

Esperança condicional (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)

Definição formal. A expectativa condicional de um resultado (neste caso, o preço de venda), dado um conjunto de preditores (as características medidas), é a média da distribuição do resultado para cada valor fixo dos preditores. Considerada como uma função dos preditores, ela é chamada de função de expectativa condicional (CEF) e é a função que a regressão visa aproximar. A regressão por mínimos quadrados ordinários (OLS) fornece a melhor aproximação linear da CEF e a reproduz exatamente quando a própria CEF é linear.

Enquadramento intuitivo. Um ajuste derivado de uma regressão parte de uma taxa: o que o mercado paga por pé quadrado, por banheiro, por ano de idade. Este artigo trata de onde vem essa taxa e a que ela se refere. A resposta curta é que se trata da inclinação da curva do preço médio de mercado à medida que uma característica muda e as demais permanecem inalteradas, e uma regressão é uma forma de medir essa inclinação a partir das vendas. A regressão OLS mede uma inclinação para toda a faixa da característica; o MARS (Multivariate Adaptive Regression Splines, Friedman 1991) mede uma inclinação para cada faixa na qual o mercado paga de maneira diferente e apresenta uma taxa por faixa. Em ambos os casos, a taxa é uma medida da mesma coisa: a forma como o preço médio se varia com a característica; e o restante desta entrada trata exatamente disso: o que é, por que a tabela de comparação de vendas também a mede e o que significa quando duas regressões sobre as mesmas vendas resultam em duas taxas diferentes.

Para o leitor que deseja conhecer o mecanismo. O “preço médio à medida que as características variam” tem um nome, a função de expectativa condicional, e o artigo passa a utilizá-la daqui em diante. Trata-se de uma superfície no espaço das características; uma taxa corresponde à sua inclinação em uma direção; o OLS ajusta o plano que melhor se aproxima da superfície e a recupera exatamente apenas quando a superfície é um plano; o MARS ajusta uma superfície composta de partes planas que se curvam nos pontos de ruptura que encontra e, assim, se aproxima mais quando a superfície verdadeira se curva. Os oito imóveis comparáveis da Pacífica da Edição 1 são o exemplo prático.

Este artigo trata mais de estatística do que de direito: aborda o que uma regressão é capaz de estimar e destina-se a avaliadores que realizam regressões ou que utilizam a tabela de comparação de vendas — que também estima o mesmo.

A regressão realizada na Edição 1 para os veículos comparáveis da Pacífica foi utilizada para derivar preços implícitos a partir de seus coeficientes (as taxas que a regressão atribui às características); um preço implícito é o valor em dólares que o mercado paga por uma unidade de uma característica. No entanto, o que a regressão estimou não foi declarado claramente, e vale a pena mencionar isso porque quase todas as questões que se seguem nesta edição — relativas a viés, endogeneidade e identificação — são questões relacionadas à diferença entre o que foi estimado e o que se desejava.

A regressão é normalmente apresentada em termos de ajuste de uma reta por meio da minimização da soma dos resíduos ao quadrado (um resíduo é o preço de venda menos o preço indicado pela equação ajustada) e, em seguida, da determinação da inclinação. Embora essa descrição esteja correta no que diz respeito à aritmética envolvida, ela é enganosa quando se trata do objetivo. A

aritmética é apenas o meio; o objetivo é uma grandeza que descreva o mercado como um todo (uma grandeza *populacional*), e não apenas as vendas em questão.

Como interpretar a notação.

P	preço de venda.
$\mathbf{Z}$	a lista das características medidas dos imóveis; uma lista como essa é chamada de <i>vetor</i> .
$\mathbf{z}$	um conjunto específico de valores para essa lista, por exemplo, 1.650 pés quadrados, um terreno de 7.200 pés quadrados, sem vista, condição 7,00.
$E[\]$	o <i>valor esperado</i> : a média de todos os valores possíveis, cada um ponderado por sua probabilidade.
$ $	“dado”.
$E[P \mid \mathbf{Z} = \mathbf{z}]$	o valor esperado do preço entre os imóveis cujas características correspondem a $\mathbf{z}$.
ε	o desvio de uma venda em relação a esse valor esperado.
u	um erro estrutural de uma venda: seu preço menos o valor que a relação causal indica que deveria ser.
$\hat{p}$	o valor fornecido pela equação de ajuste.
$z_{\text{GLA}}, \dots$	valor de uma característica: área bruta (GLA) e área do lote em pés quadrados; vista e reforma: 1 se sim, 0 se não; escore entre 0,00 e 10,00.

