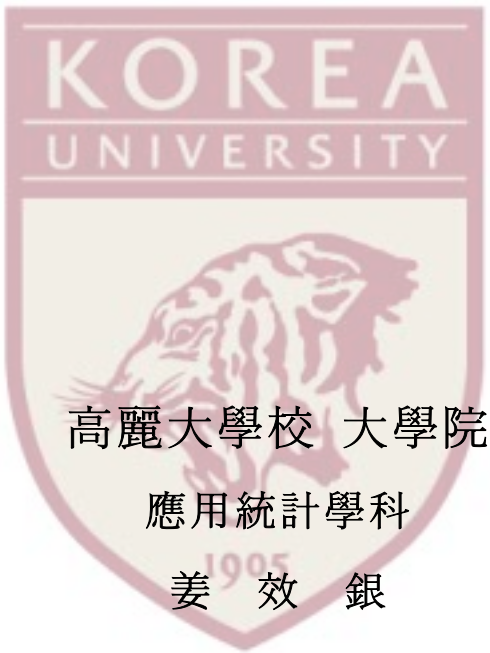
碩 士 學 位 論 文

로지스틱 회귀모형에서 SAMCIS 알고리즘을 이용한 정확추론



高麗大學校 大學院

應用統計學科

姜 效 銀

2017年 5月

全 秀 榮 敎 授 指 導

碩 士 學 位 論 文

로지스틱 회귀모형에서 SAMCIS 알고리즘을 이용한 정확추론

이 論文을 統計學 碩士學位 論文으로
提出함.

2017年 5月

高 麗 大 學 校 大 學 院

應 用 統 計 學 科

姜 效 銀 (印)

姜 效 銀의 統計學 碩士學位論文

審査를 完了함.

2017年 5月

委員長 전수영 (印)

委 員 진서윤 (印)

委 員 최보승 (印)

요 약 문

로지스틱 회귀모형은 범주형 데이터를 종속변수로 하여 종속변수와 독립변수 간의 관계를 구체적인 함수로 나타내는 예측 모형으로, 로지스틱 회귀모형에 대한 통계적 추론에서 종종 장애모수를 제외한 관심모수의 추론에 관심이 많다. 로지스틱 회귀모형에 대한 통계적 추론은 전형적으로 비조건부 우도함수에 근거한 대표본 근사를 사용한다. 하지만 소표본이거나 희박 자료(sparse data)인 경우 또는 불균형 자료(unbalanced data)인 경우 일반적으로 대표본 근사추론은 신뢰할 수 없다.

이러한 추론에는 비조건부 우도함수에 근거한 대표본 근사보다 정확추론이 사용된다. 정확 추론은 장애모수에 대한 충분통계량의 관측된 값을 고정시켰을 때 관심 있는 모수에 대한 충분통계량의 조건부 분포에 근거하는 방법으로 표본공간이 작거나 편향되어 있더라도 신뢰할 수 있다. 정확추론 연구는 Forster et al. (1996), Forster et al. (2003), Zamar et al. (2007) 등이 있지만, Markov chain Monte Carlo 방법을 이용하기 때문에 국소 트랩(local trap)의 문제점이 여전히 존재한다.

로지스틱 회귀모형에서 정확추론의 문제점을 극복하고자 본 연구는 Stochastic approximation Monte Carlo importance sampling(Cheon et al, 2014; SAMCIS) 알고리즘을 이용한 정확추론 방법을 제안하고자 한다. SAMCIS 알고리즘은 국소 트랩의 문제점을 본질적으로 가지고 있지 않은 SAMC 알고리즘을 포함하고 있고, 또한 에르고딕 성질을 만족하며 확장 표본공간을 조절할 수 있는 능력을 갖추고 있다. 본 연구에서 제안

한 SAMCIS 알고리즘을 이용한 정확추론 방법은 기존 방법과 실 자료 분석을 통해 제안된 방법이 더욱 정확한 추정치를 제공하는 우수성을 보여 주었다.

핵심어 : 로지스틱 회귀모형, 정확추론, Markov chain Monte Carlo, Stochastic approximation Monte Carlo importance sampling, 국소트랩.

목 차

표 목 차

제 1 장 서 론

로지스틱 회귀모형은 범주형 데이터를 종속변수로 하여 종속 변수와 독립 변수 간의 관계를 구체적인 함수로 나타내는 예측 모형이다. 주로 의료, 통신과 같은 다양한 분야에서 분류 및 예측을 수행한다. 로지스틱 회귀모형에 대한 통계적 추론에서 장애모수를 제외한 관심모수의 추론 시 일반적으로 비조건부 우도 함수에 근거한 대표본 근사를 사용한다.

균형 자료(balanced data) 또는 오직 몇몇 모수를 가지는 자료에서, 비조건부 최대 우도 추론은 만족스러운 접근이다. 그러나 점근적 추론은 표본공간이 작거나 희박 자료(sparse data)인 경우 또는 불균형 자료(unbalanced data)인 경우 종종 신뢰할 수 없다. 희박자료에서 모수 추정치들은 모수 공간의 가장자리에 놓이게 된다. 다시 말해서, 추정될 확률은 0 또는 1에 가깝고 이때 로짓 회귀계수는 $-\infty$ 또는 $+\infty$ 가 된다. 개념적으로, 참의 모수 값이 모수 공간의 가장자리에 있는지 또는 내부에 있는지 구별하기 위한 불충분한 정보이다. 이 불확실성은 모수 추정에서 큰 표준 오차에 반영된다. 참의 모수 값이 모수 공간의 가장자리에 놓여 있을 때, 대표본 이론(large-sample theory)은 유효하지 않다. Cox et al. (1970)는 대표본 추론의 대안으로서 정확 분포를 활용하는 빈도주의 방법을 제안했다.

정확추론은 검정통계량의 조건부 분포를 사용하는 방법이다. 즉, 장애모수에 대한 충분통계량의 관측된 값을 고정했을 때 관심모수에 대한 충분통계량의 조건부 분포에 기초한다. 정확추론을 사용하는 경우 소표본이거나 희박자료인 경우 또는 불균형자료인 경우 신뢰할 수 있다.

Oster (2002)와 Oster (2003)는 범주형 데이터 분석을 위한 정확추론을 다섯 개의 통계 소프트웨어 패키지(StatXact, LogXact, Stata, Testimate, SAS)를 사용하여 비교했다. 필요한 조건부 분포를 생성하기 위한 반복되는 알고리즘은 상용 소프트웨어 패키지인 LogXact에서 시행된다(Cytel Inc. 2006a). 그러나 이 알고리즘은 오직 적당한 표본 크기와 공변량의 수에서만 문제를 해결할 수 있

다. 분석 가능한 문제의 수준을 높이기 위해, Mehta et al. (2000)는 아크와 노드의 네트워크에 기반을 둔 다이렉트 몬테카를로(direct Monte Carlo) 샘플링 방법을 연구했고 LogXact에서 실행했다. 그러나 메모리에 저장할 수 있는 네트워크의 크기에 문제가 있다. Forster et al. (1996)는 조건부 분포로부터 종속적인 샘플을 생성하기 위해 깁스 샘플링(Gibbs sampling) 방법을 연구했고, 몬테카를로(Monte carlo) 정확 조건부 검정을 수행했다. 깁스 샘플링 방법의 단점은 장애 모수에 대한 충분통계량의 관측된 값과 관심 있는 모수에 대한 충분통계량의 값이 주어졌을 때 특정 충분통계량의 조건부 분포로부터 샘플을 뽑는다는 것이다. 너무 많은 충분통계량을 조건으로 하는 것은 과도한 조건화(overconditioning) 문제를 일으키며 결과적으로 충분통계량의 완전한 벡터에 대해 이산형(discrete) 또는 퇴화된(degenerate) 분포를 만든다. 과도한 조건화 문제는 특히 로지스틱 회귀모형에서 연속적인 공변량과 관련된 충분통계량을 조건으로 할 때 심해져 좋지 못한 혼합(mixing)을 야기한다. Forster et al. (2003)은 메트로폴리스 헤스팅스(Metropolis-Hastings, MH) 알고리즘을 제안했다. 그러나 큰 표본데이터를 다룰 때 메모리의 문제가 생긴다. Zamar et al. (2007)는 대표본에 적용하기 위해 R에서 사용하는 elrm 알고리즘을 개발하였다. 그러나 elrm 알고리즘도 MH 알고리즘을 기반으로 하므로 국소트랩(local trap)의 문제가 발생한다.

