Modelltheorie

LM

Das allgemeine Ziel der Regression besteht darin, eine Beziehung zwischen der interessierenden bzw. der Zielvariable Y und einer oder mehreren Einflussvariablen x1, …, xp herzustellen, wobei bei der linearen Regression von einem linearen Zusammenhang ausgegangen wird. Die zielvariable sowie die abhängigen Variablen sind metrische Größen. Daraus ergibt sich für eine Stichprobe von Umfang i = 1, 2, …, n das lineare Modell als:

Yi = β0 + β1 \* xi1 + β2 \* xi2 + … + βp \* xip + εi = **xtiβ** + εi

Mit xi = (1, xi1, xi2, ... , xip)talsPrädiktorvektor und β als Parametervektor.[[1]](#footnote-1) Dabei gelten folgende Annahmen. Die Beobachtungen der Zielvariable Yi sind unabhängig. Außerdem haben die Resiuduen εi E(εi)=0 und Var(εi)=σ2. Sind die Residuen zusätzlich normalverteilt, spricht man von klassischer Normalregression.[[2]](#footnote-2) Man schätzt somit den bedingten Erwartungswert von Y bei gegebenen **xti** und das Modell liefert eine Prognose der Zielvariablen.

Im hiesigen Fall würde die Modellgleichung folgendermaßen aussehen:

Getriggerte Magnitudei = β0 + β1 \* xi triggernde Magnitude + β2 \* xi heat flow + β3 \* xi strain rate +

β4 \* xi dip + β5 \* xi depth + β6 \* xi rake + β7 \* xi crustal thickness + β8 \* xi zeitdifferenz +

β9 \* xi completeness Magnitude + β10 \* xi mantle thickness

Da primär der Zusammenhang zwischen triggernder und getriggerter Magnitude betrachtet werden soll, ist die getriggerte Magnitude die Zielvariable und die triggernde Magnitude ist als Einflussvariable im Modell aufgenommen. Indem die anderen Variablen auch im Modell enthalten sind, wird deren Effekt auf die getriggerte Magnitude kontrolliert. Die Schätzung der unbekannten Koeffizienten β erfolgt anhand der Kleinste-Quadrate-Methode (KQ-Methode). Diese beruht auf der Minimierung der Summe der quadrierten Abweichungen:[[3]](#footnote-3)

Man würde folglich ein Modell erhalten, welches eine erste Prognose für den Erwartungswert der getriggerten Magnitude liefert. Bei Betrachtung der Verteilung der Zielvariable fällt auf, dass diese Exponentialverteilt ist (vgl Abb. 1). Gezeigt ist hier ein Histogramm, welches die Häufigkeitsverteilung der getriggerten Magnituden für Japan widerspiegelt. Zu beachten ist, dass die Magnituden so transformiert wurden, dass sie bei 0 beginnen, indem alle Werte mit 4 subtrahiert wurden. Dadurch liegt das Minimum der Magnituden bei 4, wodurch wird der Vergleich mit der Exponentialverteilung vereinfacht.

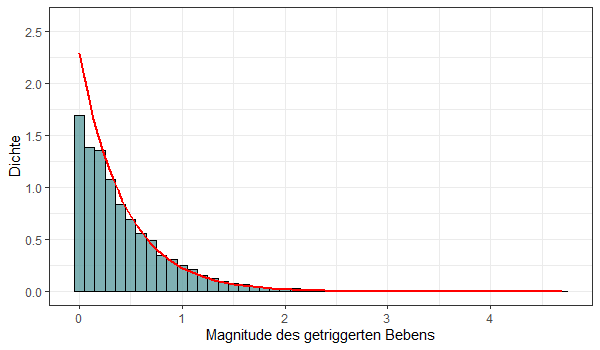


Abb. 1: Vergleich der Verteilung der getriggerten Magnituden für Japan mit der Dichte der Exponentialverteilung mit λ = 2.28 (rote Linie)

Diese Verteilung folgt dem Gutenberg-Richter-Gesetz: beobachtet man beispielsweise ein Beben der Magnitude6, treten 10-mal so viele Beben der Magnitude 5.0 auf und 100-Mal so viele der Magnitude 4.0.[[4]](#footnote-4) Man hat somit einen exponentiellen Zusammenhang. Diese Annahme wird durch die Dichte der Exponentialverteilung mit Parameter λ = 2.28 bestätigt, die hier durch eine rote Linie dargestellt wird. Der Parameter λ lässt sich folgendermaßen schätzen:

Erstellt man eine identische Graphik für Kalifornien, zeichnet sich ein vergleichbares Bild ab. Aus verschiedenen Gründen, die später genauer erläutert werden, wird im Weiteren eine Gammaverteilung angenommen. Dies ist möglich, da die Exponentialverteilung ein Spezialfall der Gammaverteilung ist, bei dem der erste Parameter auf eins gesetzt wird. Die Gammaverteilung Ga(a, b) ist definiert durch die Dichte[[5]](#footnote-5):

für alle a, b > 0 (1)

