

Aus der Forschung

Was ist Regression zur Mitte?

Regression zur Mitte: Auf einen extremen Messwert folgt meist ein weniger extremer. In einer danach gewählten Gruppe kann das wie echte Veränderung wirken.

Autor

Sinan Isoglu, MBA (Quantic)

Führungskraft für Umsatzwachstum, Dozent und Doktorand

Veröffentlicht

18. Oktober 2026

Lesezeit

18 Min.

Wörter

3.845

Management Summary

Regression zur Mitte ist die Neigung eines Werts, der bei einer Messung extrem ausfiel, bei der nächsten näher am Durchschnitt zu liegen, sofern beide Messungen gleich stark streuen und nicht vollkommen korrelieren. Der Leitfaden definiert sie (Barnett, van der Pols und Dobson; Kahneman und Tversky), erklärt, warum sie entsteht und warum sie wächst, wenn Einheiten nach einem extremen ersten Wert ausgewählt werden, zeigt Galtons Tabelle von 1886 zur Größe von Eltern und Kindern, berichtet Studien, in denen Menschen vorhersagten, als gäbe es keine Regression, und beschreibt, wie man sie von echter Veränderung trennt: mit einer zufällig zugeteilten Vergleichsgruppe, mehr als einer Ausgangsmessung und der Kovarianzanalyse. Die Quellen sind ein Tutorial mit klinischen Beispielen, eine Studie von 1886 zur Körpergröße in Familien und eine Reihe von Vorhersagestudien, meist mit Studierenden.

Inhalt

§ 1	Was ist Regression zur Mitte?	3
§ 2	Warum fällt die zweite Messung weniger extrem aus als die erste?	4
§ 3	Was fand Galton in der Körpergröße von Eltern und Kindern?	5
Abbildung 1	Galtons Elternmittel und der Median der Größe ihrer erwachsenen Kinder	7
§ 4	Warum übersehen intuitive Vorhersagen die Regression?	8
§ 5	Woran erkennt man Regression zur Mitte statt echter Veränderung?	10
§ 6	Wie rechnen Barnett und Kollegen sie aus einer Schätzung heraus?	11
§ 7	Wie sähe das für ein Vertriebsteam nach einem schwachen Quartal aus?	13

Abbildung 2	Eine Regressionsprüfung für eine ausgewählte Gruppe	13
§ 8	Was lassen ein Tutorial, eine Tabelle von 1886 und Experimente mit Studierenden offen?	15
	Referenzen	16

Was ist Regression zur Mitte?

§ 1 von 8

Sie beschreibt, was zwischen zwei Messungen derselben Größe geschieht, wenn die erste ungewöhnlich ausfiel. Barnett et al. (2005)¹ beginnen ihr Tutorial mit der Beschreibung eines statistischen Phänomens, das natürliche Schwankungen in wiederholten Daten wie echte Veränderung aussehen lassen kann; es tritt auf, wenn auf ungewöhnlich große oder kleine Messwerte tendenziell Messwerte folgen, die näher am Mittelwert liegen (S. 215). Ihr erster Abschnitt verortet es bei wiederholten Messungen an derselben Person oder Beobachtungseinheit („RTM is a statistical phenomenon that occurs when repeated“). Kahneman und Tversky (1973)³ betonen, wie verbreitet das Phänomen ist: Regressionseffekte begegnen uns überall („Regression effects are all about us“, S. 249).

¹ Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology*, 34(1), 215–220.

³ Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4), 237–251.

In diesem Leitfaden heißt Regression zur Mitte, auch Regression zum Mittelwert genannt: Ein Wert, der bei einer Messung extrem ausfiel, liegt bei der nächsten im Durchschnitt näher am Mittelwert, auch wenn dazwischen nichts auf die Einheit einwirkt, sofern beide Messungen nicht vollkommen korrelieren und über die Einheiten gleich stark streuen: Kahneman und Tversky³ formulieren den Zusammenhang für zwei Variablen gleicher Varianz, und die Formel von Barnett et al.¹ setzt für beide Messungen dieselbe Streuung an. Sie beschreibt eine Gruppe, nicht jedes ihrer Mitglieder: Manche Einheiten, die am Rand begannen, entfernen sich weiter, doch der Durchschnitt der Gruppe bewegt sich zurück. Zum praktischen Problem wird sie, wenn Einheiten ausgewählt werden, weil ihr erster Wert extrem war, und die folgende Veränderung als Wirkung dessen gelesen wird, was mit ihnen geschah.

