2005 Proceedings of the Autumn Conference, Korea Statistical Society

프랜시스 골턴 다시 읽기

조 재 근[1)]

요약

오늘날 대부분의 통계학 교과서에서 프랜시스 골턴(Francis Galton 1822-1911)은 비록 "회귀(regression)"라는 용어를 처음 사용한 인물이기는 하되, 그가 생각했던 "평균으로의 회귀(regression to the mean)"라는 것은 오늘날의 회귀분석과는 거리가 먼 것이라는 언급과 함께 짧게 소개된다.

이 글에서는 바로 그 골턴이 직접 쓴 것들을 다시 읽어보고 골턴 자신과 후세에 소개되는 골턴 사이의 거리를 살펴보려 한다. 그 결과 골턴은 통계학의 역사에서 여러 가지로 흥미로운 인물이므로 그의 이름이 통계학 교육 내용 중에 지금보다는 조금 더 등장해도 좋을 것으로 보인다.

주요용어: 골턴, 회귀, 상관.

1. 서 론

통계학의 역사에서 골턴은 20세기 통계학의 가장 직접적인 모태라 일컬어지는 19세기말 이후 영국의 통계학을 거론할 때면 에지워스, 피어슨, 율, 피셔 등에 앞서 항상 가장 먼저 등장하는 인물이다. 하지만 통계학 교과서와 통계학 강의실에서 골턴의 이름을 보고 듣기란 대단히 어렵다.

그의 이름은 겨우 회귀분석 교과서의 한 모퉁이 정도에서 찾아 볼 수 있을 정도에 그치는데 거기서 그는 "오늘날의 회귀분석과는 다른 뜻으로 그 용어를 썼는데도 불구하고 회귀라는 이름을 처음 쓴 사람으로 역사에 남은 인물"로 옹색하게 소개되고 있다. 그리고 골턴이 회귀라는 용어를 처음 쓰게 되는 부모 키와 자식 키 사이의 관계를 직선으로 나타내는 문제에서 우리가 쉽게 생각할 수 있는 해법, 즉 부모의 키를 독립변수로 자식의 키를 종속변수로 삼아 단순회귀직선을 구하는 방법을 그가 알아내기까지는 무려 이십년이 넘게 걸렸으며, 그마저 최소제곱법과 같은 좋은 방법으로 구한 것도 아닌데다가, 골턴이 최소제곱법과 그의 '회귀'를 연결시킬 생각조차 하지 않았다는 사실을 알고 나면 통계학 책에 나오는 그의 이름은 더욱 궁색해 보인다.

그런데 통계학의 역사에 등장하는 인물 가운데 통계학 바깥에서는 골턴만큼 죽은 지 백년이 가까운 지금까지 과학사 연구자들을 비롯하여 유전학, 진화생물학 등의 역사를 연구하는 학자들에 의해 끊임없이 주목 받고 재평가되고 있는 인물도 달리 없음을 확인하게 된다. 그렇다면 과연 통계학의 역사에서 골턴은 어떻게 평가되어야 하는가? 이 글에서 우리는 회귀와 상관을 비롯한 통계학의 역사에 중심을 두면서도 시야를 통계학 안팎으로 조금 넓혀서 골턴이라는 인물을 살펴볼 것이다. 그 결과로 이 흥미로운 인물의 이름을 통계학 강의실에서 조금 더 들을 수 있게 되기를 기대한다.

2. 골턴 다시 읽기

골턴은 매우 다양한 분야에서 활동한 만큼 각 분야에서 적지 않은 논문 혹은 책을 발표했는데 웹사이트 www.galton.org에 가면 그가 쓴 것들 거의 모두를 다 볼 수 있다. 그가 쓴 것들 가운데 통계학에 관련된 것만 해도 수십 편에 이르는데 그 가운데 우리가 주로 살펴볼 회귀와 상관이라는 주제와 관련이 있는 것은 Galton (1865, 1877, 1886, 1888)등의 논문과 Galton (1869, 1889)와 같은 책들이다. 당시 골턴의 연구를 조

1) 경성대학교 정보통계학과, 교수, 부산시 남구 대연동 608-736. Email: jkjo@star.ks.ac.kr.

