Semiparametric likelihood inference of rare disease associations with a genetic factor and independent continuous attribute in a case-control study

Jinko Graham, Brad McNeney and Ji-Hyung Shin
Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada

Abstract

In case-control studies, covariate information often is collected on a genetic factor and a continuous attribute such as age. In some instances, it is reasonable to assume 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. In this paper, we apply the method to investigate age-specific associations between type 1 diabetes and a variant of the glutamate-cysteine ligase catalytic subunit. The results are compared to those obtained from a standard logistic regression analysis, which does not make use of the independence assumption.

Keywords: Genetic associations; Likelihood inference; Case-control study.

Introduction

Many complex disorders with a genetic component share the feature of requiring certain key environmental exposures to manifest the disease state (e.g. Merikangas and Risch 2003). In fact, differences in key environmental exposures among study populations has been proposed as one possible explanation for the lack of replication of many genetic associations (Hoffjan et al. 2005). Within a population, an environmental exposure may increase disease risk in a genetically susceptible subgroup but have little or no effect outside this subgroup. 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. Consequently, there is increasing interest in investigating the effect of genes in conjunction with environmental exposures or non-genetic attributes, such as age and gender (e.g. Hunter 2005). This paper focusses on rare diseases (e.g. frequencies $\leq 1\%$), for which a case-control study design is standard. When the aim of a case-control study is to detect statistical interaction between genetic and nongenetic factors, adequate power is a concern, even for studies with relatively large sample sizes (Greenland 1983, Hwang et al. 1994). A useful rule-of-thumb is that the sample size required to detect interaction is at least four times that required to detect main effects of the same magnitude (Smith and Day 1984).

A number of approaches have been proposed to increase the efficiency of the statistical analysis when it is reasonable to assume that the genetic and non-genetic factors occur independently in the population. For categorical non-genetic covariates, Umbach and Weinberg (1997) developed maximum likelihood (ML) estimators of disease risk based on a log-linear model that enforces independence of the genetic and nongenetic risk factors under a rare-disease assumption. They achieved mildly enhanced precision for estimates of main effects and much enhanced precision for estimates of statistical interactions. Chatterjee and Carroll (2005) develop ML estimators of association parameters in a logistic regression model of disease penetrance by adopting a semiparametric framework that allows the distribution of nongenetic covariates to be completely nonparametric, under arbitrary disease frequencies. They show that the intercept parameter of the logistic regression model is theoretically identifiable from the retrospective case-control likelihood under independence of the genetic and nongenetic risk factors. However, for rare diseases, they note that estimation of this intercept term is problematic, unless the marginal probability of disease is known.

We develop a complementary ML approach for case-control studies of rare diseases which circumvents the need to estimate the intercept parameter in the logistic regression model. Our semi-parametric approach, like that of Chatterjee and Carroll, is applicable to continuous non-genetic covariates and may be viewed as an extension to the work of Umbach and Weinberg (1997). In order to reduce the potential bias arising from dependence between genetic and nongenetic covariates (Albert et al. 2001), we also discuss a strategy for relaxing the independence assumption. Finally, we present results from a limited simulation study investigating: statistical efficiency of the proposed method relative to ordinary logistic regression, robustness of the proposed method to the assumption of independence and the reduction in bias achieved by relaxing the independence assumption

Methods

Maximum likelihood estimation

Throughout, we use the notation "pr" to denote densities or mass functions as appropriate. A logistic model for disease status given covariates may be written as

$$\log\left[\frac{\mathrm{pr}(D=1 \mid x)}{\mathrm{pr}(D=0 \mid x)}\right] = \beta_0 + x\beta$$

where D is a binary disease indicator with value 1 indicating disease and value 0 indicating non-disease; β_0 is an intercept term; β is a column-vector of regression parameters;

and x is a row-vector of covariates. However, under case-control sampling, we sample covariates given disease status. Covariates are derived from risk factors. The risk factors we consider are categorical genetic variables G and possibly continuous non-genetic attributes A. Let $\mathcal{G} = \{g^0, \ldots, g^K\}$ denote the set of possible genetic categories and hence values of G. Let $\mathcal{A}$ denote the possible values of A. Let $x(g, a)$ denote the covariate vector corresponding to $(G = g, A = a)$ in the logistic regression model. From equation (6) of Prentice and Pyke (1979),

$$\begin{aligned} &\mathrm{pr}(G = g, A = a \mid D = i) \\ &= c_i(\xi, \beta) \exp\left(\xi(g, a) + ix(g, a)\beta\right), \end{aligned}$$

$$\text{where} \quad \xi(g, a) = \log\left[\frac{\mathrm{pr}(G = g, A = a \mid D = 0)}{\mathrm{pr}(G = g^0, A = a^0 \mid D = 0)}\right]$$

for baseline values g^0 and a^0 of the genetic and non-genetic risk factors, and where $c_i(\xi, \beta)$ is a normalizing constant. We further define

$$\begin{aligned} \delta_0 \equiv \delta_0(\xi, \beta) &= \log(c_0(\xi, \beta) n_0/n) \quad \text{and} \\ \delta \equiv \delta(\xi, \beta) &= \log(c_1(\xi, \beta) n_1/n) - \delta_0 \end{aligned} \tag{1}$$

so that

$$\begin{aligned} &\mathrm{pr}(G = g, A = a \mid D = i) \\ &= \frac{n}{n_i} \exp(\delta_0 + i\delta + \xi(g, a) + ix(g, a)\beta), \end{aligned}$$

where n_0 is the number of controls, n_1 is the number of cases and $n = n_0 + n_1$.

We start by assuming independence of G and A in controls, which for a rare disease is approximately the same as independence in the population. The independence assumption implies that

$$\xi(g, a) = \gamma(g) + \alpha(a) \tag{2}$$

where

$$\begin{aligned} \gamma(g) &= \log\left[\frac{\mathrm{pr}(G = g \mid D = 0)}{\mathrm{pr}(G = g^0 \mid D = 0)}\right] \quad \text{and} \\ \alpha(a) &= \log\left[\frac{\mathrm{pr}(A = a \mid D = 0)}{\mathrm{pr}(A = a^0 \mid D = 0)}\right]. \end{aligned}$$

Hence

$$\mathrm{pr}(G = g, A = a \mid D = i) = \frac{n}{n_i} \exp(\delta_0 + i\delta + \gamma(g) + \alpha(a) + ix(g, a)\beta), \tag{3}$$

where δ_0 and δ may now be viewed as functions of γ, α and β. When one or more of the non-genetic attributes are continuous, $\alpha(\cdot)$ is an infinite-dimensional nuisance parameter. $\gamma(\cdot)$ is a function whose domain is $\mathcal{G}$ and so can be viewed as a finite dimensional parameter. Index individuals within disease category i $(i = 0, 1)$ by $j = 1, \ldots, n_i$. Then the likelihood is proportional to

$$L(\gamma, \alpha, \beta) = \prod_{i=0}^{1} \prod_{j=1}^{n_i} \exp\left(\delta_0 + i\delta + \gamma(g_{ij}) + \alpha(a_{ij}) + ix(g_{ij}, a_{ij})\beta\right) \tag{4}$$

Maximization of (4) is complicated by the infinite-dimensional parameter $\alpha(\cdot)$, and by the dependence of δ_0 and δ on γ, α and β through the constants of integration $c_i(\xi, \beta)$ as specified in equations (1) and (2). Using arguments similar to Prentice and Pyke (1979), we show that there is a reparametrization of (4) that eliminates δ_0 from the likelihood and which allows us to treat δ as a free parameter when maximizing the likelihood. Furthermore, maximization of the reparametrized likelihood with respect to its infinite dimensional parameter turns out to be straightforward. This reparametrization is derived in Appendix A, which shows that $L(\gamma, \alpha, \beta)$ can be re-expressed as

$$\begin{aligned} L(\gamma, p_v^A, \beta) &= \left[\prod_{i=0}^{1} \prod_{j=1}^{n_i} \frac{\exp\left(i\delta + \gamma(g_{ij}) + ix(g_{ij}, a_{ij})\beta\right)}{\sum_{l=0}^{1} \sum_{g \in \mathcal{G}} \exp\left(l\delta + \gamma(g) + lx(g, a_{ij})\beta\right)}\right] \\ &\quad \times \left[\prod_{i=0}^{1} \prod_{j=1}^{n_i} p_v^A(a_{ij})\right] \\ &\equiv L_1(\gamma, p_v^A, \beta) \times L_2(p_v^A) \end{aligned}$$

where δ may be defined as the solution to equation (A-4)

$$1 = \frac{n}{n_1} \int_{\mathcal{A}} \sum_{g \in \mathcal{G}} \frac{\exp\left(\delta + \gamma(g) + x(g, a)\beta\right)}{\sum_{l=0}^{1} \sum_{g' \in \mathcal{G}} \exp\left(l\delta + \gamma(g) + lx(g, a)\beta\right)} p_v^A(a)\, da,$$

and p_v^A is the marginal distribution of A under a variant sampling scheme (VSS), discussed in more detail in Appendix A, in which a total of n subjects are sampled. Under this variant sampling scheme, a case is sampled with probability n_1/n and a control with probability n_0/n. In the reparametrized likelihood, $p_v^A(\cdot)$ replaces $\alpha(\cdot)$ as the infinite dimensional parameter. At first glance the reparametrized likelihood appears to have no advantage over the original likelihood (4), since δ still depends on the model parameters indirectly through the set of constraints (A-4), including the infinite dimensional parameter $p_v^A(\cdot)$. However, we now show that if we treat δ as a free parameter, maximization with respect to the infinite dimensional parameter p_v^A is straightforward, and the maximizer of the resulting over-parametrized likelihood satisfies the constraint that defines δ. It follows that the unconstrained maximizer gives MLEs of the odds-ratio parameters of interest.

