Analysis of Data from USA Region Using Additive Regression with Hilbertian Responses

기도형[1], 유명훈[1], 이경원[2], and 이동현[2]

[1]*Department of Mathmatics, Seoul National University*

[2]*Department of Statistics, Seoul National University*

December 15, 2019

1 Introduction

우리는 비모수함수추정론 강의를 통해 smooth backfitting 알고리즘을 사용한 additive mean regression model의 적합, 그리고 이 알고리즘의 convergence를 포함한 asymptotic property에 대해 학습하였다. 또한, mean을 추정하고자 하는 반응변수(response)가 Euclidean인 경우에서 더 나아가 Hilbertian인 경우에 대해서도 다루었다.

Simplex 데이터는 Euclidean 형태이나 모든 원소의 합이 1이라는 조건을 만족하는 데이터이다. 이러한 데이터는 주로 3개 이상의 그룹으로 나누어지는 비율의 형태로서, 실존하는 데이터 형태 중에서 non-Euclidean이면서 Hilbertian인 대표적인 형태이다. Hilbertian response에 대해 처음으로 smooth backfitting을 통한 additive regression model을 제시한 논문 Jeon and Park (2018)에서도 simplex 데이터가 대표적인 실제 데이터 예시로써 활용되었다.

이 보고서에서는 논문에서 제시된 코드를 사용하여 미국 지역별 인종 비율 데이터를 분석해 본다. 미국은 다양한 인종이 살고 있는 나라이고, 각 주마다 기온, 강수량 등의 환경적인 요소들과 평균 소득 등의 생활 관련 요소들이 서로 다르다. 이러한 요소들이 인종 비율에 어떻게 연관되는지 파악하기 위해 additive model을 적용하고, 각 설명변수의 변화에 따라 인종 비율이 어떻게 변하는지를 시각적으로 표현해 본다.

2 Data

분석에 이용한 자료는 다양한 출처를 통해 수집하였으며, 모든 자료는 미국 각 주(state)별 관측치 50개와 워싱턴 D.C.의 관측치, 총 51개의 관측치로 이루어져 있다. 각 자료에 대한 설명은 다음과 같다.

먼저, 반응변수로 선정한 각 주별 인종 분포는 미국 인구조사국(Census Bureau)에서 제공하는 State Population by Characteristics: 2010-2018 자료[1]를 바탕으로 계산하였다. 자료는 2010년부터 2018년까지 각 주별 특성별(성별, 인종, 나이 등) 인구수로 구성되어 있다. 자료에서는 인종을 (1) 백인과 그 혼혈, (2) 흑인과 그 혼혈, (3) 아시안과 그 혼혈, (4) 아메리카 원주민과 알래스카 원주민, 그리고 그들의 혼혈, (5) 하와이 원주민과 그 혼혈로 나누어 조사하였으며, 한 사람이 여러 인종의 혼혈인 경우, 그 사람을 해당되는 모든 항목에서 셈하고 있다. 여기서는 2017년 기준 자료를 사용하였으며, 분석과 시각화의 편의를 위하여 가장 많은 비율을 차지하는 백인과 그 혼혈(Y_1), 흑인과 그 혼혈(Y_2), 아시안과 그 혼혈(Y_3)의 비율을 반응변수 $\mathbf{Y} = (Y_1, Y_2, Y_3) \in S_1^3$로 사용하였다.

위의 반응변수를 설명하기 위한 변수로는 각 주별 연령, 소득, 범죄율, 그리고 기후 관련 변수를 찾았다. 각 주별 연령(X_1)은 인종 분포 데이터와 마찬가지로 State Population by Characteristics: 2010-2018 자료를 통해 계산한 2017년 기준 주별 연령 분포의 중앙값을 사용하였다. 각 주별 소득(X_2)은 2013-2017년 American Community Survey 자료를 통해 나온 것으로, 2017년 기준 주별 가구당 소득(household income) 분포의 중앙값이다. 이는 미국에서 소득 수준을 나타내는 대표적인 지표 중 하나이다.

각 주별 범죄율(X_3)은 FBI에서 제공하는 Uniform Crime Reporting Program 2017-2018의 보고서[2]를 바탕으로 계산하였다. 자료는 2017년과 2018년의 강력범죄(violent crime)과 재산 범죄(property crime)의 통계를 각 주별로 나타내고 있는데, 여기서 우리는 강력범죄 통계만을 사용하여 각 주의 범죄율을 계산하였다. 마지막으로는 기후 관련 변수로, 다른 자료들과 달리 연간 편차가 심하다는 점을 고려하였고, 결과적으로 1981년부터 2010년까지의 각 주 별 평균 기온 자료(X_4)와 평균 강수량 자료(X_5)를 사용하였다.