본 연구에서는 컴퓨터에 저장되는 메모리의 문제와 국소트랩 문제를 해결하기 위하여 SAMCIS 알고리즘을 이용하여 로지스틱 회귀모형의 정확추론을 수행하고자 한다. 제 2장에서는 로지스틱 회귀모형에 대한 정확추론 관련 연구에 대해 알아본다. 선행연구인 Forster에 의해 제안된 MH알고리즘과, Zamar et al. (2007)에 의해 제안된 elrm알고리즘을 소개한다. 제 3장에서는 SAMCIS 알고리즘을 소개하고, 이를 이용한 로지스틱 회귀에서의 추론을 알아본다. 제 4장에서는 실제 자료를 이용하여 elrm알고리즘과 SAMCIS 알고리즘을 비교한다.

제 2 장 로지스틱 회귀모형에 대한 정확추론

2.1 정확 로지스틱 회귀

정확 로지스틱 회귀에서 관심 있는 결과는 이항 반응 변수로 나타난다. Y_i는 성공확률 p_i를 가지는 m_i번 시행에서의 i번째 이항 반응변수이다. 로지스틱 회귀모형은 다음과 같다.

$$\mathrm{logit}(p_i) = w_i^T\beta + z_i^T\gamma, \qquad i = 1, \ldots, n, \tag{1}$$

β는 i번째 반응에 대한 p개의 설명변수들 $w_i = (w_{i1}, w_{i2}, \ldots, w_{ip})^T$에 대응되는 장애모수의 벡터이고, γ는 나머지 q개 설명변수들 $z_i = (z_{i1}, z_{i2}, \ldots, z_{iq})^T$에 대응되는 관심모수의 벡터이다. $p_i = P(Y_i = 1)$이고, n은 반응의 수이다. $Y = (Y_1, \ldots, Y_n)^T$이고, W는 i번째 행이 w_i^T인 $n \times p$행렬이고, Z는 i번째 행이 z_i^T인 $n \times q$행렬이다. 로지스틱 회귀모형에서 본 연구는 γ의 추론에 관심이 있다.

정확 조건부 추론은 장애모수 β에 대한 충분통계량의 관측된 값을 고정시켰을 때, 관심모수 γ에 대한 충분통계량의 분포에 기초한다. 즉, 추론은 장애모수에 대한 충분통계량을 고정시켰을 때 Y의 조건부 분포에 기초하는 것과 같다. 장애모수 β의 충분통계량을 조건으로 주었기 때문에 장애모수 β에 의존하지 않는다. 먼저, Y의 결합 분포는 식 (2)와 같다.

$$
\begin{aligned}
f(y|\beta,\gamma) &= \prod_{i=1}^{n}\binom{m_i}{y_i} p_i^{y_i}(1-p_i)^{m_i-y_i} \\
&= \prod_{i=1}^{n}\binom{m_i}{y_i} \exp\{y_i \log(p_i) + (m_i - y_y)\log(1-p_i)\} \\
&= \left[\prod_{i=1}^{n}\binom{m_i}{y_i}\right] \exp\left\{\sum_i y_i \log\left(\frac{p_i}{1-p_i}\right) + \sum_i m_i \log(1-p_i)\right\} \\
&= \left[\prod_{i=1}^{n}\binom{m_i}{y_i}\right] \exp\left\{\sum_i y_i w_i^T \beta + \sum y_i z_i^T \gamma\right\} \exp\left\{\sum_i m_i \log(1-p_i)\right\} \qquad (2)
\end{aligned}
$$

식 (2)를 이용하여 Y의 조건부 분포를 구하기 위해 다음과 같이 일부 식을 벡터형태로 변경한다.

$$
\sum y_i w_i^T \beta = (y_1, \ldots, y_n)\begin{pmatrix} w_1^T \\ \vdots \\ w_n^T \end{pmatrix}\beta = y^T W\beta = \beta^T W^T y
$$

$$
\sum y_i z_i^T \gamma = (y_1, \ldots, y_n)\begin{pmatrix} z_1^T \\ \vdots \\ z_n^T \end{pmatrix}\gamma = y^T Z\gamma = \gamma^T Z^T y \qquad (3)
$$

식 (2)에서 인수분해 정리에 의하여 관심모수 γ에 대한 충분통계량은 $T = Z^T y$이고, 장애모수 β에 대한 충분통계량은 $S = W^T y$이다.

$S = W^T y = s$, $C(s) = \{y^* : S(y^*) = s\} = \{y^* : W^T y^* = s\}$라고 가정하자. 즉, 장애모수의 충분통계량을 만족하는 y^*의 집합은 $C(s)$이다. Y의 조건부 분포 $f(y|S=s)$는 식 (4)와 같다. 여기서 $h(\beta,\gamma)$는 $\sum_i m_i \log(1-p_i)$이다.

$$
\begin{aligned}
f(y|S=s) &= \frac{\left[\prod_{i=1}^{n}\binom{m_i}{y_i}\right]\exp\{\beta^T W^T y+\gamma^T Z^T y\}\exp\{h(\beta,\gamma)\}}{\sum_{y^*\in C(s)}\left[\prod_{i=1}^{n}\binom{m_i}{y_i^*}\right]\exp\{\beta^T W^T y^*+\gamma^T Z^T y^*\}\exp\{h(\beta,\gamma)\}} \\
&= \frac{\left[\prod_{i=1}^{n}\binom{m_i}{y_i}\right]\exp\{\beta^T s\}\exp\{\gamma^T Z^T y\}}{\sum_{y^*\in C(s)}\left[\prod_{i=1}^{n}\binom{m_i}{y_i^*}\right]\exp\{\beta^T s\}\exp\{\gamma^T Z^T y^*\}} \\
&= \frac{\left[\prod_{i=1}^{n}\binom{m_i}{y_i}\right]\exp\{\gamma^T Z^T y\}}{\sum_{y^*\in C(s)}\left[\prod_{i=1}^{n}\binom{m_i}{y_{i^*}}\right]\exp\{\gamma^T Z^T y^*\}} \\
&\propto \left[\prod_{i=1}^{n}\binom{m_i}{y_i}\right]\exp\{\gamma^T Z^T y\} \qquad (4)
\end{aligned}
$$

γ에 대한 정확추론을 위해, 조건부 분포 $f(y|S=s)$의 계산이 필요하다. γ에 대한 정확추론은 $f(y|S=s)$로부터 표본을 추출하여 얻어지는 추정치에 기초한다. 그러나 $f(y|S=s)$의 분모에서 집합 $C(s)$를 만족하는 y^*의 열거가 필요하기 때문에 큰 표본데이터에서 정규화 상수의 계산이 불가능하다.

정규화 상수 계산 문제를 극복하기 위해 Forster et al.(2003)과 Zamar et al.(2007)은 각각 마코브체인 몬테카를로(Markov chain Monte Carlo, MCMC)에 기반을 둔 Forster 알고리즘과 elrm 알고리즘을 제안하였다. MCMC 접근 방법은 $f(y|S=s)$에서 오직 비례상수까지의 계산만 필요로 한다.

2.2 Forster 알고리즘

2.2.1 Metropolis-Hastings 알고리즘

Metropolis-Hastings(Metropolis et al. 1953; Hastings 1970, MH) 알고리즘은 직접 표본을 얻기 어려운 확률분포로부터 표본을 생성하는데 사용하는 대표적인 표본 추출 알고리즘이다. 확률밀도함수의 적분 값을 1로 만드는 정규화 상수를 구하기 힘든 경우에 활용할 수 있는 장점이 있다. Metropolis-Hastings 알고리즘은 다음의 과정으로 이루어진다. y는 주어진 관측값이고, $f(y)$는 y의 확률분포이다. y^*는 새롭게 생성되는 y의 값이다.

Metropolis-Hastings 알고리즘:

현재 표본을 $y_t\,(t=1,2,...,n)$라고 하자. 다음 표본 y_{t+1}은 다음과 같은 과정을 통해 생성된다.

(1) 적절한 표본생성(proposal) 분포 $q(y_t,y^*)$로부터 새로운 표본 y^*값을 생성한다.

