ÜBERSICHTSARBEIT

Lineare Regressionsanalyse

Teil 14 der Serie zur Bewertung wissenschaftlicher Publikationen

Astrid Schneider, Gerhard Hommel, Maria Blettner

ZUSAMMENFASSUNG

Hintergrund: Die Regressionsanalyse ist eine wichtige statistische Methode zur Auswertung medizinischer Daten. Sie ermöglicht es, Zusammenhänge zwischen verschiedenen Faktoren zu analysieren und aufzudecken. Des Weiteren können prognostisch wichtige Risikofaktoren identifiziert werden, die die Bildung von Risikoscores für die Erstellung von individuellen Prognosen ermöglichen.

Methoden: Die Arbeit basiert auf ausgewählten Lehrbüchern der Statistik, einer selektiven Literaturauswahl und der eigenen Expertise.

Ergebnisse: Nach einer kurzen Darstellung des univariablen und multivariablen Regressionsmodells wird anhand von Beispielen erklärt, was vor der Durchführung einer Regression zu beachten ist und wie die Ergebnisse interpretiert werden können. Der Leser soll in die Lage versetzt werden, zu beurteilen, ob Methoden korrekt angewandt wurden und wie die Resultate zu bewerten sind.

Schlussfolgerung: Die Durchführung und Interpretation einer linearen Regressionsanalyse beinhaltet zahlreiche Fallstricke, auf die hier ausführlich eingegangen wird. Darüber hinaus werden dem Leser häufige Fehler bei der Interpretation mittels Beispielen aus der Praxis verdeutlicht. Zusätzlich werden sowohl die Möglichkeiten als auch die Grenzen der linearen Regressionsanalyse aufgezeigt.

▶ **Zitierweise**
Schneider A, Hommel G, Blettner M: Linear regression analysis—part 14 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2010; 107(44): 776–82. DOI: 10.3238/arztebl.2010.0776

Institut für Medizinische Biometrie, Epidemiologie und Informatik der Johannes-Gutenberg-Universität Mainz: Dipl.-Math. Schneider, Prof. Dr. rer. nat. Hommel, Prof. Dr. rer. nat. Blettner

Bei statistischen Auswertungen von medizinischen Daten besteht das Ziel oft darin, Zusammenhänge zwischen zwei oder mehreren Variablen zu beschreiben. Es ist beispielsweise nicht nur interessant, ob Patienten einen erhöhten Blutdruck haben, sondern auch, ob dieser durch Faktoren wie Gewicht oder Alter des Patienten beeinflusst wird. Die Variable, die erklärt werden soll (Blutdruck), bezeichnet man als abhängige Variable (Zielvariable). Die erklärenden Variablen (Gewicht, Alter) werden unabhängige Variablen (Einflussvariablen) genannt. Einen ersten Eindruck von der Stärke eines Zusammenhangs bieten Zusammenhangsmaße. Falls Zielvariable und Einflussvariable stetig sind (Blutdruck und Gewicht), beschreiben Korrelationskoeffizienten die Stärke des Zusammenhangs *(Kasten 1)*.

Die Regressionsanalyse ermöglicht es, drei Aspekte zu untersuchen:

- Beschreibung: Der Zusammenhang zwischen Zielvariable und Einflussvariablen kann mittels Regressionsanalysen statistisch beschrieben werden.
- Schätzung: Die Werte der Zielvariablen können mittels der beobachteten Werte der Einflussvariablen geschätzt werden.
- Prognose: Prognostisch wichtige Risikofaktoren können identifiziert und individuelle Prognosen erstellt werden.

Dabei verwendet die Regressionsanalyse ein Modell, das den Zusammenhang zwischen der Zielvariablen und den verschiedenen Einflussvariablen vereinfacht in einer mathematischen Form beschreibt. Es kann biologische Gründe geben, bestimmte Funktionen anzunehmen oder es werden einfache Annahmen (Blutdruck steigt linear mit dem Alter) gemacht. Die bekanntesten Regressionsanalysen sind *(Tabelle 1)*:

- die lineare Regression
- die logistische Regression
- die Cox-Regression.

Dieser Artikel hat das Ziel, eine Einführung in die lineare Regression zu geben. Neben einer kurzen Erläuterung der Theorie wird anhand von Beispielen auf die Interpretation der statistischen Parameter eingegangen. Die Methoden der Regressionsanalysen werden in vielen Standardlehrbüchern ausführlich dargestellt (1–3).

Die Cox-Regression wird in einem folgenden Artikel im Deutschen Ärzteblatt erörtert.