Warum fällt die zweite Messung weniger extrem aus als die erste? § 2 von 8

Weil jede Messung Schwankungen enthält, die nicht zur gemessenen Größe gehören. Barnett et al.¹ fassen es in einem Satz: Sie entsteht, weil Werte mit zufälligem Fehler beobachtet werden („It happens because values are observed with random“), worunter sie nicht systematische Schwankungen um einen wahren Mittelwert verstehen, etwa zufällige Messfehler oder zufällige Schwankungen bei einer Person. Systematischer Fehler, bei dem die Werte durchgehend verzerrt sind, ist nicht die Ursache („Systematic error, where the observed values are consistently“). Ihr Beispiel ist eine Person mit einem wahren mittleren HDL-Cholesterin von 50 mg/dl: Ein Wert von 30 ist für sie ungewöhnlich niedrig, und der nächste Wert läge wahrscheinlich näher an 50. Weil Daten fast nie ohne zufälligen Fehler vorliegen, nennen sie das Phänomen verbreitet („a common phenomenon“, S. 215).

Auswahl verstärkt sie. Barnett et al.¹ simulieren 50 Personen aus einer hypothetischen Population und bilden aus allen, deren erster Wert unter 40 mg/dl liegt, eine Gruppe mit niedrigen Werten, die behandelt werden soll. Von den 21 Personen, die bei einer einzelnen Messung in diese Gruppe fallen könnten, haben nur 2 einen wahren Mittelwert unter 40 und 19 einen darüber; bei erneuter Messung liegen diese 19 höchstwahrscheinlich näher an ihrem wahren Mittelwert, wodurch der Durchschnitt der Gruppe steigt. Eine solche Veränderung kann fälschlich einer echten Veränderung zugeschrieben werden („Such a change in the group mean may wrongly be attributed to“), besonders wenn zwischen den beiden Messungen eine Intervention stattfindet („intervention is applied between the two measurements“, S. 216). Ihr Abstract nennt die allgemeine Regel: Der Effekt wird umso deutlicher, je größer der Messfehler ist, und wenn Folgemessungen nur in einer Teilgruppe betrachtet werden, die nach einem Ausgangswert ausgewählt wurde (S. 215).

Sie geben auch die Größe des Effekts für normalverteilte Daten an: die gesamte Standardabweichung eines einzelnen Messwerts in der Population, die die Streuung zwischen Personen und die Schwankung innerhalb jeder Person zusammenfasst, mal eins minus die Korrelation zwischen den beiden Messungen mal einen Faktor, der wächst, je weiter der Grenzwert im Randbereich der Verteilung liegt. Der Effekt nimmt also zu, wenn die Korrelation kleiner wird („becomes smaller the RTM effect increases“), er ist proportional zur Standardabweichung der Population („the RTM effect is proportional to the population“), in

der Formel also zu dieser gesamten, und er ist bei einem Grenzwert weiter im Randbereich größer. Bei einer Korrelation von 0,64 zwischen den beiden Messungen, einer gesamten Standardabweichung von 15 mg/dl und einem Grenzwert von 40 mg/dl ergibt ihr Beispiel etwa 9,6 mg/dl, einen scheinbar großen Anstieg des mittleren HDL-Cholesterins der Gruppe („seemingly large increase in the group’s mean HDL cholesterol“), ganz ohne Behandlung (S. 217).

Kahneman und Tversky³ beschreiben den Zusammenhang mit der Korrelation ohne Formel. Bei zwei Variablen gleicher Varianz ist die Aussage, die eine sei in Bezug auf die andere regressiv, logisch gleichbedeutend damit, dass ihre Korrelation kleiner als eins ist; Regression zu erklären heißt daher zu erklären, warum eine Korrelation kleiner als eins ist („why a correlation is less than unity“, S. 250). Barnett et al.¹ ergänzen, dass Regression zur Mitte bei jeder Messung auftritt, die mit Fehler beobachtet wird („that is observed with error“), nicht nur bei normalverteilten oder stetigen Daten: Bei binären Daten kann sie Einheiten die Kategorie wechseln lassen, ohne dass sich an der zugrunde liegenden Größe etwas ändert (S. 219). Der Leitfaden zum Messfehler trennt einen Indikator von der Zielgröße, für die er steht, und erklärt den klassischen Messfehler, bei dem Rauschen im Prädiktor die geschätzte Steigung einer einfachen Regression abflachen kann; Regression zur Mitte ist eine weitere Folge derselben Art von Zufallsfehler, und Barnett et al. ziehen dieselbe Verbindung (S. 216 bis 217).

Was fand Galton in der Körpergröße von Eltern und Kindern?

