

2000년 미국대선에서 플로리다주의 투표결과 분석

김현철¹⁾

요약

다면량 회귀분석을 통해 플로리다주의 2000년 미국 대선 결과를 분석하였다. 우리는 두 가지 방법으로 이상점을 탐색했는데, 기존의 연구들과 달리, 이상점이 하나가 아니라 는 결론을 얻었다. 이는 그 동안 단일변량회귀분석으로 얻은 사실에 기초해서 팜 비치만이 이상점이고, 그 원인이 나비투표지 때문이었다는 주장이 재고되어야 함을 의미한다. 이중기표로 인한 무효표의 수 분석을 곁들여야 만 그런 주장이 가능하다.

주요용어: 다변량 회귀분석, 다변량 이상점, 나비투표지, 우도비검정, 2000년 미국 대통령 선거.

1. 서론

지난 2000년에 있었던 미국의 대통령선거는 수많은 논쟁을 불러일으켰고, 전 세계 사람들의 관심속에서, 법원의 판결에 따라 결국 부시(Bush) 현 대통령이 대통령으로 확정된 역사상 보기 드문 해프닝이었다.

Gillman(2000)은 2000년 미국대선과 관련해서 법원판결로 끝난 논쟁은 크게 5가지라고 정리했다. 그 중에는 소위 '나비투표지(Butterfly Ballot)'라고 불리우는 팜 비치(Palm Beach) 카운티의 투표용지의 적법성 논쟁이 포함되어 있다. 나비투표지란 투표용지에 후보 이름을 두 옆로 나열했으면서도 기표란은 가운데 한 줄로 배치함으로써 마치 나비가 날개를 펼치고 있는 모양을 하고 있다 해서 불여진 이름이다. 나비투표지가 문제가 되는 것은 기표란을 통상적인 투표지와 달리 좌, 우의 후보를 한 사람씩 교대로 배열함으로써 왼쪽 2번째 후보인 고어(Gore)의 기표란이 3번째가 되고 오른쪽 첫번째 후보인 부캐넌(Buchanan)의 기표란이 2번째가 되었기 때문이다. 투표 직후 많은 유권자들이 고어에게 투표하려 했으나 투표용지가 혼란을 주게 만들어졌기 때문에 자기 의도와는 달리 엉뚱하게 부캐넌에게 투표한 결과가 되었다고 주장했다. 따라서 이는 '투표자의 의지'를 반영하지 못한 투표였다는 범리 논쟁이 벌어진 것이다.

이러한 논쟁이 유권자들의 항의소동에서 전국적으로 관심을 받는 핫이슈로 확산된 것은 Adams와 Fastnow(2000)가 회귀분석을 통해 팜 비치의 투표결과가 추정된 회귀선에서 크게 벗어나는 이상점이라고 이메일을 통해 여러 동료들에게 알리면서부터이다. 이 후 40개 이상의 연구결과가 인터넷에 올라왔는데 O'Keefe(2000)는 이런 연구결과들의 리스트를 링크와 함께 제공하고 있다.

1) (573-701) 전북 군산시 미룡동 산 68번지, 군산대학교 수리정보통계학부, 부교수
E-mail: ki@mhc@kunsan.ac.kr

이 연구들은 모두 회귀분석이나 여러 가지 통계그림을 통해 팜 비치의 투표결과가 이상점임을 밝히거나 확률적으로 발생 가능성이 거의 0에 가까운 값이라는 주장이었다. 다만 Shimer(2000)는 팜 비치의 결과가 그다지 이상점으로 보이지 않는다고 주장했으나, 다른 연구결과들은 일반 독자들에게 팜 비치의 투표결과가 통상적인 기대수준에서 크게 벗어난다는 점을 깨닫게 하는데 충분하다. 물론 투표행위는 매우 복잡한 정치, 사회, 심리적 배경을 갖고 있기 때문에 통계적으로 이상점이라고 해서 정말 이상점이라고 할 수 있느냐에 대해서는 이견이 있을 수 있다. 게다가 팜 비치에서 발견되는 이상현상이 과연 나비투표지때문인가 하는 의문도 제기될 수 있다.

