Abstract

Statistical tools to analyze research data are widely applied in many scientific disciplines and the need for adequate statistical models and sound statistical analyses is apparent. This thesis addresses limitations in statistical models commonly used to identify causal effects and for prediction purposes. Moreover, difficulties in the replicability of statistical results are revealed and remedies are suggested.

With regard to causality, the incorporation of penalized splines into fixed effects panel data models is proposed. Fixed effects panel data models are often used in order to establish causal effects since they control for unobserved time-invariant heterogeneity of the study entities. The inclusion of penalized splines relieves the researcher from determining functional shapes of the covariate effects. Instead, the functional forms are allowed to be flexible and are estimated based on the data at hand such that a data-driven degree of nonlinearity is identified. Simultaneous confidence bands are presented as a computationally fast and reliable uncertainty measure for the estimated functions. Furthermore, this thesis studies causal effects not only on the expectation but on all aspects of the distribution of the dependent variable. In particular, generalized additive models for location, scale and shape are introduced to (quasi-)experimental methods. A step-by-step guide demonstrates how the proposed methodology may be applied and provides insights which may go unnoticed in common regression frameworks.

In the domain of prediction, a small area prediction problem is considered. It is shown how to obtain reliable up-to-date welfare estimates when an outdated census without information on income and a more recent survey with information on income are available. Instead of using survey variables to explain income in the survey, the proposed approach uses variables constructed from the census. The underlying assumptions are less restrictive than those in commonly applied methods in this field that are tailored to situations with simultaneous census and survey collection.

As an overarching topic relating to all statistical analyses, the replicability of statistical results is considered from two viewpoints. On the one hand, the prevalence of reporting errors in statistical results is investigated. On the other hand, studies are replicated if possible by using the same data and software code as in the reference study. It is shown that replicability is frequently made impossible by reporting errors as well as by missing data and software code. At the same time, simple solutions to enhance replicability in future research are presented. Open data and software code policies together with a vivid replication culture seem to be most promising.

Zusammenfassung

Statistische Methoden zur Analyse von Forschungsdaten werden in vielen wissenschaftlichen Disziplinen eingesetzt. Der Bedarf an adäquaten statistischen Modellen und fundierten statistischen Analysen ist offensichtlich. Diese Dissertation adressiert Einschränkungen in statistischen Modellen, die üblicherweise zur Ermittlung kausaler Effekte und zu Vorhersagezwecken verwendet werden. Darüber hinaus werden Probleme hinsichtlich der Replizierbarkeit statistischer Ergebnisse aufgedeckt und Lösungen vorgeschlagen.

Im Hinblick auf Kausalität wird die Integration von pönalisierten Splines in Paneldatenmodelle mit fixen Effekten vorgeschlagen. Diese Modelle werden häufig zur Ermittlung kausaler Effekte verwendet, da sie für nicht beobachtete zeitinvariante Heterogenität der Beobachtungseinheiten kontrollieren. Die Einbeziehung von pönalisierten Splines befreit die Forscherin von der Aufgabe, die funktionalen Formen der Effekte der Kovariaten selbst festzulegen. Stattdessen dürfen die Funktionsformen flexibel sein und werden anhand der vorliegenden Daten geschätzt, sodass ein datengetriebenes Maß an Nichtlinearität bestimmt wird. Als eine rechenunaufwendige und zuverlässige Methode zur Unsicherheitsmessung für die geschätzten Funktionen werden simultane Konfidenzbänder vorgestellt. Darüber hinaus untersucht diese Arbeit kausale Effekte nicht nur auf den Erwartungswert, sondern auf alle Aspekte der Verteilung der abhängigen Variablen. Insbesondere werden generalisierte additive Modelle für Lokation, Skala und Form mit (quasi-)experimentelle Methoden verbunden. Eine Schritt-für-Schritt-Anleitung zeigt, wie die vorgeschlagene Methodik angewendet werden kann und Einblicke liefert, die in herkömmlichen Regressionsmodellen unbemerkt bleiben könnten.

Im Bereich der Prädiktion wird ein Problem der Vorhersage kleinräumiger Daten betrachtet. Es wird gezeigt, wie verlässliche und aktuelle Wohlfahrtsschätzungen erhalten werden können, wenn ein veralteter Zensus ohne Informationen über das Einkommen und neuere Surveydaten mit Informationen über das Einkommen verfügbar sind. Anstelle der Nutzung von Variablen aus dem Survey zur Vorhersage von Einkommen verwendet der vorgeschlagene Ansatz aus dem Zensus konstruierte Variablen. Die dafür notwendigen Annahmen sind weniger einschränkend als die in gewöhnlich verwendeten Verfahren, die auf Situationen mit gleichzeitiger Erhebung von Zensus und Survey zugeschnitten sind.

Als übergreifendes Thema aller statistischen Analysen wird die Replizierbarkeit statistischer Ergebnisse aus zwei Blickwinkeln betrachtet. Zum einen wird die Häufigkeit von Berichtsfehlern in statistischen Ergebnissen untersucht. Auf der anderen Seite wird versucht, Studien unter Verwendung der gleichen Daten und des gleichen Softwarecodes zu replizieren. Es wird gezeigt, dass die Replizierbarkeit häufig durch Berichtsfehler sowie durch fehlende Daten und Softwarecode unmöglich gemacht wird. Gleichzeitig werden einfache Lösungen zur Verbesserung der Replizierbarkeit in zukünftiger Forschung präsentiert. Vorschriften zur Offenlegung von Daten und Softwarecode zusammen mit einer regen Replikationskultur scheinen die vielversprechendsten zu sein.