Statistische Analyse von Kriminalitätsdaten aus North Carolina

Ausarbeitung im Kurs: Statistische Verfahren

eingereicht von

Maximillian Enderling, Max Möbius und Henning Woydt

Jena, September 28, 2021

Kurzzusammenfassung

Dieser Artikel betrachtet einen Datensatz über die möglichen Einflüsse von Verbrechen in North Carolina. Für jeden der 90 Counties wurde vom Jahr 1981 bis 1987 verschiedenste Einflüsse gemessen und datiert. Das Hauptaugenmerk des Datensatzes liegt dabei auf der absoluten Anzahl an Verbrechen (`crimes`). Außerdem enthält der Datensatz diverse weitere Variablen, die einen möglichen Einfluss auf diese Zielgröße haben. In diesem Artikel werden verschiedenste statistische Verfahren genutzt, um anhand von generellen linearen Modellen die Beziehungen zwischen den Einflussgrößen und der Zielgröße zu analysieren. Modelle werden aufgestellt, durch Likelihood-Quotienten-Tests auf ihre Legitimität überprüft und anschließend durch Kreuzvalidierung auf ihre Anpassung an den gegebenen Datensatz überprüft. Im zweiten Teil dieses Artikels werden zwei Methoden zur Bestimmung von Konfidenzintervallen verglichen. Dabei wird ein wahres Modell angenommen. Auf der Grundlage von Pseudobeobachtungen wird ein weiteres Modell geschätzt. Aus diesem werden die Konfidenzintervalle geschätzt und es wird überprüft wie häufig die wahren Koeffizienten in den Konfidenzintervallen liegen. Es wird hierbei R's *confint* Methode und eine Berechnung über die beobachtete Fisher-Informations-Matrix verglichen.

1 Einführung

Kriminalitätsbekämpfung war und ist eine wichtige Aufgabe des Staates. So ist es wichtig zu wissen, welche Faktoren einen Einfluss auf die Kriminalität haben.

Einfluss klingt nach Kausalität

1.1 Einführung in den Datensatz

Der Datensatz ist eine minimale Abwandlung des Datensatzes von RePEc:tpr:restat:v:76:y:1994:i:2:p:360-66 [**?**]. Er handelt von der absoluten Anzahl an Verbrechen in 90 Counties von North Carolina vom Jahr 1981 bis 1987. Der Datensatz enthält die Größen:

- *crimes*: Die absolute Anzahl an Verbrechen
- *prbarr*: Die Wahrscheinlichkeit, dass ein Verbrecher festgenommen wird
- *prbpris*: Die Wahrscheinlichkeit, dass ein festgenommener Verbrecher eine Gefängnisstrafe erhält
- *polpc*: Die Anzahl an Polizisten pro Kopf
- *density*: Die Anzahl an Einwohnern pro Quadratmeile
- *area*: Die flächenmäßige Größe des Counties in Quadratmeilen
- *taxpc*: Das Steueraufkommen pro Kopf
- *region*: Aufteilung in die übergeordneten Regionen „west“, „central“ und „other“.
- *pctmin*: Die prozentuale Anzahl an Minderheiten an der Gesamtbevölkerung
- *pctymale*: Die prozentuale Anzahl an jungen Männern (15 bis 25 Jahren) an der Gesamtbevölkerung
- *wcon*: Das durchschnittliche Gehalt pro Woche im Baugewerbe
- *wsta*: Das durchschnittliche Gehalt pro Woche der Staatsangestellten
- *wser*: Das durchschnittliche Gehalt pro Woche im Dienstleistungssektor
- *wtrd*: Das durchschnittliche Gehalt pro Woche im Handel
- *wfir*: Das durchschnittliche Gehalt pro Woche im Finanz-, Versicherungs- und Immobiliensektor

Die Größen *wtuc* (wöchentlicher Lohn im Transport, in der Versorgung und der Kommunikation), *wmfg* (wöchentlicher Lohn im Manufakturgewerbe), *wfed* (wöchentlicher Lohn der Bundesangestellten), *wloc* (wöchentlicher Lohn der lokalen Regierungsangestellten) und *urban* (Beteiligung am Rechtswesen) sind in diesem Datensatz nicht aufgeführt.

2 Modellwahl

Wir gehen für unsere Modelle von Poisson verteilten Zufalsgrößen aus. Dies liegt zum einem daran, dass die Zielgröße *crimes* diskret und nicht kontinuierlich ist. Zudem ist die Zielgröße immer positiv, die Poisson-Verteilung ist auch nur für positive Werte definiert, sodass eine Schätzung von negativen Werten nicht möglich ist. **TODO: Paper zitieren**

Das unterliegenden statistische Modell setzt sich aus dem stochastischen (1) und deterministischen (2) Teil zusammen.

$$Y_i \sim P(\lambda_i)\ ,\ i = 1, ..., 90\ ,\ \text{stochastisch unabhängig} \tag{1}$$

$$\mathrm{E}Y_i = \lambda_i = exp(\underline{x}_i^T \underline{\beta}) \tag{2}$$

Die Annahme dass die Zufallsgrößen stochastisch unabhängig sind, ist in der Regel nicht gegeben. Dafür müsste gelten, dass begangene Straftaten in einem County unabhängig von begangene Straftaten in

einem anderem County sind. Dies ist aber nicht gegeben, da es beispielsweise Verbrecher auf Raubzügen gibt, die in mehreren Counties stattfinden oder Verbrecher nah an einer Grenze zwischen den verschiedene Counties verkehren können.

Für die Modellwahl werden mehrere Methoden benutzt, welche hier kurz vorgestellt werden sollen.

2.0.1 Likelihood-Quotienten-Test

Der Likelihood-Quotienten-Test ist ein Hypothesentest, mit dem die Güte der Anpassung von zwei Modellen zum Datensatz verglichen wird. Dafür wird immer die Nullypothese Es gibt keinen Unterschied zwischen den Modellen gegen die Gegenhypothese Es gibt einen Unterschied zwischen den Modellen gestellt. Für diese Arbeit setzen wir immer den Likelihood-Quotienten-Chi-Quadrat-Test ein, mit dem Assozationen (Zusammenhang) zwischen Variablen ermittelt werden können. Beim Test wird die so genannte Test-Statistik $T(y)$ und dann damit dann die Überschreitungswahrscheinlichkeit bzw. der p-Wert berechnet. Der p-Wert gibt an, wie wahrscheinlich es ist die Nullhypothese anzunehmen. Wenn der p-Wert kleiner als 0,05 ist, dann weißt eine Variable eine statistisch signifikante Assozation auf und die Nullhypothese wird abgelehnt.

wieder: Assoziation?