그런데 많은 연구들이 팜 비치의 투표결과가 이상점임을 밝히거나 여기에서 한 걸음 더 나아가 정상적인 경우라면 부캐넌의 실제 득표수가 얼마나 될 것인지를 예측함으로써 고어가 잃은 표가 얼마나 되는지를 계산하는 데만 주로 초점을 두고 있다. 따라서 사용한 방법이 모두 부캐넌의 득표수를 종속변수로 하는 단일변량 회귀분석이 그 주축을 이루고 있다.

이렇게 부캐넌의 득표수만을 단일변량 회귀분석으로 분석하는 것은 암묵적으로 세 가지 가정에서 출발한다. 첫째, 팜 비치 이외에 다른 카운티에서는 이상점이 없다. 둘째, 다른 후보자들의 득표수에는 이상점이 없다. 그리고 이 가정들의 결과로서 팜 비치에서 나타난 문제의 원인은 나비투표지이다. 물론 첫번째 가정에 대해서는 다른 이상점이 방송사가 데이터를 잘 못 집계해서 생긴 오류(Monroe 2000)임을 밝히거나, 팜 비치 외의 이상점일 가능성이 있는 점에 대해서는 모형이 비선형이 될 수도 있다는 가정(Adams와 Fastnow 2000)등을 통해서 비쳐가면서 언급하기도 했다. 다만 Smith(2000)는 다른 이상점이 있을 가능성에 대해서도 검토했다.

그러나 우리는 과연 팜 비치만이 이상점이었는가와 다른 후보들에게서는 이상점이 발견되지 않는지에 대해서도 관심을 가질 필요가 있다. 만약 팜 비치 외에도 이상점으로 보이는 카운티가 있거나 다른 후보의 득표수에도 이상점이 있다면 반드시 Irons(2000)가 했던 이중기표로 인한 무효표의 분석을 곁들여야만 팜 비치에서 나타난 이상현상의 원인이 나비투표지였다고 주장할 수 있기 때문이다. 따라서 이 연구에서는 다변량 회귀모형을 통해 이상점을 찾으려고 시도하였다. 이렇게 함으로써 다른 이상점이 있는지를 살펴봄과 동시에 부캐넌 뿐 아니라 다른 후보의 득표수에 존재할지도 모르는 이상점이 동시에 고려되기 때문이다.

2. 회귀분석에 의한 선행연구들

Adams와 Fastnow(2000)는 간단한 두 가지 이론적 사실에 기초하여 부시의 득표수 혹은 총투표수를 공변량으로 삼아 부캐넌의 득표수를 회귀시키는 단순회귀모형을 제시했다. 그들은 이 모형을 통해 팜비치에서 부캐넌의 득표수가 명백한 이상점임을 보였다. Fox(2000) 역시 다른 후보들의 득표수를 공변량으로 하는 로그변환 모형을 통해 팜 비치의 투표결과가 이상점임을 밝히는 데 기여했다. Monroe(2000)도 로그변환 모형을 통해 비선형성을 고려한 회귀모형을 제시했다.

Thorson(2000)은 1996년의 예비선거 결과 데이터를 공변량으로 사용하는 모형에 1996년과 2000년 사이에 변화한 개혁당(Reform Party, 부캐넌의 소속 정당)의 당원등록수를 공변량으로 포함시킨 회귀모형을 제시하므로써 투표결과만을 사용한 회귀모형과 그 궤를 달리하고 있다. 또 Ruben(2000)은 부캐넌이 보수파의 후보임을 감안하여 특정 카운티가 얼마나 보수적인지를 나타내는 대리변수로 부시의 득표수를 사용하고, 동시에 빈곤정도를 나타내는 변수를 포함하는 회귀모형을 사용하였다. O'Keefe(2000)는 비록 회귀분석을 하거나 모형을 제시하지는 않았지만 특정 카운티가 플로리다의 어느 지리적 위치에 있느냐하는 점이 부캐넌의 득표와 관계가 있음을 발견하였다. 참고로 플로리다주는 지리적으로 북서부, 북부중앙, 북동부, 중앙서부, 중앙, 중앙동부, 남서부, 남동부 등 8개 지역으로 나뉘어 진다. 또 카운티의 크기(총투표수를 대리변수로 사용)가 커질수록 부캐넌의 득표율이 낮아지는 것을 발견했다.

