

Statistik-Übung 7 – PCA

Beispielhafter Methoden- und Ergebnisteil

Methoden

Ziel war es zu visualisieren, wie sechs sozioökonomische Variablen (Bevölkerungszahl, durchschnittliches Einkommen, Analphabetenrate, durchschnittliche Lebenserwartung, Häufigkeit von Tötungsdelikten, Anteil der Bevölkerung mit einem Studienabschluss) sowie die mittlere Anzahl an Frosttagen in 10 ausgewählten US-amerikanischen Bundesstaaten in Beziehung zueinanderstehen.

Die Analysen wurden in R, Version 4.2.2, durchgeführt (R Core Team 2024). Es wurde eine Hauptkomponentenanalyse (PCA) der sieben Variablen gerechnet (Befehl 'prcomp' in Base R). Da die sieben Variablen auf ganz unterschiedlichen Skalen gemessen worden sind, wurden sie zunächst z-transformiert, so dass alle Variablen einen Mittelwert von 0 und eine Standardabweichung von 1 haben. Mit einer «Screeplot» wurde ermittelt, wie viele Achsen der PCA bedeutsam sind.

Ergebnisse

Gemäss «Screeplot» waren nur die erste und zweite Hauptkomponente relevant, die zusammen auch 90.7 % der Varianz erklärten (PC1: 63.2%, PC2: 27.6 %). In der Visualisierung der PCA (Abb. 1) zeigt sich, dass die erste Hauptkomponente v.a. für einen Gradienten von Bundesstaaten mit hoher Analphabetenrate und vielen Tötungsdelikten (links) zu Bundesstaaten mit hoher Lebenserwartung und vielen Hochschulabsolventen (rechts) steht. Die zweite Achse repräsentiert v.a. einen Gradienten von bevölkerungsreichen (unten) zu bevölkerungsarmen Bundesstaaten (oben).

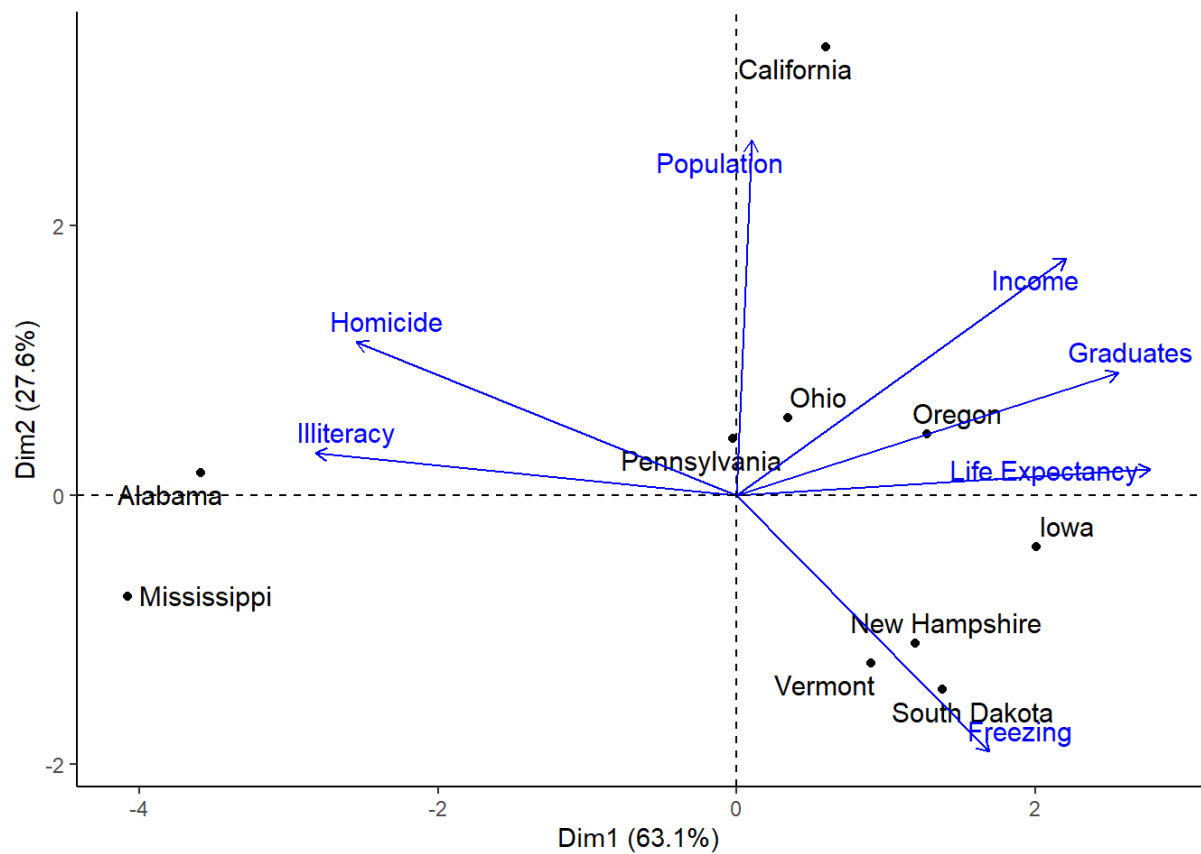


Abb. 1. Visualisierung der Hauptkomponentenanalyse von sieben Variablen in 10 US-amerikanischen Bundesstaaten.

Quellen

R Core Team. (2024) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <http://www.r-project.org/>.