Technische Universität Dortmund
Fakultät Statistik
Wintersemester 2022/2023

Fallstudien I

Projekt 4: Regressionsmodelle für Zähldaten

Prof. Dr. Guido Knapp
M. Sc. Yassine Talleb

Bericht von: Louisa Poggel

Mitglieder der Gruppe 1:
Caroline Baer
Daniel Sipek
Julia Keiter
Louisa Poggel

22.12.2022

Inhaltsverzeichnis

1 Einleitung

Im Rahmen des Sozioökonischen Panles (SOEP) werden seit dem Jahr 1984 randomisiert ausgewählte Personen aus ganz Deutschland zu politischen un gesellschaftlichen Themen befragt. In diesem Projekt handelt es sich um die Analyse einer Befragung über den Gesundheitszustand über die Jahre 1984 bis 1994. Ziel ist es, mithilfe der Daten aus dem Jahr 1984, die Zielvariablen *Anzahl der Arztbesuche (in den letzten drei Monaten)* und *Anzhal der Krankenhausbesuche (im letzten Kalenerjahr)* mithilfe eines geeigneten generalisierten Modell für Zähldaten zu modellieren. Gegebenenfalls soll dabei eine getrennte Analyse nach Geschlecht erfolgen.
Zuerst wird sich herausstellen das Overdispersion ein Problem ist. Daraufhin wird aufgrund des quadratsichen Varianz-Erwartungswert Zusammenhangs der Zielvariable, ein Modell mit der Negativ Binomialverteilung ausgewählt. Da bei der Modellierung der *Anzahl an Arztbesuchen* die Variable *Geschlecht* signifikant ist, werden dort zudem zwei Teilmodelle getrennt nach Männern und Frauen erstellt. Bei der *Anzahl von Krankenhausbesuchen* ist dies nicht der Fall. Als wichtige Variable wird sich in beiden Modellen die *Zufriedenheit mit der Gesundheit* herausstellen. Bei der *Anzahl der Arztbesuche* spielt zudem die Variable *Behinderung* eine wichtige Rolle.
Der Bericht ist so aufgebaut, dass in Kapitel 2 zunächst die Problemstellung und der Datensatz genauer erläutert wird. Darauf folgt eine Übersicht der verwendeten statistischen Methoden in Kapitel 3. Dies ist aufgeteilt die die Unterkapitel Modelle für Zähldaten (3.1), Schätzungen und Tests (3.2) und Modellbildung und Variablenselktion (3.3). In Kapitel 4 erfolgt die statistische Auswertung. Dazu wird zuerst eine deskriptive Analyse der Daten durchgeführt (4.1). Dann erfolgt die Analyse der *Anzahl an Arztbesuchen* (4.2) und die Analyse der *Anzahl an Krankenhausaufenthalten* (4.3). Zuletzt werden alle zentralen Ergebnisse in Kapitel 5 zusammengefasst.

2 Problemstellung

Die seit dem Jahr 1984 in Deutschland durchgeführte Panelstudie des Sozioökonomischen Panels (SOEP) befragt randomisiert Privathaushalte zu politischen und gesellschaftlichen Themen. Dabei werden bis auf einige Ab- und Neuzugänge immer dieselben

Haushalte befragt. Mithilfe der entstandenen repräsentativen Zufallsstichprobe, der in Deutschland lebenden Menschen, lassen sich unter anderem gesundheitswissenschaftliche Fragestellungen beantworten. In diesem Projekt wird dazu der Datensatz *Gesundheitszustand.csv* betrachtet, der Daten des SOEP zum Gesundheitszustand aus sieben Jahren des Zeitraumes von 1984 bis 1994, beinhaltet. Dabei sind 27 Variablen und 27326 Beobachtungen enthalten.
Ziel diese Projektes ist die Analyse einer Querschnittsstudie des Jahres 1984. Dazu sollen die nun 3874 Beobachtungen aus dem Jahr 1984 eine Einschätzung über den Bedarf an medizinischer Versorgung in Deutschland liefern. In diesem Projekt werden die Zähldaten *Anzahl der Arztbesuche* (in den letzten drei Monaten) und *Anzahl der Krankenhausaufenthalte* (im letzten Kalenderjahr) als Regressanden in einem generalisierten linearen Regressionsmodell verwendet. Gegenbenenfalls erfolgt dabei eine getrennte Analyse nach Männern und Frauen. Eine Analyse der gesamten Panel Studie ist in Riphanh et al. (2003) zu finden.
Als Regressoren stehen die metrischen Variablen *monatliches Haushaltsenttoeinekommen* (in DM), *Grad der Behinderung* (in Prozent) und *Anzahl der Schuljahre* zur Verfügung. Weitere sind die ordinale Variable *Zufriedenheit der Gesundheit* (auf einer Skala von 0 bis 10) und die dichotomen Variablen *Geschlecht*, *Behinderung*, *Heirat*, *Kinder*, *Beschäftigungsverhältniss*, *Angestellter*, *Beamter*, *Selbständiger*, *Arbeiter*, *Krankenversicherung* und *Zusatzkrankenversicherung*. Zudem wird der höchste Schulabschluss erfasst, welcher in die fünf dichotomen Variablen *Hauptschule*, *Realschule*, *Abitur*, *Fachhochschule* und *Hochschule* geliedert ist.
Dabei enthält keine der Variablen fehlende Werte, jedoch treten oftmals fragwürdige Werte auf, die nicht zum Skalanniveau der Variablen passen. Daher wurden einige Beobachtungen mit unpassenden Werten entfernt. Dazu gehören die vier reelen Zahlen in der Variable *Zufriedenheit der Gesundheit* und die 12 nicht dichotomen Werte der Variable *Behinderung*. Zudem wurden einige Werte der *Anzahl der Schuljahre* so gerundet, dass die Anzahl der Schuljahre auf halbe Jahre genau vorliegen. Da beim höchsten Schulabschluss teilweise mehrere Abschlüsse angekreuzt wurden, wird so umkodiert, dass der jeweilige höchste Abschluss zählt. Außerdem werden zwei Variablen des Datensatzes bezeichnet mit *hast2* und *newhsat*, aufgrund fehlender Informationen zu dessen Bedeutung, nicht weiter betrachtet.

