

Aus der Forschung

Was ist Propensity-Score-Matching?

Propensity-Score-Matching vergleicht Behandelte mit Unbehandelten, die eine ähnliche geschätzte Behandlungschance hatten. Es gleicht nur Gemessenes aus.

Autor	Veröffentlicht	Lesezeit	Wörter
Sinan Isoglu, MBA (Quantic) Führungskraft für Umsatzwachstum, Dozent und Doktorand	18. Oktober 2026	25 Min.	5.279

Management Summary Propensity-Score-Matching schätzt den Effekt einer Behandlung auf die Einheiten, die sie erhalten haben: Jede wird mit unbehandelten Einheiten verglichen, deren aus vorher gemessenen Kovariaten geschätzte Behandlungswahrscheinlichkeit nahezu gleich war. Dieser Konzeptleitfaden definiert den Propensity Score und erklärt, warum ein einziger Score viele Kovariaten ersetzen kann (Rosenbaum und Rubin), nennt die beiden Annahmen einer kausalen Lesart, keine nicht gemessenen gemeinsamen Ursachen und Überlappung, vergleicht Matching, Stratifizierung und Gewichtung auf dem Score, zeigt die Prüfung der Balance und berichtet, wie Matching und verwandte Verfahren gegenüber 15 randomisierten Werbeexperimenten bei Facebook abschnitten (Gordon et al.). Er bietet ein Blatt für eine Schätzung aus dem Matching. Die Quellen sind ein statistischer Aufsatz und eine Vergleichsstudie auf einer Plattform.

Inhalt	§ 1	Was ist Propensity-Score-Matching?	3
	§ 2	Warum kann ein einziger Score viele Kovariaten ersetzen?	4
	§ 3	Was muss gelten, damit ein Vergleich nach dem Matching als kausal gilt?	6
	§ 4	Wie unterscheiden sich Matching, Stratifizierung und Gewichtung auf dem Score?	7
	§ 5	Wie wird die Balance nach dem Matching geprüft?	10
	§ 6	Was zeigte sich, als Matching an den Experimenten von Facebook gemessen wurde?	12
	Abbildung 1	Anteil der Verzerrung des rohen Vergleichs nach der Stratifizierung auf dem Propensity Score, 14 Facebook-Studien mit Kauf als Conversion	15

§ 7	Was gehört zu einer Schätzung aus dem Matching?	16
Abbildung 2	Ein Blatt für eine Schätzung aus dem Matching	17
§ 8	Welche Teile des Facebook-Befunds lassen sich auf andere Daten übertragen?	19
	Referenzen	21

Was ist Propensity-Score-Matching?

§ 1 von 8

Es ist ein Weg, eine Vergleichsgruppe aus Daten zu bilden, in denen niemand die Behandlung per Zufall zugeteilt hat. Rosenbaum und Rubin (1983)² definieren den Score, auf dem es beruht, als die bedingte Wahrscheinlichkeit, einer bestimmten Behandlung zugeteilt zu werden, gegeben einen Vektor beobachteter Kovariaten („The propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates.“). Matching auf dem Score stellt jeder behandelten Einheit unbehandelte Einheiten mit etwa derselben geschätzten Behandlungswahrscheinlichkeit gegenüber und vergleicht ihre Ergebnisse. Matching im Allgemeinen ist bei ihnen eine Stichprobenziehung aus einem großen Vorrat möglicher Kontrollen, die eine Kontrollgruppe überschaubarer Größe ergibt, deren Kovariaten so verteilt sind wie in der behandelten Gruppe („is a method of sampling from a large reservoir of potential controls to produce a control group of modest size in which the distribution of covariates is similar to the distribution in the treated group“). Die Kovariaten stammen bei ihnen aus der Zeit vor der Behandlung („all of the measurements in x are made prior to treatment assignment“), auch wenn x nicht jede Kovariate enthalten muss, nach der die Behandlung zugeteilt wurde.

² Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.

Gordon et al. (2019)¹, die das Verfahren an Werbedaten prüfen, beschreiben, was der Vergleich schätzt: Matching-Verfahren suchen unbehandelte Personen, die den behandelten ähneln, und setzen deren Ergebnisse für die fehlenden potenziellen Ergebnisse der Behandelten ein („To estimate the ATT, matching methods find untreated individuals similar to the treated individuals and use the outcomes from the untreated individuals to impute the missing potential outcomes for the treated individuals.“). Der ATT ist der durchschnittliche Behandlungseffekt der Behandelten: der Effekt auf die Einheiten, die die Behandlung erhalten haben, nicht auf alle, die sie hätten erhalten können. Als einen Vorzug nennen die Autoren, dass Matching-Verfahren keine bestimmte Form eines Ergebnismodells unterstellen („do not assume a particular form for an outcome model“).

¹ Gordon, B. R., Zettelmeyer, F., Bhargava, N., & Chapsky, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Marketing Science*, 38(2), 193–225.

In diesem Leitfaden heißt Propensity-Score-Matching: den Effekt einer Behandlung auf die Einheiten zu schätzen, die sie erhalten haben,

indem man jeder von ihnen eine oder mehrere unbehandelte Einheiten gegenüberstellt, deren geschätzter Propensity Score, berechnet aus vor der Behandlung gemessenen Kovariaten, nahezu gleich ist, und ihre Ergebnisse vergleicht. Kausal ist das Ergebnis nur unter den beiden Annahmen, die unten folgen, und nur für die behandelten Einheiten, die vergleichbare Partner gefunden haben.

Der Leitfaden zur Inkrementalität nutzt dieselbe Facebook-Studie, um zu zeigen, wie weit ein roher Vergleich von Nutzern mit und ohne Werbekontakt den Beitrag der Werbung überzeichnen kann. Diese Seite behandelt das Verfahren selbst: was der Score leistet, was er voraussetzt, wie sich seine Varianten unterscheiden und wie gut es gegenüber randomisierten Experimenten abschnitt.

