UNIVERSITY OF CALGARY

A Comparison of Two Different Logistic Regression Models for Analyzing Data from Case-Control Studies

by

Xiaochun Wang

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF COMMUNITY HEALTH SCIENCES

CALGARY, ALBERTA

MAY, 2010

UNIVERSITY OF
CALGARY

Abstract

Case-control studies can be used to investigate the relationship between a disease and potential risk factor(s). The logistic regression analysis is one of the analytical tools used in case-control studies. There are two types of logistic regression models that can be used in case-control studies. The model for the log odds of exposure fits the case-control sampling scheme which is disease-dependent. The model for the log odds of disease contradicts the case-control sampling scheme. However, Prentice and Pyke provided the theoretical justification for using the model for the log odds of disease in case-control studies. The primary aim of this thesis is to compare the coefficients that are related to disease or exposure, as well as, their standard errors in the two types of logistic regression models. Some suggestions for future research directions are provided at the end.

Acknowledgements

There are so many people that I am indebted to. My research was made possible with their huge input. I own profound thanks to Dr. Marilynne A. Hebert, Dr. Misha Eliasziw, Dr. Gordon H. Fick, and Dr. Lynn M. Meadows for giving me the hope to pursue a graduate study. Dr. Gordon H. Fick and Dr. Misha Eliasziw, my educational supervisors guided me through my graduate study. Dr. Gordon H. Fick provided an interesting topic and a clear outline. Dr. Misha Eliasziw provided a worthy support and a great assistance. I cannot thank them enough. I would like to express my deep gratitude to Dr. Xuewen Lu for the generous share of his considerable insights. I wish to thank Dr. Colleen J. Maxwell, Dr. Hude Quan, and Dr. Christine Friedenreich for their helpful suggestions and references. Dr. Scott B. Patten, Dr. Gemai Chen, Dr. Viena Stastna, Dr. Thi Dinh and Dr. Joseph Amuah motivated me to initiate my graduate study. I also wish to thank Ms. Crystal Elliott for providing excellent administrative services and Ms. Christine Sopczak for assisting me with the completion of my thesis. In addition, I would like to thank other examination committee members Dr. Alberto Nettel−Aguirre and Dr. Brent Hagel, and the examiner Dr. Peter Ehlers, for their valuable suggestions.

I am grateful to my father, my sisters, my brother-in-laws, my relatives and my friends for their consistently support and encourage. I also want to thank my husband and my two sons for their cherishable support and patience during my graduate study.

Dedication

In Memory of My Dearest Mother

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Case-Control Studies

Much medical research is carried out to explore the relationship between the occurrence of disease and its risk factors by estimating measures of association. Cohort studies and case-control studies can be used for this purpose. In a cohort study, exposure groups are defined from a source population at the beginning of the study, and the subjects are followed over a given period of time and their disease frequencies are ascertained and compared. In contrast, a case-control study compares exposure frequencies between a group of subjects with a disease and a "comparable" [9] group of subjects without the disease from a source population (or a "hypothetical" population [9]) where the source population is defined as "were the illness now to occur, it would be identified by the defined scheme ..... [the scheme] by which cases of the illness are identified" [40].

Cohort studies are a valuable type of observational study that assess a putative casual relationship; cohort studies provide a clear temporal sequence of risk factor(s) and disease [40]. Researchers, in some circumstances, have to use a case-control study as an alternative to cohort studies "to reach the same conclusions in a [case-control] study as would have been obtained from a [cohort] study, if one had been done" [36] and "case-control studies can be conceptualized as a more efficient version of a corresponding cohort study" [49]. For example, if researchers investigate a disease which has a very low incidence, then even a large sample size may record only a few diseased individuals. Under this circumstance, case-control studies "may be the only feasible approach" [36] or

"the only useful alternative" [49] to explore the relationship between the disease and its risk factor(s). Another example would be, if a disease latency is relatively long, follow-up might prove to be difficult or impossible.

For the simplest case-control study, we can represent the data in a 2×2 table as in Table 1.1. The case-control odds ratio of exposure which is used to measure the association between the disease (D) and its risk factor (E) is $\frac{a/b}{c/d} = \frac{ad}{bc}$.

Table 1.1: The 2×2 Table (Frequencies)

	$E = 1$	$E = 0$
$D = 1$	a	b
$D = 0$	c	d

Some types of case-control study designs can be considered.

1.1.1 Density Case-Control Studies

In this design, "the sampling probability of any person as a control should be proportional to the amount of person-time that person spends at risk of disease in the source population...apart from sampling error " [49], i.e., the rate of the number of exposed controls (c) to the total exposed person-time (T_E) is the same as the rate of the number of unexposed controls (d) to the total unexposed person-time ($T_{\bar{E}}$) in the source population expressed as

$$\frac{c}{T_E} = \frac{d}{T_{\bar{E}}} \tag{1.1}$$

which means

$$\frac{T_E}{T_{\bar{E}}} = \frac{c}{d}. \tag{1.2}$$

Then the rate ratio can be estimated as

$$\frac{\frac{a}{T_E}}{\frac{b}{T_{\bar{E}}}} = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{ad}{bc}. \tag{1.3}$$

That is, with a density case-control design, a case-control odds ratio can be used to estimate the rate ratio. In this type of design, controls are selected from the noncases in the source population at the time points that each case appears.

1.1.2 Case-Cohort Studies

In this design, the source population is a cohort and everyone in the cohort has an equal chance to be selected as a control regardless of whether that person developed the disease during the study period [49], i.e., the number of exposed control (c) to the total exposed persons (N_E)) is the same as the number of unexposed control (d) to the total unexposed persons ($N_{\bar{E}}$) expressed as

$$\frac{c}{N_E} = \frac{d}{N_{\bar{E}}} \tag{1.4}$$

which means

$$\frac{N_E}{N_{\bar{E}}} = \frac{c}{d}. \tag{1.5}$$

Then the incidence proportion ratio can be estimated as

$$\frac{\frac{a}{N_E}}{\frac{b}{N_{\bar{E}}}} = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{ad}{bc}. \tag{1.6}$$

That is, with a case-cohort design, a case-control odds ratio can be used to estimate the incidence proportion ratio. In this type of design, controls are selected from the source population at the beginning of the study.

1.1.3 Cumulative Case-Control Studies

In this design, controls are selected from the noncases in the source population at the end of the study period. Suppose that a fraction f of both exposed and unexposed noncases

($N_E - a$ and $N_{\bar{E}} - b$) are included as controls. That is

$$f = \frac{c}{N_E - a} = \frac{d}{N_{\bar{E}} - b} \tag{1.7}$$

which means

$$\frac{N_E - a}{N_{\bar{E}} - b} = \frac{c}{d}. \tag{1.8}$$

Then the incidence odds ratio can be estimated as

$$\frac{\frac{a}{N_E - a}}{\frac{b}{N_{\bar{E}} - b}} = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{ad}{bc}. \tag{1.9}$$

That is, with a cumulative case-control design, a case-control odds ratio can be used to estimate the incidence odds ratio.

The incidence odds ratio will provide a good approximation of the rate ratio, provided that the disease incidence proportion is low (less than about 0.1 [49]) and it is for a closed population ($N_E + N_{\bar{E}}$) over the study period (Δt) [49]. Let

$$I_E = \frac{a}{N_E\ \Delta t} \approx \frac{a}{(N_E - a)\ \Delta t} \tag{1.10}$$

where I_E is the disease incidence rate in the exposed group. Then using formula (1.9)

$$\frac{I_E}{I_{\bar{E}}} = \frac{\frac{a}{N_E - a}}{\frac{b}{N_{\bar{E}} - b}} = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{ad}{bc}. \tag{1.11}$$

In a similar manner, Cornfield [15] showed that this incidence odds ratio can also approximate the incidence proportion ratio (the relative risk) provided the disease incidence proportion is low (see Appendix A).

1.2 Statistical Methodologies in Case-Control Studies

Case-control studies involve comparisons between different groups (diseased and non-diseased) with respect to specified characteristics. The idea about comparison of

different treatment effects emanated in the 16th and 17th centuries [34]. Louis, who was a pioneer in clinical trials [44], provided the arguments of the numerical method for such comparisons [35]. An example from these early years is a study conducted by one of his students, Guy [27], who compared the ratio of pulmonary consumption frequency to other disease frequency in various occupations in 1845. This methodology of comparison became popular during the 1920s for investigating the relationship between various cancers and their risk factors [34]. For example, Broders [11] investigated the relationship between squamous-cell epithelioma in lip and pipe smoking, showing cases had a high proportion of pipe smokers. Increased attention and development of case-control methodology began in the 1950s [8][9][34].

1.2.1 Mantel-Haenszel Method

Using the odds ratio to estimate the association can be misleading due to potential confounder(s). Miettinen [41] stated that only when the confounders are controlled, can the association measure give the unbiased estimate of the effect measure. In 1959, Mantel and Haenszel [36] presented a method for controlling for confounders. The data are first summarized into a series of 2×2 tables indexed by the cross-classification of the potential confounders (Table 1.2) where a_k, b_k, c_k, and d_k denote the observed cell frequencies in the k^{th} stratum.

Table 1.2: The k^{th} 2×2 Table (Frequencies)

	$E = 1$	$E = 0$
$D = 1$	a_k	b_k
$D = 0$	c_k	d_k

Then two steps are performed:

- A test [12][36] is conducted to see whether the odds ratio is the same and equal to

1 in each table (no association between the disease and the exposure). Under the assumption of homogeneity of the odds ratio, the common odds ratio [36] can be calculated as

$$\hat{OR}_{MH} = \frac{\sum_k \frac{a_k d_k}{n_k}}{\sum_k \frac{b_k c_k}{n_k}} \tag{1.12}$$

where $n_k = a_k + b_k + c_k + d_k$. Mantel and Haenszel did not provide a variance formula for their estimator. Methods by Cornfield [16] and Woolf [55] have been used for calculating the corresponding variance.

- A summary chi-square statistic is calculated to test the homogeneity of the odds ratio across all strata based on the above estimated Mantel-Haenszel odds ratio.

For the ensuing two decades, the Mantel-Haenszel method was the predominant analytical tool used in case-control studies to control for confounding [9]. However, this method is limited. For example, if potential confounders are measured variables, then they cannot be stratified into a series of 2×2 tables.

1.2.2 Logistic Regression Models

Regression models have been developed to describe the relationship between one or more explanatory variables and a response variable [51]. Logistic regression models are generalized linear models (GLM) with a logit link function [37] and can be used when the outcome is binary (e.g. the diseased or the non-diseased). For example, let D be the disease indicator taking on values 0 (non-diseased) and 1 (diseased) and E be an exposure variable. The logistic regression model that specifies the probability of disease that depends on the exposure status can be written as

$$Pr^*(D = 1|E) \; = \; \frac{\exp\{\alpha^* \; + \; \beta E\}}{1 + \exp\{\alpha^* \; + \; \beta E\}} \tag{1.13}$$

or in logit form as

$$\log \frac{Pr^*(D=1|E)}{1-Pr^*(D=1|E)} \;=\; \alpha^* \;+\; \beta E. \tag{1.14}$$

The reason that logistic regression models can be applied to binary outcomes is that logistic regressions do not impose constraints on parameters α^* and β, as the estimates of probabilities of disease will always be between 0 and 1. This appealing property is described in detail in Vittinghoff et al. [54] and can be illustrated in Figure 1.1 where E is a measured variable (the dose).

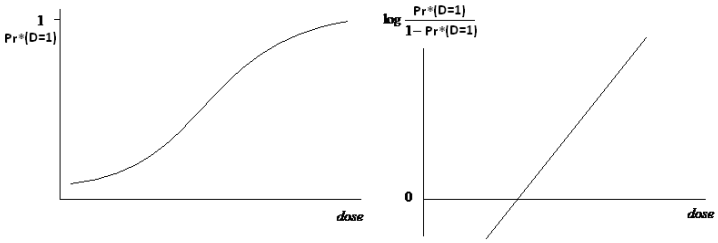

Figure 1.1: The Logistic Function

The odds ratio of disease can be estimated from the logistic regression model as $\exp(\beta)$ when E takes on values 0 and 1. In studying the relationship between a disease and its exposure, other covariates can also be considered for their potential confounding or modifying effects [39].

In the early 1960s, Cornfield et al. [17] began to apply the logistic regression model in cohort studies to deal with the issue of simultaneously assessing several risk factors [8][49]. Logistic regression models were then applied to case-control studies during the 1970s in two versions:

- The Model for the Log Odds of Exposure

  This model, based upon a case-control design, was proposed by Prentice [46] in 1976 based on the invariance of odds ratio [15]. Let G (Gender) be the covariate

that is related to both disease D and exposure E. Then the invariance of the odds ratio of exposure for $G = g$ ($g = 0$ for female, $g = 1$ for male) can be expressed as

$$\frac{\frac{Pr(E=1|D=1,G=g)}{Pr(E=0|D=1,G=g)}}{\frac{Pr(E=1|D=0,G=g)}{Pr(E=0|D=0,G=g)}} = \frac{\frac{Pr^*(D=1|E=1,G=g)}{Pr^*(D=0|E=1,G=g)}}{\frac{Pr^*(D=1|E=0,G=g)}{Pr^*(D=0|E=0,G=g)}}. \tag{1.15}$$

The term on the right side of equation (1.15) is an approximation of relative risk of disease when the disease incidence proportion is low (as mentioned by Cornfield [15]). The term on the left of equation (1.15) can be estimated from a case-control study by modeling the log odds of exposure. The model for the log odds of exposure can be written as

$$\log \frac{Pr(E=1|D,G)}{1 - Pr(E=1|D,G)} = \upsilon \ + \ \tau D \ + \ \theta G \ + \ \zeta DG. \tag{1.16}$$

- The Model for the Log Odds of Disease

  In 1979, Prentice and Pyke [47] proposed a model for the log odds of disease. In addition, they demonstrated that the model for the log odds of disease can be used in case-control studies. Let G be the covariate that is related to disease D and exposure E. Then the model for the log odds of disease can be written as

$$\log \frac{Pr(D=1|E,G)}{1 - Pr(D=1|E,G)} = \alpha \ + \ \beta E \ + \ \gamma G \ + \ \delta EG. \tag{1.17}$$

  Letting $\boldsymbol{x} = \begin{pmatrix} E \\ G \\ EG \end{pmatrix}$ and $\boldsymbol{\beta} = \begin{pmatrix} \beta \\ \gamma \\ \delta \end{pmatrix}$ and $x = (E, G)$, equation (1.17) becomes

$$\log \frac{Pr(D=1|x)}{1 - Pr(D=1|x)} = \alpha \ + \ \boldsymbol{x}\boldsymbol{\beta}. \tag{1.18}$$

Model (1.18) was actually derived by Prentice and Pyke [47] from theories corresponding to a prospective sampling scheme. They argued that if a sample is taken from a mixture of two populations (as in cohort studies), then the likelihood function for the observations can be written as

$$\begin{aligned} L^* &\propto \prod_{d=0}^{1}\prod_{x}\{Pr^*(D=d|x)Pr^*(x)\}^{n_{dx}} \\ &= \prod_{d=0}^{1}\prod_{x}\{Pr^*(D=d|x)\}^{n_{dx}} \prod_{d=0}^{1}\prod_{x}\{Pr^*(x)\}^{n_{dx}} = L_1^* L_2^* \end{aligned} \tag{1.19}$$

where n_{dx} is the sample size for the group with $D = d$ and x and

$$Pr^*(D=1|x) = \exp\{\alpha^* + \boldsymbol{x}\boldsymbol{\beta}\}Pr^*(D=0|x) \tag{1.20}$$

and

$$Pr^*(D=0|x) = \frac{1}{1+\exp\{\alpha^* + \boldsymbol{x}\boldsymbol{\beta}\}} \tag{1.21}$$

which are the Cox-Day-Kerridge formulae [18][20] for posterior probabilities. In 1967, Day and Kerridge [20] demonstrated that L_1^* in (1.19) can be estimated alone as the estimates of $\boldsymbol{\beta}$ and corresponding covariance matrix will be identical as those when L_1^* and L_2^* are estimated together. However, if separate samples are taken from each population (as in case-control studies), then using the fact that $Pr(x|D) = Pr^\sharp(x|D)$ (i.e., the probability of exposure given disease will be the same in different source populations) and

$$Pr^\sharp(x|D) = \frac{Pr^\sharp(D=d|x)Pr^\sharp(x)}{Pr^\sharp(D)}, \tag{1.22}$$

and assuming that the marginal density $Pr^\sharp(D)$ is known, Anderson [3] showed

that the case-control likelihood function is

$$\begin{aligned} L &= \prod_{d=0}^{1}\prod_{x}\{Pr(x|D=d)\}^{n_{dx}} \\ &\propto \prod_{d=0}^{1}\prod_{x}\{Pr\ (D=d|x)Pr\ (x)\}^{n_{dx}} \\ &= \prod_{d=0}^{1}\prod_{x}\{Pr\ (D=d|x)\}^{n_{dx}} \prod_{d=0}^{1}\prod_{x}\{Pr\ (x)\}^{n_{dx}} \\ &= L_1 L_2 \end{aligned} \tag{1.23}$$

where

$$Pr\ (D=1|x) = \exp\{\alpha\ +\ \boldsymbol{x\beta}\}Pr\ (D=0|x) \tag{1.24}$$

and

$$Pr\ (D=0|x) = \frac{1}{1+\exp\{\alpha\ +\ \boldsymbol{x\beta}\}}. \tag{1.25}$$

L_1 in (1.23) can be estimated alone but is subject to constraints

$$\sum_{x} Pr\ (x) = 1 \tag{1.26}$$

and

$$\sum_{x}\{Pr\ (D=d|x)Pr\ (x)\} = \pi_d \tag{1.27}$$

where π_d is the marginal disease probability and the value of it will not affect the estimated coefficients (except the intercept) so that it can be set as $\frac{n_d}{n}$. By applying Aitchison and Silvey's theory [1], Anderson demonstrated that estimates of $\boldsymbol{\beta}$ and the corresponding covariance matrix in (1.23) are identical to those involved in L_1 alone. However, his demonstrations were based on the assumption that the covariates x are discrete.

Prentice and Pyke [47] extended Anderson's work to measured covariates. They

used another likelihood factorization

$$\begin{aligned} L &= \prod_{d=0}^{1}\prod_{x}\{Pr(x|D=d)\}^{n_{dx}} \\ &\propto \prod_{d=0}^{1}\prod_{x}\{Pr(D=d|x)Pr(x)\}^{n_{dx}} \\ &= \prod_{d=0}^{1}\prod_{x}\{Pr(D=d|x)\}^{n_{dx}} \prod_{d=0}^{1}\prod_{x}\{Pr(x)\}^{n_{dx}} \\ &= L_1 L_2 \end{aligned} \tag{1.28}$$

where

$$Pr(D=1|x) = \exp\{\alpha + \boldsymbol{x}\boldsymbol{\beta}\}Pr(D=0|x), \tag{1.29}$$

$$Pr(D=0|x) = \frac{1}{1+\exp\{\alpha + \boldsymbol{x}\boldsymbol{\beta}\}}, \tag{1.30}$$

and

$$Pr(x) = \sum_{d} \frac{n_d}{n} Pr(x|D=d). \tag{1.31}$$

L_1 in (1.28) can be estimated alone but subject to the constraint

$$\int_x Pr(D=d|x)Pr(x)dx = \frac{n_d}{n}. \tag{1.32}$$

Prentice and Pyke [47] showed that estimating L_1 and L_2 separately in (1.28) will still satisfy constraint (1.32). The estimates of $\boldsymbol{\beta}$ and corresponding covariance matrix in (1.28) will be identical to those involved in L_1 alone.

Prentice and Pyke's arguments were also supported using the profile likelihood. Let the distribution function be expressed as $f(\boldsymbol{x};\alpha,\boldsymbol{\beta})$ and the maximum likelihood function be expressed as $L(\alpha,\boldsymbol{\beta})$. Let $\hat{\alpha}(\boldsymbol{\beta})$ be the maximum likelihood estimate of α for a fixed $\boldsymbol{\beta}$. The profile likelihood for $\boldsymbol{\beta}$ is

$$L_p(\boldsymbol{\beta}) = L(\boldsymbol{\beta},\hat{\alpha}(\boldsymbol{\beta})) = sup_{\alpha}L(\boldsymbol{\beta},\alpha) \tag{1.33}$$

which is "the maximum likelihood function of a subset of parameters, $\boldsymbol{\beta}$, [or] the value of the log-likelihood function when the nuisance parameter, α, is replaced by its conditional maximum likelihood estimator, $\hat{\alpha}(\boldsymbol{\beta})$" [43]. Young and Smith stated "The profile likelihood [$L_p(\boldsymbol{\beta})$] can, to a considerable content, be thought of and used as if it were a genuine likelihood. In particular, the maximum likelihood estimate of [α] equals [$\hat{\alpha}$]" [56]. Patefield [43] demonstrated, in parametric models, the inverse of the observed information matrix of the profile likelihood $L_p(\boldsymbol{\beta})$) is equal to the $\boldsymbol{\beta}$ aspect of the inverse of the observed information matrix of the likelihood $L(\alpha, \boldsymbol{\beta})$.

Roeder, Carroll, and Lindsay [48] and Murphy and Van Der Vaart [42] confirmed Prentice and Pyke's results using the profile likelihood, assuming some covariates were measured with errors. They used a semiparametric approach and demonstrated that the model for the log odds of disease could be used in case-control studies to estimate the coefficients $\boldsymbol{\beta}$ in equation (1.33). They also provided a theoretical justification for the profile likelihood-based estimates and confidence intervals in the semiparametric model.

Seaman and Richardson [50] used an alternative approach where the covariates were measured without errors. They used the multinomial-Poisson transformation, Baker [4] had shown that "we can transform the multinomial likelihood into a Poisson likelihood, with additional parameters... [which] yields identical estimates and asymptotic variances". Seaman and Richardson [50] demonstrated that the profile likelihoods were identical for the coefficients $\boldsymbol{\beta}$ in equation (1.33) for both "The logistic model for disease incidence during the defined accession period"[47] and log odds of disease model.

1.3 The Dataset

To illustrate some of the theoretical methods in this thesis, data from the North American Symptomatic Carotid Endarterectomy Trial study (NASCET) are used [21][22][23][24]. NASCET was a clinical trial which investigated the effectiveness of carotid endarterectomy to reduce the risk of stroke. Because the purpose of using data in this thesis was to demonstrate the methods, a case-control study was constructed and the relationship between history of diabetes and myocardial infarction was explored. Subjects having a myocardial infarction (MI) event during the study period were considered as cases (209 cases) and a simple random sample of the same number of noncases (209 noncases) was considered as controls. The variables include:

ANYMIDT (D): the disease indicator (0 if not developing MI and 1 if developing MI),

DIABMEL (E): the exposure indicator (0 if not having a history of diabetes and 1 if having a history of diabetes),

AGE (A): the measured age variable,

SEX (G): the sex indicator (female=0 and male=1),

HYPERL (L): the hyperlipidemia history indicator (0 if not having a history of hyperlipidemia and 1 if having a history of hyperlipidemia),

HYPERT (H): the hypertension history indicator (0 if not having a history of hypertension and 1 if having a history of hypertension),

SMOKING (S): the smoking history indicator (0 if not having a history of smoking and 1 if having a history of smoking).

1.4 Overview of Thesis

In this thesis, the estimated odds ratios and corresponding variances for modeling the log odds of disease and for modeling the log odds of exposure for case-control studies will be compared. One disease variable D and one exposure variable E will be considered. Both variables are dichotomous taking on values 1 as presence and 0 as absence. Some covariables, relating to both exposure and disease, will also be considered as potential confounders or effect modifiers. These include age A (=0 if young, 1 if old, but sometimes A is treated as a measured variable instead depending on the context) and Gender G (=0 if female, 1 if male). It is also assumed that the models depend linearly on the unknown parameters. For example, the full model for the log odds of disease can be written as

$$\log \frac{Pr(D=1|E,A,G)}{1-Pr(D=1|E,A,G)} = \alpha+\beta E+\gamma_1 A+\gamma_2 G+\gamma_3 AG+\delta_1 EA+\delta_2 EG+\delta_3 EAG, \quad (1.34)$$

and the full model for the log odds of exposure can be written as

$$\log \frac{Pr(E=1|D,A,G)}{1-Pr(E=1|D,A,G)} = \upsilon+\tau D+\theta_1 A+\theta_2 G+\theta_3 AG+\zeta_1 DA+\zeta_2 DG+\zeta_3 DAG. \quad (1.35)$$

The primary aim is to compare the coefficients that are related to disease or exposure (β and τ, or δ_k and ζ_k), as well as, their variances.

The thesis will:

- determine when the estimated odds ratios and variances relating to exposure or disease from the two types of logistic regression models are equivalent.
- determine when the estimated odds ratios and variances relating to exposure or disease from the two types of logistic regression models are not equivalent.
- provide an expansion of Breslow and Powers' arguments [10] that the odds ratio estimates relating to exposure or disease from the two types of logistic regression

models are identical when data can be summarized as a series of 2×2 contingency tables.

- provide an expansion of Breslow's arguments [7] that the estimated odds ratio variances relating to exposure or disease from the two types of logistic regression models are identical when data can be summarized in a series of 2×2 contingency tables.

- determine when the profile likelihoods with respect to the odds ratios relating to exposure or disease from the two types of logistic regression models are the same by applying the multinomial-Poisson transformation [4].

- provide an expansion of Prentice and Pyke's arguments in 1979 [47] to show why models for the log odds of disease can be used in case-control studies.

Chapter 2

Examples for Equivalence of the Two Types of Logistic Regression Models in Case-Control Studies

This chapter provides three examples when the estimated coefficients and standard errors relating to exposure and disease from the two types of logistic regression models are equivalent.

2.1 Example 1: A single 2×2 Table

When the observations in a case-control study are classified by two levels of a risk factor (exposure), the data may be summarized in a single table (Table 2.1).

Table 2.1: The Single 2×2 Table (Frequencies)

	$E = 1$	$E = 0$
$D = 1$	a	b
$D = 0$	c	d

Let $\hat{p}_1$ represent the estimated probability that exposed members are a case and $\hat{p}_0$ represent the estimated probability that non-exposed members are a case. It follows from likelihood theory corresponding to a binomial distribution based on each category of exposure that the likelihood function

$$L = \binom{a+c}{a} p_1^a (1-p_1)^c \binom{b+d}{b} p_0^b (1-p_0)^d \tag{2.1}$$

and the log likelihood function

$$l \propto a \log p_1 + c \log(1-p_1) + b \log p_0 + d \log(1-p_0). \tag{2.2}$$

Taking the first derivatives with respect p_1 and p_0 and set them equal to zero, we obtain

$$\begin{aligned}\hat{p}_1 &= \hat{Pr}(D=1|E=1) = \frac{a}{a+c},\\ \hat{p}_0 &= \hat{Pr}(D=1|E=0) = \frac{b}{b+d}.\end{aligned} \tag{2.3}$$

Let the model for the log odds of disease be expressed as

$$\log \frac{Pr(D=1|E)}{1-Pr(D=1|E)} = \alpha + \beta E. \tag{2.4}$$

Then

$$\begin{aligned}\log \frac{\hat{p}_1}{1-\hat{p}_1} &= \hat{\alpha} + \hat{\beta},\\ \log \frac{\hat{p}_0}{1-\hat{p}_0} &= \hat{\alpha}.\end{aligned} \tag{2.5}$$

A simple calculation using (2.5) yields

$$\hat{\beta} = \log \frac{\hat{p}_1}{1-\hat{p}_1} - \log \frac{\hat{p}_0}{1-\hat{p}_0} = \log \frac{ad}{bc}. \tag{2.6}$$

Similarly, let $\hat{q}_1$ represent the estimated probability that diseased members are exposed and $\hat{q}_0$ represent the estimated probability that non-diseased members are exposed. Then

$$\begin{aligned}\hat{q}_1 &= \hat{Pr}(E=1|D=1) = \frac{a}{a+b},\\ \hat{q}_0 &= \hat{Pr}(E=1|D=0) = \frac{c}{c+d}.\end{aligned} \tag{2.7}$$

Let the model for the log odds of exposure be expressed as

$$\log \frac{Pr(E=1|D)}{1-Pr(E=1|D)} = \upsilon + \tau D. \tag{2.8}$$

Then

$$\begin{aligned}\log \frac{\hat{q}_1}{1-\hat{q}_1} &= \hat{\upsilon} + \hat{\tau},\\ \log \frac{\hat{q}_0}{1-\hat{q}_0} &= \hat{\upsilon}.\end{aligned} \tag{2.9}$$

A simple calculation using (2.9) yields

$$\hat{\tau} = \log \frac{\hat{q}_1}{1-\hat{q}_1} - \log \frac{\hat{q}_0}{1-\hat{q}_0} = \log \frac{ad}{bc}. \tag{2.10}$$

This demonstrates that $\hat{\beta} = \hat{\tau}$. The estimated coefficients relating to exposure and disease in the two types of logistic regression models are equivalent.

2.2 Example 2: Two 2×2 Tables

When a dichotomous exposure variable and a dichotomous covariate (e.g. gender) are considered, the data in a case-control study may be summarized in two 2×2 tables (Table 2.2).

