Technische Universität Dortmund
Fakultät Statistik
Wintersemester 2022/2023

Fallstudien I
Projekt 4

Regressionsmodelle für Zähldaten

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

Caroline Baer

Louisa Poggel
Julia Keiter
Daniel Sipek

Gruppennummer: 1

22.12.2022

Inhaltsverzeichnis

1 Einleitung

Das sozioökonomische Panel (SOEP) ist eine repräsentative Studie zum Gesundheitszustand in Deutschland. Bei dieser werden seit 1984 jährliche Befragungen von zufällig ausgewählten und in Deutschland lebenden Personen gemacht. Die somit erhobenen Daten helfen bei der Beantwortung verschiedener soziologischer, ökonomischer, psychologischer, demographischer, gesundheitswissenschaftlicher und geographischer Fragestellungen. In diesem Bericht werden ein Teildatensatz aus dem Jahr 1984 betrachtet und für Zähldaten geeignete Regressionsmethoden zur Modellierung der Anzahl Arztbesuche in den letzten drei Monaten, sowie der Anzahl Krankenhausaufenthalte im letzten Kalenderjahr, anhand von 22 möglichen erklärenden Variablen angewendet. Nach einer genaueren Vorstellung der Variablen und Ziele der Analyse in Kapitel 2, werden in Kapitel 3 die Methoden zur Poisson-Regression, Overdispersion, Quasi-Poisson-Regression und Regression mit negativer Binomialverteilung erläutert. Anschließend werden in Kapitel 4 die genannten Methoden zur Erstellung und Auswertung von Modellen angewendet, um die beiden Zielvariablen in getrennten Regressionsmodellen möglichst optimal zu modellieren. Zum Schluss erfolgt in Kapitel 5 eine Übersicht der wichtigsten Ergebnisse und ein Ausblick hinsichtlich weiterer möglicher Untersuchungsaspekte.

2 Problemstellung

Der vorliegende Datensatz *Gesundheitszustand.csv* ist ein Ausschnitt des sozioökonomischen Panels (SOEP) in Deutschland, welches auch in Ripanh et al. (2003) thematisiert wird. Beim SOEP wurden jährlich Befragungen mit denselben Personen und Familien durchgeführt, wobei auch neue Personen hinzukamen und andere aus der Studie ausgestiegen sind. Auf Grund der zufälligen Auswahl der Befragten handelt es sich um eine repräsentative Studie, die dabei helfen soll, dass soziologische, ökonomische, psychologische, demographische, gesundheitswissenschaftliche und geographische Fragestellungen beantwortet werden können. Insgesamt enthält das SOEP gesundheitsbezogene Daten zu 28.037 Personen, die in den Jahren 1984-1995 erhoben wurden (vgl. Ripanh et al. (2003)). Mit dem hier vorliegenden Ausschnitt wird eine Analyse zu den 3874 Beobachtungen aus dem Jahr 1984 durchgeführt.
Enthalten sind dabei 25 Variablen zu *Geschlecht*, *Alter*, die Bewertung der *Zufriedenheit* der Gesundheit auf einer Skala von Null bis Zehn, eine Einstufung, ob eine *Behinderung* vorliegt und falls ja, der *Behinderungsgrad* in Prozent. Außerdem vorhanden sind Variablen zum monatlichen *Haushaltsnettoeinkommen* in D-Mark, ob *Kinder* unter 16 Jahren im

Haushalt leben, ob die befragte Person *verheiratet* ist, die Anzahl an *Schuljahren* und fünf Variablen zum höchsten *Schulabschluss* (von Hauptschulabschluss bis Hochschule). Darüberhinaus gibt es Angaben dazu, ob ein *Beschäftigungsverhältnis* vorliegt, die befragte Person als *Arbeiter*, *Angestellter*, *Selbstständiger* oder *Beamter* tätig ist. Ebenfalls erhoben wurde die Anzahl *Arztbesuche* in den letzten drei Monaten, die Anzahl *Krankenhausaufenthalte* im letzten Kalenderjahr, ob *krankenversichert* und ebenfalls *zusatzversichert*, sowie eine *Identifikationsnummer* und das *Jahr*, welches bei allen hier betrachteten 3874 Beobachtungen 1984 ist.
Davon sind die 16 Variablen *Geschlecht*, *Behinderung*, *Kinder*, *verheiratet*, *Beschäftigungsverhältnis*, *Arbeiter*, *Angestellter*, *Selbständiger*, *Beamter*, *krankenversichert*, *zusatzversichert*, sowie die fünf Variablen zum *Schulabschluss* dichotom. Die *Zufriedenheit* liegt ordinal skaliert vor, während das *Alter*, der *Behinderungsgrad*, das *Einkommen* und die Anzahl *Schuljahre* metrisch sind. Die Anzahl *Arztbesuche* und *Krankenhausaufenthalte* sind metrische Zählvariablen.
Es liegen keine fehlenden Werte vor, jedoch wurden 16 Beobachtungen aus dem Datensatz entferntm da sie unplausible, dh. von der vorgegebenen Skala abweichende Werte zur *Zufriedenheit* oder *Behinderung* beinhalteten. Dementsprechend besteht der zu untersuchende Datensatz aus 3858 Beobachtungen.
Zusätzlich wurden 9 Werte bei dem *Behinderungsgrad* auf die nächste 5%-Stufe abgerundet und 8 Werte bei der Anzahl *Schuljahren* auf komplette Angaben in halben Jahren aufgerundet. Bei insgesamt 235 Beobachtungen wurde mehr als ein höchster *Schulabschluss* angegeben, diese wurden auf ihren jeweils höchsten umkodiert.
Ziel ist die Untersuchung der Zusammenhänge zwischen den jeweiligen Zielvariablen *Arztbesuche* und *Krankenhausaufenthalte* und den restlichen Variablen, um Informationen über Einflussfaktoren bezüglich des Bedarfs an medizinischer Versorgung zu erhalten. Zu diesem Zweck werden für die Zielvariablen getrennte Regressionsmodelle mittels für Zähldaten geeigneter Poisson-Regression erstellt. Darüberhinaus findet eine nach den Geschlechtern aufgeteilte Betrachtung der Analysen statt.

