

Identifikation von Parametern (Parameter Identification)

Foundations • The Valuation Engineer

William Bert Craytor • October 1, 2026

Version 1.2.0

DOI: [10.67330/xjszh949](https://doi.org/10.67330/xjszh949)

Formale Definition. Ein Parameter gilt als identifiziert, wenn er durch die gemeinsame Verteilung der beobachtbaren Daten (d. h. wie oft jede Kombination aus Merkmalen und Preis unter allen derartigen Verkäufen auftritt) unter Berücksichtigung der Annahmen und der Spezifikation des Modells eindeutig bestimmt wird. Die Identifizierung erfolgt logisch vor der Schätzung und unterscheidet sich von dieser.

Anschauliche Einordnung. Es gibt bestimmte Preise, die sich aus den Verkaufszahlen überhaupt nicht ableiten lassen, und daran ändert auch keine Methode etwas. Wenn jedes renovierte Haus auf dem Markt in ausgezeichnetem Zustand ist und jedes Haus in ausgezeichnetem Zustand renoviert wurde, zeigen die Verkaufszahlen, was beide zusammen wert sind, und nichts über jedes für sich; eine Regression, die für jedes eine Quote angibt – wie man es tun würde, wenn beide fast immer zusammen auftreten –, gibt eine Aufteilung wieder, die der Markt kaum vorgenommen hat. Wenn jedes Vergleichsobjekt eine bessere Aussicht hat als das Bewertungsobjekt, wird der „Aussichtswert“ über das Ende der Daten hinaus interpretiert. Wenn die einzigen Verkäufe, die eine Veränderung der Marktbedingungen im Zeitverlauf zeigen, auch die einzigen Verkäufe sind, die eine Veränderung der Hypothekenzinsen zeigen, lassen sich die beiden nicht voneinander unterscheiden. Dies sind Fragen dazu, was die Verkäufe beinhalten, die vor jeder Anpassung gestellt werden, und die Antwort ist dieselbe, unabhängig davon, ob die Methode eine Rate über den gesamten Bereich eines Merkmals oder unterschiedliche Raten über verschiedene Abschnitte davon anpasst, wie es MARS (Multivariate Adaptive Regression Splines, Friedman 1991, die wichtigste Regressionsmethode des Journals) tut: MARS kann nur dort feststellen, wo sich eine Rate ändert, wo die Vergleichsobjekte den entsprechenden Kontrast aufweisen, um dies zu zeigen. Die erste Frage des Gutachters lautet daher nicht „Wie hoch ist die Rate?“, sondern „Können mir diese Vergleichsobjekte Aufschluss geben?“.

Für Leser, die sich für den Mechanismus interessieren: Der statistische Fachbegriff hierfür lautet „Identifikation“: die Frage, ob die gesamte Umsatzpopulation, sofern sie fehlerfrei gemessen würde, den Parameter überhaupt festlegen würde. Die Methode der kleinsten Quadrate macht den Extremfall sichtbar, indem sie eines von zwei Merkmalen weglässt, die sich exakt gemeinsam bewegen, und verbirgt den gefährlichen Fall, in dem sie sich fast gemeinsam bewegen und zwei Zahlen zurückkommen, die auf denselben wenigen Verkäufen beruhen. Der Beitrag behandelt beide Fälle sowie die Fälle der Extrapolation und der verzerrenden Zeiteffekte, mit denen das Raster täglich konfrontiert ist.

In diesem Beitrag geht es um Statistik: Er befasst sich damit, was ein Datensatz grundsätzlich aussagen könnte, bevor ein Schätzer (eine Methode zur Berechnung der Raten aus den Verkaufszahlen, wie z. B. OLS oder MARS) angewendet wurde, und richtet sich sowohl an diejenigen, die vergleichbare Datensätze zusammenstellen, als auch an diejenigen, die Anpassungen an Regressionsmodelle vornehmen.

Ein *Parameter* ist hier eine Zahl, die das Modell liefern soll – eine Rate oder ein Koeffizient, aus der bzw. dem eine Anpassung berechnet wird. Die Frage, die sich bei der Schätzung stellt, lautet: Was sollte ich auf der Grundlage der Stichprobe als meine beste Schätzung des Parameters

annehmen? Die Identifikation wirft eine vorgängige Frage auf: Wenn ich die gesamte Grundgesamtheit hätte – also jede Transaktion, die jemals auf diesem Markt stattfinden wird, fehlerfrei gemessen –, wäre der Parameter dann überhaupt bestimmt?

