The Korean Communications in Statistics
Vol. 13 No. 3, 2006, pp.479-490

Francis Galton in the History of Statistics[1)]

Jae Keun Jo[2)]

Abstract

Francis Galton (1822-1911) introduced the term "regression" and "correlation" in the study on human inheritance of the stature from parents to their children. In almost every statistics textbook, superficial attentions have been given to him just as the inventor of the term "regression".

Rereading his books and papers, we investigated problems he had tried to solve and the methods he had used to solve the problems. In addition, we tried to find the motivation that had led Galton to take attention to the variation rather than the central tendency of observational data that had fascinated his forerunner Adloph Quetelet.

Keywords : Regression; correlation; history of statistics.

1. 서론

19세기 후반부터 20세기 초까지 활동했던 프랜시스 골턴(1822-1911)이라는 인물은 오늘날 통계학 교과서에서 회귀(regression)라는 용어를 처음 쓰고 상관(correlation)이라는 개념을 또한 처음 만든 사람으로 등장한다. 골턴이 활동했던 시기는 통계학이 학문으로서 겨우 독립된 지위를 가지려하던 때인데, 그는 그러한 변화의 중심인물 가운데 한 사람으로 평가되기도 하지만 한편으로는 오늘날의 회귀분석과는 다른 뜻으로 그 용어를 썼는데도 불구하고 단지 그 용어를 처음 썼다는 이유만으로 역사에 이름이 남은 사람으로 소개되기도 한다. 뿐만 아니라 '피어슨의 상관계수'라는 이름에서도 알 수 있듯이 상관계수의 역사에서도 골턴의 이름은 칼 피어슨의 이름만큼 널리 회자되지는 못하고 있다. 이처럼 교과서 등에서 골턴이 소홀히 취급되는 이유는 후대 통계학의 기준으로 보았을 때 높이 평가받을 만한 업적을 그가 남기지 못했기 때문일 것이다.

그러한 후대의 평가에도 불구하고 골턴은 통계학이 약 일 세기 전 영국에서 혁신적으로 변모하기 시작했을 때 통계학이 했던 역할, 통계학과 다른 학문들과의 관계, 그리고 신생학문으로서 통계학이 가졌던 한계 등을 살펴보려면 가장 먼저 만나야하는 인물이다. 그러므로 오늘날의 통계학을 기준으로 해서 판단하기보다는 19세기 말이라

1) The Research was supported by Kyungsung University Research Year in 2004.
2) Professor, Department of Informational Statistics, Kyungsung University, Busan, 608-736. Korea.
E-mail : jkjo@ks.ac.kr

는 시대적 배경 속에서 골턴을 살펴보는 것은 통계학의 역사 연구에 있어서 의미 있는 작업일 것이다.

골턴의 연구와 골턴 이전 다른 사람들의 연구를 비교했을 때 가장 두드러진 차이는 무엇보다 골턴이 자신보다 바로 앞 세대와는 매우 다른 방식으로 자료를 보는 길을 열었다는 점이다. 골턴이 등장하기 직전까지, 즉 19세기 중반 무렵 유럽의 통계학을 대표했던 인물은 벨기에의 케틀레(Adolph Quetelet)였다. 케틀레는 통계학의 역사에서 데이터의 중심을 가장 강조한 인물로 기록되는데 그는 자연 현상에서뿐 아니라 사회 현상에서도 데이터의 중심이 전체 데이터의 전형적인 대푯값이자 하나의 고정된 모범이라고 생각하였다. 한 걸음 더 나아가 그는 중심에서 멀리 떨어진 데이터들은 오차와 같은 존재라고까지 간주하였다 (스티글러, 2005, 제5장).

그와 반면에 골턴은 중심보다는 중심으로부터 먼 자료들, 자료의 산포, 그리고 변수들 사이의 관계에 주목하였다. 회귀와 상관에 대한 연구가 바로 그러한 결과인데 이러한 연구에서 골턴이 분석한 자료는 부모 키와 자녀 키 사이의 유전 관계를 알아보기 위한 데이터였다. 그런데 뜻밖에도 그는 회귀직선의 기울기(또는 두 변수 사이의 상관계수)를 구하기 위해 19세기 초 이후부터 잘 알려져 있던 최소제곱법이라는 훌륭한 방법을 사용하지 않았을 뿐 아니라 아예 회귀와 상관에 대한 글에서 그 방법을 언급조차 하지 않았다. 이 연구에서는 먼저 그 이유를 추론해 보았는데 여기서 간략히 요약하면 다음과 같다: 골턴은 자신이 풀어야할 문제가 당시까지 최소제곱법으로 풀던 문제와는 다른 문제라고 생각했을 것이며 그렇다면 최소제곱법은 19세기말까지만 하더라도 데이터 분석을 위해 오늘날보다는 보다 더 제한적으로 사용되었던 것으로 볼 수 있다.

다음으로 이 연구에서는 오늘날의 교과서에서 골턴의 연구를 언급할 때 종종 발견되는 오류에 대해서도 알아보았으며, 골턴이 케틀레와 달랐던 점에 대해서도 살펴보았다. 그나마 골턴의 이름은 회귀분석 교과서에 남았지만 1830년대 이후 적지 않은 기간 동안 유럽 통계학계를 지배했던 케틀레의 이름은 통계학 교과서에서 찾아볼 수 없게 되었다. 그 두 사람의 차이에 대해서는 여러 학자들이 서로 다른 방향에서 연구한 바 있는데(스티글러, 2005, 제5장, 제8장; Hacking, 1990), 본 연구에서는 두 사람의 통계학적 입장 차이를 각자가 그 시대적 조건 속에서 관심을 가졌던 분야의 차이로부터 찾아보았다. 그 결과 아직 통계학을 학문 그 자체로 연구하는 사람이 거의 없다시피 했던 19세기에 통계학이 가졌던 독특한 면모를 살필 수도 있었다.