3 Statistische Methoden

Die im Folgenden beschriebenen Methoden stammen, sofern nicht anders angegeben, aus Kapitel 4.2 und 4.4 von Fahrmeir et al. (2007). Im Weiteren beschreibt n die Anzahl der Beobachtungen pro Variable, k die Anzahl der Regressoren, y den Regressand und x_{ij} die i-te Beobachtung des j-ten Regressors.

Die statistische Auswertung mit den hier aufgeführten Methoden wird mit der Software

R Core Team (2022) Version 4.2.2 durchgeführt. Zusätzlich wichtige Pakete hierfür sind `MASS` von Venables und Ripley (2002) und `AER` von Kleiber und Zeileis (2008).

3.1 Poisson-Regression

Für die Analyse von Zähldaten mit $y_i \in \mathbb{N}_0$ (für $i = 1, \ldots, n$) eignet sich insbesondere die Poisson-Regression, denn $\forall i$ wird angenommen, dass $y_i|x_i \sim Poi(\lambda_i)$ und die y_i unabhängig sind (vgl. Fahrmeir et al. (2007), Kap. 4.2). Bei der Poisson-Regression handelt es sich um ein generalisiertes lineares Modell, da die Zielvariable nicht alle Werte aus $\mathbb{R}$ annehmen kann. Dementsprechend wird der Erwartungswert $\mathbb{E}(y|x) =: \lambda$ geeignet transformiert, sodass er sich, wie in Formel (1) angegeben, durch den linearen Prädiktor η darstellen lässt.

$$g(\lambda_i) = g(\mathbb{E}(y_i|x_{i1}, \ldots, x_{ik})) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} = x_i^\top \beta =: \eta_i \quad i = 1, \ldots, n \qquad (1)$$

Wie in Kapitel 4.4 von Fahrmeir et al. (2007) beschrieben, bezeichnet die Funktion $g(\lambda) = \eta$ dabei die sogenannte Link-Funktion, die bei der Poisson-Regression der Funktion $log(\lambda)$ entspricht. Mit $h(\eta) = g^{-1}(\eta) = exp(\eta)$ wird die Response-Funktion definiert. Somit ergibt sich $\mathbb{E}(y_i|x_{i1}, \ldots, x_{ik}) = \lambda_i = exp(\eta_i) = exp(\beta_0 + \beta_1 x_{i1} + \ldots \beta_k x_{ik}) = exp(\beta_0) \cdot exp(\beta_1 x_{i1}) \cdot \cdots \cdot exp(\beta_k x_{ik})$.
In `R` kann das Modell zur Poisson-Regression mit `glm(formula, family = "poisson")` umgesetzt werden.

—> **Devianz-Residuen !!! und Devianz**

Der Parametervektor β wird bei der Poisson-Regression mithilfe der Maximum-Likelihood-Schätzung geschätzt. Dabei wird die Likelihood wie in Formel (2) durch die Dichtefunktionen der Zielvariable definiert, um anhand dieser die Parameterwerte mit der höchsten Wahrscheinlichkeit zu bestimmen (siehe **Stahel zitieren**, (S. 26) Kap. 2.3). Zu diesem Zweck wird die Likelihood-Funktion logarithmiert (vgl. Formel (3)), da diese lineare Transformation die Produkte zu Summen vereinfacht, aber die Extremstellen unverändert lässt. Mit der Ableitung in Form der Score-Funktion in Formel (4) können diese dann mit dem

Null-Setzen der Score-Funktion $s(\hat{\beta}) \stackrel{!}{=} 0$ und Auflösen nach $\hat{\beta}$ bestimmt werden.

$$\text{Likelihood:} \qquad L(\beta) := \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \frac{\lambda_i^{y_i} exp(-\lambda_i)}{y_i!} \tag{2}$$

$$\begin{aligned}\text{log-Likelihood:} \qquad l(\beta) &:= log(L(\beta)) = log\left(\prod_{i=1}^{n} f(y_i)\right) = \sum_{i=1}^{n} log(f(y_i)) \\ &= \sum_{i=1}^{n} y_i \cdot log(\lambda_i) - \lambda_i - log(y_i!) \\ &= \sum_{i=1}^{n} y_i (x_i^\top \beta) - exp(x_i^\top \beta) - log(y_i!)\end{aligned} \tag{3}$$

$$\begin{aligned}\text{Score-Funktion:} \qquad s(\beta) &:= \frac{\partial}{\partial \beta} l(\beta) = \frac{\partial}{\partial \beta} log(L(\beta)) = \frac{1}{L(\beta)} \frac{\partial}{\partial \beta} L(\beta) \\ &= \sum_{i=1}^{n} x_i (y_i - \lambda_i) = \sum_{i=1}^{n} x_i (y_i - exp(x_i^\top \beta))\end{aligned} \tag{4}$$

Für den somit ermittelten Maximum-Likelihood-Schätzer, kurz ML-Schätzer, gilt, dass $\hat{\beta}$ asymptotisch normalverteilt ist mit $\hat{\beta} \overset{a}{\sim} \mathcal{N}\left(\beta, F^{-1}(\hat{\beta})\right)$ und $\widehat{\mathbb{C}ov(\hat{\beta})} = F^{-1}(\hat{\beta})$, wobei $F^{-1}(\hat{\beta}) := \sum_{i=1}^{n} x_i x_i^\top \hat{\lambda}_i = \sum_{i=1}^{n} x_i x_i^\top exp(x_i^\top \hat{\beta})$ die Inverse der Fisher-Informationsmatrix ist (vgl. Fahrmeir et al. (2007), Kap. 4.2.2). In der Software `R` wird das, in Kapitel 5.4 von Fahrmeir et al. (2007) beschriebene, Fisher-Scoring-Iterationsverfahren angewendet, um den ML-Schätzer zu approximieren (vgl. Formel (5)). Sobald das Abbruchkriterium $\frac{||\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)}||}{||\hat{\beta}^{(k)}||} \leq \epsilon$ erreicht wird, ist $\hat{\beta}^{(k)}$ der approximierte ML-Schätzer. Beim Aufruf der `summary(glm.object)` wird standardmäßig die Anzahl benötigter Fisher-Scoring Iterationen ausgegeben.