§ 3 von 8

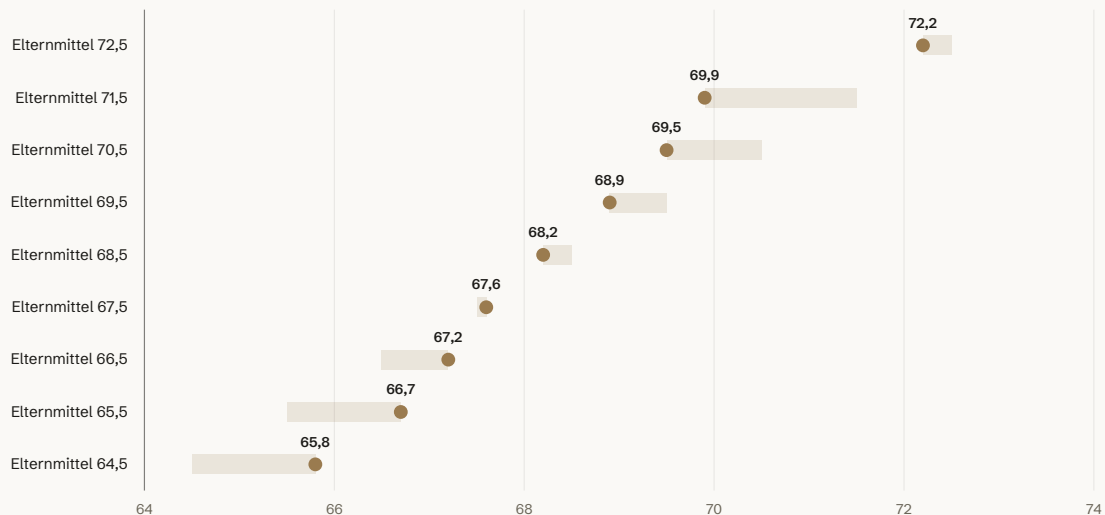
Der Name geht auf ihn zurück. Barnett et al.¹ datieren die erste Erörterung auf Galton im Jahr 1877 und schreiben ihm den Ausdruck zu (S. 215, 217); sein eigener Titel von 1886 spricht von einer Regression zur Mittelmäßigkeit. In diesem Aufsatz erinnert Galton (1886)² an Versuche mit Samen unterschiedlicher Größe derselben Art („seeds of different size but of the same species“): Die Nachkommen fielen weniger extrem aus als die Elternsamens, kleiner als große Eltern und größer als sehr kleine, und die durchschnittliche Regression war proportional zur Abweichung der Eltern von der Mittelmäßigkeit, einem Punkt deutlich unter der durchschnittlichen Größe der gekauften Samen (S. 246).

² Galton, F. (1886). Regression towards mediocrity in hereditary stature. Journal of the Anthropological Institute of Great Britain and Ireland, 15, 246–263.

Danach sammelte er Familienangaben über die Körpergröße. Seine Daten umfassten 930 erwachsene Kinder und 205 Elternpaare; jede weibliche Körpergröße multiplizierte er mit 1,08, um sie auf die männliche Skala zu bringen, und fasste beide Eltern zu einem Elternmittel zusammen, ihrer durchschnittlichen Größe (S. 247, 250). In dieser Population lag das Niveau, das er Mittelmäßigkeit nennt, bei 68 1/4 Zoll, und er formulierte sein Gesetz: Die Abweichung der Nachkommen von diesem Niveau beträgt im Durchschnitt zwei Drittel der Abweichung ihres Elternmittels (S. 252). Er fand es nach eigener Aussage mit unerwarteter Stimmigkeit und Genauigkeit („with unexpected coherence and precision“, S. 247), und seine Tafel IX hält fest, dass sich die Abweichungen der Kinder zu denen ihrer Elternmittel wie 2 zu 3 verhalten („their Mid-Parents as 2 to 3“). Abbildung 1 zeigt die Mediane aus seiner Tabelle I.

Abbildung 1 Galtons Elternmittel und der Median der Größe ihrer erwachsenen Kinder

In allen neun Klassen lag der Median der erwachsenen Kinder näher an 68 1/4 Zoll als das Elternmittel: unterhalb dieses Niveaus lag der Median über dem Elternmittel, oberhalb darunter.



Jede Zeile ist eine Klasse des Elternmittels. Das Band reicht von der Größe der Klasse bis zum Median der Körpergröße ihrer erwachsenen Kinder, markiert durch den Punkt und seinen Wert. In jeder Klasse liegt der Median näher an 68 1/4 Zoll, dem Niveau, das Galton Mittelmäßigkeit nennt (S. 252), als die Größe der Klasse. In der obersten Klasse, am weitesten von 68 1/4 Zoll entfernt, liegt der Median nur 0,3 Zoll unter der Größe der Klasse; diese Klasse beruht auf 6 Elternmitteln, die unterste auf 5. Drei der neun Abstände betragen 0,3 Zoll oder weniger, und Galton überschreibt die Spalten der Kinder mit 62,2, 63,2 und so weiter, weil angegebene Größen sich bei ganzen Zoll häufen (Anmerkung zu Tabelle I). Galtons zwei Drittel sind seine Zusammenfassung dieser Daten, als Linie in seiner Tafel IX gezeichnet, und werden hier nicht berechnet.

