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 유클리드 공간 상의 자료(Euclidean data)에 smooth backfitting 알고리듬을 사용한 가법모형(additive regression model)의 적합, 그리고 이 알고리듬의 수렴성(convergence)를 포함한 점근적 성질(asymptotic property)에 대해 학습하였다. 또한, Jeon and Park (2018)을 통해 반응변수(response variable)를 힐베르트 공간 상의 자료(Hilbertian data)로 확장한 모형을 소개하고, 그 이론적 성질들을 다루었다.

본 보고서에서는 수업시간에 다룬 모형을 실제 자료에 적용하고 그 결과를 살펴보고자 한다. 구체적으로는, Jeon and Park (2018)에 소개된 여러 모형들을 통해 미국 지역별 인종 비율 자료를 분석한다. 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 가법모형을 적합하고, 각 설명변수의 변화에 따라 인종 비율이 어떻게 변하는 지를 시각적으로 나타내보고자 한다.

미국 지역별 인종 비율 자료는 p차원 simplex S_1^p 위의 자료이다. S_1^p는 p 차원 유클리드 공간의 부분공간 중 모든 원소의 합이 1이라는 조건을 만족하는 공간, 즉, 다음과 같은 공간을 의미한다

$$S_1^p = \left\{ \mathbf{x} = (x_1, \cdots, x_p) \in \mathbb{R}^p : x_i \geq 0,\ i = 1, 2, \cdots, p,\ \sum_{i=1}^{p} x_i = 1 \right\}.$$

S_1^p에 적당한 연산을 부여하면 이 공간을 비유클리드(non-Euclidean) 힐베르트 공간으로 만들 수 있다. Jeon and Park (2018) 또한, 이러한 공간 위의 자료로 한국의 선거 결과 자료를 분석한 결과를 본문에 소개하고 있다.

각 절은 다음과 같이 구성되어있다. Section 2에서는 분석에 사용할 자료에 대해 간략히 소개한다. Section 3에서는 p차원 simplex S_1^p를 힐베르트 공간으로 나타내는 방법과 그러한 공간 위의 자료를 분석하기 위해 제안된 모형들을 소개한다. Section 4에서는 소개한 모형들을 실제 자료에 적용한 결과를 다루며 Section 5에서 그 결과에 대한 해석과 제언을 다룬다.

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 자료와 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. 기후 관련 변수는 다른 자료들과 달리 연간 편차가 심하다는 점을 고려하여 1981년부터 2010년까지의 각 주 별 평균 기온 자료(X_4)와 평균 강수량 자료(X_5)를 사용하였다.

[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 Compositional Data

3.1 Aitchison geometry

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 힐베르트 공간이 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 최소화하는 모형을 적합하는 함수들 또한 제공하고 있다.

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

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

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) α 값은 교차 검증(cross validation; CV)를 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. 비모수 함수 추정에 필요한 bandwidth는 동일한 논문에서 제안된 CBS(Coordinate-wise Bandwidth Selection) 알고리듬을 사용하여 결정할 수 있다. B-SBF 모형은 적합 과정에서 모수적 모형들보다 더 많은 시간을 필요로 하지만 유연한 구조를 갖는 함수공간을 탐색할 수 있다. 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

이 절에서는 3.2에서 소개한 모형들을 자료에 적용한 결과를 다룬다. 각 모형의 성능은 Jeon and Park (2018)에서와 같이 다음과 같이 정의되는 10-fold average squared prediction error (ASPE) 값으로 비교하였다

$$\text{ASPE} = 10^{-1} \sum_{k=1}^{10} |S_k|^{-1} \sum_{i \in S_k} \| \mathbf{Y}_i \ominus \hat{\mathbf{Y}}_i^{(-S_k)} \|^2.$$

여기서 각 S_k는 k번째 분할(partition)을, $\hat{\mathbf{Y}}_i^{(-S_k)}$은 S_k를 제외한 나머지 자료로 적합한 모형으로 계산한 반응변수의 예측값을 의미한다. 각 모형의 성능을 계산한 결과를 Table 1에 정리하였다. α-regression 모형에서 α값은 CV로 선택된 -1을 사용하였다.

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

Table 1: Comparison of ASPE

ASPE를 기준으로 하였을 때, OLS를 이용한 Multinomial logistic regression 모형을 제외한

나머지 모수적 모형들은 비교적 서로 비슷한 결과를 보여주었다. B-SBF 모형은 이들에 비해 error가 더 크게 나타났다. OLS를 이용한 Multinomial logistic 모형은 ASPE 값이 지나치게 높게 계산되었는데, 이는 Aitchison geometry의 거리가 simplex의 경계에 위치한 점에서 극단적으로 크게 계산되기 때문인 것으로 보인다.