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + F^{-1}(\hat{\beta}^{(k)}) \cdot s(\hat{\beta}^{(k)}) \quad k = 0, 1, 2, \ldots \tag{5}$$

Koeffizienten-Interpretation mit exp()

3.2 Overdispersion

Auf Grund der Annahme von $y_i|x_i \sim Poi(\lambda_i)\ \forall i = 1, \ldots, n$ mit $f(y_i|\lambda_i) = \frac{\lambda_i^{y_i} exp(-\lambda_i)}{y_i!}$ muss $\mathbb{E}(y_i|x_i) = \lambda_i$ und $\mathbb{V}ar(y_i|x_i) = \lambda_i$ gelten. Werden jedoch signifikant größere Varianzen festgestellt, wird dies Overdispersion genannt (siehe Dunn und Smyth (2018), Kap. 10.5.1). Die Overdispersion kann durch eine positive Korrelation zwischen den individuellen Beobachtungen der Zielvariable oder einer Verletzung der Unabhängigkeit resultieren.
Sie führt zu Problemen, da die Standardfehler $\text{se}(\hat{\beta}_j) = \hat{\sigma}(\hat{\beta}_j)$ für $(j = 1, \ldots, k+1)$ zwangsläufig unterschätzt werden und somit auch die Tests signifkantere Ergebnisse liefern, als die Daten rechtfertigen (vgl. Dunn und Smyth (2018), Kap. 10.5.1). Neben dem

Unterschied zwischen Erwartungswert und Varianz gibt auch eine, gegenüber der Anzahl Freiheitsgrade, deutlich erhöhte Residuen-Devianz einen Hinweis auf Overdispersion. Diese beiden Kennzahlen werden in `R` automatisch mit der Funktion `summary(glm.object)` ausgegeben.
Der Dispersionsparameter ϕ, mit dem sich die Varianz durch $\mathbb{V}ar(y_i|\lambda_i) = \phi\lambda_i$ darstellen lässt, kann mit der gemittelten Pearson-Statistik wie in Formel (6) geschätzt werden (vgl. Fahrmeir et al. (2007), Kap. 4.2). Für $\hat{\phi} > 1$ liegt Overdispersion vor.

$$\hat{\phi} = \frac{1}{n-(k+1)}\chi^2 = \frac{1}{n-(k+1)}\sum_{i=1}^{n}\frac{(y_i-\hat{\lambda}_i)^2}{\hat{\lambda}_i} \tag{6}$$

`dispersiontest`
Im Fall der vorliegenden Overdispersion kann entweder ein Quasi-Poison-Modell angepasst werden oder ein Modell, welches auf der negativen Binomialverteilung beruht (siehe Dunn und Smyth (2018), Kap. 10.5).

3.3 Quasi-Poisson-Regression

In `R` wird das Quasi-Poisson-Modell mit `glm(formula, family = "quasipoisson")` erstellt. Es erlaubt, dass sich Erwartungswert und Varianz um den Dispersionparameter ϕ unterscheiden (vgl. Fahrmeir et al. (2007), Kap. 4.5). Zur Schätzung des Parametervektors β kann hier jedoch nicht mehr die Likelihood verwendet werden, sodass auf die Quasi-Likelihood zurückgegriffen wird (vgl. Crawley (2012), Kap. 13.7). Die Quasi-Score-Funktion wird wie in Formel (7) definiert und die Quasi-ML-Schätzer $\hat{\beta}_Q$ als Lösung von $s_Q(\hat{\beta}_Q) \stackrel{!}{=} 0$ bestimmt (vgl. Fahrmeir et al. (2007), Kap. 4.5).

$$\text{Quasi-Score-Funktion:} \qquad s_Q(\beta) := \sum_{i=1}^{n} x_i \frac{d_i}{\phi\lambda_i}(y_i-\lambda_i) \;\text{ mit } d_i := \frac{\partial\, exp(x_i^\top\beta)}{\partial\, x_i^\top\beta} \tag{7}$$

Daraus ergibt sich die Quasi-Fisher-Informationsmatrix mit $F(\beta) = \sum_{i=1}^{n} x_i x_i^\top \frac{d_i}{\phi\lambda_i}$ (siehe Fahrmeir et al. (2007), Kap. 4.5). Des Weiteren verändert ist der Standardfehler, der sich beim Quasi-Poisson-Modell mit $se(\hat{\beta}_j) = \sqrt{\hat{\phi}} \cdot \hat{\sigma}(\hat{\beta}_j)$ darstellen lässt (siehe **Wolf zitieren**, Kap. 2.2). Damit folgt auch die veränderte Teststatistik $t = \frac{\hat{\beta}_j}{\sqrt{\hat{\phi}}\cdot\hat{\sigma}(\hat{\beta}_j)}$, mit der für die Hypothesen $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ für $j = 0, \ldots, k$ getestet wird, ob ein Regressor einen signifikanten Einfluss zum Niveau α hat. Dabei wird die Nullhypothese $H_0 : \beta_j = 0$ abgelehnt, falls $t > t_{1,1-\alpha}$ gilt.

Test
Modellselektion

()

Da bei der Quasi-Poisson-Regression keine echte Likelihood-Funktion vorliegt, ist das Akaike Informationskriterium (AIC) nicht definiert und kann von daher nicht zur Modellwahl verwendet werden (vgl. Dunn und Smyth (2018), Kap. 10.5.3).