Lautet die Antwort „nein“, so ist der Parameter nicht identifiziert, und nichts, was danach unternommen wird, kann dies korrigieren. Ein verbesserter Schätzer macht keinen Unterschied, eine größere Stichprobe derselben Art von Verkäufen hilft nicht weiter, und ein ausgefeilteres Modell hilft ebenfalls nicht, es sei denn, es fügt eine Annahme hinzu, die anhand der Daten nicht überprüft werden kann (der unten beschriebene Kausalfall), und es könnte das Problem sogar verschleiern. Das Versäumnis, einen Parameter zu identifizieren, hängt mit der Fragestellung und den zugrunde gelegten Daten zusammen – also damit, was der Markt hervorgebracht und was der Gutachter erfasst hat – und nicht mit der Methode des Analysten.

Ein Identifikationsfehler lässt sich am einfachsten erkennen, wenn zwei Merkmale immer zusammen auftreten. Stellen Sie sich vor, jede Immobilie, die renoviert wurde, würde auch mit „ausgezeichnet“ (9,00) bewertet, und jede Immobilie, die mit „ausgezeichnet“ bewertet wurde, wäre auch renoviert worden; in diesem Fall wären die Mengen der renovierten und der als „ausgezeichnet“ bewerteten Immobilien identisch. Der Markt zahlt für diese Gruppe einen höheren Preis – für ein „ausgezeichnetes“, renoviertes Haus gegenüber einem „guten“ Haus, das nicht renoviert wurde. Ganz gleich, wie viele Verkäufe wir beobachten, können wir diesen Preisunterschied nicht auf den erfassten Zustand und den durch die Renovierung erzielten Mehrwert aufteilen, der durch die anderen Merkmale nicht erfasst wird, da in diesem Markt das eine niemals ohne das andere variiert. Wir kennen die Gesamtsumme; die einzelnen Bestandteile kennen wir nicht.

Wenn die Methode der kleinsten Quadrate dies im Fall exakter Kollinearität (ein Merkmal ist eine exakte lineare Funktion der anderen) deutlich macht, liegt das daran, dass die Designmatrix – also die Tabelle, die sich aus den Spalten der Regressoren (erklärenden Variablen) zusammensetzt, die der Schätzer verwendet – rangmangelhaft ist: Eine ihrer Spalten ist eine exakte Kombination der anderen. Die Software lässt daher einen Koeffizienten weg oder gibt einen fehlenden Wert zurück. Dies ist lediglich eine Art Höflichkeitsgeste. Gefährlich sind die Grenzfälle – jene, in denen die beiden Variablen *fast* immer gemeinsam variieren –, da in solchen Fällen die Parameter technisch identifizierbar sind, die Software zwei Zahlen liefert und beide auf der geringen Anzahl von Beobachtungen basieren, die das Muster durchbrechen. Sie sind zwar identifiziert; da jedoch nur wenige Verkaufszahlen vom Muster abweichen, sind sie so ungenau, dass die Daten sie in der Praxis nicht voneinander unterscheiden können. Hier ähneln sich ein Identifikationsproblem und ein Präzisionsproblem. Darauf läuft Multikollinearität (Merkmale, die sich fast exakt gemeinsam bewegen) aus Sicht der Identifikation hinaus.

Die Identifizierung kann selbst bei völliger Abwesenheit von Kollinearität fehlschlagen, wenn der betreffende Parameter kausal ist und es sich um Beobachtungsdaten handelt, es sei denn, es wird eine Annahme hinzugefügt, die anhand der Daten nicht überprüft werden kann. Das Beispiel aus Eintrag 009 (*Endogenität*), bei dem dieselben sechzehn Verkäufe aus einem realen Markt stammten, ist ein Fall dieser Art: Der Zusammenhang zwischen Renovierung und Preis lässt sich identifizieren – da er ein Merkmal der gemeinsamen Verteilung ist und eine große Stichprobe ihn genau schätzen wird –, während der kausale Effekt der Renovierung nicht identifiziert werden kann, da die beobachtbare Verteilung keine Informationen enthält, anhand derer sich der durch die Renovierung geschaffene Mehrwert von dem zusätzlichen Wert unterscheiden ließe, den die Häuser bereits hatten und der ihre Eigentümer dazu veranlasste, die Arbeiten durchzuführen. Um den kausalen Parameter zu identifizieren, muss eine Annahme getroffen werden, die sich anhand der Daten selbst nicht überprüfen lässt, und die Zuverlässigkeit der Schlussfolgerung hängt daher von der Glaubwürdigkeit dieser Annahme ab.

