Signifikanz – Die Lebenslüge der empirischen Wirtschaftsforschung

Zweck eines Ökonomiestudiums sei es, die Studenten davor zu bewahren, von Ökonomen hinters Licht geführt zu werden, schrieb die berühmte Ökonomin Joan Robinson schon 1955. Das Bonmot hat Konjunktur, seit die Volkswirtschaftslehre - gemeinsam mit anderen Sozialwissenschaften - in eine sogenannte Replikationskrise geraten ist. Damit ist gemeint, dass sich sehr viele experimentelle oder statistische "Beweise" als extrem brüchig erwiesen haben.

Die fehlende Wiederholbarkeit gilt auch für viele in den renommiertesten Fachzeitseitschriften abgedruckte, häufig zitierte Ergebnisse. Sie können oft selbst mit den gleichen Daten nicht reproduziert werden, oder die Ergebnisse fallen ganz anders aus, wenn andere Forscher die Untersuchung durchführen. Die Replikationskrise ist ein herber Rückschlag für die Ambition der Ökonomen, durch datengetriebene Beweisführung den "harten" Wissenschaften wie der Physik nachzueifern. Deren vielfach experimentell bestätigte Theorien gelten als verlässlich. Andrew Chang und Phillip Li von der Notenbank Federal Reserve stellten fest, dass die Ergebnisse von 60 Studien aus 13 renommierten Fachzeitschriften mehrheitlich nicht nachvollziehbar waren. Ihr vernichtendes Resümee: "Ökonomische Forschung ist meistens nicht replizierbar." Andere derartige Querschnittsuntersuchungen brachten ähnlich verheerende Ergebnisse. So hat der Wirtschaftsstatistikprofessor Walter Krämer empirische Arbeiten der Zeitschrift "German Economic Review" aus elf Jahren untersucht und in der Mehrzahl gravierende methodische Fehler gefunden.

Zu den prominentesten Opfern der aktuellen Replikationsbewegung, also des um sich greifenden Nachrechnens, wurde Harvard-Ökonom Ken Rogoff. Dessen von Austeritätsaposteln gern angeführte und verbreitete These, dass das Wachstum ab einer Staatsschuldenquote von 90 Prozent leidet, stellte sich als Ergebnis von Rechenfehlern und fragwürdiger Datenauswahl heraus.

Die Replikationskrise betrifft keineswegs allein die Ökonomen. In der medizinischen Forschung machte zuletzt die Entdeckung Schlagzeilen, dass die jahrzehntelang als gesichert geltende Schädlichkeit von Cholesterin auf unsauberen wissenschaftlichen Methoden beruhte - die viel mit finanziellen Anreizen zu tun hatten. Noch intensiver wird die Replikationskrise in der Psychologie diskutiert, wo reihenweise als empirisch abgestützt und verlässlich geltende sozialpsychologische Theorien über den Haufen geworfen wurden.

Dass die Ökonomen - anders als die Psychologen - erst jetzt ernsthaft über die Krise ihrer empirischen Forschung zu diskutieren beginnen, liegt für Andrew Gelman, Professor für Statistik und Politik an der Columbia University, an einer geringeren Bereitschaft, problematische Entwicklungen offen zu diskutieren. Je stärker finanzielle Anreize auf die Wissenschaftler einwirkten, was in der Psychologie weniger der Fall sei, desto stärker sei das Bestreben, negative Ergebnisse zu unterdrücken und Kritiker zum Schweigen zu bringen oder zu ignorieren.

Eine Studie macht noch keinen Beweis

Statistik-Professor Roger Peng, von der Johns Hopkins Universität, sieht die Ursache der Krise in einem überzogenen Vertrauen in die Beweiskraft einzelner statistischer Untersuchungen. Die Epidemiologie führt er als Gegenbeispiel an. Hier werde einem Ergebnis erst getraut, wenn es von verschiedenen Forschern an unterschiedlichen Orten mit unterschiedlichen Stichproben bestätigt wurde. In der Ökonomie gilt bis zum Nachweis des Gegenteils als Stand der Wissenschaft, was einmal von einem Forscherteam in einem bestimmten Kontext mit einem Datensatz empirisch festgestellt wurde. Der Nachweis des Gegenteils bleibt in

aller Regel aus, weil man damit keine Meriten verdienen kann.

