(24) Modellbildung und Regression

Description

Unsere Modellierstudie „Einfluss des Fischkonsums auf die PISA-Werte von Schülern“ als Einführungsbeispiel

Ich werde demnächst unsere Modellierstudie vorstellen, die versucht, die Schwankung in den Todesraten der Covid-19 Mortalität in der ersten Welle bis zum Sommer 2020 vorzustellen. Um diese Methode zu illustrieren, verwende ich hier eine vor einiger Zeit von uns durchgeführte Studie [1]. Mein Kollege Volker Schmiedel, der sich sehr für die Bedeutung von Omega-3-Fettsäuren interessiert, gab den Anstoß zu dieser Studie. Wir stellten die einfache Frage:

Hat die Verfügbarkeit von Omega3-Fettsäuren in einem Land einen Einfluss auf den PISA-Wert von Kindern?

PISA (Programme for International Student Assessment) ist ja bekanntlich ein international durchgeführter, standardisierter Test, um Fähigkeiten von Kindern in der Schule zu untersuchen. Volker Schmiedel kam auf die Idee, den Fischkonsum eines Landes mit den PISA-Werten dieses Landes zu korrelieren und entdeckte eine signifikante Korrelation. Die einfache Korrelation zwischen Fischkonsum und PISA-Wert eines Landes beträgt r = .57 und ist damit nicht nur signifikant, sondern auch ziemlich hoch. Eigentlich sogar erstaunlich hoch. Denn warum sollte Fischkonsum mit den Kenntnissen von Schülern in der Schule zusammenhängen? Der Zusammenhang könnte eben durch die Omega-3-Fettsäuren verstehbar sein, die vor allem in fettreichem Fisch enthalten sind, aber auch in dunkelgrünen Pflanzen, Algen und allem, was sich davon ernährt. Es ist nicht leicht, die Omega-3-Werte in einer Bevölkerung zu erfassen. Man müsste dazu von einer repräsentativen Bevölkerungsstichprobe Blut entnehmen und den Omega-3-Gehalt z.B. in den Membranen der roten Blutkörperchen bestimmen. Das hat meines Wissens systematisch noch niemand über alle möglichen Länder hinweg gemacht. Da ist Fischkonsum leichter zu erfassen, eine sog. Proxy- oder Stellvertretervariable. Denn Fisch ist ein Hauptlieferant von Omega-3-Fettsäuren. Und Omgea-3-Fettsäuren sind als essenzielle Fettsäuren für uns wichtig. Wir müssen sie durch die Nahrung zu uns nehmen, weil wir sie nicht selber bilden können. Seit der industriellen Revolution Ende des 18. Jahrhunderts hat die Omega-3 Zufuhr abgenommen [2]. Omega-3 ist nicht nur zentral für das Immunsystem, weil es die Vorläufersubstanz für alle Zytokine mit entzündungshemmender Wirkung ist. Es ist vor allem wichtig für das Nervenwachstum bei Kindern und das Lernen im Alter. Es ist auch wichtig zum Erhalt von kognitiver Leistung. Beispielsweise kann der Omega-3 Gehalt in der Muttermilch die Intelligenz von Schulkindern zu einem erstaunlichen Grad vorhersagen [3, 4].

Aus all diesen Gründen war Schmiedels Überlegung natürlich sehr klug: möglicherweise hängt der PISA-Wert, als Ausdruck des kognitiven Leistungsniveaus von Kindern, ja neben anderem auch damit zusammen, wie viel Omega-3-Fettsäuren sie zu sich nehmen, grob gemessen am Fischkonsum einer Nation. Nun stellt sich aber natürlich sofort die Frage: Was beeinflusst denn den PISA-Wert vor allem? Und wenn wir das kennen, spielt dann der Fischkonsum zusätzlich dazu noch eine Rolle?

Das allgemeine Prinzip: Linearkombination von gewichtetem Einfluss von Variablen

Die allgemeine mathematische Schätzformel für eine solche Fragestellung lautet:

$y = a + ?_1x_1 + ?_2x_2 + ?_3x_3 + \ldots ?_nx_n + e$. (Gleichung 1)

“y” ist dabei, ganz allgemein, die Variable, die man klären will, also z.B. die Schwankung in den PISA-Werten von Schülern, oder in den Covid-19 Todesfällen in Europa.

„a“ ist eine Konstante, oder das sog. Interzept. Grafisch dargestellt wäre es der Punkt, an dem eine Regressionslinie die x-Achse schneidet und damit den empirischen Nullpunkt angibt. Man benötigt diesen Wert, wenn man konkrete Rechnungen für einzelne Individuen anstellen will, oder wenn man eine gefundene Regressionsgleichung in der Zukunft oder bei einem anderen Datensatz zum Ausrechnen von Werten verwenden will. Im Moment ist dieser Wert zum Verstehen des allgemeinen Prinzips der Aufklärung nicht so wichtig.

Die sog. „?“-Gewichte sind die Regressionsgewichte oder Regressionskoeffizienten. Wenn sie standardisiert sind, also eine Verteilung zwischen -1 und +1 annehmen können, werden sie meistens mit dem griechischen ?-Symbol wiedergegeben. Wenn sie unstandardisiert sind, dann wird meistens b notiert. Sie geben an, wie groß der Einfluss einer Variablen x auf das Kriterium y ist. Wäre z.B. ein Regressionsgewicht nur 0,0001, dann wäre der Einfluss der Variablen mit diesem Gewicht nachvollziehbarer Weise sehr gering. Ist ? sehr groß, z.B. 0,8, dann ist auch der Einfluss dieser Variablen relativ groß. Ist das Gewicht positiv, dann hat die Variable einen positiven Einfluss: also, je größer x, umso größer y. Ist das Gewicht negativ, dann hat die Variable einen negativen Einfluss: also, je größer x, umso kleiner y.

