

# Esperanza 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)

---

**Definición formal.** La esperanza condicional de una variable de respuesta (en este caso, el precio de venta) dada una serie de predictores (las características medidas) es la media de la distribución de la variable de respuesta para cada valor fijo de los predictores. Considerada como una función de los predictores, se denomina función de esperanza condicional (CEF), y es la función que la regresión pretende aproximar. La regresión por mínimos cuadrados ordinarios (OLS) proporciona la mejor aproximación lineal a la CEF, y reproduce la CEF exactamente cuando la propia CEF es lineal.

**Planteamiento intuitivo.** Un ajuste derivado de una regresión parte de una tasa: lo que paga el mercado por pie cuadrado, por baño y por año de antigüedad. Esta entrada trata sobre el origen de esa tasa y a qué se refiere. La respuesta breve es que se trata de la pendiente del precio medio del mercado cuando una característica varía y las demás se mantienen constantes, y una regresión es una forma de medir esa pendiente a partir de las ventas. La regresión OLS mide una única pendiente para todo el rango de la característica; MARS (Multivariate Adaptive Regression Splines, Friedman 1991) mide una pendiente para cada rango en el que el mercado paga de forma diferente, y proporciona un informe con una tasa por rango. En ambos casos, la tasa es una medida de lo mismo: la forma en que el precio medio varía en función de esa característica; y el resto de esta entrada trata precisamente de eso: qué es, por qué la cuadrícula comparativa de ventas también la mide y qué significa cuando dos regresiones sobre las mismas ventas arrojan dos tasas diferentes.

*Para el lector que quiera conocer el mecanismo:* . El «precio medio a medida que varían las características» tiene un nombre, la función de esperanza condicional, y el artículo la utiliza a partir de aquí. Se trata de una superficie en el espacio de las características; una tasa es su pendiente en una dirección; el método OLS realiza el ajuste del plano que mejor se aproxima a la superficie y la recupera con exactitud solo cuando la superficie es un plano; el método MARS realiza el ajuste de una superficie formada por fragmentos planos que se curvan en los puntos de ruptura que encuentra, por lo que se acerca más cuando la superficie real presenta curvas. Los ocho inmuebles comparables de Pacífica del número 1 sirven de ejemplo ilustrativo.

Esta entrada trata más de estadística que de derecho: aborda lo que una regresión es capaz de estimar, y está dirigida a los tasadores que realizan regresiones o que utilizan la Grid de comparación de ventas —que también estima lo mismo—.

Se utilizó la regresión realizada en el número 1 para los inmuebles comparables de Pacífica con el fin de obtener precios implícitos a partir de sus coeficientes (los valores que la regresión asigna a las características); un precio implícito es la cantidad en dólares que el mercado paga por una unidad de una característica. Sin embargo, no se especificó claramente qué estimaba la regresión, y merece la pena mencionarlo porque casi todas las cuestiones que se plantean a continuación en este número —relativas al sesgo, la endogeneidad y la identificación— se refieren a la diferencia entre lo que se estimó y lo que se deseaba.

La regresión suele presentarse en términos del ajuste de una recta mediante la minimización de la suma de los residuos al cuadrado (un residuo es el precio de venta menos el precio que da la ecuación ajustada) y, a continuación, obtener la pendiente. Aunque esta descripción es correcta en

lo que respecta a la aritmética implicada, resulta engañosa en cuanto al objetivo. La aritmética no es más que un medio; el objetivo es una magnitud que describa el mercado en su conjunto (una magnitud *poblacional*), y no solo las ventas de las que se dispone.