Maximization of an overparametrized likelihood

Write $\tilde{L}_1(\delta, \gamma, \beta)$ for the expression L_1 when δ is considered as a free parameter:

$$\tilde{L}_1(\delta, \gamma, \beta) = \prod_{i=0}^{1} \prod_{j=1}^{n_i} \frac{\exp\left(i\delta + \gamma(g_{ij}) + ix(g_{ij}, a_{ij})\beta\right)}{\sum_{l=0}^{1} \sum_{g \in \mathcal{G}} \exp\left(l\delta + \gamma(g) + lx(g, a_{ij})\beta\right)} \tag{5}$$

The overparametrized likelihood is $\tilde{L}(\delta, \gamma, p_v^A, \beta) \equiv \tilde{L}_1(\delta, \gamma, \beta) \times L_2(p_v^A)$. The factorization implies that the maximizer of $\tilde{L}$ with respect to p_v^A would be obtained by maximizing L_2 alone and that the maximizer of $\tilde{L}$ with respect to (δ, γ, β) would be obtained by maximizing $\tilde{L}_1$ alone. The term $L_2(p_v^A)$ is a marginal likelihood for a non-parametric distribution function p_v^A. Hence the maximizer of $L_2(p_v^A)$ is the empirical distribution of A from the case-control sample. Since δ, γ and β are all finite-dimensional, the maximizer of $\tilde{L}_1(\delta, \gamma, \beta)$ is obtained in the usual way, by taking derivatives of $\tilde{L}_1$ and setting the resulting equations equal to zero. Taking derivatives of $\tilde{L}_1$ is simplified by writing $i\delta + \gamma(g) + ix(g, a)\beta$ as a linear combination of

a single parameter vector θ. That is, we wish to define a row vector $\Lambda(i,g,a)$ and a column vector θ of parameters such that $i\delta + \gamma(g) + ix(g,a)\beta = \Lambda(i,g,a)\theta$. Towards this re-expression, write the function $\gamma(\cdot)$ as a linear combination of K parameters, where $K+1 = |\mathcal{G}|$. Let $z(g^k)$ be an indicator row vector of K elements for the k^{th} genetic category ($k = 1,\ldots,K$) or a vector of all zeros for the baseline genetic category when $k = 0$. Let $\vec{\gamma}$ denote the column vector $(\gamma(g^1),\ldots,\gamma(g^K))^T$. Then $z(g)\vec{\gamma}$ takes on value $\gamma(g)$ for $g \in (g^1,\ldots,g^K)$ and value 0 when $g = g^0$. If we define $\Lambda(i,g,a) = [i, z(g), ix(g,a)]$ and $\theta = (\delta, \vec{\gamma}^T, \beta^T)^T$, we have as desired that

$$i\delta + \gamma(g) + ix(g,a)\beta = \Lambda(i,g,a)\theta.$$

When genetic categories are genotypes, modelling of genotype probabilities is also possible, in which case $\vec{\gamma}$ is specified by a parameter vector ω with fewer than K elements, where $K+1$ is the number of genotypes. For example, for a genetic locus with $M+1$ alleles and with Hardy-Weinberg proportions in controls, the $K+1 = (M+1)(M+2)/2$ genotype frequencies in controls can be expressed in terms of M allele frequencies. In this case, it is possible to define $z(g)$ and ω so that $\gamma(g) = \kappa(g) + z(g)\omega$, where $\kappa(g)$ is a known constant that depends on the value of the genotype g, and ω is a vector whose ith elementis the generalized logit of the control frequency of the ith allele relative to a baseline allele, as follows. Let the alleles be $m_0, m_1, \ldots, m_M$. Take m_0 as a baseline allele, and m_0/m_0 as a baseline genotype. Let

$$\omega_i = \log\left[\frac{\mathrm{pr}(m_i \mid D=0)}{\mathrm{pr}(m_0 \mid D=0)}\right].$$

Let $H(g)$ be 1 if g is a homozygous genotype and 2 if heterozygous. Then, for a particular genotype $g = m_i/m_j$, we have

$$\begin{aligned}\gamma(m_i/m_j) &= \log\left[\frac{\mathrm{pr}(G = m_i/m_j \mid D=0)}{\mathrm{pr}(G = g^0 \mid D=0)}\right] \\ &= \log\left[\frac{H(g)\mathrm{pr}(m_i \mid D=0)\mathrm{pr}(m_j \mid D=0)}{\mathrm{pr}(m_0 \mid D=0)^2}\right] \\ &= \kappa(g) + \omega_i + \omega_j\end{aligned}$$

where $\kappa(g) = \log H(g)$. Redefine $z(g)$ to be a row vector of length M with ith element equal to the number of copies of allele i, $i = 1,\ldots,M$. Then $\gamma(g) = \kappa(g) + z(g)\omega$, where $\omega = (\omega_1,\ldots,\omega_M)^T$. With this alternate definition of $z(g)$, let $\Lambda(i,g,a) = (i\delta, z(g), ix(g,a))$ as before and define $\theta = (\delta, \omega^T, \beta^T)^T$. Then we obtain

$$i\delta + \gamma(g) + ix(g,a)\beta = \kappa(g) + \Lambda(i,g,a)\theta.$$

It is easily verified that inclusion of the "offset" term $\kappa(g)$ does not change the final expressions for the estimating equations or the asymptotic variance calculations in the developments below. Thus, assuming Hardy-Weinberg proportions in controls involves changing the definition of $z(g)$ in Λ and the definition of the parameter vector θ, but the estimating equations for the maximum-likelihood estimator of the parameters of interest and the expression for its asymptotic variance are the same functions of Λ and θ as before. For simplicity of exposition, however, we proceed as though the control genotype frequencies are not modelled and the values of G are arbitrary genetic categories. Re-write $\tilde{L}_1(\delta,\gamma,\beta) = \tilde{L}_1(\theta)$ as

$$\left[\prod_{i=0}^{1}\prod_{j=1}^{n_i}\frac{\exp(\Lambda(i,g_{ij},a_{ij})\theta)}{\sum_{l=0}^{1}\sum_{g\in\mathcal{G}}\exp(\Lambda(l,g,a_{ij})\theta)}\right] \tag{6}$$

Let

$$\begin{aligned}\tilde{l}_1(\theta) &= \log\tilde{L}_1(\theta) \\ &= \sum_{i=0}^{1}\sum_{j=1}^{n_i}\left\{\Lambda(i,g_{ij},a_{ij})\theta - \log\sum_{l=0}^{1}\sum_{g\in\mathcal{G}}\exp(\Lambda(l,g,a_{ij})\theta)\right\}\end{aligned}$$

Then the maximizer $\hat{\theta}$ of $\tilde{L}_1$ is the solution to the estimating equations

$$\begin{aligned}0 = \frac{\partial\tilde{l}_1}{\partial\theta} &= \sum_{i=0}^{1}\sum_{j=1}^{n_i}\left\{\Lambda(i,g_{ij},a_{ij})^T - \sum_{l=0}^{1}\sum_{g\in\mathcal{G}}\Lambda(l,g,a_{ij})^T\right. \\ &\qquad \left.\frac{\exp(\Lambda(l,g,a_{ij})\theta)}{\sum_{l'=0}^{1}\sum_{g'\in\mathcal{G}}\exp(\Lambda(l',g',a_{ij})\theta)}\right\}\end{aligned}$$

Or, defining

$$p_{ig}(a;\theta) = \frac{\exp(\Lambda(i,g,a)\theta)}{\sum_{l'=0}^{1}\sum_{g'\in\mathcal{G}}\exp(\Lambda(l',g',a)\theta)}, \tag{7}$$

we obtain the estimating equations

$$0 = \left.\frac{\partial\tilde{l}_1}{\partial\theta}\right|_{\hat{\theta}} = \sum_{i=0}^{1}\sum_{j=1}^{n_i}\left[\Lambda(i,g_{ij},a_{ij})^T - \sum_{l=0}^{1}\sum_{g\in\mathcal{G}}\Lambda(l,g,a_{ij})^T p_{lg}(a_{ij};\hat{\theta})\right].$$

In Appendix C, we verify that the unconstrained maximizers $\hat{\theta}$ and $\widehat{p}_v^A$ satisfy the constraints. Thus $(\hat{\theta}, \widehat{p}_v^A)$ are the ML estimators.

Asymptotic distribution of estimators

Let θ^0 denote the "true" value of θ. Recall $\theta = (\delta, \vec{\gamma}^T, \beta^T)^T$, but that $\theta_1 = \delta$ is a nuisance parameter, while $\theta_2 = (\vec{\gamma}^T, \beta^T)^T$ are the parameters of interest. A Taylor series expansion shows that

$$\sqrt{n}(\hat{\theta} - \theta^0) = I(\theta^*)S(\theta^0)$$

where θ^* is between θ^0 and $\hat{\theta}$ and

$$I(\theta^*) = -\frac{1}{n}\left.\frac{\partial^2\tilde{l}_1}{\partial\theta\partial\theta^T}\right|_{\theta^*}, \qquad S(\theta^0) = \left.\frac{1}{\sqrt{n}}\frac{\partial\tilde{l}_1}{\partial\theta}\right|_{\theta^0}.$$

Under regularity conditions $I(\theta^*)$ converges in probability to $G(\theta^0) = E(I(\theta^0))$, which can be consistently estimated by $I(\hat{\theta})$. We will show $E(S(\theta^0)) = 0$, where the expectation is taken under the true probability model at θ^0. Therefore, $S(\theta^0) \xrightarrow{d} N(0, \Sigma(\theta^0))$, where $\Sigma(\theta^0) = V(S(\theta^0))$, and the variance is taken under the true probability model at θ^0 (Prentice and Pyke 1979). Then, by Slutsky's theorem,

$$\sqrt{n}(\hat{\theta}_n - \theta^0) = I(\theta^*)S(\theta^0) \xrightarrow{d} N(0, G(\theta^0)^{-1}\Sigma(\theta^0)G(\theta^0)^{-1}).$$

Let $G = G(\theta^0)$ and $\Sigma = \Sigma(\theta^0)$ so that the asymptotic variance of $\hat{\theta}$ is $G^{-1}\Sigma G^{-1}$. We will show that the asymptotic variance matrix $[G^{-1}\Sigma G^{-1}]_{22}$ corresponding to the parameters of interest $\hat{\theta}_2 = (\widehat{\vec{\gamma}}^T, \hat{\beta}^T)^T$ is equal to $(G^{-1})_{22}$, which can be consistently estimated by $[I(\hat{\theta})^{-1}]_{22}$

Establishing $E\left(S(\theta^0)\right) = 0$

From the definition of $S(\theta^0)$,

$$E(S(\theta^0)) = 0 \iff E\left(\left.\frac{\partial \tilde{l}_1}{\partial \theta}\right|_{\theta^0}\right) = 0$$

To simplify notation, let $p_{ig}(a) = p_{ig}(a;\theta^0)$. Then, from the calculations leading up to the estimating equations, we have

$$\begin{aligned}
&E\left(\left.\frac{\partial \tilde{l}_1}{\partial \theta}\right|_{\theta^0}\right) \\
&= E\left\{\sum_{i=0}^{1}\sum_{j=1}^{n_i}\left[\Lambda(i,G_{ij},A_{ij})^T - \sum_{l=0}^{1}\sum_{g\in\mathcal{G}}\Lambda(l,g,A_{ij})^T p_{lg}(A_{ij})\right]\right\} \\
&= \sum_{i=0}^{1} E\left\{\sum_{j=1}^{n_i}\left[\Lambda(i,G_{ij},A_{ij})^T - \sum_{l=0}^{1}\sum_{g\in\mathcal{G}}\Lambda(l,g,A_{ij})^T p_{lg}(A_{ij})\right]\right\} \\
&= \sum_{i=0}^{1} n_i\, E\left[\Lambda(i,G_{i1},A_{i1})^T - \sum_{l=0}^{1}\sum_{g\in\mathcal{G}}\Lambda(l,g,A_{i1})^T p_{lg}(A_{i1})\right]
\end{aligned}$$

