Logistic Regression Under Independence of Genetic and Non-Genetic Covariates in a Case-Control Study

by

Ji-Hyung Shin

B.Sc., Simon Fraser University, 2002

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in the Department

of

Statistics and Actuarial Science

SIMON FRASER UNIVERSITY

August 2004

APPROVAL

Name: Ji-Hyung Shin

Degree: Master of Science

Title of project: Logistic Regression Under Independence of Genetic and Non-Genetic Covariates in a Case-Control Study

Examining Committee: Dr. Boxin Tang
Chair

Dr. Jinko Graham
Senior Supervisor
Simon Fraser University

Dr. Brad McNeney
Senior Supervisor
Simon Fraser University

Dr. John Spinelli
External Examiner
BC Cancer Agency and Simon Fraser University

Date Approved: _____

Abstract

In a case-control study of a rare disease such as type 1 diabetes, covariate information is often collected on a genetic factor and a continuous attribute such as age. In some instances, it is reasonable to assume that the attribute and genetic factor occur independently in the population. Under this independence assumption, we develop maximum likelihood estimators of parameters in a logistic model of disease risk. Estimates are based on data from both patients and controls and may be obtained by fitting a polychotomous regression model of joint disease and genetic status. Our results extend previous log-linear approaches to imposing independence between a genetic factor and a categorical attribute, thereby avoiding potential loss of information from discretizing a continuous attribute. We apply the method to investigate the effects of age and a variant of the glutamate-cysteine ligase catalytic subunit on type 1 diabetes. The results are compared to those obtained from a standard logistic regression analysis, which does not make use of the independence assumption.

Acknowledgements

I would like to thank my supervisors Dr. Jinko Graham and Dr. Brad McNeney for their continual guidance and support and for the suggestions and advice throughout this project. Also, I am very grateful to Jinko and Brad for their encouragement that helped me decide to carry on my education to this level. Jinko and Brad, I have learned so much from both of you, not only throughout this project but also through all these years I've known you.

I would also like to express my appreciation to Dr. John Spinelli for his suggestions and comments that provided me with new insights and ideas. And, I would like to give my thanks to all the professors and instructors, here at SFU, who have taught me courses over the years (both undergraduate and graduate). I am also grateful to Ms. Sylvia Holmes for her enormous assistance she provides graduate students.

And, many thanks to all of my fellow students and friends. Sandy, thank you for making me go for the Grouse Grind which made me feel great achievement when I finally got to the top of the mountain after 2 hours of grind (just for a reference, it took Eric only 45 minutes). Steve, thank for your diligence, kindness and patience that you've showed to everyone. I've learned so much from you. Chunfang, thank you for sharing your delicious Chinese food with me, and for being there for me whenever I needed. Wilson, Jason, L., Michael, Pritam, Crystal, Karey, Farouk, Matt, Mercedeh, Linnea, Simon, Wendell, Jeremy, Maria, Jason, N, Laura, Amy, Suman, Darcy, Eric, John and Paul, thank you all for making my school life pleasant and memorable one!!

I am also grateful to my best friend Julie for giving me ride (yes, both ways!!)

for three years. Thank you for helping me with everything, whether big or small. I know I can always count on you. I wish to give my heartful thanks to my family: my grandmother, my parents and my sister, for their endless support and prayers that make me go on. Thank you for putting up with my everyday insanity and insecurities!! Lastly, I would like to thank God for everything.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Before the development of molecular techniques that allow genetic typing, epidemiologic studies collected information on genetic factors indirectly. Now that molecular-genotyping techniques are available, there is increasing interest in genetic susceptibility to disease and the joint effect of genes and the environment or genes and the non-genetic attributes such as age and gender. For example, an environmental exposure may increase disease risk in a genetically susceptible subgroup but have little or no effect in the rest of the population. Alternatively, a genetic variant may increase disease risk in individuals with a specific set of attributes but have little or no effect in the rest of the population.

A case-control study design is often adopted to investigate associations for a rare disease. However, when a main aim of the study is to test the interaction of any two factors, a large sample size is required to obtain adequate statistical power (Smith and Day 1984). A popular tool for analyzing data from case-control studies and assessing the effect of interaction is logistic regression. For some studies, it is reasonable to assume that the genetic and non-genetic factors are independent in the population. However, standard logistic regression cannot accommodate this independence assumption. Incorporating the valid assumption of independence is expected to improve statistical efficiency and stability of estimation. Umbach and Weinberg (1997)

developed maximum likelihood estimators of disease risk based on a log-linear model that enforces independence. They achieved mildly enhanced precision for estimates of main effects and much enhanced precision for estimates of statistical interactions.

To illustrate the idea behind their approach and the approach that we will propose, let us consider a case-control study of type 1 diabetes. The data to be used throughout this thesis consists of information on age and a genetic variant of the glutamate-cysteine ligase catalytic subunit (GCLC) in 400 subjects aged 0-34 years from two overlapping Swedish case-control studies. Further details on the study design and subset selection may be found in Bekris et al. (2004).

Type 1 diabetes (T1D) has a prevalence of approximately 0.3% in Caucasians (Todd 1990). The disease is most common in children and young adults and the associated clinical symptoms (e.g., Gambelunghe et al. 2001), genetic factors (e.g.,Rotter and Rimoin 1978) and immune markers (e.g., Graham et al. 2002) vary with age at onset. T1D is an autoimmune disorder in which the insulin-producing pancreatic cells are destroyed. The concordance rate in monozygotic twins is estimated to be only 30-50% (Davies et al. 1994), indicating that both genetic and non-genetic factors jointly contribute to risk.

Glutamate-cysteine ligase (GCL) is the enzyme catalyzing the first rate-limiting step in the synthesis of glutathione. Glutathione is one of the major antioxidants produced by the cell and plays a role in the detoxification of many harmful compounds. A variant of the polymorphic trinucleotide repeat within the 5' untranslated region of the GCL catalytic subunit (GCLC) has been observed to be associated with drug sensitivity (Walsh et al. 2001). Since toxic chemical agents such as nitrosamines are known to trigger T1D (Dahlquist et al. 1990), we sought to characterize the association between this genetic variant, which we denote as GCLC8, and T1D and to assess whether or not the association is modified by age. The hypothesis of age-dependent associations between T1D and GCLC8 is equivalent to the hypothesis that T1D age-at-onset varies according to GCLC8 status when GCLC8 status is

independent of age in the general population. To investigate these hypotheses, we analyzed 186 controls and 179 cases who had complete data on GCLC and age. Throughout, GCLC8+ represents those with at least one copy of GCLC8 and GCLC8- represents those with no copies of GCLC8. Table 1.1 summarizes the data for the investigation.

Table 1.1: Contingency table without the assumption of GCLC8-age independence

	age in years									
	[0,7)		[7,14)		[14,21)		[21,28)		[28,35]	
	con[a]	cas[b]	con	cas	con	cas	con	cas	con	cas
GCLC8-	3	12	55	46	29	25	20	25	26	20
GCLC8+	3	11	23	22	10	8	8	5	9	5
OR[c]	0.92		1.14		0.93		0.50		0.72	
95% CI[d]	(.1–8.4)		(.6–2.3)		(.3–2.7)		(.1–1.8)		(.2–2.5)	

[a] "con", controls; [b] "cas", T1D patients; [c] "OR",odds-ratio;
[d] "CI", confidence interval.

The odds-ratios appear to decrease with age, but instability of stratified estimates reduces power to detect GCLC8-by-age interaction. The odds-ratio for the 0-7 year old age-group is particularly unstable, owing to there being very few controls of this age.

In the relatively stable Swedish population, genotype frequencies are expected to remain approximately constant over several generations. Moreover, in the 0-34 year age group, GCLC variation is not expected to influence mortality. Thus it seems reasonable to assume that GCLC8 is independent of age in the study population. In addition, for a rare disease such as T1D, controls can be thought of as representative of the general population. Therefore, independence between GCLC8 and age can be assumed in the controls, as well. Under the independence assumption, we may legitimately pool all the controls across ages, and assess the associations based on the following contingency table (Table 1.2). The confidence intervals are narrower than those from Table 1.1, particularly for the 0-7 year old age-group. Consequently,

Table 1.2: Contingency table under the independence assumption

	age in years					
	con	[0,7)	[7,14)	[14,21)	[21,28)	[28,35]
GCLC8-	133	12	46	25	25	20
GCLC8+	53	11	22	8	5	5
OR	1.0	2.30	1.20	0.80	0.50	0.63
95% CI	(—)	(1.0–5.5)	(0.7–2.2)	(0.3–1.9)	(0.2–1.4)	(0.2–1.8)

the overall impression of an age-dependent effect of GCLC8 is now stronger. This example illustrates intuitively how power to detect interactions can be increased by imposing independence.

In this project, we describe a method to estimate disease associations with a particular genetic variant and a *continuous* non-genetic attribute which are assumed to be independent in the population. The method extends previous log-linear approaches to imposing independence by allowing for a continuous attribute. Chapter 2 develops the statistical approach in which risk estimates are obtained by fitting a polychotomous logistic model of joint disease and genetic status. Estimators are developed by reparameterization of the case-control likelihood, using arguments similar to those of Prentice and Pyke (1979). In Chapter 3, the proposed method is applied to the T1D data, after verifying the independence assumption in the controls. Standard errors of parameter estimates are obtained by use of the nonparametric bootstrap. This enables us to compare the efficiency of the proposed approach to standard logistic regression. The last chapter comments on the approach and on limitations and possible extensions.

Chapter 2

Maximum Likelihood Estimation

Consider a case-control study of a rare disease in which n_0 controls and n_1 cases, for a total of $n = n_0 + n_1$ subjects, are selected randomly from their respective subpopulations. Let D denote the disease status such that,

$$D = \begin{cases} 1 & \text{if the individual is a diseased case} \\ 0 & \text{if the individual is a disease-free control;} \end{cases}$$

Z_g denote a binary genetic covariate such that,

$$Z_g = \begin{cases} 1 & \text{if the individual is in some genetically defined group} \\ 0 & \text{otherwise;} \end{cases}$$

$Z_a = (A, A^2)$ code for information on the continuous non-genetic attribute A and $Z = (Z_g, Z_a, Z_g \times Z_a)$, where $Z_g \times Z_a = (Z_g A, Z_g A^2)$. We assume that the continuous attribute A and the genetic covariate Z_g are independent in the general population. Since the disease of interest is rare, we can then assume that Z_g and A occur independently in controls as well.

