

Bedingter Erwartungswert (Conditional Expectation)

Foundations • The Valuation Engineer

William Bert Craytor • October 1, 2026

Version 1.2.0

DOI: [10.67330/99hhbc69](https://doi.org/10.67330/99hhbc69)

Formale Definition. Der bedingte Erwartungswert eines Ergebnisses (hier: Verkaufspreis) bei gegebener Menge von Prädiktoren (den gemessenen Merkmalen) ist der Mittelwert der Verteilung des Ergebnisses bei jedem festen Wert der Prädiktoren. Betrachtet man sie als Funktion der Prädiktoren, wird sie als bedingte Erwartungsfunktion (CEF) bezeichnet, und es ist diese Funktion, die die Regression zu approximieren versucht. Die Regression nach der Methode der kleinsten Quadrate (OLS) liefert die beste lineare Approximation der CEF und gibt die CEF exakt wieder, wenn die CEF selbst linear ist.

Anschauliche Einordnung. Eine aus einer Regression abgeleitete Anpassung geht von einem Satz aus: dem Preis, den der Markt pro Quadratfuß, pro Badezimmer und pro Baujahr zahlt. In diesem Beitrag geht es darum, woher ein solcher Satz stammt und worauf er sich bezieht. Die kurze Antwort lautet: Es handelt sich um die Steigung des durchschnittlichen Marktpreises, wenn sich ein Merkmal ändert, während die anderen unverändert bleiben, und eine Regression ist eine Methode, diese Steigung anhand von Verkaufszahlen zu ermitteln. Die OLS-Regression misst eine Steigung für den gesamten Bereich des Merkmals; MARS (Multivariate Adaptive Regression Splines, Friedman 1991) misst eine Steigung für jeden Bereich, in dem der Markt unterschiedlich zahlt, und gibt einen Satz pro Bereich an. In beiden Fällen ist die Rate ein Maß für dasselbe: die Art und Weise, wie sich der Durchschnittspreis mit dem Merkmal verändert. Der Rest dieses Eintrags befasst sich mit genau diesem Thema: Was es ist, warum das Verkaufsvergleichsraster es ebenfalls misst und was es bedeutet, wenn zwei Regressionen auf denselben Verkaufszahlen zwei unterschiedliche Raten ergeben.

Für den Leser, der sich für den Mechanismus interessiert: Der „Durchschnittspreis bei variierenden Merkmalen“ hat einen Namen, nämlich die bedingte Erwartungsfunktion, und der Beitrag verwendet diesen Begriff von nun an. Es handelt sich um eine Fläche über dem Raum der Merkmale; eine Rate ist ihre Steigung in einer Richtung; OLS führt Anpassungen an der Ebene durch, die die Fläche am besten approximiert, und gibt sie nur dann exakt wieder, wenn die Fläche eine Ebene ist; MARS führt Anpassungen an einer Fläche aus flachen Teilen durch, die sich an den von ihm gefundenen Brüchen biegen, und kommt der tatsächlichen Fläche somit näher, wenn diese sich krümmt. Die acht Pacifica-Vergleichsobjekte aus Ausgabe 1 dienen als Beispiel.

In diesem Beitrag geht es eher um Statistik als um Recht: Er befasst sich damit, was sich mit einer Regressionsanalyse schätzen lässt, und richtet sich an Gutachter, die Regressionsanalysen durchführen oder die Vergleichswertmethode anwenden – die ebenfalls dieselben Schätzungen liefert.

Die in Ausgabe 1 für die Pacifica-Vergleichsobjekte durchgeführte Regression wurde verwendet, um aus ihren Koeffizienten (den Werten, die die Regression den Merkmalen zuordnet) implizite Preise abzuleiten; ein impliziter Preis ist der Dollarbetrag, den der Markt für eine Einheit eines Merkmals zahlt. Allerdings wurde nicht klar angegeben, was mit der Regression geschätzt wurde, und es lohnt sich, dies zu erwähnen, da fast alle Fragen, die in dieser Ausgabe folgen – bezüglich Verzerrung, Endogenität und Identifikation – Fragen sind, die sich auf den Unterschied zwischen dem Geschätzten und dem Gewünschten beziehen.

Regression wird typischerweise so dargestellt, dass Anpassungen an einer Gerade durch Minimierung der Summe der quadrierten Residuen vorgenommen werden (ein Residuum ist der Verkaufspreis abzüglich des Wertes, den die angepasste Gleichung ergibt) und anschließend die Steigung angegeben wird. Diese Beschreibung ist zwar hinsichtlich der zugrunde liegenden Arithmetik korrekt, jedoch irreführend, wenn es um das eigentliche Ziel geht. Die Berechnung ist nur das Mittel; das Ziel ist eine Größe, die den gesamten Markt beschreibt (eine Grundgesamtheitsgröße), nicht nur die vorliegenden Verkaufszahlen.

Die Notation lesen.