Table 2.2: The Two 2×2 Tables (Frequencies)

	$G=0$		$G=1$	
	$E=1$	$E=0$	$E=1$	$E=0$
$D=1$	a_1	b_1	a_2	b_2
$D=0$	c_1	d_1	c_2	d_2

Let $\hat{p}_{1k}$ represent the estimated probability that exposed members in stratum k are a case and $\hat{p}_{0k}$ represent the estimated probability that non-exposed members in stratum k are a case. $k = 1$ or 2 indicates the 2×2 tables with $G = 0$ or $G = 1$. Following the same arguments as in section 2.1,

$$\begin{aligned}
\hat{p}_{11} &= \hat{Pr}(D=1|E=1,G=0) = \frac{a_1}{a_1+c_1}, \\
\hat{p}_{01} &= \hat{Pr}(D=1|E=0,G=0) = \frac{b_1}{b_1+d_1}, \\
\hat{p}_{12} &= \hat{Pr}(D=1|E=1,G=1) = \frac{a_2}{a_2+c_2}, \\
\hat{p}_{02} &= \hat{Pr}(D=1|E=0,G=1) = \frac{b_2}{b_2+d_2}.
\end{aligned} \tag{2.11}$$

Let the model for the log odds of disease be expressed as

$$\log \frac{Pr(D=1|E,G)}{1-Pr(D=1|E,G)} = \alpha + \beta E + \gamma G + \delta EG. \tag{2.12}$$

Then

$$\begin{aligned}
\log \frac{\hat{p}_{11}}{1-\hat{p}_{11}} &= \hat{\alpha} + \hat{\beta}, \\
\log \frac{\hat{p}_{01}}{1-\hat{p}_{01}} &= \hat{\alpha}, \\
\log \frac{\hat{p}_{12}}{1-\hat{p}_{12}} &= \hat{\alpha} + \hat{\beta} + \hat{\gamma} + \hat{\delta}, \\
\log \frac{\hat{p}_{02}}{1-\hat{p}_{02}} &= \hat{\alpha} + \hat{\gamma}.
\end{aligned} \tag{2.13}$$

A simple calculation using (2.13) yields

$$\begin{aligned}
\hat{\beta} &= \log \frac{\hat{p}_{11}}{1-\hat{p}_{11}} - \log \frac{\hat{p}_{01}}{1-\hat{p}_{01}} = \log \frac{a_1 d_1}{b_1 c_1}, \\
\hat{\delta} &= \log \frac{\hat{p}_{12}}{1-\hat{p}_{12}} - \log \frac{\hat{p}_{11}}{1-\hat{p}_{11}} - \log \frac{\hat{p}_{02}}{1-\hat{p}_{02}} + \log \frac{\hat{p}_{01}}{1-\hat{p}_{01}} = \log \frac{\dfrac{a_2 d_2}{b_2 c_2}}{\dfrac{a_1 d_1}{b_1 c_1}}.
\end{aligned} \tag{2.14}$$

In the same manner,

$$\begin{aligned}
\hat{q}_{11} &= \hat{Pr}(E=1|D=1,G=0) = \frac{a_1}{a_1+b_1}, \\
\hat{q}_{01} &= \hat{Pr}(E=1|D=0,G=0) = \frac{c_1}{c_1+d_1}, \\
\hat{q}_{12} &= \hat{Pr}(E=1|D=1,G=1) = \frac{a_2}{a_2+b_2}, \\
\hat{q}_{02} &= \hat{Pr}(E=1|D=0,G=1) = \frac{c_2}{c_2+d_2}.
\end{aligned} \tag{2.15}$$

Let the model for the log odds of exposure be expressed as

$$\log \frac{Pr(E=1|D,G)}{1-Pr(E=1|D,G)} = \upsilon + \tau D + \theta G + \zeta DG. \tag{2.16}$$

Then

$$\begin{aligned}
\log \frac{\hat{q}_{11}}{1-\hat{q}_{11}} &= \hat{\upsilon} + \hat{\tau}, \\
\log \frac{\hat{q}_{01}}{1-\hat{q}_{01}} &= \hat{\upsilon}, \\
\log \frac{\hat{q}_{12}}{1-\hat{q}_{12}} &= \hat{\upsilon} + \hat{\tau} + \hat{\theta} + \hat{\zeta}, \\
\log \frac{\hat{q}_{02}}{1-\hat{q}_{02}} &= \hat{\upsilon} + \hat{\theta}.
\end{aligned} \tag{2.17}$$

A simple calculation using (2.17) yields

$$\begin{aligned}
\hat{\beta} &= \log \frac{\hat{q}_{11}}{1-\hat{q}_{11}} - \log \frac{\hat{q}_{01}}{1-\hat{q}_{01}} = \log \frac{a_1 d_1}{b_1 c_1}, \\
\hat{\delta} &= (\log \frac{\hat{q}_{12}}{1-\hat{q}_{12}} - \log \frac{\hat{q}_{02}}{1-\hat{q}_{02}}) - (\log \frac{\hat{q}_{11}}{1-\hat{q}_{11}} - \log \frac{\hat{q}_{01}}{1-\hat{q}_{01}}) = \log \frac{\dfrac{a_2 d_2}{b_2 c_2}}{\dfrac{a_1 d_1}{b_1 c_1}}.
\end{aligned} \tag{2.18}$$

This demonstrates that $\hat{\beta} = \hat{\tau}$ and $\hat{\delta} = \hat{\zeta}$. The estimated coefficients relating to exposure and disease in the two types of logistic regression models are equivalent.

2.3 Example 3: Four 2×2 Tables

When a dichotomous exposure variable and two dichotomous covariates (e.g. gender and age) are considered in the model, the data may be summarized in four 2×2 tables (Table 2.3).

Table 2.3: The Four 2×2 Tables (Frequencies)

	$G=0$				$G=1$			
	$A=0$		$A=1$		$A=0$		$A=1$	
	$E=1$	$E=0$	$E=1$	$E=0$	$E=1$	$E=0$	$E=1$	$E=0$
$D=1$	a_1	b_1	a_2	b_2	a_3	b_3	a_4	b_4
$D=0$	c_1	d_1	c_2	d_2	c_3	d_3	c_4	d_4

Following directly from the previous sections,

$$\begin{aligned}
\hat{p}_{11} &= \hat{Pr}(D=1|E=1,G=0,A=0) = \frac{a_1}{a_1+c_1},\\
\hat{p}_{01} &= \hat{Pr}(D=1|E=0,G=0,A=0) = \frac{b_1}{b_1+d_1},\\
\hat{p}_{12} &= \hat{Pr}(D=1|E=1,G=1,A=0) = \frac{a_2}{a_2+c_2},\\
\hat{p}_{02} &= \hat{Pr}(D=1|E=0,G=1,A=0) = \frac{b_2}{b_2+d_2},\\
\hat{p}_{13} &= \hat{Pr}(D=1|E=1,G=0,A=1) = \frac{a_3}{a_3+c_3},\\
\hat{p}_{03} &= \hat{Pr}(D=1|E=0,G=0,A=1) = \frac{b_3}{b_3+d_3},\\
\hat{p}_{14} &= \hat{Pr}(D=1|E=1,G=1,A=1) = \frac{a_4}{a_4+c_4},\\
\hat{p}_{04} &= \hat{Pr}(D=1|E=0,G=1,A=1) = \frac{b_4}{b_4+d_4}. \qquad (2.19)
\end{aligned}$$

Let the model for the log odds of disease be expressed as

$$\log\frac{Pr(D=1|E,A,G)}{1-Pr(D=1|E,A,G)} = \alpha+\beta E+\gamma_1 A+\gamma_2 G+\gamma_3 AG+\delta_1 EA+\delta_2 EG+\delta_3 EAG. \quad (2.20)$$

Then

$$\begin{aligned}
\log\frac{\hat{p}_{11}}{1-\hat{p}_{11}} &= \hat{\alpha}+\hat{\beta},\\
\log\frac{\hat{p}_{01}}{1-\hat{p}_{01}} &= \hat{\alpha},\\
\log\frac{\hat{p}_{12}}{1-\hat{p}_{12}} &= \hat{\alpha}+\hat{\beta}+\hat{\gamma_2}+\hat{\delta_2},\\
\log\frac{\hat{p}_{02}}{1-\hat{p}_{02}} &= \hat{\alpha}+\hat{\gamma_2},\\
\log\frac{\hat{p}_{13}}{1-\hat{p}_{13}} &= \hat{\alpha}+\hat{\beta}+\hat{\gamma_1}+\hat{\delta_1},\\
\log\frac{\hat{p}_{03}}{1-\hat{p}_{03}} &= \hat{\alpha}+\hat{\gamma_1},\\
\log\frac{\hat{p}_{14}}{1-\hat{p}_{14}} &= \hat{\alpha}+\hat{\beta}+\hat{\gamma_1}+\hat{\gamma_2}+\hat{\gamma_3}+\hat{\delta_1}+\hat{\delta_2}+\hat{\delta_3},\\
\log\frac{\hat{p}_{04}}{1-\hat{p}_{04}} &= \hat{\alpha}+\hat{\gamma_1}+\hat{\gamma_2}+\hat{\gamma_3}. \qquad (2.21)
\end{aligned}$$

A simple calculation using (2.21) yields

$$\hat{\beta} = \log \frac{\hat{p}_{11}}{1-\hat{p}_{11}} - \log \frac{\hat{p}_{01}}{1-\hat{p}_{01}} = \log \frac{a_1 d_1}{b_1 c_1},$$

$$\hat{\delta_1} = (\log \frac{\hat{p}_{13}}{1-\hat{p}_{13}} - \log \frac{\hat{p}_{03}}{1-\hat{p}_{03}}) - (\log \frac{\hat{p}_{11}}{1-\hat{p}_{11}} + \log \frac{\hat{p}_{01}}{1-\hat{p}_{01}}) = \log \frac{\frac{a_3 d_3}{b_3 c_3}}{\frac{a_1 d_1}{b_1 c_1}},$$

$$\hat{\delta_2} = (\log \frac{\hat{p}_{12}}{1-\hat{p}_{12}} - \log \frac{\hat{p}_{02}}{1-\hat{p}_{02}}) - (\log \frac{\hat{p}_{11}}{1-\hat{p}_{11}} + \log \frac{\hat{p}_{01}}{1-\hat{p}_{01}}) = \log \frac{\frac{a_2 d_2}{b_2 c_2}}{\frac{a_1 d_1}{b_1 c_1}},$$

$$\begin{aligned}\hat{\delta_3} = &(\log \frac{\hat{p}_{14}}{1-\hat{p}_{14}} - \log \frac{\hat{p}_{04}}{1-\hat{p}_{04}}) - (\log \frac{\hat{p}_{12}}{1-\hat{p}_{12}} - \log \frac{\hat{p}_{02}}{1-\hat{p}_{02}}) \\ &- (\log \frac{\hat{p}_{13}}{1-\hat{p}_{13}} - \log \frac{\hat{p}_{03}}{1-\hat{p}_{03}}) + (\log \frac{\hat{p}_{11}}{1-\hat{p}_{11}} - \log \frac{\hat{p}_{01}}{1-\hat{p}_{01}}) = \log \frac{\frac{a_4 d_4}{b_4 c_4}\frac{a_1 d_1}{b_1 c_1}}{\frac{a_2 d_2}{b_2 c_2}\frac{a_3 d_3}{b_3 c_3}}. \quad (2.22)\end{aligned}$$

Similarly,

$$\begin{aligned}\hat{q}_{11} &= \hat{Pr}(E=1|D=1,G=0,A=0) = \frac{a_1}{a_1+b_1}, \\ \hat{q}_{01} &= \hat{Pr}(E=1|D=0,G=0,A=0) = \frac{c_1}{c_1+d_1}, \\ \hat{q}_{12} &= \hat{Pr}(E=1|D=1,G=1,A=0) = \frac{a_2}{a_2+b_2}, \\ \hat{q}_{02} &= \hat{Pr}(E=1|D=0,G=1,A=0) = \frac{c_2}{c_2+d_2}, \\ \hat{q}_{13} &= \hat{Pr}(E=1|D=1,G=0,A=1) = \frac{a_3}{a_3+b_3}, \\ \hat{q}_{03} &= \hat{Pr}(E=1|D=0,G=0,A=1) = \frac{c_3}{c_3+d_3}, \\ \hat{q}_{14} &= \hat{Pr}(E=1|D=1,G=1,A=1) = \frac{a_4}{a_4+c_b}, \\ \hat{q}_{04} &= \hat{Pr}(E=1|D=0,G=1,A=1) = \frac{c_4}{c_4+d_4}. \quad (2.23)\end{aligned}$$

Let the model for the log odds of exposure be expressed as

$$\log \frac{Pr(E=1|D,A,G)}{1-Pr(E=1|D,A,G)} = \upsilon + \tau D + \theta_1 A + \theta_2 G + \theta_3 AG + \zeta_1 DA + \zeta_2 DG + \zeta_3 DAG. \quad (2.24)$$

Then

$$\begin{aligned}
\log\frac{\hat{q}_{11}}{1-\hat{q}_{11}} &= \hat{\upsilon}+\hat{\tau},\\
\log\frac{\hat{q}_{01}}{1-\hat{q}_{01}} &= \hat{\upsilon},\\
\log\frac{\hat{q}_{12}}{1-\hat{q}_{12}} &= \hat{\upsilon}+\hat{\tau}+\hat{\theta}_2+\hat{\zeta}_2,\\
\log\frac{\hat{q}_{02}}{1-\hat{q}_{02}} &= \hat{\upsilon}+\hat{\theta}_2,\\
\log\frac{\hat{q}_{13}}{1-\hat{q}_{13}} &= \hat{\upsilon}+\hat{\tau}+\hat{\theta}_1+\hat{\zeta}_1,\\
\log\frac{\hat{q}_{03}}{1-\hat{q}_{03}} &= \hat{\upsilon}+\hat{\theta}_1,\\
\log\frac{\hat{q}_{14}}{1-\hat{q}_{14}} &= \hat{\upsilon}+\hat{\tau}+\hat{\theta}_1+\hat{\theta}_2+\hat{\theta}_3+\hat{\zeta}_1+\hat{\zeta}_2+\hat{\zeta}_3,\\
\log\frac{\hat{q}_{04}}{1-\hat{q}_{04}} &= \hat{\upsilon}+\hat{\theta}_1+\hat{\theta}_2+\hat{\theta}_3. \qquad (2.25)
\end{aligned}$$

A simple calculation using (2.25) yields

$$\begin{aligned}
\hat{\tau} &= \log\frac{\hat{q}_{11}}{1-\hat{q}_{11}}-\log\frac{\hat{q}_{01}}{1-\hat{q}_{01}} = \log\frac{a_1d_1}{b_1c_1},\\
\hat{\zeta}_1 &= (\log\frac{\hat{q}_{13}}{1-\hat{q}_{13}}-\log\frac{\hat{q}_{03}}{1-\hat{q}_{03}})-(\log\frac{\hat{q}_{11}}{1-\hat{q}_{11}}+\log\frac{\hat{q}_{01}}{1-\hat{q}_{01}}) = \log\frac{\dfrac{a_3d_3}{b_3c_3}}{\dfrac{a_1d_1}{b_1c_1}},\\
\hat{\zeta}_2 &= (\log\frac{\hat{q}_{12}}{1-\hat{q}_{12}}-\log\frac{\hat{q}_{02}}{1-\hat{q}_{02}})-(\log\frac{\hat{q}_{11}}{1-\hat{q}_{11}}+\log\frac{\hat{q}_{01}}{1-\hat{q}_{01}}) = \log\frac{\dfrac{a_2d_2}{b_2c_2}}{\dfrac{a_1d_1}{b_1c_1}},\\
\hat{\zeta}_3 &= (\log\frac{\hat{q}_{14}}{1-\hat{q}_{14}}-\log\frac{\hat{q}_{04}}{1-\hat{q}_{04}})-(\log\frac{\hat{q}_{12}}{1-\hat{q}_{12}}-\log\frac{\hat{q}_{02}}{1-\hat{q}_{02}})\\
&\quad-(\log\frac{\hat{q}_{13}}{1-\hat{q}_{13}}-\log\frac{\hat{q}_{03}}{1-\hat{q}_{03}})+(\log\frac{\hat{q}_{11}}{1-\hat{q}_{11}}-\log\frac{\hat{q}_{01}}{1-\hat{q}_{01}}) = \log\frac{\dfrac{a_4d_4}{b_4c_4}\dfrac{a_1d_1}{b_1c_1}}{\dfrac{a_2d_2}{b_2c_2}\dfrac{a_3d_3}{b_3c_3}}. \qquad (2.26)
\end{aligned}$$

This demonstrates that $\hat{\beta}=\hat{\tau}$ and $\hat{\delta}_k=\hat{\zeta}_k$ $(k=1,2,3)$. The estimated coefficients relating to exposure and disease in the two types of logistic regression models are equivalent.

2.4 Data Illustrations

The following data examples will illustrate the equivalence from the preceding sections. The data are from Chapter 1 Section 1.3.

- Example 1

  This example demonstrates that the estimated coefficients relating to exposure (history of diabetes) and disease (MI) in the two types of logistic regression models in Section 2.1 are equivalent. Table 2.4 reports the frequencies in each cell from a single 2×2 table. As observed from Table 2.5, the estimated odds of developing MI in the group with a history of diabetes is 2.0196 (95% CI: 1.2918, 3.1575) times the estimated odds of developing MI in the group without a history of diabetes. These results are equivalent to those in Table 2.6 using the odds of exposure (history of diabetes). That is, the estimated coefficients relating to exposure (history of diabetes) and disease (MI) in the two types of logistic regression models are the same (0.7029), as well as the standard errors (0.2280).

Table 2.4: The Single 2×2 Table (Frequencies)

	$E = 1$	$E = 0$
$D = 1$	69	140
$D = 0$	41	168
$\hat{OR}$	2.0195	

Table 2.5: Estimated Coefficients and Standard Errors from Fitting the Model for the Log Odds of Disease (The Single Table)

Variable	Coefficient	Std. Err.
exp	**0.7029**	**(0.2280)**
Intercept	-0.1823	(0.1144)

Table 2.6: Estimated Coefficients and Standard Errors from Fitting the Model for the Log Odds of Exposure (The Single Table)

Variable	Coefficient	Std. Err.
dis	**0.7029**	**(0.2280)**
Intercept	-1.4104	(0.1742)

- Example 2

  This example demonstrates that the estimated coefficients relating to exposure and disease in the two types of logistic regression models in Section 2.2 are equivalent. Table 2.7 reports the frequencies in each cell in the two 2×2 tables. For females, the estimated odds of developing MI in the group with a history of diabetes is 3.0535 (95% CI: 1.4397, 6.4763) times the estimated odds of developing MI in the group without a history of diabetes (Table 2.8). These results are equivalent to those in Table 2.9 when modeling the odds of exposure (history of diabetes). That is, the estimated coefficients relating to exposure (history of diabetes) and disease (MI) in the two types of logistic regression models are the same (1.1163), as well as the standard errors (0.3836).

  ‘

Table 2.7: The Two 2×2 Tables (Frequencies)

	$G = 0$		$G = 1$	
	$E = 1$	$E = 0$	$E = 1$	$E = 0$
$D = 1$	28	27	41	113
$D = 0$	18	53	23	115
$\hat{OR}$	3.0535		1.8141	

Table 2.8: Estimated Coefficients and Standard Errors from Fitting the Model for the Log Odds of Disease (The Two Tables)

Variable	Coefficient	Std. Err.
exp	**1.1163**	**(0.3836)**
gender	0.6569	(0.2710)
eg	**-0.5207**	**(0.4823)**
Intercept	-0.6745	(0.2364)

Table 2.9: Estimated Coefficients and Standard Errors from Fitting the Model for the Log Odds of Exposure (The Two Tables)

Variable	Coefficient	Std. Err.
dis	**1.1163**	**(0.3836)**
gender	-0.5295	(0.3558)
dg	**-0.5207**	**(0.4823)**
Intercept	-1.0799	(0.2728)

- Example 3

  This example demonstrates that the estimated coefficients relating to exposure and disease in the two types of logistic regression models in Section 2.3 are equivalent. Table 2.10 reports the frequencies in each cell in the four 2×2 tables. For young females, the estimated odds of developing MI in the group with a history of diabetes is 4.9998 (95% CI:1.2744, 19.6152) times the estimated odds of developing MI in the group without a history of diabetes (Table 2.11). As in the previous two examples, the estimated coefficients relating to exposure and disease are identical between Table 2.11 and Table 2.12.

Table 2.10: The Four 2×2 Tables (Frequencies)

	$G=0$				$G=1$			
	$A=0$		$A=1$		$A=0$		$A=1$	
	$E=1$	$E=0$	$E=1$	$E=0$	$E=1$	$E=0$	$E=1$	$E=0$
$D=1$	10	11	18	16	17	40	24	73
$D=0$	4	22	14	31	13	41	10	74
$\hat{OR}$	5		2.4911		1.3404		2.4329	

Table 2.11: Estimated Coefficients and Standard Errors from Fitting the Model for the Log Odds of Disease (The Four Tables)

Variable	Coefficient	Std. Err.
exp	**1.6094**	**(0.6974)**
agegrp	0.0317	(0.4808)
gender	0.6685	(0.4310)
ag	-0.0207	(0.5547)
ea	**-0.6967**	**(0.8415)**
eg	**-1.3165**	**(0.8195)**
eag	**1.2928**	**(1.0306)**
Intercept	-0.6931	(0.3693)

Table 2.12: Estimated Coefficients and Standard Errors from Fitting the Model for the Log Odds of Exposure (The Four Tables)

Variable	Coefficient	Std. Err.
dis	**1.6094**	**(0.6974)**
agegrp	0.9098	(0.6318)
gender	0.5561	(0.6299)
ag	-1.7627	(0.7836)
da	**-0.6967**	**(0.8415)**
dg	**-1.3165**	**(0.8195)**
dag	**1.2928**	**(1.0306)**
Intercept	-1.7047	(0.5436)

Chapter 3

Examples for Lack of Equivalence of the Two Types of Logistic Regression Models in Case-Control Studies

This chapter provides four examples when the estimated coefficients and standard errors relating to exposure and disease from the two types of logistic regression models are not equivalent.

3.1 Example 1: Excluding the Cross-product Term of Age and Gender when Age and Gender are Considered as Potential Confounders

In this example, both binary age and gender are considered as potential confounders but not joint confounders. The model for the log odds of disease can be written as

$$\log \frac{Pr(D=1|E,A,G)}{1-Pr(D=1|E,A,G)} = \alpha + \beta E + \gamma_1 A + \gamma_2 G \tag{3.1}$$

and the model for the log odds of exposure can be written as

$$\log \frac{Pr(E=1|D,A,G)}{1-Pr(E=1|D,A,G)} = \upsilon + \tau D + \theta_1 A + \theta_2 G. \tag{3.2}$$

The results in Tables 3.1 and 3.2 show that the estimated coefficients relating to exposure (history of diabetes) and disease (MI) in the two types of logistic regression models are not the same (0.7934 vs. 0.7932 respectively), and neither are the standard errors.

Table 3.1: Estimated Coefficients and Standard Errors from Fitting the Model for the Log Odds of Disease (without AG)

Variable	Coefficient	Std. Err.
exp	**0.7934**	**(0.2339)**
agegrp	0.0643	(0.2057)
gender	0.4949	(0.2222)
Intercept	-0.5923	(0.2427)

Table 3.2: Estimated Coefficients and Standard Errors from Fitting the Model for the Log Odds of Exposure (without AG)

Variable	Coefficient	Std. Err.
dis	**0.7932**	**(0.2340)**
agegrp	-0.1445	(0.2339)
gender	-0.8145	(0.2395)
Intercept	-0.8298	(0.2623)

3.2 Example 2: Excluding the Cross-product Term of Age and Gender when Age and Gender are Considered as Potential Modifiers

In this example, both binary age and gender are considered as potential modifiers but not joint modifiers. The log odds of disease can be written as

$$\log \frac{Pr(D=1|E,A,G)}{1-Pr(D=1|E,A,G)} = \alpha + \beta E + \gamma_1 A + \gamma_2 G + \delta_1 EA + \delta_2 EG \tag{3.3}$$

and the log odds of exposure can be written as

$$\log \frac{Pr(E=1|D,A,G)}{1-Pr(E=1|D,A,G)} = \upsilon + \tau D + \theta_1 A + \theta_2 G + \zeta_1 DA + \zeta_2 DG. \tag{3.4}$$

The results in Tables 3.3 and 3.4 show that the estimated coefficients and standard errors relating to exposure (history of diabetes) and disease (MI) in the two types of logistic regression models are not the same.

Table 3.3: Estimated Coefficients and Standard Errors from Fitting the Model for the Log Odds of Disease (without AG and Modifiers)

Variable	Coefficient	Std. Err.
exp	**1.0404**	**(0.4964)**
agegrp	0.0162	(0.2398)
gender	0.6560	(0.2714)
ea	**0.1071**	**(0.4731)**
eg	**-0.4993**	**(0.4874)**
Intercept	-0.6840	(0.2754)

Table 3.4: Estimated Coefficients and Standard Errors from Fitting the Model for the Log Odds of Exposure (without AG and Modifiers)

Variable	Coefficient	Std. Err.
dis	**1.0714**	**(0.4820)**
agegrp	-0.1813	(0.3570)
gender	-0.5348	(0.3562)
da	**0.0712**	**(0.4735)**
dg	**-0.5148**	**(0.4827)**
Intercept	-0.9669	(0.3499)

3.3 Example 3: Age has Three Values and is Considered as a Potential Confounder

The following example illustrates the simplest case where the potential confounder age (Am) has more than two possible values, demonstrating that the coefficients relating to exposure and disease are not equivalent. Without loss of generality, consider age to have three possible values (=0,1,2). The model for the log odds of disease can be written as

$$\log \frac{Pr(D=1|E,A_m)}{1-Pr(D=1|E,A_m)} = \alpha + \beta E + \gamma A_m \tag{3.5}$$

and the model for the log odds of exposure can be written as

$$\log \frac{Pr(E=1|D,A_m)}{1-Pr(E=1|D,A_m)} = \upsilon + \tau D + \theta A_m. \tag{3.6}$$

The results in Tables 3.5 and 3.6 show that the estimated coefficients and standard errors relating to exposure (history of diabetes) and disease (MI) in the two types of logistic

regression models are not the same.

Table 3.5: Estimated Coefficients and Standard Errors from Fitting the Model for the Log Odds of Disease (Age with 3 Values)

Variable	**Coefficient**	**Std. Err.**
exp	**0.7110**	**(0.2284)**
Am	0.1458	(0.1768)
Intercept	-0.3555	(0.2395)

Table 3.6: Estimated Coefficients and Standard Errors from Fitting the Model for the Log Odds of Exposure (Age with 3 Values)

Variable	**Coefficient**	**Std. Err.**
dis	**0.7115**	**(0.2285)**
Am	-0.1821	(0.2020)
Intercept	-1.2037	(0.2851)

However, the estimated coefficients and standard errors relating to exposure (history of diabetes) and disease (MI) in the two types of logistic regression models (3.7 and 3.8) are the same when age squared (A_m^2) is included in the models (Tables 3.7 and 3.8 respectively). In this case, the model for the log odds of disease can be written as

$$\log \frac{Pr(D=1|E,A_m)}{1-Pr(D=1|E,A_m)} = \alpha + \beta E + \gamma_1 A_m + \gamma_2 A_m^2 \tag{3.7}$$

and the model for the log odds of exposure can be written as

$$\log \frac{Pr(E=1|D,A_m)}{1-Pr(E=1|D,A_m)} = \upsilon + \tau D + \theta_1 A_m + \theta_2 A_m^2. \tag{3.8}$$

The feature of adding the Am^2 to the model illustrates the principle of "spanning the same linear space" that will be discussed further in Chapter 4.

Table 3.7: Estimated Coefficients and Standard Errors from Fitting the Model for the Log Odds of Disease (Age with 3 Values Squared)

Variable	Coefficient	Std. Err.
exp	**0.7026**	**(0.2289)**
Am	0.4599	(0.5767)
Am^2	-0.1329	(0.2319)
Intercept	-0.4971	(0.3458)

Table 3.8: Estimated Coefficients and Standard Errors from Fitting the Model for the Log Odds of Exposure (Age with 3 Values Squared)

Variable	Coefficient	Std. Err.
dis	**0.7026**	**(0.2289)**
Am	0.6458	(0.6760)
Am^2	-0.3533	(0.2709)
Intercept	-1.5788	(0.4218)

3.4 Example 4: Age has Three Values and is Considered as a Potential Effect Modifier

In this example, the age (A_m) variable has three possible values (=0,1,2) and is considered as a potential effect modifier. The model for the log odds of disease can be written as

$$\log \frac{Pr(D=1|E,A_m)}{1-Pr(D=1|E,A_m)} = \alpha + \beta E + \gamma A_m + \delta E A_m \tag{3.9}$$

and the model for the log odds of exposure can be written as

$$\log \frac{Pr(E=1|D,A_m)}{1-Pr(E=1|D,A_m)} = \upsilon + \tau D + \theta A_m + \zeta D A_m. \tag{3.10}$$

The results in Tables 3.9 and 3.10 show that the estimated coefficients and standard errors relating to exposure (history of diabetes) and disease (MI) in the two types of logistic regression models are not the same.

Table 3.9: Estimated Coefficients and Standard Errors from Fitting the Model for the Log Odds of Disease (Age with 3 Values and modifier)

Variable	**Coefficient**	**(Std. Err.)**
exp	**0.5804**	**(0.5447)**
Am	0.1218	(0.1988)
exp × Am	**0.1148**	**(0.4358)**
Intercept	-0.3269	(0.2626)

Table 3.10: Estimated Coefficients and Standard Errors from Fitting the Model for the Log Odds of Exposure (Age with 3 Values and modifier)

Variable	**Coefficient**	**(Std. Err.)**
dis	**0.6556**	**(0.5167)**
Am	-0.2094	(0.3039)
dis × Am	**0.0490**	**(0.4067)**
Intercept	-1.1733	(0.3801)

Chapter 4

Equivalence of the Estimated Coefficients relating to Exposure and Disease in the Two Types of Logistic Regression Models

This chapter will examine the theoretical basis when the estimated coefficients relating to exposure and disease in the two types of logistic regression models are equivalent and when they are not equivalent.