각 주별 연령($X1$) 또한 State Population by Characteristics: 2010-2018 자료를 통해 계산한 2017년 기준 주별 연령 분포의 중앙값을 사용하였다.
각 주별 범죄율은 FBI에서 제공하는 Uniform Crime Reporting Program 2017-2018의 보고서[3]를 바탕으로 계산하였다. 자료는 2017년과 2018년의 강력범죄(violent crime)과 재산 범죄(property crime)의 통계를 각 주별로 나타내고 있는데, 여기서 우리는 강력범죄 통계만을 사용하여 각 주의 범죄율을 계산하였다.

각 주별 소득 자료는 2013-2017년 American Community Survey를 통해 나온 2017년 주별 가구당 소득(household income) 중앙값으로, 이는 미국에서 소득 수준을 나타내는 대표적인 지표 중

[1] https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-detail.html
[2] https://ucr.fbi.gov/crime-in-the-u.s/2018/crime-in-the-u.s.-2018/tables/table-4
[3] https://ucr.fbi.gov/crime-in-the-u.s/2018/crime-in-the-u.s.-2018/tables/table-4

하나이다.

3 Compositional Data

3.1 Aitchison geometry

다음과 같은 p차원 simplex S_k^p를 생각하자

$$S_k^p = \left\{ \mathbf{x} = (x_1, \cdots, x_p) \in \mathbb{R}^p : x_i \geq 0,\ i = 1, 2, \cdots, p,\ \sum_{i=1}^{p} x_i = k \right\}.$$

분석해야 할 자료에서 반응변수 $Y = (Y_1, Y_2, Y_3)$는 유클리드 공간 상의 자료가 아닌 3차원 simplex S_1^3 위의 자료이다. 이와 같은 자료를 compositional data라 부른다. p차원 simplex S_1^p 위에 다음과 같이 연산을 정의하면 이 벡터공간은 Hilbert space가 됨이 알려져 있으며 (Aitchison, 1982) 이러한 공간을 Aitchison geometry 또는 Aitchison simplex라 한다

$$\begin{aligned}
x \oplus y &= \left(\frac{x_1 y_1}{\sum x_i y_i}, \cdots, \frac{x_p y_p}{\sum x_i y_i} \right), \\
\alpha \odot y &= \left(\frac{x_1^\alpha}{\sum x_i^\alpha}, \cdots, \frac{x_p^\alpha}{\sum x_i^\alpha} \right), \\
\langle x, y \rangle &= \frac{1}{2D} \sum_i \sum_j \log \frac{x_i}{x_j} \log \frac{y_i}{y_j}.
\end{aligned}$$

3.2 Methods

이 절에서는 compositional data를 분석하기 위해 제안된 모형들에 대해 간략히 소개한다. 먼저, 다음과 같은 *Multinomial logistic regression*

$$\begin{gathered}
\mathbb{E}[\mathbf{Y}|\mathbf{X} = \mathbf{x}] = (\alpha_1(\mathbf{x}), \alpha_2(\mathbf{x}), \cdots, \alpha_p(\mathbf{x})), \\
\alpha_1(\mathbf{x}) = \frac{1}{1 + \sum_{l=2}^{p} \exp(\beta_{l0} + \boldsymbol{\beta}_l^T \mathbf{x})},\ \alpha_j(\mathbf{x}) = \frac{\exp(\beta_{j0} + \boldsymbol{\beta}_j^T)}{1 + \sum_{l=2}^{p} \exp(\beta_{l0} + \boldsymbol{\beta}_l^T \mathbf{x})},\ \text{for } j \geq 2
\end{gathered} \tag{1}$$

을 생각해볼 수 있다. 여기서 모수 β_{j0}, $\boldsymbol{\beta}_j$의 추정으로 최소제곱 추정(least squares)이나 다음과 같이 정의되는 Kullback-Leibler divergence를 최소화하는 값을 사용할 수 있다.

$$\mathrm{KL} = \sum_{i=1}^{p} \sum_{j=1}^{p} Y_{ij} \log(Y_{ij} / \alpha_j(\mathbf{X}_i)).$$

각 모형은 R의 `Compositional` 패키지의 `ols.compreg`, `kl.compreg`를 이용해 적합할 수 있다. 해당 패키지에서는 Jensen-Shannon divergence, symmetric Kullback-Leibler divergence 등을 최

소화하는 모형을 적합하는 함수들 또한 제공하고 있다.

