Genotyped-Proband Design에서 침투율의 추정방법에 관한 연구

연세대학교 대학원

의학전산통계학협동과정

의학통계학전공

명 성 민

Genotyped-Proband Design에서 침투율의 추정방법에 관한 연구

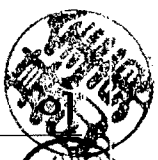
지도 김 동 기 교수

이 논문을 석사 학위논문으로 제출함

2001년 12월 일

연세대학교 대학원
의학전산통계학협동과정
의학통계학전공
명 성 민

명성민의 석사학위 논문을 인준함

심사위원 김 동 기 

심사위원 박 상 언 

심사위원 임 길 섭 

연세대학교 대학원

2001년 12월 일

감사의 글

학부생활과는 또 다른 환경과 시간에 적응하며 보낸 2년이란 짧지 않은 기간동안 보잘것없는 저의 논문이 완성되기까지의 가르침과 꾸짖음을 주신 모든 분들에게 감사의 글을 전하고자 합니다.

먼저 이 논문이 완성되기까지 물심양면으로 격려와 충고, 꾸짖음과 음우를 아끼지 않으셨던 김동기 교수님께 감사드립니다. 또한, 유전통계라는 생소한 학문의 영역에서 너무나 큰 도움을 주시고, 언제나 넉넉한 마음을 기반으로 자상함을 베푸신 임길섭 교수님께 감사의 뜻을 표하고 싶습니다.

어려운 가운데에서도 언제나 말없이 묵묵히 지켜보아 주시고 힘을 북돋아 주시는 아버님과 어머님께 안쓰러움과 감사를 함께 전합니다. 그리고 부족한 동생을 위하여 정성을 쏟고 걱정해주는 누나와 형님과 같은 온정을 전해주는 매형, 또한 사랑스러운 조카 린아 에게도 고맙다는 말을 전하고 싶습니다.

2년여의 생활동안 함께 해오며 귀찮음을 마다않고 토론과 조언을 기꺼이 해주었던 기준형, 언제나 밝은 웃음으로 따듯하게 감싸주시는 박춘선 선생님, 언제나 부드러운 미소로 편안함을 주시는 강대룡 선생님, 지금은 가까이에 없지만 격려와 따가운 질책을 마다하지 않았던 임종건 선생님, 희중형, 한 학기 선배였지만 동기 같은 애정을 나누며 지냈던 시내누나, 희철형, 윤주씨, 영진씨, 언제나 나의 짜증을 받아주면서 동기이상의 우정으로 격려와 도움을 공유한 경민, 갓 1학기를 마쳤지만 너무나 많은 도움과 편안함을 주었던 민지에게 진심으로 감사의 말씀을 드립니다. 아울러 학부에서의 연을 계속 이어가는 우선, 영선, 성재, 무영, 현철, 원열과 연구실에서 저에게 여러모로 도움을 주신 명희씨, 현선씨, 은정, 찬미에게도 감사의 말을 꼭 전하고 싶습니다. 선후배의 관계를 뛰어넘어 한 잔술에 우리들만의 철학과 고민을 토로했던 일훈형, 언제나 웃음과 특유의 해학으로 즐겁게 해주었던 영돈, 유희, 자주 보지는 못했지

만 만날 때마다 언제나 따듯한 웃음으로서 격려해주는 대황형, 비록 다른 분야의 학문을 하지만, 서로를 격려하고 질책했던 순호, 승환, 성훈, 명섭, 그의 반쪽 효미에게도 우애를 담아 감사드립니다.

끝으로 감사를 드려야 할 많은 분들에게 고마움을 드리지 못함을 안타깝게 생각하며 언제나 시종여일의 마음가짐으로 정진하겠다는 말로서 감사의 글을 맺을까 합니다.

2002년 1월

명 성 민 올림

제 목 차 례

그 림 차 례

표 차 례

국문요약

Genotyped-Proband Design에서 침투율의 추정방법에 관한 연구

본 논문에서는 유전자에서 희귀한 변이(mutation)의 침투율(penetrance)을 추정하기 위하여 지원자(volunteer), 또는 시조(proband)를 이용하는 단면연구(cross-sectional study)의 다른 형태인 GPD(Genotyped-proband design)를 이용하고, 그 중에서 Moore 등(2001)이 제시한 유사우도함수(pseudo likelihood function) 와 Chatterjee 등(2001)에 의하여 제안된 주변우도함수(marginal likelihood function)를 이용한 침투율의 추정에 관한 내용을 중점적으로 다룬다. 통계프로그램(S-PLUS)을 이용한 모의실험기법에서는 유사우도를 이용한 추정기법에 대한 평가를 하기 위하여, GPD에 기초한 모의자료를 멘델리안 계산(Li, 1976)을 적용하여 생성하였고, 유사우도추정기법과 환자-대조군 연구(case-control)에 기초한 추정기법을 비교했을 때, 다음과 같은 결론을 얻을 수 있었다.

첫째, 유사우도추정기법은 기존의 연구에 비해서, 조사해야 할 시조의 수(number of proband)에 덜 민감하다.

둘째, 변이가 드문 경우, 즉, $P(A)$가 작은 경우에도 침투율을 추정하는데 있어서 보다 정확하게 추정치를 구할 수 있다는 것이다.

핵심되는 말 : Genotyped-Proband Design, 침투율, 킨-코호트, EM 알고리즘, 시조, 유사우도, 주변우도

제 1 장 서 론

유전자에서 희귀한 변이(mutation)의 침투율(penetrance)을 측정하기 위하여 기존의 여러 가지 척도를 가지고 조사하는 환자-대조군(case-control) 연구방법 또는 단면연구(cross-sectional study)가지고는 설명하기가 힘들뿐더러, 이러한 연구를 실행하기 위해서는 경제적, 윤리적 측면에서 애로사항이 발생한다. 그렇다고, 연관성분석(linkage analysis)등에 의해 이미 수집된 가계도(pedigree)에 관해 특정한 질병에 관한 위험도를 직접 추정하는 것은 과추정(overestimate)될 경향이 있다(Walcholder et al, 1998).

Risch(1984)가 제안한 기존의 단면연구에 대하여, 표현형(phenotype)에 조건부인 유전형(genotype)을 모형화하는 방법은 위험도가 높은 가족에서는 확인에 대한 편이(ascertainment bias)없이 침투율(penetrance)을 추정할 수 있지만, 유전적 또는 환경적요인 때문에 질병이 있는 가족에서 위험도가 높다면, 추정값은 질병을 덜 가지는 가족보다 훨씬 높게 나타날 것이다(Schatzkin et al, 1997).

본 논문에서는 이러한 질병에 대한 유전자의 위험도의 효과를 측정하기 위하여 지원자(volunteers)를 이용하는 단면연구(Cross-Sectional Study)의 다른 형태인 GPD(Genotyped-Proband Design)에 관하여 다룰 것이다.

GPD에서 침투율을 추정하기 위해 알려진 통계학적 방법은 크게 네 가지로 나누어지는데, 그 첫째가 Walcholder 등(1998)이 제안한 생존시간에 대한 $F_g(t)$를 추정하는 방법으로서 Kaplan-Meier 방법을 이용한 것이다.

둘째는 Gail 등(1999)이 제안한 방법으로서, 최대우도추정치(maximum likelihood estimator)를 이용한 방법이지만, 추정치를 구하기 위한 방정식은 풀기가 어려울뿐더러, 또한 불안정(unstable)하다는 점에 있어서 단점이 존재한다.

세 번째는 Moore 등(2001)이 제시한 유사우도(pseudo likelihood)방법으로서 많은 구간을 가지는 조각별 지수모형(piecewise exponential model)에 대하여 쉽게 풀 수 있고, 또한 생존곡선(survival curve)에 대하여 완전한 비모수적인 추정치로 확장시킬 수 있다.

네 번째는 Chatterjee 등(2001)에 의해 제안된 주변우도(marginal likelihood)방법으로 잔여 가족성 상관관계(residual familial correlation)문제가 존재할 때, 다른 제안된 우도함수에 기초한 모형에서 가정하는 조건부 독립의 위반에 대하여 로버스트한 결과를 갖는다고 알려져 있다.

본 논문에서는 GPD에서 침투율을 추정하기 위한 통계학적 방법에 대하여 유사우도추정방법과 주변우도추정방법을 중심으로 다루게 된다. 또한, 기존의 환자-대조군 연구와의 침투율에 대한 비교를 이분형자료(dichotomous data)에 대하여 다루게 될 것이다.

본 논문에서 다루게 될 내용들을 소개하면 다음과 같다. 먼저 GPD의 정의와 특징, 선행가정들에 대하여 간략히 소개한다. 다음으로 GPD에서 침투율을 추정하기 위한 유사우도추정기법과 주변우도추정기법에 관하여 설명하고 구체적으로 표현형(phenotype)이 이분형일 때와 생존시간일 때에 대하여 살펴본다. 마지막으로는 모의자료를 생성하여 유사우도추정기법을 이용한 침투율의 추정치와 환자-대조군 연구를 통한 추정치를 비교하여 유사우도추정기법의 타당성을 제시하고 그 결과를 요약한다.

제 2 장 Genotyped-Proband Design(GPD)

2.1 GPD의 정의와 연구설계

2. 1. 1 개 요

유전자에서 희귀한 변이(mutation)의 침투율(penetrance)을 측정하기 위하여 기존의 코호트(cohort) 또는 환자-대조군(case-control) 연구방법으로는 설명하기가 힘들뿐더러, 이러한 연구를 실행하기 위해서는 경제적, 윤리적 측면에서 애로사항이 발생한다. 또한, 연관성분석(linkage analysis)등에 의해 이미 수집된 특정한 질병에 관한 경향성연구(disease-prone)로부터 위험도(risk)를 직접 추정하는 것은 너무 높게 나타날 수도 있다(Walcholder et al, 1998).

기존에 단면연구(Cross-Sectional Study)에서 쓰이는 방법으로, Risch(1984)가 제안한 표현형(phenotype)에 조건부인 유전형(genotype)을 모형화하는 방법은 위험도가 높은 가족에서는 확인에 대한 편이(ascertainment bias)없이 침투율(penetrance)을 추정할 수 있지만, 유전적 또는 환경적요인 때문에 질병이 있는 가족에서 위험도가 높다면, 추정값은 질병을 덜 가지는 가족보다 훨씬 높게 나타날 것이다(Schatzkin et al, 1997).

그리하여, Walcholder 등(1998)에 의해 최초로 제안된 유전형-시조 연구(Genotyped-Proband Design, 이하 GPD)또는 킨-코호트 연구(Kin-Cohort Design)는 질병에 대한 유전자의 위험도의 효과를 측정하기 위하여 지원자(volunteers)를 이용하는 단면연구(Cross-Sectional Study)이다(Walcholder et al, 1998).

질병에 대한 유전자(gene)에서 변이(mutation)의 효과를 측정하기 위한 도구는 침투율(penetrance) 또는 보균자(carrier)에서의 위험도(risk)이다 (Wacholder et al, 1998).

지원자의 친족(relative)은 출생부터 질병의 발생까지, 또는 중도절단 (censoring)시점까지 추적관찰(followed)되는 후향성 코호트(retrospective cohort)를 형성한다. 즉, 환자군과 대조군의 친족(relative)은 질병에 대한 진단 시 나이의 분포와 질병을 일으킬 위험도(risk)에서의 가족력(family history)의 효과를 추정하기 위하여 후향적인 코호트를 형성한다는 의미이다.

지원자의 친족에 대한 코호트의 구성원들은 유전형(genotype)을 조사하지 않기 때문에, 유전형을 조사한 친족으로부터 변이에 대한 정보를 부분적으로 얻을 수 있다(Wacholder et al, 1998).

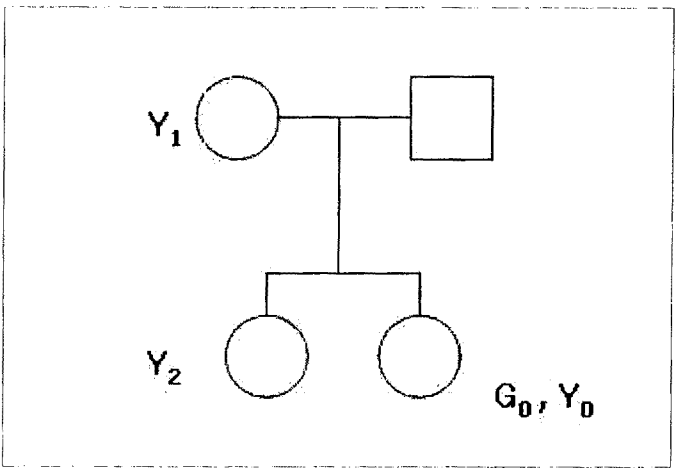
이러한 가능한 정보들을 이용하여 보균자(carrier)와 비보균자(noncarrier)들에 대한 침투율을 어떻게 추정하여야 되는가를 고려해야 한다.

2. 1. 2 GPD의 연구설계

연구설계 방법은 1명의 시조(proband)와 2명의 일촌(first-degree relative)으로 한 가계도(pedigree)를 구성한다. 여기서 시조는 유전형(genotype)과 표현형(phenotype)을 조사하고, 두 명의 친족은 병력(medical history)만을 조사하게 된다(Gail et al, 1999).