Assume a logistic regression model for disease risk with parameters α and $\beta = (\beta_g, \beta_{1a}, \beta_{2a}, \beta_{1ga}, \beta_{2ga})^T = (\beta_g, \beta_a^T, \beta_{ga}^T)^T$, where $\beta_a^T = (\beta_{1a}, \beta_{2a})$ and $\beta_{ga}^T = (\beta_{1ga}, \beta_{2ga})$. This model gives the flexibility to handle quadratic main effects and interaction terms

involving the continuous attribute. It is also straightforward to generalize the method to include higher-order polynomial terms. We can write the log odds of disease or $logit(p)$, where $p = \Pr(D = 1 \mid Z)$, as a linear function of the parameters α and β:

$$\log\left[\frac{\Pr(D=1 \mid Z)}{\Pr(D=0 \mid Z)}\right] = logit(p) = \alpha + Z\beta,$$

which implies

$$\Pr(D=1 \mid Z) = \frac{\exp(\alpha + Z\beta)}{1+\exp(\alpha + Z\beta)} \text{ and } \Pr(D=0 \mid Z) = \frac{1}{1+\exp(\alpha + Z\beta)}. \quad (2.1)$$

equation (2.1) can be generalized as in Prentice and Pyke (1979) by defining $\alpha_0 = 0$, $\alpha_1 = \alpha$, $\beta_0 = (\beta_{g0}, \beta_{1a0}, \beta_{2a0}, \beta_{1ga0}, \beta_{2ga0})^T = (0,0,0,0,0)^T$ and $\beta_1 = \beta$, and rewriting

$$\Pr(D=i \mid Z) = \frac{\exp(\alpha_i + Z\beta_i)}{\sum_{l=0}^{1}\exp(\alpha_l + Z\beta_l)} \qquad i = 0{,}1 \quad (2.2)$$

In a logistic model, the covariate vector Z is a fixed quantity and the disease response is the random quantity. However, in a case-control study, subjects are sampled conditional on their disease status, and risk assessment is based on differences in the distributions of their covariate vectors. In other words, in a case-control study we draw samples of covariate vectors from two different populations: $P(Z \mid D = 0)$ for controls, and $P(Z \mid D = 1)$ for cases. Hence it is the covariate vector that should be regarded as the random outcome rather than the disease status. Thus an important question arises: how can one interpret the coefficients of a logistic regression model which has been fit to data from a case-control study? It turns out that inferences about the risk parameters β are the same regardless of whether the data were collected prospectively or retrospectively. Prentice and Pyke (1979) demonstrated this point by reparameterizing the case-control likelihood in such a way that the β's appear in a single term of logistic form which can be maximized without regard to the other nuisance parameters. However, as in Table 1.1 of the Introduction, the resulting maximum likelihood estimates (MLE's) do not make use of the assumed independence of the genetic and non-genetic covariates. In this chapter, following the arguments of Prentice and Pyke (1979), we will illustrate how a polychotomous

logistic regression model can be adapted to fit retrospective data from a case-control study under independence of the binary genetic covariate and continuous non-genetic attribute.

2.1 Incorporating the Independence Assumption

Under independence, the covariate distribution in controls, $P(Z \mid D = 0)$, can be expressed as:

$$P(Z \mid D = 0) = P(Z_g \mid D = 0)P(Z_a \mid D = 0).$$

Assuming that $\Pr(D = 1 \mid Z)$ follows the logistic regression model (2.1), the covariate distribution $P(Z \mid D = 1)$ in cases can be written as a function of β and the control covariate distribution. In Appendix A we show that

$$P(Z \mid D = 1) = P(Z \mid D = 0) \times \frac{\exp(Z\beta)}{\mathrm{E}\{\exp(Z\beta) \mid D = 0\}}. \tag{2.3}$$

Recalling that $\beta_0 = (0, 0, 0, 0, 0)^T$ and $\beta_1 = \beta$, equation (2.3) can be generalized to

$$P(Z \mid D = i) \;\; = \;\; P(Z \mid D = 0) \times \frac{\exp(Z\beta_i)}{\mathrm{E}\{\exp(Z\beta_i) \mid D = 0\}}, \tag{2.4}$$

and under independence, (2.4) can be expressed as

$$[P(Z_a \mid D = 0)P(Z_g \mid D = 0)] \times \frac{\exp(Z\beta_i)}{\mathrm{E}\{\exp(Z\beta_i) \mid D = 0\}}. \tag{2.5}$$

Therefore, a likelihood analysis would involve fitting a product of terms of the form in (2.5) to the data.

2.2 Likelihood

In a case-control study we over-sample from the case population and under-sample from the control population, and as a result, our sample is not representative of the general population. Although artificial, we could view our sample as being selected

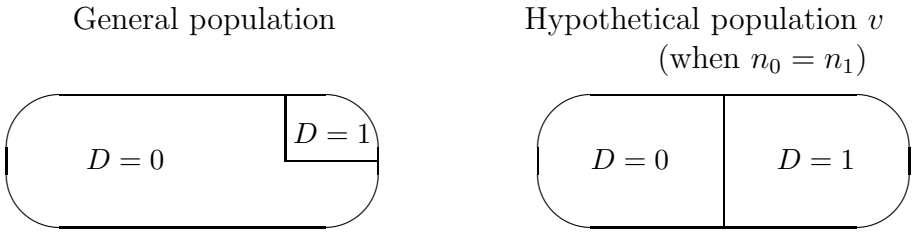

Figure 2.1: Population diagram

randomly from a hypothetical population v with a proportion of $P_v(D = 0) = n_0/n$ controls, and a proportion of $P_v(D = 1) = n_1/n$ cases. Let Z_{ij} be the covariate vector of the jth subject in disease category i. An important feature of the hypothetical population v is that the conditional probability $P_v(Z_{ij} \mid D = i)$ of the covariate vector given disease status is the same as the conditional probability $P(Z_{ij} \mid D = i)$ in the general population because the covariates for population v originate from the general population.

Define γ_k for $k = 0, 1$ such that

$$\gamma_k = \log\left[\frac{P(Z_g = k \mid D = 0)}{P(Z_g = 0 \mid D = 0)}\right]. \tag{2.6}$$

Then the case-control likelihood can be expressed as

$$\begin{aligned} L\left(\beta, \gamma_1, P(Z_a \mid D = 0)\right) &= \prod_{i=0}^{1}\prod_{j=1}^{n_i} P(Z_{ij} \mid D = i) \\ &= \prod_{i=0}^{1}\prod_{j=1}^{n_i} P_v(Z_{ij} \mid D = i) \\ &= \prod_{i=0}^{1}\prod_{j=1}^{n_i} P_v(D = i, Z_{gij} \mid Z_{aij}) P_v(Z_{aij}) / P_v(D = i) \\ &= \prod_{i=0}^{1}\prod_{j=1}^{n_i} \frac{n}{n_i} P_v(D = i, Z_{gij} \mid Z_{aij}) P_v(Z_{aij}) \end{aligned} \tag{2.7}$$

Appendix B shows that $P_v(D = i, Z_{gij} \mid Z_{aij})$ and $P_v(Z_{aij})$ can be expressed in terms

of β, γ_1, and $P(Z_a \mid D = 0)$. Hence, the likelihood is a function of the desired parameters β, γ_1, and $P(Z_a \mid D = 0)$.

2.3 Reparameterization

Since $Z_a = (A, A^2)$ contains a continuous random variable with a completely unspecified distribution function, $P(Z_a \mid D = 0)$ is an *infinite dimensional parameter*, and so maximization of the likelihood is not straightforward. Prentice and Pyke (1979) showed that a similar estimation problem could be clarified by reparameterization. Following their arguments, we will approach our problem by reparameterizing the case-control likelihood in terms of β, γ_1, and $P_v(Z_a)$ instead of the original parameters β, γ_1, and $P(Z_a \mid D = 0)$. Throughout, we assume a risk model with interaction between Z_g and A:

$$logit(P(D = 1 \mid Z)) = \alpha + Z_g\beta_g + Z_a\beta_a + Z_gZ_a\beta_{ga}.$$

Reparameterizing the original likelihood (2.7) will give

$$\begin{aligned}
&L(\beta, \gamma_1, P_v(Z_a)) \\
&= \prod_{i=0}^{1}\prod_{j=1}^{n_i} \frac{n}{n_i}\, P_v(D = i, Z_{gij} \mid Z_{aij})\, P_v(Z_{aij}) \\
&\overset{(B.6)}{=} \prod_{i=0}^{1}\left(\frac{n}{n_i}\right)^{n_i}\prod_{j=1}^{n_i} \frac{\exp\left\{\delta_i\left(\beta, \gamma_1, P_v(Z_a)\right) + \gamma_{Z_{gij}} + Z_{gij}\beta_{gi} + Z_{aij}\beta_{ai} + Z_{gij}Z_{aij}\beta_{gai}\right\} \times P_v(Z_{aij})}{\sum_{m=0}^{1}\sum_{l=0}^{1}\exp(\delta_l\left(\beta, \gamma_1, P_v(Z_a)\right) + \gamma_m + m\beta_{gl} + Z_{aij}\beta_{al} + mZ_{aij}\beta_{gal})} \\
&= \frac{n^n}{n_0^{n_0}n_1^{n_1}} \times \left[\prod_{i=0}^{1}\prod_{j=1}^{n_i} P_v(Z_{aij})\right] \times \\
&\qquad \left[\prod_{i=0}^{1}\prod_{j=1}^{n_i} \frac{\exp\left\{\delta_i\left(\beta, \gamma_1, P_v(Z_a)\right) + \gamma_{Z_{gij}} + Z_{gij}\beta_{gi} + Z_{aij}\beta_{ai} + Z_{gij}Z_{aij}\beta_{gai}\right\}}{\sum_{m=0}^{1}\sum_{l=0}^{1}\exp(\delta_l\left(\beta, \gamma_1, P_v(Z_a)\right) + \gamma_m + m\beta_{gl} + Z_{aij}\beta_{al} + mZ_{aij}\beta_{gal})}\right],
\end{aligned} \tag{2.8}$$

where β_{gi}, β_{ai} and $\beta_{gai}, i = 0, 1$ are defined as in (B.1) and (B.2) and $\delta_i(\beta, \gamma_1, P_v(Z_a))$ is definded as in (B.5). Appendix C shows that this reparameterization is legitimate because the new parameter $P_v(Z_a)$ can be written in terms of the original parameters β, γ_1 and $P(Z_a \mid D = 0)$.

