Risikofaktoren richtig einschätzen: Multiple lineare Regression (Teil 2)

Oft möchte man in Studien untersuchen, wie eine Kombination von mehreren Faktoren ein Outcome beeinflusst oder den Einfluss eines Faktors auf das Outcome (z.B. Confounder) rechnerisch eliminieren. Das leistet bei linearen Zusammenhängen die multiple lineare Regression.

Roland Müller-Waldeck

Die einfache lineare Regression findet eine mathematische Funktion, die den linearen Zusammenhang eines Risikofaktors (z.B. Strahlendosis) mit einem Outcome (z.B. Krebsrisiko) beschreibt. Die allgemeine Formel dafür lautet:

$$y = a + bx$$

Dabei ist a der Y-Achsenabschnitt und b die Steigung der Geraden (siehe Ausgabe 6/2019 des ärztlichen journals onkologie oder www.aerztliches-journal.de/medizin).

Mit multipler Regression kann man eine mathematische Funktion entwickeln, die beschreibt, wie mehrere Faktoren (z.B. Strahlendosis, Phenolbelastung, Acrylamidbelastung) gleichzeitig mit dem Outcome (z.B. Krebsrisiko) zusammenhängen. So kann man untersuchen, wie mehrere Faktoren gemeinsam das Outcome beeinflussen.

Mit Hilfe der multiplen linearen Regression kann man eine Formel entwickeln, mit der das Outcome aus allen untersuchten Risikofaktoren errechnet werden kann und mit der sich bestimmen lässt, wie stark jeder einzelne Faktor das Outcome beeinflusst. Das ist nötig, wenn man herausfinden möchte, ob ein möglicher Störfaktor (z.B. Confounder) das Outcome verändert, um zu ermitteln, wie einzelne mögliche prognostische Faktoren das Outcome beeinflussen und um einen prognostischen Index zu entwickeln, der das Outcome aus verschiedenen Variablen vorhersagen soll. Dabei können die Faktoren kontinuierlich (z.B. Gewicht), binär (z.B. Geschlecht) oder kategorial (mindestens 3 Kategorien, z.B. Herkunftsland) sein. Methoden, die zugleich mehrere Outcomes untersuchen (z.B. Lungenkrebsrisiko und Brustkrebsrisiko), werden multivariate Methoden genannt (z.B. MANOVA).

Die allgemeine Formel für multiple lineare Regression lautet:

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + b_ix_i + e$$

Dabei sind x_1 bis x_i die Werte für die Risiken 1, 2 und 3 bis i (z.B. Strahlendosis, Phenolbelastung, Acrylamidbelastung) und b_1 bis b_i beschreiben, wie stark jeder Faktor einzeln das Outcome verändert, wenn er um eine Einheit steigt (Steigung: siehe Ausgabe 6/2019 des ärztlichen journals onkologie oder www.aerztliches-journal.de/medizin).

Sehr grundsätzlich ausgedrückt ist b_ix_i das Teilrisiko, das der Risikofaktor i zu dem Gesamtrisiko y beiträgt. Das Gesamtrisiko y setzt sich im Wesentlichen aus der Summe der Teilrisiken zusammen; a (Y-Achsenabschnitt) ist dabei das Risiko, wenn alle untersuchten Faktoren Null sind. e ist eine Zufallsvariable, die der Gaußschen Normalverteilung folgt und die Streuung der Werte ausgleicht. Wie auch die einfache Regression passt die multiple Regression

Foto: Colourbox / #230685

die Werte für a und b_1 bis b_i so an, dass die Gerade möglichst nah an den Messwerten liegt.

Confounding kann nicht nur durch das Studiendesign, sondern auch in der Auswertung durch Stratifizierung und multiple Regression korrigiert werden. Dafür wird in dem Regressionsmodell berücksichtigt, wie der Confounder das Outcome beeinflusst und der Einfluss kann anschließend herausgerechnet werden:

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + b_{Confounder}x_{Confounder} + b_ix_i + e$$

Die Stichprobengröße für eine Studie mit multipler linearer Regression hängt davon ab, wie viele Faktoren gleichzeitig untersucht werden sollten: Als grobe Faustregel kann gelten, dass pro Kovariable 10 bis 40 Messergebnisse vorliegen sollten. Eine Studie, die 3 Kovariablen gleichzeitig untersucht, braucht also mindestens 30 bis 40 unabhängige Messergebnisse.

Der Computer kann für jeden untersuchten Faktor eines multiplen Regressionsmodells einen **p-Wert** errechnen, der die Nullhypothese testet. Die Nullhypothese lautet hier, dass der entsprechende Faktor keinen Einfluss auf das Regressionsmodell hat.