Defina o vetor de características $\mathbf{z}$ — ou seja, 1.650 pés quadrados, um terreno de 7.200 pés quadrados, sem vista e em “bom” estado (7,00 na escala de estado do Journal, explicada no exemplo ilustrativo). Entre os imóveis no mercado que possuem exatamente esse vetor de características, os preços variam. Os preços não são iguais porque os imóveis apresentam diferenças que não são capturadas pelo vetor, porque compradores e vendedores são diferentes e porque a negociação nem sempre resulta no mesmo preço. Nesse mercado, há uma distribuição completa de preços correspondente a $\mathbf{z}$. A média dessa distribuição é a expectativa condicional do preço, dado $\mathbf{z}$, escrita como $E[P \mid \mathbf{Z} = \mathbf{z}]$; a definição geral é de Kolmogorov (1933). A média das vendas em andamento é uma estimativa desse valor esperado, não o próprio valor esperado.

Como as características dos imóveis abrangem todo o espaço de características, a expectativa condicional traça uma superfície: em cada ponto há um número, ou seja, o preço médio dos imóveis que apresentam essas características. Essa superfície é conhecida como *função de expectativa condicional*, ou CEF. Trata-se de uma característica do mercado e não de algo pertencente a um conjunto de dados específico — ela ainda existiria mesmo que nenhum imóvel tivesse sido submetido a uma avaliação. A Figura 1 ilustra isso para o exemplo em questão.

Dois fatos tornam a CEF o alvo natural da estimativa. Primeiro, dentre todas as funções de $\mathbf{z}$, a CEF é aquela que minimiza o valor esperado do erro de previsão ao quadrado: quando é preciso estimar um preço com base apenas em $\mathbf{z}$ e a avaliação é feita pelo erro ao quadrado, $E[P \mid \mathbf{Z} = \mathbf{z}]$ é a melhor estimativa possível. Em segundo lugar, qualquer resultado pode ser expresso na forma

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

, sendo que a segunda equação não é uma suposição, mas uma identidade, uma vez que o desvio da média condicional tem, por definição, uma média condicional igual a zero (Angrist & Pischke 2009, cap. 3).

Esta entrada vem antes das entradas 008 (*Viés de variável omitida*) a 010 (*Identificação de parâmetros*), pois cada uma delas questiona se o que uma regressão recupera é o CEF de interesse. A identidade $E[\varepsilon \mid \mathbf{Z}] = 0$, por si só, sempre se mantém, pois ε é definida como a distância da média condicional. O que pode falhar é a mesma condição — uma média condicional igual a zero — para um erro diferente: o erro estrutural, denotado por u na entrada 009 (*Endogeneidade*) — a diferença entre o preço e a relação causal que o avaliador deseja medir. Esse erro pode variar