2.4 Overparameterization

Since $P(Z \mid D=1)$ is a probability distribution, $\delta_1 = \delta_1(\beta, \gamma_1, P_v(Z_a))$ is constrained by the following equation:

$$\begin{aligned}
1 &= \int \sum_{k=0}^{1} P(Z_g = k, Z_a = z_a \mid D=1) dz_a \\
&= \int \sum_{k=0}^{1} P_v(Z_g = k, Z_a = z_a \mid D=1) dz_a \\
&= \int \sum_{k=0}^{1} \frac{P_v(D=1, Z_g = k \mid z_a) P_v(z_a)}{P_v(D=1)}\, dz_a \\
&= \frac{n}{n_1} \int \sum_{k=0}^{1} P_v(D=1, Z_g = k \mid z_a) P_v(z_a)\, dz_a \\
&\overset{(B.6)}{=} \frac{n}{n_1} \int \sum_{k=0}^{1} \frac{\exp\{\delta_1 + \gamma_k + k\beta_{g1} + z_a\beta_{a1} + kz_a\beta_{ga1}\}}{\sum_{m=0}^{1}\sum_{l=0}^{1} \exp\{\delta_l + \gamma_m + m\beta_{gl} + z_a\beta_{al} + mz_a\beta_{gal}\}}\, P_v(z_a)\, dz_a \\
&= \frac{n}{n_1} \int \frac{\sum_{m=0}^{1} \exp\{\delta_1 + \gamma_m + m\beta_{g1} + z_a\beta_{a1} + mz_a\beta_{ga1}\}}{\sum_{m=0}^{1}\sum_{l=0}^{1} \exp\{\delta_l + \gamma_m + m\beta_{gl} + z_a\beta_{al} + mz_a\beta_{gal}\}}\, P_v(z_a)\, dz_a \quad (2.9)
\end{aligned}$$

As in Prentice and Pyke (1979), we ignore this constraint and view the reparameterized likelihood (2.8) as a function of β, γ_1, $P_v(Z_a)$ and δ_1, pretending that δ_1 is a free parameter. The corresponding overparameterized likelihood $\tilde{L}(\beta, \gamma_1, \delta_1, P_v(Z_a))$ has a maximum that must be at least as large as that of the true likelihood with the constraint (2.9). That is, we maximize

$$\begin{aligned}
&\tilde{L}(\beta, \gamma_1, \delta_1, P_v(Z_a)) \\
&= \frac{n^n}{n_0^{n_0} n_1^{n_1}} \times \left[\prod_{i=0}^{1} \prod_{j=1}^{n_i} \frac{\exp\left\{\delta_i + \gamma_{Z_{gij}} + Z_{gij}\beta_{gi} + Z_{aij}\beta_{ai} + Z_{gij}Z_{aij}\beta_{gai}\right\}}{\sum_{m=0}^{1}\sum_{l=0}^{1} \exp(\delta_l + \gamma_m + m\beta_{gl} + Z_{aij}\beta_{al} + mZ_{aij}\beta_{gal})} \right] \times \\
&\qquad \left[\prod_{i=0}^{1} \prod_{j=1}^{n_i} P_v(Z_{aij}) \right] \\
&\propto L_1(\beta, \gamma_1, \delta_1) \times L_2(P_v(Z_a)) \quad (2.10)
\end{aligned}$$

as a function of the new parameters $\beta, \gamma_1, \delta_1$ and $P_v(Z_a)$. Appendix D shows that the resulting estimators satisfy the constraint (2.9).

Because $\tilde{L}(\beta, \gamma_1, \delta_1, P_v(Z_a))$ is factorized as in (2.10), it suffices to maximize L_1 as a function of β, γ_1, and δ_1 and to maximize L_2 as a function of $P_v(Z_a)$. We will focus only on maximizing L_1 because we are interested in the maximum likelihood estimates of the risk parameters β; γ_1 and $P_v(Z_a)$ are nuisance parameters.

If we look at the numerator of

$$L_1(\beta, \gamma_1, \delta_1) = \left[\prod_{i=0}^{1} \prod_{j=1}^{n_i} \frac{\exp\left\{\delta_i + \gamma_{Z_{gij}} + Z_{gij}\beta_{gi} + Z_{aij}\beta_{ai} + Z_{gij}Z_{aij}\beta_{gai}\right\}}{\sum_{m=0}^{1}\sum_{l=0}^{1}\exp(\delta_l + \gamma_m + m\beta_{gl} + Z_{aij}\beta_{al} + mZ_{aij}\beta_{gal})} \right] ,$$

it is

$$\begin{array}{ll} 1 & \text{when } D = 0, Z_g = 0; \\ \exp(\gamma_1) & \text{when } D = 0, Z_g = 1; \\ \exp(\delta_1 + Z_a\beta_a) & \text{when } D = 1, Z_g = 0; \text{ and} \\ \exp(\gamma_1 + \delta_1 + \beta_g + Z_a(\beta_a + \beta_{ga})) & \text{when } D = 1, Z_g = 1. \end{array}$$

Now let C denote a joint disease-genetic category such that

$$C = \begin{cases} 0 & \text{for } D{=}0,\ Z_g{=}0 \\ 1 & \text{for } D{=}0,\ Z_g{=}1 \\ 2 & \text{for } D{=}1,\ Z_g{=}0 \\ 3 & \text{for } D{=}1,\ Z_g{=}1, \end{cases}$$

and let $Z_{aij} = (A_{ij}, A_{ij}^2)$ where A_{ij} is the value of the non-genetic attribute for the jth subject in category i. Let us also define $\eta_0 = 0$, $\eta_1 = \gamma_1$, $\eta_2 = \delta_1$, $\eta_3 = \gamma_1 + \delta_1 + \beta_g$, $\boldsymbol{\xi}_0 = \boldsymbol{\xi}_1 = (0,0)^T$, $\boldsymbol{\xi}_2 = \beta_a = (\beta_{1a}, \beta_{2a})^T$, and $\boldsymbol{\xi}_3 = \beta_a + \beta_{ga} = (\beta_{1a} + \beta_{1ga}, \beta_{2a} + \beta_{2ga})^T$ and relabel the subjects according to their categories. Then, we can rewrite $L_1(\beta, \gamma_1, \delta_1)$ in terms of $\eta_1, \eta_2, \eta_3, \boldsymbol{\xi}_2$ and $\boldsymbol{\xi}_3$ such that

$$L_1(\eta_1, \eta_2, \boldsymbol{\xi}_2, \eta_3, \boldsymbol{\xi}_3) = \prod_{i=0}^{3} \prod_{j=1}^{m_i} \frac{\exp(\eta_i + Z_{aij}\boldsymbol{\xi}_i)}{\sum_{k=0}^{3}\exp(\eta_k + Z_{aij}\boldsymbol{\xi}_k)}, \tag{2.11}$$

where m_i is the number of subjects in category $C = i$. Notice that equation (2.11) is in the same form as the likelihood from a prospective generalized logistic regression

model of C regressed on the attribute information coded in $Z_a = (A, A^2)$. (See for example equation (4) in Prentice and Pyke (1979).) Therefore, a polychotomous logistic regression model can be employed when analyzing the data.

2.5 Parameter Estimation

The equivalent polychotomous logistic regression model to Equation (2.11) can be written as

$$\log\left[\frac{P_v(C=i \mid Z_a)}{P_v(C=0 \mid Z_a)}\right] = \eta_i + Z_a\boldsymbol{\xi}_i,$$

where $i = 1, \ldots, 3; \quad \eta_0 = 0; \quad \boldsymbol{\xi}_0 = \boldsymbol{\xi}_1 = (0,0)^T$ and P_v denotes the probability for the hypothetical population v. More explicitly,

$$\begin{aligned}
\log\left[\frac{P_v(C=3 \mid Z_a)}{P_v(C=0 \mid Z_a)}\right] &= \eta_3 + Z_a\boldsymbol{\xi}_3 \\
\log\left[\frac{P_v(C=2 \mid Z_a)}{P_v(C=0 \mid Z_a)}\right] &= \eta_2 + Z_a\boldsymbol{\xi}_2 \\
\log\left[\frac{P_v(C=1 \mid Z_a)}{P_v(C=0 \mid Z_a)}\right] &= \eta_1.
\end{aligned}$$

At first glance, it appears that $(\eta_1, \eta_2, \boldsymbol{\xi}_2, \eta_3, \boldsymbol{\xi}_3)$ could be estimated by fitting three separate binary logistic regressions: one for each of the above equations. However, this only approximates the maximum likelihood solution. In order to get the actual values that maximize the likelihood (2.11), we consider the Newton-Raphson method.

Notice that the conditional probability $P_v(C = i \mid Z_{aij})$ of joint disease-gene category C conditioned on the attribute information coded in Z_a can be expressed in terms of η_i and $\boldsymbol{\xi}_i$ such that,

$$P_v(C=i \mid Z_{aij}) = \frac{\exp(\eta_i + Z_{aij}\boldsymbol{\xi}_i)}{\sum_{k=0}^{3}\exp(\eta_k + Z_{aij}\boldsymbol{\xi}_k)} \qquad i = 1, \ldots, 3.$$

Let $\theta \equiv (\eta_1, \eta_2, \boldsymbol{\xi}_2^T, \eta_3, \boldsymbol{\xi}_3^T)^T$, and $l_1(\theta) \equiv l_1(\eta_1, \eta_2, \boldsymbol{\xi}_2, \eta_3, \boldsymbol{\xi}_3) = \log L_1(\eta_1, \eta_2, \boldsymbol{\xi}_2, \eta_3, \boldsymbol{\xi}_3)$. Then, the maximum likelihood estimate $\hat{\theta}$ must satisfy

$$\dot{l}_1(\theta) = \frac{\partial l_1}{\partial \theta} = (0,0,0,0,0,0,0)^T \tag{2.12}$$

where the first, second, and fifth elements of the score vector $\dot{l}_1(\theta)$ can be shown to be

$$\frac{\partial l_1}{\partial \eta_k} = m_k - \sum_{i=0}^{3} \sum_{j=1}^{m_i} P_v(C = k \mid Z_{aij}) \quad k = 1, 2, 3 \ ,$$

the third and fourth elements to be

$$\frac{\partial l_1}{\partial \boldsymbol{\xi}_2} = \sum_{j=1}^{m_2} Z_{a2j}^T - \sum_{i=0}^{3} \sum_{j=1}^{m_i} Z_{aij}^T P_v(C = 2 \mid Z_{aij}) \ ,$$

and the sixth and seventh elements to be

$$\frac{\partial l_1}{\partial \boldsymbol{\xi}_3} = \sum_{j=1}^{m_3} Z_{a3j}^T - \sum_{i=0}^{3} \sum_{j=1}^{m_i} Z_{aij}^T P_v(C = 3 \mid Z_{aij}).$$

The solutions to equation (2.12) can be obtained by a Newton-Raphson algorithm and, for this, we need the Hessian matrix.

