건강검진 빅데이터를 이용한 선형 및 다중회귀분석 기반 헤모글로빈 추정 방법에 관한 연구

홍상훈*, 홍광석*
*성균관대학교 전자전기컴퓨터공학과
hshjjang10@g.skku.edu, kshong@skku.ac.kr

A Study on the Estimation Method of Hemoglobin Based on Linear and Multiple Regression Analysis Using Health Examination Big Data

Sang-Hoon Hong*, Kwang-Seok Hong*
*Dept. of Electrical and Computer Engineering, Sungkyunkwan University

요 약

빈혈의 유병률은 매년 증가하고 있으나 이를 가벼운 질병으로 인식해 치료 시기를 놓치는 환자들이 존재한다. 빈혈의 발생원인으로 혈액 내 헤모글로빈 및 헤모글로빈 내 철 부족이 있으며, 헤모글로빈 측정기술의 경우 채혈 이외에 사람의 신체 및 건강 정보를 적용한 사례는 찾아보기 어렵다. 본 논문에서는 신체(키, 몸무게 및 허리둘레) 및 건강 정보(혈청지오티, 이완기 혈압 및 감마지티피 등)가 포함된 건강검진 빅데이터를 이용하여 단일 특징에 대해 선형회귀분석을 수행하고, 다중 특징에 대해 다중회귀분석을 수행하여 회귀분석 식을 산출, 산출된 회귀분석 식을 통해 헤모글로빈을 추정하여 실제 헤모글로빈값과 오차율을 계산하고 비교한다. 실험 결과, 선형회귀분석 식을 통해 헤모글로빈을 추정하였을 때 평균 8.124%의 오차율이 계산되었으며, 다중회귀분석의 경우 선형회귀분석보다 낮은 6.767%의 오차율이 계산되었다.

1. 서론

최근 빈혈에 관한 보건복지부와 질병관리본부의 국민건강영양조사에 따르면 만 10세 이상 국내 빈혈 유병률은 11.5%로 여성이 남성보다 약 4.4배 정도 높게 나타났으며, 특히 70세 이상에서 남성은 11.1%, 여성은 18.0%로 여성 환자 수가 7% 가까이 많은 것으로 조사되었다[1]. 이러한 빈혈은 적혈구에서 철을 포함하며 산소 운반을 담당하는 붉은 색 단백질인 헤모글로빈(hemoglobin 또는 haemoglobin)과 관련이 있고, 헤모글로빈 및 헤모글로빈 내 성분 중 하나인 철이 부족하게 되면 헤모글로빈이 완전하게 기능할 수 없어 빈혈(anemia)이 일어나게 된다[2]. 하지만 빈혈을 가벼운 질병으로 인식해 치료 시기를 놓치는 환자들이 많으며, COVID-19로 인하여 건강관리에 대한 인식이 강화되고 디지털·비대면으로의 스마트 헬스케어 산업에 대한 기대가 확대되고 있지만, 헤모글로빈 측정기술의 경우 채혈 이외에 사람의 신체 및 건강 정보를 헤모글로빈 측정기술의 향상에 적용한 사례를 찾아보기 어렵다[3][4].

본 논문에서는 신체(키, 몸무게 및 허리둘레 등) 및 건강 정보(혈청지오티, 이완기 혈압 및 감마지티피 등)가 포함된 국민 건강검진 데이터를 이용하여 독립변수와 종속변수 간의 상관성을 통계적으로 분석하고, 이를 통해 헤모글로빈 추정 실험에 사용될 특징들을 선별하여 선형 및 다중회귀분석을 통해 회귀분석 식을 산출하였다. 산출된 회귀분석 식을 통해 헤모글로빈을 추정하여 실제 헤모글로빈값과 오차율을 계산하고 비교하였다.

2. 관련연구

기존 헤모글로빈 측정기술로는 채혈 측정방법과 광 산란 측정방법이 있다. 채혈 측정의 예로는 헤모글로빈 측정기가 있으며, 녹십자 세라첵 Hb 플러스와 베네첵 헤모글로빈 측정기 두 가지 제품이 시장을 선점하고 있고, 마이크로니들(needle)을 사용하여 채혈을 통해 헤모글로빈을 측정한다. 이는 보정 및 통증이 존재하는 최소침습 방식이라는 한계를 가진다[5]. 광 산란 헤모글로빈 측정의 경우 혈중 헤모글

로빈 농도를 측정하는 가장 대표적인 기기로, 혈구 검사기(Hematology Analyzer)가 있다. 병원이나 대형 실험실에서 많이 사용되는 이 장비는 매회 측정 시 50~200㎕의 혈액이 필요하며, 상대적으로 매우 높은 정확도를 가지고 있지만, 기기의 가격과 매회 측정에 발생 되는 비용이 크고, 큰 부피와 무게를 차지해 일반인 혹은 비입원 환자가 정기적인 모니터링을 하기에는 부적합하다. 훈련받은 전문 인력 외의 일반인에게 접근성이 떨어지며, 측정 때마다 독성물질인 포타슘 시안 산화물(청산가리)이 사용되는 시안화법(Cyaniding)을 적용하는 문제가 있다[6].

