figure 5.6: Titanic example predicted probabilities for exact, Firth and hidden



from the previous two examples which support the 100% correct classification at $\widehat{\pi} = 0.5$. Even though the results have improved for exact logistic regression from the previous two examples, the curve of the predicted probabilities under exact logistic regression for the Titanic example continue to deviate from the curve in Figure 1.1. This could be due to the small sample size, a large sample will be used in Chapter 6 to clarify if a better approximation occurs.

The curves for the predicted probabilities under Firth's method and hidden logistic regression still resembles the graphs obtained for the previous two examples. Firth's method yields a model which is adequate and has a good classification rate. Under Firth's method however not all the coefficient estimates are significant and can therefore not be interpreted with as much confidence as in the previous two cases. The predicted probabilities obtained under the hidden logistic regression model still resembles the actual data and this explains the coefficient estimates, goodness-of-fit tests and the classification table.

5.5 Conclusion

Each of the three methods discussed to deal with complete or quasi-complete separation has its own positive and negative attributes. For each method discussed one must ensure that the correct method is used according to the covariate structure of the data set to obtain reliable results.

For exact logistic regression to be valid one can only use a data set where each of the covariates used is of the same data type. When using a combination of both categorical and continuous covariates in exact logistic regression, it is almost certain that one would obtain degenerate results. Exact logistic regression can be applied when the independent variables only consist of continuous data although this is not recommended as there are

models which give better results. Exact logistic regression performs the best when only categorical covariates are used to predict the dependent variable.

When considering Firth's method all covariate types, including a combination of categorical and continuous, can be used. For all three cases of different covariates, most of the coefficient estimates obtained were significant. By considering the overall goodness-of-fit of Firth's model, it was given as an adequate model according to Pearson's chi-square, deviance and Hosmer-Lemeshow's test statistic. For the classification tables, 100% correct classification is obtained for a predefined estimated probability of 0.5.

Finally the hidden logistic regression model performed very well according to the classification tables and to Pearson's chi-square, deviance and Hosmer-Lemeshow's test statistic. A drawback of hidden logistic regression is the fact that it does not give significant coefficient estimates and that there is no predefined function in SAS (or any of the computer programs mentioned in Chapter 2) available to calculate the coefficient estimates. The coefficient estimates which are not significant could be due to the small sample size and will be investigated in Chapter 6 with a larger sample size.

Since all the examples considered in Chapter 5 are based on a small sample, the results for Pearson's chi-square and deviance test statistic should be interpreted with caution as an approximate chi-square distribution might not have been obtained.

When considering the Hosmer-Lemeshow's statistic other than the result depending on the group sizes, it also depends on how many groups are chosen. If a few number of groups are chosen then according to Xie (2008) it will almost always indicate that the model fits the data well. This can also be confirmed for all examples where there is only a few number of groups. For all cases where v_a is 1 or 2 the model is shown to be an adequate model except for the exact logistic regression model applied to the breast cancer data. Therefore when interpreting Hosmer-Lemeshow's statistic one should rather consider number of groups of four and higher.

When interpreting any of the models one should keep the statement of Box (1987) in mind "All models are wrong, some models are useful". Therefore if the object of the study is to resemble the true model as close as possible the hidden logistic regression model will be suitable, if the object of the model is to eliminate most of the separation present to the data set Firth's model will work well. Each method is useful in its own way and the model chosen depends on the desired outcome of the researcher.

Chapter 6

Quasi-complete separation (large sample)

6.1 Introduction

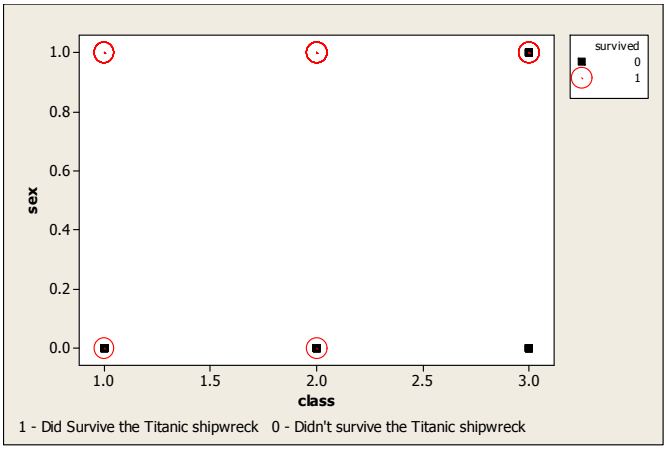
When coefficient estimates of a logistic regression model are unobtainable due to non-convergence for a practical application, the chances that quasi-complete separation is present is much more likely than complete separation. This is especially true when the sample size is large and continuous covariates or a combination of continuous and categorical covariates are used. Since complete separation is a more severe case than quasi-complete separation, all the approaches to deal with complete separation that have been discussed can be used to deal with quasi-complete separation. To demonstrate non-convergence of coefficient estimates for a large sample, the Titanic example of Kaggle (2012) will be revisited. When the sample size is increased to 100, quasi-complete separation is observed in the data set.