Hansen(2000)은 부캐넌의 득표가 각 카운티의 인구통계적 변인들과 강한 상관관계가 있음을 밝히고, 이런 관계를 회귀모형에 반영시킨 최초의 연구이다. 그가 사용한 인구통계적 변인들에는 65세 이상 인구비율, 흑인비율, 남미계 인구비율, 대학졸업자비율, 가계소득의 중앙값 등이 포함된다. 특히 분산이 인구(N)의 로그값과 반비례 관계에 있을 것으로 보고 이분산문제의 해결을 시도했다.

Wand 등(2000)은 투표 결과가 보고된 46개 주에 걸쳐 4317개 지역(카운티와 도시들)의 데이터를 바탕으로 한 분석에서 팜 비치 카운티가 두 번째로 이상한 결과를 보여주는 카운티이며, 투표자가 10,000명을 넘은 지역 중에서는 가장 이상한 점이라고 주장했다.

Smith(2000)는 두 집단의 데이터를 바탕으로 심도 있는 분석을 하고 있다. 하나는 투표 결과 데이터이고 다른 하나는 인구통계적 변인들이다. 또한 이분산의 문제를 해결하기 위한 여러 가지 변환을 분석한 후, 부캐넌 득표수(y_i)의 제곱근($\sqrt{y_i}$)이 가장 적절한 변환인 것으로 주장하고 있다. 그는 변수선택 기법에 의해 적절한 모형을 결정하고 이 모형에서 얻은 스튜던트화 잔차를 이용하여 이상점을 탐색했다. 그는 팜 비치 이외의 다른 이상점에도 관심을 보였으나 팜 비치를 제외하고 얻은 회귀분석 결과를 바탕으로 이상점은 팜 비치 하나뿐이라고 주장했다.

3. 이 연구에서 사용한 데이터

우선 종속변수로 4후보의 득표수를 사용하기로 했다. 여기에서 4후보란 공화당의 부시 및 같은 보수파인 개혁당에 속하는 부캐넌, 민주당의 고어 및 같은 진보파에 속하는 녹색당(Green Party)의 네이더(Nader) 등이다. 네 후보를 선정한 것은 다른 후보들의 득표수를 공변량으로 삼은 기준의 연구들이 대부분이 네 후보의 득표수에 관심을 보이고 있기 때문이다. 이 네 후보의 득표수를 다변량 회귀모형으로 분석함으로써 앞서 이야기한 다른 후보의 득표수에도 있을지 모르는 이상점을 탐색할 수 있다.

우리는 종속변수 행렬 Y 가 다변량 정규분포가 되는 변환을 찾기 위해 원래의 데이터행렬 Y 와 다른 저자들이 제안한 제곱근변환 행렬, 로그변환행렬에 대해서 카이제곱그림(χ^2 plot)을 그려 보았다. 이 결과 다음 그림 (3.1), (3.2), (3.3)에서 확인할 수 있듯이 로그변환

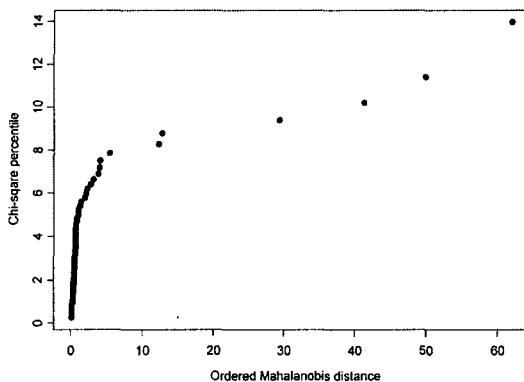


그림 3.1: 원래 데이터에 대한 카이제곱 그림

이, 다변량 정규분포에서 크게 벗어나지 않아, 가장 적절한 변환인 것으로 나타났다. 따라서 우리는 Y의 로그변환을 사용하였다.

