

기획논문

사회과학연구에서 사건계수 방법

이해와 적용

이 상 환

사건계수(Event-Count) 방법은 사건진계 자료에 대한 적절한 통계적 분석을 가능하게 하는 최선의 통계적 분석기법이라고 할 수 있다. 이 방법은 경험적인 사회과학자들에게 자료가 숫자로만 되어 있으면 무조건 선형회귀분석을 하는 우를 범하지 않도록 하여 준다. 사건계수 방법은 이산(discrete)적인 종속변수를 위한 분석기법이며, 그 대표적인 기법으로 포아송(Poisson) 분석방법과 음이항(Negative Binomial: NB) 분석방법이 있다. 이러한 분석기법의 적용 이유는 특정한 연구가 다루는 종속변수가 사건 발생횟수를 나타내고, 그 값이 음수가 아니며, 따라서 그 데이터가 정규분포 되어 있지 않고, 더욱이 각 사건간에 독립성을 가정(Poisson)하거나 혹은 그 연관성을 고려(NB)해야 함에 있는 것이다. 이와 같이 변수와 자료(data)의 성격을 적절히 고려한 통계적 방법의 선택과 이것의 경험적 사회과학연구에의 적절한 활용은 정확한 이론 검증을 위해 필요한 것이다.

1. 서론: 사건계수 방법의 소개

경험적 사회과학연구에는 다양한 종류의 분석 방법들이 사용되어 왔으며, 그 중에서도 회귀분석(*regression analysis*)의 사용이 가장 흔한 것이었다. 그러나 “이러한 회귀분석이 적절하게 행해지고 있는가?” 라는 질문에 자신 있게 답하는 사회과학자들이 그리 많지 않을지도 모른다. 사회현상에 대한 경험적 연구에서 회귀분석이 가장 널리 사용되는 유의한 통계학적 도구의 역할을 해온 것을 인정하면서도 상당수의 경험적 연구에서 이것이 오용되어 왔음도 인정해야 한다. 회귀분석, 특히 선형회귀분석(*linear regression analysis*) 방법을 사용하기 위해서 여러가지 전제조건 내지는 가정들이 필요하고, 이들이 충족되지 못하면 아주 부적절한 결과가 발생함에도 불구하고, 많은 연구자들이 습관적

이상환은

미국 미시간주립
대학교에서 정치학
박사학위를 받고
현재 창원대학교
국제관계학과
교수로 재직중이다.

leepol@sarim.chang-
won.ac.kr

으로 회귀분석을 행하고 그 결과를 맹목적으로 믿고 결론을 내린다. 이러한 오류들 중 대표적인 것의 하나는 종속변수가 연속(*continuous*) 변수가 아님에도 불구하고 단지 구간(*interval*) 변수라는 이유만으로 습관적으로 선형회귀분석을 함으로써 범하는 통계적 오류¹⁾이다.

여기서 우리의 의문은 “선형회귀분석의 사용을 위한 여러 전제조건들 중의 하나인 종속변수가 연속변수이어야 한다는 조건이 맞지 않는 경우에 과연 어떠한 통계적 방법이 바람직한가?” 하는 것이다. 왜냐하면, 종속변수가 이산변수인 경우의 선형회귀분석은 독립변수들의 효과의 크기를 심각하게 잘못 평가할 수 있기 때문이다. 종속변수가 이산변수, 즉 불연속적인(*discrete*) 값인 0, 1, 2, 3, ... 을 갖는 경우는 주로 지속기간(*duration*)이나 사건집계(*event count*)의 예로, 정치지도자의 집권연수, 시위의 지속일수, 쿠데타가 일어난 횟수, 국제사회에서의 협력과 갈등의 횟수, 그리고 전쟁의 횟수 등이 이에 해당된다. 이러한 경우 선형회귀분석은 다른 적절한 기법으로 대체되어야 한다. 사건 집계를 통한 자료, 즉 특정한 사건의 발생 빈도수를 의미하는 종속변수를 다루는 경우 우리는 사건계수 분석기법(*event-countanalysis*)을 적용해야 한다. 그 기법의 종류에는 포아송(*Poisson*), 일반화된 사건계수(*Generalized Event Count: GEC*), 그리고 음이항분산함수(*Negative Binomial Variance Function, NBVF* 혹은 *NB*)의 분석방법이 있다. 따라서 이 논문에서는 종속변수가 이산변수인 경우에 적합한 모델들에 대한 소개와 더불어, 그 모델들을 이용한 사례연구를 통하여 통계적 결과들을 어떻게 해석하는가를 보여줄 것이다(Allison, 1984; Yamaguchi, 1991; Londregan, 1996).

선형회귀분석과 같은 방법은 연구자로 하여금 본래 불연속적인 개념을 가진 변수를 연속적으로 정의를 내리게 하여, 결과적으로 측정의 오류나 편견을 발생하게 하고 심각한 분량의 정보를 잃게 한다. 이러한 문제들을 극복하기 위해서 종속변수가 불연속적인 성격을 띠는 경우에는 주로 사건계수 방법이 사용되는 것이 바람직하다. 물론 이에 앞서서 연구의 목적이 무엇인가도 중요하고, 연구에 쓰이는 독립

변수들의 성질 즉 독립변수가 연속변수인가 불연속변수인가도 살펴보아야 한다.