3. 제안방법

본 연구를 수행하기 위해 수집한 데이터는 '국민건강보험공단'[7]에서 제공하는 3개년(2014, 2015, 2017년)에 대한 국민 건강검진 DB이며, 신체 및 건강 정보에 해당하는 총 26개의 특징으로 이루어져 있다. 수집한 데이터에 대하여 전처리(결측치 및 이상치 데이터 제거)를 수행하였으며, 단일 특징에 대해 선형회귀분석을 수행하고, 다중 특징에 대해 다중회귀분석을 수행하여 회귀분석 식을 산출, 헤모글로빈 추정을 통하여 실제 헤모글로빈값과 오차율을 계산하고 비교하였다.

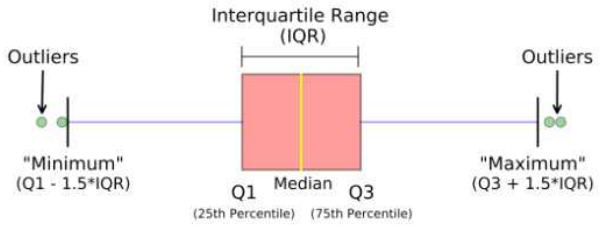
3.1 데이터 전처리

본 연구에서는 총 3,000,000명의 신체 및 건강 정보 데이터의 결측치에 대하여 행 제거를 수행하였으며, 헤모글로빈 추정을 위한 회귀분석 식 산출에 이용될 특징을 선별하기 위하여 결측치가 제거된 데이터에 대해 상관 관계분석을 수행하였고, 상관계수가 ±0.2 이상으로 계산된 변수는 표 1과 같다.

<표 1> 상관 관계분석 결과(상관계수 ±0.2 이상)

No.	상관계수	변수명
1	0.51	HEIGHT(키)
2	0.48	WEIGHT(몸무게)
3	0.32	WAIST(허리둘레)
4	0.24	TRIGLYCERIDE (트라이글리세라이드)
5	0.24	SGOT_ALT (혈청지오티(ALT))
6	0.23	BP_LWST(이완기 혈압)
7	0.22	GAMMA_GTP (감마지티피)

데이터 불균형 방지를 위해 특징별 정규분포의 99.3%를 벗어나는 데이터를 이상치 데이터(Outlier)로 판단하는 IQR(Inter Quantile Range, 사분 범위) 방식을 이용하여 이상치 데이터를 제거하였으며[8], 이는 그림 1 및 식(1)과 같다.



(그림 1) IQR

$$IQR = Q3 - Q1$$
$$Minimum = Q1 - 1.5 \times IQR$$
$$Maximum = Q3 + 1.5 \times IQR \quad (1)$$

- $Q1$ = 특징별 정규분포의 $\frac{1}{4}$ 지점
- $Q3$ = 특징별 정규분포의 $\frac{3}{4}$ 지점

이상치 데이터 제거를 수행한 후, 상관계수 변화를 확인하기 위하여 상관 관계분석을 재수행하였으며, 변수별 계산된 상관계수는 표 2와 같다. 또한, 결측치 및 이상치 데이터 제거를 통한 데이터 개수의 변화는 표 3과 같다.

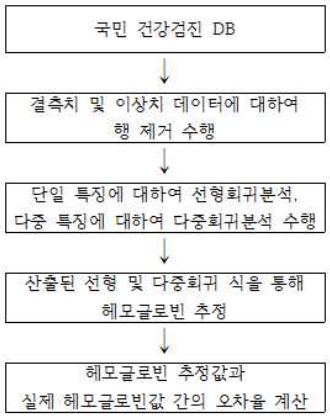
<표 2> 이상치 데이터 제거를 수행한 데이터에 대한 상관 관계분석 결과

No.	상관계수	변수명
1	0.52	HEIGHT(키)
2	0.46	WEIGHT(몸무게)
3	0.31	WAIST(허리둘레)
4	0.20	TRIGLYCERIDE (트라이글리세라이드)
5	0.32	SGOT_ALT (혈청지오티(ALT))
6	0.19	BP_LWST(이완기 혈압)
7	0.36	GAMMA_GTP (감마지티피)