Wo Bewerter ihr begegnen. Die Anforderung der Einordnung in einen Vergleichsrahmen. Die Regel, wonach das Bewertungsobjekt in einen Vergleichsrahmen eingeordnet werden muss – das heißt, dass für jedes wichtige Merkmal sowohl über- als auch unter ihm liegende Vergleichswerte einbezogen werden müssen –, wird als Faustregel dargestellt und beantwortet zwei unterschiedliche Fragen, die man voneinander trennen sollte. Die erste betrifft die Identifizierung: Weisen diese Verkäufe hinsichtlich des Merkmals Unterschiede auf, die sich von den anderen unterscheiden? Wenn alle Vergleichsobjekte die gleiche Aussicht haben, lässt sich daraus kein Preis für die Aussicht ableiten. Die zweite Frage betrifft die Absicherung: Liegt das Bewertungsobjekt innerhalb des Bereichs, für den der Wert ermittelt wurde? Wenn jedes Vergleichsobjekt eine bessere Aussicht hat als das Bewertungsobjekt, mag zwar durchaus ein Wert existieren – aus diesen Verkäufen oder aus einem größeren Datensatz –, doch seine Anwendung auf das Bewertungsobjekt würde eine Extrapolation über die Daten hinaus bedeuten. Die Einordnung in eine Spanne sichert den zweiten Punkt ab; nur ein solcher Unterschied sichert den ersten ab. Ein in eine Spanne eingeordnetes Bewertungsobjekt kann sich immer noch in einem Bereich befinden, in dem sich zwei Merkmale nie voneinander unterscheiden, und ein nicht in eine Spanne eingeordnetes Bewertungsobjekt kann immer noch einen an anderer Stelle ermittelten Preis erhalten – dies ist die Abhilfe durch den „größeren Datensatz“, die der Abschnitt zur Vertretbarkeit zulässt.

Ein Beispiel für Merkmale, die innerhalb eines Teilmarktes nicht unabhängig voneinander variieren: ein Gebiet, in dem jedes Haus auf der Seite der Straße mit Aussicht über einen größeren Grundriss verfügt; Aussicht und Größe können in diesem Gebiet nicht als separate Faktoren betrachtet werden. Ein Gutachter, der für beide Faktoren separate Anpassungen vornimmt, liefert faktisch zwei Zahlen, die anhand der Gebietsdaten nicht voneinander unterschieden werden können, und wie sich der Gesamtbetrag auf diese beiden Faktoren aufteilt, hängt vom Ermessen des Gutachters oder von einem anderen Markt ab. Das mag zwar angemessen sein, ist jedoch etwas anderes als „marktbasiert“.

Alter, Zustand und Bauzeit. Diese Faktoren hängen oft eng miteinander zusammen. Ältere Häuser befinden sich in der Regel in einem schlechteren Zustand und wurden nach anderen Standards gebaut. Die Verwendung eines Anpassungsrasters mit getrennten Zeilen für Alter und Zustand legt nahe, dass der Markt diese Faktoren getrennt bewertet und dass die Daten zwischen ihnen unterscheiden können. Ob die vorliegenden Daten diese Unterscheidung tatsächlich treffen können, muss überprüft werden und darf nicht einfach angenommen werden – und oft bestehen die Daten diese Überprüfung nicht.

Zeit und Marktbedingungen. Hier verbergen sich zwei unterschiedliche Fehlerquellen. Wenn das Raster eine Zeile für die Marktbedingungen enthält und keine Zeile für die anderen Faktoren, die sich im Laufe des Zeitraums verändert haben – Zinssätze, Lagerbestände, ein neuer Arbeitgeber –, wird der Zeitkoeffizient zwar identifiziert, ist jedoch zusammengesetzt und entspricht nicht allein dem Effekt der Marktbedingungen: Die gesamte Veränderung wird auf die einzige verfügbare Bezeichnung übertragen, was dem Fall der ausgelassenen Variablen in Eintrag 008 entspricht (*Verzerrung durch ausgelassene Variablen*). Sind beide Zeilen vorhanden, zeigen jedoch nur die Umsätze, bei denen sich die Marktbedingungen geändert haben, auch eine Änderung der Zinssätze, ist die Aufteilung zwischen beiden nicht identifiziert – was bei diesem Eintrag der Fall ist.

Warum sie für die Belastbarkeit zählt. Wenn Sie die Identifikationsfrage stellen, müssen Sie dies vor der Schätzungsfrage tun, da die Reihenfolge entscheidend ist. Wenn ein Gutachter sich zunächst fragt: „Was lässt sich anhand meiner Daten unterscheiden?“, und erst danach: „Wie lässt sich dies am besten schätzen?“, dann sind die daraus resultierenden Anpassungen solche, die durch die Daten gestützt werden können. Wird die Identifikationsfrage hingegen zuletzt gestellt, führt dies zu verlässlichen Zahlen für Größen, die niemand hätte messen können.

