

Vergleich des Fay-Harriot (FH) und des Battese-Harter-Fuller (BHF) Models

Niklas, Lorenz

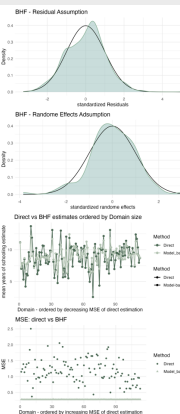
Institut für Statistik | Universität Bamberg



1) Motivation

Wir vergleichen die Schätzgenauigkeit von Area-Level- (FH) und Unit-Level-Modellen (BHF) bei sehr kleinen Stichproben. Ein Simulationsansatz ermöglicht den direkten Vergleich der Schätzungen mit der bekannten wahren Verteilung der Zielvariable.

2) Methoden



Modellbildung

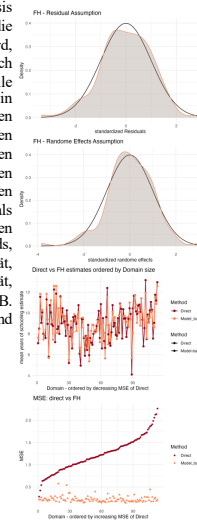
Zunächst eine einzelne Stichprobe aus den simulierten Daten herangezogen. Die Variablenselektion erfolgt schrittweise: Variablen mit hohem Anteil fehlender Werte werden ausgeschlossen, ebenso Variablen mit sehr vielen Faktorstufen, stark korrelierte Variablen sowie Variablen, die inhaltlich zu nah an der Zielvariable Bildungsjahre liegen. Die finale Variablenauswahl erfolgt mittels Stepwise-Regression (backward, forward, both) auf Basis des BIC-Kriteriums, wodurch die Modellkomplexität reduziert wird, ohne die erklärte Varianz wesentlich zu verringern. Für nicht-faktorielle Variablen erfolgt eine Umkodierung in Dummy-Variablen. Die endgültigen FH- und BHF-Modelle werden anschließend mit den ausgewählten Kovariate geschätzt. Eingeschlossen sind sowohl individuelle Variablen wie Alter, Beruf und Lesefähigkeit, als auch haushaltsbezogene Indikatoren des Wohn- und Lebensstandards, darunter Urbanität, Krankenversicherung, Wohnqualität, Ausstattung der Wohnung (z. B. Küche, Warmwasserbereitung) und Autobesitz.

Daten

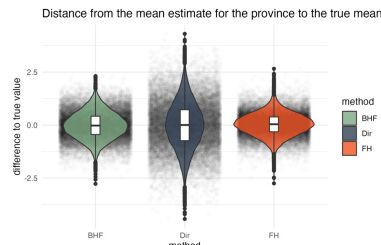
Vollerhebung (Census Bolivien 2024) mit Personen- und Haushaltsdaten. Personen über 18 Jahre.

Simulation

Simulation von 200 Stichproben mittels SRS mit $n = 20$ pro Domäne für 113 Provinzen. Die Formel, die für das erste Sample ermittelt wurden dann auf diese angewandt



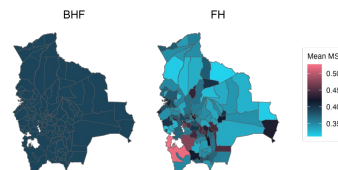
3) Ergebnisse



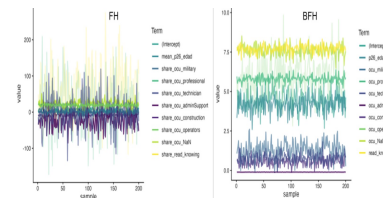
In diesem Abschnitt möchte ich dan nvor allem darüber reden, wie viel Variablen, durch die Regression ausgeschlossen wurden und welche variable, dann am ende übrig geblieben sind, außerdem, wie häufig welche variable bei dem BHF Modell dann für die Analyse ausgeschlossen wurde

Obwohl die Koeffizienten des BHF-Modells stabiler sind, erzielt das FH-Modell im Mittel niedrigere MSEs und liegt näher an den wahren Werten. Dies zeigt, dass größere Konsistenz der Parameter nicht automatisch höhere Genauigkeit der Schätzungen bedeutet, insbesondere bei sehr kleinen Stichproben pro Domäne.

Mittlerer MSE auf provinzieller Ebene



Obwohl die Koeffizienten des BHF-Modells stabiler sind, erzielt das FH-Modell im Mittel niedrigere MSEs und liegt näher an den wahren Werten. Dies zeigt, dass größere Konsistenz der Parameter nicht automatisch höhere Genauigkeit der Schätzungen bedeutet, insbesondere bei sehr kleinen Stichproben pro Domäne.



4) Diskussion

Obwohl die Koeffizienten des BHF-Modells stabiler sind, erzielt das FH-Modell im Mittel niedrigere MSEs und liegt näher an den wahren Werten.

Dies zeigt, dass größere Konsistenz der Parameter nicht automatisch höhere Genauigkeit der Schätzungen bedeutet, insbesondere bei sehr kleinen Stichproben pro Domäne. Der FH-Ansatz profitiert von der direkten Aggregation auf Domänenebene, während das BHF-Modell auf individuelle Daten angewiesen ist, die bei kleinen n stärker schwanken.

5) Limitationen

Die Verteilung der Residuen ist nicht optimal, und auch die gewählten Variablen weisen keine ideale Verteilung auf. Insbesondere bei korrelierten Variablen wäre es sinnvoll gewesen, zu prüfen, welche Variablen einen besseren Modell-Fit ermöglichen. Darüber hinaus würde ein aussagekräftiger Vergleich der Modelle erfordern, dass für jedes Modell eine separate Modellsuche durchgeführt wird, um die besten Prädiktoren zu identifizieren.

Die Verteilung der Residuen ist nicht optimal, und auch die gewählten Variablen weisen keine ideale Verteilung auf. Insbesondere bei korrelierten Variablen wäre es sinnvoll gewesen, zu prüfen, welche Variablen einen besseren Modell-Fit ermöglichen. Darüber hinaus würde ein aussagekräftiger Vergleich der Modelle erfordern, dass für jedes Modell eine separate Modellsuche durchgeführt wird, um die besten Prädiktoren zu identifizieren.

References

- Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data.
- Fay III, R. E., & Harriot, R. A. (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data.
- Harmening, S., Kreuzmann, A.-K., Schmidt, S., Salvati, N., & Schmid, T. (2023). A framework for producing small area estimates based on area-level models in R.
- INE Bolivia. (2024). Census (Dataset).
- Kreuzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Tempel, M., & Tzavidis, N. (2019). The R package emdi for estimating and mapping regionally disaggregated indicators.
- UN Office for the Coordination of Humanitarian Affairs (OCHA). (2025). Bolivia Administrative Boundaries [Dataset].
- Wurz, N. (2025). *saetra: Transformations for unit-level small area models* [Manual].