Technische Universität Dortmund
Fakultät Statistik
Wintersemester 2022/23

Fallstudien I: Projekt 4

Regressionsmodelle für Zähldaten

Dozenten:
Prof. Dr. Guido Knapp
Yassine Talleb, M. Sc.

Verfasserin:
Julia Keiter

Gruppe 1:
Caroline Baer
Julia Keiter
Louisa Poggel
Daniel Sipek

22.12.2022

Inhaltsverzeichnis

1 Einleitung

Um allgemeingültige Aussagen über politische und gesellschaftliche Veränderungen in Deutschland zu treffen, wird seit 1984 eine jährliche Befragung von Privathaushalten von sozioökonomische Panel (SOEP) durchgeführt. Das SOEP ist die größte und am längsten laufende multidisziplinäre Langzeitstudie für angewandte Sozialforschung in Deutschland (Deutsches Institut für Wirtschaftsforschung e.V., 2022). Die Datensammlung des SOEPs hilft, soziologische, ökonomische, psychologische, demographische, gesundheitswissenschaftliche und geographische Fragestellungen zu beantworten. Besonders ist, dass jedes Jahr dieselben Personen und Familien im Rahmen des SOEPs befragt werden und dass Kinder, die in den befragten Haushalten leben, ab dem 16. Lebensjahr als Befragte nachrücken. Die befragten Personen und Haushalte wurden zufällig ausgewählt, so dass sie die in Deutschland lebenden Menschen repräsentieren.

In diesem Projekt wird die SOEP Datensammlung aus dem Jahr 1984 betrachtet, um anhand der Daten, Informationen über den Bedarf an medizinischer Versorgung der damaligen Bevölkerung zu erhalten. Dazu werden die Zusammenhänge zwischen 21 unabhängigen Variablen und der abhängigen Variable *Arztbesuche in den letzten drei Monaten* bzw. der abhängigen Variable *Krankenhausaufenthalte im letzten Kalenderjahr* und mittels Regressionsverfahren für Zähldaten untersucht. Nach einer ausführlichen Vorstellung des Datensatzes und der Zielsetzung in Kapitel 2 wird in Kapitel 3 erläutert, warum für die Auswertung von Zähldaten alternative Regressonsmodelle verwendet werden wie diese im Vergleich zum allgemeinen linearen Modell aufgebaut sind. Mithilfe der statistischen Methoden erfolgt in Kapitel 4 die statistische Auswertung des Datensatzes in Hinblick auf die oben genannten Zusammenhänge. In Kapitel (5) wird die statistische Auswertung zusammengefasst und diskutiert.

2 Problemstellung

2.1 Datenmaterial

Der zur Verfügung gestellte Datensatz *Gesundheitszustand.csv* enthält Daten aud den Jahren 1984 bis 1994 des SOEP zum Gesundheitszustand in Deutschland. In diesem Projekt wird ausschließlich das Jahr 1984 betrachtet. Zu $n = 3874$ Beobachtungen sind die 23 interessierenden Variablen in Tabelle 1 mit dem jeweiligen Skalenniveau und möglichen Merkmalsausprägungen angegeben.

Die beiden Zielvariablen *Anzahl der Arztbesuche in den letzten drei Monaten*, im Folgenden abgekürzt mit *Arztbesuche*, und *Krankenhausaufenthalte im letzten Kalenderjahr*, im Folgenden abgekürzt mit *Krankenhausaufenthalte*, stehen in Abhängigkeit zu 21 unabhängigen Variablen. Die metrischen unabhängigen Variablen *Alter*, *Nettoeinkommen*, *Bildungsjahre* und *Behinderungsgrad* wurden diskret erhoben. Die Variable *Bildungsjahre* gibt an, wie viele Jahre die befragte Person in einer Bildungseinrichtung besucht hat. Für die nominale Variable *Zufriedenheit* sollten die Befragten auf einer Skala von Null bis

Zehn angeben, wie zufrieden sie mit ihrer Gesundheit sind. Dabei steht der Wert Null für eine niedrige und der Wert Zehn für eine hohe Zufriedenheit.

Tabelle 1: Erhobene interessierende Variablen mit Messniveau und Ausprägung

Variable	Skalenniveau	Ausprägungen
Arztbesuche	metrisch, diskret	$\{0, ..., 121\}$
Krankenhausaufenthalte	metrisch, diskret	$\{0, ..., 17\}$
Alter (in Jahren)	metrisch, diskret	$\{25, ..., 64\}$
Nettoeinkommen (pro Monat, in DM)	metrisch, diskret	$\{15, ..., 25000\}$
Bildungsjahre	metrisch, diskret	$\{7, ..., 18\}$
Behinderungsgrad (in Prozent)	metrisch, diskret	$\{0, ..., 100\}$
Zufriedenheit	nominal	$\{0, ..., 10\}$
Weiblich	nominal, dichotom	$\{0, 1\}$
Behinderung	nominal, dichotom	$\{0, 1\}$
Kinder	nominal, dichotom	$\{0, 1\}$
Verheiratet	nominal, dichotom	$\{0, 1\}$
Hauptschulabschluss	nominal, dichotom	$\{0, 1\}$
Realschulabschluss	nominal, dichotom	$\{0, 1\}$
Abitur	nominal, dichotom	$\{0, 1\}$
Fachhochschulabschluss	nominal, dichotom	$\{0, 1\}$
Hochschulabschluss	nominal, dichotom	$\{0, 1\}$
Beschäftigungsverhältnis	nominal, dichotom	$\{0, 1\}$
Arbeitend	nominal, dichotom	$\{0, 1\}$
Angestellt	nominal, dichotom	$\{0, 1\}$
Selbstständig	nominal, dichotom	$\{0, 1\}$
Verbeamtet	nominal, dichotom	$\{0, 1\}$
Krankenversichert	nominal, dichotom	$\{0, 1\}$
Zusatzversichert	nominal, dichotom	$\{0, 1\}$