예를 들어, 아래 그림 1과 같이 한 가계도에서 유방암(breast cancer)에 관한 연구를 한다고 하자. 시조는 유전형과 표현형을 조사하고, 그녀의 어머니(Y_1)와 자매(Y_2)는 표현형, 즉, 병력만을 조사하게 된다. 그러나, 그녀의 아버지는 표현형을 조사하지 않는데, 이 질병은 남자에게는 드문 침투율

(penetrance)을 가지기 때문이다.



[그림 2] GPD 연구설계

2.2 침투율의 기본적인 추정기법

질병에 대한 누적위험도를 보균자집단과 비보균자집단의 두 집단에 관하여 직관적으로 추정할 수 있다(Wacholder et al, 1998).

보균자와 비보균자인 친족들에서의 누적위험도(cumulative risk)는 보균자와 비보균자들에서의 위험도의 가중평균(weighted average)으로 나타낸다. 물론 보균자에 대한 위험도의 가중치는 비보균자에 비해서 더욱 높을 것이다.

GPD접근방법은 이러한 가중평균치들을 각각의 구성부분으로 분해함으로서, 변이(mutation)의 나이에 따른 침투율(age-specific penetrance)을 추정할 수 있도록 한다. 이러한 가중치는 알려진 유전형질(mode of inheritance) 집단에서 변이의 유병율(prevalance)에 의존한다(Wacholder et al., 1998).
연구집단에서 상염색체(autosomal)에 대하여 우성인 변이의 대립형질(

dominant mutant allele)에 대하여 모든 나이에서 p 의 빈도를 갖는다고 가정한다면, 즉, p가 충분히 작거나, 변이에 대한 동종(homozygosity)이 드물다면, 보균자의 비율이 $2p$ 라고 가정할 때, 보균자에 대한 친족의 구성원은 $\frac{p}{2}+\frac{1}{2}$ 의 변이를 유발할 확률값을 가진다. 반대로, 비보균자 친족의 구성원은 보균자에 대한 확률값 p 가 된다(Wacholder et al, 1998).

이는 다음과 같이 증명할 수 있다.

어머니가 특정한 변이(mutation)를 유발한다고 할 때, A^* 라 하고, 다른 한쪽은 열성형질(wild type) A^w 를 가진다고 하고 아버지의 경우는 알려져 있지 않다고 가정하자. 그렇다면 어머니의 유전형(genotype)은 A^*A^w 로, 아버지는 $A^{a_1}A^{a_2}$로 정의할 수 있다. 그렇다면 자식의 유전형은 $A^*A^{a_1}$, $A^*A^{a_2}$, $A^wA^{a_1}$, $A^wA^{a_2}$가 될 것이며 각각의 확률값은 1/4 이 된다. 그렇다면 변이가 우성형질일 경우에, 보균자가 될 확률값은 각각 1, 1, p, p 가 될 것이다. 여기서 p는 변이가 일어나는 대립형질의 빈도이다. 그렇다면 어머니가 보균자라면 자식이 보균자가 될 확률값은 $(1+1+p+p)/4$, 즉, $p/2+1/2$이 된다.

R_+ 와 R_-를 보균자/비보균자인 친족(kin)에서 나이 t 시점이전에 병이 발생할 개인의 비율(proportion)이라고 정의하자. 그렇다면 R_+ 와 R_- 는 보균자/비보균자에서 나이 t 시점이전에 병이 발생할 개인의 누적위험도(cumulative risk)인 S_+ 와 S_- 의 가중평균(weighted average)이 될 것이다. 근사적인 방정식(approximate equations)으로 나타내면 다음과 같다.

$$R_- = pS_+ + (1-p)S_- \qquad (2.1)$$

$$R_+ = (\frac{p}{2} + \frac{1}{2})S_+ + (\frac{1}{2} - \frac{p}{2})S_- \qquad (2.2)$$

위의 (2.1),(2.2)식은 두 개의 식을 아래와 같이 품으로써 S_+ 와 S_- 를 추정하는데 쓰일 수 있다.

$$S_- = \frac{1+p}{1-p} R_- - 2\frac{p}{1-p} R_+ \qquad (2.3)$$

$$S_+ = 2R_+ - R_- \qquad (2.4)$$

위 (2.4)의 식에서는 p 를 모르고서도 나이에 따른 침투율(age-specific penetrance) S_+ 를 추정할 수 있다. 즉, 다시 말하면 S_-인 경우는 자료로부터 추정되어지거나 알려져 있는 p 값이 필요하다. 여기서, 만약 p 값이 작다면, 즉, 변이(mutation)에 의해 주어진 질병의 특정한 확률값(extra probability) $S_+ - S_-$ 는 아래와 같은 간단한 관계식으로 근사(approximate)된다.

$$S_+ - S_- = 2\frac{(R_+ - R_-)}{1-p} \qquad (2.5)$$

$$\approx 2(R_+ - R_-) \qquad (2.6)$$

R_+ 와 R_-는 Kaplan-Meier 방법을 이용하여 추정되기 때문에 코호트구성원에서 질병진단시의 나이때 점프가 존재한다. 그러므로, R_+ 와 R_-간의 가중차이(weighted difference)인 S_- 와 S_+의 추정치는 단조(monotonic)하지 않다. 예를 들어, 나이 t 시점의 보균자인 친족에서 사건(event)이 존재

친족에서는 존재하지 않는다면, S_+에서 감소의 결과로 식(2.4)에 의해, R_+는 R_-가 감소할 동안에는 변하지 않을 것이다.

또한 분할된 유전적, 환경적 요인에 기인한 하나의 지원자 또는 관련된 지원자들의 친족들 간의 위험도에 대한 독립성 때문에 지원자의 가족이 단위가 되는 S_-, S_+, $S_+ - S_-$ 의 붓스트랩(bootstrap) 신뢰구간을 나타낸다 (Wacholder et al, 1998).

2.3 GPD에서의 선행가정들

GPD에서 선행되는 몇 가지 가정들은 다음과 같다.(Wacholder et al, 1998)

2. 3. 1 질병에 관한 감수성의 유전경향

예를 들어 BRCA1 과 BRCA2 변이로부터 췌장암이나 유방암에 대한 감수성(susceptibility)은 상염색체(autosomal)에 대하여 우성(dominant)의 경향으로 유전된다는 것이다.
유전형을 조사한 개인의 친족에 대한 위험도를 반영하는 식 (2.1),(2.2)에 가중치를 대체함으로써, 다른 유전적인 경향에 대해서도 이와 유사한 분석방법으로 사용되어질 수 있다.

2. 3. 2 변이 대립형질빈도의 일관성

위의 식 (2.1), (2.2), (2.3), (2.5)에서, p 는 출생시 변이 대립형질(mutant allele)의 빈도를 나타낸다. 만약 변이(mutation)가 전체 사망률(mortality)에

강한 영향을 미친다면, 나이에 따른 변이빈도(age-specific mutation frequency)는 나이든 사람에 대해서는 작아질 것이다.

2. 3. 3 위험도의 동질성

위의 식 (2.1)~(2.4)는 다른 유전자나 환경적 요인과 관련된 위험도에서 어떤 이질성에 대해서는 설명하지는 못한다. 예를 들어, 유방암과 췌장암같은 경우에 대해서는, 이러한 변이에 연관된 초과되는 위험도(excess risk)는 다른 알려진 위험요인(risk factor)들의 효과를 압도할 수 있다.
Satten과 Kupper(1993)에 의해 제시된 노출에 대한 확률모형(probability of exposure)모형은 가능하다면, 출산(parity)이나 초경연령(age at menarche)과 같은 친족간의 각각에 대한 위험요인(risk factor)을 통합하여 나타내곤 한다. 물론, 보균자에 관한 질병의 경로(pathway)는 드문 질병(sporadic disease)을 가진 것들하고는 다를 수 있다.

2. 3. 4 지원자의 영향

아마도 GPD에서 주요관심사는 지원자(proband)들에서의 신뢰성이라 할 수 있다. 침투율의 타당한 추정은 지원자들이 우리가 추론하는 집단으로서 질병에 대한 과거력이 같은 분포를 가지는 것이 필요하다. 만약 질병에 대한 가족력을 가진 사람들이 거의 참여하게 되면, 당연히 변이에 관한 유병율(prevalance)과 R_+, R_-의 추정치가 높아지는 경향을 가질 것이다.
가족력의 기능으로서 자발적참여(volunteering)에 대한 경향의 인식은 S_+, S_-의 추정에서 지원자의 영향을 확인하는 것이 필요하기 때문에, 특히 침투율 S_+의 추정이 높게 편의(bias) 되어지는 것처럼 보일수도 있다.

대부분의 환자-대조군(case-control)연구에서 가족력의 영향을 평가하는 것과는 달리, 지원자들의 변이상태를 알지 못하는 지원자들의 사용이 같은 유전자에서 다른 대립형질(alleles)이나 또는 다른 유전자에서의 변이의 침투율을 비교하는 데에 영향을 미쳐서는 안 된다. (Khoury MJ, 1995)

2. 3. 5 자료의 질(Data Quality)

GPD에서는 지원자의 가족병력에 대하여 지원자로부터의 정보에 의존하므로, 더욱 정확한 정보를 얻기 위해서는 친족들과 직접 면담하여 얻어져야 한다.

2.4 GPD에서의 우도함수의 설정

2. 4. 1 선행가정들

GPD에서 일반적인 우도함수를 이용한 침투율을 추정하기 위해서 몇 가지 가정들을 전제한다(Gail et al, 2001).

(A1) 위험도는 상염색체에서 우성의 형태(autosomal dominance pattern)를 따라오는데, 변이 대립형질(mutant allele)의 보균자(carrier)는 병이 발생할 φ_1 라는 침투율(penetrance)를 가지고, 비보균자(non-carrier)는 φ_0 의 침투율을 갖는다.

(A2) A(변이 대립형질), a(정상 대립형질)는 하아디-와인버그 평형(Hardy-Weinberg Equilibrium;이하 HWE)을 따르고, 집단에서 임의교배(random mating)를 한다.

(A3) 유전형(genotype)에 조건부인 친족의 표현형(phenotype)은 지원자

(proband)나 다른 친족의 표현형에 독립이다.

(A4) 지원자(proband)들은 같은 표현형(phenotype)을 가지는 집단의 나머지 구성원들을 대표한다.

(A5) 질병의 상태는 오차(error)없이 결정된다.

(A6) 표본수는 근사이론(asymptotic theory)을 정당화시킬정도로 충분히 커야 한다.

2. 4. 2 표현형이 이분형인 경우

GPD에서 표현형은 양적형질(qunantitative trait)로 나타날 수 있다. 즉, 이분형(dichotomous)형태나 시간에 따른 질병의 발병(time-to-disease-onset)의 형태의 두 가지로 나타난다. 이분형인 경우 우도함수의 설정방법에 대하여 알아보면 다음과 같다.

지원자(proband)의 유전형(genotype)을 g_0 라 정의하고 표현형을 Y_0, 친족의 표현형을 각각 Y_1, Y_2 라 정의하자 그렇다면, $g_0=1$ 인 경우는 지원자가 변이 대립형질 A 를 가지고 있다는 것이며, $g_0=0$ 인 경우는 대립형질 A를 가지고 있지 않다는 의미이다. 또한 표현형 $Y_0=1$인 경우는 질병이 발생한 경우, $Y_0=0$인 경우는 질병이 발생하지 않는 경우가 된다.

q를 인구집단에서의 대립형질 A의 빈도(frequency)라고 한다면, 아래의 식이 성립할 것이다.

$$P(g_0=1)=q^2+2q(1-q)$$

$$P(g_0=0)=(1-q)^2$$

주어진 가족에 대한 우도함수는 $P(Y_{11}, Y_{12}, g_0 \mid Y_0)$ 인데, 위 2. 4. 1.에서의 가정(A3)으로부터 아래와 같이 축약시킬 수 있다.

$$P(g_0 \mid Y_0)P(Y_{11}, Y_{12} \mid g_0)$$

$P(g_0 \mid Y_0)$부분은 $P(g_0, Y_0)$를 대신하여 쓰인것인데, 지원자(probands)가 대표될 수 있지만, 질병의 상태에는 조건부(conditional)하다고 가정할 수 있기 때문이다.(Gail et al, 2001)

예를 들면, 이것은 $Y_0 = 1$을 가지는 모든 가능한 지원자들을 포함하지만, $Y_0 = 0$을 가지는 지원자들인 경우에는 낮은 비율을 포함한다는 것이다.

$P(Y_{11}, Y_{12} \mid g_0)$ 는 $P(Y_{11}, Y_{12} \mid g_0, Y_0)$대신에 쓰인 것이다. 왜냐하면, 조건부(condtional)하게 독립이라는 가정 (A3)하에, Y_{11}과 Y_{12}는 지원자의 유전형을 통하여, 오직 지원자에게만 의존하기 때문이다.

베이즈 정리에 의하여,

$$P(g_0 = 1 \mid Y_0 = 1) = \{q^2 + 2q(1-q)\}\varphi_1 / [\{q^2 + 2q(1-q)\}\varphi_1 + (1-q^2)\varphi_0]$$
$$= \{q^2 + 2q(1-q)\}\varphi_1 / [\{q^2 + 2q(1-q)\}\varphi_1 + (1-q^2)\varphi_0]$$

$P(g_0 = 1 \mid Y_0 = 0)$ 또한 위의 방법과 같은 식으로 구할 수 있다.