포아송 분석기법은 종속변수가 특정한 사건이 발생한 횟수와 같은 발생 건수이고, 만약 그 자료(*data*)가 매우 낮은 값들을 가지고 있다면 특별히 유용한 것이다. 포아송 분포는 각 사건이 서로 독립적이고 일정한 발생 확률을 가지고 있다는 전제하에 일정기간에 어떤 사건이 몇 번 일어날지의 확률을 묘사한다. 포아송 분포의 모양은 그 평균의 값에 달려 있으며, 이는 분산(分散, *variance*)과 일치한다. 만약 평균이 0에 접근하면, 그 분포는 수직축에 크게 치우친 모양의 그래프를 보여주고, 평균이 매우 크고 수직축으로부터 멀리 떨어져 있으면, 포아송 분포는 대략 정규분포와 유사하게 될 수 있다.

포아송 모델에서 두 가지의 특징을 주목할 만한데, 첫째로, 회귀분석과는 다르게 포아송 모델은 그 데이터가 등분산적(*homoscedastic*)이라고 가정하지 않는다. 그러므로 종속변수의 값들과 관련된 분산들은 동일하지 않으며 정규분포가 될 수 없다. 둘째로, 종속변수의 값이 매 개변수 λ_i 를 가진 포아송 분포에 의해 파생된 독립된 사건들의 발생 건수이다. 독립된 사건들이란 한 사건의 발생이 다른 사건의 발생 확률을 증가 혹은 감소시키지 않는다는 것을 의미하며, 이를 위반시 포아송 모델은 부적절하다는 것이다. 예를 들어, 전염병의 발생에 포아송 모델을 적용한다는 것은 부적절한 것이다. 따라서 포아송 모델의 타당성은 종속변수가 일정기간 동안 주어진 사건의 발생 횟수를 나타낸 것이고, 그 값이 크지 않으며, 즉 과잉산포(過剩散布)화 되어 있지 않은 경우, 그리고 그 사건들이 서로 독립적으로 발생하는 경우에 가장 적합하다. 이 모델에서 분산이 평균보다 크면, 그 데이터는 '과잉산포'(*overdispersion*)의 문제를 안고 있는 것이다(Lovett & Flowerdew, 1989:191).

포아송 모델에서는 사건간의 독립성을 전제로 한다. 그러나 만약 이와 같은 가정이 틀리다면 포아송 모델의 적용에 문제가 생기는 것이다. 이러한 경우, 각 사건간의 상호연관성을 알아보기 위해, 두 가

지 종류의 '전파'(contagion)를 고려해 보고 이에 적합한 모델을 적용하여야 한다. '음전파'(negative contagion)는 한 사건이 다른 사건의 발생 가능성을 감소시켜 주는 경우를 말하며, '양전파'(positive contagion)는 한 사건이 다른 사건의 발생 가능성을 증가시켜 주는 경우를 말한다. 이러한 전파의 성격을 파악하기 위해 일반화된 사건계수(Generalized Event Count) 모델이 적용되며, 그 성격이 양전파인 경우 음이항(Negative Binomial) 모델은 통계적으로 그 전파성을 교정해 준다. 즉, 포아송 모델에서는 $E(Y_i) = \lambda_i$ 그리고 $\text{Var}(Y_i) = \lambda_i$, 그러나 음이항 모델에서는 $E(Y_i) = \lambda_i$, 그리고 $\text{Var}(Y_i) = \lambda_i[1 + \exp(\gamma)]$ 이다. 음이항 분포에는 γ 값이 포함되는데, 그 값이 음수인 경우 Y_i 의 분산은 감소되고, 그 값이 양수인 경우 Y_i 의 분산은 증가된다. 그리고 절대값 γ 의 크기가 크다면 이는 '과잉산포'를 의미하고, 이는 '전파'가 존재함을 의미한다. 음이항 모델은 사건간에 '전파'가 존재하는가를 보여주고, 이를 통계적으로 교정해 줌으로써 올바른 분석결과를 이끈다.

2. 사건계수 모델들에 대한 이해 : 포아송 모델과 음이항 모델

사건계수 모델들은 그 자료의 분포형태와 사건계수의 상한선에 따라서 크게 네 가지 유형으로 나눌 수 있다. 첫째 유형은 자료 분포의 형태가 과잉산포가 아니며 사건계수의 상한선이 높거나 상한선이 없는 경우로 포아송 모델이며, 둘째 유형은 자료 분포의 형태가 과잉산포이며 사건계수의 상한선이 높거나 없는 경우로 음이항 모델이 있다. 이들 두 가지 유형이 사건계수 모델의 대표적인 유형들이며, 일반화된 사건계수 모델의 변형들에 해당된다. 이외에 이항(binomial) 모델²⁾과 베타-이항(beta-binomial) 모델³⁾이 있으나 자주 사용되지 않는다(Achen, 1996:170~172).

이 논문에서 소개되는 모델은 사건계수 분석을 위한 두 가지 모델인 포아송 모델과 음이항 모델이다. 우선 이들 모델에 대한 구체적인 내용들을 설명하고자 한다.