Warum kann ein einziger Score viele Kovariaten ersetzen? § 2 von 8

Weil behandelte und unbehandelte Einheiten bei gleichem Score dieselbe Verteilung der Kovariaten haben, die in ihn eingegangen sind. Rosenbaum und Rubin (1983)² nennen eine Funktion der Kovariaten einen Balancing Score, wenn behandelte und Kontrolleinheiten bei jedem ihrer Werte dieselbe Verteilung der Kovariaten haben; mit solchen Scores lassen sich behandelte und Kontrolleinheiten so gruppieren, dass direkte Vergleiche aussagekräftiger werden („Balancing scores, defined here, can be used to group treated and control units so that direct comparisons are more meaningful.“). Die Gesamtheit der Kovariaten ist der feinste Balancing Score, der Propensity Score der gröbste („is the finest balancing score and the propensity score is the coarsest“): Eine Zahl leistet die Arbeit vieler. Daraus folgt: Sind eine Subklasse oder ein Paar aus behandelter und Kontrolleinheit im Score homogen, haben beide Seiten dieselbe Verteilung der Kovariaten („if a subclass of units or a matched treatment-control pair is homogeneous in $e(x)$, then the treated and control units in that subclass or matched pair will have the same distribution of x “); dabei ist $e(x)$ der Score und x der Vektor der Kovariaten. Gordon et al. (2019)¹ nennen den praktischen Grund, warum das zählt: Exaktes Matching ist nur mit wenigen diskreten Merkmalen machbar („Exact matching is only feasible using a small set of discrete observables.“).

Ausgeglichene Kovariaten sind noch keine kausale Schätzung. Dafür braucht es die Annahme im nächsten Abschnitt. Gilt sie, können Einheiten mit gleichem Balancing Score, aber unterschiedlicher Behandlung einander als Kontrollen dienen („units with the same value of the balancing score $b(x)$ but different treatments can act as controls

for each other“), und Paar-Matching, Subklassifikation und Kovarianzadjustierung auf einem Balancing Score können unverzerrte Schätzungen des durchschnittlichen Behandlungseffekts liefern („pair matching, subclassification and covariance adjustment on a balancing score can produce unbiased estimates of the average treatment effect“); der Durchschnitt läuft dort über die ganze Population, beim Matching jeder behandelten Einheit, wie Gordon et al.¹ es tun, dagegen über die Behandelten. Außerhalb einer randomisierten Studie ist die Funktion des Propensity Score fast immer unbekannt („the propensity score function is almost always unknown“) und wird aus den Daten geschätzt, etwa mit einem Logit-Modell.

Das Schätzen kostet in der Simulation der Autoren wenig; sie übernehmen sie aus früheren Arbeiten von Rubin. Bei 50 behandelten Einheiten, bivariat normalverteilten Kovariaten mit gemeinsamer Kovarianzmatrix und linearen, in beiden Gruppen parallelen Beziehungen zwischen Kovariaten und Ergebnis beseitigte Matching auf dem geschätzten Score, das unter diesen Bedingungen dem Matching auf einer linearen Diskriminante entspricht, zwischen 67 und 92 Prozent der anfänglichen Verzerrung, je nachdem, wie weit die Gruppen anfangs auseinanderlagen, wenn der Vorrat möglicher Kontrollen doppelt so groß war wie die behandelte Gruppe, und zwischen 90 und 98 Prozent bei vierfacher Größe (Tabelle 1, S. 51). Die Autoren folgern, dass geschätzter und wahrer Score die Verzerrung annähernd gleich gut beseitigen und das Schätzen daher kaum etwas kostet („the population and sample propensity scores are approximately equally effective in removing bias, so that no substantial loss is incurred by having to estimate the propensity score“).

Was muss gelten, damit ein Vergleich nach dem Matching als kausal gilt?

§ 3 von 8

Zwei Annahmen über die übliche Bedingung hinaus, dass jede Einheit nur eine Version der Behandlung erhält und die Behandlung einer Einheit das Ergebnis einer anderen nicht verändert, und nur eine davon lässt sich an den Daten prüfen. Gordon et al. (2019)¹ nennen beide. Die erste ist die Annahme bedingter Unabhängigkeit, bei ihnen *unconfoundedness*: Gegeben die beobachteten Kovariaten, sind die potenziellen Ergebnisse unabhängig von der Behandlung; es gibt also keine unbeobachteten Merkmale, die mit Behandlung und potenziellen Ergebnissen zusammenhängen („no unobserved characteristics of individuals associated with the treatment and potential outcomes exist“). Sie gilt als die umstrittenste Annahme und ist ohne Experiment nicht prüfbar („This particular assumption is considered the most controversial and is untestable without an experiment.“). Die zweite ist die Überlappung: Bei jedem Wert der Kovariaten liegt die Behandlungswahrscheinlichkeit echt zwischen 0 und 1, jede Art von Einheit hätte also behandelt oder unbehandelt sein können. Anders als die erste lässt sie sich prüfen, vor und nach der Adjustierung („Overlap can be assessed before and after adjustments are made to each group.“).

Zusammen ergeben sie, was Rosenbaum und Rubin (1983)² stark ignorierbare Behandlungszuteilung nennen. In einer randomisierten Studie mit sauber erhobenen Daten enthalten die Kovariaten bekanntermaßen alles, was über die Zuteilung entschied und mit dem Ergebnis zusammenhängen kann; außerhalb einer solchen Studie ist das in der Regel nicht bekannt („This condition is usually not known to hold in a nonrandomized experiment.“). Eine randomisierte Studie gibt zudem jeder Einheit eine Chance auf jede Behandlung. Unter starker Ignorierbarkeit ist nach Gordon et al.¹ bei zwei Personen mit gleichem Propensity Score der Werbekontakt so gut wie zufällig („Given two individuals with the same propensity scores, exposure status is as good as random.“).