6.2 Categorical covariates

6.2.1 Titanic example: large sample

By considering the Titanic example in Appendix D as discussed in Section 5.4.1 from Kaggle (2012) the logistic regression model stays the same as given in equation (5.3) with indicators discussed in Table 5.31. For 100 randomly selected cases, quasi-complete separation is present in the data set, which means that each covariate set is not unique to either $y_i = 1$ (survived) or $y_i = 0$ (did not survive) since some covariate sets present both $y_i = 1$ and $y_i = 0$, as represented in Figure 6.1.

Figure 6.1: Large sample scatter plot of x_{SEX} vs. x_{CLASS} according to the observed value of y



6.2.2 General logistic regression model revisited

From SAS, PROC LOGISTIC the coefficient estimates for the logistic regression model in equation (5.3) is represented in Table 6.1 for 8, 25 and 100 iterations respectively. From Table 6.1 one can note that the coefficient estimates still did not converge after 100 iterations, but the standard errors for the coefficients are smaller than for the case of complete separation. The coefficient estimates and the standard error are also very similar for all coefficient estimates whereas for complete separation the standard error differed by a noticeable amount for the different coefficient estimates. The standard error for the coefficient estimates are still larger than the coefficient estimate itself for all coefficient estimates.

Table 6.1: Logistic regression coefficient estimates for Titanic: large sample

	After 8 iterations		After 25 iterations		After 100 iterations	
Variable	$\widehat{\boldsymbol{\beta}}$	**SE of** $\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\beta}}$	**SE of** $\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\beta}}$	**SE of** $\widehat{\boldsymbol{\beta}}$
Intercept	-11.2245	12.977	-28.2247	38599	-103.32	162225
$x_{d_{11}}$	8.7396	12.9884	25.7398	38599	100.83	162225
$x_{d_{12}}$	8.6596	12.9685	25.6598	38599	100.75	162225
x_{SEX}	11.3299	12.9695	28.3301	38599	103.42	162225

Nevertheless, even if the standard errors are smaller than for complete separation, the coefficient estimates still do not converge after 100 iterations. To investigate significance of the coefficient estimates obtained after 8 iterations the Wald and LR statistic are studied in Table 6.2. From this it can be noted both tests reject the null hypothesis, which implies that both tests indicate that the coefficient estimates used in the model are significant. This conclusion is obtained for 8 iterations and is conflicting with the fact that the coefficient estimates did not converge. From this it can be observed that for a

large sample both tests give the same result (which wasn't the case for a small sample) but the result is not reliable.

Table 6.2: LR and Wald statistics for Titanic: large sample

	After 8 iterations	
Variable	**Statistic**	**p-value**
LR	98.0747	< 0.0001
Wald	9.7488	0.0208

From these results it is apparent that one should make several attempts, trying different methods to obtain coefficient estimates that converge.

6.2.3 Exact logistic regression

From the exact function in SAS, PROC LOGISTIC, the coefficient estimates for exact logistic regression in Table 6.3 are obtained. It can be noted that all the coefficient estimates are highly significant in estimating the survival status of a passenger who was Section the RMS Titanic.

Table 6.3: Exact logistic regression coefficient estimates for Titanic: large sample

Variable	$\widehat{\boldsymbol{\beta}}$	**p-value**
Intercept	-5.619	< 0.0001
$x_{d_{11}}$	3.3772	0.0006
$x_{d_{12}}$	3.5117	0.0001
x_{SEX}	5.8443	< 0.0001

The logistic regression model defined in equation (5.3) with the coefficient estimates given in Table 6.3 leads to the goodness-of-fit statistics given in Tables 6.4 and 6.5.

Table 6.4: Pearson and deviance test statistic for Titanic under exact logistic regression: large sample

Test	**Test Statistic**	**DOF**	**p-value**	**Conclusion**
Pearson chi-square	1.3358	2	0.513	The model is adequate
Deviance	2.3541	2	0.308	The model is adequate

The Pearson chi-square and deviance test statistics both specify that the model is an adequate model since all p-values are greater than 0.05. When comparing the results for Pearson chi-square and deviance for a small sample to a large sample two immediate

Table 6.5: Hosmer-Lemeshow test statistics for Titanic under exact logisitic regression: large sample

v_a	a	**Hosmer-Lemeshow test statistic**	**DOF**	**p-value**	**Conclusion**
1	100	40.764	98	0.999	The model is adequate
2	50	25.401	48	0.997	The model is adequate
5	20	8.875	18	0.963	The model is adequate
10	10	6.506	8	0.591	The model is adequate
20	5	3.109	3	0.375	The model is adequate

differences can be observed. The first is that for a large sample the two values are much closer in value (with an absolute difference of 1.02 between the two statistics) as opposed to the small sample case where there was an absolute difference of 3.88 between the value of the Pearson chi-square and deviance test statistic.