Ein Parameter, der nicht identifiziert wurde, sollte als solcher angegeben werden. Wenn die Daten zwar einen kombinierten Effekt, nicht aber dessen Aufschlüsselung ergeben, ist es sinnvoll, die Anpassung nicht wegzulassen, sondern ehrlich darzulegen, wie sie aufgeschlüsselt wurde: nämlich, dass die Vergleichsgruppe diese beiden Effekte nicht voneinander trennen kann, dass die verwendete Aufteilung auf einem größeren Datensatz oder auf einer Einschätzung beruht und worin der kombinierte Effekt besteht, da es der kombinierte Effekt ist, den die Daten tatsächlich bestimmen. Der Leser kann dann die Aufteilung selbst beurteilen. Im Gegensatz dazu kann der Leser zwei Zahlen, die als vom Markt abgeleitet angegeben werden, nicht beurteilen, wenn tatsächlich nur deren Summe vom Markt abgeleitet ist.

Bei der Identifizierung treffen Methodik und USPAP aufeinander. USPAP steht für „Uniform Standards of Professional Appraisal Practice“ (Einheitliche Standards für die professionelle Wertermittlung). Danach ist der Gutachter verpflichtet, alle Informationen zu erheben, zu überprüfen und zu analysieren, die für glaubwürdige Ergebnisse der Wertermittlung erforderlich sind (Standardregel 1-4). Die Richtlinien der Kreditgeber bringen das Thema Anpassungen direkt auf den Punkt: Der „Selling Guide“ von Fannie Mae verlangt Anpassungen, um die Reaktion des Marktes auf Unterschiede zwischen den Immobilien widerzuspiegeln. Keine der beiden Regeln besagt, dass jede Zeile der Vergleichstabelle separat anhand der Vergleichsobjekte vor dem Leser identifiziert werden muss; die Reaktion des Marktes kann sich aus gepaarten Verkäufen außerhalb der Tabelle, aus einer breiteren Stichprobe oder aus einer offengelegten Beurteilung ergeben – die oben genannte Abhilfe. Nach Auffassung des Journals machen die Vorschriften die Frage der Identifizierung zwingend: Ein Gutachter, der eine marktgestützte Anpassung geltend macht, muss angeben können, welche Daten diese identifizieren. So formuliert ist die Frage überprüfbar – man kann die zugrunde gelegten Daten einsehen und prüfen, ob sie den Unterschied enthalten –, ohne den mathematischen Identifikationstest in einen Konformitätstest zu verwandeln.

Die Lösung liegt im Design, nicht in der Berechnung. Wenn ein Parameter nicht identifiziert wird, besteht die Abhilfe darin, die verwendeten Daten zu ändern – das heißt, den geografischen Umfang zu erweitern, den Zeitraum zu verlängern, die Eigenschaften zu finden, die das Muster durchbrechen – oder die Aussage anzupassen. Kein Schätzer kann Informationen wiederherstellen, die in den Daten nicht enthalten sind.

Durchgerechnetes Bewertungsbeispiel. Betrachten Sie den Pacifica-Datensatz mit 16 Immobilien unter Eintrag 008 (*Omitted Variable Bias*). Der Zustand wird hier anhand des „Condition Score“ des Journals erfasst: einer Einschätzung des Zustands des Hauses im Vergleich zum Zustand der umliegenden Häuser in seinem definierten Marktgebiet. Er basiert auf der Einschätzung des Gutachters hinsichtlich dieses Marktgebiets und sollte nicht wesentlich im Widerspruch zur Einschätzung anderer Gutachter stehen, die im selben Marktgebiet tätig sind. Der Wert reicht von 0,00 bis 10,00: Der Prozentsatz der Häuser im Marktgebiet, die sich in einem schlechteren Zustand befinden, geteilt durch zehn. Der Datensatz erfasst drei Stufen: durchschnittlich (5,00), gut (7,00) und ausgezeichnet (9,00), und die konstruierten Daten zahlen pro Punkt \$20,000. Wie in den Einträgen 008 (*Verzerrung durch ausgelassene Variablen*) und 009 (*Endogenität*) entspricht die Einbeziehung des Zustands in die Regression der traditionellen Bewertungsmethode, nicht der Methode des Journals, der RCA-basierten Bewertung (Eintrag 007, *Bedingte Erwartung*). Beschränken Sie den Datensatz auf die zehn Immobilien, die einem Muster entsprechen – „gut“ und nicht renoviert oder „ausgezeichnet“ und renoviert: Vergleichsobjekte A, B, C, D, G, I, K, M, N und O. Zwei Filter bilden diese Teilstichprobe, und sie erfüllen unterschiedliche Aufgaben. Wenn man E, H und J (das „ausgezeichnete“ Haus, das nie renoviert wurde, und die beiden „guten“ renovierten Häuser) herausnimmt, bleibt ein realistisches Muster übrig: So sieht ein Markt aus, wenn die Renovierung der Hauptgrund dafür ist, dass ein Haus die höchste Zustandsbewertung in den Daten erreicht (hier

