statistische Verfahren WS 2017/2018

Projekt 7 - Kriminalität

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

Jena, am 28. März 2018

Inhaltsverzeichnis

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 Wie Osgood in [Osgood2000][5] schreibt, ist es von Vorteil Poissonverteilungen zu nutzen, um Kriminalitätsraten zu analysieren. Poissonbasierte Regressionsmodelle sind bei Beobachtungen von Verbrechensdelikten eine gute Wahl, da sie anhand von Annahmen über Fehlerverteilungen gebaut werden, die mit der Art der Ereignisanzahl konsistent sind[6].
Osgood empfiehlt daher die Verwendung der negativen Binomialverteilung. Diese wurde von Poisson selber in den 1820-er Jahren entwickelt, um Verbrechen zu analysieren[7]. Daher wurde in dieser Arbeit nicht das `OLS`-Verfahren (ordinary least-squares) verwendet, welches eigentlich die Standardmethode in solchen Untersuchungen ist.
Eine Normalverteilung oder eine symmetrische Fehlerverteilung kann hier nicht angenommen werden, da die Verbrechensanzahl sehr gering sein kann. Die kleinstmögliche Anzahl an Verbrechen in einem County ist Null. Daher müsste eine Fehlerverteilung immer mehr verzerrt werden[8]. In Abbildung 1 sind 100 Zufallszahlen auf Basis der negativen Binomialverteilung mit $\theta = 0.7$ dargestellt.

Um die Annahme der negativen Binomialverteilung zu rechtfertigen wurden zu Beginn der Arbeit zwei Modelle verglichen, welche die selben Daten und die selbe Formel nutzen, aber unterschiedliche Verteilungen annehmen. Es wurde der gesamte Datensatz von allen 90 Counties genutzt. Die verwendete Formel war:

$$crimes = \beta_0 \cdot prbarr : prbpris + \beta_1 \tag{12}$$

Un zu entscheiden, welches der Modelle besser schätzt, wurden die Akaike-Werte der beiden Modelle miteinander verglichen. In Tabelle 1 finden sich diese Zahlen wieder. Der Akaikewert von Modell M1NB ist wesentlich geringer als der von m1. Dies sagt eigentlich nicht aus, dass M1NB ein besseres Modell als M1 ist. Vor allem aber ist die Devienz des ersten Modells viel größer, als die des binomialverteilten. Dies spricht für eine deutliche Verbesserung.

Während der Untersuchung wurden natürlich mehrere Gleichungen angenommen als diese eine. Die Ergebnisse haben aber alle zu diesem selben Schluss geführt. Eine graphische Veranschaulichung der Gegenüberstellung der beiden Verteilungen findet sich

[5] Vgl.: D.Wayne Osgood: Poisson-Based Regression Analysis of Aggregate Crime Rates, Journal of Quantitative Criminology, Vol. 16, No. 1, 2000
[6] [Osgood2000] S.21
[7] Maltz, M. D. (1994). Operations research in studying crime and justice: Its history and accomplishments. In Pollock, S. M., Rothkopf, M. H., and Barnett, A. (eds.), Operations Research and the Public Sector, Volume 6 of Handbooks in Operations Research and Management Science, North-Holland, Amsterdam, pp. 200 - 262.
[8] Vgl.: [Osgood2000] S. 21 f.

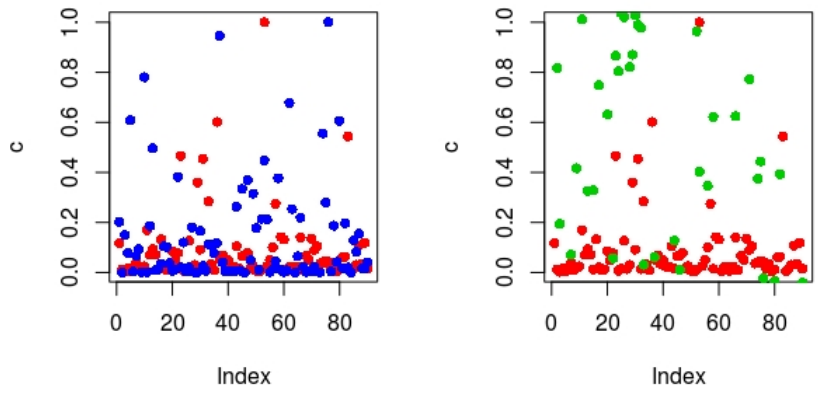

Abbildung 1: In beiden Diagrammen sind die relativierten Verbrechenszahlen rot dargestellt. In dem linken Diagramm wurden dazu in blau mit Hilfe von `rnegbin()` berechnete Zufallszahlen hinzugefügt. Das rechte Diagramm enthält stattdessen zusätzlich Zufallszahlen, die mit `rnorm()` berechnet wurden.