The second difference when considering a large sample, is that the two test statistics lead to the same conclusion whereas with the small sample the test statistics were conflicting. One can therefore conclude that the sample size has a considerable effect on the Pearson chi-square and deviance test statistic. For the Hosmer-Lemeshow statistics given in Table 6.5 the model is adequate for all group sizes.

Finally the classification table for the predicted probabilities under the exact logistic regression model is given in Table 6.6.

Table 6.6: Classification table for Titanic under exact logisitic regression for cut-off probability from 0.1 to 0.9: large sample

			Exact: large sample		
$\widehat{\pi}$			**Predicted values**		**Percentage correct**
			S	**D**	
0.1	Observed	S	49	1	98%
	Values	D	22	28	56%
	overall %				77%
0.2 to 0.5	Observed	S	48	2	96%
	Values	D	9	41	82%
	overall %				89%
0.6 to 0.9	Observed	S	38	12	76%
	Values	D	0	50	100%
	overall %				88%

At a predicted probability of 0.5, 89% of the cases are correctly classified. Over all

predicted probability cut-off values the exact logistic regression model gives a high percentage of values which are correctly classified. The lowest correctly classified percentage of 77% is at a cut-off estimated probability of 0.1.

6.2.4 Firth's method

The coefficient estimates obtained for equation (5.3) given Firth's method are given in Table 6.7. As for exact logistic regression for a large sample, all the coefficient estimates are significant. When using the coefficient estimates acquired in Table 6.7 in equation (5.3) the overall goodness-of-fit statistics of the model are given in Tables 6.8 and 6.9.

Table 6.7: Firth's method coefficient estimates for Titanic: large sample

Variable	$\widehat{\beta}$	**p-value**
Intercept	-6.3106	< 0.0001
$x_{d_{11}}$	4.1497	0.008
$x_{d_{12}}$	4.1018	0.0051
x_{SEX}	6.4116	< 0.0001

Table 6.8: Pearson and deviance test statistic for Titanic under Firth's method: large sample

Test	**Test Statistic**	**DOF**	**p-value**	**Conclusion**
Pearson chi-square	0.804	2	0.669	The model is adequate
Deviance	1.402	2	0.496	The model is adequate

Table 6.9: Hosmer-Lemeshow test statistics for Titanic under Firth's method: large sample

v_a	a	**Hosmer-Lemeshow test statistic**	**DOF**	**p-value**	**Conclusion**
1	100	40.177	98	0.999	The model is adequate
2	50	21.094	48	0.999	The model is adequate
5	20	8.364	18	0.973	The model is adequate
10	10	6.099	8	0.636	The model is adequate
20	5	2.603	3	0.457	The model is adequate

Even though the Pearson chi-square and deviance test statistic gave the same conclusion for both a small and large sample, there is still a difference between the two values. There is a bigger absolute difference between Pearson's chi-square and deviance for the small sample of 1.675 than when using a large sample where the absolute difference is 0.598.

Given the goodness-of-fit tests, including the Hosmer-Lemeshow test statistics for any number of groups, the overall fit of Firth's model is adequate.

Table 6.10: Classification table for Titanic under Firth's method for cut-off probability from 0.1 to 0.9: large sample

			Firth: large sample		
$\widehat{\pi}$			Predicted values		Percentage correct
			S	**D**	
0.1	Observed Values	S	49	1	98%
		D	21	29	58%
	overall %				78%
0.2 to 0.5	Observed Values	S	48	2	96%
		D	9	41	82%
	overall %				89%
0.6 to 0.9	Observed Values	S	38	12	76%
		D	0	50	100%
	overall %				88%

Once again when considering the classification table given in Table 6.10 one can note that at a predicted probability of 0.5, the number of correctly classified observations is 89 out of a total of 100, identical to the classification of exact logistic regression.

6.2.5 Hidden logistic regression

The final model to be considered for the large sample case of predicting the survival status of a passenger who was Section the RMS Titanic is the hidden logistic regression model. The coefficient estimates given in Table 6.11 are all significant according to the Wald test statistic, in accordance with the rest of the models using a large sample. This is the first example considered where the coefficient estimates obtained under the hidden logistic regression model are all significant.

The overall goodness-of-fit of the model is evaluated by interpreting the Pearson chi-square, deviance and Hosmer-Lemeshow test statistics. The absolute difference between Pearson chi-square and deviance test statistic as seen from Table 6.12 is 0.268. The difference between these two values is the smallest for all the methods mentioned using a large sample.

Table 6.11: Hidden logistic regression coefficient estimates for Titanic: large sample

Variable	$\widehat{\beta}$	**p-value**
Intercept	-7.309	0.0005
$x_{d_{11}}$	4.906	0.0225
$x_{d_{12}}$	4.851	0.0181
x_{SEX}	7.428	0.0003

Table 6.12: Pearson and deviance test statistic for Titanic under hidden logistic regression: large sample

Test	**Test Statistic**	**DOF**	**p-value**	**Conclusion**
Pearson chi-square	0.287	2	0.866	The model is adequate
Deviance	0.554	2	0.758	The model is adequate

The Hosmer-Lemeshow test statistics values in Table 6.13 also confirm that the hidden logistic regression model for a large sample is adequate.

