Multivariate Analyse sozioökonomischer Einflüsse auf die Sterblichkeit

Lucas Reck

Fragestellung

Das Ziel der Arbeit ist es, ein Modell für die menschliche Sterblichkeit unter Berücksichtigung diverser sozioökonomischer Einflüsse zu entwickeln. Die Relevanz dieser Fragestellung zeigt sich beispielsweise darin, dass bereits vor mehreren hundert Jahren erste Sterbetafeln entwickelt wurden und untersucht wurde, welche Faktoren das Sterbeverhalten beeinflussen. Mit Hilfe einer detaillierteren Datengrundlage und hoher Rechenleistung können heutzutage jedoch deutlich komplexere und genauere Vorhersagen getroffen werden. Mit dieser Analyse werden also Treiber für das Sterberisiko identifiziert und deren Einfluss quantifiziert.

Methodologie

Die Fragestellung lässt sich im Wesentlichen der Ereigniszeitanalyse (Survival Analysis) zuordnen. Die größte Herausforderung stellt der Umgang mit zensierten Daten dar. Konkret bedeutet dies für diese Arbeit, dass die Zielgröße – nämlich der Todeszeitpunkt – nicht immer bekannt ist. Grund hierfür ist einerseits, dass eine Person vorzeitig aus der Studie austreten kann und andererseits, dass viele Personen zum Ende der Studie noch leben. In beiden Fällen ist also lediglich die Mindestüberlebenszeit bekannt, nicht jedoch der genaue Todeszeitpunkt. Deshalb können viele statistische Standardgrößen, wie beispielsweise der Durchschnitt, nicht berechnet werden. Es bedarf spezieller Methoden, um die Daten angemessen beschreiben zu können. Das bekannteste Verfahren ist das sogenannte Cox-Modell. Hierbei wird die Risikorate unter Berücksichtigung mehrerer Kovariablen modelliert. Auf die Grundrate (Baseline Hazard), die für alle Personen gilt und nur vom Alter abhängt, werden einzelne, personenspezifische Schätzer multipliziert, um das Gesamtrisiko einer Person zu berechnen.

Datengrundlage

Die Daten für diese Analyse stammen vom Sozio-ökonomischen Panel (SOEP), einer Abteilung des Deutschen Instituts für Wirtschaftsforschung (DIW). Es handelt sich bei den Daten um Ergebnisse von repräsentativen Befragungen deutscher Privatpersonen und Haushalte. Die Umfragen finden seit 1984 im jährlichen Rhythmus statt. Die Fragebögen beziehen sich auf die verschiedensten Aspekte im Leben der Befragten.

Um eine angemessene Datengrundlage für diese Analyse zu gewährleisten, müssen die Originaldaten noch angepasst werden. Zu den wesentlichen Aufgaben hierzu zählen unter anderem:

- Festlegung des Beobachtungszeitraums
- Auswahl der Kovariablen
- Umgang mit fehlenden Werten & Imputation
- Anpassung einzelner Kovariablen

Insgesamt entsteht somit ein Datensatz mit 23.885 Personen aus 12.690 Haushalten, von denen 2.229 im Beobachtungszeitraum versterben. Es werden folgende Kovariablen untersucht:

- Geschlecht
- Geburtsjahr
- Rauchen
- Arbeitslos
- Verheiratet
- Kinder unter 16 leben im Haushalt
- Private Krankenversicherung
- Private Zusatzkrankenversicherung
- Unbefristeter Arbeitsvertrag
- Voll erwerbstätig
- In Ausbildung
- Monatliches Bruttogehalt
- Monatliches Haushaltnettoeinkommen

Bevor das eigentliche Modell erstellt wird, werden die Daten deskriptiv untersucht. Hierzu zählt beispielsweise eine Korrelationsanalyse oder das Schätzen der Überlebensfunktion für unterschiedliche Personengruppen. Variablen, die zu sehr korreliert sind, werden aus dem Modell entfernt, da sie ansonsten die Ergebnisse verzerren könnten.