금 더 세부적으로 살펴보고 오늘날의 관점에서 몇 가지 짚어보자.

(1) 먼저 그가 생각했던 회귀는 오늘날처럼 변수들 사이의 일반적인 함수관계를 찾는 것이 목적이 아니었다. 골턴은 유전 법칙의 하나로 부모의 특징(가령 키)와 자식의 특징 사이를 설명할 수 있는 선형함수관계를 찾으려 하였고 그의 목적은 그 선형함수의 기울기, 즉 일종의 비례상수에 해당하는 단 하나의 상수였다. 골턴이 '회귀 비(ratio of regression)이라고 부르고 기호로는 'r'이라고 나타낸 이 상수는 그의 '회귀 법칙(복귀 법칙)'에 따라 1보다 클 수 없으므로 반드시 0과 1 사이의 상수여야만 했다.

돌이켜볼 때, 골턴의 회귀는 절편과 기울기를 가진 회귀선을 구하여 독립변수의 변화에 따른 종속변수의 변화 관계를 알아보려는 것과는 거리가 멀었다. 그가 회귀라는 이름으로 구하려한 것을 오늘날의 표현으로 나타내면 두 변수 사이의 상관계수를 구하려한 것이라고 볼 수 있고 엄밀히 말하자면 그것도 양의 상관계수에 국한된 것이었다. 오늘날 통계학의 역사에서는 골턴의 업적 가운데 상관계수에 대한 연구가 가장 중요한 것으로 평가되는데 사실상 두 변수 간의 관계를 재는 하나의 숫자라는 의미에서의 상관계수라는 개념은 골턴 자신의 연구에서는 극히 작은 역할밖에 하지 못했다(스티글러, 2005, p. 551).

(2) 대부분의 통계학 교과서에서 골턴의 '평균으로의 회귀'는 통계학적으로 이변량 정규분포에서 조건부 기댓값의 선형관계로 설명된다. 즉 $(X, Y) \sim N_2(\mu_1, \mu_2, 1, 1, r)$라고 할 때

$$\mathrm{E}(Y \mid X = x) = rx, \quad \mathrm{E}(X \mid Y = y) = ry$$

라는 관계를 골턴이 밝혔다는 것이다.

하지만 이러한 설명은 엄밀히 말해서 골턴이 한 것과 다른 것이다. 골턴은 그의 논의 과정에서 시종일관 기댓값이 아니고 그가 M이라는 기호로 나타낸 중앙값을 썼다. 그리고 표준편차가 아니고 Q라는 기호로 나타낸 probable error(사분위수범위(inter-quartile range)의 반을 뜻하는데 이에 대해서는 스티글러, 2005, pp. 206-207의 옮긴이 주를 참조할 것)를 썼다. 두 표현 사이의 관계는 정규분포에서 평균과 중앙값이 같고 표준편차와 probable error의 관계는 $1Q = 0.6745\sigma$의 관계가 있으므로 계산과정이나 결과는 평균과 표준편차를 썼을 때와 마찬가지이다. 그러나 후세 사람들이 중앙값과 분위수로 표현된 골턴의 회귀를 평균과 표준편차를 이용한 것으로 바꾸어버림으로 인해 골턴의 표현이 갖는 개념상의 직관이나 장점을 가려버리게 되었다는 지적 또한 제기되고 있다(Gilchrist, 2005). 예컨대 골턴의 모형은 median regression model과 같은 모형까지 다 포괄하는 훌륭한 모형이었음에도 불구하고 왜곡되어 전달되어온 결과 오늘날까지도 제대로 이해되지 않고 있다는 것이다.