The Hessian matrix may be expressed in terms of blocks $B_{11}(1 \times 1), B_{12}$ and $B_{13}(1 \times 3)$, and B_{22}, B_{23} and $B_{33}(3 \times 3)$ as

$$\ddot{l}_1(\theta) = \frac{\partial^2 l_1}{\partial \theta \partial \theta^T} = \begin{bmatrix} B_{11} & B_{12} & B_{13} \\ B_{12}^T & B_{22} & B_{23} \\ B_{13}^T & B_{23}^T & B_{33} \end{bmatrix} .$$

The blocks are defined by the three sets of parameters η_1 (one element), $(\eta_2, \boldsymbol{\xi}_2^T)$ (three elements) and $(\eta_3, \boldsymbol{\xi}_3^T)$ (three elements) as

$$B_{11} = \frac{\partial^2 l_1}{\partial \eta_1^2}, \qquad B_{1l} = \left[\frac{\partial^2 l_1}{\partial \eta_1 \partial \eta_l} \ , \ \frac{\partial^2 l_1}{\partial \eta_1 \partial \boldsymbol{\xi}_l^T} \right] \quad \text{for } l = 2, 3 \ ,$$

$$\text{and} \qquad B_{kl} = \begin{bmatrix} \frac{\partial^2 l_1}{\partial \eta_k \partial \eta_l} & \frac{\partial^2 l_1}{\partial \eta_k \partial \boldsymbol{\xi}_l^T} \\ \frac{\partial^2 l_1}{\partial \boldsymbol{\xi}_k \partial \eta_l} & \frac{\partial^2 l_1}{\partial \boldsymbol{\xi}_k \partial \boldsymbol{\xi}_l^T} \end{bmatrix} \quad \text{for } (k, l) = (2, 2), (2, 3), (3, 3) \ .$$

The second derivatives that make up these blocks can be written as

$$\frac{\partial^2 l_1}{\partial \eta_k \partial \eta_l} = \begin{cases} \sum_{i=0}^{3} \sum_{j=1}^{m_i} P_v(C = k \mid Z_{aij}) P_v(C = l \mid Z_{aij}) & k \neq l \\ -\sum_{i=0}^{3} \sum_{j=1}^{m_i} P_v(C = k \mid Z_{aij}) \left[1 - P_v(C = k \mid Z_{aij})\right] & k = l \end{cases}$$

$$\frac{\partial^2 l_1}{\partial \eta_k \partial \boldsymbol{\xi}_l^T} = \begin{cases} \sum_{i=0}^{3} \sum_{j=1}^{m_i} Z_{aij} P_v(C = k \mid Z_{aij}) P_v(C = l \mid Z_{aij}) & k \neq l \\ -\sum_{i=0}^{3} \sum_{j=1}^{m_i} Z_{aij} P_v(C = k \mid Z_{aij}) \left[1 - P_v(C = k \mid Z_{aij})\right] & k = l \end{cases}$$

$$\frac{\partial^2 l_1}{\partial \boldsymbol{\xi}_k \partial \eta_l} = \left(\frac{\partial^2 l_1}{\partial \eta_l \partial \boldsymbol{\xi}_k^T} \right)^T$$

$$\frac{\partial^2 l_1}{\partial \boldsymbol{\xi}_k \partial \boldsymbol{\xi}_l^T} = \begin{cases} \sum_{i=0}^{3} \sum_{j=1}^{m_i} Z_{aij}^T Z_{aij} P_v(C = k \mid Z_{aij}) P_v(C = l \mid Z_{aij}) & k \neq l \\ -\sum_{i=0}^{3} \sum_{j=1}^{m_i} Z_{aij}^T Z_{aij} P_v(C = k \mid Z_{aij}) \left[1 - P_v(C = k \mid Z_{aij})\right] & k = l \end{cases}$$

Initial estimates for the Newton-Raphson algorithm may be obtained as follows. Estimate η_i and $\boldsymbol{\xi}_i$ (with $\boldsymbol{\xi}_1 = (0,0)^T$) from three separate binary logistic regressions in which subjects with $C = i$ (i = 1,2,3) are treated as "successes" and subjects with $C = 0$ are treated as "failures". Then proceed with the usual Newton-Raphson updates where the parameter values θ^{t+1} at step $t+1$ are obtained from parameters θ^t at step t via the formula

$$\theta^{t+1} = \theta^t - \left[\ddot{l}(\theta^t)\right]^{-1} \dot{l}(\theta^t)$$

The estimates of the original parameters $(\gamma_1, \delta_1, \beta_g, \beta_a^T, \beta_{ga}^T)$ can be obtained by linear transformation of $\theta^T = (\eta_1, \eta_2, \boldsymbol{\xi}_2^T, \eta_3, \boldsymbol{\xi}_3^T)$ with $\gamma_1 = \eta_1$, $\delta_1 = \eta_2$, $\beta_g = \eta_3 - \eta_2 - \eta_1$,

$\beta_a = \boldsymbol{\xi}_2$ and $\beta_{ga} = \boldsymbol{\xi}_3 - \boldsymbol{\xi}_2$. Hence estimation of $\theta^T = (\eta_1, \eta_2, \boldsymbol{\xi}_2^T, \eta_3, \boldsymbol{\xi}_3^T)$ is equivalent to estimation of $(\gamma_1, \delta_1, \beta_g, \beta_a^T, \beta_{ga}^T)$.

Chapter 3

Application: Analysis of T1DM Data

Logistic regression is a powerful analytic tool in epidemiologic studies. However, a standard logistic regression model cannot incorporate the assumption of independence between genetic and non-genetic factors. As we saw in Chapter 1, imposing the valid assumption of independence is expected to improve estimation precision. In Chapter 2, we enforced the independence assumption and developed maximum-likelihood estimators of disease risk which can be obtained by fitting a polychotomous logistic model of joint disease and genetic status. In this chapter, we apply the proposed method to the T1D data and compare the results to those from a standard logistic regression. Throughout the section, let D denote whether or not a subject is an incident patient with type 1 diabetes, Z_g the GCLC8 status of the genotype, and Z_a the information coded for age. Specifically, D=1 represents the incident patients with type 1 diabetes, and D=0, the controls; $Z_g = 1$, those with at least one copy of GLCL8 (i.e., GCLC8+), and Z_g=0 those without any copy of GCLC8.

3.1 Independence Assumption

Umbach and Weinberg (1997) developed maximum-likelihood methods for inference of disease associations with categorical covariates. They imposed the independence assumption through a log-linear model. Albert et al. (2001) showed that the resulting inference is very sensitive to validity of the independence assumption. Validity of the independence assumption is therefore a concern for our extension of the log-linear approach to continuous covariates. Hence, before we proceed with the application, we use the control data to verify whether or not GCLC8 frequencies can be assumed to be independent of age in the Swedish population. As mentioned in the Introduction (Chapter 1), it is biologically reasonable to assume these frequencies are independent of age in the Swedish population and in the controls. Figure 3.1 shows an exploratory plot of the trend in the frequencies of GCLC8+ as a function of age in controls. The plot was produced by fitting a generalized additive logistic model to the data with a smoothing term for age. (Hastie and Tibshirani 1990). The apparent lack of trend in the figure is consistent with our assumption of independence.

3.2 Descriptive Summaries

Prior to the formal analyses, we made some descriptive summaries to familiarize ourselves with the data. Table 3.1 compares the marginal distribution of GCLC8 in patients and controls, unadjusted for age. There is no statistical evidence that the distribution of GCLC8+ genotypes differs between patients and controls, according to a Chi-squared test on 1 degree of freedom ($P = 0.91$). To compare the marginal distribution of age in patients and controls (unadjusted for GCLC8 status) we fit a generalized additive logistic model with a smoothing term for age to case-control status. Figure 3.3 shows the resulting smoothed curve, which suggests that controls are under-represented among the younger age groups. This under-representation may reflect the difficulty of recruiting younger controls for this study in which controls were

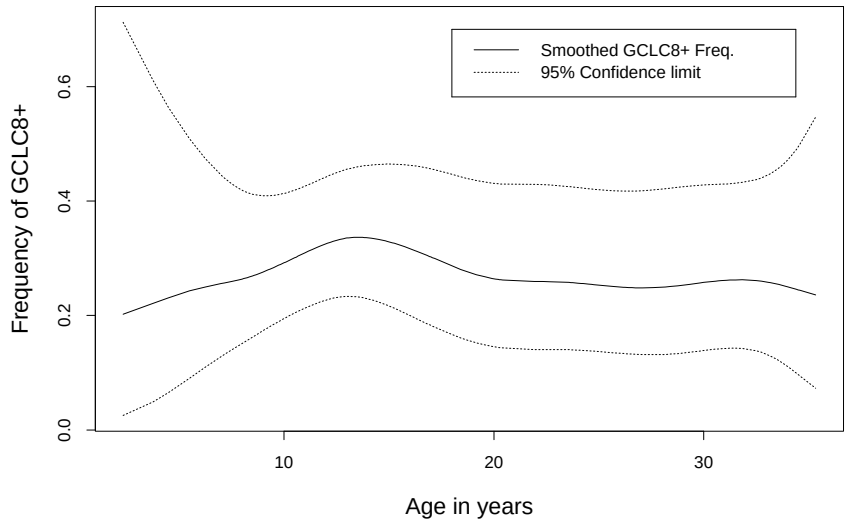

Figure 3.1: The fitted frequencies of GCLC8+ by age in the controls with 95% pointwise confidence limits

initially to be age-matched to cases (Graham et al. 1999). To explore the possibility of age-by-GCLC8 interaction, we plotted the age distribution by case-control and GCLC8 status (Figure 3.2). From the plot, GCLC8+ cases appeared to have younger age-at-onset.

3.3 Model Fitting

Throughout, we assume that the conditional probability $Pr(D = 1 \mid Z_g, Z_a)$ of developing T1D given GCLC8 genotype and age follows a binary logistic regression model. Since one of the main aims is to study statistical interaction between GCLC8 and age, we start by fitting a simple linear trend in age and a linear GCLC8-by-age interaction term. We fit the standard binary logistic model (2.1) and the logistic model imposing independence (2.11), with $Z_a = Age$, $\beta_a = \beta_{1a}$ and $\beta_{ga} = \beta_{1ga}$. For model (2.11),

Table 3.1: Distribution of GCLC8 status in T1D patients and control subjects

	Number (%)		
	GCLC8 +	GCLC8 -	Total
Controls	53 (28.5)	133 (71.5)	186 (100)
Patients	51 (28.5)	128 (71.5)	179 (100)

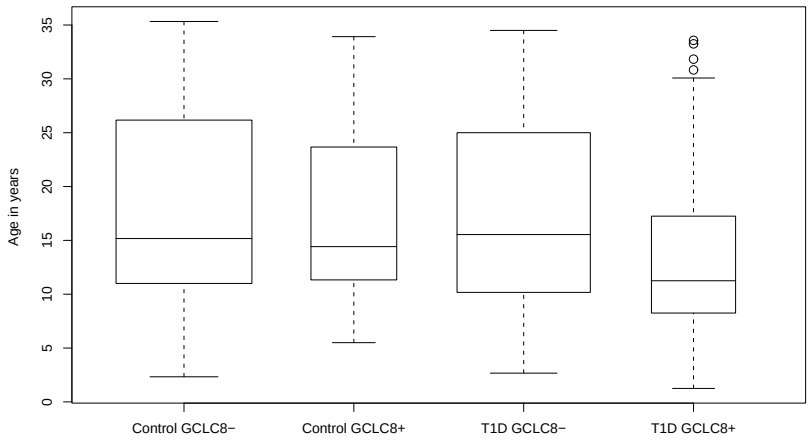

Figure 3.2: Age distribution by case-control and GCLC8 status

the joint disease and genetic status C is defined as C=0 for the GCLC8- controls; C=1 for the GCLC8+ controls; C=2 for the GCLC8- T1D patients; and C=3 for the GCLC8- patients. The freely available statistical software R (www.r-project.org) is used for all of our model fitting.

