statistische Verfahren WS 2017/2018

Projekt 7 - Kriminalität

Reda Ihtassine (155 685) Ingo Schäfer (165 220)

Jena, am 27. März 2018

Inhaltsverzeichnis

1 Einleitung **1**

2 Material und Methoden **2**
2.1 Material 2
2.2 Methoden 3

3 Resultate **6**
3.1 Modellwahl 6
3.2 Simulationsaufgabe 8

4 Diskussion **9**

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Statistiken sind ein wichtiges Mittel, um die Werte und Trends der Kriminalität zu schätzen, die Kosten für Auswirkungen auf die Gesellschaft zu bewerten und darüber die Strafverfolgungsansätze zu optimieren, um die Kriminalität im folgenden zu verhindern. Um ein ökonomisches Kriminalitätsmodell zu schätzen, können die Eigenschaften der Counties nicht ignoriert werden. Diesem Projekt liegen solche Daten des US-amerikanischem Bundesstaats North Carolina zugrunde, welche in dem Zeitraum von 1981 bis 1987 erhoben wurden. Sie wurden u.a. in der Arbeit von Baltagli[1] sowie von Cornwell und Trumbull[2] veröffentlicht.

Der übliche Hausman-Test, der auf dem Unterschied zwischen fixierten und zufälligen Effekten basiert, kann zu einer irreführenden Inferenz führen, wenn es endogene Regressoren des konventionellen simultanen Gleichungstyps gibt[1]. Daher ist es das Ziel dieser Projektarbeit ein geeignetes statistisches Modell für die Zahl der Verbrechen mithilfe von anderen Kriterien zu entwickeln. Dabei betrachten wir insbesondere die qualitative Einflussgröße *region* und deren mögliche Wechselwirkungen mit anderen Prädiktoren. Der zweite Teil dieser Arbeit beschäftigt sich mit der Untersuchung des Einflusses des Stichprobenumfangs auf die Genauigkeit der Approximation der tatsächlichen Kovarianzmatrix des Maximum-Likelihood-Schätzers durch die asymptotische Kovarianzmatrix.

Diese Arbeit gliedert sich in drei Kapitel: Im Kapitel Material und Methoden wird zunächst das Material aus der Datei *crimes.csv* und die verwendeten Methoden beschrieben. Im Kapitel Resultate werden die numerischen Ergebnisse vorgestellt. Im letzten Kapitel erfolgt die Diskussion und Interpretation der Ergebnisse hinsichtlich der Aufgabenstellung und der praktischen Anwendbarkeit der ausgewählten Modelle.

[1] Vgl.: Badi H. Baltagli, estimating an economic model of crime using panel data from North Carolina, journal of applied econometrics, S.: 543

[2] Vgl.: Cornwell C, Trumbull WN. 1994. Estimating the economic model of crime with panel data. Review of Economics and Statistics 76: 360 - 366.

2 Material und Methoden

2.1 Material

Cornwell und Trumbull (1994)[2], schätzten ein Wirtschaftsmodell der Kriminalität unter Verwendung von Daten aus 90 Counties in North Carolina zwischen 1981 und 1987. Becker (1963) und Ehrlich (1973)[3] unter anderem folgen dem empirische Modell, welches die Kriminalitätsrate misst und sich dabei auf eine Reihe von Variablen bezieht. Dazu gehören auch solche Variablen, wie z.B. die Angst eine Straftat zu begehen oder auch Variablen, die messen wie oft der Täter danach wieder straffrei geblieben ist. Diese Kriminalitätsrate ist ein FBI-Index, der das Verhältnis zwischen Anzahl der Verbrechen und der Kreisbevölkerung berechnet[4].
In dieser Arbeit jedoch werden nicht alle diese Daten zur Ermittlung eines geeigneten Modells genutzt. Hier folgt eine Beschreibung des Datensatzes:

Der Datensatz ist in einer .csv-Datei gespeichert. In ihr sind die unterschiedlichen 90 Counties von North Carolina zeilenweise aufgelistet. Die Spalten sind (mögliche) Eigenschaftsvektoren. Alle Eigenschaftsvektoren sind logarithmisch mit Ausnahme der Region, die eine Dummy-Variable ist.
Die erste Spalte beinhaltet die Zielgröße *crimes*, also die Anzahl aller Straftaten in dem jeweiligen County über den Zeitraum von 1981-1987.