Nun sieht man an der Gleichung (1) unmittelbar, dass es sich dabei um eine lineare Kombination von Variablen x_1 bis x_n handelt, also im Grunde um beliebig viele Variablen oder Prädiktoren, die man zur Erklärung von y, dem Kriterium, heranziehen kann. Das ist eben der Charme der Modellbildung: Man kann so viele Variablen wie man erheben kann zur Erklärung heranziehen. Eine Grenze ist vor allem praktisch gegeben: Da man diese Regressionsgewichte nicht einfach auf der grünen Wiese findet, sondern durch ein rechenaufwändiges Iterationsverfahren schätzen muss, benötigt man dazu entsprechend viele Datensätze, um diese Schätzung stabil vornehmen zu können.

Technisch gesprochen wird dazu meistens das Verfahren der kleinsten Quadrate herangezogen: Der Computer bildet Mittelwerte der einzelnen Variablen, setzt dann unterschiedliche Regressionsgewichte ein, während alles andere konstant gehalten wird, quadriert die Differenz zwischen Mittelwert und gewichtetem Mittelwert und macht das iterativ so lange, bis die Differenz ein Minimum beträgt. Als man solche Verfahren noch von Hand rechnen musste, war das sehr zeitaufwändig und begrenzte schon allein deswegen die Anzahl der Variablen. Heute können Computer das in Bruchteilen von Sekunden. Aber man muss sich trotzdem der Tatsache bewusst sein, dass auch der Computer nur mit dem rechnet, was vorhanden ist. Und um eine Schätzung stabil durchführen zu können, benötigt der Computer – Faustregel – etwa 10 Fälle pro zu schätzender Einflussvariable bzw. dem zugehörigen Regressionsgewicht ? [5].

Ganz zum Schluss von Gleichung (1) sehen wir dann noch das „e“, manchmal auch als griechisches Epsilon – e – dargestellt. Das ist statistische Universalsprache für „Fehlerterm“ oder „Residuum“. Das ist der Anteil der Schwankung, der nicht von diesen Variablen aufgeklärt werden kann.

Dieses allgemeine Prinzip der linearen Kombination von gewichteten Einflussvariablen zur „Vorhersage“, also zur Erklärung, eines individuellen Wertes gilt für alle Modellbildungen. Bei manchen Regressionsverfahren ist die Verbindung der einzelnen Vorhersage-Terme komplizierter. Bei nicht-linearen Regressionen etwa kommen eben quadratische, kubische oder andere Funktionsterme vor. Bei logistischen Regressionen sind diese

Regressionselemente Exponenten der Eulerschen Zahl e. Aber das Prinzip ist immer das Gleiche: Eine Reihe von Variablen wird dazu verwendet, um in einer optimalen Kombination eine zu erklärende Variable, das Kriterium oder die abhängige Variable, „vorherzusagen“, das heißt in ihrer Schwankungsbreite oder Varianz so weit als möglich aufzuklären.

Konkret am Beispiel der PISA-Studie

Wir haben uns ans Werk gemacht und PISA-Werte von 64 Ländern gesammelt, von denen wir auch Informationen zum Fischkonsum hatten. Zusätzlich zogen wir Daten zur wirtschaftlichen Entwicklung heran, in diesem Fall das Bruttoinlandsprodukt (weil dieses ja indirekt auch bestimmt, wie viel Mittel einem Land zur Verfügung stehen), Daten zur Verfügbarkeit des Internets in einem Land, als Indikator für die technische Entwicklung, und die Stillquote. All diese Daten sind aus öffentlichen Quellen erhältlich und sind intuitiv und theoretisch plausible Einflussfaktoren, deren Einfluss auf den PISA-Wert eines Landes abzuschätzen ist.

Man erkennt an dieser Stelle: Es hängt durchaus auch von der Fragestellung ab, welche Variablen man in ein solches Modell einspeist. Dies ist wiederum abhängig von theoretischer Kenntnis und von konzeptuellen Vorannahmen und nicht selten, wie in unserem Fall, auch von der Verfügbarkeit von Daten.

Die einzelnen Einheiten, also Fälle, sind übrigens in dieser Studie nicht einzelne Kinder, sondern Länder mit ihren PISA-Durchschnittswerten. Meistens sind in solchen Studien einzelne Personen die Analyseeinheit. In der PISA-Studie und auch in unserer Covid-19-Modellierung sind Länder die Analyseeinheiten oder „Fälle“.

Wir haben nun ein lineares Regressionsmodell gerechnet, wie oben beschrieben. Ich gebe die originale Tabelle III der Publikation hier als Tabelle 1 wieder und erläutere sie dann:

Table III. Results of regression analysis – dependent variable: PISA Mean Score Adjusted R^2 = .72; p < 0.0001; significant predictors in italics.

	Parameter	Std.Err	*t*-value	*p*-value	ß-weights	–95.00%-Cnf.Lmt	+95.00%-Cnf.Lmt
Intercept	117.1	44.3	2.64	0.01			
GDP	5.65	8.0	0.70	0.5	0.10	–0.18	0.37
Internet coverage	*62.3*	*12.6*	*4.93*	*>0.0001*	*0.65*	*0.38*	*0.91*
Breastfeeding	0.1	0.3	0.35	0.73	0.03	–0.13	0.18
Fish consumption	*9.8*	*4.3*	*2.28*	*0.03*	*0.20*	*0.02*	*0.38*

GDP: gross domestic product in million USD; log-transformed.
Internet coverage in percent; log-transformed.
Breastfeeding: Exclusive breastfeeding for the first 3–6 months in percent.
Fish consumption: 1: 2–5 kg fish per year and person; 2: 5–10 kg fish; 3: 10–20 kg; 4: 20–30 kg; 5: 30–60 kg; 6 = < 60 kg.

