R. Bender[1]
A. Ziegler[2]
St. Lange[3]

Multiple Regression

- Artikel Nr. 13 der Statistik-Serie in der DMW -

Einfache lineare Regression

Mit Hilfe der **einfachen linearen Regression** (engl.: simple linear regression) lässt sich der Zusammenhang zwischen zwei stetigen Merkmalen statistisch untersuchen (6). Hierbei wird unterschieden zwischen der **erklärenden Variable** X (z.B. X = Gewicht in kg) und der **Zielvariable** Y (z.B. Y = systolischer Blutdruck in mmHg). Der Zusammenhang wird mit Hilfe der Geradengleichung

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

untersucht, d.h. man beschränkt sich auf die Untersuchung linearer Zusammenhänge. Ist die Annahme der Linearität verletzt, d.h. liegen die Punkte (X,Y) im Mittel gar nicht auf einer Geraden, müssen die Variablen entweder so transformiert werden, dass zwischen den transformierten Variablen ein linearer Zusammenhang besteht, oder es muss ein entsprechendes nichtlineares Regressionsmodell angewendet werden. Der Parameter von Interesse ist i.d.R. der **Regressionskoeffizient** β; er gibt den Anstieg von Y bei einem Anstieg von X um eine Einheit an: je größer der Betrag von β ist, desto größer ist der Einfluss von X auf Y. Der **Achsenabschnitt** (engl.: intercept) α gibt den Y-Wert bei $X = 0$ an. Im obigen Beispiel bedeutet $\beta = 1{,}31$, dass im Mittel mit jedem Anstieg des Gewichts um 1 kg der systolische Blutdruck um 1,31 mmHg ansteigt (6).

Multiple lineare Regression

Das einfache lineare Regressionsmodell lässt sich formal leicht auf ein multiples Modell mit mehreren erklärenden Variablen $X_1,...,X_m$ verallgemeinern durch

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

Mit Hilfe dieses Modells der **multiplen linearen Regression** (engl.: multiple linear regression) lässt sich der gemeinsame lineare Einfluss der erklärenden Variablen $X_1,...,X_m$ auf die Zielvariable Y untersuchen. Dieses allgemeine Modell hat in der medizinischen Statistik eine große Bedeutung, da sehr viele Fragestellungen zur Anwendung multipler Regressionsmodelle führen.

Zunächst einmal sind reine bivariate Zusammenhänge in der medizinischen Forschung eher die Ausnahme. Zwar besteht ein Zusammenhang zwischen dem Gewicht als erklärender Variable X und dem systolischen Blutdruck als Zielvariable Y, aber in aller Regel gibt es weitere Variablen, die einen Einfluss auf Y haben, z.B. X_2 = Alter, X_3 = Geschlecht und X_4 = Rauchen. Durch ein multiples lineares Regressionsmodell lässt sich also der gemeinsame Einfluss der Variablen Gewicht, Alter, Geschlecht und Rauchen auf dem systolischen Blutdruck untersuchen. Zu beachten ist hierbei, dass in die Modellgleichung nur erklärende Variablen mit stetigem und/oder binären Messniveau betrachtet werden können. Erklärende kategorielle Variablen mit mehr als zwei Kategorien können durch Kodierungstechniken berücksichtigt werden. Am häufigsten werden die Variablen mit Hilfe der so genannten Dummy-Kodierung berücksichtigt. Das bedeutet, dass man eine Kategorie als Referenzkategorie wählt und die anderen Kategorien im Vergleich hierzu als binäre Variablen darstellt. Auf diese Weise lässt sich der Einfluss einer erklärenden Variable mit k Kategorien durch k-1 Regressionskoeffizienten darstellen.

In vielen Anwendungen der medizinischen Statistik interessiert man sich zwar zunächst nur für den Einfluss einer erklärenden Variable X_1 auf die Zielvariable Y, jedoch muss man andere Variablen im Modell berücksichtigen, um den Zusammenhang zwischen X_1 und Y möglichst unverzerrt schätzen zu können. Ein häufiges Beispiel ist der Vergleich von 2 Gruppen (z.B. exponierte und nicht exponierte Personen) bezüglich der Zielvariable Y (wie

Institut
[1] AG Epidemiologie und Medizinische Statistik (Leitung: Prof. Dr. M. Blettner), Fakultät für Gesundheitswissenschaften , Universität Bielefeld
[2] Institut für Medizinische Biometrie und Statistik (Direktor: Prof. Dr. A. Ziegler), Universitätsklinikum Lübeck, Medizinische Universität zu Lübeck
[3] Abteilung für Medizinische Informatik, Biometrie u. Epidemiologie (Direktor: Prof. Dr. H.J. Trampisch), Ruhr-Universität Bochum