3.4 Modell mit negativer Binomialverteilung

Beim Modell mit negativer Binomialverteilung wird angenommen, dass die Zielvariable y_i, wie in Formel (8) dargestellt, die Dichte der negativen Binomialverteilung aufweist, bei der die „Anzahl Misserfolge“ gezählt wird (vgl. **Zeileis zitieren**, Kap. 2.1). Die Wahrscheinlichkeit p für das „Eintreten eines Erfolges im Einzelversuch“ wird dabei mithilfe des Erwartungswerts $\mathbb{E}(y) = \frac{r(1-p)}{p}$ und der Varianz $\mathbb{V}ar(y) = \frac{r(1-p)}{p^2} = \frac{1}{p}\mathbb{E}(y)$ festgelegt und mit dem Quotienten aus arithmetischen Mittel und empirischer Varianz geschätzt.

$$y \sim \text{negBin}(r,p) \;\Rightarrow\; f(y) = \frac{\Gamma(y+r)}{\Gamma(r)y!}p^r(1-p)^y \;\text{ mit } p = \frac{\mathbb{E}(y)}{\mathbb{V}ar(y)} \text{ und } r = \frac{p\cdot\mathbb{E}(y)}{1-p} \tag{8}$$

$$f(y|\theta,\mu) = \frac{\Gamma(\theta+y)}{\Gamma(\theta)y!}\left(\frac{\mu}{\mu+\theta}\right)^y\left(\frac{\theta}{\mu+\theta}\right)^\theta \;\text{ mit }\; \mathbb{E}(y)=\mu,\;\; \mathbb{V}ar(y) = \mu + \frac{\mu^2}{\theta} \tag{9}$$

In `R` kann das Modell mit der Funktion `glm.nb(formula)` aus dem Paket `MASS` erstellt werden. Zur Parameterschätzung kann hier wieder das in Abschnitt 3.1 bereits erläuterte Prinzip der ML-Schätzung, sowie das Fisher-Scoring-Iterationsverfahren angewendet werden.

Bedeutung von theta im R-Output

Um den Einfluss eines Regressors auf den Regressanden zu prüfen, werden die Hypothesen $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ für $j = 0,\ldots,k$ getestet. In der Software `R` wird unter Verwendung der `summary()` dazu automatisch der Wald-Test verwendet. Bei diesem wird die Wald-Statistik $w = \hat{\beta}_j^\top F^{-1}(\hat{\beta})\hat{\beta}_j = \frac{\hat{\beta}_j^2}{a_{jj}}$ aufgestellt, wobei a_{jj} dem j-ten Diagonalelement der Fisher-Informationsmatrix, also der Varianzschätzung von $\hat{\beta}_j$, entspricht (vgl. Fahrmeir et al. (2007), Kap. 4.4). Mit $\sqrt{w} = \frac{\hat{\beta}_j}{\sqrt{a_{jj}}} = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)}$ und $\hat{\beta} \overset{a}{\sim} \mathcal{N}\left(\beta, F^{-1}(\hat{\beta})\right)$ folgt, dass $w \overset{a}{\sim} \chi^2_1$ bzw. $\sqrt{w} \overset{a}{\sim} \mathcal{N}(0,1)$ gilt. Das bedeutet, dass die Nullhypothese $H_0 : \beta_j = 0$ zum Niveau α abgelehnt wird, falls $w > \chi^2_{1,\,1-\alpha}$ gilt. Für die Auswertung in diesem Bericht wird das Signifkanzniveau $\alpha = 0.05$ festgelegt.

Da die Likelihood-Funktion für die negative Binomialverteilung definiert ist, kann mit dem Akaiken Informationskriterium AIC := $2\cdot l(\hat{\beta})+2(k+1)$ und der Funktion `step(glm.object)`

eine schrittweise Regression zur Variablenselektion durchgeführt werden (siehe Dunn und Smyth (2018), Kap. 7.8).

4 Statistische Auswertung

4.1 Deskriptive Zusammenfassung

Die Anzahl der *Arztbesuche* liegt bei den insgesamt 3858 befragten Personen zwischen keinem und 121 Besuchen und im arithmetischen Mittel bei 3.16 Besuchen in den letzten drei Monaten (vgl. Tabelle 8). Abbildung 1 ist zu entnehmen, dass keine oder wenige Arztbesuche weitaus häufiger angegeben wurden. Die Anzahl der *Krankenhausaufenthalte* liegt zwischen keinem und 17 Aufenthalten und im arithmetischen Mittel bei 0.12 Aufenthalten im letzten Kalenderjahr (siehe Tabelle 8). Wie in Abbildung 2 zu erkennen ist, haben über 80% der Befragten null *Krankenhausaufenthalte* angegeben.

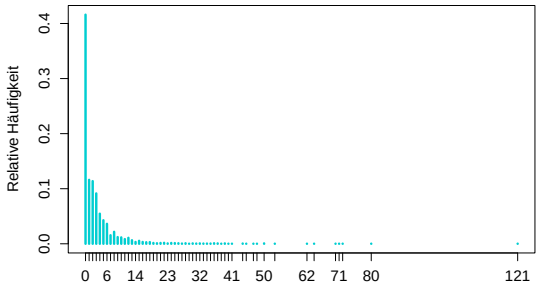

Abbildung 1: Anzahl der *Arztbesuche*

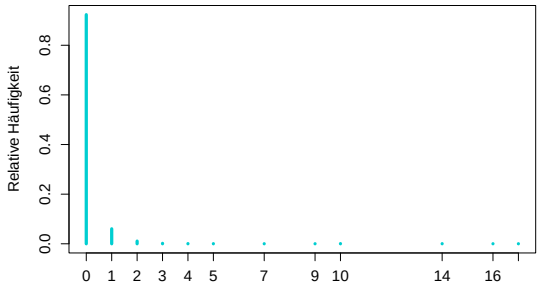

Abbildung 2: Anzahl der *Krankenhausaufenthalte*

Das *Alter* der untersuchten Personen ist im arithmetischen Mittel bei 43.98 Jahren und insgesamt zwischen 25 und 64 Jahren (vgl. Tabelle 8). Das höchste monatliche *Nettohaushaltseinkommen* wurde mit 25000 DM und das niedrigste mit 15 DM genannt. Tabelle 9 beinhaltet die absoluten und relativen Häufigkeitsangaben zu den dichotomen Variablen. So wurden 1850 Frauen und 2008 Männer befragt. Von den 3858 Befragten sind 78.85% *verheiratet* und 44.84% leben mit *Kindern* unter 16 Jahren in einem Haushalt. Der höchste *Bildungsabschluss* ist bei 6.09% ein Universitätsabschluss, bei 3.73% das Abitur, bei 3.21% der Fachhochschulabschluss, bei 17.47% der Realschulabschluss und bei 68.33% der Hauptschulabschluss. Wie in Tabelle 9 dargestellt ist, sind 3485 Personen, also 90.33%, *krankenversichert* und 14 Personen, dh. 0.36%, zusätzlich *zusatzversichert*.