P	Verkaufspreis.
$\mathbf{Z}$	die Liste der gemessenen Merkmale einer Eigenschaft; eine solche Liste wird als <i>Vektor</i> bezeichnet.
$\mathbf{z}$	ein bestimmter Wertesatz für diese Liste, zum Beispiel 1.650 Quadratfuß, ein 7.200 Quadratfuß großes Grundstück, keine Aussicht, Zustand 7,00.
$E[\]$	der <i>Erwartungswert</i> : der Durchschnitt aller möglichen Werte, wobei jeder Wert mit seiner Wahrscheinlichkeit gewichtet wird.
$ $	„gegeben“.
$E[P \mid \mathbf{Z} = \mathbf{z}]$	der erwartete Preiswert unter den Immobilien, deren Merkmale mit $\mathbf{z}$ übereinstimmen.
ε	die Abweichung eines Verkaufs von diesem Erwartungswert.
u	Ein struktureller Fehler bei einem Verkauf: der Preis abzüglich des Betrags, den er gemäß dem kausalen Zusammenhang haben sollte.
$\hat{p}$	den Wert, den die angepasste Gleichung liefert.
$z_{\text{GLA}}, \dots$	Wert eines Merkmals: Bruttogeschossfläche (GLA) und Grundstücksfläche in Quadratfuß; Aussicht und Renovierung: 1 bei „Ja“, 0 bei „Nein“; Bewertung im Bereich von 0,00 bis 10,00.

Legen Sie den Merkmalsvektor fest: $\mathbf{z}$ — das heißt: 1.650 Quadratfuß, ein 7.200 Quadratfuß großes Grundstück, keine Aussicht und ein Zustand von „gut“ (7,00 auf der Zustandsskala des Journals, wie im Beispiel erläutert). Unter den auf dem Markt angebotenen Immobilien, die genau diesen Merkmalsvektor aufweisen, unterscheiden sich die Preise. Die Preise sind nicht identisch, da sich die Immobilien in Aspekten unterscheiden, die durch den Vektor nicht erfasst werden, da Käufer und Verkäufer unterschiedlich sind und da Verhandlungen nicht immer mit demselben Preis enden. Auf diesem Markt gibt es eine vollständige Preisverteilung, die dem Merkmalsvektor $\mathbf{z}$ entspricht. Der Mittelwert dieser Verteilung ist der bedingte Erwartungswert des Preises bei gegebenem $\mathbf{z}$, geschrieben als $E[P \mid \mathbf{Z} = \mathbf{z}]$; die allgemeine Definition stammt von Kolmogorov (1933). Der Durchschnitt der vorliegenden Verkaufspreise ist eine Schätzung dieses Erwartungswerts, nicht der Erwartungswert selbst.

Da sich die Merkmale der Immobilien über den gesamten Merkmalsraum erstrecken, bildet der bedingte Erwartungswert eine Fläche ab: An jedem Punkt befindet sich eine Zahl, nämlich der Durchschnittspreis der Immobilien, die diese Merkmale aufweisen. Diese Fläche wird als *bedingte Erwartungsfunktion* (oder CEF) bezeichnet. Sie ist ein Merkmal des Marktes und gehört nicht zu einem bestimmten Datensatz – sie würde auch dann existieren, wenn noch nie eine Immobilie bewertet worden wäre. Abbildung1 stellt sie für das laufende Beispiel dar.

Zwei Tatsachen machen die CEF zum naheliegenden Ziel der Schätzung. Erstens ist die CEF unter allen Funktionen von $\mathbf{z}$ jene, die den Erwartungswert des quadrierten Vorhersagefehlers minimiert: Wenn man einen Preis allein auf der Grundlage von $\mathbf{z}$ schätzen muss und anhand des quadrierten Fehlers bewertet wird, ist $E[P \mid \mathbf{Z} = \mathbf{z}]$ die beste Schätzung, die man vornehmen kann. Zweitens lässt sich jedes Ergebnis in der Form

$$P = E[P \mid \mathbf{Z}] + \varepsilon, \quad E[\varepsilon \mid \mathbf{Z}] = 0,$$