조건부하게 독립이라는 가정으로부터, 다음과 같은 식으로 나타낼 수 있다.

$$P(Y_{11}, Y_{12} \mid g_0) = \sum_{g_{11}, g_{12}} P(Y_{11} \mid g_{11})P(Y_{12} \mid g_{12})P(g_{11}, g_{12} \mid g_0)$$

2. 4. 3 표현형이 생존시간인 경우

관심있는 질병의 age t 시점까지의 $g=1$(보균자) 또는 $g=0$(비보균자)인 유전형(genotype)에 대한 누적위험도(cumulative risk)를 고려해보면 다음과 같다.

$$F_g(t) = 1 - S_g(t) = \varphi_g[1 - exp\{-(\lambda_g t)^{\alpha_g}\}] \qquad (2.7)$$

식 2.13 에 관련된 family는 융통성 있고(flexible), $t \to \infty$로서의 침투율 φ_g 를 가지는 와이블분포(Weibull distribution)와 일치한다. (Gail et al, 1999)

특정시점 C 에서 뽑힌 지원자(proband)를 가정하고 a_0 와 a_1 을 지원자와 친족(relative)의 출생부터 C시점까지의 각각의 시간(respective time)으로 놓는다.

또한 d_0 와 d_1 은 지원자와 친족의 사망시 나이라고 정의한다. 그런데, $d_0 > a_0$ 인데, 왜냐하면 지원자는 반드시 뽑히기 전에 생존되어야 하나, $d_1 < a_1$ 은 가능하다. v_0 와 v_1 은 지원자와 친족에서 관심 있는 질병이 발생했을 때의 나이가 된다.

$t_i = min(a_i, d_i, v_i)$ 라 하고, 만약 $t_i = v_i$라 하면 $\delta_i = 1$ 로 놓고, $t_i < v_i$ 라면, $\delta_i = 0$ 으로 놓는다.

g_0에 조건부(conditional)하고, 다른 사망이유가 g_0와는 독립이라 가정하면, 친족으로부터의 우도함수는 아래와 같다.

$$\sum_{g_1} S_{g_1}(t_1)\{h_{g_1}(t_1)\}^{\delta_1} P(g_1 \mid g_0;q)\, G(t_1) \tag{2.8}$$

$$h_g(t) = \phi_g \alpha_g \lambda_g^{\alpha_g t^{\alpha_g - 1}} exp\{-(\lambda_g t)^{\alpha_g}\}/S_g(t) \tag{2.9}$$

또한, $G(t)$는 시점 t 까지의 모든 특정한 질병에 의해서 사망한 경우가 아닌 확률값을 의미한다. G는 g_i에 독립이기때문에, (2.8)를 통한 침투율의 예측에 영향을 미칠 수가 없다. 표현형(phenotype) $Y_0 = (t_0 = a_0, \delta_0 = 0)$를 가지는 대조군 지원자(control proband)를 생각해보면, 지원자의 유전형을 조사한 것으로부터의 우도(likelihood)는 아래와 같이 나타낼 수 있다.

$$P(g_0 = 1 \mid Y_0)^{g_0} P(g_0 = 0 \mid Y_0)^{1-g_0} \tag{2.10}$$

$$P(g_0, Y_0) = P(g_0) G(t_0) S_{g_0}(t_0) \tag{2.11}$$

여기서 $G(\cdot)$ 은 유전형 g_0 에 독립이라 가정되어지고, S_{g_0} 는 독립에 대한 가정없이 사망이유(cause of death)의 존재에서 특정한 이유의 질병에 관한 발생율(cause-specific disease incidence rate)로부터 아래와 같이 예측되어질 수 있다.

$$P\{g_0 = 1 \mid Y_0 = (t_0, \delta_0 = 0)\} \tag{2.12}$$
$$= P(g_0 = 1) S_1(t_0)\{P(g_0 = 1) S_1(t_0) + P(g_0 = 0) S_0(t_0)\}^{-1}$$

지원자(proband)가 나이 $t_0 = v_0$ 이전에 관심 있는 질병을 가질 수 있지만, 나이가 a_0 인 시점에서 생존한 경우 다음과 같이 나타낼 수 있다.

$$P(g_0, Y_0) = P(g_0)\, G(t_0) S_{g_0}(t_0) h_{g_0}(t_0) J(a_0 - t_0; t_0) \qquad (2.13)$$

$J(u, v)$ 는 나이 v 에서 관심있는 질병이 발생한 사람의 나이가 $v+u$ 까지 살아남을 확률값을 의미한다. 단, G 와 J 는 g_0 에 독립이라고 가정한다. 물론, 후자의 가정은 항상 유지되는 것은 아니다(Gail et al, 1999).

식 (2.13)에 의해, 다음과 같이 식(2.14)를 나타낼수 있다.

$$P\{g_0 = 1 | Y_0 = (t_0, \delta_0 = 1)\} \qquad (2.14)$$
$$= P(g_0 = 1) S_1(t_0) h_1(t_0) \{P(g_0 = 1) S_1(t_0) h_1(t_0) + P(g_0 = 0) S_0(t_0) h_0(t_0)\}^{-1}$$

각각의 가족에 대하여 우도(likelihood)는 (2.12) 또는 (2.14) 로부터 $P(g_0 | Y_0)$을 계산함으로서 얻어지고, 식 (2.10)과 (2.8)를 곱하면 전체 우도함수(full likelihood)를 구할 수 있다.

제 3 장 GPD에서 침투율의 추정기법

3. 1 유사우도(Pseudo likelihood)를 이용하는 방법

3. 1. 1 개 요

Wacholder등은(1998) 상염색체 우성유전자와 연관된 병이 발생할 확률값(penetrance)를 추정하는 방법을 킨-코호트연구(kin-cohort design)라 명명하였다. Gail등(1999)은 킨-코호트라는 용어대신 GPD(Genotyped-proband design)라는 용어를 사용하였는데, 침투율을 인구집단에 기초한 추정치를 얻기 위하여 표현형(phenotype)에 조건부(conditional)한, 다시 말하면 지원자(proba nd)의 대표적인 표본을 얻기 위함의 중요성과 지원자에 대하여 유전형을 조사(genotyping)했다는 것을 강조하기 위함이었다.

변이의 유/무에 대한 질병에 관한 생존분포(1-cumulative incidence function)를 추정하기 위하여 Walcholder(1998)는 다음과 같이 방법을 이용하였다.

변이(mutation)를 유발하는 지원자의 친족(first-degree relative)에 관한 생존분포(survival distribution)는 변이가 희귀하게 일어나는 것에 대하여 50:50의 혼합비율(mixing proportion)을 가지는 변이의 유/무에 관한 생존분포의 혼합형태라는 것이다. 마찬가지로, 변이가 없는 지원자 친족의 생존분포는 0:100의 변이의 유/무 분포의 혼합형태로서 나타난다는 것이다.

실제의 혼합비율(mixing proportions)은 대립형질의 빈도(allele frequency) $q = P(A)$ 로서 나타난다. 여기서 A 는 변이의 대립형질(mutant allele)을 의미한다.

Wacholder등(1998)과 Struewing등(1997)은 변이의 유/무인 지원자들의 친족에 대하여 생존분포(survival dist)를 Kaplan-Meier 추정치를 각각 구하고 각각의 보균자/비보균자에 대한 생존분포를 구하기 위해 혼합되어 있는 형태의 2개의 선형방정식을 풀었다. 이러한 추정치들은 일치적인(consistent) 성격을 갖지만, 표본이 작은 경우에서의 생존률(survival) 의 추정치가 반드시 단조(monotone)하지가 않았다는 것이다.

Gail등(1999)은 질병이 발생할 확률값을 나타내는 모수를 포함하는 부적절 와이블모형(improper Weibull model)에 대하여 보균자/비보균자의 자료로부터 생존분포와 대립형질빈도 q 의 모수적인 최대우도추정치(parametric MLE)를 어떻게 구하는가에 대하여 설명하였다. 이러한 접근방법의 장점은 누적발생율(cumulative incidence)의 추정치가 단조하게 증가하는 것이 나타날 것이라는 점이다. 그러나 최대우도추정치를 구하기 위한 방정식은 풀기가 어려울뿐더러, 또한 불안정(unstable)하다는 점이다.

그래서 단점들을 보완하여 Moore등(2001)이 제시한 유사우도접근법(Pseudo-likelihood procedure)은 많은 구간을 가지는 조각별 지수모형(piecewise exponential model)에 대하여 쉽게 풀 수 있고, 또한 생존곡선(survival curve)에 대하여 완전한 비모수적인 추정치로 확장시킬 수 있다.

3. 1. 2 선행가정 및 우도함수의 설정

Y_0 를 지원자의 표현형(phenotype)으로 정의하고, 친족의 표현형의 배열로서

$Y_1{}^T = (Y_{11}, Y_{12}, \cdots, Y_{1m})$ 로 놓는다. 본 논문에서는 친족(어머니와 누이 또는 동생)으로 제한하지만, 일반적으로 많은 식이 적용될 것이다.

이분형자료(dichotomous data)에서는 Y_0 가 1 인 경우 지원자가 병에

걸림을 의미하고 0 인경우는 병에 걸리지 않을 경우를 의미하며, 양적자료(quantitative data)에 대해서는 Y_0 가 측정되어질 수 있는 자료, 즉 생존자료(survival data)같은 것을 의미한다. $Y_0=(T,\delta)$ 는 추적기간이 끝났을 때의 나이(T)와 지원자가 $\delta=0$ 또는 1 의 질병상태를 나타내는 한 쌍으로 구성되어 있다.

추적관찰 T 는 질병이 발생하거나 중도절단(censoring)의 바로 전 시점에서 끝나게 된다. Y_{1j} 또한 마찬가지로 정의할 수 있다.

우리는 변이 대립형질(mutant allele) A 와 야생형 대립형질(wildtype allele) a 를 가지는 상염색체 우성질병모형(autosomal dominant disease model) 에 관하여 가정할 것이다. 또한, 하아디-와인버그 평형(Hardy-Weinberg equilibrium)을 가정하는데, 즉, 임의로 선택된 개체의 유전형이 AA, Aa, aa 가 나타날 확률은 각각 q^2, $2q(1-q)$, $(1-q)^2$ 이다. 여기서 $q=P(A)$ 이다.

상염색체 우성모형(autosomal dominant model)하에서는 질병의 확률값이 한 개체가 보균자(AA 또는 Aa) / 비보균자(aa) 여부에 의존하게 된다. 그러므로, 실제 계산에서는 q 대신에 $\pi=P(AA\ or\ Aa)=1-(1-q)^2$ 의 보균자가 될 확률값을 이용하는 것이 편리하다. 더욱이, 지원자가 보균자인지의 여부는 $g_0=1$ 또는 0 에 의해서 유전형이 특정지어질 수 있고, 친족에 관한 $m\times 1$ 유전형 $g_{\cdot}$ 은 g_0 와 유사한 구조를 가지는 성분으로 이루어져 있다.

하아디-와인버그 평형의 가정하에, 일반적인 조건부 확률질량함수(conditional mass function) $p(g_1|g_0;\pi)$를 얻기 위하여, Li(1976)가 제시한 멘델리안 계산방법을 이용할 수 있다. 또한 하아디-와인버그 평형가정은 가계도에서의 각각의 유전형 확률값을 필요로 한다. Gail 등(1999)은 우리가

고려하는 경우의 가계도가 작게 구성되어있는 경우에 대하여 $p(g_1 | g_0;\pi)$를 계산하기 위한 목록의 간단한 방법을 기술하였다.

i 번째 가족을 색인 할 때, 즉, $i=1, \cdots, I$ 일 때, y_{0i}, $\tilde{y_{1i}}$, g_{0i}, g_{1i}, g_{1ij}의 기호를 사용한다.

우리의 주요관심사는 유전형이 주어졌을 때, 표현형의 조건부 밀도함수, 또는 질량함수 $f(y_0|g_0;\varphi)$ 을 추정하는데 있다. 예를 들면, 이분형자료(dichotomous data)에 관한다면 다음과 같은 식으로 나타낼 수 있을 것이다.

$$f(y_0|g_0=1\,;\varphi_0,\varphi_1)= \varphi_1^{\,y_0}(1-\varphi_1)^{\,1-y_0}$$

$$f(y_0|g_0=0\,;\varphi_0,\varphi_1)= \varphi_0^{\,y_0}(1-\varphi_0)^{\,1-y_0}$$

여기서 $\tilde{\varphi}=(\varphi_0,\varphi_1)$ 은 보균자/비보균자 침투율에 관한 모수이다.

시간이 고려된 자료의 경우에 f 는 비보균자에 대한 모수 φ_0 와 보균자에 대한 φ_1 으로 특징지어지는 생존곡선의 밀도함수가 된다.

GPD에 대한 우도함수(likelihood)를 기술하기 위해, 몇 가지 가정을 한다.(Moore et al, 2001)

1. 가족의 구성원에 관한 표현형은 주어진 그들의 유전형과 조건부 독립성(conditionally independent)을 가진다.

2. 또한 하아디-와인버그 평형을 가정함으로써, 개개인의 표현형이 다음 세대에 영향을 미칠 가능성을 무시한다는 것이다.