Für ein aufgestellten Modelle bedeutet das, dass es zuerst in Untermodelle eingeteilt werden muss. Beispielmodell:

- $m \coloneqq crimes = \beta_0 + \beta_1 \cdot regionother + \beta_2 \cdot regionwest + \beta_3 \cdot prbarr + \beta_4 \cdot prbpris$

Bei diesem Modell werden 4 Untermodelle erstellt:

- $m_0 \coloneqq crimes = \beta_0$
- $m_1 \coloneqq crimes = \beta_0 + \beta_1 \cdot regionother + \beta_2 \cdot regionwest$
- $m_2 \coloneqq crimes = \beta_0 + \beta_1 \cdot regionother + \beta_2 \cdot regionwest + \beta_3 \cdot prbarr$
- $m_3 \coloneqq crimes = \beta_0 + \beta_1 \cdot regionother + \beta_2 \cdot regionwest + \beta_3 \cdot prbarr + \beta_4 \cdot prbpris$

Das erste Untermodell enthält noch keine Variable und die folgenden immer eine Variable mehr. Der Test bezieht sich nämlich immer nur auf die nächste zu einem Modell hinzukommende Variable, weshalb die Unterteilung der Modelle wichtig ist. Als nächstes wird die Test-Statistik und damit dann auch der p-Wert jedes Untermodells bestimmt:

Modell	Test-Statistik	p-Wert
m_0	NA	NA
m_1	73.466,892	0
m_2	91.079,802	0
m_3	8.676,222	0

Table 1: Ergebnisse des Likelihood-Quotienten-Tests

Das Modell m_0 ist das Startmodell, weshalb noch kein Vergleichsmodell für die Berechnung vorliegt und somit auch Test-Statistik und p-Wert nicht bestimmt werden können. ´Bei Modell m_1, m_2 und m_3 ist sichtbar der p-Wert kleiner als 0,05, damit weisen die neu eingefügten Variablen eine statistisch signifikante Assoziation auf. Da alle Modelle einen p-Wert unter 0,05 haben, stellt das Modell m ein legitimes Modell für den Datensatz dar.

2.0.2 Kreuzvalidierung

Für die Auswahl eines konkreten und gut an den Datensatz angepasst Modells, wird die d-fache Kreuzvalidierung angewendet. Hier wird der Datensatz in d ungefähr gleich große zufällige disjunkte Teile aufgeteilt, in diesem Fall sind es 9 Teile . Daraus ergeben sich insgesamt 9 Trainings- und Test-Datensätze. Die Modelle werden mit jedem Trainings-Datensatz gelernt und mit dem Test-Datensatz verglichen. Dazu wird der $SEP = \sum_{i=1}^{k}(y_i - \hat{y}_i)^2$ (Squared error of prediction), $MSEP = \frac{SEP}{k}$ (Mean squared error of prediction) und $RMSEP = \sqrt{MSEP}$. (Root mean squared error of prediction). Der SEP ist dabei der Prognosefehler bzw. die Summe der quadratischen Abweichungen zukünftiger Beobachtungen eines Modells. Um so geringer der SEP um so besser ist das Modell an den Datensatz angepasst. Den MSEP nutzt man um eine Vergleichbarkeit über verschieden große Datensätze zu ermöglichen. Durch die Quadratbildung im SEP werden die Einheiten verändert ($m \rightarrow m^2$). Um dem entgegenzuwirken wird im $RMSEP$ die Wurzel gezogen, wodurch alle Werte wieder in ihrer ursprünglichen Dimension sind.

wieso 9?

2.0.3 Schrittweise Regression

Mit schrittweiser Regression kann geprüft werden ob es optimiertere Modelle eines ausgewählten Modells gibt. Dazu kann die Rückwärtsselektion, Vorwärtsselektion und schrittweise Selektion genutzt werden. Zuerst werden aus dem ausgewählten Modell, also aus den damit ausgewählten Einflussgrößen alle mögliche Modelle erstellt. Dann wird für jedes Modell der RMSE (Root mean square error) bestimmt und das Modell mit dem geringsten RMSE ist dann das am besten an den Datensatz angepasste Modell (in Abhängigkeit vom ausgewählten Modell).

2.1 Modellwahl über Expertenwissen

Das Aufstellen eines Modells mithilfe von Expertenwissen bedeutet, die Bedeutungen bzw. Inhalte der einzelnen Einflussgrößen/Variablen zu analysieren und zu bestimmen, was sie für einen möglichen Einfluss auf die Zielgröße haben. In diesem Fall ist es die Zielgröße `crimes`. Eine Beeinflussung besteht dabei nur, wenn durch Veränderungen einer Einflussgröße sich `crimes` sich auch mit verändert.

2.1.1 Betrachtung der Einflussgrößen

Die Einflussgrößen werden mithilfe ihrer Bedeutung aus Kapitel **??** analysiert. Die folgenden Korrelationen sind immer nur in einer Richtung (steigende oder fallende Anzahl) dargestellt, in der jeweils andere Richtung ist der Einfluss umgekehrt:

- **prbarr:** Mit steigender Anzahl an Straftäter die arrestiert werden, sinkt potenziell die Anzahl an Strafttaten. Das liegt daran, dass für potenzielle Straftäter mit dem Risiko auch die Angst steigt, für eine Strafttat arrestiert zu werden.
- **prbpris:** Wenn viele Verbrecher ins Gefängnis eingesperrt werden, herrschen wahrscheinlich strenge Regelungen in einem County. So steigt für potenzielle Verbrecher wieder das Risiko und die Angst eine Straftat zu begehen, weshalb die Zahl an Straftaten sinkt. Zusätzlich können Verbrecher die im Gefängnis sind keine weiteren Strafttaten begehen.
- **polpc:** Mit mehr Polizisten pro Kopf sinkt die Anzahl an Straftaten. Da mehr Polizisten Straftaten feststellen können, werden mehr Straftäter eingesperrt und gleichzeitig steigt das Risiko bei einer Strafttat verhaftet zu werden . So können weniger Verbrechen begangen werden und gleichzeitig gibt es weniger neue Strafttäter.
- **density:** Je mehr Einwohner ein County pro Quadratmeile hat, desto mehr Einwohner wohnen insgesamt in dem County . Damit steigt auch die Anzahl an potenziellen Verbrechern.

wo ist der Unterschied zwischen den beiden Sachen?

vlt bei hohen Dichten auch mehr Straßenkriminalität

- **area:** Aus größerer oder kleinerer Fläche allein kann kein direkter Zusammenhang geschlossen werden, ob mehr oder weniger Straftaten begangen werden.
- **taxpc:** Je niedriger die durchschnittlichen Steuern eines Counties, desto weniger verdient der Großteil der Einwohner dort an Geld. Somit können immer weniger ihre Rechnungen bezahlen und werden potenziell eher kriminell, um Geld zu verdienen und zu überleben (Zahl an Strafttaten steigt).
- **region:** In region sind die 3 Werte: west, central und other enthalten und können jeweils einen unterschiedlichen Einfluss auf die Anzahl der Straftaten eines Counties haben. So begehen in der Region **central** z.B. grundsätzliche mehr Menschen Straftaten als in den anderen (siehe 1).