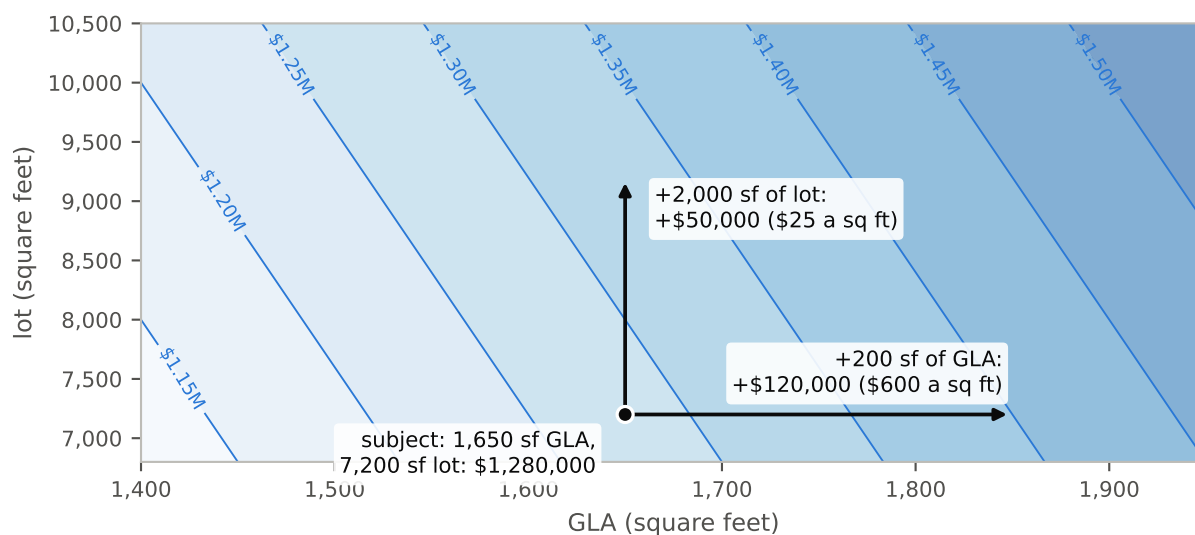


Figure 1: A função de expectativa condicional do mercado Pacifica construído (exemplo ilustrativo), para casas sem vista, com escore de condição de 7,00 e sem reforma, representada como um mapa topográfico: cada linha une as combinações de GLA e terreno cujo preço médio é o mesmo, e o sombreado fica mais escuro à medida que o preço médio aumenta. O ponto marcado é a casa desta entrada, com 1,650 pés quadrados de GLA em um lote de 7,200 pés quadrados, no endereço \$1,280,000 . As duas setas representam as duas inclinações, que correspondem às taxas: 200 pés quadrados a mais de GLA aumentam o preço médio em \$120,000 , ou \$600 por pé quadrado; 2,000 pés quadrados a mais de lote aumentam-no em \$50,000 , ou \$25 por pé quadrado.

de acordo com as características. Isso ocorre, por exemplo, quando uma característica que afeta o preço é deixada de fora; quando o preço exerce efeito reverso sobre uma característica — os dias no mercado (DOM) geralmente aumentam à medida que o preço sobe; assim, uma regressão que usa o DOM para explicar o preço capta parte do próprio efeito do preço sobre o DOM, razão pela qual poucos analistas incluem o DOM em uma regressão do preço de venda líquido; ou quando uma característica é medida com erro. *O viés de variável omitida* (entrada 008) e *a endogeneidade* (entrada 009) tratam dessas formas; *a identificação de parâmetros* (entrada 010) aborda a questão prévia: se as vendas podem, de fato, indicar a taxa.

Onde o avaliador a encontra. Um avaliador que afirma que “uma casa com área de 1.650 pés quadrados e sem vista nesta região é vendida por cerca de \$1.28 million” está expressando uma expectativa condicional. A palavra usada aqui é “cerca de”. A afirmação não sugere que todas as casas desse tipo sejam vendidas por \$1.28 million; significa, ao contrário, que \$1.28 million é a média dos preços pelos quais as casas desse tipo são vendidas. Ao dizer “cerca de”, o avaliador reconhece que os preços individuais variam em torno desse valor.

Uma grade de comparação de vendas é um método para estimar o valor do CEF no vetor característico do imóvel objeto da avaliação; cada imóvel comparável é uma observação extraída da distribuição de preços no vetor característico desse imóvel comparável; cada ajuste desloca essa observação ao longo da superfície do CEF na direção do vetor do imóvel objeto da avaliação; a indicação reconciliada fornece, então, uma estimativa da altura da CEF no imóvel objeto da avaliação. Estruturalmente, quando seus ajustes são extraídos das vendas, a grade pode ser vista como um estimador não paramétrico simples (que não pressupõe nenhuma fórmula para a superfície) de uma média condicional — que emprega média local e inclinações especificadas manualmente em vez do MÉG; os ajustes extraídos do custo ou do julgamento são opiniões estruturadas, não estimativas desse tipo.

