# Project 4 :
# Consumption of Cigarette Prediction and Analysis

2014122006 허승민

2019년 5월 30일

## 1 분석 개요

주어진 데이터는 National Insurance Organizaition 에서 담배판매량의 패턴을 분석하고자 한 데이터이다. 데이터 분석의 최종 목적은 담배 판매량을 설명할 수 있는 회귀모형을 찾고 관련 현상의 분석, 예측이다.

## 2 데이터 구성

| Variable | Definition |
|---|---|
| Age | Median age of a person living in a state |
| HS | Percentage of people over 25 years of age in a state who had completed high school |
| Income | Per capita personal income for a state(income in dollars) |
| Black | Percentage of blacks living in a state |
| Female | Percentage of females living in a state |
| Price | Weighted average price(in cents) of a pack of cigarettes in a state |
| Sales | Number of packs of cigarettes sold in a state on a per capita basis |

표 1 : 변수 설명

데이터에서 설명하고자 하는 종속변수는 Sales 이며, 설명변수는 Age, HS, Income, Black, Female, Price로 사회, 경제적 요소가 반영되어있다. 기본적으로 제공되는 변수가 상당히 제한적이므로 새로운 파생변수를 고려하고자 하였다. 판매량 자체가 경제적인 요소와 관련이 있으므로 경제적인 변수를 추가하였다.

- 구매력지수(소득으로 구매할 수 있는 담배의 개수, IPP) = Income / Price

이외에도 사회적인 요소를 반영하고자 하였으나 사회적인 요소(교육, 인종, 성별 등)의 경우 당시 차별적인 요소( 흑인과 여성은 고등교육 비율이 음의 상관관계)가 있지만 주요하게 만들 요소가 없어 추가할 수 없었다.

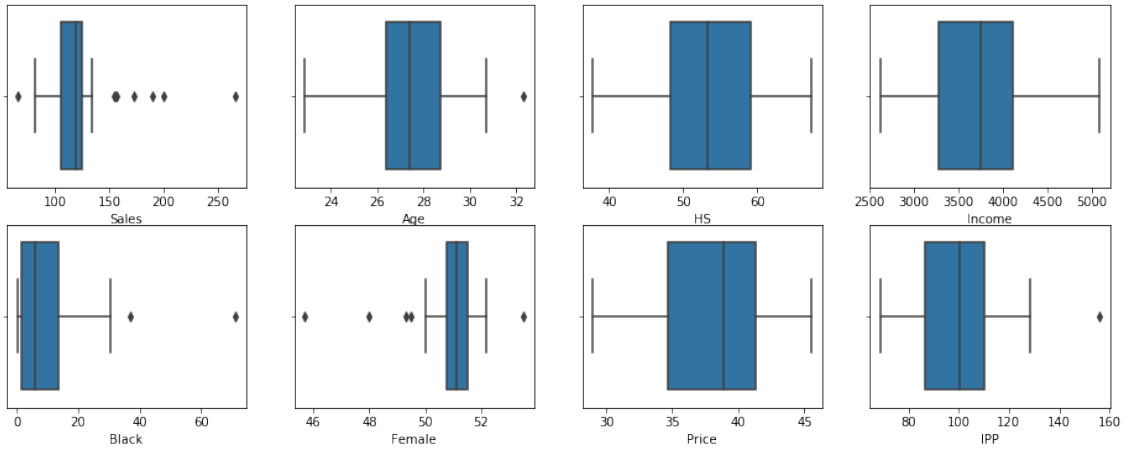
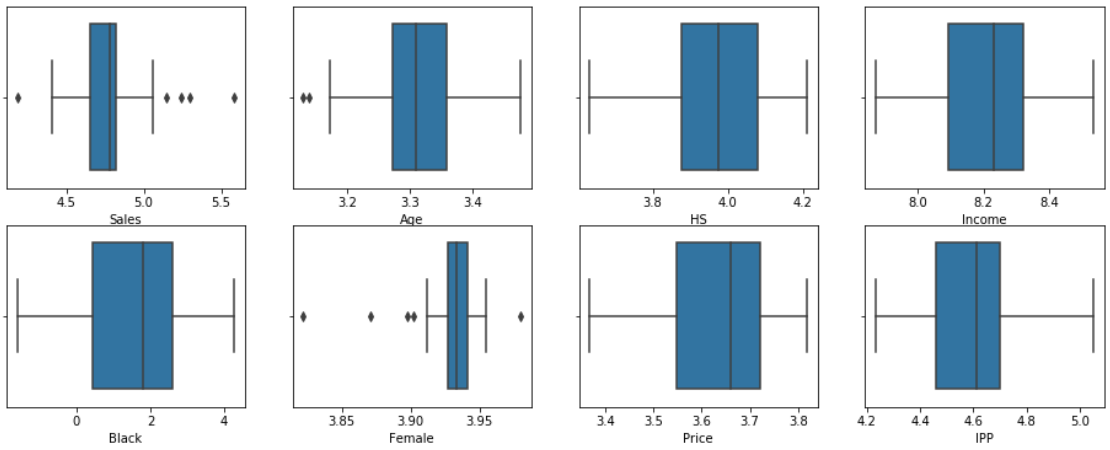