포아송 분석기법은 이미 설명하였듯이 종속변수가 일정기간 동안 주어진 사건의 발생 횟수를 나타낸 것이고, 그 값이 과잉산포(*overdispersed*) 되어 있지 않으며, 그리고 그 사건들이 서로 독립적으로 발생하는 경우에 가장 적합하다. 여기서는 1차적으로 각 사건간의 독립성을 전제로 포아송 모델을 다루고, 그 다음에 그 사건간의 '전파성'의 존재여부 확인과 그 교정을 위해 음이항 모델 분석을 행할 것이다.

1) 포아송 모델

포아송 분포는 일정한 시간, 공간단위 범위 내에서 확률변수가 무한개의 정수값 0, 1, 2, 3, ... 을 갖는 실험에 적용되는 분포를 말한다. 일반적으로 포아송 분포는 다음의 수식으로 표현될 수 있다(Gourieroux, 1984:702~703; Lee, 1986:690~691).

$$\Pr(Y_i = y_i) = f(y_i) = \exp(-\lambda_i) \lambda_i^{y_i} / y_i! \quad (y_i = 0, 1, 2, \dots).$$

여기서, $\lambda_i = \exp(x_i b) = \exp(\sum_{k=1}^n x_{ik} b_k)$

포아송 분포에서 평균과 분산은 λ_i (사건발생률)와 같은 값이며

$$[\lambda_i = E(Y_i) = \text{Var}(Y_i)],$$

그 log 우도함수(*likelihood function*)는

$$L(b) = -\sum_{i=1}^n \lambda_i + \sum_{i=1}^n y_i \log \lambda_i - \sum_{i=1}^n \log(y_i!)$$

$$= \text{constant} - \sum_{i=1}^n \exp(x_i b) + \sum_{i=1}^n y_i x_i b$$

$L(b)$ 를 극한까지 증가시키면,

$$dL/db = -\sum_{i=1}^n x_i' \exp(x_i b) - y_i = 0 \text{ or } \sum_{i=1}^n x_i' [\lambda_i - y_i] = 0$$

여기서 b 의 최대우도추정값(MLE), b^{MLE} 은 다음의 등식과 관련된다.

$$\sum_{i=1}^n x_i' y_i = \sum_{i=1}^n x_i' \exp(x_i b^{MLE})$$

$$d^2L / db \cdot db' = -\sum_{i=1}^n x_i' x_i \exp(x_i b)$$

b^{MLE} 는 종속변수에 대한 독립변수들의 영향을 나타내며 최대우도 추정(MLE)의 분석기법을 사용함으로써 측정될 수 있다. 포아송 분석기법은 사건발생수로 기록된 데이터에 대한 각 이론의 경험적 타당성을 검증하는데 유용한 통계기법이다. 그런데 그 기본 포아송 모델은

그 데이터에 있는 선택편의(selection bias)를 제거함으로써 더욱 타당성을 갖게 된다. 즉, 검증할 데이터에 0이라는 값이 존재치 않으면 이에 대한 표준(standard) 포아송 모델의 적용은 문제가 있다. 이러한 경우에 0에서 절단된(truncated-at-zero) 포아송 모델을 적용함이 타당한 것이다(Martin, 1992:75~76). 이 새로운 모델에서 $y_i > 0$ 이다.

2) 음이항 모델

포아송 모델은 사건간의 독립성을 전제로 한다. 그러나 만약 이와 같은 가정이 틀리다면 포아송 모델의 적용에 문제가 생기는 것이다. 이러한 경우, 각 사건간의 상호연관성을 고려하여, 두 가지 종류의 '전파'(contagion) — 양전파 혹은 음전파 — 를 생각해 보고 이에 적합한 모델을 적용하여야 한다. 이와 같은 점을 고려하는 것이 음이항(Negative Binomial) 분석기법이다. 즉, 포아송 모델에서는 $E(Y_i) = \lambda_i$ 그리고 $Var(Y_i) = \lambda_i$ 이다. 그러나 음이항 모델에서는 $E(Y_i) = \lambda_i$, 그리고 $Var(Y_i) = \lambda_i[1 + \exp(\gamma)]$ 이다. 음이항 분포에는 γ 값이 포함되는데, 절대값 γ 의 크기는 '산포'(dispersion) 혹은 '전파'(contagion)의 정도를 보여주고, 그 양·음이 '전파성'의 성격을 결정한다. 음이항 모델은 각 사건간에 '전파'(contagion)가 존재하는가를 보여주고, 이를 통계적으로 교정해 줌으로써 올바른 분석결과를 이끈다.

음이항 분포는 다음의 수식으로 표현될 수 있다(King, 1989:136~140).

$$\begin{aligned} \Pr(Y_i = y_i) &= f(y_i) \\ &= [\Gamma(\lambda_i/\theta + y_i) / y_i! \Gamma(\lambda_i/\theta)] \theta y_i (1 + \theta)^{-(\lambda_i/\theta + y_i)} \\ &\quad (y_i = 0, 1, 2, \dots). \end{aligned}$$

여기서, $\delta^2 = 1 + \theta$ (NB 모델의 분산 - 평균 비율), $\theta = \exp(\gamma)$, γ (dispersion 수준 혹은 각 사건간의 contagion 정도).