Im Datensatz gibt es 16 nominale, dichotome Variablen, die wegen ihrer binären Ausprägung Dummy-Variablen sind. Die Ausprägung Null gibt jeweils an, dass die Variable nicht zutrifft und wird daher im Folgenden als "nein"bezeichnet. Die Ausprägung Eins gibt jeweils an, dass die Variable zutrifft und wird daher im Folgenden als "ja"bezeichnet. Ist die Ausprägung der Variable *Kinder* "ja ", so heißt dies, dass *Kinder* unter 16 Jahren im selben Haushalt wie die befragte Person wohnen. Ist die Ausprägung der Variablen *Hauptschulabschluss, Realschulabschluss, Abitur, Fachhochschulabschluss* oder *Hochschulabschluss* "ja"bedeutet dies, dass die jeweilige Variable der höchste Bildungsabschluss der befragten Person ist und die *Bildungsjahre* bis zu diesem Zeitpunkt gemessen wurden. Ist die Ausprägung der Variable *Behinderung* "ja", liegt bei der befragten Person eine Behinderung vor. Für diese Personen gibt die metrisch, diskrete Variable *Behinderungsgrad* in Prozent den Grad der Behinderung in Fünferschritten an. Liegt keine *Behinderung* vor, so ist der *Behinderungsgrad* gleich Null.

Der Datensatz enthält keine fehlenden Einträge. Jedoch liegen in den Variablen *Zufriedenheit*, *Bildungsjahre*, *Behinderungsgrad* und *Behinderung* und in den Dummy-Variablen zum höchsten Bildungsabschluss unplausible Werte vor. Vier Beobachtungen des Daten-

satzes haben eine *Zufriedenheit* von "6.877203 ". Die *Zufriedenheits-* angaben dieser Beobachtungen werden an die Skala von Null bis Zehn insofern angeglichen, dass sie auf "7 "gerundet werden. 21 Beobachtungen enthalten in den Variablen *Behinderung* und *Behinderungsgrad* Messfehler. Diese Beobachtungen werden insofern angeglichen und neu codiert, indem *Behinderungs*-ausprägungen, die größer als Null sind, als Eins codiert werden. *Behinderungsgrade*, die größer als Null und kleiner als Fünf sind, werden als Fünf codiert, *Behinderungsgrade*, die größer als Fünf und kleiner als Zehn sind, werden als Zehn codiert. 235 Beobachtungen gaben mehr als eine Dummy-Variable als höchstens Bildungsabschluss an. Beispielsweise gaben 221 Personen sowohl das *Abitur* als auch einen *Hochschulabschluss* als höchsten Bildungsabschluss an. Da die Frage nach dem höchsten Bildungsabschluss gestellt war, werden diese 235 Messfehler insofern neu codiert, dass der jeweils niedrigere Bildungsabschluss als „nein" codiert wird. Neun Beobachtungen geben eine Anzahl an *Bildungsjahren* an, die mehr als eine Nachkommastelle hat. Da davon ausgegangen werden muss, dass diese *Bildungsjahren* den Bildungshalbjahresabschluss oder -ganzjahresabschluss nicht erreicht haben, werden sie der Skalierung angeglichen, indem sie entsprechend abgerundet werden. Insgesamt ist die Qualität des Datensatzes als mäßig gut zu beschreiben. Es ist durch sachlogische Umcodierung möglich, für eine maximale Auswertung der Daten alle 3874 Beobachtungen zu erhalten. Jedoch ist die Umcodierung der Daten durch subjektivem Verständnis der Sachlage begründet, die immer auch fehlerhaft sein kann.

2.2 Ziele des Projekts

Ziel dieses Projektes ist es, zwei getrennte Regressionsmodelle für die jeweils eine der beiden Zielvariablen *Arztbesuche* und *Krankenhausaufenthalte* adäquat zu erstellen. Da es sich bei den Merkmalsausprägungen der Zielvariablen um Zähldaten handelt, kommt für die Regressionsmodellbildung kein allgemeines lineares Modell in Frage. Stattdessen werden spezielle Regressionsmethoden für Zähldaten in Betracht gezogen. Diese werden in Kapitel 3 ausführlich vorgestellt und zur Modellbildung in der statistischen Auswertung in Kapitel 4 genutzt. Bei der statistischen Auswertung wird zudem geprüft, ob es einen signifikanten Unterschied für die Anzahl der *Arztbesuche* oder *Krankenhausaufenthalte* macht, ob die befragte Person *weiblich* ist. Falls dem so sein sollte, werden nach Geschlecht getrennte Analysen durchgeführt. Abschließend werden die Rückschlüsse, die aus den gebildeten Regressionsmodelle gezogen werden können in Kapitel 5 zusammengefasst und die Ergebnisse diskutiert.

3 Statistische Methoden

Stellen Variablen die Häufigkeit bestimmter Ereignissen in einem festgelegten Zeitintervall dar, so werden sie **Zähldaten** genannt. Für die Beschreibung der Anzahl der Fälle, in denen das jeweilige Ereignis eintritt, wird die **Poissonverteilung** verwendet. Ist eine Variable Y poissonverteilt mit Parameter λ, so ist laut Fahrmeir et al., 2007 die diskrete

Dichte dieser Variable gegeben durch

$$f(y|\lambda) = P(Y = y) = \frac{\lambda^y \exp(-\lambda)}{y!}, \quad \text{für } y = 0, 1, 2, ..., \ \lambda > 0 \tag{1}$$