(2) 채택확률을 다음과 같이 구한다. $\alpha_{y_t,y^*}=\min\left\{1,\dfrac{f(y^*)q(y^*,y_t)}{f(y_t)q(y_t,y^*)}\right\}$.

(3) 채택확률 α_{y_t,y^*}의 확률로 $Y_{t+1}=y^*$을 채택하고, $1-\alpha_{y_t,y^*}$의 확률로 $Y_{t+1}=y_t$을 채택한다.

2.2.2 Forster 알고리즘

Forster et al.(2003)은 MH알고리즘의 (1)에서 y^*를 생성하는 방법을 $y^* = y + d \bullet v$ 형태로 제안하였다. 선택된 정수 r에 대하여 v는 집합

$$V = \left\{ v : \sum |v_i| \le r \text{ and } v_i \; coprime \text{ for } i = 1,...,n \text{ and } W^T v = 0 \right\} \qquad (5)$$

로부터의 벡터이다. 여기서 W는 i번째 행이 $w_i^T = (w_{i1}, w_{i2}, \ldots, w_{ip})$인 $n \times p$행렬이다. V은 열거 가능한 집합

$$V' = \left\{ v : \sum |v_i| \le r \text{ and } v_i \; coprime \text{ for } i = 1,...,n \right\} \qquad (6)$$

로부터의 부분집합이다. $W^T v = 0$의 조건을 만족하면 장애모수에 대한 충분통계량 $S = W^T y^*$가 유지된다. 즉,

$$W^T y^* = W^T (y + dv) = W^T y + d \bullet W^T v = W^T y \qquad (7)$$

이다. d는 $i = 1,...,n$에 대해 $0 \le y_i + dv_i \le m_i$를 만족하는 정수이다. 보통 선형모형에 상수항이 포함된 것처럼 행렬 W의 열 공간에도 1로 구성된 벡터가 존재한다. $W^T v = 0$을 만족하기 위해 $\sum_{i=1}^{n} v_i = 0$이고 $\sum_{i=1}^{n} |v_i|$와 r은 짝수이어야 한다. 선택된 r 값은 마코브 체인(Markov chain)의 혼합을 통제하는데, 큰 r 값은 마코브 체인에서 더 큰 전이와 더 나은 혼합을 할 수 있도록 한다. 그러나 r 값은 v와 y의 관찰된 값이 조건일 때 d의 샘플링에 영향을 미치기 때문에 큰 r 값은 제약조건 $0 \le y_i + dv_i \le m_i$를 만족하는 d의 유일한 정수가 0이라는 확률을 높일 것이다. 만약 $d = 0$이면, 마코브 체인은 현재 상태로의 전이를 나타내기

때문에 좋지 못한 혼합을 할 것이다. 또한 작은 r 값은 V의 열거를 가능하게 하지만 마코브 체인은 지역 덫(local neighborhood)에 빠지게 된다. Forster et al.(2003)은 r로 4, 6, 8을 제안하였다.

샘플링은 크게 두 가지 단계로 진행된다. 첫째, y를 고려하지 않고 v를 균일하게 생성한다. 둘째, v와 y를 사용하여 $0 \le y_i + dv_i \le m_i \left(= -\frac{y_i}{m_i} \le d \le \frac{m_i - y_i}{m_i} \right)$를 만족하는 모든 정수 d를 구한다. 생성된 v, d와 관찰치 y를 이용한

$$q(d|v,y) \propto \exp\{\gamma^T Z^T (y + dv)\} \prod_{i=1}^{n} \binom{m_i}{y_i + dv_i} = \eta(d|v,y) \tag{8}$$

를 사용하여 여러 개의 정수 중 하나의 d를 선택한다. $q(d|v,y)$의 비례 상수까지를 $\eta(d|v,y)$라고 정의한다.

$y^*(= y + d \bullet v)$ 생성 단계를 요약하면 다음과 같다.

(단계 1) r을 이용하여 집합 V를 열거한다.
(단계 2) 동일한 확률로 $v \in V$를 선택한다.
(단계 3) $0 \le y_i + dv_i \le m_i$를 만족하는 모든 정수 d를 찾는다.
여기서 찾은 d의 개수를 k라 한다.
(단계 4) $\sum_{i=1}^{k} \eta(d_i|v,y)$를 계산한 후, $i = 1,2,...,k$에 대하여 $P(d_i) = \frac{\eta(d_i|v,y)}{\sum_{i=1}^{k} \eta(d_i|v,y)}$를 부여한다.
(단계 5) $P(d_i)$에 따라 d를 선택한다.
(단계 6) $y^* = y + d \bullet v$ 값을 생성한다.

y에서 y^*로 가는 전이확률 $q(y,y^*)$는 y가 주어졌을 때 y^*의 조건부 분포 $f(y^*|y)$와 같다. 관측치 y는 주어지므로 새로 생성하는 y^*의 분포는 균일하게 생성한 v와 y가 주어졌을 때 생성하는 d의 결합분포이다. 즉, y^*의 조건부 분포는 d와 v의 분포와 비례한다. v는 균일분포로부터 생성되기 때문에 $q(v) \propto 1$이 되므로 $q(y,y^*)$는 다음과 같이 d의 조건부분포와 비례한다.

$$q(y,y^*) = f(y^*|y) \propto q(d|v)q(v) \propto q(d|v) \tag{9}$$

또한 (4)와 (8)을 비교했을 때 $q(d|v) \propto f(y^*)$ 임을 알 수 있고, 따라서 모든 y, y^*에 대해 $q(y,y^*) \propto f(y^*)$이다. y^*에서 y로의 전이는 오직 $-(dv)$이므로 $q(y^*,y) \propto f(y)$가 성립한다. 이런 이유로 채택확률

$$\alpha(y,y^*) = \min\left\{\frac{f(y^*)q(y^*,y)}{f(y)q(y,y^*)},1\right\} = \min\left\{\frac{f(y^*)f(y)}{f(y)f(y^*)},1\right\} = 1 \tag{10}$$

이다. 즉, $a(y,y^*) = 1$이고 y^*를 1의 확률로 받아들인다.

Forster 알고리즘:

(1) 위에 언급한 방법으로 새로운 $y^*(= y + d \bullet v)$를 생성한다.

(2) 채택확률 $\alpha_{y_t y^*}$는 1이므로 생성된 y^*를 $Y_{t+1} = y^*$로 채택한다.

Forster 알고리즘의 제약은 V의 완전한 열거가 필요하다는 것이다. 그러나 초기집합 V'의 크기는 반응 벡터 길이에 따라 급속하게 커지므로 부분집합 V의 열거가 불가능하여 큰 표본데이터를 다룰 때 메모리의 문제가 발생한다.

2.3 elrm 알고리즘

Forster et al., (2003) 알고리즘은

$$V=\left\{v:\sum|v_i|\le r \text{ and } v_i \; coprime \text{ for } i=1,...,n \text{ and } W^Tv=0\right\} \qquad (11)$$

를 열거하고 메모리에 저장 한 후 벡터들의 균일 샘플링을 제안한다. 그러나 V를 생성시키는데 사용되는 초기 집합 V'의 크기는 반응 벡터의 길이에 따라 급속하게 커진다. 그러므로 대량데이터 집합에서, V'의 열거는 불가능 할 수 있다. 추가적으로, V가 너무 크면 메모리에 저장을 하지 못한다. 대량데이터를 다루기 위해, Zamar et al.(2007)은 Forster et al.(2003)에 의해 제안된 알고리즘에 두 개의 제약조건을 추가하였다.

조건 1. $1\le i\le n$에 대해 제약조건 $|v_i|\le m_i$를 만족하는 V의 부분집합 V_A로부터 균일하게 샘플을 뽑는다. 식 (11)의 부분집합

$$V_A=\left\{v:\sum|v_i|\le r \text{ and } v_i \; coprime \text{ for } i=1,...,n \text{ and } W^Tv=0, \left|v_i\right|\le m_i\right\} \qquad (12)$$

로부터 v를 생성함으로써 혼합을 향상시킨다. 왜냐하면 $|v_i|\ge m_i$인 벡터들은 오직 $d=0$일 때만 제약조건 $0\le y_i+dv_i\le m_i$를 만족하기 때문이다.