Korrespondenz
Priv.-Doz. Dr. rer. biol. hum. Ralf Bender · AG Epidemiologie und Medizinische Statistik
Fakultät für Gesundheitswissenschaften
Universität Bielefeld · Postfach 100131 · 33501 Bielefeld · E-Mail: Ralf.Bender@uni-bielefeld.de

Bibliografie
Dtsch Med Wochenschr 2002; 127: T 8–T 10 · © Georg Thieme Verlag Stuttgart · New York · ISSN 0012-0472

bisher z.B. Y = systolischer Blutdruck) in einer Beobachtungsstudie. Wenn die Gruppenzugehörigkeit nicht durch Randomisierung zugewiesen werden konnte, kann man nicht davon ausgehen, dass alle weiteren für Y wichtigen erklärenden Variablen in den Gruppen gleich verteilt sind. Würde man einfach den t-Test (7) zum Vergleich der Gruppen anwenden, könnte ein signifikanter Unterschied zwischen den Gruppen sowohl auf einen Effekt der Exposition, als auch auf systematische Unterschiede zwischen den beiden Gruppen bezüglich anderer Variablen (z.B. Alter, Geschlecht und Rauchen) zurückzuführen sein. Um eine solche **Verzerrung** (engl.: bias) bei der Schätzung des Expositionseffekts zu reduzieren (im Idealfall auf Null), müssen die wichtigen, d.h. die prognostisch relevanten Einflussvariablen berücksichtigt werden. Dies ist, als Erweiterung des t-Tests, mit Hilfe eines multiplen Regressionsmodells möglich, in dem die erklärenden Variablen X_1 = Exposition, X_2 = Alter, X_3 = Geschlecht und X_4 = Rauchen gemeinsam in einem Modell betrachtet werden. Durch ein solches Modell erhält man den interessierenden Expositionseffekt durch den Regressionskoeffizienten β_1. Da im multiplen Modell die anderen erklärenden Variablen und damit mögliche systematische Unterschiede bezüglich dieser Variablen berücksichtigt sind, spricht man hier von einem nach Alter, Geschlecht und Rauchen **adjustierten Regressionskoeffizienten**.

Eine solche multifaktorielle Analyse kann in Interventionsstudien kein Ersatz für eine Randomisierung sein. Die Berechnung adjustierter Effekte stellt aber in Fällen, in denen aus ethischen oder praktischen Gründen keine Randomisierung durchgeführt werden kann, eine wesentlich adäquatere Auswertungsstrategie dar als die einfache Schätzung der rohen nicht adjustierten Effekte.

Beispiel: Effektivität eines ambulanten Gewichtsreduktionsprogramms

In einer Beobachtungsstudie zur Effektivität eines ambulanten Gewichtsreduktionsprogramms wurde anhand einer Stichprobe von n = 294 übergewichtigen Patienten untersucht, welche Faktoren mit einer Gewichtsabnahme assoziiert sind (5). Eine Fragestellung war, ob die Gewichtsabnahme bei Männern und Frauen unterschiedlich ist. Als Zielvariable wurde die relative Gewichtsänderung zwischen Therapieende und Therapieanfang in %

$$Y = 100 \times (\text{Gewicht am Ende} - \text{Anfangsgewicht}) / \text{Anfangsgewicht}$$

betrachtet, d.h. bei negativen Werten für Y liegt eine Gewichtsabnahme vor. Es zeigte sich, dass Männer im Durchschnitt (-8,83%) mehr abgenommen haben als Frauen (–7,16%). Der Unterschied von –1,67% ist aber nicht signifikant (t-Test: $p = 0{,}0731$). Genau das gleiche Resultat erhält man, indem eine einfache lineare Regression mit der binären erklärenden Variable Geschlecht (1 = männlich, 0 = weiblich) durchgeführt wird. Der Regressionskoeffizient entspricht dann gerade der mittleren Differenz zwischen Männern und Frauen (**Tab.1**).

Die Formulierung als Regressionsmodell hat den Vorteil, dass es sich auf den Fall mehrerer erklärender Variablen verallgemeinern lässt. Potenzielle erklärende Variablen für die Gewichtsabnahme sind hier u.a. die Dauer der Behandlung und der Bildungsstand. Der Einfachheit halber beschränken wir uns in diesem Beispiel auf die Betrachtung dieser Variablen. Die Berücksichtigung der Behandlungsdauer ist hier besonders wichtig, da diese Variable einen starken Einfluss auf die Gewichtsabnahme besitzt und bei Männern und Frauen unterschiedlich verteilt ist. Während Männer im Mittel 5,8 Monate am Programm teilnahmen, lag diese Zahl bei Frauen im Mittel bei 7,2 Monaten. Da die Behandlungsdauer mit einer höheren Gewichtsabnahme assoziiert ist, ergibt sich bei der einfachen Betrachtung des Unterschieds zwischen Männern und Frauen ein Bias. Dieser kann durch ein adäquates multiples Modell ausgeglichen werden. In einer multiplen linearen Regression mit den erklärenden Variablen Geschlecht, Behandlungsdauer (in Monaten) und Bildungsstand (1 = hoch, 0 = niedrig) zeigt sich ein signifikanter Einfluss des Geschlechts (**Tab.2**).