포아송 모델에서와 마찬가지로 bMLE(종속변수에 대한 독립변수의 영향), 그리고 γ (산포수준) 값은 최대우도추정법을 사용함으로써 측

정될 수 있다. 음이항 모델 또한 그 데이터상의 0의 값 존재여부에 따라 표준(*standard*) 혹은 0에서 절단된(*truncated-at-zero*) 음이항 모델을 적용한다.

3. 사건계수 모델들의 적용 사례

한국과 미국간의 무역갈등에 대한 경험적 사례연구

구체적 사례로서 소개되는 모델은 국제무역분쟁과 관련된 것으로서 사건집계로 된 종속변수를 가지고 있다. 한미간의 무역분쟁에 대한 결정요인들의 일반적 영향을 살펴보기 위해서 그리고 선형회귀분석이 직면한 한계점을 해결하기 위해서 이 연구는 포아송 분석기법을 사용한다. 그 기법은 사건 발생수(무역분쟁의 횟수)로서 기록된 종속변수를 평가하는 유용한 분석기술인 것이다.

국제무역분쟁과 관련된 두 가지 주요한 보완적인 이론적 시각들이 있는데, 이들은 패권안정 이론과 잉여능력 이론이다. 전자는 국제무역관계에 있어 패권국의 역할을 강조한 것이고, 후자는 변화하는 수요패턴의 견지에서 전반적 수준의 생산능력에 있어 급격한 조정 가능성에 그 관심을 둔 것이다. 다른 말로 바꾸면, 패권안정 이론은 국가들간의 관계(한미 무역관계)가 국가들 사이의 힘의 특정한 배분(한국에 대한 미국의 영향력)에 의해 결정된다고 주장을 하고, 잉여능력 이론은 무역갈등을 과도한 생산에 의해 초래된 어려운 경제상황(한국의 증가된 수입침투와 관련한 악화되어 가는 미국의 전반적 경제상황)과의 함수관계로 파악한다.

요약컨대, 패권안정이론에 의하면 한미 무역갈등은 경제적 그리고 군사적 힘의 견지에서 미국의 한국에 대한 상대적 영향력에 의해 결정되는 것이며, 동북아시아에서의 동서긴장은 그 지역에서 미국의 군사적 입지를 강화함으로써 무역갈등과 관계가 있다는 것이다.⁴⁾ 또한 잉여능력이론에 의하면, 한미 무역분쟁은 미국의 경제상황에 의해 설

명될 수 있는 것이다. 즉, 악화되어 가는 미국의 국내경기가 한미 무역분쟁을 야기하고, 개선되어 가는 경제적 상황이 그 분쟁을 완화시키는 것이다. 그리고 미국에 대한 한국의 수입침투의 증감이 한미 무역갈등에 영향을 끼치는 한국에 대한 미국 무역정책의 보호적 혹은 개방적 방향을 결정하는 것이다. 여기서 미국의 경제상황은 미국의 실업률과 GNP 성장률로 측정되며, 미국에 대한 한국의 수입침투는 미국의 총수입액 중 한국산품이 차지하는 비율로 측정된다. 이들 자료는 Citi Bank Data Base와 U. S. General Imports (by the U. S. Department of Commerce) 등으로부터 얻어진다.

이 실험을 위한 연구에서는 미국과 한국 사이의 무역분쟁의 정도를 측정하기 위해서, 한국에 대해 1970~1989년 기간 동안 미국 국제무역위원회(USITC)에 제소된 분쟁사례를 사용한다. 즉, 한미 무역분쟁의 심화는 제소된 분쟁사례의 증가와 관련되고, 그 분쟁의 약화는 제소된 분쟁사례의 감소와 관련된다. 이 연구에서 선형회귀분석기법을 적용할 시 20년의 기간은 상관관계의 검증을 위해 다소 불충분한 기간이 되나, 사건계수 분석기법의 적용은 바로 그와 같은 경우에 적합한 분석기법이 된다. 아울러 선형회귀분석기법을 사용할 경우 시계열 분석에 따른 자동상관관계 문제를 고려해야 하나, 사건계수분석기법은 최대우도추정법의 사용으로 그러한 문제를 고려할 필요가 없고 단지 사건간의 독립성의 문제(포아송 모델의 경우)만을 고려한다. 이 연구는 한미 무역분쟁에 두 가지 주요한 주장을 적용하여 경험적으로 비교 분석한다. 처음에는 포아송 모델을, 뒤이어 음이항 모델을 적용하여 각 이론의 타당성을 검증해 보는 것이다. 아울러 이 논문은 패권안정 모델과 잉여능력 모델을 다루고 있으며, 그 경험적 결과를 얻기 위해 각각 포아송과 음이항 분석기법이 사용된다.⁵⁾

1) 포아송 모델

포아송 분석기법에 근거한 각 모델은 다음과 같다.

$$\Pr(Y_i = y_i) = f(y_i) = \exp(-\lambda_i) \lambda_i^{y_i} / y_i! \quad (y_i = 0, 1, 2, \dots).$$

$$\lambda_i = \exp(x_i b) = \exp(\sum_{k=1}^n x_{ik} b_k).$$

y_i = 한미간의 무역분쟁

$x_{i1}, \dots, x_{ik} = i_{ih}$ 관찰과 관련된 k 외생변수들 = (패권안정 모델)