Gordon et al.¹ begegneten fehlender Überlappung, indem sie Einheiten mit extremen geschätzten Scores herausnahmen, für die es kaum vergleichbare Partner gibt. Sie entfernten Nutzer mit einem geschätzten Score unter 0,05 oder über 0,95, nach dem Vorgehen von Imbens, und schätzten den Score für den Rest neu; diese einfache Trimmregel sollte die Überlappung verbessern („apply a simple trimming rule to improve overlap in the covariate distributions“). Nach der Lesart dieses Leitfadens ändert das Trimmen, wen die Schätzung

beschreibt: Sie gilt dann für die verbliebenen Behandelten, nicht für alle.

In der Werbung setzt schon das Targeting die Überlappung unter Druck. Kennten sie die genauen Daten und die Struktur der Targeting-Systeme, schreiben Gordon et al.¹, fiel die Verteilung des Propensity Score auf Punktmassen bei 0 und 1 zusammen („the propensity score distribution would collapse to discrete masses at 0 and 1“), denn eine deterministische Auspielung lässt keinen gemeinsamen Wertebereich von Behandelten und Unbehandelten zu („a deterministic exposure system implies that common support in observables between treated and untreated observations cannot exist“). Ihre Daten enthielten noch zufällige Variation im Werbekontakt, aus der Steuerung der Auslieferung jeder Kampagne durch Facebook, den Geboten von Werbetreibenden, deren Interessen mit dem Ergebnis der Kampagne nichts zu tun haben, und dem Zeitpunkt der Besuche, sodass die Verfahren nicht scheitern mussten; mit ausgefeilteren Targeting-Systemen wird ihr Scheitern aber immer wahrscheinlicher („However, as ad-targeting systems become more sophisticated, such failure is increasingly likely.“).

Wie unterscheiden sich Matching, Stratifizierung und Gewichtung auf dem Score?

§ 4 von 8

Darin, wie sie den Score für den Vergleich nutzen, und darin, welchen Durchschnitt sie liefern. Rosenbaum und Rubin (1983)² wenden den Score in drei Standardverfahren der Adjustierung in Beobachtungsstudien an („three standard techniques for adjustment in observational studies“): Matching, Subklassifikation und Kovarianzadjustierung.

Matching stellt jeder behandelten Einheit die unbehandelten Einheiten mit dem nächstgelegenen Score gegenüber. Exakte Paare sind selten: Selbst auf einem einzigen Score sind genaue Treffer oft nicht zu erreichen, sodass Verfahren für näherungsweise Treffer nötig sind („Unfortunately, exact matches even on a scalar balancing score are often impossible to obtain, so methods which seek approximate matches must be used.“). Gordon et al. (2019)¹ nahmen für jeden Nutzer mit Werbekontakt die M Nutzer ohne Werbekontakt mit den nächstgelegenen geschätzten Log-Odds des Scores, und zwar mit Zurücklegen, weil das die Verzerrung verringern kann, nicht von der Sortierung der Daten abhängt und weniger Rechenaufwand erfordert („because it can reduce bias, does not depend on the sort order of the data, and is less computationally burdensome“). Da die zugeordneten Einheiten für die fehlenden Ergebnisse der Behandelten eintreten,

schätzt das Verfahren den Effekt auf die Behandelten. Rosenbaum und Rubin² nennen Gründe für das Matching, zuerst, dass Paare aus behandelter und Kontrolleinheit auch Forschenden ohne tiefe statistische Ausbildung die Gleichwertigkeit beider Gruppen unmittelbar vor Augen führen („Matched treated and control pairs allow relatively unsophisticated researchers to appreciate immediately the equivalence of treatment and control groups“), und sie halten fest, dass Matching auf dem Score auch seltene binäre Merkmale ausgleichen kann, für die sich einzelne Paare nicht passend bilden lassen („can effectively balance rare binary variables for which it is not possible to match treated and control units adequately on an individual basis“).

Stratifizierung, bei Rosenbaum und Rubin Subklassifikation², teilt den Score in Bänder und vergleicht behandelte und unbehandelte Einheiten innerhalb jedes Bandes. Wer nach vielen Kovariaten zugleich gruppiert, erhält bald Bänder, die nur Behandelte oder nur Unbehandelte enthalten; die Subklassifikation auf dem Propensity Score umgeht dieses Problem („Subclassification on the propensity score is a natural way to avoid this problem.“). In ihrem medizinischen Beispiel bildeten die Autoren fünf gleich große Bänder an den Quintilen des Scores mit je 303 Patienten („each containing 303 patients“), gestützt auf Cochrans Beobachtung, dass fünf Subklassen bei vielen stetigen Verteilungen mindestens 90 Prozent der Verzerrung beseitigen („at least 90% of the bias for many continuous distributions“). Die Schätzungen der Bänder werden dann gemittelt. Gewichtet mit dem Anteil jedes Bandes an der Population, schätzt der Durchschnitt den Effekt für alle; Gordon et al.¹ gewichteten mit dem Anteil der Behandelten im Band an allen Behandelten („corresponding to the fraction of treated users in the stratum relative to all treated users“), was den Effekt auf die Behandelten ergibt. In der Praxis sind Bänder im Score nicht ganz einheitlich, sodass die direkt adjustierte Schätzung eine Restverzerrung enthalten kann („the directly adjusted estimate may contain some residual bias due to x“).