$\mathbf{X}$가 주어져 있을 때, $\mathbf{Y}$의 조건부 분포를 디리클레(Dirichlet) 분포로 가정한 *Dirichlet regression model*

$$\mathbf{Y}|\mathbf{X} = \mathbf{x} \sim \text{Dirichlet}(\phi\alpha_1(\mathbf{x}), \phi\alpha_2(\mathbf{x}), \cdots, \phi\alpha_p(\mathbf{x})),\ \phi > 0 \tag{2}$$

을 적합할 수도 있다. 이때 ϕ, β_{j0}, $\boldsymbol{\beta}_j$의 값의 추정은 최대가능도 추정(maximum likelihood estimation)을 이용한다. `Compositional` 패키지의 `diri.reg` 함수를 이용해 적합할 수 있다.

위의 모형들은 $Y_i = 0$인 자료가 존재할 때, 모형이 제대로 추정되지 않는 문제가 발생한다. 이 경우 M. Tsagris (2015)가 제안한 *Regression with compositional data using the α-transformation* (α-regression)을 고려해볼 수 있다. α-regression 모형은 M. T. Tsagris, Preston, and Wood (2011)가 제안한 α-transform을 통해 자료를 변환하고, 변환된 자료에 적당한 모수적 모형을 적합한다. 이때, 조율 모수(tunning parameter) $|\alpha| < 1$ 값은 cross validation을 통해 결정할 수 있다. α-regression 모형은 `Compositional` 패키지의 `alfa.reg` 함수를 이용해 적합할 수 있다.

앞서 언급한 모형들은 모두 모수적 모형으로 일반적인 형태의 함수를 추정하기엔 제약이 있다. 이러한 문제를 해결할 수 있는 모형으로 Jeon and Park (2018)이 제안한 *Bochner smooth backfitting estimators* (B-SBF)을 고려해볼 수 있다. B-SBF 모형은 앞서 언급한 모형들에 비해 모형의 형태에 큰 제약을 두지 않는다. 이러한 특징으로 적합에는 앞서 언급한 모형들보다 더 많은 시간을 필요로 하지만 유연한 구조를 갖는 함수를 추정할 수 있는 강점을 갖는다. B-SBF 모형은 github repository `https://github.com/jeong-min-jeon/Add-Reg-Hilbert-Res`의 소스코드를 이용해 적합할 수 있다.

4 Results

4.1 Model Comparison

Method	ASPE
B-SBF with CBS	5.905169
Alpha transformation method (M. Tsagris, 2015)	0.512279
Dirichlet regression	0.7118787
Multinomial logistic regression (KL divergence)	0.7314985
Multinomial logistic regression (OLS)	100.0437

Table 1: Comparison of ASPE

Section 3.2에서 설명한 다양한 방법으로 적합하였으며, cross-validation을 통해 구한 average squared prediction error(ASPE)은 Table 1에 제시되어 있다. 여기서 ASPE를 계산하는 데 필요한 거리 d는 Aitchison geometry에서 계산하였다. Alpha transformation method에서 alpha값은 10-fold CV로 계산된 -1을 사용하였다.

ASPE를 기준으로 하였을 때, multinomial logistic을 제외한 나머지 모수적 모형들은 비교적 서로 비슷한 결과를 보여주었다. Bochner smooth backfitting을 활용한 비모수적 모형은 이들에 비해 error가 더 크게 나타났다. Multinomial logistic의 경우 error가 지나치게 높게 계산되었는데, 이는 simplex 데이터에 극단적으로 치우친 비율이 있는 경우 수치가 조금만 차이가 나도 서로 거리가 매우 크게 계산되기 때문인 것으로 보인다.