Enquadrar a grade dessa maneira traz à tona as suposições tácitas do avaliador. Sempre que são utilizados imóveis comparáveis, presume-se que a CEF seja suficientemente suave para que os pontos próximos sejam relevantes para o valor no ponto do imóvel objeto da avaliação. Os ajustes lineares pressupõem que a inclinação da CEF para uma das características específicas seja aproximadamente constante em toda a faixa abrangida. Atribuir peso igual a três indicações pressupõe que cada uma estime a mesma média condicional, que nenhuma esteja sistematicamente deslocada e que todas tenham precisão semelhante. Nenhuma dessas suposições é incomum; todas dizem respeito à forma de uma superfície e todas se tornam temas de discussão assim que a superfície em questão é identificada.

Essa abordagem também esclarece uma confusão comum sobre intervalos. Existem dois tipos de incerteza. Um diz respeito à localização do CEF: quão distante a superfície ajustada pode estar da média verdadeira. O outro diz respeito à dispersão dos preços individuais em torno dessa média. Um intervalo de confiança abrange apenas o primeiro; um intervalo de previsão abrange ambos e, portanto, é sempre o mais amplo dos dois. Com o MARS, o principal método de regressão do Journal, o pacote earth (Milborrow 2026a) fornece um intervalo de previsão para um imóvel objeto da avaliação a partir de um segundo modelo, seu modelo de variância (Milborrow 2026b), que estima como a dispersão dos preços varia ao longo do mercado. As condições são: que haja vendas suficientes e que os erros sejam aproximadamente normais e independentes uns dos outros; o pacote informa com que frequência as vendas se enquadram no intervalo, e essa é a verificação. Na avaliação baseada na RCA, o valor de mercado do imóvel objeto da avaliação, de acordo com a Abordagem de Comparação de Vendas, é o preço de venda estimado pelo modelo mais um resíduo estimado pelo avaliador. O modelo de variância fornece o intervalo dentro do qual, com probabilidade de 95%, uma venda do imóvel objeto da avaliação se situaria, desde que essas condições sejam atendidas.

Quando perguntam a um avaliador “qual é o seu grau de certeza?”, geralmente estão perguntando qual é a dispersão dos preços individuais (o intervalo de previsão), e ele costuma responder qual é o seu grau de certeza em relação à média (o intervalo de confiança).

Nomear o estimando como — ou seja, as características da população que está sendo estimada — é um ato defensável, pois uma conclusão não pode ser avaliada sem isso. A afirmação “Os ajustes são \$100,000” ainda não é algo que possa ser considerado certo ou errado. Somente quando dizemos “A diferença no preço de venda condicional entre imóveis com vista e aqueles sem vista, mantendo-se constantes a área habitável bruta (AHB), o lote e as condições, é de \$100,000 neste segmento de mercado” é que fazemos uma afirmação que pode ser contestada com base no que é mantido constante, na população em questão e nas unidades envolvidas.

Por que importa para a solidez. Daí decorrem três consequências em termos de defensabilidade.

A escolha de um conjunto de condicionamento (as características mantidas constantes) é uma questão metodológica: ela define o estimando. Se os dados podem estimar esse estimando é uma questão à parte, objeto da *Identificação de Parâmetros*. As expressões $E[P \mid \text{GLA, view}]$ e $E[P \mid \text{GLA, view, condition}]$ representam funções diferentes, que podem ter valores e inclinações distintos. Quando um avaliador inclui uma variável na grade, ele não está simplesmente melhorando a estimativa; está alterando o estimando. A postura razoável é indicar o conjunto de condicionamento no qual se baseia a opinião de valor.

A expressão “mantendo as demais variáveis constantes” descreve o raciocínio do avaliador na avaliação e possui um significado definido. No âmbito da estrutura do CEF, ela significa comparar as médias condicionais em dois vetores característicos que diferem em apenas uma característica. Trata-se de uma afirmação relacionada à população e pode ser verificada com base em dados de mercado, quando houver dados suficientes. A análise subjacente a ela geralmente é incluída no arquivo de trabalho — o registro da avaliação que a Regra de Manutenção de Registros da USPAP exige que o avaliador mantenha —, e não no relatório.