4.1 Equivalence when both Models can be Expressed as a Classical Stratified Analysis

Suppose the data are summarized in s 2×2 tables categorized by the cross-classification of covariates. Table 4.1 is the k^{th} table where $k = 1, ..., s$.

Table 4.1: The k^{th} 2×2 Table (Frequencies)

	$E = 1$	$E = 0$	Row Total
$D = 1$	a_k	b_k	$a_k + b_k$
$D = 0$	c_k	d_k	$c_k + d_k$
Column Total	$a_k + c_k$	$b_k + d_k$	

Let $\hat{p}_{1k}$ represent the estimated probability that exposed members in stratum k are a case and $\hat{p}_{0k}$ represent the estimated probability that non-exposed members in stratum k are a case. It follows from section 2.1 that

$$\begin{aligned} \hat{p}_{1k} &= \hat{Pr}(D = 1|E = 1, k) = \frac{a_k}{a_k + c_k}, \\ \hat{p}_{0k} &= \hat{Pr}(D = 1|E = 0, k) = \frac{b_k}{b_k + d_k}. \end{aligned} \tag{4.1}$$

If each covariate pattern is designated by the indicator variable I_k for $k = 1, ..., s$, where

$$I_k = \begin{cases} 1 & \text{if it is the } k^{th} \text{ covariate pattern} \\ 0 & \text{otherwise.} \end{cases}$$

Then the model for the log odds of disease can be expressed as

$$\log \frac{Pr(D=1|E,k)}{1-Pr(D=1|E,k)} = \beta E + \sum_{k=1}^{s} \gamma_k I_k + \sum_{k=1}^{s} \delta_k E I_k \tag{4.2}$$

where at least one δ_k is zero (the baseline). Then

$$\begin{aligned} \log \frac{\hat{p}_{1k}}{1-\hat{p}_{1k}} &= \hat{\beta} + \hat{\gamma}_k + \hat{\delta}_k, \\ \log \frac{\hat{p}_{0k}}{1-\hat{p}_{0k}} &= \hat{\gamma}_k. \end{aligned} \tag{4.3}$$

Suppose $\delta_1 = 0$ is the baseline, a simple calculation using (4.3) yields

$$\begin{aligned} \hat{\beta} &= \log \frac{\hat{p}_{11}}{1-\hat{p}_{11}} - \log \frac{\hat{p}_{01}}{1-\hat{p}_{01}} = \log \frac{a_1 d_1}{b_1 c_1}, \\ \hat{\delta_k} &= (\log \frac{\hat{p}_{1k}}{1-\hat{p}_{1k}} - \log \frac{\hat{p}_{0k}}{1-\hat{p}_{0k}}) - (\log \frac{\hat{p}_{11}}{1-\hat{p}_{11}} - \log \frac{\hat{p}_{01}}{1-\hat{p}_{01}}) = \log \frac{\frac{a_k d_k}{b_k c_k}}{\frac{a_1 d_1}{b_1 c_1}}. \end{aligned} \tag{4.4}$$

Similarly,

$$\begin{aligned} \hat{q}_{1k} &= \hat{Pr}(E=1|D=1,k) = \frac{a_k}{a_k+b_k}, \\ \hat{q}_{0k} &= \hat{Pr}(E=1|D=0,k) = \frac{c_k}{c_k+d_k}. \end{aligned} \tag{4.5}$$

The model for the log odds of exposure can be expressed as

$$\log \frac{Pr(E=1|D,k)}{1-Pr(E=1|D,k)} = \tau D + \sum_{k=1}^{s} \theta_k I_k + \sum_{k=1}^{s} \zeta_k D I_k. \tag{4.6}$$

Then

$$\begin{aligned} \log \frac{\hat{q}_{1k}}{1-\hat{q}_{1k}} &= \hat{\tau} + \hat{\theta}_k + \hat{\zeta}_k, \\ \log \frac{\hat{q}_0}{1-\hat{q}_0} &= \hat{\theta}_k. \end{aligned} \tag{4.7}$$

Suppose $\delta_1 = 0$ is the baseline, a simple calculation using (4.7) yields

$$\hat{\beta} = \log \frac{\hat{q}_{11}}{1-\hat{q}_{11}} - \log \frac{\hat{q}_{01}}{1-\hat{q}_{01}} = \log \frac{a_1 d_1}{b_1 c_1},$$

$$\hat{\delta_k} = (\log \frac{\hat{q}_{1k}}{1-\hat{q}_{1k}} - \log \frac{\hat{q}_{0k}}{1-\hat{q}_{0k}}) - (\log \frac{\hat{q}_{11}}{1-\hat{q}_{11}} - \log \frac{\hat{q}_{01}}{1-\hat{q}_{01}}) = \log \frac{\frac{a_k d_k}{b_k c_k}}{\frac{a_1 d_1}{b_1 c_1}}. \tag{4.8}$$

This demonstrates that $\hat{\beta} = \hat{\tau}$ and $\hat{\delta_k} = \hat{\zeta_k}$. The estimated coefficients relating to exposure and disease in the two types of logistic regression models are equivalent. This equivalence is verified in the following example, whereby the estimated coefficients and standard errors corresponding to exposure in Table 4.2 are identical to the estimated coefficients and standard errors corresponding to disease in Table 4.3.

Table 4.2: Estimated Coefficients and Standard Errors from Fitting the Model for the Log Odds of Disease (A Series of Tables)

Variable	Coefficient	Std. Err.
exp	**1.6094**	**(0.6974)**
young female (I_1)	-0.6931	(0.3693)
old female (I_2)	0.0317	(0.4808)
young male (I_3)	0.6685	(0.4310)
old male (I_4)	0.6795	(0.4044)
exp $\times$ old female (I_2)	**-0.6967**	**(0.8415)**
exp $\times$ young male (I_3)	**-1.3165**	**(0.8195)**
exp $\times$ old male (I_4)	**-0.7204**	**(0.8095)**

Table 4.3: Estimated Coefficients and Standard Errors from Fitting the Model for the Log Odds of Exposure (A Series of Tables)

Variable	Coefficient	Std. Err.
dis	**1.6094**	**(0.6974)**
young female (I_1)	-1.7047	(0.5436)
old female (I_2)	0.9098	(0.6318)
young male (I_3)	0.5561	(0.6299)
old male (I_4)	-0.2967	(0.6395)
dis $\times$ old female (I_2)	**-0.6967**	**(0.8415)**
dis $\times$ young male (I_3)	**-1.3165**	**(0.8195)**
dis $\times$ old male (I_4)	**-0.7204**	**(0.8095)**

4.2 Definition of Saturation for Models Expressed in Terms of Design Matrices

This section will use the following two theorems in proving equivalence and non-equivalence. From Graybill [25]

- Let $\boldsymbol{V} = [\boldsymbol{v_1}, \boldsymbol{v_2}, ..., \boldsymbol{v_m}]$ be a matrix consisting of a set of vectors that is a basis for $\mathbb{V}_n$ and let $\boldsymbol{U} = [\boldsymbol{u_1}, \boldsymbol{u_2}, ..., \boldsymbol{u_q}]$ be a matrix that is any set of vectors in $\mathbb{V}_n$. The set of vectors in $\boldsymbol{U}$ is a basis set for $\mathbb{V}_n$ if and only if $m = q$ and there exists a nonsingular $m \times m$ matrix $\boldsymbol{A}$ such that $\boldsymbol{U} = \boldsymbol{V}\boldsymbol{A}$, Theorem. 5.4.5.

- Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be $n \times m$ matrices. There exists a nonsingular $m \times m$ matrix $\boldsymbol{C}$ such that $\boldsymbol{AC} = \boldsymbol{B}$ if and only if $\boldsymbol{A}$ and $\boldsymbol{B}$ have the same column space, Theorem 2.5.6.

In studying the relationship between a disease and its exposure, covariates need to be considered for their potential confounding or modifying effects [39]. Suppose that A (=0 if young, 1 if old) and G (=0 if female, 1 if male) are covariates. Let

$$I_1 = \begin{cases} 1 & \text{if } A = 0, G = 0 \\ 0 & \text{otherwise,} \end{cases}$$

$$I_2 = \begin{cases} 1 & \text{if } A = 1, G = 0 \\ 0 & \text{otherwise,} \end{cases}$$

$$I_3 = \begin{cases} 1 & \text{if } A = 0, G = 1 \\ 0 & \text{otherwise,} \end{cases}$$

$$I_4 = \begin{cases} 1 & \text{if } A = 1, G = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Suppose the model for the log odds of disease can be expressed as

$$\log \frac{P_(D=1|E,k)}{1-Pr(D=1|E,k)} \;=\; \beta E \;+\; \sum_{k=1}^{4} \gamma_k I_k \;+\; \sum_{k=1}^{4} \delta_k E I_k \tag{4.9}$$

where at least one δ_k is zero (the baseline). The design matrix of the model for the log odds of disease (4.9) with $\delta_1 = 0$ can be expressed as

$$\begin{array}{c} E \;\; I_1 \;\; I_2 \;\; I_3 \;\; I_4 \;\; EI_2 \;\; EI_3 \;\; EI_4 \\ \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ . & . & . & . & . & . & . & . \end{bmatrix} \end{array}$$

where each column shows the status of the designated variables or the cross-product of them. Each row is for one observation so that there are ($\sum_{k=1}^{4} n_k = n$) rows (n_k is the sample size for stratum k).

Suppose that $\delta_k = 0$ for $k = 1, ..., 4$, then the model (4.9) becomes

$$\log \frac{Pr(D=1|E,k)}{1-Pr(D=1|E,k)} \;=\; \beta E \;+\; \sum_{k=1}^{4} \gamma_k I_k. \tag{4.10}$$

An abbreviated version of the matrix relating to columns I_k ($k = 1, ..., 4$) in the design matrix for model (4.10) can be written as

$$\begin{array}{c} I_1 \;\; I_2 \;\; I_3 \;\; I_4 \\ \boldsymbol{I} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

where each row represents the membership status to one of the four covariate patterns formulated by the cross-classification of covariates A and G (young female, old female, young male, and old male). Breslow and Powers made a pivotal observation by stating that "the covariate effects are saturated with parameters" [10] when "the number of independent parameters [γ_1, γ_2, γ_3, and γ_4] equals the number of covariate patterns" [10]. For example, there are four covariate patterns in model (4.10).

Breslow and Powers' observation can also be applied to other forms of covariate coding systems. For example, let

$$Z_1 = \begin{cases} 1 & \text{if } A = 1, G = 0 \\ 0 & \text{otherwise,} \end{cases}$$

$$Z_2 = \begin{cases} 1 & \text{if } A = 0, G = 1 \\ 0 & \text{otherwise,} \end{cases}$$

$$Z_3 = \begin{cases} 1 & \text{if } A = 1, G = 1 \\ 0 & \text{otherwise.} \end{cases}$$

The corresponding abbreviated matrix $\boldsymbol{Z}$ with columns 1, Z_1, Z_2, and Z_3 can be expressed as

$$\boldsymbol{Z} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

where each row represents the membership status to one of the four covariate patterns. Using the theorems from Graybill [25], $\boldsymbol{Z} = \boldsymbol{I}\boldsymbol{R}$ where $\boldsymbol{R}$ is a nonsingular matrix (because $\boldsymbol{R} = \boldsymbol{Z}$ and $\boldsymbol{Z}$ is a nonsingular matrix). Therefore, the column spaces of the two matrices span the same linear space. Thus, $\boldsymbol{Z}$ can be used to define yet another parameterization

of covariate effects. Model (4.9) can be rewritten as

$$\log \frac{Pr(D=1|E,k)}{1-Pr(D=1|E,k)} = \alpha + \beta E + \sum_{k=1}^{3} \gamma_k Z_k. \tag{4.11}$$

or equivalently as

$$\log \frac{Pr(D=1|E,A,G)}{1-Pr(D=1|E,A,G)} = \alpha + \beta E + \gamma_1 A + \gamma_2 G + \gamma_3 AG. \tag{4.12}$$

Suppose the four covariate patterns in model (4.12) can be expressed by a matrix given by

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix},$$

where each row of $\boldsymbol{X}$ represents the membership status to one of the four covariate patterns. The columns of $\boldsymbol{X}$ correspond to 1, A, G, and AG. Notice that $\boldsymbol{X} = \boldsymbol{I}\,\boldsymbol{R}$ where $\boldsymbol{R}$ is a nonsingular matrix (because $\boldsymbol{R} = \boldsymbol{X}$ and $\boldsymbol{X}$ is a nonsingular matrix). Therefore, the column spaces of the two matrices span the same linear space.

Because the column spaces of $\boldsymbol{I}$, $\boldsymbol{Z}$, and $\boldsymbol{X}$ span the same linear space, this implies that

- the three types of covariate coding systems saturate the model with parameters for covariate effects.
- the three types of covariate coding systems yield equivalent estimated odds ratios and variance for a given stratum.

Breslow and Powers' "The covariate effects are saturated with parameters" [10] concept is examined further in the following two examples that remove specific parameters from

the model. In the first example, removal of a parameter results in a saturated model (with respect to covariate effects), whereas in the second example it does not.

4.2.1 When $\gamma_2 = 0$ (Corresponding to Gender)

When $\gamma_2 = 0$, model (4.12) is written as

$$\log \frac{Pr(D=1|E,A,G)}{1-Pr(D=1|E,A,G)} \;=\; \alpha \;+\; \beta E \;+\; \gamma_1 A \;+\; \gamma_3 AG. \tag{4.13}$$

The corresponding matrix $\boldsymbol{X}$ will be

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}.$$

Each row of $\boldsymbol{X}$ represents the membership status to one of the three covariate patterns (young, old female, and old male). The columns of $\boldsymbol{X}$ correspond to 1, A, and AG. Notice that $\boldsymbol{X} = \boldsymbol{I}\,\boldsymbol{R}$ where $\boldsymbol{R}$ is a nonsingular matrix (because $\boldsymbol{R} = \boldsymbol{X}$ and $\boldsymbol{X}$ is a nonsingular matrix). Therefore, the column spaces of the two matrices $\boldsymbol{X}$ and $\boldsymbol{I}$ span the same linear space. Alternatively, let

$$Z_1 = \begin{cases} 1 & \text{if } A = 1, G = 0 \\ 0 & \text{otherwise,} \end{cases}$$

$$Z_2 = \begin{cases} 1 & \text{if } A = 1, G = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then the corresponding matrix $\boldsymbol{Z}$ can be written as

$$\boldsymbol{Z} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

where each row of $\boldsymbol{Z}$ represents the membership status to one of the three covariate patterns. The columns of $\boldsymbol{Z}$ correspond to 1, Z_1, and Z_2. Notice $\boldsymbol{Z} = \boldsymbol{I}\,\boldsymbol{R}$ where $\boldsymbol{R}$ is a nonsingular matrix (because $\boldsymbol{R} = \boldsymbol{Z}$ and $\boldsymbol{Z}$ is a nonsingular matrix). Therefore, the column spaces of the two matrices span the same linear space. As column spaces of $\boldsymbol{Z}$, and $\boldsymbol{X}$ span the same linear space as the column space of $\boldsymbol{I}$, the covariate effects are saturated.

4.2.2 When $\gamma_3 = 0$ (Corresponding to the Cross-Product of Age and Gender)

The term AG in model (4.12) is the cross-product between A and G. When $\gamma_3 = 0$, model (4.12) is written as

$$\log \frac{Pr(D=1|E,A,G)}{1-Pr(D=1|E,A,G)} \;=\; \alpha \;+\; \beta E \;+\; \gamma_1 A \;+\; \gamma_2 G. \tag{4.14}$$

The corresponding matrix $\boldsymbol{X}$ will become

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

Each row of $\boldsymbol{X}$ represents the membership status to one of the four covariate patterns (young female, old female, young male, and old male). The columns of $\boldsymbol{X}$ correspond to 1, A, and G. Alternatively, let

$$Z_1 = \begin{cases} 1 & \text{if } A = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$Z_2 = \begin{cases} 1 & \text{if } G = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then the matrix related to stratum indicators Z_1 and Z_2 can be written as

$$\boldsymbol{Z} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

where each row represents the membership status to one of the four covariate patterns. Columns are 1, Z_1, and Z_2. In effect, $\boldsymbol{X} = \boldsymbol{Z}$ and

$$\underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}}_{\boldsymbol{X} \text{ or } \boldsymbol{Z}} = \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}}_{\boldsymbol{I}} \underbrace{\begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \\ r_{41} & r_{42} & r_{43} \end{pmatrix}}_{\boldsymbol{R}}. \tag{4.15}$$

For equality to hold in equation (4.15), $\boldsymbol{R}$ has to be a 4×3 matrix, therefore $\boldsymbol{R}$ cannot be a nonsingular matrix. Therefore, according to Graybill's theorems [25] at the beginning of section 4.2, the column spaces of the two matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ do not span the same linear space as the column space of the matrix $\boldsymbol{I}$. This implies, according to Breslow and Powers [10], that the covariate effects are not saturated with these two parameterizations.

The results from section 4.2 can be summarized as follows. The covariate effects are saturated with parameters in the model if and only if there exists a nonsingular matrix $\boldsymbol{R}$ such that $\boldsymbol{X} = \boldsymbol{IR}$ or $\boldsymbol{Z} = \boldsymbol{IR}$ where $\boldsymbol{X}$ and $\boldsymbol{Z}$ correspond to different covariate coding systems for the model and the covariate strata indicated by the stratum indicators I_k are mutually exclusive.

4.3 Examination of Equivalence through Likelihood Equations

This section expands the work of Breslow and Powers [10] to show that the estimated coefficients relating to exposure and disease in the two types of logistic regression models are identical when covariate effects are saturated with parameters using likelihood equations. Let the observed cell frequencies in the k^{th} 2×2 table ($k = 1, ..., s$) be presented as in Table 4.1. Suppose that the probability of disease follows a logistic form, then with Table 4.1, model (4.2)can be written as

$$Pr(D = 1|E, k) = \frac{\exp\{\beta E + \sum_{k=1}^{s} \gamma_k I_k + \sum_{k=1}^{s} \delta_k E I_k\}}{1 + \exp\{\beta E + \sum_{k=1}^{s} \gamma_k I_k + \sum_{k=1}^{s} \delta_k E I_k\}}. \tag{4.16}$$

For a specified stratum k,

$$\begin{aligned} p_{1k} &= Pr(D = 1|E = 1, k) = \frac{\exp\{\beta + \gamma_k + \delta_k\}}{1 + \exp\{\beta + \gamma_k + \delta_k\}}, \\ p_{0k} &= Pr(D = 1|E = 0, k) = \frac{\exp\{\gamma_k\}}{1 + \exp\{\gamma_k\}}. \end{aligned} \tag{4.17}$$

Note that $\delta_1 = 0$ is the baseline and δ_k are allowed to be zero for $k = 2, ..., s$.

The likelihood function relating the observed frequencies to the parameters p_{1k} and p_{0k} is

$$L_p = \prod_{k=1}^{s} \Big\{ \underbrace{\binom{a_k + c_k}{a_k} p_{1k}^{a_k} (1 - p_{1k})^{c_k}}_{\text{the exposed}} \underbrace{\binom{b_k + d_k}{b_k} p_{0k}^{b_k} (1 - p_{0k})^{d_k}}_{\text{the non-exposed}} \Big\}, \tag{4.18}$$

and the log likelihood function can be written as

$$l_p \propto \sum_{k=1}^{s} \{a_k \log p_{1k} + c_k \log(1 - p_{1k}) + b_k \log p_{0k} + d_k \log(1 - p_{0k})\}. \tag{4.19}$$

Differentiating (4.17) with respect to γ_k,

$$\begin{aligned} \frac{\partial p_{1k}}{\partial \gamma_k} = \frac{\exp\{\beta + \gamma_k + \delta_k\}}{1 + \exp\{\beta + \gamma_k + \delta_k\}} - \frac{[\exp\{\beta + \gamma_k + \delta_k\}]^2}{[1 + \exp\{\beta + \gamma_k + \delta_k\}]^2} = p_{1k}(1 - p_{1k}) \\ \Rightarrow \frac{\partial \log p_{1k}}{\partial \gamma_k} = 1 - p_{1k}, \end{aligned}$$

$$\frac{\partial(1-p_{1k})}{\partial\gamma_k} = -\frac{\exp\{\beta + \gamma_k + \delta_k\}}{[1+\exp\{\beta + \gamma_k + \delta_k\}]^2} = -p_{1k}(1-p_{1k})$$
$$\Rightarrow \frac{\partial \log(1-p_{1k})}{\partial\gamma_k} = -p_{1k},$$

$$\frac{\partial p_{0k}}{\partial\gamma_k} = \frac{\exp\{\gamma_k\}}{1+\exp\{\gamma_k\}} - \frac{[\exp\{\gamma_k\}]^2}{[1+\exp\{\gamma_k\}]^2} = p_{0k}(1-p_{0k})$$
$$\Rightarrow \frac{\partial \log p_{0k}}{\partial\gamma_k} = 1-p_{0k},$$

$$\frac{\partial(1-p_{0k})}{\partial\gamma_k} = -\frac{\exp\{\gamma_k\}}{[1+\exp\{\gamma_k\}]^2} = -p_{0k}(1-p_{0k})$$
$$\Rightarrow \frac{\partial \log(1-p_{0k})}{\partial\gamma_k} = -p_{0k}. \tag{4.20}$$

The score function for γ_k is written as

$$\frac{\partial l_p}{\partial\gamma_k} = a_k(1-p_{1k}) - c_k p_{1k} + b_k(1-p_{0k}) - d_k p_{0k}. \tag{4.21}$$

Following the same procedure for deriving the derivatives with respect to β and δ_k, the score functions for β and δ_k can be written as

$$\frac{\partial l_p}{\partial\beta} = \sum_{k=1}^{s} \{a_k(1-p_{1k}) - c_k p_{1k}\}, \tag{4.22}$$
$$\frac{\partial l_p}{\partial\delta_k} = a_k(1-p_{1k}) - c_k p_{1k}. \tag{4.23}$$

Setting the score functions (4.21)-(4.23) equal to zero, the following equations are obtained.

$$a_k + b_k = (a_k + c_k)\hat{p}_{1k} + (b_k + d_k)\hat{p}_{0k}, \tag{4.24}$$
$$\text{and } c_k + d_k = (a_k + c_k)(1-\hat{p}_{1k}) + (b_k + d_k)(1-\hat{p}_{0k}), \tag{4.25}$$
$$\sum_{k=1}^{s} a_k = \sum_{k=1}^{s} \{(a_k + c_k)\hat{p}_{1k}\}, \tag{4.26}$$
$$a_k = (a_k + c_k)\hat{p}_{1k} \text{ and } b_k = (b_k + d_k)\hat{p}_{0k} \text{ if and only if } \delta_k \neq 0. \tag{4.27}$$

Table 4.4: The k^{th} 2×2 Table (Fitted Frequencies)

	$E=1$	$E=0$	Row Total
$D=1$	$(a_k + c_k)\hat{p}_{1k}$	$(b_k + d_k)\hat{p}_{0k}$	$a_k + b_k$
$D=0$	$(a_k + c_k)(1-\hat{p}_{1k})$	$(b_k + d_k)(1-\hat{p}_{0k})$	$c_k + d_k$
Column Total	$a_k + c_k$	$b_k + d_k$	

Using the above equations, the fitted frequencies for Table 4.1 are shown in Table 4.4.

Suppose that the probability of exposure also follows a logistic form, then with Table 4.1, model (4.6) can be written as

$$Pr(E=1|D,k) = \frac{\exp\{\tau D + \sum_{k=1}^{s}\theta_k I_k + \sum_{k=1}^{s}\zeta_k D I_k\}}{1+\exp\{\tau D + \sum_{k=1}^{s}\theta_k I_k + \sum_{k=1}^{s}\zeta_k D I_k\}}. \tag{4.28}$$

For a specified stratum k,

$$\begin{aligned} q_{1k} &= Pr(E=1|D=1,k) = \frac{\exp\{\tau + \theta_k + \zeta_k\}}{1+\exp\{\tau + \theta_k + \zeta_k\}}, \\ q_{0k} &= Pr(E=1|D=0,k) = \frac{\exp\{\theta_k\}}{1+\exp\{\theta_k\}}. \end{aligned} \tag{4.29}$$

Note that $\zeta_1 = 0$ is the baseline and ζ_k are allowed to be zero for $k = 2, ..., s$.

The likelihood function relating the observed frequencies to the parameters q_{1k} and q_{0k} is

$$L_q = \prod_{k=1}^{s} \underbrace{\binom{a_k+b_k}{a_k} q_{1k}^{a_k}(1-q_{1k})^{b_k}}_{\text{the diseased}} \underbrace{\binom{c_k+d_k}{c_k} q_{0k}^{c_k}(1-q_{0k})^{d_k}}_{\text{the non-diseased}}, \tag{4.30}$$

and the log likelihood function can be written as

$$l_q \propto \sum_{k=1}^{s}\{a_k \log q_{1k} + b_k \log(1-q_{1k}) + c_k \log q_{0k} + d_k \log(1-q_{0k})\}. \tag{4.31}$$

Following the preceding procedures for calculating derivatives, the corresponding score

functions for θ_k, τ, and ζ_k are

$$\frac{\partial l_q}{\partial \theta_k} = a_k(1-q_{1k}) - b_k q_{1k} + c_k(1-q_{0k}) - d_k q_{0k}, \tag{4.32}$$

$$\frac{\partial l_q}{\partial \tau} = \sum_{k=1}^{s} \{a_k(1-q_{1k}) - b_k q_{1k}\}, \tag{4.33}$$

$$\frac{\partial l_q}{\partial \zeta_k} = a_k(1-q_{1k}) - b_k q_{1k}. \tag{4.34}$$

Setting the score functions (4.32)-(4.34) equal to zero, the following equations are obtained

$$a_k + c_k = (a_k + b_k)\hat{q}_{1k} + (c_k + d_k)\hat{q}_{0k}, \tag{4.35}$$

$$\text{and } b_k + d_k = (a_k + b_k)(1-\hat{q}_{1k}) + (c_k + d_k)(1-\hat{q}_{0k}), \tag{4.36}$$

$$\sum_{k=1}^{s} a_k = \sum_{k=1}^{s} (a_k + b_k)\hat{q}_{1k}, \tag{4.37}$$

$$a_k = (a_k + b_k)\hat{q}_{1k} \text{ and } c_k = (c_k + d_k)\hat{q}_{0k} \text{ if and only if } \zeta_k = 0. \tag{4.38}$$

Using the above equations, the fitted frequencies for Table 4.1 are shown in Table 4.5.

Table 4.5: The k^{th} 2×2 Table (Fitted Frequencies)

	$E=1$	$E=0$	Row Total
$D=1$	$(a_k + b_k)\hat{q}_{1k}$	$(a_k + b_k)(1-\hat{q}_{1k})$	$a_k + b_k$
$D=0$	$(c_k + d_k)\hat{q}_{0k}$	$(c_k + d_k)(1-\hat{q}_{0k})$	$c_k + d_k$
Column Total	$a_k + c_k$	$b_k + d_k$	

In summary, section 4.3 showed that both the model for the log odds of disease and the model for the log odds of exposure have the same set of score functions (written in terms of observed and fitted frequencies) [10] through maximum likelihood (ML) fitting. Specifically, the row and column marginal totals are the same from both logistic regression models and

$$\sum_{k=1}^{s} a_k = \sum_{k=1}^{s} (a_k + c_k)\hat{p}_{1k} = \sum_{k=1}^{s} (a_k + b_k)\hat{q}_{1k}. \tag{4.39}$$

When the number of strata is large, an approximate estimation method can be used. Breslow [7] showed that, for the "diseased" approach, the asymptotic mean $\tilde{a}_k$ for a_k given $a_k + b_k$, $a_k + c_k$ and $b_k + d_k$ can be obtained as the solution of the quadratic equation

$$\frac{\tilde{a}_k\{[b_k + d_k] - [(a_k + b_k) + \tilde{a}_k]\}}{\{[a_k + b_k] - \tilde{a}_k\}\{[a_k + c_k] - \tilde{a}_k\}} = e^{(\hat{\beta}+\hat{\delta}_k)}, \tag{4.40}$$

or equivalently

$$\frac{\tilde{a}_k\{d_k - a_k + \tilde{a}_k\}}{\{a_k + b_k - \tilde{a}_k\}\{a_k + c_k - \tilde{a}_k\}} = e^{(\hat{\beta}+\hat{\delta}_k)}. \tag{4.41}$$

Similarly, for the "exposed" approach, the asymptotic mean $\breve{a}$ for a_k given $a_k + b_k$, $a_k + c_k$ and $c_k + d_k$ can be obtained as the solution of the quadratic equation

$$\frac{\breve{a}_k\{[c_k + d_k] - [(a_k + c_k) + \breve{a}_k\}}{\{[a_k + b_k] - \breve{a}_k\}\{[a_k + c_k] - \breve{a}_k\}} = e^{(\hat{\tau}+\hat{\zeta}_k)}, \tag{4.42}$$

or equivalently

$$\frac{\breve{a}_k\{d_k - a_k + \breve{a}_k\}}{\{a_k + b_k - \breve{a}_k\}\{a_k + c_k - \breve{a}_k\}} = e^{(\hat{\tau}+\hat{\zeta}_k)}, \tag{4.43}$$

Noticing that the row and column marginal totals are the same for both approaches. One can substitute $\tilde{a}_k$ in equation (4.41) and $\breve{a}_k$ in equation (4.43) into equation (4.39), which will lead to $\hat{\beta} = \hat{\tau}$ and $\hat{\delta}_k = \hat{\zeta}_k$.