9,00), wie es vorkommen kann. Das Außer-Betracht-Lassen der drei durchschnittlichen Häuser F, L und P ist künstlich und dient dazu, den Fehler exakt statt annähernd darzustellen; der annähernde Fall, der der gefährliche ist, wird nach der Tabelle behandelt. Betrachten Sie die zehn Immobilien als die gesamte Datenwelt des Gutachters – einen Teilmarkt, in dem Renovierung und ein erstklassiger Zustand niemals getrennt voneinander auftreten. Bezogen auf diese Grundgesamtheit ist die Aufteilung nicht identifizierbar; die gesamten sechzehn Objekte zeigen, dass der Fehler aus den Daten des Teilmarkts stammte, nicht aus dem Gesamtmarkt.

Innerhalb dieser Teilstichprobe stehen der erfasste Zustand und der Umbau in einer exakten Identität zueinander: „ $\text{cond} = 7 + 2 \cdot \text{remodel}$ “ für jede einzelne der zehn Eigenschaften. Bei den Anpassungen der vollständigen Spezifikation – Bruttowohnfläche (GLA), Grundstück, Aussicht, Zustand, Umbau – muss die Software eine der beiden kollinearen Spalten entfernen, und welche davon entfernt wird, entscheidet darüber, was ausgegeben wird. Beide Fälle werden gezeigt:

	Teilstichprobe (10)		Vollständige Stichprobe (16)
	Umbau gestrichen	Bedingung entfällt	
Schnittpunkt	–240,000	110,000	–30,000
GLA	600	600	600
Los	25	25	25
Anzeigen	100,000	100,000	100,000
Zustand (pro Punkt)	50,000	<i>fallen gelassen</i>	20,000
Umbau	<i>fallen gelassen</i>	100,000	60,000

Drei Punkte in dieser Tabelle verdienen besondere Beachtung.

Erstens ist weder der Zustandskoeffizient noch der Renovierungskoeffizient lediglich ungenau; je nachdem, welcher weggelassen wird, fehlt dieser, und der andere nimmt einen Wert an, der ausschließlich vom Weglassen abhängt. Die Entwurfsmatrix hat sechs Spalten (eine Spalte mit Einsen für den Schnittpunkt, die Bruttogeschossfläche (GLA), das Grundstück, die Aussicht, den Zustand und den Renovierungsgrad), aber nur fünf, die separate Informationen enthalten (Rang 5) und die Methode der kleinsten Quadrate liefert keine eindeutige Antwort: Jedes Paar $(\beta_{\text{cond}}, \beta_{\text{remodel}})$ auf der Geraden $\beta_{\text{cond}} + \frac{1}{2} \beta_{\text{remodel}} = 50,000$, jeweils mit seinem eigenen Schnittpunkt, führt zu gleichwertigen und genauen Anpassungen an den zehn Verkaufszahlen. Die beiden Teilstichproben-Spalten der Tabelle sind zwei Punkte auf dieser Geraden. Die Achsenabschnitte unterscheiden sich aus demselben Grund: Jeder ist der angepasste Preis eines Hauses ohne Wohnfläche, ohne Grundstück und ohne Aussicht – bei einem Zustandswert von Null bei den Anpassungen und in der anderen als nicht renoviert –, was auf kein Haus zutrifft und auch keine Marktrealität darstellt.

Zweitens hängt der Preisunterschied zwischen einem hervorragenden, renovierten Haus und einem guten, nicht renovierten Haus nicht von diesem Rückgang ab: \$100,000, gelesen als \$50,000 – ein Punkt über zwei Punkten in einer Spalte – und als \$100,000 – bei der Renovierung in der anderen Spalte –, und in beiden Fällen genau $2 \cdot 20,000 + 60,000$ – die beiden Punkte des erfassten Zustands plus alles, was die Renovierung mit sich brachte. Ein Leser, der die „\$50,000“ neben die „\$20,000“ des Marktes stellt und dies als Übertreibung bezeichnet, hat eine willkürliche Darstellungsweise des Ergebnisses (diejenige, bei der „Remodel“ weggelassen wurde) so gelesen, als handle es sich um eine Quote; die Spaltenüberschrift sagt jedoch, was es ist. Die zehn Verkäufe legen auf den Dollar genau die Preisdifferenz zwischen einem ausgezeichneten, renovierten Haus und einem guten, nicht renovierten Haus fest und sagen nichts darüber aus, wie sich diese Differenz aufteilt. Beachten Sie, dass es sich hierbei nicht um eine Verzerrung im Sinne von Eintrag 008 (*Verzerrung durch ausgelassene Variablen*) handelt. Der „\$100,000“ hat vollkommen Recht, was die Frage betrifft, die die Teilstichprobe beantworten kann; es handelt sich lediglich um eine andere Frage.