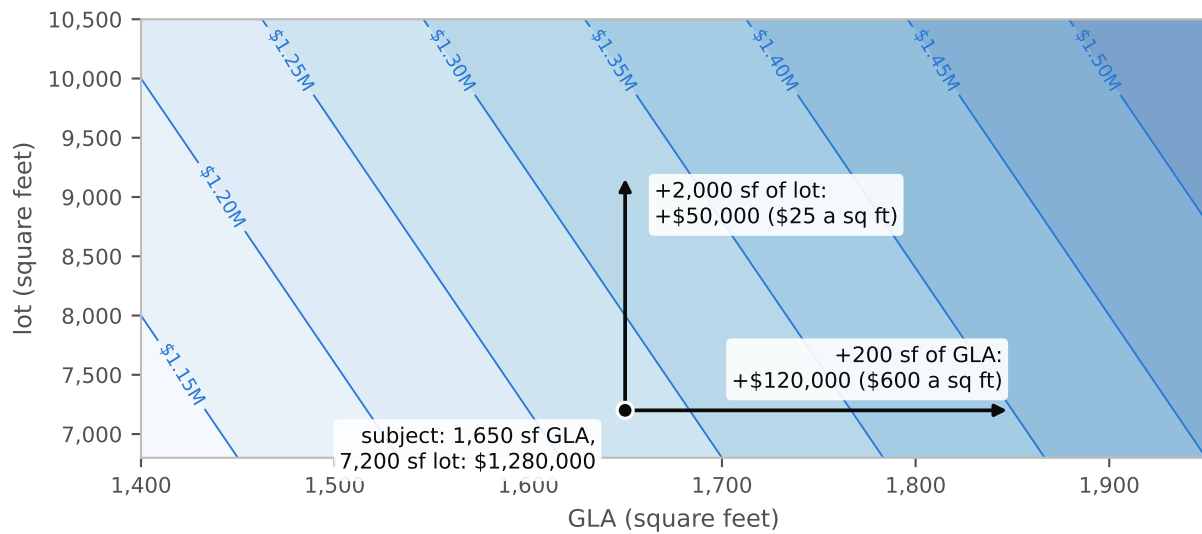


Figure 1: Die bedingte Erwartungsfunktion des konstruierten Pacifica-Marktes (Beispielrechnung) für Häuser ohne Aussicht, mit einem Zustandswert von 7,00 und ohne Renovierung, dargestellt wie eine topografische Karte: Jede Linie verbindet die Kombinationen aus Bruttomietfläche (GLA) und Grundstücksgröße, deren Durchschnittspreis identisch ist, und die Schattierung wird mit steigendem Durchschnittspreis dunkler. Der markierte Punkt ist das Haus dieses Eintrags: 1,650 Quadratfuß Bruttogeschossfläche (GLA) auf einem 7,200 Quadratfuß großen Grundstück, unter \$1,280,000 . Die beiden Pfeile stellen die beiden Steigungen dar, die den Prozentsätzen entsprechen: 200 zusätzliche Quadratfuß Bruttogeschossfläche erhöhen den Durchschnittspreis um \$120,000 , also um \$600 pro Quadratfuß; 2,000 zusätzliche Quadratfuß Grundstücksfläche erhöhen ihn um \$50,000 , also um \$25 pro Quadratfuß.

ausdrücken, wobei die zweite Gleichung keine Annahme, sondern eine Identität ist, da die Abweichung vom bedingten Mittelwert per Definition einen bedingten Mittelwert von Null hat (Angrist&Pischke 2009, Kap. 3).

Dieser Eintrag steht vor den Einträgen 008 (*Verzerrung durch ausgelassene Variablen*) bis 010 (*Parameteridentifikation*), da in jedem dieser Einträge die Frage gestellt wird, ob das, was eine Regression ermittelt, der betreffende CEF ist. Die Identität $E[\varepsilon | \mathbf{Z}] = 0$ selbst gilt immer, da ε als Abstand vom bedingten Mittelwert definiert ist. Was jedoch fehlschlagen kann, ist dieselbe Bedingung – ein bedingter Mittelwert von Null – für einen anderen Fehler: den strukturellen Fehler, der in Eintrag 009 (*Endogenität*) als „ u “ bezeichnet wird – die Differenz zwischen dem Preis und der kausalen Beziehung, die der Gutachter messen möchte. Dieser Fehler kann sich mit den Merkmalen verschieben. Dies geschieht beispielsweise, wenn ein Merkmal, das den Preis beeinflusst, unberücksichtigt bleibt; wenn der Preis auf ein Merkmal zurückwirkt – die Vermarktungsdauer (DOM) steigt in der Regel mit steigendem Preis, sodass eine Regression, die die DOM zur Erklärung des Preises heranzieht, einen Teil des preisinhärenten Effekts auf die DOM erfasst, weshalb nur wenige Analysten die DOM in eine Regression des Nettoverkaufspreises einbeziehen; oder wenn ein Merkmal fehlerhaft gemessen wird. Die *Verzerrung durch ausgelassene Variablen* (Eintrag 008) und *Endogenität* (Eintrag 009) befassen sich mit diesen Aspekten; die *Parameteridentifikation* (Eintrag 010) stellt die vorgängige Frage, ob die Verkaufszahlen die Rate überhaupt abbilden können.

Wo Bewerter ihr begegnen. Ein Gutachter, der angibt, dass „ein Haus mit einer Fläche von 1.650 Quadratfuß und ohne Aussicht in dieser Gegend für etwa \$1.28 million verkauft wird“, gibt damit eine bedingte Erwartung an. Das hier verwendete Wort ist „etwa“. Die Aussage behauptet nicht, dass jedes solche Haus für \$1.28 million verkauft wird; vielmehr bedeutet sie, dass \$1.28 million der Durchschnitt der Preise ist, zu denen Häuser dieses Typs verkauft werden. Durch die Verwendung von „etwa“ räumt der Gutachter ein, dass die einzelnen Preise um diesen Wert schwanken.