(3) 그의 놀라운 직관과 달리 세부적인 계산법은 매우 거칠고 조악한 것이었다. 오늘날의 회귀분석 방법과 비교할 때 가장 다른 점은 골턴이 그의 '회귀 법칙'을 유도해내는 과정에서 상수 r을 얻기 위한 추정방법이 없었다는 점이다. 그는 최소제곱법은 물론 어떠한 '객관적인' 추정방법도 사용하지 않았는데, 사실 그가 이용한 방법은 그림을 그린 다음 대충 눈으로 보고 편리한 수를 얻는 완전한 '주먹구구' 방법이었다. Galton(1889, pp. 95-99)을 보면 부모의 평균키와 자식 키 사이의 회귀계수를 얻기 위해 그는 데이터를 점으로 나타낸 그림(Galton, 1889, Fig. 10, p. 96)을 그린 다음, 그 점들을 잘 지나는 듯 보이는 '적당한' 직선을 하나 그어 그 직선의 기울기를 골랐다고 한다. "처음에는 3/5으로 할까 하다가 2/3가 좀 더 간단해서"(Galton, 1889, p. 98) 그 값을 택했다고 한다. 그러므로 통계학의 역사에서 골턴의 '평균으로의 회귀'는 그 계산과정만을 본다면 1805년 르장드르가 최소제곱법을 만들어 낸 이후 이미 천문학을 비롯한 여러 분야에서 널리 활용되고 있던 선형모형의 계수를 추정하는 방법과는 매우 동떨어진 것이었다.

(4) 그렇다면 왜 골턴은 최소제곱법을 비롯한 훌륭한 추정방법이 있는데도 주먹구구식으로 그의 회귀계수를 추정했을까? 골턴 혹은 그의 주변 사람들이 최소제곱법을 몰랐을 리가 없으므로 결국 골턴이 자신의 문제에는 최소제곱법을 비롯한 기존의 추정방법을 적용할 수 없다고 믿었던 것이 분명하다. 그랬다면 그 이유, 즉 기존의 방법들이 적용되었던 문제와 골턴의 문제는 본질적으로 어디서 달랐을까? 이 질문은 알고 보면 19세기 초에 르장드르가 최소제곱법을 발표한 이후 19세기 통계학에서 오랫동안 이어진 과제 중 하나였다. 거칠게 표현하자면 르장드르와 가우스 그리고 라플라스 등의 연구에서는 우리가 오늘날 회귀모형

이라 부르는 모형에서 종속변수는 랜덤하다고 볼 수 있었지만 독립변수는 그 수가 몇 개이든 오차 없는 정확한 값들이었다. 좀 더 정확하게 표현해서 르장드르의 최소제곱법에는 사실 독립, 종속변수라는 것이 없었고 반복 관측 횟수에 해당하는 개수만큼의 서로 일치하지 않는, 미지수를 여럿 가진, 방정식들이 있을 뿐이었다. 여기서 19세기 초의 수학자들이 랜덤하게 둔 것은 종속변수가 아니라 각 방정식의 '오차'들이었다(스티글러, 2005, p. 97). 따라서 그들의 방정식에 필요한 미지수를 추정한 상수 값들을 가지고 모형을 얻고 나면 독립변수의 특정값에 대해서는 종속 변수의 특정값이 유일하게 정해졌던 것이다. 즉 결과는 물리학적인 법칙과 마찬가지였던 것이다.

하지만 골턴의 자료는 전혀 그렇지 않았다. 골턴이 연구한 부모 키와 자녀 키와 같은 경우에는 양쪽 모두 랜덤한 변수들이었고, 사람마다 서로 다른 키들과 그들의 평균키와의 차이를 도저히 '오차(error)'라고 여길 수는 없었을 것이다. 뿐만 아니라 모형을 추정하고 나서도 어떤 독립변수 값에 대해 종속변수 값이 법칙처럼 단일한 값으로 정해질 수 없었을 것이다. 결국 수학적으로 말하자면 골턴의 회귀는 이전까지 다뤄지지 않았던 문제, 독립변수까지 모두 랜덤한 다변량(실제로는 이변량) 모형을 처음으로 연구한 셈이었다.

(5) 골턴의 회귀와 상관은 단순히 이변량(골턴 자신은 다루지 않았지만 나아가 다변량)분포를 따르는 자료를 분석할 수 있게 되었다는 수학적인 측면에서의 중요성만 갖는 것이 아니라 그 결과 통계학의 적용분야가 크게 넓어질 수 있었기 때문에 더욱 중요하다.