As a final numerical example, suppose that all the cross-product terms in models (4.16) and (4.28) are set to zero. Then the model for the log odds of disease reduces to

$$\log \frac{Pr(D = 1|E, k)}{1 - Pr(D = 1|E, k)} = \beta E + \sum_{k=1}^{4} \gamma_k I_k \tag{4.44}$$

and the model for the log odds of exposure reduces to

$$\log \frac{Pr(E = 1|D, k)}{1 - Pr(E = 1|D, k)} = \tau D + \sum_{k=1}^{4} \theta_k I_k. \tag{4.45}$$

The numerical results in the following two tables show that the estimated coefficients and standard errors relating to exposure (history of diabetes) and disease (MI) in the two types of logistic regression models are equivalent.

Table 4.6: Estimated Coefficients and Standard Errors from Fitting the Model for the Log Odds of Disease (A Series of Tables 2)

Variable	**Coefficient**	**Std. Err.**
exp	**0.8103**	**(0.2355)**
young female (I1)	-0.4586	(0.3068)
old female (I2)	-0.6177	(0.2525)
young male (I3)	-0.1592	(0.2018)
old male (I4)	-0.0008	(0.1557)

Table 4.7: Estimated Coefficients and Standard Errors from Fitting the Model for the Log Odds of Exposure (A Series of Tables 2)

Variable	**Coefficient**	**Std. Err.**
dis	**0.8103**	**(0.2355)**
young female (I1)	-1.2519	(0.3474)
old female (I2)	-0.7472	(0.2580)
young male (I3)	-1.4473	(0.2600)
old male (I4)	-1.9490	(0.2475)

Chapter 5

Equivalence of the Profile Likelihood Functions for the Two Types of Logistic Regression Models in Case-Control Studies

An alternative approach to demonstrating equivalence between the two types of logistic regression models is to use the multinomial-Poisson transformation [4] and the profile likelihood [42]. As before, let the observed cell frequencies in the k^{th} stratum ($k = 1, ..., s$) be displayed as in Table 5.1.

Table 5.1: The k^{th} 2×2 Table (Frequencies)

	$E = 1$	$E = 0$
$D = 1$	a_k	b_k
$D = 0$	c_k	d_k

Suppose that a_k, b_k, c_k, and d_k independently follow Poisson distributions $Po(\lambda_{de})$ [50] where λ_{de} are defined in Table 5.2.

Table 5.2: The k^{th} 2×2 Table (Assuming a Poisson Distribution)

	$E = 1$	$E = 0$
$D = 1$	$\lambda_{11} = \exp\{\log \mu_k + \log \xi_k + \log \omega_k + \eta_k\}$	$\lambda_{10} = \exp\{\log \mu_k + \log \xi_k\}$
$D = 0$	$\lambda_{01} = \exp\{\log \mu_k + \log \omega_k\}$	$\lambda_{00} = \exp\{\log \mu_k\}$

The Poisson likelihood function is derived from a model which considers the frequencies data as 4 independent Poisson distributions corresponding to each cell. Specifically,

$$L_{Po} = \prod_{k=1}^{s} \left\{ \frac{\lambda_{11}^{a_k} \exp\{-\lambda_{11}\}}{a_k!} \frac{\lambda_{01}^{b_k} \exp\{-\lambda_{01}\}}{b_k!} \frac{\lambda_{10}^{c_k} \exp\{-\lambda_{10}\}}{c_k!} \frac{\lambda_{00}^{d_k} \exp\{-\lambda_{00}\}}{d_k!} \right\}. \tag{5.1}$$

The corresponding Poisson log likelihood function is

$$\begin{aligned} l_{Po} \propto & \sum_{k=1}^{s} \Big\{ a_k\{\log\mu_k + \log\xi_k + \log\omega_k + \eta_k\} + b_k\{log\mu_k + \log\xi_k\} \\ & \qquad + c_k\{\log\mu_k + \log\omega_k\} + d_k\{\log\mu_k\} \Big\} \\ & - \sum_{k=1}^{s} \Big\{ \exp\{\log\mu_k + \log\omega_k + \log\xi_k + \eta_k\} + \exp\{\log\mu_k + \log\xi_k\} \\ & \qquad + \exp\{\log\mu_k + \omega_k\} + \exp\{\log\mu_k\} \Big\} \\ = & \sum_{k=1}^{s} \Big\{ (a_k + b_k + c_k + d_k)\log\mu_k + (a_k + b_k)\log\xi_k + (a_k + c_k)\log\omega_k + a_k\eta_k \Big\} \\ & - \sum_{k=1}^{s} \Big\{ \exp\{\log\mu_k + \log\omega_k + \log\xi_k + \eta_k\} + \exp\{\log\mu_k + \log\xi_k\} \\ & \qquad + \exp\{\log\mu_k + \omega_k\} + \exp\{\log\mu_k\} \Big\}. \end{aligned} \tag{5.2}$$

5.1 The Profile Likelihood Function for modeling the Log Odds of Disease

Using Poisson parameters from Table 5.2, the probabilities of disease given exposure can be written as

$$\begin{aligned} Pr(D=1|E=1,k) &= \frac{\exp\{\log\mu_k + \log\xi_k + \log\omega_k + \eta_k\}}{\exp\{\log\mu_k + \log\omega_k\} + \exp\{\log\mu_k + \log\xi_k + \log\omega_k + \eta_k\}} \\ &= \frac{\exp\{\log\xi_k + \eta_k\}}{1 + \exp\{\log\xi_k + \eta_k\}}, \end{aligned} \tag{5.3}$$

$$Pr(D=0|E=1,k) = 1 - Pr(D=1|E=1,k) = \frac{1}{1 + \exp\{\log\xi_k + \eta_k\}}, \tag{5.4}$$

$$Pr(D=1|E=0,k) = \frac{\exp\{\log\mu_k + \log\xi_k\}}{\exp\{\log\mu_k + \log\xi_k\} + \exp\{\log\mu_k\}} = \frac{\exp\{\log\xi_k\}}{1 + \exp\{\log\xi_k\}}, \tag{5.5}$$

and

$$Pr(D=0|E=0,k) = 1 - Pr(D=1|E=0,k) = \frac{1}{1 + \exp\{\log\xi_k\}} \tag{5.6}$$

where $\beta + \delta_k = \eta_k$ and $\gamma_k = \log \xi_k$. Note that β, δ_k and γ_k are the parameters of interest in model (4.16). Using equations (5.3)-(5.6), the odds ratio is

$$\frac{\dfrac{Pr(D=1|E=1,k)}{Pr(D=0|E=1,k)}}{\dfrac{Pr(D=1|E=0,k)}{Pr(D=0|E=0,k)}} = \exp\{\eta_k\}. \tag{5.7}$$

Consider, now, a binomial likelihood function as a product of two binomial distributions for each category of exposure (one for the exposed and one for the non-exposed). Then

$$\begin{aligned} L_p &= \prod_{k=1}^{s} \binom{a_k + c_k}{a_k} \big[Pr(D=1|E=1,k)\big]^{a_k} \big[Pr(D=0|E=1,k)\big]^{c_k} \\ &\qquad \binom{b_k + d_k}{b_k} \big[Pr(D=1|E=0,k)\big]^{b_k} \big[Pr(D=0|E=0,k)\big]^{d_k} \\ &= \prod_{k=1}^{s} \binom{a_k + c_k}{a_k} \Big[\frac{\exp\{\log \xi_k + \eta_k\}}{1+\exp\{\log \xi_k + \eta_k\}}\Big]^{a_k} \Big[\frac{1}{1+\exp\{\log \xi_k + \eta_k\}}\Big]^{c_k} \\ &\qquad \binom{b_k + d_k}{b_k} \Big[\frac{\exp\{\log \xi_k\}}{1+\exp\{\log \xi_k\}}\Big]^{b_k} \Big[\frac{1}{1+\exp\{\log \xi_k\}}\Big]^{d_k} \end{aligned} \tag{5.8}$$

and the corresponding binomial log likelihood function is

$$\begin{aligned} l_p \propto \sum_{k=1}^{s} & a_k \log \frac{\exp\{\log \xi_k + \eta_k\}}{1+\exp\{\log \xi_k + \eta_k\}} + c_k \log \frac{1}{1+\exp\{\log \xi_k + \eta_k\}} \\ & + b_k \log \frac{\exp\{\log \xi_k\}}{1+\exp\{\log \xi_k\}} + d_k \log \frac{1}{1+\exp\{\log \xi_k\}} \quad . \end{aligned} \tag{5.9}$$

Taking the first derivatives of the Poisson log likelihood function (5.2) with respect to μ_k and ω_k, the following expressions are obtained

$$\frac{\partial l_{po}}{\partial \mu_k} = \frac{(a_k + b_k + c_k + d_k)}{\mu_k} - \quad \exp\{\log \omega_k + \log \xi_k + \eta_k\} + \exp\{\log \omega_k\} + \exp\{\log \xi_k\} + 1 \quad , \tag{5.10}$$

$$\frac{\partial l_{po}}{\partial \omega_k} = \frac{a_k + c_k}{\omega_k} - \quad \exp\{\log \mu_k + \log \xi_k + \eta_k\} + \exp\{\log \mu_k\} \quad . \tag{5.11}$$

Setting these derivatives equal to zero and solving the two equations for μ_k and ω_k, we

obtain

$$\hat{\mu}_k = \frac{b_k + d_k}{1 + \exp\{\log \xi_k\}} \quad \text{and}$$
$$\hat{\omega}_k = \frac{(a_k + c_k)(1 + \exp\{\log \xi_k\})}{(b_k + d_k)(1 + \exp\{\log \xi_k + \eta_k\})}. \tag{5.12}$$

Substituting of $\hat{\mu}_k$ and $\hat{\omega}_k$ into the Poisson log likelihood function (5.2) yields

$$\begin{aligned} l_{po} \propto \sum_{k=1}^{s} & a_k \log \frac{\exp\{\log \xi_k + \eta_k\}}{1 + \exp\{\log \xi_k + \eta_k\}} + c_k \log \frac{1}{1 + \exp\{\log \xi_k + \eta_k\}} \\ & + b_k \log \frac{\exp\{\log \xi_k\}}{1 + \exp\{\log \xi_k\}} + d_k \log \frac{1}{1 + \exp\{\log \xi_k\}}\} \\ & + (a_k + c_k)\log(a_k + c_k) + (b_k + d_k)\log(b_k + d_k) - (a_k + b_k + c_k + d_k) \\ \propto \sum_{k=1}^{s} & a_k \log \frac{\exp\{\log \xi_k + \eta_k\}}{1 + \exp\{\log \xi_k + \eta_k\}} + c_k \log \frac{1}{1 + \exp\{\log \xi_k + \eta_k\}} \\ & + b_k \log \frac{\exp\{\log \xi_k\}}{1 + \exp\{\log \xi_k\}} + d_k \log \frac{1}{1 + \exp\{\log \xi_k\}}\} \end{aligned} \tag{5.13}$$

which is the same as the binomial log likelihood function (5.9) for modeling the log odds of disease.

5.2 The Profile Likelihood Function for modeling the Log Odds of Exposure

Using the Poisson parameters from Table 5.2, the probabilities of exposure given disease can be written as

$$\begin{aligned} & Pr(E = 1 | D = 1, k) \\ & = \frac{\exp\{\log \mu_k + \log \xi_k + \log \omega_k + \eta_k\}}{\exp\{\log \mu_k + \log \xi_k + \log \omega_k + \eta_k\} + \exp\{\log \mu_k + \log \xi_k\}} \\ & = \frac{\exp\{\log \omega_k + \eta_k\}}{1 + \exp\{\log \omega_k + \eta_k\}}, \end{aligned} \tag{5.14}$$

$$Pr(E=0|D=1,k) = 1 - Pr(E=1|D=1,k) = \frac{1}{1+\exp\{\log\omega_k + \eta_k\}}, \tag{5.15}$$

$$Pr(E=1|D=0,k) = \frac{\exp\{\log\mu_k + \log\omega_k\}}{\exp\{\log\mu_k + \log\omega_j\} + \exp\{\mu_k\}} = \frac{\exp\{\log\omega_k\}}{1+\exp\{\log\omega_k\}}, \tag{5.16}$$

and

$$Pr(E=0|D=0,k) = 1 - Pr(E=1|D=0,k) = \frac{\exp\{\log\omega_k\}}{1+\exp\{\log\omega_k\}} \tag{5.17}$$

where $\tau + \zeta_k = \eta_k$ and $\theta_k = \log\omega_k$. Note that τ, ζ and θ_k are the parameters of interest in model (4.28). Using equations (5.14)-(5.17), the odds ratio is

$$\frac{\dfrac{Pr(E=1|D=1,k)}{Pr(E=0|D=1,k)}}{\dfrac{Pr(E=1|D=0,k)}{Pr(E=1|D=0,k)}} = \exp\{\eta_k\}. \tag{5.18}$$

Consider again a binomial likelihood function as a product of two binomial distributions for each category of disease (one for the diseased and one for the non-diseased). Then

$$\begin{aligned} L_q =& \prod_{k=1}^{s} \binom{a_k+b_k}{a_k} \Big[Pr(E=1|D=1,k)\Big]^{a_k} \Big[Pr(E=0|D=1,k)\Big]^{b_k} \\ & \binom{c_k+d_k}{c_k} \Big[Pr(E=1|D=0,k)\Big]^{c_k} \Big[Pr(E=0|D=0,k)\Big]^{d_k} \\ =& \prod_{k=1}^{s} \binom{a_k+b_k}{a_k} \Big[\frac{\exp\{\log\omega_k+\eta_k\}}{1+\exp\{\log\omega_k+\eta_k\}}\Big]^{a_k} \Big[\frac{1}{1+\exp\{\log\omega_k+\eta_k\}}\Big]^{b_k} \\ & \binom{c_k+d_k}{c_k} \Big[\frac{\exp\{\log\omega_k\}}{1+\exp\{\log\omega_k\}}\Big]^{c_k} \Big[\frac{1}{1+\exp\{\log\omega_k\}}\Big]^{d_k} \end{aligned} \tag{5.19}$$

and the corresponding log likelihood function is

$$\begin{aligned} l_q \propto \sum_{k=1}^{s} & a_k \log\frac{\exp\{\log\omega_k+\eta_k\}}{1+\exp\{\log\omega_k+\eta_k\}} + b_k \log\frac{1}{1+\exp\{\log\omega_k+\eta_k\}} \\ & + c_k \log\frac{\exp\{\log\omega_k\}}{1+\exp\{\log\omega_k\}} + d_k \log\frac{1}{1+\exp\{\log\omega_k\}} \quad . \end{aligned} \tag{5.20}$$

Taking the first derivatives of the Poisson log likelihood function (5.2) with respect to μ_k and ξ_k, the following expressions are obtained

$$\frac{\partial l_{po}}{\partial \mu_k} = \frac{(a_k + b_k + c_k + d_k)}{\mu_k} - \quad \exp\{\log \omega_k + \log \xi_k + \eta_k\} + \exp\{\log \omega_k\} + \exp\{\log \xi_k\} + 1 \quad , \tag{5.21}$$

$$\frac{\partial l_{po}}{\partial \xi_k} = \frac{(a_k + b_k)}{\xi_k} - \quad \exp\{\log \mu_k + \log \omega_k + \eta_k\} + exp\{\log \mu_k\} \quad . \tag{5.22}$$

Setting these derivatives equal to zero and solving the two equations for μ_k and ξ_k, we obtain

$$\begin{aligned} \hat{\mu}_k &= \frac{c_k + d_k}{1 + \exp\{\log \omega_k\}} \quad \text{and} \\ \hat{\xi}_k &= \frac{(a_k + b_k)(1 + \exp\{\log \omega_k\})}{(c_k + d_k)(1 + \exp\{\log \omega_k + \eta_k\})}. \end{aligned} \tag{5.23}$$

Substituting of $\hat{\mu}_k$ and $\hat{\xi}_k$ into the Poisson log likelihood function (5.2) yields

$$\begin{aligned} l_{po} \propto \sum_{k=1}^{s} & a_k \log \frac{\exp\{\log \omega_k + \eta_k\}}{1 + \exp\{\log \omega_k + \eta_k\}} + b_k \log \frac{1}{1 + \exp\{\log \omega_k + \eta_k\}} \\ & + c_k \log \frac{\exp\{\log \omega_k\}}{1 + \exp\{\log \omega_k\}} + d_k \log \frac{1}{1 + \exp\{\log \omega_k\}} \\ & + (a_k + b_k) \log(a_k + b_k) + (c_k + d_k) \log(c_k + d_k) - (a_k + b_k + c_k + d_k) \\ \propto \sum_{k=1}^{s} & a_k \log \frac{\exp\{\log \omega_k + \eta_k\}}{1 + \exp\{\log \omega_k + \eta_k\}} + b_k \log \frac{1}{1 + \exp\{\log \omega_k + \eta_k\}} \\ & + c_k \log \frac{\exp\{\log \omega_k\}}{1 + \exp\{\log \omega_k\}} + d_k \log \frac{1}{1 + \exp\{\log \omega_k\}} \end{aligned} \tag{5.24}$$

which is the same as the binomial log likelihood function (5.20) for modeling the log odds of exposure.

From the results in Chapter 5, the following summary statements can be made:

I. The Poisson profile likelihood function of the Poisson likelihood function for ξ_k and η_k, after maximizing with respect to μ_k and ω_k, is identical to the binomial likelihood function for modeling the log odds of disease for ξ_k and η_k.

II. The Poisson profile likelihood function of the Poisson likelihood function for ω_k and η_k, after maximizing with respect to μ_k and ξ_k, is identical to the binomial likelihood function for modeling the log odds of exposure for ω_k and η_k.

Baker [4] provided a similar argument to the above statements, that is, "we can transform the multinomial likelihood (i.e. in (5.9) or (5.20)) into a Poisson likelihood (i.e. in (5.2)), with additional parameters... [which] yields identical estimates and asymptotic variances" [4]. This demonstrates, in an alternative way, that the estimated coefficients and standard errors relating to exposure and disease are the same in the models for the log odds of disease and log odds of exposure.

Chapter 6

Equivalence of the Standard Errors for the Estimated Coefficients relating to Exposure and Disease for the Two Types of Logistic Regression Models

In the preceding chapters, identical estimated coefficients relating to exposure and disease have been demonstrated for the two types of logistic regression models when the covariate effects are saturated in the models. In this chapter, it will be shown that their corresponding standard errors are also identical. This chapter is an expansion of the appendix of Breslow's paper "Regression Analysis of the Log Odds Ratio: A Method for Retrospective Studies".

6.1 The Poisson Regression

Suppose the frequencies in the s 2×2 tables (as in Table 5.1 where $k = 1, ..., s$) arise from independent Poisson distributions with corresponding parameters shown in Table 6.1.

Table 6.1: The k^{th} 2×2 Table (Assuming a Poisson Distribution)

	$E = 1$	$E = 0$
$D = 1$	$\exp\{\gamma_{1k} + \beta + \delta_k\}$	$\exp\{\gamma_{1k} + \gamma_{2k}\}$
$D = 0$	$\exp\{\gamma_{3k}\}$	$\exp\{\gamma_{2k} + \gamma_{3k}\}$

The odds ratio will be

$$OR = \frac{\exp\{\gamma_{1k} + \beta + \delta_k\} \exp\{\gamma_{2k} + \gamma_{3k}\}}{\exp\{\gamma_{3k}\} \exp\{\gamma_{1k} + \gamma_{2k}\}} = \exp\{\beta + \delta_k\}. \tag{6.1}$$

The Poisson regression model can be written as

$$\boldsymbol{y} = \exp\{\boldsymbol{X}\boldsymbol{B}\} \tag{6.2}$$

where

$$\boldsymbol{y} = \begin{pmatrix} a_1 \\ \vdots \\ a_s \\ b_1 \\ \vdots \\ b_s \\ c_1 \\ \vdots \\ c_s \\ d_1 \\ \vdots \\ d_s \end{pmatrix}$$

and

$$\boldsymbol{XB}$$

$$= \begin{pmatrix}
1 & & 0 & \vdots & 0 & \dots & 0 & \vdots & 0 & \dots & 0 & \vdots & 1 & 0 & 0 \\
 & \ddots & & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & 1 & \ddots & \\
0 & & 1 & \vdots & 0 & \dots & 0 & \vdots & 0 & \dots & 0 & \vdots & 1 & 0 & 1 \\
\dots & \dots & \dots & \dots\dots & \dots & \dots & \dots\dots & \dots & \dots & \dots\dots & \dots & \dots & \dots & \dots & \dots \\
1 & & 0 & \vdots & 1 & & 0 & \vdots & 0 & \dots & 0 & \vdots & 0 & \dots & 0 \\
 & \ddots & & \vdots & & \ddots & & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
0 & & 1 & \vdots & 0 & & 1 & \vdots & 0 & \dots & 0 & \vdots & 0 & \dots & 0 \\
\dots & \dots & \dots & \dots\dots & \dots & \dots & \dots\dots & \dots & \dots & \dots\dots & \dots & \dots & \dots & \dots & \dots \\
0 & \dots & 0 & \vdots & 0 & \dots & 0 & \vdots & 1 & & 0 & \vdots & 0 & \dots & 0 \\
\vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & & \ddots & & \vdots & \vdots & & \vdots \\
0 & \dots & 0 & \vdots & 0 & \dots & 0 & \vdots & 0 & & 1 & \vdots & 0 & \dots & 0 \\
\dots & \dots & \dots & \dots\dots & \dots & \dots & \dots\dots & \dots & \dots & \dots\dots & \dots & \dots & \dots & \dots & \dots \\
0 & \dots & 0 & \vdots & 1 & & 0 & \vdots & 1 & & 0 & \vdots & 0 & \dots & 0 \\
\vdots & & \vdots & \vdots & & \ddots & & \vdots & & \ddots & & \vdots & \vdots & & \vdots \\
0 & \dots & 0 & \vdots & 0 & & 1 & \vdots & 0 & & 1 & \vdots & 0 & \dots & 0
\end{pmatrix}
\begin{pmatrix}
 \\
\gamma_{11} \\
\vdots \\
\gamma_{1s} \\
\gamma_{21} \\
\vdots \\
\gamma_{2s} \\
\gamma_{31} \\
\vdots \\
\gamma_{3s} \\
\dots\dots \\
\beta \\
\delta_2 \\
\vdots \\
\delta_s
\end{pmatrix}$$

$$= \begin{pmatrix}
\boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{Z} \\
\boldsymbol{I} & \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\
\boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} \\
\boldsymbol{0} & \boldsymbol{I} & \boldsymbol{I} & \boldsymbol{0}
\end{pmatrix}
\begin{pmatrix}
\boldsymbol{\gamma} \\
\boldsymbol{\delta}
\end{pmatrix}.$$

$\boldsymbol{I}$ and $\boldsymbol{Z}$ are both $s \times s$ submatrices and $\boldsymbol{\gamma} = \begin{pmatrix} \gamma_{11} \\ \vdots \\ \gamma_{1s} \\ \gamma_{21} \\ \vdots \\ \gamma_{2s} \\ \gamma_{31} \\ \vdots \\ \gamma_{3s} \end{pmatrix}$ and $\boldsymbol{\delta} = \begin{pmatrix} \beta \\ \delta_2 \\ \vdots \\ \delta_s \end{pmatrix}$.

For a classical linear model, the Gauss-Markov theorem [28] states that the least squares estimator is the best linear unbiased estimator (BLUE) and the errors are uncorrelated and have a mean of zero and equal variances (σ^2). When the errors are heterogeneous, Aitken [2] demonstrated that the estimators are BLUE by using a weight which is equal to the inverse of the variance of the variable. Define $\boldsymbol{V_y}$ as the inverse of $cov[\log(\boldsymbol{y})]$ then

$$\boldsymbol{B} = (\boldsymbol{X^T V_y X})^{-1} \boldsymbol{X^T V_y} \log(\boldsymbol{y}) \tag{6.3}$$

and

$$cov(\boldsymbol{B}) = (\boldsymbol{X^T V_y X})^{-1} \boldsymbol{X^T V_y V_y^{-1} V_y^T X} \ (\boldsymbol{X^T V_y X})^{-1}\big]^{\boldsymbol{T}} = (\boldsymbol{X^T V_y X})^{-1}. \tag{6.4}$$

6.2 The Covariance Matrix for $\hat{\boldsymbol{\delta}}$

Suppose a random variable (e.g. a_1) has a Poisson distribution with estimated mean and variance $\hat{a}_1$. By a first-order Taylor series expansion [32],

$$\begin{aligned}\log(a_1) &= \log(\hat{a}_1)|_{a_1=\hat{a}_1} + \frac{\partial \log(a_1)}{\partial a_1}|_{a_1=\hat{a}_1}(a_1 - \hat{a}_1) \\ &= \log(\hat{a}_1) + \frac{1}{\hat{a}_1}(a_1 - \hat{a}_1) \\ &= \log(\hat{a}_1) + \frac{1}{\hat{a}_1}a_1 - 1 \end{aligned} \tag{6.5}$$

so that

$$var[\log(a_1)] = (\frac{1}{\hat{a}_1})^2 var(a_1) \tag{6.6}$$

and

$$\widehat{var}[\log(a_1)] = (\frac{1}{\hat{a}_1})^2 \hat{a}_1 = \frac{1}{\hat{a}_1}. \tag{6.7}$$

Thus, the inverse of the covariance matrix for $\log(\boldsymbol{y})$ can be estimated by

$$\boldsymbol{V}_{\hat{\boldsymbol{y}}} = \begin{pmatrix} \hat{a}_1 & & & & & & & & & & 0 \\ & \ddots & & & & & & & & & \\ & & \hat{a}_s & & & & & & & & \\ & & & \hat{b}_1 & & & & & & & \\ & & & & \ddots & & & & & & \\ & & & & & \hat{b}_s & & & & & \\ & & & & & & \hat{c}_1 & & & & \\ & & & & & & & \ddots & & & \\ & & & & & & & & \hat{c}_s & & \\ & & & & & & & & & \hat{d}_1 & \\ & & & & & & & & & & \ddots \\ 0 & & & & & & & & & & \hat{d}_s \end{pmatrix} = \begin{pmatrix} \boldsymbol{V}_{\hat{\boldsymbol{a}}} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_{\hat{\boldsymbol{b}}} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{V}_{\hat{\boldsymbol{c}}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{V}_{\hat{\boldsymbol{d}}} \end{pmatrix}$$

such that

$$\widehat{cov}(\boldsymbol{B}) = [\boldsymbol{X^T V_{\hat{y}} X}]^{-1} = \left[\begin{pmatrix} \boldsymbol{V_{\hat{a}} + V_{\hat{b}}} & \boldsymbol{V_{\hat{b}}} & \boldsymbol{0} & \boldsymbol{V_{\hat{a}} Z} \\ \boldsymbol{V_{\hat{b}}} & \boldsymbol{V_{\hat{b}} + V_{\hat{d}}} & \boldsymbol{V_{\hat{d}}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V_{\hat{d}}} & \boldsymbol{V_{\hat{c}} + V_{\hat{d}}} & \boldsymbol{0} \\ \boldsymbol{Z^T V_{\hat{a}}} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{Z^T V_{\hat{a}} Z} \end{pmatrix} \right]^{-1}.$$

Let

$$[\boldsymbol{X^T V_{\hat{y}} X}]^{-1} = \left[\begin{pmatrix} \boldsymbol{A_{11}} & \boldsymbol{A_{12}} & \boldsymbol{A_{13}} & \boldsymbol{A_{14}} \\ \boldsymbol{A_{21}} & \boldsymbol{A_{22}} & \boldsymbol{A_{23}} & \boldsymbol{A_{24}} \\ \boldsymbol{A_{31}} & \boldsymbol{A_{32}} & \boldsymbol{A_{33}} & \boldsymbol{A_{34}} \\ \boldsymbol{A_{41}} & \boldsymbol{A_{42}} & \boldsymbol{A_{43}} & \boldsymbol{A_{44}} \end{pmatrix} \right]^{-1}.$$