Weiterhin wurde die Arrestwahrscheinlichkeit P_A hinzugefügt. Sie berechnet sich aus

$$P_A = \frac{\text{Arrestierungen}}{\text{Delikte}} \tag{1}$$

Sie wird abgekürzt *prbarr* geschrieben.

Daneben gibt es auch die Überzeugungswahrscheinlichkeit P_C. Sie gibt das Verhältnis zwischen tatsächlichen Arrestierungen und den gestandenen Straftaten an und wird daher berechnet mit

$$P_C = \frac{\text{Anzahl tatsächlicher Arrestierungen}}{\text{Anzahl gestandener Straftaten}} \tag{2}$$

Sie wird bezeichnet als *prbpris*.

Eine weitere Eigenschaft ist die Fähigkeit des Countys ein Verbrechen auch zu ermitteln. In dem Datensatz spiegelt sich dies in der Variable *polpc* wieder. Sie gibt das Verhältnis zwischen Anzahl der Polizisten zu der Bevölkerungsanzahl an.

[3] Ehrlich I. 1973. Participation in illegitimate activities: a theoretical and empirical investigation. Journal of Political Economy 81: 521â567.
[4] Vgl.: Badi H. Baltagli, estimating an economic model of crime using panel data from North Carolina, journal of applied econometrics, S.: 543 f.

Ein weiteres wichtiges Merkmal ist die Bevölkerungsdichte (*density*). Sie stellt das Verhältnis zwischen der Bevölkerungsanzahl und der Fläche des Countys (in square miles) dar.

Darüberhinaus wird das Verhältnis von Minderheiten zu der Gesamtanzahl Einwohner in der Variable *pctmin* ausgedrückt.
pctymale ist eine Eigenschaft, die den Anteil der jungen männlichen Bevölkerung zur Gesamtbevölkerung anzeigt.

Die letzten fünf Variablen geben den durschnittlichen Bruttolohn in den Bereichen Baugewerbe (*wcon*), Staatsangestelle (*wsta*), Dienstleistungssektor (*wser*), Handel (*wtrd*) und Bankgeschäften (*wfir*) wieder.

2.2 Methoden

Um ein geeignetes Modell aus den oben beschriebenen Merkmalen zu finden, wurden fünf unterschiedliche Herangehensweisen vorgeschlagen, um ein Modell zu finden, das möglichst geringe Fehler aufweist.

- explorative Herangehensweise (ausprobieren)
- Vergleich aller Modelle mit nur einem Merkmal
- Verwendung von step() und anschließende Minimierung des Modells
- strukturierte Suche nach einem geeigneten Modell
- Verwendung von cor()

Am Ende einer jeden Herangehensweise wurde ein bestes Modell vorgschlagen. Diese wurden dann anschließend miteinander verglichen, um ein bestmögliches Modell zu bestimmen.
Hauptsächlich wurden zwei Gütekriterien verwendet, um verschiedene Modelle miteinander vergleichen zu können.

Zum einen *Akaikes Information Criterion* (AIC), welches die logarithmische Fehlerabweichung des Schätzers $ln(\hat{\underline{\Theta}}_n)$ mit der Anzahl der verwendeten Merkmale p bestraft.

$$\text{AIC} := -2 * ln(\hat{\underline{\Theta}}_n) + 2p \tag{3}$$

Je kleiner also der erhaltene Wert ist, desto besser sei das untersuchte Modell.
Der Faktor 2, der hier in Formel (3) auftritt, kann mit einem beliebigen Wert $n, n \in N$ belegt werden. In dieser Arbeit wurde er allerdings dauerhaft auf 2 belassen.
AIC spiegelt den Kompromiss zwischen Verbesserung der Modellanpassung durch erhöhte p und erhöhte Ungenauigkeit durch Schätzung vieler Parameter wider.

In einigen Fällen wurde auch die *Devienz* betrachtet, um die Güte mehrerer Modelle miteinander zu vergleichen.
Hier geht man von einem saturierten Modell aus. Dies ist das komplexeste Modell für einen Datensatz, dass durch Erhöhung der Parameterzahl erzeugt werden kann. In vielen Fällen hat das saturierte Modell daher so viele Parameter wie Beobachtungen. Falls Einflusvektoren mehrfach vorkommen, besitzt das saturierte Modell weniger Parameter. Das ist typischerweise der Fall für Experimente mit qualitativen Einflussgrößen.
Hier wird die Likelihood-Quotienten-Statistik zum Vergleich eines Modells M mit dem saturierten Modell