From equation (A-2),

$$\text{pr}(G=g, A=a \mid D=i) = \frac{n}{n_i}\, p_{ig}(a)\, p_v^A(a).$$

Hence

$$\begin{aligned}
E\left(\left.\frac{\partial \tilde{l}_1}{\partial \theta}\right|_{\theta^0}\right) &= \sum_{i=0}^{1} n_i \int_{\mathcal{A}} \sum_{g\in G}\Bigg[\Lambda(i,g,a)^T - \sum_{l=0}^{1}\sum_{g'\in G}\Lambda(l,g',a)^T \\
&\qquad p_{lg'}(a)\Bigg]\frac{n}{n_i}\, p_{ig}(a)\, p_v^A(a)\, da \\
&= n\sum_{i=0}^{1}\int_{\mathcal{A}}\Bigg[\sum_{g\in G}\Lambda(i,g,a)^T p_{ig}(a) - \sum_{g\in G}\sum_{l=0}^{1}\sum_{g'\in G}\Lambda(l,g',a)^T \\
&\qquad p_{lg'}(a)\, p_{ig}(a)\Bigg] p_v^A(a)\, da \\
&= n\sum_{i=0}^{1}\int_{\mathcal{A}}\Bigg[\sum_{g\in G}\Lambda(i,g,a)^T p_{ig}(a) - \sum_{l=0}^{1}\sum_{g'\in G}\Lambda(l,g',a)^T \\
&\qquad p_{lg'}(a)\sum_{g\in G} p_{ig}(a)\Bigg] p_v^A(a)\, da \\
&= n\int_{\mathcal{A}}\Bigg\{\sum_{i=0}^{1}\sum_{g\in G}\Lambda(i,g,a)^T p_{ig}(a) - \sum_{i=0}^{1}\sum_{l=0}^{1}\sum_{g'\in G} \\
&\qquad \Lambda(l,g',a)^T p_{lg'}(a)\sum_{g\in G} p_{ig}(a)\Bigg\} p_v^A(a)\, da \\
&= n\int_{\mathcal{A}}\Bigg\{\sum_{i=0}^{1}\sum_{g\in G}\Lambda(i,g,a)^T p_{ig}(a) \; - \sum_{l=0}^{1}\sum_{g'\in G} \\
&\qquad \Lambda(l,g',a)^T p_{lg'}(a)\sum_{i=0}^{1}\sum_{g\in G} p_{ig}(a)\Bigg\} p_v^A(a)\, da \\
&= n\int_{\mathcal{A}} 0\; p_v^A(a)\, da \;\; = \;\; 0
\end{aligned}$$

as claimed since $\sum_{i=0}^{1}\sum_{g\in G} p_{ig}(a) = 1$.

Expression for $G(\theta^0)$

It can be shown that

$$\left.\frac{\partial^2 \tilde{l}_1}{\partial\theta\partial\theta^T}\right|_{\theta^0} = -\sum_{i=0}^{1}\sum_{j=1}^{n_i}\Bigg\{\sum_{l=0}^{1}\sum_{g\in\mathcal{G}}\Bigg[\Lambda(l,g,a_{ij})^T - \sum_{l'=0}^{1}\sum_{g'\in\mathcal{G}}\Lambda(l',g',a_{ij})^T p_{l'g'}(a_{ij})\Bigg]^{\otimes 2} p_{lg}(a_{ij})\Bigg\},$$

where $v^{\otimes 2} = vv^T$ and $p_{lg}(a)$ is as defined above. Hence,

$$\begin{aligned}
G(\theta^0) &= E\left(-\frac{1}{n}\left.\frac{\partial^2 \tilde{l}_1}{\partial\theta\partial\theta^T}\right|_{\theta^0}\right) \\
&= \frac{1}{n}E\sum_i\sum_j\Bigg\{\sum_{l=0}^{1}\sum_{g\in\mathcal{G}}\Bigg[\Lambda(l,g,A_{ij})^T - \sum_{l'=0}^{1}\sum_{g'\in\mathcal{G}}\Lambda(l',g',A_{ij})^T p_{l'g'}(A_{ij})\Bigg]^{\otimes 2} p_{lg}(A_{ij})\Bigg\} \\
&= \frac{1}{n}\sum_i n_i E\Bigg\{\sum_{l=0}^{1}\sum_{g\in\mathcal{G}}\Bigg[\Lambda(l,g,A_{i1})^T - \sum_{l'=0}^{1}\sum_{g'\in\mathcal{G}}\Lambda(l',g',A_{i1})^T p_{l'g'}(A_{i1})\Bigg]^{\otimes 2} p_{lg}(A_{i1})\Bigg\} \\
&= \frac{1}{n}\sum_i n_i \int_{\mathcal{A}}\sum_{g\in\mathcal{G}}\Bigg\{\sum_{l=0}^{1}\sum_{g'\in\mathcal{G}}\Bigg[\Lambda(l,g',a)^T - \sum_{l'=0}^{1}\sum_{g''\in\mathcal{G}}\Lambda(l',g'',a)^T p_{l'g''}(a)\Bigg]^{\otimes 2} \\
&\qquad p_{lg'}(a)\Bigg\}\frac{n}{n_i} p_{ig}(a)\, p_v^A(a)\, da
\end{aligned}$$

after substituting

$$\text{pr}(G=g, A=a \mid D=i) = \frac{n}{n_i}\, p_{ig}(a)\, p_v^A(a)$$

from equation (A-2) in the last line. In the above expression for $G(\theta^0)$, nothing in the curly brackets depends on the indices i and g, and so we can move the corresponding outer sums past this term. After appropriate cancelling of the n and n_i terms, this gives

$$\begin{aligned}
&G(\theta^0) \\
&= \int_{\mathcal{A}}\Bigg\{\sum_{l=0}^{1}\sum_{g'\in\mathcal{G}}\Bigg[\Lambda(l,g',a)^T - \sum_{l'=0}^{1}\sum_{g''\in\mathcal{G}}\Lambda(l',g'',a)^T p_{l'g''}(a)\Bigg]^{\otimes 2} \\
&\qquad p_{lg'}(a)\Bigg\}\sum_{i=0}^{1}\sum_{g\in\mathcal{G}} p_{ig}(a)\, p_v^A(a)\, da \\
&= \int_{\mathcal{A}}\sum_{i=0}^{1}\sum_{g\in\mathcal{G}}\Bigg[\Lambda(i,g,a)^T - \sum_{l=0}^{1}\sum_{g'\in\mathcal{G}}\Lambda(l,g',a)^T p_{lg'}(a)\Bigg]^{\otimes 2} p_{ig}(a)\, p_v^A(a)\, da
\end{aligned}$$

since $\sum_{i=0}^{1}\sum_{g\in\mathcal{G}} p_{ig}(a) = 1$.

Expression for $\Sigma(\theta^0)$

We have

$$\Sigma(\theta^0) = V(S(\theta^0)) = \frac{1}{n}\sum_{i=0}^{1}\sum_{j=1}^{n_i} V\left(\left.\frac{\partial \tilde{l}_{1ij}}{\partial\theta}\right|_{\theta^0}\right),$$

where the last equality follows by the independence of subjects and where

$$\left.\frac{\partial \tilde{l}_{1ij}}{\partial\theta}\right|_{\theta^0} = \Lambda(i,G_{ij},A_{ij})^T - \sum_{l=0}^{1}\sum_{g\in G}\Lambda(l,g,A_{ij})^T p_{lg}(A_{ij})$$

is the contribution of the jth subject in the ith disease category to $\frac{\partial \tilde{l}_1}{\partial \theta}\Big|_{\theta^0}$. Within disease categories, the subjects are iid and so

$$
\begin{aligned}
&\Sigma(\theta^0) \\
&= \frac{1}{n}\sum_{i=0}^{1} n_i\, V\left(\Lambda(i,G_{i1},A_{i1})^T - \sum_{l=0}^{1}\sum_{g\in G}\Lambda(l,g,A_{i1})^T p_{lg}(A_{i1})\right) \\
&= \sum_{i=0}^{1}\frac{n_i}{n}\Bigg\{ E\Bigg(\big[\,\Lambda(i,G_{i1},A_{i1})^T - \sum_{l=0}^{1}\sum_{g\in G}\Lambda(l,g,A_{i1})^T p_{lg}(A_{i1})\,\big]^{\otimes 2}\Bigg) \\
&\quad - \Bigg(E\big[\,\Lambda(i,G_{i1},A_{i1})^T - \sum_{l=0}^{1}\sum_{g\in G}\Lambda(l,g,A_{i1})^T p_{lg}(A_{i1})\,\big]\Bigg)^{\otimes 2}\Bigg\} \\
&= \sum_{i=0}^{1}\frac{n_i}{n}\Bigg\{ \int_{\mathcal{A}}\sum_{g\in G}\big[\,\Lambda(i,g,a)^T - \sum_{l=0}^{1}\sum_{g'\in G}\Lambda(l,g',a)^T p_{lg'}(a)\,\big]^{\otimes 2} \\
&\quad \frac{n}{n_i}\,p_{ig}(a)\,p_v^{\mathcal{A}}(a)\,da \;-\Bigg(\int_{\mathcal{A}}\sum_{g\in G}\big[\,\Lambda(i,g,a)^T - \sum_{l=0}^{1}\sum_{g'\in G} \\
&\quad \Lambda(l,g',a)^T p_{lg'}(a)\,\big]\frac{n}{n_i}\,p_{ig}(a)\,p_v^{\mathcal{A}}(a)\,da\Bigg)^{\otimes 2}\Bigg\} \\
&= \sum_{i=0}^{1}\int_{\mathcal{A}}\sum_{g\in G}\big[\,\Lambda(i,g,a)^T - \sum_{l=0}^{1}\sum_{g'\in G}\Lambda(l,g',a)^T p_{lg'}(a)\,\big]^{\otimes 2} \\
&\quad p_{ig}(a)\,p_v^{\mathcal{A}}(a)\,da \;-\sum_{i=0}^{1}\frac{n}{n_i}\Bigg(\int_{\mathcal{A}}\sum_{g\in G}\big[\,\Lambda(i,g,a)^T - \\
&\quad \sum_{l=0}^{1}\sum_{g'\in G}\Lambda(l,g',a)^T p_{lg'}(a)\,\big]\,p_{ig}(a)\,p_v^{\mathcal{A}}(a)\,da\Bigg)^{\otimes 2} \\
&= G(\theta^0) \;-\; \sum_{i=0}^{1}\frac{n}{n_i}\Bigg(\int_{\mathcal{A}}\sum_{g\in G}\big[\,\Lambda(i,g,a)^T - \sum_{l=0}^{1}\sum_{g'\in G} \\
&\quad \Lambda(l,g',a)^T p_{lg'}(a)\,\big]\,p_{ig}(a)\,p_v^{\mathcal{A}}(a)\,da\Bigg)^{\otimes 2} \\
&= G(\theta^0) \;-\; \sum_{i=0}^{1}\frac{n}{n_i}\,B_i^{\otimes 2},
\end{aligned}
$$

where $B_i = \int_{\mathcal{A}}\sum_{g\in G}[\,\Lambda(i,g,a)^T - \sum_{l=0}^{1}\sum_{g'\in G}\Lambda(l,g',a)^T p_{lg'}(a)\,]\,p_{ig}(a)\,p_v^{\mathcal{A}}(a)\,da$