Using the method for inverting partitioned symmetric matrix [33], the lower right hand $s \times s$ corner of the matrix $[\boldsymbol{X^T V_{\hat{y}} X}]^{-1}$ can be written as

$$\widehat{cov}(\boldsymbol{\delta}) = \boldsymbol{A_{44.123}} = \boldsymbol{A_{44.23} - A_{41.23} A^{-1}_{11.23} A_{14.23}} \tag{6.8}$$

where

$$\begin{aligned} \boldsymbol{A_{44.23}} &= \boldsymbol{A_{44.3} - A_{42.3} A^{-1}_{22.3} A_{24.3}} \\ &= (\boldsymbol{A_{44} - A_{43} A_{33}}^{-1} \boldsymbol{A_{34}}) \\ &\quad - (\boldsymbol{A_{42} - A_{43} A_{33}}^{-1} \boldsymbol{A_{32}})\ (\boldsymbol{A_{22} - A_{23} A_{33}}^{-1} \boldsymbol{A_{32}})\big]^{-1} (\boldsymbol{A_{24} - A_{23} A_{33}}^{-1} \boldsymbol{A_{34}}) \\ &= \boldsymbol{Z^T V_{\hat{a}} Z}, \end{aligned} \tag{6.9}$$

$$\begin{aligned} \boldsymbol{A_{41.23}} &= \boldsymbol{A_{41.3} - A_{42.3} A^{-1}_{22.3} A_{21.3}} \\ &= (\boldsymbol{A_{41} - A_{43} A_{33}}^{-1} \boldsymbol{A_{31}}) \\ &\quad - (\boldsymbol{A_{42} - A_{43} A_{33}}^{-1} \boldsymbol{A_{32}})\ (\boldsymbol{A_{22} - A_{23} A_{33}}^{-1} \boldsymbol{A_{32}})\big]^{-1} (\boldsymbol{A_{21} - A_{23} A_{33}}^{-1} \boldsymbol{A_{31}}) \\ &= \boldsymbol{Z^T V_{\hat{a}}}, \end{aligned} \tag{6.10}$$

$$\begin{aligned}\boldsymbol{A_{11.23}^{-1}} &= [\boldsymbol{A_{11.3}} - \boldsymbol{A_{12.3}}\boldsymbol{A_{22.3}^{-1}}\boldsymbol{A_{21.3}}]^{-1} \\ &= (\boldsymbol{A_{11}} - \boldsymbol{A_{13}}\boldsymbol{A_{33}}^{-1}\boldsymbol{A_{31}}) \\ &- (\boldsymbol{A_{12}} - \boldsymbol{A_{13}}\boldsymbol{A_{33}}^{-1}\boldsymbol{A_{32}})[(\boldsymbol{A_{22}} - \boldsymbol{A_{23}}\boldsymbol{A_{33}}^{-1}\boldsymbol{A_{32}})]^{-1}(\boldsymbol{A_{21}} - \boldsymbol{A_{23}}\boldsymbol{A_{33}}^{-1}\boldsymbol{A_{31}})\big]^{-1} \\ &= (\boldsymbol{V_{\hat{a}}} + \boldsymbol{V_{\hat{b}}}) - \boldsymbol{V_{\hat{b}}}\Big[(\boldsymbol{V_{\hat{b}}} + \boldsymbol{V_{\hat{d}}}) - \boldsymbol{V_{\hat{d}}}(\boldsymbol{V_{\hat{c}}} + \boldsymbol{V_{\hat{d}}})^{-1}\boldsymbol{V_{\hat{d}}}\Big]^{-1}\boldsymbol{V_{\hat{b}}}^{\quad -1}, \end{aligned} \tag{6.11}$$

and

$$\begin{aligned}\boldsymbol{A_{14.23}} &= \boldsymbol{A_{14.3}} - \boldsymbol{A_{12.3}}\boldsymbol{A_{22.3}^{-1}}\boldsymbol{A_{24.3}} \\ &= (\boldsymbol{A_{14}} - \boldsymbol{A_{13}}\boldsymbol{A_{33}}^{-1}\boldsymbol{A_{34}}) \\ &- (\boldsymbol{A_{12}} - \boldsymbol{A_{13}}\boldsymbol{A_{33}}^{-1}\boldsymbol{A_{32}})\ (\boldsymbol{A_{22}} - \boldsymbol{A_{23}}\boldsymbol{A_{33}}^{-1}\boldsymbol{A_{32}})\big]^{-1}(\boldsymbol{A_{24}} - \boldsymbol{A_{23}}\boldsymbol{A_{33}}^{-1}\boldsymbol{A_{34}}) \\ &= \boldsymbol{V_{\hat{a}}}\boldsymbol{Z}. \end{aligned} \tag{6.12}$$

It then follows from (6.8) that

$$\begin{aligned}&\widehat{cov}(\boldsymbol{\delta}) \\ &= \boldsymbol{Z^T}\ \ \boldsymbol{V_{\hat{a}}} - \boldsymbol{V_{\hat{a}}}\ \ (\boldsymbol{V_{\hat{a}}} + \boldsymbol{V_{\hat{b}}}) - \boldsymbol{V_{\hat{b}}}\ (\boldsymbol{V_{\hat{b}}} + \boldsymbol{V_{\hat{d}}}) - \boldsymbol{V_{\hat{d}}}(\boldsymbol{V_{\hat{c}}} + \boldsymbol{V_{\hat{d}}})^{-1}\boldsymbol{V_{\hat{d}}}\big]^{-1}\boldsymbol{V_{\hat{b}}}\ \ ^{-1}\boldsymbol{V_{\hat{a}}}\ \ \boldsymbol{Z}. \end{aligned} \tag{6.13}$$

The k^{th} diagonal element from the $s \times s$ matrix sandwiched between $\boldsymbol{Z^T}$ and $\boldsymbol{Z}$ in (6.13) is equal to

$$\begin{aligned}&\hat{a}_k - \hat{a}_k\ \ (\hat{a}_k + \hat{b}_k) - \hat{b}_k\Big[(\hat{b}_k + \hat{d}_k) - \hat{d}_k(\hat{c}_k + \hat{d}_k)^{-1}\hat{d}_k\Big]^{-1}\hat{b}_k\ \ ^{-1}\hat{a}_k \\ =&\hat{a}_k - \hat{a}_k\ \ (\hat{a}_k + \hat{b}_k) - \hat{b}_k^2\frac{\hat{c}_k + \hat{d}_k}{\hat{b}_k\hat{c}_k + \hat{b}_k\hat{d}_k + \hat{c}_k\hat{d}_k}\ \ ^{-1}\hat{a}_k \\ =&\hat{a}_k - \hat{a}_k^2\frac{\hat{b}_k\hat{c}_k + \hat{b}_k\hat{d}_k + \hat{c}_k\hat{d}_k}{\hat{a}_k\hat{b}_k\hat{c}_k + \hat{a}_k\hat{b}_k\hat{d}_k + \hat{a}_k\hat{c}_k\hat{d}_k + \hat{b}_k\hat{c}_k\hat{d}_k} \\ =&\frac{\hat{a}_k\hat{b}_k\hat{c}_k\hat{d}_k}{\hat{a}_k\hat{b}_k\hat{c}_k + \hat{a}_k\hat{b}_k\hat{d}_k + \hat{a}_k\hat{c}_k\hat{d}_k + \hat{b}_k\hat{c}_k\hat{d}_k} \\ =&\frac{1}{\frac{1}{\hat{a}_k} + \frac{1}{\hat{b}_k} + \frac{1}{\hat{c}_k} + \frac{1}{\hat{d}_k}}. \end{aligned} \tag{6.14}$$

As the $\hat{b}_k$ and $\hat{c}_k$ are interchangeable, (6.14) proves that the covariance matrix for $\hat{\boldsymbol{\beta}}$ will be identical for the two types of logistic regression models.

Chapter 7

Prentice and Pyke's Theoretical Justification for Modeling the Log Odds of Disease in Case-Control Studies

In 1979, Prentice and Pyke [47] suggested a logistic regression model which treats the disease as the outcome for case-control studies. Let G (Gender) be the covariate relating to disease D and exposure E, the model for the log odds of disease can be written as

$$\log \frac{Pr(D=1|E,G)}{1-Pr(D=1|E,G)} = \alpha \ + \ \beta E \ + \ \gamma G \ + \ \delta EG. \tag{7.1}$$

The next section will introduce an important equivalence which Prentice and Pyke used to obtain the model for the log odds of disease. Subsequently, Prentice and Pyke's theoretical justification [47] for treating disease as the outcome in case-control studies will be provided.

7.1 The Logistic Model for Disease Incidence during a Defined Accession Period and the Corresponding Multinomial Logistic Regression Model

Letting G (Gender) be the covariate relating to disease D and exposure E, then the frequencies can be summarized as in Table 7.1.

According to Prentice and Pyke, "The logistic model for disease incidence during the defined accession period" [47] is given by

$$Pr^*(D=1|E,G) = \frac{\exp\{\alpha^* \ + \ \beta E \ + \ \gamma G \ + \ \delta EG\}}{1+\exp\{\alpha^* \ + \ \beta E \ + \ \gamma G \ + \ \delta EG\}} \tag{7.2}$$

Table 7.1: The Two 2×2 Tables (Frequencies)

	$G=0$		$G=1$	
	$E=1$	$E=0$	$E=1$	$E=0$
$D=1$	a_1	b_1	a_2	b_2
$D=0$	c_1	d_1	c_2	d_2

where α^* is the log odds of disease for the baseline category ($E = 0$ and $G = 0$) in "the logistic model [(7.2)] for disease incidence during the defined accession period" [47]. Without loss of generality, the frequencies shown in Table 7.1 can also be summarized in terms of a multinomial logistic regression model. Using $Pr(E = 0, G = 0)$ as the baseline, the corresponding multinomial logistic regression model can be written as

I.

$$\log \frac{Pr(E=1,G=0|D)}{Pr(E=0,G=0|D)} = \psi_e \ + \ \beta D. \tag{7.3}$$

II.

$$\log \frac{Pr(E=0,G=1|D)}{Pr(E=0,G=0|D)} = \psi_g \ + \ \gamma D. \tag{7.4}$$

III.

$$\log \frac{Pr(E=1,G=1|D)}{Pr(E=0,G=0|D)} = \psi_{eg} \ + \ (\beta+\gamma+\delta) D. \tag{7.5}$$

If $\delta = 0$, a constraint is imposed on the multinomial logistic regression model during the parameter estimation. The number of constraints needed in the multinomial logistic regression model depends on the degree of saturation with respect to the number of parameters in the logistic regression model.

Though only two dichotomous variables (E and G) are considered in this section, the congruence between the two types of logistic regression models can be readily applied to more than two dichotomous variables or even measured variables. Specifically, measured variables are treated (or pretended) as "only... values actually observed may be observed"

[26] in the maximum likelihood estimation for the multinomial logistic regression model. The following section shows such a simplified example.

Assume that the covariate (A_m) is a variable that has only three observed values(0, 1, and 2), "The logistic model for disease incidence during the defined accession period" [47]) can be written as

$$\log \frac{Pr^*(D=1|E,A_m)}{1-Pr^*(D=1|E,A_m)} = \alpha^* + \beta E + \gamma A_m + \delta E A_m \tag{7.6}$$

and the corresponding multinomial logistic regression model consists of the following five equations

I.
$$\log \frac{Pr(E=0,A_m=1|D)}{Pr(E=0,A_m=0|D)} = \psi_g + \gamma D.$$

II.
$$\log \frac{Pr(E=0,A_m=2|D)}{Pr(E=0,A_m=0|D)} = \psi_{g2} + 2\gamma D.$$

III.
$$\log \frac{Pr(E=1,A_m=0|D)}{Pr(E=0,A_m=0|D)} = \psi_e + \beta D.$$

IV.
$$\log \frac{Pr(E=1,A_m=1|D)}{Pr(E=0,A_m=0|D)} = \psi_{eg} + (\beta+\gamma+\delta) D.$$

V.
$$\log \frac{Pr(E=1,A_m=2|D)}{Pr(E=0,A_m=0|D)} = \psi_{eg2} + (\beta+2\gamma+2\delta) D. \tag{7.7}$$

There are two constraints involved in this simplest case, namely, $2(\beta+\gamma+\delta) = \beta + (\beta+2\gamma+2\delta)$ and $2(\gamma) = 2\gamma$.

7.2 Prentice and Pyke's Theoretical Justification

The key point in this section is to illustrate that the disease indicator D can be treated as the outcome in a case-control study [47]. Suppose that "the logistic model for disease incidence during the defined accession period" [47] is

$$Pr^*(D=1|E,G) = \frac{\exp\{\alpha^* + \beta E + \gamma G + \delta EG\}}{1+\exp\{\alpha^* + \beta E + \gamma G + \delta EG\}}. \tag{7.8}$$

For any given E and G, the following equivalence exists (as described in section 7.1)

$$\frac{\dfrac{Pr^*(D=1|E,G)}{Pr^*(D=0|E,G)}}{\dfrac{Pr^*(D=1|E=0,G=0)}{Pr^*(D=0|E=0,G=0)}} = \frac{\dfrac{Pr(E,G|D=1)}{Pr(E=0,G=0|D=1)}}{\dfrac{Pr(E,G|D=0)}{Pr(E=0,G=0|D=0)}} = \exp\{\beta E+\gamma G+\delta EG\}, \tag{7.9}$$

and an induced model [47] can be written as

$$\begin{aligned}
&Pr(E,G|D=d)\\
&= Pr(E=0,G=0|D=d)\,\frac{Pr(E,G|D=0)}{Pr(E=0,G=0|D=0)}\,\exp\{(\beta E + \gamma G + \delta EG)\,d\}\\
&= Pr(E=0,G=0|D=d)\,\exp\{\log\frac{Pr(E=e,G=g|D=0)}{Pr(E=0,G=0|D=0)} + (\beta E + \gamma G + \delta EG)\,d\}\\
&= c_d\,\exp\{\phi + (\beta E + \gamma G + \delta EG)\,d\}
\end{aligned} \tag{7.10}$$

where $c_d = Pr(E=0,G=0|D=d)$ and $\phi = \log\dfrac{Pr(E=e,G=g|D=0)}{Pr(E=0,G=0|D=0)}$.

Let

$$q(E,G) = \sum_{d=0}^{1} P_s(D=d)\,Pr(E,G|D=d) \tag{7.11}$$

where $P_s(D=d)$ is the probability of $D=d$ under the case-control sampling scheme. Let n be the sample size and n_d be the sample size for $D=d$, then

$$\begin{aligned}
q(E,G) &= \sum_{d=0}^{1} \frac{n_d}{n}\,Pr(E,G|D=d)\\
&= \sum_{d=0}^{1} \frac{n_d}{n}\,c_d\,\exp\{\phi + (\beta E + \gamma G + \delta EG)\,d\}\\
&= \exp\{\phi\}\sum_{d=0}^{1} \frac{n_d}{n}\,c_d\,\exp\{(\beta E + \gamma G + \delta EG)\,d\}.
\end{aligned} \tag{7.12}$$

$q(E,G)$ is actually the joint marginal density function [47] for E and G under the case-control sampling scheme. By reexpressing $\exp\{\phi\}$ in terms of $q(E,G)$ in the model (7.10), the induced model (7.10) becomes

$$\begin{aligned}
&Pr(E,G|D=d)\\
&= c_d\ \exp\{(\beta E\ +\ \gamma G\ +\ \delta EG)\ d\}\frac{q(E,G)}{\sum_{l=0}^{1}\ \frac{n_l}{n}\ c_l\ \exp\{(\beta E\ +\ \gamma G\ +\ \delta EG)\ l\}}\\
&= \frac{\frac{n_d}{n}\ c_d\ \exp\{(\beta E\ +\ \gamma G\ +\ \delta EG)\ d\}}{\sum_{l=0}^{1}\ \frac{n_l}{n}\ c_l\ \exp\{(\beta E\ +\ \gamma G\ +\ \delta EG)\ l\}}\ q(E,G)\ \frac{n}{n_d}\\
&= \frac{\exp\{\alpha_d + (\beta E\ +\ \gamma G\ +\ \delta EG)\ d\}}{\sum_{l=0}^{1}\exp\{\alpha_l + (\beta E\ +\ \gamma G\ +\ \delta EG)\ l\}}\ q(E,G)\ \frac{n}{n_d}\\
&= p_d(E,G)\ q(E,G)\ \frac{n}{n_d}\\
&= p_d(x)\ q(x)\ \frac{n}{n_d}\\
&= Pr(x|D=d) \qquad (d=0,1)
\end{aligned} \tag{7.13}$$

where $\alpha_d = \log c_d\frac{n_d}{n}$, $p_d(x) = p_d(E,G) = \frac{\exp\{\alpha_d + (\beta E\ +\ \gamma G\ +\ \delta EG)\ d\}}{\sum_{l=0}^{1}\exp\{\alpha_l + (\beta E\ +\ \gamma G\ +\ \delta EG)\ l\}}$, and $q(x) = q(E,G)$. Notice that $p_d(x)$ or $p_d(E,G)$ has a logistic form with parameters β, γ, and δ. Actually $p_d(x)$ is $Pr(D=d|x)$ in equation (1.28) and $q(x)$ is $Pr(x)$ in equation (1.28).

7.2.1 The Maximum Likelihood Estimators

This section will show how a constraint on defining $Pr(E,G|D=d)$ to be a probability is satisfied automatically when $p_d(\boldsymbol{x})$ and $q(\boldsymbol{x})$ are estimated separately. As the likelihood function for induced model (7.13) can be written as

$$L\ \propto\ \prod_{d=0}^{1}\prod_{h=1}^{n_d} p_d(x_{dh}) \qquad \prod_{d=0}^{1}\prod_{h=1}^{n_d} q(x_{dh})\ \ =\ \ L_1 \qquad L_2\ . \tag{7.14}$$

Parameter estimations with respect to β, γ, δ and α_d in L_1 and $q(\boldsymbol{x})$ in L_2 of (7.14) will separately be subject only to the constraint that $Pr(E, G|D = d)$ in equations (7.10) and (7.13) is a probability distribution [30]. That is

$$\text{the constraint:} \quad \int_x p_d(x)\, q(x)\, dx = \frac{n_d}{n} \quad (d = 0, 1). \tag{7.15}$$

The log likelihood function l_1 of L_1 in (7.14) is

$$\begin{aligned} l_1 &= \sum_{d=0}^{1} \sum_{h=1}^{n_d} \log\, p_d(x_{dh}) \\ &= \sum_{d=0}^{1} \sum_{h=1}^{n_d} (\alpha_d + (\beta E_h \,+\, \gamma G_h \,+\, \delta E_h G_h)\, d) \\ &\qquad + \log \sum_{l=0}^{1} \exp\{\alpha_l + (\beta E_h \,+\, \gamma G_h \,+\, \delta E_h G_h)\, l\}\Big] \end{aligned} \tag{7.16}$$

where E_h is the exposure status and G_h is the gender status for the h^{th} observation. The first partial derivative with respect to α_d is

$$\frac{\partial\, l_1}{\partial \alpha_d} = n_d - \sum_{m=0}^{1} \sum_{g=1}^{n_l} p_d(x_{mg}). \tag{7.17}$$

After setting the above score function equal zero, the following equation is obtained.

$$\begin{aligned} & \sum_{m=0}^{1} \sum_{g=1}^{n_l} E[p_d(x)] = n_d \\ \Rightarrow \quad & E[p_d(x)] = \frac{n_d}{n} \\ \Rightarrow \quad & \int_x p_d(x)\, q(x)\, dx = \frac{n_d}{n} \\ & (d = 0, 1). \end{aligned} \tag{7.18}$$

As the last equation in (7.18) shows that the constraint is satisfied even if the parameters included in L_1 and L_2 (equation 7.14) are estimated separately, this means

that β, γ, δ and α_d can be estimated without estimating $q(x)$. Therefore, the estimated odds ratios and variances can be obtained only through L_1 in the log likelihood function (7.14). Specifically, the model for the log odds of disease can be expressed in terms of $p_d(E,G)$ from (7.13). That is

$$\begin{aligned}
\text{D=1:}\quad p_1(E,G) &= \frac{\exp\{\alpha_1 + \beta E + \gamma G + \delta EG\}}{\exp\{\alpha_0\} + \exp\{\alpha_1 + \beta E + \gamma G + \delta EG\}} \\
&= \frac{\exp\{(\alpha_1 - \alpha_0) + \beta E + \gamma G + \delta EG\}}{1 + \exp\{(\alpha_1 - \alpha_0) + \beta E + \gamma G + \delta EG\}} \\
&= \frac{\exp\{\alpha + \beta E + \gamma G + \delta EG\}}{1 + \exp\{\alpha + \beta E + \gamma G + \delta EG\}}, \\
\text{D=0:}\quad p_0(E,G) &= \frac{1}{1 + \exp\{\alpha + \beta E + \gamma G + \delta EG\}}.
\end{aligned} \tag{7.19}$$

model (7.19) re-expressed as the model (7.1), i.e.

$$Pr(D = 1|E,G) = \frac{\exp\{\alpha + \beta E + \gamma G + \delta EG\}}{1 + \exp\{\alpha + \beta E + \gamma G + \delta EG\}} \tag{7.20}$$

where $\alpha = \alpha_1 - \alpha_0 = \alpha^*$ if models (7.1)(7.20) and the model (7.8) are compared. It is noted that α cannot be interpreted as the log odds of disease for the baseline ($E = 0$ and $G = 0$) in models (7.1)(7.20) for the log odds of disease as "the [case-control] study gives no information about the marginal probability of disease [for the source population]" [19]. But the odds ratio $\exp\{\beta + \delta\}$ in models (7.1)(7.20) for the log odds of disease will have the same interpretation as the odds ratio in "the logistic model (7.8) for disease incidence during the defined accession period" [47]. Thus, data from case-control studies can be analyzed by the model for the log odds of disease to obtain the estimates for β, γ, and δ.

The premise in the preceding section is that the disease can be treated as the outcome in case-control studies. It is important to note that it does not matter which terms are included on the right hand side of the model (7.8), that is, the part ($\beta E + \gamma G + \delta EG$) is not involved in the preceding demonstration of treating the disease indicator as the

outcome variable. γ and/or δ can be zero. In addition, the right-hand side can consist of more than one covariate or even measured variables.

7.2.2 The Asymptotic Distribution of $\hat{\boldsymbol{\beta}}$

A first-order Taylor series expansion [32] of the score function for l_1 in (7.16) about the 'true' parameter $\boldsymbol{\theta^0}$ is

$$0 = \frac{\partial\, l_1}{\partial \hat{\boldsymbol{\theta}}} = \frac{\partial\, l_1}{\partial \boldsymbol{\theta^0}}|_{\boldsymbol{\theta^*}=\boldsymbol{\theta^0}} + \frac{\partial^2\, l_1}{\partial \boldsymbol{\theta^\star}\partial \boldsymbol{\theta^\star}}|_{\boldsymbol{\theta^*}=\boldsymbol{\theta^\star}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta^0}) \tag{7.21}$$

where $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_s \\ \dots \\ \beta \\ \gamma \\ \delta \end{pmatrix}$, i.e., $\boldsymbol{\alpha} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_s \end{pmatrix}$ and $\boldsymbol{\beta} = \begin{pmatrix} \beta \\ \gamma \\ \delta \end{pmatrix}$.

$\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator and $\boldsymbol{\theta^\star}$ is between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta^0}$, so that

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta^0}) = (-n^{-1}\frac{\partial^2\, l_1}{\partial \boldsymbol{\theta^\star}\partial \boldsymbol{\theta^\star}})^{-1}(n^{-\frac{1}{2}}\frac{\partial\, l_1}{\partial \boldsymbol{\theta^0}}) = \boldsymbol{I}(\boldsymbol{\theta^\star})^{-1}\boldsymbol{S}(\boldsymbol{\theta^0}). \tag{7.22}$$

7.2.2.1 The Asymptotic Distribution of $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta^0})$

I. The Asymptotic Distribution of $\boldsymbol{S}(\boldsymbol{\theta^0})$

The contributions to the score statistic $\boldsymbol{S}(\boldsymbol{\theta^0})$ of L_1 in equation (7.14) from the individual samples (i.e., an individual disease group) do not in general have mean zero, that is, $E\quad {}^{n_d}\frac{\partial \log p_d(x)}{\partial \boldsymbol{\theta^0}}]$ will not in general be zero. Therefore, the variance for $\boldsymbol{S}(\boldsymbol{\theta^0})$ is not $\boldsymbol{G}(\boldsymbol{\theta^0})$ where $\boldsymbol{G}(\boldsymbol{\theta^0})$ $(= E[\boldsymbol{I}(\boldsymbol{\theta^0})])$ is the variance

for the prospective sampling score statistic. Let $\boldsymbol{\mu_m^0}$ be the contribution to the expectation $E\ \boldsymbol{S(\theta^0)}\big]$ from the m^{th} disease group ($m = 0, ..., s$). Since $E\ \boldsymbol{S(\theta^0)}\big] = \mathbf{0}$ (that is, if one were to repeatedly sample from a distribution and the mean of the score with the 'true' $\boldsymbol{\theta^0}$ would tend to zero as the number of repeated sample approaches infinity), $\sum_{m=0}^{s} n_m \boldsymbol{\mu_m^0} = \mathbf{0}$. Then

$$\begin{aligned}
\boldsymbol{S(\theta^0)} =& n^{-\frac{1}{2}} \frac{\partial l_1}{\partial \boldsymbol{\theta^0}} \\
=& n^{-\frac{1}{2}} \sum_{m=0}^{s} \sum_{g=1}^{n_m} \frac{\partial \log p_m(x_{mg})}{\partial \boldsymbol{\theta^0}} \\
=& \sum_{m=0}^{s} (\frac{n}{n_m})^{-\frac{1}{2}} \ n_m^{\frac{1}{2}} \sum_{g=1}^{n_m} \frac{\partial \log p_m(x_{mg})}{\partial \boldsymbol{\theta^0}} - \boldsymbol{\mu_m^0}\big] \ + n^{-\frac{1}{2}} \sum_{m=0}^{s} n_m \boldsymbol{\mu_m^0}.
\end{aligned} \tag{7.23}$$

By applying the Central Limit Theorem [30] to the term in the curly brackets, $\boldsymbol{S(\theta^0)}$ is asymptotically normal distributed with mean 0 and variance matrix

$$\boldsymbol{\Sigma} = var\ \ \boldsymbol{S(\theta^0)}\ \ = \frac{1}{n} E\ \ (\frac{\partial l_1}{\partial \boldsymbol{\theta^0}})(\frac{\partial l_1}{\partial \boldsymbol{\theta^0}})^T\ \ . \tag{7.24}$$

II. $\boldsymbol{I(\theta^\star)}$ is the consistent estimator of $\boldsymbol{G(\theta^0)}$ $\boldsymbol{\hat{\theta}}$ is a consistent estimator of $\boldsymbol{\theta^0}$ [30] and $\boldsymbol{\theta^\star}$ lies between $\boldsymbol{\theta^0}$ and $\boldsymbol{\hat{\theta}}$ so that $\boldsymbol{\theta^\star}$ is a consistent estimator of $\boldsymbol{\theta^0}$ too. This implies that $\boldsymbol{G(\theta^\star)}$ is a consistent estimator of $\boldsymbol{G(\theta^0)}$ [30]. Also by the strong law of large numbers [52], $\boldsymbol{I(\theta^\star)}$ will almost surely converge to its expectation $\boldsymbol{G(\theta^\star)}$. It follows then that $\boldsymbol{I(\theta^\star)}$ is a consistent estimator of $\boldsymbol{G(\theta^0)}$.

With equation (7.22) and by Slutsky's Theorem [30], we have

$$\begin{aligned}
var\ \ n^{\frac{1}{2}}(\boldsymbol{\hat{\theta}} - \boldsymbol{\theta^0})\ \ = var\ \ \boldsymbol{I(\theta^\star)}^{-1}\boldsymbol{S(\theta^0)}\ \ &= var\ \ \boldsymbol{G(\theta^0)}^{-1}\boldsymbol{S(\theta^0)} \\
&= \boldsymbol{G(\theta^0)}^{-1} var\ \ \boldsymbol{S(\theta^0)}\ \ \boldsymbol{G(\theta^0)}^{-1} \\
&= \boldsymbol{G}^{-1}\Sigma\boldsymbol{G}^{-1}
\end{aligned} \tag{7.25}$$

where $\boldsymbol{G} = \boldsymbol{G}(\boldsymbol{\theta^0})$. Now $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta^0})$ in equation (7.22) has an asymptotic distribution with mean zero and covariance matrix $\boldsymbol{G}^{-1}\boldsymbol{\Sigma}\boldsymbol{G}^{-1}$.