Der Parameter λ der Poissonverteilung stellt sowohl den Erwartungswert μ als auch die Varianz σ^2 dar. Neben der Annahme unabhängiger Zielvariablen y_i stellt die Annahme $\mu = \sigma^2 = \lambda$, die wichtigste Modellannahme für eine **Regression mit der Poisson-Methode** dar. Die Poisson-Regression ist eine spezielle Variante der multiplen linearen Regression für generalisierte lineare Modelle. Falls $y_i|x_i \sim Poi(\lambda_i) \ \forall \ i$ gilt, so wird die Modellgleichung (2) durch eine geeignete Transformierung des Erwartungswertes der Zielgröße $\mathbb{E}(y|x) = \lambda$ dargestellt, indem die lineare Funktion η exponiert wird (Fahrmeir et al., 2007, S. 210).

$$\mathbb{E}(y_i|x_{i1}, ..., x_{ik}) = \lambda_i = \exp(\eta_i) = \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}), \quad \text{für } i = 1, \ldots, n \tag{2}$$

mit

$y_i \ \hat{=}$ Zielvariable

$\eta_i \ \hat{=}$ Linearer Prädiktor

$g(\lambda) \ \hat{=}$ Linkfunktion: $g(\lambda) = \log(\lambda) = \eta = x^T\beta$

$h(\eta) \ \hat{=}$ Responsefunktion: $h(\eta) = \lambda = \exp(\eta)$, d.h. $g(\lambda) = h^{-1}$

Die Linkfunktion $g(\lambda)$ verknüpft den Erwartungswert $\mathbb{E}(y_i|x_{i1}, ..., x_{ik})$ mit dem linearen Prädiktor η_i. Da bei der Poisson-Regression $g(\lambda)$ den Parameter λ logarithmiert, wird sie Log-Link-Funktion genannt. Die Responsefunktion $h(\eta)$ ist die Umkehrfunktion von der Log-Linkfunktion, hier also $h(\eta) = \exp(\log(\lambda)) = \lambda$ (Fahrmeir et al., 2007, S. 217). Um den Regressionskoeffizientenvektor $\beta = (\beta_0, ..., \beta_k)^T$ zu bestimmen, wird die Dichtefunktion der poissonverteilten Zufallsvariable (Formel 1) genutzt, um die Log-Likelihood-Funktion einer Poisson-Verteilung aufzustellen. In dieser Log-Likelihood-Funktion findet sich $log(y_i)$ als die Log-Link Funktion $g(\lambda)$ und λ als die Responsefunktion $h(\eta)$ wieder, sodass die Log-Likelihood-Funktion der Poisson-Verteilung in Abhängigkeit des Regressionskoeffizientenvektors β wie folgt geschrieben werden kann

$$l(\beta) = \sum_{i=1}^{n} y_i \cdot \log(\lambda_i) - \lambda_i - \log(y_i!) = \sum_{i=1}^{n} y_i \cdot x_i^T\beta - \exp(x_i^T\beta) - \log(y_i!) \tag{3}$$

Durch erste Ableitung dieser Funktion erhält man die Score-Funktion $s(\beta)$ als Vektor der partiellen ersten Abteilungen der Log-Likelihood-Funktion. Eine zweite Ableitung ergibt die Fisher Informationsmatrix $F(\beta)$, die ein lokales Maß für die Information ist, welche die Log-Likelihood-Funktion über die Parameter des Modells liefert (Fahrmeir et al., 2007, S. 199 f.). Die computergestützte Schätzung von $\hat{beta}$ erfolgt in der Poisson-Regression durch das **Fisher-Scoring-Verfahren** verwendet. Das Fisher-Scoring-Verfahren ist ein Iterationsverfahren (Formel 4) mit gegebenem Startwert $\beta^(0)$, der beispielsweise der KQ-Schätzer sein kann. Das Fisher-Scoring-Verfahren bricht ab, sobald das Kriterium Ξ eintritt, wobei ϵ als ein „cut-off-Wert“ vordefiniert ist (Fahrmeir et al., 2007, S. 223).

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + F^{-1}(\hat{\beta}^{(k)}) \cdot s(\hat{\beta}^{(k)}), \;\; k = 0, 1, 2..., \qquad \Xi = \frac{|| \, \hat{\beta}^{(k+1)} - \hat{\beta}^{(k)} \, ||}{|| \, \hat{\beta}^{(k)} \, ||} \leq \epsilon \qquad (4)$$

In `R` erfolgt die Modellbildung eines generalisierten linearen Modells mit der `glm` Funktion, in die die gleichen Argumente eingegeben werden wie in die `lm` Funktion für allgemeine lineare Modelle. Um die Regression nach Poisson zu spezifizieren wird die Option `family= "poisson"` der `glm` Funktion genutzt. Die Ausgabe des Fisher-Scoring-Verfahrens mit der `summary` Funktion, in die das mit `glm` erzeugte Objekt eingesetzt wird listet die einzelnen Koeffizientenschätzungen $\hat{\beta}_j$ auf und zeigt ebenfalls die Anzahl der Fisher-Scoring Iterationen an. Außerdem wird in der `summary` Ausgabe eines `glm`-Objekts zu jedem Regressor ein z-Wert aufgelistet. Dieser z-Wert ist der Wert der Teststatistik des **Wald-Test**

$$\frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)} \qquad (5)$$

In dieser Teststatistik steht $\hat{\sigma}(\hat{\beta}_j)$ für den geschätzten Standardfehler von $\hat{\beta}_j$ aus der Maximum-Likelihood-Theorie (Groß, 2010, S.227). Diese Testverfahren erkennen an, dass die durch das Fisher-Scoring-Verfahren erhaltenen Maximum-Likelihood-Schätzer $\hat{\beta}_j$ „approximativ“, d.h. für große Stichproben, normalverteilt sind (Formel 6).

$$\hat{\beta} \overset{a}{\sim} \mathcal{N}(\beta, F^{-1}(\hat{\beta})) \qquad \text{mit} \quad \widehat{\mathbb{C}ov(\hat{\beta})} = F^{-1}(\hat{\beta}) \qquad (6)$$