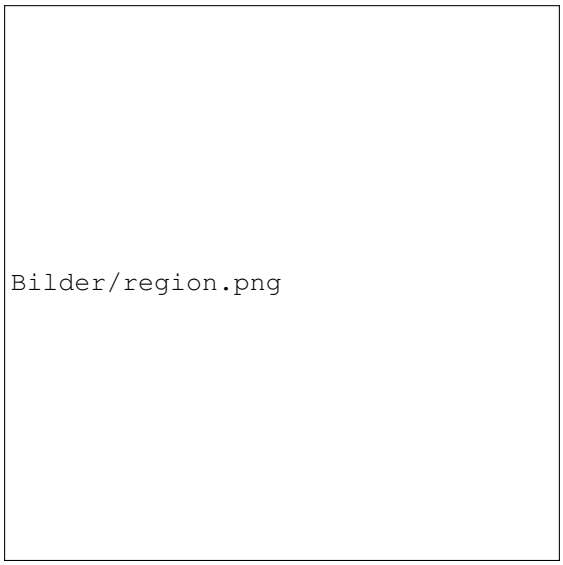

Figure 1: Plot zwischen crimes und region

- **pctmin:** Wenn der Anteil an Einwohnern aus Minderheiten in einem County hoch ist, weiß man nicht aus welchen Minderheiten diese stammen und Menschen die zu einer Minderheit gehören, neigen nicht automatisch mehr oder weniger zu Verbrechen neigen. Somit ist kein Zusammenhang feststellbar.
- **pctymale:** Männer neigen im Gegensatz zu Frauen eher zu Straftaten. Das sieht man z.B. daran, dass mehr Männern in Gefängnissen sind als Frauen. Außerdem haben Männer im jungen Alter vielleicht noch keine richtige Perspektive. So haben diese vielleicht noch keine Job, nicht den richtigen gefunden oder verdienen nur wenig Geld. Junge Männer können durch Geldmangel oder durch Freunden und auch Gangs beeinflusst werden Straftaten zu begehen. Ein hoher Anteil an jungen Männern in der Bevölkerung eines Counties könnte eine Erhöhung der Anzahl der Straftaten mit sich ziehen.
- **wcon, wser und wtrd:** In diesen Sektoren könnte der Anteil an ungelernten und niedrig bezahlten Arbeitern höher als bei den anderen Sektoren sein. Wenn der Anteil der in einem County arbeitenden Personen in diesen Sektoren sehr hoch ist, könnten dort auch mehr Straftaten stattfinden. Die Arbeiter verdienen wenig Geld und finden schlecht bessere Jobs, womit Kriminalität immer attraktiver wirkt, da diese einfacher und auch lukrativer als ihre Arbeit ist.

Mögliche Interaktionen zwischen den Einflussgrößen. Die folgenden Korrelationen sind immer nur in einer Richtung (steigende oder fallende Anzahl) dargestellt, in der jeweils andere Richtung ist der Einfluss umgekehrt:

- **polpc:area:** Je größer die Fläche eines Counties, desto weniger Polizisten gibt es pro Quadratmeile. Das Risiko, das ein Straftäter nicht für ein Verbrechen verhaftet wird, steigt. Somit sind potenzielle Straftäter eher bereit ein Verbrechen zu begehen.
- **area:density:** Je höher die Einwohnerdichte, desto höher ist die Population eines Counties unabhängig von der Größe der Fläche. Somit steigt für eine Fläche auch die Anzahl an potenziellen Verbrechnern.

2.1.2 Modelle aufstellen

- $me_0 \coloneqq crimes = \beta_0 + \beta_1 \cdot crimes + \beta_2 \cdot regionother + \beta_3 \cdot regionwest$
- $me_1 \coloneqq crimes = \beta_0 + \beta_1 \cdot regionother + \beta_2 \cdot regionwest + \beta_3 \cdot prbarr + \beta_4 \cdot prbpris + \beta_5 \cdot polpc + \beta_6 \cdot density + \beta_7 \cdot taxpc + \beta_8 \cdot pctymale + \beta_9 \cdot wcon + \beta_{10} \cdot wser + \beta_{11} \cdot wtrd$
- $me_2 \coloneqq crimes = \beta_0 + \beta_1 \cdot regionother + \beta_2 \cdot regionwest + \beta_3 \cdot prbarr + \beta_4 \cdot prbpris + \beta_5 \cdot polpc + \beta_6 \cdot density + \beta_7 \cdot taxpc + \beta_8 \cdot pctymale + \beta_9 \cdot wcon + \beta_{10} \cdot wser + \beta_{11} \cdot wtrd + \beta_{12} \cdot area + \beta_{13} \cdot polpc \cdot area + \beta_{14} \cdot area \cdot density$

Das Modell me_0 ist das Startmodell und ist aus der Aufgabenstellung entstanden, da `region` in das Modell auf alle Fälle einfließen sollte. Modell me_1 ist aus der Analyse der Einflussgrößen entstanden (siehe 2.1.1). Hierbei fließen lediglich die Einflussgrößen an sich ein und auch nur linear, da noch keine Analyse der Zusammenhänge zwischen den Einflussgrößen und crimes einbezogen wird. Modell me_2 erweitert Modell me_1 um die in der Analyse festgestellten Interaktionen (siehe 2.1.1).

Bei der Auswertung der nach dem Likelihood-Quotienten-Test, hat kein Modell den p-Wert von 0,05 überschritten und so sind alle 3 Modelle legitime Modelle für den gegebenen Datensatz.

Kreuzvalidierung der Modelle:

Modell	SEP	MSEP	RMSEP
me_0	**2.170.530.515**	**241.170.057**	**15.528,77**
me_1	7.163.568.325	795.952.036	26.816,92
me_2	5.377.690.266	597.521.141	23.746,69

Table 2: Ergebnisse der 9-fachen Kreuzvalidierung. Das beste Modell ist in **fett** hervorgehoben.

Das Startmodell me_0 stellt sichtbar mit allen 3 Werten immer noch das beste Modell da. Die Anpassung des Modells rein durch Expertenwissen scheint also keine großen Nutzen gebracht zu haben. Grund dafür kann natürlich sein, das das Expertenwissen zu gering war um wirklich gute Modelle aufstellen zu können. Mit Analyse der Daten der Einflussgrößen und verschiedenen Verfahren, können diese Modelle jedoch noch weiter verbessert werden (siehe später im Kapitel 2.3).