7.2.2.2 The asymptotic covariance matrix $[\boldsymbol{G^{-1}\Sigma G^{-1}}]_{\boldsymbol{22}}$ (i.e., the covariance matrix of $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta^0})$ corresponding to $\boldsymbol{\beta}$) is equal to $\boldsymbol{G_{22}^{-1}}$

Suppose that we have s disease groups. From equation (7.18), we have

$$
\begin{aligned}
n_i &= n \int_x p_i(x)\, q(x)\, dx \\
&= n \int_x p_i(x) \Big[\sum_{m=0}^{s} p_m(x)\Big]\, q(x)\, dx \\
&= n \int_x p_i(x)p_0(x)\, q(x)\, dx + n \int_x p_i(x)p_1(x)\, q(x)\, dx + ... + n \int_x p_i(x)p_s(x)\, q(x)\, dx \\
&= \sum_{g=1}^{n_0} E\ p_i(x_{0g})\Big] + \sum_{g=1}^{n_1} E\ p_i(x_{1g})\Big] + ... + \sum_{g=1}^{n_s} E\ p_i(x_{sg})\Big] \\
&= n_0 E\ p_i(x_{0g})\Big] + n_1 E\ p_i(x_{1g})\Big] + ... + n_s E\ p_i(x_{sg})\Big] \\
&\qquad\qquad (i, m = 0, ..., s) \qquad (7.26)
\end{aligned}
$$

which implies

$$
\begin{aligned}
E\ p_i(x_{mg})\Big] &= \frac{n}{n_m} \int_x p_i(x)p_m(x)\, q(x)\, dx = \frac{n}{n_m} a_{im} \\
&\Rightarrow \\
\sum_{m=0}^{s}\sum_{g=1}^{n_m} E\ p_i(x_{mg})\Big] &= \sum_{m=0}^{s}\sum_{g=1}^{n_m} \int_x p_i(x)p_m(x)\, q(x)\, dx \\
&= \sum_{m=0}^{s}\sum_{g=1}^{n_m} \frac{n}{n_m} a_{im} \\
&= n \sum_{m=0}^{s} a_{im} \\
&= n_i \qquad (7.27)
\end{aligned}
$$

where $a_{im} = \int_x p_i(x)p_m(x)\, q(x)\, dx$. Similarly,

$$
\begin{aligned}
E\ p_i(x_{mg})p_j(x_{mg})\big] &= \frac{n}{n_m} \int_x p_i(x)p_j(x)p_m(x)\, q(x)\, dx \\
&\Rightarrow \\
\sum_{m=0}^{s}\sum_{g=1}^{n_m} E\ p_i(x_{mg})p_j(x_{mg})\big] &= \sum_{m=0}^{s}\sum_{g=1}^{n_m} \frac{n}{n_m} \int_x p_i(x)p_j(x)p_m(x)q(x)\, dx \\
&= n \int_x p_i(x)p_j(x) \Big[\sum_{m=0}^{s} p_m(x)\Big] q(x)\, dx \\
&= n \int_x p_i(x)p_j(x)q(x)\, dx \\
&= na_{ij}. \qquad (7.28)
\end{aligned}
$$

The elements in $\boldsymbol{G_{\alpha,\alpha}}$ will be

$$
\begin{aligned}
\boldsymbol{G_{\alpha_i,\alpha_i}} = E\ -\frac{1}{n}\frac{\partial^2\, l_1}{\partial(\alpha_i)^2}\ &= -\frac{1}{n} \sum_{m=0}^{s}\sum_{g=1}^{n_m} E\ -\ p_i(x_{mg}) - p_i(x_{mg})p_i(x_{mg}) \\
&= -\frac{1}{n}\ -\ n_i - na_{ii} \\
&= \frac{n_i}{n} - a_{ii} \qquad (7.29)
\end{aligned}
$$

and

$$
\begin{aligned}
\boldsymbol{G_{\alpha_i,\alpha_j}} = E\ -\frac{1}{n}\frac{\partial^2\, l_1}{\partial\alpha_i\partial\alpha_j}\ &= -\frac{1}{n} \sum_{m=0}^{s}\sum_{g=1}^{n_m} E\ -\ -\ p_i(x_{mg})p_j(x_{mg}) \\
&= -\frac{1}{n}\ na_{ij} \\
&= -a_{ij}. \qquad (7.30)
\end{aligned}
$$

$\boldsymbol{G_{\alpha,\alpha}}$ can then be expressed in matrix notation. Let

$$
\boldsymbol{A} = \begin{pmatrix} a_{11} & \dots & a_{1s} \\ \vdots & \ddots & \vdots \\ a_{s1} & \dots & a_{ss} \end{pmatrix}
$$

and

$$\boldsymbol{N} = \begin{pmatrix} \frac{n_1}{n} & & 0 \\ & \ddots & \\ 0 & & \frac{n_k}{n} \end{pmatrix},$$

then

$$\boldsymbol{G_{\alpha,\alpha}} = \boldsymbol{N} - \boldsymbol{A}. \tag{7.31}$$

From equations (7.17) and (7.24), it follows that

$$\frac{\partial\, l_1}{\partial \alpha_i} = n_i - \sum_{m=0}^{s} \sum_{g=1}^{n_m} p_i(x_{mg}), \tag{7.32}$$

so that

$$\begin{aligned}
\boldsymbol{\Sigma_{\alpha_i,\alpha_j}} &= \frac{1}{n} E\ \frac{\partial l_1}{\partial \alpha_i}\frac{\partial l_1}{\partial \alpha_j} \\
&= \frac{1}{n} E\ (n_i - \sum_{m=0}^{s} \sum_{g=1}^{n_m} p_i(x_{mg}))(n_j - \sum_{m=0}^{s} \sum_{g=1}^{n_m} p_j(x_{mg})) \\
&= \frac{1}{n} E \sum_{m=0}^{s} \sum_{g=1}^{n_m} p_i(x_{mg}) \sum_{m=0}^{s} \sum_{g=1}^{n_m} p_j(x_{mg}) - n_i n_j \\
&= \frac{1}{n} E \sum_{m=0}^{s} \sum_{g=1}^{n_m} p_i(x_{mg}) p_j(x_{mg}) + \sum_{l=0}^{s} \sum_{t=1}^{n_l} \sum_{l\neq r \ \text{or}\ t\neq q} p_i(x_{lt}) p_j(x_{rq}) - n_i n_j \\
&= \frac{1}{n}\ n a_{ij} + \sum_{l=0}^{s} \sum_{t=1}^{n_l} \sum_{l\neq r \ \text{or}\ t\neq q} E\ p_i(x_{lt}) p_j(x_{rq})\big] - n_i n_j\ .
\end{aligned} \tag{7.33}$$

Because x_{lt} and x_{rq} are independent when $l = r$ or $t = q$, it means that

$$E\ p_i(x_{lt}) p_j(x_{rj})\big] = E\ p_i(x_{lt})\big] E\ p_j(x_{rq})\big] = \ \frac{n}{n_i} a_{il}\big]\ \frac{n}{n_j} a_{jr}\big], \tag{7.34}$$

so that

$$\begin{aligned}
&\sum_{l=0}^{s}\sum_{t=1}^{n_l}\sum_{r=l \text{ or } t=q} E\ p_i(x_{lt})p_j(x_{rq})\Big] \\
&= \sum_{l=0}^{s}\sum_{t=1}^{n_l} E\ p_i(x_{lt}) \sum_{r=0}^{s}\sum_{q=1}^{n_l} E\ p_j(x_{rq}) \ - \sum_{m=0}^{s}\sum_{g=1}^{n_m} E\ p_i(x_{mg})\ E\ p_j(x_{mg}) \\
&=(n\sum_{l=0}^{s} a_{il})(n\sum_{r=0}^{s} a_{jr}) - \sum_{m=0}^{s}\sum_{g=1}^{n_m} \frac{n}{n_m}a_{im}\frac{n}{n_m}a_{jm} \\
&=n_in_j - \sum_{m=0}^{s}(\frac{1}{n_m}na_{im}na_{jm}).
\end{aligned} \tag{7.35}$$

It follows that

$$\begin{aligned}
\boldsymbol{\Sigma_{\alpha_i,\alpha_j}} &= \frac{1}{n}\ na_{ij} + \sum_{l=0}^{s}\sum_{t=1}^{n_l}\sum_{r=l \text{ or } q=t} E\ p_i(x_{lt})p_j(x_{rq})\Big] - n_in_j \\
&= \frac{1}{n}\ na_{ij} + n_in_j - \sum_{m=0}^{s}(\frac{1}{n_m}na_{im}na_{jm}) - n_in_j \\
&= \frac{1}{n}\ na_{ij} - \sum_{m=0}^{s}\frac{1}{n_m}na_{im}na_{jm} \\
&= a_{ij} - n\sum_{m=1}^{s}(\frac{1}{n_m}a_{im}a_{jm}) - n(\frac{1}{n_0}a_{i0}a_{j0}) \\
&= a_{ij} - n\sum_{m=1}^{s}(\frac{1}{n_m}a_{im}a_{jm}) - \frac{n}{n_0}(\frac{n_i}{n} - \sum_{m=1}^{s} a_{im})(\frac{n_j}{n} - \sum_{m=1}^{s} a_{jm}).
\end{aligned} \tag{7.36}$$

$\boldsymbol{\Sigma_{\alpha,\alpha}}$ can then be expressed in matrix notation. Let

$$x = \begin{pmatrix} \frac{n}{n_0}+\frac{n}{n_1} & & \frac{n}{n_0} \\ & \ddots & \\ \frac{n}{n_0} & & \frac{n}{n_k}+\frac{n}{n_0} \end{pmatrix},$$

then

$$\begin{aligned}
\boldsymbol{\Sigma_{11}} = \boldsymbol{\Sigma_{\alpha,\alpha}} &= \boldsymbol{A} - \boldsymbol{AXA} - (\boldsymbol{NXN} - \boldsymbol{N}) + (\boldsymbol{AXN} - \boldsymbol{A}) + (\boldsymbol{NXA} - \boldsymbol{A}) \\
&= \boldsymbol{N} - \boldsymbol{A} - (\boldsymbol{N} - \boldsymbol{A})\boldsymbol{X}(\boldsymbol{N} - \boldsymbol{A}) \\
&= \boldsymbol{G_{\alpha,\alpha}} - \boldsymbol{G_{\alpha,\alpha}}\boldsymbol{X}\boldsymbol{G_{\alpha,\alpha}} \\
&= \boldsymbol{G_{11}} - \boldsymbol{G_{11}}\boldsymbol{X}\boldsymbol{G_{11}}.
\end{aligned} \tag{7.37}$$

Following the same procedure, we can also obtain

$$\begin{aligned}
\boldsymbol{\Sigma_{12}} &= \boldsymbol{\Sigma_{\alpha,\beta}} = \boldsymbol{G_{12}} - \boldsymbol{G_{11}XG_{12}}, \\
\boldsymbol{\Sigma_{21}} &= \boldsymbol{\Sigma_{\beta,\alpha}} = \boldsymbol{G_{21}} - \boldsymbol{G_{21}XG_{11}}, \\
\boldsymbol{\Sigma_{22}} &= \boldsymbol{\Sigma_{\beta,\beta}} = \boldsymbol{G_{22}} - \boldsymbol{G_{21}XG_{12}}.
\end{aligned} \tag{7.38}$$

Then

$$\begin{aligned}
\boldsymbol{G^{-1}\Sigma G^{-1}} &= \boldsymbol{G^{-1}} \begin{pmatrix} \boldsymbol{G_{11} - G_{11}XG_{11}} & \boldsymbol{G_{21} - G_{21}XG_{11}} \\ \boldsymbol{G_{11} - G_{11}XG_{11}} & \boldsymbol{G_{22} - G_{21}XG_{12}} \end{pmatrix} \boldsymbol{G^{-1}} \\
&= \boldsymbol{G^{-1}}(\boldsymbol{G} - \begin{pmatrix} \boldsymbol{G_{11}XG_{11}} & \boldsymbol{G_{21}XG_{11}} \\ \boldsymbol{G_{11}XG_{11}} & \boldsymbol{G_{21}XG_{12}} \end{pmatrix})\boldsymbol{G^{-1}} \\
&= \boldsymbol{G^{-1}}(\boldsymbol{G} - \boldsymbol{G} \begin{pmatrix} \boldsymbol{X} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \boldsymbol{G})\boldsymbol{G^{-1}} \\
&= \boldsymbol{G^{-1}GG^{-1}} - \boldsymbol{G^{-1}G} \begin{pmatrix} \boldsymbol{X} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \boldsymbol{GG^{-1}} \\
&= \boldsymbol{G^{-1}} - \begin{pmatrix} \boldsymbol{X} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}.
\end{aligned} \tag{7.39}$$

That is $[\boldsymbol{G^{-1}\Sigma G^{-1}}]_{\boldsymbol{22}} = \boldsymbol{G_{22}^{-1}}$. $\boldsymbol{G_{22}^{-1}}$ can be consistently estimated by $\boldsymbol{I(\hat{\theta})_{22}^{-1}}$. "An asymptotic distribution for $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta^0})$ [corresponding to $\hat{\boldsymbol{\beta}}$] with mean zero and variance matrix $\boldsymbol{I(\hat{\theta})_{22}^{-1}}$ is precisely the distributional statement that would arise if the prospective model (7.20) were directly applied to the case-control data, as if a prospective study had been conducted" [47].

Chapter 8

A Case Study

In this chapter, data from a case-control study will again be used to compare differences between the two models (the model for the log odds of disease and the model for the log odds of exposure) on the basis of the overall fit and diagnostic statistics . The data are from Chapter 1 Section 1.3.

8.1 Building Models

Tavani et al. provided evidence that diabetes mellitus is a contributor to the risk of acute myocardial infarction [53]. History of hypertension, hyperlipidemia and smoking status are adjusted for biologic rationales [6][29]. These variables, as well as age and gender, were considered as covariates and retained in the logistic regression models. Effect modifications were assessed in the two models. Tests for nonlinearity and collinearity were performed for the two models. The final model for the log odds of disease was derived to be

$$\log \frac{Pr(D=1|E,A,G,L,H,S)}{1-Pr(D=1|E,A,G,L,H,S)} = \alpha + \beta E + \gamma_1 A + \gamma_2 G + \gamma_3 L + \gamma_4 H + \gamma_5 S \quad (8.1)$$

and the final model for the log odds of exposure was derived to be

$$\log \frac{Pr(E=1|D,A,G,L,H,S)}{1-Pr(E=1|D,A,G,L,H,S)} = \upsilon + \tau D + \theta_1 A + \theta_2 G + \theta_3 L + \theta_4 H + \theta_5 S. \quad (8.2)$$

Table 8.1 reports the estimates of the coefficients and standard errors in the model for the log odds of disease and Table 8.2 reports the estimates of the coefficients and standard

errors in the model for the log odds of exposure. As expected, the estimated coefficients and their standard errors are not identical because covariate effects are not saturated with parameters in both models. The estimated odds of developing MI with a history of diabetes is 2.1663 (95% CI: 1.3579, 3.4558) times the estimated odds of developing MI without a history of diabetes after controlling the confounding. The estimated odds of having a history of diabetes in the group with MI is 2.1596 (95% CI: 1.3532, 3.4465) times the estimated odds of having a history of diabetes in the group without MI after controlling the confounding. The results are different. However, the difference is small.

Table 8.1: Estimated Coefficients and Standard Errors from Fitting the Model for the Log Odds of Disease

Variable	Coefficient	Std. Err.
exp	**0.7730**	**(0.2383)**
age	0.0136	(0.0132)
gender	0.5498	(0.2321)
hypertension_history	0.5703	(0.2107)
hyperlipidemia_history	-0.0463	(0.2231)
smoking_history	-0.0407	(0.2205)
Intercept	-0.9001	(0.3006)

Table 8.2: Estimated Coefficients and Standard Errors from Fitting the Model for the Log Odds of Exposure

Variable	Coefficient	Std. Err.
dis	**0.7699**	**(0.2385)**
age	-0.0245	(0.0147)
gender	-0.8244	(0.2510)
hypertension_history	0.0864	(0.2465)
hyperlipidemia_history	-0.0066	(0.2533)
smoking_history	-0.6980	(0.2594)
Intercept	-0.6733	(0.3192)

8.2 Assessing Differences

Differences between the two logistic regression models can be examined through the overall fit and diagnostic statistics.

8.2.1 Overall Fit

Sensitivity and specificity can be used in model assessment. Sensitivity measures the proportion of diseased subjects identified through a specified classification rule and specificity measures the proportion of nondiseased subjects identified through a classification rule. In the model for the log odds of disease, the fitted disease/nondisease probabilities and a "disease probability cutoff" were used to classify each individual as either "diseased" or "not diseased" and all subjects' actual disease statuses were compared with this classification to determine the disease sensitivity and specificity. Similarly, in the model for the log odds of exposure, the fitted exposure/nonexposure probabilities and an "exposure probability cutoff" were used to classify each individual as either "exposed" or "not exposed" and all subjects' actual exposure statuses were compared with this classification to determine the exposure sensitivity and specificity. Figure 8.1 shows the disease sensitivity and specificity versus various cutoff values and Figure 8.2 shows the exposure sensitivity and specificity versus various cutoff values. Sensitivity versus 1-specificity can also be used for model assessment. The Area Under the Curve (AUC) in Figure 8.3 and 8.4 increases when sensitivity increases and 1-specificity decreases. The difference between the observed AUC values (0.6331 versus 0.6619) illustrates that the curves are generated from different fitted values arising from two different regression models. It's interesting to note that even though the coefficients relating to exposure and disease may be equivalent, the other coefficients are numerically different between the two regression models.

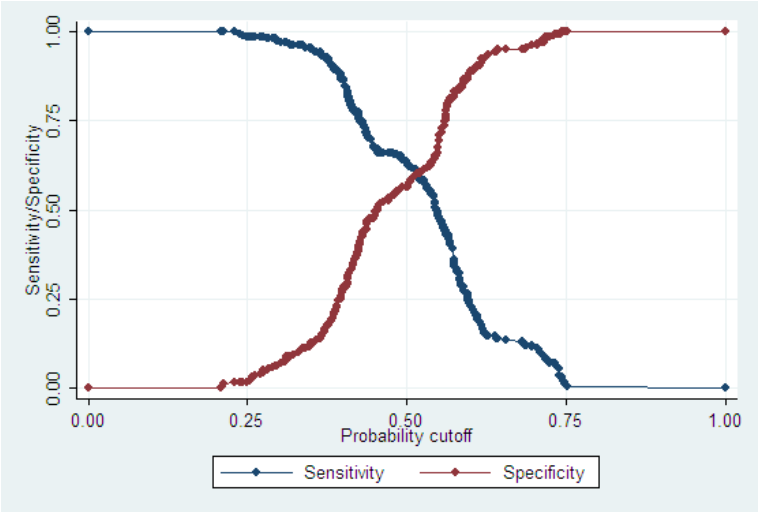

Figure 8.1: Plot of Sensitivity and Specificity versus all Possible Disease Cutoffs from Fitting the Model for the Log Odds of Disease

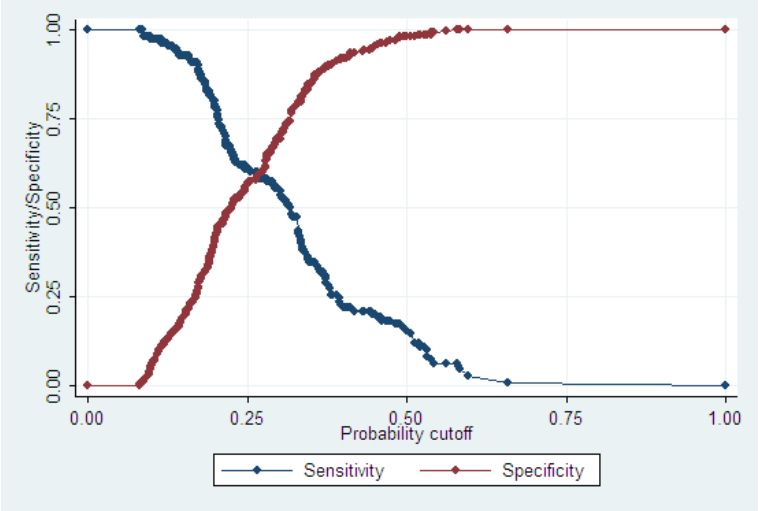

Figure 8.2: Plot of Sensitivity and Specificity versus all Possible Exposure Cutoffs from Fitting the Model for the Log Odds of Exposure

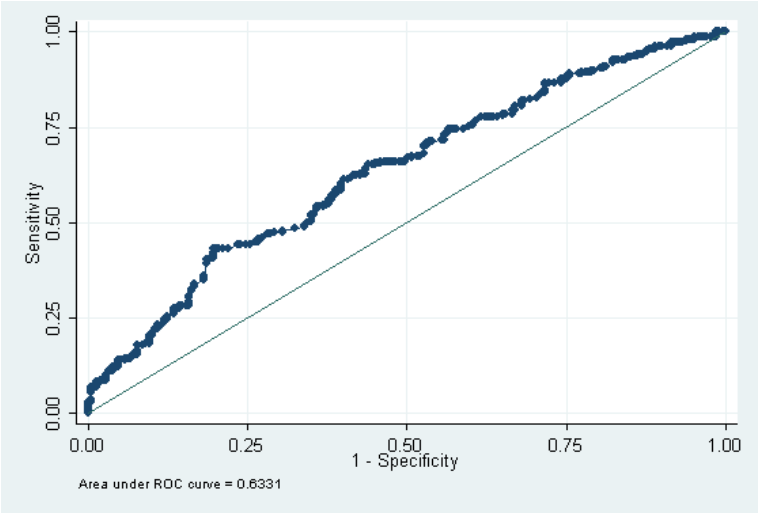

Figure 8.3: Plot of Sensitivity versus 1- Specificity versus all Possible Disease Cutoffs from Fitting the Model for the Log Odds of Disease

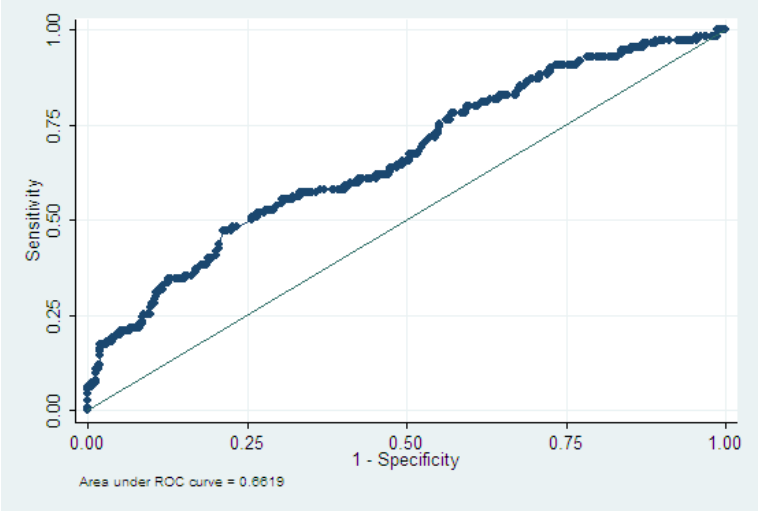

Figure 8.4: Plot of Sensitivity versus 1- Specificity versus all Possible Exposure Cutoffs from Fitting the Model for the Log Odds of Exposure

8.2.2 Diagnostic Statistics

The analysis of influential observations is an important step in regression diagnostics as these influential observations may be data entry errors or they may be of interest to investigate.

8.2.2.1 Outliers

In the following set of examples, it will be shown that the standardized residuals reveal two very different sets of extreme values, those beyond two standard deviations of the mean. It will also be shown that the model for log odds of disease hides extreme standardized residuals whereas the model for log odds of exposure uncovers standardized residuals beyond two standard deviations.

(a) Pearson Residual

The Pearson residual measures the difference between the observed and fitted frequency for a group of observations (j^{th}) with the same observed independent variable values [31][45]. It can be expressed as

$$r_j = \frac{y_j - m_j\hat{p}_j}{\sqrt{m_j\hat{p}_j(1-\hat{p}_j)}}. \tag{8.3}$$

The standardized Pearson residual is written as

$$r_{sj} = \frac{r_j}{\sqrt{1-h_{jj}}} \tag{8.4}$$

where h_{jj} is the h^{th} diagonal element in the projection matrix. The summation of the standardized Pearson residuals

$$\chi^2 = \sum_{j=1}^{J} r_{sj}^2 \tag{8.5}$$

has a limiting χ^2 distribution with degrees of freedom $J-(\pi+1)$ where J is the number of distinct groups with the same observed independent variable

values and $\pi + 1$ is the number of parameters in the model. Figures 8.5 and 8.6 are the plots of standardized Pearson residuals vs. predicted disease and exposure probabilities. It is interesting to note that the two graphs have different patterns and that different cuts of standardized Pearson residuals arise from the two different models (Table 8.3). The model for the log odds of exposure has a few more poorly fitted points (with a 2 or larger absolute value) than the model for the log odds of disease.

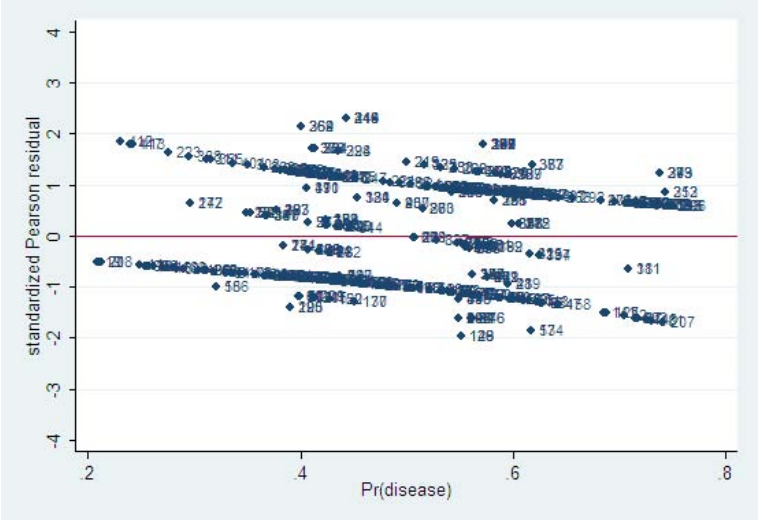

Figure 8.5: Plot of Standardized Pearson Residuals vs. Predicted Disease Probabilities from Fitting the Model for the Log Odds of Disease

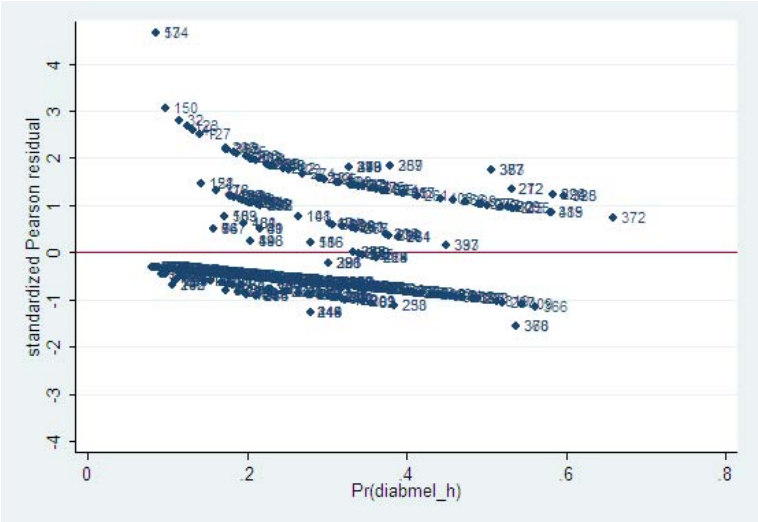

Figure 8.6: Plot of Standardized Pearson Residuals vs. Predicted Exposure Probabilities from Fitting the Model for the Log Odds of Exposure

Table 8.3: Extreme Standardized Pearson residuals

The Model for the Log Odds of Disease		The Model for the Log Odds of Exposure	
Index	Standardized Pearson Residual	Index	Standardized Pearson Residual
244	2.2955	23	2.1060
245	2.2955	32	2.7976
259	2.1519	45	2.5882
349	2.2955	53	4.666
362	2.1519	123	2.6688
364	2.1519	127	2.4973
416	2.2955	150	3.0668
		174	4.666
		175	2.1323
		203	2.0010
		232	2.2141
		302	2.1865
		368	2.2012
		408	2.0293

(b) Deviance Residual

Deviance residual "measures the disagreement between the maxima of the observed and fitted log likelihood functions" [45]. The formula is

$$
\begin{aligned}
d_j &= sgn \quad 2 \; l(\tilde{p}_j; y_j) - l(\hat{p}_j; y_j) \quad ^{\frac{1}{2}} \\
&= sgn \quad 2 \quad \log[\; \frac{y_j}{m_j} \quad ^{y_j} \quad \frac{m_j - y_j}{m_j} \quad ^{m_j - y_j}] - \log[\hat{p}_j^{y_j}(1-\hat{p}_j)^{m_j - y_j}] \quad ^{\frac{1}{2}} \\
&= sgn \quad 2 \quad y_j \log \frac{y_j}{m_j \hat{p}_j} + (m_j - y_j) \log \frac{m_j - y_j}{m_j(1-\hat{p}_j)} \quad ^{\frac{1}{2}}
\end{aligned} \tag{8.6}
$$

where the sign is the same as the sign of $(y_j - m_j\hat{p}_j)$ and $\tilde{p}_j = \frac{y_j}{m_j}$. The deviance $\sum_{j=1}^{J} d_j^2$ has a limiting distribution χ^2 with degrees of freedom $J - (\pi + 1)$. Figures 8.7 and 8.8 are the plots of deviance residuals vs. predicted disease and exposure probabilities. Table 8.4 displays all deviance residuals with a 2 or larger absolute value. The model for the log odds of exposure has a few more poorly fitted points than the model for the log odds of disease.

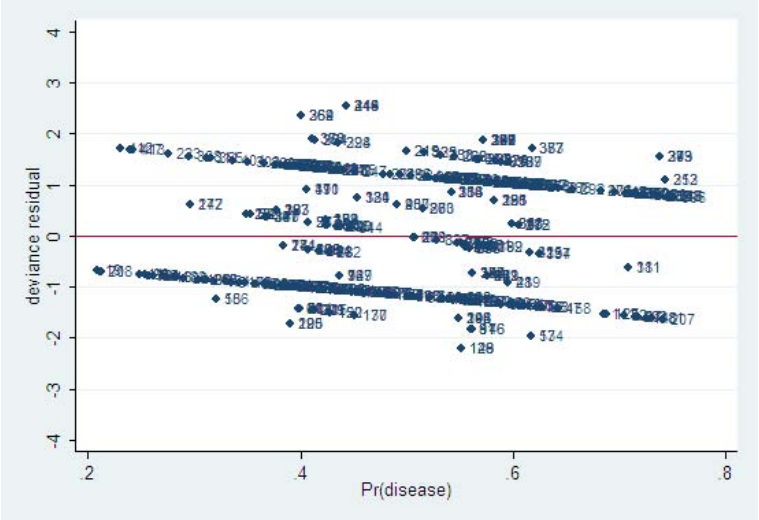

Figure 8.7: Plot of Deviance Residuals vs. Predicted Disease Probabilities from Fitting the Model for the Log Odds of Disease

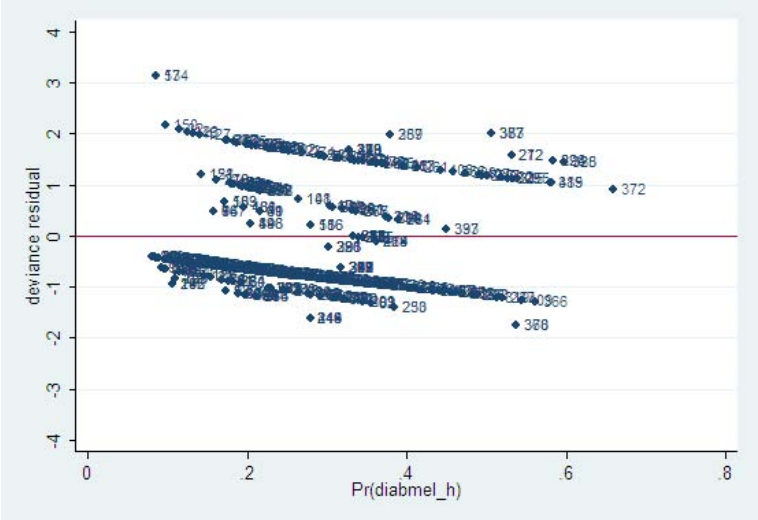

Figure 8.8: Plot of Deviance Residuals vs. Predicted Exposure Probabilities from Fitting the Model for the Log Odds of Exposure

Table 8.4: Extreme Deviance Residuals

The Model for the Log Odds of Disease		The Model for the Log Odds of Exposure	
Index	Deviance Resudual	Index	Deviance Residual
244	2.5498	32	2.0818
245	2.5498	45	2.0139
259	2.3428	53	3.1365
349	2.5498	123	2.0427
362	2.3428	150	2.1596
364	2.3428	174	3.1365
416	2.5498	357	2.0199
		377	2.0199
		383	2.0199

8.2.2.2 The influential points in the design space (Leverage)

In classical linear models, large leverages can reveal the points in the design space at which the value of the outcome variable has a large impact on the regression fit [45]. That is, leverages can identify extreme values of independent variables. However, it is slightly different in the logistic regression model[31]. Leverages are the diagonal elements in the projection matrix. The projection matrix arises as a consequence of the Iteratively Reweighted Least Square (IRLS) (see Appendix B) and is a $J \times J$ matrix that can be expressed as

$$\boldsymbol{H} = \boldsymbol{V}^{\frac{1}{2}}\boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{V}^{\frac{1}{2}}. \tag{8.7}$$

The upper bound for each leverage is 1. However, if an observation is far away from other leverages, then it can be considered as an extreme leverage. Figures 8.9 and 8.10 show the plots of leverage versus predicted disease and exposure probabilities [13]. As observed in the previous diagnostic examples, the plots and leverages shown in Table 8.5 are different between the two models. There are a few points that are far away from the mean of the data (leverage=0.0915) in the model for the log odds of exposure. This implies that those points have considerable influence on the estimates.

