Logistische Regression

– Artikel Nr. 14 der Statistik-Serie in der DMW –

Logistic regression

Autoren R. Bender [1] A. Ziegler [2] S. Lange [1]

Institut
[1] Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, Köln
[2] Institut für Medizinische Biometrie und Statistik, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Universität zu Lübeck

Schlüsselwörter
- Logistische Regression
- Binäre Daten
- Chancenverhältnis
- Modellgüte

Key words
- Logistic regression
- Binary data
- Odds ratio (OR)
- Goodness-of-fit

Lineare Regression

Mit Hilfe der linearen Regression lässt sich der Einfluss einer oder mehrerer erklärender Variablen $X_1,...,X_m$ (z. B. X_1 = Alter, X_2 = Geschlecht und X_3 = Rauchen) auf eine stetige Zielvariable Y (z. B. Y = systolischer Blutdruck) statistisch untersuchen [3]. Liegt nur eine erklärende Variable X vor, spricht man von der **einfachen linearen Regression** (engl.: simple linear regression) und verwendet die Geradengleichung [5]

$$Y = \alpha + \beta X .$$

Im Fall mehrerer erklärender Variablen $X_1,...,X_m$ liegt das Modell der **multiplen linearen Regression** (engl.: multiple linear regression) vor, das durch die Gleichung

$$Y = \alpha + \beta_1 X_1 + ... + \beta_m X_m$$

beschrieben wird [3]. Die Bedeutung der multiplen Regressionsmodelle in der medizinischen Statistik liegt zum einen darin, den gemeinsamen Einfluss mehrerer Variablen auf eine Zielvariable untersuchen zu können und zum anderen in der Möglichkeit, den interessierenden Effekt einer Variable bezüglich anderer Variablen zu **adjustieren**, um eine **Verzerrung** (engl.: bias) bei der Effektschätzung zu reduzieren [3].

Logistische Regression

Die **logistische Regression** (engl.: logistic regression) kommt als Auswertungsmethode in Frage, wenn man den Einfluss erklärender Variablen $X_1,...,X_m$ auf eine Zielvariable Y untersuchen möchte, und Y **binäres Messniveau** besitzt (z. B. Y = Krankheit ja/nein). Da Y nur die beiden Werte 1 = ja und 0 = nein annehmen kann, ist die Anwendung der linearen Regression in der Regel nicht sinnvoll. Betrachten wir zur Modellentwicklung zunächst den einfachen Fall von nur einer erklärenden Variable X. Der Schlüssel zur quantitativen Beschreibung eines Zusammenhangs zwischen Y und X liegt darin, anstelle von Y die Wahrscheinlichkeit für den Eintritt des Zielereignisses p = P(Y = 1) zu modellieren. In medizinischen Anwendungen ist die Wahrscheinlichkeit p meist ein Risiko für eine bestimmte Krankheit. Während Y nur die beiden Ausprägungen 1 und 0 besitzt, kann das Risiko p jede beliebige Zahl zwischen 0 und 1 annehmen. Die **Chance** (engl.: odds) p/(1-p) kann jede beliebige positive Zahl annehmen [2] und der Logarithmus der Chance ***log***[p/(1-p)], genannt ***logit***, besitzt die ganze reelle Zahlenmenge als Wertebereich. Damit ist es häufig sinnvoll, eine lineare Beziehung zwischen dem logit von p und X anzunehmen, d. h.

$$\textit{logit}\,(p) = \textit{log}[\,p/(1\text{-}p)] = \alpha + \beta X ,$$

was mathematisch äquivalent ist mit

$$p = \frac{exp(\alpha + \beta X)}{1 + exp(\alpha + \beta X)}$$

Der Term „***exp***“ bezeichnet hierbei die Exponentialfunktion. Der rechte Term obiger Gleichung stellt die so genannte logistische Funktion dar, daher erklärt sich die Bezeichnung „logistische Regression“. Die Erweiterung auf ein multiples Modell mit mehreren erklärenden Variablen erhält man wie bei der linearen Regression, indem βX ersetzt wird durch die Linearkombination $\beta_1 X_1 + ... + \beta_m X_m$. Zur Schätzung der logistischen Regressionskoeffizienten werden in der Praxis iterative Algorithmen eingesetzt.

Da in der medizinische Forschung oftmals binäre Zielvariablen auftreten, wird die logistische Regression in der Praxis sehr häufig angewendet. Eine besondere Stellung erhält das logistische Regressionsmodell dadurch, dass man sowohl

Bibliografie
DOI 10.1055/s-2007-959037
Dtsch Med Wochenschr 2007; 132: e33–e35 · © Georg Thieme Verlag KG Stuttgart · New York · ISSN 0012-0472

Korrespondenz
Privatdozent Dr. rer. biol. hum. Ralf Bender
Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)
Dillenburger Straße 27
51105 Köln
eMail Ralf.Bender@iqwig.de

Tab. 1 Einfache logistische Regressionsanalyse für die Entwicklung einer diabetischen Nephropathie nach 6 Jahren bei 480 Typ 1 Diabetikern.