in Abbildung 1. Hier sieht man recht gut, dass die Verbrechenszahlen eher zu einer negativen Binomialverteilung passen, als zu einer normalen Gaußverteilung.

	df	AIC	Devienz
m1	3.00	1789.10	2118963611
m1Nb	3.00	1598.57	106.3874

Tabelle 1: Gegenüberstellung der Akaike-Werte zweier Modelle. m1 nimmt eine Gaußverteilung an, während m1Nb eine negative Binomialverteilung annimmt.

Die besondere Rolle von der Einflussgröße region Bei der Einflussgröße *region* handelt es sich um eine dichotome Dummy-Variable. Sie gibt an in welchem Bereich des Bundesstaates (= die Region) sich das County befindet. Es gibt drei mögliche Werte, welche diese Variable annehmen kann: *central*, *west* und *other*.

Anhand von Abbildung 2 sieht man, über welche Bereiche des Bundesstaates sich diese drei Regionen erstrecken.
Bei der Untersuchung dieser Einflussgröße ist aufgefallen, dass diese Variable wohl hinzugefügt wurde, um die stärker besiedelten Regionen des Bundesstaates zu kennzeichnen. So sind die Counties, die sich z.B in der Region *central* befinden wesentlich stärker bevölkert, als die Counties in der Region *west*. Die Region *other* ist flächenmäßig die größte, ihre Counties haben aber trotzdem eine größere Bevölkerungsdichte als die Counties der kleinsten Region *west*. Daher handelt es sich bei *west* wohl auch um eine stärker besiedelte Region. Dies ist deswegen interessant zu betrachten, da schon sehr früh in den

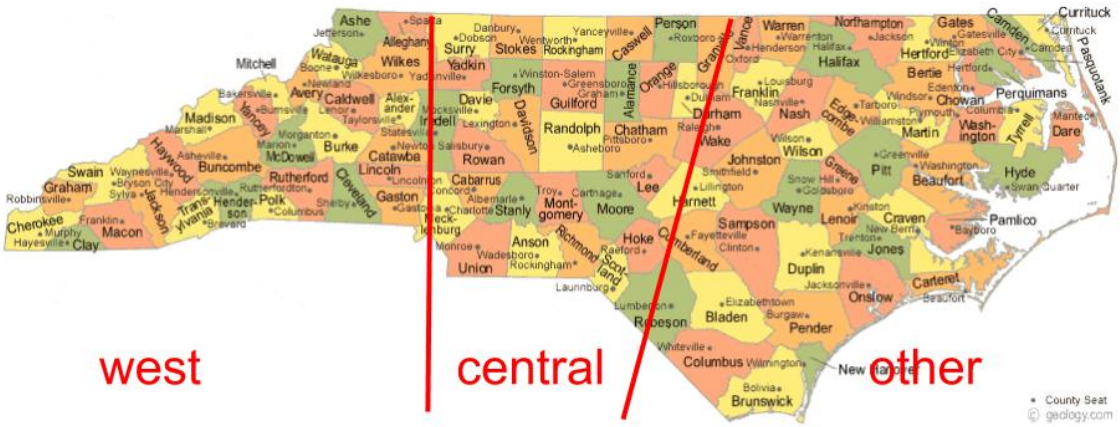

Abbildung 2: Diese Übersicht zeigt in etwa an wo sich die drei unterschiedlichen Regionen befinden [9]

Untersuchungen aufgefallen ist, dass die Variable *density* sehr gut dazu geeignet ist, ein gutes Modell zu finden.

region	Anzahl Counties	Durchschnitt *crimes*	Median *crimes*	*density*
central	34	4764	2172	196
west	21	1027	513	86
other	35	2250	1235	101

Tabelle 2: Vergleich der durchscnittlichen Eigenschaften der Counties aus den jeweiligen Regionen.