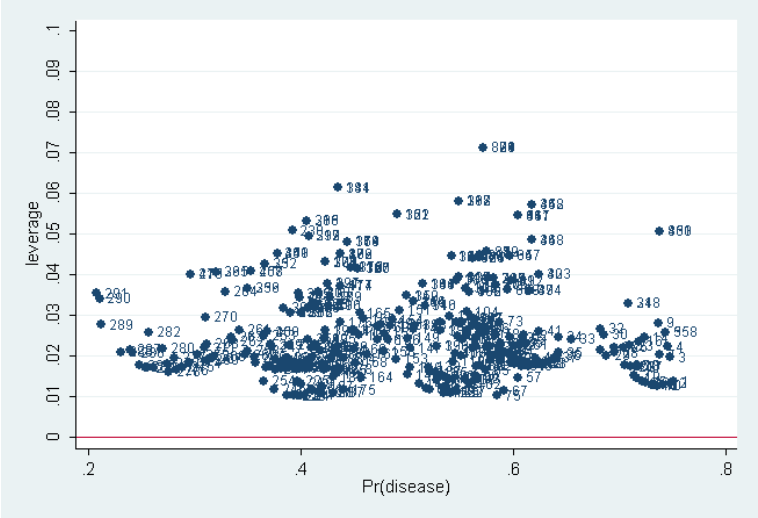

Figure 8.9: Plot of Leverages vs. Predicted Disease Probabilities from Fitting the Model for the Log Odds of Disease

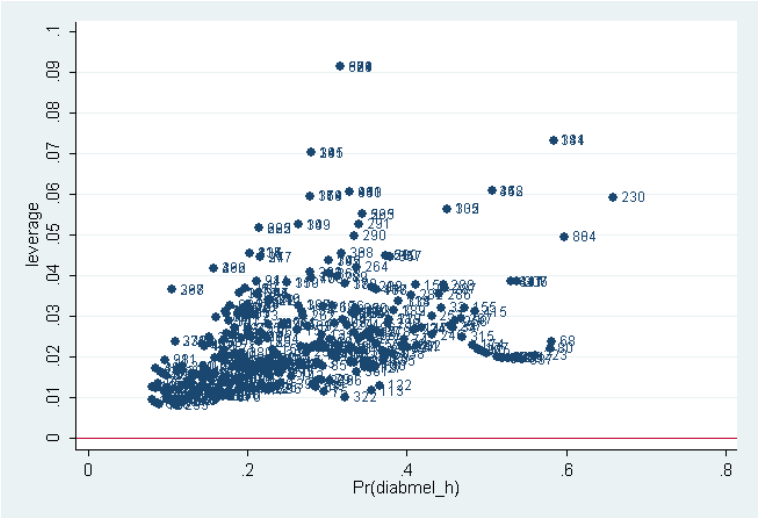

Figure 8.10: Plot of Leverages vs. Predicted Exposure Probabilities from Fitting the Model for the Log Odds of Exposure

Table 8.5: Extreme Leverages

The Model for the Log Odds of Disease		The Model for the Log Odds of Exposure	
Index	Leverage	Index	Leverage
10	0.0713	55	0.0703
229	0.0713	116	0.0703
236	0.0713	186	0.0703
267	0.0713	229	0.0915
291	0.0713	236	0.0915
307	0.0713	252	0.0915
342	0.0713	267	0.0915
370	0.0713	291	0.0915
		298	0.0733
		307	0.0915
		313	0.0915
		324	0.0733
		342	0.0915
		370	0.0915

8.2.2.3 Influential Diagnostics

Influential diagnostics are conducted by first removing one or more data points from the model and relevant statistics are used to determine how the absence of the observations changes the analysis. Appendix D outlined the theoretical backgrounds behind influential diagnostic statistics covered in this section.

(a) Coefficient Sensitivity Tests

Pregibon's $\Delta_l \hat{\boldsymbol{B}}^{\mathbf{1}}$ [45] statistic measures the impact (or the change) of a group of observations with the same observed independent variable values on the selected estimated coefficient(s). Figures 8.11 and 8.12 show the plots of $\Delta_l \hat{\boldsymbol{B}}^{\mathbf{1}}$ relating to exposure ("DF diabme") and disease ("DF disease") versus predicted disease and exposure probabilities.

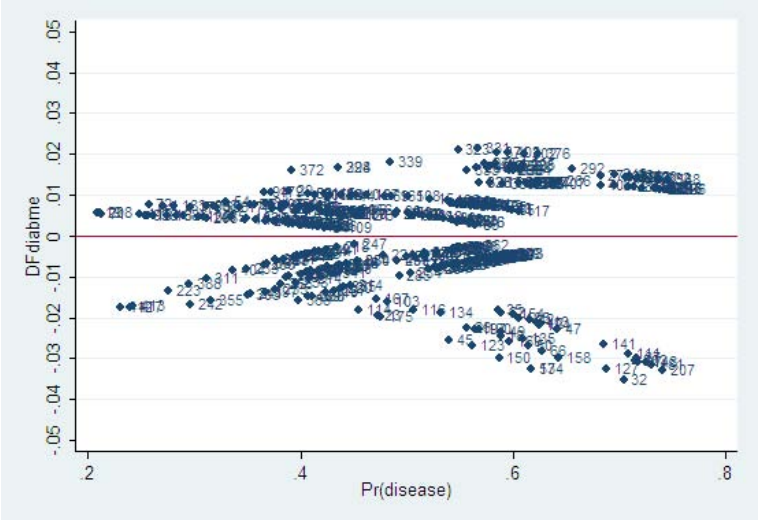

Figure 8.11: Plot of $\Delta_i \hat{\boldsymbol{B}}^{(1)}$ ("DF diabme") vs. Predicted Disease Probabilities from Fitting the Model for the Log Odds of Disease

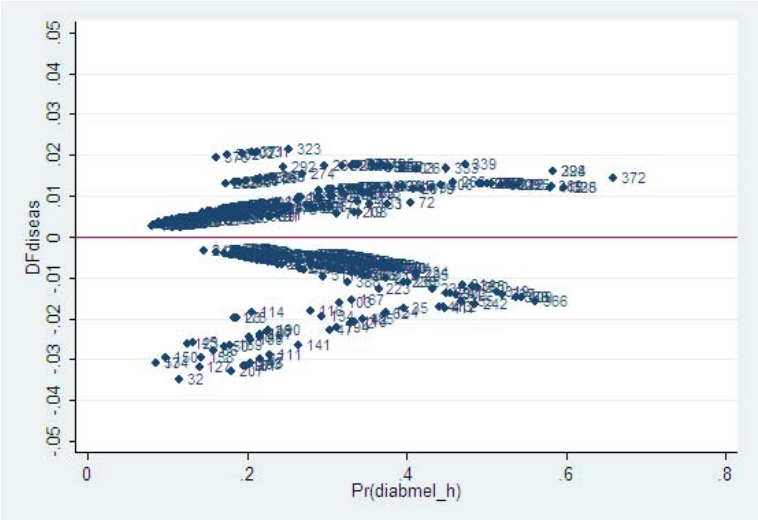

Figure 8.12: plot of $\Delta_i \hat{\boldsymbol{B}}^{(1)}$ ("DF disease") vs. Predicted Exposure Probabilities from Fitting the Model for the Log Odds of Exposure

Pregibon's c_l^1 statistic measures the overall change in fitted logits due to deleting a group of observations with the same observed independent variable values for all observations [45]. Figures 8.13 and 8.14 show the plots of Pregibon's c_l^1 versus predicted disease and exposure probabilities. Large values of c_l^1 (especially greater than 1 [38]) require investigation (None from both models).

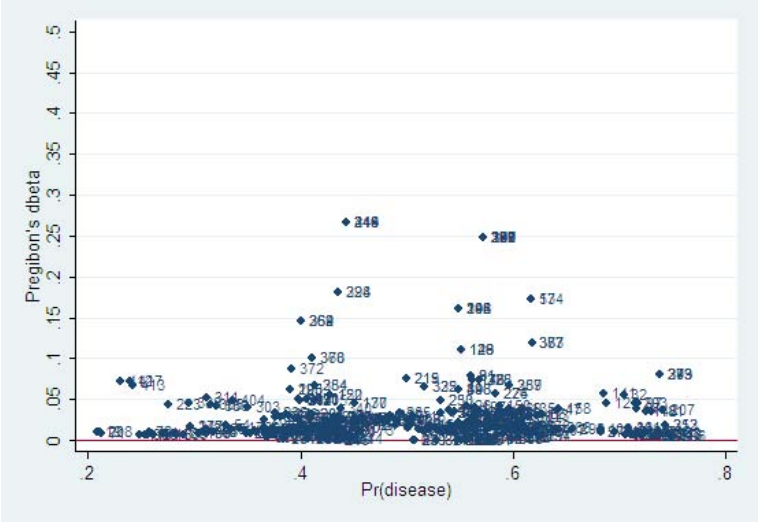

Figure 8.13: Plot of c_l^1 vs. Predicted Disease Probabilities from Fitting the Model for the Log Odds of Disease

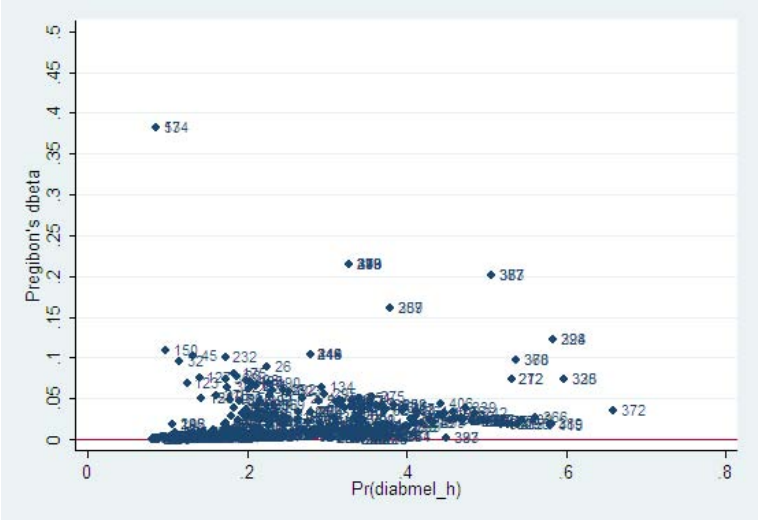

Figure 8.14: Plot of c_l^1 vs. Predicted Exposure Probabilities from Fitting the Model for the Log Odds of Exposure

(b) Goodness-of-Fit Sensitivity Tests

Pregibon's $\Delta_l\chi^2$ influence statistic (see Figures 8.15 and 8.16) and $\Delta_l D$ influence statistic (see Figures 8.17 and 8.18) [45][31] measure the impact of deleting l^{th} group of observations with the same observed independent variable values on χ^2 and Deviance. That is, these measures assess whether the group of observations is influential on the overall likelihood function. Tables 8.6 and 8.7 show measures that are greater than 4 [38] .

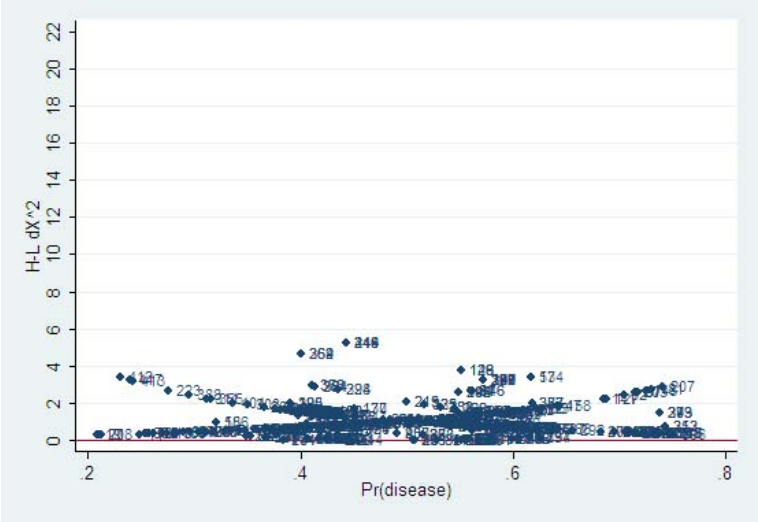

Figure 8.15: Plot of $\Delta\chi^2$ vs. Predicted Disease Probabilities from Fitting the Model for the Log Odds of Disease

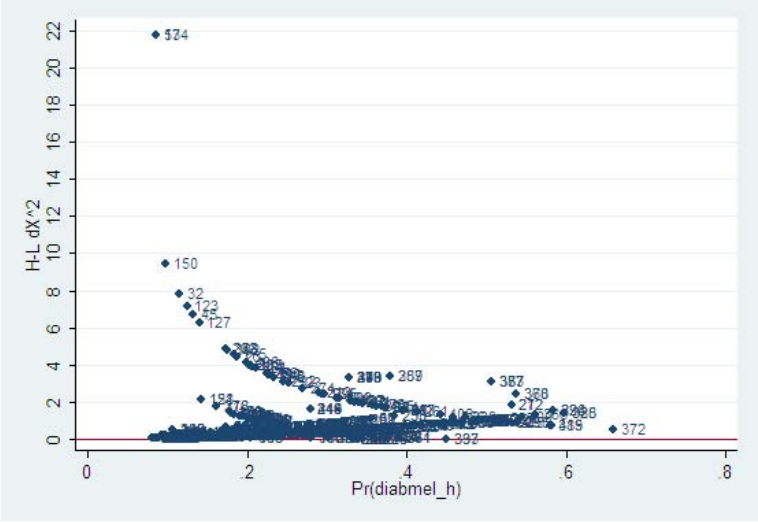

Figure 8.16: Plot of $\Delta\chi^2$ vs. Predicted Exposure Probabilities from Fitting the Model for the Log Odds of Exposure

Table 8.6: Extreme $\Delta\chi^2$

The Model for the Log Odds of Disease		The Model for the Log Odds of Exposure	
Index	$\Delta\chi^2$	Index	$\Delta\chi^2$
244	5.2693	23	4.4354
245	5.2693	32	7.8265
259	4.6306	45	6.6988
349	5.2693	53	21.7716
362	4.6306	123	7.1224
364	4.6306	127	6.2365
416	5.2693	150	9.4051
		174	21.7716
		175	4.5467
		203	4.0042
		232	4.9024
		302	4.7807
		368	4.8451
		408	4.1179

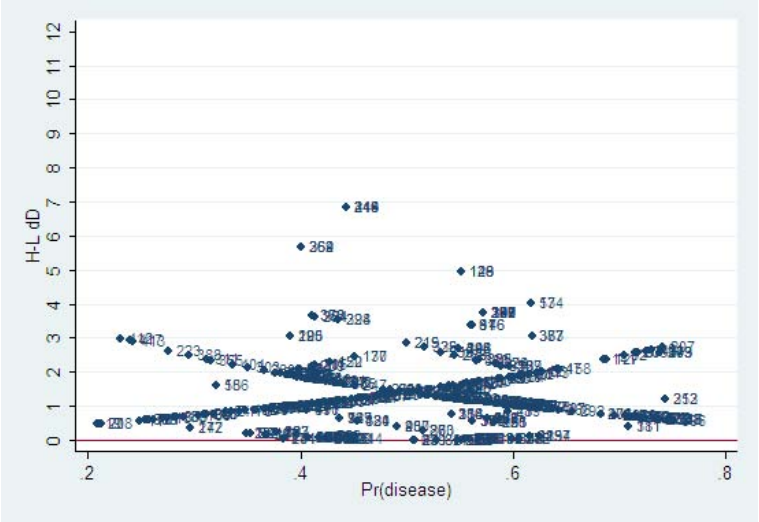

Figure 8.17: Plot of ΔD^2 vs. Predicted Disease Probabilities from Fitting the Model for the Log Odds of Disease

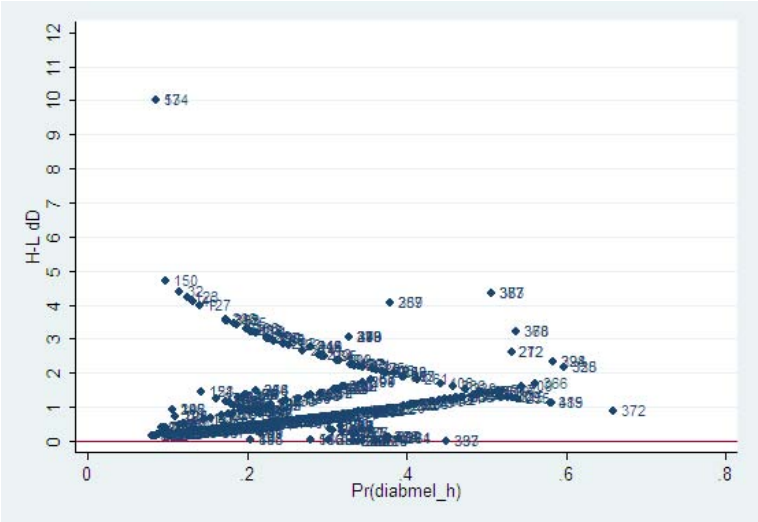

Figure 8.18: Plot of ΔD^2 vs. Predicted Exposure Probabilities from Fitting the Model for the Log Odds of Exposure

Table 8.7: Extreme ΔD^2

The Model for the Log Odds of Disease		The Model for the Log Odds of Exposure	
Index	ΔD^2	Index	ΔD^2
53	4.0400	32	4.3870
126	4.9557	45	4.1177
128	4.9557	53	10.0104
149	4.9557	123	4.2126
174	4.0400	150	4.7180
244	6.8302	174	10.0104
245	6.8302	287	4.0614
259	5.6620	357	4.3445
349	6.8302	359	4.0614
362	5.6620	377	4.3445
364	5.6620	383	4.3445
416	6.8302		

Chapter 9

Conclusions and Future Work

In this thesis two logistic regression models for case-control studies, the model for the log odds of disease and the model for the log odds of exposure have been compared and contrasted on many levels. In the simplest case, when the regression models were exact mathematical representations of the classical stratified analysis, all regression coefficients and standard errors corresponding to disease and exposure were identical. Breslow and Powers [10] coined the phrase"the covariate effects are saturated with parameters" to characterize the relationship between fully stratified analysis and logistic regression models. When the regression models omitted parameters, so that the covariate effects were no longer saturated, the regression coefficients differed between the two types of models. Two specific examples of non-equivalence were the omission of the joint confounding parameter and the case of a measured potential confounder. When the condition of parameter saturation was not satisfied, Breslow and Powers [10] argued that "as one adds terms to describe more fully the effects of the covariates", the two estimated coefficients or standard errors relating to exposure and disease converge toward the same value until covariate effects are fully adjusted.

The aforementioned examples of equivalence and non-equivalence were proved with the use of design matrices. In doing so, a theoretical definition of saturation was put forth, namely, "the covariate effects are saturated with parameters" when the number of independent parameters [relating to covariates in the model] equals the number of covariate patterns. Equivalence was also examined through the use of likelihood equations and profile likelihood equations. It was shown that in the case of equivalence, the

score functions arising from both logistic regression models were the same, and that the multinomial Poisson log likelihood could be written as a product of two binomial log likelihood functions. Subsequently, Prentice and Pyke's theoretical justification [47] for using disease as the outcome in a case-control study was explored, by relating a multinomial logistic regression model to the model for the log odds of disease. Finally, it was shown that not only were the coefficients and standard errors different between the two models in the case of non-equivalence, but so were the tests of fits and regression diagnostics.

It's uncommon in practice to build a logistic regression model that will satisfy the saturation condition. Therefore, it's expected that the coefficients from the two logistic regression models will yield different numerical results. However, as the numerical examples in this thesis have shown, the differences are small but they are real. Therefore, it's recommended that both models be fit to the same case-control data and the results compared. At the present time, it's unknown which model yields more efficient parameter estimates. There is room in future research to delineate when to use each model in terms of parameter efficiency. In this thesis, the simplest case of a binary disease and a binary exposure variable was considered. It would be of interest in future research to explore the extent of non-equivalence in the case of ordinal and measured variables.

Bibliography

[1] J. Aitchison and S. D. Silvey. Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics*, 29:813–828, 1958.

[2] A. C. Aitken. On least squares and linear combinations of observations. *Proceedings of the Royal Society of Edinburgh*, 55:42–48, 1935.

[3] J. A. Anderson. Separate sample logistic discrimination. *Biometrika*, 59:19–35, 1972.

[4] S. G. Baker. The multinomial-poisson transformation. *Statistician*, 43:495–504, 1994.

[5] R. J. Beckman and H. J. Trussell. The distribution of an arbitrary studentized residual and the effects of updating in multiple regression. *Journal of the American Statistical Association*, 69:199–201, 1974.

[6] W. Bennett, D. Lombardi, A. Eisenhart, J. Acosta, D. Cerbone, L. McCoy, and D. Yens. Risk factors for acute myocardial infarction in our patient population: A retrospective pilot study. *New York Medical Journal*, 3, 2008.

[7] N. E. Breslow. Regression analysis of the log odds ratio: A method for retrospective studies. *Biometrics*, 32:409–416, 1976.

[8] N. E. Breslow. Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association*, 91:14–28, 1996.

[9] N. E. Breslow and N. E. Day. *Statistical Methods In Cancer Research: The Analysis of Case-control Studies*. International Agency for Research on Cancer, 1980.

[10] N. E. Breslow and W. Powers. Are there two logistic regressions for retrospective studies? *Biometrics*, 34:100–105, 1978.

[11] A. C. Broders. Squamous-cell epithelioma of the lip. *The Journal Of the American Medical Association*, 74:656–664, 1920.

[12] W. G. Cochran. Some methods for strengthenin the common χ^2 tests. *Biometrics*, 10:417–451, 1954.

[13] D. Collett. *Modeling Binary Data (Second Edition)*. Chapman & Hall/crc, 2003.

[14] R. D. Cook. Detection of influential observation in linear regression. *Technometrics*, 19:15–18, 1977.

[15] J. Cornfield. A method of estimating comparative rates from clinical data. applications to cancer of the lung, breast and cervix. *Journal of the National Cancer Institute*, 11:1269–1275, 1951.

[16] J. Cornfield. *A statistical Problem Arising from Retrospective Studies (in Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability)*. University of California Press, 1956.

[17] J. Cornfield, T. Gordon, and W. W. Smith. Quantal response curves for experimentally uncontrolled variables. *Bulletin of the International Statistical Institute*, 38:97–115, 1961.

[18] D. R. Cox. Some procedures associated with the logistic qualitative response curve. In *In Research Papers in Statistics: Festschrift for J. Neyman*, pages 55–71, 1966.

[19] D. R. Cox and E. J. Snell. *Analysis of Binary Data*. Chapman and Hall/CRC; 2 edition, 1989.

[20] N. E. Day and D. F. kerridge. A general maximum likelihood discrimination. *Biometrics*, 23:313–323, 1967.

[21] M. Eliasziw, R. N. Rankin, A. J. Fox, R. B. Haynes, and H. J. Barnett. Accuracy and prognostic consequences of ultrasonography in identifying severe carotid artery stenosis. North American Symptomatic Carotid Endarterectomy Trial (NASCET) Group. *Stroke*, 10:1747–1752, 1995.

[22] M. Eliasziw, R. F. Smith, N. Singh, D. W. Holdsworth, A. J. Fox, and H. J. Barnett. Further comments on the measurement of carotid stenosis from angiograms. North American Symptomatic Carotid Endarterectomy Trial (NASCET) Group. *Stroke*, 12:2445–2449, 1994.

[23] M. Eliasziw, J. D. Spence, and H. J. Barnett. Carotid endarterectomy does not affect long-term blood pressure: observations from the NASCET. North American Symptomatic Carotid Endarterectomy Trial. *Cerebrovascular Diseases*, 8:20–24, 1998.

[24] M. Eliasziw, J. Y. Streifler, J. D. Spence, A. J. Fox, V. C. Hachinski, and H. J. Barnett. Prognosis for patients following a transient ischemic attack with and without a cerebral infarction on brain CT. North American Symptomatic Carotid Endarterectomy Trial (NASCET) Group. *Neurology*, 45:428–431, 1995.

[25] F. A. Graybill. *Introduction to matrices with applications in statistics*. Wadsworth Pub. Co., 1969.

[26] P. Gustafson, N. D. Le, and M. Vallee. A bayesian approach to case-control studies with errors in covariables. *Biostatistics*, 3:229–243, 2002.

[27] W. A. Guy. On the causes which determine the choice of an employment: being

an addition to the essays on the influence of employments upon health. *The Royal Statistical Society*, 8:351–353, 1843.

[28] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[29] C. Heidemanna, K. Hoffmanna, K. Klipstein-Grobuscha, C. Weikert, T. Pischona, H.-W. Hensed, and H. Boeinga. Potentially modifiable classic risk factors and their impact on incident myocardial infarction: results from the epicpotsdam study. *European Journal of Cardiovascular Prevention and Rehabilitation*, pages 65–71, 2007.

[30] R. V. Hogg, J. W. Mckean, and A. T. Craig. *Introduction to Mathematical Statistics*. Prentice Hall; 6 edition, 2005.

[31] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression (Second Edition)*. John Wiley & Sons, 2000.

[32] A. I. Khuri. *Advanced Calculus with Applications in Statistics*. John Wiley & Sons, 2003.

[33] D. A. Kikuchi. The inverse of a partitioned positive-definite symmetric matrix. *Communications in Statistics - Theory and Methods*, 12:1889–1900, 1983.

[34] A. M. Lilienfeld and D. E. Lilienfeld. A century of case-control studies: Progress? *The Journal of Chronic Disease*, 32:5–13, 1979.

[35] P. C. A. Louis. *Essay on Clinical Instruction, Translated by Peter Martin*. London: S. Highley, 1834.

[36] N. Mantel and W. Haenszel. Statistical aspects of the analysis of data from

retrospective studies of disease. *Journal of the National Cancer Institute*, 22:719–748, 1959.

[37] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, 1989.

[38] S. W. Menard. *Applied Logistic Regression Analysis*. Sage Publications, Inc., 2002.

[39] O. S. Miettinen. Confunding and effect-modification. *American Journal of Epidemiology*, 100:350–353, 1974.

[40] O. S. Miettinen. Etiologic research. *Scandinavian Journal of Work, Environment & Health*, 25:484–490, 1999.

[41] O. S. Miettinen. Epidemiology: Quo vadis? *European Journal of Epidemiology*, 19:713–718, 2004.

[42] S. A. Murphy and A. W. Van Der Vaart. On profile likelihood. *Journal of the American Statistical Association*, 95:449–465, 2000.

[43] W. M. Patefield. On the maximized likelihood function. *The Indian Journal of Statistics*, 39:92–96, 1977.

[44] R. Porter. *The Cambridge Illustrated History of Medicine*. Cambridge University Press, 1996.

[45] D. Pregibon. Logistic regression diagnostics. *The Annals of Statistics*, 9:705–724, 1981.

[46] R. L. Prentice. Use of the logistic model in retrospective studies. *Biometrics*, 32:599–606, 1976.

[47] R. L. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66:403–411, 1979.

[48] K. Roeder, R. J. Carroll, and B. G. Lindsay. A semiparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association*, 91:722–732, 1996.

[49] K. J. Rothman, S. Greenland, and T. L. Lash. *Modern Epidemiology*. Lippincott Williams and Wilkins; 3 edition, 2008.

[50] S. R. Seaman and S. Richardson. Equivalence of prospective and retrospective models in the bayesian analysis of case-control studies. *Biometrika*, 91:15–25, 2004.

[51] S. R. Searle. *Linear Models*. John Wiley & Sons, 1971.