3 Statistische Methoden

Alle folgenden statistischen Methoden werden in der Version 4.1.1 der Software R durchgeführt (R Core Team (2021)). Dabei wird bei Ergebnissen, wenn nicht anderes angegeben, auf zwei Nachkommastellen gerundet.

3.1 Modelle für Zähldaten

Für einen Regressand y mit Beobachtungen $y_1, \ldots, y_n$ in Form von Zähldaten sollte, aufgrund der Verletzung der Normalverteilungsannahme, ein generalisiertes lineares Modell angewendet werden. Passende Verteilungen zur Modellierung sind die Poisson Verteilung oder die Negative Binomialverteilung.
Problem bei Modellierung mit der Poisson Verteilung ist, dass gilt $\mathbb{E}(y_i) = \mathbb{V}ar(y_i) = \lambda_i$ mit $\lambda_i > 0$ und $i = 1, \ldots, n$. Oftmals ist die Varianz von y_i jedoch deutlich größer. Dieses Problem nennt man Overdispersion (Dunn und Symth (2018), S. 397). Eine Lösung dafür ist das Quasi Poisson Modell, bei dem $\mathbb{E}(y_i) = \lambda_i$ und $\mathbb{V}ar(y_i) = \phi\lambda_i$ gilt. Das ϕ ist ein Dispersionsparameter mithilfe dessen die Varianz von y größer modelliert werden kann. Unter Verwendung der Link Funktion $log(\lambda) = \eta$ lässt sich dann ein generalisiertes lineares Modell (1) ausftellen, dass die zufällige Komponente $\mathbb{E}(y|x)$ mit dem linearen Prediktor η verbindet. Dabei bezeichnet $k+1 \in \mathbb{N}$ die Anzahl der Parameter des Modells (Fahrmeir et al. (2009), S. 226, 227).

$$log(\ \mathbb{E}(y_i|x_{i1}, \ldots, x_{ik})\ ) = log(\lambda_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} \quad i = 1, \ldots, n \qquad (1)$$

In R lässt sich dieses Modell mithilfe der `glm` Funktion unter Angabe des Argumentes `family = "quasipoison"` umsetzten.
Alternativ lassen sich Zähldaten mithife eines generalisierten Modells auf Basis der negativen Binomialverteilung unter Verwendung eines log-Links modellieren (Dunn und Symth (2018) S. 399 bis 401 und Hilbe (2011) S. 185 ff.). Zusätzlich zur Verteilungsannahme $y_i|\lambda_i \sim Pois(\lambda_i)$ wird angenommen, dass $\lambda_i \sim Gam(\mu_i, \psi)$ ist. Es ergibt sich, dass $\mathbb{E}(y_i) = \mu_i$ und $\mathbb{V}ar(y_i) = \mu_i + \psi\mu_i^2$ gilt. Das heißt auch mit diesem Modell lässt sich Overdispersion modellieren. Die Wahrscheinlichkeitsdichte von y_i folgt dann der einer negativen Binomialverteilung und lautet wie in 2. Und die Modellgleichung lautet wie

in 3.

$$f(y_i, \mu_i, m) = \frac{\Gamma(y_i + m)}{\Gamma(y_i + 1)\Gamma(m)} \left(\frac{\mu_i}{\mu_i + m}\right)^{y_i} \left(1 - \frac{\mu_i}{\mu_i + m}\right)^{m} \qquad m = \frac{1}{\psi} \tag{2}$$

$$log(\ \mathbb{E}(y_i|x_{i1}, \ldots, x_{ik})\ ) = log(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} \quad i = 1, \ldots, n \tag{3}$$

Dabei bezeichnet $\Gamma()$ die Gamma Funktion. In R lässt sich das Modell mit der Funktion `glm.nb` aus dem Paket `MASS` implementieren (Venables und Ripley (2002)). Der Unterschied zwischen dem Quasi Poisson Modell und dem Modell der negativen Binomialverteilung liegt in der verschiedenen Modellierung der Dispersion. Beim Quasi Poisson Modell liegt eine linearer und beim Modell der Negativen Bimonialverteilung ein quadratischer Zusammenhang zwischen Varainz und Erwartungswert vor (Dunn und Symth (2018) S. 403).

3.2 Schätzungen und Tests

Da später lediglich als finales Modell das Negativ Binomialmodell verwendet wird, konzentriert sich dieser Abschnitt auf eben dieses. Denn dort werden simultane Maximum Likelihood Schätzungen verwendet, die das m und die Koeffizienten des Parametervektors β simultan schätzen (Dunn und Symth (2018), S. 400). Dabei wird zunächst die Likelihood Funktion wie in 4 als Produkt aller einzelnen Dichten von $y_i, \ldots, y_n$ gebildet. Zur Erleichterung der Differenzierung in 6 wird die Log-Likelihood wie in 5 gebildet. Um Parameterschätzungen für m und β zu erhalten wird sowohl die Ableitung der Log-Likelihood nach m als auch nach β gebildet (6) (Hilbe (2011), S. 191, 192). Diese sogenannten Score-Funktionen werden gleich Null gesetzt um das Maximum der Likelihood Funktion zu berechnen.

$$L(\beta, m; y) = \prod_{i=1}^{n} f(y_i, \mu_i, m) \tag{4}$$

$$l(\beta, m; y) = \log(L(\beta, m; y)) = \sum_{i=1}^{n} \log(f(y_i, \mu_i, m)) \tag{5}$$

$$s_1(\beta, m; y) = \frac{\partial}{\partial \beta} l(\beta, m; y) \quad ; \quad s_2(\beta, m; y) = \frac{\partial}{\partial m} l(\beta, m; y) \tag{6}$$

Eine Lösung dieses Problems erfolgt numerisch in R über ein wechselndes iteratives Verfahren („alternating iteration process“). Dieses ist in der Funktion `glm.bn` implementiert