A dispersão do resíduo é uma informação, não um motivo de constrangimento. A dispersão dos preços em torno da função de expectativa condicional é uma característica do mercado; em um mercado onde os preços se dispersam amplamente em torno dela, qualquer estimativa de um único valor — independentemente do método utilizado — estará acompanhada de grande incerteza. Apresentar um valor sem fazer referência a essa dispersão faz com que a conclusão pareça mais precisa do que o mercado realmente permite.

Exemplo de avaliação resolvido. Os oito preços comparáveis do Pacifica na Edição 1 constituem uma amostra sem ruído: eles foram construídos para satisfazer exatamente uma única equação linear. Quando esses oito preços são submetidos a uma regressão com base na área bruta locável (GLA), no lote, na vista, nas condições e no indicador de reforma, cada preço é reproduzido ao dólar — nada fica de fora — e a regressão retorna a equação a partir da qual os preços foram construídos, taxa por taxa:

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

O “estado” aqui refere-se ao escore de estado desta revista: uma estimativa do estado da casa em relação ao estado das residências vizinhas em sua área de mercado definida. Baseia-se na compreensão do avaliador sobre essa área de mercado e não deve entrar em conflito substancial com a compreensão de outros avaliadores que atuam na mesma área de mercado. O escore varia de 0,00 a 10,00: a porcentagem de casas no mercado que estão em pior estado, dividida por dez. Um escore de 5,00 significa que metade das casas no mercado está em pior estado; 9,00 significa que nove em cada dez estão. A Edição 1 descreveu o estado em palavras e o codificou de 1 a 3 para sua regressão; no escore desta Revista, as descrições correspondem a “média” 5,00, “boa” 7,00 e

“excelente” 9,00; assim, um nível da Edição 1 equivale a dois pontos, e o mercado paga \$20,000 por ponto. O último termo é a diferença de preço que os dados construídos atribuem a uma reforma: \$60,000 a mais do que uma casa com a mesma área bruta locável (GLA), terreno, vista e escore de condição. Não se trata de um preço pelo ato de reformar; representa tudo o que a reforma trouxe de mais que as outras características não registram (entrada 008, *Viés de Variável Omitida*).

Este é, precisamente, o CEF para o mercado Pacífica construído. É importante deixar claro que tipo de objeto acaba de ser descrito e por que ele pode ser descrito com tanta exatidão. O conjunto de dados foi construído, e um conjunto de dados construído pode, de fato, estar isento de ruído; um conjunto de dados real nunca está. A vantagem de construí-lo é que ele remove completamente o ruído aleatório dos preços — cada preço se encaixa exatamente na fórmula —; portanto, nas entradas 008 (*Viés de Variável Omitida*) a 010 (*Identificação de Parâmetros*), nenhuma diferença entre uma estimativa e seu valor-alvo pode ser atribuída ao ruído: cada uma delas é um viés ou, na *Identificação de Parâmetros*, uma falha de identificação.

O exemplo trata a condição como uma característica com uma taxa em dólares, assim como faz a avaliação tradicional e como foi feito na Edição 1, e as entradas 008 (*Viés de Variável Omitida*), 009 (*Endogeneidade*) e 010 (*Identificação de Parâmetros*) mostram o que dá errado com essa abordagem. A metodologia de avaliação tradicional não é o método adotado pela revista, que é a avaliação baseada na residual Constraint (residual Constraint). Na residual Constraint, a regressão considera apenas as características que podem ser medidas; o estado de conservação e a qualidade são propositalmente deixados de fora. O que a regressão deixa sem explicação em relação a um preço de venda é o resíduo. Nos mercados da Califórnia, onde atua o editor, as características medidas geralmente explicam cerca de 80% da variação nos preços de venda, deixando cerca de um quinto para o resíduo. O avaliador divide o resíduo entre condição, qualidade e as outras características que os dados não medem, usando seus próprios escores como orientação, sob uma restrição: a soma das contribuições de valor dos componentes residuais deve ser igual ao resíduo do imóvel comparável. A conclusão sobre o valor não depende de como o avaliador a divide — Prova II do artigo da RCA (Craytor 2026, p. 36): alterar a composição do resíduo sob a restrição de que todas as contribuições dos componentes do valor residual somem-se ao resíduo do imóvel comparável não tem impacto na conclusão final sobre o valor. A taxa de condição, portanto, nunca é estimada nem reivindicada; o que o avaliador atribui à condição é a condição além daquilo que as características medidas, entre elas a idade, já representam.