Methoden

Mittels linearer Regression wird der lineare Zusammenhang zwischen einer Zielvariablen Y (Blutdruck) und einer oder mehreren Einflussvariablen X (Gewicht, Alter, Geschlecht) untersucht.

Die Zielvariable Y muss stetig sein, die Einflussvariablen können stetig (Alter), binär (Geschlecht) oder kategorial (Sozialstatus) sein. Für die erste Beurteilung eines möglichen Zusammenhangs zwischen zwei stetigen Variablen sollte die Darstellung immer durch ein Streudiagramm (Punktwolke) erfolgen. Hier wird sichtbar, ob es sich um einen linearen *(Grafik 1)* oder einen nichtlinearen Zusammenhang *(Grafik 2)* handelt.

Nur im Falle eines linearen Zusammenhangs ist die Durchführung einer linearen Regression sinnvoll. Zur Untersuchung von nichtlinearen Zusammenhängen müssen andere Methoden herangezogen werden. Oft bieten sich Variablentransformationen oder andere komplexere Methoden an, auf die hier nicht eingegangen wird.

Univariable lineare Regression

Die univariable lineare Regression untersucht den linearen Zusammenhang zwischen der Zielvariablen Y und nur einer Einflussvariablen X. Das lineare Regressionsmodell beschreibt die Zielvariable durch eine Gerade $Y = a + b \times X$, mit a = Achsenabschnitt und b = Steigung der Geraden. Zunächst werden aus den Werten der Zielvariablen Y und der Einflussvariablen X die Parameter a und b der Regressionsgerade mit Hilfe statistischer Methoden geschätzt. Die Gerade ermöglicht, Werte der Zielvariablen Y durch Werte der Einflussvariablen X vorherzusagen. Nach Durchführung einer linearen Regression könnte beispielsweise das Gewicht (Zielgröße) einer Person mittels ihrer Körpergröße (Einflussvariable) geschätzt werden *(Grafik 3)*.

Die Steigung b der Regressionsgeraden wird als Regressionskoeffizient bezeichnet. An ihm lässt sich der Beitrag der Einflussvariablen X für die Erklärung der Zielgröße Y ablesen. Bei einer stetigen Einflussgröße (zum Beispiel Körpergröße in cm) beschreibt der Regressionskoeffizient die Veränderung der Zielvariablen (Körpergewicht in kg) pro Maßeinheit der Einflussvariablen (Körpergröße in cm). Bei der Interpretation ist es daher wichtig, die Maßeinheiten zu berücksichtigen. Das nachfolgende Beispiel verdeutlicht diese Beziehung:

Betrachtet werden fiktive Daten von 135 Frauen und Männern zwischen 18 und 27 Jahren. Die Körpergröße der Personen liegt zwischen 1,59 m und 1,93 m. Untersucht wird der Zusammenhang zwischen Größe und Gewicht. Das Gewicht (kg) ist die Zielvariable, die mittels der Einflussvariable Körpergröße (cm) geschätzt werden soll. Aus den Daten ergibt sich die Regressionsgerade $Y= -133{,}18 + 1{,}16 \times X$, wobei X der Größe in cm entspricht. Der y-Achsenabschnitt a = –133,18 ist der Wert der Zielgröße Y bei X = 0, der in diesem Fall nicht vorkommen kann. Die

KASTEN 1

Interpretation des Korrelationskoeffizienten (r)

Spearman's Koeffizient:
Betrachtung eines monotonen Zusammenhangs
Man spricht von einem monotonen Zusammenhang, falls gilt, dass mit wachsenden Werten der einen Variablen, die andere Variable stetig wächst oder sinkt.

Korrelationskoeffizient nach Pearson:
Betrachtung eines linearen Zusammenhangs.

Interpretation/Bedeutung:
Korrelationskoeffizienten geben Auskunft über die Stärke und Richtung eines Zusammenhangs zwischen zwei stetigen Variablen. Eine Differenzierung zwischen „erklärender" und „zu erklärender" Variable ist nicht erforderlich. Es gilt:

- r = ± 1: perfekter linearer beziehungsweise monotoner Zusammenhang. Je näher r betragsmäßig bei 1 liegt, desto stärker ist der Zusammenhang.
- r = 0: kein linearer beziehungsweise monotoner Zusammenhang
- r < 0: negativer, umgekehrter Zusammenhang (große Werte der einen Variablen treten eher bei kleineren Werten der anderen auf)
- r > 0: positiver Zusammenhang (große Werte der einen Variablen treten eher bei großen Werten der anderen auf)

Grafische Darstellung eines linearen Zusammenhangs:
Streudiagramm mit Regressionsgerade
Ein negativer Zusammenhang wird durch eine fallende Regressionsgerade (Regressionskoeffizient b < 0) beschrieben – ein positiver Zusammenhang mit einer steigenden Gerade (b > 0).