Ein Verkaufsvergleichsraster ist eine Methode zur Schätzung des Werts der CEF am Charakteristikvektor des Bewertungsobjekts; jeder Vergleichswert ist eine Beobachtung aus der Preisverteilung am Charakteristikvektor dieses Vergleichswerts; jede Anpassung verschiebt diese Beobachtung entlang der CEF-Fläche in Richtung des Vektors des Bewertungsobjekts; der abgeglichenen Wert liefert dann eine Schätzung der Höhe der CEF am Bewertungsobjekt. Strukturell betrachtet kann das Raster, wenn seine Anpassungen aus den Verkaufszahlen extrahiert werden, als einfacher nicht-parametrischer Schätzer (der keine Formel für die Oberfläche annimmt) eines bedingten Mittelwerts angesehen werden – einer, der lokale Mittelwertbildung und manuell festgelegte Steigungen anstelle von OLS verwendet; Anpassungen, die aus Kosten oder aus einer Einschätzung abgeleitet werden, sind strukturierte Meinungen und keine Schätzungen dieser Art.

Diese Art der Darstellung des Rasters bringt die unausgesprochenen Annahmen des Gutachters ans Licht. Wann immer Vergleichsen herangezogen werden, wird davon ausgegangen, dass die CEF glatt genug ist, damit nahegelegene Punkte für den Wert am Standort des Bewertungsobjekts relevant sind. Eine lineare Anpassung setzt voraus, dass die Steigung der CEF für dieses bestimmte Merkmal über den gesamten erfassten Bereich hinweg annähernd konstant ist. Die gleichgewichtete Berücksichtigung von drei Indikatoren setzt voraus, dass jeder den gleichen bedingten Mittelwert schätzt, keiner systematisch abweicht und alle eine ähnliche Genauigkeit aufweisen. Keine dieser Annahmen ist ungewöhnlich; sie alle betreffen die Form einer Oberfläche und werden alle zu Diskussionsthemen, sobald die betreffende Oberfläche benannt wurde.

Diese Sichtweise beseitigt zudem eine häufige Verwirrung bezüglich der Intervalle. Es gibt zwei Arten von Unsicherheit. Die eine betrifft die Lage des CEF: Wie weit die angepasste Oberfläche vom wahren Mittelwert entfernt sein könnte. Die andere betrifft die Streuung der einzelnen Preise um diesen Mittelwert herum. Ein Konfidenzintervall deckt nur die erste ab; ein Vorhersageintervall

deckt beide ab und ist daher immer das breitere der beiden. Bei MARS, der wichtigsten Regressionsmethode des Journals, liefert das „earth“-Paket (Milborrow 2026a) ein Vorhersageintervall für ein Bewertungsobjekt aus einem zweiten Modell, dem Varianzmodell (Milborrow 2026b), das abschätzt, wie sich die Streuung der Preise über den Markt hinweg verändert. Voraussetzungen sind, dass genügend Verkäufe vorliegen und dass die Fehler annähernd normalverteilt und voneinander unabhängig sind; das Paket gibt an, wie oft die Verkäufe innerhalb des Intervalls liegen, und dies dient als Überprüfung. Bei der RCA-basierten Wertermittlung entspricht der Marktwert des Bewertungsobjekts gemäß dem Vergleichswertverfahren dem vom Modell geschätzten Verkaufspreis zuzüglich eines vom Gutachter geschätzten Residuums. Das Varianzmodell liefert das Intervall, in das ein Verkauf des Bewertungsobjekts mit einer Wahrscheinlichkeit von 95 Prozent fallen würde, sofern diese Bedingungen erfüllt sind. Wenn ein Gutachter gefragt wird: „Wie sicher sind Sie?“, wird er in der Regel danach gefragt, wie weit sich die einzelnen Preise verteilen (das Vorhersageintervall), und er antwortet in der Regel, wie sicher er sich hinsichtlich des Durchschnitts ist (das Konfidenzintervall).

Die Benennung des Schätzgrößen – also des zu schätzenden Merkmals der Grundgesamtheit – ist ein Akt der Rechtfertigung, da eine Schlussfolgerung ohne diese Angabe nicht beurteilt werden kann. Die Aussage „Die Anpassung beträgt \$100,000“ kann noch nicht als richtig oder falsch angesehen werden. Erst wenn wir sagen: „Der Unterschied im bedingten durchschnittlichen Verkaufspreis zwischen Immobilien mit und ohne Aussicht, bei gleichbleibender Bruttowohnfläche (GLA), Grundstücksgröße und Substanz, beträgt \$100,000 in diesem Marktsegment“, stellen wir eine Behauptung auf, die hinsichtlich der konstant gehaltenen Variablen, der betreffenden Grundgesamtheit und der einbezogenen Einheiten angefochten werden kann.

Warum sie für die Belastbarkeit zählt. Daraus ergeben sich drei Konsequenzen hinsichtlich der Vertretbarkeit.

Die Wahl eines Konditionierungssatzes (die konstant gehaltenen Merkmale) ist eine methodische Entscheidung: Sie definiert den Schätzgrößenwert. Ob sich dieser Schätzgrößenwert anhand der Daten schätzen lässt, ist eine separate Frage, die Bewertungsobjekt der *Parameteridentifikation* ist. Die Ausdrücke $E[P \mid \text{GLA, view}]$ und $E[P \mid \text{GLA, view, condition}]$ stellen unterschiedliche Funktionen dar, die unterschiedliche Werte und unterschiedliche Steigungen aufweisen können. Wenn ein Gutachter eine Variable in das Raster aufnimmt, verbessert er nicht einfach nur die Schätzung, sondern verändert den Schätzgrößenwert. Es ist daher sinnvoll, die Konditionierungsmenge anzugeben, auf der die Wertermittlung beruht.