Gewichtung nutzt den Score, um Einheiten zu gewichten, statt sie zu Paaren oder Gruppen zu ordnen. Das einzige Gewichtungsverfahren in diesen Quellen ist die von Gordon et al.¹ geprüfte inverse Wahrscheinlichkeitsgewichtung mit Regressionsadjustierung (IPWRA), die den Propensity Score nutzt, um die Beobachtungen im Regressionsmodell zu gewichten („uses the propensity score to weigh observations in the regression model“): Aus den geschätzten Scores werden Gewichte gebildet, die den Zusammenhang zwischen Behandlungsstatus und Kovariaten kontrollieren sollen („The

estimated propensity scores are used to form weights to help control for correlation between treatment status and the covariates.“). Die Autoren zählen das Verfahren zu den doppelt robusten: Die Schätzung bleibt konsistent, auch wenn eines der beiden zugrunde liegenden Modelle, das Propensity-Modell oder das Ergebnismodell, falsch spezifiziert ist („the estimator is consistent even if one of the underlying models“). Gordon et al. nutzen es wie ihr Matching und ihre Stratifizierung für den Effekt auf die Behandelten: Die gewichtete Regression sagt voraus, was die Nutzer mit Werbekontakt ohne ihn getan hätten. Gewichtung ohne Regression prüft keine der beiden Quellen, und dieser Leitfaden vergleicht sie nicht. Die einfache Regressionsadjustierung kommt ohne Score aus und stützt sich dort, wo sich Behandelte und Unbehandelte stark unterscheiden, auf Extrapolation, die wahrscheinlich verzerrte Schätzungen liefert („relies heavily on extrapolation, which is likely to produce biased estimates of the treatment effect“). Stratifizierung auf dem Score mit einer Regression innerhalb jedes Bandes verbindet beides, sodass die Regression innerhalb der Bänder weniger extrapoliert („the within-stratum regression is less prone to extrapolate“). Rosenbaum und Rubin² warnen zudem, dass die Regression auf den Kovariaten nur dann verlässlich ist, wenn die lineare Diskriminante eng mit dem Propensity Score korreliert („cannot be relied upon to perform well unless the linear discriminant is highly correlated with the propensity score“).

Welche Variante am besten abschneidet, haben die Facebook-Daten nicht entschieden: Keine dominierte durchgängig („We do not find that one method consistently dominates“). In der Studie, die die Autoren zuerst durchgehen, schien es unerheblich, ob ein Verfahren ein Ergebnismodell nutzt („whether a method relies on an outcome model seems unimportant“).

Wie wird die Balance nach dem Matching geprüft?

§ 5 von 8

Indem man die Behandelten und die zugeordneten Unbehandelten am Score und an jeder Kovariate vergleicht, vorher und nachher. Rosenbaum und Rubin (1983)² werten verbliebenes Ungleichgewicht als Warnsignal: Anzeichen von Restverzerrung im Propensity Score sind Anzeichen möglicher Verzerrung der geschätzten Effekte („evidence of residual bias in the propensity score is evidence of potential bias in estimated treatment effects“). In ihrer Studie zur Bypass-Operation gegenüber medikamentöser Therapie bei koronarer Herzkrankheit verglichen sie beide Gruppen bei jeder von 74 Kovariaten mit einer F-Statistik, vor und nach der Bildung der fünf Bänder; der Median sank von 10,9 auf 0,2 (Tabelle 2, S. 52). Innerhalb der Bänder fanden sie mehr Balance, als bei zufälliger Zuteilung innerhalb jedes Bandes zu erwarten gewesen wäre („within the constructed subclasses there is greater balance than would have been expected if treatments had been assigned at random within each subclass“). Die Prüfung muss nicht aufwendig sein: Die Vergleichbarkeit innerhalb der Bänder lässt sich mit einfachsten Mitteln zeigen, etwa mit Balkendiagrammen der Mittelwerte („The comparability of units within subclasses can be verified by the simplest methods, such as bar charts of means.“).

Gordon et al. (2019)¹ prüften zweierlei. Für den Score betrachteten sie Histogramme der geschätzten Scores nach Gruppe („we inspect the histograms of the estimated propensity scores by treatment group“). Für die Kovariaten untersuchten sie standardisierte Differenzen zwischen Nutzern mit und ohne Werbekontakt, vor und nach dem Matching („we examine standardized differences in covariates between the exposed and unexposed users, before and after matching“), jeweils die Differenz der Mittelwerte geteilt durch die gepoolte Standardabweichung. Einer t-Statistik ziehen sie das vor, weil diese in großen Stichproben groß ausfallen kann, auch wenn der Unterschied in der Sache klein ist („might be large if the sample is large, even if the difference is substantively small“), und sie nennen eine von Stuart und Rubin vorgeschlagene Obergrenze von 0,25 Standardabweichungen. In ihrer Studie 4 drückte ein Score aus der ersten Variablengruppe Differenzen von bis zu etwa 0,4 Standardabweichungen unter 0,1, glich die übrigen Variablen aber kaum aus, bei denen manche Differenzen fast eine Standardabweichung erreichten („balancing on the FB variables does little to reduce the mean differences for the other variables, some of which have differences approaching 1 standard deviation“). Nach

diesem Befund gleicht ein Score verlässlich nur aus, was in ihn eingegangen ist; Variablen außerhalb des Modells können weit auseinander bleiben.

An der Balance scheiterten die Facebook-Studien nicht. Die Balancierung auf dem Propensity Score verringerte die Mittelwertdifferenzen in allen Studien deutlich und erreichte eine gute Balance der beobachteten Merkmale („Balancing on the propensity score substantially reduces the mean differences in all the studies and achieves good balance on the underlying observables“), und dennoch konnten viele der Beobachtungsmodelle den Effekt des Experiments nicht reproduzieren („Many of the observational models are unable to recover the RCT treatment effect, even though matching on the propensity score achieves good balance on the propensities themselves and on the underlying distribution of covariates.“). Balance zeigt, dass sich die gemessenen Kovariaten gleichen; die Annahme bedingter Unabhängigkeit verlangt zusätzlich, dass es keine unbeobachteten Größen gibt, die mit Behandlung und potenziellen Ergebnissen korrelieren („the unconfoundedness assumption requires not only balance on observables but also that no unobservables exist that might be correlated with treatment and potential outcomes“).