한편 우리는 이 연구를 위하여 기존의 연구(Thorson 2000, O'Keefe 2000, Hansen 2000, Smith 2000 등)에서 사용했던 공변량들을 가능한 모두 포함시키기로 했다. 물론 기존의 연구들과 달리 우리는 다변량 회귀모형을 사용할 것이기 때문에 각 카운티가 얼마나 보수적인가를 나타내는 대리변수였던 다른 후보의 득표수는 포함시킬 수 없었다. 그것 자체가 종속변수가 되기 때문이다. 대신 우리는 데이터를 획득할 수 있는 다른 공변량들을 추가했다. 이렇게 해서 얻은 공변량들은 다음과 같다.

지리적 위치1(loc) : 각 카운티가 북동부, 북부중앙, 북서부, 중앙동부, 중앙, 중앙서부, 남동부, 남서부 중 어디에 속하는가에 따라 1, …, 8의 값을 가짐.

지리적 위치2(region) : 위의 지리적 위치1을 북부, 중앙부, 남부의 세 지역으로 구분하여 각각 1, 2, 3의 값을 부여

인구(pop) : 각 카운티의 크기를 나타내기 위한 2000년 센서스 인구, U.S. Census Bureau(2001)

로그인구(logpop) : 인구의 자연로그값

65세이상 인구비율(age65) : 2000년 센서스 결과, U.S. Census Bureau(2001)

18세이상 인구비율(age18) : 2000년 센서스 결과, U.S. Census Bureau(2001)

흑인 인구비율(black) : 2000년 센서스 결과, U.S. Census Bureau(2001)

남미계 인구비율(hispanic) : 2000년 센서스 결과, U.S. Census Bureau(2001)

대출자 인구비율(coll) : 1990년 센서스 결과, U.S. Census Bureau(2001)

가계소득의 중앙값(income) : 모형에 기초한 추정 1997년, U.S. Census Bureau(2001)

가계소득비율(incratio) : , 모형에 기초한 추정 1997년, U.S. Census Bureau(2001)

주택보유비율(houseown) : 주택보유 세대비율, 2000년 센서스 결과, U.S. Census Bu-

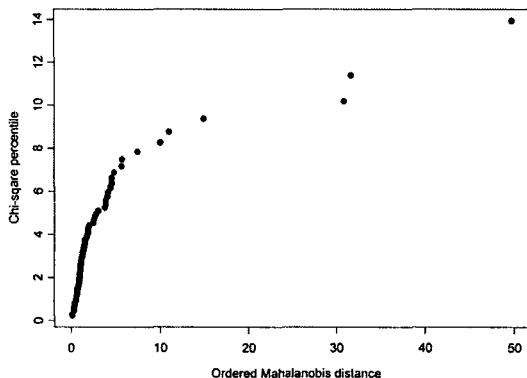


그림 3.2: 제곱근변환 후의 데이터에 대한 카이제곱 그림

reau(2001)

생계곤란자 비율(poverty) : 생계가 곤란한 절대빈곤층 비율, 모형에 기초한 추정 1997년,
U.S. Census Bureau(2001)

민주당원의 수(democ) : 2000년 10월 현재 등록되어 있는 민주당원의 수, Florida State(2001)

공화당원의 수(repub) : 2000년 10월 현재 등록되어 있는 공화당원의 수, Florida State(2001)

개혁당원의 수(reform) : 2000년 10월 현재 등록되어 있는 개혁당원의 수, Florida State(2001)

녹색당원의 수(green) : 2000년 10월 현재 등록되어 있는 녹색당원의 수, Florida State(2001)

총투표율(vote) : 투표자수 / 18세이상 인구, Florida State(2001)

여기서 인구의 자연로그값을 고려한 이유는 2장(회귀분석에 의한 선행연구들)에서 밝혔듯이, Hansen(2000)이 부캐넌의 득표수가 인구의 로그값에 반비례하는 이분산성을 갖고 있다고 밝혔기 때문이다.