Von den 3858 Befragten gaben 442 eine *Behinderung* an. Zusätzlich erfasst wurden die *Behinderungsgrade* von 0% bis 100%, deren absolute und relative Häufigkeiten Tabelle 10 entnommen werden können. Dabei gibt es *Behinderungsgrade* von mindestens 70% mit einer relativen Häufigkeit von 4.22%.
Tabelle 1 zeigt die Bewertung der *Zufriedenheit* mit der Gesundheit auf einer Skala von Null bis Zehn. Mit 19.34% haben die meisten befragten Personen die *Zufriedenheit* mit ihrer Gesundheit mit einer Acht und die zweitmeisten, mit 19.26%, mit einer Zehn bewertet. Eine Bewertung bis maximal Drei gaben insgesamt 10.55% der Personen an, wohingegen 61.80% mindestens eine Sieben angegeben haben.

Tabelle 1: Bewertung der *Zufriedenheit* (0 = niedrig, ..., 10 = hoch)

	abs. Häufigkeit	rel. Häufigkeit		abs. Häufigkeit	rel. Häufigkeit
0	101	0.0262	6	276	0.0715
1	41	0.0106	7	502	0.1301
2	110	0.0285	8	746	0.1934
3	155	0.0402	9	393	0.1019
4	157	0.0407	10	743	0.1926
5	634	0.1643			

4.2 Modellierung der Zielvariable Arztbesuche

Zuerst betrachtet wird das Modell **glm-arzt-poisson**, welches mittels Poisson-Regression die Zielvariable *Anzahl Arztbesuche* anhand der restlichen 21 Variablen ohne *Krankenhausaufenthalte*, *Jahr* und *Identifikationsnummer* modelliert.
Damit ergeben sich die in Tabelle 2 beschriebenen Devianz-Residuen. Diese streuen zischen 19.7845 und -6.4875 und im Median um -1.1780 herum. Die p-Werte bezüglich des Wald-Tests zeigen an, dass die Variablen *Geschlecht*, *Alter*, *Zufriedenheit*, *Behinderung*, *Behinderungsgrad*, *Einkommen*, *Kinder*, *Schuljahre*, *Fachhochschulabschluss*, *Abitur*, *Universitätsabschluss* und *zusatzversichert* mit p-Werten zwischen 0.0109 und $< 2 \cdot 10^{-16}$ zum Niveau $\alpha = 5\%$ einen signifikanten Einfluss auf den Regressanden haben.
Das AIC liegt bei 25131 und die ML-Schätzung der Koeffizienten $\hat{\beta}$ erfolgt nach 6 Iterationen des Fisher-Scoring-Verfahrens.

Tabelle 2: Verteilung der Devianz-Residuen im Modell **glm-arzt-poisson**

Residuen	Minimum	1. Quartil	Median	3. Quartil	Maximum
	-6.4875	-1.7430	-1.1780	0.4456	19.7845

Der Dispersionsparameter wird bei der Poisson-Regression als $\phi = 1$ angenommen, aber die Devianz ist mit 18023 weitaus größer als die Anzahl Freiheitsgrade mit 3827. Dies ist ein Hinweis auf Overdispersion, sodass als nächstes die Schätzungen von Erwartungswert und Varianz der Zielvariable betrachtet werden.
Der arithmetische Mittel der Zielvariable ist mit 3.16 deutlich verschieden zur empirischen Varianz mit 39.45. Eine Betrachtung des Vergleichs von der Verteilung der *Anzahl Arztbesuche* mit der Dichte der Poisson-Verteilung in Abbildung 3 und mit der Dichte der negativen Binomialverteilung in Abbildung 4, zeigt, dass die Zielvariable durch die negative Binomialverteilung erheblich besser beschrieben wird als durch die Poisson-Verteilung.

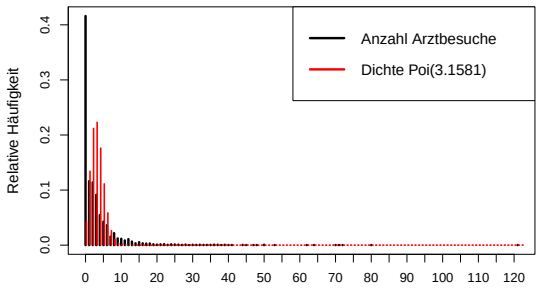

Abbildung 3: Vergleich mit der Dichte der Poisson-Verteilung

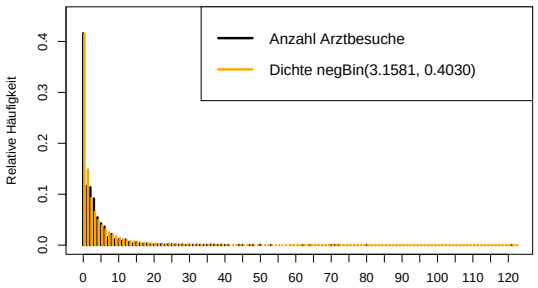

Abbildung 4: Vergleich mit der Dichte der neg. Binomialverteilung

Schätzung des Dispersionsparameters ϕ

Deswegen erfolgt nun die Betrachtung des Modell **glm-arzt-negBin**, welches mittels negativer Binomialverteilungsannahme die Zielvariable *Anzahl Arztbesuche* anhand der restlichen 21 Variablen ohne *Krankenhausaufenthalte*, *Jahr* und *Identifikationsnummer* modelliert.
Hierbei ergeben sich die in Tabelle 3 angegeben Devianz-Residuen. Sie streuen bei diesem Modell weniger als bei dem Modell mit Poisson-Regression und sind im Median mit -0.5007 viel näher um Null herum verteilt.

