로지스틱 회귀분석을 이용한 우울증과 경제 활동 간의 상관관계 분석

조성운[1], 김규일[2], 이서엘[3], 권혜정[4], 정경용[5*]

*경기대학교[1,2,3,4,5]

jodory9524@gmail.com[1], ran6491@kyonggi.ac.kr[2], lse_1031@kyonggi.ac.kr[3],
rnjsgpwjd121@kyonggi.ac.kr[4], *dragonhci@gmail.com[5]

A Correlation Analysis between Depression and Economic Activity Using Logistic Regression

Seong-Un Cho [1], Kyu-Il Kim [2], Seo-El Lee[3], Hye-Jeong Kwon[4], Kyungyong Chung[5*]

*Kyonggi Univ[1,2,3,4,5]

요 약

최근 우울증과 불안장애에 대한 진단이 꾸준하게 증가하고 있다. 이러한 사회적인 요인 중 하나로 감소하고 있는 취업률과 우울증과의 상관관계를 분석한다. 본 논문에서는 우울증과 경제 활동 상태 간에 상관관계를 분석하기 위해 질병관리청에서 제공하는 국민건강영양조사 데이터를 전처리하여 사용한다. 전처리를 마친 데이터를 활용하여 우울증 여부를 종속변수로 설정하고 그 외의 경제 활동 관련 데이터들을 독립변수로 설정하여 로지스틱 회귀분석을 진행한다. 로지스틱 회귀분석과 각 독립변수들의 오즈비를 활용해 우울증과의 상관관계를 파악할 수 있다. 이를 통해 경제 활동 상태 정보 중 어떤 요인이 우울증에 큰 영향을 주는지 파악 가능하며 이 모델을 통해 개인의 경제 활동 정보로부터 우울증이 발병할지 예측 가능하며 예방 가능하다.

Ⅰ. 서 론

최근 우울증과 불안장애에 대한 진단이 꾸준하게 증가하고 있다. 건강보험심사평가원의 자료에 따르면 우울증 및 불안장애 환자 수가 2021년 기준 93만 3,481명으로 5년 전인 2017년 비해 35.1%가 증가했다[1]. 이러한 현상의 이유로 정부 기관과 언론매체에서는 2019년에 도래한 코로나 바이러스의 확진 휴유증이 주된 이유라고 얘기한다. 하지만 해당 바이러스의 종식을 앞두고 있는 현재까지 우울증과 불안장애에 대한 진단이 지속적으로 증가하고 있다. 또한 최근 영국 통계청 자료에 따르면 해당 바이러스에 대한 코로나의 휴유증 기간이 확진일로부터 최대 6개월이라는 것을 미루어 볼 때 지속적으로 증가하는 우울증의 요인이 코로나 바이러스 외에 다른 요인이 존재한다고 판단 할 수 있다.

본 논문에서는 지속적으로 증가하는 우울증에 대한 진단이 취업률, 취업 형태, 근무시간 등의 요인과 상관 관계가 있을 것으로 예측한다, 이에 대한 데이터와 우울증 데이터 셋은 질병관리청에서 제공하는 원시자료를 사용한다. 그 후 해당 원시데이터를 전처리 과정을 통해 유의미한 데이터로 추출한다. 이렇게 우울증과 상관 관계가 있을 것이라고 생각한 유의미한 데이터들을 독립변수로 설정하고 종속변수를 우울증 진단 여부로 설정한다. 이와 같이 변수를 설정 한 후 로지스틱 회귀(Logistic Regression)분석을 통해 오즈(Odds)와 오즈비(Odds Ratio; OR)를 구하여 각각의 독립변수들이 종속변수에 미치는 영향을 파악 할 수 있다. 이를 통해 우울증과 경제 활동 간에 상관관계를 분석한다. 이러한 과정을 통해 우울증과 취업률, 취업 형태, 근무 시간과의 상관관계를 분석하였고 이 모델을 통해 개인의 경제활동 정보로부터 우울증이 발병할지 예측 가능하며 예방 가능한 방법을 제안한다.