*Cómo interpretar la notación.*

|                                     |                                                                                                                                                           |
|-------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|
| $P$                                 | precio de venta.                                                                                                                                          |
| $\mathbf{Z}$                        | la lista de las características medidas de un inmueble; una lista como esta se denomina <i>vector</i> .                                                   |
| $\mathbf{z}$                        | un conjunto concreto de valores para esa lista; por ejemplo, 1.650 pies cuadrados, una parcela de 7.200 pies cuadrados, sin vistas, estado 7,00.          |
| $E[ \ ]$                            | el <i>valor esperado</i> : la media de todos los valores posibles, cada uno con un peso según su probabilidad.                                            |
| $ $                                 | «dado».                                                                                                                                                   |
| $E[P \mid \mathbf{Z} = \mathbf{z}]$ | el valor esperado del precio entre los inmuebles cuyas características coinciden con las de $\mathbf{z}$ .                                                |
| $\varepsilon$                       | la desviación de una venta respecto a ese valor esperado.                                                                                                 |
| $u$                                 | Un error estructural de una venta: su precio de venta menos lo que, según la relación causal, debería ser.                                                |
| $\hat{p}$                           | el valor que da la ecuación de ajuste.                                                                                                                    |
| $z_{\text{GLA}}, \dots$             | Valor de una característica: superficie bruta (GLA) y superficie del solar en pies cuadrados; vistas y reforma: 1 si sí, 0 si no; puntuación: 0,00–10,00. |

Fijemos el vector de características  $\mathbf{z}$  — es decir, 1.650 pies cuadrados, una parcela de 7.200 pies cuadrados, sin vistas y en «buen» estado (7,00 en la escala de estado de la revista, tal y como se explica en el ejemplo resuelto). Entre los inmuebles del mercado que presentan exactamente este vector de características, los precios varían. Los precios no son iguales porque los inmuebles presentan diferencias que no quedan reflejadas en el vector, porque los compradores y vendedores son distintos, y porque la negociación no siempre culmina en el mismo precio. En ese mercado existe una distribución completa de precios correspondiente a  $\mathbf{z}$ . La media de esta distribución es la esperanza condicional del precio dado  $\mathbf{z}$ , que se escribe  $E[P \mid \mathbf{Z} = \mathbf{z}]$ ; la definición general se debe a Kolmogorov (1933). La media de las ventas en curso es una estimación de este valor esperado, no el valor esperado en sí mismo.

Dado que las características de los inmuebles abarcan todo el espacio de características, la esperanza condicional traza una superficie: en cada punto hay un número, que es el precio medio de los inmuebles que presentan esas características. Esta superficie se conoce como la *función de esperanza condicional*, o CEF. Es una característica del mercado y no algo propio de ningún conjunto de datos concreto; seguiría existiendo aunque nunca se hubiera tasado ningún inmueble. La figura 1 la representa para el ejemplo que estamos viendo.

Hay dos hechos que convierten a la CEF en el objetivo natural de la estimación. En primer lugar, de todas las funciones de  $\mathbf{z}$ , la CEF es la que minimiza el valor esperado del error de predicción al cuadrado: cuando hay que estimar un precio basándose únicamente en  $\mathbf{z}$  y se evalúa en función del error al cuadrado,  $E[P \mid \mathbf{Z} = \mathbf{z}]$  es la mejor estimación que se puede hacer. En segundo lugar, cualquier resultado puede expresarse de la forma

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

, no siendo la segunda ecuación una hipótesis, sino una identidad, ya que la desviación respecto a la media condicional tiene, por definición, una media condicional igual a cero (Angrist & Pischke 2009, cap. 3).

Esta entrada precede a las entradas 008 (*Sesgo de variable omitida*) a 010 (*Identificación de parámetros*), ya que cada una de ellas plantea si lo que recupera una regresión es la CEF de interés.