2.2 Modellwahl über Analyse der Zusammenhänge

Bei der Modellwahl über Zusammenhänge, wurde die Korrelation aller Einflussgrößen mit `crimes` auf Zusammenhänge analysiert. Hierzu wurde betrachtet, ob ein linearer, logarithmischer, wurzel, quadratischer oder kubischer Zusammenhang besteht.

Für jede Kombination mit `crimes` wurde ein generelles lineares Modell (GLM) erstellt und in einem Plot auf die Daten gemapped. So wurden die Korrelationen grafisch analysiert, wie in diesem Beispiel:

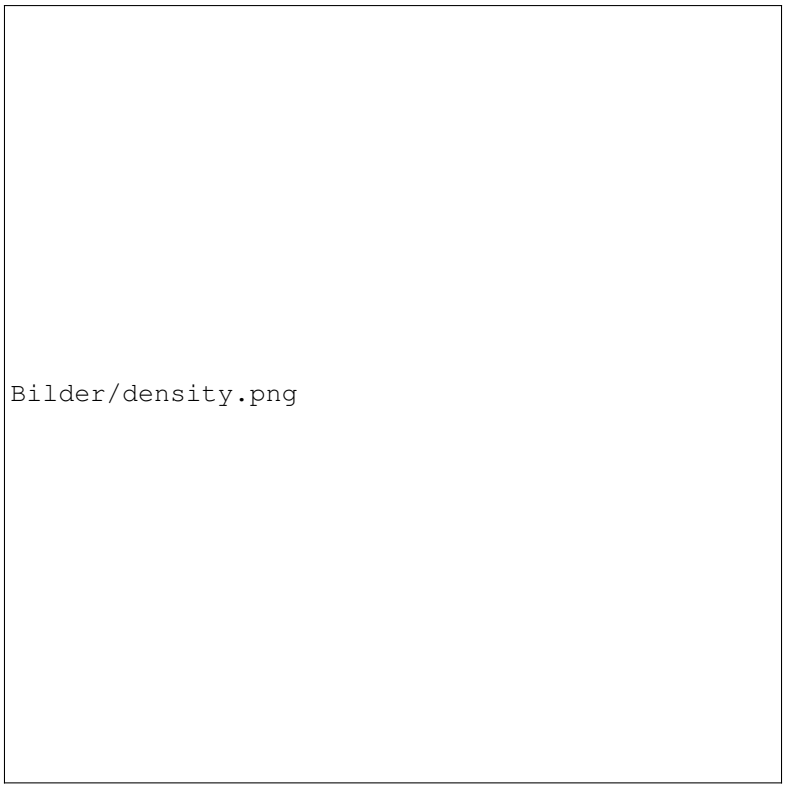

Figure 2: Plot zwischen crimes und density

Die Modelle hatten dabei unterschiedliche Farben: Linear (Schwarz), Logarithmus (Rot), Wurzel (Grün), Quadrat (Dunkelblau) und Kubisch (Hellblau). Herauszufinden welches Modell am besten an die Daten angepasst ist, ist schwierig herauszufinden, da natürlich nur eine Auswahl an Modellen betrachtet wird, so ist nicht immer das beste dabei und mehrere Modelle können „gute Ergebnisse“ liefern. Ein Ergebnis ist gut, wenn es nach subjektiver Betrachtung gut an die gegeben Datenpunkte angepasst ist. In diesem Beispiel liefert die Wurzel ein relativ gutes Ergebnis.

Liste aller festgestellten Zusammenhänge:

- **prbarr:** Logarithmus
- **prbpris:** Quadrat
- **polpc:** Quadrat
- **density:** Wurzel
- **area:** Linear
- **taxpc:** kein Zusammenhang

- **region:** (wird als Faktor einbezogen, dementsprechend ist kein Zusammenhang feststellbar)
- **pctmin:** Quadrat
- **pctymale:** kein Zusammenhang
- **wcon:** Wurzel
- **wsta:** kein Zusammenhang
- **wser:** Wurzel
- **wtrd:** Quadrat
- **wfir:** Logarithmus

2.2.1 Modelle aufstellen

Die folgenden Modelle enthalten noch keine Interaktionen, da zuerst rein die Zusammenhänge der Variablen mit `crimes` betrachtet werden.

- $mz_0 \coloneqq crimes = \beta_0 + \beta_1 \cdot region$
- $mz_1 \coloneqq crimes = \beta_0 + \beta_1 \cdot region + \beta_2 \cdot area + \beta_2 \cdot density + \beta_3 \cdot pctmin + \beta_4 \cdot polpc + \beta_5 \cdot prbarr + \beta_6 \cdot prbpris + \beta_7 \cdot wcon + \beta_8 \cdot wfir + \beta_9 \cdot wser + \beta_{10} \cdot wtrd$
- $mz_2 \coloneqq crimes = \beta_0 + \beta_1 \cdot region + \beta_2 \cdot area + \beta_3 \cdot \sqrt{density} + \beta_4 \cdot \sqrt{pctmin} + \beta_5 \cdot polpc + \beta_6 \cdot polpc^2 + \beta_7 \cdot \log prbarr + \beta_8 \cdot prbpris + \beta_9 \cdot prbpris^2 + \beta_{10} \cdot \sqrt{wcon} + \beta_{11} \cdot \log wfir + \beta_{12} \cdot \sqrt{wser} + \beta_{13} \cdot wtrd + \beta_{14} \cdot wtrd^2$
- $mz_3 \coloneqq crimes = \beta_0 + \beta_1 \cdot region + \beta_2 \cdot area + \beta_3 \cdot \sqrt{density} + \beta_4 \cdot \sqrt{pctmin} + \beta_5 \cdot polpc + \beta_6 \cdot polpc^2 + \beta_7 \cdot \log prbarr + \beta_8 \cdot \sqrt{wcon} + \beta_9 \cdot wtrd + \beta_{10} \cdot wtrd^2$

Das Modell mz_0 ist das Startmodell wie in Kapitel 2.1.2. Modell mz_1 enthält alle Variablen aus der Liste aller Zusammenhänge, die einen Zusammenhang bzw. Korrelation mit `crimes` haben, jedoch fließen diese nur linear ein. Modell mz_2 enthält die gleichen Variablen wie mz_1, jedoch werden die Zusammenhänge aus der Liste mit einbezogen. Für Modell mz_3 wurde überprüft ob Variablen untereinander einen linearen Zusammenhang haben. Bei Variablen mit linearem Zusammenhang muss in ein Modell nur eine der beiden eingefügt werden, da die andere nicht mehr Information liefert. Das Modell mz_3 enthält also alle Variablen aus mz_2, außer `prbpris`, `wfir` und `wser`.