Das System, mit dem über die Qualität von Studien und damit auch Karrieren geurteilt wird, prämiert positive Ergebnisse. Die wichtigste Währung sind Veröffentlichungen in führenden Fachzeitschriften. Um dort publiziert zu werden, muss man viele Mitbewerber aus dem Feld schlagen. Das geht nur, wenn man interessante und vermeintlich durchschlagende Ergebnisse reklamiert. Die Fachzeitschriften wiederum stehen in scharfer Konkurrenz um aufsehenerregende Artikel, die viel zitiert werden. Denn welche Fachzeitschriften als "führend" gelten, bestimmt sich nach Zitationsrankings. Mit dem Abdruck eines kritischen Kommentars zu einem vielzitierten Aufsatz macht sich eine Zeitschrift selbst das Zitationsranking kaputt. Die Bereitschaft dazu ist gering.

Ein vielsagendes Beispiel ist eine Studie, die 2007 den Ehrenplatz des Aufmachers im "Journal of Political Economy" (JPE) bekam, eine der fünf führenden Zeitschriften. Er wies anhand des Zusammenhangs der Anzahl deutscher Schüler in Ferien und Albumverkäufen nach, dass Tauschbörsen am Niedergang der CDs unschuldig seien. Erst neun Jahre später verteidigen sich die Autoren, Felix Oberholzer-Gee und Koleman Strumpf, die ihre Daten nicht offenlegten, in der nachrangigen Zeitschrift "Information Economics and Policy" gegen die immer lauter werdenden Vorwürfe, ihre Daten seien völlig unplausibel. Das JPE jedoch sah nie einen Grund, Kommentare zu diesem fragwürdigen Artikel abzudrucken. Dabei wurde der damalige Herausgeber, Steve Levitt, schon vor Abdruck auf massive Probleme hingewiesen.

Wer nicht früh verinnerlicht, wie man aufsehenerregende Ergebnisse produziert, wird in diesem System gar nicht erst Professor. Und so wird das ganze Arsenal der Techniken angewandt und toleriert, das vielzitierte Ergebnisse verspricht. "Data-Mining"-Spezialisten lassen Computerprogramme massenhaft Datenreihen analysieren und denken sich, wenn sie interessante Muster entdeckt haben, hinterher eine Theorie dazu aus. Dann erzählen sie die Geschichte umgekehrt - als hätten sie eine vorher entwickelte Hypothese anhand der Daten gezielt überprüft. Man nennt das "HARKing" - für Hypothesizing After Results are Known. Oder es wird so lange mit verschiedenen Daten und statistischen Methoden herumprobiert, bis sich das gewünschte Ergebnis einstellt, ohne später zu berichten, wie viele vorher ausprobierte Varianten das gewünschte Ergebnis nicht erbrachten.

„Angesichts der von der Wissenschaft selbst diagnostizierten Reproduktionskrise ist es nachvollziehbar, wenn die Öffentlichkeit angeblich wissenschaftlich untermauerten ‚statistisch signifikanten' Ergebnissen nicht traut", räumen die Ökonomen Norbert Hirschauer, Oliver Mußhoff und Sven Grüner im aktuellen Wirtschaftsdienst ein. Unter dem Titel „False Discoveries und Fehlinterpretationen wissenschaftlicher Ergebnisse" fragen sie, ob es überhaupt moralisch vertretbar wäre, über bessere Kommunikation Vertrauen in einen nicht vertrauenswürdigen Gegenstand herzustellen:

„Ist das Problem der Wissenschaftskommunikation also gar nicht in erster Linie ein Problem der Kommunikation, sondern ein Problem der Wissenschaft – einer Wissenschaft, die immer wieder Aufsehen erregende „statistisch signifikante Studienergebnisse" produziert, der Überprüfung (Reproduzierbarkeit) von Einzelstudien aber zu wenig Aufmerksamkeit schenkt und die Nicht-Bestätigung vorher gefeierter Ergebnisse oft „laut" beschweigt?"

Der Kult der statistischen Signifikanz

Ein gern gemachter logischer Fehler besteht darin, aus der Tatsache, dass eine Nullhypothese nicht mit hinreichend hohem Signifikanzwert verworfen wird, zu schließen, dass sie (wahrscheinlich) stimmt.