Folglich ergeben sich Erwartungswert und Varianz zu:

und (2)

Die Daten erfüllen durch ihren Definitionsbereich auf den positiven Reellen zahlen und ihre Rechtsschiefe außerdem die Eigenschaften der Gammaverteilung. Möchte man diese Information in das Modell miteinfließen lassen, betrachtet man kann einfaches lineares Modell mehr, sondern ein generalisiertes lineares Modell (GLM). Dieses erlaubt es, eine Verteilungsannahme für die Zielvariable zu treffen. Die angenommene Verteilung unterliegt hierbei einer einparametrigen linearen Exponentialfamilie. Die Überprüfung dieser Voraussetzung ist hier jedoch nicht notwendig, da das finale Modell diese Anforderung nicht hat. Ein GLM zeichnet sich durch den linearen Prädiktor ηi = xtiβ und eine stetige, zwei Mal differenzierbare Linkfunktion aus. Diese verbindet den linearen Prädiktor mit dem Erwartungswert und hängt von der getroffenen Verteilungsannahme ab.[[6]](#footnote-6) In dem Fall der Gammaverteilung ergibt sich ein log-Link. Der Zusammenhang sieht somit wie folgt aus:

g(E(Y)) = xtiβ ⬄ E(Y) = g-1(xtiβ)

wobei g der Logarithmus und g-1 somit die Exponentialfunktion ist. Die Modellgleichung bleibt dadurch unverändert, lediglich die dahintersteckenden Annahmen passen sich an: Für die Zielvariable Y wird eine Gammaverteilung angenommen, wodurch der log-Link bei der Auswertung der Effekte berücksichtigt werden muss.

Ziel ist es nun, die Form des Einflusses der Kovariablen besser abzubilden, da eine lineare Modellierung eventuell nicht ausreichend ist. Die Idee hierbei ist es, die Effekte nicht mehr rein parametrisch zu schätzen, sondern über eine Summe glatter Polynomialfunktionen. Dafür wird der Wertebereich der Einflussvariable unterteilt und für jeden Abschnitt ein Polynom gerechnet. Die Grenzen der Bereiche werden als Knoten bezeichnet. Man spricht dann von einem generalisierten additiven Modell (GAM). Man erhält dann folgende Modellgleichung:

g(µi) = f1(zi1) + … + fq(ziq) + xtiβ

Hierbei sind z1, …, zq stetige Variablen, denn nur diese können glatt geschätzt werden. Alle vorigen Modellannahmen des GLM gelten weiterhin.

Es gibt verschiedene Ansätze für die Anwendung solcher Polynomialfunktionen. Einer davon sind B-Splines. Diese stellen den Anspruch, dass die Enden der Polynome kontinuierlich verlaufen und die Funktion somit insgesamt stetig ist. [[7]](#footnote-7)Hier werden die einzelnen Polynome als Basisfunktionen bezeichnet. Deren Summe ergibt sich nun folgendermaßen:

d bezeichnet die Anzahl der Basisfunktionen, welche Anhand der Zahl der inneren Knoten festgelegt wird. l gibt den Polynomgrad der Basisfunktion an. In der Praxis werden größtenteils kubische Polynome, also Polynome vom Grad 3 verwendet, da sie eine ausreichende Flexibilität bieten ohne Überanpassung an die Daten (genannt Overfitting). γ stellt den Vektor der Gewichte der einzelnen Basisfunktionen dar. Diese Summe kann somit in Form eines linearen Modells dargestellt werden und eine Schätzung des Gewichtsvektors ist über die Kleinste-Quadrate-Methode möglich. Das Problem hierbei ist, dass die Flexibilität des Glätters stark von der Anzahl innerer Knoten abhängt. Wählt man jedoch eine zu hohe Anzahl an Knoten, droht Overfitting. Um dem entgegenzuwirken, bieten die sogenannten penalisierten Splines (P-Splines) eine Möglichkeit zur automatisierten Wahl der Komplexität des Glätters. Dafür wird im ersten Schritt die zu schätzende Funktion f(x) durch Polynom-Splines mit einer sehr hohen Knotenanzahl approximiert, wodurch man eine große Flexibilität erreicht. Anschließend wird eine zu raue Schätzung durch einen Strafterm penalisiert. Man versucht dadurch eine zu starke Anpassung an die Daten zu vermeiden und gleichzeitig die Glättung des Effekts beizubehalten. Dafür wird der Gewichtsvektor γ hier über die Minimierung eines penalisierten KQ- Kriterium geschätzt:

P(γ) ist der Strafterm und der Glättungsparameter λ steuert hier den oben genannten Trade-Off: lässt man λ gegen unendlich gehen, so wird die Schätzung des Gewichtsvektors von dem Strafterm bestimmt und die Funktion f(x) passt sich zu stark an die Daten an. Hält man λ hingegen klein, so geht der Strafterm nur mit sehr geringem Gewicht in die Schätzung ein und es ergibt sich eine nahe am KQ-Schätzer gelegene Schätzung von γ. Der Vorteil der penalisierten Splines liegt somit darin, dass die Glättung anhand des Parameters λ gesteuert werden kann und unabhängig von Anzahl und Position der Knoten erfolgt.[[8]](#footnote-8) Es existieren zudem noch weitere Spline-Arten, insbesondere zyklische Splines, welche bei dem finalen Modell Anwendung finden. Sie beruhen auf der Annahme, dass Minimum und Maximum der Variable dieselbe inhaltliche Bedeutung haben und deren Effekte somit identische geschätzt werden.[[9]](#footnote-9)

Ziel der Auswertung ist es nun, die Verteilung der getriggerten Magnituden zu betrachten. Deshalb ist es notwendig nicht ausschließlich den Mittelwert zu betrachten, sondern auch weitere Verteilungsparameter in die Analyse miteinzubeziehen. Dementsprechend fiel die Entscheidung auf ein GAMLSS, ein generalisiertes additives Modell für Lage, Skalen- und Formparameter, welches nicht nur Mittelwertsregression, sondern eine flexible Modellierung der Zielvariable mithilfe weiterer Verteilungsparameter erlaubt. Es handelt sich hierbei um eine semiparametrische Regression: man hat die Möglichkeit die Effekte mithilfe von nicht-parametrischen Splinefunktionen glatt zu schätzen, jedoch wird, wie beim GLM eine parametrische Verteilungsannahme benötigt. Wie zuvor beschrieben, wird im Folgenden eine Gammaverteilung für unsere Zielvariable, die getriggerte Magnitude angenommen. Zusätzlich zu den bereits erwähnten Gründen, bietet diese Verteilung den Vorteil zwei Verteilungsparameter schätzen zu können. Es ergibt sich dadurch folgende Parametrisierung der Dichte Ga(σ - 2, (σ2µ)-1) aus (1):

wobei µ und σ die Bezeichnungen für die zu schätzenden Parameter des GAMLSS sind. Es gilt außerdem µ, σ > 0. Dadurch erhält man folgenden Erwartungswert und Varianz (siehe (2)):

und = σ2µ2

Letztendlich ergibt sich das finale Modell, jeweils für die Parameter µ und σ folgendermaßen:

Log(µ) = β0 + f1(triggernde Magnitude) + f2(heat Flow) + f3(strain rate) + f4(dip) + f5(depth) +

f6(rake) + f7(crustal thickness) + f8(Zeitdifferenz) + f9(completeness Magnitude) +

f10(mantle thickness)

Log(σ) = β0 + f1(triggernde Magnitude) + f2(heat Flow) + f3(strain rate) + f4(dip) + f5(depth) +

f6(rake) + f7(crustal thickness) + f8(Zeitdifferenz) + f9(completeness Magnitude) +

f10(mantle thickness)

wobei f1, …, f10 wie zuvor beschrieben für nicht-parametrische Splinefunktionen stehen. Hier wurden penalisierte Splines verwendet, bis auf f6 welche für zyklische Splines steht. Um die daraus geschätzten Werte sinnvoll als Effekte auf den Erwartungswert beziehungsweise die Varianz zu interpretieren, sind folgende Transformationen notwendig:

Effekt auf Erwartungswert = exp(µ) Effekt auf Varianz = exp(σ)2 \* exp(µ)2

Zur Erinnerung, das Anwenden der Exponentialfunktion ist aufgrund des log-Links erforderlich. Dadurch lassen sich die erhaltenen Effekte als multiplikative Effekte auf den Erwartungswert, beziehungsweise die Varianz der triggernden Magnitude interpretieren.

1. Masterarbeit [↑](#footnote-ref-1)
2. Fahrmeir [↑](#footnote-ref-2)
3. Fahrmeir 90 [↑](#footnote-ref-3)
4. Manthei, Gerd: Anwendung des Gutenberg-Richter-Gesetzes in der Schallemissionsanalyse, S. 1 [↑](#footnote-ref-4)
5. Meintrup seite 275 [↑](#footnote-ref-5)
6. Masterarbeit 11 [↑](#footnote-ref-6)
7. Fahrmeir 303 [↑](#footnote-ref-7)
8. Fahrmeir 308 [↑](#footnote-ref-8)
9. Simpson, Gavin: Modelling seasonal data with GAMs, 2014 https://fromthebottomoftheheap.net/2014/05/09/modelling-seasonal-data-with-gam/ [↑](#footnote-ref-9)