Quelle Galton (1886), S. 248, Tabelle I: Zahl der erwachsenen Kinder verschiedener Körpergröße von 205 Elternmitteln, Größen in Zoll, alle weiblichen Größen mit 1,08 multipliziert; die Zeilenköpfe und die Spalte der Mediane. Die Klassen Above und Below nennen keinen Median und sind weggelassen. Von der Klasse 72,5 abwärts umfassen die Klassen 6, 11, 22, 41, 49, 33, 20, 12 und 5 Elternmittel. Vom Seitenbild der gescannten Ausgabe abgelesen.

Galtons eigene Erklärung betraf die Vererbung, nicht die Messung: Das Kind erbt teils von den Eltern, teils von seinen Vorfahren („The child inherits partly from his parents, partly from his ancestry“), und eine weit zurückreichende Ahnenreihe gleicht im Mittel der Bevölkerung, sodass sie mittelmäßig ist („in other words, it will be mediocre“, S. 252). Er fand die Beziehung auch in umgekehrter Richtung. Liest man

dieselbe Tabelle nach Spalten statt nach Zeilen, weicht das wahrscheinlichste Elternmittel eines Mannes nur ein Drittel so stark ab wie der Mann selbst; und weil Menschen nahe der Mittelmäßigkeit so zahlreich sind, ist ein außergewöhnlicher Mann häufiger der etwas außergewöhnliche Sohn recht durchschnittlicher Eltern als der durchschnittliche Sohn sehr außergewöhnlicher Eltern (S. 253 bis 254).

Nach der Lesart dieses Leitfadens trennt dieser zweite Befund die Regression von einer ursächlichen Wirkung. Nichts, was von Eltern auf Kinder übergeht, kann die Eltern großer Kinder kleiner machen, und doch liegen die Eltern großer Kinder näher an der Mitte als ihre Kinder, so wie die Kinder großer Eltern näher an der Mitte liegen als ihre Eltern. Spiegelbildlich ist das nicht: Galton betont, dass die Umkehrung seines Gesetzes weit von dessen zahlenmäßigem Gegenstück entfernt ist, nach dem das Elternmittel anderthalbmal so stark abweicht wie der Sohn, während die Tabelle ein Drittel ergibt (S. 253 bis 254). Beides beruht auf einer unvollkommenen Korrelation, allerdings ohne die gleichen Varianzen, die Kahneman und Tversky³ voraussetzen: Nach Galton unterscheidet sich die Streuung der Elternmittel von der Streuung der Körpergröße Erwachsener um den Faktor Quadratwurzel aus 2 (S. 251), sodass die beiden Anteile nicht gleich sein müssen. Galtons Zahlen haben Bestand. Die späteren Quellen behandeln die Körpergröße von Eltern und Kindern als bekannten Fall von Regression und erklären Regression allgemein mit unvollkommener Korrelation und zufälliger Schwankung; auf die weit zurückreichende Ahnenreihe, mit der Galton argumentierte, gehen beide nicht ein.

Warum übersehen intuitive Vorhersagen die Regression?

§ 4 von 8

Weil Menschen nach Ähnlichkeit vorhersagen, und eine so gebildete Vorhersage ist eine nicht regressive Vorhersage: so extrem wie die Evidenz, auf der sie beruht. Kahneman und Tversky (1973)³ formulieren ihre These ohne Umschweife: Menschen sagen nach Repräsentativität vorher („people predict by representativeness“), das heißt, sie wählen das Ergebnis, das der Evidenz am ähnlichsten erscheint. Deshalb, so ihr Argument, sind intuitive Vorhersagen unempfindlich dafür, wie verlässlich die Evidenz ist, und für die A-priori-Wahrscheinlichkeit des Ergebnisses, entgegen der Logik statistischer Vorhersage, und falsche Intuitionen über statistische Regression führen sie auf dieselbe Repräsentativitätsheuristik zurück (S. 237).