Tabelle 1 – Die Tabelle III der Originalpublikation mit den Modellparametern der Regressionsanalyse

Man erkennt: Wir haben fünf Variablen für die Vorhersage verwendet, das Bruttosozialprodukt (GDP-Gross Domestic Product), die Internetabdeckung eines Landes, den Prozentsatz der Kinder in einem Land, die gestillt wurden, und am Schluss den Fischkonsum eines Landes, grob gemessen in 6 einigermaßen kontinuierlich steigenden Kategorien (2-5 kg pro Person und Jahr, 5-10 kg, 10-20 kg, 20-30 kg, 30-60 kg und mehr als 60 kg).

Letzteres ist deswegen wichtig, weil lineare Regressionsmodelle verschiedene Voraussetzungen haben. Eine davon ist, dass die Kriteriumsvariablen und alle anderen Variablen einigermaßen normalverteilt sein müssen und dass die Variablen, die man zur Vorhersage verwendet, kontinuierliche Variablen sein müssen. Wenn sie nicht kontinuierlich, sondern kategorial sind, dann muss man sie umkodieren in sog. Dummy-Variablen, also 1-0-

Kodierungen (oder -1 und +1) für einzelne Kategorien, die dann wieder kontinuierlich sind. In unserem Falle habe ich die Fischkonsum-Variable sowohl als kontinuierliche Variable als auch als Dummy-kodierte Variable für die einzelnen Kategorien verwendet. Der Unterschied ist vernachlässigbar. Daher berichte ich in der Publikation das Modell für die kontinuierliche Variable und diskutiere das mögliche Problem in der Diskussion, weil ein Gutachter darauf bestanden hatte.

Wir sehen: Das Modell ist hochsignifikant und kann mit $R^2 = .72$ sogar 72 % der Varianz aufklären. Diese Modell-Statistik ist die erste wichtige Erkenntnis. Sie sagt uns, ob das statistische Modell erstens signifikant ist und zweitens wie hoch die multiple Korrelation R, also die Korrelation aller Variablen gemeinsam mit dem Kriterium ist. Quadriert ergibt jeder Korrelationskoeffizient die aufgeklärte Varianz. Beispiel: die Intelligenz eines Menschen sei mit seinem späteren Einkommen etwa r = .3 korreliert – was übrigens in etwa den empirischen Verhältnissen entspricht; dann wäre die dadurch aufgeklärte Varianz $r^2 = .3^2 = .09$ oder 9 %.

In unserem Fall ist $R^2 = .72$ (der multiple Korrelationskoeffizient, der den Einfluss mehrere Variablen gleichzeitig beschreibt, wird immer großgeschrieben). Die Varianzaufklärung mit 72 % ist erheblich. Denn man benötigt dazu nur 2 Variablen: die Internetabdeckung, die als Stellvertreter für die wirtschaftlich-technische Entwicklung eines Landes steht und den Fischkonsum. Diese detailliertere Einsicht ist die zweite wichtige Erkenntnis, die eine statistische Modellierung liefert. Sie sagt uns, welche Variablen, die wir in unserer Modellierung verwenden, wie stark zu dieser Varianzaufklärung beitragen.

Man sieht an der Tabelle 1 oben, dass das Beta-Gewicht für die Internetabdeckung mit 0.65 recht hoch ist. Diese Variable ist auch hoch signifikant, während das Bruttosozialprodukt als Prädiktor irrelevant bleibt. Das liegt daran, dass Internetabdeckung und Bruttosozialprodukt mit r = .87 sehr hoch untereinander korrelieren (das ist in Tabelle 2 der Publikation erläutert) und das Modell in diesem Fall die Variable verwendet, die ein besserer Prädiktor ist. Dadurch fällt die andere automatisch aus der Gleichung. Ich habe auch Analysen nur mit Bruttosozialprodukt gerechnet. Diese haben aber leicht geringere Varianzaufklärung.

Nun wäre die analytische Idee dieser Analyse: Wenn der PISA-Wert eines Landes durch diese sozialen Variablen (GDP, Internetabdeckung, Stillquote) erklärbar ist, dann dürfte der Fischkonsum als Prädiktor irrelevant sein. Was wir aber sehen, ist: Die Stillquote spielt kaum eine Rolle. Das Beta-Gewicht ist mit 0.03 sehr klein und nicht signifikant. Aber der Fischkonsum ist mit beta = .20 ein signifikanter Prädiktor.

Man kann in solchen Analysen nämlich quasi-experimentell vorgehen und z.B. die Frage stellen: Wenn man alle sozialen Variablen kontrolliert, ist dann der Fischkonsum immer noch ein signifikanter Prädiktor? In einem solchen Fall, geht man schrittweise vor bzw. forciert das System, zuerst die sozialen Variablen einzuschließen und danach, an letzter Stelle, oder auch im letzten Schritt, die interessierende Variable. Das ist hier der Fischkonsum. Das habe ich hier so gemacht und man sieht: Auch wenn man alle anderen Variablen vorher einschließt, dann ist der Fischkonsum immer noch ein signifikanter Prädiktor. Er klärt zusätzliche 4 % der Varianz auf. Ein Modell ohne den Prädiktor „Fischkonsum“ hätte also nur ein $R^2 = .68$.

