Wissenschaftliche Fragestellungen in der Medizin brauchen statistische Modelle: Lineare und logistische Regression

Held, U

Wissenschaftliche Fragestellungen in der Medizin brauchen statistische Modelle

Lineare und logistische Regression

Ulrike Held,
Horten Zentrum, UniversitätsSpital, Zürich

In der Forschung ist es häufig von Interesse, eine Ursache-Wirkungs-Beziehung herzuleiten, also den Einfluss von einem Parameter, wie z.B. dem Körpergewicht, auf eine Zielgrösse, z.B. den systolischen Blutdruck, zu untersuchen. Oder aber man möchte wissen, ob bestimmte Ernährungsgewohnheiten das Auftreten von Krankheiten beeinflussen. Zur Beantwortung dieser oder ähnlicher Fragestellungen können statistische Regressionsmodelle verwendet werden, mit denen der Einfluss von einer Einflussgrösse oder mehreren -grössen auf eine Zielvariable untersucht werden kann. In Abhängigkeit von der Verteilung der Zielgrösse, also ob sie z.B. auf einer kontinuierlichen Skala (metrisch) als «ja/nein»-Variable (dichotom) oder als Zeit bis zum Auftreten eines bestimmten Ereignisses, vorliegt wird ein entsprechendes statistisches Modell verwendet. Bei der Wahl des geeigneten Modells ist einerseits die klinische Kompetenz des Arztes sowie auch biostatistisches Know-how gefragt.
In diesem Artikel beginnen wir mit der Beschreibung der vielleicht häufigsten statistischen Modelle in der medizinischen Forschung: dem linearen Regressionsmodell und dem logistischen Regressionsmodell.

Lineare Regression

Unter einem linearen Modell verstehen wir, dass der Zusammenhang zwischen Einfluss- und Zielgrösse typischerweise folgendermassen beschrieben werden kann:

$$\text{Zielgrösse} = a + b \times \text{Einflussgrösse}.$$

Die Interpretation dieser Beziehung ist die folgende: Liegt eine feste, nicht zufällige Einflussgrösse vor, hier z.B. das Körpergewicht, so ergibt sich die Zielgrösse, hier der systolische Blutdruck, im Prinzip durch die oben angegebene Formel. Linearität bezieht sich hier nur auf die Parameter a und b und nicht auf die Einflussgrösse selbst. Man kann also mit einem linearen Regressionsmodell auch den Zusammenhang von $a + b \times (\text{Einflussgrösse})^2$ mit der Zielgrösse untersuchen.
Ziel der Regressionsanalyse ist nun die Schätzung der unbekannten Grössen a und b, welche dem Achsenabschnitt und der Steigung der sog. Regressionsgeraden entsprechen.
Wir betrachten einen fiktiven Datensatz mit 20 Beobachtungen von Gewicht, systolischem Blutdruck und Alter bei Männern wie in Tabelle 1 dargestellt.
Die 18 vollständigen Datenpaare von Gewicht und Blutdruck können gegeneinander in einem sog. Streudiagramm oder Scatterplot aufgetragen werden, wie in Abbildung 1 zu sehen ist. Jeweils eine Beobachtung von Gewicht und Blutdruck muss ausgeschlossen werden, da der zugehörige Wert in der jeweils anderen Variable fehlend ist. Anschliessend wurden der Achsenabschnitt a und die Steigung b der Regressionsgeraden geschätzt, woraus sich die folgende Regressionsgleichung ergibt:

$$\text{Systolischer Blutdruck} = 41{,}2 + 0{,}95 \times \text{Gewicht}.$$

Die daraus resultierende Regressionsgerade wurde ebenfalls in die Abbildung 1 eingefügt.
Man kann anhand der Schätzung von b = 0,95, aber auch anhand von Abbildung 1 erkennen, dass der Zusammenhang von Blutdruck und Gewicht positiv ist, da die Steigung der Regressionsgeraden grösser Null ist, andernfalls würde man von einem negativen Zusammenhang sprechen.
Es gibt natürlich über den linearen Zusammenhang hinausgehend eine gewisse zufällige Abweichung für jedes einzelne Individuum. Die Differenz zwischen dem durch das Modell vorhergesagten Wert und der tatsächlichen Beobachtung bezeichnen wir als Fehlerterm. Dieser Fehlerterm ist im Durchschnitt über alle Patienten Null, hat aber eine Variation. Exemplarisch wurde für die Person mit dem geringsten Gewicht der Fehlerterm abgetragen. Er besteht aus dem vertikalen Abstand der tatsächlichen Beobachtung und der Regressionsgeraden an der entsprechenden Stelle. Konkret ist in diesem Fall der tatsächlich beobachtete Wert des Blutdrucks um 17,6 mm Hg höher, als man durch das Regressionsmodell schätzen würde.
Möchte man nun noch wissen, ob das Körpergewicht einen *signifikanten* Einfluss auf die Zielgrösse hat, d.h. also, ob der Einfluss «statistisch gesichert» von Null verschieden ist, betrachtet man die Schätzung der Steigung und deren *p-Wert*. Die Steigung, oder die Schätzung des Einflusses des Gewichts, ist gerade 0,95. Diese Aussage kann als statistisch gesichert angesehen werden, denn der zugehörige p-Wert ist 0,04 (also <0,05) und somit signifikant auf dem 5%-Niveau.
Bei der Verwendung eines statistischen Modells für einen medizinisch relevanten Zusammenhang sollte man sich immer fragen, ob die Wahl des Modells «richtig» war bzw. wie gross der Abstand zwischen der tatsächlichen Beobachtung und dem geschätzten Wert für alle Beobachtungspaare tatsächlich ist. Im Fall eines linearen Regressionsmodells ist es deshalb sinnvoll, eine Masszahl R^2 anzugeben, die diese Anpassungsgüte beschreibt. R^2 gibt den Anteil der Variation in der Zielgrösse (Blutdruck) an, der durch die Variation der Ein-


Ulrike Held

flussgrösse (Gewicht) erklärt wird. Will man nun die Anpassungsgüte eines Modells beurteilen, so kann man sagen, dass ein R^2 nahe an Null eine unbefriedigende Anpassung bedeutet, wohingegen ein hohes R^2 nahe an Eins eine gute lineare Anpassung bedeutet. In unserem Beispiel liegt R^2 bei 0,23, also 23%.

Tabelle 1. Gewicht, Blutdruck und Alter der 20 Männer.

Patient	Gewicht (kg)	Syst. Blutdruck (mm Hg)	Alter (Jahre)
1	92,8	153,6	70
2	82,9	116,2	72
3	101,6	157,8	74
4	97,7	153,5	55
5	111,3	153,2	57
6	fehlend	123,3	59
7	73,3	128,4	61
8	87,2	fehlend	59
9	114,7	126,1	71
10	113,1	167,9	66
11	97,4	130,5	67
12	90,2	141,2	70
13	95,0	109,9	54
14	89,4	89,8	53
15	90,4	137,6	62
16	92,2	114,1	74
17	105,1	130,6	69
18	89,4	138,1	80
19	88,7	109,4	79
20	86,3	103,5	58