Drittens werden GLA, Lot und View mit ihren ursprünglichen Werten wiederhergestellt — 600, 25 und 100,000 — unverändert gegenüber der vollständigen Stichprobe. Die Identifikation erfolgt Parameter für Parameter. Ein Modell kann die meisten seiner Koeffizienten identifizieren und scheitert an drei — hier „Zustand“, „Renovierung“ und, mit ihnen, der Achsenabschnitt, dessen ausgegebener Wert nur davon abhängt, welche der beiden kollinearen Spalten die Software verwirft — und das Scheitern ist in den Koeffizienten, die erfolgreich waren, nicht erkennbar. Was die zehn Eigenschaften jedoch bestimmen, ist die kombinierte Preisdifferenz von zwei Punkten des erfassten Zustands und des Umbaus: \$100,000.

Der Kontrast zum Eintrag 008 (*Omitted Variable Bias*) ist der Kernpunkt dieses Eintrags. Dort waren die benötigten Informationen in den Daten vorhanden, und das Modell hat es versäumt, sie zu nutzen: Durch Einbeziehung des Remodel-Indikators wurde das Problem behoben. Hier ändert die Einbeziehung des „Remodel“-Indikators nichts, da die Informationen in der Teilstichprobe überhaupt nicht enthalten sind. Gleiche Spezifikation, gleicher Schätzer, gleiche Software — und ein Fehler ganz anderer Art, der nur durch Eigenschaften außerhalb der „ $\text{cond} = 7 + 2 \cdot \text{remodel}$ “-Linie behoben werden kann. Die sechs beiseite gelegten Objekte tun dies auf zwei verschiedene Arten. E, ein ausgezeichnetes, nie renoviertes Haus, sowie H und J, renovierte Häuser in gutem Zustand, trennen die beiden Spalten jeweils direkt: Jeder dieser Kontraste identifiziert die Trennung anhand der Verkaufszahlen selbst, allein aufgrund der Form des Modells, ohne Annahmen über Raten außerhalb des Datenbereichs. F, L und P, die durchschnittlichen Häuser, durchbrechen die Identität ebenfalls, trennen die Spalten jedoch nur durch eine Annahme: dass die von ihnen angegebene Zustandsrate zwischen 5,00 und 7,00 auch zwischen 7,00 und 9,00 gilt, wo sich jedes renovierte Haus befindet. Setzt man nur F, L und P wieder ein, gibt die Software erneut zwei Zahlen zurück — hier genau die generierten Werte, da die Daten kein Rauschen enthalten —, doch beide beruhen auf drei Verkäufen und auf dieser Annahme. Das ist eng verwandt mit dem Fall, den der Eintrag als „gefährlich“ bezeichnet: formal identifiziert, aber basierend auf den drei Verkäufen, die das Muster durchbrechen, und auf der Annahme, dass die Rate zwischen 5,00 und 7,00 auch zwischen 7,00 und 9,00 gilt. Bei echtem Rauschen würden die beiden Zahlen gemeinsam entlang der Linie $\beta_{\text{cond}} + \frac{1}{2}\beta_{\text{remodel}} = 50,000$ schwanken. Erst durch das Hinzufügen von E, H und J lässt sich die Abweichung von den reinen Verkaufszahlen erkennen. Die Korrektur war nie statistischer Natur.

References.

- Fisher, F. M. (1966). *The Identification Problem in Econometrics*. McGraw-Hill.
- Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1), 1–67.
<https://doi.org/10.1214/aos/1176347963>
- Koopmans, T. C. (1949). Identification Problems in Economic Model Construction. *Econometrica*, 17(2), 125–144.
- Manski, C. F. (1995). *Identification Problems in the Social Sciences*. Harvard University Press.

Cross-references: endogeneity; omitted variable bias; conditional expectation.