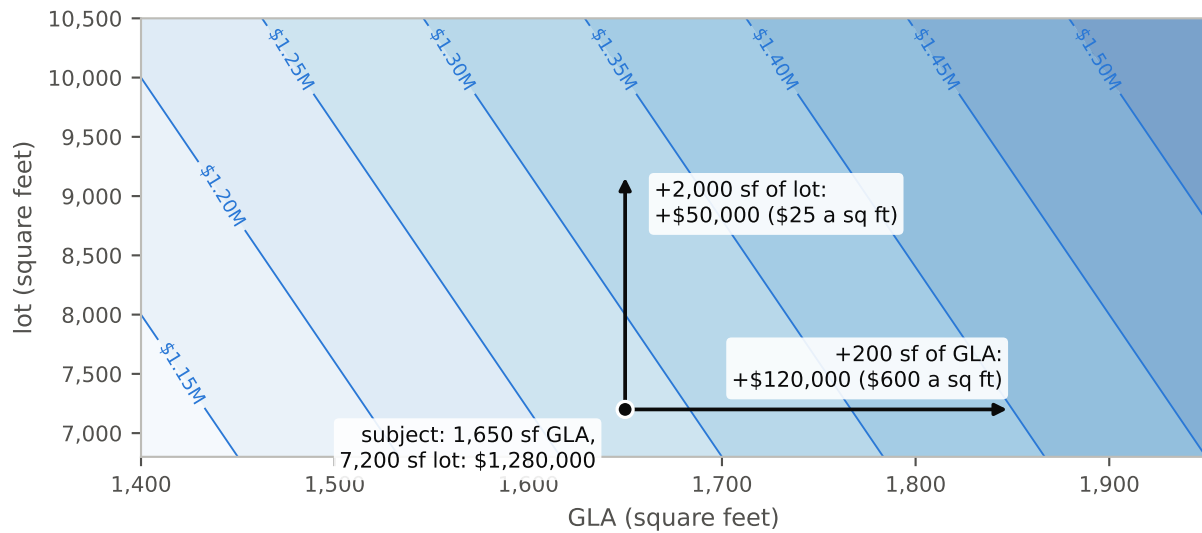


Figure 1: La función de esperanza condicional del mercado de Pacifica construido (ejemplo práctico), para viviendas sin vistas, con una puntuación de condiciones de 7,00 y sin reformas, representada a modo de mapa topográfico: cada línea une las combinaciones de superficie bruta alquilable (GLA) y parcela cuyo precio medio es el mismo, y el sombreado se oscurece a medida que aumenta el precio medio. El punto marcado es la vivienda de esta entrada, 1,650 pies cuadrados de superficie bruta alquilable (GLA) en una parcela de 7,200 pies cuadrados, situada en \$1,280,000 . Las dos flechas representan las dos pendientes, que son las tasas: 200 pies cuadrados más de superficie bruta alquilable (GLA) elevan el precio medio en \$120,000 , es decir, \$600 por pie cuadrado; 2,000 pies cuadrados más de parcela lo elevan en \$50,000 , es decir, \$25 por pie cuadrado.

La identidad  $E[\varepsilon \mid \mathbf{Z}] = 0$  se cumple siempre, ya que  $\varepsilon$  se define como la distancia respecto a la media condicional. Lo que puede fallar es esa misma condición —una media condicional igual a cero— para un error diferente: el error estructural, denominado « $u$ » en la entrada 009 (*Endogeneidad*), es decir, la diferencia entre el precio y la relación causal que el tasador desea medir. Ese error puede variar en función de las características. Esto ocurre, por ejemplo, cuando se omite una característica que afecta al precio; cuando el precio influye a su vez en una característica —los días en el mercado (DOM) suelen aumentar a medida que sube el precio, por lo que una regresión que utilice el DOM para explicar el precio capta parte del propio efecto del precio sobre el DOM, razón por la cual pocos analistas incluyen el DOM en una regresión del precio de venta neto—; o cuando una característica se mide con error. *El sesgo por variable omitida* (entrada 008) y *la endogeneidad* (entrada 009) se refieren a estos casos; *la identificación de parámetros* (entrada 010) plantea la cuestión previa de si las ventas pueden reflejar la tasa en absoluto.

**Dónde la encuentra el valorador.** Un tasador que afirma que «una vivienda con una superficie de 1.650 pies cuadrados y sin vistas en esta zona se vende por unos \$1.28 million» está expresando una expectativa condicional. La palabra utilizada aquí es «unos». La afirmación no pretende afirmar que todas las viviendas de este tipo se vendan por \$1.28 million; más bien, significa que \$1.28 million es la media de los precios a los que se venden las viviendas de ese tipo. Al decir «aproximadamente», el tasador reconoce que los precios individuales oscilan en torno a ese valor.

Una cuadrícula de comparación de ventas es un método para estimar el valor de la CEF en el vector característico del sujeto; cada comparable es una observación extraída de la distribución de precios en el vector característico de dicho comparable; cada ajuste desplaza esa observación a lo largo de la superficie de la CEF en la dirección del vector del sujeto; la indicación conciliada proporciona entonces una estimación de la altura de la CEF en el inmueble objeto de estudio. Estructuralmente, cuando sus ajustes se extraen de las ventas, la cuadrícula puede considerarse como un estimador no paramétrico simple (que no asume ninguna fórmula para la superficie) de una media condicional —uno que emplea promedios locales y pendientes especificadas manualmente en lugar de OLS—; los ajustes tomados del coste o del criterio son opiniones estructuradas, no estimaciones de ese tipo.