X_1 = 미국의 한국에 대한 경제적 패권

X_2 = 미국의 한국에 대한 군사적 패권

X_3 = 동북아의 동서긴장

(잉여능력 모델)

X_4 = 미국의 경제적 상황(실업률)

X_5 = 미국의 경제적 상황(GNP 성장률)

X_6 = 한국의 수입침투

$b_1, \dots, b_k = k$ unknown parameters

= $b_1 - b_3$ (패권안정모델 모수)

$b_4 - b_6$ (잉여능력모델 모수)

이 모델분석에서 데이터상의 선택 편의(selection bias)를 고려하여 데이터상 0의 값을 갖고 있지 않은 경우 truncated Poisson 모델이 적용되고, 0의 값을 갖고 있는 경우 standard Poisson 모델이 적용된다. 한미간의 무역분쟁은 그 데이터의 성격상 표준 포아송 모델이 적용된다.

2) 음이항 모델

음이항 분석기법에 근거한 각 모델은 다음과 같다.

$$\Pr(Y_i = y_i) = f(y_i) = [\Gamma(\lambda_i / \theta + y_i) / y_i! \Gamma(\lambda_i / \theta)] \theta^{y_i} (1 + \theta)^{-(\lambda_i / \theta + y_i)}$$

($y_i = 0, 1, 2, \dots$).

$$\lambda_i = \exp(x_i b) = \exp(\sum_{k=1}^n x_{ik} b_k).$$

$y_i, x_{i1}, \dots, x_{ik}, b_1, \dots, b_k$ 는 포아송 모델과 동일하다.

이 음이항 모델에서 포아송 모델에서의 마찬가지로 한미간의 무역 분쟁은 표준 음이항 모델이 적용된다.

3) 결과의 이해방법

한미 무역분쟁을 설명함에 있어 보다 중요한 모델 혹은 요인들이 무엇인가를 파악하기 위해 이 실례 연구는 log 우도비 (*log likelihood ratio*, LLR), *t* 값, 그리고 기타 통계적 추론을 통해 분석한다.

한국과 미국 사이의 무역분쟁은 미국의 경제적 패권과 한국의 동서 긴장과 같은 변수들을 포함하는 패권안정 모델에 의해서 설명될 수 있다. 그러나 단지 한 변수, 즉 미국의 한국에 대한 경제적 패권만이 그 이론의 기본적 주장과 일치한다. 미국의 실업률과 한국의 수입침투를 포함하는 잉여능력 모델은 한미무역관계에 보다 잘 적용된다. 그 경험적 결과는 기본적으로 잉여능력 가설에 일치한다. 구체적으로 말하여, 한국의 수입침투에 관한 회귀계수(.89)와 한미 무역분쟁 발생건수의 평균(3.30)은 주목할 필요가 있다. 각 수입침투 변수의 무역분쟁 변수에 대한 영향을 고려하기 위해 등식 $\lambda_i b_i$ 가 사용된다. 즉, 이는 한국의 수입침투에 있어 1%의 증가가 한미간의 무역분쟁을 2.94번 더 증가시킨다는 것을 의미한다. 한미 무역갈등은 그 수입침투에 가장 민감하다. 한국의 수입침투의 증가가 한국과 미국간의 무역분쟁을 심화시킨다는 것은 자명한 일이다. 결론적으로 잉여능력 모델이 근본적으로 그 이론의 가설과 일치된 결과를 보여주었다는 점에서 한미 무역분쟁을 이해하는 데 패권안정이론보다 나은 것이다. 패권안정 모델의 결과는 부분적으로 그 이론의 주장과 일치하며, 잉여능력이론과 더불어 미국과 한국 사이의 쌍무적 비대칭적인 무역관계에 어느 정도 적용될 수 있는 것이다.

요약컨대, 다음의 결론을 지을 수 있다. 첫째로, 잉여능력모델이 한미 무역분쟁을 설명함에 적합성을 가진다. 미국에 대한 한국의 수입침투 그리고 미국의 실업률이 가장 중요한 변수이고, 이는 미국의 불경기 속에 한국의 증가된 수입침투가 무역분쟁을 심화시킴을 의미

한다. 이러한 경제적 시각은 1970~1989년 시기 동안 한미 무역관계를 이해함에 적절한 설명을 제공한다.

둘째로, 패권안정 모델은 한미간의 무역마찰에 잘 적용되는 듯 보이나, 그 관계의 방향이 그 이론의 기본 주장과 불일치하는 경향이 있다. 즉, 미국의 군사적 패권의 증가가 오히려 한미 무역분쟁을 증가시키는 경우가 있으며, 동북아의 동서긴장은 예상과 다르게 그 무역분쟁을 증가시키는 경향이 있다. 이는 패권안정의 기본 아이디어와 상치되며, 단지 미국의 경제적 패권만이 그 타당성을 유지한다.

표 1
한미간의 무역분쟁에 대한
패권안정 및 잉여능력
Poisson 모델 분석결과

종속변수		독립변수	Estimate	Standard Error	t-score	LLR
한미 무역 분쟁	패권안정	미·한 GNP비율	-.09	.03	-3.59*	2.52
		미·한 군사비비율	.02	.02	1.02	
		동서협력갈등지수	-.00	.00	-5.00*	
	잉여능력	미국 실업률	.59	.10	6.08*	
		미국 GNP 성장률	.04	.05	.80	
	한국의 수입침투율	.89	.17	5.33*	2.22	

* p < .05, 총계 (N) = 20
출처 : 이상환(1995:58).