Maximum likelihood estimates from the standard binary logistic model (2.1) were obtained through the usual iteratively reweighted least squares method; those from the logistic model imposing independence were obtained through the Newton-Raphson iterative algorithm, which we implemented in R. Appendix E contains the

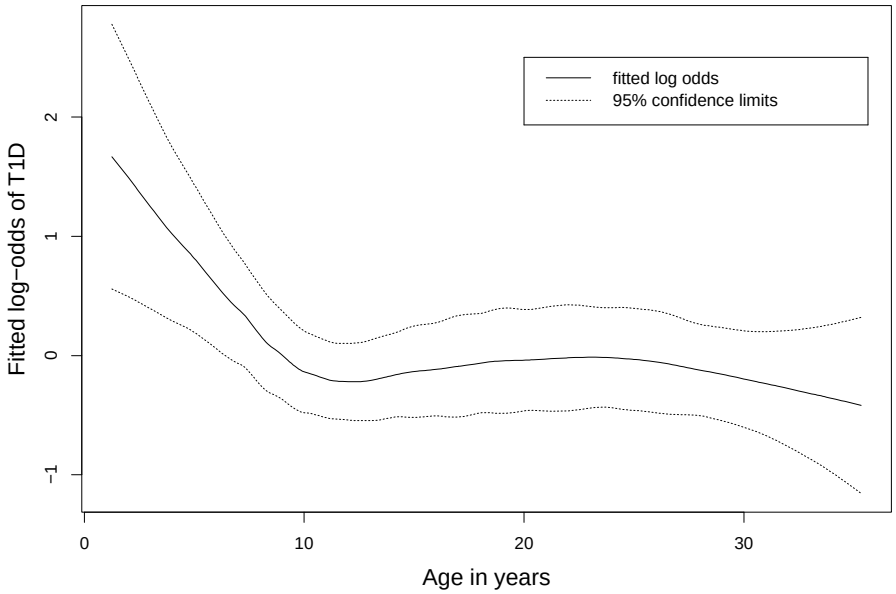

Figure 3.3: Smoothed log-odds of T1D by age

documentation for our R function implementing logistic regression under independence. In this application, our function took about four Newton-Raphson iterations to converge. At convergence, we took a linear transformation of η and $\boldsymbol{\xi}$ to calculate the risk parameters β of interest. Corresponding standard errors of the estimates to model (2.10) were obtained via the non-parametric bootstrap resampling method (Efron and Tibshirani 1993). Bootstrap samples were drawn separately for cases and controls. Large sample theory for maximum-likelihood estimators could also be used to derive the asymptotic variances and covariances as in Prentice and Pyke (1979); however, this is beyond the scope of this master's project.

In model building, we next added quadratic terms for age and its interaction with GCLC8 (i.e., $Age^2, Age^2 \times GCLC8$). The significance of both terms was assessed simultaneously on the basis of Wald tests with significance level 10% for inclusion. In order to assess significance, we fit models (2.1) and (2.10), with $Z_a = (Age, Age^2)$ and

$\beta_a = (\beta_{1a}, \beta_{2a})^T$, and $\beta_{ga} = (\beta_{1ga}, \beta_{2ga})^T$ defined as in (A.1) and (A.2) in Appendix A. For model (2.1), the Wald test of the hypothesis $H_0 : \beta_{2a} = \beta_{2ga} = 0$ has $\chi^2 = 3.06$ on 2 degrees of freedom ($P = 0.22$), and for model (2.10), the Wald test of the same hypothesis has $\chi^2 = 2.86$ (P= 0.24). Thus neither of the tests support the addition of quadratic terms in age to the model.

Based on these results, we estimated the disease risk parameters from the initial models. Table 3.2 gives maximum likelihood estimates with corresponding standard errors from fitting (2.1) and (2.10), with linear age and GCLC8-by-age interaction terms. As shown in Table 3.2, estimates from the two approaches were roughly

Table 3.2: Parameter estimates and standard errors from standard logistic regression (2.1) and logistic regression under independence (2.10)

	Estimate (Standard Error*)		Wald Statistic (p-value)	
parameter	Equation (2.1)	Equation (2.10)	Equation (2.1)	Equation (2.10)
β_g	0.687 (0.507)	0.876 (0.407)	1.36 (0.18)	2.15 (0.03)
β_a	-0.008 (0.014)	-0.006 (0.013)	-0.60 (0.55)	-0.44 (0.66)
β_{ga}	-0.046 (0.028)	-0.057 (0.024)	-1.62 (0.10)	-2.41 (0.02)

* Standard errors from model (2.10) were calculated via the nonparametric bootstrap

comparable for all parameters. However, standard errors of the estimates were smaller under model (2.10) which incorporated independence between GCLC8 and age. As anticipated, there was a noticeable enhancement in precision to detect GCLC8-by-age interaction; with model (2.1), we could not detect this even with the liberal 10% significance level.

For better illustration, in Figure 3.4 we plotted the fitted odds of T1D against age for individuals who are GCLC8+ relative to those who are GCLC8-, with the 95% pointwise confidence limits. From panel B of Figure 3.4, we can see the narrowed confidence intervals for the age-specific odds of T1D at younger and older ages, which contributes to the increased precision to detect GCLC8-by-age interaction.

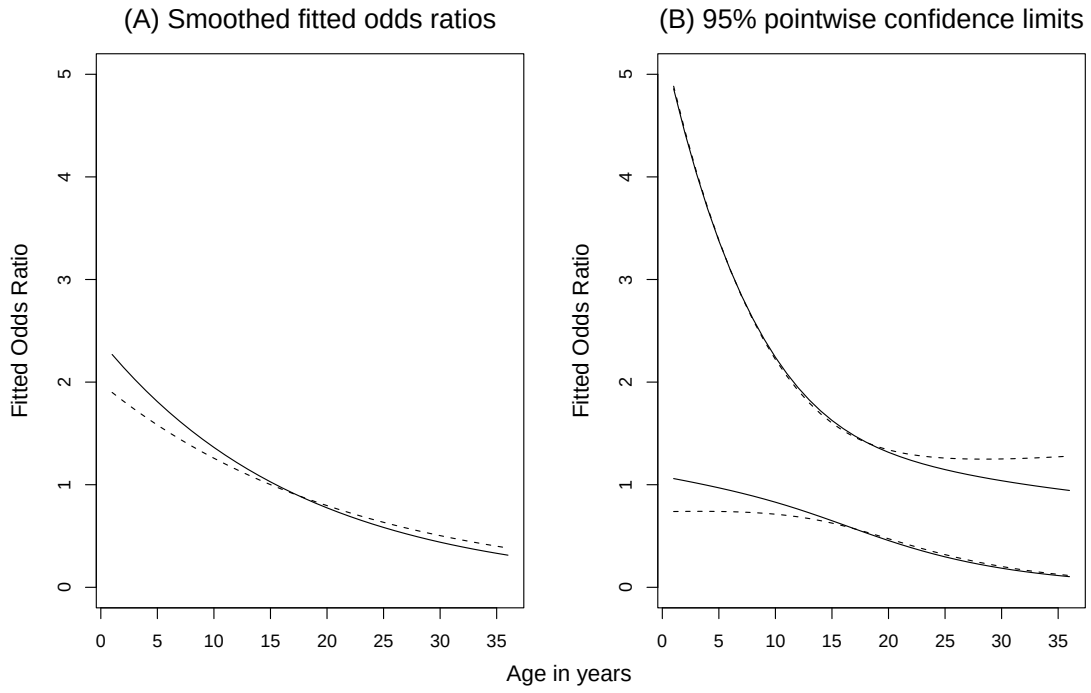

Figure 3.4: Fitted odds ratios (A) and their 95% confidence limits (B) from standard logistic regression (dotted lines) and logistic regression incorporating the independence assumption (solid lines).

Overall, this analysis illustrates that incorporating a valid assumption of independence between a genetic variant and a non-genetic attribute provides improved efficiency. Unlike standard logistic regression, the proposed method gave us the ability to detect not only the interaction between the two factors ($P = 0.016$ versus 0.104) but also the overall effect of the genetic variant. For the overall effect of GCLC8 (i.e., the test of the hypothesis $\beta_g = \beta_{ga} = 0$), the Wald test for model (2.10) gave a marginally significant p-value of 0.05, whereas the Wald test for model (2.1) gave an insignificant p-value of 0.26.

The results from model (2.10) show that the effect of GCLC8 on the risk of T1D is

significantly modified by age. Specifically, it confirms the pattern revealed in Tables 1.1 and 1.2 of the Introduction: the odds of developing T1D in the individuals with at least one copy of GCLC8 relative to those without a copy of GCLC8 is attenuated with age (Figure 3.4A). Note that the estimated odds ratios at ages over 30 do not differ significantly from 1 (Figure 3.4B). As to the overall effect of GCLC8, as shown in Figure 3.5, the risk of developing T1D decreases rapidly with age in individuals with GCLC8. By contrast, in individuals without GCLC8, the risk of T1D does not seem to be affected by age.

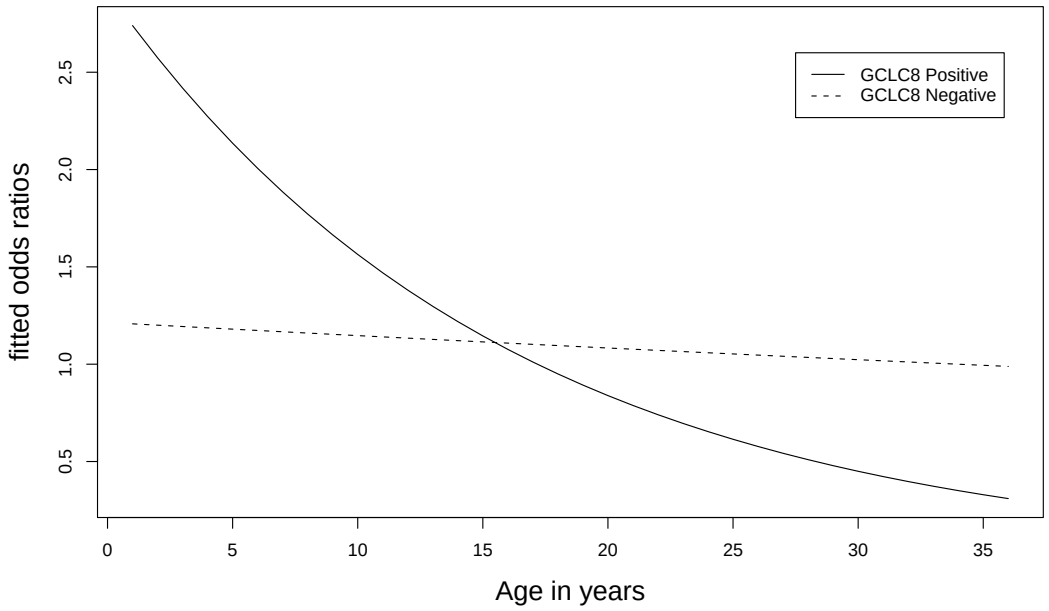

Figure 3.5: Fitted odds of T1D relative to GCLC8- individuals of age 34

Chapter 4

Conclusions

Our interest has been in evaluating the joint effect of a genetic variant and a non-genetic attribute on the risk of T1DM from case-control data. A standard approach to analyzing data from case-control studies is to fit a regression model of disease risk as if the data had arisen prospectively. Prentice and Pyke (1979) justified this practice through theory involving an unspecified covariate distribution. However, in case-control studies, sample sizes required to obtain reasonable power to detect statistical interaction between two factors can be much larger than those required to detect marginal effects of comparable size (Smith and Day 1984). Thus power is a consideration when using standard logistic regression to assess the statistical significance of interaction between a genetic variant and a non-genetic attribute. When a genetic factor and a binary non-genetic factor occur independently in the population, and the disease is rare, Piegorsch et al. (1994) proposed a case-only analysis to estimate statistical interaction between them and demonstrated that the approach had precision that was greater than or equal to a logistic regression analysis with the same number of cases and any number of controls. Umbach and Weinberg (1997) extended this work to categorical non-genetic attributes by developing maximum likelihood estimators based on a log-linear model which enforces the independence assumption. Although their approach provides higher power than standard logistic

regression under independence, it has some drawbacks.