이 연구에서 인용하거나 언급한 골턴의 연구는 발표 당시의 원문 그대로 http://galton.org 사이트에서 모두 볼 수 있다.

2. 본론

영국의 빅토리아 시대를 망라하는 긴 생애 동안 골턴은 아프리카 탐험, 기상학, 지문 연구, 인류학, 유전학, 우생학, 그리고 통계학 등 대단히 많은 분야에서 다양한 활동을 펼쳤던 인물이다. 하지만 후세의 평가에서도 드러나듯 그는 어느 분야에서 특정 주제에 대해 깊이 있게 연구하기보다는 다양한 문제에 대해 독특한 아이디어를 제시

하기만 했던 경우가 많았다(Gillham, 2001; Bulmer, 2003; Brookes, 2004). 회귀와 상관이라는 주제에 대해서도 골턴은 아이디어만 제공하였을 뿐이었으므로 이론적인 연구는 결국 후세 사람들의 몫일 수밖에 없었다. 따라서 통계학사에서 그의 역할은 인색한 평가를 받을 수밖에 없었지만 한편으로 20세기 초 피어슨, 율, 피셔를 비롯한 영국 통계학자들에 의해 통계학이 혁신적으로 발달하는 데에는 골턴의 아이디어가 중요한 바탕이 되었다고 인정하는 연구들도 적지 않다(스티글러, 2005; Stigler, 1999; Keynes, 1993). 골턴이 쓴 것들 가운데 우리가 중점적으로 살펴볼 회귀, 상관과 관련이 있는 것은 Galton (1865, 1877, 1886a, 1886b, 1888, 1890) 등의 논문과 Galton (1869, 1889)과 같은 책들이다.

이 연구에서는 먼저 골턴의 회귀와 오늘날의 회귀는 서로 목적이 다르며, 따라서 문제를 해결하기 위한 방법도 다르다는 것, 나아가 골턴이 자신의 문제를 풀기위해 당시 널리 쓰이던 방법, 즉 최소제곱법을 쓰지 않았던(혹은 못했던) 이유가 무엇인가를 알아보았다. 다음으로 오늘날 교과서 등에서 골턴이 어떻게 소개되는지 살펴보고 마지막으로 케틀레와 차별성을 갖는 골턴의 연구가 나오게 된 시대적 배경의 하나로 그들이 통계학적 방법을 적용하려 했던 분야들, 즉 사회학과 유전학이 당시에 어떠한 단계에 있었는지 살펴보았다.

2.1 골턴과 최소제곱법

Farebrother (1999)에서도 볼 수 있듯 사실 변수들 사이의 함수관계를 알아보려하는 회귀 혹은 선형, 비선형모형의 역사를 따진다면 골턴의 연구보다 백여 년이나 앞선 연구들, 즉 적어도 18세기 중엽의 연구에까지도 거슬러 올라갈 수 있다. 회귀분석에서 중요하게 이용되는 최소제곱법 역시 1805년 르장드르(A. M. Legendre)에 의해 발표된 이후 가우스(C. F. Gauss)와 라플라스(P. S. Laplace)에 의해 1800년대 초부터 1820년대 사이에 여러 가지 측면에서 그 성질이 연구되었고 이 방법으로 얻은 결과가 바람직한 성질을 가진 것이 증명되면서 그 이후로 널리 쓰이게 되었다 (스티글러, 2005, 제1장-제4장).

한편, 최소제곱법이 발표된 지 약 60년 후인 19세기 후반에 골턴이 새로운 개념과 방법을 창안하게 되었던 분야는 유전학(heredity)이었는데 역시 유전학을 뜻하는 genetics라는 용어는 아직 만들어지기도 전이었다. 골턴의 연구에서 통계학적인 방법들이 사용되고 새로운 방법이 등장하게 된 것은 그가 콩의 일종인 스위트피(sweet pea)의 크기, 또는 사람의 키와 같은 계량적인 자료를 수집, 분석하기 시작하면서 부터였다. 그가 회귀(평범성으로의 회귀, regression to the mediocrity)라고 지칭한 현상은 부모 세대와 자녀 세대 사이의 변이(variation)의 관계인데 골턴은 스위트피의 유전에 대해 설명할 때는 그 관계를 'reversion'이라는 다른 이름으로 부르기도 했다 (Galton 1877). 그런데 1870년대 당시만 하더라도 골턴은 실제로 그러한 관계를 나타내는 어떤 값을 구하려는 시도는 하지 않았다. 그러다가 골턴은 부모 키와 자녀 키 사이의 유전 관계를 발표하면서 회귀라는 용어를 처음 썼고(Galton 1886b: 1885년에 했던 연설을 이듬해에 논문으로 발표한 것이다), 일종의 산점도 위에 회귀 직선을 그

린 다음 그 직선의 기울기 값도 구했다. 또한 골턴은 그 작업을 한 단계 더 발전시켜 두 변수 사이의 관련성을 나타내는 대칭적인 측도로서 상관이라는 개념을 만들고 그 이름도 붙이게 된다(Galton 1888).