이상의 변수들을 가지고 일단 단일변량 회귀분석을 통해 어느 공변량이 의미있는 변수로 나타나는지를 조사하였다. 즉, 각 종속변수에 대해 모든 독립변수를 포함하는 회귀모형을 만든 다음, 단계별회귀방법(진입 F -확률은 0.05, 제거 F -확률은 0.1)에 의해 최적의 모형을 찾았다. 이렇게 해서 각 모형에 포함된 공변량들을 전부 모으면 다음과 같다.

loc, logpop, age18, black, coll, poverty, incratio, houseown, democ, reform, repub, green, vote

우리가 얻은 각각의 단일변량 회귀분석 결과에서도 이상점의 영향이 나타나는 것을 제외하고는 정규확률도표(normal probability plot)를 볼 때 정규성가정에서 크게 벗어나 보이지 않으며, 잔차의 산점도 역시 등분산성을 의심할만한 모습은 발견할 수 없었다.

한편 우리는 어느 한 종속변수에 대해서라도 의미있는 공변량은 모두 다변량 회귀모형에 포함시켰다. 이렇게 한 것은 다변량 회귀모형에서 각 종속변수별로 공통의 독립변수 집합을 통해 모형을 추정함으로써 각 후보의 득표수가 동일한 인구통계적 변인들과 관계를

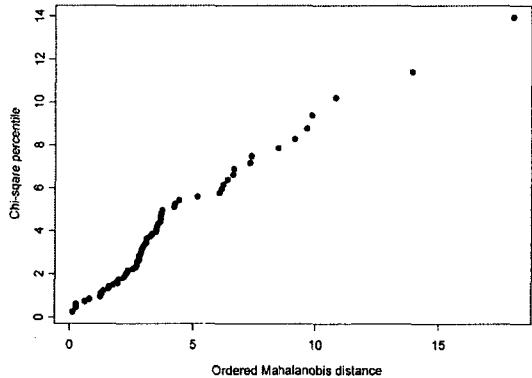


그림 3.3: 로그변환 후의 데이터에 대한 카이제곱 그림

갖는다는 가정에 충실하기 위해서이다. 동시에 우리의 목적이 이상점을 찾는 것이기 때문에 어느 한 종속변수에 대해서라도 의미있는 공변량이라면 종속변수들을 최대한 설명하게 함으로써 다른 숨은 이유가 있기 때문에 이상점으로 나타나는 점(이런 점은 이상점이라 할 수 없다)을 최소화하기 위해서이다. 또 같은 이유에서 다른 통계적인 문제들(다중공선성과 이로 인한 부호의 문제 등)은 논의에서 제외시켰다.

4. 이상점 탐색방법

우리는 이상점을 탐색하기 위해서 두 가지 방법을 사용하였다. 하나는 다변량 회귀분석에서 얻은 잔차들이 다시 다변량 데이터가 되는 점에 착안하여 이 잔차들에 다변량 이상점 탐색방법을 적용하는 것이고, 두번째 방법은 다변량 회귀분석에서 이상점을 검정하는 Srivastava와 Rosen(1998)의 우도비검정방법을 반복적으로 적용하는 것이다.

먼저 p 개의 모수와 k 개의 회귀모형들로 구성되는 다변량 회귀모형은 관측값의 수를 n 이라 하면 다음과 같이 나타낼 수 있다.

$$\mathbf{Y} = \mathbf{BX} + \mathbf{E} \quad (4.1)$$

여기서 \mathbf{Y} 는 $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ 으로 $(p \times n)$ 인 반응값들의 랜덤벡터들로 구성된 행렬이고, \mathbf{X} 는 $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ 으로 $(k \times n)$ 인 알려진 벡터들의 행렬, \mathbf{B}' 는 $\{\beta_1, \dots, \beta_p\}$ 로 $(k \times p)$ 인 미지의 회귀계수들의 행렬이고, 끝으로 \mathbf{E} 는 $\{\epsilon_1, \dots, \epsilon_n\}$ 으로 $(p \times n)$ 인 확률난수벡터들의 행렬이다. ϵ_i ($i = 1, \dots, n$) 들은 서로 독립이고 동일한 정규분포, $N(\mathbf{0}, \Sigma)$ 를 따른다. 즉 동일한 단일변량 회귀모형의 오차항은 모든 관측값에 대해 서로 독립이지만, 동일한 관측값에 대해 서로 다른 단일변량 회귀모형의 오차항은 서로 독립이 아니다.