Ⅱ. 관련 연구

2.1 로지스틱 회귀분석

로지스틱 회귀는 대표적인 지도 학습 방식 중 하나이다. 해당 기술은 독립변수의 선형결합을 사용하여 0부터 1 사이의 범주에 데이터가 속할 확률을 예측하는 수학적 기법이다[2]. 로지스틱 회귀 분석은 직접적인 값을 예측하는 선형 분석과는 다르게 종속변수가 특정 범주에 속할 확률을 이용해 두 개의 클래스 중 한 개의 클래스로 예측할 수 있다. 이 때 해당 확률에 따라 분류 기준 값(cut-off)을 적용하여 더 높은 가능성이 있는 범주에 속하는 것으로 분류하는 방법이다. 이러한 분석 방법은 원시적이고 기본적인 방법이지만 여전히 여러 분야에서 분류 및 예측을 위해 사용되고 있으며 예측하고자 하는 데이터가 두 개의 클래스를 갖는 범주형 데이터일 때 사용할 수 있다.

2.2 오즈와 오즈비

로지스틱 회귀분석에서 종속변수를 확률로 표현하여 사용할 때 올바른 추정결과를 얻기 어렵다는 문제가 있다. 이를 해결하기 위해 오즈를 사용한다. 오즈는 특정한 사건이 발생하지 않을 확률 대비 해당 사건이 발생할 확률을 의미한다. 식(1)은 오즈의 공식을 나타낸다[3].

$$Odds = \frac{(\text{사건발생확률})}{(\text{사건미발생확률})} = \frac{p}{1-p} \qquad \text{식 (1)}$$

위의 공식으로 구한 오즈는 0부터 무한대까지의 구간을 갖는데, 이것은 여전히 비선형적 관계이므로 적절한 모델을 얻기가 어렵다는 문제가 있다. 이를 해결하기 위해 오즈비를 사용한다. 오즈비란 특정 원인에 따른 사건 발생 확률을 비교할 때 사용되는 척도이다. 말 그대로 서로 다른 오

즈들 간에 비율을 의미한다. 이렇게 계산하게 되면 0 또는 1인 두 개의 종속변수 범주 중 어떤 특정한 범주에 속할 확률을 파악하기 쉬워진다. 오즈비의 결과 값이 1보다 크다면 독립변수가 증가하는 방향으로 종속변수에 영향을 주고 1보다 작다면 감소하는 방향으로 영향을 준다. 또한 오즈비는 1과 오즈비의 결과 값 사이의 거리로 독립변수와 종속변수간에 관계를 파악할 수 있다. 오즈비의 결과 값이 1을 기준으로 멀리 있을수록 독립변수와 종속변수간의 관계가 강하고 영향을 주는 강도가 크다는 것을 의미한다.

Ⅲ. 본 론

3.1 데이터 수집 및 전처리

본 논문에서는 우울증과 경제활동 상태 간에 상관관계를 분석하기 위해 질병관리청에서 제공하는 국민건강영양조사 원시자료를 사용한다[4]. 하지만 원시자료에 데이터에는 불필요한 데이터가 많다. 만약 이러한 데이터를 사용하여 로지스틱 회귀분석을 하게 되면 결측 값들이 많아 적절하지 않은 모델의 결과가 나오게 된다[5]. 이를 방지하기 위해 데이터 전처리 과정은 필수불가결하다. 이에 따라 유의미한 데이터를 추출하기 위해 불필요한 데이터 삭제, 데이터 변환, K-means 클러스터링을 활용한 이상값 처리, 데이터 표준화 과정 등을 통해 데이터 수집 및 전처리를 한다. 그 후 변수 5개와 21,462개의 데이터를 사용한다. 본 논문에는 우울증 진단 여부를 종속변수로 설정하고, 경제 활동 여부, 취업형태(정규직, 비정규직(임시직, 일용직)), 근로 시간을 독립변수로 설정하였다. 표 1은 설정한 변수들에 대한 설명이다.

표1. 데이터 프레임

종속변수	**변수 설명**
DF2_dg	우울증 진단 여부
독립변수	**변수 설명**
EC1_1	경제 활동 여부
EC_wh1	정규직 여부
EC_wh1	비정규직(임시직, 일용직) 여부
EC_wht_23	근로 시간

3.2 로지스틱 회귀를 이용한 상관관계 분석 및 예측 모델

전처리한 데이터를 활용해 우울증과 경제활동의 상관관계를 분석하기 위해 로지스틱 회귀분석을 진행한다. 표 1처럼 종속변수와 독립변수를 설정한 후 분석을 진행한다. 표 2는 각 독립변수들의 회귀계수(Regression Coefficient)와 오즈비를 나타낸다.