먼저 그가 회귀라고 부른 것과 오늘날의 회귀가 어떻게 다른지 알아보자. 골턴은 부모의 특징과 자녀의 특징 사이에는 선형함수관계가 성립해야한다고 믿었으며 그 관계를 하나의 유전 법칙(law of heredity)이라고 여겼으므로 이를 회귀법칙(law of regression)이라고 불렀다. 그는 사람의 여러 특징 중에 키를 측정한 데이터를 가지고 회귀법칙을 구했는데, 사람 키의 변이 폭이 여러 세대에 걸쳐 일정하게 나타나는 이유가 바로 평균보다 크거나 작은 키는 다음 세대에 평균 쪽으로 돌아간다는 회귀법칙 때문이라고 주장했다(Galton, 1889, p. 104 외 여러 곳). 그리고 그는 회귀법칙이 선형함수라고 생각했으므로 부모 키와 자녀의 키 사이의 관계를 나타내는 선형함수의 기울기, 즉 일종의 비례상수에 해당하는 단 하나의 상수만 찾으려 하였다. 그가 회귀 비(ratio of regression)라고 부르고 기호로는 'r'이라고 나타낸 그 상수는 그의 회귀법칙에 따라 1보다 클 수 없었으므로 반드시 0과 1 사이의 상수여야만 했다. 따라서 그가 회귀라는 이름으로 구하려했던 것은 두 변수 사이의 상관계수였으며 엄밀히 말하자면 그것도 양의 상관계수에 국한된 것이었다.

그런데 흥미롭게도 그는 그 값을 구한 뒤 똑같은 회귀법칙이 사람의 모든 특징에 다 적용된다고 믿고는 그 상수 값을 자녀의 키에 대한 부모의 키, 형제 사이의 키, 부모 자녀 사이의 예술적 재능 (artistic faculty) 등을 설명하는데 계속 이용했다. 결국 골턴은 회귀법칙을 단지 유전학에만 국한된 것으로 생각했고 또한 단 하나의 상수로 결정되는 법칙으로 생각했기 때문에 그의 회귀는 지금의 회귀와는 그 목적이나 성격이 크게 달랐던 셈이다. 아마 그러한 차이는 당시 적지 않은 사람들이 그러했듯 골턴 역시 근본적으로 단일한 법칙으로 아주 많은 현상을 설명할 수 있다는 결정론적 과학관과 세계관을 갖고 있었기 때문일 것이다.

그가 쓴 글을 세부적으로 볼 때 그림으로부터 회귀 비를 구해내는 그의 놀라운 직관과는 달리 회귀 비의 값을 구하는 과정에서 드러나는 계산법은 매우 조악한 것이었다. 그의 회귀를 오늘날의 회귀와 비교할 때 가장 두드러져 보이는 차이점은 그가 r을 얻기 위해 이용했던 추정방법이 없었다는 점이다. 그는 최소제곱법은 물론 어떠한 객관적인 추정방법도 사용하지 않았는데, 놀랍게도 그가 이용한 방법은 그림을 그린 다음 대충 눈으로 보고 편리한 수를 얻는 완전한 주먹구구 방법이었다. Galton (1889, pp. 95-99)을 보면 골턴은 부모의 평균키와 자녀 키 사이의 회귀계수를 얻기 위해 데이터를 점으로 나타낸 그림을 그린 다음, 점들을 잘 지나는 듯 보이는 적당한 직선을 하나 그어 그 직선의 기울기를 골랐다고 한다. 그의 설명에 따르면 "처음에는 3/5으로 할까 하다가 2/3가 좀 더 간단해서" 그 값을 택했다는 것이다. 이변량정규분포에 해당하는 타원을 그려서 회귀선의 기울기를 구할 때 그가 쓴 방법도 결과적으로 주먹구구이기는 마찬가지였다. 그러므로 골턴의 회귀는 그 목적은 차치하고 추정방법만 보더라도 1805년 르장드르가 최소제곱법을 만들어 낸 이후 이미 천문학을 비롯한 여러 분야에서 널리 활용되고 있던 선형모형의 계수를 추정하는 방법과는 매우 동떨어진 것이었다.

여기서 과연 최소제곱법이라는 방법이 무엇을 하기 위한 것이었는지 살펴보면 골턴이 왜 자신의 문제에 최소제곱법을 적용할 생각을 하지 않았는지 추리해 볼 수 있을 것이다. 최소제곱법의 성격은 수학적인 것과 통계학적인 것으로 나누어 볼 수 있다. 오차의 제곱 합을 최소로 만드는 미지수를 구하는 방법으로 본다면 최소제곱법은 최적화(optimization) 방법의 하나라고 볼 수 있는데 미분을 통해 풀든 직교투영을 통한 기하학적 방법으로 풀든 이 문제는 순수한 수학적인 문제이다. 한편 최소제곱법의 통계학적 측면을 보기 위해서는 19세기 초에 최소제곱법과 그와 유사한 추정 방법들이 어떤 이름으로 불렸는지 살펴볼 필요가 있다. 예컨대 가우스의 책 제목 *Theory of the Combination of Observations Least Subject to Errors* (1995)에서 볼 수 있듯 그 방법들은 "관측 결과들을 결합하는" 방법으로 불렸다. 당시 사람들이 최소제곱법으로 풀어보려 했던 문제들은 데이터들이 만족해야하는 관계식이 이미 주어져 있고 또 한편 서로 일치하지 않는 관측 데이터들이 있을 때, 데이터 가운데 일부를 버리는 대신 모든 데이터를 다 써서 그 관계식의 미지수들을 찾는 문제였던 것이다.