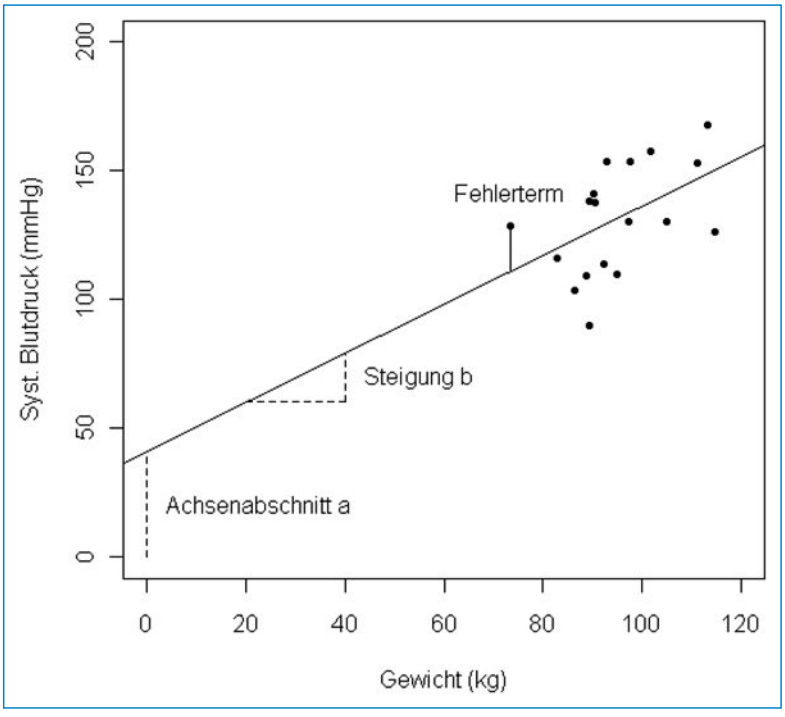

Abbildung 1
Streudiagramm der 18 Beobachtungspaare mit der linearen Regressionsgeraden.

Falls man ein sehr geringes R^2 erhält, könnte dies einerseits daran liegen, dass der Zusammenhang zwischen Gewicht und Blutdruck nicht linear ist. Allerdings deutet die graphische Darstellung des Zusammenhangs nicht darauf hin. Weiterhin könnte es sein, dass wichtige weitere Variablen zur Erklärung der Zielgrösse «systolischer Blutdruck» noch fehlen. Möchte man das lineare Regressionsmodell auf mehrere Einflussgrössen ausweiten, spricht man anstelle von einfacher Regression nun von multipler Regression.
Betrachtet man in diesem Beispiel noch das zugehörige Alter der Patienten als Einflussgrösse, ergibt sich das folgende Regressionsmodell:

$$\text{Systolischer Blutdruck} = a + b \times \text{Gewicht} + c \times \text{Alter}.$$

Ziel ist nun, die Grössen a, b, und c zu schätzen bzw. die Stärke des Einflusses von Gewicht und Alter gemeinsam auf den systolischen Blutdruck zu quantifizieren.
Daraus ergibt sich eine Schätzung für den Einfluss von Gewicht (also b) von 0,95, der zugehörige p-Wert ist 0,047. Die Schätzung für c, also den Koeffizienten für Alter, ist 0,38, und der p-Wert ist 0,50. Hier wurde nun also ein multiples Regressionsmodell angepasst, wobei der Koeffizient für Gewicht weiterhin statistisch signifikant ist – und fast unverändert im Vergleich zu dem einfachen Regressionsmodell. Der Koeffizient des zusätzlichen Faktors Alter ist nicht signifikant auf dem 5%-Niveau. Es kann in manchen Fällen vorkommen, dass sich der Einfluss von einzelnen Regressionskoeffizienten deutlich verändert, wenn man zusätzliche Variable in das Modell aufnimmt. Wenn wir nun wieder die Anpassungsgüte betrachten, stellen wir fest, dass sich das R^2 unter Hinzunahme des Alters auf 26% verbessert.
Ein anderer Aspekt, der bisher noch nicht angesprochen wurde, ist die mögliche Verallgemeinerbarkeit von linearen Zusammenhängen eines Regressionsmodells. In unserem fiktiven Datensatz mit 20 Patienten ist der Datenbereich *(Range)*, also das Intervall von vorkommenden Werten für Gewicht, gerade von 73 bis 115 kg. Die zugehörigen Werte für den systolischen Blutdruck liegen im Intervall von 90 bis 168 mm Hg. Wir haben ein lineares Regressionsmodell auf diese Daten angepasst, aber die Ergebnisse dieses Modells sind nicht unbedingt auf alle anderen Bereiche dieser Variablen extrapolierbar. So könnte man mit Hilfe dieses Modells berechnen, dass eine Person mit einem Gewicht von 60 kg im Mittel einen vorhergesagten Wert von 98 mm Hg besitzt, allerdings könnte die Art des Zusammenhangs in unteren Bereichen der Variablen Gewicht auch z.B. ganz anders als linear sein.