조건 2. 초기집합 V'로부터 열거 없이 균일하게 샘플을 뽑고, V_A에 속하지 않는 모든 벡터들을 기각시킨다. 즉, $W^Tv\ne 0$이거나 $|v_i|>m_i$인 모든 v_i를 기각한다.

Forster et al.(2003)은 실제로 V를 열거하는 절차를 설명하고 있지 않는다. Zamar et al. (2007)는 V_A으로부터 균일하게 샘플을 뽑는 방법을 설명하였다. 다음과 같이 r개 원소를 가지는 벡터를 고려해보자.

$$R = \left\{ r : \sum_{i=1}^{r} |r_i| \le r \text{ and } r_i \textit{ coprime for } i = 1,...,r, \sum_{i=1}^{r} r_i = 0 \right\} \qquad (13)$$

행렬 W에 1로 이루어진 열벡터가 존재함으로 $W^T v = 0$ 인 조건은 $\sum_{i=1}^{n} r_i = 0$와 같이 나타낼 수 있다. r이 주어졌을 때 R에 속하는 원소를 열거하기 위한 절차는 다음과 같다.

1. $r/2$보다 작거나 같고 서로소인 양의 정수를 나열한다.
2. $1 \le j \le r/2$에 대해 크기 j인 모든 가능한 조합을 나열한다.
3. $0 \le k \le r-j$개의 1을 각 원소 앞에 붙인다.
4. $\sum r_i > r$이거나 $\sum r_i$가 홀수인 벡터 또는 중복인 벡터를 삭제한다.
5. 합이 0이 되도록 - 부호를 할당한다.
6. r개의 원소를 가지도록 적절한 수의 0을 할당한다.

예로 r이 4일 때 R에 속하는 원소를 열거하면 다음과 같다.

1. $r/2$보다 작거나 같고 서로소인 양의 정수를 나열한다.
   r=4이므로 1,2
2. $1 \le j \le r/2$에 대해 크기 j인 모든 가능한 조합을 나열한다.
   j=1 : {1}, {2}
   j=2 : {1,2}

3. $0 \le k \le r-j$개의 1을 각 원소 앞에 붙인다.

   j=1 : {1}, {1,1}, {1,1,1}, {1,1,1,1}, {2}, {1,2}, {1,1,2}, {1,1,1,2}

   j=2 : {1,2}, {1,1,2}, {1,1,1,2}

4. $\sum r_i > r$이거나 $\sum r_i$가 홀수인 벡터 또는 중복인 벡터를 삭제한다.

   j=1 : ~~{1}~~, {1,1}, ~~{1,1,1}~~, {1,1,1,1}, ~~{2}~~, ~~{1,2}~~, {1,1,2}, ~~{1,1,1,2}~~

   j=2 : ~~{1,2}~~, ~~{1,1,2}~~, ~~{1,1,1,2}~~

5. 합이 0이 되도록 - 부호를 할당한다.

   {1,-1}, {1,1,-1,-1}, {-1,-1,2}, {1,1,-2}

6. r개의 원소를 가지도록 적절한 수의 0을 할당한다.

   {1,-1,0,0}, {1,1,-1,-1}, {-1,-1,2,0}, {1,1,-2,0}

따라서 R에 속하는 벡터는 {1,-1,0,0}, {1,1,-1,-1}, {-1,-1,2,0}, {1,1,-2,0}이다. 데이터가 주어진다면 열거된 R의 벡터에 제약조건 $|v_i| \le m_i$을 추가한 V_A의 열거가 가능하다. V_A에 속하는 벡터 중 동일한 확률로 하나를 선택하여 v로 사용한다. 그러나 elrm알고리즘은 여전히 MH 알고리즘을 사용하기 때문에 국소 트랩(local trap)에 빠지기 쉽다는 문제점이 존재한다.

R에서의 elrm 알고리즘의 사용법을 간단히 설명한다. elrm 알고리즘은 R의 CRAN으로부터 elrm 패키지를 다운받아 사용 할 수 있으며 elrm의 기본 형식은 다음과 같다.

elrm(formula, interest, r=4, iter=1000, dataset, burnIn=0, alpha=0.05)

여기서 formula에는 모형 식을 쓴다. 주의해야 할 점은 이항 반응은 success/trials 형태로 써야한다는 것이다. interest에는 정확 조건부 추론에서 관심 있는 모수의 변수를 쓴다. r은 v를 생성할 때 영향을 주는 것으로 사용자

가 지정한다. iter은 반복수이고 burnIn은 예열 시간(burning time)이다. dataset에는 분석할 데이터의 이름을 쓴다. alpha=0.05의 의미는 유의수준 5%에서 정확추론을 수행하라는 의미이다. 핵심 함수 elrm()을 호출하면 관심모수에 대한 샘플링 된 충분통계량의 마코브 체인(Markov chain)을 생성시키고 추론을 수행한다.

제 3 장 로지스틱 회귀모형에서 SAMCIS 알고리즘을 이용한 정확추론

Stochastic approximation Monte Carlo importance sampling (SAMCIS) 알고리즘 (Cheon et al, 2014)은 Stochastic approximation Monte Carlo (SAMC) 알고리즘 (Liang et al, 2007)과 importance sampling을 결합한 것이다.

3.1 SAMCIS 알고리즘

우선 표본공간 분할을 통해 국소트랩(local trap) 문제를 해결하는 Stochastic approximation Monte Carlo (SAMC) 알고리즘 (Liang et al, 2007)을 간략히 설명한다. 다음과 같은 형태의 분포로부터 표본을 추출하는 데에 관심이 있다고 하자.

$$f(x)=\frac{1}{Z}\psi(x),\ x\in\chi \tag{14}$$

Z는 정규화 상수이고, χ는 표본 공간, $\psi(x)$는 비음(non-negative) 함수이다. 표본 공간은 미리 정해진 상수 $u_0,...,u_{m-1}$에 따라 함수 $U(x)$가 $m+1$개의 분리된 하위영역으로 분할된다. $E_0=\{x:U(x)\le u_0\}, E_1=\{x:u_0<U(x)\le u_1\},...$ $E_{m-1}=\{x:u_{m-2}<U(x)\le u_{m-1}\}, E_m=\{x:U(x)>u_{m-1}\}$이다. SAMC는 미리 정해진 빈도를 이용하여 각 영역에서 표본을 추출한다. $m+1$개의 하위영역은 non-empty라고 가정한다. 즉, $i=0,...,m$에 대해 $w_i=\int_{E_i}\psi(x)dx>0$이다.

$\pi=(\pi_0,\pi_1,...,\pi_m)$를 $0<\pi_i<1$와$\sum_{i=0}^{m}\pi_i=1$을 만족하는 $m+1$-벡터인 각 부분 영

역에서의 원래의 표본 함수라고 정의한다. $i=0,...,m$에 대해 $\theta_i = \log(\int_{E_i} \psi(x)dx/\pi_i) = \log(\frac{w_i}{\pi_i})$, $\theta=(\theta_0,\theta_1,...,\theta_m)$, Θ는 θ의 공간이라 정의한다. $\psi(x)$의 일반적인 선택을 위해 $\Theta = R^{m+1}$이라고 한다. $\theta^{(t)} = (\theta_0^{(t)},\theta_1^{(t)},...,\theta_m^{(t)})$는 t번째 반복에서 얻어진 θ의 추정치이다. (14)번 식을 (15)번 식과 같이 다시 표현할 수 있다.

$$f_{\theta^{(t)}}(x) \propto \sum_{i=0}^{m} \frac{\psi(x)}{e^{\theta_i^{(t)}}} I(x \in E_i) \tag{15}$$

$\{\gamma_t\}$는 어떤 $\tau \in (1,2)$에 대해

$$(a)\ \sum_{t=1}^{\infty} \gamma_t = \infty,\ \ (b)\ \sum_{t=1}^{\infty} \gamma_t^{\tau} < \infty \tag{16}$$

를 만족하는 양수이고 단조 증가하는 수열이라고 정의한다. 본 연구에서는 $t_0 > 1$인 어떤 정해진 값에 대해 $\gamma_t = \frac{t_0}{\max(t_0,t)}$, $t=1,2,...$을 사용한다. $J(x)$는 표본 x가 속해있는 하위영역의 인덱스이다. 이와 같은 가정을 이용하여 SAMC 알고리즘을 정리하면 다음과 같다.