Ergebnisse

Das Cox-Modell liefert eine Parameterschätzung für jede Kovariable. Hierbei lässt sich sowohl ablesen, ob sich das Risiko erhöht oder verringert, als auch zu welchem Faktor dies geschieht.
Der Koeffizient für “Geschlecht” ist negativ, was bedeutet, dass das Sterberisiko für Frauen geringer ist als das für Männer. Der Faktor für diese Variable ist 0.62, sodass das Risiko für eine Frau nur 0.62-mal so groß ist wie das eines Mannes, wenn alle anderen Kovariablen gleich sind.
Rauchen hingegen hat einen positiven Koeffizienten und erhöht somit das Risiko. In diesem Fall ist das Risiko ungefähr doppelt so groß für Raucher im Vergleich zu Nicht-Rauchern.
Auch die Variable “Arbeitslos” hat einen positiven Koeffizienten und das Risiko für Arbeitslose ist ungefähr 1.74-mal so groß wie das Risiko für Nicht-Arbeitslose.
Das monatliche Haushaltsnettoeinkommen hat einen senkenden Effekt auf das Sterberisiko mit zugehörigem Faktor 0.93. Das bedeutet, dass eine Erhöhung des monatlichen Haushaltsnettoeinkommens um Tausend Euro einer Verringerung des Risikos um den Faktor 0.93 entspricht.
Auch die Variablen in Bezug auf Versicherung, nämlich “Private Krankenversicherung” und “Private Zusatzkrankenversicherung” haben einen senkenden Einfluss auf das Risiko. Der Faktor ist jeweils ungefähr 0.71.
Das monatliche Bruttogehalt hat einen positiven Koeffizienten und damit erhöht ein höheres Gehalt das Risiko. Dies ist zunächst überraschend und widerspricht auf den ersten Blick auch dem Ergebnis für “Monatliches Haushaltsnettoeinkommen”. Dies lässt sich jedoch wahrscheinlich auf die Korrelation zurückführen. Das monatliche Bruttogehalt ist mit dem Geschlecht korreliert, sodass sich die Interpretation für diese Variable ändert. Ein höheres Gehalt ist
in diesem Datensatz also tendenziell bei Männern zu sehen, die wiederum ein erhöhtes Risiko haben. Somit ist der Schätzer für das monatliche Bruttogehalt in diesem Modell positiv, obwohl das Gehalt selbst vermutlich eher einen senkenden Effekt auf das Sterberisiko
hat, wie der Schätzer für das monatliche Haushaltsnettoeinkommen zeigt. Von diesem sind nämlich Männer und Frauen im gleichen Maße betroffen, sodass das monatliche Haushaltsnettoeinkommen in diesem Modell die geeignetere Variable ist. Die letzte Variable “Verheiratet” ist nicht signifikant, sodass die Interpretation hierfür nicht aussagekräftig
ist.

Da immer noch gewisse Abhängigkeiten zwischen den Variablen herrschen, ist bei der Interpretation nicht klar, welche Variable der

Haupttreiber für das Sterberisiko ist. Beispielsweise ist bei “Private Krankenversicherung” nicht klar, ob der Einfluss daher ruht, dass privatversicherte Personen ein größeres Bewusstsein für Krankheiten und dadurch beispielsweise für Vorsorgetermine bei Ärzten haben, oder ob sich der Einfluss auf die Korrelation mit dem Einkommen zurückführen lässt. Auch beim Einkommen ist nicht eindeutig, ob ein höheres Einkommen zu einer besseren medizinischen Versorgung führt, oder ob ein höheres Einkommen lediglich mit einem höheren Bildungsniveau und einem gesunden Lebensstil verknüpft ist. Geld allein schützt schließlich auch nicht vor allen Krankheiten oder Unfällen.

Natürlich muss bei dem Modell auch überprüft werden, inwiefern die zugrundeliegenden Annahmen erfüllt sind. Eine wichtige Annahme ist die der nicht-informativen Zensierung. Für diese Analyse bedeutet das, dass die Austrittswahrscheinlichkeit einer Person unabhängig von der Sterbewahrscheinlichkeit sein muss. Entscheiden sich beispielsweise viele Personen krankheitsbedingt dafür, aus der Studie auszutreten, würde die Sterblichkeit generell unterschätzt.
Die vermutlich schwerwiegendsten Annahmen dieses Modells sind die Multiplikativitäts- und Proportionalitätsannahme des Cox-Modells. Es lässt also weder Interaktionseffekte noch Effekte in Abhängigkeit der Zeit zu. Wenn also beispielsweise das Rauchen eher für Männer als für Frauen gesundheitlich bedenklich ist oder das Risiko des Rauchens vom
Alter abhängt, würde das Modell dies nicht zulassen. Das Modell gibt lediglich einen einzigen Schätzwert für das Risiko des Rauchens. Bekannte Interaktionen können aber durch Feature-Engineering abgebildet werden.

Fazit

In dieser Arbeit wurden statistische Verfahren der Survival Analysis auf Daten des SOEPs angewandt. Das Augenmerk lag hierbei auf der Analyse von Sterblichkeit in Abhängigkeit sozioökonomischer Faktoren. Zu den wesentlichen Einflüssen auf die Sterblichkeit zählen in diesem Modell das Geschlecht, wobei Frauen ein geringeres Risiko haben, Rauchen und Arbeitslosigkeit mit einem erhöhten Risiko, sowie das monatliche Haushaltsnettoeinkommen, wobei ein höheres
Einkommen das Risiko senkt. Diese Resultate sind intuitiv und konnten dank des Cox-Modells auch quantifiziert werden. Darauf aufbauend lassen sich diverse Erweiterungen implementieren, die zum Teil auch in der Arbeit thematisiert wurden.