GPD 샘플링 계획으로부터 주어진 가족들에 대한 우도함수(likelihood function)를 다음과 같이 기술할 수 있다. (Moore et al, 2001)

$$f_0(g_0|y_0;\tilde{\varphi},\pi)\,f_1(\tilde{y_1}|g_0;\tilde{\varphi},\pi) \qquad (3.1)$$

여기서 f_0 는 g_0 에 관한 조건부 확률질량함수(conditional p.m.f) 이며, f_1 은 친족 $\tilde{y}_1$ 들의 표현형에 벡터에 관한 조건부 밀도 또는 질량함수이다.

식 (3.1)에서 첫 번째 요인(f_0)은 지원자가 랜덤하게 선택되어졌을 때, 그것들이 표현형(phenotype)에 조건부(conditional)하면, 베이즈정리에 의하여 아래와 같은 식으로 나타낼 수 있다.

$$f_0(g_0 \mid y_0;\pi, \tilde{\varphi}) = \frac{\pi^{g_0}(1-\pi)^{1-g_0} f(y_0 \mid g_0;\tilde{\varphi})}{\sum_u \pi^u (1-\pi)^{1-u} f(y_0 \mid u;\tilde{\varphi})} \qquad (3.2)$$

(3.1)식에서 두 번째 요인(f_2)은 조건부독립 가정으로부터 나타난다. 왜냐하면, g_0 와 y_0 가 주어진 $\tilde{y}_1$ 의 조건부 밀도함수는 아래와 같기 때문이다.

$$f_1(\tilde{y}_1 \mid g_0;\tilde{\varphi}, \pi) = \sum_{\tilde{g}_i} \prod_{j=1}^{m} f(y_{1j} \mid g_{1j};\tilde{\varphi}) P(\tilde{g}_i \mid g_0;\pi) \qquad (3.3)$$

전체 우도함수(full likelihood)는 $e^l = e^{l_1} \cdot e^{l_0}$ 가 되며, e^{l_1} 과 e^{l_0}는 $f_1(\tilde{y}_1 \mid g_0;\tilde{\varphi}, \pi)$ 와 $f_0(g_j \mid y_0;\tilde{\varphi}, \pi)$ 의 가족들의 곱으로 나타난다. 원칙적으로, 로그 우도함수(log-likelihood) l 은 π 와 $\tilde{\varphi}$ 에 대하여 최대화되어질 수 있으며, $\hat{\pi}$ 와 $\hat{\tilde{\varphi}}$ 의 분산은 관찰된 피셔의 정보행렬(Fisher's information matrix)로부터 얻어질 수 있다. 그러나 실제로는, 모수추정치에서 평가된 로

그 우도함수의 수치적 차이(numerical differentiation)의 의해 관찰된 정보행렬의 평가를 하는 것이 더욱 용이하다.(Moore et al, 2001)

전체 최대우도점수방정식(full maximum likelihood score equations)은 불안정한 추정치와 많은 모수를 가지는 조각지수 생존모형(piecewise exponential survivial model)에 대한 수렴(convergence)의 실패를 유도할 수 있기 때문에 유사우도기법을 고려하여 본다.(Moore et al, 2001)

3. 1. 3 유사우도를 이용한 추정방법

우리는 다음의 추정방정식을 풀어야 한다.(Godambe, 1991)

$$U_{1\varphi}(\tilde{\varphi};\pi)=\frac{\partial l_1}{\partial \tilde{\varphi}}=0 \qquad (3.4)$$

$$U_{0\varphi}(\tilde{\varphi};\pi)=\frac{\partial l_0}{\partial \pi}=0 \qquad (3.5)$$

식 (3.4)는 고정된 π 에 대하여 $\tilde{\varphi}$ 의 관점에서 l_1 을 최대화하는 것을 의미하며, 식 (3.5)는 고정된 $\tilde{\varphi}$ 에 대하여 π 의 관점에서 l_0 를 최대화 하는 것을 의미하며 모수의 추정치가 수렴(converge)할 때까지 계산한다.

로그우도함수로서 l_1 을 보게 되면, 우리는 $\hat{\pi}$ 의 대입(substitution)을 유사우도(pseudo-likelihood procedure)로 볼 수 있다 (Gong and Samaniego, 1981).

마찬가지로, l_0 를 볼 때 $\hat{\hat{\varphi}}_0$ 와 $\hat{\hat{\varphi}}_1$ 의 대입 또한 마찬가지로 유사우도라 할 수 있다.

일반적인 테일러급수를 이용하여 가족의 수가 무한대로 갈수록, 유사우도

과정을 통한 추정치는 $\hat{B}^{-1}\hat{\Omega}(\hat{B}^{-1})'$ 에 의한 추정치가 되는 분산-공분산 행렬을 가지는 점근적인 정규적 분포(asymptotically normally distribute)가 될 것이다. 여기서,

$$\hat{B} = \left| \begin{array}{cc} \dfrac{\partial^2 l_0}{\partial \pi^2} & \dfrac{\partial^2 l_0}{\partial \pi \partial \tilde{\varphi}} \\ \dfrac{\partial^2 l_1}{\partial \tilde{\varphi} \partial \pi} & \dfrac{\partial^2 l_1}{\partial \tilde{\varphi} \partial \tilde{\varphi}'} \end{array} \right|_{\hat{\pi}PLE, \hat{\tilde{\varphi}}PLE}$$

$$\hat{\Omega} = \left| \begin{array}{cc} \hat{\Omega}_{11} & 0 \\ 0 & \hat{\Omega}_{22} \end{array} \right| = \left| \begin{array}{cc} \dfrac{\partial^2 l_0}{\partial \pi^2} & 0 \\ 0 & \dfrac{\partial^2 l_1}{\partial \tilde{\varphi} \partial \tilde{\varphi}'} \end{array} \right|_{\hat{\pi}PLE, \hat{\tilde{\varphi}}PLE}$$

고정된 π 에 대해서는 EM 알고리즘을 이용하여 $\tilde{\varphi}$ 에 대하여 $U_{1\varphi} = 0$ 을 풀 수 있다. 만약, $\tilde{g_1}$ 이 알려져 있다면, 일반적인 알고리즘은 완전자료에 대한 우도함수 $L_1 = \Pi f(\tilde{y_1} | \tilde{g_1}; \varphi)$ 을 $\tilde{\varphi}$에 대하여 최대화시킬 수 있다 (M-Step).

$f(\tilde{y_1} | \tilde{g_1}; \tilde{\varphi})$는 $f(y_{ij} | g_{1j}; \tilde{\varphi})$의 친족들을 곱한 것이다. 'E-step'에 관해서는 $\tilde{g_{1i}}$와 $\tilde{g_{0i}}$ 가 주어졌을 때, $\tilde{g_{1i}}$ 의 기대값이 필요하다(Mclachlan and Krishan, 1997)

첫 번째로, 특정한 가족에 관한 결합 조건부확률밀도함수(joint conditional density)를 계산하는 것이 필요하다.

$$h(\tilde{g_1} | \tilde{y_1}, g_0) = \frac{f(\tilde{y_1} | \tilde{g_1}) p(\tilde{g_1} | g_0)}{\sum_{\tilde{g}} f(\tilde{y_1} | \tilde{g}) p(\tilde{g} | g_0)} \qquad (3.6)$$

$\tilde{g_1}$ 의 j번째 요소의 조건부기대값은 다음과 같이 주어진다.

$$E(g_{1j}|\tilde{y_1}, g_0) = h(g_{1j}|\tilde{y_1}, g_0) = \sum_{u_l: l \neq j} h(u_1, \cdots, g_{1j}, \cdots, u_m | \tilde{y_1}, g_0) \qquad (3.7)$$

고정된 π에 대해서 식 (3.4)를 풀기 위하여 M-step 과 E-step를 반복한다.

유사우도함수의 추정치 $(\hat{\pi}, \hat{\varphi})$ 를 얻기 위하여, 식 (3.4) 를 풀기 위하여 EM 알고리즘을 사용하고 고정된 φ되고 π에 대해서 식 (3.5)를 풀고 이를 반복시킨다.

3. 2 주변우도(marginal likelihood)를 이용하는 방법

3. 2. 1 개 요

로쿠스에서 개개인의 유전형이 완전히 그 사람에 대한 질병여부를 결정한다고 하면, 로쿠스의 관점에서 질병은 완전 침투(fully penetrant)한다고 말할 수 있다. 그러나, 복잡한 질병의 발생은 다른 많은 유전적 요인뿐만 아니라 환경적요인 또한 포함할 수 있다. 그러므로, 질병의 표현(disease expression)에서 변이(variation)는 다른 위험요인(risk factor)들의 관점을 배경으로 하여, 그것들의 차이에 기인하는 주어진 로쿠스에서 같은 유전형을 가지는 개개인간으로 나타낼 수 있다.

복잡한 질병(complex disease)에 관한 불확실성(uncertainty)은 것은 침투율(penetrance)이라는 용어로 나타낼 수 있다. 이는 즉, 주어진 질병의 로쿠스(disease locus)에서 위험요인의 유전형(risk genotype)을 가지는 질병의

확률값이다 (Chatterjee et al, 2001)

특정한 로쿠스에서 변이(mutation)가 질병에 관한 위험요인(risk factor)으로서 확인되어진다면, 조사자는 변이를 유발하는 것과 관련된 침투율(penetrance)을 인구집단에 기초하여 추정하는 것이 주관심사가 될 것이다.

연관성분석(linkage study)같은, 질병의 유전자(disease gene)를 확인하기 위한 연구는 일반적으로는 유전자를 찾아낼 수 있는 검정력(power)을 높이기 위하여 병에 걸린 사람이 많은 가족들에 관한 자료를 수집한다.

병에 걸린 가족에 의하여 침투율을 추정하는 것은 표본에서 병에 걸린 사람들이 많이 나타났기 때문에, 일반 집단에서는 위험도가 과추정(overestimate)될수 있다는 것이다.

Wacholder등(1998)의 모멘트를 이용한 접근방법보다 주변우도(marginal approach)를 이용한 방법의 장점은 많은 경우의 정보를 분석에 통합할 수가 있다는 점이다. 예를 들면, 알려진 누적위험도의 모수적인 형태, 공변수(covariates), 친족들의 과거력, 지원자들의 질병상태등과 같은 것들이다. 또한, Wacholder등(1998)이 제시한 모멘트를 이용한 방법으로부터 얻어진 누적위험도(cumulative risk)의 추정이 감소하지 않는다는 문제는 우도함수를 이용하는 방법을 고려함으로써 해결될 수 있다. (Gail et al, 1999) 그러나, 다른 공유된 유전적이나 환경적 요인으로부터 발생하는 잔여 가족성 상관관계(residual familial correlation)문제 때문에, 일반적인 우도함수를 이용한 방법이 내포하고 있는 가정에 위반될 수도 있다는 점이다.

Carroll 등(1999)은 우도함수로부터 추정된 침투율이 이러한 상관관계의 존재에 의해 편의(bias)가 나타난다는 것을 발견하였다.

주변우도(marginal likelihood)에 기초한 추정방법은 계산이 간단하고, 속도가 빠르며, 다른 제안된 우도함수에 기초한 모형에서 가정하는 조건부 독립의 위반에 대하여 로버스트한 결과를 갖는다(Chatterjee et al, 2001).

명확하게, 이는 다음과 같이 나타낼 수 있다. 만약, 지원자가 주어진 집단에서의 임의로 뽑힌 표본으로 볼 수 있다면, 주변우도함수에 의한 추정치는 잔여 가족성 상관관계(residual familial correlation)의 존재에 관계없이 일치적인 침투율의 추정치를 구할 수 있다. 그러나, 잔여 가족성 상관관계(residual familial correlation)가 존재하지 않는다고 전제했을 경우에는 주변우도함수에 의한 추정치는 덜 효율적이다(Chatterjee et al, 2001).

3. 2. 2 선행가정 및 우도함수의 설정

적절한 샘플링 계획을 이용하여 조사하고자 하는 인구집단으로부터 뽑힌 K 명의 지원자(proband)에게서 DNA 샘플을 제공받고, 또한 개인력과 병에 대한 가족력을 조사했다고 가정하자. 여기서 우리는 다음과 같은 사항을 가정한다(Chatterjee et al, 2001).

관심의 대상이 되는 로쿠스가 상염색체에 있다. 또한, 각각의 개인들은 두개의 대립형질 중 어느 한쪽을 받는다. 즉, 각각의 부모로부터 변이형질 A, 또는 야생형질 a 를 받는다는 것이다. 그렇다면, 각각의 유전형은 3개의 가능한 조합, AA, Aa, aa 들 중 하나를 가질 것이다.

여기서 질병이 우성의 경향으로 나타날 것이라고 가정한다. 즉, 질병에 관한 위험도가 변이 대립형질(mutant allele)의 관점에서 동일하게 나타나거나 (AA) 또는 이질적으로 나타는 경우(Aa)에 상관없이 위험도는 동일하다는 것이다.

g_i^p 는 K 지원자들 중 i번째 사람이 보균자(AA 또는 Aa)인지, 아니면 비보균자(aa)인지의 여부에 관한 표시변수(indicator variable)이다.

i 번째 지원자는 그들의 n_i 친족에 대한 관심 있는 질병의 가족력에 대한 정보를 제공한다. 즉, $y_i^R = (y_{i1}^R, \cdots, y_{in_i}^R)$ 은 i 번째 지원자의 가족력에 대

한 정보라 정의하고, $g_i^R = (g_{i1}^R, \cdots, g_{in_i}^R)$은 친족의 유전형이지만, 관찰되는 것은 아니다. $(y_1^P, \cdots, y_K^P)$는 지원의 질병력을 의미한다.