Der Wald-Test testet zum Signifikanzniveau α die speziellen Hypothesen $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ (Fahrmeir et al., 2007, S.250). Die H_0-Hypothese für den Regressionskoeffizienten β_j kann abgelehnt werden, wenn der in Formel 7 definierte Fall eintritt, wobei $q_{1-\frac{\alpha}{2}}$ das $1 - \frac{\alpha}{2}$ Quantil der Standardnormalverteilung darstellt.

$$z > q_{1-\frac{\alpha}{2}} \qquad (7)$$

Da die Regressionsmethode auf dem Maximum-Likelihood-Ansatz beruht, kann zur Modellauswahl und Variablenselektion eine **Rückwärtselimination** mit dem AIC-Kriterium verwendet werden (Cameron und Trivedi, 2013, S. 233). Bei dem Rückwärtseliminationsverfahren wird zunächst ein generalisiertes lineares Modell mit allen in Frage kommenden Regressoren aufgestellt. Die mit der `R` Funktion `step` ausgegebenen AIC-Werte der einzelnen Regressoren werden der Größe nach geordnet. Der Regressor mit dem kleinsten AIC-Wert wird aus dem Modell entfernt, da die Entfernung dieses Regressors den geringsten Informationsverlust für das Modell bedeutet. Im nächsten *Schritt* wird das generalisierte lineare Modell mit allen Regressoren außer dem im ersten Schritt entfernten Regressor gebildet, woraufhin wieder die AIC Werte nach ihrer Größe angeordnet werden und der Regressor mit dem kleinsten AIC-Wert entfernt wird. Dieses Verfahren wird so lange fortgesetzt bis eine Herausnahme eines weiteren Regressors zu keiner Verbesserung, d.h. Verkleinerung des Modell-AIC Wertes führt (Hedderich und Sachs, 2015, S.841). Das Variablenselektionsverfahren mit AIC Kriterium entfernt außerdem eventuell vorhandene

Multikolinearität zwischen den Variablen(Tripathi, 2019).
Die oben erwähnte wichtigste Modellannahme in der Poisson-Regression $\mu = \sigma^2 = \lambda$ wird oftmals nicht eingehalten. Liegt eine im Vergleich zum Erwartungswert μ signifikant größere Varianz σ^2 der poissonverteilten Variable Y vor, spricht man von **Overdispersion**. Gründe für Overdispersion können in einer positiven Korrelation zwischen den unabhängigen Beobachtungen und der Zielvariable oder in einer Verletzung der Unabhängigkeitsannahme der Zielvariable begründet sein(Toutenburg, 2013, S. 380 ff.). Liegt Overdispersion vor, ist die Varianz wie in Formel 8 definiert (Fahrmeir und Tutz, 1994, S. 35, wobei ϕ der Dispersionsparameter ist, der wie in Formel 9 angegeben berechnet wird (Fahrmeir et al., 2007, S. 213).

$$\mathbb{V}ar(y_i|x_i) = \phi \cdot \lambda_i \qquad (8) \qquad \hat{\phi}_P = \frac{1}{n-k-1} \sum_{i=1}^{n} \frac{(y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i / n} \qquad (9)$$

Durch diese Veränderung in der Varianz ergibt sich eine transformierte Kovarianzstruktur der Koeffizientenschätzer $\hat{\beta}_j$

$$\widehat{\mathbb{C}ov}(\hat{\beta}) = \hat{\phi}_P \cdot F^{-1}(\hat{\beta}) \qquad (10)$$

sodass die Koeffizientenschätzungen nicht mehr auf einer *echten* Maximum-Likelihood-Theorie sondern auf einer **Quasi-Likelihood-Theorie** beruhen (Fahrmeir et al., 2007, S. 214). Dieser Veränderung wird die Option `family=quasipoisson` in der Modellbildung mit `glm` gerecht. Die `summary`-Ausgabe dieser Quasi-Poisson-Regressionsmethode listet analog zur Poisson-Methode die Koeffizientenschätzungen $\hat{\beta}_j$ der einzelnen Regressoren auf. Hier wird statt des Wald-Tests ein t-Test mit (n-k-1)-Freiheitsgraden zum Testen der speziellen Hypothesen verwendet. Wird die Overdispersion nicht im Regressionsverfahren berücksichtigt, so werden die Standardfehler unterschätzt, sodass die Teststatistiken des Wald-Test falsche Ergebnisse liefern könnten (Toutenburg, 2013, S. 381). Während bei der Wahl der Quasi-Poisson-Verteilung die Varianzwerte σ_i^2 im Vergleich zum Erwartungswerte $\mu_i = \lambda_i$ durch den Dispersionsparameter ϕ linear ansteigen (siehe Abbildung 5, Anhang) ist der Anstieg der Varianzwerte in anderen Fällen noch stärker und nichtlinear. Der Zusammenhang von μ und σ^2 in der **Negativ-Binomialverteilung** wird grafisch durch eine Parabel beschrieben. Dies ist darin begründet, dass die Schätzung der Varianz in der Negativ-Binomialverteilung der Parameter λ quadriert wird und so ein quadratischer Zusammenhang zwischen μ und σ^2 entsteht (Hilbe, 2011, S. 187). Ist die Zielvariable Y negativ-binomialverteilt, d.h. $Y \sim nBin(k, p)$ mit $k = 1/\psi$ und $p = 1/(1 + \psi\lambda)$, so ergibt sich die Dichtefunktion für diese Zielvariable wie in Formel 11 angegeben, die von den Parametern λ und ψ abhängt (Hilbe, 2011, S. 189). ψ ist als die Inverse des Parameters θ der Gammafunktion definiert.

$$f(y_i, \lambda_i, \psi) = \binom{y_i + (1/\psi) - 1}{(1/\psi) - 1} \Big(\frac{1}{1+\psi\lambda_i}\Big)^{1/\psi} \Big(\frac{\psi\lambda_i}{1+\psi\lambda_i}\Big)^{1/\psi} \qquad (11)$$