이는 골턴의 회귀와 상관을 케틀레의 "평균적인 사람(average man)"과 비교해볼 때 잘 드러난다. 케틀레는 그의 앞 시대부터 천문학 분야에서 쓰이던 통계적 방법을 사회에서 수집된 자료에 적용시키려 하였다(스티글러, 2005, pp. 335-430). 하지만 그는 통계적 방법뿐만 아니라 천문학과 물리학적인 개념 즉 '하나의 참값과 그 주변의 오차'라는 사고 역시 사회 데이터에 적용하려 하였다. 즉 그의 입장은 통계자료에서 극단적인 것들은 그 사회를 대표하거나 표현하는데 적절하지 않은 것들이기 때문에 극단으로 치우치지 않고 집단을 대표할 수 있는 값, 즉 평균을 이상적으로 생각하는 것이었다. 나아가 그는 국가나 사회에서 개별적인 개인이란 모두 결점을 지닌 존재에 지나지 않으므로 가장 보편적인 "평균적인 사람"(average man)을 그 집단의 전형으로 내세웠다. 따라서 케틀레에게 있어서 개인이란 평균적인 사람이라는 이상적인 중심에서 벗어난 '오차'와 같은 존재였던 셈이다. 골턴 역시 케틀레처럼 정규분포를 대단히 중요하게 여겼다는 면에서 케틀레의 후계자라고 할 수 있다. 하지만 골턴의 경우 천문학이나 물리학적인 사고틀을 벗어나 하나의 중심보다는 중심으로부터 벗어난 사람들에 대한 관심에서 출발하였을 뿐 아니라 여러 세대에 걸친 유전의 법칙을 찾겠다는 목적 때문에 자연스럽게 두 변수가 각각 갖는 여러 값들 사이의 관계에 초점을 맞추게 되었던 것이다. 따라서 골턴의 연구는 기존의 통계적 방법들이 주로 적용되던 천문학, 측지학 등의 좁은 분야를 넘어 생물학과 사회과학 등의 분야로 통계학의 적용분야를 넓히는 획기적인 것이 되었다. 이러한 이유로 골턴은 사회과학의 역사 분야에서 지속적인 연구대상이 되고 있다.

(6) 사소한 문제도 하나 지적하자. 적지 않은 회귀분석 교과서들이 골턴이 회귀라는 용어를 만들어낸 과정을 설명하면서 '아버지의 키와 아들의 키' 사이의 관계를 연구하는 과정에서 이 용어가 처음 등장한다고 설명하고 있다. 뿐만 아니라 외국에서 나온 상당수의 교과서들에서도 이러한 설명을 종종 볼 수 있다. 예컨대 통계학 교과서의 고전 가운데 하나인 Yule and Kendall (1950) 역시

Galton found that the sons of fathers who deviate x inches from the mean height of all fathers themselves deviate from the mean height of all sons by less than x inches, I.e. there is what Galton called a "regression to mediocrity." (p. 213).

라고 설명하고 있는데 엄밀히 말해서 그러한 설명은 틀린 것이다. 앞서 언급했듯 Galton(1889)를 비롯하여 회귀라는 용어가 나오는 연구는 모두 유전학을 위한 연구였는데 골턴이 키와 같은 현상을 다룰 때에는 부모 가운데 한 쪽만 고려하지 않고 유전이라는 과정에 기여하는 정도는 부모 양쪽이 공평하게 똑같다고 생각했었다. 그러므로 골턴 자신이 상세히 밝혔듯(Galton, 1889, pp. 5-7) 그가 다룬 변수는 '아버지의 키'가 아니고 부모의 평균키, 더 정확하게 말하면 어머니의 키에는 1.08이라는 가중치가 부여된 부모 키의 가중평균(그는 이를 "midparent"라고 불렀다)이었다. 그는 자식의 키를 다룰 때에도 역시 성인이 된 딸의 키에는

같은 가중치를 부여했다.