Table 6.13: Hosmer-Lemeshow test statistics for Titanic under hidden logisitic regression: large sample

v_a	a	**Hosmer-Lemeshow test statistic**	**DOF**	**p-value**	**Conclusion**
1	100	44.220	98	0.999	The model is adequate
2	50	21.493	48	0.999	The model is adequate
5	20	7.963	18	0.979	The model is adequate
10	10	5.377	8	0.717	The model is adequate
20	5	1.859	3	0.602	The model is adequate

Finally the classification table in Table 6.14 revealed that the hidden logistic regression model could correctly classify 89% of the observations at $\widehat{\pi} = 0.5$ which is the same as for exact logistic regression and Firth's method for the large sample case.

6.3 Conclusion

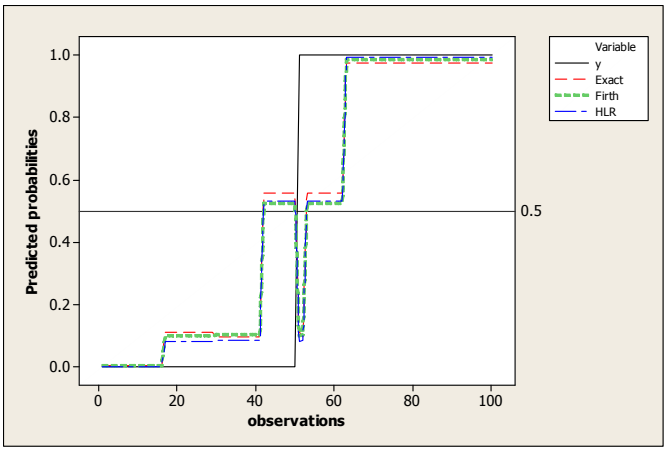
As examined for all the small sample cases, a visual representation of the predicted probabilities for exact logistic regression, Firth's method and hidden logistic regression are shown in Figure 6.2.

The first noticeable difference between the small sample case and the large sample case, is that for the large sample case the curve for the predicted probabilities for all three methods are almost exactly similar. From Figure 6.2 one can explain the 89% correctly classified values given for all methods and the similar results obtained from the goodness-of-fit tests. For the large sample case the difference between the coefficient estimates for

Table 6.14: Classification table for Titanic under hidden logisitic regression for cut-off probability from 0.1 to 0.9: large sample

			Hidden: large sample		
$\widehat{\pi}$			**Predicted values**		**Percentage correct**
			S	**D**	
0.1 to 0.5	Observed Values	S	48	2	96%
		D	9	41	82%
	overall %				77%
0.6 to 0.9	Observed Values	S	38	12	76%
		D	0	50	100%
	overall %				88%

Figure 6.2: Large sample Titanic example predicted probabilities for exact, Firth and hidden



all three methods are much smaller than for the small sample case. From this one can conclude that for the case where only categorical covariates are used, as the sample size increases, the coefficient estimates obtained for exact logistic regression, Firth's method and hidden logistic regression approximates to the same value.

When considering the Titanic example other differences between the small and the large sample case were observed.

The first noticeable difference is the fact that the Wald, Pearson chi-square and deviance test statistics gave more reliable results, due to the fact that with a large sample one can be more certain of the fact that an approximate chi-square distribution is obtained.

The second noticeable difference is that the percentage of correctly classified observations has decreased to 89% for all methods at a cut-off predicted probability of 0.5 where it

was 100% for all methods for the small sample case. Even though this is the case, for the large sample all the coefficient estimates obtained under all methods are significant which enables one to interpret the results with more confidence than for the small sample case.

When considering the different methods, exact logistic regression and Firth's method increased in performance from the Pearson chi-square, deviance and Hosmer-Lemeshow test statistics are considered. All of the corresponding p-values for the above mentioned tests increased for both exact logistic regression and for Firth's method. From the goodness-of-fit measures for the two above-mentioned methods, one also notices that the difference between Pearson's chi-square and the deviance test statistic became smaller for the large sample case than it was for the small sample. Given all these improvements for the large sample case it must also be mentioned that complete or quasi-complete separation is much more unlikely in a large sample than it is in a small sample.

Chapter 7

Summary and Conclusion

A logistic regression model predicts the outcome of a dichotomous variable based on a set of covariates. When deriving a logistic regression model it is very important to investigate the type of covariates present in your data set. When interpreting goodness of fit tests it is important to bear in mind that Pearson's chi-square and the deviance test statistic perform best when the sample size is large and when categorical covariates are considered. Hosmer-Lemeshow's statistic is influenced by the group sizes and the number of groups chosen

When constructing a logistic regression model the coefficient estimates do not always exist. Therefore it is essential to test whether complete or quasi-complete separation is present in a data set when the dependent variable is dichotomous, especially when the sample size is small. When complete or quasi-complete separation is present in the data set it is imperative not to continue with the general approach of a logistic regression model, but to follow a different tactic.