Umbach and Weinberg's log-linear approach allows hybrid designs such as a design with non-genetic information on controls and non-genetic and genetic information on cases. When no genetic information is collected on controls, there is a temptation to abuse the method for inference of interactions between genetic and non-genetic factors without checking the key independence assumption. It has been shown that interaction assessment can be anticonservative when there is dependence between the two factors in the population (Albert et al. 2001). Another drawback is that, when there is no genetic information on controls, estimation of the marginal effect of the genetic variant is impossible. Lastly, since one cannot avoid doing some arbitrary grouping on a continuous attribute, loss of information is also expected when one is interested in interaction between a genetic factor and a continuous attribute.

Considering these drawbacks, we developed maximum likelihood estimators for risk parameters using the information collected from both cases and controls, and allowing for a continuous non-genetic attribute assumed to be independent of the genetic factor in the general population. Our method involves reparameterization of the case-control likelihood under the independence assumption. This reparameterization, which is similar to that in Prentice and Pyke (1979), allows us to have disease risk parameters of interest (i.e, β's) appear in a single term that can be maximized without regard to the infinite-dimensional nuisance parameter describing the distribution of the non-genetic attribute. It turns out that maximization of this single term, which we force to be "unconstrained", pretending the constrained parameter $\delta_1(\beta, \gamma_1, P_v(Z_a))$ is a free parameter δ_1, is equivalent to maximization of a polychotomous logistic model of the joint disease and genetic status. Furthermore, the maximum-likelihood estimators from the "unconstrained" single term satisfy the constraint for δ_1.

In order to investigate potential improvement in precision, we applied the proposed method to the T1DM data from a Swedish case-control study and compared

the results to those from standard logistic regression, after verifying that GCLC8 frequencies are not associated with age in the controls. Although sizes of the estimates for disease risk were similar for both approaches, standard errors were smaller from the proposed approach than those from a standard logistic regression, suggesting that enforcing a valid independence assumption indeed improves the precision of estimation. Most importantly, the proposed approach allowed us to detect interaction between GCLC8 and age while the standard approach was unable to provide sufficient evidence to support this interaction. However, the conclusions are based on the results from the application to only a single real data set. Simulation studies will be required to better assess statistical properties of the proposed method such as power and Type 1 error.

One concern regarding the proposed method is the computational time currently required to estimate the variance-covariance matrix for parameter estimators by bootstrapping. Each time a bootstrap sample is generated, it has to be passed through the Newton-Raphson algorithm which takes three or four iterations for convergence, and over 1000 bootstrap replicates are required for reasonable variance estimation. Running 2000 bootstrap replicates takes about 35 minutes on the Statistics department Sun Ultra 5's. However, this problem will be resolved once we derive the asymptotic variance-covariance matrix using large-sample theory. Another concern is that, like the case-only approach, we expect our approach to lack robustness to deviations from the independence assumption. Nevertheless, for our particular application, it was reasonable to assume that the genetic variant and the non-genetic attribute were independent in the population of interest, and control data were available to check the key assumption.

It should be possible to extend the method to handle more complex data sets than the T1DM data used for this thesis project. Currently, the method is implemented to handle a binary response and a binary genetic variable. It would be useful to generalize the method to accommodate categorical genetic variables (genotypes). Once we

can handle genotypes, we could build in the assumption of population Hardy Weinberg Equilibrium (if valid) for more precision. Another possibility for future work is to extend the R software so that it can control simultaneously for other risk factors which are independent of the genetic risk factor. We could also look into extending the results to allow for categorical responses with more than two values.

In conclusion, our analyses suggest that imposing a valid assumption of independence between a genetic factor and a continuous non-genetic attribute can increase the power and efficiency to detect and estimate interaction between the genetic factor and the attribute in case-control studies for rare diseases. We believe that this method has the same power advantages over standard logistic regression as the log-linear approach since we make use of the same independence assumption. The major potential advantage of our method over the log-linear approach is that, by allowing for a continuous attribute, we avoid potential loss of information that can result from discretizing the attribute. Any subsequent loss of power could be investigated in future simulation studies comparing our method to the log-linear approach.

Appendix A

$P(Z \mid D = 1)$ in terms of β and $P(Z \mid D = 0)$

In equation (2.3), the conditional distribution of covariates $P(Z \mid D = 1)$ in cases is expressed in terms of β and $P(Z \mid D = 0)$. In order to show this, we will look at equation (6) of Prentice and Pyke (1979):

$$P(Z \mid D = i) = c_i \exp\{\gamma(Z) + Z\beta_i\},$$

where $\gamma(Z)$ is defined below and c_i is a normalization constant. For simplicity, we will consider a binary disease response as defined in the beginning of Chapter 2 (i.e., i = 0,1). We will also let $Z = (Z_g, Z_a, Z_g \times Z_a)$, $\beta_0 = (0, 0, 0, 0, 0)^T$ and $\beta_1 = \beta = (\beta_g, \beta_{1a}, \beta_{2a}, \beta_{1ga}, \beta_{2ga})^T$, as defined in Chapter 2.

Define

$$\gamma(Z) = \log\left[\frac{P(Z \mid D = 0)}{P(Z_0 \mid D = 0)}\right],$$

where $Z_0 = (Z_{g0}, Z_{a0}, Z_{g0} \times Z_{a0})$ denotes some arbitrary reference (baseline) covariate vector. Then we have

$$P(Z \mid D = i) = c_i \frac{P(Z \mid D = 0)}{P(Z_0 \mid D = 0)} \exp(Z\beta_i). \tag{A.1}$$

Since $P(Z \mid D = i)$ is a probability distribution for each i,

$$\begin{aligned} 1 &= \int P(Z = z \mid D = i)\, dz \\ &= \int c_i \frac{P(Z = z \mid D = 0)}{P(Z_0 \mid D = 0)} \exp(z\beta_i)\, dz \\ &= \frac{c_i}{P(Z_0 \mid D = 0)} \int P(Z = z \mid D = 0) \exp(z\beta_i)\, dz \\ &= \frac{c_i}{P(Z_0 \mid D = 0)} \mathrm{E}\{\exp(Z\beta_i) \mid D = 0\}. \end{aligned}$$

(Note that the integral sign denotes both integration and summation because Z contains both continuous and discrete random variables.)

Hence

$$c_i = \frac{P(Z_0 \mid D = 0)}{\mathrm{E}\{\exp(Z\beta_i) \mid D = 0\}}, \tag{A.2}$$

which will give us

$$\begin{aligned} P(Z \mid D = i) &\overset{(A.1,A.2)}{=} \left[\frac{P(Z_0 \mid D = 0)}{\mathrm{E}\{\exp(Z\beta_i) \mid D = 0\}}\right] \left[\frac{P(Z \mid D = 0)}{P(Z_0 \mid D = 0)} \exp(Z\beta_i)\right] \\ &= P(Z \mid D = 0) \frac{\exp(Z\beta_i)}{\mathrm{E}\{\exp(Z\beta_i) \mid D = 0\}} \end{aligned}$$

Specifically, when $i = 1$,

$$\begin{aligned} P(Z \mid D = 1) &= P(Z \mid D = 0) \frac{\exp(Z\beta_1)}{\mathrm{E}\{\exp(Z\beta_1) \mid D = 0\}} \\ &= P(Z \mid D = 0) \frac{\exp(Z\beta)}{\mathrm{E}\{\exp(Z\beta) \mid D = 0\}}. \end{aligned}$$

Therefore, the covariate distribution $P(Z \mid D = 1)$ in cases can be expressed in terms of β and $P(Z \mid D = 0)$ as in equation (2.3).

Appendix B

Parametrization of $P_v(Z_a)$ and $P_v(D, Z_g \mid Z_a)$

The likelihood in equation (2.7) contains $P_v(D = i, Z_{gij} \mid Z_{aij})$ and $P_v(Z_{aij})$, which we will show are functions of γ_1, β and $P(Z_a \mid D = 0)$. Let

$$\beta_{g0} = 0, \;\; \beta_{a0}^T = (0, 0) \;\text{ and }\; \beta_{ga0}^T = (0, 0).$$

Then

$$\beta_0 = (\beta_{g0}, \beta_{a0}^T, \beta_{ga0}^T)^T = (0, 0, 0, 0, 0)^T. \tag{B.1}$$

Similarly, let

$$\beta_{g1} = \beta_g, \;\; \beta_{a1}^T = (\beta_{1a}, \beta_{2a}) \;\text{ and }\; \beta_{ga1}^T = (\beta_{1ga}, \beta_{2ga}).$$

Then

$$\beta_1 = (\beta_{g1}, \beta_{a1}^T, \beta_{ga1}^T)^T = (\beta_g, \beta_a^T, \beta_{ga}^T)^T = \beta. \tag{B.2}$$

Since

$$\begin{aligned} P_v(Z) &= P_v(Z \mid D=0)P_v(D=0) + P_v(Z \mid D=1)P_v(D=1) \\ &\overset{(2.3)}{=} P(Z \mid D=0) \times \frac{n_0}{n} + P(Z \mid D=0) \times \frac{\exp(Z\beta)}{\mathrm{E}\{\exp(Z\beta) \mid D=0\}} \times \frac{n_1}{n} \\ &= P(Z \mid D=0)\left[\frac{n_0}{n} + \frac{\exp(Z\beta)}{\mathrm{E}\{\exp(Z\beta) \mid D=0\}}\frac{n_1}{n}\right], \end{aligned} \tag{B.3}$$