Die Varianz einer negativ-binomialverteilten Zielvariable Y ergibt sich als

$$\mathbb{V}ar(Y_i) = \sigma_i^2 = \frac{1}{\psi}\frac{\psi\lambda_i}{1+\psi\lambda_i}(1+\psi\lambda_i)^2 = \lambda_i(1+\psi\lambda_i) = \lambda_i + \psi\lambda_i^2 \tag{12}$$

Im dritten Term von Formel 12 wird der quadratische Zusammenhang zwischen σ_i^2 und $\mu = \lambda$ ersichtlich. Die Schätzung der Koeffizienten erfolgt analog zum Schätzungsverfahren in der Poissonregression mit Fisher-Scoring Verfahren (Hilbe, 2011, S. 211). In `R` wird die Regression einer negativ-binomialverteilten Zielvariable mit der Funktion `glm.nb` aus dem Paket `MASS` (Venables und Ripley, 2002) ausgeführt. Da die Koeffizientenschätzung also wie in der Poisson-Regression auf der Maximum-Likelihood-Methode beruht, kann für die Variablenselektion eines mit Negativ-Binomialverteilungs-Regression gebildeten Modells eine Rückwärtselimination mit dem AIC Kriterium verwendet werden.
Zur Bewertung der Modellgüte wird bei Regressionsmodellen für Zähldaten die **Devianz** verwendet. Die Devianz ist als Summe der Abweichungsquadrate definiert (Formel 13).

$$\mathrm{D} = 2\sum_{i=1}^{n} y_i \cdot log\left(\frac{y_i}{exp\left(x_i^\top \hat{\beta}\right)}\right) - \left(y_i - exp\left(x_i^\top \hat{\beta}\right)\right) \tag{13}$$

Wenn das Modell gut an die Daten angepasst ist, liegen die beobachteten Werte y_i nah an ihren vorhergesagten Mittelwerten $\hat{\mu}_i = exp(x_i^\top \hat{\beta})$, sodass beide Terme in D klein sind und somit auch die Abweichung klein ist. In der Modellbildung eines generalisierten linearen Modells entspricht die Devianz der Summe der Abweichungsquadrate bei linearen Regressionsmodellen (vgl. Hedderich und Sachs, 2015, S. 834). Demnach ist die Güte eines Modells umso höher einzuschätzen, je kleiner die Devianz ist. In der `R` Funktion `summary(glm.objekt)` ist die Devianz des gebildeten Modells mit `Residual deviance` angegeben. Die `Null deviance` bezeichnet die Devianz für das Nullmodell, welches nur β_0 also keine erklärenden Variablen enthält.

4 Statistische Auswertung

4.1 Deskriptive Auswertung

Um einen ersten Überblick über die Verteilungen der interessierenden Variablen des Datensatzes zu erhalten, sind in Tabelle 2 univariate Kenngrößen für die metrischen und nominalen Variablen angegeben. Während die Verteilungen der Variablen *Alter*, *Nettoeinkommen* und *Bildungsjahre* aufgrund der Gleich- oder Ähnlichkeit der Lagemaße arithmetisches Mittel und Median bzw. Streuungsmaße Standardabweichung und MAD als ausgeglichen verteilt angenommen können, unterscheiden sich die Kenngrößen für die Variablen *Arztbesuche*, *Krankenhausaufenthalte* und *Behinderungsgrad* stark. Besonders von Interesse ist, ob die heterogenen Verteilungen der Zielvariablen *Arztbesuche* und *Krankenhausaufenthalte* anders ausfallen würden, wenn man eine diese Variablen geschlechtergetrennt betrachten würde.

Tabelle 2: univariate Kenngrößen für metrische Variablen

	Arzt-besuche	Krankenhaus-aufenthalte	Alter	Nettoein-kommen	Grad der Behinderung	Bildungs-jahre
arithm. Mittel	3.16	0.12	44.00	2968.79	6.60	11.09
Median	1.00	0.00	44.00	2800.00	0.00	10.50
Minimum	0.00	0.00	25.00	15.00	0.00	7.00
Maximum	121.00	17.00	64.00	25000.00	100.00	18.00
Spannweite	121.00	17.00	39.00	24985.00	100.00	11.00
1.Quartil	0.00	0.00	35.00	2000.00	0.00	10.00
3.Quartil	4.00	0.00	54.00	3590.00	0.00	11.50
IQR	4.00	0.00	19.00	1590.00	0.00	1.50
Standardabw.	6.28	0.70	11.24	1477.31	19.93	2.22
MAD	1.48	0.00	14.83	1186.08	0.00	1.48

Dazu ist eine nach den Geschlechtern getrennte grafische Darstellung der *Krankenhausaufenthalte* ist in Abbildung 1 in Form von Histogrammen gegeben. Die *Krankenhausaufenthalte* unterscheiden sich in ihrer Verteilung nicht stark zwischen den Geschlechtern. Der Großteil der Beobachtungen gaben sowohl in der Subgruppe der weiblichen als auch in der Subgruppe der männlichen Befragten eine Anzahl an *Krankenhausaufenthalten* zwischen Null und Fünf. Eine geschlechtergetrennte Analyse der *Krankenhausaufenthalte* ist also nicht gerechtfertigt.