Many different solutions for complete and quasi-complete separation exist in practice. Each method has its own benefits and disadvantages and these should be investigated before applying the method to your data set. For the methods which were investigated, exact logistic regression performs well with categorical covariates but caution should be applied when using this method to a sparse data set. Firth's method gives significant coefficient estimates for most cases and transforms the data to represent a logistic curve which gradually increases to an estimated probability from 0 to 1. Finally the hidden logistic regression model gives perfect classification for all cases, but still closely resembles a model under complete or quasi complete separation.

For exact logistic regression one can only use a data set where each of the covariates used are of the same data type. When using a combination of both categorical and continuous covariates in exact logistic regression, it is almost certain that one would obtain unreliable results. Exact logistic regression can be used when the independent variables only consist

of continuous data although this is not recommended. Exact logistic regression performs the best when only categorical covariates are used to predict the dependent variable.

When considering Firth's method or hidden logistic regression all covariate types, including a combination of categorical and continuous, can be used. The coefficient estimates for exact logistic regression and Firth's method can be obtained with available functions in SAS. A drawback of hidden logistic regression is that there is no predefined function in SAS (or any of the computer programs mentioned in Chapter 2) available to calculate the coefficient estimates.

For the large sample case there was a remarkable improvement in exact logistic regression, Firth's method and hidden logistic regression compared to a small sample.

Future work could comprise of including more covariates in the model and analysing the model fit statistics to compare the effect of including and excluding covariates in the different models; multicollinearity in the models can be explored, and the presence of outliers evaluated. The different methods to deal with complete or quasi-complete separation mentioned in Section 3.2, not covered in this dissertation, can be investigated for different covariate types and sample sizes.

The scope which have been covered in this dissertation expresses the importance in revaluating the model when either complete or quasi-complete separation is obtained in the data. It also highlights the differences among the three methods investigated (exact logistic regression, Firth's method and hidden logistic regression) and the fact that each method cannot be used in any scenario. Finally it was shown how the goodness of fit methods can differ based on the type of covariates used, sample size and group sizes.

References

ABRAHANTES, J. C. and AERTS, M. (2012). A solution to separation for clustered binary data. Statistical Modelling 12, 3-27.

AGRESTI, A. (1990). Categorical data analysis. Wiley.

AGRESTI, A. (1996). An Introduction to Categorical Data Analysis. Wiley.

ALBERT, A. and ANDERSON, J. (1984). On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71, 1-10.

ALLISON, P., ALTMAN, M., GILL, J. and MCDONALD, M. P. (2004). Convergence problems in logistic regression. Numerical issues in statistical computing for the social scientist, 238-252.

ALLISON, P. D. (2012). Logistic regression using SAS: Theory and application. SAS Institute.

BEGGS, S., CARDELL, S. and HAUSMAN, J. (1981). Assessing the potential demand for electric cars. Journal of econometrics 17, 1-19.

BERTOLINI, G., DAMICO, R., NARDI, D., TINAZZI, A. and APOLONE, G. (2000). One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. Journal of epidemiology and biostatistics 5, 251-253.

BOX, G. E. P. and DRAPER, N. R. (1987). Empirical model-building and response surfaces. Wiley.

CHAPMAN, R. G. (1984). An Approach to Estimating Logit Models of a Single Decision Maker's Choice Behavior. Advances in Consumer Research 11, 656-661.

CHRISTMANN, A. and ROUSSEEUW, P. J. (2001). Measuring overlap in binary regression. Computational Statistics and Data Analysis 37, 65-75.

CLOGG, C. C., RUBIN, D. B., SCHENKER, N., SCHULTZ, B. and WEIDMAN, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. Journal of the American Statistical Association 86, 68-78.

COLLETT, D. (2003). Modelling Binary Data. Chapman and Hall/CRC.

COPAS, J. (1988). Binary regression models for contaminated data. Journal of the Royal Statistical Society. Series B (Methodological), 225-265.

COX, D. R. and HINKLEY, D. V. (1979). Theoretical Statistics. Chapman and Hall.

COX, D. R. and SNELL, E. J. (1989). Analysis of Binary Data. Chapman and Hall.

DAY, N. E. and KERRIDGE, D. F. (1967). A general maximum likelihood discriminant. Biometrics, 313-323.

DYKE, G. and PATTERSON, H. (1952). Analysis of factorial arrangements when the data are proportions. Biometrics 8, 1-12.

EKHOLM, A. and PALMGREN, J. (1982). A model for a binary response with misclassifications. Proceedings of the GLIM 82: Proceedings of the international conference on generalised linear models, 128-143.

FIRTH, D. (1992a). Bias reduction, the Jefferys prior and GLIM. Advances in GLIM and statistical modelling: proceedings of the GLIM92 Conference and the 7th International Workshop on Statistical Modelling, Munich, 13-17 July 1992: Springer-Verlag.