In Tabelle 2 befindet sich eine Gegenüberstellung der durchschnittlichen Werte von der Zielgröße *crimes* und der durchschnittlichen Bevölkerungsdichte (*density*) der drei Regionen. Bei der Betrachtung wird ersichtlich, dass es scheinbar einen Zusammenhang zwischen *region* und *density* gibt. In ländlichen Gegenden sind die *crimes*-Werte geringer. Befindet sich in dieser ländlichen Gegend aber eine größere Stadt, ist *crimes* auch erhöht.

In Abbildung 3 wird ersichtlich, dass es in der Region *central* einige Counties gibt, die wesentlich mehr Delikte gemeldet haben, als die Counties aus anderen Regionen. Dies scheint an der höheren Bevölkerungsdichte *density* zu liegen, wie die Darstellung 4 aufzeigt.

Tatsächlich wird sich herausstellen, dass diese Wechselwirkung zwischen *region* und *density* (also *density:region*) Bestandteil von vielen guten Modellen sein wird. Daher wurde diese Dummy-Variable nicht weiter verändert. R behandelt solche Variablen automatisch mithilfe von `as.factor()` als diskrete Zahlen.

Herangehensweisen

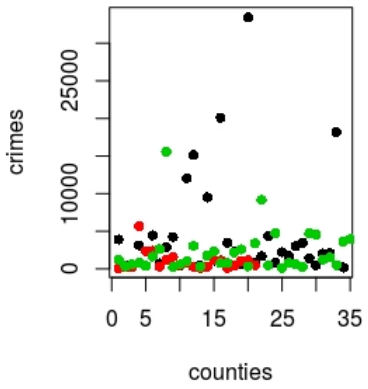

Abbildung 3: Anhand der Farbgebung lässt sich erkennen, in welchen Regionen die Counties wieviele Verbrechen gemeldet haben. Schwarz sind Counties aus der Region *central*, rot die Counties aus der Region *west* und grün die Counties aus der Region *other*.

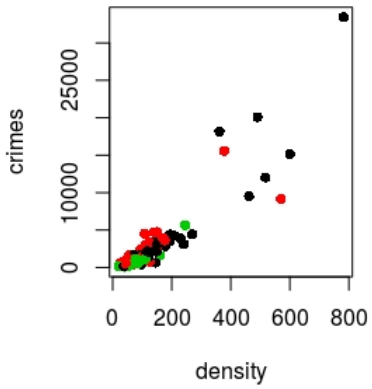

Abbildung 4: Hier wird das Veräl̈tnis *density* zu *crimes* aufgezeigt.

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:

$$\begin{aligned} crimes = \beta_0 \cdot prbarr + \beta_1 \cdot prbpris + \beta_2 \cdot polpc + \beta_3 \cdot density + \\ \beta_4 \cdot area + \beta_5 \cdot taxpc + \beta_6 \cdot region + \beta_7 \cdot pctmin + \\ \beta_8 \cdot pctymale + \beta_9 \cdot wcon + \beta_{10} \cdot wsta + \beta_{11} \cdot wser + \\ \beta_{12} \cdot wtrd + \beta_{13} \cdot wfir \end{aligned} \tag{13}$$

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 wurden Modelle aus diesen beiden Einflussgrößen betrachtet. Das Modell m1 hat den Formelaufruf glm.nb(formula = crimes 1 + prbarr:prbpris. Modell m2 besitzt die Formel glm.nb(formula = crimes (1 + prbarr + prbpris)^2).$DieAIC - WertedieserModellesindinTabelle??einsehbar.WieerwartetsagenhieralleKriterienhnlichesaus$: $DieseModellem1undm2sindsichetwashnlich, wobeidasetwaskomplexereModellm2etwasbesserabschn$

	df	AIC	SPSE	Devienz
mAll	17.00	1432.29	3003095120	92.88015
m1	3.00	1598.57	3003572113	106.3874
m2	5.00	1580.94	3003529863	103.9891

Auffällig in Abbildung 1 ist, dass beide Wahrscheinlichkeiten sich gewissermassen ähnlich verhalten: Die meisten Wahrscheinlichkeiten haben nur relativ geringe Verbrechensanzahlen. Jedoch gibt es bei beiden Wahrscheinlichkeiten zwei Bereiche, in denen relativ ohe Verbrechensanzahlen vorliegen. Für die Arrestierungswahrscheinlichkeit ist dies das Intervall $[0.15, 0.3]$ und für die Verurteilungswahrscheinlichkeit