표 2
한미간의 무역분쟁에 대한
잉여능력 및 패권안정
NB 모델 분석결과

종속변수		독립변수	Estimate	Standard Error	t-score	LLR
한미 무역 분쟁	패권 안정	미·한 GNP 비율	-.08	.03	-2.80*	2.55
		미·한 군사비 비율	.01	.02	.59	1 +
		동서협력갈등지수	-.00	.00	-4.00*	$\exp(\gamma)$
		γ	-.64			= 1.53
	잉여 능력	미국 실업률	.58	.13	4.53*	2.35
		미국 GNP 성장률	.05	.08	.62	1 +
		한국의수입침투율	.83	.21	3.91*	$\exp(\gamma)$
	γ	.16			= 2.17	

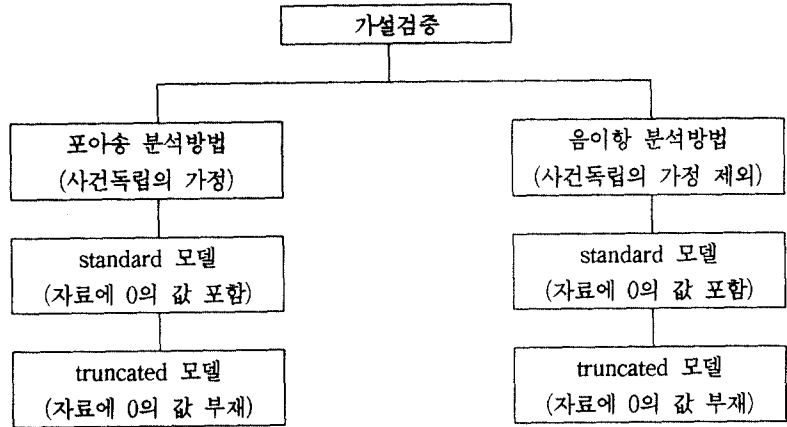
* p < .05, 총계 (N) = 20.
출처 : 이상환(1995: 61).

포아송 결과와 비교하여, 패권안정 모델의 경우 일반적으로 t 값이 음이항 결과에서 내려갔고, \log 우도비(LLR)가 포아송 결과(2.52)에 비해 음이항 결과(2.55)에서 약간 올라갔는데, 이는 음이항 모델이 포아송 모델보다 한미 무역분쟁을 설명함에 약간 더 적합함을 의미한다. 즉, '전파'(contagion)를 고려함이 그 모델의 타당성을 약간 증가시켜 준 것이다. 다른 한편, 분산이 그 평균에 비해 1.53배 큰 데, 이는 단지 약간의 '양전파'(positive contagion)가 한미간의 무역분쟁에 존재함을 의미한다.

패권안정 결과와 함께 잉여능력 결과에서, 일반적으로 t 값이 음이항 결과에서 내려갔고, \log 우도비(likelihood ratio)가 포아송 결과에 비해 음이항 결과에서 약간 올라갔는데, 이는 음이항 모델이 포아송 모델보다 한미 무역분쟁을 설명함에 약간 나음을 의미하며, 패권안정 결과에서와 마찬가지로 '전파'(contagion)를 고려함이 그 모델의 적합성을 약간 증가시켜 준 것이다. 다른 한편, 분산이 그 평균에 비해 2.17배 큰 데, 이 역시 단지 약간의 '전파'(contagion)가 한미간의 무역분쟁에 존재함을 의미한다. 따라서 이러한 무역분쟁의 사건들은 대체로 독립적인 것이며, 이는 포아송 결과가 음이항 결과와 매우 흡사함을 말한다.

요약컨대, Poisson과 NB 결과 사이에는 큰 차이가 없고, 다만 LLR만이 NB 모델에서 약간 커지고 있는데, 이는 그 모델이 포아송 모델보다 개선된 것임을 의미한다. 산포(dispersion) 수준은 약간의 전파(contagion)가 한미 무역분쟁에 존재함을 의미하나, 이것이 포아송 모델의 사건독립의 가정을 부정할 정도는 되지 못한다.⁶⁾

그림 1
가설검증 절차



4) 선형회귀분석기법, 즉 보통최소제곱(Ordinary Least Squares: OLS) 모델의 적용과 통계적 오류

위에 언급한 사건계수 분석결과 우리는 잉여능력모델이 한미 무역분쟁을 설명함에 적합성을 가지며, 미국에 대한 한국의 수입침투와 미국의 실업률이 한미 무역분쟁을 설명하는 가장 중요한 변수라는 점을 파악하였다. 이는 미국의 불경기 속에서 한국의 증가된 수입침투가 무역분쟁을 심화시킨다는 것을 의미한다. 이러한 경제적인 해석은 1970~1989년 동안 한미 무역관계를 이해함에 적절한 설명을 제공한다고 할 수 있다. 한편, 패권안정 모델은 한미간의 무역마찰에 잘 적용되는 듯 보이나, 그 관계의 방향이 이론의 기본 주장과 불일치 하는 경향이 있음을 알았다. 즉, 미국의 군사적 패권의 증가가 오히려 한미 무역분쟁을 증가시키는 경우가 있으며, 동북아의 동서긴장은 예상과는 다르게 그 무역분쟁을 증가시키는 경향이 있다. 이는 패권안정의 기본 아이디어와 상치되며, 단지 미국의 경제적 패권만이 그 타당성을 유지함을 보여주고 있다.