Was zeigte sich, als Matching an den Experimenten von Facebook gemessen wurde?

§ 6 von 8

Die Schätzungen lagen oft weit daneben, in beide Richtungen. Gordon et al. (2019)¹ nutzten 15 große, als randomisierte Experimente durchgeführte Werbekampagnen auf Facebook aus dem Jahr 2015 („15 large-scale advertising campaigns conducted on Facebook as RCTs in 2015“), gesammelt zwischen Januar und September 2015, alle in den USA und beschränkt auf Nutzer ab 18 Jahren. In jeder wurden die Nutzer der Zielgruppe per Zufall einer Testgruppe zugeteilt, die die Kampagne sehen konnte, oder einer Kontrollgruppe, die sie nie sah; die Kontrollgruppe bekam die Anzeigen zu sehen, die sie ohne die Kampagne gesehen hätte („shown the full distribution of ads they would have seen if the advertiser“). Zwei der Autoren wählten die Studien nach Kriterien aus, die sie für den Vergleich von Messverfahren geeignet machten, aus Handel, Finanzdienstleistungen, E-Commerce, Telekommunikation und Technologie, mit jeweils etwa 2 bis 140 Millionen Nutzern, und die Autoren halten fest, dass die Auswahl nicht für die gesamte Werbung auf Facebook steht und das auch nicht soll („the sample of studies is not representative of all Facebook advertising (nor is it intended to be)“). Die Ergebnisse werden als Lift angegeben: die zusätzlichen Conversions der Nutzer mit Werbekontakt in Prozent der Conversions, die sie ohne die Anzeigen gehabt hätten.

Um einen Werbetreibenden ohne Experiment nachzubilden, ließen die Autoren die Kontrollgruppe beiseite und verglichen innerhalb der Testgruppe Nutzer mit und ohne Werbekontakt. Dieser Vergleich ist verzerrt, weil der Werbekontakt nicht zufällig ist. Den Aktivitätsbias erklären die Autoren damit, dass ein Nutzer Facebook während der Kampagne besuchen muss, um die Anzeige zu sehen („activity bias arises because a user must visit Facebook during the campaign to be exposed“); das Ausspielungssystem lernt, welche Nutzer das Kampagnenziel am ehesten erfüllen, und bevorzugt sie („learns which types of users are most likely to meet the objective“); und die Gebote anderer Werbetreibender verschieben, wer welche Anzeige sieht. Sie prüften sechs Verfahren: exaktes Matching nach Alter und Geschlecht, Stratifizierung auf dem Score, Propensity-Score-Matching, Regressionsadjustierung, inverse Wahrscheinlichkeitsgewichtung mit Regressionsadjustierung und Stratifizierung mit Regression. Außer dem exakten Matching lief jedes auf vier ineinander geschachtelten Variablengruppen: gängigen Facebook-Variablen wie Alter, Geschlecht und Zahl der Freunde; knapp 40 Variablen aus dem US-Zensus und dem American Community Survey nach Postleitzahl („nearly 40

variables drawn from the most recent Census and American Communities Surveys“); Dezilen der Aktivität auf Facebook in der Woche vor der Studie; und einem Match-Score, einer zusammengesetzten Kennzahl, die Tausende Verhaltensvariablen bündelt („a composite metric of Facebook data that summarizes thousands of behavioral variables“). Es fehlte die eigene Ergebnishistorie des Werbetreibenden: Ergebnisse aus der Zeit vor der Kampagne, die Messdienstleister vom Werbetreibenden erhalten könnten, lagen den Autoren nicht vor („we do not observe precampaign outcome histories that vendors might have received from an advertiser“), sodass ihre Übung am ehesten der Lage bei der Neukundengewinnung entspricht („most comparable to what an advertiser would have at their disposal for new customer acquisition“).

Drei der Muster, die sie aus ihrer Übersicht ablesen: Die Beobachtungsverfahren überschätzten den Lift des Experiments meist („the observational methods we study mostly overestimate the RCT lift“), unterschätzten ihn in manchen Fällen aber auch signifikant („in some cases they can significantly underestimate RCT lift“); bei 7 der 14 Studien mit Kauf als Conversion lagen die Punktschätzungen durchgehend um mehr als den Faktor drei daneben („in 7 of the 14 studies with a checkout-conversion outcome are consistently off by more than a factor of three“); und für Registrierungen und Seitenaufrufe kamen die Verfahren dem Experiment näher als für Käufe. In Studie 4, dem Fall, den der Leitfaden zur Inkrementalität berichtet, kam Propensity-Score-Matching mit allen vier Variablengruppen zusammen auf 95 Prozent gegenüber 73 im Experiment (Abbildung 8, S. 21 der Articles-in-Advance-Fassung), ein Unterschied, den die Autoren für das Matching nicht testen konnten. In Studie 9, mit einem Lift von 2,4 Prozent im Experiment, lag die nächstgelegene Schätzung bei 1.306 Prozent („The closest lift estimate is 1,306% (RA4), a massive overestimate.“), was die Autoren nach eigener Vermutung teils auf eine Kontaktquote von nur 6,6 Prozent zurückführen. In Studie 15 unterschätzten alle Verfahren außer dem exakten Matching nach Alter und Geschlecht den Lift („Except for exact matching on age and gender, all methods underestimate lift.“).