4.1. 다변량 이상점 탐색방법

위 모형 (4.1)에서 오차항이 다변량 정규분포를 따른다는 가정을 만족하면 반응값 행렬도 역시 다변량 정규분포를 따라야 한다. 우리는 앞에서 이 가정을 충족시키기 위해 반응값 행렬을 로그변환했음을 밝힌 바 있다. 다변량 회귀모형을 추정하여 얻은 잔차에 대해 다시 다변량 정규분포를 따르는지 확인하기 위해 카이제곱그림을 그려보았다. 이 그림 (4.1)로부터 우리는 잔차 행렬이 다변량 정규분포에서 크게 벗어나는 것으로 의심할 만한 점을 발견할 수 없었다.

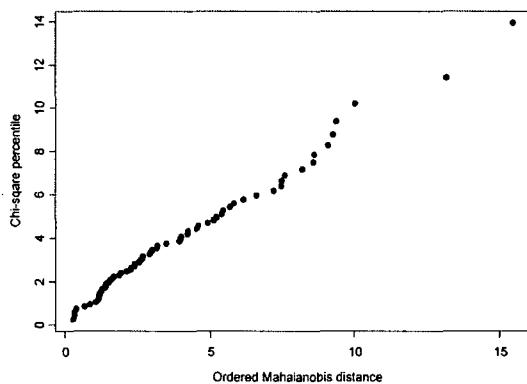


그림 4.1: 다변량 회귀분석 후 잔차에 대한 카이제곱 그림

따라서 우리는 이 잔차행렬에 다변량 이상점 탐색방법을 적용시킬 수 있다. 여기에서 Cook(1986)이 단일변량 회귀분석에서 제안한 국지적 영향(local influence) 방법을 다변량 정규모형으로 확장한 김명근과 정강모(2000)의 방법을 사용하였다. 이 방법으로 나타낸 2차원그림은 그림 (4.2)에 있는데 이 그림을 보면 우리는 쉽게 50, 7, 39번 관측값이 이상점임을 알 수 있다. 이 점들은 순서대로 팜 비치 카운티, 칼하운(Calhoun) 카운티, 그리고 메디슨(Madison) 카운티이다. 이 외에도 34번, 56번 관측값이 추가로 이상점으로 판정될 가능성이 보이는데 이들은 레이크(Lake) 카운티와 세인트 루시(Saint Lucie) 카운티이다.

4.2. 우도비에 의한 다변량 이상점 검정방법

Srivastava와 Rosen(1998)은 모형 (4.1)에서 이상점이 없는 경우의 모형을 모형 H 라고 하고 i 번째 관측값이 평균이 이동한(mean-shifted) 이상점인 경우의 모형을 H_i 라고 했을 때 각각의 공분산 행렬을 추정하여 두 공분산의 비를 이용한 이상점 검정방법을 제안하였다. 모형 H 하에서는

$$\mathbf{E}[\mathbf{Y}] = \mathbf{BX} \quad (4.2)$$

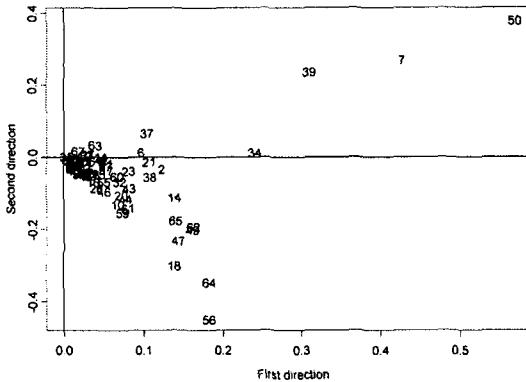


그림 4.2: 국소적 영향 방법에 의한 이상점 탐색

이며, Σ 의 MLE는 $n^{-1}\mathbf{S}$ 이다. 여기서 $\mathbf{S} = \mathbf{Y}(\mathbf{I} - \mathbf{R})\mathbf{Y}'$ 이며, $\mathbf{R} = \mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}$ 이다. 또 모형 H_i 하에서는

$$\mathbf{E}[\mathbf{Y}] = \mathbf{B}\mathbf{X} + \delta \mathbf{a}_i' = \mathbf{B}^* \mathbf{X}_i^* \quad (4.3)$$