그림1 : 변수들의 분포 형태

기본적인 데이터를 살펴보면, 전체적으로 분포가 고르지 못한 경우가 많고 이상치로 추정되는 값들이 있는 것으로 보인다. 따라서 조금 더 좋은 성질의 변수로 사용하기 위해 모두 로그 변환을 실시하여 보고 상관계수를 확인해보았다.




| | Sales | Age | HS | Income | Black | Female | Price | IPP |
|---|---|---|---|---|---|---|---|---|
| **Sales** | 1 | 0.31 | 0.04 | 0.38 | 0.13 | 0.16 | -0.31 | 0.54 |
| **Age** | 0.31 | 1 | -0.07 | 0.27 | 0.12 | 0.56 | 0.21 | 0.1 |
| **HS** | 0.04 | -0.07 | 1 | 0.58 | -0.58 | -0.38 | 0.08 | 0.47 |
| **Income** | 0.38 | 0.27 | 0.58 | 1 | 0.09 | -0.09 | 0.19 | 0.78 |
| **Black** | 0.13 | 0.12 | -0.58 | 0.09 | 1 | 0.4 | 0.05 | 0.04 |
| **Female** | 0.16 | 0.56 | -0.38 | -0.09 | 0.4 | 1 | 0.01 | -0.09 |
| **Price** | -0.31 | 0.21 | 0.08 | 0.19 | 0.05 | 0.01 | 1 | -0.47 |
| **IPP** | 0.54 | 0.1 | 0.47 | 0.78 | 0.04 | -0.09 | -0.47 | 1 |

그림 2 : 로그변환 후의 변수들의 형태와 상관계수 행렬

로그 변환한 후 결과를 보면 비대칭의 문제는 상당히 해소 되었지만 종속변수인 Sales와 설명변수 Female의 이상치 문제는 전혀 변함이 없는 결과물이다. 이는 변수의 분산이 매우 작아 이상치들이 중심으로부터 매우 떨어지게 되어서 발생한 현상이다. 이를 해결하기 위한 마땅한 방법이 없으므로

그대로 사용하여 모델을 사용하고, 모델의 문제를 일으키는 경우 삭제하는 것을 고려하도록 한다. 최종적으로는 모두 로그 변환한 변수를 사용하도록 한다.

# 3 모델링

모델링을 함에 있어 가장 주안점을 두고 바라보는 부분은 이상치의 작용 형태를 파악하면서 적절한 회귀모형을 찾아내는 것이다. 따라서 가장 기본적으로 모든 변수를 이용하여 회귀모형 탐색을 실시하였다. 기본 선형회귀를 실시하되, 상관계수 행렬을 보면 변수간의 상관 관계가 높은 조합의 변수들이 보이므로 다중공선성에 대한 대안을 제시하는 LASSO, Ridge 모형도 함께 사용하여 확인하도록 한다.

| 모형 | 변수 | $R^2$ |
|---|---|---|
| 선형회귀 | Age, HS, Income, Black, Female, Price, IPP | 0.405 |
| LASSO | Age, HS, Income, Black, Female, Price, IPP | 0.394 |
| Ridge | Age, HS, Income, Black, Female, Price, IPP | 0.398 |

표 2 : 이상치 제거 후 회귀 모형의 탐색.
선형회귀의 경우 유의미하지 않은 변수는 빨간색, LASSO의 경우 지워지는 변수를 빨간색으로 표시

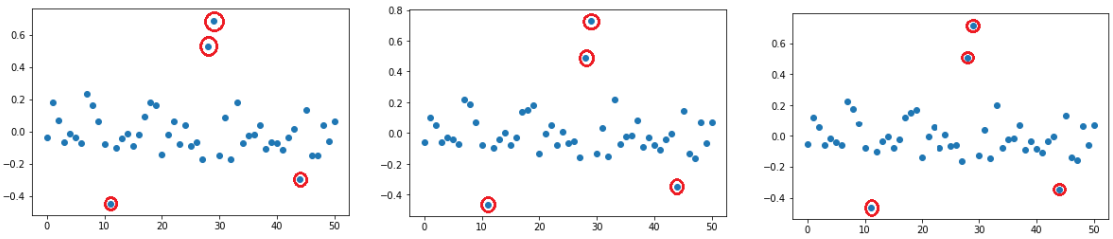


그림 3 : 기초 탐색한 모형들의 잔차도. 좌측부터 선형회귀, LASSO, Ridge.