표 3
한미간의 무역분쟁에 대한
패권안정 및 잉여능력
OLS 모델 분석결과

종속변수		독립변수	Estimate	Sig. T	Adjusted R-Square
한미 무역 분쟁	패권안정	미·한 GNP 비율	-.16	.08	.38
		미·한 군사비비율	.052	.14	
		동서협력갈등지수	-.01	.06	
	잉여능력	미국 실업률	1.52	.04*	.42
		미국 GNP 성장률	.10	.74	
		한국의 수입침투율	1.89	.06	

*p < .05, 총계 (N) = 20

이 논문은 앞서 종속변수가 사건집계 자료인 경우 선형회귀분석기법의 적용은 심각한 통계적 오류를 낳을 수 있다고 언급하였다. 실제로 어떠한 통계적 결과가 나오는지 알아보기 위하여 같은 자료를 이용하여 선형회귀분석을 행하였다. 예상한 대로 OLS (Ordinary Least Squares, 보통최소제곱) 결과는 사건계수 분석결과와 상이하였다. <표 3>이 보여주는 바와 같이 미국 실업률을 제외하고는 한미 무역분쟁을 설명해 주는 변수가 없으며, 패권안정 모델은 물론 잉여능력 모델마저 한미 통상관계를 설명해주지 못한다는 결론에 이르게 한다. 이는 심각한 통계적 문제점이 초래한 결과이다. 한 마디로 말하여, OLS가 전제하는 정규분포의 가정이 그 결과를 왜곡시킨 것이라고 할 수 있다. 물론 사건계수 자료의 분포가 정규분포와 흡사하다면 OLS 분석결과와 사건계수 분석결과, 즉 포아송 및 음이항 분석결과는 크게 다르지 않을 것이다. 그러나 위에서 다룬 예에서처럼 데이터 포인트가 20개에 불과하고 그마저 비대칭으로 산포되어 있다면, 그 분석결과는 커다란 차이점을 보이게 된다. 더욱이 OLS의 적용은 앞서 언급한 문제점과 함께 시계열 자료가 갖고 있는 자동상관관계 문제와 25개 미만의 데이터 포인트를 선형회귀분석하는 한계점 등 여러가지 부수적인 통계적 한계를 드러내게 된다. 따라서 사건계수 분석기법의 적용이 요망된다고 할 수 있는 것이다.

4. 결론 : 정확한 이론 검증을 위한 사건계수 방법의 활용

이 논문에서 소개된 사건계수 분석방법들은 대부분의 경험적 사회과학연구에서 간과되어 왔던 사건집계 자료에 대한 적절한 통계적 분석을 가능하게 하는 해결책의 역할을 할 수 있다. 특히 자료가 숫자로만 되어 있으면 무조건 선형회귀분석을 하는 우를 범하지 않도록 경각심을 주는 것이다. 즉, 자료의 부적절한 사용으로 인하여 연구의 결과가 왜곡되게 나오는 경우를 막고자 하는 것이다.

이 논문에서 소개한 연구방법인 Event-Count 분석기법은 이산적인 종속변수를 위한 분석기법이며, 그 대표적인 기법의 하나는 포아송 분석방법이고, 다른 하나는 음이항 분석방법이다. 이러한 분석기법의 적용 이유는 특정한 연구가 다루는 종속변수가 사건 발생횟수를 나타내고, 그 값이 음수가 아니며, 따라서 그 데이터가 정규 분포되어 있지 않고, 더욱이 각 사건간에 독립성을 가정(Poisson)하거나 혹은 그 연관성을 고려(NB)해야 하기 때문이다. 이와 같이 변수와 자료(data)의 성격을 적절히 고려한 통계방법의 선택과 이것의 경험적 사회과학 연구에의 적절한 적용은 정확한 이론 검증을 위해 하나의 지침이 되는 것이다.

■ 주

- 1) 선형회귀분석의 5가지 전제조건들(Kennedy, 1985:40~50) 중의 하나는 종속변수와 독립변수들과의 관계가 선형적(linear)이어야 한다는 것인데, 종속변수가 불연속변수인 경우 이를 위반하게 되며 따라서 선형회귀분석의 적용은 통계적으로 잘못된 결과를 얻게 된다.
- 2) 자료의 형태가 과잉산포가 아니며 사건계수의 상한선이 낮은 경우에 적용가능한 모델이며 일반화된 사건 계수 모델의 변형에 해당된다.
- 3) 자료의 형태가 과잉산포이며 사건계수의 상한선이 낮은 경우에 적용가능한 모델이며 일반화된 사건 계수 모델의 변형에 해당되지 않는다.
- 4) 미국의 한국에 대한 상대적인 경제적·군사적 영향력은 미국의 GNP 및 군비지출의 한국에 대한 상대적 비율로 측정되며, 동서긴장관계는 WEIS 국제협력력동지수로 측정된다. 이들 자료는 World

Tables, *SIPRI Yearbook*, 그리고 *Three-Way Street* (by Goldstein & Freeman) 등에 근거한다.