Se compararmos a regressão da Edição 1 com essa CEF, a regressão pode ser avaliada — embora seja necessário fazer uma distinção primeiro. De modo geral, uma regressão OLS não reconstrói a função de expectativa condicional (CEF) a que se destina; em vez disso, ela encontra a melhor aproximação linear a ela, ou seja, a projeção linear, e essas duas só coincidem no caso em que a CEF é linear. Um método não paramétrico, como o MARS, não precisa que a CEF seja linear: ele aproxima a CEF por partes, com uma articulação (um ponto onde a taxa muda) onde a superfície se curva e um termo de interação (uma taxa que depende de outra característica) onde ela se torce, seguindo assim uma CEF não linear que uma projeção linear achataria. Ainda assim, trata-se de uma aproximação do mesmo objeto, e as questões relacionadas à variável omitida, endogeneidade e identificação, abordadas nas próximas entradas, se aplicam a ela exatamente da mesma maneira. A CEF de cinco variáveis mencionada acima é linear, razão pela qual a regressão que inclui o indicador de reforma é capaz de recuperá-la exatamente. A regressão da Edição 1, condicionada a quatro características — excluindo o indicador de remodelação —, estima a melhor aproximação linear a $E[P \mid \text{GLA, lot, view, cond}]$ e retorna, com a condição reafirmada com base no score da Revista (a Edição 1 publicou um intercept de 80,777 e 49,824 por etapa de sua codificação de 1 a 3; trata-se do mesmo ajuste),

$$\hat{p}(\mathbf{z}) = 6,041 + 575 \cdot z_{\text{GLA}} + 22.70 \cdot z_{\text{lot}} + 86,179 \cdot z_{\text{view}} + 24,912 \cdot z_{\text{cond}}.$$

Cada coeficiente é diferente do que está acima dele. Uma vez que a regressão tenha sido descrita em termos de ajuste de reta, não há mais nada a acrescentar: temos dois modelos e dois conjuntos de números. No entanto, na abordagem CEF, duas afirmações exatas estão disponíveis. Primeiro, as duas regressões dependem de conjuntos diferentes de variáveis e, portanto, aproximam funções de expectativa condicional diferentes:

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

Em um mercado em que alguns imóveis reformados e outros não reformados compartilham as mesmas quatro características, as duas funções apresentam valores diferentes nesses pontos. Nesse pequeno mercado artificial, não há dois imóveis com as mesmas quatro características; portanto, as duas funções coincidem para cada imóvel, mas diferem na forma: a segunda é uma fórmula linear em suas características, enquanto a primeira não é. Em cada caso, a regressão fornece apenas a melhor aproximação linear ao seu alvo. Ambos os alvos são objetos válidos. Em segundo lugar, apenas aquela com cinco variáveis possui coeficientes que podem ser interpretados como preços implícitos — sendo o coeficiente de reforma a diferença de preço entre uma casa reformada e uma não reformada com a mesma área bruta locável (GLA), terreno, vista e escore de condição — e o item 008 (*Viés de Variável Omitida*) explica o motivo.

O resíduo do Comp H da Edição 1 ilustra exatamente o quão significativa é essa distinção. Se considerarmos as oito propriedades da Edição 1 como representativas de todo o mercado, então cada vetor de características de quatro variáveis corresponde exatamente a uma das oito propriedades e, portanto, o CEF de quatro variáveis passa exatamente por todos os oito preços — sendo o Comp H um deles. Comp H apresenta desvio zero em relação ao CEF de quatro variáveis; seu resíduo da Edição 1, \$46,455 é o valor pelo qual ele difere da regressão linear de quatro variáveis ajustada — a projeção linear desse mercado de oito imóveis — e é precisamente na lacuna entre a projeção e o CEF que o efeito da remodelação está oculto. Com a especificação de cinco variáveis, a projeção e o CEF tornam-se idênticos, e o desvio desaparece. O resíduo não veio do próprio Comp H; ele veio da especificação.

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')).
% Scryer 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).
```