아주 간단하게 탐색을 실시해본 결과 전체적으로 $R^2$의 값이 0.4를 간신히 넘기고 있다. 이를 해석할 때 0.4가 무조건적으로 낮은 수치라고는 할 수 없다. 하지만 변수탐색 결과 이상치가 있었고, 잔차도에도 이상치가 있으므로 이상치의 영향이라는 가정을 할 수 있다. 이상치 제거 후 회귀모형을 다시 확인해보도록 한다. 이상치 제거의 기준은 잔차의 절대값이 log(Sales)의 표준편차보다 큰 경우로 정의하고 제거하도록 하였다.

| 모형 | 변수 | $R^2$ |
|---|---|---|
| 선형회귀 | Age, HS, Income, Black, Female, Price, IPP | 0.670 |
| LASSO | Age, HS, Income, Black, Female, Price, IPP | 0.661 |
| Ridge | Age, HS, Income, Black, Female, Price, IPP | 0.660 |
| 후진제거 | HS, Income, Price, IPP | 0.648 |

표 3 : 이상치 제거 후 회귀 모형의 탐색.
선형회귀의 경우 유의미하지 않은 변수는 빨간색, LASSO의 경우 지워지는 변수를 빨간색으로 표시

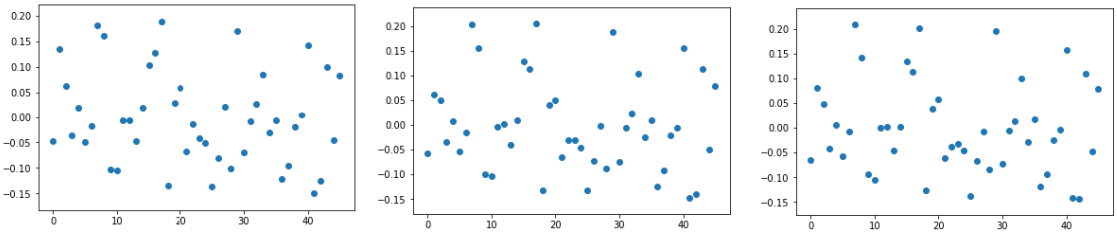


그림 4 : 추가 탐색한 모형들의 잔차도. 좌측부터 선형회귀, LASSO, Ridge.

이상치를 제거하고 탐색한 결과 $R^2$의 값이 상당히 커졌다는 것을 알 수 있다. 따라서 어느 정도의 설명력을 갖추게 된 모형이라고 할 수 있다. 하지만 LASSO의 결과를 보면 변수 중 Income과 Price가 삭제된 것을 알 수 있다. 그 이유는 IPP 변수가 만들어진 공식이 Income/Price 이므로 상관계수가 높을 수 밖에 없는 관계였다. 또한 상관계수 행렬 상 서로 상관관계가 높은 관계들이 많았으므로 다중공선성에 대한 진단을 하고 넘어가도록 한다.

| 모형 | 변수 | 특징 |
|---|---|---|
| 주성분분석 | $PC_1, \cdots, PC_7$ | $\lambda_1 = 109, \lambda_2 = 89, \lambda_3 = 66, \lambda_4 = 39, \lambda_5 = 14, \lambda_6 = 5, \lambda_7 = 0$ |
| PC Regression | $PC_1, \cdots, PC_6$ | $R^2 = 0.67$ |

표 4 : 다중공선성 진단을 위한 주성분 분석 및 고유값 계산 후 주성분 회귀분석 실시

선형회귀의 경우 유의미하지 않은 변수는 빨간색, LASSO의 경우 지워지는 변수를 빨간색으로 표시

다중공선성 여부를 확인하기 위해서 주성분 분석을 통해 고유값을 계산하였다. 결과적으로 완전한 다중공선성을 나타내는 $\lambda_7 = 0$ 이 등장하므로7번째 주성분은 제거하도록 한다. 다음으로 작은 5 고유값의 경우는 작은 값이 아니라는 판단 하에 6번째 주성분까지를 설명변수로 하여 회귀모형을 적합시켜 결과를 도출하였다. 결론적으로 다중공선성이 없는 좋은 모형이 된 것으로 보인다.

| 모형 | 변수 | $R^2$ |
|---|---|---|
| 선형회귀 (Full) | Age, HS, Income, Black, Female, Price, IPP | 0.670 |
| LASSO($\alpha = 0.0006$) | Age, HS, Income, Black, Female, Price, IPP | 0.661 |
| Ridge($\alpha = 0.1$) | Age, HS, Income, Black, Female, Price, IPP | 0.660 |
| 후진제거 | HS, Income, Price, IPP | 0.648 |
| PC Regression | $PC_1, \cdots, PC_6$ | 0.670 |

표 5 : 최종 모형 비교, $\alpha$의 경우 Cross Validation을 통해 찾아내었다.