$q_0(y;\theta_0)$ 와 $q_1(y;\theta_1)$은 모수 θ_0, θ_1 를 갖는 Y의 확률질량/밀도함수를 의미한다. 즉, 비보균자와 보균자의 집단간의 질병에 관한 분포를 의미한다.

또하나의 가정은, 유전형에 조건부하면, 친족의 집단에서 위험도는 표본추출된 지원자들의 집단에 대한 위험도(risk)와 동일하다는 것이다.
유전형의 확률분포에 관한 이후에 나오는 모든 계산은 일반적인 멘델리안 유전가정을 따른다(Li, 1978). 특히 비근친교배(noinbreeding), 임의교배(random mating), 와 하아디-와인버그 평형(HWE)을 가정한다.

변이 대립형질 A의 대립형질의 빈도는 f 로 정의한다.

우도함수를 설정하는 방법은 잔여 가족성 상관관계(residual familial correlation)가 존재하는 경우와 존재하지 않는 경우로 나눌 수 있다. 먼저 잔여 가족성 상관관계가 존재하지 않는 경우는 Gail 등(1999)이 제시한 우도함수 접근방법이라 할 수 있다.

먼저 가족들의 구성원에 대한 유전형이 조건부 하면, 그들의 표현형 Y는 독립이라고 가정하면, 아래와 같은 식으로 나타낼 수 있다.

$$\begin{aligned} & pr(y_{i1}^R, \cdots, y_{in_i}^R, y_i^P \mid g_{i1}^R, \cdots, g_{in_i}^R, g_i^P) \\ &= pr(y_{i1}^R \mid g_{i1}^R) \cdots pr(y_{in_i}^R \mid g_{in_i}^R) pr(y_i^P \mid g_i^P) \end{aligned} \tag{3.8}$$

위의 가정 하에, 색인된 지원자(participant)들의 유전형에 조건부한 친족들의 자료에 대한 우도함수는 식 (3.9)와 같이 기술할 수 있다.

$$L^R = \prod_{i=1}^{K} \sum_{g_{i1}^R, \cdots, g_{in_i}^R} q_{g_{i1}}(y_{i1}^R;\theta_{g_{i1}^R}) \cdots q_{g_{in_i}}(y_{in_i}^R;\theta_{g_{in_i}^R}) \times pr(g_{i1}^R, \cdots, g_{in_i}^R | g_i^P) \quad (3.9)$$

여기서 $pr(g_{i1}^R, \cdots, g_{in_i}^R | g_i^P)$ 부분, 즉, 지원자들의 유전형이 주어진 가족 구성원들의 유전형의 결합분포(joint distribution)는 멘델리안 유전적 방식(mode of inheritance)을 이용하여 대립형질 빈도의 함수로서 계산되어질 수 있다(Gail et al, 1999).

지원자들의 우도함수에 대한 기여도(likelihood contribution) 는 임의표본(random sample)의 형태라고 가정할 때 아래와 같은 식으로 나타난다.

$$\begin{aligned} L^P &= \prod_{i=1}^{K} pr(y_i^P | g_i^P) pr(g_i^P) \quad (3.10) \\ &= \prod_{i=1}^{K} q_{g_i^P}(y_i^P;\theta_{g_i^P}) pr(g_i^P) \end{aligned}$$

표현형에 조건부한 지원자들의 샘플링을 적합시키기 위하여, Gail등(1999)은 아래의 식으로서 지원자들의 우도함수를 제시하였다.

$$L^P = \prod_{i=1}^{K} pr(g_i^P | y_i^P) \quad (3.11)$$

지원자들의 주어진 샘플링 방법에 의존하여, 식 (3.10),(3.11)을 이용할 수 있다.

지원자들이 랜덤샘플링이 되어진 주변 우도함수(marginal likelihood)는 L^R 식을 수정하는 것으로서 정의된다. 즉, 지원자의 친족들 사이의 어떤 관계(relationship)도 무시함으로써, 각각 i 번째 지원자의 n_i 친족들을 통제한다. 그러므로 1명의 지원자와 n_i 명의 친족을 가지는 n_i+1의 구성원을 가

지는 가족이 각각의 2개의 구성(지원자와 친족)된 n_i의 유사-가족집단으로 분리된다(Chatterjee et al, 2001).

i 번째 지원자의 유전형이 주어졌을 때, j 번째 친족에 관한 표현형의 조건부 확률값은 아래와 같다.

$$pr(y_{ij}^R \mid g_i^P) = \sum_{g_{ij}^R}^{1} q_{g_{ij}^R}(y_{ij}^R;\theta_{g_{ij}^R}) pr(g_{ij}^R \mid g_i^P)$$

친족들의 자료에 관한 주변우도(marginal likelihood)는 아래와 같이 정의된다(Chatterjee et al, 2001).

$$L_M^R = \Pi_{i=1}^{K} \Pi_{j=1}^{n_i} pr(y_{ij}^R \mid g_i^P) \qquad (3.12)$$

우도함수 L^R 과 L_M^R 을 설정하는 데에서 유전적 방식(mode of inheritance)의 사용에서의 차이를 보게 되면, 지원자의 유전형에 조건부하다면, 우리는 L_M^R 에 대해서는 개개인의 친족들의 유전형 분포를 계산할 수 있지만, L^R에 관해서는 결합적이다(Chatterjee et al, 2001). 이는 가족의 수가 많은 경우에 관한 결합분포의 계산은 상당히 번거로운 작업이기 때문에, L_M^R을 계산하는 것에 대하여 상당한 이득이 주어진다.

3. 2. 3 주변우도를 이용한 추정방법

자료로부터 대립형질에 관한 빈도를 추정하는 방법을 살펴보면 다음과 같다. 하아디-와인버그 평형의 가정으로부터, 인구집단에서 알고 있는 사실은, 변이가 일어날 확률이 $1-(1-f)^2$ 이라는 것이다. 그러므로 만약 p를

지원자들이 변이가 일어날 확률이라고 정의하면, 지원자가 랜덤하게 뽑힌다는 가정 하에, $1-(1-f)^2=p$ 의 방정식을 품으로써, f 를 추정할 수 있다.

변이가 희귀한 경우, 즉, $f^2 \approx 0$, 이라면, $\hat{f}=p/2$ 가 된다(Wacholder et al., 1998).

다른 방법으로서, θ와 f의 관점에서 결합적으로 주변우도함수를 최대화시킬 수 있다. 친족들로부터 f에 대한 정보는 비록 작기는 하지만, 두 번째 접근방법에서 통합되어질 수 있다.

추정치의 점근적 정규성은 관련된 자료에 대한 존재정리(existing theory)로부터 성립되어질 수 있다(Diggle, Liang and Zeger, 1996).

i 번째 가족에 대한 θ-점수와 f-점수의 합을 아래와 같은 식으로 정의한다.

$$S_{i+}(\theta,f)=\sum_{j=1}^{n_i} S_{ij}^R(\theta,f)+S_i^P(\theta,f)$$

$$u_{i+}(\theta,f)=\sum_{j=1}^{n_i} u_{ij}^R(\theta,f)+u_i^P(\theta,f)$$

이는 $K\to\infty$, $K^{1/2}\{(\hat{\theta},\hat{f})-(\theta,f)\}$는 다변량정규분포로 수렴하는데 평균은 0 이고 분산-공분산 행렬은 다음과 같다.

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & a_{22} \end{bmatrix}^{-1} \begin{bmatrix} B_{11} & B_{12} \\ B_{12}^T & b_{22} \end{bmatrix} \begin{bmatrix} A_{11} & A_{21}^T \\ A_{12}^T & a_{22} \end{bmatrix}^{-1}$$

여기서,

$$A_{11}=-\lim_{K\to\infty} K^{-1}\frac{\partial}{\partial\theta^T}\sum_{i=1}^{K} S_{i+}(\theta,f)$$

$$A_{12} = -\lim_{K\to\infty} K^{-1}\frac{\partial}{\partial f}\sum_{i=1}^{K} S_{i+}(\theta, f)$$

$$A_{21} = -\lim_{K\to\infty} K^{-1}\frac{\partial}{\partial \theta^T}\sum_{i=1}^{K} u_{i+}(\theta, f)$$

$$a_{22} = -\lim_{K\to\infty} K^{-1}\frac{\partial}{\partial f}\sum_{i=1}^{K} u_{i+}(\theta, f)$$

$$B_{11} = \lim_{K\to\infty} K^{-1}\sum_{i=1}^{K} var\{S_{i+}(\theta, f)\}$$

$$B_{12} = \lim_{K\to\infty} K^{-1}\sum_{i=1}^{K} cov\{S_{i+}(\theta, f), u_{i+}(\theta, f)\}$$

$$b_{22} = \lim_{K\to\infty} K^{-1}\sum_{i=1}^{K} var\{u_{i+}(\theta, f)\}$$

f가 $1-(1-f)^2$으로 간단히 풀어져서 추정되어질 때, 수식이 간단해 진다. $A_{21}=0, B_{12}=0$, $a_{22}=b_{22}=4/\{1-(1-f)^2\}$ 이 되는 사실을 이용하여, $\hat{\theta}$의 근사적인 분산은 아래의 식으로 나타낼 수 있다.

$$A_{11}^{-1}\{B_{11} + \frac{1-(1-f)^2}{4} A_{12}A_{12}^T\}A_{11}^{-1}$$

그러므로 f의 추정치가 $\hat{\theta}$의 분산을 증가시킨다는 것을 알 수 있는데 즉, 분산의 증가는 희귀한 변이가 작다는 것을 의미한다. (Chatterjee et al, 2001).

f의 값을 고정시킨 것에 대하여, 주변우도(marginal likelihood)는 EM 알고리즘을 이용하여 $\theta=(\theta_0, \theta_1)$ 의 관점에서 최대화를 할 수 있다. 단순하

게 하기 위하여, y의 밀도함수 $q(y;\theta)$가 비보균자 θ_0 와 보균자 θ_1에 대하여 같은 모수적인 형태를 갖고, 또한, 모수값은 독립적으로 변화한다고 가정한다.

제 4 장 표현형의 분포에 따른 추정기법

4. 1 이분형(dichotomous)자료에 대한 추정기법

4. 1. 1 유사우도를 이용한 추정방법

표현형 Y_0 와 Y_1 이 1 또는 0 (질병의 유/무)을 가지는 이분형자료(dichotomous data)라고 한다면, 침투율 φ_0 와 φ_1을 아래와 같은 식으로 나타낼 수 있다(Moore et al, 2001).

$$f(y|g=0) = \varphi_0{}^y(1-\varphi_0)^{1-y}$$
$$f(y|g=1) = (\varphi_1)^y(1-\varphi_1)^{1-y}$$

EM 알고리즘의 M-step 은 추정치를 다음과 같이 산출한다.

$$\hat{\varphi}_0 = \frac{\sum_i \sum_j y_{1ij}(1-\hat{g_{1ij}})}{\sum_i \sum_j (1-\hat{g_{1ij}})} \tag{4.1}$$

$$\hat{\varphi}_1 = \frac{\sum_i \sum_j y_{1ij}\hat{g_{1ij}}}{\sum_i \sum_j \hat{g_{1ij}}} \tag{4.2}$$

여기서 $\hat{g_{1ij}} = E(g_{1ij}|\ \tilde{y_{1i}}, g_{0i}; \hat{\pi}, \hat{\varphi})$는 E-Step에서 계산된 것으로서 식 (3.7)에

서 도출된 값이 쓰인다.

4. 1. 2 주변우도를 이용한 추정방법

f의 값을 고정시킨 것에 대하여, 주변우도함수는 EM 알고리즘을 이용하여 $\theta=(\theta_0,\theta_1)$의 관점에서 최대화를 할 수 있다(Chatterjee et al, 2001). 단순하게 하기 위하여, y의 밀도함수 $q(y;\theta)$가 비보균자 θ_0와 보균자 θ_1에 대하여 같은 모수적인 형태를 갖고, 또한, 모수값은 독립적으로 변화한다고 가정한다.

그렇다면 EM 알고리즘의 i 번째 반복은 아래와 같은 단계를 포함하게 된다(Chatterjee et al, 2001).

E-Step은 상수를 정규화 할 때까지 가중치에 대한 두 집합을 정의한다.

$$W_{0i}^{P}=I(g_i^{P}=0)$$

$$W_{0ij}^{R}=pr(g_{ij}^{R}=0\mid g_i^{P})q(y_{ij}^{R};\hat{\theta}_0^{(i-1)})$$
$$\div\,[pr(g_{ij}^{R}=0\mid g_i^{P})q(y_{ij}^{R};\hat{\theta}_0^{(i-1)})$$
$$+\,pr(g_{ij}^{R}=1\mid g_i^{P})q(y_{ij}^{R};\hat{\theta}_0^{(i-1)})]$$

$$W_{1i}^{P}=1-W_{0i}^{P}\quad,\quad W_{1ij}^{R}=1-W_{0ij}^{R}$$

여기서, $W_{0i}^{R}=(W_{0i1}^{R},\cdots,W_{0in_i}^{R})$, $W_{1i}^{R}=(W_{1i1}^{R},\cdots,W_{1in_i}^{R})$ 이다.