Tabelle 3: Verteilung der Devianz-Residuen im Modell **glm-arzt-negBin**

Residuen	Minimum	1. Quartil	Median	3. Quartil	Maximum
	-2.1399	-1.1858	-0.5007	0.1968	5.9994

Die p-Werte bezüglich des Wald-Tests weisen nun nur noch die Variablen *Geschlecht*, *Zufriedenheit*, *Behinderung* und *Behinderungsgrad* als zum Niveau $\alpha = 5\%$ signifikante Regressoren aus. Ihre p-Werte liegen zwischen 0.0295 und $< 2 \cdot 10^{-16}$. Das AIC ist mit 15838 im Vergleich zu vorher 25131 deutlich verringert und die Koeffizienten wurden mit

einer einzigen Fisher-Scroing-Iteration bestimmt. Die Devianz weicht mit 3960.6 nun nicht mehr so erheblich von den 3827 Freiheitsgraden ab.

Interpretation von theta und 2 × log-likelihood

Alles in allem hat sich die Modellierung der *Arztbesuche* durch die Verwendung der negativen Binomialverteilung statt der Poisson-Regression verbessert, sodass die folgende Modellselektion auf Grundlage des Modells glm-arzt-negBin stattfindet.

Die schrittweise Regression anhand des AIC's resultiert im Modell **glm-arzt-negBin-final**, welches die Regressoren *Geschlecht*, *Zufriedenheit*, *Behinderung*, *Behinderungsgrad*, *Einkommen*, *Kinder* und *zusatzversichert* beinhaltet. Die Devianz-Residuen von glm-arzt-negBin-final weisen mit 5.7735 ein etwas kleineres Maximum auf als bei glm-arzt-negBin und sind in Bezug auf Median, Quartile und Minimum nur geringfügig schlechter beziehungsweise nahezu gleich (vgl. Tabelle 4).

Tabelle 4: Verteilung der Devianz-Residuen im Modell **glm-arzt-negBin-final**

Residuen	Minimum	1. Quartil	Median	3. Quartil	Maximum
	-2.1438	-1.1919	-0.5181	0.1906	5.7735

Die p-Werte der Wald-Tests für die zum 5%-Niveau nicht-signifikanten Variablen *Nettohaushaltseinkommen* und *zusatzversichert* sind 0.0845 und 0.0929. Die restlichen enthaltenen Variablen weisen einen signifikaten Einfluss auf die *Anzahl Arztbesuche* auf. Die Devianz ist im Vergleich zum Modell glm-arzt-negBin um 0.4 gesunken und das AIC hat bei diesem Modell den leicht verringerten Wert von 15817. Erneut wurde nur eine Iteration zur Koeffizientbestimmung benötigt.

Interpretation von theta und 2 × log-likelihood

4.3 Modellierung der Zielvariable Krankenhausaufenthalte

Zuerst betrachtet wird das Modell **glm-krankenhaus-poisson**, welches mittels Poisson-Regression die Zielvariable *Anzahl Krankenhausaufenthalte* anhand der übrigen 21 Variablen ohne *Arztbesuche*, *Jahr* und *Identifikationsnummer* modelliert.

Hierbei ergibt sich die in Tabelle 5 aufgeführte Verteilung der Devianz-Residuen. Sie streuen zwischen -1.6030 und 10.3721 und im Median um -0.3920 herum. Die p-Werte bezüglich des Wald-Tests zeigen an, dass die Variablen *Zufriedenheit*, *verheiratet*, *Arbeiter*, *Angestellter* und *Beamter* mit p-Werten zwischen 0.0152 und $2 \cdot 10^{-16}$ zum Niveau $\alpha = 5\%$ einen signifikanten Einfluss auf den Regressanden haben. Das AIC hat den Wert 3127.1

und die ML-Schätzung der Koeffizienten $\hat{\beta}$ erfolgt nach 13 Iterationen des Fisher-Scoring-Verfahrens.

Tabelle 5: Verteilung der Devianz-Residuen im Modell **glm-krankenhaus-poisson**

Residuen	Minimum	1. Quartil	Median	3. Quartil	Maximum
	-1.6030	-0.5077	-0.3920	-0.3021	10.3721

Obwohl die Devianz mit 2413 kleiner ist als die 3827 Freiheitsgrade, und somit kein Hinweis auf Overdispersion vorliegt, zeigt die Betrachtung des arithmetischen Mittels und der empirischen Varianz, dass mit 0.1210 im Vergleich zu 0.4855 ein erkennbarer Unterschied vorliegt, der für eine vorliegende Overdispersion spricht.

dispersionstest()

Daher wird nun das mit Quasi-Poisson-Regression erstellte Modell **glm-krankenhaus-quasi** mit dem mit negativer Binomialverteilung erstellten Modell **glm-krankenhaus-negBin** verglichen, um die bestmögliche Modellierung der Zielvariablen zu identifizieren. Die Devianz vom Modell glm-krankenhaus-negBin ist mit 1012.4 weniger als halb so groß wie die von glm-krankenhaus-quasi mit 2413. Auch die Verteilung der Devianz-Residuen (siehe Tabelle 6) deutet darauf hin, dass die Modellierung mit der negativen Binomialverteilung besser geeignet ist. Denn die Streuung hier schmaler ist und die Residuen streuen im Median geringfügig näher um Null herum, wobei die Devianz-Residuen vom Quasi-Poisson-Modell identisch mit denen vom Poisson-Modell sind.

Tabelle 6: Verteilung der Devianz-Residuen in den Modellen **glm-krankenhaus-quasi** und **glm-krankenhaus-negBin**

Residuen	Minimum	1. Quartil	Median	3. Quartil	Maximum
glm-krankenhaus-quasi	-1.6030	-0.5077	-0.3920	-0.3021	10.3721
glm-krankenhaus-negBin	-0.8791	-0.4287	-0.3521	-0.2820	4.9156

Die in den Abbildungen 5 und 6 dargestellten Vergleiche zwischen der echten Datenverteilung und den theoretischen Dichten der Poisson- oder negativen Binomialverteilung zeigen, dass die Poisson-Regression die *Krankenhausaufenthalte* zwar schon recht gut modelliert, aber die negative Binomialverteilung noch besser passt. Dementsprechend wird die Modellselektion auf Grundlage des Modells glm-krankenhaus-negBin durchgeführt.