we have

$$\begin{aligned} P_v(Z_a) &= \sum_{m=0}^{1} P_v(Z_a, Z_g = m) \\ &\overset{(B.3)}{=} \sum_{m=0}^{1} P_v(Z_a, Z_g = m \mid D=0)\left[\frac{n_0}{n} + \frac{\exp(m\beta_g + Z_a\beta_a + mZ_a\beta_{ga})}{\mathrm{E}\{\exp(Z\beta) \mid D=0\}}\ \frac{n_1}{n}\right] \\ &= P_v(Z_a \mid D=0)\sum_{m=0}^{1} P_v(Z_g = m \mid D=0)\left[\frac{n_0}{n} + \frac{\exp(m\beta_g + Z_a\beta_a + mZ_a\beta_{ga})}{\mathrm{E}\{\exp(Z\beta) \mid D=0\}}\ \frac{n_1}{n}\right] \\ &\overset{(B.1,B.2)}{=} P_v(Z_a \mid D=0)\sum_{m=0}^{1}\sum_{l=0}^{1} P_v(Z_g = m \mid D=0)\frac{n_l}{n}\ \frac{\exp(m\beta_{gl} + Z_a\beta_{al} + mZ_a\beta_{gal})}{\mathrm{E}\{\exp(Z\beta_l) \mid D=0\}} \\ &= P_v(Z_a \mid D=0)\frac{n_0}{n}P_v(Z_g = 0 \mid D=0) \times \\ &\qquad \sum_{m=0}^{1}\sum_{l=0}^{1}\frac{P_v(Z_g = m \mid D=0)}{P_v(Z_g = 0 \mid D=0)}\frac{n_l}{n_0}\ \frac{\exp(m\beta_{gl} + Z_a\beta_{al} + mZ_a\beta_{gal})}{\mathrm{E}\{\exp(Z\beta_l) \mid D=0\}} \\ &\overset{(2.6)}{=} P_v(Z_a \mid D=0)\frac{n_0}{n}P_v(Z_g = 0 \mid D=0) \times \\ &\qquad \sum_{m=0}^{1}\sum_{l=0}^{1}\frac{n_l}{n_0}\ \frac{\exp(\gamma_m + m\beta_{gl} + Z_a\beta_{al} + mZ_a\beta_{gal})}{\mathrm{E}\{\exp(Z\beta_l) \mid D=0\}} \\ &= P_v(Z_a \mid D=0)\frac{n_0}{n}P_v(Z_g = 0 \mid D=0) \times \\ &\qquad \sum_{m=0}^{1}\sum_{l=0}^{1}\exp(\delta_l\left(\beta, \gamma_1, P(Z_a \mid D=0)\right) + \gamma_m + m\beta_{gl} + Z_a\beta_{al} + mZ_a\beta_{gal}), \end{aligned} \tag{B.4}$$

where $\delta_0 = 0$ and, in general,

$$\delta_l\left(\beta, \gamma_1, P(Z_a \mid D=0)\right) = \log\left[\frac{1}{\mathrm{E}\{\exp(Z\beta_l) \mid D=0\}} \times \frac{n_l}{n_0}\right]. \tag{B.5}$$

We also have

$$
\begin{aligned}
P_v(D=i, Z_g=k \mid Z_a) &= \frac{P_v(D=i, Z_g=k, Z_a)}{P_v(Z_a)} \\
&= \frac{P_v(Z_g=k, Z_a \mid D=i)P_v(D=i)}{P_v(Z_a)} \\
&\overset{(2.5)}{=} \frac{P(Z_a \mid D=0)P(Z_g=k \mid D=0)}{P_v(Z_a)} \frac{\exp(k\beta_{gi}+Z_a\beta_{ai}+kZ_a\beta_{gai})}{\mathrm{E}\{\exp(Z\beta_i) \mid D=0\}} P_v(D=i) \\
&= \frac{P(Z_a \mid D=0)P(Z_g=k \mid D=0)}{P_v(Z_a)} \frac{\exp(k\beta_{gi}+Z_a\beta_{ai}+kZ_a\beta_{gai})}{\mathrm{E}\{\exp(Z\beta_i) \mid D=0\}} \frac{n_i}{n} \\
&= \frac{P(Z_a \mid D=0)\frac{n_0}{n}P(Z_g=k \mid D=0)}{P_v(Z_a)} \frac{\exp(k\beta_{gi}+Z_a\beta_{ai}+kZ_a\beta_{gai})\frac{n_i}{n_0}}{\mathrm{E}\{\exp(Z\beta_i) \mid D=0\}} \\
&\overset{(B.5)}{=} \frac{P(Z_a \mid D=0)\frac{n_0}{n}P(Z_g=k \mid D=0)}{P_v(Z_a)} \times \\
&\qquad \exp(\delta_i(\beta, \gamma_1, P(Z_a \mid D=0)) + k\beta_{gi} + Z_a\beta_{ai} + kZ_a\beta_{gai}) \\
&\overset{(B.4)}{=} \frac{P(Z_a \mid D=0)\frac{n_0}{n}P(Z_g=k \mid D=0)}{P(Z_a \mid D=0)\frac{n_0}{n}P(Z_g=0 \mid D=0)} \times \\
&\qquad \frac{\exp(\delta_i(\beta, \gamma_1, P(Z_a \mid D=0)) + k\beta_{gi} + Z_a\beta_{ai} + kZ_a\beta_{gai})}{\sum_{m=0}^{1}\sum_{l=0}^{1}\exp(\delta_l(\beta, \gamma_1, P(Z_a \mid D=0)) + \gamma_m + m\beta_{gl} + Z_a\beta_{al} + mZ_a\beta_{gal})} \\
&\overset{(2.6)}{=} \frac{\exp(\delta_i(\beta, \gamma_1, P(Z_a \mid D=0)) + \gamma_k + k\beta_{gi} + Z_a\beta_{ai} + kZ_a\beta_{gai})}{\sum_{m=0}^{1}\sum_{l=0}^{1}\exp(\delta_l(\beta, \gamma_1, P(Z_a \mid D=0)) + \gamma_m + m\beta_{gl} + Z_a\beta_{al} + mZ_a\beta_{gal})}
\end{aligned}
\tag{B.6}
$$

Therefore, since $\gamma_0 = 0$, both $P_v(D=i, Z_{gij} \mid Z_{aij})$ and $P_v(Z_{aij})$ can be written in terms of β, γ_1 and $P(Z_a \mid D=0)$. Hence the likelihood is a function of the desired parameters β, γ_1 and $P(Z_a \mid D=0)$.

Appendix C

Reparameterization Justification

We wish to reparameterize the case-control likelihood in terms of β, γ_1 and $P_v(Z_a)$. It turns out that the reparameterized likelihood will be easier to maximize for the regression parameters β of interest because it consists of a term involving only the infinite-dimensional parameter $P_v(Z_a)$ and a separate term involving only β and γ_1. If $P(Z_a \mid D = 0)$ can be written in terms of β, γ_1 and $P_v(Z_a)$, we can say that this reparameterization is legitimate.

From the third line leading to Equation (B.4)

$$P_v(Z_a) = P_v(Z_a \mid D = 0) \sum_{m=0}^{1} P_v(Z_g = m \mid D = 0) \left[\frac{n_0}{n} + \frac{\exp(m\beta_g + Z_a\beta_a + mZ_a\beta_{ga})}{\mathrm{E}\{\exp(Z\beta) \mid D = 0\}} \frac{n_1}{n} \right],$$

which implies that

$$P(Z_a \mid D = 0) = \frac{nP_v(Z_a)}{\sum_{m=0}^{1} P_v(Z_g = m \mid D = 0) \left\{ n_0 + n_1 \frac{\exp(m\beta_g + Z_a\beta_a + mZ_a\beta_{ga})}{\mathrm{E}\{\exp(Z\beta) | D=0\}} \right\}}. \quad \text{(C.1)}$$

From Equation (2.6), we can show that $P(Z_g = m \mid D = 0) = \frac{\exp(\gamma_m)}{1+\exp(\gamma_1)}$ where $m = 0, 1$ and $\gamma_0 = 0$. Thus $P_v(Z_g = m \mid D = 0)$ is a function of the parameter γ_1 only. All that remains is to show that $\mathrm{E}\{\exp(Z\beta) \mid D = 0\}$ is also a function of β, γ_1 and $P_v(Z_a)$. In fact, from integrating (C.1), we get that $\mathrm{E}\{\exp(Z\beta) \mid D = 0\}$ is the

solution s to the integral equation

$$1 = n \int \frac{P_v(z_a)}{\sum_{m=0}^{1} P_v(Z_g = m \mid D = 0) \left\{ n_0 + n_1 \frac{\exp(m\beta_g + z_a\beta_a + mz_a\beta_{ga})}{s} \right\}} \, dz_a.$$

We see that this solution s depends only on the parameters $P_v(z_a)$, β and $P_v(Z_g = 1 \mid D = 0)$. Since $P_v(Z_g = 1 \mid D = 0)$ is equivalent to γ_1 by Equation (2.6), we obtain the conclusion that $\mathrm{E}\{\exp(Z\beta) \mid D = 0\}$ is a function of β, γ_1 and $P_v(Z_a)$. Hence, $P(Z_a \mid D = 0)$ can be written in terms of β, γ_1 an $P_v(Z_a)$ and so the reparameterization is justified.

Appendix D

Satisfying the Constraint

Let Z_{ij} be the covariate vector of the jth subject in disease category i. Prentice and Pyke (1979) argued that, because of their choice of reparameterization, it happened that the constraints were satisfied by the unconstrained maximum likelihood estimators. We use analogous arguments to show that our unconstrained maximum likelihood estimators $\hat{\beta}$, $\hat{\delta}_1$, $\hat{\gamma}_1$ and $\hat{P}_v(Z_a)$ also satisfy the constraint (2.9).

From the second line of Equation (2.10), the log of L_1 is

$$\begin{aligned} l_1 &= \log L_1 = \log \left[\prod_{i=0}^{1} \prod_{j=1}^{n_i} \frac{\exp\left\{\delta_i + \gamma_{Z_{gij}} + Z_{gij}\beta_{gi} + Z_{aij}\beta_{ai} + Z_{gij}Z_{aij}\beta_{gai}\right\}}{\sum_{l=0}^{1}\sum_{m=0}^{1}\exp(\delta_l + \gamma_m + m\beta_{gl} + Z_{aij}\beta_{al} + mZ_{aij}\beta_{gal})} \right] \\ &= \sum_{i=0}^{1}\sum_{j=1}^{n_i} \left[\delta_i + \gamma_{Z_{gij}} + Z_{gij}\beta_{gi} + Z_{aij}\beta_{ai} + Z_{gij}Z_{aij}\beta_{gai} - \right. \\ &\qquad \left. \log\left\{ \sum_{l=0}^{1}\sum_{m=0}^{1} \exp(\delta_l + \gamma_m + m\beta_{gl} + Z_{aij}\beta_{al} + mZ_{aij}\beta_{gal}) \right\} \right] \end{aligned}$$

The derivative of l_1 with respect to δ_1 is

$$\frac{\partial l_1}{\partial \delta_1} = n_1 - \sum_{i=0}^{1}\sum_{j=1}^{n_i} \frac{\sum_{m=0}^{1}\exp(\delta_1 + \gamma_m + m\beta_{g1} + Z_{aij}\beta_{a1} + mZ_{aij}\beta_{ga1})}{\sum_{l=0}^{1}\sum_{m=0}^{1}\exp(\delta_l + \gamma_m + m\beta_{gl} + Z_{aij}\beta_{al} + mZ_{aij}\beta_{gal})}$$

Hence the argmaxes $\hat{\gamma}_1$, $\hat{\delta}_1$ and $\hat{\beta}$ of L_1 must satisfy

$$n_1 = \sum_{i=0}^{1}\sum_{j=1}^{n_i} \frac{\sum_{m=0}^{1}\exp(\hat{\delta}_1 + \hat{\gamma}_m + m\hat{\beta}_{g1} + Z_{aij}\hat{\beta}_{a1} + mZ_{aij}\hat{\beta}_{ga1})}{\sum_{l=0}^{1}\sum_{m=0}^{1}\exp(\hat{\delta}_l + \hat{\gamma}_m + m\hat{\beta}_{gl} + Z_{aij}\hat{\beta}_{al} + mZ_{aij}\hat{\beta}_{gal})} \qquad \text{(D.1)}$$