그렇다면 여기서 우리는 혹시 골턴이 최소제곱법에 대해 알지 못하였을지도 모른다는 생각을 할 수 있다. 그런데 Merriman (1884)을 비롯하여 최소제곱법을 소개하는 문헌들은 당시에 쉽게 구할 수 있었으며 골턴은 자신이 쓴 책에서 Merriman의 책을 언급하기까지 하였으므로(Galton, 1889, p. 202) 그가 최소제곱법을 몰랐을 리는 없다. 그렇다면 결국 골턴은 자신이 원하는 상수를 추정하기 위해서는 최소제곱법을 비롯한 기존의 추정방법을 적용할 수 없다고 믿었던 것이 분명하다. 즉 골턴은 기존의 추정방법들이 적용되던 문제와 자신이 해결해야할 문제는 본질적으로 다른 문제라고 생각했을 것이다. 간단히 표현해서 르장드르와 가우스 그리고 라플라스 등의 최소제곱법에는 독립, 종속변수라는 것이 없었고 서로 일치하지 않으면서 미지수를 여럿 가진 방정식들이 반복 관측 횟수에 해당하는 개수만큼 있을 뿐이었다. 그리고 19세기 초의 수학자들이 랜덤하게 둔 것은 방정식의 '오차'들이었다. 따라서 그들의 방정식에 필요한 미지수를 추정하고 나면 독립변수의 어떤 값에 대해서는 종속 변수의 값이 유일하게 하나로 정해졌고 결과는 천문학이나 물리학의 법칙과 마찬가지가 되었던 것이다. 골턴의 경우에도 앞서와 마찬가지로 부모 키와 자녀 키 사이의 관계식은 그의 회귀법칙에 따라 일차함수로 이미 주어져 있었다. 그런데 골턴은 부모 키, 자녀 키가 모두 랜덤한 변수들이므로 사람들의 키와 평균키 사이의 차이를 하나의 오차항으로 나타낼 수는 없었을 것이다. 나아가 다음 인용문에서 보듯 골턴은 자신이 분석하는 데이터에 대해서는 오차라는 용어가 적절하지 못하다고 명시적으로 밝히고 있다(강조는 필자). Moreover the term Probable Error is absurd when applied to the subjects now in hand, such as Stature, Eye-colour, Artistic Faculty, or Disease. I shall therefore usually speak of Prob. Deviation (Galton, 1889, p. 58). 즉 천문학에서의 오차란 거의 관측오차(measurement error)를 일컫는데 사람에 대해 얻은 측정값을 그처럼 생각할 수는 없다는 말이다. 최소제곱법을 최초로 발표한 르장드르의 글(스티글러, 2005, pp. 94-101, p, 166)이나 가우스의 책 Gauss (1995)을 보면 어디서든 관계식에 'E'라는 기호로 표현된 오차항이 명시적으로 나타나있다. 그리고 그러한 항들은 그 당시부터 오차라고 불렸다. 그리고 가우스와 라플라스에 의해 오차의 분포로 이용된 정규분포

또한 오차의 법칙(Law of Error)이라고 불렸다. 하지만 골턴은 변수들 사이의 관계를 오차항이 포함된 회귀식으로 표현한 적이 없었다. 그가 편차(deviation)라고 불렀던 것은 최소로 만들거나 없애야할 대상이 아니라 설명해야할 대상이었다. 다시 말해 그는 이변량 자료의 편차들 사이의 관계를 설명하기 위해 오차의 제곱합을 최소로 만들기 위한 최소제곱법을 적용할 수는 없었을 것이다. 즉 데이터도 달랐고 그 데이터를 가지고 분석하는 목적 또한 달랐던 것이다. 한편, 골턴이 19세기 중반의 통계학자들과 다른 방식으로 정규분포를 이용할 수 있게 된 과정에 대해서는 스티글러 (2005, 제8장)가 상세히 다룬 바 있다. 그런데, 골턴은 편차라고 부른 것들의 분포를 다룰 때 여전히 정규분포라는 틀 속에서 벗어나지 못했고 회귀라고 일컬었던 두 변수 사이의 관계도 이변량 정규분포의 틀 속에서만 생각했었다. 하지만 지금까지의 논의에 따르면 그가 생각한 정규분포는 그의 앞 세대 사람들이 오랫동안 참값, 즉 평균을 중심으로 한 오차의 법칙이라고 믿어왔던 분포가 더 이상 아니었음이 분명하다.