Der Ausdruck „unter sonst gleichen Bedingungen“ beschreibt die Argumentationsweise des Gutachters bei der Analyse und hat eine genau definierte Bedeutung. Im Rahmen des CEF-Konzepts bedeutet er den Vergleich der bedingten Mittelwerte bei zwei Merkmalsvektoren, die sich nur in einem Merkmal unterscheiden. Dies ist eine Aussage, die sich auf die Grundgesamtheit bezieht und anhand von Marktdaten überprüft werden kann, sofern diese in ausreichender Menge vorliegen. Die zugrunde liegende Analyse wird in der Regel nicht im Bericht, sondern in der Arbeitsdatei festgehalten – also in der Dokumentation des Auftrags, zu deren Führung der Gutachter gemäß der USPAP-Vorschrift zur Aufbewahrung von Unterlagen verpflichtet ist.

Die Residuen sind Informationen und kein Grund zur Verlegenheit. Die Streuung der Preise um die bedingte Erwartungsfunktion herum ist ein Merkmal des Marktes; in einem Markt, in dem die Preise weit um diese Funktion herum streuen, ist jede Schätzung mit einer einzigen Zahl – unabhängig davon, welche Methode verwendet wird – mit großer Unsicherheit behaftet. Die Angabe eines Wertes ohne Hinweis auf diese Streuung lässt die Schlussfolgerung präziser erscheinen, als es der Markt rechtfertigt.

Durchgerechnetes Bewertungsbeispiel. Die acht Vergleichspreise für Pacifica in Ausgabe 1

stellen eine „rauschfreie“ Stichprobe dar: Sie wurden so konstruiert, dass sie eine einzige lineare Gleichung exakt erfüllen. Wenn diese acht Preise auf Bruttogeschossfläche (GLA), Grundstücksgröße, Aussicht, Zustand und den Renovierungsindikator regressiert werden, lässt sich jeder Preis auf den Dollar genau reproduzieren – es bleibt nichts übrig – und die Regression liefert die Gleichung, auf deren Grundlage die Preise ermittelt wurden, Rate für Rate:

$$E[P \mid \mathbf{Z}] = -30,000 + 600 \cdot z_{\text{GLA}} + 25 \cdot z_{\text{lot}} + 100,000 \cdot z_{\text{view}} + 20,000 \cdot z_{\text{cond}} + 60,000 \cdot z_{\text{remodel}}.$$

Der Zustand bezieht sich hier auf den Zustandswert dieses Journals: eine 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 von der Einschätzung anderer Gutachter abweichen, 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. Ein Wert von 5,00 bedeutet, dass die Hälfte der Häuser im Marktgebiet in einem schlechteren Zustand ist; 9,00 bedeutet, dass neun von zehn Häusern in einem schlechteren Zustand sind. In Ausgabe 1 wurde der Zustand in Worten beschrieben und für die Regression mit den Werten 1–3 kodiert; bei der Bewertung dieses Journals entsprechen die Begriffe „durchschnittlich“ 5,00, „gut“ 7,00 und „ausgezeichnet“ 9,00, sodass eine Stufe aus Ausgabe 1 zwei Punkten entspricht und der Markt pro Punkt \$20,000 zahlt. Der letzte Term ist der Preisunterschied, den die konstruierten Daten einer Renovierung zuweisen: \$60.000 mehr als ein Haus mit derselben Bruttowohnfläche (GLA), demselben Grundstück, derselben Aussicht und demselben Zustandswert. Es handelt sich nicht um einen Preis für den Umbau selbst; er steht für alles, was der Umbau mit sich gebracht hat, was die anderen Merkmale nicht erfassen (Eintrag 008, *Omitted Variable Bias*).

Dies ist genau genommen der CEF für den konstruierten Pacifica-Markt. Es ist wichtig, klarzustellen, um welche Art von Objekt es sich bei der soeben gegebenen Beschreibung handelt und warum es so genau beschrieben werden kann. Der Datensatz wurde konstruiert, und ein konstruierter Datensatz kann tatsächlich rauschfrei sein; ein realer Datensatz ist dies niemals. Der Vorteil der Konstruktion besteht darin, dass zufälliges Rauschen vollständig aus den Preisen entfernt wird – jeder Preis entspricht exakt der Formel –, sodass in den Einträgen 008 (*Verzerrung durch ausgelassene Variablen*) bis 010 (*Parameteridentifikation*) kein Unterschied zwischen einer Schätzung und ihrem Zielwert auf Rauschen zurückzuführen sein kann: In jedem Fall handelt es sich um eine Verzerrung oder, bei der *Parameteridentifikation*, um einen Identifikationsfehler.