Bei der Auswertung der Modelle mit dem Likelihood-Quotienten-Test, wurde der p-Wert von 0,05 nicht überschritten und so sind alle Modelle legitim für den Datensatz.

Kreuzvalidierung der Modelle:

Modell	SEP	MSEP	RMSEP
mz_0	2.170.530.515	241.170.057	15.528,768
mz_1	8.692.690.305	965.854.478	30.480,822
mz_2	304.348.797	33.816.533	5.765,326
mz_3	**267.308.964**	**29.700.996**	**5.369,317**

Table 3: Ergebnisse der 9-fachen Kreuzvalidierung. Das beste Modell ist in **fett** hervorgehoben.

Abgesehen von Modell mz_1 zeigen mz_2 und mz_3 deutliche Verbesserungen zum Startmodell. Das beste Modell ist mz_3 mit einem Prognosefehler der 8 mal kleiner ist als der vom Startmodell. Wie schon erwähnt fließen dabei noch keine Interaktionen mit ein und die Modelle können noch verfeinert werden (siehe später im Kapitel **??**).

2.3 Verfeinerung der Modelle aus Expertenwissen

Durch Überprüfung linearer Zusammenhänge zwischen den Variablen und der Liste der Zusammenhänge aus Kapitel 2.2 können die aufgestellten Modelle verbessert bzw. erweitert werden um mögliche Verbesserungen zu erzielen. Zur Verbesserung wird hierbei Modell me_2 genutzt.

2.3.1 Verfeinerte Modelle aufstellen

- $me_3 \coloneqq crimes = \beta_0 + \beta_1 \cdot regionother + \beta_2 \cdot regionwest + \beta_3 \cdot prbarr + \beta_4 \cdot polpc + \beta_5 \cdot density + \beta_6 \cdot taxpc + \beta_7 \cdot pctymale + \beta_8 \cdot wcon + \beta_8 \cdot wser + \beta_9 \cdot wtrd + \beta_9 \cdot polpc \cdot area + \beta_{10} \cdot area \cdot density$
- $me_4 \coloneqq crimes = \beta_0 + \beta_1 \cdot regionother + \beta_2 \cdot regionwest + \beta_3 \cdot \log prbarr + \beta_4 \cdot prbpris + \beta_5 \cdot prbpris^2 + \beta_6 \cdot polpc + \beta_7 \cdot polpc^2 + \beta_8 \cdot \sqrt{density} + \beta_9 \cdot taxpc + \beta_{10} \cdot pctymale + \beta_{11} \cdot \sqrt{wcon} + \beta_{12} \cdot \sqrt{wser} + \beta_{13} \cdot wtrd + \beta_{14} \cdot wtrd^2 + \beta_{15} \cdot area + \beta_{15} \cdot polpc \cdot area + \beta_{16} \cdot area \cdot density$
- $me_5 \coloneqq crimes = \beta_0 + \beta_1 \cdot regionother + \beta_2 \cdot regionwest + \beta_3 \cdot \log prbarr + \beta_4 \cdot polpc + \beta_5 \cdot polpc^2 + \beta_6 \cdot \sqrt{density} + \beta_7 \cdot taxpc + \beta_8 \cdot pctymale + \beta_9 \cdot \sqrt{wcon} + \beta_{10} \sqrt{wser} + \beta_{11} \cdot wtrd + \beta_{12} \cdot wtrd^2 + \beta_{13} \cdot polpc \cdot area + \beta_{14} \cdot area \cdot density$

Modell me_3 ist die gekürzte Version von me_2. Hierbei wurden die Einflussgrößen wieder auf lineare Zusammenhänge untereinander untersucht und bei möglicher Korrelation wird eine der beiden Einflussgrößen aus dem Modell entfernt (hier `prbpris` und `area`). Das Modell me_4 erweitert me_2 um die erwähnten Zusammenhänge aus der Liste der Zusammenhänge. Das selbe gilt auch für Modell me_5, jedoch erweitert es das Modell me_3.

Alle Modelle überschreiten nach dem Likelihood-Quotienten-Test (siehe 2.0.1) nicht den p-Wert von 0,05 und stellen somit legitime Modelle für den Datensatz da.

Kreuzvalidierung der Modelle:

Modell	SEP	MSEP	RMSEP
me_3	7.366.934.146	818.548.238	27.842,869
me_4	**189.134.487**	**21.014.943**	**4.488,199**
me_5	218.432.097	24.270.233	4.869,731

Table 4: Ergebnisse der 9-fachen Kreuzvalidierung. Das beste Modell ist in **fett** hervorgehoben.

Das Modell me_3 ist keine Verbesserung des Modells me_2. Die Modelle me_4 und me_5 erweitern me_2 und me_3 und zeigen auch eine deutliche Verbesserung. Das beste Modelle stellt dabei me_4 dar.

2.3.2 Weitere Verbesserung des besten Modells

Mithilfe von schrittweise Regression wird aus dem besten Modell ein optimaleres Modell gesucht. Hierbei wird die Rückwärtsselektion angewendet.
Es hat sich herausgestellt, das Interaktionen eines Modells nicht in die Regression nicht einfließen sollten,

da dann die optimalen Modelle keine besseren Ergebnisse liefern. Aufgrund dessen wird Modell me_4 ausgewählt, jedoch ohne die darin befindlichen Interaktionen (diese werden im Nachhinein wieder eingefügt).

Die durch Regression entstehenden Modelle:

- $me_6 \coloneqq crimes = \beta_0 + \beta_1 \cdot prbpris^2 + \beta_2 \cdot \sqrt{density} + \beta_3 \cdot \sqrt{wser} + \beta_4 \cdot wtrd + \beta_5 \cdot wtrd^2 + \beta_6 \cdot area$
- $me_7 \coloneqq crimes = \beta_0 + \beta_1 \cdot prbpris^2 + \beta_2 \cdot \sqrt{density} + \beta_3 \cdot \sqrt{wser} + \beta_4 \cdot wtrd + \beta_5 \cdot wtrd^2 + \beta_6 \cdot area + \beta_7 \cdot polpc \cdot area + \beta_8 \cdot area \cdot density$

Modell me_6 ist das Modell aus me_4 (ohne Interaktionen) mit dem geringsten RMSE Wert. Das Modell me_7 erweitert me_6 dann nur noch um die fehlenden Interaktionen aus me_4,

Bei beide Modelle liegt der p-Wert unter 0.05, somit sind es legitime Modelle für den Datensatz (Likelihood-Quotienten-Test).