(Venables und Ripley (2002)).
Es lässt sich zeigen, dass Maximum Likelihood Schätzer $\hat{\beta}$ unter Regularitätsbedingungen konsistent und asymptotisch normalverteilt sind. Somit gilt $\hat{\beta} \overset{a}{\sim} N(\beta, F^{-1}(\hat{\beta}))$, wobei $F^{-1}(\hat{\beta})$ die Inverse der Fisher-Informationsmatrix an der Stelle $\hat{\beta}$ ist (Fahrmeir et al. (2009), S. 224). Unter diesen Annahmen lässt sich der Wald-Test/z-Test herleiten (Fahrmeir et al. (2009) S. 225 und Groß (2010) S. 227). Denn für allgemeine Hypothesen $H_0 : C\beta = 0 \quad v.s. \quad H_1 : C\beta \neq d$ mit $d \in \mathbb{R}^k, C \in \mathbb{R}^{rxk}$ und $r = rg(C)$ gilt, dass die Teststatistik g in (7) approximativ χ^2_r verteilt ist. Somit ergibt sich in (8) für den Signifikanztest der Parameter in Form der Hypothesen $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ für ein $j \in \{1, \ldots, n\}$ die Teststatistik z, die approximativ Standardnormalverteilt ist.

$$g = (C\hat{\beta} - d)^T (CF^{-1}(\hat{\beta})C^T)^{-1}(C\hat{\beta} - d) \sim \chi^2_r \tag{7}$$

$$g = \hat{\beta_j}^T F_j^{-1}(\hat{\beta})\hat{\beta_j} = \frac{\beta_j^2}{a_{jj}} \Rightarrow z := \sqrt{g} = \frac{\hat{\beta}_j}{\sqrt{a_{jj}}} = \frac{\hat{\beta}_j}{\sqrt{\mathbb{V}ar(\hat{\beta}_j)}} \overset{a}{\sim} N(0,1) \tag{8}$$

Dabei bezeichnet a_{jj} das j-te Diagonalelement der Fisher-Informationsmatrix, dass der geschätzten Varianz von $\hat{\beta}_j$ entspricht. Der z-Test lehnt H_0 ab, falls $z > q_{1-\frac{\alpha}{2}}$ ist. Da das Modell der negativen Binomialverteilung auf der Maximum Likelihood Schätzung basiert, lassen sich dort z-Tests anwenden.
Zur Analyse des Varianz Erwartungswert Zusammenhangs, im Rahmen der Modellbildung, wird zudem die in R implementierte Schätzung für den Dispersionsparamer ϕ im Quasi-Poisson Modell genutzt. Das ϕ wird wie in (9) mithilfe der Person Statistik X^2 geschätzt (Dunn und Symth (2018) S. 255).

$$\hat{\phi}_P = \frac{X^2}{n-p} \qquad X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu})/\omega_i} \overset{Quasi-Poisson}{\Rightarrow} \sum_{i=1}^{n} \frac{(y_i - \hat{\lambda}_i)^2}{\phi\hat{\lambda}_i} \tag{9}$$

Die in (9) verwendete Funktion $v(\hat{\mu}_i)$ bezeichnet die geschätzte Varianz von y. Die geschätzten Werte $\hat{\mu}_i$ lassen sich mithilfe der Umkehrfunktion der Link-Funktion berechnen und lauten im Quasi-Poisson Modell $\hat{\lambda}_i = exp(\hat{\eta}_i)$. Das Gewicht ω_i kann bei Individualdaten als eins gesetzt werden. Die Person Statistik lässt sich also in diesem Projekt konkret wie in 9 spezifizieren.

3.3 Modellbildung und Variablenselktion

Als Maß für die Güte der Anpassung des Modells wird die Devianz genutzt. Diese entspricht im Kontext von Modellbildung der Summe der Abweichungsquadrate (Hedderich und Sachs (2016), S. 795) und ist wie in 10 definiert.

$$D = -2\phi \sum_{i=1}^{n} \{l_i(\hat{\mu}_i) - l_i(y_i)\} \tag{10}$$

Dabei entspricht $l_i(y_i)$ der Log-Likelihood der i-ten Beobachtung des gesättigten („saturated“) Modells. Dieses hat genauso viele erklärende Variablen wie y-Werte. Analog ist $l_i(y_i)$ die Log-Likelihood der i-ten Beobachtung des interessierenden Modells. Da ein perfekt angepasstes Modell eine Devianz von null hätte (Hedderich und Sachs (2016), S. 795), ist eine möglichst geringe Devianz gewünscht.
Ein weiteres Kriterium für die Anpassung des Modells und ein Hilfsmittel für die Variablenselektion ist das Akaike Informationskriterium (AIC), dass wie in 11 definiert ist (Fahrmeir et al. (2009), S. 226).

$$AIC = -2l(\hat{\beta}) + 2p \tag{11}$$

Dabei bezeichnet das p die Anzahl der zu schätzenden Parameter $(k + 1)$ und ist Teil eines Strafterms, der die Hinzunahme von weiteren Variablen bestraft. Der AIC lässt sich so interpretieren, dass je kleinerer der AIC ist, desto besser ist die Modellanpassung ist. Somit lässt sich anhand des AIC eine schrittweise Variablenselektion durchführen. Da es jedoch möglich ist das das daraus resultierende Modell nicht signifikante Variablen beeinhaltet, wird in diesem Projekt gefordert, dass alle Variablen zumindest zum Niveau $\alpha = 0.1$ signifikant sein sollen. Ist dies nicht der Fall erfolgt zusätzlich eine Rückwärtselimination anhand der p-Werte. In Bezug auf den AIC ist außerdem noch zu erwähnen, dass dieser nicht beim Quasi Poisson Modell angewendet werden kann, da dort nur eine Quasi Likelihood Funktion existiert (Fahrmeir et al. (2009), S. 226).
Ein weiteres zu prüfendes Problem, dass bei der Modellierung von Zähldaten auftreten kann, ist die Zero Inflation. Das heißt, dass in den Daten mehr Nullen vorhanden sind als das Modell prognostiziert. Dazu wird die Funktion `check_zeroinflation` aus dem Paket `performance` genutzt (Lüdecke et al. (2021)). Dort wird die Anzahl der durch das Modell prognostizierten nullen durch die Anzahl der tatsächlich im Datensatz vorhandenen nullen geteilt. Liegt dieses Verhältnis nahe 1 liegt keine Zero Inflation vor. Als

Daumenregel wird in diesem Projekt verwendet, dass für eine Abweichungen von 5% keine Zero Inflation vorliegt und das Abeichungen bis 10% zumindest auf eine leichte Zero Inflation hindeuten.