In diesem Beispiel wird der Zustand als Merkmal mit einem Dollarkurs behandelt, wie es bei der traditionellen Bewertung und in Ausgabe 1 der Fall war, und die Einträge 008 (*Verzerrung durch ausgelassene Variablen*), 009 (*Endogenität*) und 010 (*Parameteridentifikation*) zeigen, was dabei schiefgeht. Die traditionelle Bewertungsmethodik entspricht nicht der Methode der Zeitschrift, der RCA-basierten Bewertung. Beim Residual Constraint Approach berücksichtigt die Regression nur die Merkmale, die gemessen werden können; der Zustand und die Qualität werden bewusst außer Acht gelassen. Was die Regression vom Verkaufspreis ungeklärt lässt, ist der Residualwert. Auf den kalifornischen Märkten des Herausgebers erklären die gemessenen Merkmale in der Regel etwa 80 Prozent der Schwankungen bei den Verkaufspreisen, sodass etwa ein Fünftel auf den Residualwert entfällt. Der Gutachter teilt den Residualwert unter Berücksichtigung seiner Bewertungsnoten als Orientierungshilfe auf den Zustand, die Qualität und die anderen Merkmale auf, die in den Daten nicht gemessen werden, wobei eine Einschränkung gilt: Die Summe der Wertbeiträge der Residuen muss dem Residualwert des Vergleichsobjekts entsprechen. Die Wertfeststellung hängt nicht davon ab, wie der Gutachter diese aufteilt – Beweis II der RCA-Veröffentlichung (Craytor 2026, S. 36): Eine Änderung der Aufschlüsselung der Residuen unter der Einschränkung, dass sich die

Wertbeiträge aller Residuen zum Residuum des Vergleichsobjekts summieren, hat keinen Einfluss auf die endgültige Wertfeststellung. Der Zustandsfaktor wird daher niemals geschätzt und niemals angegeben; was der Gutachter dem Zustand zuordnet, ist der Zustand, der über das hinausgeht, was die gemessenen Merkmale – darunter das Alter – bereits erklären.

Vergleicht man die Regression aus Ausgabe 1 mit dieser CEF, lässt sich die Regression beurteilen – allerdings muss zuvor eine Unterscheidung getroffen werden. Allgemein gesprochen rekonstruiert eine OLS-Regression nicht die bedingte Erwartungsfunktion, auf die sie abzielt; stattdessen ermittelt sie die beste lineare Annäherung daran, d. h. die lineare Projektion, und diese beiden stimmen nur dann überein, wenn die CEF linear ist. Eine nichtparametrische Methode wie MARS benötigt keine lineare CEF: sie approximiert die CEF stückweise, mit einem Knickpunkt (einem Punkt, an dem sich die Rate ändert), an dem sich die Fläche krümmt, und einem Interaktionsterm (einer Rate, die von einem anderen Merkmal abhängt), an dem sie sich verdreht, und folgt somit einer nichtlinearen CEF, die eine lineare Projektion abflachen würde. Es handelt sich dennoch um eine Annäherung an dasselbe Objekt, und die in den nächsten Einträgen behandelten Fragen zu ausgelassenen Variablen, Endogenität und Identifikation gelten dafür in genau derselben Weise. Die oben erwähnte CEF mit fünf Variablen ist linear, weshalb die Regression, die den Remodel-Indikator einbezieht, diese exakt wiedergeben kann. Die Regression aus Ausgabe 1, die auf vier Merkmalen unter Ausschluss des Remodel-Indikators konditioniert ist, schätzt die beste lineare Annäherung an $E[P \mid \text{GLA, lot, view, cond}]$ und liefert – bei einer Neuformulierung der Kondition auf den Score des Journals (Ausgabe 1 gab einen Achsenabschnitt von 80,777 und 49,824 pro Schritt seiner 1–3-Kodierung; es handelt sich um dieselben Anpassungen),

$$\hat{p}(\mathbf{z}) = 6,041 + 575 \cdot z_{\text{GLA}} + 22,70 \cdot z_{\text{lot}} + 86,179 \cdot z_{\text{view}} + 24,912 \cdot z_{\text{cond}}.$$

Jeder Koeffizient unterscheidet sich von dem darüber liegenden. Sobald die Regression anhand der Anpassungen der Geraden beschrieben wurde, gibt es nichts weiter hinzuzufügen: Wir haben zwei Modelle und zwei Zahlenmengen. Im Rahmen des CEF-Ansatzes lassen sich jedoch zwei konkrete Aussagen treffen. Erstens basieren die beiden Regressionen auf unterschiedlichen Variablenmengen und approximieren daher unterschiedliche bedingte Erwartungsfunktionen:

$$E[P \mid \text{GLA, lot, view, cond}] \quad \text{and} \quad E[P \mid \text{GLA, lot, view, cond, remodel}].$$

In einem Markt, in dem einige renovierte und nicht renovierte Immobilien dieselben vier Merkmale aufweisen, haben die beiden Funktionen an diesen Punkten unterschiedliche Werte. In diesem kleinen künstlichen Markt weisen keine zwei Immobilien dieselben vier Merkmale auf, sodass die beiden Funktionen bei jeder einzelnen Immobilie übereinstimmen, sich jedoch in ihrer Form unterscheiden: Die zweite ist in ihren Merkmalen eine lineare Formel, die erste hingegen nicht. In jedem Fall liefert die Regression lediglich die beste lineare Annäherung an ihr Ziel. Beide Ziele sind gültige Objekte. Zweitens weist nur die Funktion mit fünf Variablen Koeffizienten auf, die als implizite Preise interpretiert werden können – wobei der Renovierungskoeffizient die Preisdifferenz zwischen einem renovierten und einem nicht renovierten Haus mit derselben Bruttomietfläche, Grundstücksgröße, Aussicht und Zustandsbewertung darstellt –, und Eintrag 008 (*Verzerrung durch ausgelassene Variablen*) liefert den Grund dafür.