2.2 교과서 속의 골턴

적지 않은 회귀분석 교과서들이 아버지의 키와 아들의 키 사이의 관계를 연구하는 과정에서 회귀라는 용어가 처음 나타났다고 설명하고 있다. 뿐만 아니라 외국에서 나온 상당수의 교과서들에서도 이러한 설명을 종종 볼 수 있다. 예컨대 통계학 교과서의 고전 가운데 하나라 할 수 있는 Yule and Kendall (1950)에서도 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 골턴이 연구한 바와 다른 것이다. 앞서 언급했듯 회귀라는 용어가 나오는 골턴의 연구는 모두 유전학을 위한 것이었는데 골턴은 유전이라는 과정에 기여하는 정도는 부모 양쪽이 공평하게 똑같다고 생각했었다. 그러므로 골턴 자신이 상세히 밝혔듯(Galton, 1889, pp. 5-7 외 여러 곳) 그가 다룬 변수는 아버지와 아들의 키가 아니고 부모의 평균키와 자녀의 키였다. 더 정확하게 말하면 골턴은 부모의 역할을 공평하게 만들기 위해 어머니의 키에는 1.08이라는 가중치를 부여하여 부모 키의 가중평균(골턴은 이렇게 계산한 값을 'midparent'라고 불렀다)을 계산하였다. 그는 자녀의 키를 다룰 때에도 역시 성인이 된 딸의 키에는 같은 가중치를 부여했다. 그의 데이터는 205쌍의 부모에게서 태어난 928명의 자녀 키를 측정한 것이었다. 많은 교과서들에서 골턴이 아버지와 아들의 키를 분석하는 과정에서 회귀라는 용어를 처음 썼다고 잘못 설명하게 된 이유는 아마 칼 피어슨이 *Biometrika*에 실은 논문 (Pearson and Lee, 1903)에서 아버지와 아들의 키 1,078쌍을 분석한 바 있고 그것이 20세기 중반까지 널리 쓰이던 교과서로서 앞서 언급한 Yule and Kendall (1950, p. 202) 등에 그대로 인용되어 실렸기 때문일 것이다. 그런데 골턴이 여러 곳 (Galton, 1886a, 1886b, 1889)에서 부모 키의 가중평균을 썼다고 강조하였음에도 불구하고 Galton (1890)을 보면 골턴 자신이 Galton (1886b)에 실었던 회귀에 대해 설명하면서 아버지의 키를 분석했다고 잘못 썼고, 심지어는 Galton (1890)을

*Statistical Science*에 재수록하면서 스티글러가 덧붙인 논문(Stigler, 1989)에는 아버지의 키(p. 74)라고도 되어있고 부모의 키(p. 75)라고도 되어있다. 이러한 결과는 아무래도 남성 학자들이 지닌 뿌리 깊은 부계 혈통주의적인 사고 탓으로 보인다.

한편 회귀에 대해 설명하는 대부분의 통계학 교과서에서는 이변량 정규분포에서 조건부 기댓값의 선형관계 즉 $(X, Y) \sim N_2(\mu_1, \mu_2, \sigma_1, \sigma_2, r)$일 때

$$\mathrm{E}(Y \mid X=x)=\mu_2+r\frac{\sigma_2}{\sigma_1}(x-\mu_1), \quad \mathrm{E}(X \mid Y=y)=\mu_1+r\frac{\sigma_1}{\sigma_2}(y-\mu_2)$$

라는 관계를 골턴이 밝혔다고 설명한다. 하지만 이러한 설명은 엄밀히 말해서 골턴이 한 것과 다른 것이다. 골턴은 논의 과정에서 시종일관 기댓값이 아니고 M이라는 기호로 나타낸 중앙값을 썼는데 그 이유로는 Median is practically the same as the mean, but is a more convenient value to find (Galton, 1886b, p. 261) 라고 밝혔다. 그리고 그는 표준편차가 아니라 Q로 나타낸 probable error(사분위수범위의 반을 뜻하는 것으로서 당시까지 표준편차 σ 대신 널리 쓰였다)를 썼다. 골턴의 표현과 오늘날의 설명 사이의 관계는, 정규분포에서 평균과 중앙값이 같으며 $1Q = 0.6745\sigma$의 관계가 있으므로 결과적으로 서로 마찬가지이다. 하지만 Gilchrist (2005)처럼 기댓값보다는 중위수 등에 바탕을 둔 통계량을 선호했던 골턴의 의도가 나름대로의 장점을 갖는데도 불구하고 오늘날 왜곡되어 버렸다고 강조하는 학자도 있다.

2.3 케틀레와 골턴

골턴은 그의 다양한 경력에서 볼 수 있듯이 수학이나 통계학만을 깊이 연구한 사람이 아니었다. 그럼에도 불구하고 골턴이 당시의 지배적인 통계학, 즉 케틀레의 통계학으로부터 벗어나 그 나름대로의 연구를 하게 된 배경에 대해 알아보려한다. 먼저 케틀레에 대해 살펴보고 다음으로 골턴이 케틀레로부터 벗어날 수 있었던 이유에 대해 알아볼 것이다. 그 결과로부터 또한 19세기에 통계학이 다른 학문들과의 관계 속에서 발전해온 사례에 대해서도 알아볼 것이다.

19세기 통계학의 역사를 연구하는 학자들은 케틀레를 다양한 방향에서 연구해 왔다. 그 가운데 스티글러가 특별히 강조했던 것은 동질성에 대한 케틀레의 검정, 즉 사람들에 대한 어떤 데이터가 있을 때 그 데이터가 과연 동질적인 집단을 측정해서 얻은 것인가에 대한 검정이었다. 스티글러의 주장에 따르면 아직 적절한 적합도검정법이 없는 상태에서 데이터의 분포가 정규분포이면 동질적인 집단의 것이라고 판단하는 케틀레의 검정은 결과적으로 실패였다는 것이다. 그런데 케틀레가 왜 데이터의 동질성 여부를 밝히는데 그처럼 몰두하게 되었는지에 대해서는 통계학사 연구자들이 별달리 언급하고 있지 않다. 스티글러의 경우에는 단지 케틀레, 렉시스, 에빙하우스 등의 연구가 라플라스의 확률이론을 사회과학에 적용하기 위한 시도들이었다고만 언급했을 뿐이다(스티글러, 2005, 제5, 6, 7장).