4 Statistische Auswertung

4.1 Deskriptive Analyse

In Tabelle 1 sind die Kennzahlen der metrischen Variablen zu finden. Im Mittel liegt die *Anzahl an Arztbesuchen* bei etwa 3.16 mit einer Standardabweichung von 6.28 (Varianz $\approx$ 39.44). Bei der *Anzahl der Krankenhausaufenthalte* liegt das arithmetische Mittel deutlich niedriger bei etwa 0.12. Die Standardabweichung liegt bei etwa 0.70 (Varianz $\approx$ 0.49). Somit beträgt die Varianz bei der *Anzahl an Arztbesuchen* etwa das 12-fache und bei der *Anzahl der Krankenhausaufenthalte* etwa das 4-fache des arithmetischen Mittels. Außerdem ist auffällig, dass der Interquartilabstand bei der *Anzahl an Krankenhausaufenthalte* und beim *Grad der Behinderung* bei Null liegt. Dies liegt daran, dass sehr viele Nullen vorhanden sind, da andere Merkmalsausprägungen bei den Individuen selten vorkommen. In Tabelle 5 auf Seite 16 im Anhang ist erkennbar, das auch einige Merkmalsausprägungen der dichotomen Merkmale recht selten vorkommen. Dazu gehört die *Zusatzkrankenverischerung* (0.4%), jeweils höhere Bildungsabschlüsse als der *Hauptschulabschluss* (Gesamtanteil höherer Abschlüsse: 31.3%), *Selbständige* (6.1%), *Beamte* (7.4%) und keine *Krankenversicherung* (9.7%). Bei der Variable *Zufriedenheit mit der Gesundheit* ist in Tabelle 4 auf Seite 16 erkennbar, dass auf der Skala Werte größer als 4 häufiger gewählt werden.

Tabelle 1: Lage- und Streuungsmaße der metrischen Variablen (IQA = Interquartilsabstand, A. = Anzahl)

	Arithm. Mittel	Standardabweichung	IQA
A. Arztbesuche	3.16	6.28	4.00
A. Krankenhausaufenthalte	0.12	0.70	0.00
Hushaltsnettoeinkommen	2969.08	1479.28	1595.22
Grad der Behinderung	6.59	19.97	0.00
A. Schuljahre	11.09	2.23	1.50

4.2 Analyse der *Anzahl der Arztbesuche*

Wie bereits festgestellt ist die Varianz der Variable *Anzahl der Arztbesuche* mit etwa 39.44 ungefähr das 12-fache des arithmetischen Mittels (6.28). Somit muss Overdispersion modelliert werden.
Dazu wird der Varianz-Erwartungswert Zusammenhang in Abbildung 1 geschätzt. Es werden die 3858 Beobachtungen in 205 Gruppen mit je 19 Beobachtungen aufgeteilt und das arithmetische Mittel sowie die empirische Varianz berechnet. Durch Anpassung eines Quasi-Poisson und Negativ Binomial Modells mit allen möglichen Regressoren werden Schätzungen für $\hat{\phi}$ (7.24) und $\hat{\psi}$ (1.61) gewonnen. Mithilfe dessen lässt sich der Varianz-Erwartungswert Zusammenhang im Quasi-Poisson (violett) und Negativ Binomialmodell (blau) schätzen. Da die blaue Kurve den Verlauf der Daten am besten beschreibt, wird für die Modellierung der *Anzahl an Arztbesuchen* ein Negativ Binomialmodell gewählt. Dafür spricht auch die Devianz des Modells mit der Negativ Binomialverteilung (3960.6), welche deutlich geringer ist, als bei der Quasi-Poisson Regression (18022).

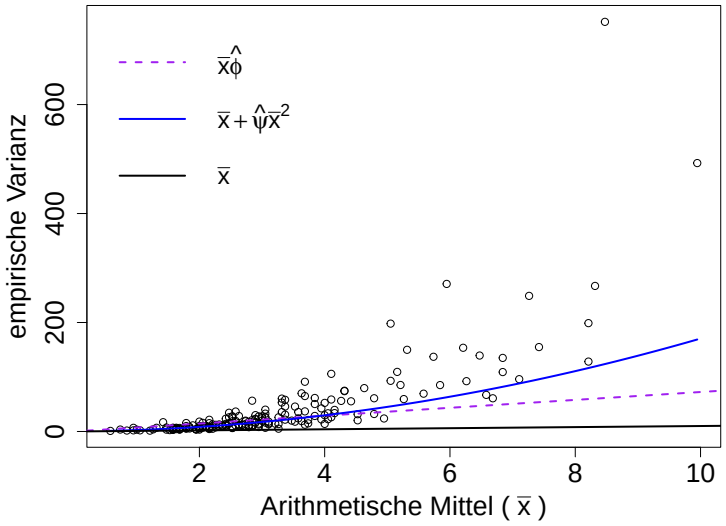

Abbildung 1: Geschätzter Zusammenhang zwischen Varianz und Erwartungswert

Anhand des AIC erfolgt dann eine Variablenselektion, bei der alle Regressoren bis auf *Geschlecht*, *Zufriedenheit mit der Gesundheit*, *Behinderung*, *Grad der Behinderung*, *monatliches Haushaltsnettoeinkommen*, *Kinder* und *Zusatzkrankenversicherung* eliminiert.