FIRTH, D. (1992b). Generalized linear models and jeffreys priors: an iterative weighted l east-squares approach. Computational statistics 1, 553-557.

FIRTH, D. (1993). Bias reduction of maximum likelihood estimates. Biometrika 80, 27-38.

GAO, S. and SHEN, J. (2007). Asymptotic properties of a double penalised maximum likelihood estimator in logistic regression. Statistics and probability letters 77, 925-930.

HABERMAN, S. (1973). Generalized Residuals for Log-Linear Models. Proceedings of the 9th International Biometrics Conference. Boston, pp. p. 104-122.

HEINZE, G. (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data. Statistics in medicine 25, 4216-4226.

HEINZE, G. and SCHEMPER, M. (2002). A solution to the problem of separation in logistic regression. Statistics in medicine 21, 2409-2419.

HOSMER, D. W., LEMESHOW, S. and COOK, E. D. (2001). Applied Logistic Regression, Second Edition: Book and Solutions Manual Set. Wiley.

JEFFREYS, H. (1946). An invariant form for the prior probability in estimation problems. Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences 186, 453-461.

KAGGLE (2012).Go from Big Data to Big Analytics. [ONLINE] Available at: https://www.kaggle.com/c/titanic-gettingStarted/data. [Accessed 01 June 13]

KING, E. N. and RYAN, T. P. (2002). A preliminary investigation of maximum likelihood logistic regression versus exact logistic regression. The American Statistician 56, 163-170.

LIU, Y. (2007). On Goodness-of-Fit of Logistic Regression Model. Department of Statistics. Kansas: Kansas State University.

LLOYD, C. J. (1999). Statistical analysis of categorical data. Wiley.

MCCULLAGH, P. and NELDER, J. A. (1989). Generalized Linear Models. Chapman and Hall.

MEHTA, C. R. and PATEL, N. R. (1995). Exact logistic regression: theory and examples. Statistics in medicine 14, 2143-2160.

PEARSON, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 50, 157-175.

PHILIPPEOS, C., STEFFENS, F. and MEYER, D. (2009). Comparative 1H NMR-based metabonomic analysis of HIV-1 sera. Journal of biomolecular NMR 44, 127-137.

QUENOUILLE, M. H. (1949). Approximate tests of correlation in time-series 3. Proceedings of the Mathematical Proceedings of the Cambridge Philosophical Society, 483-484.

QUENOUILLE, M. H. (1956). Notes on bias in estimation. Biometrika 43, 353-360.

RAO, C. R. (1973). Linear Statistical Inference and Its Applications. Wiley.

ROUSSEEUW, P. J. and CHRISTMANN, A. (2003). Robustness against separation and outliers in logistic regression. Computational Statistics and Data Analysis 43, 315-332.

SANTNER, T. J. and DUFFY, D. E. (1986). A note on A. Albert and JA Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. Biometrika 73, 755-758.

SILVAPULLE, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. Journal of the Royal Statistical Society. Series B (Methodological), 310-313.

XIE, X.-J., PENDERGAST, J. and CLARKE, W. (2008). Increasing the power: A practical approach to goodness-of-fit test for logistic regression models with continuous predictors. Computational Statistics and Data Analysis 52, 2703-2713.

ZORN, C. (2005). A solution to separation in binary response models. Political Analysis 13, 157-170.

Appendix A: Data on HIV status patients

HIV Status	V3.74	V1.18
1	−0.4	1.63
1	−0.42	1.62
1	−0.44	1.79
1	−0.45	1.8
1	−0.46	1.99
1	−0.35	1.44
1	−0.45	1.61
1	−0.38	1.6
1	−0.5	1.85
1	−0.45	1.61
0	−0.29	1.73
0	−0.48	2.36
0	−0.36	1.95
0	−0.02	1.26
0	−0.26	1.8
0	−0.28	1.74
0	−0.24	2.27
0	−0.26	1.49
0	−0.27	1.71
0	−0.27	1.54

Appendix B: Data on breast cancer patients

Outcome	+ nodes	age group
1	2	2
1	3	2
1	0	1
1	2	2
1	0	1
1	0	1
1	0	1
1	2	2
1	3	2
1	2	2
0	22	3
0	12	2
0	7	2
0	11	2
0	7	2
0	6	2
0	7	2
0	19	2
0	6	2
0	20	2

Appendix C: Data on Titanic passengers: small sample

Survived	Class	Sex
1	1	1
1	2	1
1	1	1
1	1	1
1	1	1
1	1	1
1	2	1
1	1	1
1	1	0
1	1	0
0	3	0
0	2	0
0	3	0
0	2	0
0	3	0
0	2	0
0	2	0
0	3	1
0	3	0
0	3	1