4.2 Visualization

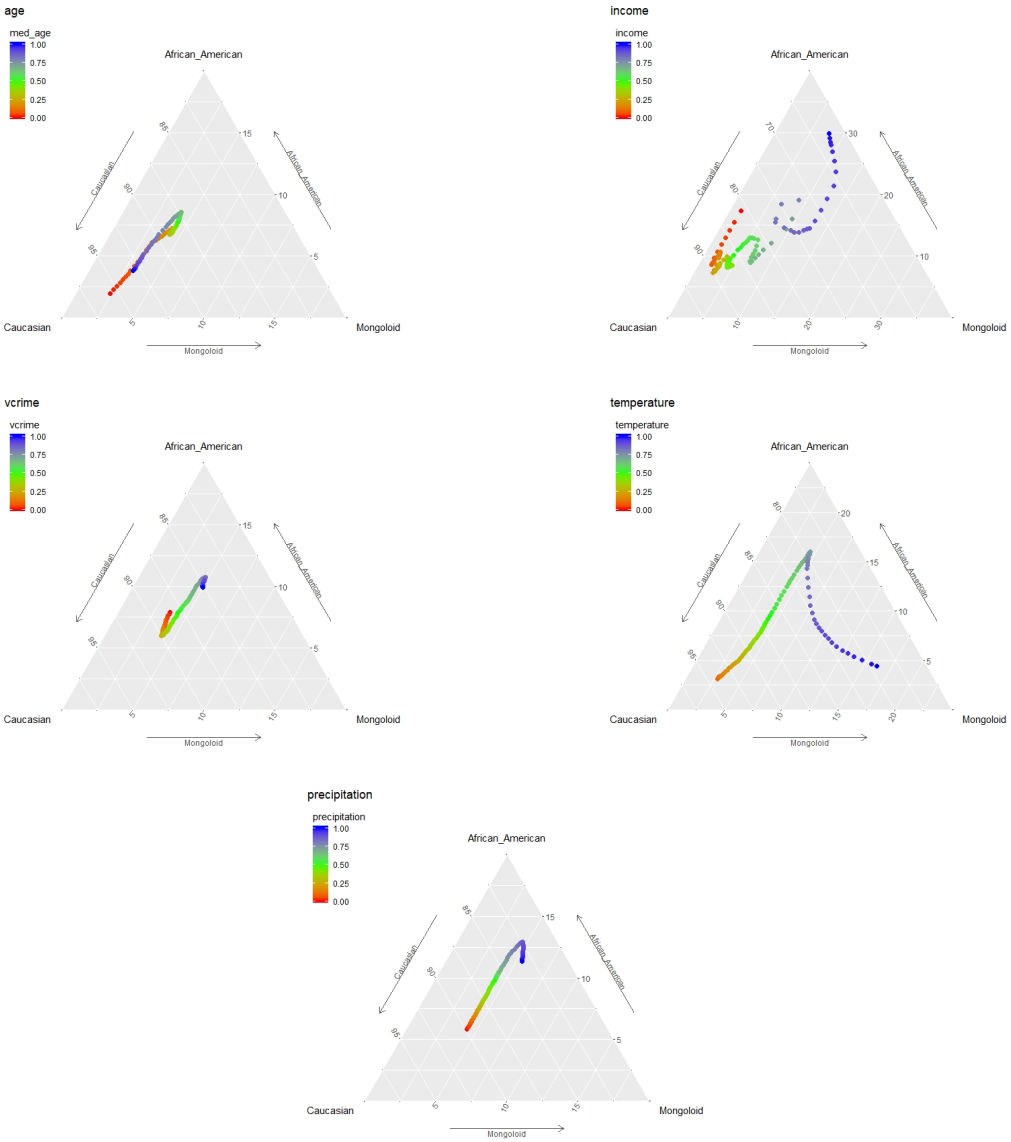

Figure 1: Jeon and Park (2018)의 방법을 이용하여 적합한 인구 구성비 결과를 나타낸 2-simplex (ternary diagram) visualization.

Figure 1은 Jeon and Park (2018)에서 제시된 방법을 이용해서 적합한 성분비의 추정 값들이 각 요인들의 움직임에 따라 어떻게 움직이는 지를 시각화한 것이다. 각각의 요인들의 영향력을 알아보기 위해서 우리는 성분비의 추정 값을 계산할 때 정해진 한 변수를 제외하고 다른 변수들은 모두 평균값으로 세팅하였고, 한 변수 값의 움직임에 따라 어떻게 성분비의 추정 값이 변화하는 지를 분석해 보았다.

첫 번째 그림에서는 연령대가 변화함에 따라서 인구 구성비 중에서 흑인과 백인의 비율이 두드러지게 변화한다는 것을 알 수 있다. 연령대가 높아짐에 따라서 처음에는 백인의 비율이 감소하고 흑인의 비율은 증가하다가 일정 나이대를 지나면 다시 백인의 비율이 증가하고 흑인의 비율은 증가하는 경향을 보이고 있다. 여기서 동양인의 비율은 평균 연령대가 변화해도 크게 변동하지 않는다는 흥미로운 사실을 발견하였다.