Dabei wird der AIC des vollen Modells von 15838 auf 15817 verringert und auch die Devianz ist mit 3960.2 etwas geringer. Alle enthalten Variablen bis auf die Dummy Variable *Zufriedenheit der Gesundheit (Bewertung: 1)* sind signifikant zum Niveau $\alpha = 0.1$. Jedoch prognostiziert das Modell mit 1523 weniger Nullen als tatsächlich in der Variable *Anzahl der Arztbesuche* vorkommen. Das Verhältnis beträgt etwa 0.95, welches auf eine leichte Zero Inflation hindeuten könnte. Da das Verhältnis aber nicht zu stark von eins abweicht wird das Negativ Binomialmodell beibehalten.
In Tabelle 2 sind die Koeffizientenschätzungen und p-Werte des finalen Modells abgebildet. Zur Besseren Interpretation wurden die Koeffizienten mit der Exponentialfunktion multipliziert. Somit kann man beispielsweise interpretieren, dass falls eine *Behinderung*

Tabelle 2: Transformierte Koeffizientenschätzungen und p-Werte des finalen Modells für die *Anazhl an Arztbesuchen* (gerundet auf fünf Nachkommastellen)

Variable	$\exp(\hat{\beta}_i)$	p-Wert
Intercept	8.92645	$< 2 \cdot 10^{-16}$
Geschlecht (weiblich)	1.63126	$< 2 \cdot 10^{-16}$
Zufriedenheit mit der Gesundheit (Bewertung: 1)	0.94012	0.798788
Zufriedenheit mit der Gesundheit (Bewertung: 2)	0.65710	0.020998
Zufriedenheit mit der Gesundheit (Bewertung: 3)	0.56965	0.000899
Zufriedenheit mit der Gesundheit (Bewertung: 4)	0.39056	$4.17 \cdot 10^{-08}$
Zufriedenheit mit der Gesundheit (Bewertung: 5)	0.34872	$2.45 \cdot 10^{-13}$
Zufriedenheit mit der Gesundheit (Bewertung: 6)	0.32727	$1.88 \cdot 10^{-12}$
Zufriedenheit mit der Gesundheit (Bewertung: 7)	0.24537	$< 2 \cdot 10^{-16}$
Zufriedenheit mit der Gesundheit (Bewertung: 8)	0.18336	$< 2 \cdot 10^{-16}$
Zufriedenheit mit der Gesundheit (Bewertung: 9)	0.11051	$< 2 \cdot 10^{-16}$
Zufriedenheit mit der Gesundheit (Bewertung: 10)	0.09995	$< 2 \cdot 10^{-16}$
Behinderung (vorhanden)	2.53287	$6.23 \cdot 10^{-08}$
Grad der Behinderung	0.99302	0.010824
monatliches Haushaltsnettoeinkommen (in DM)	0.99997	0.084488
Kinder unter 16 Jahren (vorhanden)	0.89538	0.022061
Zusatzkrankenversicherung (vorhanden)	0.47231	0.092969

vorhanden ist, die mittlere *Anzahl an Arztbesuchen* um etwa den Faktor 2.54 steigt. dabei wird angenommen, dass alle weiteren Variablen konstant im Modell enthalten sind. Zur einzigen weitere Variable mit erhöhendem Faktor zählt das weibliche *Geschlecht*. Alle anderen Faktoren sind kleiner als eins, das heißt sie verringern die mittlere *Anzahl an Arztbesuchen* um einen gewissen Faktor. Der Intercept lässt sich nicht sinnvoll interpretieren, da es in der Regel kein *monatliches Haushaltsnettoeinkommen* von null gibt.
Da die Variable *Geschlecht* einen signifikanten Einfluss auf die *Anzahl der Arztbesuche*

hat, sind auch getrennte Analysen sinnvoll. Auch hier wird das volle Negativ Poisson Modell und das Negativ Binomial Modell angepasst und verglichen.
Bei den Frauen ist die Devianz im Quasi-Poisson Modell mit 9345.9 deutlich größer als im Negativ Binomialmodell (2014.7). Somit wird eine Varaiblenselktion anhand des AIC am Negativ Binomialmodell durchgeführt. Es resultiert ein Modell mit den Variablen *Zufriedenheit mit der Gesundheit*, *Behinderung*, *Grad der Behinderung*, *monatliches Haushaltsnettoeinkommen*, *Kinder*, *Hauptschulabschluss*, *Angestellter* und *Selbständiger* mit einem AIC von 8525.4. Da die Variable *monatliches Haushaltsnettoeinkommen* mit etwa 0.14 einen p-Wert größer als 0.1 hat. Daraufhin ergibt sich jedoch eine schrittweise Elimination der Variablen *Heirat* und *Angestellte* wegen zu großen p-Werten. Dann sind alle Variablen, bis auf die Dummy Variable *Zufriedenheit der Gesundheit (Bewertung: 1)* signifikant zum Niveau $\alpha = 0.1$. Dieses Modell führt zu einem AIC von 8525.8 und einer Devianz von 2014.6. Das Verhältnis von vorhergesagten und tatsächlich vorhandenen Nullen liegt bei 0.93. Dies könnte auf eine leichte Zero Inflation hindeuten.
Bei den Männern führt ein analoges Vorgehen zur Auswahl des negativen Binomialmodells (Devianz Quasi Poisson: 8373.2, Devianz Negative Binomialverteilung: 1912.2). Eine Variablenselektion anhand des AIC führt zu einem Modell mit den Variablen *Alter*, *Zufriedenheit mit der Gesundheit*, *Behinderung*, *Grad der Behinderung*, *Heirat*, *Realschulabschluss*, *Hochschulabschluss* und *Zusatzkrankenversicherung*. Alle Variablen bis auf die Dummy Variablen *Zufriedenheit der Gesundheit (Bewertung: 1 und 2)* sind signifikant zum Niveau $\alpha = 0.05$. Der AIC wurde von 7283.9 auf 7267.7 reduziert und die Devianz liegt mit 1911.3 etwas geringer als im vollen Modell. Da das Modell 994 nullen prognostiziert und in den Daten tatsächlich 960 nullen vorhanden sind (Verhältnis: 0.97), liegt vermutlich kein Problem mit Zero Inflation vor.