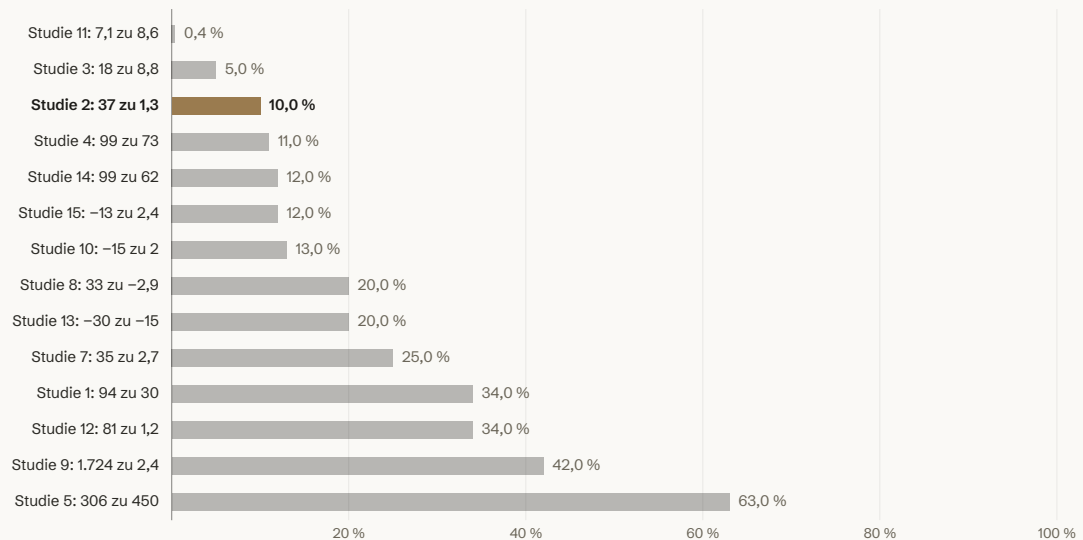
Der Maßstab hat sein eigenes Rauschen. Sechs der 14 Experimente mit Kauf als Conversion ergaben keinen auf dem 5-Prozent-Niveau signifikanten Lift („Of the 14 studies with a checkout conversion, six failed to produce statistically significant lifts at the 5% significance level“), und die Autoren warnen, dass Unterschiede zwischen Verfahren groß wirken können, wenn die Basis-Conversion-Rate der Behandelten klein ist („One downside of using lift is that differences

between methods can seem large when the treated group's baseline conversion rate is small.“). Für die 14 Studien mit Kauf als Conversion berichten sie außerdem, wie viel des Selektionseffekts nach der Stratifizierung auf dem Score übrig blieb: den Abstand zwischen stratifiziertem und experimentellem Lift als Anteil am Abstand zwischen dem rohen Lift aus dem Vergleich von Nutzern mit und ohne Werbekontakt und dem experimentellen Lift. Abbildung 1 zeigt diesen Anteil.

Die Autoren geben dieses Verhältnis für die Stratifizierung auf dem Score an, nicht für das Matching, und rechnen auch ihre Sensitivitätsanalyse mit der Stratifizierung, weil sie sich schnell genug über viele simulierte Ziehungen wiederholen lässt. Ihre Übersicht nach Studien stellt beide Verfahren nebeneinander: Mit allen vier Variablengruppen zusammen lag die Schätzung aus dem Matching in 10 der 14 Studien mit Kauf als Conversion höchstens vier Prozentpunkte von der stratifizierten entfernt, etwa 95 gegenüber 99 Prozent in Studie 4, und in keiner Studie mehr als 14 Prozentpunkte; den größten Abstand hatte Studie 9 mit 1.710 gegenüber 1.724 Prozent (Abbildung 10, S. 24 der Articles-in-Advance-Fassung).

Abbildung 1 Anteil der Verzerrung des rohen Vergleichs nach der Stratifizierung auf dem Propensity Score, 14 Facebook-Studien mit Kauf als Conversion

Die Stratifizierung auf dem Propensity Score ließ zwischen 0,4 und 63 Prozent der Verzerrung des rohen Vergleichs übrig. Auch ein kleiner Rest konnte den Effekt übersteigen: In Studie 2 waren 10 Prozent Rest 37 gegenüber 1,3.



Waagerechte Achse: Anteil der Verzerrung des rohen Vergleichs, der nach der Stratifizierung auf dem Propensity Score übrig blieb, von 0 bis 100 Prozent. Beschriftungen: Lift aus der Stratifizierung auf dem Score zu Lift im Experiment, in Prozent. Die 95-Prozent-Intervalle in Tabelle 4 des Aufsatzes schließen für die experimentellen Lifts der Studien 2, 7, 8, 9, 10 und 13 die Null ein. In den Studien 5, 10, 11, 13 und 15 liegt der stratifizierte Lift auf der anderen Seite des experimentellen Lifts als der rohe; der Anteil misst dort eine Überkorrektur, keinen Rest. Tabelle 6 nennt für Studie 7 einen stratifizierten Lift von 35; Abbildung 10 und Abbildung A.1 des Aufsatzes (S. 24 und 29 der Articles-in-Advance-Fassung) nennen -35 und -35,3, womit der Anteil dieser Studie bei etwa 30 Prozent läge, auf der Seite der Überkorrektur.

Quelle Gordon et al. (2019), Tabelle 6, S. 27 (Seiten der vorliegenden Articles-in-Advance-Fassung, S. 1 bis 33): das Verhältnis der Verzerrungsminderung für die 14 Studien mit Kauf als Conversion, in der Reihenfolge der Tabelle. Das Verhältnis ist der Abstand zwischen dem durch Stratifizierung auf dem Propensity Score geschätzten Lift und dem Lift im Experiment, geteilt durch den Abstand zwischen dem rohen Lift aus dem Vergleich von Nutzern mit und ohne Werbekontakt und dem Lift im Experiment. Die Beschriftungen nennen den stratifizierten und den experimentellen Lift aus derselben Tabelle, in Prozent. Die Verhältnisse stehen mit einer Nachkommastelle, weil die Tabelle für Studie 11 den Wert 0,4 angibt; die übrigen sind dort ganze Zahlen. Werte der Autoren, keine Neuschätzung.