Die statistische Regel lautet, dass Vorhersagen nur so stark streuen sollten, wie die Evidenz es rechtfertigt. Bei perfekter Genauigkeit streuen die Vorhersagen so stark wie die Ergebnisse; bei Genauigkeit

null wird jedes Mal derselbe Wert vorhergesagt; dazwischen sind Vorhersagen regressiv, und je größer die Unsicherheit, desto kleiner ihre Streuung (S. 243). Die Versuchspersonen folgten dieser Regel nicht. In zwei Studien mit Beschreibungen von Studienanfängern bewertete eine Gruppe jede Beschreibung, eine andere sagte den Notendurchschnitt jedes Studierenden vorher; beide Gruppen urteilten gleich extrem („groups produced equally extreme“), mit mittleren Standardabweichungen von 25,7 und 24,0 in der einen Studie und 22,2 und 21,4 in der anderen, und keiner der Unterschiede war signifikant (S. 245).

Eine dritte Studie gab drei Gruppen dieselben zehn Perzentilwerte und fragte nach dem Notendurchschnitt. Gruppe 1 erfuhr, die Werte seien Perzentile des Notendurchschnitts selbst; Gruppe 2, sie stammten aus einem Test der geistigen Konzentration, dessen Ergebnisse als von Tag zu Tag schwankend beschrieben wurden; Gruppe 3, sie stammten aus einem Test des Sinns für Humor. Die Gruppen, die aus den entfernteren Angaben vorhersagten, hätten regressiver urteilen sollen. Die Streuung der Vorhersagen betrug 0,91 in Gruppe 1 und 0,87 in Gruppe 2, kein signifikanter Unterschied, und 0,69 in Gruppe 3, signifikant weniger als in Gruppe 2 (Tabelle 3, S. 247). Mit den Worten der Autoren sind Vorhersagen von Leistung aus einem Maß der Fähigkeit nicht regressiv, so unzuverlässig dieses Maß auch ist („people fail to regress when predicting“); die Vorhersagen aus dem Humortest waren regressiv, nach Darstellung der Autoren aber nicht genug.

Dasselbe Muster zeigte sich außerhalb des Hörsaals. Offiziere eines Auswahlgremiums der israelischen Armee, die wussten, dass ihre Vorhersagevalidität nur mäßig war („predictive validity was only moderate“, in der Größenordnung von 0,20 bis 0,40), sagten die Abschlussnoten der Offiziersschule für über 200 Kandidaten in einer Verteilung vorher, die der tatsächlichen nahezu glich, nur dass sie weniger Misserfolge vorhersagten, als tatsächlich eintraten: Ihre Vorhersagen waren nicht regressiv („their predictions were nonregressive“, S. 247). Weil intuitive Vorhersagen oft nicht regressiv genug sind, ergänzen die Autoren, ist der Abstand zwischen Vorhersagen und Ergebnissen an den Extremen am größten (S. 249).

Woran erkennt man Regression zur Mitte statt echter Veränderung?

§ 5 von 8

Indem man zunächst annimmt, dass sie vorliegt. Barnett et al.¹ schreiben, man solle von Regression zur Mitte ausgehen, solange die Daten nichts anderes zeigen („One should assume that RTM has taken place unless the data“), und ihre Diskussion ergänzt, sie müsse als Ursache einer beobachteten Veränderung ausgeschlossen werden („needs to be ruled out as a cause of an observed“), bevor man eine andere Erklärung sucht. Ihr erster Prüfschritt ist eine Grafik, ein Streudiagramm der Veränderung („a scatterplot of change“), also Folgemessung minus Ausgangswert, gegen den Ausgangswert. Ihre Kernaussagen nennen die Frage, die man daran stellt („is there more change at the tails of the baseline measurements?“): Zeigt sich an den Rändern der Ausgangswerte mehr Veränderung? Sie nennen auch einen Fall aus der Gesundheitsforschung, der ihrer Ansicht nach übersehen wurde: Ein Teil des Rückgangs der Meningitis-C-Fälle in Großbritannien um 90 Prozent nach einem Impfprogramm könnte darauf zurückgehen, dass auf ein sehr schlechtes Jahr wahrscheinlich ein besseres folgt (S. 217, 219, 220).

Kahneman und Tversky³ erklären, warum sie so leicht übersehen wird. Menschen erwarten Regression in vielen Situationen nicht, in denen sie zwangsläufig auftritt („it is bound to occur“), lernen das Konzept nur schwer und erfinden, wenn sie Regression doch bemerken, Erklärungen über eine dynamische Ursache. Gefragt nach einem 95-Prozent-Intervall um den wahren IQ einer Person, die in einem Test, beschrieben als wahrer Wert plus Zufallsfehler, 140 Punkte erreicht hatte, gaben 73 von 108 fortgeschrittenen Psychologiestudierenden Intervalle an, die symmetrisch um 140 lagen; 24 gaben regressive Intervalle an, 11 solche, die in die Gegenrichtung wiesen. Die meisten sagten vorher, als sei der Wert von 140 ein wahrer Wert („the value of 140 was a true score“, S. 250). Unter den Scheinerklärungen in den Sozialwissenschaften nennen sie dynamische Prinzipien, mit denen erklärt wurde, warum Unternehmen, die zu einem Zeitpunkt außergewöhnlich gut abschnitten, später schlechter dastehen, und sie verweisen auf frühere Diskussionen über Regressionsfehlschlüsse in der Forschung (S. 250).