M-step에서는, 가중치의 특정한 집합을 갖는 독립적인 관찰치들의 집합으로부터 모형 $q(y,\theta)$와 일치하는 θ의 최대우도추정치 $\hat{\theta}_0^{(i)}$ 와 $\hat{\theta}_1^{(i)}$를 얻기 위하여 다음과 같은 방법을 사용한다.

가중치 $W_0=(W_{01}^{R},W_{01}^{P},\cdots,W_{0K}^{R},W_{0k}^{P})$,$W_1=(W_{11}^{R},W_{11}^{P},\cdots,W_{1K}^{R},W_{1K}^{P})$

을 이용하여 $Y=(Y_1^R, Y_1^P, \cdots, Y_K^R, Y_K^P)$의 자료로부터 최대우도추정치를 얻는다.

4. 2 생존시간에 대한 자료인 경우의 추정기법

4. 2. 1 유사우도를 이용한 추정방법

$t_{1ij}(i=1,\cdots,I\ ,\ j=1,\cdots,m)$는 질병의 발생여부 또는 중도절단의 시점에서 친족의 나이로 정의한다. δ_{1ij}는 질병의 발생여부의 가변수(indicator)로 정의한다.

지원자에 관해서는 t_{0i} , δ_{0i} 로 정의한다. 그러면 표현형 y는 (t,δ)쌍으로 나타낼 수 있다.

절단점(cut-point) $\nu_0=0, \nu_1, \nu_2, \cdots, \nu_k$ 를 가지는 조각별 상수(piecewise constant)로서 보균자/비보균자들에 관한 질병의 위험도(disease hazard)를 모델링 할 수 있다.

이러한 절단점(cut-point)는 $[\nu_0,\nu_1), [\nu_1,\nu_2), \cdots, [\nu_k,\infty)$ 과 같은 구간으로서 정의할 수 있다. 다만, 모수적 경우에서, 절단점(cut-point)의 수와 위치는 사전에 정의해야 한다.

$\lambda_1^g, \lambda_2^g, \cdots, \lambda_k^g$ 는 보균자/비보균자$(g=1/0)$에 대한 위험도라 한다.

$\gamma_i(t)$는 각각의 개인 i 가 나이가 t시점인 경우 위험도의 유무에 따른 표시함수(indicator function)를 의미한다. 그렇다면, 그 사람이 나이 t 시점 전에 질병을 가지고 있다고 관찰되지 않을 확률값은 다음과 같은 식으로 나타낼 수 있다(Moore et al, 2001).

$$S_i^g(t_i) \equiv S(t_i;\tilde{\lambda}_0, \tilde{\lambda}_1, g_i) = exp(-\int_0^{\infty}\gamma_i(t)[g_i\lambda^1(t_i)+(1-g_i)\lambda^0(t_i)]dt)$$

$$= exp(-\sum_{l=1}^{k}\int_{\nu_{l-1}}^{\nu_l}\gamma_i(t)[g_i\lambda_l^1+(1-g_i)\lambda_l^0]dt$$

$$= exp(-\sum_{l=1}^{k}[g_i\lambda_l^1+(1-g_i)\lambda_l^0]V_{il}) \qquad (4.3)$$

여기서 $V_{il}=\int_{\nu_{l-1}}^{\nu_l}\gamma_i(t)dt$ 는 i 번째 개인이 l 번째 구간에서 소비한 연도(person year)의 수라고 정의할 수 있다. 여기서, 어떠한 중도절단(censoring)도 유전형에 독립이라고 가정한다.

위의 가정 하에서 , 식 (3.3)에서의 $f(y_{1j}\mid g_{1j};\varphi)$를 $\lambda^{g_{1j}}(t_{1j})^{\delta_{1j}}S^{g_{1j}}(t_{1j})$로서 대체할 수 있다. 또한, 식 (3.2)에서의 $f(y_0\mid g_0;\varphi)$를 $\lambda^{g_0}(t_0)^{\delta_0}S^{g_0}(t_0)$로서 대체할 수 있다. 덧붙여서, 질병의 발생이 보균자인 상태에 독립이라는 것을 따른다(Gail et al., 1999).

보균자/비보균자($g_{ij}=0/g_{ij}=1$)에 대한 l 번째 구간을 1사람당 단위년(person-year)의 식, PY_l^{1NC} 와 PY_l^{1C} 으로 정의하고, 구간당 사망을 D_l^{1NC} , D_l^{1C} 라 정의한다. 즉,

$$PY_l^{1NC}=\sum_i\sum_j(1-g_{1ij})V_{1ijl}$$

$$D_l^{1NC}=\sum_i\sum_j\delta_{1ij}(1-g_{1ij})I[\nu_{l-1}<t_{1ij}\leqq\nu_l]$$

$$PY_l^{1C}=\sum_i\sum_j g_{1ij}V_{1ijl}$$

$$D_l^{1C} = \sum_i \sum_j \delta_{1ij} g_{1ij} I[\nu_{l-1} < t_{1ij} \leqq \nu_l] \tag{4.4}$$

식 (4.4)에서 $V_{1ijl} = \int_{\nu_{l-1}}^{\nu_l} \gamma_{1ij}(t)dt$ 이다.

M-step은 위험도(hazard)의 완전자료에 대한 최대우도함수(단, 친족에 기초한)에 의해 $\widehat{g_{1ij}}$ 를 이용하여 아래와 같이 정의되어진다(Moore et al, 2001).

$$\widehat{\lambda_{0l}} = \frac{D_l^{1NC}}{PY_l^{1NC}} \tag{4.5}$$

$$\widehat{\lambda_{1l}} = \frac{D_l^{1C}}{PY_l^{1C}} \tag{4.6}$$

E-Step에 대해서는 식 (3.7)에 의해서 정의한대로 보균자 상태의 조건부 기대값(conditional expection)을 이용한다.

4. 2. 2 주변우도를 이용한 추정방법

질병이 발생하는 시점(나이)에 대한 자료일 경우 특정시점에서의 누적위험도(age-specific cumulative risk)를 추정하기 위하여 다음과 같이 고려한다(Chatterjee et al, 2001).

보균자/비보균자에대한 나이 t 시점까지의 누적위험도를 다음과 같이 정의한다.

$$F_g(t) = 1 - S_g(t) \ , \ g = 0, 1$$

또한 $h_g(t)$, $g=0,1$ 은 위험함수(hazard function)라 정의한다.

$Y=(T,\delta)$, 즉, T 는 병이 발생하는 나이 또는 추적관찰종료시의 나이를 의미하며, δ는 추적관찰종료가 될 때까지의 병의 발생여부에 대한 표시변수이다. 각각의 개인에 대한 중도절단시점(censoring time)의 분포가 그들의 유전형에 의존하지 않는다고 가정하면, L_M^R과 L^P는 다음의 식에 비례하게 나타난다.

$$\Pi_{i=1}^{K}\Pi_{j=1}^{n_i}\sum_{g_{ij}^R=0}^{1} pr(g_{ij}^R \mid g_i^P) S_{g_{ij}^R}(t_{ij}^R) h_{g_{ij}^R}^{\delta_{ij}^R}(t_{ij}^R)$$

$$\Pi_{i=1}^{K} S_{g_i^P}(t_i^P)\{h_{g_i^P}(t_i^P)\}^{\delta_i^P}$$

누적위험도에 관한 모수적 모형(parametric model)은 위에 기술한 EM 알고리즘으로부터 추정되어질 수 있다.

Wacholder 등(1998)은 $F_g(t)$를 추정하는 방법으로 Kaplan-Meier 방법을 제시하였다. 즉, 드문 변이에 대하여,

$$R_0(t) = (1-f)F_0(t) + fF_1(t)$$
$$R_1(t) = (0.5-f/2)F_0(t) + (0.5+f/2)F_1(t)$$

여기서 $R_g(t)$, $g=0,1$은 비보균자와 보균자의 친족(first-degree relative)들에 대한 누적위험도라고 정의한다. 여기서, 그들이 제안한 방법은 근사방정식(approximate equation)에서 R_0 와 R_1 대신에 Kaplan-Meier 추정치를 사용하여 $F_0(t)$와 $F_1(t)$를 풀었다는 점이다. 그러나 이 방법은, 누적위험도(cumulative risk)의 결과추정치가 유한한 표본(finite sample)에서

단조(monotone)할 것이라는 점을 보장할 수 없다는 것이다(Chetterjee et al., 2001).

실제로, 변이가 드문 경우, 이 방법은 F_1이 비증가하는 추정치가 발생하였다(Struewing et al., 1997; Wacholder et al., 1998).

S_0와 S_1 의 적절한 비모수적 추정치를 얻기 위하여, Chatterjee등(2001)은 주변우도함수를 비모수적으로 나타낼 것을 제안하였다.

$M = \{t_1 < t_2 < \cdots < t_M\}$ 을 자료에서 관찰된 사건이 발생하는 시점(event time)으로 정의한다. 어떤 T값에 관하여, $l(t)$는 T 와 같거나 또는 T 보다 적은 가장 큰 사건발생시점의 색인이라 정의하자.

$\hat{S_0}$, $\hat{S_1}$은 관찰된 사건발생시점 M 내에서 잠재적인 jump가 일어날 것이다. 또한, $\hat{S_g}$ 의 위험요인(hazard component)으로서 $\{\hat{\lambda_{g1}}, \cdots, \hat{\lambda_{gM}}\}$ 으로 정의한다.

비모수적 최대주변우도함수추정치(nonparametric maximum marginal likelihood; NPMMLE)를 구하기 위해, 전에 기술한 EM 알고리즘을 적용시킨다(Chetterjee et al., 2001).

E-step에서는 $f(y, \theta_g)$대신에 아래의 식을 넣는다.

$$\{\Pi_{m \leq l(T)} (1 - \hat{\lambda_{gm}})\}^{1-\delta} \{\hat{\lambda_{gl(T)}} \Pi_{m < l(T)} (1 - \hat{\lambda_{gm}})\}^{\delta}$$

M-step은 $\hat{\lambda}_{gm} = \sum_{i \in \epsilon(t_m)} w_{gi} / \sum_{i \in R(t_m)} w_{gi}$ 을 가지는 폐쇄형의(closed-form) 근으로서 나타낼 수 있다(Chetterjee et al., 2001). 여기서 $R(t_m)$은 위험도로서, 개인에 대한 색인의 집합, $T \geq t_m$ 인 경우를 의미한다. 또한, $\epsilon(t_m)$은 시점 t_m에서 사건(event)을 가지는 개개인의 집합을 의미한다.

여기서 주목할 점은, 만약 우리가 실패시점(faiure time)에 관한 자료가 이산형의 형태를 갖는다면, 추정되는 모수의 수는 비록 크게는 가능하지만, 고정된 상태로 유지되어야 한다. 그러나, 만약 연속형 이라면, 사건에 대한 시점(event time)의 수와 향후에 추정되어지는 모수의 수가 표본크기가 증가할수록 커지게 된다(Chetterjee et al, 2001)

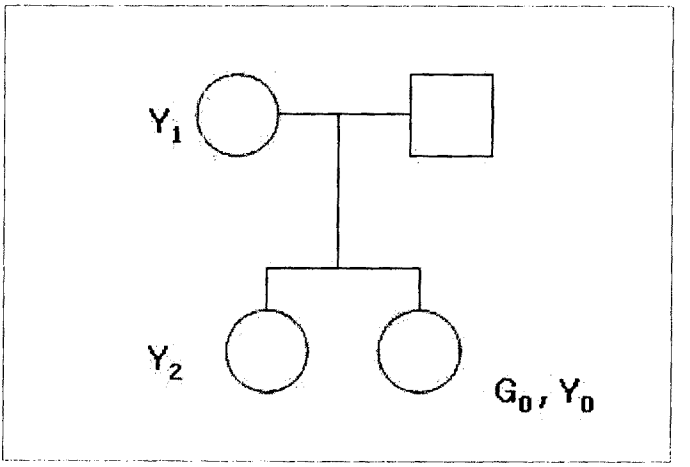
이러한 비모수적 방법에 관한 근사적인 성격은 추후 연구과제로 제시될 수 있다.

제 5 장 모의실험기법과 모형

5. 1 모의 가계도자료의 작성

모의실험을 통한 유사우도함수와 기존의 환자-대조군 연구와 비교를 하기 위해서 표현형(phenotype)이 이분형(dichotomous)이라고 가정하고 가계도(pedigree)자료를 생성하였다.

가계도 자료를 생성하기 위해서는 먼저 가계도를 가정해야 한다. 가계도는 아래 그림과 같은 형태로 가정한다. 즉, 유전형(g_0)과 표현형(Y_0)을 모두 조사하는 지원자(proband)는 자식이 되고, 표현형(Y_1, Y_2)만을 조사하는 친족(first-degree relative)은 어머니와 형제가 된다.