[52] J. Shao. *mathematical Statistics*. Springer, 2003.

[53] A. Tavani, M. Bertuzzi, S. Gallus, E. Negri, and C. La Vecchia. Diabetes mellitus as a contributor to the risk of acute myocardial infarction. *Journal of Clinical Epidemiology*, 55:1082–1087, 2002.

[54] E. Vittinghoff, S. C. Shiboski, D. V. Glidden, and C. E. McCulloch. *Regression Methods in Biostatistics: Linear, logistic, Survival, and Repeated Measures Models*. Springer, 2007.

[55] B. Woolf. On estimating the relationship between blood group and disease. *Annals of Human Genetics*, 19:251–253, 1955.

[56] G. A. Young and R. L. Smith. *Essentials of statistical inference*. Cambridge University Press, 2005.

Appendix A

Cornfield's Formula

Cornfield [15] showed that the incidence odds ratio can approximate the incidence proportion ratio (the relative risk) in the cumulative case-control studies provided the disease incidence is low ($\frac{a+b}{N} \approx 0$ in the denominator) and the study is conducted in a closed population. First

$$\begin{aligned}
\frac{a}{N_E} =& \frac{\frac{a}{N}}{\frac{a}{N} + \frac{N_E - a}{N}} \\
=& \frac{\frac{a}{a+b}\frac{a+b}{N}}{\frac{a}{a+b}\frac{a+b}{N} + \frac{N_E - a}{N_{\bar{D}}}\frac{N_{\bar{D}}}{N}} = \frac{\frac{a}{a+b}\frac{a+b}{N}}{\frac{a}{a+b}\frac{a+b}{N} + \frac{N_E - a}{N_{\bar{D}}}(1 - \frac{a+b}{N})} \\
=& \frac{\frac{a}{a+b}\frac{a+b}{N}}{\frac{N_E - a}{N_{\bar{D}}} + (\frac{a}{a+b} - \frac{N_E - a}{N_{\bar{D}}})\frac{a+b}{N}} \approx \frac{\frac{a}{a+b}\frac{a+b}{N}}{\frac{N_E - a}{N_{\bar{D}}}}
\end{aligned} \tag{A.1}$$

where N is the total number of population and $N_{\bar{D}}$ is the number of noncases in the population. Similarly,

$$\begin{aligned}
\frac{b}{N_{\bar{E}}} =& \frac{\frac{b}{N}}{\frac{b}{N} + \frac{N_{\bar{E}} - b}{N}} \\
=& \frac{\frac{b}{a+b}\frac{a+b}{N}}{\frac{b}{a+b}\frac{a+b}{N} + \frac{N_{\bar{E}} - b}{N_{\bar{D}}}\frac{N_{\bar{D}}}{N}} = \frac{\frac{b}{a+b}\frac{a+b}{N}}{\frac{b}{a+b}\frac{a+b}{N} + \frac{N_{\bar{E}} - b}{N_{\bar{D}}}(1 - \frac{a+b}{N})} \\
=& \frac{\frac{b}{a+b}\frac{a+b}{N}}{\frac{N_{\bar{E}} - b}{N_{\bar{D}}} + (\frac{b}{a+b} - \frac{N_{\bar{E}} - b}{N_{\bar{D}}})\frac{a+b}{N}} \approx \frac{\frac{b}{a+b}\frac{a+b}{N}}{\frac{N_{\bar{E}} - b}{N_{\bar{D}}}}.
\end{aligned} \tag{A.2}$$

Finally,

$$\frac{\frac{a}{N_E}}{\frac{b}{N_{\bar{E}}}} \approx \frac{\frac{\frac{a}{a+b}\frac{a+b}{N}}{\frac{N_E - a}{N_{\bar{D}}}}}{\frac{\frac{b}{a+b}\frac{a+b}{N}}{\frac{N_{\bar{E}} - b}{N_{\bar{D}}}}} = \frac{\frac{a}{N_E - a}}{\frac{b}{N_{\bar{E}} - b}} = \frac{ad}{bc}. \tag{A.3}$$

Appendix B

Iteratively Reweighted Least Square (IRLS)

Suppose that $y_1, ..., y_J$ are observed values of independent variables $Y_1, ...Y_J$ each having a binomial distribution $B_j(m_j, p_j)$. Then the log likelihood function can be written as

$$\begin{aligned} l(p; y) &\propto \sum_{j=1}^{J} \log \; p_j^{y_j} \; 1 - p_j \; ^{m_j - y_j}\big] \\ &= \sum_{j=1}^{J} y_j \log \frac{p_j}{1-p_j} + m_j \log(1-p_j)\big]. \end{aligned} \tag{B.1}$$

The relationship between the independent variables and the logit takes the form

$$g(p_j) = \eta_j = \sum_{r} X_{jr} B_r. \tag{B.2}$$

The derivative of the log likelihood function (B.1) is

$$\frac{\partial l}{\partial p_j} = \frac{y_j - m_j p_j}{p_j(1-p_j)}. \tag{B.3}$$

By the chain rule [32], the derivative with respect to B_r is

$$\begin{aligned} \frac{\partial l}{\partial B_r} &= \sum_{j=1}^{J} \frac{y_j - m_j p_j}{p_j(1-p_j)} \frac{\partial p_j}{\partial B_r} \\ &= \sum_{j=1}^{J} \frac{y_j - m_j p_j}{p_j(1-p_j)} \frac{\partial p_j}{\partial \eta_j} \frac{\partial \eta_j}{\partial B_r} \\ &= \sum_{j=1}^{J} \frac{y_j - m_j p_j}{p_j(1-p_j)} \frac{\partial p_j}{\partial \eta_j} X_{jr} \end{aligned} \tag{B.4}$$

which can be written in matrix notation as $\boldsymbol{u(B)}$. As the likelihood estimates $\hat{\boldsymbol{B}}$ must satisfy $\boldsymbol{u}(\hat{\boldsymbol{B}}) = 0$, by a first-order Taylor series expansion [32]

$$\boldsymbol{u}(\hat{\boldsymbol{B}}) = \boldsymbol{u}(\boldsymbol{B^*}) + \boldsymbol{H}(\boldsymbol{B^*})(\hat{\boldsymbol{B}} - \boldsymbol{B^*}) \tag{B.5}$$

where B^* is near to $\hat{B}$, so that

$$\hat{\boldsymbol{B}} = \boldsymbol{B^*} - \boldsymbol{H}(\boldsymbol{B^*})^{-1}\boldsymbol{u}(\boldsymbol{B^*}). \tag{B.6}$$

Through a iterative process, the estimates of $\hat{\boldsymbol{B}}$ are

$$\hat{\boldsymbol{B}}^{\boldsymbol{t+1}} = \hat{\boldsymbol{B}}^{\boldsymbol{t}} - \boldsymbol{H}(\hat{\boldsymbol{B}}^{\boldsymbol{t}})^{-1}\boldsymbol{u}(\hat{\boldsymbol{B}}^{\boldsymbol{t}}). \tag{B.7}$$

The value of $\boldsymbol{H}(\boldsymbol{B})$ is usually replaced by the expected value of $\boldsymbol{H}(\boldsymbol{B})$ in the above equation [13] so that

$$\begin{aligned} E\ (\frac{\partial l}{\partial B_r})(\frac{\partial l}{\partial B_k})] &= E\ (\sum_{j=1}^{J} \frac{y_j - m_j p_j}{p_j(1-p_j)}\frac{\partial p_j}{\partial \eta_j}X_{jr})(\sum_{j'=1}^{J} \frac{y_{j'} - m_{j'} p_{j'}}{p_{j'}(1-p_{j'})}\frac{\partial p_{j'}}{\partial \eta_{j'}}X_{j'k})] \\ &= \sum_{j=1}^{J} \frac{m_j}{p_j(1-p_j)}(\frac{\partial p_j}{\partial \eta_j})^2 X_{jr}X_{jk} \\ &= [\boldsymbol{X^T V X}]_{\boldsymbol{rk}} \end{aligned} \tag{B.8}$$

because

$$\begin{aligned} &j = j'\text{:}\ \ E[(y_j - m_j p_j)(y_{j'} - m_{j'} p_{j'})] = cov(y_j, y_{j'}) = 0, \\ &j = j'\text{:}\ \ E[(y_j - m_j p_j)(y_{j'} - m_{i'} p_{j'})] = var(y_j) = m_j p_j (1-p_j). \end{aligned} \tag{B.9}$$

Equation (B.8) implies $-E[\boldsymbol{H}(\hat{\boldsymbol{B}})] = \boldsymbol{X^T V X}$ and for equation (B.4)

$$\begin{aligned} \frac{\partial l}{\partial B_r} &= \sum_{j=1}^{J} \frac{y_j - m_j p_j}{p_j(1-p_j)}\frac{\partial p_j}{\partial \eta_j}X_{jr} \\ &= \sum_{j=1}^{J} \frac{m_j}{p_j(1-p_j)}(\frac{\partial p_j}{\partial \eta_j})^2 \frac{y_j - m_j p_j}{m_j}(\frac{\partial p_j}{\partial \eta_j})^{-1} X_{jr} \\ &= [\boldsymbol{X^T V q}]_{\boldsymbol{r}} \end{aligned} \tag{B.10}$$

which means $\boldsymbol{u}(\boldsymbol{B}) = \boldsymbol{X^T V q}$. Finally,

$$\begin{aligned}
\hat{\boldsymbol{B}}^{\boldsymbol{t+1}} &= \hat{\boldsymbol{B}}^{\boldsymbol{t}} - \boldsymbol{H}(\hat{\boldsymbol{B}}^{\boldsymbol{t}})^{-1}\boldsymbol{u}(\hat{\boldsymbol{B}}^{\boldsymbol{t}}) \\
&= \hat{\boldsymbol{B}}^{\boldsymbol{t}} + (\boldsymbol{X^T V^t X})^{-1}\boldsymbol{X^T V^t q^t} \\
&= (\boldsymbol{X^T V^t X})^{-1}\boldsymbol{X^T V^t}(\boldsymbol{X}\hat{\boldsymbol{B}}^{\boldsymbol{t}} + \boldsymbol{q^t}) \\
&= (\boldsymbol{X^T V^t X})^{-1}\boldsymbol{X^T V^t z^t}. \qquad \text{(B.11)}
\end{aligned}$$

Appendix C

Regression Diagnostics for Classical Linear Models

Suppose the classical linear model can be expressed as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{B}, \tag{C.1}$$

so that the log likelihood function is

$$l = \sum_{i=1}^{n} l(\boldsymbol{X_i}\boldsymbol{B}; \boldsymbol{y_i}). \tag{C.2}$$

Two important quantities for regression diagnostics are the residuals $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{B}}$ and the projection matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T}$. In order to assess the impact of each individual on the aspects of the fit, two approaches are described below.

C.1 Assessment by Deletion

For a single deletion (e.g. l^{th} observation), the change of the estimates is given by [14]

$$\begin{aligned} \Delta_l \hat{\boldsymbol{B}} &= \hat{\boldsymbol{B}}(\mathbf{1}) - \hat{\boldsymbol{B}}(\mathbf{0}) \\ &= \frac{(\boldsymbol{X^T X})^{-1}\boldsymbol{X_l}(\boldsymbol{y_l} - \boldsymbol{X_l}\hat{\boldsymbol{B}})}{1 - \boldsymbol{X_l}(\boldsymbol{X^T X})^{-1}\boldsymbol{X_l}} \\ &= \frac{(\boldsymbol{X^T X})^{-1}\boldsymbol{X_l}\boldsymbol{r_l}}{1 - h_{ll}} \end{aligned} \tag{C.3}$$

where $\hat{\boldsymbol{B}}(\mathbf{1})$ are the estimates with the l^{th} observation and $\hat{\boldsymbol{B}}(\mathbf{0})$ are the estimates without the l^{th} observation. h_{ll} is the l^{th} diagonal element in the projection matrix $\boldsymbol{H}$. A measure that summarizes the change of all coefficients is given by [45]

$$\begin{aligned} c_l &= \boldsymbol{\Delta_l}\hat{\boldsymbol{B}}^{\boldsymbol{T}}\boldsymbol{X^T X}\boldsymbol{\Delta_l}\hat{\boldsymbol{B}} \\ &= \frac{\boldsymbol{r_l}^2 h_{ll}}{(1 - h_{ll})^2}. \end{aligned} \tag{C.4}$$

Another one is the change of the residual sum of squares (RSS) from deleting the l^{th} observation [5][45].

$$\Delta_l RSS = RSS(1) - RSS(0) = \frac{\boldsymbol{r_l}^2}{1 - h_{ll}}. \tag{C.5}$$

C.2 Assessment by Perturbation

Model perturbation is another approach used to assess the impact of each observation on the fitted model. The log likelihood function for the classical regression model can be reexpressed as

$$l_w = \sum_{i=1}^{n} w_i l(\boldsymbol{X_i B}; \boldsymbol{y_i}) \tag{C.6}$$

and the infinitesimal perturbation approach is obtained by specifying

$$w_i = \begin{cases} w & \text{for } i = l,\ 0 \leq w \leq 1 \\ 1 & \text{otherwise.} \end{cases}$$

The estimates for $\boldsymbol{B(w)}$ are

$$\hat{\boldsymbol{B}}(\boldsymbol{w}) = (\boldsymbol{X^T W X})^{-1} \boldsymbol{X^T W y}. \tag{C.7}$$

Then the quantities in Section C.2 can be obtained in a similar fashion. For example, the quantity as in equation (C.3) would be [45]

$$\hat{\boldsymbol{B}}(\boldsymbol{1}) - \hat{\boldsymbol{B}}(\boldsymbol{w}) = \frac{(\boldsymbol{X^T X})^{-1} \boldsymbol{X_l}(1-w)\boldsymbol{r_l}}{1 - (1-w)h_{ll}} \tag{C.8}$$

which is equation (C.3) when $w = 0$ and can also be expressed as

$$\hat{\boldsymbol{B}}(\boldsymbol{w}) = \hat{\boldsymbol{B}}(\boldsymbol{1}) - \frac{(\boldsymbol{X^T X})^{-1} \boldsymbol{X_l}(1-w)\boldsymbol{r_l}}{1 - (1-w)h_{ll}} = \hat{\boldsymbol{B}} - \frac{(\boldsymbol{X^T X})^{-1} \boldsymbol{X_l}(1-w)\boldsymbol{r_l}}{1 - (1-w)h_{ll}}. \tag{C.9}$$

Now the change of the estimates through model perturbation is

$$\begin{aligned} \Delta_l \hat{\boldsymbol{B}}^{\boldsymbol{1}} &= \hat{\boldsymbol{B}}(\boldsymbol{w})|_{w=1} - \hat{\boldsymbol{B}}(\boldsymbol{w})|_{w=0} \\ &= \frac{(\boldsymbol{X^T X})^{-1} \boldsymbol{X_l r_l}}{1 - h_{ll}}. \end{aligned} \tag{C.10}$$

Appendix D

Model Perturbation and One-Step Estimates in the Logistic Regression Model

With the background information provided in Appendic B and Appendix C, the estimates for $B(w)$ in the logistic regression would be

$$\begin{aligned}\hat{\boldsymbol{B}}^{\boldsymbol{t+1}}(\boldsymbol{w}) &= \hat{\boldsymbol{B}}^{\boldsymbol{t}}(\boldsymbol{w}) + \left(\boldsymbol{X^T}(\boldsymbol{V^t})^{\frac{1}{2}}\boldsymbol{W}(\boldsymbol{V^t})^{\frac{1}{2}}\boldsymbol{X}\right)^{-1}\boldsymbol{X^T}(\boldsymbol{V^t})^{\frac{1}{2}}\boldsymbol{W}(\boldsymbol{V^t})^{\frac{1}{2}}\boldsymbol{q^t} \\ &= \left(\boldsymbol{X^T}(\boldsymbol{V^t})^{\frac{1}{2}}\boldsymbol{W}(\boldsymbol{V^t})^{\frac{1}{2}}\boldsymbol{X}\right)^{-1}\boldsymbol{X^T}(\boldsymbol{V^t})^{\frac{1}{2}}\boldsymbol{W}(\boldsymbol{V^t})^{\frac{1}{2}}\left(\boldsymbol{X}\hat{\boldsymbol{B}}^{\boldsymbol{t}}(\boldsymbol{w}) + \boldsymbol{q^t}\right) \\ &= \left(\boldsymbol{X^T}(\boldsymbol{V^t})^{\frac{1}{2}}\boldsymbol{W}(\boldsymbol{V^t})^{\frac{1}{2}}\boldsymbol{X}\right)^{-1}\boldsymbol{X^T}(\boldsymbol{V^t})^{\frac{1}{2}}\boldsymbol{W}(\boldsymbol{V^t})^{\frac{1}{2}}\boldsymbol{z^t}. \end{aligned} \tag{D.1}$$

When $w = 1$, equation (D.1) becomes equation (B.11).

Pregibon [45] proposed that terminating the iterative scheme after one step can help identify the individual effects with a minimal effort. That is

$$\begin{aligned}\hat{\boldsymbol{B}}_{\boldsymbol{1}}(\boldsymbol{w}) &= (\boldsymbol{X^T}\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{W}\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{X})^{-1}\boldsymbol{X^T}\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{W}\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{z} \\ &= (\boldsymbol{X^T}\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{W}\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{X})^{-1}\boldsymbol{X^T}\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{W}\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{q} \end{aligned} \tag{D.2}$$

where $\boldsymbol{z} = \boldsymbol{z_0}$ and $\boldsymbol{q} = \boldsymbol{q_0}$. Because $\hat{\boldsymbol{B}}_{\boldsymbol{0}} = \boldsymbol{0}$ so that $\boldsymbol{z_0} = \boldsymbol{q_0}$. Pregibon mentioned that "this equation is identical to the corresponding exact noniterative solution for the standard linear model with $\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{X}$ as the design matrix for the response variable $\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{z}$" [45] so that the formula in Appendix C can be directly applied to the one-step estimation.

D.1 Coefficient Sensitivity Tests

The quantity as in equation (C.9) would be [45]

$$\hat{\boldsymbol{B}}^{\boldsymbol{1}}(\boldsymbol{w}) = \hat{\boldsymbol{B}} - \frac{(\boldsymbol{X^T V X})^{-\mathbf{1}}\boldsymbol{X_l}(1-w)\boldsymbol{s_l}}{1-(1-w)h_{ll}}$$

where $\boldsymbol{s} = \boldsymbol{y} - \boldsymbol{m}\hat{\boldsymbol{p}}$. Notice that l represents a distinct groups with the same observed independent variable values ($l = 1, ..., J$).

The quantity as in equation (C.10) would be

$$\Delta_l \hat{\boldsymbol{B}}^{\boldsymbol{1}} = \frac{(\boldsymbol{X^T V X})^{-\mathbf{1}}\boldsymbol{X_l s_l}}{1-h_{ll}}. \tag{D.3}$$

Similarly, the corresponding quantity (C.4) from Deletion method can also be obtained by model perturbation

$$\begin{aligned} c_l^1 &= (\boldsymbol{\Delta_l}\hat{\boldsymbol{B}}^{\boldsymbol{1}})^{\boldsymbol{T}}\boldsymbol{X^T V X}\boldsymbol{\Delta_l}\hat{\boldsymbol{B}}^{\boldsymbol{1}} \\ &= \frac{\boldsymbol{s_l X_l^T}(\boldsymbol{X^T V X})^{-\mathbf{1}}\boldsymbol{X_l s_l}}{(1-h_{ll})^2} \\ &= \frac{\boldsymbol{s_l} v_{ll}^{-\frac{1}{2}} v_{ll}^{\frac{1}{2}}\boldsymbol{X_l^T}(\boldsymbol{X^T V X})^{-\mathbf{1}}\boldsymbol{X_l} v_{ll}^{\frac{1}{2}} v_{ll}^{-\frac{1}{2}}\boldsymbol{s_l}}{(1-h_{ll})^2} \\ &= \frac{\boldsymbol{s_l} v_{ll}^{-\frac{1}{2}} h_{ll} v_{ll}^{-\frac{1}{2}}\boldsymbol{s_l}}{(1-h_{ll})^2} \\ &= \frac{\boldsymbol{r_l}^2 h_{ll}}{(1-h_{ll})^2} \end{aligned} \tag{D.4}$$

as $\boldsymbol{r_l}^2 = \frac{\boldsymbol{s_l}^2}{v_{ll}}$ and v_{ll} is the l^{th} diagonal element in matrix $\boldsymbol{V}$.

D.2 Goodness-of-Fit Sensitivity Tests

- The change in D due to deleting the lth group with the same observed independent variables

Define

$$D_w(\boldsymbol{X}\hat{\boldsymbol{B}}(w);\boldsymbol{y}) = 2\sum_{j=1}^{J} w_j \;\; l(\tilde{\boldsymbol{p}}_{\boldsymbol{j}};\boldsymbol{y}_{\boldsymbol{j}}) - l(\hat{\boldsymbol{p}}_{\boldsymbol{j}};\boldsymbol{y}_{\boldsymbol{j}}) \tag{D.5}$$

where $\tilde{\boldsymbol{p}}_{\boldsymbol{j}} = \frac{\boldsymbol{y}_{\boldsymbol{j}}}{\boldsymbol{m}_{\boldsymbol{j}}}$. Taking a second-order Taylor series expansion [32] about $\hat{\boldsymbol{B}}$, we have

$$\begin{aligned}
&D_w(\boldsymbol{X}\hat{\boldsymbol{B}}^{1}(w);\boldsymbol{y}) \\
=&D_w(\boldsymbol{X}\hat{\boldsymbol{B}}^{1}(w);\boldsymbol{y})|_{\hat{\boldsymbol{B}}^{1}(w)=\hat{\boldsymbol{B}}} + \frac{\partial D_w(\boldsymbol{X}\hat{\boldsymbol{B}}^{1}(w);\boldsymbol{y})}{\partial\hat{\boldsymbol{B}}^{1}(w)}|_{\hat{\boldsymbol{B}}^{1}(w)=\hat{\boldsymbol{B}}} \;\; \hat{\boldsymbol{B}}^{1}(w) - \hat{\boldsymbol{B}}\big] \\
&\quad + \;\; \hat{\boldsymbol{B}}^{1}(w) - \hat{\boldsymbol{B}}\big]^{T} \frac{\partial^2 D_w(\boldsymbol{X}\hat{\boldsymbol{B}}^{1}(w);\boldsymbol{y})}{2\partial(\hat{\boldsymbol{B}}^{1}(w))^2}|_{\hat{\boldsymbol{B}}^{1}(w)=\hat{\boldsymbol{B}}} \;\; \hat{\boldsymbol{B}}^{1}(w) - \hat{\boldsymbol{B}}\big] \\
=&D_w(\boldsymbol{X}\hat{\boldsymbol{B}};\boldsymbol{y}) + \;\; \hat{\boldsymbol{B}}^{1}(w) - \hat{\boldsymbol{B}}\big]^{T} \frac{\partial^2 D_w(\boldsymbol{X}\hat{\boldsymbol{B}}^{1}(w);\boldsymbol{y})}{2\partial(\hat{\boldsymbol{B}}^{1}(w))^2}|_{\hat{\boldsymbol{B}}^{1}(w)=\hat{\boldsymbol{B}}} \;\; \hat{\boldsymbol{B}}^{1}(w) - \hat{\boldsymbol{B}}\big] \\
=& \;\; D(\boldsymbol{X}\hat{\boldsymbol{B}};\boldsymbol{y}) - (1-w)\boldsymbol{d}_{\boldsymbol{l}}{}^2 \;\; - \;\; [\hat{\boldsymbol{B}}^{1}(w) - \hat{\boldsymbol{B}}]^{T}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{V}\boldsymbol{X} \;\; 1-(1-w)h_{ll}\big][\hat{\boldsymbol{B}}^{1}(w) - \hat{\boldsymbol{B}}] \\
=& \;\; D(\boldsymbol{X}\hat{\boldsymbol{B}};\boldsymbol{y}) - (1-w)\boldsymbol{d}_{\boldsymbol{l}}{}^2 \\
-& \;\; \frac{(1-w)\boldsymbol{s}_{\boldsymbol{l}}\boldsymbol{X}_{\boldsymbol{l}}^{\boldsymbol{T}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{V}\boldsymbol{X})^{-\mathbf{1}}}{1-(1-w)h_{ll}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{V}\boldsymbol{X} \;\; 1-(1-w)h_{ll}\big]\frac{(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{V}\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}_{\boldsymbol{l}}(1-w)\boldsymbol{s}_{\boldsymbol{l}}}{1-(1-w)h_{ll}} \\
=& \;\; D(\boldsymbol{X}\hat{\boldsymbol{B}};\boldsymbol{y}) - (1-w)\boldsymbol{d}_{\boldsymbol{l}}{}^2 \;\; - \frac{\boldsymbol{r}_{\boldsymbol{l}}{}^2(1-w)^2h_{ll}}{1-(1-w)h_{ll}}.
\end{aligned} \tag{D.6}$$

where $\boldsymbol{d_l}$ is the l^{th} element in the deviance residual vector. The above equation uses another important equation [5]

$$
\begin{aligned}
&(\boldsymbol{X(0)^T V X(0)})^{-1} \\
=&(\boldsymbol{X^T V X} - \boldsymbol{X_l} v_{ll} \boldsymbol{X_l^T})^{-1} \\
=&(\boldsymbol{X^T V X})^{-1} + \frac{(\boldsymbol{X^T V X})^{-1} \boldsymbol{X_l} v_{ll} \boldsymbol{X_l^T} (\boldsymbol{X^T V X})^{-1}}{1 - (\boldsymbol{X_l} v_{ll}^{\frac{1}{2}})^{\boldsymbol{T}} (\boldsymbol{X^T V X})^{-1} \boldsymbol{X_l} v_{ll}^{\frac{1}{2}}} \\
=&(\boldsymbol{X^T V X})^{-1} + \frac{(\boldsymbol{X_l} v_{ll}^{\frac{1}{2}})^{-\boldsymbol{T}} (\boldsymbol{X_l} v_{ll}^{\frac{1}{2}})^{\boldsymbol{T}} (\boldsymbol{X^T V X})^{-1} \boldsymbol{X_l} v_{ll}^{\frac{1}{2}} (\boldsymbol{X_l} v_{ll}^{\frac{1}{2}})^{\boldsymbol{T}} (\boldsymbol{X^T V X})^{-1}}{1 - (\boldsymbol{X_l} v_{ll}^{\frac{1}{2}})^{\boldsymbol{T}} (\boldsymbol{X^T V X})^{-1} \boldsymbol{X_l} v_{ll}^{\frac{1}{2}}} \\
=&(\boldsymbol{X^T V X})^{-1} + \frac{(\boldsymbol{X_l} v_{ll}^{\frac{1}{2}})^{-\boldsymbol{T}} h_{ll} (\boldsymbol{X_l} v_{ll}^{\frac{1}{2}})^{\boldsymbol{T}} (\boldsymbol{X^T V X})^{-1}}{1 - h_{ll}} \\
=&(\boldsymbol{X^T V X})^{-1} + \frac{h_{ll} (\boldsymbol{X^T V X})^{-1}}{1 - h_{ll}} \\
=&\frac{(\boldsymbol{X^T V X})^{-1}}{1 - h_{ll}},
\end{aligned}
$$

so that

$$(\boldsymbol{X(0)^T V X(0)}) = (\boldsymbol{X^T V X})(1 - h_{ll}). \tag{D.7}$$

Then

$$
\begin{aligned}
\text{When } w = 0\text{: } & \frac{\partial^2 D_w(\boldsymbol{X}\hat{\boldsymbol{B}}^1(w); \boldsymbol{y})}{2\partial(\hat{\boldsymbol{B}}^1(w))^2}|_{\hat{\boldsymbol{B}}^1(w)=\hat{\boldsymbol{B}}} = (\boldsymbol{X^T V X})(1 - h_{ll}), \\
\text{When } w = 1\text{: } & \frac{\partial^2 D_w(\boldsymbol{X}\hat{\boldsymbol{B}}^1(w); \boldsymbol{y})}{2\partial(\hat{\boldsymbol{B}}^1(w))^2}|_{\hat{\boldsymbol{B}}^1(w)=\hat{\boldsymbol{B}}} = (\boldsymbol{X^T V X})^{-1}.
\end{aligned} \tag{D.8}
$$

This implies

$$\frac{\partial^2 D_w(\boldsymbol{X}\hat{\boldsymbol{B}}^1(w); \boldsymbol{y})}{2\partial(\hat{\boldsymbol{B}}^1(w))^2}|_{\hat{\boldsymbol{B}}^1(w)=\hat{\boldsymbol{B}}} = (\boldsymbol{X^T V X})[1 - (1 - w)h_{ll}]. \tag{D.9}$$

Finally with (D.6), we obtain

$$\Delta_l D \approx D_1(\boldsymbol{X}\hat{\boldsymbol{B}}^1(1); \boldsymbol{y}) - D_0(\boldsymbol{X}\hat{\boldsymbol{B}}^1(0); \boldsymbol{y}) = \boldsymbol{d_l}^2 + \frac{\boldsymbol{r_l}^2 h_{ll}}{1 - h_{ll}}. \tag{D.10}$$

- The change in χ^2 due to deleting the lth group with the same observed independent variables

  Beckman showed detailed information about how to obtain (C.5) [5]. This quantity is obtained in a similar fashion as (C.5).

$$\chi^2(1) = \chi^2(0) + \frac{\boldsymbol{r_l}^2}{1 - h_{ll}} \tag{D.11}$$

and

$$\Delta\chi^2 = \chi^2(1) - \chi^2(0) = \frac{\boldsymbol{r_l}^2}{1 - h_{ll}} = \chi_l^2. \tag{D.12}$$