Enmarcar la cuadrícula de esta manera saca a la luz los supuestos tácitos del tasador. Siempre que se utilizan «» comparables, se da por supuesto que la CEF es lo suficientemente suavizada como para que los puntos cercanos sean relevantes para el valor en el punto del inmueble objeto de tasación. Un ajuste lineal supone que la pendiente de la CEF para esa característica concreta es aproximadamente constante en todo el rango que se está cubriendo. Otorgar el mismo peso a tres indicaciones supone que cada una estima la misma media condicional, que ninguna se desvía sistemáticamente y que todas tienen una precisión similar. Ninguna de estas suposiciones es inusual; todas se refieren a la forma de una superficie y todas se convierten en temas de debate una vez que se ha identificado la superficie en cuestión.

Este enfoque también aclara una confusión habitual sobre los intervalos. Existen dos tipos de incertidumbre. Uno se refiere a la ubicación del CEF: la distancia a la que la superficie ajustada puede encontrarse del promedio real. El otro se refiere a la dispersión de los precios individuales en torno a ese promedio. Un intervalo de confianza abarca solo el primero; un intervalo de predicción abarca ambos y, por lo tanto, siempre es el más amplio de los dos. Con MARS, el principal método de regresión de la revista, el paquete «earth» (Milborrow 2026a) proporciona un intervalo de predicción para un inmueble objeto a partir de un segundo modelo, su modelo de varianza (Milborrow 2026b), que estima cómo varía la dispersión de los precios en el mercado. Las condiciones son que haya suficientes ventas y que los errores sigan una distribución aproximadamente normal y sean independientes entre sí; el paquete informa con qué frecuencia las ventas se sitúan dentro del

intervalo, y esa es la verificación. En la tasación basada en el RCA, el valor de mercado del inmueble objeto de tasación, según el método de comparación de ventas, es el precio de venta estimado por el modelo más un residuo estimado por el tasador. El modelo de varianza proporciona el intervalo dentro del cual, con una probabilidad del 95 %, se situaría una venta del inmueble objeto de tasación, siempre que se cumplan dichas condiciones. Cuando se le pregunta a un tasador «¿qué grado de certeza tiene?», normalmente se le está preguntando cuál es la dispersión de los precios individuales (el intervalo de predicción), y él suele responder qué grado de certeza tiene respecto a la media (el intervalo de confianza).

Denominar al estimando «» —es decir, las características de la población que se está estimando— es un acto defendible, ya que sin ello no se puede juzgar una conclusión. La afirmación «El ajuste es \$100,000» aún no puede considerarse correcta o incorrecta. Solo cuando decimos «La diferencia en el precio de venta condicional entre los inmuebles con vistas y los que no las tienen, manteniendo constantes la superficie habitable bruta (GLA), la parcela y el estado, es \$100,000 en este segmento de mercado» es cuando hacemos una afirmación que puede ser rebatida en función de lo que se mantiene constante, la población en cuestión y las unidades implicadas.

**Por qué importa para la solidez.** De ello se derivan tres consecuencias en cuanto a la defendibilidad.

La elección de un conjunto de condicionamiento (las características que se mantienen constantes) es una cuestión metodológica: define el estimando. Que los datos puedan estimar ese estimando es otra cuestión, que forma parte de *la identificación de parámetros*. Las expresiones  $E[P \mid \text{GLA, view}]$  y  $E[P \mid \text{GLA, view, condition}]$  representan funciones diferentes, que pueden tener valores y pendientes distintos. Cuando un tasador incluye una variable en la cuadrícula, no se limita a mejorar la estimación, sino que modifica el estimando. Lo razonable es indicar el conjunto de condiciones en el que se basa la valoración.

La expresión «manteniendo constantes las demás variables» describe el razonamiento del tasador en el análisis y tiene un significado concreto. En el marco del CEF, significa comparar las medias condicionales en dos vectores característicos que difieren únicamente en una característica. Se trata de una afirmación relativa a la población y puede contrastarse con los datos de mercado cuando estos sean suficientes. El análisis subyacente suele incluirse en el expediente de trabajo —el registro del encargo que la norma de mantenimiento de registros de la USPAP exige que conserve el tasador— en lugar de en el informe.