[그림 2] 연구설계

먼저, 지원자의 유전형(genotype) g_0 를 생성한다. g_0 는 대립형질빈도(allele frequency) π 를 갖는 베르누이 분포를 고려할 수 있다. 즉, 다음과

같은 식으로 생성할 수 있다.

$$g_0 = 1 \ or \ 0 \ \sim B(1,1,\pi)$$

지원자의 유전형이 생성되면 표현형(phenotype) Y_0를 생성해야 한다. 표현형은 침투율의 정의에 따라 $g_0 = 0$ 또는 $g_0 = 1$ 에 따라 아래와 같이 생성할 수 있다.

$$Y_0 = \begin{cases} B(1,1,\varphi_1) & \text{if} \quad g_0 = 1 \\ B(1,1,\varphi_0) & \text{if} \quad g_0 = 0 \end{cases}$$

Y_1 과 Y_2 를 생성하기 위해서는 g_0가 주어졌을때 Y_1, Y_2의 조건부확률(conditional probability)을 구해야 하는데 이를 도출하기 위해서는 먼저 각 가족에 대한 결합확률(joint probability)을 구해야 한다.

g_F 는 아버지의 유전형(genotype), g_F는 어머니의 유전형, g_s 는 딸(sister)의 유전형을 의미한다.

$$P(g_F, g_M, g_s, g_0) = P(g_F, g_M)\, P(g_s, g_0 \mid g_F, g_M) \qquad (5.1)$$

여기서 g_F 와 g_M 은 서로 독립이기 때문에,

$$= P(g_F)\, P(g_M)\, P(g_s \mid g_F, g_M)\, P(g_0 \mid g_F, g_M) \qquad (5.2)$$

$P(g_s \mid g_F, g_M)$ 은 다음과 같이 멘델리안 계산법으로 나타낼 수 있다(Li,

1976).

아래 [표 1] 를 보면, 확률값이 의미하는 것은 아버지가 AA 이고 어머니가 Aa 일 때, 자식은 AA 또는 Aa가 나오므로, AA가 될 확률값은 0.5를 의미한다. 즉, $P(g_S = AA \mid g_F = AA, g_M = Aa) = 0.5$ 를 만족한다는 것이다.

Genotype			Prob
Father	Mother	Sister	
AA	AA	AA	1.0
AA	AA	Aa	0.0
AA	AA	aa	0.0
AA	Aa	AA	0.5
AA	Aa	Aa	0.5
AA	Aa	aa	0
AA	aa	AA	0
AA	aa	Aa	1.0
AA	aa	aa	0
Aa	AA	AA	0.5
Aa	AA	Aa	0.5
Aa	AA	aa	0
Aa	Aa	AA	0.25
Aa	Aa	Aa	0.50
Aa	Aa	aa	0.25
Aa	aa	AA	0
Aa	aa	Aa	0.5
Aa	aa	aa	0.5
aa	AA	AA	0
aa	AA	Aa	1.0
aa	AA	aa	0.0
aa	Aa	AA	0
aa	Aa	Aa	0.5
aa	Aa	aa	0.5
aa	aa	AA	0
aa	aa	Aa	0
aa	aa	aa	1.0

표 1 . 각 가족구성원의 유전형에 따른 확률값

그러나 g_F 는 우리가 가정했던 가계도(pedigree)에서는 실질적으로 알 수 없기 때문에 주변적으로(marginal) 계산 할 수 있다.

예를 들면 , $P(g_S = AA \mid g_M = AA)$ 는 아래와 같이 계산할 수 있다.

$$P(g_S = AA \mid g_F = AA, g_M = AA) + \cdots + P(g_S = AA \mid g_F = aa, g_M = AA)$$

식 (5.2)의 $P(g_0 \mid g_F, g_M)$ 의 경우도 역시 지원자(volunteer)와 형제와 동일한 식으로 나타나게 된다.

위의 결합 확률에 기초하여 조건부확률 $P(g_1 \mid g_0)$는 다음과 같이 나타낼 수 있다.

$$P(g_1 \mid g_0) = P(g_F, g_M, g_s \mid g_0) = \frac{P(g_F, g_M, g_s, g_0)}{P(g_0)}$$

조건부확률 $P(g_1 \mid g_0)$가 대립형질 π 의 함수형태로 나타나면 Y_1, Y_2를 생성할 수 있다.

5. 2 모의자료를 이용한 투과율의 추정분석

유사우도함수를 이용하여 추정하기 위해서는 먼저 지원자(proband)에 대한 우도함수를 설정해야 한다. 즉, 다시 말하면 $\tilde{\varphi}$ 를 고정시킨 상태에서, π 에 관한 최대우도함수를 구해야 한다는 것이다.

지원자의 유전형과 표현형에 관한 우도함수 l_1 은 아래와 같다.

$$l_1(\pi) = \prod_{i=1}^{n} \left[\frac{\pi^{g_0}(1-\pi)^{1-g_0} f(y_0 \mid g_0;\varphi)}{\pi f(y_0 \mid g_0=1) + (1-\pi) f(y_0 \mid g_0=0)} \right]$$

우도함수에 로그를 취하면,

$$\ln l_1(\pi) = \sum_{i=1}^{n} ln \left[\frac{\pi^{g_0}(1-\pi)^{1-g_0} f(y_0 \mid g_0)}{\pi f(y_0 \mid g_0=1) + (1-\pi) f(y_0 \mid g_0=0)} \right]$$

위의 로그우도함수(log-likelihood)를 π에 대하여 아래와 같이 미분하고, 최대값을 찾기 위해서 Newton-Raphson 방법을 이용하여 π의 추정치를 구한다.

$$\frac{\partial \ln l_1(\pi)}{\partial \pi}$$
$$= \sum_{i=1}^{n} \left[\frac{g_0}{\pi} - \frac{1-g_0}{1-\pi} - \frac{f(y_0 \mid g_0=1) - f(y_0 \mid g_0=0)}{\pi f(y_0 \mid g_0=1) + (1-\pi) f(y_0 \mid g_0=0)} \right] = 0$$

π의 추정치를 구하면, 친족(first-degree relative)에 대한 자료를 가지고 EM 알고리즘을 적용시켜서 $\tilde{\varphi} = (\varphi_0, \varphi_1)$을 추정한다.

E-step은 식 (3.7)을 이용하여 구할 수 있고, M-Step은 식 (4.1), (4.2)를 이용하여 $\tilde{\varphi}$를 구할 수 있다. 이렇게 구한 추정식에 대하여 수렴여부를 판정하여, 수렴이 될 때까지, 계속 반복시키는데, 허용한계는 10^{-7} 까지, 최대 반복회수는 100회를 두었다.

5. 3 환자-대조군 연구 추정기법

환자-대조군 연구란 특정한 질병에 관한 노출(exposure)의 가능한 관계(relationship)을 살펴보기 위해, 질병을 가진 집단(환자군)과, 비교를 목적으로, 질병을 가지지 않은 집단(대조군)을 설정하는 것을 의미한다(Leon Gordis, 1996).

인구집단에서 $P(Y=1)$이 알려져 있고, $\epsilon_{ij}=P(g=i \mid Y=j)$ 라 정의하면, 베이즈 정리에 의해 φ_1을 다음과 같이 추정할 수 있다(Cornfiled, 1951).

$$\varphi_1 = \frac{P(Y=1)\epsilon_{11}}{P(Y=1)\epsilon_{11}+P(Y=0)\epsilon_{10}}$$

ϵ_{11} 과 ϵ_{10} 는 n명의 환자군과 m명의 대조군의 랜덤표본에 대한 유전형 조사를(genotyping)함으로서 추정되어질 수 있으나, 인구집단에서 질병이 발생할 확률값이 알려져 있지 않다면, 대립형질의 빈도(allele frequency)를 통하여 φ_0, φ_1을 추정할 수 있다(Gail et al, 1999).

$P(Y=1)$을 모르는 경우, 다음과 같은 식으로 나타낼 수 있다.

$$P(Y=1)=\pi\varphi_1+(1-\pi)\varphi_0$$

실제로 φ_0, φ_1을 추정하기 위해서는 다음 2개의 방정식을 풀어야 한다.

$$\begin{aligned}\varphi_1 = P(Y=1 \mid g=1) &= \frac{P(Y=1)P(g=1 \mid Y=1)}{P(g=1)} \\ &= \frac{P(Y=1)\epsilon_{11}}{\pi}\end{aligned} \tag{5.1}$$

$$1-\varphi_1 = P(Y=0 \mid g=1) = \frac{P(Y=0)\epsilon_{10}}{\pi} \qquad (5.2)$$

식 (5.1), (5.2)를 풀면,

$\varphi_1 = \frac{\epsilon_{11}(\pi-\epsilon_{10})}{\pi(\epsilon_{11}-\epsilon_{10})}$, $\varphi_0 = \frac{\epsilon_{01}((1-\pi)-\epsilon_{00})}{(1-\pi)(\epsilon_{01}-\epsilon_{00})}$ 을 구할 수 있다.

5. 4 모의실험을 통한 추정치 비교결과

5. 4. 1 지원자의 수의 변화에 대한 비교

시조의 수(number of probands)를 500에서 5000까지 생성하면서, 각각의 경우에 따라 유사우도함수 방법과 환자-대조군 연구방법으로 각각의 침투율을 모의실험으로 추정하였다. 모의실험의 반복횟수는 100번으로 하였고, 유사우도함수 추정기법에서 π 및 φ_0, φ_1 을 추정하기 위한 EM 알고리즘의 수렴여부를 확인하기 위한 최대 허용한계는 10^{-7}까지, 최대 반복회수는 100회로 두었다. 생성한 가계도자료의 실제 값(true value)은 변이가 희귀한(rare)경우로서, $P(A)=0.01$로 두고, $\varphi_0=0.1$, $\varphi_1=0.9$ 로 생성하였다.

각 추정기법에 따른 모의실험의 평균 및 표준편차는 [표 2]와 같다.

모의실험을 한 결과, 시조의 수가 500일 경우의 유사우도함수를 이용한 투과율을 추정하였을 때에는 $\hat{\varphi_1}=0.8791\pm0.1084$ 이나, 환자-대조군 방법을 이용한 추정치는 0.8409±0.1337 로서 유사우도함수를 이용한 추정방법이 환자-대조군 방법보다 정확하게 추정되는 것을 보여주고 있다.

그러나, 시조의 수가 3000개 이상인 경우에는 두 방법을 이용한 추정치의 차이가 거의 없음을 알 수 있다.

표 2. 변이가 희귀한 경우 시조의 수에 따른 투과율의 추정결과

number of probands	Pseudo-likelihood			Case-control	
	Estimate Mean ± SD			Estimate Mean ± SD	
	$\hat{\varphi_1}$	$\hat{\varphi_0}$	$P(A)$	$\hat{\varphi_1}$	$\hat{\varphi_0}$
500	0.8792±0.1084	0.0965±0.0121	0.0065±0.0019	0.7883±0.1785	0.0600±0.0222
1000	0.9029±0.0690	0.1113±0.0399	0.0801±0.0024	0.9303±0.0135	0.1369±0.0229
2000	0.9036±0.0589	0.1020±0.0542	0.0580±0.0035	0.9332±0.0097	0.1438±0.0170
3000	0.9085±0.0526	0.1025±0.0543	0.0543±0.0041	0.9333±0.0063	0.1473±0.0134
4000	0.8952±0.0434	0.1141±0.0403	0.0100±0.0580	0.9333±0.0088	0.1473±0.0113
5000	0.8845±0.0187	0.1282±0.0051	0.0590±0.0030	0.9338±0.0062	0.1479±0.0098

* number of iteration=100, true paramater $\varphi_0 = 0.1$, $\varphi_1 = 0.9$, $P(A) = 0.01$

표 3. 변이가 희귀하지 않은 경우 시조의 수에 따른 투과율의 추정결과

number of probands	Pseudo-likelihood			Case-control	
	Estimate Mean ± SD			Estimate Mean ± SD	
	$\hat{\varphi_1}$	$\hat{\varphi_0}$	$P(A)$	$\hat{\varphi_1}$	$\hat{\varphi_0}$
500	0.8791±0.1084	0.0965±0.1211	0.0065±0.0019	0.8409±0.1337	0.0837±0.0305
1000	0.9164±0.0533	0.0986±0.0077	0.0063±0.0012	0.8525±0.0855	0.0805±0.0188
2000	0.8982±0.1006	0.1007±0.0034	0.0060±0.0010	0.8638±0.0889	0.0849±0.0191
3000	0.8796±0.0871	0.1011±0.0047	0.0058±0.0009	0.8659±0.0589	0.0871±0.0146
4000	0.8787±0.0795	0.1017±0.0028	0.0058±0.0006	0.8741±0.0563	0.0872±0.0107
5000	0.8852±0.0607	0.1021±0.0033	0.0058±0.0006	0.8833±0.0450	0.0856±0.0098

* number of iteration=100, true paramater $\varphi_0 = 0.1$, $\varphi_1 = 0.9$, $P(A) = 0.1$

[표 3]은 변이가 드물지 않은 경우에 대하여 시조의 수에 따른 투과율의 추정결과를 나타낼 것이다. 전체적으로 유사우도함수를 이용한 방법이 환자-대조군 연구를 이용한 방법보다 추정치의 결과가 실제 값과 근사함을 알 수 있다.

5. 4. 2 변이 대립형질 빈도의 변화에 대한 비교

변이 대립형질에 따른 투과율의 추정결과는 아래 [표 4]와 같다.