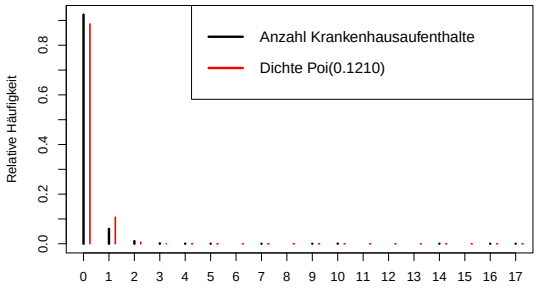

Abbildung 5: Vergleich mit der Dichte der Poisson-Verteilung

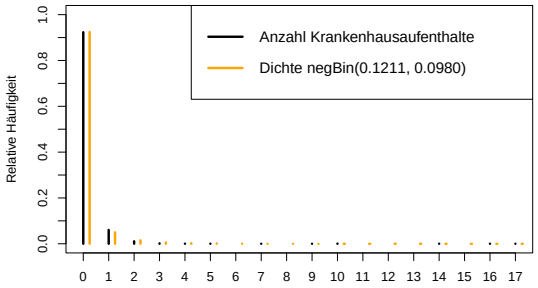

Abbildung 6: Vergleich mit der Dichte der neg. Binomialverteilung

Die schrittweise Regression anhand des AIC's liefert das Modell **glm-krankenhaus-negBin-final**, welches die Regressoren *Zufriedenheit*, *Behinderungsgrad*, *Einkommen*, *Schuljahre*, *verheiratet*, *Realschulabschluss*, *Fachhochschulabschluss*, *Beschäftigungsverhältnis*, *Arbeiter*, *Angestellter* und *Beamter* beinhaltet. Die Devianz-Residuen von glm-krankenhaus-negBin-final weisen im Vergleich zu denen von glm-krankenhaus-negBin nur eine sehr geringfügige Verschlechterung bezüglich ihrer Verteilung auf (vgl. Tabelle 7). Das AIC ist mit 2555.6, ebenso wie die Devianz mit 1011.3, hier jedoch am kleinsten und somit besser als bei den zuvor betrachteten Modellen.

Tabelle 7: Verteilung der Devianz-Residuen im Modell **glm-krankenhaus-negBin-final**

Residuen	Minimum	1. Quartil	Median	3. Quartil	Maximum
	-0.8554	-0.4269	-0.3547	-0.2848	4.9965

Hinsichtlich des Wald-Tests haben die Variablen *Zufriedenheit*, *verheiratet*, *Beschäftigungsverhältnis*, *Arbeiter* und *Beamter* mit p-Werten zwischen 0.0172 und $9.69 \cdot 10^{-11}$ einen zum Niveau 5% signifikanten Einfluss auf den Regressanden. Die restlichen im Modell enthaltenen und nicht-signifikanten Variablen besitzen p-Werte zwischen 0.0634 und 0.5801.
Interpretation theta

4.4 Analyse nach Geschlechtern getrennt

Da das *Geschlecht* weder einen signifikanten Einfluss auf die *Anzahl Krankenhausaufenthalte* hat noch ins endgültige Modell glm-krankenhaus-negBin-final aufgenommen wurde, wird eine genauere Analyse hinsichtlich eines Unterschieds in der Modellierung dieser Zielvariable

nicht durchgeführt. Bei der Modellierung der *Anzahl Arztbesuche* hingegen, wurde die erklärende Variable *Geschlecht* ins endgültige Modell glm-arzt-negBin-final aufgenommen und zusätzlich als signifikanter Regressor identifiziert.
Bei den folgenden Modellen werden alle möglichen erklärenden Variablen außer *Geschlecht*, *Jahr*, *Identifikationsnummer* und *Krankenhausaufenthalte* aufgenommen.

Die Betrachtung des Modells **glm-arzt-female-poisson** mit Poisson-Regression bezogen auf alle Daten der weiblichen Befragten liefert auch hier mit der, gegenüber den 1820 Freiheitsgraden, deutlich erhöhten Devianz von 9347 einen Hinweis auf Overdispersion. Dies bestätigt sich durch den Unterschied zwischen der Schätzung des Erwartungswertes anhand des arithmetischen Mittels mit 3.9168 und der empirischen Varianz mit 44.4614.
dispersionstest()
Der Vergleich mit dem Quasi-Poisson-Modell **glm-arzt-female-quasi** und dem negativen Binomialverteilungs-Modell **glm-arzt-female-negBin** zeigt, dass wie auch im Gesamtdatensatz die negative Binomialverteilung eine bessere Modellierung der *Arztbesuche* darstellt. Die Modellselektion auf Grundlage von schrittweiser Regression und dem AIC liefert das Modell **glm-arzt-female-negBin-final**, welches die zwölf Variablen *Zufriedenheit*, *Behinderung*, *Behinderungsgrad*, *Einkommen*, *Kinder*, *Schuljahre*, *verheiratet*, *Abitur*, *Beschäftigungsverhältnis*, *Arbeiterin*, *Angestellte* und *Beamtin* enthält. Das AIC liegt bei 8530.2 und die Devianz bei 2014.8. Signifikant zum Niveau $\alpha = 5\%$ sind die Regressoren *Zufriedenheit*, *Behinderung*, *Kinder*, *Beschäftigungsverhältnis* und *Angestellte* mit p-Werten zwischen 0.0290 und $2 \cdot 10^{-16}$ bezüglich des Wald-Tests.