Der Residuenwert für Comp H aus Ausgabe 1 verdeutlicht, wie bedeutend dieser Unterschied ist. Wenn wir die acht Eigenschaften aus Ausgabe 1 als repräsentativ für den gesamten Markt betrachten, dann entspricht jeder viervariablen Merkmalsvektor genau einer der acht Eigenschaften, und somit verläuft die viervariablen CEF exakt durch alle acht Preise – wobei Comp H einer davon ist. Comp H weist keine Abweichung vom viervariablen CEF auf; sein Residuen-Wert aus Ausgabe 1 von \$46,455 ist der Betrag, um den er von den Anpassungen an die lineare Regression mit vier

Variablen abweicht – der linearen Projektion dieses Marktes mit acht Eigenschaften –, und genau in der Lücke zwischen der Projektion und der CEF verbirgt sich der Effekt der Neumodellierung. Bei der Spezifikation mit fünf Variablen werden die Projektion und der CEF identisch, und die Abweichung verschwindet. Der Residualwert stammte nicht von Comp H selbst, sondern von der Spezifikation.

References.

- Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
- Craytor, W. B. (2026). Residual Constraint Approach (RCA): Framework & Protocol. *The Valuation Engineer*, 1(1). <https://doi.org/10.67330/kgbxf46>
- Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1), 1–67. <https://doi.org/10.1214/aos/1176347963>
- Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Julius Springer.
- Milborrow, S. (2026a). *earth: Multivariate Adaptive Regression Splines* (R package version 5.3.6; derived from mda:mars by T. Hastie and R. Tibshirani). <https://doi.org/10.32614/CRAN.package.earth>
- Milborrow, S. (2026b, August 7). *Variance Models in earth*. <https://www.milbo.org/doc/earth-varmod.pdf>

Cross-references: hedonic regression; implicit price; omitted variable bias; parameter identification.

Addendum: Prolog Formalization

The following machine-readable companion (007-conditional-expectation.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>.

%% 007-conditional-expectation.pl
%%
%% Prolog companion to Foundations Vol. 1, Issue 2, Entry 007
%% Title: Conditional Expectation
%% Author: Bert Craytor
%% Concept: conditional_expectation
%% Version: 1.2.0
%%
%% v1.2.0: entry text revised for readers who are not statisticians;
%% no Prolog clause changes.
%%
%% 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 generating function lives in
%% kb/pacifica_comps_extended.pl, which is restated in those units with this
%% release; nothing here changes but the units it reads.
%%
```

```

%% This file is the Prolog galley for the entry: it encodes the entry's
%% main claims, its worked example, and the generating function of the
%% running dataset in a form other Prolog code in the journal can
%% consult, query, and verify. Read alongside entry.tex / entry.pdf.

% =====
% DICTIONARY IMPORT
% =====

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

% --- From the dictionary (quoted for reference) -----
% term(conditional_expectation, "Conditional Expectation",
%     econometric_method,
%     "The conditional expectation of an outcome given a set of
%     predictors is the mean of the outcome's distribution at each
%     fixed value of the predictors. Viewed as a function of the
%     predictors, it is called the conditional expectation function
%     (CEF), and it is the function that regression aims to
%     approximate. Ordinary least squares (OLS) regression gives the
%     best linear approximation to the CEF, and it reproduces the CEF
%     exactly when the CEF itself is linear.", published,
%     'dictionary-2026.1').
%
% depends_on(conditional_expectation, hedonic_regression).
% -----

uses_terms([ conditional_expectation, hedonic_regression,
             implicit_price, hedonic_price_function,
             characteristics_space, residual,
             ordinary_least_squares,
             sales_comparison_approach, adjustment,
             comp_selection, reconciliation ]).
uses_relations([depends_on/2, related/2]).

% =====
% ENTRY SUBJECT
% =====

article_concept(conditional_expectation).
article_issue(volume(1), number(2), year(2026)).
article_id('foundations.2026.007').

% =====
% ENTRY-LOCAL VOCABULARY
% =====
%
% Field names for the worked-example records and for the two properties
% of the CEF the entry relies on. Too narrow to promote to the
% dictionary.

local_term(cef).                % conditional expectation function
local_term(estimand).
local_term(conditioning_set).
local_term(prediction_interval).
local_term(confidence_interval).

```

```

local_term(linear_projection).      % best linear approximation to the CEF

% =====
% THE TWO PROPERTIES OF THE CEF THE ENTRY USES
% =====

%! cef_property(?Name, ?Statement) is nondet.

cef_property(best_predictor,
    "Among all functions of the conditioning variables, the CEF minimizes expected squared
    ↪ prediction error.").
cef_property(orthogonal_deviation,
    "P = E[P|Z] + eps with E[eps|Z] = 0 holds by construction, not by assumption.").