$$T(\underline{Y}) = 2(l_{\text{saturiert}} - l_M) \tag{4}$$

betrachtet.
Die Likelihood-Quotienten-Statistik ist asymptotisch $\mathcal{X}^2$ - verteilt. Dabei ist r die Differenz der Parameterzahlen. Deswegen funktioniert hier der Likelihood-Quotienten-Test nicht, da für $n \to \infty$ die Anzahl der Freiheitsgrade auch typischerweise unbeschränkt wächst.
Die Größe

$$D(M) = 2(l_{\text{saturiert}} - l_M) \tag{5}$$

heißt Devienz des Modells M.
Dabei ist zu beachten, dass ein Modell M ein geeignetes Modell ist, falls die Devienz von M ungefähr so groß ist wie die ungefähre Anzahl Parameter von M.

$$D(M) \approx n - |M| \tag{6}$$

Als anderes Gütekriterium wurde das Quadrat der erwarteten Fehlerabweichungen (*SPSE*) im Kreuzvalidierungsverfahren berechnet.
Dazu wurde der gesamte ausgewählte Datensatz in einen Trainings- und einen Testdatensatz aufgeteilt. Das "beste Modell" ist dasjenige, dass im Mittel den kleinsten geschätzten erwarteten Prognosefehler liefert. Dabei wird typischerweise eine l-fache Kreuzvalidierung durchgeführt:
Es gibt einen Testdatensatz $I = 1...n$. Dieser wird in l etwa gleichgroße Indexmengen $I_1, ..., I_l$ zerlegt.
In jedem j-ten Schritt wird ein I_j als Testdatensatz gewählt. Alle anderen Indexmengen bilden den Trainingsdatensatz.
Nun wird der erwartete Prognosefehler geschätzt:

$$\sum_{i \in I_j} (y_i - \underline{x}_i^{(M)^T} \hat{\underline{\beta}}^{(m-j)})^2 = \hat{SPSE}_j^{(M)} \tag{7}$$

Dabei ist $\hat{\underline{\beta}}^{(m-j)}$ die auf I/I_j basierende Schätzung.
Zuletzt werden alle Teilschätzungen zu einer Schätzung für SPSE zusammen kombiniert:

$$\hat{SPSE}^{(M)} := \sum_{j=1}^{l} (\hat{SPSE}_j^{(M)}) \tag{8}$$

Zu bemerken ist, dass jede Beobachtung einmal in einem Testdatensatz verwendet wird. Außerdem ist die Abhängigkeit von der konkreten Zerlegung nur reduziert, aber nicht verschwunden. Es gibt einen Spezialfall, wenn $l = n$. Das heißt, dass der gestamte Testdatensatz in n Teildatensätze zerlegt wird. Jede Beobachtung wird mit der Prognose basierend auf ($n-1$ Beobachtungen verglichen. Dies ist auch bekannt als *leave-one-out-cross-validation*. Als Faustregel empfiehlt es sich $l \approx 10$ zu wählen.

In der Simulationsaufgabe des Projektes sollte der Einfluss des Stichprobenumfangs auf die Genauigkeit der Approximation der tatsächlichen Kovarianzmatrix des Maximum-Likelihood-Schätzers durch die asymptotische Kovarianzmatrix untersucht werden. Dazu wurde zunächst ein möglichst einfaches wahres Modell angenommen. Anhand dessen wurde aus dem gesamten Datensatz eine beliebig große Teilmenge T entnommen. Aus den Daten von T wurde dann eine Designmatrix gebildet, die Grundlage für die darauf folgendenden Generierungen für Pseudozufallszahlen war. Standarmäßig wurde die Größe dieser zu untersuchenden Teilmenge auf 30 gesetzt. (Der gesamte Datensatz umfasst 90 Subjekte.) Jedoch wurden auch viele andere Größen überprüft. Aus der auf diese Art und Weise gebildeten Designmatrix, wurden nun wiederum unterschiedlich große Stichproben ausgewählt und die daraus berechneten β_0 und β_1 Werte in einer weiteren Matrix gespeichert. Aus dieser Matrix wurden dann die Varianz und die Kovarianz für die tatsächliche Kovarianzmatrix berechnet.