Kreuzvalidierung der Modelle:

Modell	SEP	MSEP	RMSEP
me_6	167.827.279	18.647.475	4.179,796
me_7	**113.866.024**	**12.651.780**	**3.457,341**

Table 5: Ergebnisse der 9-fachen Kreuzvalidierung. Das beste Modell ist in **fett** hervorgehoben.

Sowohl Modell me_6, als auch me_7 sind bessere Modelle als me_4. Zusätzlich ist me_7 mit den Interaktionen deutlich besser als me_6. Damit ist Modell me_7 das beste Modell, welches durch Expertenwissen aufgestellte und dann noch verfeinerte wurde.

2.4 Verfeinerung der Modelle aus Zusammenhängen

Mithilfe von schrittweise Regression, hinzufügen von Interaktionen und neu erstellten Einflussgrößen, können die durch Zusammenhänge aufgestellten Modelle noch weiter verbessert werden. Zur Verfeinerung wird das Modell mz_3 genutzt.

2.4.1 Verbesserung des Modells durch schrittweise Regression

Anhand von Rückwärtsselektion wird aus mz_3 ein optimaleres Modell erstellt:

- $mz_4 \coloneqq crimes = \beta_0 + \beta_1 \cdot area + \beta_2 \cdot \sqrt{density} + \beta_3 \cdot \sqrt{pctmin} + \beta_4 \cdot wtrd + \beta_5 \cdot wtrd^2$

Erweitern der Modelle mz_3 und mz_4 durch Interaktion:

- $mz_5 \coloneqq crimes = \beta_0 + \beta_1 \cdot regionother + \beta_2 \cdot regionwest + \beta_3 \cdot area + \beta_4 \cdot \sqrt{density} + \beta_5 \cdot \sqrt{pctmin} + \beta_6 \cdot polpc + \beta_7 \cdot polpc^2 + \beta_8 \cdot \log prbarr + \beta_9 \cdot \sqrt{wcon} + \beta_{10} \cdot wtrd + \beta_{11} \cdot wtrd^2 + \beta_{12} \cdot area \cdot density + \beta_{13} \cdot polpc \cdot area$
- $mz_6 \coloneqq crimes = \beta_0 + \beta_1 \cdot area + \beta_2 \cdot \sqrt{density} + \beta_3 \cdot \sqrt{pctmin} + \beta_4 \cdot wtrd + \beta_5 \cdot wtrd^2 + \beta_6 \cdot area \cdot density + \beta_7 \cdot polpc \cdot area$
- $mz_7 \coloneqq crimes = \beta_0 + \beta_1 \cdot regionother + \beta_2 \cdot regionwest + \beta_3 \cdot area + \beta_4 \cdot \sqrt{density} + \beta_5 \cdot \sqrt{pctmin} + \beta_6 \cdot polpc + \beta_7 \cdot polpc^2 + \beta_8 \cdot \log prbarr + \beta_9 \cdot \sqrt{wcon} + \beta_{10} \cdot wtrd + \beta_{11} \cdot wtrd^2 + \beta_{12} \cdot area \cdot density$

- $mz_8 \coloneqq crimes = \beta_0 + \beta_1 \cdot area + \beta_2 \cdot \sqrt{density} + \beta_3 \cdot \sqrt{pctmin} + \beta_4 \cdot wtrd + \beta_5 \cdot wtrd^2 + \beta_6 \cdot area \cdot density$
- $mz_9 \coloneqq crimes = \beta_0 + \beta_1 \cdot regionother + \beta_2 \cdot regionwest + \beta_3 \cdot area + \beta_4 \cdot \sqrt{density} + \beta_5 \cdot \sqrt{pctmin} + \beta_6 \cdot polpc + \beta_7 \cdot polpc^2 + \beta_8 \cdot \log prbarr + \beta_9 \cdot \sqrt{wcon} + \beta_{10} \cdot wtrd + \beta_{11} \cdot wtrd^2 + \beta_{12} \cdot polpc \cdot area$
- $mz_{10} \coloneqq crimes = \beta_0 + \beta_1 \cdot area + \beta_2 \cdot \sqrt{density} + \beta_3 \cdot \sqrt{pctmin} + \beta_4 \cdot wtrd + \beta_5 \cdot wtrd^2 + \beta_6 \cdot polpc \cdot area$
- mz_5 = `crimes ~ region + area + I(sqrt(density)) + I(sqrt(pctmin)) + polpc + I(`$polpc^2$`) + I(log(prbarr)) + I(sqrt(wcon)) + wtrd + I(`$wtrd^2$`) + area:density + polpc:area`
- mz_6 = `crimes ~ area + I(sqrt(density)) + I(sqrt(pctmin)) + wtrd + I(`$wtrd^2$`) + area:density + polpc:area`
- mz_7 = `crimes ~ region + area + I(sqrt(density)) + I(sqrt(pctmin)) + polpc + I(`$polpc^2$`) + I(log(prbarr)) + I(sqrt(wcon)) + wtrd + I(`$wtrd^2$`) + area:density`
- mz_8 = `crimes ~ area + I(sqrt(density)) + I(sqrt(pctmin)) + wtrd + I(`$wtrd^2$`) + area:density`
- mz_9 = `crimes ~ region + area + I(sqrt(density)) + I(sqrt(pctmin)) + polpc + I(`$polpc^2$`) + I(log(prbarr)) + I(sqrt(wcon)) + wtrd + I(`$wtrd^2$`) + polpc:area`
- mz_{10} = `crimes ~ area + I(sqrt(density)) + I(sqrt(pctmin)) + wtrd + I(`$wtrd^2$`) + polpc:area`

Modelle mit ungerader Zahl sind Erweiterungen vom Modell mz_3 und die mit gerader Zahl vom Modell mz_4. Die Modelle mz_5 bis mz_{10} wurden mit den Interaktionen `area:density` und `polpc:area` erweitert. Diese beiden Interaktionen wurden gewählt, da diese sich schon bei den Modellen durch Expertenwissen als sehr sinnvoll erwiesen haben.

Keines der 7 Modelle überschreitet beim Likelihood-Quotienten-Test den p-Wert von 0.05 und somit sind sie alle legitimes Modelle für den Datensatz.

Kreuzvalidierung der Modelle:

Modell	SEP	MSEP	RMSEP
mz_4	157.053.141	17.450.349	4.020,416
mz_5	136.728.331	15.192.037	3.784,481
mz_6	165.860.241	18.428.916	4.143,400
mz_7	**133.909.669**	**14.878.852**	**3.724,919**
mz_8	143.615.471	15.957.275	3.884,638
mz_9	165.657.990	18.406.443	4.235,226
mz_{10}	222.195.668	24.688.408	4.764,043

Table 6: Ergebnisse der 9-fachen Kreuzvalidierung. Das beste Modell ist in **fett** hervorgehoben.