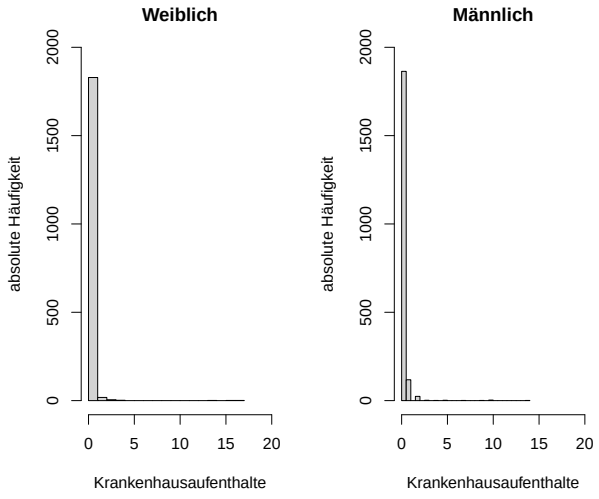

Abbildung 1: Histogramme zu *Krankenhausaufenthalte* getrennt nach Geschlecht

Da sich die Standardabweichung der *Arztbesuche* besonders stark von dem auf extreme Werte weniger empfindlicheren Streuungsmaß MAD unterscheidet wird der Einfluss eines oder mehrerer Ausreißer vermutet. Die grafische Darstellung in Abbildung 4 (siehe Anhang) bestätigt diese Annahme, sodass die Beobachtung mit Anzahl *Arztbesuche*=121 entfernt wird. Um zu entscheiden, ob Unterschiede in der Anzahl der *Arztbesuche* zu erwarten sind, je nachdem ob die befragte Person *weiblich* ist, sind in Abbildung 2 nach Geschlechtern getrennte Boxplots dargestellt. Es ist ersichtlich, dass die Verteilungen der *Arztbesuche* zwischen den Geschlechtern stark unterschiedlich ausfallen. Eine geschlechtergetrennte Regressionsanalyse der Variable *Arztbesuche* ist gerechtfertigt.

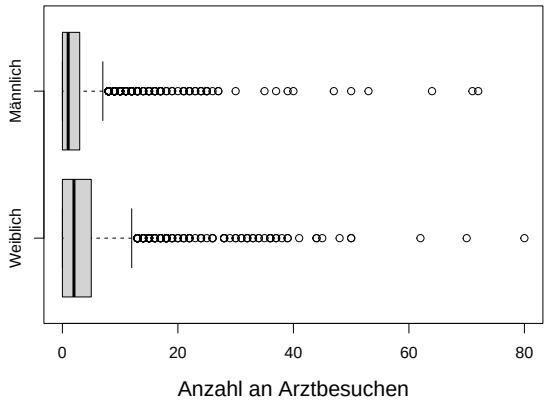

Abbildung 2: Boxplot zu Variable *Arztbesuche* getrennt nach Geschlecht

In Tabelle 3 sind die absoluten und relativen Häufigkeiten der 16 Dummy-Variablen aufgeführt. 2017 Befragte (52 %) sind männlich, 3420 Befragte (88%) haben keine *Behinderung*. Dies begründet die heterogene Verteilung der metrischen Variable *Behinderungsgrad*, die durch die starken Unterschiede der Lage- und Streuungsmaße dieser Variable in Tabelle 2 bereits aufgefallen ist. 2134 Befragte (55 %) gaben an, dass *Kinder* unter 16 Jahren mit ihnen im Haushalt wohnen. 3057 befragte Personen (79 %) sind *verheiratet*. 2647 Befragte (68 %) haben ihre *Bildungszeit* mit dem *Hauptschulabschluss*, 683 Befragte (18%) mit dem *Realschulabschluss*, 365 Befragte (9 %) mit dem *Abitur*, 133 Befragte (3 %) mit einem *Fachhochschulabschluss* und 235 Befragte (6%) mit einem *Hochschulabschluss* abgeschlossen. 2454 (63%) der Befragten befindet sich in einem *Beschäftigungsverhältnis*, 994 (26%) der Befragten sind *arbeitend*, 1044 (27%) *angestellt*, 237 (6%) *selbstständig* und 286 (7%) der Befragten sind *verbeamtet*. 3497 der Befragten (90%) sind *krankenversichert* und 14 befragte Personen (1 %) sind *zusatzversichert*.

Tabelle 3: Absolute und relative Häufigkeiten (abs./rel. Häufig.) der Dummy-Variablen

	abs. Häufig. "ja "	abs. Häufig. "nein "	rel. Häufig. "ja "	rel. Häufig. "nein "
Weiblich	1857.00	2017.00	0.48	0.52
Behinderung	454.00	3420.00	0.12	0.88
Kinder	1740.00	2134.00	0.45	0.55
Verheiratet	3057.00	817.00	0.79	0.21
Hauptschule	2647.00	1227.00	0.68	0.32
Realschule	678.00	3196.00	0.17	0.83
Abitur	144.00	3730.00	0.04	0.96
Fachhochschule	124.00	3750.00	0.03	0.97
Hochschule	235.00	3639.00	0.06	0.94
Beschäftigungsv.	2454.00	1420.00	0.63	0.37
Arbeitend	994.00	2880.00	0.26	0.74
Angestellt	1044.00	2830.00	0.27	0.73
Selbstständig	237.00	3637.00	0.06	0.94
Verbeamtet	286.00	3588.00	0.07	0.93
Krankenversichert	3497.00	377.00	0.90	0.10
Zusatzversichert	14.00	3860.00	0.01	0.99

Die absoluten und relativen Häufigkeiten der Merkmalsausprägungen der nominalen Variablen *Zufriedenheit* und *Behinderungsgrad* sind in Tabelle 5 (siehe Anhang) aufgeführt. Die meisten (jeweils 19.3 %) Befragten geben eine *Zufriedenheit* mit ihrer Gesundheit von „8“ (749 Befragte) oder „10“ (746 Befragte) an.