Risikofaktor	Regressionskoeffizient	Standardfehler	p-Wert	Differenz für Odds Ratio	Odds Ratio	95% Konfidenzintervall
Achsenabschnitt	- 5,089	0,731	0,0001			
HbA_{1c}	+ 0,457	0,089	0,0001	1%	1,58	1,33 – 1,88

Tab. 2 Multiple logistische Regressionsanalyse für die Entwicklung einer diabetischen Nephropathie nach 6 Jahren bei 480 Typ 1 Diabetikern.

Risikofaktor	Regressionskoeffizient	Standardfehler	*p*-Wert	Differenz für Odds Ratio	Odds Ratio	95% Konfidenzintervall
Achsenabschnitt	- 8,980	1,736	0,0001			
HbA_{1c}	+ 0,464	0,091	0,0001	1%	1,59	1,33 – 1,90
diast. Blutdruck	+ 0,048	0,019	0,0148	5 mm Hg	1,27	1,05 – 1,54
Diabetesdauer	+ 0,004	0,018	0,8220	5 Jahre	1,02	0,85 – 1,22
Geschlecht	- 0,025	0,249	0,9212	männl. vs. weibl.	0,98	0,60 – 1,59

für **prospektive Kohortenstudien** als auch für **retrospektive Fall-Kontroll Studien** sinnvoll interpretierbare Effektschätzer erhält. Das gebräuchliche Effektmaß in der Epidemiologie ist das **Odds Ratio** (OR), das als Verhältnis der Chancen zwischen exponierten und nicht exponierten Personen definiert ist [2]. Aus dem Regressionskoeffizient β einer logistischen Regression kann direkt das Odds Ratio berechnet werden durch $OR = \boldsymbol{exp}(\beta)$. In einem multiplen Modell kann für die Beziehung zwischen Y und einer erklärenden Variablen X_j das aus β_j berechnete $OR_j = \boldsymbol{exp}(\beta_j)$ als das nach allen anderen erklärenden Variablen **adjustierte Odds Ratio** betrachtet werden. Bei stetigen erklärenden Variablen bezieht sich der Wert des Odds Ratios auf die Erhöhung der erklärenden Variablen um jeweils 1 Einheit bzw. auf den Anstieg einer vorher definierten klinisch relevanten Differenz (siehe Beispiel).

Wie bei der linearen Regression muss auch bei der logistischen Regression die **Modellgüte** (engl.: goodness-of-fit) untersucht werden. Auf die entsprechenden Methoden können wir hier nicht eingehen. Der interessierte Leser sei auf die Literatur verwiesen [5]. Außer der logistischen Regression für binäre Zielvariablen gibt es Modellerweiterungen für nominale und ordinale Daten. Das bekannteste Modell ist hierbei das **proportionale Odds Modell** für ordinale Zielvariablen [1].

Beispiel

Mit Hilfe der logistischen Regression wurde der Einfluss von Risikofaktoren auf die Entwicklung der diabetischen Nephropathie bei Typ 1 Diabetikern untersucht [7]. Betrachten wir zunächst nur das glykierte Hämoglobin (HbA_{1c}) als Risikofaktor. In der einfachen logistischen Regressionsanalyse ist das HbA_{1c} ein signifikanter Risikofaktor (**Tab. 1**). Die Stärke des Effekts lässt sich mit Hilfe des Odds Ratios angeben. Pro Einheit des HbA_{1c} (1%) steigt die Chance nach 6 Jahren eine diabetische Nephropathie zu entwickeln um den Faktor von OR = 1.6 (95% Konfidenzintervall 1,3 – 1,9).

Dieser Zusammenhang lässt sich auch grafisch veranschaulichen, indem das Risiko als Funktion des Risikofaktors dargestellt wird (**Abb. 1**). Für HbA_{1c}-Werte im Normalbereich (4,3 – 6,1%) liegt das Risiko, eine diabetische Nephropathie zu entwickeln, unter 10%, während es bei extrem hohen HbA_{1c}-Werten von 16% und höher auf über 90% ansteigt.

Diese Ergebnisse verdeutlichen die starke Assoziation zwischen der Stoffwechseleinstellung und dem Risiko diabetischer Spätschäden bei Typ 1 Diabetes. Um zu zeigen, dass eine Reduktion des HbA_{1c} auch zu einer Reduktion des Risikos für diabetische Spätschäden führt, benötigt man allerdings entsprechende Ergebnisse einer randomisierten klinischen Therapiestudie, wie z. B. den Diabetes Control and Complications Trial (DCCT, [4]).