Grundsätzlich sind all diese aufgestellten Modelle Verbesserungen des Modells mz_3 und das Modell mz_7 liefert das beste Ergebnis. Außerdem zeigt sich, das die Erweiterungen des Regressionsmodells mz_4, letztendlich doch schlechtere Ergebnisse liefern als die Modelle, die mz_3 erweitern. Somit zeigt sich, Modelle mit einer schlechteren Anpassung an einen Datensatz, können trotzdem stark verbessert werden und sollten deshalb nie direkt missachtet werden.

2.5 Letzte Verfeinerung durch neue Einflussgrößen

Aus den Einflussgrößen `area` und `density` lässt sich die Größe `population` = `area` * `density` erstellen. Diese gibt für jedes County die gesamte Einwohnerzahl an. Außerdem lässt sich aus `population` die Größe `urban` (enthält boolesche Werte) ableiten, die festlegt ob ein County eher eine urbane oder ländliche Umgebung hat. Damit ein County urbane Umgebung besitzt, muss es mindestens eine Einwohnerzahl von 50.000 aufweisen.Die Variable `population` könnte einen möglichen Einfluss auf `crimes` haben, da die Variablen `area` und `density` auch schon in den besten Modellen enthalten sind. Zum Vergleich sollte somit auch getestet werden, welchen Einfluss die Einteilung eines Counties in urbane und ländliche Umgebung hat.

Es werden keine neuen Modelle erstellt, lediglich werden die besten Modell erweitert. Das beste Modell aus Expertenwissen ist me_7 und aus den Zusammenhängen mz_7. Da beide Modelle schon die Interaktion `area:density` enthalten, muss `population` nicht mehr in die Modelle einfließen, da es mathematisch das selbe bedeutet und somit den selben Einfluss hat. Somit bleibt nur noch die Erweiterungen der beiden Modelle mit der Variable `urban`:

- $mn_1 \coloneqq crimes = \beta_0 + \beta_1 \cdot prbpris^2 + \beta_2 \cdot \sqrt{density} + \beta_3 \cdot \sqrt{wser} + \beta_4 \cdot wtrd + \beta_5 \cdot wtrd^2 + \beta_6 \cdot area + \beta_7 \cdot area \cdot density + \beta_8 \cdot polpc \cdot area + \beta_9 \cdot urban$
- $mn_2 \coloneqq crimes = \beta_0 + \beta_1 \cdot regionother + \beta_2 \cdot regionwest + \beta_3 \cdot area + \beta_4 \cdot \sqrt{density} + \beta_5 \cdot \sqrt{pctmin} + \beta_6 \cdot polpc + \beta_7 \cdot polpc^2 + \beta_8 \cdot \log prbarr + \beta_9 \cdot \sqrt{wcon} + \beta_{10} \cdot wtrd + \beta_{11} \cdot wtrd^2 + \beta_{12} \cdot area \cdot density + \beta_{13} \cdot urban$
- mn_1 `= crimes ~ I(`$prbpris^2$`) + I(sqrt(density)) + I(sqrt(wser)) + wtrd + I(`$wtrd^2$`) + area + polpc:area + area:density + urban`
- mn_2 `= crimes ~ region + area + I(sqrt(density)) + I(sqrt(pctmin)) + polpc + I(`$polpc^2$`) + I(log(prbarr)) + I(sqrt(wcon)) + wtrd + I(`$wtrd^2$`) + area:density + urban`

Das Modell mn_1 erweitert das Modell me_7 und mn_2 erweitert mz_7.

Beide Modelle stellen nach dem Likelihood-Quotienten-Test legitime Modelle dar.

Kreuzvalidierung der Modelle:

Modell	SEP	MSEP	RMSEP
mn_1	176.902.539	19.655.838	4.286,544
mn_2	**61.489.122**	**6.832.125**	**2.536,939**

Table 7: Ergebnisse der 9-fachen Kreuzvalidierung. Das beste Modell ist in **fett** hervorgehoben.

Für beide Modelle zeigt sich wie unterschiedlich neue Variablen ein Modell beeinflussen können. Das vorher bessere Modelle me_7 hat sich stark durch den Einfluss von `urban` verschlechtert, dahingegen hat

sich mz_7 fast um den selben Wert verbessert.

Das Modell mn_2 ist damit insgesamt von den aufgestellten Modellen, das Modell, welches am besten an den Datensatz angepasst ist.

3 Vergleich von Konfidenzintervallen

Ein Konfidenzintervall ist ein Intervall, in dem ein unbekannter Parameter mit einer Wahrscheinlichkeit von $100(1-\alpha)\%$ liegt. Diese Abschätzung kann nicht nur auf den Erwartungswert einer Zufallsgröße angewendet werden, sondern auch auf die gesuchten Parameter eines Modells.

Definition 1 (Konfidenzintervall)**.** *Gegeben sei ein statistisches Modell* $(\underline{Y}, A, \{P_\theta, \theta \in \Theta\})$*. Ein zufälliges Intervall* $(C_u, C_o) = (C_u(\underline{Y}, C_o(\underline{Y})))$ *heißt Konfidenzintervall zum Niveau* $1-\alpha$ *(oder* $(1-\alpha)$*-Konfidenzintervall) für einen unbekannten Parameter* θ*, falls gilt:*

$$\forall \theta \in \Theta : P_\theta(C_u(\underline{Y}) < \theta < C_o(\underline{Y})) = 1 - \alpha$$

Es gibt mehrere Methoden um Konfidenzintervalle zu bestimmen und mit dieser Simulation möchten wir Zwei vergleichen. Die erste Methode approximiert die Konfidenzintervalle über die beobachtete Fisher-Informations-Matrix, die zweite Methode sucht ein Supremum über die Loglikelihood. Die zweite Methode ist in viele Statistik-ProgrammierSprachen

3.1 Bestimmung des Konfidenzintervalls

Die hier benutzten Modelle gehören alle den Poisson verteilten verallgemeinerten linearen Modellen an. Da die Poisson-Verteilung aus der Exponential-Dispersions-Familie ist gelten für sie auch die Eigenschaften dieser Familie. So gilt für Maximum-Likelihood-Schätzer in verallgemeinerten linearen Modell und unter Regularitätsbedingung:

$$\lim_{n\to\infty} F_n^{T/2}(\underline{\beta})(\underline{\hat{\beta}}_n - \underline{\beta}) = \mathcal{N}(\mathbf{0}_k, \mathbf{1}_k)$$

Für eine endliche Stichprobe n kann es nur approximiert werden.