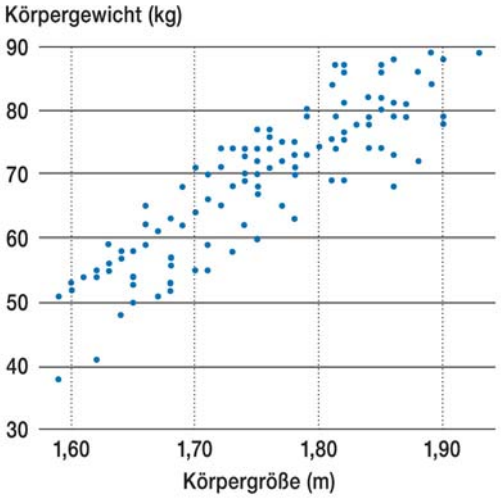
TABELLE 1

Regressionsmodelle

	Anwendungsgebiet	abhängige Variablen	unabhängige Variablen
lineare Regression	Beschreibung eines linearen Zusammenhanges	stetig (Gewicht, Blutdruck)	stetig und/oder kategorial
logistische Regression	Prognose der Zugehörigkeitswahrscheinlichkeit zu den Gruppen (Ereignis: ja/nein)	dichotom (Behandlungserfolg: ja/nein)	
Proportional-Hazard-Regression (Cox-Regression)	Modellierung von Überlebenszeitdaten	Überlebenszeiten (Zeitspanne, Diagnose bis Ereignis)	
Poisson-Regression	Modellierung von Zählprozessen	Zähldaten ganzzahlige geordnete Ereignisse von Prozessen (Anzahl der Geburten einer Frau innerhalb eines Zeitintervalls)	

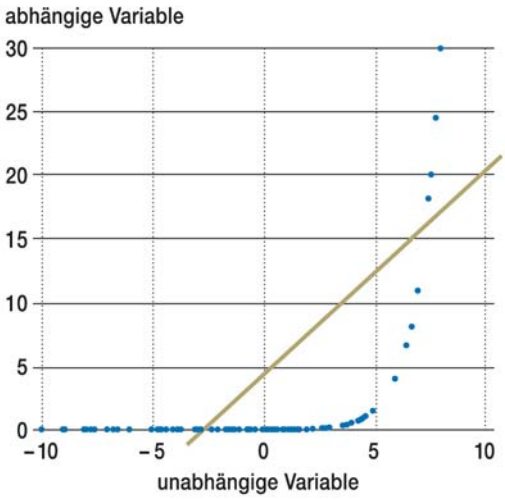
Darstellung eines linearen Zusammenhangs anhand einer Punktwolke.

GRAFIK 1



Darstellung eines exponenziellen Zusammenhangs durch eine Punktwolke. Die Angabe von Determinationskoeffizient und Regressionsgerade ist nicht geeignet.

GRAFIK 2



Interpretation der Konstante a ist oft nicht sinnvoll. Generell sollten nur Werte aus dem Beobachtungsbereich der Einflussvariablen eingesetzt werden. Je weiter der eingesetzte Wert von diesem Bereich entfernt ist, desto unsicherer ist die Schätzung der Zielvariablen.

Der Regressionskoeffizient von 1,16 bedeutet, dass nach diesem Modell mit jedem zusätzlichen Zentimeter das Gewicht um 1,16 kg steigt. Eine Angabe der Körpergröße in Metern ergibt einen Regressionskoeffizienten von b = 115,91. Die Konstante a bleibt von der gewählten Maßeinheit der Einflussvariablen unbeeinflusst. Für die Interpretation ist es daher wichtig, dass der Regressionskoeffizient zusammen mit der gewählten Maßeinheit der betreffenden Variablen betrachtet wird. Dies ist besonders zu beachten, wenn in internationalen Arbeiten unterschiedliche Maßeinheiten (inch, feet, pound) verwendet wurden.

Grafik 3 zeigt die Regressionsgerade, die den linearen Zusammenhang zwischen Größe und Gewicht darstellt.

Für eine 1,74 m große Person wird der Wert 68,50 kg (y = –133,18 + 115,91 × 1,74 m) für ihr Gewicht vorhergesagt. Dies bedeutet, dass der geschätzte Wert von Personen mit der Größe von 1,74 m bei 68,50 kg liegt. In diesem Datensatz existieren 6 Personen von dieser Größe. Das Gewicht dieser Personen variiert zwischen 63 kg und 75 kg.