Die Stratifizierung auf dem Score beseitigte in 13 der 14 Studien mit Kauf als Conversion mehr als die Hälfte der Verzerrung des rohen Vergleichs, doch der Rest konnte immer noch ein Vielfaches des Lifts im Experiment betragen. Das Beispiel der Autoren: Obwohl in Studie 2

bereits 90 Prozent der Verzerrung beseitigt waren, hätten sie Daten gebraucht, um den Rest von 37 auf 1,3 Prozent zu senken („eliminated 90% of the bias in study 2 (brr = 10%), we need data that would allow us to reduce the remaining bias from 37% to 1.3%“). In Studie 9 entsprachen 42 Prozent Rest einem Lift von 1.724 gegenüber 2,4 im Experiment.

Wie stark müsste eine fehlende Variable sein? Die Autoren simulierten eine unbeobachtete Variable, die die verbliebene Verzerrung beseitigt, und verglichen ihre Erklärungskraft mit der ihrer Variablen. In Studie 4 genügte eine relativ schwache unbeobachtete Variable („a relatively weak unobservable is required to remove the bias entirely“), etwa so stark wie die Zensusvariablen; in Studie 1 eine so starke wie ihre vierte Variablengruppe; in Studie 9 bräuchte sie die fünf- bis zehnfache Erklärungskraft ihrer beobachteten Variablen („the unobservable would need to have between 5 and 10 times as much explanatory power as the observables in our data“). Ihr Schluss: Die Verzerrung zu beseitigen, wäre selbst für Insider der Branche mit Zugang zu zusätzlichen Daten schwer („This suggests that eliminating bias from observational methods would be hard, even for industry insiders with access to additional data.“).

Was gehört zu einer Schätzung aus dem Matching?

§ 7 von 8

Sechs Felder, in einem Eintrag je Schätzung: das Estimand (die Schätzgröße), die Kovariaten und wann sie gemessen wurden, die Überlappung und was das Trimmen entfernt hat, die Balance jeder Kovariate vorher und nachher, die nicht gemessenen Ursachen, die Behandlung und Ergebnis zugleich bestimmen könnten, und der Maßstab: ein Experiment, an dem die Schätzung gemessen wurde, oder, wenn es fehlt, eine Sensitivitätsanalyse dazu, wie stark eine nicht erfasste Ursache sein müsste. Die Felder folgen den Bedingungen und Prüfungen von Rosenbaum und Rubin² und den Schritten, die Gordon et al.¹ gingen und berichteten. Der Beitrag zur Uplift-Spezifikation bietet ein Blatt dafür, worauf sich ein Uplift bezieht; dieser Eintrag ergänzt, was eine Schätzung aus dem Matching darüber hinaus braucht.

Abbildung 2 Ein Blatt für eine Schätzung aus dem Matching

Estimand, Kovariaten, Überlappung, Balance, nicht erfasste Ursachen und Maßstab festhalten, bevor eine Schätzung aus dem Matching als Effekt gilt.

Estimand	Kovariaten	Überlappung	Balance	Nicht erfasst	Maßstab
Effekt auf welche Behandelten, in welchem Zeitraum?	Vor der Behandlung gemessen? Beeinflussbar?	Wie viele Behandelte wurden getrimmt?	Standardisierte Differenzen vorher und nachher?	Was bestimmt Behandlung und Ergebnis ohne Feld?	Experiment oder Sensitivitätsanalyse, Ergebnis?

Ohne Experiment einen adjustierten Zusammenhang berichten und wie stark eine nicht erfasste Ursache sein müsste, um ihn aufzuheben.

Eine Zeile je Schätzung. Die Balance für jede Kovariate angeben, auch für die, die nicht ins Score-Modell eingingen.

Quelle Entscheidungsblatt des Autors. Die Felder für Estimand und Kovariaten folgen Rosenbaum und Rubin (1983), S. 42 und 43, und Gordon et al. (2019), S. 7 bis 10; das Feld für die Überlappung folgt Gordon et al. (2019), S. 10, 12 und 16; das Feld für die Balance folgt Rosenbaum und Rubin (1983), S. 48 und 52, und Gordon et al. (2019), S. 16 bis 19; die Felder für nicht erfasste Ursachen und den Maßstab folgen Gordon et al. (2019), S. 22 bis 27. Die Seiten von Gordon et al. sind die der vorliegenden Articles-in-Advance-Fassung. Die Felder sind Eingaben der Lesenden; es werden keine Unternehmensdaten geliefert.

Ein Beispiel ohne Unternehmensdaten: Das Marketingteam eines B2B-Softwareanbieters will wissen, was eine Webinarreihe für Kunden zum Erweiterungsumsatz beigetragen hat, und teilgenommen haben nur Accounts, die sich selbst dafür entschieden. Der Eintrag nennt als Estimand den Effekt auf die teilnehmenden Accounts in den zwei Quartalen nach der Teilnahme. Seine Kovariaten sind Lizenzen, Tarif, Kundendauer, Produktnutzung und Supportfälle aus dem Quartal vor der Einladung; die Nutzung nach der Teilnahme bleibt außen vor, weil die Webinare sie verändert haben könnten. Er hält fest, wie viele teilnehmende Accounts keinen vergleichbaren nicht teilnehmenden Account hatten und deshalb entfielen, sodass die Schätzung nur den Rest beschreibt. Er führt standardisierte Differenzen für jede Kovariate auf, auch für die, die nicht ins Score-Modell eingingen. Unter den nicht erfassten Ursachen nennt er Accounts, die bereits ein Upgrade prüfen:

Sie nehmen eher teil und erweitern eher, und kein Feld hält das fest. Solange kein Experiment gelaufen ist, nennt das Feld Maßstab stattdessen eine Sensitivitätsanalyse, und die Schätzung wird als adjustierter Zusammenhang berichtet, zusammen damit, wie stark eine nicht erfasste Ursache sein müsste, um ihn aufzuheben. Das Team könnte Einladungen für zwei Quartale per Los einem Teil der infrage kommenden Accounts vorenthalten, sofern Accounts ohne Einladung nicht teilnehmen können, den Erweiterungsumsatz beider Gruppen im Quartal davor als Ausgangswert erfassen und die Schätzung aus dem Matching mit der des Experiments vergleichen. Gordon et al.¹ rechneten einen solchen Vergleich in den Effekt auf die Behandelten um, indem sie ihn durch den Anteil mit Werbekontakt teilten; das setzt voraus, dass die Zuteilung nur über die tatsächliche Behandlung wirkt („only through receipt of the treatment“). Ihre Nutzer kannten ihre Zuteilung nicht; eine Einladung ist dagegen selbst eine Botschaft an den Account. Das Team müsste also begründen, dass eine Einladung ohne Teilnahme den Erweiterungsumsatz nicht verändert, oder den Effekt der Einladung, den Intention-to-Treat-Effekt, für sich berichten. Bei einigen hundert Accounts werden beide Schätzungen weit ungenauer sein als in Studien mit Millionen Nutzern. Das sind Beispiele des Autors, sie sind nicht geprüft.

Welche Teile des Facebook-Befunds lassen sich auf andere Daten übertragen?

§ 8 von 8

Die Autoren ziehen die Grenze selbst: Ihre Ergebnisse seien nicht so zu lesen, als könnten Verfahren der kausalen Inferenz experimentelle Befunde nirgends reproduzieren („Of course, our results should not be interpreted to suggest that causal inference methods are unable to replicate experimental findings in all settings.“). Sie berichten, dass Eckles und Bakshy, die mit Facebook-Daten Peer-Effekte statt Werbung maßen, mit hochdimensionalen Modellen den größten Teil der Verzerrung naiver Vergleiche beseitigten, und dass die Verfahren bei Werbung, die weniger gezielt ausgespielt wird als auf Facebook, funktionieren könnten („may work to recover the causal effects of advertising that is not subject to the same degree of targeting as advertising on Facebook“). In ihrem Fazit beschränken sie das Scheitern auf die Facebook-Daten, die ihnen vorlagen („the failure of observational methods is specific to the Facebook data we had at our disposal“), halten auf einer anderen Plattform andere Ergebnisse für möglich („performing this analysis on a different advertising platform may yield different results“) und konnten bei ihrer kleinen Zahl von Studien keine Merkmale von Kampagnen finden, die mit starken Verzerrungen einhergehen („With our small number of studies, we could not identify campaign characteristics that are associated with strong biases.“). Dem stehen ihre reichhaltigen Daten gegenüber und das geschlossene System von Facebook, das Selektionsprobleme leichter lösbar macht als viele andere Messaufgaben in der Werbung („makes solving selection issues an easier problem than many other ad-effectiveness applications“). Zwei der Autoren nennen Facebook als ihre Institution; die beiden anderen wurden laut Aufsatz für die Studie nicht von Facebook bezahlt und waren mit einem symbolischen Gehalt, das sie spendeten, bei Facebook beschäftigt.

Die Frage ist älter als die digitale Werbung. Gordon et al.¹ berichten, dass LaLonde Beobachtungsverfahren mit randomisierten Experimenten zu Beschäftigungs- und Weiterbildungsprogrammen verglich und viele dieser Verfahren die experimentellen Ergebnisse nicht reproduzierten, und dass eine spätere Anwendung verbesserter Verfahren auf dieselben Daten durch Imbens sie reproduzierte. Ihr Grund, in der Werbung größere Schwierigkeiten zu erwarten: Durch das Targeting sind die Endogenitätsprobleme digitaler Werbung womöglich gravierender („Because of targeting, the endogeneity problems associated with digital advertising are potentially more severe“). Rosenbaum und Rubin (1983)² liefern die Theorie des Verfahrens, mit einem Anwendungsbeispiel aus einer

Beobachtungsstudie zur koronaren Herzkrankheit und ohne experimentellen Maßstab.

Nach der Lesart dieses Leitfadens lassen sich zwei Punkte mit einiger Zuversicht auf B2B-Marketingdaten übertragen: Balance bei gemessenen Kovariaten kann nicht zeigen, dass ein Vergleich frei von Confounding ist, also von Verzerrung durch gemeinsame Ursachen, und eine Schätzung aus dem Matching braucht einen Maßstab, bevor sie eine Budgetentscheidung trägt; der Beitrag zum Attributionsmodell macht denselben Punkt für die Ergebnisse von Attributionsmodellen. Wie groß die Verzerrung ist, wenn Accounts statt Nutzer einander zugeordnet werden, wenn Kaufzyklen lang sind und wenn die Kovariaten frühere Käufe enthalten, die den Facebook-Daten fehlten, 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** Gordon, B. R., Zettelmeyer, F., Bhargava, N., & Chapsky, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Marketing Science*, 38(2), 193–225.
<https://doi.org/10.1287/mksc.2018.1135>
- 2** Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.
<https://doi.org/10.1093/biomet/70.1.41>

Zitierweise

Isoglu, S. (2026, 18. Oktober). Was ist Propensity-Score-Matching? Sinan Isoglu.
<https://isoglu.com/de/insights/journal/was-ist-propensity-score-matching/>
Online-Fassung:
<https://isoglu.com/de/insights/journal/was-ist-propensity-score-matching/>

Sinan Isoglu, MBA (Quantic)

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