Der umgekehrte Fehler ist noch häufiger: aus der erfolgreichen „statistisch signifikanten" Ablehnung einer Nullhypothese wird geschlossen, die bevorzugte These sei bewiesen. Kritiker sprechen von einem

"Signifikanzkult". Dazu beigetragen haben dürfte das Aufkommen billiger und mächtiger statistischer Softwarepakete, die es auch Wissenschaftlern mit geringem methodischen Know-how ermöglichen, entsprechende Auswertungen durchzuführen. Die American Statistical Association, die Vereinigung der amerikanischen Statistiker, sah sich von der Diskussion über dieses Phänomen Anfang 2016 zu einem für sie sehr ungewöhnlichen Schritt genötigt. In einer öffentlichen Erklärung warnte sie Wissenschaftler aller Fachrichtungen vor falschen oder missbräuchlichen Interpretationen von Signifikanztests. Stephen Ziliak, der daran beteiligt war, schreibt:

„Wir waren uns einig, dass die gegenwärtige Kultur der Signifikanztests, ihrer Interpretation und ihrer Kommunikation verschwinden muss.“

Eine Fachzeitschrift auf dem Gebiet der Psychologie ergriff sogar die drastische Maßnahme, Beiträge, die auf Signifikanzwerte Bezug nehmen, zu verbannen.

Auf empirisch arbeitende Ökonomen scheint diese Diskussion bisher jedoch wenig Eindruck zu machen. Das hat sechs deutsche Ökonomen um den Hallenser Agrarökonom Norbert Hirschauer veranlasst, auf dem Portal "Ökonomenstimme" einen vielbeachteten Artikel über die "Mangelhafte Rezeption der p-Wert Debatte in den Wirtschaftswissenschaften" zu veröffentlichen.

Der p-Wert ist ein Maß für die statistische Signifikanz, gern auch als "Irrtumswahrscheinlichkeit" bezeichnet. Dieser Ausdruck legt die falsche Interpretation nahe, dass bei einem p-Wert unter fünf Prozent, also einem signifikanten Ergebnis, das Gegenteil mit 95 Prozent Wahrscheinlichkeit zutrifft. Der Dortmunder Wirtschaftsprofessor Walter Krämer, seit längerem ein scharfer Kritiker des Signifikanzkults, zitiert sogar führende Lehrbücher, in denen diese falsche Interpretation verbreitet wird.

Um den Denkfehler verständlich zu machen, nutzt das Hirschauer-Team das Beispiel eines fünffachen Münzwurfs. Wenn die Münze nicht manipuliert ist (Nullhypothese), dann ist die Wahrscheinlichkeit, dass bei fünf Würfen fünfmal Kopf erscheint, nur rund drei Prozent. Das entspricht dem p-Wert. Es ist jedoch offenkundig, dass durch ein entsprechendes Ergebnis eines Münzwurftests nicht bewiesen wurde, dass die Münze manipuliert war. Man kann auch nicht folgern, dass die Münze mit 97-Prozent Wahrscheinlichkeit manipuliert ist. Um dazu etwas sagen zu können, braucht man viel mehr Informationen. Man müsste etwa wissen, mit welcher Wahrscheinlichkeit eine manipulierte Münze Kopf zeigt. Außerdem bräuchte man Kontextwissen. Hole ich die Münze zufällig aus dem Geldbeutel, werde ich kaum aus einer unwahrscheinlichen Wurffolge auf Manipulation schließen. Hat mir dagegen ein Hütchenspieler auf der Straße eine Münzwurf-Wette vorgeschlagen, schon eher.

Probleme von empirisch arbeitenden Ökonomen mit Signifikanztests sind nicht die Ausnahme, sondern die Regel. Wirtschaftsstatistiker Krämer hat eine große Anzahl in Deutschland veröffentlichter empirischer Arbeiten untersucht. Er stellte fest, dass mehr als die Hälfte den Fehler machten, dass sie ökonomisch signifikante und theoretisch plausible Zusammenhänge wegen schwacher statistischer Signifikanz verwarfen, oder dass sie umgekehrt Zusammenhänge in ökonomisch unbedeutender Größenordnung herausstellten, nur weil sie statistisch signifikant waren. Mehr als zwei Drittel unterließen die notwendige Überprüfung, ob das theoretische Modell, das den Signifikanzwerten zugrunde lag, auch korrekt war.

Immerhin lobt Krämer, dass es in den letzten Jahren stärker üblich geworden sei, die ökonomische Signifikanz relativ zur statistischen explizit zu diskutieren. Es gibt jedoch noch sehr viele Fehlerarten bei dieser Art von Signifikanztest, die weiterhin sehr häufig vorkommen.

