Risikofaktoren richtig einschätzen: Einfache lineare Regression (Teil 1)

Epidemiologen interessiert oft, wie ein Risikofaktor (unabhängige Variable, x) das Krankheitsrisiko (Outcome, abhängige Variable, y) beeinflusst. Mit Hilfe von linearer Regression kann aus Messwerten ermittelt werden, wie Risikofaktor und Krankheitsrisiko zusammenhängen – am Beispiel Strahlendosis und Krebsrisiko.

Roland Müller-Waldeck

In einer Studie, die den Einfluss einer Strahlendosis auf das Krebsrisiko untersucht, sind die Bedingungen im idealen Fall konstant gehalten, nur die Strahlendosis (als Prädiktor-Variable/ unabhängige Variable = x) variiert. Entsprechend wird das Krebsrisiko eines Studienteilnehmers nur noch durch die Strahlendosis beeinflusst.

Das Studienergebnis sind Wertepaare für Strahlendosis und Krebsrisiko für jeden Teilnehmer. Diese Wertepaare trägt man in ein Koordinatensystem ein, auf der x-Achse die Strahlendosis und auf der y-Achse das Krebsrisiko (Abb. 1, S. 48).

Den Epidemiologen interessiert besonders die mathematische Formel, welche die Beziehung zwischen Strahlendosis (unabhängige Variable) und Krebsrisiko (abhängige Variable) möglichst genau beschreibt, um in Zukunft aus der Strahlendosis das Krankheitsrisiko vorhersagen zu können. Um die Formel zu ermitteln, die den linearen Zusammenhang zwischen Strahlendosis und Krebsrisiko wiedergibt, wird durch die Punktewolke eine Gerade gelegt, die möglichst nah an allen Messpunkten liegt.

Die zu Grunde liegende mathematische Funktion dieser Geraden beschreibt den linearen Zusammenhang zwischen x und y. Man bezeichnet sie auch als ein Modell, das beschreibt, wie die unabhängige Variable die abhängige beeinflusst. Die Gerade, die den linearen Zusammenhang für die einzelnen Messpunkte am besten beschreibt, wird durch die Methode der kleinsten Quadrate errechnet. Im Prinzip wird mathematisch nach der Geraden gesucht, die dem y-Wert aller Messpunkte am nächsten liegt. Jede lineare Gleichung, also auch jede so ermittelte, lässt sich mit der mathematischen Gleichung beschreiben:

$$y = a + bx$$

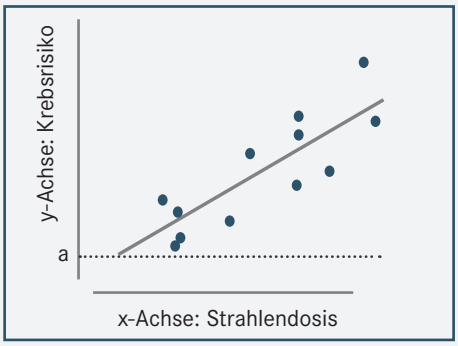
Aus dieser Gleichung lassen sich zwei Werte ablesen: Die Steigung b beschreibt, wie der Wert für das Krebsrisiko (y) steigt, wenn die Strahlendosis (x) um eine Einheit ansteigt. Der Wert des y-Achsenabschnitts a gibt an, wie hoch das Krebsrisiko ist, wenn die Strahlenbelastung 0 ist.

Weil Epidemiologen in der Regel wissen wollen, wie stark ein Faktor das Outcome beeinflusst, sind sie besonders an dem Wert für b interessiert.

Foto: Colourbox/Nataliiap

Das 95% Konfidenzband ist essenziell für die Beurteilung der ermittelten Funktion: Das Regressionsmodell schätzt den y-Wert für die gesamte Population anhand der Werte aus einer Stichprobe. Das 95% Konfidenzband schließt mit 95%iger Wahrscheinlichkeit die wahren Werte ein. Anders ausgedrückt liegt die Gerade, die die wahren Verhältnisse wiedergibt, mit 95%iger Wahrscheinlichkeit innerhalb des Konfidenzbandes (Abb. 2, S. 48). Schließt das Band die Null auf der y-Achse ein, ist die Beziehung zwischen x und y durch zufällige Schwankungen entstanden.