4.3 Analyse der *Anzahl der Krankenhausaufenthalte*

Da bereits in der deskriptiven Analyse feststellt wurde, dass bei der Variable *Anzahl der Krankenhausaufenthalte* Overdispersion ein Problem ist (Varianz ≈ 0.49, Arithmetische Mittel ≈ 0.12), gilt es nun den korrekten Zusammenhang zwischen Erwartungswert und Varianz zu ermitteln. Dieser wird in Abbilung 2 geschätzt.

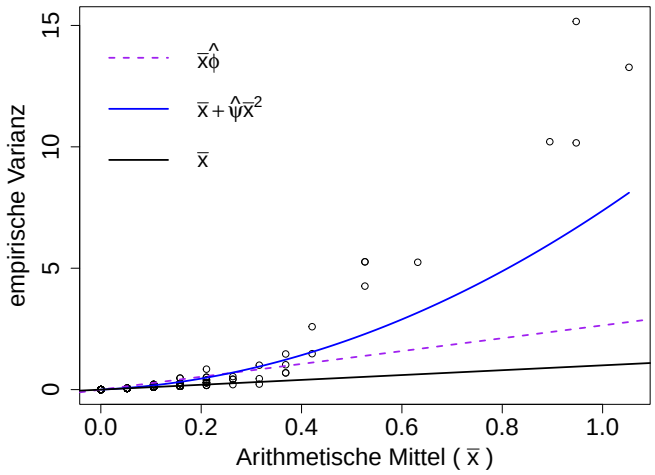

Abbildung 2: Geschätzter Zusammenhang zwischen Varianz und Erwartungswert

Dabei wurde zunächst der Dispersionsparameter ϕ aus einem Quasi Poisson Modell mit allen Regressoren als etwa 2.65 geschätzt. Analog ergibt sich aus dem vollen Negativ Binomialmodell die Schätzung für ψ als etwa 6.37. Dann ist in Abbildung 2 erkennbar, dass die blaue Kurve am besten zu dem Verlauf der Punkte passt. Da diese der geschätzten Varianz bei Modellierung mit der Negativ Binomialverteilung entspricht, wird das Negativ Binomial Modell verwendet. Diese Wahl lässt sich auch über die Devianz begründen, da diese mit 2413 im Quasi-Poisson Modell größer ist als im Negativ Binomialmodell (Devianz = 1012.4, AIC = 2570.3).
Nun erfolgt eine schrittweise Variablenselektion anhand des AIC, bei der ein Modell mit den elf Variablen *Zufriedenheit mit der Gesundheit*, *Grad der Behinderung*, *monatliches Haushaltsnettoeinkommen*, *Anzahl an Schuljahre*, *Heirat*, *Realschule*, *Fachhochschule*, *Beschäftigungsverhältniss*, *Arbeiter*, *Angestellter* und *Beamter* resultiert. Dieses hat eine Devianz von 1011.3 und einen AIC von 2555.6. Um sicherzustellen, dass alle Variablen das Signifikanzniveau von $\alpha = 0.1$ einhalten, wird zudem die Variable *Grad der Behinderung* (p-Wert ≈ 0.14) entfernt. Nun sind alle Variablen signifikant zum Niveau $\alpha = 0.1$ bis auf die zwei Dummy Variablen *Zufriedenheit der Gesundheit (Bewertung 1 und 2)*.Das finale Modell hat dann einen einen geringfügig höheren AIC (2555.7) und eine Devianz von 1012.5.
Ein Problem mit Zeroinflation liegt nicht vor, da das Modell mit 3566 nullen etwa genauso viele nullen prognostiziert wie in den Daten sind (Anzahl = 3562). Zudem ist eine Analyse getrennt nach Frauen und Männern nicht sinnvoll, da das *Geschlecht* bei der

Variablenselektion aufgrund fehlendem Einfluss auf die Zielvarible eliminiert wird.
In Tabelle 3 sind die transformierten Koeffizientenschätzungen und p-Werte der Variablen des finalen Modells dargestellt. Analog zum Modell in Kapitel 4.2 man beispiels-

Tabelle 3: Transformierte Koeffizientenschätzungen und p-Werte des finalen Modells der *Anzahl an Krankenhausaufenthalten* (gerundet auf fünf Nachkommastellen)

Variable	$\exp(\hat{\beta}_i)$	p-Wert
Intercept	0.29919	0.01144
Zufriedenheit mit der Gesundheit (Bewertung: 1)	1.24703	0.67490
Zufriedenheit mit der Gesundheit (Bewertung: 2)	0.60767	0.22681
Zufriedenheit mit der Gesundheit (Bewertung: 3)	0.40819	0.02273
Zufriedenheit mit der Gesundheit (Bewertung: 4)	0.25446	0.00105
Zufriedenheit mit der Gesundheit (Bewertung: 5)	0.25896	$3.47 \cdot 10^{-05}$
Zufriedenheit mit der Gesundheit (Bewertung: 6)	0.19982	$2.14 \cdot 10^{-05}$
Zufriedenheit mit der Gesundheit (Bewertung: 7)	0.18948	$1.50 \cdot 10^{-06}$
Zufriedenheit mit der Gesundheit (Bewertung: 8)	0.08386	$2.78 \cdot 10^{-12}$
Zufriedenheit mit der Gesundheit (Bewertung: 9)	0.07725	$3.03 \cdot 10^{-10}$
Zufriedenheit mit der Gesundheit (Bewertung: 10)	0.13055	$2.13 \cdot 10^{-09}$
monatliches Haushaltsnettoeinkommen (in DM)	0.99990	0.08306
Anzahl der Schuljahre	1.06276	0.08527
Heirat (vorhanden)	1.48410	0.02721
Höchster Schulabschluss: Realschule	0.72008	0.09766
Höchster Schulabschluss: Fachhochschule	0.42635	0.08587
Beschäftigungsverhältniss (vorhanden)	1.62883	0.03717
Arbeiter (ja)	0.54625	0.01545
Angestellter (ja)	0.63491	0.06540
Beamter (ja)	0.42848	0.02161