4.2 Visualization

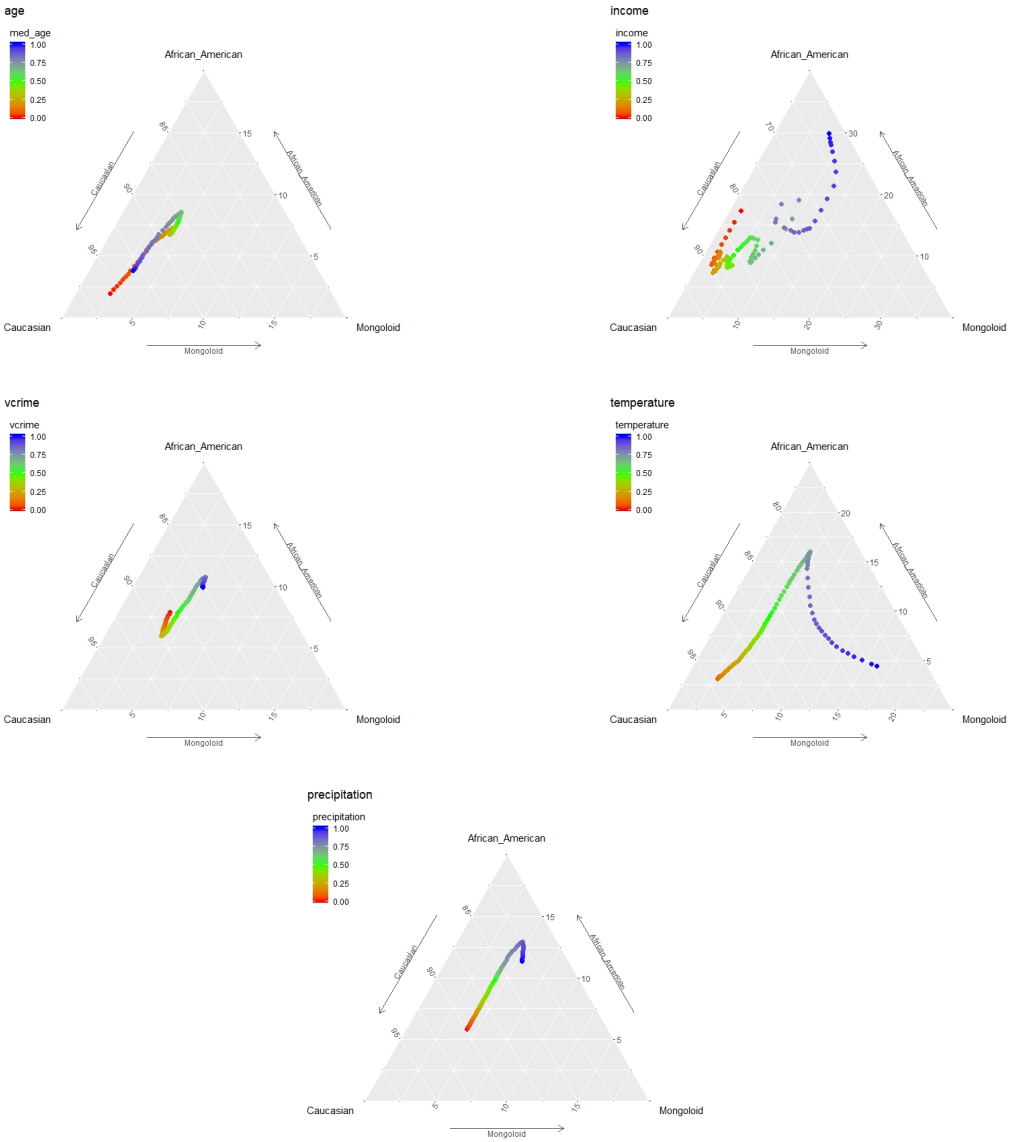

Figure 1: Jeon and Park (2018)의 방법을 이용하여 적합한 인구 구성비 결과를 나타낸 2-simplex (ternary diagram) visualization.

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 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 위의 자료에 적용할 수 있는 비모수 가법모형을 이해하고, 다양한 미국의 지역별 자료를 통해 다양한 변수에 따라 지역의 인종 비율이 어떻게 바뀌는 지를 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.

Section 4.1에서 알 수 있듯이 분석 대상인 미국 지역의 자료에 대해 비모수적 모형은 모수적

모형과 비교하였을 때 충분히 만족스럽게 적합되지 못하였다. Section 3에서 언급했듯이, 비모수적 모형은 모수적 모형에 비해 그 형태에 제약이 더 작다는 강점을 가지나, 실제 모형(true model)이 모수적 모형일 때 과적합 등으로 인해 더 낮은 성능을 가질 수 있다. 본 보고서에서 사용한 자료의 크기가 충분히 크지 않았던 점을 고려해보면 이러한 이유로 인해 비모수적 모형이 적합 과정에서는 더 오랜 시간이 걸렸음에도 성능은 더 떨어지는 결과를 보였을 것이라 추측할 수 있다. 또한, 반응변수로 선정한 미국 인종 비율 자료는 전체적으로 백인의 비율이 높은 불균형한(unbalanced) 자료였는데, 그 결과 추정된 값이 simplex의 경계에 위치했을 때, ASPE 값이 극단적으로 높게 계산되는 문제가 발생했다. α-regression은 적당한 변환을 통해 이러한 문제를 해결한 것으로 생각된다.

마지막으로 본 보고서의 발전 방향을 몇 가지 언급하고자 한다. 먼저, 더 나은 자료를 사용할 것을 제안한다. 기존 계획은 미국의 각 county 별 자료를 수집하는 것이었으나, 자료의 출처마다 관측치의 정확도의 차이가 심하고 분포가 매우 불균형하여 주 단위로 자료를 수집하였다. 향후의 분석에서는 더 많은 자료를 통해 계산한 안정적인 결과에 대한 비교를 제안한다. 다음으로, 자료에 맞는 모형의 최적화(fine-tunning)을 제안한다. 기존 분석에서는 Jeon and Park (2018)에서 제공하는 함수만을 사용하였는데, 이를 문제에 맞게 세부적으로 조절할 수 있다면 더 나은 결과를 얻을 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*.