두 번째 그림은 수입이 변화함에 따라서 인구 구성비가 어떻게 변화하는지를 보여주고 있다. 다른 4개의 요인에 비해 불규칙한 경향성을 보이고 있다. 다만 전반적으로 평균 수입이 늘어남에 따라서 동양인의 비율이 늘어나고 흑인의 비율은 줄어드는 경향을 보이고 있다. 또한, 일정 수준 이상부터는 수입이 늘어남에 따라 백인의 비율이 줄어드는 경향성을 보인다. 수입 최상위층 구간에서 흑인의 비율이 조금 늘어나고 동양인의 비율은 오히려 줄어드는 경향을 보이는 것이 또 하나의 흥미로운 사실이었다.

세 번째 그림에서는 범죄율이 변함에 따라서 흑인과 백인의 비율이 두드러지게 변화하는 것을 알 수 있다. 양 극단을 제외하고는 범죄율이 높을수록 흑인의 비율이 늘어나고 백인의 비율은 줄어들지만, 극단에서는 이 관계가 역전이 되고 있다. 한편 동양인의 비율은 범죄율에 거의 영향이 없는 것으로 보였다.

네 번째 그림, 다섯 번째 그림은 각각 온도, 강수량의 변화에 따른 인구 구성비의 변화를 나타낸다. 일정 온도 이하까지는 온도가 높아질수록 흑인의 비율은 늘어나고 백인의 비율은 줄어들다가, 일정 온도 이상이 되면 흑인의 비율은 줄어들고 동양인의 비율은 두드러지게 늘어남을 알 수 있다. 이는 알래스카와 같은 척박한 곳에는 백인의 비율이 높지만 일정 온도 이상이 되면 온도가 오를 수록, 즉 살기에 더 쾌적할수록 동양인의 비율이 늘어나고 있음을 보여주고 있다. 강수량에 따른 변화 역시 비슷한 경향을 보인다. 즉, 강수량이 낮은 구간에서는 백인의 비율이 높지만 일정 강수량 이상이 되는 지점부터는 강수량이 늘어날수록 동양인의 비율이 높아지고 있다.

5 Conclusion and Discussion

이 보고서에서는 simplex 데이터에 적용할 수 있는 비모수적 additive regression 모형을 이해하고, 다양한 미국의 지역별 데이터를 통해 다양한 변수에 따라 지역의 인종 비율이 어떻게 바뀌는지를 분석하였다. 그리고 이 결과를 다양한 모수적 모형과 비교하여 성능을 확인하였다. 또한, 각 변수들의 변화에 따른 구체적인 인종 비율의 변화를 시각적으로 표현하였으며, 이는 Section 4.2에서 보듯 해석 가능하고 흥미로운 분석 결과를 제공하였다.

한편 데이터의 처리 과정에서의 어려움이 다소 있었다. 우선 데이터를 얻고 취합하는 과정에서, 최대한 일치된 시기에 수집된 데이터를 사용하려고 하였으나 강수량 등의 데이터는 수집에 일정 기간이 필요하기 때문에 다소 한계가 있었다. 또한, 원래 데이터를 수집하는 지역 단위는 county로 하고자 하였으나 county별 인구수 등의 차이가 매우 심하고 일부 변수는 수집하는데 어려움이 있어, 주(state) 단위로 데이터를 수집하였고 이로 인해 데이터의 개수가 다소 적다는 문제가 발생하였다. 이에 더해 반응변수로 선정한 미국 인종 비율 데이터는 전체적으로 백인의 비율이 높은 unbalanced 데이터이므로 적합에 어려움이 있었다.

Section 4.1에서 알 수 있듯이 분석 대상인 미국 지역의 데이터에 대해 비모수적 모형은 다른 모수적 모형과 비교하였을 때 충분히 만족스럽게 적합되지는 못하였다. Section 3.2에서 언급했듯이, 비모수적 모형의 경우 다른 모형에 비해 그 형태에 제약이 더 작은 대신 분석의 대상인 미국 지역 데이터에 대한 최적화가 충분하지 못했던 것이 하나의 이유라고 할 수 있다. 기존 분석에서는 Jeon and Park (2018)에서 제공한 소스코드를 사용하였는데, 이를 문제에 맞게 세부적으로 조절할 수 있다면 조금 더 개선할 수 있을 것으로 생각된다.

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *44*(2), 139–160.

Jeon, J. M., & Park, B. U. (2018). *Additive regression with hilbertian responses* (Unpublished doctoral dissertation). PhD thesis. Seoul National University (cit. on pp. 2999–3001).

Tsagris, M. (2015). Regression analysis with compositional data containing zero values. *arXiv preprint arXiv:1508.01913*.

Tsagris, M. T., Preston, S., & Wood, A. T. (2011). A data-based power transformation for compositional data. *arXiv preprint arXiv:1106.1451*.