Addendum: Prolog Formalization

The following machine-readable companion (`010-identification.pl`) formalizes this entry in the journal’s Prolog knowledge base. It is also published as a browsable galley alongside this article.

```

% SPDX-License-Identifier: AGPL-3.0-or-later
% Copyright (C) 2026 William Bert Craytor. Published by The Valuation Engineer
% Journal (Pacific Vista Net). This program is free software: you can
% redistribute it and/or modify it under the terms of the GNU Affero General
% Public License as published by the Free Software Foundation, either version 3
% of the License, or (at your option) any later version. It is distributed
% WITHOUT ANY WARRANTY, without even the implied warranty of MERCHANTABILITY or
% FITNESS FOR A PARTICULAR PURPOSE; see the COPYING file beside it or
% <https://www.gnu.org/licenses/agpl-3.0.html>.

%% 010-identification.pl
%%
%% Prolog companion to Foundations Vol. 1, Issue 2, Entry 010
%% Title: Parameter Identification
%% Author: Bert Craytor
%% Concept: identification
%% Version: 1.2.0
%%
%% v1.1.0: condition on the Journal's 0.00-10.00 score (average 5.00, good
%% 7.00, excellent 9.00; 20,000 a point). The knowledge base carries the
%% score and the generating function in those units.
%%
%% The galley reconstructs the entry's identification failure from the
%% dataset itself: not_identified_in_subsample/0 proves the collinearity
%% rather than asserting it.

% The dictionary (Issues/tools/dictionary) is not loaded at run time, so
% that the file runs unchanged in Trealla, SWI-Prolog and Scyer. The
% declarations below are plain facts, read by the dictionary linter.
dictionary_release_pinned('dictionary-2026.1').

% --- From the dictionary -----
% term(identification, "Identification", econometric_method,
% "A parameter is identified when it is uniquely determined by the
% joint distribution of the observable data given the assumptions
% of the model; identification concerns what the data could reveal
% in principle and is logically prior to, and separate from,
% estimation.", published, 'dictionary-2026.1').
%
% depends_on(identification, conditional_expectation).
% depends_on(identification, endogeneity).
% -----

uses_terms([ identification, endogeneity, omitted_variable_bias,
               conditional_expectation, multicollinearity,
               hedonic_regression, comp_selection,
               paired_sales_analysis, adjustment, uspap,
               standards_rule ]).
uses_relations([depends_on/2, related/2, introduced_by/2,
               formalized_by/2, generalized_by/2]).

article_concept(identification).
article_issue(volume(1), number(2), year(2026)).
article_id('foundations.2026.010').

:- initialization(consult('../..kb/pacifica_comps_extended')).
% Scyer Prolog loads forall/2 and the list predicates on request;
% SWI-Prolog and Trealla have them built in. The guarded loads below
% let the file run unchanged in all three.

```

```

:- initialization(catch(use_module(library(lists)), _, true)).
:- initialization(catch(use_module(library(iso_ext)), _, true)).

local_term(rank_deficiency).
local_term(design_matrix).
local_term(bracketing).
local_term(estimand).

% =====
% THE DISTINCTION
% =====

asks(estimation,      "Given this sample, what is the best guess at the parameter?").
asks(identification, "Would the entire population pin the parameter down at all?").

prior_to(identification, estimation).

% What does NOT repair an identification failure.
does_not_repair(better_estimator).
does_not_repair(larger_sample).
does_not_repair(more_sophisticated_model).

% =====
% THE SUBSAMPLE ON WHICH REMODEL IS NOT IDENTIFIED
% =====
%
% Condition score from the dataset's three levels: average 5.00, good 7.00,
% excellent 9.00.
cond_score(P, S) :- attribute_value(P, cond_score, S).

% The generating function, from the knowledge base.
true_coefficient_score(K, V) :- true_coefficient(K, V).

% Restrict to good (7.00) or excellent (9.00) properties, in a submarket
% where every excellent house was remodeled and no good one was (the three
% average homes are set aside to keep the illustration exact). Within the
% restriction, cond = 7 + 2*remodel identically, so the two columns are
% perfectly collinear given an intercept.

in_subsample(P) :-
    property(P),
    cond_score(P, C),
    attribute_value(P, remodel_numeric, R),
    ( C == 9.0 -> R == 1 ; C == 7.0, R == 0 ).

subsample(Ps) :- findall(P, in_subsample(P), Ps).

%! collinear_in_subsample is semidet.
%
% Proves cond = 7 + 2*remodel at every property of the subsample. This is
% the identification failure, derived from the data rather than asserted.

collinear_in_subsample :-
    forall(in_subsample(P),
        ( cond_score(P, C),
          attribute_value(P, remodel_numeric, R),
          C == 7 + 2*R ) ).

%! not_identified_in_subsample(?ParamA, ?ParamB) is semidet.