$$F_n^{T/2}(\underline{\beta})(\underline{\hat{\beta}}_n - \underline{\beta}) \approx \mathcal{N}(\mathbf{0}_k, \mathbf{1}_k)$$

und so gilt auch für den Parametervektor nur eine approximative Abschätzung:

$$\underline{\hat{\beta}}_n \approx \mathcal{N}(\underline{\beta}, I_n^{-1}(\underline{\hat{\beta}}))$$

wobei die unbekannte erwartete Fisher-Informations-Matrix F_n durch die beobachtete Fisher-Information-Matrix I_n ersetzte werden kann: $F_n(\underline{\beta}) \approx I_n(\underline{\hat{\beta}}) = X^T \hat{V} X$. X ist hier dies Designmatrix und $\hat{V} = V(\underline{\hat{\beta}}) = diag(Var(Y_1), ..., Var(Y_n))$. Da der Parametervektor nur approximativ geschätzt werden kann, gilt auch für die Konfidenzintervall, dass sie nur approximativ bestimmt werden können. Für einen Komponete $\hat{\beta}_{j,n}$ des Parametervektors gilt:

$$P\Big(\hat{\beta}_{j,n} - z_{1-\frac{\alpha}{2}}\sqrt{I_n(\underline{\hat{\beta}}_n)^{-1}_{jj}} < \beta_j < \hat{\beta}_{j,n} + z_{1-\frac{\alpha}{2}}\sqrt{I_n(\underline{\hat{\beta}}_n)^{-1}_{jj}}\Big) \approx 1 - \alpha \tag{3}$$

und

$$\lim_{n\to\infty} P(\hat{\beta}_{j,n} - z_{1-\frac{\alpha}{2}}\sqrt{I_n(\underline{\hat{\beta}}_n)^{-1}_{jj}} < \beta_j < \hat{\beta}_{j,n} + z_{1-\frac{\alpha}{2}}\sqrt{I_n(\underline{\hat{\beta}}_n)^{-1}_{jj}}) = 1 - \alpha \tag{4}$$

3.2 R's `confint` Methode

R verwendet *Profile Likelihood* zur Schätzung von Konfidenzintervallen. Dabei wird der Parametervektor $\underline{\beta} = (\gamma, \psi)$ aufgeteilt in den zu schätzenden Parameter γ und den Rest ψ. Die `Profile Likelihood`

Funktion ist dann definiert als $\mathcal{L}_P(\gamma \vdots \underline{y}) = \sup_\psi \mathcal{L}((\gamma, \psi) \vdots \underline{y})$. Es wird der größte Wert der Likelihood unter allen verschiedenen Kombinationen von ψ gesucht, während γ fest definiert ist.

3.3 Simulation

Für den Vergleich zwischen R's `confint` Methode und Formel **??** wird das 0.95-Konfidenzintervall gewählt, es gilt also $\alpha = 0.05$. Um zu überprüfen, wie genau die Konfidenzintervalle geschätzt werden wird der wahre Parametervektor $\underline{\beta}$ des wahren Modells benötigt. Dieses Modell ist in der Regel nicht bekannt und auch nur schwer bis gar nicht ermittelbar. Um dennoch einen wahren Parametervektor zu haben, wird eine pseudowahres Modell festgelegt. Dieses ist me_7 aus der Modellwahl. $\text{me}_7 \coloneqq crimes = \beta_0 + \beta_1 \cdot prbpris^2 + \beta_2 \cdot \sqrt{density} + \beta_3 \cdot \sqrt{wser} + \beta_4 \cdot wtrd + \beta_5 \cdot wtrd^2 + \beta_6 \cdot area + \beta_7 \cdot polpc \cdot area + \beta_8 \cdot area \cdot density$

Die approximativen 0.95-Konfidenzintervalle gelten nur im Unendlichen, daher wird die Genauigkeit der Methoden über einen stetig wachsenden Stichprobenumfang simuliert. Da der Datensatz allerdings nur 90 Einträge beinhaltet, werden Einträge aus den original Daten stetig neu hinzugenommen, sodass schnell viele doppelte Einträge vorliegen. Mit diesem Stichprobenumfang wird dann das pseudowahre Modell geschätzt. Man erhält den Erwartungswertvektor $\underline{EY}$ der Zufallsvektoren, sowie den geschätzten Parametervektor $\underline{\beta}$ des Modells.

Nun werden Pseudobeobachtungen durch den Erwartungswertvektor generiert. Für jeden Eintrag im Stichprobenumfang liegt ein Erwartungswert μ vor. Über die Poisson-Verteilung mit Erwartungswert μ wird dann eine Beobachtung für diesen Eintrag simuliert. Das normale Modell wird nun mit den Pseudobeobachtungen und der Designmatrix des Stichprobenumfangs geschätzt. Man erhält den geschätzten Parametervektor $\hat{\underline{\beta}}$ des Modells. Mit diesem kann man das Intervall (C_u, C_o) für jeden Parameter wie in **??** aufstellen. Auch R's `confint` Methode kann nach dem Schätzen des Modells ein 0.95-Konfidenzintervall für jeden Parameter aufstellen. Über den Vergleich $C_u < \beta_j < C_o$ wird dann ermittelt, ob der pseudowahre Parameter in dem approximierten Konfidenzintervall liegt. Um eine aussagekräftige wahrscheinlichkeit zu bekommen wird dieser Prozess mehrmals mit neuen Pseudobeobachtungen wiederholt. Insgesamt wird erwartet, dass sich die Wahrscheinlichkeit immer näher an 95% annähert umso größer der Stichprobenumfang wird.

4 Ergebnisse

Die Modellwahl erfolgt einerseits aus dem Expertenwissen über die Variablen des Datensatzes und anderseits aus aus der graphischen Analyse des Datensatzes. Die dabei entstandenen Modellen wurden durch schrittweise Regression, hinzufügen von Interaktion und neu erstellten Einflussgrößen immer weiter verfeinert um letztendlich zu diesem Modell zu kommen:

- $mn_2 \coloneqq crimes = \beta_0 + \beta_1 \cdot regionother + \beta_2 \cdot regionwest + \beta_3 \cdot area + \beta_4 \cdot \sqrt{density} + \beta_5 \cdot \sqrt{pctmin} + \beta_6 \cdot polpc + \beta_7 \cdot polpc^2 + \beta_8 \cdot \log prbarr + \beta_9 \cdot \sqrt{wcon} + \beta_{10} \cdot wtrd + \beta_{11} \cdot wtrd^2 + \beta_{12} \cdot area \cdot density + \beta_{13} \cdot urban$

5 Diskussion

6 Fazit

7 Literaturverzeichnis

[heading=none]