시군구별 코로나19 사망자 수 예측을 위한 선형회귀분석

김건*, 배지훈*, 이종혁**

Linear regression analysis for predicting the number of deaths due to COVID-19 by city, province, and district

Geon Kim*, Ji-Hoon Bae*, and JongHyuk Lee**

요 약

본 논문은 코로나19에 따른 사망률을 감소시키기 위해 사망 위험에 영향을 미치는 주요 변수들을 도출하고 선형회귀분석 기법을 이용한 분석 방법을 제안한다. 제안하는 분석 방법은 보건 분야 예산액, 의료진 현황, 음주율 등의 데이터들을 사용한다. 본 논문에서는 지역별 특성과 같은 데이터들을 적용한 선형회귀 모형을 개발하고, 사망 위험에 영향을 미치는 유의미한 변수들을 발견하여 시군구별 코로나19 사망자 수를 예측한다.

Abstract

In this paper, we propose a linear regression model by deriving key variables affecting the risk of death to reduce the mortality rate of COVID-19. The proposed linear regression analysis uses data such as health sector budget amounts, health care staff status, and alcohol consumption rates. This paper develops a linear regression model that applies data such as regional characteristics, finds significant variables affecting mortality risk, and predicts the number of deaths by city, province, and district.

Key words

COVID-19, number of deaths, linear regression analysis

Ⅰ. 서 론

코로나감염증-19(COVID-19, 이하 코로나19)는 2020년 12월 22일 현재 전 세계 220개국으로 확산되어, 확진자 77,705,355명을 기록하고 있으며, 사망자 수는 1,708,610명에 이르고 있다[1]. 코로나19 감염자가 사망에 이르기까지 많은 복합적 요인이 존재하겠지만 현재까지 위험 요인으로는 고령, 면역 저하, 비만, 만성 신장 질환, 암, 당뇨병, 심장질환, 폐질환 등이 중증 감염 및 사망의 위험인자로 발표된 바 있다[2]. 따라서 본 논문에서는 기존의 개인에 국한된 사망 위험 요인에서 보다 통합적으로 지역별 특성에 따른 사망 위험 요인을 분석한다. 이를 위해 시군구별 보건 분야 예산액, 인구밀도, 의료시설 현황 등에 관한 데이터를 수집한 후, 선형회귀분석을 적용하여 시군구별 사망자 수를 예측하고자 한다.

* 대구가톨릭대학교 소프트웨어융합대학 인공지능 · 빅데이터공학과
** 교신저자: 대구가톨릭대학교 이종혁(jonghyuk@cu.ac.kr)

※ 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업의 연구결과로 수행되었음(2019-0-01056).

표 1. 실험 데이터 세트
Table 1. Experimental data sets

활용데이터	내용	출처
보건분야 예산액	시군구별 보건 의료 부문 정책사업 예산 총계	지방재정 365 (2020)
나이	시군구별 평균 연령	KOSIS (2020)
인구밀도	시군구 면적당 인구비율	e-나라지표 (2020)
흡연율	시군구별 흡연 여부 정보	KOSIS (2020)
음주율	시군구별 음주 여부 정보	KOSIS (2020)
의료시설 현황	시군구별 일반입원실, 중환자실, 격리병실 현황	KOSIS (2020)
의료 인력	시군구별 일반의, 전문의, 간호사 현황	KOSIS (2020)
고령화 비율	시군구별 고령화 비율	KOSIS (2020)
사망자	시군구별 covid-19 사망자 수	질병관리본부

II. 코로나19 사망자 수 선형회귀분석

본 논문에서는 코로나19 사망과 관련된 다양한 독립변수들(보건분야 예산액, 나이, 인구밀도 등)을 사용하여 시군구별 코로나19 사망자 수 예측모형을 개발하기 위해 여러 독립변수를 사용하는 다중회귀모형을 적용하였다.

종합적인 다양한 분석 실험 결과, 본 논문에서 도출한 시군구별 코로나19 사망자 수 예측모형은 다음의 표 2와 같이 주어진다. 본 논문에서 도출된 위험 요인은 나이, 인구밀도, 1인 대비 의료시설 현황이다. 제안된 선형회귀 모형에 대한 검정통계량에서 결정계수(R-squared)는 0.749, adjusted 결정계수는 0.737, F-검정통계량(F-statistic)은 60.34로 높게 나왔으며, 이는 제안된 모형이 통계적으로 적절하였음을 의미한다.

표 2. 코로나19 사망자 수 예측모형 결과
Table 2. Result of Covid-19 death predictin Model development

Proposed linear regression model
$\ln(Y) = 0.1174(X_1) + 0.1875(X_2) - 0.713(X_3)$

주) : Y=시군구별 코로나19 사망자(명)
X_1=나이
X_2=인구밀도(명/㎢)
X_3=1인 대비 의료시설 현황

III. 결 론

본 논문은 다중 선형회귀분석 기법을 적용하여 시군구별 코로나19 사망자 수 예측모형을 제안하였으며, 제안된 모델에 대하여 선형회귀의 통계적 관점에서 신뢰성 있는 검정통계량을 도출하였다. 또한, 코로나19 사망에 영향을 미치는 유의미한 변수들을 발견하였다.

참 고 문 헌

[1] 류종수. "코로나19 팬데믹에 대한 국가별 대응 및 보건정책 고찰 및 비교." 국내박사학위논문 연세대학교 일반대학원, 2021. 서울

[2] 조용탁. "체질량지수가 국내 코로나19 사망률에 미치는 영향." 국내석사학위논문 한양대학교, 2021. 서울