% The second property is the hinge for entries 008-010: each asks when
% the recovered object is the CEF of interest.
hinge_for(orthogonal_deviation, omitted_variable_bias).
hinge_for(orthogonal_deviation, endogeneity).
hinge_for(orthogonal_deviation, identification).

% =====
% CENTRAL CLAIM: CONDITIONING SET DETERMINES THE ESTIMAND
% =====
%
% Two regressions over the same data with different conditioning sets
% target DIFFERENT population objects, and each recovers only the best
% linear approximation to its target (the linear projection); OLS
% recovers the CEF itself exactly when the CEF is linear. Neither is
% wrong; they answer different questions.

conditioning_set(issue1_model,
    [gla_sqft, lot_sqft, view_numeric, cond_score]).
conditioning_set(issue2_model,
    [gla_sqft, lot_sqft, view_numeric, cond_score, remodel_numeric]).

%! different_estimand(?ModelA, ?ModelB) is nondet.
%
% True when two models condition on different variable sets and
% therefore approximate different conditional expectation functions.

different_estimand(A, B) :-
    conditioning_set(A, SA),
    conditioning_set(B, SB),
    A \== B,
    sort(SA, X), sort(SB, Y),
    X \== Y.

% What OLS recovers, relative to the CEF it conditions on.
ols_recovers(best_linear_approximation_to_cef).
ols_recovers_cef_exactly_when(cef_is_linear).

% =====
% WORKED EXAMPLE - THE GENERATING FUNCTION
% =====
%
% The running dataset is defined in ../../kb/pacifica_comps_extended.pl.
% Its generating function IS the conditional expectation function for
% the constructed market, exactly: the data carry no noise.

```

```

:- 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)).

%! cef_exact(+Property, -Price) is det.
%
% The CEF evaluated at a property's characteristic vector. Identical to
% true_price/2 in the dataset file; named here to make the entry's claim
% ("the generating function IS the CEF") explicit in the galley.

cef_exact(P, Price) :- true_price(P, Price).

%! noiseless_dataset is semidet.
%
% Verifies the entry's central factual claim about the data: every
% observed sale price equals the CEF at that property, so the residual
% is zero everywhere.

noiseless_dataset :-
    forall(property(P),
        ( cef_exact(P, Price), sale_price(P, Price) ) ).

% =====
% THE COMP H POINT
% =====
%
% Issue 1 reported a $46,455 residual for Comp H under the four-variable
% model. Taking issue 1's eight properties as the market, every
% four-variable characteristic vector belongs to exactly one property, so
% the four-variable CEF interpolates all eight prices and Comp H deviates
% from it by ZERO. The 46,455 is Comp H's deviation from the fitted
% four-variable LINEAR regression --- the linear projection --- and the
% projection-CEF gap is where the remodel effect hides. Under the
% five-variable specification, projection and CEF coincide and the
% deviation vanishes.

issue1_residual(comp_h, 46455).

% deviation_from(Property, Object, Dollars).
deviation_from(comp_h, four_variable_projection, 46455).
deviation_from(comp_h, four_variable_cef, 0).
deviation_from(comp_h, five_variable_cef, 0).

residual_under(issue2_model, comp_h, 0) :-
    cef_exact(comp_h, Price),
    sale_price(comp_h, Price).

%! residual_is_artifact_of_specification(?Property) is nondet.
%
% The issue-1 residual is a property of the specification (a linear
% projection under a smaller conditioning set), not of the property.

residual_is_artifact_of_specification(P) :-
    issue1_residual(P, R), R =\= 0,
    deviation_from(P, four_variable_cef, 0),
    residual_under(issue2_model, P, 0).

```

```

% =====
% APPRAISER ENCOUNTERS
% =====

grid_assumption(smoothness,
    "Using comparables at all assumes the CEF is smooth enough that nearby points inform the
    ↪ subject's point.").
grid_assumption(local_linearity,
    "Adjusting linearly assumes the CEF's slope in that characteristic is roughly constant over the
    ↪ range crossed.").
grid_assumption(unbiased_indications,
    "Averaging indications assumes each is an unbiased draw around the same conditional mean.").

% The grid is structurally a nonparametric estimator of a conditional
% mean, using local averaging and hand-specified slopes.
estimator_role(sales_comparison_approach, conditional_mean_estimator).

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

defensibility_consequence(conditioning_set,
    "Adding a variable to a grid changes the estimand, not merely the precision of an estimate.").
defensibility_consequence(ceteris_paribus,
    "'Holding other things equal' means comparing conditional means at vectors differing in one
    ↪ coordinate.").
defensibility_consequence(residual_spread,
    "Conditional dispersion is a market property; a point estimate reported without it overstates
    ↪ precision.").

% The confidence/prediction interval confusion the entry names.
interval_answers(confidence_interval, "uncertainty about where the CEF is").
interval_answers(prediction_interval, "spread of individual prices around the CEF").
commonly_confused(confidence_interval, prediction_interval).

```