Der R^2-Wert gibt an, wie viel Prozent der Varianz der gemessenen y-Werte durch das lineare Regressionsmodell erklärt werden kann. Beträgt R^2 beispielsweise 59%, dann sind 41% der Varianz nicht durch das Regressionsmodell, sondern



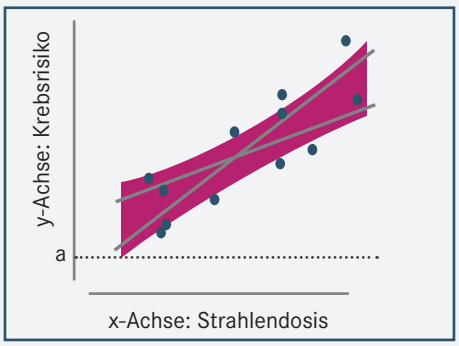


Abb. 1: Jeder Punkt gibt die Strahlendosis und das Krebsrisiko eines Studienteilnehmers wieder. Das Krebsrisiko steigt, wenn die Strahlendosis steigt. „a" gibt den Wert wieder, wie hoch das Krebsrisiko ohne Strahlenbelastung ist.

Abb. 2: Das rot unterlegte Konfidenzband weitet sich an den äußeren Enden. So können mehrere verschiedene Regressionsgeraden innerhalb des Konfidenzbandes liegen.

z.B. durch Messfehler, biologische Variation oder nichtlineare Beziehung zwischen x und y verursacht.

Der p-Wert für ein lineares Regressionsmodell beantwortet die folgende Frage: Wie wahrscheinlich ist es, dass die lineare Regression zu der ermittelten Steigung führt, wenn die Nullhypothese zutrifft? Die Nullhypothese lautet, dass es keinen linearen Zusammenhang gibt zwischen x und y (die Steigung des Regressionsmodells ist Null).

Voraussetzungen

Damit eine lineare Gleichung den Zusammenhang zwischen Outcome und unabhängier Variablen zutreffend wiedergeben kann, müssen einige Voraussetzungen gegeben sein:

- Der Zusammenhang zwischen x und y muss tatsächlich linear sein. Nichtlineare Zusammenhänge können nur mit nichtlinearer Regression beschrieben werden.
- Die Daten müssen nach der Gaußschen Normalverteilung streuen, sonst können Konfidenzintervalle oder p-Werte nicht interpretiert werden.
- Die Standardabweichung der Werte muss über den ganzen Messbereich gleich sein. Diese Bedingung ist verletzt, wenn die Punkte an einem Ende der Skala weiter von der errechneten Geraden entfernt liegen als am anderen Ende oder der Mitte.
- Die Messwerte beeinflussen sich nicht gegenseitig.
- Die Werte für x (unabhängige Variable) müssen genau bekannt sein und nur die Werte für y dürfen streuen.

Typische Fehler

- **Wenn der Wert für R^2 sehr niedrig ist,** bedeutet das nicht, dass es keine Korrelation zwischen x und y gibt, es bedeutet, dass die lineare Korrelation nur schwach ist. Eine nichtlineare Korrelation könnte stärker sein.
- Eine Voraussetzung für lineare Korrelation ist, dass jeder Punkt unabhängige Informationen beinhaltet. **Geglättete Daten** (smoothed data) erfüllen diese Voraussetzung nicht. Wenn solche Daten eingesetzt werden, führt die Regressionsgerade in die Irre, p-Wert und R^2 sind bedeutungslos.
- **Wenn x- und y-Wert nicht unabhängig voneinander sind,** sondern zusammenhängen, kann keine Regression durchgeführt werden. Das ist der Fall, wenn auf der x-Achse ein Ausgangswert (z.B. Baseline-Wert) und auf der y-Achse die Veränderung des Ausgangswerts aufgetragen werden.
- Die in der Studie kontrollierte Variable ist die unabhängige Variable (x), die gemessene Variable (outcome) ist y. **Werden x und y vertauscht,** kann keine Regression mehr durchgeführt werden.
- In biologischen Systemen ist die **Varianz** in der Regel proportional zu y und verletzt damit die Voraussetzung, dass die Varianz über alle Werte von y konstant ist. Statistische Programme können das mit einer unterschiedlichen Gewichtung der Messpunkte berücksichtigen.
- Das lineare Regressionsmodell gilt nur innerhalb der tatsächlichen Messpunkte und sollte nicht nach rechts oder links über die Messpunkte hinaus angewendet werden.
- **Ein niedriger p-Wert** sagt nur, dass die Regressionsgerade besser zu den Daten passt als eine waagerechte Gerade. Er beantwortet nicht die Frage, wie gut die Regressionsgerade zu den gemessenen Daten passt. Das kann mit dem R^2-Wert beurteilt werden. Daher kann es in die Irre führen, nur den p-Wert zu betrachten. ■

In der nächsten Ausgabe Februar 2020:
Teil 2: Multiple lineare Regression

Literatur: 1. Tripepi G et al. Linear and logistic regression analysis. Kidney International (2008) 73,806-810. **2.** Motulsky H. Intuitive Biostatistics. 2018 Oxford University Press