Das Bestimmtheitsmaß der linearen Regression

Von der Vielzahl an Gütemaßen ist das Bestimmtheitsmaß oder R^2 das bekannteste. Es gibt an, wie gut die durch ein Regressionsmodell vorhergesagten Werte mit den tatsächlichen Beobachtungen übereinstimmen.

Interpretation des R^2 in der linearen Regression

Formal ist das Bestimmtheitsmaß der Anteil der Varianz der abhängigen Variable, der durch die unabhängige(n) Variable(n) erklärt wird. Es kann insofern Werte zwischen 0 und 1 annehmen.

Abbildung 1 zeigt verschiedene Konstellationen der Beobachtungen einer unabhängigen Variable X und einer abhängigen Variable Y. Die lineare Regressionsanalyse bestimmt in diesem einfachen Fall mit den Regressionskoeffizienten den Achsenabschnitt und die Steigung einer Geraden, die möglichst gut alle Beobachtungen widerspiegelt. Wie gut dies gelingt, beschreibt das R^2. Ist $R^2 = 1$, so liegen alle Beobachtungen genau auf der Regressionsgeraden. Zwischen X und Y besteht dann ein perfekter linearer Zusammenhang. Je kleiner R^2 ist, desto geringer ist der lineare Zusammenhang. Ein $R^2 = 0$ bedeutet, dass zwischen X und Y kein linearer Zusammenhang vorliegt. Die Regressionsgerade ist eine horizontale Linie, die die Y-Achse in Höhe des Mittelwertes der Beobachtungen der abhängigen Variable schneidet. Aus $R^2 \approx 0$ lässt sich jedoch nicht zwangsläufig folgern, dass gar kein Zusammenhang besteht. Er kann zum Beispiel quadratisch sein.

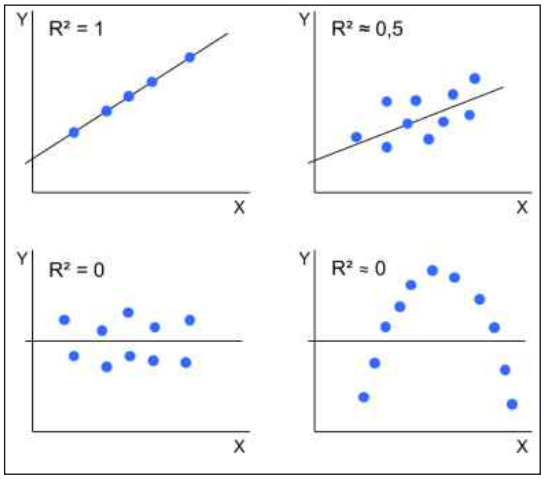

Abbildung 1: Beispiele geschätzter Regressionsgeraden

Beurteilung der Höhe des Bestimmtheitsmaßes

Grundsätzliche Empfehlungen, wie hoch das Bestimmtheitsmaß mindestens sein sollte, lassen sich nicht geben. Das R^2 hängt von der Höhe der Varianz ab, die überhaupt erklärbar, das heißt nicht durch den Zufall bedingt ist, und damit von der untersuchten Fragestellung.

Zudem tendiert das Bestimmtheitsmaß dazu, mit größerem Stichprobenumfang zu sinken. Dies lässt sich anhand des beispielhaften Streudiagramms in Abbildung 2 veranschaulichen. Gleichgültig ob (a) nur die drei roten oder (b) alle neun roten und blauen Beobachtungen zur Schätzung der Regressionskoeffizienten herangezogen werden, ergibt sich dieselbe dargestellte Regressionsfunktion. In (a) ist $R^2 = 0{,}79$, in (b) dagegen ist $R^2 = 0{,}56$. Je größer der Stichprobenumfang, desto eher gibt es zu demselben Wert der unabhängigen Variable bzw. derselben Kombination von Werten der unabhängigen Variablen unterschiedliche Werte der abhängigen Variable, so dass sich das R^2 verringert.

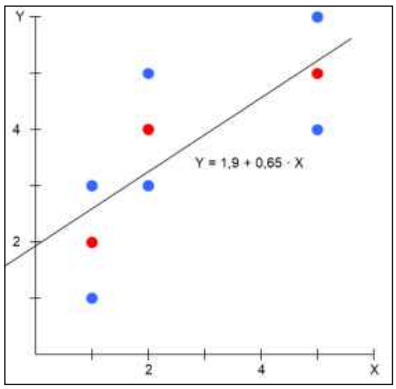

Abbildung 2: Stichprobenumfang und R^2

Nichtsdestoweniger machen Autoren aus der Marketingforschung Angaben zu Grenzwerten. Für Strukturgleichungsmodelle nennen Homburg/Baumgartner (1995) 0,4 oder Hermann et al. (2006) 0,3, wenn das Ziel die möglichst gute Erklärung der abhängigen Variable ist.

Aber selbst Regressionsanalysen mit geringem R^2 können wertvolle Infomationen liefern. Der Einfluss einzelner unabhängiger Variablen kann statistisch signifikant sein, das heißt es werden Variablen identifiziert, mit denen die abhängige Variable verändert werden kann.

Relative Wichtigkeit einzelner Variable

Die Höhe der geschätzten Regressionskoeffizienten hängt auch vom Skalenniveau der Variablen ab. Der standardisierte Regressionskoeffizient β dagegen gibt unbeeinflusst vom Skalenniveau die Stärke des linearen Zusammenhangs zwischen einer unabhängigen und der abhängigen Variable an. Im Fall einer einfachen Regression entspricht er dem Korrelationskoeffizienten r. Dann ist das Bestimmtheitsmaß $R^2 = \beta \cdot r = r^2$. Bei mehreren unabhängigen Variablen X_i ist $R^2 = \sum(\beta_i \cdot r_i)$. Demnach ist der Beitrag einer Variablen X_i zum R^2 gleich $\beta_i \cdot r_i$ und damit in Treiberanalysen ein Maß für die relative Wichtigkeit einer unabhängigen Variable für die abhängige Variable.

In Ausgabe 4/2019: Nicht-lineare Regression

Johannes Lüken, Diplom-Psychologe, ist Leiter des Bereichs Data Sciences bei IfaD.

jlueken@ifad.de

Prof. Dr. Heiko Schimmelpfennig, Diplom-Kaufmann, ist Projektleiter für Data Sciences bei IfaD.

hschimmelpfennig@ifad.de

Literatur

Hermann, A.; Huber, F.; Kressmann, F.: *Varianz- und kovarianzbasierte Strukturgleichungsmodelle.* In: zfbf, Nr. 1/2006, S. 34-66.

Homburg, C.; Baumgartner, H.: *Beurteilung von Kausalmodellen.* In: Marketing ZFP, Nr. 3/1995, S. 162-176.