La dispersión del residuo es información, no un motivo de vergüenza. La dispersión de los precios en torno a la función de esperanza condicional es una característica del mercado; en un mercado en el que los precios se dispersan ampliamente en torno a ella, cualquier estimación de una sola cifra —independientemente del método que se utilice— irá acompañada de una gran incertidumbre. Indicar un valor sin hacer referencia a esta dispersión hace que la conclusión parezca más precisa de lo que el mercado justifica.

**Ejemplo de valoración resuelto.** Los ocho precios comparables de Pacifica que aparecen en el número 1 de constituyen una muestra «sin ruido»: se construyeron para satisfacer exactamente una única ecuación lineal. Cuando se realiza una regresión de estos ocho precios en función de la superficie bruta alquilable (GLA), la parcela, las vistas, el estado y el indicador de reforma, cada precio se reproduce al dólar —no queda nada sin explicar— y la regresión devuelve la ecuación a partir de la cual se construyeron los precios, tasa por tasa:

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

El «estado» a que se hace referencia aquí es la puntuación de estado de esta revista: una estimación del estado de la vivienda en relación con el estado de las viviendas circundantes en su

zona de mercado definida. Se basa en el conocimiento que tiene el tasador de dicha zona de mercado y no debería diferir sustancialmente del conocimiento de otros tasadores que operan en la misma zona de mercado. La puntuación oscila entre 0,00 y 10,00: el porcentaje de viviendas de la zona de mercado que se encuentran en peor estado, dividido entre diez. Una puntuación de 5,00 significa que la mitad de las viviendas de la zona de mercado se encuentran en peor estado; 9,00 significa que nueve de cada diez lo están. El número 1 describía el estado con palabras y lo codificaba del 1 al 3 para su regresión; en la puntuación de esta revista, las descripciones se corresponden con «normal» (5,00), «buena» (7,00) y «excelente» (9,00), por lo que un nivel del Número 1 equivale a dos puntos, y el mercado paga \$20,000 por punto. El último término es la diferencia de precio que los datos construidos atribuyen a una reforma: \$60,000 más que una vivienda con la misma superficie bruta alquilable (GLA), parcela, vistas y puntuación de estado. No se trata de un precio por el acto de reformar, sino que representa todo aquello que la reforma ha aportado y que las demás características no registran (entrada 008, *sesgo de variable omitida*).

Este es, precisamente, el CEF del mercado «Pacífica» construido. Es importante dejar claro qué tipo de objeto se acaba de describir y por qué puede describirse con tanta precisión. El conjunto de datos ha sido construido, y un conjunto de datos construido puede, de hecho, estar libre de ruido; uno real nunca lo está. La ventaja de construirlo es que elimina por completo el ruido aleatorio de los precios —cada precio se ajusta exactamente a la fórmula—, por lo que en los apartados 008 (*sesgo de variable omitida*) a 010 (*identificación de parámetros*) ninguna diferencia entre una estimación y su valor objetivo puede deberse al ruido: cada una de ellas es un sesgo o, en el caso de la *identificación de parámetros*, un fallo de identificación.

En el ejemplo se trata el estado como una característica con una tasa en dólares, tal y como se hace en la valoración tradicional y como se hizo en el Número 1, y las entradas 008 (*sesgo de variable omitida*), 009 (*endogeneidad*) y 010 (*identificación de parámetros*) muestran qué falla en este enfoque. La metodología de valoración tradicional no es el método de la revista, la valoración basada en el Constraint Approach. En el Constraint Approach, la regresión solo tiene en cuenta las características que se pueden medir; el estado y la calidad se excluyen de ella a propósito. Lo que la regresión deja sin explicar del precio de venta es el residuo. En los mercados californianos del editor, las características medidas suelen explicar alrededor del 80 % de la variación en los precios de venta, dejando aproximadamente una quinta parte al residuo. El tasador divide el residuo entre el estado, la calidad y las demás características que los datos no miden, utilizando sus propias puntuaciones como guía, bajo una única restricción: la suma de las contribuciones de valor de los componentes residuales debe ser igual al residuo del inmueble comparable. La conclusión sobre el valor no depende de cómo la reparta el tasador —Demostración II del artículo sobre el RCA (Craytor 2026, p. 36): modificar el desglose del residuo bajo la restricción de que la suma de las contribuciones de valor de todos los componentes del residuo sea igual al residuo del inmueble comparable no tiene ningún impacto en la conclusión final sobre el valor. Por lo tanto, la tasa correspondiente al estado nunca se estima ni se afirma; lo que el tasador asigna al estado es el estado que excede lo que ya explican las características medidas, entre ellas la antigüedad.