Tab.**1** Einfache lineare Regressionsanalyse für die Assoziation zwischen prozentualer Gewichtsabnahme und Geschlecht bei 294 übergewichtigen Patienten.

	Regressionskoeffizient	*Standardfehler*	*95% Konfidenzintervall*	*p-Wert*
Achsenabschnitt	– 7,158	0,473		0,0001
Geschlecht (männl. vs. weibl.)	– 1,675	0,931	– 3,50 bis +0,15	0,0731

Tab.**2** Multiple lineare Regressionsanalyse für die Assoziationen zwischen prozentualen Gewichtsabnahme und Geschlecht, Behandlungsdauer und Bildungsstand bei 294 übergewichtigen Patienten.

	Regressionskoeffizient	*Standardfehler*	*95% Konfidenzintervall*	*p-Wert*
Achsenabschnitt	– 3,152	0,594		0,0001
Geschlecht (männl. vs. weibl.)	– 2,416	0,819	– 4,02 bis – 0,81	0,0034
Behandlungsdauer (Monate)	– 0,530	0,059	– 0,65 bis – 0,41	0,0001
Bildungsstand (hoch vs. niedrig)	– 4,566	1,886	– 8,26 bis – 0,87	0,0161

Der nach Behandlungsdauer und Bildungsstand adjustierte durchschnittliche Unterschied zwischen Männern und Frauen ist identisch mit dem Regressionskoeffizienten des Geschlechts (–2,42%, $p = 0{,}0034$), der deutlich höher ist als der rohe nicht adjustierte Unterschied. Durch ein multiples Regressionsmodell lassen sich auch adjustierte Mittelwerte für die einzelnen Gruppen schätzen. Bei gleicher Behandlungsdauer und gleichem Bildungsstand beträgt die relative Gewichtsveränderung bei Männern im Mittel –9,383% und bei Frauen –6,967%; die Differenz dieser beiden Werte ergibt gerade den Wert des Regressionskoeffizienten.

Modellbildung und Modellgüte

Die sinnvolle Anwendung der multiplen Regressionsanalyse in der Praxis ist sehr viel komplizierter als hier in Kürze dargestellt werden kann. Außer der Auswahl der Zielvariablen und der erklä-

renden Variablen sollte zunächst eine konkrete Modellgleichung entwickelt werden, welche die untersuchten Zusammenhänge adäquat beschreibt. Dazu gehört die Betrachtung von möglichen **Transformationen** sowohl der Zielvariablen als auch der erklärenden Variablen, die Untersuchung möglicher nichtlinearer Zusammenhänge durch quadratische oder kubische Effekte und Überlegungen zu möglichen **Wechselwirkungen** (engl.: interactions) zwischen den erklärenden Variablen. Zur Modellbildung und Untersuchung der **Modellgüte** (engl.: goodness-of-fit) gibt es eine Reihe von Verfahren, die als **Regressionsdiagnostiken** (engl.: regression diagnostics) bezeichnet werden. Auf diese Methoden kann im Rahmen dieses Artikel nicht eingegangen werden. Der interessierte Leser sei auf die Literatur verwiesen (3, 4, 8).

Ein Maß für den prädiktiven Wert eines multiplen linearen Regressionsmodells ist das multiple **Bestimmtheitsmaß R^2** (engl.: coefficient of determination). Es stellt für die Untersuchung von Zusammenhängen zwischen mehr als zwei Variablen eine Verallgemeinerung des quadrierten Korrelationskoeffizienten (6) dar. Das Bestimmtheitsmaß R^2 gibt den Anteil der Varianz der Zielvariablen an, der durch alle erklärenden Variablen im multiplen Regressionsmodell gemeinsam erklärt werden kann. Im betrachteten Beispiel der Assoziationen zwischen relativer Gewichtsabnahme und den erklärenden Variablen Geschlecht, Behandlungsdauer und Bildungsstand ergibt sich der Wert $R^2 = 0{,}25$, d.h. durch alle 3 Faktoren gemeinsam lässt sich 25% der Variabilität der Gewichtsabnahme erklären. Ein großer Anteil der Variabilität wird durch andere Faktoren erklärt, so dass sich die Gewichtsabnahme eines übergewichtigen Patienten aus der Kenntnis der 3 erklärenden Variablen vermutlich nicht mit genügender Genauigkeit ableiten lässt.