Die Qualität des Regressionsmodells

Auch bei multipler linearer Regression gibt der **R^2-Wert** für einen Faktor an, wie viel Prozent der Variabilität durch das Modell erklärt wird. Bei der einfachen linearen Regression kann man mit einem Blick auf ein Koordinatensystem mit den Messpunkten und dem Grafen für das Modell erkennen, wie gut das Modell zu den Messwerten passt.

Bei multipler Regression ist das nicht so einfach, weil schon zwei Faktoren zu einem dreidimensionalen Koordinatensystem führen und jeder weitere Faktor eine weitere Dimension erfordert. Trotzdem kann grafisch einfach gezeigt werden, wie gut ein Regressionsmodell zu den Messwerten passt. Dafür wird auf der X-Achse eines Koordinatensystems für jeden Studienteilnehmer das gemessene Outcome (Krebsrisiko) für den untersuchten Risikofaktor (z.B. Strahlungsrisiko) aufgetragen, auf der Y-Achse wird der mit dem Regressionsmodell errechnete Wert für das Outcome aufgetragen. Gibt das Regressionsmodell die tatsächlichen Verhältnisse perfekt wieder, entsteht eine gerade Punktereihe mit einer Steigung von 1.

Der R^2-Wert wird gern genutzt, um die Qualität eines multiplen linearen Regressionsmodells zu beurteilen. Besonders bei kleinen Stichproben ist seine Aussagekraft jedoch begrenzt, denn selbst, wenn die untersuchten Faktoren y nicht vorhersagen können, wird der Wert über 0 liegen. Je mehr Messpunkte hinzukommen, desto zuverlässiger wird der R^2-Wert.

Automatische Variablenselektion

Programme können automatisch Variablen (Risikofaktoren) identifizieren und solche aus der Auswertung ausschließen, die das Outcome nicht oder nur schwach beeinflussen. Dafür ermittelt der Computer, welches Regressionsmodell den Messergebnissen am besten entspricht. Bei einem Modell mit 2 Faktoren vergleicht er, welche der folgenden drei Gleichungen die Messergebnisse am besten beschreiben:

$$y = a + b_1x_1 + b_2x_2 + e$$
$$y = a + b_2x_2 + e$$
$$y = a + b_1x_1 + e$$
$$y = a + e$$

So entsteht ein vom Computer optimiertes Modell, dass „zu gut“ passt und möglicherweise die tatsächlichen Verhältnisse nicht mehr wiedergibt: Der R^2-Wert ist zu hoch, die Werte für die am besten passenden Variablen sind zu hoch, die Konfidenzintervalle sind zu eng und suggerieren höhere als die tatsächliche Präzision, der p-Wert ist zu niedrig.

Freedman (1983) simulierte eine Studie mit 100 Teilnehmern, indem er die Daten von 50 unabhängigen Variablen erfand, die vollkommen zufällig waren. Das Outcome konnte somit nicht von den Daten abhängig sein. Der Gesamt-p-Wert für die multiple Regression war hoch, wie auch die meisten p-Werte für die einzelnen unabhängigen Variablen. Freedman suchte die 15 Variablen aus, die die niedrigsten p-Werte hatten (<0,25) und führte eine multiple Regression mit diesen Variablen durch. Der Gesamt-p-Wert für diese Regression betrug 0,0005, die Verteilungen für 6 der 15 untersuchten Variablen war statistisch signifikant.

Obwohl die Werte vollkommen zufällig und nicht geeignet waren, das Outcome vorherzusagen, würde man aus dem niedrigen p-Wert schließen, dass die Variablen geeignet sind, das Outcome vorherzusagen. Das ist das Problem der multiplen Vergleiche.

Grundsätzlich ist es kein Fehler, mit Hilfe von statistischen Methoden ein oder wenige sorgfältig ausgewählte Variablen in die Regression ein- oder auszuschließen, aber es ist unsinnig, ein Statistikprogramm dutzende oder gar tausende von Modellen vergleichen zu lassen. Wenn in einer Veröffentlichung viele der ursprünglich für die multiple Regression erhobenen Variablen nicht berücksichtigt werden, sollten Sie die Ergebnisse sehr skeptisch betrachten. ■

Literatur: 1. Tripepi G et al. Linear an logistic regression analysis. Kidney International (2008) 73,806-810. 2. Harvey Motulsky. Intuitive Biostatistics. 2018 Oxford University Press. 3. Altmann, Practical Statistics for medical research, 1991 Chapman & Hall CRC.