Immer noch hoch, aber niedriger. Das erlaubt uns den Schluss: Wenn man den sozial-ökonomischen Fortschritt in Rechnung stellt, dann ist der Fischkonsum, und damit vermutlich Omega-3-Verfügbarkeit, ein zusätzlicher, wichtiger Prädiktor. Die Tatsache, dass man mit diesen Variablen gemeinsam 72 % der Varianz aufklären kann, ist aus meiner Sicht erstaunlich. Natürlich spielen auch noch andere Faktoren eine Rolle: wie gut das Schulsystem ist, wie gut die Lehrerausbildung, wie motiviert die Lehrer, wie groß die Klassen, wie lange Kinder schlafen, etc. Aber all das haben wir nicht erfasst bzw. hatten dazu keine Daten. Wir hatten Daten zur Schulzufriedenheit von einigen Ländern und haben für diese Länder die Analyse mit Schulzufriedenheit wiederholt. Das Bild änderte sich aber nicht und Schulzufriedenheit war kein signifikanter Prädiktor.

Ich habe an erster Stelle in Tabelle 1 oben die Parameter oder die rohen Regressionsgewichte angegeben. Diese sind nicht standardisiert und geben Auskunft darüber, wie stark eine Variable in einer aktuellen Vorhersagerechnung zu gewichten wäre.

Dann folgt der Standardfehler dieser Schätzung. Dieser wird zur Signifikanzberechnung benötigt, die freundlicherweise das Statistikprogramm mitliefert. Die Verteilung dieser Kennwerte folgt der T-Verteilung, einer statistischen Verteilung, die so ähnlich ist wie die Normalverteilung, nur steiler in Abhängigkeit von der Beobachtungsanzahl. Aus ihr kann man die Irrtumswahrscheinlichkeit p gewinnen. Sie sagt uns, ob ein Regressionsgewicht signifikanten, also statistisch überzufälligen Einfluss, hat. Es kann durchaus vorkommen, dass ein relativ großes Regressionsgewicht nicht signifikant ist und umgekehrt ein sehr kleines signifikant. Das bedeutet dann: Der Einfluss ist vorhanden, aber statistisch schwer von einer Zufallsschwankung zu unterscheiden. Oder: Der Einfluss ist sehr klein, aber deutlich überzufällig.

Die standardisierten Beta-Gewichte, die dann in der nächsten Spalte folgen, kann man interpretieren als Partialkorrelationskoeffizienten. Sie stellen die Korrelation der entsprechenden Variablen mit dem Kriterium dar, also in dem Fall mit dem PISA-Wert eines Landes, wenn man den Einfluss aller anderen Variablen statistisch konstant hält oder herausrechnet. (Denn die Variablen haben ja auch Korrelationen untereinander, die dann kontrolliert werden.)

Man kann sich das Prinzip auch grafisch in einem sog. Venn-Diagramm verdeutlichen, das ich hier in Abb. 1 wiedergebe.

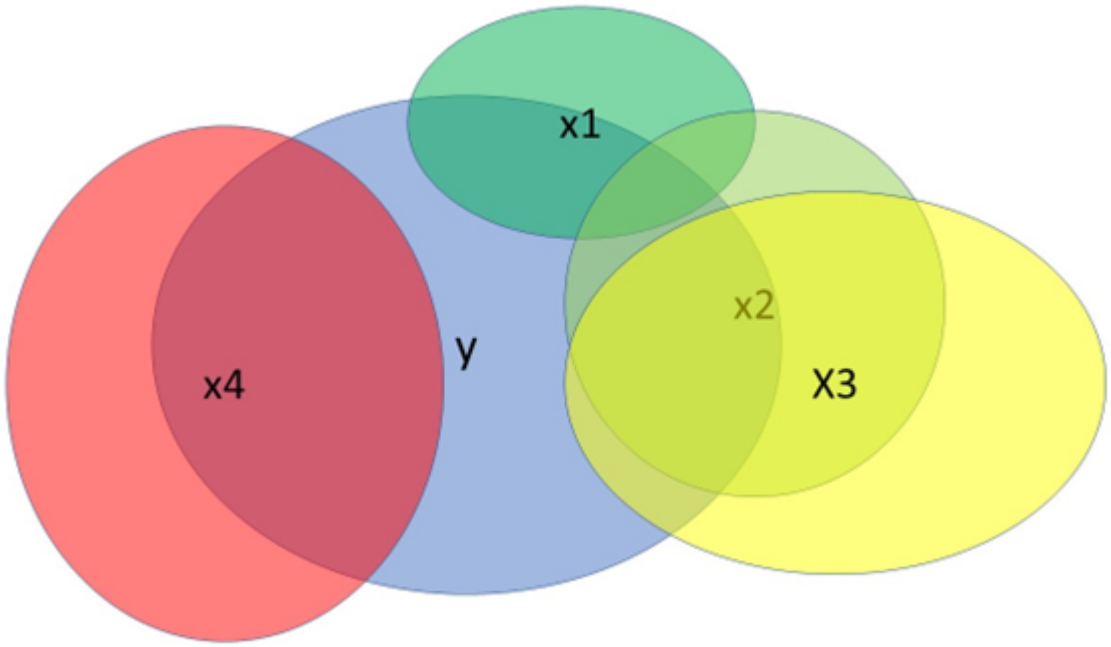

Abb. 1 – Venn Diagramm der Zusammenhänge von verschiedenen Prädiktoren x_1-x_4 mit einer aufzuklärenden Variablen y