이며, Σ 의 MLE는 $n^{-1}\mathbf{S}_i$ 이다. 여기서 $\mathbf{S}_i = \mathbf{Y}(\mathbf{I} - \mathbf{R}_i)\mathbf{Y}'$ 이며, $\mathbf{R}_i = \mathbf{X}_i^{*\prime}(\mathbf{X}_i^* \mathbf{X}_i^{*\prime})^{-1} \mathbf{X}_i^*$ 이다. 또 $\mathbf{B}^* = (\delta, \mathbf{B})$, $\mathbf{X}_i^* = \begin{pmatrix} \mathbf{a}_i' \\ \mathbf{X} \end{pmatrix}$, 그리고 δ 는 미지의 상수벡터, \mathbf{a}_i 는 n 차원 단위행렬(identity matrix)의 i 번째 열이다.

정규모형에서 가설 H 대 H_i 에 대한 우도비 검정(likelihood ratio test)은

$$\lambda_i = \frac{|\mathbf{S}|}{|\mathbf{S}_i|}$$

이다. Srivastava등은 이 검정통계량이 다음과 같음을 보였다.

$$\frac{f-p+1}{p}(\lambda_i - 1) \sim F_{p, f-p+1} \quad (4.4)$$

여기서 $f = n - k + 1$ 이다.

이 방법은 하나의 이상점을 검정하는 방법이므로 우리는 이 방법을 반복적으로 적용하는 다단계 방법을 채택하였다. 즉 하나의 이상점이 있는지를 검정하고, 이상점이 발견되면 그 이상점을 제외한 나머지 관측값에 대해서 다시 다변량 회귀분석을 한 뒤 이상점 검정을 수행하는 반복 작업을 더 이상 이상점으로 판정되는 관측값이 나오지 않을 때까지 반복 적용하였다. 이런 방법은 동반효과(swamping effect)나 가장효과(masking effect)로 부터 자유롭지 못할 수 있다.

이 방법에 의해 이상점으로 판정된 점들은 표 (4.1)과 같다. 우리는 이상점이 아닌데도 이상점으로 판정되는 제1종의 오류를 가능한 작게 하기 위해 이 검정의 유의수준을 1%로 하였다. 참고로 4개의 이상점이 식별된 뒤에 그 다음으로 이상점일 가능성의 높은 점은 34번

표 4.1: 반복적 Srivastava방법에 의해 식별된 이상점들

단계	번호	카운티 이름	유의확률
1	50	팜 비치	0.0001
2	7	칼하운	0.0059
3	39	메디슨	0.0049
4	56	세인트 루시	0.0008

관측값(유의확률은 0.0118)이었는데, 이는 앞에서 첫 번째 방법으로 얻은 결과와 크게 유사한 결과를 보여주는 셈이다. 또 64번 관측값의 유의확률은 0.1192였다.

5. 결론

이상에서 우리는 지난 2000년 미국대선에서 있었던 플로리다주의 투표결과의 분석을 시도하였다. 이 분석에서는 팜 비치 카운티의 투표결과가 이상점이라는 기준의 연구결과들이 부캐넌의 득표수에 대한 단일변량 회귀모형에 근거해서 이루어졌음을 발견했다. 그리고 이런 연구들이 대부분 무의식적이든 의식적이든 일정한 전제 위에 세워진 것임을 발견했다. 이런 전제들이 만약 사실이 아닐 때에는 그 연구의 결과 자체가 의심받을 수도 있기 때문에 이런 전제의 타당성을 확인할 필요가 있었다.

따라서 이 연구에서는 다변량 회귀분석을 이용하여 팜 비치 이외에 다른 이상점이 있는지를 두 가지 방법으로 확인했다. 그 결과 두 방법의 결과가 정확히 일치하지는 않았으나 적어도 팜 비치 이외에도 통계적으로 이상점이라 말할 수 있는 점이 더 있음을 확인할 수 있었다. 이 사실은 기존의 연구들이 부캐넌의 득표수에만 집착해서 팜 비치만이 이상점이라 주장했던 것과는 상당한 차이가 있는 것이다. 또 기존의 연구들은 잘못된 전제 위에서 출발함으로써 팜 비치만이 이상점이고 따라서 당연히 그 원인은 나비투표지에 있다고 결론내렸으나, 우리의 연구결과는 팜 비치 이외에도 이상점은 더 있었고 따라서 이중기표로 인한 무효표의 수를 함께 분석해야만 팜 비치의 문제가 나비투표지 때문이었다고 주장할 수 있다는 결론을 얻었다.