이 연구에서는 케틀레가 동질적인 집단이라는 주제에 집중한 이유를 밝히는 실마리를 19세기 중엽에 사회학(sociology)이라는 새로운 학문이 탄생하였다는 역사에서 찾아보려한다. 사회학이라는 학문 명칭을 만든 콩트(Auguste Comte)로 대표되는 사람

들은 당시에 사회를 여러 개인들의 모임으로 보는 대신 사회 자체를 하나의 연구 단위로, 즉 사회를 통일성 있는 독자적인 개체로 보고 개인의 특성이 제거된 사회의 법칙을 과학적으로 찾아야한다고 주장하고 있었다.

오랫동안 천문대장 일을 하면서 오차이론에 밝았던 케틀레는 어떤 집단이 사회학에서 연구할 하나의 단위가 될 수 있는가 없는가를 오차의 법칙, 즉 정규분포를 이용하여 판단할 수 있다고 보았을 것이다. 그런데 그가 주장한 방법은 라플라스의 이론을 뒤집은 것이었다. 라플라스의 이론이란 천문학에서 관측을 반복하면 유일한 참값을 중심으로 정규분포를 따르는 데이터를 얻게 되고 그런 데이터는 동질적인 집단으로 간주할 수 있다는 것이었다. 그런데 케틀레의 방법은 데이터가 나온 조건을 따지기 전에 먼저 데이터의 분포를 보고 그 분포가 정규분포라면 그 데이터는 단일한 평균을 갖는 동질적인 집단의 데이터라고 판단하자는 것이었다(Quetelet, 1849).

또한 그는 평균적인 사람에 대해 the average man of any one period represents the type of development of human nature for that period; ... the average man was always such as was conformable to and necessitated by time and place; in perfect harmony, alike removed from excess or defect of every kind, so that, in the circumstances in which he is found, he should be considered as the type of all which is beautiful-of all which is good (Quetelet, 1842, p. 100)와 같이 주장했는데 여기서 그의 목적은 사람들이 모인 집단이되 그 집단을 구성하는 개인들의 잡다한 특성은 모두 제거된 뒤에 얻어지는, 사회의 일관되고 독자적인 특성을 일컫기 위한 것이라 할 수 있을 것이다. 즉 그의 평균적인 사람은 물리학의 무게중심처럼 동질적인 사회 집단의 중심이자 통계라는 과학적인 방법을 사회집단에 적용하여 얻은 것으로서 사회학이라는 신생 학문이 연구 목적으로 삼아야할 대상이므로 위의 인용문에서처럼 "모든 아름답고 선한 것들의 전형"으로 일컬을 수 있었을 것이다. 만일 동질적인 사람의 집단을 찾기 위한 사회학 문제를 해결하려는 목적이 없었다면 평균적인 사람이라는 주제, 그리고 정규분포를 이용해서 동질성을 검정하는 문제에 케틀레가 1830년대 이후 상당히 오랫동안 집중할 필요는 없었을 것으로 보인다.

그런데 골턴의 경우에는 데이터에서 얻으려는 목표가 케틀레의 것과 달랐다. 골턴의 연구 분야는 유전학이었다. 그는 다윈의 『종의 기원』을 읽고 자연 선택이 인간 사회에서도 적용됨을 보이기 위해 사람들의 특성이 세대 사이에 어떻게 유전되는지 연구하기 시작하였는데 골턴이 유전을 연구하기 시작한 1860년대는 유전학을 일컫는 'genetics'라는 말이 생기기도 전이었다. 그러한 이유로 회귀라는 용어가 처음 나온 논문 Galton (1886b)은 인류학 학술지에 발표되었다. 골턴이 케틀레와 달랐던 점을 잘 보여주는 사례로는 케틀레가 데이터의 중심을 대단히 강조했음에도 불구하고 골턴은 다음 인용문에서처럼 당시 통계학자들이 평균에만 주목하는 것을 명시적으로 반박하였던 점을 들 수 있다. It is difficult to understand why statisticians commonly limit their inquiries to Averages, and do not revel in more comprehensive views. Their souls seem as dull to the charm of variety as that of the native of one of our flat English counties, whose retrospect of Switzerland was that, if its mountains could be thrown into its lakes, two nuisances would be got rid of at once (Galton, 1889,

p. 62). 케틀레는 물론 평균이라는 단순한 통계학적 방법을 대단히 지나칠 만큼 강조하였다. 그러나 그처럼 평균을 강조하는 사람을 골턴이 마치 촌뜨기 취급하며 비난한 것 역시 지나쳐보인다. 평균을 둘러싼 두 사람의 이와 같은 편향은 그들이 통계학적 방법 그 자체보다는 각자가 관심을 가진 서로 다른 문제에 더 집중했음을 보여주는 사례라 하겠다.

또한 케틀레와 골턴 두 사람은 정규분포에 대해서도 서로 다른 생각을 갖고 있었다. 스티글러는 다른 통계학사가들에 비해 골턴이 만든 퀸컹커스(quincunx)라는 장치를 매우 높이 평가한 바 있는데 그의 주장에 따르면 골턴은 퀸컹커스라는 장치를 가지고 여러 정규분포가 혼합될 경우에도 정규분포가 생긴다는 사실을 보여주었다는 것이다. 다시 말해서 케틀레의 생각처럼 정규분포가 단일한 중심에 오차가 덧붙을 경우에만 생기는 것이 아니므로 결국 라플라스의 이론은 정규분포가 생기기 위한 충분조건이 아님을 골턴이 보였다는 것이다(스티글러, 2005, 제8장).