Simplified expression for $G^{-1}\Sigma G^{-1}$

Following Prentice and Pyke, we wish to show

$$\sum_{i=0}^{1}\frac{n}{n_i}B_i^{\otimes 2} = G\begin{bmatrix} X & 0\\ 0 & 0\end{bmatrix}G = \begin{bmatrix} G_{11}XG_{11} & G_{11}XG_{12}\\ G_{21}XG_{11} & G_{21}XG_{12}\end{bmatrix},$$

where

$$X = \sum_{i=0}^{1}\frac{n}{n_i} \quad \text{and} \quad \mathrm{G} = \begin{bmatrix} G_{11} & G_{12}\\ G_{21} & G_{12}\end{bmatrix} = \mathrm{G}(\theta^0).$$

X is a scalar quantity and G, the expectation of a normalized Hessian, is symmetric. Hence,

$$G\begin{bmatrix} X & 0\\ 0 & 0\end{bmatrix}G = X\begin{bmatrix} G_{11}G_{11} & G_{11}G_{12}\\ G_{21}G_{11} & G_{21}G_{12}\end{bmatrix} = X\begin{bmatrix} G_{11}\\ G_{21}\end{bmatrix}[G_{11}G_{12}] = X\begin{bmatrix} G_{11}\\ G_{21}\end{bmatrix}^{\otimes 2}.$$

Therefore, we wish to show

$$\sum_{i=0}^{1}\frac{n}{n_i}B_i^{\otimes 2} = X\begin{bmatrix} G_{11}\\ G_{21}\end{bmatrix}^{\otimes 2}.$$

Then, setting $\Sigma(\theta^0) = \Sigma$, we have

$$\Sigma = G - G\begin{bmatrix} X & 0\\ 0 & 0\end{bmatrix}G,$$

so that

$$
\begin{aligned}
G^{-1}\Sigma G^{-1} &= G^{-1}(G - G\begin{bmatrix} X & 0\\ 0 & 0\end{bmatrix}G)G^{-1} \\
&= G^{-1}GG^{-1} - G^{-1}G\begin{bmatrix} X & 0\\ 0 & 0\end{bmatrix}GG^{-1} \\
&= G^{-1} - \begin{bmatrix} X & 0\\ 0 & 0\end{bmatrix}
\end{aligned}
$$

From this equality, it follows that the asymptotic variance matrix $[G^{-1}\Sigma G^{-1}]_{22}$ for the parameter estimators $\hat{\theta}_2 = (\hat{\widetilde{\gamma}}^T, \hat{\beta}^T)^T$ of interest is equal to $(G^{-1})_{22}$, which is consistently estimated by $[I(\hat{\theta})^{-1}]_{22}$.

Here, we will first show that $B_1 = -B_0$. In deriving $E(S(\theta^0)) = 0$, we established that

$$\sum_{i=0}^{1} B_i = \frac{1}{n}E\left(\frac{\partial \tilde{l}_1}{\partial \theta}\Big|_{\theta^0}\right) = 0.$$

Thus, $B_1 = -B_0$, so that

$$\sum_{i=0}^{1}\frac{n}{n_i}B_i^{\otimes 2} = \sum_{i=0}^{1}\frac{n}{n_i}B_1^{\otimes 2} = B_1^{\otimes 2}\sum_{i=0}^{1}\frac{n}{n_i} = XB_1^{\otimes 2}.$$

Next, we show that small

$$B_1 = \begin{bmatrix} G_{11}\\ G_{21}\end{bmatrix},$$

the first column of G. For any column vector v, the first column of $v^{\otimes 2}$ is

$$[v^{\otimes 2}]_{\cdot 1} = v_1 v,$$

where v_1 is the first element of v. Taking

$$v = \Lambda(l,g,a)^T - \sum_{l'=0}^{1}\sum_{g'\in\mathcal{G}}\Lambda(l',g',a)^T p_{l'g'}(a),$$

we get that the first element is

$$v_1 = l - \sum_{l'=0}^{1}\sum_{g'\in\mathcal{G}} l' p_{l'g'}(a)$$

because the row vector $\Lambda(l,g,a) = [\; l,\; z(g),\; lx(g,a)\;]$. It

follows that the first column of G is

$$\begin{aligned}
\begin{bmatrix} G_{11} \\ G_{21} \end{bmatrix} =& \int_{\mathcal{A}} \sum_{l=0}^{1} \sum_{g\in\mathcal{G}} \left(\left[l - \sum_{l'=0}^{1} \sum_{g'\in\mathcal{G}} l' p_{l'g'}(a) \right] \left[\Lambda(l,g,a)^T - \right. \right. \\
& \left. \left. \sum_{l'=0}^{1} \sum_{g'\in\mathcal{G}} \Lambda(l',g',a)^T p_{l'g'}(a) \right] p_{lg}(a) \right) p_v^A(a)\, da \\
=& \int_{\mathcal{A}} \sum_{l=0}^{1} \sum_{g\in\mathcal{G}} \left(\left[l - \sum_{g'\in\mathcal{G}} p_{1g'}(a) \right] \left[\Lambda(l,g,a)^T - \sum_{l'=0}^{1} \sum_{g'\in\mathcal{G}} \right. \right. \\
& \left. \left. \Lambda(l',g',a)^T p_{l'g'}(a) \right] p_{lg}(a) \right) p_v^A(a)\, da \\
=& \int_{\mathcal{A}} \left\{ \left[-\sum_{g'\in\mathcal{G}} p_{1g'}(a) \right] \sum_{g\in\mathcal{G}} \left[\Lambda(0,g,a)^T - \sum_{l=0}^{1} \sum_{g'\in\mathcal{G}} \right. \right. \\
& \Lambda(l,g',a)^T p_{lg'}(a) \Big] p_{0g}(a) + \left[1 - \sum_{g'\in\mathcal{G}} p_{1g'}(a) \right] \sum_{g\in\mathcal{G}} \\
& \left. \left[\Lambda(1,g,a)^T - \sum_{l=0}^{1} \sum_{g'\in\mathcal{G}} \Lambda(l,g',a)^T p_{lg'}(a) \right] p_{1g}(a) \right\} p_v^A(a)\, da \\
=& \int_{\mathcal{A}} \left\{ \left[-\sum_{g'\in\mathcal{G}} p_{1g'}(a) \right] \sum_{l=0}^{1} \sum_{g\in\mathcal{G}} \left[\Lambda(l,g,a)^T - \sum_{l'=0}^{1} \sum_{g'\in\mathcal{G}} \right. \right. \\
& \left. \Lambda(l',g',a)^T p_{l'g'}(a) \right] p_{lg}(a) + \sum_{g\in\mathcal{G}} \left[\Lambda(1,g,a)^T - \sum_{l'=0}^{1} \sum_{g'\in\mathcal{G}} \right. \\
& \left. \left. \Lambda(l',g',a)^T p_{l'g'}(a) \right] p_{1g}(a) \right\} p_v^A(a)\, da
\end{aligned}$$

Within the integral, the expression in the first summand

$$\sum_{l=0}^{1} \sum_{g\in\mathcal{G}} \left[\Lambda(l,g,a)^T - \sum_{l'=0}^{1} \sum_{g'\in\mathcal{G}} \Lambda(l',g',a)^T p_{l'g'}(a) \right] p_{lg}(a) = 0$$

because $\sum_{l=0}^{1} \sum_{g\in\mathcal{G}} p_{lg}(a) = 1$. Hence, as required,

$$\begin{aligned}
\begin{bmatrix} G_{11} \\ G_{21} \end{bmatrix} &= \int_{\mathcal{A}} \sum_{g\in\mathcal{G}} \left[\Lambda(1,g,a)^T - \sum_{l'=0}^{1} \sum_{g'\in\mathcal{G}} \right. \\
&\qquad \left. \Lambda(l',g',a)^T p_{l'g'}(a) \right] p_{1g}(a)\, p_v^A(a)\, da \\
&= B_1.
\end{aligned}$$

Extension to Hardy-Weinberg proportions in controls

In the special case of Hardy-Weinberg proportions in controls,

$$\tilde{L}_1 = \prod_{i=0}^{1} \prod_{j=1}^{n_i} p_{ig_{ij}}(a_{ij};\theta),$$

where $p_{ig}(a;\theta)$ is re-defined as

$$p_{ig}(a;\theta) = \frac{\exp(\kappa(g) + \Lambda(i,g,a)\theta)}{\sum_{l'=0}^{1} \sum_{g'\in\mathcal{G}} \exp(\kappa(g') + \Lambda(l',g',a)\theta)},$$

and the offset term $\kappa(g)$ is a known constant. With this alternate definition of $p_{ig}(a;\theta)$, the calculations of the estimating equations, hessian $I(\theta)$, expectations $G(\theta^0)$ and $\Sigma(\theta^0)$, and variance-covariance matrix $G^{-1}\Sigma G^{-1}$ all remain exactly the same. To see this, the key points are as follows. First, the derivative of $\kappa(g) + \Lambda(i,g,a)\theta$ with respect to θ is the same as the derivative of $\Lambda(i,g,a)\theta$ with respect to θ. Thus derivatives of $p_{ig}(a;\theta)$ with respect to θ are of the same form as before. Since $\tilde{L}_1$ is a product of terms of the form $p_{ig_{ij}}(a_{ij};\theta)$, we obtain the same score-like equations and the same hessian as before. Moreover, for $G(\theta^0)$ and $\Sigma(\theta^0)$, the expected values are taken with respect to the probability

$$\text{pr}(G = g, A = a \mid D = i) = \frac{n}{n_i}\, p_{ig}(a)\, p_v^A(a).$$

We use a new definition of $p_{ig}(a)$, but the expression for $\text{pr}(G = g, A = a \mid D = i)$ continues to be of the same form in $p_{ig}(a)$. Hence expressions for expectations stay of the same form as well. Similarly, nothing changes in the algebra to simplify the variance-covariance matrix $G^{-1}\Sigma G^{-1}$