```

```

not_identified_in_subsample(cond_score, remodel_numeric) :-
    collinear_in_subsample.

% The six properties that BREAK the pattern. Two kinds: E, H and J separate
% the columns directly (an excellent house never remodeled; remodeled houses
% at good); F, L and P separate them only under the assumption that the
% condition rate from 5.00 to 7.00 holds from 7.00 to 9.00.
breaks_pattern(P) :- property(P), \+ in_subsample(P).
direct_contrast(P) :- breaks_pattern(P), cond_score(P, C), C >= 7.0.
linearity_contrast(P) :- breaks_pattern(P), cond_score(P, C), C < 7.0.

% =====
% WHAT THE SUBSAMPLE DOES AND DOES NOT DETERMINE
% =====

% Fitted on the subsample (n = 10) with the remodel column dropped; see
% entry.tex. With the condition column dropped instead: intercept 110,000,
% remodel 100,000, the same three rates. Both lie on
%  $\beta_{\text{cond}} + \beta_{\text{remodel}}/2 = 50,000$ .
subsample_coefficient(intercept,      -240000).
subsample_coefficient(gla_sqft,       600).
subsample_coefficient(lot_sqft,       25).
subsample_coefficient(view_numeric,   100000).
subsample_coefficient(cond_score,     50000). % per point
% remodel_numeric: not estimable (rank deficiency).

not_estimable(remodel_numeric, subsample).
alternative_normalization(condition_dropped, intercept, 110000).
alternative_normalization(condition_dropped, remodel_numeric, 100000).

%! normalizations_agree_on_the_difference is semidet.
% 2 points * 50,000 = 100,000 on the remodel = 2*20,000 + 60,000.
normalizations_agree_on_the_difference :-
    subsample_coefficient(cond_score, S),
    alternative_normalization(condition_dropped, remodel_numeric, M),
    2 * S =:= M,
    true_coefficient_score(cond_score, A), true_coefficient_score(remodel_numeric, B),
    M =:= 2 * A + B.

%! identified_sum_is_correct is semidet.
%
% The subsample's condition coefficient, per point, equals the recorded
% condition's rate plus the remodel's 60,000 spread over the two points
% that separate good from excellent: 20,000 + 60,000/2 = 50,000. Across
% those two points, 100,000 = 2*20,000 + 60,000, the price difference the
% data determine and cannot divide.

identified_sum_is_correct :-
    subsample_coefficient(cond_score, S),
    true_coefficient_score(cond_score, A),
    true_coefficient_score(remodel_numeric, B),
    S =:= A + B / 2,
    2 * S =:= 2 * A + B.

%! identification_is_parameter_by_parameter is semidet.
%
% GLA, lot and view are recovered at their generating values in the
% subsample even though two other parameters are not identified.

```

```

identification_is_parameter_by_parameter :-
    forall(member(R, [gla_sqft, lot_sqft, view_numeric]),
        ( subsample_coefficient(R, X), true_coefficient_score(R, X) )).

% =====
% CONTRAST WITH ENTRY 008
% =====

failure_kind(omitted_variable_bias,
    "The information is present in the data and the model fails to use it. Including the variable
    ↪ fixes it.").
failure_kind(identification,
    "The information is absent from the data relied upon. Including the variable fixes nothing.").

remedy(omitted_variable_bias, include_the_variable).
remedy(identification,      change_the_data_or_the_claim).

% =====
% APPRAISAL ENCOUNTERS
% =====

encounter(bracketing,
    "Bracketing answers two questions: contrast varying apart from the other characteristics (is
    ↪ the rate identified at all?) and support (does the subject lie inside the range where it was
    ↪ found?). Bracketing guards support; only such contrast guards identification.").
encounter(tract_subdivision,
    "Where every view-side home is also the larger plan, view and size cannot be separated within
    ↪ that tract.").
encounter(age_and_condition,
    "Age, condition and era of construction often move together; separate grid lines assert a
    ↪ separation the data may not support.").
encounter(time_adjustment,
    "A market-conditions coefficient is identified but composite: it attributes to time everything
    ↪ else that changed over the period, and is not the effect of market conditions alone.").

% Near-misses are the dangerous case: formally identified, practically not.
dangerous_case(near_collinearity,
    "Two coefficients technically identified but determined by the handful of observations that
    ↪ break the pattern; this is multicollinearity seen from the identification side.").

% =====
% DEFENSIBILITY
% =====

defensibility_consequence(order_of_questions,
    "Ask what the data can distinguish before asking how best to estimate it; the reverse order
    ↪ produces confident numbers for unmeasurable quantities.").
defensibility_consequence(report_the_sum,
    "Where two effects are not separately identified, report the combined effect the data do
    ↪ determine and disclose the basis of any split applied.").
defensibility_consequence(uspap_link,
    "In the Journal's reading, USPAP SR 1-4 (credible results) and the Fannie Mae Selling Guide (
    ↪ market-based adjustments) make the identification question obligatory -- which data identify
    ↪ this adjustment? -- without requiring each grid line to be identified from the comparables in
    ↪ front of the reader.").
defensibility_consequence(remedy_is_design,
    "Widen the geography, extend the period, or find the properties that break the correlation; no
    ↪ estimator recovers absent information.").

```