<표 3> 결측치 및 이상치 데이터가 제거된 데이터 개수

구 분	데이터 개수
원데이터	3,000,000
결측치 제거	2,972,311
이상치 데이터 제거	2,361,110

4. 실험 및 결과



(그림 2) 헤모글로빈 추정 실험절차

상기 그림 2의 절차에 따라 실험을 진행하였으며, 실험 결과 선형회귀분석을 통하여 헤모글로빈을 추정하여 실제 헤모글로빈값과 비교하였을 때 평균 8.124%의 오차율이 계산되었으며, 다중회귀분석의 경우 선형회귀분석보다 더 낮은 6.767%의 오차율이 계산되었고, 다중회귀분석의 오차율은 FDA와 CLIA에서 승인한 자동헤모글로빈측정기의 오차율 허용기준인 7% 이내에 해당하는 결과를 보였다[9]. 선형 및 다중회귀 식을 통해 산출된 헤모글로빈 추정값과 실제 헤모글로빈값 간의 오차율 계산 결과는 표 4와 같다.

<표 4> 오차율 계산 결과

선형 및 다중회귀 식을 통해 산출된 헤모글로빈 추정값과 실제 헤모글로빈값 간의 오차율			
독립변수		선형회귀	다중회귀
		오차율(%)	
x1	키	7.313	6.767
x2	몸무게	7.603	
x3	허리둘레	8.271	
x4	이완기 혈압	8.645	
x5	트라이글리세라이드	8.633	
x6	혈청지오티(ALT)	8.296	
x7	감마지티피	8.105	

5. 결론 및 향후 과제

본 논문은 총 2,361,110명의 신체 및 건강 정보 데이터를 이용하여 선형 및 다중회귀분석을 수행해 산출된 회귀분석 식을 통하여 헤모글로빈을 추정하고 오차율을 계산 및 비교하였다. 오차율 계산 및 비교 결과, 선형 회귀분석보다 다중회귀분석을 통해 산출된 회귀분석 식을 이용하여 실제 헤모글로빈값과 오차율을 계산하였을 때 1.357% 더 낮은 오차율이 계산되는 것을 확인하였으며, 회귀분석에 적용하는 특징(키, 몸무게 및 감마지티피 등)에 따라 오차율이 변화하는 것을 확인하였고, 이를 통해 신체 및 건강 정보를 이용한 헤모글로빈 추정 가능성을 확인하였다. 향후 성별, 나이 및 흡연상태 등을 이용한 클래스 세분화를 통하여 강인한 헤모글로빈 추정 방법에 관한 연구를 수행할 계획이며, 회귀분석 식 산출 과정에서 데이터 분리를 통하여 신뢰도 높은 연구를 수행할 계획이다. 위 과정을 통해 최종적으로 사용자가 신체 및 건강 정보를 입력하여 헤모글로빈을 추정할 수 있는 형태의 애플리케이션을 개발해 볼 계획이다.

ACKNOWLEDGMENT

본 논문은 2021년도 정부(교육부)의 재원으로 한국 연구 재단의 지원(NRF-2018R1D1A1B07042422)을 받아 수행된 것임

참고문헌

[1] KDCA(Korea Disease Control and Prevention Agency), “2019 국민건강통계”, 2020

[2] MJ Guzmán Llanos, “Significance of anaemia in the different stages of life”, Enfermería global, vol. 15, no. 3, pp. 419-430, 2016

[3] KISTEP(Korea Institute of S&T Evaluation and Planning), “스마트 헬스케어”, 2020

[4] NIPA(National IT Industry Promotion Agency), “스마트 헬스케어 서비스 분야 도입사례 분석집”, 2017

[5] S Yoon, “Differences in Hemoglobin Levels as Measured by Blood Gas and Auto Blood Cell Count Analyzers”, vol. 20, no. 3, pp. 242-246, 2009

[6] 김의한, “광열 광 각도 산란을 이용한 적혈구내 헤모글로빈 농도 측정”, 대한기계학회 춘추학술대회, 2014, 2613-2615

[7] 공공데이터포털[웹사이트], (2020.12.20.), URL: https://www.data.go.kr/

[8] Whaley, Dewey Lonzo, "The Interquartile Range: Theory and Estimation.", Electronic Theses and Dissertations. Paper 1030, 2005

[9] NIFDS(National Institute of Food and Drug Safety Evaluation), “자동헤모글로빈측정기 평가 가이드라인 개발 연구”, 2016