Extension to allow dependence between G and A

The assumption that genetic susceptibility and environmental exposures or other attributes are independent in the population is a strong one. If an exposure or attribute, such as smoking or body-mass index, depends on a subject's behaviour, the independence assumption is questionable (Chatterjee and Carroll 2005). Albert et al. (2001) showed that methods that incorrectly impose independence can yield anticonservative inference about multiplicative interaction. Recall that in the development of the likelihood, the probability of covariates given disease status was described by a model that included the parameter

$$\xi(g,a) = \log\left[\frac{\text{pr}(G = g, A = a \mid D = 0)}{\text{pr}(G = g^0, A = a^0 \mid D = 0)} \right].$$

Independence of G and A in controls implies that

$$\xi(g,a) = \log\left[\frac{\text{pr}(G = g \mid D = 0)}{\text{pr}(G = g^0 \mid D = 0)} \right] + \log\left[\frac{\text{pr}(A = a \mid D = 0)}{\text{pr}(A = a^0 \mid D = 0)} \right].$$

To allow for dependence, write

$$\begin{aligned}
\xi(g,a) &= \log\left[\frac{\text{pr}(G = g \mid A = a, D = 0)}{\text{pr}(G = g^0 \mid A = a^0, D = 0)} \right] + \log\left[\frac{\text{pr}(A = a \mid D = 0)}{\text{pr}(A = a^0 \mid D = 0)} \right] \\
&= \log\left[\frac{\text{pr}(G = g \mid A = a, D = 0)}{\text{pr}(G = g^0 \mid A = a^0, D = 0)} \frac{\text{pr}(G = g^0 \mid A = a, D = 0)}{\text{pr}(G = g^0 \mid A = a, D = 0)} \right] + \\
&\qquad \log\left[\frac{\text{pr}(A = a \mid D = 0)}{\text{pr}(A = a^0 \mid D = 0)} \right] \\
&= \log\left[\frac{\text{pr}(G = g \mid A = a, D = 0)}{\text{pr}(G = g^0 \mid A = a, D = 0)} \frac{\text{pr}(G = g^0 \mid A = a, D = 0)}{\text{pr}(G = g^0 \mid A = a^0, D = 0)} \right] + \\
&\qquad \log\left[\frac{\text{pr}(A = a \mid D = 0)}{\text{pr}(A = a^0 \mid D = 0)} \right] \\
&= \log\left[\frac{\text{pr}(G = g \mid A = a, D = 0)}{\text{pr}(G = g^0 \mid A = a, D = 0)} \right] + \\
&\qquad \log\left[\frac{\text{pr}(G = g^0 \mid A = a, D = 0)}{\text{pr}(G = g^0 \mid A = a^0, D = 0)} \frac{\text{pr}(A = a \mid D = 0)}{\text{pr}(A = a^0 \mid D = 0)} \right] \\
&= \log\left[\frac{\text{pr}(G = g \mid A = a, D = 0)}{\text{pr}(G = g^0 \mid A = a, D = 0)} \right] + \log\left[\frac{\text{pr}(G = g^0, A = a \mid D = 0)}{\text{pr}(G = g^0, A = a^0 \mid D = 0)} \right] \\
&= \gamma_a(g) + \alpha_0(a)
\end{aligned}$$

where

$$\gamma_a(g) = \log\left[\frac{\text{pr}(G = g \mid A = a, D = 0)}{\text{pr}(G = g^0 \mid A = a, D = 0)} \right]$$

and

$$\alpha_0(a) = \log\left[\frac{\text{pr}(G = g^0, A = a \mid D = 0)}{\text{pr}(G = g^0, A = a^0 \mid D = 0)} \right].$$

Under independence of G and A in controls, $\gamma_a(g)$ and $\alpha_0(a)$ reduce to $\gamma(g)$ and $\alpha(a)$, respectively. Other than the substitutions of $\gamma_a(g)$ for $\gamma(g)$ and $\alpha_0(a)$ for $\alpha(a)$ in equation (4), the remaining development of the likelihood up to the overparametrized likelihood $\tilde{L}_1(\delta, \gamma_a, \beta)$ in equation (5) remains the same. In order to keep $\tilde{L}_1(\delta, \gamma_a, \beta)$ parametric, we introduce a parametric model for $\gamma_a(g)$. In particular, we propose the polychotomous regression

$$\gamma_a(g) = \log\left[\frac{\text{pr}(G = g \mid A = a, D = 0)}{\text{pr}(G = 0 \mid A = a, D = 0)}\right] = \nu_g + a\tau_g.$$

We now show that this polychotomous regression holds if the attribute A is a continuous, count or categorical variable that, in controls, has a conditional distribution given G from the exponential family. The mean but not the dispersion parameter of this conditional distribution may depend on the level of the genetic factor. In quantitative genetics, it is common to assume that a continuous trait has constant dispersion across genotypic groups.

Justification of the polychotomous regression model

Case 1: The attribute is a count or continuous variable

When A is a continuous or count variable, we suppose that the conditional density of A given $G = g$ in controls is from the exponential family, as defined in McCullagh and Nelder (1989; page 28):

$$\text{pr}(A = a \mid G = g, D = 0) = \exp\{[a\vartheta_g - b(\vartheta_g)]/\alpha(\phi) + c(a, \phi)\},$$

where the canonical parameter ϑ_g (which relates to the conditional mean of A) but not the dispersion parameter ϕ depends on the level g of G. The joint distribution of A and G in controls is thus

$$\text{pr}(A = a, G = g \mid D = 0) = \exp\{[a\vartheta_g - b(\vartheta_g)]/\alpha(\phi) + c(a, \phi)\}\,\text{pr}(G = g \mid D = 0).$$

The conditional mass function for G given $A = a$ and $D = 0$ is

$$\text{pr}(G = g \mid A = a, D = 0) = \frac{\exp\{[a\vartheta_g - b(\vartheta_g)]/\alpha(\phi) + c(a, \phi)\}\,\text{pr}(G = g \mid D = 0)}{\text{pr}(A = a \mid D = 0)}$$

and so

$$\log\ \text{pr}(G = g \mid A = a, D = 0) = \frac{a\vartheta_g - b(\vartheta_g)}{\alpha(\phi)} + c(a, \phi) + \log \text{pr}(G = g \mid D = 0) - \log \text{pr}(A = a \mid D = 0).$$

Let g^0 be a baseline genetic category with corresponding canonical parameter ϑ_0. Then

$$\log\left[\frac{\text{pr}(G = g \mid A = a, D = 0)}{\text{pr}(G = g^0 \mid A = a, D = 0)}\right]$$

$$= \frac{[a\vartheta_g - b(\vartheta_g)]}{\alpha(\phi)} + c(a, \phi) + \log\ \text{pr}(G = g \mid D = 0) - \left\{\frac{[a\vartheta_0 - b(\vartheta_0)]}{\alpha(\phi)} + c(a, \phi) + \log\ \text{pr}(G = g^0 \mid D = 0)\right\}$$

$$= a\frac{\vartheta_g - \vartheta_0}{\alpha(\phi)} + \log\ \text{pr}(G = g \mid D = 0) - \log\ \text{pr}(G = g^0 \mid D = 0) - \frac{b(\vartheta_g) - b(\vartheta_0)}{\alpha(\phi)}$$

$$= \nu_g + a\tau_g,$$

where

$$\nu_g = \log\ \text{pr}(G = g \mid D = 0) - \log\ \text{pr}(G = g^0 \mid D = 0) - [b(\vartheta_g) - b(\vartheta_0)]/\alpha(\phi)$$

and

$$\tau_g = (\vartheta_g - \vartheta_0)/\alpha(\phi).$$

Hence the conditional distribution of G given $A = a$ and $D = 0$ follows a polychotomous regression model that is linear in the attribute a. In summary, so long as the conditional distribution of A given G is in the exponential family with a constant dispersion parameter across the levels of G, the proposed polychotomous regression model should capture the dependence between G and A.

Case 2: The attribute is a categorical variable

Suppose now that A is categorical with categories $a^0, a^1, \ldots, a^p$, where a^0 is a baseline category. Then A has a multinomial distribution one trial and $p + 1$ categories. Write the conditional probability for A given G in controls in exponential family from. Let $\pi_{ig} = \text{pr}(A = a^i \mid G = g, D = 0)$, $i = 1, \ldots, p$, and $\pi_{0g} = \text{pr}(A = a^0 \mid G = g, D = 0) = 1 - \sum_{i=1}^{p} \pi_{ig}$. For an observed value a of A, let $\vec{a} = (a_1, \ldots, a_p)$ denote an indicator row-vector with ith element 1 if $a = a^i$, or a vector of all zeros if $a = a^0$. Further define $a_0 = 1 - \sum_{i=1}^{p} a_i$. Then the conditional mass function for A given $G = g$ in controls is

$$\begin{aligned}
\text{pr}(A = a \mid G = g, D = 0) &= \prod_{i=0}^{p} \pi_{ig}^{a_i} \\
&= \exp\left(\sum_{i=0}^{p} a_i \log \pi_{ig}\right) \\
&= \exp\left((1 - \sum_{i=1}^{p} a_i) \log \pi_{0g} + \sum_{i=1}^{p} a_i \log \pi_{ig}\right) \\
&= \exp\left(\log \pi_{0g} + \sum_{i=1}^{p} a_i [\log \pi_{ig} - \log \pi_{0g}]\right) \\
&= \exp\left(\log \pi_{0g} + \sum_{i=1}^{p} a_i \log[\pi_{ig}/\pi_{0g}]\right)
\end{aligned}$$

Let $\vartheta_{ig} = \log \pi_{ig}/\pi_{0g}$, $\vartheta_g = (\vartheta_{1g}, \ldots, \vartheta_{pg})^T$, $b(\vartheta_g) = -\log \pi_{0g} = -\log(1 - \sum_{i=1}^{p} \pi_{ig})$, $\alpha(\phi) = 1$ and $c(a, \phi) = 0$. Then

$$\begin{aligned}
\text{pr}(A = a \mid G = g, D = 0) &= \exp\left\{\sum_{i=1}^{p} a_i \vartheta_{ig} - b(\vartheta_g)\right\} \\
&= \exp\{\vec{a}\vartheta_g - b(\vartheta_g)\} \\
&= \exp\{[\vec{a}\vartheta_g - b(\vartheta_g)]/\alpha(\phi) + c(a, \phi)\}
\end{aligned}$$

Keeping in mind that now $\vec{a}$ is a row vector and ϑ_g is a column vector, we obtain

$$\text{pr}(G = g \mid A = a, D = 0) = \nu_g + \vec{a}\tau_g$$

where, using the fact that $\alpha(\phi) = 1$,

$$\nu_g = \log\ \text{pr}(G = g \mid D = 0) - \log\ \text{pr}(G = g^0 \mid D = 0) - [b(\vartheta_g) - b(\vartheta_0)]$$

and

$$\tau_g = \vartheta_g - \vartheta_0.$$

This is a saturated model with different probabilities for each value of G given A. The form of the model for $\text{pr}(G = g \mid A = a, D = 0)$ when A is categorical is nearly identical to the model when A is a count or continuous variable. The only exception is that, in the categorical case, an indicator vector $\vec{a}$ appears in the linear model instead of a itself. In fact, adopting the convention that the observed attribute *is* the indicator vector $\vec{a}$, the model may be viewed as the same for count, continuous or categorical data. For simplicity, however, we use the notation from the case where A is a count or continuous variable.