Neben dem glykierten Hämoglobin gibt es noch weitere Risikofaktoren, die hier in Betracht gezogen werden müssen, vor allem Blutdruck, Diabetesdauer und möglicherweise das Geschlecht. Die Ergebnisse einer multiplen logistischen Regressionsanalyse zeigen, dass das HbA_{1c} und der diastolische Blutdruck signifikante Risikofaktoren darstellen, während ein Effekt der Diabetesdauer und des Geschlechts nicht nachweisbar ist (**Tab. 2**).

Zur Darstellung des Odds Ratios wurde für den diastolischen Blutdruck eine Differenz von 5 mm Hg und für die Diabetesdauer von 5 Jahren gewählt, da eine Erhöhung dieser Risikofaktoren um jeweils eine Einheit (1 mm Hg bzw. 1 Jahr) nicht als klinisch relevante Änderung angesehen wird. Es lässt sich somit darstellen, dass bei einem Anstieg des diastolischen Blutdrucks um 5 mm Hg die Chance, nach 6 Jahren eine diabetische Nephropathie zu entwickeln, um den Faktor von OR = 1,3 (95% Konfidenzintervall 1,1 – 1,5) erhöht ist. Für das HbA_{1c} erhält man ähnliche

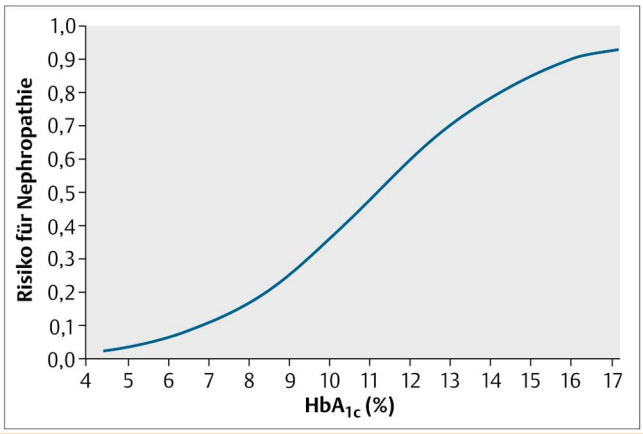

Abb. 1 Risiko für die Entwicklung einer diabetischen Nephropathie nach 6 Jahren in Abhängigkeit vom HbA_{1c} bei Typ 1 Diabetes, berechnet mit Hilfe einfacher logistischer Regressionsanalyse (n = 480).

Tab. 3 Übersetzung (deutsch – englisch).

erklärende Variable	explanatory variable
Zielvariable	response variable
einfache lineare Regression	simple linear regression
multiple lineare Regression	multiple linear regression
adjustieren	adjust
Verzerrung	bias
logistische Regression	logistic regression
binär	binary
Chance	odds
Kohortenstudie	cohort study
Fall-Kontroll Studie	case-control study
Regressionskoeffizient	regression coefficient
adjustiertes Odds Ratio	adjusted odds ratio
Modellgüte	goodness-of-fit
proportionales Odds Modell	proportional odds model

Resultate wie im einfachen Modell, d. h. in diesem Fall gibt es kaum Unterschiede zwischen den rohen und den adjustierten Resultaten bezüglich des Zusammenhangs zwischen der Stoffwechseleinstellung und dem Risiko einer diabetischen Nephropathie. Die englischen Bezeichnungen der hier diskutierten Begriffe zeigt **Tab. 3**.

kurzgefasst

Mit Hilfe der multiplen logistischen Regression lässt sich der Einfluss erklärender Variablen (Risikofaktoren) auf eine binäre Zielvariable (z. B. Krankheit ja/nein) untersuchen. Aus den Regressionskoeffizienten lassen sich adjustierte Odds Ratios als Maß für die Stärke des Zusammenhangs berechnen.

Dieser Beitrag ist eine überarbeitete Fassung aus dem Supplement Statistik aus dem Jahr 2002.

Literatur

1 *Bender R, Grouven U.* Ordinal logistic regression in medical research. J R Coll Physic London 1997; 31: 546–551
2 *Bender R, Lange S.* Die Vierfeldertafel. Dtsch Med Wochenschr 2007; 132: e12–e14
3 *Bender R, Ziegler A, Lange S.* Multiple Regression. Dtsch Med Wochenschr 2007; 132: e30–e32
4 *The Diabetes Control and Complications Trial Research Group.* The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. N Engl J Med 1993; 329: 977–986
5 *Hosmer DW, Lemeshow S.* Applied Logistic Regression (2nd Ed). Wiley, New York, 2000
6 *Lange S, Bender R.* (Lineare) Regression/Korrelation. Dtsch Med Wochenschr 2007; 132: e9–e11
7 *Mühlhauser I, Bender R, Bott U, Jörgens V, Grüsser M, Wagener W, Overmann H, Berger M.* Cigarette smoking and progression of retinopathy and nephropathy in type 1 diabetes. Diabet Med 1996; 13: 536–543