Ihr letztes Beispiel ist eine Geschichte aus der Erfahrung eines der Autoren. Fluglehrer, die jedes gelungene Manöver lobten, berichteten, dass das nächste meist schlechter ausfiel, und schlossen daraus, dass Lob schade. Die Leistung ist nicht vollkommen verlässlich und verbessert sich von Manöver zu Manöver nur langsam, also schnitten

Flugschüler, die bei einem Versuch außergewöhnlich gut waren, beim nächsten wahrscheinlich schlechter ab, gleich was der Lehrer sagte. Die Lehrer entdeckten die Regression, schrieben sie aber dem Schaden durch Lob zu („discovered the regression but attributed it“). Weil Menschen gutes Verhalten eher belohnen und schlechtes bestrafen, folgern die Autoren, dass sich andere allein wegen der Regression am ehesten nach einer Strafe verbessern und nach einer Belohnung verschlechtern („deteriorate after being rewarded“). Keiner der Studierenden, denen das Problem vorgelegt wurde, nannte Regression als Ursache (S. 250 bis 251).

Zusammengenommen sind die Warnzeichen nach der Lesart dieses Leitfadens vier: Die Einheiten wurden ausgewählt, weil ein erster Wert extrem war; dieser Wert ist eine einzelne Messung; nichts, was auf dieselbe Weise ausgewählt wurde, blieb unbehandelt; und die Veränderung ist dort am größten, wo die ersten Werte am extremsten waren. Treffen alle vier zu und liegt eine Geschichte über die Ursache bereit, ist diese Geschichte noch nicht gegen Regression geprüft, und sie ungeprüft zu übernehmen, ist der Regressionsfehlschluss.

Wie rechnen Barnett und Kollegen sie aus einer Schätzung heraus?

§ 6 von 8

In zwei Phasen: bei der Planung der Studie und bei der Auswertung der Daten. Das erste Studiendesign ist die zufällige Zuteilung. Werden Personen zufällig auf Vergleichsgruppen verteilt („If subjects are randomly allocated to comparison groups the“), sollte die Regression zur Mitte alle Gruppen gleich treffen; die mittlere Veränderung einer unbehandelten Gruppe, etwa einer Placebogruppe, schätzt dann die durch Regression zur Mitte bewirkte Veränderung („provides an estimate of the change caused by“) zuzüglich eines etwaigen Placeboeffekts, und der Unterschied zwischen den mittleren Veränderungen der Gruppen schätzt den um sie bereinigten Behandlungseffekt (S. 217). Der Leitfaden zum A/B-Test behandelt randomisierte Experimente dieser Art; lässt sich eine Vergleichsgruppe nicht randomisieren, muss ein Difference-in-Differences-Vergleich begründen, dass sich beide Gruppen ohne Behandlung parallel entwickelt hätten, was der Leitfaden zu parallelen Trends darlegt. Nach der Lesart dieses Leitfadens ist eine Gruppe, die wegen eines extremen ersten Werts ausgewählt wurde, ein schwieriger Fall für dieses Argument, weil gerade von ihr Regression zu erwarten ist.

Das zweite Studiendesign wählt nach mehr als einer Messung aus. Um die Schwankung zu verringern, lassen sich Personen nach zwei oder mehr Ausgangsmessungen auswählen („To reduce the variability we

can select subjects using two or more baseline measurements“), wobei der Grenzwert auf ihren Mittelwert oder auf eine spätere Messung angewandt wird. Die Verringerung ist von der ersten zur zweiten Ausgangsmessung am größten und danach kleiner. In einem realistischeren Modell, das die Autoren zitieren und in dem sich zwei Messungen derselben Person ähnlicher sind als Messungen zweier Personen, beseitigen weitere Ausgangsmessungen den Effekt nie ganz, und die Kernaussagen halten fest, dass dieses Studiendesign das Problem wohl nicht vollständig ausschaltet (S. 218, 220).