Ein Strauß möglicher Modelle

Arbeitnehmer mit adlig klingendem Nachnamen wie Kaiser oder König haben Karrierevorteile. Das war eines

dieser schlagzeilenträchtigen Ergebnisse empirischer Wirtschaftsforschung, in diesem Fall von zwei Business-School-Professoren. Doch das Ergebnis schien Uri Simonsohn von der Marketing-Fakultät der Wharton Business School suspekt. Er kritisierte die Methode der Kollegen. Danach passierte etwas, was sonst so gut wie nie passiert: Die Autoren, Raphael Silberzahn und Eric Uhlmann, setzten sich mit Simonsohn auseinander und räumten in einem gemeinsamen Aufsatz mit diesem den Fehler ein.

Die beiden hatten gemessen, welcher Anteil der in der Datenbank Xing gelisteten Menschen mit Namen wie Baron oder Herzog eine Führungsposition bekleidet. Der Anteil war deutlich höher als in der Vergleichsgruppe derer mit den 100 am häufigsten vorkommenden Namen. Es stellte sich jedoch heraus, dass dafür allein eine Besonderheit der Datenbank bei selten vorkommenden Namen verantwortlich war. Tatsächlich ist der Führungskräfte-Anteil bei adlig klingenden Namen nicht höher als bei anderen.

Silberzahl und Uhlmann machten ein Forschungsprojekt aus dieser ernüchternden Erfahrung - und gewannen 29 Forscherteams für ein Experiment. Darin bekamen alle die gleiche Leitfrage: Werden dunkelhäutige Profifußballer durch Schiedsrichter diskriminiert? Das Ergebnis war bemerkenswert: Aus dem gleichen Datensatz ermittelten 19 der 29 Teams eine statistisch signifikante und ökonomisch relevante Diskriminierung, sechs einen relevanten, aber nicht statistisch signifikanten Effekt und vier gar keinen Effekt in ökonomisch relevanter Größe. Die Schätzergebnisse reichten von einem leichten Nachteil für hellhäutige Spieler bis zu einem dreifach erhöhten Risiko dunkelhäutiger Spieler, für ein Foul vom Platz gestellt zu werden; sie häuften sich aber im Bereich einer um rund 30 Prozent erhöhten Wahrscheinlichkeit.

Bei Kenntnis der Ergebnisse aller Teams kann man damit zu der Einschätzung kommen, dass eine moderate Diskriminierung im Profisport wohl stattfindet. Hätte nur eines der Teams die Untersuchung durchgeführt und veröffentlicht, hätte das Ergebnis recht zufallsgetrieben entweder lauten können, dass es keine Diskriminierung gibt, dass es eine moderate Diskriminierung gibt oder auch, dass massiv gegen dunkelhäutige Spieler diskriminiert wird.

Anders als im realen Forscherleben hatten die Teams keinen Anreiz, statistisch zu tricksen, um die Aufnahmehürden der Zeitschriften zu nehmen. "Die Tatsache, dass man so viele konkurrierende analytische Ansätze der Datenanalyse vertreten kann, sollte zu denken geben. Es bedeutet, dass es ein Fehler wäre, einzelnen Ergebnissen zu viel Bedeutung beizumessen", folgern die Autoren. Sie sind damit in guter Gesellschaft. Schon John Maynard Keynes machte wegen der unvermeidlichen Willkür bei der Festlegung des verwendeten statistischen Modells aus seiner Skepsis gegenüber der Ökonometrie keinen Hehl.

In der Praxis haben die Forscher einen sehr starken Anreiz, gezielt signifikante Ergebnisse zu produzieren. Denn sonst haben sie kaum Chancen bei den renommierten Zeitschriften. Wer will es den Forschern also verdenken, wenn sie von zwei oder drei Methoden, die man vertreten kann, das auswählen, das günstige Ergebnisse auswirft. Oder wenn sie den Untersuchungszeitraum etwas früher oder später beginnen lassen, falls sich das positiv auswirkt, oder wenn sie den Ausschluss oder Nichtausschluss von extremen Werten davon abhängig machen, wie es das Ergebnis beeinflusst. Berichtet werden die Ergebnisse danach auf jeden Fall so, als habe man aus sachlichen Gründen diese Spezifikation des Modells gewählt und keine andere. Doch die Maße für die statistische Signifikanz der Ergebnisse sind wertlos, wenn so vorgegangen wird.