여기서 골턴이 그런 증명을 필요로 했던 이유에서도 우리는 유전학의 영향력을 찾아볼 수 있다. 골턴은 평균이라는 것이 평범해서 관심 대상이 되지 못한다고 하면서 평균 대신 집단에서 특출한 재능을 가진 사람과 같이 분포의 끝부분에 있는 데이터에 주목하였다. Galton (1869)에서 그는 저명인사들의 족보를 추적하여 그러한 집안에서 역시 저명인사가 나오는 경우가 많다고 주장하였으며, 나아가 뛰어난 재능을 가진 좋은 혈통을 지닌 사람들의 출산은 장려하고 그렇지 못한 사람들의 출산은 억제함으로써 사회와 국가가 진보하리라는 신념, 즉 우생학(eugenics)을 평생 주장하였다. 따라서 그에게는 사회 전체 사람의 재능이 정규분포 모양의 분포를 가질 때 그 분포가 라플라스나 케틀레의 생각처럼 집단 전체의 단일한 평균을 중심으로 작은 오차들이 모여서 만들어진 것이어서는 쓸모가 없었을 것이다. 핏줄에 따라 재능이 유전된다고 주장했던 그의 입장에서는 부모로부터의 유전이라는 뚜렷한 원인에 의해 여러 작은 정규분포 집단들이 생기고 그 분포들이 혼합되어 전체 사회의 정규분포가 생겨야 했던 것이다. 따라서 그가 얻은 결과들은 통계학적 방법이 유전학 이론을 유도해낸 것이라기보다는, 그가 주장하고 싶어 했던 유전학적 개념과 모형에 의해 새로운 통계학적 개념과 방법이 나온 경우라고도 할 수 있겠다.

이처럼 통계학과 밀접하게 관련성을 갖고 태어난 유전학에서는 통계학과 유전학의 관계를 둘러싼 논쟁이 몇 십년간 진행되기도 하였다. 그 한쪽 편에는 칼 피어슨(K. Pearson)과 웰던(W. F. Weldon)으로 대표되는 생물측정학파(biometricians)가, 다른 한 편에는 베이트슨(W. Bateson)으로 대표되는 멘델학파(Mendellians)가 있었는데 20세기로 넘어와서 그 논쟁을 종식시킨 사람 가운데 한 사람이 피셔(R. A. Fisher)였다(Provine, 1971; MacKenzie, 1981; Kim, 1994). 그런데 그 논쟁에서 골턴의 위치가 흥미롭다. 그는 논쟁의 주역은 아니었지만 양쪽에 모두 소속될 수도 있는 독특한 위치에 있었던 것이다. 그는 생물측정학파의 창시자였을 뿐 아니라 다른 한편으로 멘델학파의 주장 중 하나였던 유전의 불연속성에 동조하기도 하였기 때문이다.

이와 같이 어떤 분야의 문제를 해결하기 위해 통계학적 방법을 적용하는 동안 새로운 통계학 이론과 방법이 개발되어 통계학이 새로운 단계로 발전하고 또한 그 방법이 적용되는 분야 역시 변모하게 되는 역사는 통계학이 갖는 가장 큰 장점이고 어쩌면

통계학의 역사만이 갖는 독특한 면모인 듯 보인다.

3. 결론

골턴은 1911년에 죽었는데 그 해에 University College London에 세계 최초의 통계학과라 할 수 있는 Department of Applied Statistics가 생겼다. 그 학과가 생긴 것은 다름 아닌 골턴이 남긴 유산 덕분이었으며 최초의 교수는 피어슨(K. Pearson)이었다. 따라서 통계학사에서 골턴은 비단 회귀와 상관으로만 기록되지 않고 통계학을 대학에 제도적으로 자리 잡게 했던 업적으로도 기록될 만하다. 또한 골턴은 뛰어난 통계그래픽을 남긴 인물로도 통계학사에 남아있다. 그 대표적인 사례로 그가 회귀를 다루면서 부모의 평균키와 자녀의 키를 조사한 통계표를 그림으로 나타낸 것을 들 수 있다 (Galton, 1886b, Table 1, Plate X). 그 그림을 잘 살펴보면 좌표의 위치를 나타내는 점들이 보통 그림들처럼 수평 수직선의 교점이 아니고 숫자들로 되어있다. 그 숫자들은 바로 표에 나타나있는 빈도 값들인데 이처럼 좌표의 위치에 흔히 하듯 선이나 점을 그려 넣는 대신 데이터를 채워 넣는 방법은 골턴의 뛰어난 아이디어 가운데 하나라고 평가된다(Tufte, 2001, pp. 145-148).

본 연구에서는 골턴이 생각했던 회귀는 그 목적이나 성격이 오늘날의 회귀분석과는 많이 달랐고 그 용어를 처음 만들 때 골턴이 분석한 데이터는 그가 생각하기에 최소제곱법을 적용할 수 없는 데이터였으며, 결과적으로 최소제곱법은 당시에 그 용도가 지금보다 제한적이었다는 결론을 유도할 수 있었다. 또한 오늘날의 교과서에서 골턴을 소개할 때 볼 수 있는 오류에 대해서도 살펴보았으며 마지막으로 케틀레와 골턴의 연구가 나오게 된 배경에 대해서도 알아보았다. 본문에서 주장한 것은 골턴이 평균이나 정규분포 등을 케틀레와 매우 다르게 생각하고 다른 방향으로 연구하게 된 것은 사회학이 필요로 하는 바를 얻으려했던 케틀레에 비해 그가 유전학이라는 새로운 분야에 관심을 가졌기 때문이라는 것이다.

