Executive Summary

Worldbank Datensatz - verschiedene Fragestellungen

Nikolai German, Wenxuan Liang, Thomas Witzani, Yanyu Zhao

Zusammenfassung

Bei der Analyse des Worldbank Datensatzes konnten wir einige klare Zusammenhänge identifizieren. Andere Fragestellungen konnten wir dagegen nicht, oder nur teilweise beantworten. Zusätzliche Länder, ein längerer Beobachtungszeitraum und weitere Variablen könnten bei diesen Fragen neue Erklärungsansätze liefern.

Fragestellungen

Wir haben uns mit fünf Fragekomplexen beschäftigt:
(i) Die Korrelation zwischen Zugang zu Elektrizität und Nettonationaleinkommen. **(ii)** Die Bildungsquote der Erwerbspersonen eines Landes und möglicher Zusammenhang zur Staatsverschuldung beziehungsweise zur Anzahl von Schülern je Lehrer. **(iii)** Die Beziehung zwischen HIV-Prävalenz und Alkoholkonsum und der Zusammenhang von Bildungsquote und HIV-Prävalenz. **(iv)** Die Korrelation zwischen Bruttoinlandsprodukt und Tabakkonsum. **(v)** Die Frage nach einer Beziehung zwischen landwirtschaftlicher Nutzfläche und CO2-Emissionen pro Kopf.

Datenlage und Methodik

Der uns vorliegende Datensatz enthält jährliche Beobachtungen von 18 verschiedenen Merkmalen und 25 Ländern über einen Zeitraum vom Jahr 2000 bis zum Jahr 2021. Die Datenlage ist je nach betrachtetem Merkmal sehr unterschiedlich. Beispielsweise sind die Beobachtungen zum Nettonationaleinkommen pro Kopf fast vollständig, während Beobachtungspaare von HIV-Prävalenz und Alkoholkonsum für sieben Länder gar nicht vorliegen. Zusätzlich nutzen wir die Zuordnung der Länder nach Kontinent und die Klassifikation in Einkommensklassen (Download) der Weltbank. Um Zusammenhänge zweier Merkmale zu quantifizieren verwendeten wir durchgängig den Spearman-Korrelationskoeffizient. Eine Berechnung dieser Rangkorrelation ist nicht möglich, falls eines der Merkmale einen konstanten Wert aufweist. Bei unvollständigen Beobachtungen entschieden wir uns dafür, die Daten zu aggregieren, um dann länderübergreifende Trends durch Regressionsgeraden zu visualisieren. Durch die begrenzte Anzahl an Ländern im Datensatz und die teilweise große Spannweite der Daten (beispielsweise Bevölkerungsanzahl von Aruba und VR China) wurden für diese Regressiongeraden sehr große Konfidenzintervalle geschätzt. Das hat zur Folge, dass für viele Fragestellungen keine eindeutigen Tendenzen bestimmt werden konnten.

Ergebnisse

Im Einzelnen kamen wir zu den folgenden Ergebnissen:
(i) Der Zugang zu Elektrizität korreliert im Datensatz stark positiv mit dem Nettonationaleinkommen pro Kopf. Einen Einfluss von Landes- oder Bevölkerungsgröße darauf konnten wir nicht feststellen. **(ii)** Einen Zusammenhang zwischen Staatsverschuldung und Bildungsquote ließ sich in den Daten nicht feststellen. Ebenso wenig ob Länder mit hoher Bildungsquote ein niedrigeres Schüler-Lehrer-Verhältnis halten können. **(iii)** Zwischen HIV-Prävalenz und dem Alkoholkonsum der erwachsenen Bevölkerung konnten wir keine Beziehung feststellen. Ein Effekt der Bildungsquote auf die HIV-Prävalenz war ebenfalls in den Daten nicht zu sehen. **(iv)** Die Prävalenz des Tabakkonsums korreliert negativ mit dem Bruttoinlandsprodukt pro Kopf. **(v)** Aus den Daten lässt sich nicht erschließen, ob es einen Zusammenhang zwischen dem Anteil der landwirtschaftlich genutzten Landesfläche und den CO2 Emissionen pro Kopf gibt.

Ausblick

Grundsätzlich gibt es aus unserer Sicht zwei mögliche Ansätze um die Ergebnisse zu verbessern:
(a) Erhöhung der Datenmenge, durch Hinzunahme zuästzlicher Länder und/oder Betrachtung eines größeren Zeitraums. Ersteres könnte die Unsicherheit bei der Betrachtung aggregierter Werte auf Länderebene verringern, Zweiteres die Unsicherheit bei der Betrachtung von Zusammenhängen innerhalb eines Landes. **(b)** Nicht erfasste Variablen, wie beispielsweise die mehrheitliche Religionszugehörigkeit der Bevölkerung eines Landes, oder die Leistungsfähigkeit des Gesundheitssystems, könnten neue Erklärungsansätze liefern.