Mittels linearer Regression kann das Gewicht jeder Person geschätzt werden, deren Körpergröße im betrachteten Bereich (1,59 m bis 1,93 m) liegt. Es ist nicht erforderlich, dass der Datensatz eine Person mit dieser Größe enthält. Rein rechnerisch ist es möglich, das Gewicht einer Person zu schätzen, deren Körpergröße außerhalb des beobachteten Bereichs liegt. Eine solche Extrapolation ist jedoch im Allgemeinen nicht sinnvoll.

Sind die Einflussvariablen kategorial oder binär, muss bei der Interpretation die Kodierung der Ausprägungen beachtet werden. Bei binären Variablen empfiehlt es sich, zwei aufeinanderfolgende ganze Zahlen zu wählen (meistens 0/1 oder 1/2). Bei der Interpretation muss man berücksichtigen, welche Ausprägung der höheren Kodierung entspricht. Der Regressionskoeffizient gibt die Änderung der Zielvariablen bei der höher kodierten Ausprägung im Vergleich zu der niedrig kodierten an.

Untersucht man den Zusammenhang zwischen Geschlecht und Gewicht, erhält man die Regressionsgerade Y = 47,64 + 14,93 × X mit X = Geschlecht (1 = weiblich, 2 = männlich). Der Regressionskoeffizient von 14,93 bedeutet, dass Männer durchschnittlich 14,93 kg mehr wiegen als Frauen.

Bei kategorialen Variablen ist zunächst die Referenzkategorie zu definieren, alle anderen Kategorien werden im Verhältnis zu dieser Gruppe betrachtet.

Wie gut das Regressionsmodell die Daten beschreibt, kann anhand des Bestimmtheitsmaßes (Determinationskoeffizient, r^2) bewertet werden *(Kasten 2)*. Bei der univariablen Regressionsanalyse entspricht r^2 dem Quadrat des Korrelationskoeffizienten von Pearson. Für den Zusammenhang von Körpergröße und Gewicht erhält man einen Determinationskoeffizient von 0,785. Es sind 78,5 % der Variation des Gewichts auf die Größe zurückzuführen. Die restlichen 21,5 % sind individuelle Abweichungen und könnten durch andere, nicht betrachtete Einflussgrößen wie Ernährungsgewohnheiten, Sport, Geschlecht oder Alter erklärt werden.

Formal kann die Hypothese β = 0 (das heißt, Regressionskoeffizient = 0, es gibt keinen Zusammenhang) mittels eines t-Tests untersucht werden. Zusätzlich kann man das 95-%-Konfidenzintervall für den Regressionskoeffizienten angeben (4).

Multivariable lineare Regression

Oft reicht der Beitrag einer Variablen zur Erklärung der Zielvariablen Y nicht aus. In diesen Fällen ist es möglich, im Rahmen einer multivariablen linearen Regression den gemeinsamen Einfluss mehrerer Variablen auf die Zielvariable zu untersuchen.

Die Zielvariable wird durch eine lineare Funktion $Y = a + b_1 \times X_1 + b_2 \times X_2 + \ldots + b_n \times X_n$ der erklärenden Variablen X_i beschrieben. Für jede Einflussgröße X_i schätzt man durch das Regressionsmodell einen Regressionskoeffizient b_i *(Kasten 3)*.

Wie bei der univariablen Regression beschreibt auch hier das Bestimmtheitsmaß den Gesamtzusammenhang zwischen den Einflussvariablen X_i (Gewicht, Alter, BMI) und der Zielvariablen Y (Blutdruck). Es entspricht dem Quadrat des multiplen Korrelationskoeffizienten, also der Korrelation zwischen Y und $b_1 \times X_1 + \ldots + b_n \times X_n$.

Man sollte jedoch besser das korrigierte Bestimmtheitsmaß angeben *(Kasten 2)*. Die einzelnen Koeffizienten b_i spiegeln den Einfluss der jeweils zugehörigen unabhängigen Variablen X_i auf Y, unter Berücksichtigung des Einflusses der anderen unabhängigen Variablen, wider. Betrachtet man eine multivariable Regressionsanalyse mit Alter und Geschlecht als Einflussvariablen und Gewicht als Zielgröße, so gibt der adjustierte Regressionskoeffizient für das Geschlecht den Anteil an der Variation des Gewichtes unter Berücksichtigung des Alters wieder. Der Einfluss des Alters wurde durch die Altersadjustierung aus dem Einfluss des Geschlechts herausgerechnet *(Kasten 4)*.

