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 13, 2019

1 Introduction

2 Data

분석에 이용한 자료는 다양한 출처를 통해 수집하였다. 먼저, 각 주(State)별 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.

각 주별 연령($X1$) 또한 State Population by Characteristics: 2010-2018 자료를 통해 계산한 2017년 기준 주별 연령 분포의 중앙값을 사용하였다.

각 자료는 미국 각 주별 관측치와 워싱턴 D.C.의 관측치, 총 51개의 관측치로 이루어져 있다.

[1] `https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-detail.html`

3 Methods

반응변수 $Y = (Y_1, Y_2, Y_3)$는 유클리드 공간 상의 자료가 아닌 3차원 simplex S_1^3 위의 자료이다. 이 절에서는 p차원 simplex S_1^p 위의 자료(compositional data)를 분석하기 위해 제안된 모형들에 대해 간략히 소개한다.

가장 간단한 모형으로, *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

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

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*

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

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

5 Conclusion

References

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*.