MSc. Research Methods – Statistikteil
Lösungstext

- Übung 6.1: Kombination aus PCA und multipler Regression –

(Hier nur die geforderte Lösung mit PCA-Achsen und nicht die im Skript auch enthaltene Lösung mit den originalen Umweltvariablen, die sich letztlich als besser erwiesen hat)

Methoden

Es wurde die Fischartenzusammensetzung zusammen mit 11 Umweltdaten (Tabelle 1) an 30 Untersuchungsstellen des Flusses Doubs im Jura erhoben.

Tab. 1: Einfluss der Bewässerung auf den Ertrag.

Abkürzung	Parameter und Einheit
dfs	Distance from source (km)
ele	Elevation (m a.s.l.)
slo	Slope (‰)
dis	Mean annual discharge (m^3 s^{-1})
pH	pH of water
har	Hardness (Ca concentration) (mg L^{-1})
pho	Phosphate concentration (mg L^{-1})
nit	Nitrate concentration (mg L^{-1})
amm	Ammonium concentration (mg L^{-1})
oxy	Dissolved oxygen (mg L^{-1})
bod	Biological oxygen demand (mg L^{-1})

Der Artenreichtum der Fische sollte in einer multiplen Regression mittels dieser Umweltvariablen erklärt werden. Wegen der teilweise hohen Korrelation zwischen den Umweltvariablen, wurden diese zunächst einer Hauptkomponentenanalyse (PCA) mit standardisierten Variablen unterzogen. Von den Hauptkomponenten wurden die n ersten, die zusammen mehr als 90 % der Varianz in den standardisierten Umweltdaten erklären als orthogonale Prädiktoren genommen.

Es wurde ein globales Regressionsmodell als lineares Modell und als Poisson-GLM mit diesen ersten PC-Achsen (als synthetischen Variablen) gebildet und diese dann schrittweise solange vereinfacht, bis nur noch signifikante Variablen verblieben. Die Validität (Varianzhomogenität, Normalverteilung der Residuen) der resultierenden minimal adäquaten Modelle wurde dann visuell in Residualplots geprüft.

Ergebnisse

Die ersten vier PC-Achsen erklärten zusammen 90.3 % der Varianz (54.3 %, 19.7 %, 9.7 %, 6.7 %) und wurden daher in das globale Regressionsmodell aufgenommen. Sowohl im linearen Modell als auch im GLM verblieben die ersten drei Achsen im minimal adäquaten Modell. Da die Modellvoraussetzungen im linearen Modell besser erfüllt waren als im GLM, wird das lineare Modell als finales Modell genommen (Tab. 1). Das sich ergebende Modell lautet:

Artenreichtum = 12.5 + 4.0 PC1 + 6.7 PC2 – 3.0 PC3

Dabei hat die zweite Achse den höchsten Erklärungswert, gefolgt von der ersten und der dritten (Tab. 2). Die erste Achse, welche 54.3% der Varianz im Umweltdatensatz erklärt und im Prinzip den Abstand von der Quelle und die damit einhergehenden Umweltveränderungen beschreibt, hat für die Erklärung der Artenzahl nur an zweiter Stelle Bedeutung. Danach nimmt die Fischartenzahl mit zunehmender Distanz von der Quelle und niedrigerer Meereshöhe zu, wobei zugleich Nährstoffgehalte (nit, pho, amm) höher sind. Am wichtigsten für die Artenzahl ist aber die zweite Achse, die einen negativen Einfluss von biologischem Sauerstoffbedarf, Ammonium- und Phosphorgehalt zeigt. Auf der dritten Achse zeigt sich schliesslich ein negativer Einfluss des pH-Wertes und ein positiver des Gefälles.

Tab. 2: Minimal adäquates lineares Modell zur Erklärung des Fischartenreichtums durch die Hauptkomponenten der 11 Umweltvariablen. Für die Hauptkomponenten sind jeweils die betragsmässig am höchsten landenden Umweltparameter genannt (Ladungen aus der PCA).

Hauptkomponente	**Effekt**	***p***	**Hochladende Umweltvariablen**
PC1	4.40	0.002	nit (+1.15), dfs (+1.09), pho (+1.06), ele (-1.04), amm (+1.01), oxy (-0.97), bod (+0.97), dis (+0.95), har (+0.90)
PC2	6.73	< 0.001	bod (-0.71), amm (-0.70), dis (+0.66), ele (-0.61), pho (-0.60)
PC3	-2.96	0.026	pH (+1.01), slo (-0.63)