SAMC 알고리즘:

(a) (샘플링) target 분포를 가지고 한번의 Metropolis-Hastings(MH) 갱신에 의해 표본 $x^{(t+1)}$추출한다.

(a.1) Proposal 분포 $q(x^{(t)},y)$ 에 따라 y를 생성시킨다.

(a.2) 다음의 비율을 계산한다.

$$r = e^{\theta^{(t)}_{J(x^{(t)})} - \theta^{(t)}_{J(y)}} \frac{\psi(y)q(y,x^{(t)})}{\psi(x^{(t)})q(x^{(t)},y)}$$

(a.3) 확률 $\min(1,r)$ 을 가지고 생성된 y를 받아들인다.

만약 받아들이면 $x^{(t+1)} = y$이고 그렇지 않다면 $x^{(t+1)} = x^{(t)}$이다.

(b) (θ 갱신) $\theta^* = \theta_t + \gamma_{t+1}(e_{t+1} - \pi)$라고 정의한다.

여기서 $e_{t+1} = (e_{t+1,1},...,e_{t+1,m})$이다. 만약 $x^{(t)} \in E_i$ 이면 $e_{t+1,i} = 1$, 그렇지 않으면 $e_{t+1,i} = 0$ 이다. 만약 $\theta^* \in \Theta$ 이면 $\theta_{t+1} = \theta^*$, 그렇지 않으면 $\theta_{t+1} = \theta^* + c^*$ 이다. 즉, y가 뽑힌 지역에 낮은 가중치를 줘서 다른 영역에서 y가 뽑힐 수 있도록 θ를 갱신한다. 여기서 $c^* = (c^*,...,c^*)$ 이고 $\theta^* + c^* \in \Theta$를 만족하는 임의의 벡터이다.

다음으로 Stochastic approximation Monte Carlo importance sampling (SAMCIS) 알고리즘 (Cheon et al, 2014)을 설명한다.

식 (4)의 $f(y|S)$로부터 직접 표본을 추출하는 대신 아래의 시도 함수(trial density) $g(y|S)$로부터 표본을 생성시키는 방법을 제안한다.

$$g(y|S) \propto \left[\prod_{i=1}^{n} \binom{m_i}{y_i}\right] \exp\{\gamma^T Z^T y\} \exp\{-[S(y) - S(y_{obs})]^T [S(y) - S(y_{obs})]\} \qquad (17)$$

$S(y)$는 새로운 y에서의 장애모수의 충분통계량 $S = W^T y^*$ 값이고, $S(y_{obs})$는

관측된 y에서의 장애모수의 충분통계량 $S = W^T y$ 값이다.

$$U(x) = [S(y) - S(y_{obs})]^T [S(y) - S(y_{obs})] \tag{18}$$

라 하자. 여기서 $U(x)$는 명시된 조건을 만족하지 않는 영역에 패널티를 준다. 영역은 0부터 m이다. β의 충분통계량을 유지시키기 위해 $\Omega = \{x : U(x) = 0\}$인 영역만 선택한다. 이는 주표집(Importance sampling) 기법이 Ω에서의 표본을 가지고 $f(y|S)$를 추론하는 데 사용되어 질 수 있음을 알 수 있다.

SAMC 이론에 따라 θ_t는 수렴한다. $\left(x^{(1)}, \theta_1\right), ..., \left(x^{(N)}, \theta_N\right)$을 확장된 집합 χ로부터 SAMCIS 알고리즘에 의해 추출된 표본이라 하자. Liang (2009)에 따르면 SAMC 알고리즘은 dynamic importance sampling 알고리즘이며, 임의의 함수 $\rho(x)$에 대해 $N \to \infty$ 에 따라

$$\frac{\sum_{t=1}^{N} e^{\theta^{(t)}_{\mathcal{M}(x^{(t)})}} \rho(x^{(t)})}{\sum_{t=1}^{N} e^{\theta^{(t)}_{\mathcal{M}(x^{(t)})}}} \to E_g \rho(x), \ a.s. \tag{19}$$

여기서 $E_g \rho(x)$ 는 $g(x)$에 대해 $\rho(x)$ 의 기대치를 말한다. 따라서 조건부 p값은 다음과 같이 추정된다.

$$\widehat{p_h} = \frac{\sum_{t=1}^{n} I_{\{h(y^{(t)}) \geq h_{obs}\}} e^{\vartheta^{(t)}_{\mathcal{M}(y^{(t)})}}}{\sum_{t=1}^{n} e^{\vartheta^{(t)}_{\mathcal{M}(y^{(t)})}}} \tag{20}$$

SAMCIS 알고리즘의 효율적인 이용을 위해 다음을 고려해 볼 필요가 있다.

1. 표본 분할

본 연구에서는 m을 3으로 정하였다. 표본공간을 함수 $U(x)$에 따라 다음과 같이 분할한다.

$$E_0 = \{x : U(x) = 0\}, E_1 = \{x : U(x) = 1\}, E_2 = \{x : U(x) = 2\}, E_3 = x : U(x) \geq 3 \quad (21)$$

2. 표본 분포 π 의 선택

본 연구의 목적이 $E_0 = \Omega$ 로부터 표본을 추출하는 것이기 때문에 π 는 참조 집합에 따라 샘플링이 편의가 되도록 선택될 수 있다. 본 논문에서는 다음과 같이 정한다.

$$\pi_i \propto \frac{1}{(i+1)^{\xi}}, \ i = 0,1,...,m \quad (22)$$

여기서 ξ 는 1또는 2이다. 예로 $\xi = 2$, $m = 5$ 이라면,
$(\pi_0, \pi_1, \pi_2, \pi_3, \pi_4, \pi_5) = (0.671, 0.168, 0.075, 0.042, 0.027, 0.019)$가 된다.

3. gain factor γ_t 의 선택과 반복 수

조건 (16)를 충족시키기 위해 다음과 같이 정의한다.

$$\gamma_t = \left(\frac{T_0}{\max(T_0, t)} \right)^{\eta}, \ t = 0,1,2,... \tag{23}$$

$T_0 > 1$ 와 $\eta \in (0.5,1]$는 미리 정해진다. 본 논문에서는 모든 예제에 $\eta = 1.0$ 으로 정하였다. T_0 의 큰 값은 m이 매우 큰 값이라 하더라도 샘플러들이 모든 부분공간에 빠르게 도달하게 해준다. N을 반복의 총 횟수라 하자. T_0와 N의 적절한 선택은 추정치 $\hat{\theta}$, $\hat{\pi}$ 의 수렴을 검사함으로써 정해진다. 만약 수렴을 하지 않는다면, T_0와 N의 더 큰 값을 통해 다시 모의실험을 진행한다.

3.2 SAMCIS 알고리즘을 이용한 정확추론

먼저 오직 γ_i의 추론에 관심이 있다고 하자.

$$H_o : \gamma_i = 0 \quad \text{vs} \quad H_1 : \gamma_i \neq 0$$

이때, $r_1, ..., r_{i-1}, r_{i+1}, ..., r_q$와 $\beta_1, ..., \beta_p$를 장애모수로 간주한다. 추론은

$$f_{T_i}(t_i | S_1 = s_1, ..., S_p = s_p, T_{-i} = t_{-i}, \gamma_i = 0) \tag{24}$$

로부터 추출된 샘플에 기초한다. $S_1, ..., S_p$는 장애모수 $\beta_1, ..., \beta_p$에 대한 충분통계량이고, $T_1, ..., T_q$는 관심모수 $\gamma_1, ..., \gamma_q$에 대한 충분통계량이다. T_{-i}와 t_{-i}는 각각 i번째 행을 제외한 나머지 행에 대한 관심모수의 충분통계량과 충분통계량의 관측치 이다. t_i 분포의 열거가 불가능하므로 본 연구는 근사적으로 χ^2분포를 이용한다.