Appendix D: Data on Titanic passengers: large sample

S	Class	Sex	S	Class	Sex	S	Class	Sex	S	Class	Sex	S	Class	Sex
1	1	1	1	2	1	1	3	1	0	1	0	0	2	0
1	1	1	1	2	1	1	3	1	0	1	0	0	2	0
1	1	1	1	2	1	1	3	1	0	1	0	0	2	0
1	1	1	1	2	1	1	3	1	0	1	0	0	2	0
1	1	1	1	2	1	1	3	1	0	1	0	0	3	0
1	1	1	1	2	1	1	3	1	0	1	0	0	3	0
1	1	1	1	2	1	1	3	1	0	1	0	0	3	0
1	1	1	1	2	1	1	3	1	0	1	0	0	3	0
1	1	1	1	2	1	1	1	0	0	1	0	0	3	0
1	1	1	1	2	1	1	2	0	0	1	0	0	3	0
1	1	1	1	2	1	0	3	1	0	1	0	0	3	0
1	1	1	1	2	1	0	3	1	0	2	0	0	3	0
1	1	1	1	2	1	0	3	1	0	2	0	0	3	0
1	1	1	1	2	1	0	3	1	0	2	0	0	3	0
1	1	1	1	2	1	0	3	1	0	2	0	0	3	0
1	2	1	1	2	1	0	3	1	0	2	0	0	3	0
1	2	1	1	2	1	0	3	1	0	2	0	0	3	0
1	2	1	1	2	1	0	3	1	0	2	0	0	3	0
1	2	1	1	3	1	0	3	1	0	2	0	0	3	0
1	2	1	1	3	1	0	1	0	0	2	0	0	3	0

Appendix E: SAS program code for penalised likelihood function

```
proc iml;
y={input values};
x1={input values};
n=nrow(y);
x0=J(n,1,1);
intercept=x0*b0;
X=x0||x1;
values=J(20,3,0);
logl=J(n,1,0);
W=J(n,n,0);

do beta=0.1 to 2 by 0.1;
x1beta=x1*beta;
fullx=intercept||x1beta;

do i=1 to n;
xirow=fullx[i,+];
W[i,i]=((exp(xirow))/(1+exp(xirow)))*(1/(1+exp(xirow)));
logl[i,1]=y[i,1]*xirow-log(1+exp(xirow));
total=logl[+];
end;

Xtrans=X‘;
Fisher=Xtrans*W*X;
detF=det(Fisher);
print fisher detF;
plogl=total+(0.5)*log(detF);
ind=beta*10;
```

```
values[ind,1]=beta;
values[ind,2]=total;
values[ind,3]=plogl;
end;

print  values;

cn = {'beta' 'LogL' 'PenLogL'};
create outputval from values[colname=cn];
append from values;
run;

goptions reset=all i=join;
proc gplot data=outputval;
plot PenLogL*beta;
run;

quit;
```

Appendix F: SAS program code for classification tables

```
proc iml;
Beta={input values};
y={input values};
x={inut values as matrix};

n=nrow(y);
intercept=J(n,1,1);
fullx=intercept||x;
print firthbeta y x fullx;
logodds=fullx*firthbeta;
results=J(10,5,0);

do prob=0.1 to 0.9 by 0.1;
ind=prob*10;
count=J(n,1,0);
cc=0;
Obs1pred0=0;
obs0pred1=0;
obs1pred1=0;
obs0pred0=0;
logoddsC0=log(prob/(1-prob));
testlogodds=J(n,1,logoddsC0);
diff=logodds-testlogodds;

do i=1 to n;
if diff[i,1]>0 then count[i,1]=1;
groups=y-count;
if groups[i,1]=0 then cc=cc+1;
if groups[i,1]=1 then obs1pred0=obs1pred0+1;
```

```
if groups[i,1]=-1 then obs0pred1=obs0pred1+1;
end;

do j=1 to (n/2);
if groups[j,1]=0 then obs1pred1=obs1pred1+1;
end;

obs0pred0=cc-obs1pred1;
results[ind,1]=cc;
results[ind,2]=obs1pred1;
results[ind,3]=obs0pred0;
results[ind,4]=obs1pred0;
results[ind,5]=obs0pred1;
end;

cn={'correctly_classified' 'observed 1 predicted 1' 'observed 0 predicted 0'
'observed 1 predicted 0' 'observed 0 predicted 1'};
create outputval from results[colname=cn];
append from results;
run;

proc print data=outputval;
run;

quit;
```