Der blaue Kreis y stellt unsere Zielvariable, das Kriterium, dar. Die Variablen x_1 bis x_4 sind mögliche Prädiktoren. Sie haben eine bestimmte Korrelation mit dem Kriterium – der Überschneidungsbereich – und oft auch eine Korrelation mit anderen Variablen. Beispielsweise wäre der eigene Beitrag von x_2 jener Bereich, der weder von x_1 noch von x_3 abgedeckt ist, relativ gering. Auch der eigene Beitrag von x_3 ist nicht so hoch, wie es

zunächst scheint, weil nämlich der Zusammenhang mit x_2 sehr hoch ist. Das nennt man „Kollinearität“, ein gemeinsamer hoher Zusammenhang. Intelligente Modellbildung prüft dies und verwendet von 2 möglichen Variablen die, mit dem höchsten eigenen Aufklärungswert. In unserer Analyse war dies die Internetabdeckung. Variable x_4 hingegen hätte in diesem grafischen Beispiel einen relativ hohen Aufklärungswert und eine eigene, unabhängige Korrelation mit dem Kriterium y, ohne mit den anderen Variablen zusammenzuhängen. Das reine Blau von y, das nicht von anderen, überlappenden Kreisen bedeckt ist, das wäre der Anteil nicht erklärter Varianz bzw. im individuellen Fall, die Residuen.

Um Residuen zu verstehen, ist es nützlich, eine konkrete Regressionsgleichung durchzurechnen. Wir tun das für die Beispiele China und Qatar aus unserem Datensatz:

China, der Ausreißer nach oben in Abb. 3 unten, hat mit 567,66 den höchsten PISA-Wert in unserem Datensatz und Qatar mit 308 den niedrigsten. Die Stillquoten sind ähnlich, auch die Internetabdeckung, in Qatar 87 %, in China 74 %, aber das Bruttoinlandsprodukt unterscheidet sich stark und liegt für Qatar bei 100.260 Millionen USD und in China bei 6.747 Millionen USD (Daten aus 2013). Nun sieht man an Tabelle 2: Der Wert für das Bruttosozialprodukt und für die Internetabdeckung habe ich log-transformiert, weil die Daten zu schief verteilt waren und habe damit eine annähernde Normalverteilung erreicht. Der Fischkonsum ist eine 6-stufige, annähernd kontinuierliche Variable.

Land	PISA Wert	Fischkonsum	Stillquote	GDP transf.	Intern. transf.
China	567,66	5	28 %	8,816	4,304
Qatar	398	4	29 %	11,515	4,466

Tabelle 2 – Originale Daten für 2 Länder aus unserer PISA-Studie

Wir verwenden nun Gleichung (1) und die Daten aus Tabelle 1, die die originalen Regressionsgewichte angeben:

$$y_{China} = 117,1 + 5,65*8,816 + 62,3*4,304 + 0,1*28 + 9,8*5 + e =$$

$$117,1 + 49,81 + 268,14 + 2,8 + 49 + e =$$

$$486,85 + e$$

$$y_{China} - 486,85 = e$$

$$567,66 - 486,85 = e$$

$$e = 80,81$$

Die Regressionsgleichung für China ergibt also einen um 80,81 Punkte niedrigeren PISA-Wert, als er in Wirklichkeit ist. Das ist in Abb. 3 unten der Ausreißer nach oben, der ziemlich genau bei 80 Punkten liegt, bzw. im Histogramm in Abb. 2 der Wert ganz rechts außen in der Verteilung.

Wer will, kann das Gleiche nun mit den Daten für Qatar tun und wird finden, dass die Gleichung einen negativen Fehler oder ein Residuum von etwa -100 Punkten ergibt, d.h. die PISA-Werte von Qatar werden durch die Gleichung um etwa 100 Punkte höher eingeschätzt, als sie in Wirklichkeit sind. („Wirklichkeit“ heißt hier: empirische Wirklichkeit.)

Es wäre nun eine Frage der differenzierteren Analyse, warum das bei diesen Ausreißern so ist. Es könnte etwa sein, dass chinesische Daten unzuverlässig sind. Dass das Schulsystem wesentlich besser ist, etc.

Jedenfalls sieht man auf diese Weise: Regressionsgleichungen können zur individuellen Vorhersage, etwa neuer Datensätze, verwendet werden, was in der Industrie oft in der Prozesskontrolle verwendet wird. Und auf diese Weise versteht man auch die Funktion und arithmetische Größenordnung von Fehlertermen oder Residuen e. Sie stellen im individuellen Fall den Fehler, im Fall eines Gesamtdatensatzes die unerklärte Varianz dar.

Voraussetzungen beachten

Nun ist bei einer solchen Analyse zu berücksichtigen, dass sie nur dann gültige Analyseergebnisse liefert, wenn die Voraussetzungen gegeben sind. Ich erwähnte schon zwei, die man vor der Analyse prüfen muss: Sind die Variablen einigermaßen normalverteilt? Waren sie in unserem Fall. Ich sage „einigermaßen“, weil die Routinen gegenüber einer Verletzung dieser Annahme relativ robust reagieren. Wenn die Normalverteilung, vor allem der Kriteriumsvariable stark verletzt ist, kann man einen Trick anwenden und sie logarithmisch transformieren. Dann wird sie oft normalverteilt. Das Gleiche kann man mit den anderen Variablen tun.