한편, 과학사 분야에서 케틀레는 다윈이 진화이론을 만드는데 중요한 영향을 미쳤다고 평가되고, 다윈은 골턴에게 또한 많은 영향을 미친 것으로 평가된다. 따라서 케틀레-다윈-골턴 이 세 사람 사이의 관계도 향후 통계학사, 생물학사, 그리고 생물철학 연구자들의 공동 연구에 의해 여러 측면에서 연구될 수 있을 것이다. 이러한 작업은 골턴을 비롯한 세 사람의 학문적, 사회적 공과를 후대의 시각으로 평가한다는 의미보다는 통계학이 하나의 독립된 학문으로서 탄생하게 되는 시대적 배경을 알아보고 좀 더 넓은 지평에서 통계학의 정체성을 파악하는 기회를 제공할 수 있을 것이다.

참고문헌

[1] 스티글러, S.M. (2005). 『통계학의 역사』, 조재근 옮김, 한길사, 경기도 파주시; Stigler, S.M. *The History of Statistics*, Belknap Press of Harvard University Press, Cambridge, Massachusetts, (1986).

[2] Bulmer, M.G. (2003). *Francis Galton: Pioneer of Heredity and Biometry*, Johns Hopkins University Press, Baltimore.

[3] Brookes, M. (2004). *Extreme Measures: The Dark Visions and Bright Ideas of Francis Galton*, Bloomsbury.

[4] Farebrother, R.W. (1999). *Fitting Linear Relationships-A History of the Calculus of Observations, 1750-1900*, Springer, New York..

[5] Galton, F. (1865). Hereditary talent and character, *Macmillan's Magazine*, Vol. 12, 157-166, 318-327.

[6] Galton, F. (1869). *Hereditary Genius*, Macmillan, London.

[7] Galton, F. (1877). Typical laws of heredity. *Proceedings of the Royal Institution*, Vol. 8, 282-301.

[8] Galton, F. (1886a). Family likeness in stature. *Proceedings of the Royal Society of London,* Vol. 40, 42-73.

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

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

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

[12] Galton, F. (1890). Kinship and correlation. *North American Review,* Vol. 150, 419-431.

[13] Gauss, C.F. (1995). *Theory of the Combination of Observations Least Subject to Errors*, translated by G.W. Stewart, SIAM, Philadelphia; *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae.* Göttingen: Dieterich, 1823.

[14] Gilchrist, W.G. (2005). Galton misrepresented, *Significance*, Vol. 2, 136-137.

[15] Gillham, N.W. (2001). *A Life of Sir Francis Galton: From African Exploration to the Birth of Eugenics*, Oxford University Press, Oxford.

[16] Hacking, I. (1990). *The Taming of Chance*, Cambridge University Press, Cambridge.

[17] Keynes, M. (1993). *Sir Francis Galton, FRS: The Legacy of His Ideas-Proceedings of the twenty-eighth annual symposium of the Galton Institute, London,* (1991), Macmillan, London.

[18] Kim, K.-M. (1994). *Explaining Scientific Consensus: The Case of Mendelian Genetics*, Guilford Press, New York.

[19] MacKenzie, D. (1981). *Statistics in Britain, 1865-1930: The Social Constr -uction of Scientific Knowledge,* Edinburgh University Press, Edinburgh.

[20] Merriman, M. (1884). *A Text-Book on the Method of Least Squares*, 8th edition, Wiley, New York, (first edition의 제목은 *Elements of the Method of Least Squares*, 1877).

[21] Pearson, K. and Lee, A. (1903). On The Laws Of Inheritance In Man: I. Inheritance Of Physical Characters. *Biometrika,* Vol. 2, 357-462.

[22] Provine, W. B. (1971). *The Origins of Theoretical Population Genetics*, University of Chicago Press, Chicago.

[23] Quetelet, M.A. (1842). *A Treatise on Man and the Development of His Faculties*. Edinburgh: Chambers; *Sur l'homme et le développement de ses facultés, ou Essai de physique sociale*. Paris: Bachelier, (1835).

[24] Quetelet, M.A. (1849). *Letters Addressed to H. T. H. the Grand Duke of Saxe Coburgand Gotha, on the Theory of Probabilities as Applied to the Moral and Political Sciences*, trans. O. G. Downes. London: Layton; *Lettres* à *S. A. R. le Duc Régnant de Saxe-Cobourget Gotha, sur la théorie des probabilités, appliquée aux sciences morales et ploitiques*. Brussels: Hayez, (1846).

[25] Seal, H.L. (1967). The Historical Development of the Gauss Linear Model, *Biometrika*, Vol. 54, 1-24.

[26] Stigler, S.M. (1989). Francis Galton's Account of the Invention of Correlation, *Statistical Science*, Vol. 4, 73-80.

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

[28] Tufte, E. (2001). *The Visual Display of Quantitative Information,* second edition, Graphics Press, Cheshire, Connecticut.

[29] Yule, G.U. and Kendall, M.G. (1950). *An Introduction to the Theory of Statistics,* 14th edition, Hafner, New York.

[Received October 2005, Accepted August 2006]