weise interpretieren, dass verheiratete Personen im Mittel eine um etwa den Faktor 1.48 erhöhte *Anzahl an Krankenhausbesuchen* haben. Weitere Umstände, die die *Anzahl an Krankenhausbesuchen* im Mittel um einen gewissen Faktor erhöhen sind die *Zufriedenheit mit der Gesundheit* mit Bewertung 1 und ein vorhandenes *Beschäftigungsverhältniss*. Zudem wirkt sich eine höhere Anzahl an Schuljahren erhöhend auf die *Anzahl an Krankenhausbesuchen* aus. Um einen gewissen Faktor verringernd wirken sich Bewertungen zwischen 2 und 10 der *Zufriedenheit mit der Gesundheit*, pro DM *Hauhaltsnettoeinkommen*, Höchste Schulabschlüsse der Schulformen *Realschule* und *Fachhochschule* und ein *Beschäftigungsverhältniss* in Form eines *Arbeiters*, *Angestellten* oder *Beamten* aus. Zudem wirkt sich ein höheres *Hauhaltsnettoeinkommen* minimal verringernd auf die *Anzahl an Krankenhausbesuchen* aus. Der Intercept wird in Regelfall nicht interpretierbar sein, da das *Haushaltsnettoeinkommen* und die *Anzahl an Schuljahren* nicht null werden.

5 Zusammenfassung

Ziel dieses Projektes ist es jeweils, die Zielvariablen *Anzahl der Arztbesuche (in den letzten drei Monaten)* und *Anzahl der Krankenhausbesuche (im letzten Kalenderjahr)*, mithilfe eines geeigneten gneralisierten Modells für Zähldaten zu analysieren. Gegebenenfalls soll dabei eine getrennte Anayse nach Männern und Frauen erfolgen. Die dazu verwendeten Daten des Datensatzes *Gesundheitszustand.csv* beinhalten Informationen bezüglich des Gesundheiszustandes von randomisiert ausgewählten Personen in Deutschland aus dem Jahr 1984. Erhoben wurden die Daten im Rahmen des Sozioökonomischen Panel (SOEP).
Zur Modellierung stellte sich für beide Zielvariablen, aufgrund der Overdispersion in Form eines quadratischen Erwartungswert-Varianz Zusammenhangs, das Modell mit der Negativen Binomialverteilung als geeignet heraus. Nach der Modellwahl erfolgte eine Variablenselektion anhand des AIC. Dabei stellte sich heraus, dass die Variable *Zufriedenheit der Gesundheit* eine zentrale Rolle spielt. Denn diese ist in allen finalen Modellen enthalten. Dabei sind in vielen Modellen die hohen Bewertungsstufen mit einem p-Wert nahe null sehr signifikant. Generell lässt sich sagen, dass hohe Bewertungen der *Zufriedenheit der Gesundheit* zu weniger *Arztbesuchen und Krankenhausaufenthalten* führen. Sehr niedrige Bewertungen hingegen zu einer höheren Anzahl. Dieser Effekt könnte an einer möglichen Korrelation zwischen der *Zufriedenheit der Gesundheit* mit dem wahren Gesundheitszustand liegen.
Bei der *Anzahl der Arztbesuche* ist erkennbar, dass die *Behinderung* eine wichtige Rolle spielt und sich steigernd auf die Anzahl der Arztbesuche auswirkt. Denn die Variable hat einem Faktor von etwa 2.53 und hat mit $6.23 \cdot 10^{-16}$ einen sehr kleinen p-Wert. Außerdem sind signifikante Unterschiede zwischen Frauen und Männern zu ernennen (p-Wert: $< 10^{-16}$). Mit einem Faktor von etwa 1.63 wirkt sich ein weibliches *Geschlecht* steigernd auf die *Anzahl an Arztbesuchen* aus. Weitere enthaltene Variablen des Modells sind *Grad der Behinderung*, *Haushaltsnettoeinkommen*, *Kinder (unter 16 Jahren)* und *Zusatzkrankenversicherung*.
Aufgrund der Geschlechterunterschiede werden zudem zwei Teilmodelle erstellt. In diesen stellt sich heraus, dass in beiden Modellen, zusätzlich zur *Behinderung*, die *Höchsten Schulabschlüsse* relevant werden. Ansonsten hat bei den Männern in Gegensatz zu den Frauen das *Alter* einen steigernden Effekt auf die *Anzahl an Arztbesuchen*. Bei den Frauen lässt sich erkennen, dass sich *Kinder (unter 16 Jahren)* und die *Selbständigkeit*

verringernd auf die *Anzahl an Arztebesuchen* auswirken.
Bei der *Anzahl an Krankenhausaufenthalten* spielt die *Behinderung* keine relevante Rolle mehr und ist nicht mehr im Modell enthalten. Dort wirken sich zusätzlich zur *Zufriedenheit der Gesundheit* unter anderem die Variablen *Heirat* und *Beschäftigungsverhältniss* steigernd auf die *Anzahl der Krankenhausaufenthalte* heraus. Verschiedenste Arten von *Beschäftigungsverhältnissen* (*Arbeiter*, *Angestellter*, *Beamter*) wirken sich wiederum verringernd auf die Anzahl aus. Zudem wirkt sich eine höhere *Anzahl an Schuljahren* erhöhend auf die *Anzahl an Krankenhausbesuchen* aus. Einzelne *Höchste Schulabschlüsse* wirken jedoch wiederum verringernd und relativieren den Effekt.
Bezüglich der Modelle der *Anzahl an Arztbesuchen*, sollte zudem bemerkt werden, dass die Diskrepanz der durch das Modell prognostizierten Nullen und tatsächlich in den Daten enthaltenden Nullen teilweise erhöht war. Somit ist das Negativ Binomialmodell nicht total unplausibel, aber es könnte teilweise die *Anzahl an Arztbesuchen* überschätzen. Alternativ könnte auch der Ansatz eines Zero Inflated Modell gewählt werden.

Literatur

Dunn, P. und G. Symth (2018). *Generalized Linear Models With Examples in R*. Springer Texts in Statistics. Springer New York.

Fahrmeir, L., T. Kneib und S. Lang (2009). *Regression: Modelle, Methoden und Anwendungen*. Statistik und ihre Anwendungen. 2. Auflage. Springer Berlin Heidelberg.