Appendix G: R program for hidden logistic regression

```
y=c(1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0)

{HIV}
x1=c(-0.4,-0.42,-0.44,-0.45,-0.46,-0.35,-0.45,-0.38,-0.5,-0.45,
-0.29,-0.48,-0.36,-0.02,-0.26,-0.28,-0.24,-0.26,-0.27,-0.27)
x2=c(1.63,1.62,1.79,1.8,1.99,1.44,1.61,1.6,1.85,1.61,1.73,
2.36,1.95,1.26,1.8,1.74,2.27,1.49,1.71,1.54)

{Breast Cancer}
dj1=c(0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0)
dj2=c(1,1,0,1,0,0,0,1,1,1,0,1,1,1,1,1,1,1,1,1)
nodes=c(2,3,0,2,0,0,0,2,3,2,22,12,7,11,7,6,7,19,6,20)

{Titanic}
dj1=c(1,0,1,1,1,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0)
dj2=c(0,1,0,0,0,0,1,0,0,0,0,1,0,1,0,1,1,0,0,0)
sex=c(1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,1,0,1)

delta=0.01
epsilon=0.000001
maxit=100

pihat = max(delta, min(1-delta,mean(y)))
delta0 = (pihat*delta) / (1+delta)
delta1 = (1+pihat*delta) / (1+delta)
ytilde = delta0*(1-y) + delta1*y
response = cbind(ytilde,1-ytilde)
print(ytilde)
print(response)
```

```
func=glm(response ~ x1+x2, family=binomial,
control=glm.control(epsilon=epsilon, maxit=maxit))
print(func)
summary(func)
confint(func)
```

Appendix H: SAS program code for Pearson chi-square, deviance and Hosmer-Lemeshow statistics

```
proc iml;
y={input values};
x1={input values};
x2={input values};
x3={input values};
fullx=x1||x2||x3||y;
n=nrow(y);
model=J(n,1,0);
predp=J(n,1,0);
vc=J(n,8,0);

/*The values for b0,b1,b2 and b3 needs to be manually inputted*/
intercept=J(n,1,(b0));
model=intercept+b1*x1+b2*x2+b3*x3;

do i=1 to n;
predp[i,1]=exp(model[i,1])/(1+exp(model[i,1]));
end;

fullx=x1||x2||sex||y||predp;
call sort( fullx, {1 2 3}, {2} );
unique_loc = uniqueby( fullx,{1 2 3}, 1:nrow(fullx) );
q=nrow(unique_loc);
Test_unique=J(q,1,(n+1));
Test_unique[1:(q-1)]=unique_loc[2:q];
yval=J(n,q,0);
sum=J(1,q,0);
```

```
do i=1 to q;
vc[i,2:4]=fullx[unique_loc[i,1],1:3];
vc[i,5]=Test_unique[i,1]-unique_loc[i,1];
vc[i,6]=unique_loc[i,1]+vc[i,5]-1;
if vc[i,5]>1 then;
yval[(unique_loc[i,1]:vc[i,6]),i]=fullx[(unique_loc[i,1]:vc[i,6]),4];
sum=yval[+,];
vc[i,7]= sum[1,i];
vc[i,8]=fullx[unique_loc[i,1],5];
end;

call sort( vc, {2 1}, {2} );
vc=vc[1:q,];

do i=1 to q;
vc[i,1]=i;
end;

resid=J(q,1,0);
residsq=J(q,1,0);
dev=J(q,1,0);
devsq=J(q,1,0);
ind=J(q,1,0);

do i=1 to nrow(vc);
resid[i,1]=(vc[i,7]-vc[i,5]*vc[i,8])/sqrt(vc[i,5]*vc[i,8]*(1-vc[i,8]));
residsq[i,1]=resid[i,1]##2;
log1=(vc[i,7]/(vc[i,5]*vc[i,8]));
log2=((vc[i,5]-vc[i,7])/(vc[i,5]*(1-vc[i,8])));
if log1>0 & log2>0 then ;
dev[i,1]=sqrt(2*(vc[i,7]*log(log1)+(vc[i,5]-vc[i,7])*log(log2)));
if log1=0 then; dev[i,1]=-sqrt(2*vc[i,5]*abs(log(1-vc[i,8])));
if log2=0 then; dev[i,1]=sqrt(2*vc[i,5]*abs(log(vc[i,8])));
devsq[i,1]=dev[i,1]##2;
end;

pearson=residsq[+];
deviance=devsq[+];
```

```
HLderive=predp||y;

call sort(HLderive, {1}, );

/*manually input the number of groups va*/
va=4;
a=n/va;
groups=J(n,a,0);
yvalues=J(n,a,0);
sum2=J(1,a,0);
HLval=J(1,a,0);
indic=J(a,1,0);
from=J(a,1,1);

do i=1 to a;
indic[i,1]=i*va;
if i>1 then from[i,1]=(indic[i-1,1]+1);
groups[1:va,i]=HLderive[from[i,1]:indic[i,1],1];
yvalues[1:va,i]=HLderive[from[i,1]:indic[i,1],2];
oi=yvalues[+,];
ei=groups[+,];
pi=ei/va;
HLval[1,i]=(oi[1,i]-va*pi[1,i])##2/(va*pi[1,i]*(1-pi[1,i]));
end;

dof_pd=q-(m+1);
dof_h=a-2;
HLstat=HLval[+];

p_pvalue=1-PROBCHI(pearson,dof_pd);
d_pvalue=1-PROBCHI(deviance,dof_pd);
hl_pvalue=1-PROBCHI(HLstat,dof_h);

print  pearson p_pvalue deviance d_pvalue dof_pd HLstat hl_pvalue dof_h;
quit;
```