Außerdem wirft man einen Blick auf die Residuen, also die unerklärten Anteile, jene 28 % der Varianz, in unserem Fall, der nicht durch diese Variablen erklärbar ist. Sie müssen nämlich einigermaßen normalverteilt sind. Publikationen zeigen das oft grafisch in den Anhängen. Ebenso sollte ein Plot der Residuen gegen die vorhergesagten Werte kein Muster erkennen lassen. Denn sind Muster erkennbar, ist die Annahme wahrscheinlich, dass der Zusammenhang nichtlinear ist.

Ich gebe hier in den Abbildungen 2 und 3 das Histogramm der Residuen und den Plot der Residuen vs. den vorhergesagten Werten wieder:

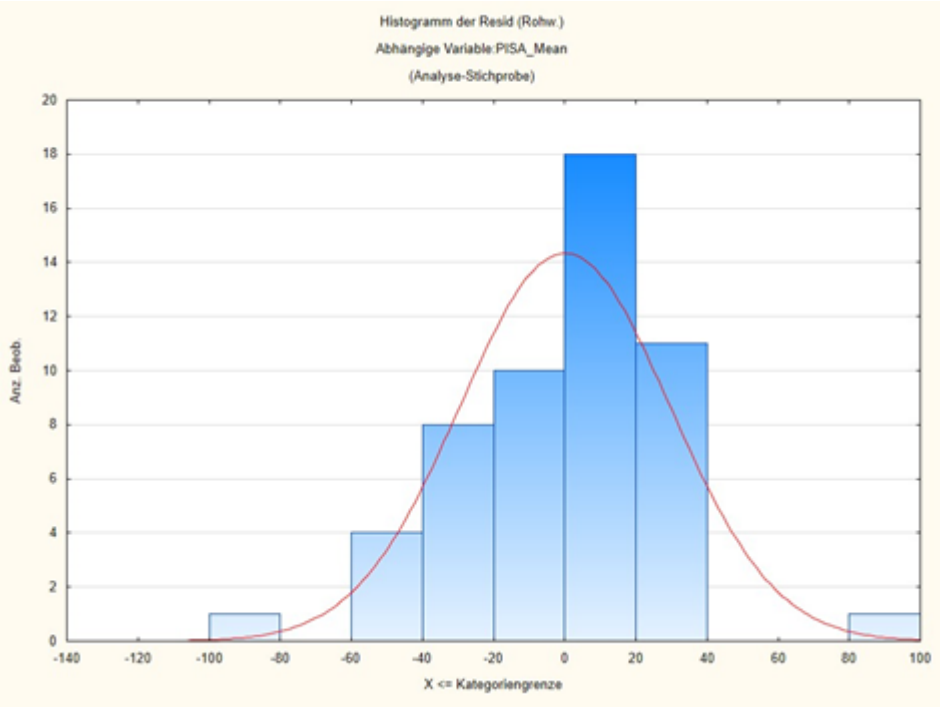

Abb. 2 – Histogramm der Residuen

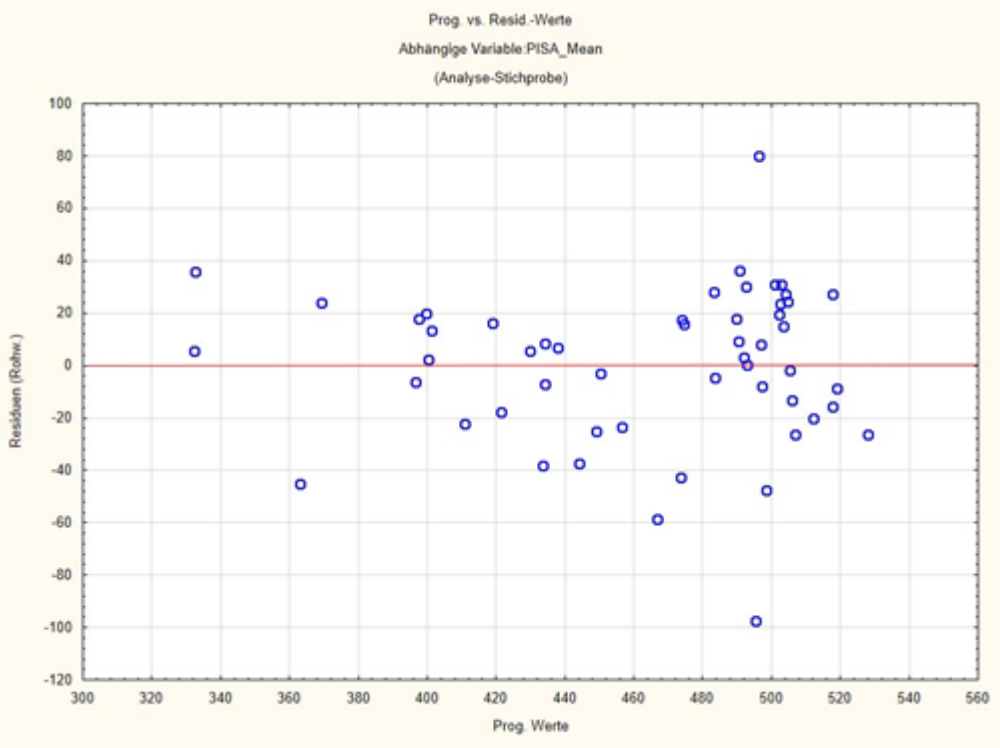

Abb. 3 – Plot der Residuen vs. vorhergesagte Werte

Man erkennt an Abb. 2: die Residuen sind einigermaßen normalverteilt um 0 herum. Es gibt einige Ausreißer, bei denen der vorhergesagte PISA-Wert beinahe 100 Punkte zu hoch oder zu niedrig ist. Aber ansonsten passt das Modell recht gut. Diese Ausreißer erkennt man auch in Abb. 3. Man kann sich mit Statistikprogrammen auch die Ausreißer ansehen, in unserem Falle ist der Ausreißer nach unten Qatar und der Ausreißer nach oben China. Aber ansonsten ist in diesem Plot kein Muster erkennbar. Ein Muster wäre etwa eine kontinuierlich nach einer Seite ansteigende Wolke.