[표 4]를 보게 되면 변이 대립형질(mutation allele frequency) $P(A)$가 0.03 인 경우, 유사우도함수기법을 이용한 추정치는 φ_0=0.1079, 환자-대조군 연구인 경우 φ_0=0.0918 로서 거의 비슷하나, φ_1인 경우에는 유사우도함수기법은 0.8982, 환자-대조군인 경우, 0.8786으로서 유사유도함수의 추정치가 더욱 실제 값과 비슷함을 알 수 있다.

$P(A)$가 0.04 이후에는 환자-대조군 연구에서 추정된 투과율과 거의 비슷함을 알 수 있다.

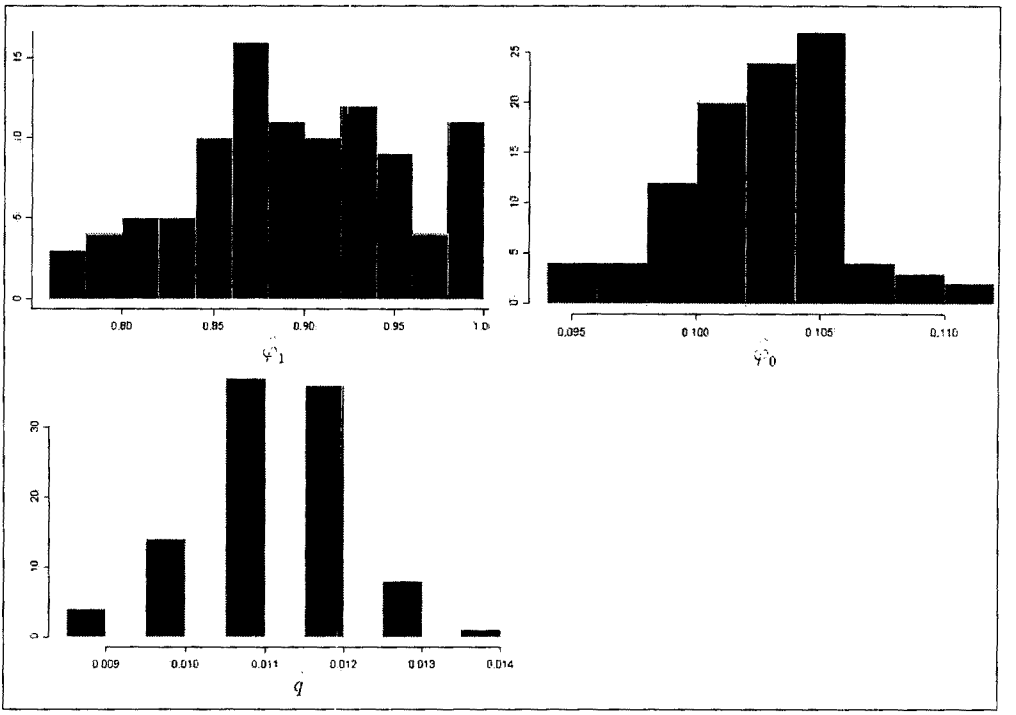
아래 [그림 3]은 유사우도함수기법으로 $P(A)=0.01$, $\varphi_0=0.1$, $\varphi_1=0.9$, $n=5000$ 으로 모의자료를 생성하고, 100번 반복하여 추정한 투과율 및 변이 대립형질의 빈도를 나타낸 것이다.

φ_1, φ_0, q 의 빈도가 대체적으로 실제 값과 비슷하게 분포되어 있음을 알 수 있다.

표 4 . $P(A)$에 따른 투과율의 추정결과

allele frequency	Pseudo-likelihood			Case-control	
	Estimate Mean ± SD			Estimate Mean ± SD	
	$\hat{\varphi_1}$	$\hat{\varphi_0}$	$P(A)$	$\hat{\varphi_1}$	$\hat{\varphi_0}$
0.01	0.8852±0.0607	0.1021±0.0032	0.0058±0.0006	0.8755±0.0480	0.0801±0.0093
0.02	0.9028±0.0455	0.1057±0.0041	0.0115±0.0011	0.8782±0.0284	0.0855±0.0104
0.03	0.8982±0.0472	0.1079±0.0059	0.0172±0.0010	0.8786±0.0183	0.0918±0.0066
0.04	0.8978±0.0398	0.1115±0.0040	0.0226±0.0014	0.8914±0.0176	0.0932±0.0068
0.05	0.9161±0.0405	0.1024±0.0364	0.0301±0.1849	0.8964±0.0174	0.0962±0.0085
0.06	0.8941±0.0207	0.1166±0.0039	0.0349±0.0025	0.8950±0.0124	0.0936±0.0078
0.07	0.8916±0.0205	0.1191±0.0039	0.0408±0.0028	0.8939±0.0094	0.0953±0.0077
0.08	0.8956±0.0352	0.1219±0.0042	0.0465±0.0014	0.8965±0.0010	0.0961±0.0053
0.09	0.8840±0.0252	0.1240±0.0055	0.0527±0.0028	0.8954±0.0098	0.0962±0.0065
0.10	0.8845±0.0187	0.1282±0.0051	0.0591±0.0030	0.8969±0.0096	0.0968±0.0065

* number of iteration=100, true paramater $\varphi_0 = 0.1$, $\varphi_1 = 0.9$, $n = 5000$



[그림 4] 유사우도함수추정기법을 이용한 추정치의 빈도

위의 결과를 종합하여 볼 때, 유사우도함수추정기법을 이용하여 투과율을 추정하는 것은 기존의 환자-대조군 연구와 비교해 볼 때, 조사해야 할 시조의 수(number of proband)에 덜 민감하고, 변이가 드문 경우, 즉, $P(A)$가 작은 경우에도 투과율을 추정하는데 있어서 보다 정확한 추정치를 구할 수 있다는 것을 위의 모의실험결과를 통하여 뒷받침 할 수 있다.

제 6 장 토의 및 결론

지금까지 GPD에서 침투율(penetrance)를 추정하기 위하여 기존의 환자-대조군 연구 및 유사우도함수, 주변우도함수를 이용한 추정기법을 소개하였고, 또한 모의실험을 통한 이분형 자료(dichotomous data)에 대한 유사우도함수기법과 환자-대조군 방법을 이용한 추정치에 대하여 설명하였다.

GPD(Genotyped-Proband Design)란, 우선 시조(proband)에 대해서는 유전형과 표현형을 모두 조사하고, 친족(first degree relative)에 관해서는 표현형만을 조사하여 가계도(pedigree)를 형성하는 연구를 의미한다.

GPD에서 침투율을 추정하기 위해서 Gail 등(1999)이 제안한 최대우도함수를 이용하는 방법부터, Moore 등(2001)이 제안한 유사우도(pseudo likelihood)를 이용한 추정방법, 또한 Chetterjee 등(2001)이 제안한 주변우도(marginal likelihood)를 이용한 추정 방법 등이 제시되었다.

또한 표현형의 형태에 따라, 즉, 이분형(dichotomous)인 경우와 생존시간이 고려된 경우에 따른 유사우도추정방법과 주변우도추정방법에 대하여 설명하였다.

본 논문의 주 관심인 유사우도를 이용한 추정기법에 대한 평가를 하기 위하여, GPD에 기초한 모의자료를 멘델리안 계산(Li, 1976)에 기초하여 S-PLUS를 이용하여 생성하였고, 추정결과를 환자-대조군 연구방법을 이용한 추정치와 비교를 하였다.

모의실험결과, 유사우도함수추정기법을 이용하여 투과율을 추정하는 것은 기존의 환자-대조군 연구에서의 추정기법과 비교해 볼 때, 조사해야 할 시조의 수(number of proband)에 덜 민감하고, 변이가 드문 경우, 즉, $P(A)$가 작은 경우에도 투과율을 추정하는데 있어서 보다 정확한 추정치를 구할 수 있다는 결론을 얻게 되었다.

본 연구에서는 이분형 자료에 대해서만 고려를 하였고, 생존시간은 고려를 하지 않았다. 선행연구의 예로서, Washington Ashkenazi 연구가 있었으나, 이러한 분석을 더욱 심화하기 위해서는 실재자료를 통하여 구체적으로 이분형과 생존시간에 대한 접근이 필요하다고 생각된다.

이외에도 GPD에서 발생할 수 있는 잔여 가족성 상관관계(residual familial correlation)의 존재나 하아디-와인버그 평형을 만족하지 않는 경우에 대한 추정기법 등에 대한 연구가 계속 진행되어져야 할 것이다.

참 고 문 헌

강동민 역, Larry Gonick, Mark Wheelis, 생물 유전학 길잡이, 도서판 국제, 1995

박종구, 현대역학, 연세대학교 출판부, 1999

송혜향, 정갑도, 이원철, 생존분석, 청문각, 1996

유종영,이승천,차경준,허문열, S-PLUS를 이용한 통계계산, 박영사, 1997

김정숙,나종화, S-PLUS 사용법 및 프로그래밍, 자유아카데미, 2000

Bruce S.Weir, Genetic data analysis II: Methods for discrete population genetic data, Sinauer Associates, 1996

Ching Chun Li, First course in population genetics, Pacific Grove, California: Boxwood, 1978

Dempster, A.P., Laird, N.M., and Rubin, D.B. , Maximum likelihood from incomplete data via the EM algorithm (with discussion), *JRSS Series B*, 1977;39;1-38

Dirk F. Moore, Nilanjan Chatterjee, David Pee, Mitchell H. Gail, Pseudo-likelihood estimates of the cumulative risk of an autosomal dominant disease from a kin-cohort study, *Genetic Epidemiology*, 2001;20;210-227

Douglas F.Easton, Deborah Ford, D. Timothy Bishop, and the Breast Cancer Linkage Consortium, Breast and Ovarian Cancer incidence in BRCA1-Mutation carriers, *Am. J. Hum. Genet*, 1995;56;265-271

Geoffrey J. McLachlan, Thyiyambakam Krishnan, The EM Algorithm and Extensions, A John Wiley & Sons, 1997

Jeffery P. Struewing et al, The risk of cancer associated with specific

mutations of BRCA1 and BRCA2 among Ashkenazi Jews, *New England Journal of Medicine*, 1997;Vol. 336;No.20;1401-1 408

Khoury MJ, Flanders WD., Bias in using family history as a risk factor in case-control studies of disease, *Epidemiology* 1995;6:511-19

Kung-Yee Liang, Scott L. Zeger, Bahjat Qaqish, Multivariate regression analyses for categorical data, *JRSS Series B*, 1992;54;No. 1;3-40

Leon Gordis, Epidemiology, W.B. Saunders Company, 1996

Li, H., Thompson, E., Semiparametric estimation of a major gene and family-specific random effects for age of onset, *Biometrics*, 1997;53;282 -293

Michell H. Gail, David Pee, Raymond Carroll, Kin-cohort designs for gene characterization, *Journal of National Cancer Institute Monographs*, 1999;26;55-60

Michell H. Gail, David Pee, Raymond Carroll, Effects of violations of assumptions on likelihood methods for estimating the penetrance of an autosomal dominant mutation from kin-cohort studies, *Journal of statistical planning and inference* , 2001; 167-177

Michell H. Gail et al., Designing Studies to estimate the penetrance of an identified autosomal dominant mutation: cohort, case-control, and genotyped- proband designs, *Genetic epidemiology* 1999;16:15-39

N. Chatterjee and S. Wacholder, A marginal likelihood approach for estimating penetrance from Kin-Cohort Designs, *Biometrics*, 1999; Vol. 57; No.1; 245-252

Risch N., Segregation analysis incorporating linkage markers. 1.Single-locus marker with an application to type I diabetes., *Am J*

Human Genet 1984;36:363-86.

Satten GA, Kupper LL., Inferences about exposure-disease associations using probability of exposure information., *J Am Stat Assoc* 1993;88:200-8.

Schatzkin A, Goldstein A, Freedman LS., What does it mean to be a cancer gene carrier? Problems in establishing causality from the molecular genetics of cancer., *J Natl Cancer Inst* 1995;87:1126-30.

Sholom Wacholder et al, The kin-cohort study for estimating penetrance, American Journal of Epidemiology, 1998;Vol. 148;No. 7;623-630

Steve Selvin, Modern applied biostatistical methods using S-PLUS, Oxford University Press, 1998

W.N.Venables, B.D.Ripley, Modern applied statistics with S-PLUS, Springer, 1999

ABSTRACT

Methods for estimating penetrance from the Genotyped-Proband Design

Myung, Sung Min
Dept. of Biostatistics and Computing
The Graduate School
Yonsei University

In the thesis, we developed the Genotyped-Proband Design(GPD), an alternative cross-sectional study that used volunteer or probands, for estimating the penetrance of a rare mutation. Especially, pseudo-likelihood (Moore et al, 2001), marginal-likelihood(Chatterjee et al, 2001) were main parts of the thesis about estimating penetrance.

For the analysis of simulated data using S-PLUS, we generate pedigree data with respect to GPD, using Mendelian calculation in order to evaluate estimating method of pseudo-likelihood. Comparing to estimating method based on pseudo-likelihood and case-control, we found two conclusions.

First, Pseudo-likelihood estimating method is less sensitive number of probands than case-control estimating method.

Second, Pseudo-likelihood estimating method is more consistent

estimate of penetrance than based on case-control approach in a rare mutation, small allele frequency $P(A)$.

Key Words : Genotyped-Proband Design, penetrance, kin-cohort, EM algorithm, proband, pseudo-likelihood, marginal-likelihood