Groß, J. (2010). *Grundlegende Statistik mit R: Eine anwendungsorientierte Einführung in die Verwendung der Statistik Software R*. Vieweg+Teubner Verlag.

Hedderich, J. und L. Sachs (2016). *Angewandte Statistik. Methodensammlung mit R. 15. Auflage*. Springer Verlag: Berlin.

Hilbe, J. M. (2011). *Negative Binomial Regression. Second Edition*. Cambridge University Press: New York.

Lüdecke, D., M. S. Ben-Shachar, I. Patil, P. Waggoner und D. Makowski (2021). „performance: An R Package for Assessment, Comparison and Testing of Statistical Models“. In: *Journal of Open Source Software* 6(60), S. 3139. DOI: 10.21105/joss.03139.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Riphanh, R., A. Wambach und A. Million (2003). „Inventive effects in the demand for health care: a bivariate panel count data estimation.“ In: *Journal of Applied Econometrics,* **18**, S. 387–405.

Venables, W. N. und B. D. Ripley (2002). *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. Springer: New York. URL: https://www.stats.ox.ac.uk/pub/MASS4/.

Anhang

Tabelle 4: Relative Häufigkeiten der Faktorstufen der *Zufriedenheit mit der Gesundheit*

0	1	2	3	4	5	6	7	8	9	10
0.03	0.01	0.03	0.04	0.04	0.16	0.07	0.13	0.19	0.10	0.19

Tabelle 5: Relative Häufigkeiten der dichotomen Merkmale - gerundet auf drei Nachkommastellen

	0 ($\hat{=}$ liegt nicht vor)	1 ($\hat{=}$ liegt vor)
Geschlecht	0.520	0.480
Behinderung	0.885	0.115
Kinder	0.552	0.448
Heirat	0.212	0.788
Hauptschule	0.317	0.683
Realschule	0.825	0.175
Abitur	0.963	0.037
Fachhochschule	0.968	0.032
Hochschule	0.939	0.061
Beschäftigungsverhältniss	0.366	0.634
Arbeiter	0.743	0.257
Angestellter	0.730	0.270
Selbständiger	0.939	0.061
Beamter	0.926	0.074
Krankenversicherug	0.097	0.903
Zusatzkrankenversicherung	0.996	0.004

Tabelle 6: Transformierte Koeffizientenschätzungen un p-Werte des finalen Modells der *Anzahl an Arztbesuchen* im Teilmodell der weiblichen Personen (gerundet auf fünf Nachkommastellen)

Variable	$\exp(\hat{\beta}_i)$	p-Werte
Intercept	15.20814	$< 2 \cdot 10^{-16}$
Zufriedenheit mit der Gesundheit (Bewertung: 1)	0.65611	0.18270
Zufriedenheit mit der Gesundheit (Bewertung: 2)	0.67868	0.09517
Zufriedenheit mit der Gesundheit (Bewertung: 3)	0.57443	0.01111
Zufriedenheit mit der Gesundheit (Bewertung: 4)	0.37971	$1.70 \cdot 10^{-05}$
Zufriedenheit mit der Gesundheit (Bewertung: 5)	0.33751	$7.02 \cdot 10^{-09}$
Zufriedenheit mit der Gesundheit (Bewertung: 6)	0.30526	$8.00 \cdot 10^{-09}$
Zufriedenheit mit der Gesundheit (Bewertung: 7)	0.21138	$3.85 \cdot 10^{-15}$
Zufriedenheit mit der Gesundheit (Bewertung: 8)	0.19376	$< 2 \cdot 10^{-16}$
Zufriedenheit mit der Gesundheit (Bewertung: 9)	0.12511	$< 2 \cdot 10^{-16}$
Zufriedenheit mit der Gesundheit (Bewertung: 10)	0.10132	$< 2 \cdot 10^{-16}$
Behidnerung (vorhanden)	2.36578	0.00068
Grad der Behinderung	0.99189	0.04155
Kinder unter 16 Jahren (vorhanden)	0.89584	0.08245
Höchster Schulabschluss: Hauptschule	0.88016	0.06192
Selbständige (ja)	0.67699	0.02705

Tabelle 7: Transformierte Koeffizientenschätzungen un p-Werte des finalen Modells der *Anzahl an Arztbesuchen* im Teilmodell der männlichen Personen (gerundet auf fünf Nachkommastellen)

Variable	$\exp(\hat{\beta}_i)$	p-Werte
Intercept	6.65120	$3.38 \cdot 10^{-13}$
Alter	1.00742	0.045953
Zufriedenheit mit der Gesundheit (Bewertung: 1)	1.17195	0.672177
Zufriedenheit mit der Gesundheit (Bewertung: 2)	0.67439	0.171625
Zufriedenheit mit der Gesundheit (Bewertung: 3)	0.54651	0.023274
Zufriedenheit mit der Gesundheit (Bewertung: 4)	0.41179	0.000712
Zufriedenheit mit der Gesundheit (Bewertung: 5)	0.36077	$4.68 \cdot 10^{-06}$
Zufriedenheit mit der Gesundheit (Bewertung: 6)	0.35352	$2.41 \cdot 10^{-05}$
Zufriedenheit mit der Gesundheit (Bewertung: 7)	0.27901	$2.40 \cdot 10^{-08}$
Zufriedenheit mit der Gesundheit (Bewertung: 8)	0.17534	$7.97 \cdot 10^{-15}$
Zufriedenheit mit der Gesundheit (Bewertung: 9)	0.10407	$< 2 \cdot 10^{-16}$
Zufriedenheit mit der Gesundheit (Bewertung: 10)	0.09835	$< 2 \cdot 10^{-16}$
Behinderung (vorhanden)	2.63116	$6.21 \cdot 10^{-05}$
Grad der Behinderung	0.99274	0.062178
Heirat (vorhanden)	0.84637	0.079726
Höchster Schulabschluss: Hauptschule	0.82541	0.061157
Höchster Schulabschluss: Hochschule	0.71489	0.018085
Zusatzkrankenversicherung (vorhanden)	0.18728	0.070899