statistische Verfahren WS 2017/2018

Projekt 7 - Kriminalität

Reda Ihtassine (Matrikelnummer) Ingo Schäfer (165 220)

Jena, am 26. März 2018

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Dem Projekt liegen KriminalitÃtsdaten des US-amerikanischem Bundesstaats North Carolina zugrunde, welche in dem Zeitraum von 1981 bsi 1987 erhoben wurden. Diese Daten wurden schon mehrfach mit Hilfe von verschiedenen Methoden (Hausmann, 2SLS, ...) von Anderen untersucht [1].
Diese Arbeit versucht ein geeignetes allgemeines lineares Modell zu erarbeiten, mit dem gute AbschÃtzungen erzielt werden kÃ¶nnen.

[1] 1

2 Material und Methoden

2.1 Material

Der Datensatz besteht aus einer .csv-Datei. In ihr sind die unterschiedlichen 90 Counties von North Carolina zeilenweise aufgelistet. Die Spalten sind (mögliche) Eingeschaftsvektoren. In der Arbeit von Baltagli [2] werden noch einige Eigenschaften mehr aufgelistet, als in dieser Arbeit betrachtet wurden. Daher hier eine kleine Übersicht Ã¼ber alle mögichen Einflussgrößen:
Alle Eigenschaftsvektoren sind logarithmisch mit Ausnahme der Region und der Zeit. 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 $P_A = \frac{\text{Arrestierungen}}{textDelikte}$. Sie wird abgekürzt *prbarr* geschrieben. Daneben gibt es auch die Überzeugungswahrschleinlichkeit P_C. Sie gibt das Verhältnis zwischen tatsächlichen Arrestierungen und den gestandenen Straftaten an und wird daher berechnet mit $P_P = \frac{\text{Anzahl tatsächlicher Arrestierungen}}{\text{Anzahl gestandener Straftaten}}$. Sie wird bezeichnet als *prbpris*.
Eine weitere Eigenschaft ist die FÃ¤higkeit des Countys ein Verbrechen auch zu ermitteln. In dem Datensatz spiegelt sich dies in der Variable *polpc* wieder. Sie gibt das Polizei-pro-Kopf-VerhÃ¤ltnis an.
Ein weiteres wichtiges Merkmal ist die Bevölkerungsdichte (*density*). Sie stellt das VerhÃ¤ltnis $frac$anzahl bevölkerungFläche des Countys in square miles dar.
DarÃ¼berhinaus wird das Verhältnis von Minderheiten zu der Gesamtanzahl Einwohner in der Variable *pctmin* ausgedrÃ¼ckt.
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 (textitwfir) 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 mÃ¶glichst geringe Fehler aufweist.

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

[2] 1

- Verwendung von cor()

Am Ende einer jeden Herangehensweise wurde ein bestes Modell vorgschlagen. Diese wurden dann anschlieÃend miteinander verglichen, um ein bestmÃ¶gliches Modell zu bestimmen.

Hauptsächlich wurden zwei Gütekriterien verwendet.

Zum einen *Akaikes Information Criterion* (AIC), welches die logarithmische Fehlerabweichung des Schätzers mit der Anzahl der verwendeten Merkmale bestraft.

$$\text{AIC} := -2 * ln(\hat{\underline{\Theta}}_n) + 2p \tag{1}$$

AIC spiegelt den Kompromiss zwischen Verbesserung der Modellanpassung durch erhÃ¶hte Parameteranzahl und erhÃ¶hte Ungenauigkeit durch SchÃtzung vieler Parameter wieder.

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

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

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

$$D(M) = 2(l_{\text{saturiert}} - l_M) \tag{3}$$

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.

$$D(M) \approx n - |M| \tag{4}$$

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:

$$\sum_{i \in I_j} (y_i - \underline{x}_i^{(M)^T} \hat{\underline{\beta}}^{(m-j)})^2 = \hat{SPSE}_j^{(M)} \tag{5}$$

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:

$$\hat{SPSE}^{(M)} := \sum_{j=1}^{l} (\hat{SPSE}_j^{(M)}) \tag{6}$$

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 fÃ¼r Pseudozufallszahlen war. Meistens wurde die Größe dieser zu untersuchenden Teilmenge auf 30 gesetzt. (Der gesamte Datensatz umfasst 90 Subjekte.) Jedoch wurden auch 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

$$F^{\frac{T}{2}}(\hat{\underline{\beta}})(\hat{\underline{\beta}}_n - \underline{\beta}) \xrightarrow[n \to \infty]{d} N(0, I) \tag{7}$$

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

$$\hat{\underline{\beta}}_n \approx N(\underline{\beta, I^{-1}(\beta)}) \tag{8}$$

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

$$I(\underline{\beta}) = \mathbb{X}^T V \mathbb{X} \tag{9}$$

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.

3 Resultate

1:
Vorgehensweise:
- negative binomialverteilung statt gauÃ-verteilung, begrÃ$\frac{1}{4}$ndug - siehe quelle! - 5 unterschiedliche Herangehensweisen um ein geeignetes Modell zu finden, alle kurz erklÃren
- besondere Rolle von *region*
- vergleich der modelle funktionsweise knapp erlÃutern(aic, cross$_v$$alidation, cor()$)
$- vorstellender5gewinnermodelle, dengewinner$
2 :
$- beschreibungtest() - funktion(ggffunktioneinengriffigererennamengeben)$
$- welcheeinstellungenerzielenguteergebnisse?$
$-einmalmiteinfachemmodellzeigen : mDensity...(< -warummDensity?)-einmalmitgewinner-modell - - > wiegutistergebnis?\frac{1}{4}berleitungzudiskussion...$

4 Diskussion