(7) 골턴의 '평균으로의 회귀'라는 개념이 '평균으로 돌아간다'는 말 그대로의 의미로 중요하게 간주되는 경우도 볼 수 있는데 그 대표적인 보기를 통계학의 역사를 쓴 것이라 해도 손색이 없는 책 Bernstein (1996, 번역서는 1997)에서 찾아볼 수 있다. 그는 위험관리(risk management)라는 측면을 강조하면서 특히 골턴에 대해 많은 지면을 할애하였는데 평균으로의 회귀가 실제로 작용하고 있다는 사례로 미국 주식시장 주식 수익률 분석 결과, 뮤추얼 펀드의 수익률 분석 결과, 장기 주식 투자 수익률 분석, 장기간에 걸친 국가별 생산성 비교 분석 등을 들고 있다(번역서 239-288쪽).

3. 결론

골턴은 1911년에 죽었는데 그 해에 University College London에 세계 최초의 통계학과라 할 수 있는 Department of Applied Statistics가 만들어진다. 그 학과가 생긴 것은 다름 아닌 골턴이 남긴 유산 덕분이었으며 최초의 교수는 칼 피어슨이었다. 따라서 통계학에서 골턴의 업적은 비단 회귀와 상관에만 국한되는 것이 아니라 통계학이 대학에서 제도화하는 데에까지 이르렀던 셈이다.

한편 19세기말부터 20세기 초에 이르는 시기는 생물학 가운데 유전학의 역사에서도 중요한 시기였다. 당시에 칼 피어슨과 웰던으로 대표되는 생물측정학파(biometricians)와 베이트슨(W. Bateson)으로 대표되는 멘델학파(Mendellians) 사이의 논쟁이 수십 년간 진행된 탓이다. 생물측정학파가 20세기 초 통계학의 역사에서 매우 중요한 역할을 한 만큼 그들과 멘델학파 사이의 논쟁을 통계학의 역사 속에서 재조명해보는 연구도 향후의 과제 가운데 하나일 것이다. 그런데 골턴은 그 논쟁에서 주역은 아니었지만 양쪽에 모두 소속될 수도 있는 독특한 위치에 있었다. 그 이유는 그가 생물측정학파의 창시자인 한편, 멘델학파의 주장 중 하나인 유전의 불연속성에 동조하기도 하였기 때문이다.

이와 같이 골턴이라는 인물은 지금 우리가 회귀라는 용어를 처음 썼던 인물로만 기억하기 보다는 통계학 강의실과 교과서에서 보다 더 많이 등장해도 좋을 인물로 보인다.

참고문헌

Bernstein, P. (1996). *Against the Gods: The Remarkable Story of Risk*, John Wiley & Sons, New York, 「리스크- 리스크 관리의 놀라운 이야기」, 안진환, 김성우 옮김, 한국경제신문사, 1997.

Galton, F. (1865). Hereditary talent and character, *Macmillan's Magazine*, pp. 157-166, 318-327.

Galton, F. (1877). Typical laws of heredity, *Proceedings of the Royal Institution*, viii, pp. 282-301.

Galton, F. (1886). Regression towards mediocrity in hereditary stature, *Journal of the Anthropological Institute,* Vol. 15, pp. 246-263.

Galton, F. (1888). Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society,* Vol. 45, pp. 135-45.

Galton, F. (1889). *Natural Inheritance*, Macmillan, London, (Facsimile edition: Genetics Heritage Press, Placitas, New Mexico, 1997).

Gilchrist, W.G. (2005). Galton Misrepresented, *Significance*, Vol. 2, pp. 136-137.

Stigler, S. M. (1986). *The History of Statistics*, Belknap Press of Harvard University Press, Cambridge, Massachusetts, 「통계학의 역사」 조재근 옮김, 한길사, 2005.

Stigler, S. M. (1999). *Statistics on the Table: The History of Statistical Concepts and Methods*, Harvard University Press, Cambridge, Massachusetts.

Yule, G. U. and Kendall, M. G. (1950). *An Introduction to the Theory of Statistics*, Hafner, New York.