Das Modell **glm-arzt-male-poisson**, welches mit der Poisson-Regression zu allen Daten der männlichen Befragten erstellt wird, hat ebenfalls eine Devianz, die mit einem Wert von 8373, deutlich die Anzahl Freiheitsgrade mit 1978 übersteigt. Eine Overdispersion liegt auch hier vor, denn das arithmetische Mittel ist mit 2.4592 erheblich kleiner als die empirische Varianz mit 33.8269.
dispersionstest()
Beim Vergleichen des Modells **glm-arzt-male-quasi** mit Quasi-Poisson-Regression und **glm-arzt-male-negBin** mit negativer Binomialverteilung ergibt sich genauso wie beim Gesamtdatensatz oder dem Teildatensatz der Frauen, dass die Regression mit negativer Binomialverteilung die Daten besser approximiert. Dies zeigt sich erneut daran, dass bei glm-arzt-male-negBin die Devianz mit 1912.2 und das AIC mit 7283.9 am geringsten sind. Mit dem Modell mit negativer Binomialverteilung als Grundlage wird die schrittweise Regression mittels AIC angewendet. Das so entstehende Modell **glm-arzt-male-negBin-final** beinhaltet die sieben Variablen *Zufriedenheit*, *Behinderung*, *Behinderungsgrad*, *Kinder*, *Realschulabschluss*, *Abitur* und *zusatzversichert*. Davon sind die Regressoren *Zufriedenheit* und *Behinderung* signifikant zum Niveau $\alpha = 5\%$. Die restlichen erklärenden Variablen

haben p-Werte zwischen 0.9494 (*Abitur*) und 0.0615 (*Kinder*).

Hinsichtlich der Zielvariable *Anzahl Arztbesuche* erweist sich eine nach *Geschlechtern* getrennte Regression als sinnvoll, da diese Variable nicht nur einen signifikanten Einfluss hat, sondern die Modelle glm-arzt-female-negBin-final und glm-arzt-male-negBin-final unterschiedliche Regressoren beinhalten. In beiden Modellen befinden sich die fünf Variablen *Zufriedenheit*, *Behinderung*, *Behinderungsgrad*, *Kinder* und *Abitur*, aber die sieben Regressoren *Einkommen*, *Schuljahre*, *verheiratet*, *Beschäftigungsverhältnis*, *Arbeiterin*, *Angestellte* und *Beamtin* wurden nur ins Modell der Frauen aufgenommen. Das Regressionsmodell der Männer enthält hingegen auch noch die Variablen *Realschulabschluss* und *zusatzversichert*.

5 Zusammenfassung

Literatur

Crawley, Michael J. (2012). „Analysis of Variance“. In: *The R Book*. 2. Aufl. John Wiley & Sons, Ltd, S. 498–536. ISBN: 9781118448908. DOI: https://doi.org/10.1002/9781118448908.ch11. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118448908.ch11.

Dunn, Peter K. und Gordon K. Smyth (2018). *Generalized Linear Models With Examples in R*. 1. Aufl. Springer New York, NY. ISBN: 978-1-4419-0117-0. DOI: https://doi.org/10.1007/978-1-4419-0118-7.

Fahrmeir, Ludwig, Thomas Kneib und Stefan Lang (2007). *Regression. Modelle, Methoden und Anwendungen*. 1. Aufl. Springer Berlin, Heidelberg. DOI: https://doi.org/10.1007/978-3-540-33933-5.

Kleiber, Christian und Achim Zeileis (2008). *Applied Econometrics with R*. ISBN 978-0-387-77316-2. Springer-Verlag: New York. URL: https://CRAN.R-project.org/package=AER.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Ripanh, R. T., A. Wambach und A. Million (2003). „Inventive effects in the demand for health care: a bivariate panel count data estimation“. In: 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 8: Deskriptive Kenngrößen der metrischen Variablen

	Alter	Einkommen	Bildungsjahre	Arztbesuche	Krankenhaus
arithm. Mittel	43.98	2969.08	11.09	3.16	0.12
Median	44.00	2800.00	10.50	1.00	0.00
emp. Varianz	126.56	2188276.90	4.96	39.45	0.49
Standardabw.	11.25	1479.28	2.23	6.28	0.70
IQR	19.00	1595.22	1.50	4.00	0.00
1.Quartil	35.00	2000.00	10.00	0.00	0.00
3.Quartil	54.00	3595.22	11.50	4.00	0.00
Maximum	64.00	25000.00	18.00	121.00	17.00
Minimum	25.00	15.00	7.00	0.00	0.00
Spannweite	39.00	24985.00	11.00	121.00	17.00

Tabelle 9: Häufigkeiten der dichotomen Variablen (die Bildungsabschlüsse beziehen sich hier auf den höchsten erworbenen Abschluss)

	abs. Häufigkeit		relative Häufigkeit	
Variable	ja	nein	ja	nein
weiblich	1850	2008	0.4795	0.5205
Behinderung	442	3416	0.1146	0.8854
Kinder	1730	2128	0.4484	0.5516
verheiratet	3042	816	0.7885	0.2115
Hauptschulabschluss	2636	1222	0.6833	0.3167
Realschulabschluss	674	3184	0.1747	0.8253
Fachhochschulabschluss	124	3734	0.0321	0.9679
Abitur	144	3714	0.0373	0.9627
Universitätsabschluss	235	3623	0.0609	0.9391
Beschäftigungsverhältnis	2446	1412	0.6340	0.3660
Arbeiter	992	2866	0.2571	0.7429
Angestellter	1040	2818	0.2696	0.7304
selbstständig	237	3621	0.0614	0.9386
Beamter	284	3574	0.0736	0.9264
krankenversichert	3485	373	0.9033	0.0967
zusatzversichert	14	3844	0.0036	0.9964

Tabelle 10: Häufigkeiten der Behinderungsgrade (0% bis 100%)

	absolut	relativ
0%	3416	0.8854
5%	9	0.0023
10%	5	0.0013
15%	2	0.0005
20%	19	0.0049
25%	5	0.0013
30%	43	0.0111

	absolut	relativ
35%	2	0.0005
40%	33	0.0086
45%	0	0.0000
50%	104	0.0270
55%	2	0.0005
60%	55	0.0143
65%	0	0.0000

	absolut	relativ
70%	51	0.0132
75%	0	0.0000
80%	55	0.0143
85%	0	0.0000
90%	19	0.0049
95%	0	0.0000
100%	38	0.0098