Dass solches Vorgehen gang und gäbe ist, zeigt eine Untersuchung von Abel Brodeur und drei Co-Autoren mit dem schönen Titel „Star Wars: The Empirics strike back“. Sie haben 50.000 empirische Arbeiten aus Top-Zeitschriften untersucht. Knapp nichtsignifikante Ergebnisse waren stark unterrepräsentiert gegenüber signifikanten und klar nichtsignifikanten Ergebnissen. "Unsere Interpretation ist, dass die Forscher versucht sein könnten, die Signifikanzwerte aufzublasen, indem sie die Spezifikation aussuchen, die die höchsten Werte auswirft", schlussfolgern sie.

"Manchmal lassen Autoren etwas durchblicken, wenn sie etwa über die Ergebnisse ihrer 'bevorzugten Spezifikation' berichten", weiß Wirtschaftsstatistik-Professor Walter Krämer, der viele Aufsätze gesichtet hat. Eigentlich müssten Gutachter da hellhörig werden. Doch die Zeitschriften haben selbst ein Interesse an solide erscheinenden Ergebnissen und viele Gutachter Verständnis für Karrierenöte ihrer Kollegen. Es soll vorkommen, dass Gutachter vorschlagen, man möge andere Modellspezifikationen ausprobieren, um die statistischen Maßzahlen zu verbessern. Und Statistikprogramme bieten Funktionen an, die die "p-Hacking" genannte Praxis automatisieren.

Tricksen mit Pseudo-Tests

HARKing ist letztlich eine verschärfte Form des p-Hacking". Dabei nimmt ein Forscher viele alternative Datenreihen, zum Beispiel die Entwicklung verschiedenster Börsenindizes, und kombiniert diese mit anderen Datensätzen, etwa zum Wetter oder zu einer Vielzahl von Sportergebnissen. Ein Statistikprogramm sucht aus allen möglichen Kombinationen "statistisch signifikante" Auffälligkeiten heraus. Welch ebenso eindrucksvolle wie unsinnige Korrelationen sich damit zeigen lassen, sehen sie hier (Dank an Werner Menne). Wenn man genügend Datenreihen untersucht, findet man immer zufällige Muster, die aussehen wie statistisch valide Zusammenhänge. Zu diesen denkt sich der Forscher eine Theorie aus. Wenn man etwa bei einem der Börsen-Indexe einen scheinbaren Zusammenhang zu bestimmten Wetterphänomenen oder Sportergebnissen entdeckt hat, lässt sich daraus eine Hypothese über den Einfluss von Stimmungsschwankungen der Händler auf die Kurse basteln und "beweisen".

Abhilfe ist möglich aber unattraktiv

Hirschauer und Kollegen fordern: "Es sollte explizit offengelegt und unterschieden werden, ob es sich um eine explorative Studie zur Identifizierung von statistischen Zusammenhängen handelt, mit der Hypothesen generiert werden sollen, oder um eine Studie zur Überprüfung von Hypothesen." Im ersten Fall wären p-Hacking und HARKing unproblematisch. Das Problem sei, dass beides regelmäßig vermischt werde.

Im Idealfall würde der Prozess mehrstufig ablaufen. Ein Forscher sucht statistische Zusammenhänge mit plausiblen theoretischen Erklärungen und publiziert die Ergebnisse. Auf dieser früheren Stufe hält Krämer Signifikanztests für sinnvoll, um Erklärungsmodelle mit den Daten abzugleichen und zu verbessern. Das dürfte bis dahin nur als Theorie ohne empirische Bestätigung gelten. Andere Forscher würden dann in einem zweiten Schritt anhand verschiedener anderer Datensätze die angebotene Theorie überprüfen. Erst dann dürfte sie als empirisch fundiert gelten.

Im derzeitigen Publikationssystem für wissenschaftliche Erkenntnisse hat dieser aufwendige Prozess jedoch keinen Platz. Exploration und Überprüfung finden standardmäßig gemeinsam statt und machen dadurch regelmäßig die Beweisführung fragwürdig.

Weitere Literatur:

Walter Krämer (2011): The Cult of Statistical Significance – What Economists Should and Should Not Do to Make their Data Talk.

Deirdre McCloskey, Stephen Ziliak (1996): The Standard Error of Regressions

[9.3.17]