Die analytischen Konzepte von linearen Modellen

Lineare Modelle können also verschiedene Zwecke erfüllen:

1. Sie dienen dazu, die Bedeutung möglicher Prädiktorvariablen oder unabhängigen Variablen und damit deren Einfluss auf das Kriterium oder die unabhängige Variable abzuschätzen. Man kann dies etwa in klinischen Studien und Experimenten auch dazu nutzen, den Einfluss einer experimentellen Manipulation zu erkennen. Diese wird dann durch eine kategoriale Dummy-Variable abgebildet, die 1/0-kodiert ist. Der Einfluss einer Variable zeigt sich an der Größe (und natürlich auch der Signifikanz) der Regressionsgewichte. Bei standardisierten Regressionsgewichten, die mit ? bezeichnet werden, kann man das unmittelbar tun. Denn die Regressionsgewichte können als partielle Korrelationskoeffizienten interpretiert werden, also als Zusammenhang der Prädiktorvariablen mit der Kriteriumsvariable, wenn die Einflüsse aller anderen Variablen statistisch kontrolliert werden. In unserem Beispiel: Fischkonsum in einem Land korreliert mit dem PISA-Wert des Landes (und umgekehrt) mit 0.20, wenn all die anderen Variablen in der Gleichung statistisch kontrolliert sind. D.h., wenn deren Einfluss auf den Fischkonsum herausgerechnet wurde. Man kann also die Größe von ? als Schätzer für den Einfluss einer Variablen verwenden. Im Bild von Abb. 1: Es sind die Überlappungen eines Kreises mit dem y-Kreis ohne den Anteil anderer überlappender Kreise. Wenn es sich, wie bei anderen Regressionsmodellen oft der Fall ist, nicht

um standardisierte Gewichte handelt, dann kann man sich an der relativen Größe orientieren, also an der Größe relativ zu allen anderen Regressionsgewichten.

2. Man kann eine Regressionsgleichung dazu verwenden, für einzelne Fälle Vorhersagen zu machen. Das wird vor allem in der Prozesskontrolle verwendet, wenn man aus standardisierten Datensätzen eine Regressionsgleichung ermittelt hat, die man dann auf neue Datensätze anwenden kann. Für die analytische Forschung ist das eher weniger von Bedeutung. Ich habe diesen Ansatz oben verwendet, um klarzumachen, welche Rolle Residuen spielen.
3. Wenn man die gesamte Gleichung über alle Datensätze löst und das statistische Modell als Ganzes abschätzt, dann erkennt man, wie gut das Modell insgesamt auf die Daten passt. Wir sahen an dem Modell der Aufklärung des PISA-Wertes, dass eine relativ hohe Varianzaufklärung mit diesem Modell möglich ist. Dieser analytische Schritt wird als „Anpassungsgüte“, oder „Modellgüte“, oder Vorhersagekraft des Modells bezeichnet. Sie hat vor allem zwei Komponenten: einen R^2-Wert und F- oder Chi^2-Wert mit einem assoziierten p-Wert oder einer Irrtumswahrscheinlichkeit. Der R^2-Wert ist der quadrierte multiple Korrelationskoeffizient, also die Korrelation aller in der Gleichung verwendeten Variablen gemeinsam mit dem Kriterium oder der abhängigen Variablen. Er wird quadriert, weil ein quadrierter Korrelationskoeffizient interpretiert werden kann als der Anteil an erklärter Varianz oder aufgeklärter Schwankung. Der multiple Korrelationskoeffizient R^2 erklärt also, wie viel Varianz oder Schwankungsbreite, z.B. in den PISA-Werten einzelner Länder, wir mit den vorgegeben Variablen erklären können, in unserem Beispiel eben 72 % der Variation in den PISA-Werten. Die Tatsache, dass wir nicht alle Variablen kennen und erfasst haben, die einen möglichen Einfluss haben, drückt sich eben in der nicht erklärten Varianz aus und auf individueller Ebene in den Fehlertermen oder Residuen e.
   Dieser R^2-Wert verteilt sich je nach Modell entsprechend der F- oder der Chi-Quadrat-Verteilung. Diese Verteilungen kennt man. Daher man auch sie normieren. Dann kann man die Fläche unter der Kurve als „1“ definieren und damit als Wahrscheinlichkeit. Dann kann man die Fläche ab einer bestimmten Ordinate ebenfalls als Wahrscheinlichkeit definieren, und wenn ein bestimmter Wert eine Grenze überschreitet bzw. die Fläche rechts davon sehr klein ist, dann ist die Wahrscheinlichkeit eines solchen Wertes sehr klein. Dies lässt sich dann zum Bestimmen der Irrtumswahrscheinlichkeit eines empirisch gefundenen R^2-Wertes verwenden.
   Das Gesamtmodell hat also zwei wichtige Kennziffern: den R^2-Wert, die Größe der Varianzaufklärung, und die Signifikanz oder statistische Irrtumswahrscheinlichkeit dieses Wertes.
   Es hängt nämlich von der Größe des Zusammenhanges, aber auch von der Größe des Datensatzes ab, ob ein multipler Korrelationskoeffizient R^2 signifikant ist. Das habe ich schon öfter unter dem Thema „Power“ oder „statistische Mächtigkeit“ abgehandelt. Dies gilt auch hier: Man kann mit sehr vielen Fällen oder Datensätzen auch sehr kleine und irrelevante Zusammenhänge, z.B. $R^2 = 0.002$, also 0,2 % der Varianzaufklärung, signifikant bekommen. Umgekehrt kann ein großer Zusammenhang die Signifikanz verfehlen, wenn der Datensatz klein ist. Idealerweise erwarten wir uns hohe Varianzaufklärung, die gleichzeitig signifikant ist.