4.2 Regression der Krankenhausaufenthalte

Um die Wahl der Regressionsmethode zur Modellierung der Zähldaten zu *Krankenhausaufenthalten* der befragten Personen im letzten Kalenderjahr zu treffen, werden Erwartungswert und Varianz dieser Variable verglichen (Formel 14). Durch die circa viermal größere Varianz ist die Modellannahme der Poissonverteilung $\sigma^2{=}\mu{=}\lambda$ nicht eingehalten, sodass die *Krankenhausaufenthalte* nicht mit Poissonregression modelliert werden können.

$$\mathbb{E}(Krankenhausaufenthalte) = \mu = 0.12, \mathbb{V}ar(Krankenhausaufenthalte) = \sigma^2 = 0.48 \quad (14)$$

Unklar ist, ob der Zusammenhang zwischen Erwartungswert und Varianz der Krankenhausaufenthalte durch einen skalierten linearen oder einen quadratischen Zusammenhang besser beschrieben werden. Um dieser Frage nachzugehen, wird das volle Modell mit der Variable Krankenhausaufenthalte Regressanden und allen anderen unabhängigen Variablen als Regressoren jeweils mit Quasi-Poisson-Regressionsmethode und mit Negativ-Binomialverteilungs-Regressionsmethode untersucht. Zum Modellvergleich sind in Tabelle 4 die Kenngrößen Devianz und Dispersionsparameter der beiden Modelierungen aufgeführt. Durch die mehr als doppelt so große Devianz der Modellierung mit der Quasi-Poissonregressionsmethode zeichnet sich im Vergleich zur Negativ-Binomialverteilungs-Regressionsmethode bereits ab, dass letztere zu bevorzugen ist.

Tabelle 4: Vergleich der Quasi-Poisson- und der Negativ-Binomialverteilungs-Regressionsmethode zur Zielvariable Krankenhausaufenthalte

	Devianz	Dispersionsparameter
Quasi-Poissonregression	2426.5	2.635
Negativ-Binomialverteilungs-Regression	1018.5	0.156

Eine grafische Veranschaulichung des Zusammenhangs zwischen Erwartungswerten und Varianzen von 242 Gruppen, in die die 3873 Krankenhausaufenthalte-Beobachtungen mit jeweils 16 Einträgen sortiert wurden, ist in Abbildung 3 gegeben.

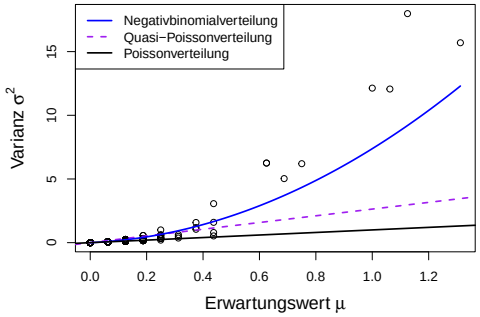

Abbildung 3: Zusammenhang zwischen Varianz σ^2 und Erwartungswert μ der gruppierten Krankenhausaufenthalte

Es ist ersichtlich, dass ein quadratischer Zusammenhang zwischen den Erwartungswerten und den Varianzen der gruppierten Beobachtungen der Krankenhausaufenthalten anzunehmen ist und demnach Regressionsmethode mit Negativ-Binomialverteilung zu wählen ist. Das so gebildete generalisierte lineare Modell ($\mathbf{M1}_{KH}$) mit der Variable Krankenhausaufenthalte als Zielvariable und allen unabhängigen Variablen als Regressoren lautet:

$$\begin{aligned}\widehat{\mathbb{E}(y_i)}^{(1)}_{KH} = exp(&-36.59 + 0.390 \cdot Zufried._1 - 0.443 \cdot Zufried._2 - 0.857 \cdot Zufried._3 \\ &- 1.246 \cdot Zufried._4 - 1.230 \cdot Zufried._5 - 1.443 \cdot Zufried._6 \\ &- 1.556 \cdot Zufried._7 - 2.328 \cdot Zufried._8 - 2.443 \cdot Zufried._9 \\ &- 1.905 \cdot Zufried._{10} + 0.7947 \cdot Arbeitend + 1.096 \cdot Verbeamtet)\end{aligned}$$

Dass laut Wald-Teststatistik kein signifikanter Einfluss festgestellt werden kann, ob die befragte Person weiblich ist, rechtfertigt die durch Abbildung 1 getroffene Entscheidung, einer nicht-geschlechtergetrennten Betrachtung der Krankenhausaufenthalte. Die Parameter $Zufriedenheit_1$ bis $Zufriedenheit_{10}$ (in der Gleichung abgekürzt durch $Zufried._1$ bis $Zufried._{10}$) geben die Ausprägungen Eins bis Zehn der Variable *Zufriedenheit* an. Zwar sind nur die Parameter $Zufriedenheit_3$ bis $Zufriedenheit_{10}$ laut Wald-Test für die Erklärung der Krankenhausaufenthalte signifikant, nach dem „alles-oder-nichts-Prinzip“ werden jedoch wegen des Anteils von 80 % signifikanter Ausprägungen alle Merkmalsausprägungen aufgeführt. Der AIC dieses Modells beträgt 2586.1, die Devianz wie oben bereits erwähnt 1018.5. Eine schrittweise Rückwärtselimination mit dem AIC Kriterium ergibt nach zwölf „Schritten“ das generalisierte lineare Modell ($\mathbf{M2}_{KH}$) mit der Variable Krankenhausaufenthalte als Zielvariable und den unabhängigen Variablen Zufriedenheit, Verheiratet, Beschäftigungsverhältnis, Arbeitend, Angestellt, Verbeamtet.

$$\begin{aligned}\widehat{\mathbb{E}(y_i)}^{(2)}_{KH} = exp(&-22.46 + 0.589 \cdot Arbeitend + 0.525 \cdot Angestellt + 0.916 \cdot Verbeamtet \\ &+ 0.2771 \cdot Zufried._1 - 0.492 \cdot Zufried._2 - 0.876 \cdot Zufried._3 \\ &- 1.306 \cdot Zufried._4 - 1.271 \cdot Zufried._5 - 1.492 \cdot Zufried._6 \\ &- 1.553 \cdot Zufried._7 - 2.327 \cdot Zufried._8 - 2.453 \cdot Zufried._9 \\ &- 1.888 \cdot Zufried._{10} - 0.407 \cdot Verheiratet - 0.534 \cdot Beschäftigt)\end{aligned}$$