Mehta et al.(1995)은 양측 p-value를 구하기 위한 방법으로 조건부 확률 검정(conditional probability test)과 조건부 스코어 검정(conditional score test)을 사용했다. 먼저 조건부 확률 검정의 양측 p-value는 기각역

$$E_{cp} = \{t : f(t|\beta = 0) \leq f(t_{obs}|\beta = 0)\}$$

에 속하는 추정치들의 합을 통해 구한다. t_{obs}는 관심모수의 충분통계량(=$Z'y$)의 관측된 값이고 t는 충분통계량(=$Z'y$)의 확률변수이다. 기각역 E_{cp}는 t_{obs}의 조건부 확률보다 이하인 조건부 확률을 생성하는 검정통계량의 모든 값을 포함한다. 다음으로 조건부 스코어 검정의 양측 p-value는 기각역

$$E_{cs} = \left\{ t : \frac{(t-\hat{\mu})^2}{\widehat{\sigma^2}} \geq \frac{(t_{obs}-\hat{\mu})^2}{\widehat{\sigma^2}} \right\}$$

에 속하는 추정치들의 합을 통해 구한다. t_{obs}는 관심모수의 충분통계량(=$Z'y$)의 관측된 값이고 t는 충분통계량(=$Z'y$)의 확률변수이다. μ와 σ^2는 T의 평균과 분산이고, $\hat{\mu}$와 $\widehat{\sigma^2}$는 추정치이다. 기각역 E_{cs}는 조건부 스코어 (conditional scores)가 검정통계량의 관측된 값에서 조건부 스코어(conditional score)와 같거나 큰 검정통계량의 모든 값을 포함한다. $(t_{obs}-\hat{\mu})^2/\widehat{\sigma^2}$는 자유도 1인 카이제곱 분포를 따른다. 조건부 스코어 검정의 양측 p-value는 다음과 같다.

$$\hat{p} = \frac{number\ of\ t \in E_{cs}}{N}$$

제 4 장 실증분석

4.1 골육종에 대한 무병 생존 자료

골육종이란 뼈에 발생하는 악성 종양 중 가장 흔히 발생하는 질병이다. 골육종의 발병 원인으로는 과거에 어떤 질환으로 인해 방사선 치료를 받은 경우나 암이 잘 발생하는 특정 질환 증후군이 존재하는 경우가 있다. 골율종의 치료가 종결된 후 첫 3년간은 약 2~4개월에 한 번씩, 그 이후로는 6개월에 한 번씩 정기적으로 진료를 받게 된다. Goorin et al. (1987)는 비전이성 골육종을 겪은 46명의 환자를 조사하여 3년동안 무병여부(DFI3)에 대한 예측 인자를 결정하는데 관심을 보였다. 관심 있는 공변량과 데이터는 다음과 같다.

· 림프구 침윤 (lymphocytic infiltration; LI)
· 성별 (SEX)
· 골육병리 (any osteoid pathology; AOP)

<표 1> 골육종에 대한 무병 생존 자료 결과

LI	SEX	AOP	Proportion DFI3
0	0	0	3/3 (100%)
0	0	1	2/2 (100%)
0	1	0	4/4 (100%)
0	1	1	1/1 (100%)
1	0	0	5/5 (100%)
1	0	1	3/5 (60%)
1	1	0	5/9 (56%)
1	1	1	6/17 (35%)

림프구란 백혈구의 약 25%를 차지하는 세포로 면역 반응을 담당한다. 림프구 침윤이 존재하면 1이고 그렇지 않으면 0이다. 성별이 남자이면 1이고 그렇지 않으면 0이다. 골육병리가 존재하면 1이고 그렇지 않으면 0이다. 데이터 첫 번째 행의 의미는 '림프구 침윤이 존재하지 않고, 성별이 여자이고, 골육병리가 존재하지 않는 3명중 3명이 3년 동안 무병이다'이다.

DFI3과 LI로 이루어진 2×2 분할표 <표2>에 대해 피셔의 정확검정을 수행해 보았다.

<표 2> DIF3과 LI에 대한 분할표

	LI=1	LI=0
DIF3=1	19	10
DIF3=0	17	0

유의수준은 0.0075로 5% 유의수준에서 통계적으로 유의하다. 따라서 LI는 DIF3에 유의한 영향을 미친다고 결론내릴 수 있다. SEX와 AOP에 대해서도 동일한 방법으로 피셔의 정확검정 p-value를 구해보면 각각의 양측 p-value는 0.0259, 0.0322이다. DFI3에 대한 SEX, AOP의 주변 효과 모두 5% 유의수준에서 통계적으로 유의하므로 각 SEX, AOP는 DIF3에 유의한 영향을 미친다고 결론내릴 수 있다.

그러나 세 가지 공변량의 효과를 로지스틱 회귀모형을 통해 동시에 연구하는 것에 관심이 있다.

$$\log\left(\frac{p_i}{1-p_i}\right)=\beta+\sum_{j=1}^{3}\gamma_1 x_{ji}$$

림프구 침윤이 없는 모든 대상자는 3년 무병 기간을 가지고 있기 때문에 DFI3과 LI의 2×2 분할표에는 0의 셀이 포함된다. 이 경우 로그우도함수는 최대

가 될 수 없지만 γ_1가 $-\infty$로 수렴함에 따라 유한 상한에 접근한다. 따라서 로그우도의 1차 및 2차 미분을 계산 할 수 없고, 전통적인 최대우도방법으로 γ_1 및 γ_1의 신뢰구간에 대한 추정이 불가능하다.

이 경우는 정확추론을 사용한다. 변수 LI의 추론에 관심이 있다고 하자.

$$H_o : \gamma_{LI} = 0 \quad \text{vs} \quad H_1 : \gamma_{LI} \neq 0$$

관심모수는 LI이고, 장애모수는 절편, SEX, AOP이다. 우선 정확 양측 p-value를 구하는 방법은 다음과 같다. (t_0, t_1, t_2, t_3)는 $(\beta, \gamma_1, \gamma_2, \gamma_3)$에 대응되는 충분통계량이다. 그러므로 다음과 같이 $t_0 = 29, t_1 = 19, t_2 = 16, t_3 = 12$이다.

$$t_0 = W_1^T y = (1\,1 \cdots 1\,1)\begin{pmatrix} 3 \\ 2 \\ \vdots \\ 5 \\ 6 \end{pmatrix} = 29$$

$$t_1 = Z_1^T y = (0\,0 \cdots 1\,1)\begin{pmatrix} 3 \\ 2 \\ \vdots \\ 5 \\ 6 \end{pmatrix} = 19$$

$$t_2 = W_2^T y = (0\,0 \cdots 1\,1)\begin{pmatrix} 3 \\ 2 \\ \vdots \\ 5 \\ 6 \end{pmatrix} = 16$$

$$t_3 = W_3^T y = (0\,1 \cdots 0\,1)\begin{pmatrix} 3 \\ 2 \\ \vdots \\ 5 \\ 6 \end{pmatrix} = 12$$

γ_1에 대한 추론에 관심이 있으므로 $\beta, \gamma_2, \gamma_3$를 장애모수로 간주한다. 장애모수의 충분통계량이 주어졌을 때 t_1이 될 수 있는 모든 가능한 값, $c(29, t_1, 16, 12)$의 분포는 다음과 같다.

<표 3> 골육종 데이터에 대한 정확 조건부 분포

t_1	$c(29, t_1, 16, 12)$	$\frac{(t-\hat{\mu})^2}{\hat{\sigma^2}}$
19	29,445,360	4.544
20	147,312,480	1.495
21	271,271,448	0.098
22	231,819,344	0.354
23	95,325,644	2.263
24	17,473,144	5.825
25	1,204,008	11.041
26	19,448	17.908
Total	793,870,896	

$(y_1, y_2, ..., y_{46})$에 대해 약 8억 개의 이진 배열이 있다. 그러나 많은 이진 배열은 충분통계량에 대해 동일한 값을 산출하기 때문에 $(t_0 = 29, t_1, t_2 = 16, t_3 = 12)$에 대해 오직 8개의 구분된 벡터가 존재한다. T_1의 정확 조건부 분포는 극히 비대칭이기 때문에 정규 근사는 적합하지 않다. 관측된 값 $t_1 = 19$는 충분통계량 범위의 최솟값이기 때문에 회귀계수의 추정치를 구하기 위해 최대우도방법을 사용할 수 없다. t_1의 평균은 21.345이고, 분산은 1.21이다.