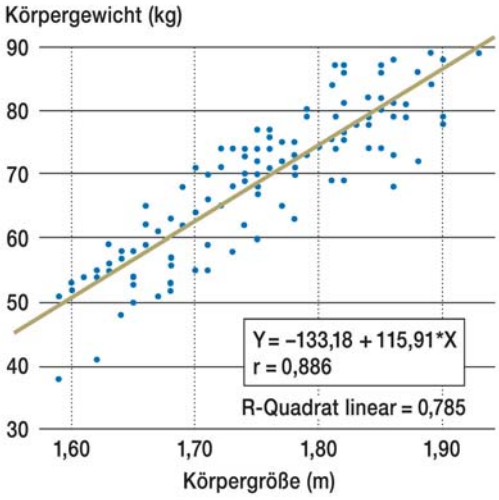
Die multivariable Regressionsanalyse ermöglicht, neben der gleichzeitigen Betrachtung von mehreren Einflussgrößen, die Regressionskoeffizienten der interessierenden Einflussgrößen bezüglich möglicher Störgrößen zu adjustieren.

Neben der Beschreibung des Zusammenhangs erlaubt das multivariable Modell individuelle Prognosen und die Beurteilung des Gesundheitszustands eines bestimmten Patienten. Durch ein lineares Regressionsmodell können beispielsweise Sollwerte für die Atemfunktion unter Berücksichtigung von Alter, Body-Mass-Index (BMI) und Geschlecht erstellt werden. Durch den Vergleich des ermittelten Wertes für einen Patienten mit dem Sollwert kann man Schlussfolgerungen hinsichtlich seines Gesundheitszustandes ziehen.

Bei medizinischen Fragestellungen liegen oft sehr viele Einflussvariablen vor. Ziel der Analyse ist es, herauszufinden, welche der Faktoren tatsächlich einen Einfluss auf die Zielvariable haben. Die Kunst der statistischen Auswertung besteht darin, die Variablen zu finden, die die Zielvariable am besten erklären.

Eine Möglichkeit, eine multivariable Regression durchzuführen, besteht darin, alle potenziellen Einflussvariablen in das Modell aufzunehmen (vollständiges Modell). Problematisch bei dieser Methode ist, dass häufig zu wenige Beobachtungen vorliegen, um ein solches Modell zu untersuchen. Die Zahl der Beobachtungen sollte etwa 20-mal größer sein als die Zahl der untersuchten Variablen.

GRAFIK 3



Punktwolke mit Regressionsgerade und Regressionsgleichung für den Zusammenhang zwischen der Zielgröße Körpergewicht (kg) und der unabhängigen Variable Körpergröße (m).
r = Korrelationskoeffizient nach Pearson
R-Quadrat linear = Bestimmtheitsmaß

Werden zudem viele irrelevante Variablen ins Modell eingeschlossen, kommt es zu einer Überanpassung: das heißt, irrelevante unabhängige Variablen zeigen aufgrund von Zufallseffekten scheinbar einen Einfluss. Durch die Aufnahme von irrelevanten Einflussgrößen wird das Modell zwar besser an den vorliegenden Datensatz angepasst, jedoch kann es nicht mehr auf die Allgemeinheit übertragen werden (1). Des Weiteren wird durch die Aufnahme von irrelevanten Einflussvariablen die tatsächliche Anpassungsgüte stark verzerrt *(Kasten 2)*.

Im Folgenden wird gezeigt, wie die oben beschriebenen Probleme umgangen werden können.

Variablenselektion

Um ein robustes Regressionsmodell zu erhalten, das Y möglichst gut erklärt, sollen nur solche Variablen in das Modell eingeschlossen werden, die einen großen Anteil von Y erklären. Mittels einer Variablenselektion können diese Einflussvariablen selektiert werden (1).

Die Variablenselektion sollte mit medizinischem Fachwissen und guten biometrischen Kenntnissen durchgeführt werden, am besten in Zusammenarbeit von Statistiker und Mediziner. Es gibt verschiedene Methoden der Variablenselektion:

Vorwärtsselektion

Die Vorwärtsselektion schließt schrittweise Variablen ins Modell ein, die einen zusätzlichen Beitrag zur Erklärung von Y leisten. Dies geschieht solange, bis es keine Variablen mehr gibt, die einen wesentlichen Beitrag an der Erklärung von Y liefern.