참고문헌

- [1] 김명근, 정강모. (1999). <S-PLUS를 이용한 다변량자료분석>, 교우사.
- [2] Adams, G. D. and Fastnow, C. (2000). A Note on the Voting Irregularities in Palm Beach, FL, <http://madison.hss.cmu.edu/>
- [3] Cook, R. D. (1986). Assessment of Local Influence, *Journal of the Royal Statistical Society(B)*, 48, 133-169.

- [4] Florida State (2001). Division of Elections, <http://election.dos.state.fl.us/online/index.shtml>
- [5] Fox, C. R. (2000). A vote for Buchanan is for a vote for Gore? - An analysis of the 2000 presidential election results in Palm Beach, Florida, <http://faculty.fuqua.duke.edu/~cfox/Bio/election2000note.pdf>
- [6] Gillman, H. (2000). Materials Relating to the Role of Courts, Law, and Politics in Election 2000, <http://www.usc.edu/dept/polsci/gillman/election2000.html>
- [7] Hansen, B. E. (2000). Who Won Florida? - Are the Palm Beach Votes Irregular?, <http://www.ssc.wisc.edu/~bhansen/vote/vote.html>
- [8] Irons, J. S. (2000). A preliminary look at the vote count in Palm Beach, Florida, <http://www.amherst.edu/~jsirons/election/>
- [9] Monroe, B. L. (2000). Did votes intended for Gore go to Buchanan?, <http://www.indiana.edu/playpol/pbmodel.pdf>
- [10] O'Keefe, J. (2000). Palm Beach County Election Irregularities, <http://www.bestbookmarks.com/election/>
- [11] Ruben, M. (2000). Statistical Analysis of the Vote in Palm Beach, Florida, <http://weber.ucsd.edu/~mruben/florida.htm>
- [12] Shimer, R. (2000). Election 2000, <http://www.princeton.edu/~shimer/election.html>
- [13] Smith, R. L. (2000). A Statistical Assessment of Buchanan's Vote in Palm Beach County, <http://www.stat.unc.edu/faculty/rs/palmbeach.html>
- [14] Srivastava, M. S. and von Rosen, D. (1998). Outliers in Multivariate Regression Models, *Journal of Multivariate Analysis*, **65**, 195-208.
- [15] Thorson, G. (2000). Estimating the Expected Vote for Buchanan in Palm Beach County, Florida: An Alternative Model Based on Prior Support for Buchanan, <http://www.mrs.umn.edu/~Egthorson/floridadispute.htm>
- [16] U.S. Census Bureau (2001). State and County QuickFacts, <http://quickfacts.census.gov/>
- [17] Wand, J. N., Shotts, K.W., Sekhon, J. S., Mebane Jr., W. R., and Herron, M. C. (2000). Voting Irregularities in Palm Beach County, <http://elections.fas.harvard.edu/wssmh.old/>

[2001년 10월 접수, 2002년 2월 채택]

Statistical Outliers in Florida Counties at the Presidential Election 2000

Hyun Chul Kim¹⁾

ABSTRACT

We searched outliers in the votes data of the State of Florida at presidential election 2000. We used a multivariate regression analysis. We got there were several outliers including Palm Beach County. It means that we should analyze the number of disqualified ballots which were double-punched as well as the votes, to insist the "Butterfly Ballot" made Palm Beach outlier.

Keywords: Multivariate Regression; Multivariate Outliers; Butterfly Ballot; Likelihood Ratio Test; Presidential Election of U.S.A. 2000.

1) Associate Professor, Faculty of Mathematics, Informatics and Statistics, Kunsan National University.
E-mail: kimhc@kunsan.ac.kr