Für die Auswertung nennen sie zwei Verfahren. Sind Mittelwert und Standardabweichung der Population und die Standardabweichung innerhalb der Person bekannt oder schätzbar, lässt sich die nach ihrer Formel erwartete Regression von der beobachteten Veränderung abziehen. Oft praktikabler („often more practical“) ist die Kovarianzanalyse (ANCOVA), die die Folgemessung jeder Person an ihre Ausgangsmessung anpasst („adjusts each subject’s follow-up measurement according to their“): Die Folgemessung wird auf den zentrierten Ausgangswert als Kovariate und auf die Gruppenzugehörigkeit regressiert, und der Koeffizient der Gruppe ist der um Regression zur Mitte bereinigte Behandlungseffekt. Sie empfehlen die ANCOVA überall dort, wo ein t-Test infrage käme (S. 218, 219).

Ihr Rechenbeispiel nutzt logarithmierte Betacarotin-Werte im Blut aus einer Teilstichprobe einer randomisierten Studie, insgesamt 96 Messpaare, 52 mit Nahrungsergänzung und 44 mit Placebo. Beschränkt auf die 49 Paare mit einem Ausgangswert unter $0,5 \mu\text{M/l}$ zeigte der Mittelwert der Placebogruppe einen möglichen Anstieg um 0,24 (95-Prozent-KI $-0,04$ bis $0,53$); die Formel setzte die erwartete Regression bei 0,29 an, sodass die Autoren in der Placebogruppe keine echte Veränderung sehen („that there was no real“). Der Behandlungseffekt in dieser Teilgruppe betrug 0,84 (0,27 bis 1,41) mit t-Tests auf die Veränderungen und 0,86 (0,37 bis 1,36) mit der ANCOVA, deren Intervall schmaler war, weil der Ausgangswert einen Teil der Varianz im Ergebnis erklärte (Tabellen 1 und 2, S. 219). Weil in der Studie zufällig zugeteilt wurde, war sie schon durch ihr Studiendesign gegen Regression zur Mitte geschützt.

Wie sähe das für ein Vertriebsteam nach einem schwachen Quartal aus?

§ 7 von 8

Zuerst kommt ein Arbeitsblatt. Es richtet aus den Empfehlungen von Barnett et al.¹ abgeleitete Fragen an jede Veränderung, die in Einheiten gemessen wird, die wegen eines extremen Werts ausgewählt wurden: wonach sie ausgewählt wurden, wie viele Messungen hinter der Auswahl standen, wer auf dieselbe Weise ausgewählt und nicht behandelt wurde, wie viel Veränderung daher ohne Maßnahme zu erwarten war, was sich verändert hat und was bleibt, wenn erwartete Veränderung und Ausgangswert berücksichtigt sind.

Abbildung 2 Eine Regressionsprüfung für eine ausgewählte Gruppe

Auswahl, Ausgangswerte, Vergleich, erwartete Veränderung, beobachtete Veränderung und bereinigten Effekt festhalten, bevor man der Maßnahme die Wirkung zuschreibt.

Auswahl	Ausgangswerte	Vergleich	Erwartet	Beobachtet	Bereinigt
Welcher erste Wert, welcher Grenzwert?	Wie viele Messungen vorher?	Gleich ausgewählt, ohne Maßnahme?	Änderung ohne Maßnahme, und warum?	Änderung nach der Maßnahme?	Effekt nach Vergleich und Ausgangswert?

Bewegte sich die Vergleichsgruppe so stark wie die ausgewählte Gruppe, ist nicht gezeigt, dass die Maßnahme wirkt.

Eine Zeile je Aussage, eine Maßnahme habe eine ausgewählte Gruppe verändert. Eine Zeile mit einer einzigen Ausgangsmessung und ohne Vergleich kann die Maßnahme nicht von Regression zur Mitte trennen.

Quelle Entscheidungsblatt des Autors. Die Felder für Auswahl und Ausgangswerte folgen Barnett et al. (2005), S. 216 bis 218; die Felder für Vergleich und Erwartung folgen Barnett et al. (2005), S. 217; das Feld für den bereinigten Effekt folgt Barnett et al. (2005), S. 218 und 219. Die Anwendung auf Vertriebs- oder Kundendaten ist eine Überlegung dieses Leitfadens. Die Felder sind Eingaben der Lesenden; es werden keine Unternehmensdaten geliefert.