- 5) 사건계수 (*Event-Count*) 분석을 위한 컴퓨터 프로그램은 GAUSS 3.1버전 (Aptech Systems, Inc., 1993)이 있다. 사회과학을 위한 통계처리 프로그램의 상당수가 최대우도추정 (MLE) 분석을 위한 부분을 가지고 있다. 그러나 사용되는 자료의 크기와 얻고자 하는 통계적 결과에 따라 그 프로그램의 선택을 달리한다. 이 논문에서는 포아송 (*Poisson*)/음이항 (*Negative Binomial*) 모델을 위해 GAUSS 프로그램이 사용되었다. GAUSS 프로그램은 자료의 크기에 관계없이 다양한 통계적 결과를 상세하게 제공한다는 점에서 사건계수 분석을 위한 통계처리에 가장 적합한 프로그램이라고 할 수 있다.
- 6) 각 사건간의 전파 (*contagion*)의 정도를 나타내는 절대값 γ 가 0에 가깝다.

■ 참고 문헌

- 김용진 편. 1994. 《방법, 방법론과 한국정치학》. 서울: 도서출판 한울.
- 이상환. 1995. "미국과 동북아 3국간의 무역분쟁: 패권안정이론과 잉여능력이론의 고찰." 《국제정치논총》 35(1): 27~62.
- 주미영 · 이상환. 1995. "Probit/Ordered Probit과 Event-Count 분석을 위한 방법: 이해와 적용." 《국제정치논총》 35(2): 101~120.
- 한국통계학회 편. 1994. 《통계용어사전》. 서울: 자유아카데미.
- Achen, Christopher H. 1996. "Implicit Substantive Assumptions Underlying the Generalized Event Count Estimator." In John R. Freeman, ed. *Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association* 6: 155~173.
- Amato, Timothy W. 1996. "On Difference Equations, Probability Models and the 'Generalized Event Count' Distribution." In John R. Freeman, ed. *Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association* 6: 175~212.
- Allison, Paul D. 1984. *Event History Analysis*. Beverly Hills: Sage Publication.
- Eliason, Scott R. 1993. *Maximum Likelihood Estimation: Logic and Practice*. A Sage University Paper.
- Gourieroux, C., A. Monfort, and A. Trognon. 1984. "Pseudo Maximum Likelihood Methods: Applications to Poisson Models." *Econometrica* 52: 701~720.
- Hausman, Jerry, Bronwyn H. Hall, and Zvi Griliches. 1984. "Econometric Models for Count Data with an Application to the Patents-R & D Relationship." *Econometrica* 52: 909~938.
- Kennedy, Peter. 1985. *A Guide to Econometrics*. Boston: The MIT Press.
- King, Gary. 1989. "Event Count Models for International Relations: Generalizations and Applications." *International Studies Quarterly* 33: 123~147.

- King, Gary. 1989. *Unifying Political Methodology: The likelihood Theory of Statistical Inference*. Cambridge: Cambridge University Press.
- King, Gary and Curtis S. Signorino. 1996. "The Generalization in the Generalized Event Count Model, with Comments on Achen, Amato, and Londregan." In John R. Freeman, ed. *Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association* 6: 225~252.
- Londregan, John. 1996. "Some Remarks on the 'Generalized Event Count' Distribution." In John R. Freeman, ed. *Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association* 6: 213~223.
- Maddala, G. S. 1983. *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.
- Martin, L. Lisa. 1992. *Coercive Cooperation*. Princeton: Princeton University Press.
- Lee, Lung-Fei. 1986. "Specification Test for Poisson Regression Models." *International Economic Review* 27: 689~705.
- Lovett, Andrew and Robin Flowerdew. 1989. "Analysis of Count Data Using Poisson Regression." *Professional Geographer* 41: 190~198.
- Winkelmann, Rainer, Curtis S. Signorion, and Gary King. 1996. "A Correction for an Underdispersed Event Count Probability Distribution." In John R. Freeman, ed. *Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association* 5: 215~228.
- Yamaguchi, Kazuo. 1991. *Event History Analysis*. Beverly Hills: Sage Publication.

Journal of Governmental Studies

Vol.6, No.1

(June 2000)

Abstracts

1. Event-Count Methods in Social Science Research : Comprehension and Application

Sang-Hwan Lee

What can be the most appropriate research method for the study analyzing event-counting dependent variables? Linear regression analysis has made a mistake in analyzing count dependent variables, since its assumptions are not valid in the variables with discrete events and non-negative integers. The analysis produces large inefficiencies, nonsensical results, and other problems. Event-count methods such as Poisson and Negative Binomial (NB) analyses are more adequately applicable to the special nature of event count data. The applicability of Poisson regression analysis depends on the assumption that the event counts are the dependent variable with a Poisson distribution and are independent of one another. A popular alternative method, when the events are not independent of one another, is the Negative Binomial analysis. Additionally, to correct for selection bias present in the data set, we need to use the standard models for the data with zero cases and the truncated-at-zero models for the data with positive cases only.