Estimating equations and asymptotic variances under dependence

Under independence of G and A, the derivation of estimating equations and the asymptotic variance of estimators was facilitated by the definition of a row vector $\Lambda(i, g, a)$ of observed variables and a column vector θ of parameters such that $i\delta + \gamma(g) + ix(g, a)\beta = \Lambda(i, g, a)\theta$. Specifically, we had $\Lambda(i, g, a) = [i, z(g), ix(g, a)]$, where $z(g^k)$ is an indicator row vector of K elements for the k^{th} genetic category ($k = 1, \ldots, K$) or a vector of all zeros for the baseline genetic category when $k = 0$, and $\theta = (\delta, \vec{\gamma}^T, \beta^T)^T$, where $\vec{\gamma}$ denotes the column vector $[\gamma(g^1), \ldots, \gamma(g^K)]^T$. The definition of $z(g)$ and $\vec{\gamma}$ were such that $z(g)\vec{\gamma} = \gamma(g)$. We now wish to re-define $\Lambda(i, g, a)$ and θ such that $i\delta + \gamma_a(g) + ix(g, a)\beta = \Lambda(i, g, a)\theta$. Define a column vector $\eta = (\nu_1, \tau_1, \ldots, \nu_K, \tau_K)^T$. Then, for a particular value a of the scalar attribute,

$$\Big[z(g^k) \otimes (1, a)\Big]\,\eta = \nu_{g^k} + a\tau_{g^k} = \gamma_a(g^k),$$

where, for row vectors a and b with p and q elements respectively, $a \otimes b$ denotes the Kronecker product $(a_1 b, a_2 b, \ldots, a_p b)$. Hence, re-defining $\Lambda(i, g, a) = [i, z(g) \otimes (1, a), ix(g, a)]$ and $\theta = (\delta, \eta^T, \beta^T)^T$, we obtain

$$i\delta + \gamma_a(g) + ix(g, a)\beta = \Lambda(i, g, a)\theta.$$

When A is categorical and the observed values are indicator vectors $\vec{a}$, we may still define a row-vector $\Lambda(i, g, \vec{a})$ and a column-vector of parameters θ such that

$$i\delta + \gamma_{\vec{a}}(g) + ix(g, \vec{a})\beta = \Lambda(i, g, \vec{a})\theta.$$

For categorical A with $p + 1$ categories, the observed data are taken to be the indicator vector $\vec{a} = (a_1, \ldots, a_p)$ and the column vector τ_g of regression coefficients is defined as $\tau_g = (\tau_{1g}, \ldots, \tau_{pg})^T$. Let $\eta = (\nu_1, \tau_1^T, \nu_2, \tau_2^T, \ldots, \nu_K, \tau_K^T)^T$. Then

$$\Big[z(g^k) \otimes (1, \vec{a})\Big]\,\eta = \nu_{g^k} + \vec{a}\tau_{g^k} = \gamma_a(g^k).$$

Hence, re-defining $\Lambda(i, g, \vec{a}) = [i, z(g) \otimes (1, \vec{a}), ix(g, \vec{a})]$ and $\theta = (\delta, \eta^T, \beta^T)^T$, we obtain

$$i\delta + \gamma_{\vec{a}}(g) + ix(g, \vec{a})\beta = \Lambda(i, g, \vec{a})\theta.$$

Hence, regardless of whether the attribute is continuous, count or categorical, independence between G and A in controls can be seen to be a special case of the dependence model with each τ_g equal to zero and $\nu \equiv (\nu_1, \ldots, \nu_K)^T = \vec{\gamma}$. Other than the re-definition of Λ and θ, the derivations of estimating equations and asymptotic variances of estimators are unchanged.

Hardy-Weinberg proportions and dependence between G and A

Under independence of G and A in controls, and when the values of G were genotypes, we observed a reduction in the number of model parameters if, in addition, it was possible to assume genotype frequencies in controls followed Hardy-Weinberg proportions. It is of interest to see if a similar reduction in the number of model parameters results from assuming Hardy-Weinberg proportions in controls under dependence. Under the proposed dependence model,

$$\log\left[\frac{\text{pr}(G = g \mid A = a, D = 0)}{\text{pr}(G = g^0 \mid A = a, D = 0)}\right] = \nu_g + a\tau_g$$

where $\nu_g = \log\ \text{pr}(G = g \mid D = 0) - \log\ \text{pr}(G = g^0 \mid D = 0) - [b(\vartheta_g) - b(\vartheta_0)]/\alpha(\phi)$ and $\tau_g = (\vartheta_g - \vartheta_0)/\alpha(\phi)$. The control genotype frequencies $\text{pr}(G = g \mid D = 0)$ appear only through the intercept terms ν_g. Therefore, a model for genotype frequencies can at best reduce the number of parameters needed to express the ν_g terms. However, because the terms $[b(\vartheta_g) - b(\vartheta_0)]/\alpha(\phi)$ in ν_g involve the genotype-specific parameters ϑ_g, it is not generally possible to specify ν_g as a function of allele-specific parameters. Therefore, the assumption of Hardy-Weinberg proportions in controls does not lead to a reduction in the number of parameters in the dependence model between G and A.

Simulation study

Proposed structure of this section: Description of study goals and simulation parameters. Relate selected parameters for the simulation study back to Ji-Hyung's application to real data. Cite her thesis. Simulation results under G-A independence – compare logistic regression (LR) to LUCA:Ind and LUCA:Ind+HWP Simulation results under general G-A dependence – compare LR to LUCA:Ind and LUCA:Dep. Can mention that LUCA:Ind+HWP similar to LUCA:Ind Simulation results under conditional G-A independence given a confounder C – compare LR to LUCA:Dep, and maybe also LUCA:Ind Conclusions from the study. Can follow the structure of Ji-Hyung's JSM talk

Simulation study goals and parameters

Simulation parameters are as follows: Penetrance model parameters $\beta_G = 0.7$, $\beta_A = 0.1$, $\beta_{GA} = 0.2$, and β_0 chosen so that the disease prevalence is 0.0009; in models with the confounder C, $\beta_C = 0.1$ and there is no $G \times C$ interaction; risk allele frequency is 25%; $n_{cas} = n_{con} = 500$; and the number of replications is 10,000.

Results under G-A independence

We will put description of comparisons of interest here. Preliminary version of the table showing results is Table 1. Incorporating valid G-A independence assumption improves power and precision to detect $G \times A$ interaction.

Results under G-A dependence

The gain in precision over standard logistic regression is very small under general dependence structures.

Table 1: Bias (in parameter estimates, standard errors and 95% confidence interval coverage probabilities) and efficiency results for simulations under Hardy-Weinberg proportions (HWP) or not ($\overline{\text{HWP}}$)

			Bias			Efficiency	
Config	Param	Method	Estimate	Std.Err	Cov.Prob	Std.Dev.	Power
HWP	β_G	LR*	0.004	−0.002	−0.007	0.105	1.000
		LUCA:Ind†	0.002	−0.002	−0.004	0.104	1.000
		LUCA:Ind+HWP‡	−0.000	−0.001	−0.007	0.100	1.000
	β_A	LR	0.000	−0.001	0.001	0.093	0.193
		LUCA:Ind	0.000	0.000	0.001	0.083	0.226
		LUCA:Ind+HWP	0.002	−0.001	0.000	0.083	0.228
	β_{GA}	LR	0.002	−0.001	0.002	0.104	0.495
		LUCA:Ind	−0.000	0.000	0.001	0.065	0.879
		LUCA:Ind+HWP	−0.001	0.001	0.001	0.064	0.881
$\overline{\text{HWP}}$	β_G	LR	0.006	0.000	0.000	0.094	1.000
		LUCA:Ind	0.004	0.000	0.002	0.093	1.000
		LUCA:Ind+HWP	0.132	−0.009	−0.234	0.108	1.000
	β_A	LR	−0.001	−0.001	−0.003	0.091	0.190
		LUCA:Ind	0.000	0.000	0.000	0.083	0.231
		LUCA:Ind+HWP	−0.034	−0.004	−0.036	0.089	0.129
	β_{GA}	LR	0.004	−0.002	−0.008	0.097	0.575
		LUCA:Ind	0.001	−0.001	−0.002	0.059	0.936
		LUCA:Ind+HWP	0.040	−0.006	−0.075	0.070	0.956

* LR – Logistic Regression.
† LUCA:Ind – Assume independence of genetic and non-genetic factors in the controls.
‡ LUCA:Ind+HWP – Besides independence, assume Hardy-Weinberg proportions for control genotypes.

Results under conditional G-A independence given a confounder C

Incorporating valid conditional independence assumption between genetic and non-genetic factors given a third confounding variable C improves power and precision to detect $G \times A$ interaction.

Simulation conclusions

Incorporating valid $G-A$ assumptions improves power and precision to detect $G \times A$ interaction for the method under independence assumption (LUCA:Ind) and the one under conditional independence assumption (LUCA:Dep). However, the methods are not robust to dependence of genetic and non-genetic factors. Lastly, incorporating HWP (LUCA:Ind+HWP), in addition to $G - A$ independence, gains little in precision over LUCA:Ind and may cost anti-conservative bias if there are departures from HWP.

Discussion

Comparison of GMS and CC dependence models

We have proposed a simple model for dependence between G and A. Chatterjee and Carroll (2005) also allow for dependence between genetic factors and non-genetic factors. We wish to compare our approach to theirs. However, we make assumptions about the distribution of covariates in controls, while Chatterjee and Carroll impose assumptions on the distribution of covariates in the population. Hence our methodology is not directly comparble to theirs, except under a rare disease assumption. Therefore, to compare our approach to allow for dependence to that of Chatterjee and Carroll, suppose a rare disease. We first describe the specific form of dependence considered by Chatterjee and Carroll. Then we discuss our dependence model under this form of dependence. Under the rare disease assumption, our model is seen to include that of Chatterjee and Carroll as a special case.