표 2. 독립변수들의 회귀계수와 오즈비의 결과 값

독립변수명	**회귀계수**	**오즈비**
EC1_1	-2.0780	0.125186
EC_wh1	.-1.1193	0.326509
EC_wh1	-0.9587	0.383407
EC_wht_23	-0.0091	0.990904

각 독립변수들의 회귀계수가 종속변수인 우울증 여부에 어떤 영향을 주는지 분석할 수 있다. 회귀계수의 값이 양수면 우울증 여부가 '1'일 확률이 높다는 뜻이고 반대로 음수면 우울증 여부가 '0'일 확률이 높아진다는 뜻으로 분석할 수 있다. 그러므로 각각의 독립변수들의 회귀계수의 값이 음수이므로 모든 독립변수들은 해당 값이 낮을수록 우울증일 확률이 높다. 예를 들어 표 2에서 독립변수인 EC1_1(경제 활동 여부)은 음의 회귀계수를 갖는다. 그러므로 경제 활동 여부가 낮을수록 우울증일 확률이 높다고 분석할 수 있다. 하지만 회귀계수의 값 자체로는 각 독립변수들이 종속변수들에게 주는 영향의 크기를 알 수 없기 때문에 오즈비를 구하여 확인한다. 오즈비는 각 독립변수의 변화에 따른 종속변수 확률의 변화율에 따라 독립변수와 종속변수의 상관관계를 파악한다. 오즈비를 통해 분석하면 종속변수에 영향을 주는 방향뿐만 아니라 크기까지 확인 할 수 있다. 오즈비의 값이 1을 기준으로 크고 작음에 따라 영향을 주는 방향을 파악할 수 있고 1을 기준으로 가까이 있는지 멀리 있는지에 따라 영향의 크기를 파악할 수 있다. 표 2에서는 독립변수인 EC1_1(경제 활동 여부)의 값이 1에서 가장 멀리 있으므로 종속변수인 우울증에 가장 큰 영향을 주는 변수라는 것을 파악할 수 있다. 또한 EC_wht_23(근로 시간)의 값은 1과 매우 가까우므로 근로 시간은 우울증과의 관계가 거의 없다고 분석할 수 있다. 이를 통해 오즈비를 통해 우울증에 영향을 주는 여러 요인들을 확인할 수 있다. 이와 같은 분석을 통해 우울증이 발병할 확률이 경제 활동을 하지 않는 사람들과 비정규직인 사람들에게 더 높다는 것이 확인되었으며 근무 시간과 우울증 사이에는 서로 관계가 없다고 분석할 수 있다.

Ⅳ. 결 론

본 논문에서는 질병관리청에서 제공하는 국민건강영양조사 데이터를 활용하여 우울증과 경제 활동 간의 상관관계를 분석하였다. 상관관계를 분석하는 방법으로는 먼저 데이터의 수집과 전처리를 통해 우울증과 경제활동 정보로 사용할 유의미한 데이터를 추출하는 과정을 진행한다. 그 후에 로지스틱 회귀분석을 통해 회귀계수와 오즈비 값을 구해 각 독립변수(경제 활동 변수들)와 종속변수(우울증)의 관계를 분석한다. 이를 통해 근무 시간과 우울증과는 상관관계가 없다는 것을 파악할 수 있었고 우울증에 걸릴 확률은 경제 활동을 하지 않는 사람일수록 높다는 것을 확인할 수 있다. 또한 취업 형태가 비정규직인 사람들이 정규직인 사람들보다 우울증에 걸릴 가능성이 조금 더 높다고 분석할 수 있다. 향후 연구에서는 이번 연구를 통해 알게 된 정보를 기반으로 경제 활동을 하지 않는 사람들 중에서 어떤 요인이 우울증과의 관계를 갖는지에 대해 연구할 예정이다.

ACKNOWLEDGMENT

This work was supported by the GRRC program of Gyeonggi province. [GRRC KGU 2020-B03, Industry Statistics and Data Mining Research]

참 고 문 헌

[1] Health Insurance Review and Assessment Service, 2022, (https://www.hira.or.kr/")

[2] Connelly, L. "Logistic regression", Medsurg Nursing, 29(5), 353-354, 2020.

[3] Wald, N. J., & Old, R. "The illusion of polygenic disease risk predictio,". Genetics in Medicine, 21(8), 1705-1707, 2019.

[4] The seventh Korea National Health and Nutrition Examination Survey (KNHANES Ⅶ-3), 2018, Korea Centers for Disease Control and Prevention

[5] Moh, Y., Han, S., Yeon, K., Kang, H. "A study on variable importance measures in multiple regression analysis", Journal of the Korean Data Analysis Society, 17(6B), 2981-2990, 2015.