최종적으로 모형 및 변수선택을 위한 비교를 하도록 한다. 주어진 데이터는 생각보다 다중공선성의 문제가 있는 데이터임을 파악할 수 있었고 그에 따른 진단이 필요하였다. 기본적인 선형회귀 Full model 과 후진제거법을 보면, $R^2$의 관점에서는 문제가 크게 없지만 여전히 다중공선성의 문제가 지속적으로 발생한다.(고유값에서 0이 지속적으로 등장.) 따라서 변수 선택을 실시하더라도 보다 근본적인 다중공선성의 해결방법이 필요하므로 제외하도록 한다. 나머지 3가지 방법을 선택해야하는데, 지표로서 사용되는 $R^2$기준으로 본다면 PC Regression이 가장 좋은 방식이라고 할 수 있다. 단순히 예측이 목적이라면 PC Regression이 좋은 방법인 것은 틀림 없지만 이 데이터는 해석이

필요하다고 할 수 있다. 주성분의 경우 해석이 어려운 부분이 있지만 LASSO의 경우 적절한 변수 제거과정과 변수자체의 해석이 가능한 장점이 있다. 또한 LASSO는 상관계수 행렬에서 상관관계가 높았던 Income과 Price 변수를 삭제함과 동시에 페널티를 부여함으로서 다중 공선성을 해결하였다. 따라서 결론적으로는 LASSO 방식을 택하는 것이 좋은 방법이라고 생각할 수 있다.

# 4 결론

$$log(Sales) = 2.45 + 0.16log(Age) - 0.33log(HS) - 0.004log(Black) - 0.09log(Female) + 0.75log(IPP)$$

위의 결과가 LASSO에 의한 최종모형이다. 변수 선택과정에서 기준은 다중공선성의 해결과 $R^2$에 따른 선택이었다. 다른 통계량을 계산하지 않은 것은 이 모형은 분석 뿐만 아니라 예측을 해야할 필요가 있는 데이터이기 때문이다. 또한 다중공선성의 문제를 사전에 예상할 수 있지만 IPP 변수를 제거하지 않은 이유가 IPP의 상관계수가 높아 변수의 중요도가 높을 것으로 예상되었기 때문이다. 결과적으로 IPP변수는 영향력이 큰 변수임을 확인할 수 있다. Age와 IPP 변수는 판매량 증가에 영향을 주는, 그리고 고교졸업율, 흑인비율, 여성비율은 판매량이 감소하는 역할을 하지만 흑인, 여성과 같은 변수는 영향력이 미미함을 확인할 수 있다. 담배 판매량은 사회적인 영향보다는 경제적인 영향이 더 강하다는 사실을 알 수 있었다.

| State | Predicted | Sales | Age | HS | Income | Black | Female | Price | IPP |
|---|---|---|---|---|---|---|---|---|---|
| DE | 123.04 | 155.0 | 26.8 | 54.6 | 4524.0 | 14.3 | 51.3 | 41.3 | 109.54 |
| HI | 132.47 | 82.1 | 25.0 | 61.9 | 4623.0 | 1.0 | 48.0 | 36.7 | 125.97 |
| NV | 111.75 | 189.5 | 27.8 | 65.2 | 4563.0 | 5.7 | 49.3 | 44.0 | 103.70 |
| NH | 125.90 | 265.7 | 28.0 | 57.6 | 3737.0 | 0.3 | 51.1 | 34.1 | 109.59 |
| UT | 97.60 | 65.5 | 23.1 | 67.3 | 3227.0 | 0.6 | 50.6 | 36.6 | 88.17 |

표 6 : 제거된 이상치 판단

마지막으로, 제거된 이상치에 살펴보도록 한다. 종속변수인 Sales와 최종모형을 통해 예측한 Predicted 값을 비교해보면 대부분 예측보다 Sales가 매우 크거나 작게 나온 것을 확인할 수 있다. 이 5개 주의 특성을 살펴보면 대부분 서비스산업이 발달하여 호텔산업 등이 발달하였고, 미 공군, 해군 기지들이 주로 위치한 지역이다. 따라서 담배판매량이 많은 이유가 담배를 소비하는 대상이 몰리게 되어 판매량이 다른 주들의 경향성에 비해 높게 나타난 것으로 보인다. 하지만 하와이의 경우 미국 내에서도 가장 각력한 금연법을 시행하는 주이며 유타주의 경우 종교적 신념이 강해 미국 내에서 흡연율이 가장 낮은 지역이라고 한다. 따라서 일반적인 근거로 해석한다면 서비스산업이 발달하고 군 관련 시설이 있으면 더 높은 경향성을 띠지만 사회적, 정부의 영향에 의해 흡연율이 낮아져 판매량이 낮아지는 경우도 있음을 확인할 수 있다.