Obwohl wieder nur 8 von 10 Merkmalsausprägungen der Variable Zufriedenheit nach dem Wald-Test signifikant sind, werden alle Zufriedenheitsparameter in das Modell ($\mathbf{M2}_{KH}$) mit aufgenommen. Der AIC liegt bei diesem Modell bei 2572.2 und die Devianz bei 1017.4. Da beide Größen kleiner als im Modell ($\mathbf{M1}_{KH}$) sind, hat die Rückwärtselimination das Modell sowohl nach den AIC als auch nach dem Devianzkriterium verbessert.
INTERPRETATION

Es folgt die in der deskriptiven Auswertung beschlossene geschlechtergetrennte Analyse des Zusammenhangs der Zielvariable Arztbesuche und der unabhängigen Variablen

Literatur

Cameron, A. Colin und Pravin K. Trivedi (2013). *Regression Analysis of Count Data*. Cambridge University Press.

Deutsches Institut für Wirtschaftsforschung e.V. (2022). *Sozio-oekonomisches Panel (SOEP)*. besucht am 22.12.2022. URL: `https://www.diw.de/de/diw_01.c.412809.de/sozio-oekonomisches_panel__soep.html`.

Fahrmeir, L., T. Kneib und S. Lang (2007). *Regression: Modelle, Methoden und Anwendungen*. Springer Berlin Heidelberg.

Fahrmeir, L. und G. Tutz (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag.

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 (2015). *Angewandte Statistik: Methodensammlung mit R*. Springer Berlin Heidelberg.

Hilbe, Joseph M. (2011). *Negative Binomial Regression*. Cambridge University Press.

Toutenburg, H. (2013). *Lineare Modelle: Theorie und Anwendungen*. Physica-Verlag HD.

Tripathi, A. (2019). *WHAT IS STEPAIC IN R?* besucht am 22.12.2022. URL: `https://ashutoshtripathi.com/2019/06/10/what-is-stepaic-in-r/#comments`.

Venables, W. N. und B. D. Ripley (2002). *Modern Applied Statistics with S*. Springer.

Anhang

Tabelle 5: Häufigkeitsangaben zu nominale Variablen

	abs. Häufig	rel. Häufig.
Zufriedenheit "0"	101	0.026
Zufriedenheit "1"	41	0.011
Zufriedenheit "2"	112	0.029
Zufriedenheit "3"	155	0.040
Zufriedenheit "4"	159	0.041
Zufriedenheit "5"	636	0.164
Zufriedenheit "6"	276	0.071
Zufriedenheit "7"	506	0.131
Zufriedenheit "8"	749	0.193
Zufriedenheit "9"	393	0.101
Zufriedenheit "10"	746	0.193
Behinderungsgrad "0"	3420	0.883
Behinderungsgrad "5"	2	0.001
Behinderungsgrad "10"	24	0.006
Behinderungsgrad "15"	2	0.001
Behinderungsgrad "20"	19	0.005
Behinderungsgrad "25"	5	0.001
Behinderungsgrad "30"	43	0.011
Behinderungsgrad "35"	2	0.001
Behinderungsgrad "40"	33	0.008
Behinderungsgrad "45"	0	0.000
Behinderungsgrad "50"	104	0.027
Behinderungsgrad "55"	2	0.001
Behinderungsgrad "60"	55	0.014
Behinderungsgrad "65"	0	0.000
Behinderungsgrad "70"	51	0.013
Behinderungsgrad "75"	0	0.000
Behinderungsgrad "80"	55	0.014
Behinderungsgrad "85"	0	0.000
Behinderungsgrad "90"	19	0.005
Behinderungsgrad "95"	0	0.000
Behinderungsgrad "100"	39	0.009

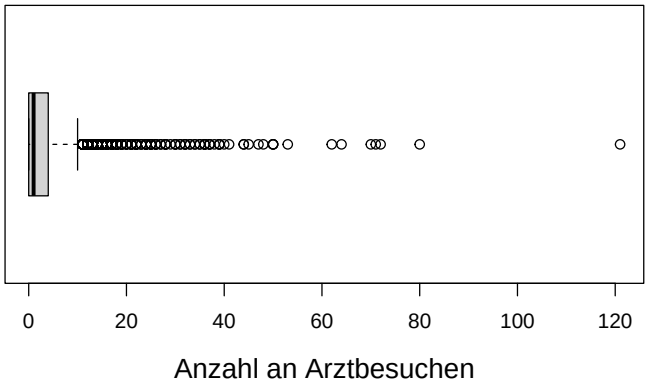

Abbildung 4: Boxplot zu Variable *Arztbesuche* getrennt nach Geschlecht

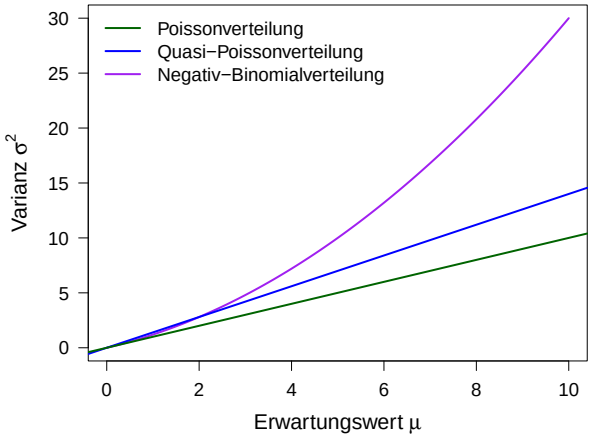

Abbildung 5: Zusammenhang zwischen Varianz σ^2 und Erwartungswert μ in unterschiedlichen Verteilungen