Ein Beispiel ohne Unternehmensdaten. Die Vertriebsleitung ordnet vierzig Vertriebsmitarbeitende nach ihrem Auftragseingang im dritten Quartal, nimmt die zehn schwächsten in ein Coachingprogramm auf und zeichnet die zehn stärksten aus. Im vierten Quartal steigt der Durchschnitt der unteren zehn, der der oberen zehn sinkt, und die naheliegende Lesart ist, dass Coaching wirkt und Auszeichnungen träge machen, die Lesart der Fluglehrer über Lob. Doch der Auftragseingang eines Quartals kann Schwankung enthalten, die nichts mit Können zu tun hat: Ein Abschluss kann über das Quartalsende rutschen, ein großer Auftrag früher kommen. Korreliert der Auftragseingang eines Quartals deutlich unter eins mit dem des nächsten und streut er ähnlich stark, war für beide Gruppen ohne Coaching und ohne Auszeichnung eine Bewegung zum Teamdurchschnitt zu erwarten, und wie weit, hängt von dieser Korrelation ab, die die Vertriebsleitung nicht gemessen hat. In Abbildung 2 lautet die Zeile: ausgewählt nach einem Quartal, die unteren zehn; eine Ausgangsmessung; kein Vergleich; also keine Schätzung der erwarteten Veränderung und kein bereinigter Effekt.

Dasselbe Budget erlaubt ein saubereres Design. Von den zehn, die infrage kommen, erhalten fünf zufällig ausgewählte das Coaching in diesem Quartal und die anderen fünf im nächsten, und man vergleicht die Veränderung beider Hälften im selben Quartal: Mit fünf je Hälfte ist der Vergleich verrauscht, aber von der Regression zur Mitte ist zu erwarten, dass sie beide Hälften gleich trifft. Ausgewählt wird nach dem Durchschnitt von zwei oder mehr Quartalen statt nach einem. Und ausgewertet wird der Auftragseingang des nächsten Quartals mit dem bisherigen Auftragseingang als Kovariate statt als einfache Vorher-nachher-Differenz. Dieselbe Zeile lässt sich für einen Kunden ausfüllen, der nach einem starken Einbruch seiner Bestellungen auffällt und an ein Team für Kundenbindung übergeben wird. Das sind Beispiele des Autors, sie sind nicht geprüft; keine der drei Quellen hat Vertriebs- oder Kundendaten untersucht.

Was lassen ein Tutorial, eine Tabelle von 1886 und Experimente mit Studierenden offen?

§ 8 von 8

Wie groß Regression zur Mitte in einer kommerziellen Kennzahl ist. Barnett et al. (2005)¹ ist ein Tutorial: Seine Formel setzt normalverteilte Daten voraus, seine Beispiele sind klinisch, und sein Rechenbeispiel ist eine zufällige Teilstichprobe einer einzigen Studie. Galton (1886)² arbeitete mit Familienangaben, die er über ausgelobte Preise sammelte, und mit Körpergrößen, deren Fehler nach seinem Urteil mindestens so oft über zwei Drittel Zoll lag wie darunter, auch weil unklar war, ob mit oder ohne Schuhe gemessen wurde (S. 252); seine Tabelle I zählt 928 Kinder, ebenso S. 254, während seine Beschreibung der Daten auf S. 247 930 nennt, und die späteren Quellen erklären Regression allgemein mit zufälligem Fehler und unvollkommener Korrelation, nicht mit der Abstammung. Kahneman und Tversky (1973)³ untersuchten Urteile, meist von Studierenden, über hypothetische Personen, mit einem einzigen realen Auswahlgremium; ihr Fall aus der Flugschule ist eine einzelne Anekdote.

Nach der Lesart dieses Leitfadens lassen sich zwei Dinge mit einiger Zuversicht übertragen: Jede Veränderung, die in Einheiten gemessen wird, die wegen eines extremen ersten Werts ausgewählt wurden, enthält Regression zur Mitte, sofern das Studiendesign sie nicht beseitigt, und Menschen, auch Analysten, neigen dazu, diese Regression mit einer Ursache zu erklären. Wie groß sie im quartalsweisen Auftragseingang, im Kundenumsatz oder in Kundenbewertungen ist, hängt davon ab, wie stark eine Periode jeweils mit der nächsten korreliert und ob beide gleich stark streuen, und das sagen die Quellen nicht.

 Ende des Beitrags

Referenzen

Zitate stehen in der Sprache des Originals; die deutschen Wiedergaben im Fließtext sind sinngemäß und stammen vom Autor.

- 1** Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology*, 34(1), 215–220. <https://doi.org/10.1093/ije/dyh299>
- 2** Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 246–263. <https://doi.org/10.2307/2841583>
- 3** Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4), 237–251. <https://doi.org/10.1037/h0034747>

Zitierweise

Isoglu, S. (2026, 18. Oktober). Was ist Regression zur Mitte? Sinan Isoglu. <https://isoglu.com/de/insights/journal/was-ist-regression-zur-mitte/>
Online-Fassung:
<https://isoglu.com/de/insights/journal/was-ist-regression-zur-mitte/>

Sinan Isoglu, MBA (Quantic)

Führungskraft für Umsatzwachstum, Dozent und Doktorand
© 2026 Sinan Isoglu · isoglu.com · Alle Rechte vorbehalten