Logistisches Regressionsmodell

Im Gegensatz zum linearen Regressionsmodell wird ein logistisches Regressionsmodell angewendet, wenn die Zielgrösse keine kontinuierliche Verteilung hat, sondern nur zwei Ausprägungen haben kann (z.B. ja/nein, oder liegt vor / liegt nicht vor). Die Theorie zur logistischen Regression ist ebenso umfangreich wie die der linearen Regression, aber für den Anwender sind die

Unterschiede nicht besonders gross und liegen hauptsächlich in der Interpretation der Ergebnisse: Bei der linearen Regression kann der Effekt der Einflussgrössen auf die Zielgrösse direkt aus dem geschätzten Koeffizienten abgelesen werden. In unserem Beispiel war der Effekt von Gewicht ja 0,95, also bei einem Patienten mit 10 Kilogramm mehr Gewicht erhöht sich der systolische Blutdruck um 9,5 mm Hg. Bei der logistischen Regression hingegen wird die Zielvariable (mit ihren zwei möglichen Ausprägungen) nicht direkt modelliert, sondern eine Funktion der Wahrscheinlichkeit, dass das Ereignis unter gegebenen Risiko-Bedingungen auftritt. Die geschätzten Effekte der Einflussgrösse(n) sind dann als Odds-Ratios zu interpretieren, auf die wir in einem späteren Artikel noch gesondert eingehen werden.

Neben der linearen und der logistischen Regression werden wir in Zukunft auch noch auf andere wichtige Modellklassen wie die Cox-Regression für die Analyse von Überlebenszeiten, Varianzanalysemodelle und Zeitreihenmodelle eingehen.

Glossar

p-Wert

Nachdem man einen statistischen Test zur Untersuchung einer Nullhypothese auf Basis der Daten verwendet hat, erhält man den Wert der Teststatistik. Der zugehörige p-Wert gibt an, wie «extrem» dieser Wert ist, und entspricht der Wahrscheinlichkeit, den errechneten oder einen noch extremeren Wert der Teststatistik zu erhalten, wenn in Wirklichkeit die Nullhypothese gilt.

R^2

Das (multiple) Bestimmtheitsmass R^2 gibt den Anteil der Variabilität in der Zielvariablen an, der durch das Modell (also durch die Einflussgrössen) erklärt werden kann. Je mehr Variabilität erklärt werden kann, umso besser wird die Zielvariable durch die Einflussgrössen angepasst.

Range

Mit dem englischen Begriff «*Range*» bezeichnet man die tatsächlich vorkommende Spannbreite der Daten von Minimum bis Maximum.

Statistische Signifikanz

Unterschiede zwischen Statistiken heissen signifikant, wenn die Wahrscheinlichkeit, dass sie durch Zufall zustande gekommen sind, also ohne dass ein tatsächlicher Unterschied vorliegt, sehr klein ist. Typischerweise möchte man erreichen, dass die Irrtumswahrscheinlichkeit kleiner als 5% oder 1% ist, und spricht in solchen Fällen von einem signifikanten Ergebnis. Der p-Wert (siehe oben) ist dann <5% oder <0,05 bzw. <1% oder <0,01.

Korrespondenz:
Dr. rer. nat. Ulrike Held
Horten Zentrum
UniversitätsSpital Zürich
Postfach Nord
CH-8091 Zürich
ulrike.held@usz.ch

Weiterführende Literatur

– Held L, Rufibach C, Seifert B. Einführung in die Biostatistik. 4. Auflage. Zürich: Abteilung Biostatistik, Institut für Sozial- und Präventivmedizin der Universität Zürich; Juli 2009. http://www.biostat.uzh.ch.
– Hüsler J, Zimmermann H. Statistische Prinzipien für medizinische Projekte. 4. Auflage. Bern: Huber-Verlag; 2006.
– Kreienbrock L, Schach S. Epidemiologische Methoden. 4. Auflage. München: Elsevier-Verlag; 2005.
– R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2008. ISBN 3-900051-07-0, URL http://www.R-project.org.