In der Forschung interessiert uns meistens 1. – Größe der Zusammenhänge von Prädiktoren mit der abhängigen Variablen oder dem Outcome – und 3. – Höhe der Varianzaufklärung durch ein Modell.

In der medizinischen und sozialwissenschaftlichen Forschung findet man selten Modelle, die mehr als ein Drittel bis die Hälfte der Varianz aufklären und benötigt dazu meistens irgendwas zwischen 3 und 10 Variablen mindestens – und um den Faktor 10 bis 20 mehr Fälle.

Große epidemiologische Erhebungen haben meistens viele Tausende von Fällen und können daher auch eine große Zahl von möglichen Einflussvariablen oder Prädiktoren modellieren. Das Problem bei all diesen Studien ist immer: Man weiß nie, ob man die wirklich interessierenden und wichtigen Variablen erfasst hat und ob nicht eine

wichtige Einflussgröße fehlt. Man hat nur eine indirekte Möglichkeit, dies abzuschätzen, nämlich R^2, die Menge der aufgeklärten Varianz. Ist diese hoch, ist die Wahrscheinlichkeit, dass man etwas Wichtiges übersehen hat, gering.

In unserem Beispiel oben hatten wir 5 Variablen und 64 Fälle, also ausreichend Power zur Abschätzung der Parameter.

Wir haben jetzt die klassische lineare Regression anhand dieses Beispiels besprochen. Dies ist die Grundstruktur. Sie kann sehr unterschiedlich erweitert werden, und das Prinzip ist im Grunde immer das gleiche.

Wenn man klinische Studien oder Experimente auswertet, dann führt man meistens neben interessierenden Prädiktoren die Variable, die die Intervention kodiert, als zusätzlichen Prädiktor ein. Ist diese signifikant, dann weiß man, dass die Intervention einen Einfluss hatte und kann auch die Stärke des Einflusses abschätzen.

Wenn die Verteilung der Kriteriums- oder Zielvariable nicht der Normalverteilung folgt, dann werden die Regressionsmodelle etwas anders formalisiert. Man spricht dann vom „verallgemeinerten linearen oder nicht-linearen Modell“. Man kann etwa Regressionen auf Variablen rechnen, die einer Poissonverteilung, einer Gamma-Verteilung, einer logistischen oder anderen Verteilung folgen. Dann werden die Prädiktoren nicht mit einer einfachen linearen Kombination verkoppelt, sondern werden entweder erst mit einer logarithmischen Transformation transformiert und dann additiv verbunden. Bei den Regressionen, die einer logistischen Verteilung folgen, sind die Regressionselemente linear verbundene Exponenten der natürlichen Zahl e. Im Falle von nichtlinearen Regressionen werden die Regressionsglieder in einer passenden Potenz eingebaut.

Aber wichtig ist das Verständnis des Prinzips, das ich hier vermitteln wollte: Es handelt sich immer um eine lineare, oder nicht-lineare, Kombination gewichteter Vorhersageglieder, um Varianz in einem Kriterium aufzuklären. Irgendwann in den 60er Jahren wurde auch arithmetisch gezeigt, dass die bis dahin so beliebte Varianzanalyse und die Regressionsanalyse konzeptuell äquivalent sind [6]. Seither spricht man vom „Allgemeinen linearen Modell“ oder vom „Verallgemeinerten linearen Modell“. Es ist das vielleicht mächtigste Instrument zum Aufklären von vielfältigen Einflüssen auf eine interessierende Variable.

Quellen und Literatur

1. Schmiedel V, Vogt H, Walach H. Are pupil’s “Programme for International Student Assessment (PISA)” scores associated with a nation’s fish consumption? Scandinavian Journal of Public Health. 2017;46:675-9. doi: https://doi.org/10.1177/1403494817717834.
2. Moffett JR, Ives JA, Namboodiri AM. Fatty acids and lipids in neurobiology: A brief overview. In: Watson RR, editor. Fatty Acids in Health Promotion and Disease Causation Urbana, IL: AOCS Press; 2009. p. 517-43.
3. Weiser M, Butt CM, Mohajeri MH. Docosahexaenoic acid and cognition throughout the lifespan. Nutrients. 2016;8(99). doi: https://doi.org/10.3390/nu8020099.
4. Hibbeln JR, Davis JM, Steer C, Emmett P, Rogers I, Williams C, et al. Maternal seafood consumption in pregnancy and neurodevelopmental outcomes in childhood (ALSPAC study): an observational cohort study. The Lancet. 2007;369(9561):578-85. doi: https://doi.org/10.1016/S0140-6736(07)60277-3.
5. Barth M. Konzeption und Evaluation multipler Regressionsanalysen in der anwendungsorientierten klinisch-psychologischen Forschung. In: Strauss B, Bengel J, editors. Forschungsmethoden in der Medizinischen Psychologie. Jahrbuch der Medizinischen Psychologie 14. Göttingen: Hogrefe; 1997. p. 146-60.
6. Wittmann W. Evaluationsforschung. Aufgaben, Probleme und Anwendungen. Berlin: Springer; 1985 1985.

Date Created

Juli 2022