Allowing for dependence: Chatterjee and Carroll

Chatterjee and Carroll consider dependence between G and environmental exposures E through a stratum variable S; that is, G and E are dependent, but are conditionally independent given S. The joint distribution of G, S and E is then $\text{pr}(G = g, S = s, E = e) = \text{pr}(G = g \mid S = s)\text{pr}(S = s, E = e)$. The term $\text{pr}(S = s, E = e)$ in the above expression is left unspecified. The term $\text{pr}(G = g \mid S = s)$ is considered in more detail. When S is categorical with only a few categories, no modelling of $\text{pr}(G = g \mid S = s)$ is necessary; i.e, $\text{pr}(G = g \mid S = s)$ is described by a saturated model with separate probabilities for each value of G given S. Otherwise, Chatterjee and Carroll note that modelling of $\text{pr}(G = g \mid S = s)$ is required. They suggest a logistic regression model if G is binary, e.g.

$$\log\left[\frac{\text{pr}(G = 1 \mid S = s)}{\text{pr}(G = 0 \mid S = s)}\right] = \nu + s\tau,$$

but do not discuss more general models for non-binary G.

To summarize, Chatterjee and Carroll allow for dependence between G and E through a stratum variable S, and the model for $\text{pr}(G \mid S)$ is either a saturated model if S is categorical with few categories, or a logistic regression model if G is binary.

Allowing for dependence: Graham, McNeney, Shin

Our model for dependence (see section **Extension to allow dependence between G and A**) is a polychotomous regression for G given A in controls with intercept and slope that may depend on the level of G:

$$\gamma_a(g) = \log\left[\frac{\text{pr}(G = g \mid A = a, D = 0)}{\text{pr}(G = 0 \mid A = a, D = 0)}\right] = \nu_g + a\tau_g.$$

We now discuss this model under the form of dependence considered by Chatterjee and Carroll; that is, when G and E are conditionally independent given S. In our context, the analogous conditional independence is of G and E given S in controls. Let $A = (S, E)$ and suppose conditional independence of G and E given S in controls. Then

$$\text{pr}(G = g \mid A = a, D = 0) = \text{pr}(G = g \mid S = s, E = e, D = 0) = \text{pr}(G = g \mid S = s, D = 0).$$

Hence, the appropriate polychotomous regression model to account for dependence between G and A is

$$\log\left[\frac{\text{pr}(G = g \mid S = s, D = 0)}{\text{pr}(G = 0 \mid S = s, D = 0)}\right] = \nu_g + s\tau_g.$$

For categorical S, the polychotomous regression is a saturated model that allows separate probabilities for each value of G given S, as discussed in section ?.?, and therefore parallels that of Chatterjee and Carroll. Furthermore, our model of dependence includes logistic regression for a binary outcome G given information on S in controls. To summarize, our model of dependence can allow for conditional independence of G and E given a stratum variable S in an analogous way to Chatterjee and Carroll's model of dependence.

References

Albert PS, Ratnasinghe D, Tangrea J, Wacholder S (2001) Limitations of the case-only design for identifying gene-environment interactions. Am J Epidemiol 154:687–693

Chatterjee N, Carroll RJ (2005) Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika 92:In press

Greenland S (1983) Tests for interaction in epidemiologic studies: a review and a study of of power. Stat Med 2:243–251

Hoffjan S, Nicolae D, Ostrovnaya I, Roberg K, Evans M, Mirel D, Steiner L, Walker K, Shult P, Gangnon R, Gern J, Martinez F, Jr. RL, Ober C (2005) Gene-environment interaction effects on the development of immune responses in the 1st year of life. Am J Hum Genet 76:696–704

Hunter DJ (2005) Gene-environment interactions in human diseases. Nat Rev Genet 6:287–298

Hwang S, Beaty TH, Liang KY, Coresh J, Khoury MJ (1994) Minimum sample size estimation to detect gene-environment interaction in case-control designs. Am J Epidemiol 140:1029–1037

McCullagh P, Nelder JA (1989) Generalized Linear Models. , Chapman and Hall: New York 2nd edn

Merikangas KR, Risch N (2003) Genomic priorities and public health. Science 302:599–601

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

Smith P, Day N (1984) The design of case-control studies: the influence of confounding and interaction effects. Int J Epidemiol 13:356–365

Umbach DM, Weinberg CR (1997) Designing and analysing case-control studies to exploit independence of genotype and exposure. Statist in Med 16:1731–1743

Appendix A: Reparametrization of the likelihood

In this appendix we justify the reparametrization of the likelihood in (4) in terms of γ, α and β to the likelihood in equation (5) in terms of γ, p_v^A and β.

Overview

The original likelihood in equation (4) is,

$$L(\gamma, \alpha, \beta) = \prod_{i=0}^{1}\prod_{j=1}^{n_i} \exp\left(\delta_0 + i\delta + \gamma(g_{ij}) + \alpha(a_{ij}) + ix(g_{ij}, a_{ij})\beta\right).$$

Up to a constant, this expression is a product of terms of the form given in equation (3):

$$\begin{aligned}&\text{pr}(G = g_{ij}, A = a_{ij} \mid D = i)\\ &= \frac{n}{n_i}\exp(\delta_0 + i\delta + \gamma(g_{ij}) + \alpha(a_{ij}) + ix(g_{ij}, a_{ij})\beta). \quad \text{(A-1)}\end{aligned}$$

The reparametrized likelihood in equation (5)

$$L(\gamma, p_v^A, \beta) = \left[\prod_{i=0}^{1}\prod_{j=1}^{n_i} \frac{\exp\left(i\delta + \gamma(g_{ij}) + ix(g_{ij}, a_{ij})\beta\right)}{\sum_{l=0}^{1}\sum_{g\in\mathcal{G}}\exp\left(l\delta + \gamma(g) + lx(g, a_{ij})\beta\right)}\right] \times \left[\prod_{i=0}^{1}\prod_{j=1}^{n_i} p_v^A(a_{ij})\right]$$

is based on two claims. The first of these claims is that

$$\begin{aligned}&\text{pr}(G = g_{ij}, A = a_{ij} \mid D = i)\\ &= \frac{n}{n_i}\,\frac{\exp\left(\delta_0 + i\delta + \gamma(g_{ij}) + ix(g_{ij}, a_{ij})\beta\right)}{\sum_{l=0}^{1}\sum_{g\in\mathcal{G}}\exp\left(\delta_0 + l\delta + \gamma(g) + lx(g, a_{ij})\beta\right)}\, p_v^A(a_{ij})\\ &= \frac{n}{n_i}\,\frac{\exp\left(i\delta + \gamma(g_{ij}) + ix(g_{ij}, a_{ij})\beta\right)}{\sum_{l=0}^{1}\sum_{g\in\mathcal{G}}\exp\left(l\delta + \gamma(g) + lx(g, a_{ij})\beta\right)}\, p_v^A(a_{ij}) \quad \text{(A-2)}\end{aligned}$$

Comparing (A-1) to (A-2), the first claim amounts to showing

$$\exp(\alpha(a_{ij})) = \frac{p_v^A(a_{ij})}{\sum_{l=0}^{1}\sum_{g\in\mathcal{G}}\exp(\delta_0 + l\delta + \gamma(g) + lx(g,a_{ij})\beta)}$$

or

$$p_v^A(a_{ij}) = \exp(\alpha(a_{ij}))\sum_{l=0}^{1}\sum_{g\in\mathcal{G}}\exp(\delta_0 + l\delta + \gamma(g) + lx(g,a_{ij})\beta) \quad \text{(A-3)}$$

The second claim is that the reparametrization is valid; i.e. the mapping $(\delta,\gamma,\alpha,\beta) \mapsto (\delta,\gamma,p_v^A,\beta)$ is one-to-one, and both parametrizations are subject to the same constraint

$$1 = \int_{\mathcal{A}}\sum_{g\in\mathcal{G}}\text{pr}(G=g, A=a \mid D=1)\, da$$

For the parametrization $(\delta,\gamma,p_v^A,\beta)$ the constraint is

$$1 = \frac{n}{n_1}\int_{\mathcal{A}}\sum_{g\in\mathcal{G}}\frac{\exp(\delta+\gamma(g)+x(g,a)\beta)}{\sum_{l=0}^{1}\sum_{g'\in\mathcal{G}}\exp(l\delta+\gamma(g)+lx(g,a)\beta)}p_v^A(a)da \quad \text{(A-4)}$$

and the parameter δ defined as the solution is a function of γ, p_v^A and β. We establish each of these two claims in the following subsections.

Derivation of p_v^A

The derivation of p_v^A relies on a hypothetical variant sampling scheme used implicitly throughout the arguments of Prentice and Pyke 1979. This is a two-stage sampling design with random sampling of disease status and covariates, but in which the total number of subjects is fixed to n. The first step is Bernoulli sampling of disease status on each of n subjects, with probability n_1/n of sampling a case and n_0/n of sampling a control. Thus the expected number of cases sampled is n_1 and the expected number of controls sampled is $n - n_1 = n_0$. In the second step, covariates are sampled from the appropriate conditional distributions of covariates given disease status. The conditional distributions in this second step are the same conditional distributions as in the true case-control sampling scheme.

Under this variant scheme, disease-covariate pairs are sampled jointly. However, covariates are sampled independently conditional on disease status. By contrast, in true case-control sampling (basic stratified sampling), disease status is fixed rather than random. However, covariates are still sampled independently conditional on disease status. Adopt the general convention that pr_v denotes probabilites or densities under the variant sampling scheme (VSS) and pr denotes probabilities or densities under sampling from the true population (population sampling). By definition, the distribution of risk factors given disease status is the same under VSS and population sampling; that is, $\text{pr}_v(G,A \mid D) = \text{pr}(G,A \mid D)$. Hence

$$\begin{aligned}\text{pr}(G=g, A=a \mid D=i) &= \frac{\text{pr}_v(G=g,A=a,D=i)}{\text{pr}_v(D=i)} \\ &= \frac{n}{n_i}\text{pr}_v(G=g,A=a,D=i),\end{aligned}$$

which, from the likelihood equation (A-1), is also

$$\frac{n}{n_i}\exp(\delta_0 + i\delta + \gamma(g) + \alpha(a) + ix(g,a)\beta).$$

Thus

$$\begin{aligned}&\text{pr}_v(D=i, G=g_{ij}, A=a_{ij}) \\ &= \exp(\delta_0 + i\delta + \gamma(g_{ij}) + \alpha(a_{ij}) + ix(g_{ij},a_{ij})\beta). \quad \text{(A-5)}\end{aligned}$$

The above model can be shown to reduce to the log-linear model of Umbach and Weinberg (1997) when the attribute A is categorical. This connection is discussed in Appendix B. It then follows that

$$\begin{aligned}p_v^A(a_{ij}) &= \sum_{l=0}^{1}\sum_{g\in\mathcal{G}}\text{pr}_v(G=g, A=a_{ij}, D=l) \\ &= \sum_{l=0}^{1}\sum_{g\in\mathcal{G}}\exp(\delta_0 + l\delta + \gamma(g) + \alpha(a_{ij}) + lx(g,a_{ij})\beta). \\ &= \exp(\alpha(a_{ij}))\sum_{l=0}^{1}\sum_{g\in\mathcal{G}}\exp(\delta_0 + l\delta + \gamma(g) + lx(g,a_{ij})\beta).\end{aligned}$$

This is equation (A-3), which, as noted in the Overview, establishes the first claim (equation A-2).