조건부 스코어 검정에 기반을 둔 기각역

$$
\begin{aligned}
E_{cs} &= \left\{ t : \frac{(t-\hat{\mu})^2}{\hat{\sigma^2}} \geq \frac{(t_{obs}-\hat{\mu})^2}{\hat{\sigma^2}} \right\} \\
&= \left\{ t : \frac{(t-21.345)^2}{1.21^2} \geq \frac{(19-21.345)^2}{1.21^2} \right\} \\
&= \left\{ t : \frac{(t-21.345)^2}{1.21^2} \geq 4.544 \right\}
\end{aligned}
$$

에 속하는 t를 이용하여 정확(exact) 양측 p-value를 구한 결과는 다음과 같다.

$$
\begin{aligned}
\hat{p} &= \frac{umber of t \in E_{cs}}{N} \\
&= \frac{29,445,360+17,473,144+1,204,008+19,448}{793,870,896} \\
&= 0.0601
\end{aligned}
$$

변수 SEX와 AOP에 대해서도 위와 같은 방법으로 정확 p-value를 얻을 수 있다. SEX에 대한 정확 p-value는 0.117이고, AOP에 대한 정확 p-value는 0.154이다.

elrm 알고리즘의 적용은 다음과 같다. 실행은 1×10^6번 반복하며 처음 1×10^4번은 burn-in 과정으로 버리고 나머지 반복들을 사용하여 추론을 수행한다.

R > elrm(y/m ~ LI + SEX + AOP, interest = ~ LI, iter = 1000000, burnIn = 10000, dataset = sarcoma)

SAMCIS 알고리즘의 적용은 다음과 같다. $u_{\max}^{(1)}=3$과 $u_{\max}^{(2)}=0$으로 설정함으로써 표본공간 χ를 제한한다. 여기서 $m=3$이고 $\chi=\bigcup_{i=0}^{3} E_i$이므로

$E_0 = \{x : U(x) = 0, x \in \chi\}$, $E_1 = \{x : U(x) = 1, x \in \chi\}$, $E_2 = \{x : U(x) = 2, x \in \chi\}$, $E_3 = \{x : U(x) \geq 3, x \in \chi\}$이다. 실행은 1×10^6번 반복하며 처음 1×10^4번은 burn-in 과정으로 버리고 나머지 반복들을 사용하여 추론을 수행한다. gain factor γ_t의 선택 시 $\eta = 1.0$과 $T_0 = 1000$을 선택하였다. 표본 분포 π를 선택할 때 ξ는 2로 선택하였고, 이는 $(\pi_0, \pi_1, \pi_2, \pi_3) = (0.48, 0.24, 0.16, 0.12)$이다. 즉, 유효한 비율은 48%이다. 종합적으로 정확 p-value와 F-test, elrm, SAMCIS를 비교한 결과는 다음과 같다.

<표 4> 골육종에 대한 무병 생존 자료 결과

변수명		Mean of p-value	SE of p-value	Bias	MSE
LI	exact p-value	0.061			
	F-test	0.039			
	elrm	0.059	0.000988	0.0011	2.186×10^{-6}
	SAMCIS	0.060	0.000986	0.0006	1.332×10^{-6}
SEX	exact p-value	0.117			
	F-test	0.059			
	elrm	0.116	0.00185	0.0008	4.063×10^{-6}
	SAMCIS	0.116	0.00174	0.0008	3.668×10^{-6}
AOP	exact p-value	0.154			
	F-test	0.087			
	elrm	0.152	0.00239	0.0015	7.962×10^{-6}
	SAMCIS	0.153	0.00220	0.0007	5.33×10^{-6}

근사적인 방법을 쓴 F-test는 모든 변수에서 정확 p-value와 큰 차이를 보이고 있어서 정확추론을 수행해야 한다. 정확추론에서는 elrm 알고리즘보다 SAMCIS 알고리즘이 모든 변수에서 정확 p-value에 더 가까운 p-value를 제공하는 것을 확인할 수 있다. 또한 모든 변수에서 SAMCIS 알고리즘의 bias와 MSE가 elrm 알고리즘 보다 더 작으므로 SAMCIS 알고리즘이 더 정확한 추론

을 했음을 알 수 있다. 각 변수에서 수행된 피셔의 정확검정과 달리, LI의 p-value는 0.061로 유의수준 1%에서 유의하다. 반면 SEX의 p-value는 0.117, AOP의 p-value는 0.154로 유의하지 않다. 즉, 3년 무병 생존기간에 대해 림프구 침윤은 유의수준 1%에서 유의하게 영향을 미치며, 성별과 골육병리는 유의한 영향을 미치지 않는다는 결론을 내린다.

4.2 당뇨병 자료

당뇨병은 인슐린의 분비량이 부족하거나 정상적인 기능이 이루어지지 않는 등의 대사질환의 일종으로 제 1형과 제 2형으로 구분된다. 제 1형 당뇨병은 인슐린을 전혀 생산하지 못하여 발생하는 질환이고, 제 2형 당뇨병은 인슐린이 상대적으로 부족하여 발생하는 질환이다. 제 1형 당뇨병은 여러 가지 단백질의 항체의 존재에 의해 특정지어진다. 항체는 제 2형 당뇨병에 존재하지 않으므로 당뇨병의 종류를 구분하는데 유용하다. Graham et al. (1999)는 20대 남성의 제 1형 당뇨병 환자를 대상으로 섬 항원-2 항체(IA2A)의 농도수준(높고 낮음)과 몇 개의 공변량 사이의 관계를 연구하는데 관심이 있었다. 관심 있는 공변량과 데이터는 다음과 같다. HLA-DQ는 인간백혈구항원의 종류이다.

· HLA-DQ2의 수 (nDQ2)
· HLA-DQ8의 수 (nDQ8)
· HLA-DQ6.2의 수 (nDQ6.2)

<표 5> 당뇨병 데이터

nDQ2	nDQ8	nDQ6.2	IA2A	N
1	1	0	12	25
0	0	0	2	11
0	1	1	0	2
1	0	0	4	21
0	2	0	2	8
0	1	0	17	30
2	0	0	2	6
1	0	1	0	2

HLA-DQ2, HLA-DQ8, HLA-DQ6.2 수의 범위는 0,1,2이다. 데이터 첫 번째 행의 의미는 'DQ2가 존재하고, DQ8이 존재하고, DQ6.2가 존재하지 않는 25명 중 12명이 IA2A의 농도가 높다'이다. 변수 nDQ2의 추론에 관심이 있다고 하자.

$$H_o : \gamma_{nDQ2} = 0 \quad \text{vs} \quad H_1 : \gamma_{nDQ2} \neq 0$$

관심모수는 nDQ2이고, 장애모수는 절편, nDQ8, nDQ6.2이다.

elrm 알고리즘의 적용은 다음과 같다. 실행은 1×10^6번 반복하며 처음 1×10^4번은 burn-in 과정으로 버리고 나머지 반복들을 사용하여 추론을 수행한다.

R > elrm(IA2A/n ~ nDQ2 + nDQ8 + nDQ6.2, interest = ~ nDQ2, iter = 1000000, burnIn = 10000, dataset = diabDat)

SAMCIS 알고리즘의 적용은 다음과 같다. $u_{\max}^{(1)} = 3$과 $u_{\max}^{(2)} = 0$으로 설정함으로써 표본공간 χ를 제한한다. 여기서 $m = 3$이고 $\chi = \bigcup_{i=0}^{3} E_i$이므로 $E_0 = \{x : U(x) = 0, x \in \chi\}$, $E_1 = \{x : U(x) = 1, x \in \chi\}$, $E_2 = \{x : U(x) = 2, x \in \chi\}$, $E_3 = \{x : U(x) \geq 3, x \in \chi\}$이다. 실행은 1×10^6번 반복하며 처음 1×10^4번은 burn-in 과정으로 버리고 나머지 반복들을 사용하여 추론을 수행한다. gain factor γ_t의 선택 시 $\eta = 1.0$와 $T_0 = 1000$을 선택하였다. 표본 분포 π를 선택할 때 ξ는 2로 선택하였고, 이는 $(\pi_0, \pi_1, \pi_2, \pi_3) = (0.48, 0.24, 0.16, 0.12)$이다. 즉, 유효한 비율은 48%이다. 종합적으로 elrm, SAMCIS를 비교한 결과는 다음과 같다.