KASTEN 2

Bestimmtheitsmaß (Determinationskoeffizient, R-Quadrat)

Definition:
Es sei:

- n die Anzahl der Beobachtungen (Personen)
- $\hat{y}_i$ die Schätzung der Zielgröße für die i-te Beobachtung mittels der Regressionsgleichung
- y_i der beobachtete Wert der Zielgröße für die i-te Beobachtung und
- $\bar{y}$ der Mittelwert über alle n Beobachtungen der Zielgröße

Dann ist das Bestimmungsmaß wie folgt definiert:

$$r^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\text{erklärte Varianz}}{\text{gesamte Varianz}} = \frac{\text{erklärte Variation}}{\text{gesamte Variation}}$$

→ r^2 ist der Anteil der erklärten Varianz an der Gesamtvarianz. Je näher die durch das Regressionsmodell geschätzten Werte $\hat{y}_i$ an den beobachteten Werten y_i liegen, desto näher liegt das Bestimmungsmaß bei dem Wert 1 und desto genauer ist das Regressionsmodell.

Bedeutung: In der Praxis wird der Determinationskoeffizient oft als Maß für die Güte eines Regressionsmodells beziehungsweise einer Regressionsschätzung verwendet. Er spiegelt den durch die Regressionsgerade aufgeklärten Anteil an der Variation der Y-Werte wider.

Problem: Der Determinationskoeffizient kann leicht künstlich in die Höhe getrieben werden, wenn man in das Modell viele unabhängige Variablen aufnimmt. Es gilt: Je größer die Anzahl der unabhängigen Variablen ist, desto höher ist der Determinationskoeffizient. Dies hat jedoch negative Auswirkungen auf die Genauigkeit des Schätzers (Schätzung der Regressionskoeffizienten b_i).

Lösung: Angabe des korrigierten Determinationskoeffizienten, der die Anzahl der erklärenden Variablen im Modell mitberücksichtigt. Anders als beim unkorrigierten Determinationskoeffizienten steigt der korrigierte Determinationskoeffizient nur, wenn die unabhängigen Variablen einen genügend großen Einfluss zeigen.

Rückwärtsselektion
Die Rückwärtsselektion beginnt mit einem Modell, das alle interessierenden unabhängigen Variablen enthält. Schrittweise werden nun die Variablen aus dem Modell entfernt, durch deren Verlust die Vorhersage der abhängigen Variablen am wenigsten verschlechtert wird. Dies wird so lange wiederholt, bis keine Einflussvariable ausgeschlossen werden kann, ohne die Vorhersage deutlich zu verschlechtern.

Schrittweise Selektion
Die schrittweise Selektion verbindet Aspekte der Rückwärts- und der Vorwärtsselektion. Wie die Vorwärtsselektion startet sie mit einem Nullmodell und schließt schrittweise die Variable ins Modell ein, die den größten Anteil an der Erklärung der Zielvariablen hat. Zusätzlich wird überprüft, ob eine Variable aufgrund ihrer Beziehung zu den anderen Variablen überflüssig geworden ist und entfernt werden kann.

KASTEN 3

Regressionsgerade einer multivariablen Regression

$Y = a + b_1 \times X_1 + b_2 \times X_2 + \ldots + b_n \times X_n$ mit
Y = Zielvariable
X_i = Einflussvariablen
a = Konstante, Schnittpunkt mit der y-Achse
b_i = Regressionskoeffizient der Variablen X_i

Bsp: Regresssionsgerade einer multivariablen Regression
$Y = -120{,}07 + 100{,}81 \times X_1 + 0{,}38 \times X_2 + 3{,}41 \times X_3$ mit
X_1 = Größe in Meter
X_2 = Alter in Jahren
X_3 = Geschlecht mit der Kodierung 1 = weiblich, 2 = männlich
Y = das zu schätzende Gewicht in kg

Blockweise Einschluss

Oft gibt es Variablen, die auf jeden Fall ins Modell aufgenommen werden sollen – beispielsweise der Einfluss einer bestimmten Therapie oder bereits bekannte Einflussgrößen. Eine Möglichkeit, dies zu berücksichtigen, ist der blockweise Einschluss der Variablen ins Modell. Durch diese Vorgehensweise kann der Einschluss von bestimmten Variablen erzwungen und zusätzliche Variablen können mittels Variablenselektion aufgenommen werden.

Für die Beurteilung des Regressionsmodells ist es notwendig, sowohl eine Vorwärts- als auch eine Rückwärtsselektion durchzuführen. Falls bei beiden Verfahren die gleiche Variablenkombination ausgewählt wird, kann das Modell als robust angesehen werden. Andernfalls sollte ein Statistiker hinzugezogen werden.