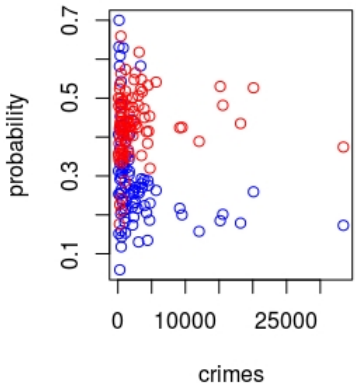

Abbildung 5: Die Arrestierungswahrscheinlichkeiten sind blau gefärbt, während die Verurteilungswahrscheinlichkeiten rot gefärbt sind.

liegt dieser Peak in $[0.4, 0.55]$. Dies sind relativ geringe Wahrscheinlichkeiten. Trotzdem ist es interessant, dass die Counties, in denen es sehr viele gemeldete Delikte gab, sich in solchen Intervallen sammeln.

Die Einflussgrössen *density* und *area* sind beides räumliche Merkmale. Auch sie wurden in einem Modell zusammengefasst. Wie in Kapitel 3.1 bereits erwähnt wurde, liefern Modelle, welche die Variablen *density* verwenden mit die besten Ergebnisse. In diesem Zusammenhang wurde auch das dichotome Merkmal *region* mit zu dieser Grupperiung gezählt. Folgende Modelle, die in Liste ?? aufgeführt sind, wurden betrachtet. Die Ergebnisse dieser Untersuchung liegen in Tabelle ?? dar. Auch hier ist die selbe Beobachtung wieder zu machen, dass, mit Zunahme der Komplexität der Modelle, sich die Kriterien verbessern.

- mSpatial1: glm.nb(crimes (density+area))
- mSpatial2: glm.nb(crimes (density+area)^2)mSpatial3: glm.nb(crimes (density+ari
- mSpatial4: glm.nb(crimes (density+area+region)^2)

	df	AIC	SPSE	Devienz
mSpatial1	4.00	1462.89	3003028556	95.0647
mSpatial2	5.00	1463.57	3003008479	95.00924
mSpatial3	6.00	1460.37	3003031244	94.74483
mSpatial4	11.00	1444.15	3003082475	93.6144

Als dritte Gruppierung besteht aus *pctymin* (der Anteil von Minderheiten an der Gesamtbevölkerung) und *pctymale* (der Anteil der jungen

männlichen Bevölkerung (15-24 Jahre)) . In der vierten Gruppierung wurden alle Löhne miteinander kombiniert. Hier wurden jedoch keine weiteren Untersuchungen an der Komplexitätszunahme der Modelle gemacht. Grund dafür ist, dass anhand der ersten beiden Gruppierungsbetrachtungen geschlussfolgert wurde, dass komplexere Modelle auch immer die besseren Akaike-, SPSE- und Devienzwerte haben. Ausnahme hierbei ist manchmal die Summe Ã¼ber die erwarteten Fehlerquadrate. Stattdessen wurden in diesen beiden Gruppierungen nur die jeweils komplexesten Modelle betrachtet, wie in Liste ?? und Tabelle ?? entnommen werden kann.

- mPct: `glm.nb(crimes (pctmin+pctymale)^2)` $mTrade : glm.nb(crimes \ (1 + wsta + wser + wtrd +$

	df	AIC
mPct	5.00	1612.85
mTrade	12.00	1541.39

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

$$[t]0.5 = 30/\text{image}_30 - 4/Rplot.pdf = 30/image_30 - 4/Rplot.eps \quad [t]0.5 = 30/\text{image}_30 - 4/Rplot01.pdf = 30/image_30 - 4/Rplot01.eps$$

$$[t]0.5 = 30/\text{image}_30 - 5/Rplot.pdf = 30/image_30 - 5/Rplot.eps \quad [t]0.5 = 30/\text{image}_30 - 5/Rplot01.pdf = 30/image_30 - 5/Rplot01.eps$$

$$[t]0.5 = 30/\text{image}_30 - 8/Rplot.pdf = 30/image_30 - 8/Rplot.eps \quad [t]0.5 = 30/\text{image}_30 - 8/Rplot01.pdf = 30/image_30 - 8/Rplot01.eps$$

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