<표 6 > 당뇨병 자료 결과

변수 명		Mean of p-value	SE of p-value
nDQ2	elrm	0.2913	0.00432
	SAMCIS	0.3021	0.00395

elrm 알고리즘과 SAMCIS 알고리즘의 p-value의 평균값은 비슷하지만, SAMCIS 알고리즘의 p-value의 SE(standard error) 값이 더 작으므로 SAMCIS 알고리즘이 더 정확한 추정을 한다고 결론내릴 수 있다. DQ2의 수가 IA2A의 농도에 유의한 영향을 미친다고 할 근거가 충분하지 않다.

4.3 가상 약물실험 자료

R의 elrm 패키지에 내장되어 있는 대조와 처리를 비교하는 가상 약물실험 자료이다. 반응변수 recovered는 환자가 컨디션을 회복했는지 여부이다. 관심있는 공변량과 데이터는 다음과 같다.

· 성별 (sex)
· 처리 (treatment)

<표 7> 가상 약물실험 자료

sex	treatment	recovered	n
1	1	16	27
0	1	10	19
1	0	13	32
0	0	7	21

성별이 남자이면 1이고 그렇지 않으면 0이다. 약물 처리 집단은 1이고 대조군은 0이다. 데이터 첫 번째 행의 의미는 '성별이 남자이고, 처리를 한 27명중 16명이 컨디션을 회복하였다' 이다. 변수 treatment의 추론에 관심이 있다고 하자.

$$H_o : \gamma_{treatment} = 0 \quad \text{vs} \quad H_1 : \gamma_{treatment} \neq 0$$

관심모수는 treatment이고, 장애모수는 절편, sex이다.

elrm 알고리즘의 적용은 다음과 같다. 실행은 1×10^6번 반복하며 처음 1×10^4번은 burn-in 과정으로 버리고 나머지 반복들을 사용하여 추론을 수행한다.

R > elrm(recovered/n ~ sex + treatment, interest = ~ treatment, iter = 1000000, burnIn = 10000, dataset = drugDat)

SAMCIS 알고리즘의 적용은 다음과 같다. $u_{\max}^{(1)}=3$과 $u_{\max}^{(2)}=0$으로 설정함으로써 표본공간 χ를 제한한다. 여기서 $m=3$이고 $\chi=\bigcup_{i=0}^{3}E_i$이므로 $E_0=\{x:U(x)=0, x\in\chi\}$, $E_1=\{x:U(x)=1, x\in\chi\}$, $E_2=\{x:U(x)=2, x\in\chi\}$, $E_3=\{x:U(x)\geq 3, x\in\chi\}$이다. 실행은 1×10^6번 반복하며 처음 1×10^4번은 burn-in 과정으로 버리고 나머지 반복들을 사용하여 추론을 수행한다. gain factor γ_t의 선택 시 $\eta=1.0$와 $T_0=1000$을 선택하였다. 표본 분포 π를 선택할 때 ξ는 2로 선택하였고, 이는 $(\pi_0,\pi_1,\pi_2,\pi_3)=(0.48,0.24,0.16,0.12)$이다. 즉, 유효한 비율은 48%이다. 종합적으로 정확 p-value와 elrm, SAMCIS를 비교한 결과는 다음과 같다.

<표 8> 가상 약물실험 자료 결과

변수 명		Mean of p-value	SE of p-value
treatment	elrm	0.0722	0.00051
	SAMCIS	0.0723	0.00045

elrm과 SAMCIS의 p-value의 평균값은 비슷하지만, SAMCIS의 p-value의 SE 값이 더 작으므로 SAMCIS가 더 정확한 추정을 한다고 결론내릴 수 있다. 환자의 컨디션 회복에 대해 가상 약물 처리는 유의수준 1%에서 유의하게 영향을 미친다고 결론을 내릴 수 있다.

제 5 장 결론

로지스틱 회귀모형의 통계적 추론에서 장애모수를 제외한 관심모수의 통계적 추론에는 정확추론이 사용된다. 기존의 마코브체인 몬테카를로(Markov chain Monte Carlo) 방법을 이용한 정확추론은 국소 트랩(local trap)의 문제점이 발생한다. 본 논문에서는 전체 표본공간 분할을 통해 국소 트랩의 문제를 해결하는 SAMCIS 알고리즘을 이용하여 로지스틱 회귀모형에서 정확추론을 수행해보았다.

정확 p-value와 SAMCIS 알고리즘의 p-value를 비교한 결과 거의 비슷한 결과 값을 가져 SAMCIS 알고리즘이 정확추론을 잘 수행하는 것을 확인하였다. 또한 SAMCIS 알고리즘의 p-value는 elrm 알고리즘보다 더 낮은 bias와 MSE를 제공하기 때문에 SAMCIS 알고리즘이 더 정확한 추정치를 제공한다.

SAMCIS 알고리즘은 기존의 정확추론 알고리즘의 문제점들을 개선한 알고리즘으로, 편향된 자료이거나 불균형자료 또는 표본의 크기에 상관없이 국소트랩의 문제점을 가지지 않고 정확추론에 대한 정확한 추정치를 제공할 것으로 판단된다.

참 고 문 헌

[1] Cheon, S., Liang, F., Chen, Y., Yu, K. (2014). Stochastic approximation Monte Carlo importance sampling for approximating exact conditional probabilities, *Statistics and Computing*, 24(4), 505-520.

[2] Cox, D. R., Snell, J. E. (1970). *Analysis of Binary Data Second Edition*, Chapman and Hall, New York, USA.

[3] Cytel Inc. (2006a). *LogXact 8: discrete Regression Software Featuring Exact Methods*, Cytel Software Corporation, Cambridge, MA. URL http://www.cytel.com/.

[4] Forster, J. J., McDonald, J. W., Smith, P. W. F. (1996). Monte Carlo Exact Conditional Tests for Log-Linear and Logistic Models, *Journal of the Royal Statistical Society B*, 58(2), 445-453.

[5] Forster, J. J., McDonald, J. W., Smith, P. W. F. (2003). Markov chain Monte Carlo exact Inference for binomial and multinomial logistic regression models, *Statistics and Computing*, 13(2), 169-177.

[6] Goorin, A. M., Perez-Atayde, A., Gebhardt, M., Andersen, J. W., Wilkinson, R. H., Delorey, M. J., Watts, H., Link, M., Jaffe, N., Frei, E. (1987). Weekly high-dose methotrexate and doxorubicin for osteosarcoma, *Journal of Clinical Oncology*, 5(8), 1178-1184.

[7] Graham, J., Kockum, I., Sanjeevi, C. B., Landin-Olsson, M., Nysrom, L., Sundkvist, G., Arnqvist, H., Blohme, G., Lithner, F., Littorin, B., Schersten, B., Wibell, L., Ostman, J., Lernmark, A., Breslow, N., Dahlquist, G. (1999). Negative Association Between Type 1 Diabets and HLA DQB1*0602-DQA1*0102 is Attenuated with Age at Onset, *European Journal of Immunogenetics*, 26(2-3), 117-127.

[8] Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications, *Biometrika*, 57(1), 97-109.

[9] Liang, F., Liu, C., Carroll, R. (2007). Stochastic approximation in Monte Carlo computation, *Journal of the American Statistical Association*, 102(447), 305-320.

[10] Mehta, C. R., Patel, N. R. (1995). Exact logistic regression: theory and examples. *Statistics in Medicine*, 14(19), 2143-2160.

[11] Mehta, C. R., Patel, N. R., Senchaudhuri, P. (2000). Efficient Monte Carlo Methods for Conditional Logistic Regression, *Journal of the American Statistical Association*, 95(449), 99-108.

[12] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E. (1953). Equations of State Calculations by Fast Computing Machines, *Journal of Chemical Physics*, 21(6), 1087-1092.

[13] Oster, R. A. (2002). An Examination of Statistical Software Packages for Categorical Data Analysis Using Exact Methods, *The American Statistician*, 56(3), 235-246.

[14] Oster, R. (2003). An Examination of Statistical Software Packages for Categorical Data Analysis Using Exact Methods - PartⅡ, *The American Statistician*, 57(3), 201-213.

[15] Zamar, D., McNeney, B., Graham, J. (2007). elrm: Software Implementing Exact-like Inference for Logistic Regressioin Models, *Journal of Statistical Software*, 21(3).

[16] Zamar, D. (2006). *Markov Chain Monte Carlo Exact Inference for Binomial and Multinomial Logistic Regression Models,* Master's thesis, Statistics and Actuarial Science : Simon Fraser University.