Diskussion

Die Untersuchung von Zusammenhängen und die Erstellung von Risikoscores sind in der medizinischen Forschung sehr bedeutsam. Eine Regressionsanalyse erfordert es jedoch, verschiedene Faktoren zu beachten und zu überprüfen:

1. Kausalität

Bevor eine Regressionsanalyse durchgeführt wird, muss im Vorfeld die Kausalität zwischen den zu betrachtenden Variablen aufgrund von inhaltlichen oder zeitlichen Überlegungen geklärt werden. Ein signifikanter Faktor sagt nichts über die Kausalität aus, gerade in Beobachtungsstudien muss diese geprüft werden (5).

2. Fallzahlplanung

Die für eine Regressionsanalyse erforderliche Fallzahl hängt von den zu erwartenden Effekten (Stärke des Zusammenhanges) und der Anzahl der Einflussfaktoren ab. Ist die Stichprobe zu klein, lassen sich nur sehr starke Zusammenhänge nachweisen. Für eine Fallzahlplanung können sowohl Informationen über das zu erwartende Bestimmtheitsmaß (r^2) als auch über den Regressionskoeffizienten (b) genutzt werden. Zusätzlich sollte die Anzahl der Beobachtungen nicht das 20-fache der zu untersuchenden Einflussvariablen unterschreiten. Möchte man zwei Einflussvariablen betrachten, sollten also mindestens 40 Beobachtungen vorliegen.

3. Fehlende Werte

Fehlende Werte sind ein häufiges Problem bei medizinischen Daten. Sobald bei einer der Einflussfaktoren oder bei der abhängigen Variable ein Wert fehlt, wird diese Beobachtung aus der Regressionsanalyse ausgeschlossen. Fehlen viele Werte im Datensatz, verringert sich die betrachtete Fallzahl beträchtlich. Die erforderliche Fallzahl wird trotz guter Fallzahlplanung nicht eingehalten. In solchen Fällen könnten existierende Zusammenhänge übersehen und die Übertragbarkeit auf die Allgemeinheit gefährdet werden. Zudem ist mit Selektionseffekten zu rechnen. Es gibt verschiedene Methoden mit fehlenden Werten umzugehen (6).

KASTEN 4

Begriffe

- Confounder (Störvariablen, bei nicht randomisierten Studien): Variablen, die sowohl mit der Zielgröße als auch mit anderen Einflussvariablen assoziiert sind. Der Effekt der anderen Einflussvariablen kann verzerrt werden. Häufige Confounder sind Alter und Geschlecht.
- Adjustierung: Statistisches Verfahren zur Bereinigung des Effektes einer oder mehrerer Störgrößen auf einen Behandlungseffekt. Beispiel: Untersuchung des Therapieeffekts auf eine bestimmte Zielgröße unter Berücksichtigung des Alters (Alter = Störvariable). Durch die Altersadjustierung wird rein rechnerisch so getan, als ob die Frauen und Männer in dem betrachteten Datensatz gleich alt wären. Somit kann der Einfluss des Alters aus dem Einfluss der Therapie herausgerechnet werden.

KASTEN 5

Worauf ist bei der Interpretation einer Regressionsanalyse zu achten?

1. Wie groß ist die Fallzahl?
2. Ist eine Kausalität inhaltlich oder zeitlich nachweisbar und plausibel?
3. Wurde bezüglich möglicher Störgrößen adjustiert?
4. Wurden die unabhängigen Variablen inhaltlich begründet?
5. Wie hoch ist der korrigierte Determinationskoeffizient (R-Quadrat)?
6. Ist das Stichprobenkollektiv homogen?
7. In welcher Maßeinheit wurden die möglichen Einflussvariablen dargestellt?
8. Wurde eine Variablenselektion bezüglich der unabhängigen Variablen (mögliche Einflussgrößen) durchgeführt, und wenn ja, welche?
9. Falls eine Variablenselektion erfolgte, wurde das Ergebnis durch ein anderes Selektionsverfahren bestätigt?
10. Beruhen Vorhersagen der Zielvariablen auf extrapolierten Daten?

4. Stichprobenkollektiv

Neben der Fallzahl ist auch die Stichprobenzusammensetzung ein wichtiger Punkt, der beachtet werden muss. Bestehen in der Stichprobe Subkollektive, die sich bezüglich der Einflussvariablen verschieden verhalten, könnten Effekte unentdeckt bleiben. Um dieses Problem zu verdeutlichen, wird im Folgenden der Einfluss des Geschlechts auf das Gewicht in einem Kollektiv untersucht, das zu gleichen Teilen aus Kindern (unter 8 Jahren) und Erwachsenen besteht. Eine lineare Regressionsanalyse über das gesamte Kollektiv zeigt einen Einfluss des Geschlechts auf das Gewicht. Wertet man Kinder und Erwachsene getrennt voneinander

aus (Subgruppenanalyse), so zeigt sich nur bei den Erwachsenen ein Einfluss des Geschlechts auf das Gewicht, nicht jedoch bei den Kindern. Bei Subgruppenanalysen sollten die Subkollektive jedoch in Abhängigkeit des Wissenschaftsstands und der Fragestellung bereits vor der Auswertung feststehen. Des Weiteren muss das multiple Testen berücksichtigt werden (7, 8).