The constraint (2.9) requires that

$$1 = \frac{n}{n_1} \int \frac{\sum_{m=0}^{1} \exp\left\{\hat{\delta}_1 + \hat{\gamma}_m + m\hat{\beta}_{g1} + z_a\hat{\beta}_{a1} + mz_a\hat{\beta}_{ga1}\right\}}{\sum_{l=0}^{1}\sum_{m=0}^{1} \exp\left\{\hat{\delta}_l + \hat{\gamma}_m + m\hat{\beta}_{gl} + z_a\hat{\beta}_{al} + mz_a\hat{\beta}_{gal}\right\}} \hat{P}_v(z_a)\, dz_a$$

But an integral over the empirical distribution $\hat{P}_v(z_a)$ becomes a sum over all observed values of Z_a, weighted by $1/n$. Hence the above equation simplifies to

$$\begin{aligned} 1 &= \frac{n}{n_1} \sum_{i=0}^{1} \sum_{j=1}^{n_i} \frac{\sum_{m=0}^{1} \exp(\hat{\delta}_1 + \hat{\gamma}_m + m\hat{\beta}_{g1} + Z_{aij}\hat{\beta}_{a1} + mZ_{aij}\hat{\beta}_{ga1})}{\sum_{l=0}^{1}\sum_{m=0}^{1} \exp(\hat{\delta}_l + \hat{\gamma}_m + m\hat{\beta}_{gl} + Z_{aij}\hat{\beta}_{al} + mZ_{aij}\hat{\beta}_{gal})} \times \frac{1}{n} \\ &= \frac{1}{n_1} \sum_{i=0}^{1} \sum_{j=1}^{n_i} \frac{\sum_{m=0}^{1} \exp(\hat{\delta}_1 + \hat{\gamma}_m + m\hat{\beta}_{g1} + Z_{aij}\hat{\beta}_{a1} + mZ_{aij}\hat{\beta}_{ga1})}{\sum_{l=0}^{1}\sum_{m=0}^{1} \exp(\hat{\delta}_l + \hat{\gamma}_m + m\hat{\beta}_{gl} + Z_{aij}\hat{\beta}_{al} + mZ_{aij}\hat{\beta}_{gal})} \end{aligned}$$

which holds by (D.1).

Appendix E

Software Documentation

```
Maximum-Likelihood Estimation Under Independence

Description:

     "logistic.UI" is used to fit logistic regression models which
     incorporate the assumption of independence between a genetic
     variant and a continuous non-genetic attribute. The models are
     fitted to data from case-control studies of rare diseases, for
     which genetic and non-genetic factors can be assumed to be
     independent in the study population.  Standard errors for
     estimates of the association parameters are obtained by
     non-parametric bootstrapping.  The bootstrap samples are drawn
     separately for cases and controls.

Usage:

     logistic.UI (data, response, case="NULL", genetic, attribute,
     poly.degree=1, R=5000, maxit=20, tol=0.000001)

Arguments:

    data: The data from a case-control study, as a matrix or data
          frame.  Each row is considered as one multivariate
          observation for a subject.

response: A character string with the name of the binary disease response
          in data.
```

```
    case: A character string of the response variable value which
          represents the cases. If it is ‘‘NULL’’ (default), the higher
          numeric value or character string of the longer length will be
          considered to represent cases.

 genetic: A character string with the name of a binary covariate in
          data representing the genetic variant.

attribute: A character string with the name of the continuous non-genetic
          attribute in data.

poly.degree: A numeric value which represents the polynomial
          relationship between disease response and non-genetic
          attribute.  The default value is 1 for a model with linear
          terms in the non-genetic attribute and in its interaction with
          the genetic variant.

       R: The number of bootstrap replicates; default=5000.

   maxit: Maximum iteration number of Newton-Raphson algorithm;
          default=20.

     tol: Maximum difference in parameter estimates between iterations
          before declaring convergence; default=0.000001.

Details:

     The vectors in data specified by response and genetic are
     typically numeric with two values (e.g., 0 or 1).  Unless a user
     specifies which response value will represent cases (in argument
     'case'), the function will take the higher numeric value, factor
     level or longer character string as a case by default.  Likewise,
     the genetic predictor value with a higher numeric value, factor
     level or longer character string will be considered as the
     genetically-defined group of interest.

Value:

     A list with components

coef.info: A p x 4 matrix (p = 2*poly.degree + 2) with columns for the
```

```
          estimated coefficient, its standard error, Wald t-statistic
          and corresponding p-value (based on asymptotic chi-squared
          distribution).

cov.estimates: A p x p matrix of covariances of the estimated coef[j];
          j=1, ..., p = 2*poly.degree + 2

Examples:

     example <- logistic.UI(GCLCdat, "type", case="IDDM", "GCLC8",
     "age", R=2000)
     example
     ## $coef.info
     ##              Estimate        S.E  Wald Stat      P.value
     ## gamma.1 -0.920057215 0.16258248 -5.6590182 1.522414e-08
     ## betag    0.876017610 0.40745571  2.1499702 3.155758e-02
     ## betaa   -0.005708465 0.01313443 -0.4346183 6.638395e-01
     ## betaga  -0.056590761 0.02350258 -2.4078531 1.604664e-02

     ## $cov.estimates
     ##               gamma.1        betag         betaa        betaga
     ## gamma.1  2.643306e-02 -0.027025579 -3.801833e-05  7.194698e-05
     ## betag   -2.702558e-02  0.166020152  1.889549e-03 -7.845471e-03
     ## betaa   -3.801833e-05  0.001889549  1.725134e-04 -1.031971e-04
     ## betaga   7.194698e-05 -0.007845471 -1.031971e-04  5.523713e-04

     example2 <- logistic.UI(GCLCdat, "type", case="IDDM", "GCLC8",
     "age", poly.degree=2, R=2000)
     round(example2$coef.info,4)
     ##          Estimate    S.E Wald Stat P.value
     ## gamma.1   -0.9201 0.1603   -5.7395  0.0000
     ## betag      1.4181 0.7414    1.9127  0.0558
     ## betaa.1   -0.0541 0.0640   -0.8458  0.3977
     ## betaa.2    0.0012 0.0016    0.7650  0.4443
     ## betaga.1  -0.1411 0.0919   -1.5354  0.1247
     ## betaga.2   0.0024 0.0024    1.0070  0.3140

     ## Wald test of significance of the quadratic terms in
     ## age (betaa.2) and age-by-GCLC8 interaction (betaga.2)
     mycoef <- example2$coef.info[,"Estimate"][c(4,6)]
     mycov <- example2$cov.estimates[c(4,6),c(4,6)]
     my.wald <- t(mycoef)
```

```
1-pchisq(my.wald,2) # p-value = 0.24

## Estimate the frequency of GCLC8+ controls.
my.gamma1 <- example$coef.info["gamma.1","Estimate"]
my.freq <- exp(my.gamma1)/(1+exp(my.gamma1))
## estimated freq. of GCLC8+ controls = 0.28
```

Bibliography

Albert, P. S., D. Ratnasinghe, J. Tangrea, and S. Wacholder (2001). Limitations of the case-only design for identifying gene-environment interactions. *American Journal of Epidemiology 154*, 687–693.

Bekris, L., C. Shephard, J. Graham, B. McNeney, J. Shin, M. Janer, M. Zarghami, F. Farin, A. Kavanagh, and Å. Lernmark (2004). Glutamate cysteine ligase catalytic subunit trinucleotide repeat polymorphism and type 1 diabetes age-at-onset. Unpublished.

Dahlquist, G. G., L. G. Blom, L.-Å. Persson, A. I. M. Sandström, and S. G. I. Wall (1990). Dietary factors and the risk of developing insulin dependent diabetes in childhood. *British Medical Journal 300*, 1302–1306.

Davies, J. L., Y. Kawaguchi, S. T. Bennett, J. B. Copeman, H. J. Cordell, L. E. Pritchard, P. W. Reed, S. C. Gough, S. C. Jenkins, and S. M. Palmer (1994). A genome-wide search for human type 1 diabetes susceptibility genes. *Nature 371*, 130–136.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the bootstrap*. New York: Chapman & Hall.

Gambelunghe, G., M. Ghaderi, C. Tortoioli, A. Falorni, F. Santeusanio, P. Brunetti, C. B. Sanjeevi, and A. Falorni (2001). Two distinct MICA gene markers discriminate major autoimmune diabetes types. *The Journal of Clinical Endocrinology & Metabolism 86*, 3754–3760.

Graham, J., W. A. Hagopian, I. Kockum, L. S. Li, C. B. Sanjeevi, R. M. Lowe, J. B. Schaefer, M. Zarghami, H. L. Day, M. Landin-Olsson, J. P. Palmer, M. Janer-Villanueva, L. Hood, G. Sundkvist, Å. Lernmark, N. Breslow, G. Dahlquist, and G. Blohmé (2002). Genetic effects on age-dependent onset and islet cell autoantibody markers in type 1 diabetes. *Diabetes 51*, 1346–1355.

Graham, J., I. Kockum, C. Sanjeevi, M. Landin-Olsson, L. Nystrom, G. Sundkvist, H. Arnqvist, G. Blohmé, F. Lithner, B. Littorin, B. Schersten, L. Wibell, J. Ostman, Å. Lernmark, N. Breslow, G. Dahlquist, and the Swedish Childhood Diabetes Study Group (1999). Negative association between type 1 diabetes and HLA DQB1*0602-DQA1*0102 is attenuated with age-at-onset. *European Journal of Immunogenetics 26*, 117–127.

Hastie, T. J. and R. J. Tibshirani (1990). *Generalized Additive Models*. London: Chapman & Hall.

Piegorsch, W. W., C. R. Weinberg, and J. A. Taylor (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-base case-control studies. *Statistics in Medicine 13*, 153–162.

Prentice, R. L. and R. Pyke (1979). Logistic disease incidence models and case-control studies. *Biometrika 66*, 403–411.

Rotter, J. I. and D. L. Rimoin (1978). Heterogeneity in diabetes mellitus – update, 1978. *Diabetes 27*, 599–605.

Smith, P. G. and N. E. Day (1984). The design of case-control studies: the influence of confounding and interaction effects. *International Journal of Epidemiology 13*, 356–365.

Todd, J. A. (1990). Genetic control of autoimmunity in type 1 diabetes. *Immunology Today 11*, 122–129.

Umbach, D. M. and C. R. Weinberg (1997). Designing and analyzing case-control studies to exploit independence of genotype and exposure. *Statistics*

in Medicine 16, 1731–1743.

Walsh, A. C., J. A. Feulner, and A. Reilly (2001). Evidence for functionally significant polymorphism of human glutamate cysteine ligase catalytic subunit: association with glutathione levels and drug resistance in the national cancer institute tumor cell line panel. *Toxicological Sciences 61*, 218–223.