Validity of the reparametrization

To establish the second claim, that $(\delta,\gamma,\alpha,\beta) \mapsto (\delta,\gamma,p_v^A,\beta)$ is one-to-one, we must be able to write p_v^A as a function of $(\delta,\gamma,\alpha,\beta)$ and conversely α as a function of $(\delta,\gamma,p_v^A,\beta)$. In establishing equation (A-3), we have already shown that p_v^A can be written as a function of $(\delta,\gamma,\alpha,\beta)$. We now show that α can be written as a function of $(\delta,\gamma,p_v^A,\beta)$. Start by rearranging (A-3):

$$\exp(\alpha(a)) = p_v^A(a)\left[\sum_{i=0}^{1}\sum_{g\in\mathcal{G}}\exp(\delta_0 + i\delta + \gamma(g) + ix(g,a)\beta)\right]^{-1}.$$

We now show that δ_0 can also be written as a function of $(\delta,\gamma,p_v^A,\beta)$. From (A-1)

$$\text{pr}(G=g^0, A=a^0 \mid D=0) = \frac{n}{n_0}\exp(\delta_0)$$

From (A-2)

$$\begin{aligned}&\text{pr}(G=g^0, A=a^0 \mid D=0) \\ &= \frac{n}{n_0}\left[\sum_{l=0}^{1}\sum_{g\in\mathcal{G}}\exp(l\delta+\gamma(g)+lx(g,a)\beta)\right]^{-1} p_v^A(a^0)\end{aligned}$$

Thus, we have

$$\exp(\delta_0) = \left[\sum_{l=0}^{1}\sum_{g\in\mathcal{G}}\exp(l\delta+\gamma(g)+lx(g,a)\beta)\right]^{-1} p_v^A(a^0)$$

Hence δ_0 is also a function of $(\delta,\gamma,p_v^A,\beta)$. Conclude that α can be written as a function of δ, γ, p_v^A and β. The original parametrization was constrained so that $\int_{\mathcal{A}}\sum_{g\in\mathcal{G}}\text{pr}(G = g, A = a \mid D = 1)\, da = 1$. By definition, the new parametrization must also satisfy this constraint, given by equation (A-4).

Appendix B: Connections to log-linear models

Overview

From equation (3), the jth individual in the ith disease category contributes the term

$$\begin{aligned}&\text{pr}(G=g_{ij}, A=a_{ij} \mid D=i) \\ &= \frac{n}{n_i}\exp(\delta_0 + i\delta + \gamma(g_{ij}) + \alpha(a_{ij}) + ix(g_{ij},a_{ij})\beta)\end{aligned}$$

to the likelihood. Since the constants n/n_i can be ignored, we end up maximizing equation (4)

$$L(\gamma,\alpha,\beta)=\prod_{i=0}^{1}\prod_{j=1}^{n_i}\exp\left(\delta_0+i\delta+\gamma(g_{ij})+\alpha(a_{ij})+ix(g_{ij},a_{ij})\beta\right).$$

From equation (A-5), the terms of the above product may be interpreted as joint probabilities of disease status, the genetic factor and the attribute under a variant sampling scheme (VSS) in which a fixed number n of subjects is sampled, cases with probability n_1/n and controls with probability n_0/n:

$$\begin{aligned}&\mathrm{pr}_v(D=i,G=g_{ij},A=a_{ij})\\ &\quad=\quad\exp\left(\delta_0+i\delta+\gamma(g_{ij})+\alpha(a_{ij})+ix(g_{ij},a_{ij})\beta\right).\end{aligned}\quad\text{(B-1)}$$

We now show that the log-linear model of Umbach and Weinberg (1997) is a special case of the above model for $\mathrm{pr}_v(D=i,G=g,A=a)$ for discrete attributes.

The general model with binary G and A

Without loss of generality, we consider the simplest possible case of binary disease status D taking value 1 for diseased and 0 for non-diseased, binary genetic factor G with baseline value $g^0=0$ and non-baseline value $g^1=1$, and binary non-genetic factor A with baseline value $a^0=0$ and non-baseline value $a^1=1$.

Recall that $\gamma(g^0)=\alpha(a^0)=0$. Let $\gamma=\gamma(g^1)$ so that $\gamma(g)=g\gamma$, for $g=0,1$. Similarly, let $\alpha=\alpha(a^1)$ so that $\alpha(a)=a\alpha$, for $a=0,1$.

Let the penetrance model have main effects for G and A, as well as a term for G-by-A interaction; i.e.,

$$\log\left[\frac{\mathrm{pr}(D=1\mid g,a)}{\mathrm{pr}(D=0\mid g,a)}\right]=\beta_0+g\beta_G+a\beta_A+ga\beta_{GA}=\beta_0+x(g,a)\beta$$

where $x(g,a)=(g,a,ga)$ and $\beta=(\beta_G,\beta_A,\beta_{GA})^T$. Then the general model (B-1) for $\mathrm{pr}_v(D=i,G=g,A=a)$ simplifies to

$$\begin{aligned}&\log\left[\mathrm{pr}_v(D=i,G=g,A=a)\right]\\ &\quad=\quad\delta_0+i\delta+g\gamma+a\alpha+ix(g,a)\beta\\ &\quad=\quad\delta_0+i\delta+g\gamma+a\alpha+i(g\beta_G+a\beta_A+ga\beta_{GA})\\ &\quad=\quad\delta_0+i\delta+g\gamma+a\alpha+ig\beta_G+ia\beta_A+iga\beta_{GA}\end{aligned}\quad\text{(B-2)}$$

Log-linear models for contingency tables

Log-linear models are typically formulated for expected cell counts in a contingency table but may be equivalently used to model cell probabilities since these are expected cell counts divided by the sample size n. For binary D, G and A, the observed data may be arranged in a $2\times2\times2$ contingency table.

Equation (B-2) is a model for the cell probabilities in such a table with δ_0 as an intercept term, δ as the main effect for disease status, γ as the main effect for the genetic variable, α as the main effect for the attribute, β_G as the two-way interaction between D and G, β_A as the two-way interaction between D and A, and β_{GA} as the three-way interaction between D, G and A. The log-linear model (B-2) thus has no term for two-way interaction between G and A. Hence, model (B-2) is a saturated log-linear model with the two-way interaction between G and A constrained to equal zero; this is precisely the log-linear model in equation (4) of Umbach and Weinberg (1997).

Appendix C: Unconstrained maximizer satisfies the constraint

The constraint

The constraint (A-4) that the parameters $\theta=(\delta,\gamma,\beta)$ and p_v^A must satisfy is

$$1=\frac{n}{n_1}\int_{\mathcal{A}}\sum_{g\in\mathcal{G}}\frac{\exp\left(\delta+\gamma(g)+x(g,a)\beta\right)}{\sum_{l=0}^{1}\sum_{g'\in\mathcal{G}}\exp\left(l\delta+\gamma(g)+lx(g,a)\beta\right)}p_v^A(a)\,da.$$

From the definition of $p_{ig}(a;\theta)$ in equation (7), we can rewrite the constraint as

$$1=\frac{n}{n_1}\int_{\mathcal{A}}\sum_{g\in\mathcal{G}}p_{1g}(a;\theta)p_v^A(a)\,da.$$

The estimators

The estimator $\hat\theta$ satisfies

$$0=\left.\frac{\partial\tilde l_1}{\partial\theta}\right|_{\hat\theta}=\sum_{i=0}^{1}\sum_{j=1}^{n_i}\left[\Lambda(i,g_{ij},a_{ij})^T-\sum_{l=0}^{1}\sum_{g\in\mathcal{G}}\Lambda(l,g,a_{ij})^Tp_{lg}(a_{ij};\hat\theta)\right],$$

and, in particular, from the first element i of $\Lambda(i,g,a)$,

$$\begin{aligned}0=\left[\left.\frac{\partial\tilde l_1}{\partial\theta}\right|_{\hat\theta}\right]_1&=\sum_{i=0}^{1}\sum_{j=1}^{n_i}\left[i-\sum_{l=0}^{1}\sum_{g\in\mathcal{G}}lp_{lg}(a_{ij};\hat\theta)\right]\\ &=\ n_1-\sum_{i=0}^{1}\sum_{j=1}^{n_i}\sum_{g\in\mathcal{G}}p_{1g}(a_{ij};\hat\theta).\end{aligned}\quad\text{(C-1)}$$

The estimator $\widehat{p}_v^A(a)$ of $p_v^A(a)$ is the empirical distribution that puts mass $1/n$ at each observed value of A.

The estimators satisfy the constraint

Evaluating the constraint at $(\hat\theta,\widehat{p}_v^A(a))$ gives

$$\begin{aligned}1\ &=\ \frac{n}{n_1}\int_{\mathcal{A}}\sum_{g\in\mathcal{G}}p_{1g}(a;\hat\theta)\widehat{p}_v^A(a)\,da=\frac{n}{n_1}\sum_{i=0}^{1}\sum_{j=1}^{n_i}\sum_{g\in\mathcal{G}}p_{1g}(a_{ij};\hat\theta)\frac{1}{n}\\ &=\ \frac{1}{n_1}\sum_{i=0}^{1}\sum_{j=1}^{n_i}\sum_{g\in\mathcal{G}}p_{1g}(a_{ij};\hat\theta),\end{aligned}$$

since an integral with respect to $\widehat{p}_v^A(a)$ is simply a sum over the observed values a_{ij}, $i=0,1$, $j=1,\ldots,n_i$, with weight $1/n$ assigned to each value. The above equation is equivalent to (C-1). Hence the unconstrained maximizers $(\hat\theta,\widehat{p}_v^A(a))$ satisfy the constraint (A-4).