Ein limitierender Faktor bei der Anwendung multipler Regressionsmodelle in der Praxis ist häufig der Stichprobenumfang. Einerseits müssen in der Regressionsgleichung alle wichtigen erklärenden Variablen enthalten sein, andererseits benötigt man mit steigender Zahl der erklärenden Variablen auch größere Stichproben. Der benötigte Stichprobenumfang hängt natürlich immer von der konkreten Situation ab. Als Faustregel gilt jedoch, dass man in einer multiplen linearen Regression pro Modellparameter mindestens 10 Beobachtungen benötigt, um ein einigermaßen stabiles Modell zu erhalten (4).

Übersicht über Regressionsmethoden

Die multiple lineare Regression ist eine spezielle Klasse der Regressionsmethoden, die in Frage kommt, wenn die betrachtete Zielvariable *stetiges Messniveau* besitzt. Je nach Zahl und Messniveau der erklärenden Variablen lassen sich auch der *t*-Test (7) und die Methoden der Varianzanalyse (1) in die Klasse der linearen Regression einbetten. Ein lineares Regressionsmodell mit genau einer erklärenden Variablen mit binärem Messniveau ist äquivalent zum *t*-Test (Vergleich von 2 Gruppen). Liegt eine erklärende Variable mit nominalem Messniveau (Vergleich mehrerer Gruppen) vor, ergibt sich das Varianzanalysemodell der Einfachklassifikation. Bei mehreren nominal skalierten erklärenden Variablen, erhält man die Varianzanalysemodelle der Mehrfachklassifikation.

Tab. **3** Übersetzung (deutsch – englisch).

Deutsch	*Englisch*
einfache lineare Regression	simple linear regression
erklärende Variable	explanatory factor
Zielvariable	response variable
Regressionskoeffizient	regression coefficient
Achsenabschnitt	intercept
multiple lineare Regression	multiple linear regression
Verzerrung	bias
adjustiert	adjusted
Wechselwirkung	interaction
Modellgüte	goodness-of-fit
Regressionsdiagnostiken	regression diagnostics
Bestimmtheitsmaß	coefficient of determination
logistische Regression	logistic regression
proportionales Hazards Modell	proportional hazards model

Hat die betrachtete Zielvariable kein stetiges Messniveau, so kann die Klasse der linearen Regressionsmodelle nicht sinnvoll angewendet werden. Bei *binären* Zielvariablen (Ereignis ja/nein), kommt die **logistische Regression** (engl.: logistic regression), bei *Überlebenszeiten* (9) als Zielgröße das **proportionale Hazards Modell** von Cox in Frage. Auf diese Modelle werden wir in weiteren Artikeln eingehen (2, 10). Die englischen Bezeichnungen der hier diskutierten Begriffe zeigt **Tab. 3**.

kurzgefasst: **Mit Hilfe der multiplen linearen Regression lassen sich Assoziationen zwischen einer stetigen Zielvariablen und mehreren erklärenden Variablen untersuchen. Der Regressionskoeffizient einer erklärenden Variable stellt ein nach den anderen Variablen adjustiertes Effektmaß dar.**

Literatur

1 Altman DG, Bland JM. Comparing several groups using analysis of variance. Br med J 1996; 312: 1472–1473
2 Bender R, Ziegler A, Lange S. Logistische Regression. Dtsch Med Wochenschr 2002; 127: T11–T13
3 Draper NR, Smith H. Applied Regression Analysis (3rd Ed). Wiley, New York, 1998
4 Harrell FEJr., Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996; 15: 361–387
5 Heise T, Kimmerle R, Heinemann L, Schubert H, Bender R, Pußkailer M, Berger M. Weight reduction in an out-patient obesity clinic: Which factors are associated with success? Int J Obes 1995; 19 (Suppl 2); 155
6 Lange S, Bender R. (Lineare) Regression/Korrelation. Dtsch Med Wochenschr 2001; 126: T33–T35
7 Lange S, Bender R. Was ist ein Signifikanztest? Dtsch Med Wochenschr 2001; 126: T42–T44
8 Ziegler A, Arminger G. Individualdaten-Regressionsanalyse. Vorlesungsskript (Kurs-Nr. 00887), FernUniversität-Gesamthochschule Hagen, 2000
9 Ziegler A, Lange S, Bender R. Überlebenszeitanalyse: Eigenschaften und Kaplan-Meier Methode. Dtsch Med Wochenschr 2002; 127: T14–T16
10 Ziegler A, Lange S, Bender R. Das Cox-Modell. Dtsch Med Wochenschr 2002; 127 (in Vorbereitung)