Da

$$F^{\frac{T}{2}}(\underline{\hat{\beta}})(\underline{\hat{\beta}}_n - \underline{\beta}) \xrightarrow[n\to\infty]{d} N(0, I) \tag{9}$$

gilt, gilt für die Approximation von $\underline{\hat{\beta}}_n$ bei festem n:

$$\underline{\hat{\beta}}_n \approx N(\underline{\beta, I^{-1}(\beta)}) \tag{10}$$

Die Kovarianzmatrix $\mathbb{X}$ ist die inverse Fisher-Matrix I. Daher hat $I(\underline{\beta})$ die kanonische Linkfunktion einfachen Gestalts

$$I(\underline{\beta}) = \mathbb{X}^T V \mathbb{X} \tag{11}$$

Dabei ist V eine Diagonalmatrix, welche in der Spur die Varianzen hält. Anhand dessen wurde die asymptotische Kovarianzmatrix berechnet. Die daraus herausgehenden Resultate wurden dann bei einem kleiner werdenden n auch immer geringer, sodass die Ergebnisse immer in Relation zueinander verglichen wurden.
Daher hat es sich angeboten eine Hilfsfunktionen `simulation()` zu schreiben, welche eine solche Simulation durchführt.
Außerdem wurde eine weitere Hilfsfunktion `compare()` geschrieben, die zwei solche Simulationen in Relation zueinander vergleicht.

3 Resultate

3.1 Modellwahl

Wie bereits erwähnt, wurden fünf unterschiedliche Herangehensweisen betrachtet, um ein geeignetes Modell zu finden.

Wahl der Verteilung negative binomialverteilung statt gauÃ-verteilung, begrÃ¼ndug - siehe quelle!

Die besondere Rolle von der Einflussgröße region

Herangehensweisen

explorative Herangehensweise Um ein gutes Gefühl für die Merkmalsvektoren zu bekommen, wurden zunächst einige Modelle ausprobiert und mittels AIC verglichen. Damit ein Vergleichswert nach dem Akaike-Maß vorhanden war, wurde ein komplettes Modell angenommen, das aus allen vorhandenen Merkmalen besteht. Dieses Modell heißt *mAll*. Die entsprechende Formel sieht so aus:

$$crimes = prbarr+prbpris+polpc+density+area+taxpc+region+pctmin+pctymale+wcon+wsta+ws \tag{12}$$

Es wurde bewusst darauf verzichtet in diesem 'gesamten' Modell die Intersections (Wechselwirkungen) der einzelnen Merkmale zu betrachten. Grund dafür ist, dass das Akaike-Maß Modelle mit vielen Einflussgrößen mehr bestraft, als solche die weniger besitzen. Da bei dieser Untersuchung das Akaike-Maß das am häufigsten verwendete Kriterium war, sollte also das erste Modell, mit dem die anderen verglichen wurden, nicht einen großen negativen Wert aufweisen, so wie das in diesem Fall der Fall gewesen wäre. (Der Akaike-Wert des Modells, das alle Merkmale und alle Wechselwirkungen zwischen diesen betrachtet, beträgt -3441.465. In diesem Modell gibt es 91 Freiheitsgrade.) Die Daten `crimes.data`, welche der Funktion `glm.nb(formula, data = crimes.data)` während der gesamten Untersuchung gegeben wurden, wurden nicht verändert. Es handelt sich hierbei immer um den gesamten Datensatz aus der Datei *crimes.csv*.

Im Folgenden wurde bemerkt, dass diese Merkmale durchaus gruppiert betrachtet werden können. Daher bestand die erste Idee darin, die unterschiedlichen Gruppierungen je Modell zu betrachten: Die ersten beiden Merkmale (*prbarr* und *prbpris*) geben beide Verhältnisse zum Anteil aller Straftäter in einem County an. Daher wurde ein Modell aus diesen beiden Einflussgrößen betrachtet.

$$crimes = prbarr : prbpris \tag{13}$$

Die Einflussgrößen *density* und *area* sind beides räumliche Merkmale. Auch sie wurden in einem Modell zusammengefasst. Wie in 3.1

Vergleich aller Modelle mit jeweils nur einem Merkmal

Verwendung von step() und anschlieÃende Minimierung des Modells

strukturierte Suche nach einem geeigneten Modell

Verwendung von cor()

Die Gewinnermodelle

3.2 Simulationsaufgabe

Beschreibung simulation()

Auswertung der Ergebnisse

einfaches Modell: mDensity

Ergebnisse mit Gewinnermodell aus der ersten Aufgabe Hier leite ich zur Diskussion über.

4 Diskussion

- 'sieger'modelle sind recht gut, aber immer noch sehr groÃe abweichungen zu den tatsÃchlichen werten. -

Literatur