5. Variablenselektion

Werden bei einer multivariablen Regression mehrere Einflussfaktoren betrachtet, kann eine wechselseitige Abhängigkeit zwischen Einflussfaktoren existieren. Variablen, die innerhalb eines univariablen Regressionsmodells einen starken Einfluss zeigen, könnten bei einer multivariablen Regression mit Variablenselektion nicht ins Modell aufgenommen werden. Eine Erklärung dafür ist, dass aufgrund des starken Zusammenhangs zwischen den Einflussvariablen die jeweils andere Variable keinen zusätzlichen Beitrag zur Erklärung der Zielvariablen leistet. Daher könnten, je nach verwendetem Variablenselektionsverfahren, unterschiedliche Einflussfaktoren ins Modell eingeschlossen werden.

Fazit

Die lineare Regression ist ein wichtiges Werkzeug in der statistischen Auswertung. Sie umfasst mit der Beschreibung von Zusammenhängen, der Schätzung und der Prognose ein weites Einsatzspektrum. Trotz ihrer vielfältigen Anwendungsgebiete muss man bei der Interpretation die Limitationen und Voraussetzungen berücksichtigen *(Kasten 5)*.

Interessenkonflikt
Die Autoren erklären, dass kein Interessenkonflikt im Sinne der Richtlinien des International Committee of Medical Journal Editors besteht.

Manuskriptdaten
eingereicht: 11. 5. 2010, revidierte Fassung angenommen: 14. 7. 2010

LITERATUR

1. Fahrmeir L, Kneib T, Lang S: Regression – Modelle, Methoden und Anwendungen. 2nd edition. Berlin, Heidelberg: Springer 2009
2. Bortz J: Statistik für Human-und Sozialwissenschaftler. 6th edition. Heidelberg: Springer 2004.
3. Selvin S: Epidemiologic Analysis. Oxford University Press 2001.
4. Bender R, Lange S: Was ist ein Konfidenzintervall? Dtsch Med Wschr 2001; 126: T41.
5. Sir Bradford Hill A: The environment and disease: Association or Causation? Proc R Soc Med 1965; 58: 295–300.
6. Carpenter JR, Kenward MG: Missing Data in Randomised Controlled Trials: A practical guide. Birmingham, Alabama: National Institute for Health Research; 2008. http://www.pcpoh.bham.ac.uk/publichealth/methodology/projects/RM03_JH17_MK.shtml. Publication RM03/JH17/MK.
7. EMEA: Poiints to consider on multiplicity issues in clinical trials; www.emea.europa.eu/pdfs/human/ewp/090899en.pdf
8. Horn M, Vollandt R: Multiple Tests und Auswahlverfahren. Stuttgart: Gustav Fischer Verlag 1995.

Anschrift für die Verfasser
Prof. Dr. rer. nat. Maria Blettner
Institut für Medizinische Biometrie
Epidemiologie u. Informatik der
Johannes Gutenberg-Universität
Obere Zahlbacher Straße 69
55131 Mainz

SUMMARY

Linear Regression Analysis—Part 14 of a Series on Evaluation of Scientific Publications

Background: Regression analysis is an important statistical method for the analysis of medical data. It enables the identification and characterization of relationships among multiple factors. It also enables the identification of prognostically relevant risk factors and the calculation of risk scores for individual prognostication.

Methods: This article is based on selected textbooks of statistics, a selective review of the literature, and our own experience.

Results: After a brief introduction of the uni- and multivariate regression models, illustrative examples are given to explain what the important considerations are before a regression analysis is performed, and how the results should be interpreted. The reader should then be able to judge whether the method has been used correctly and interpret the results appropriately.

Conclusion: The performance and interpretation of linear regression analysis contains numerous pitfalls, which are discussed here in detail. The reader is made aware of common errors of interpretation by means of practical examples. Both the opportunities for applying linear regression analysis and its limitations are presented.

Zitierweise
Schneider A, Hommel G, Blettner M: Linear regression analysis—part 14 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2010; 107(44): 776–82. DOI: 10.3238/arztebl.2010.0776

The English version of this article is available online:
www.aerzteblatt-international.de