Si comparamos la regresión del Número 1 con esta CEF, se puede evaluar la regresión, aunque primero hay que hacer una distinción. En términos generales, una regresión OLS no reconstruye la función de esperanza condicional a la que se dirige; en su lugar, halla la mejor aproximación lineal a la misma, es decir, la proyección lineal, y ambas coinciden únicamente en el caso de que la CEF sea lineal. Un método no paramétrico como MARS no requiere que la CEF sea lineal: aproxima la CEF por tramos, con una bisagra (un punto en el que cambia la tasa) donde la superficie se curva y un término de interacción (una tasa que depende de otra característica) donde se retuerce, por lo que sigue una CEF no lineal que una proyección lineal aplanaría. Sigue siendo una aproximación al mismo objeto, y las cuestiones relativas a las variables omitidas, la endogeneidad y la identificación

que se tratan en las siguientes entradas se aplican a ella exactamente de la misma manera. La CEF de cinco variables mencionada anteriormente es lineal, razón por la cual la regresión que incluye el indicador de remodelación es capaz de recuperarla con exactitud. La regresión del Número 1, condicionada a cuatro características —excluyendo el indicador de remodelación—, estima la mejor aproximación lineal a  $E[P \mid \text{GLA, lot, view, cond}]$  y arroja, con la condición reformulada en función de la puntuación de la revista (el Número 1 publicó una intersección de 80,777 y 49,824 por paso de su codificación de 1 a 3; se trata del mismo 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 es diferente del que le precede. Una vez descrita la regresión en términos de ajuste de rectas, no hay nada más que añadir: tenemos dos modelos y dos conjuntos de números. Sin embargo, según el enfoque CEF, se pueden formular dos afirmaciones exactas. En primer lugar, las dos regresiones se condicionan a diferentes conjuntos de variables y, por lo tanto, aproximan diferentes funciones de esperanza condicional:

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

En un mercado en el que algunos inmuebles reformados y otros sin reformar comparten las mismas cuatro características, las dos funciones tienen valores diferentes en esos puntos. En este pequeño mercado artificial, no hay dos inmuebles que tengan las mismas cuatro características, por lo que las dos funciones coinciden en cada inmueble, pero difieren en su forma: la segunda es una fórmula lineal en sus características, mientras que la primera no lo es. En cada caso, la regresión proporciona únicamente la mejor aproximación lineal a su objetivo. Ambos objetivos son objetos válidos. En segundo lugar, solo la que tiene cinco variables presenta coeficientes que pueden interpretarse como precios implícitos —siendo el coeficiente de reforma la diferencia de precio entre una vivienda reformada y otra sin reformar con la misma superficie bruta alquilable (GLA), parcela, vistas y puntuación de estado— y la entrada 008 (*sesgo de variable omitida*) explica el motivo.

El residuo de Comp H del número 1 ilustra lo significativa que es esta distinción. Si consideramos que las ocho propiedades del Número 1 representan la totalidad del mercado, entonces cada vector de características de cuatro variables se corresponde exactamente con una de las ocho propiedades y, por lo tanto, la CEF de cuatro variables pasa exactamente por los ocho precios, siendo Comp H uno de ellos. Comp H presenta un residuo nulo respecto a la CEF de cuatro variables; su residuo del Número 1, \$46,455 es la cantidad en la que difiere de la regresión lineal de cuatro variables con ajuste —la proyección lineal de ese mercado de ocho inmuebles— y es precisamente en la diferencia entre la proyección y la CEF donde se oculta el efecto de la remodelación. Con la especificación de cinco variables, la proyección y el CEF se vuelven idénticos, y la desviación desaparece. El residuo no procedía de Comp H en sí mismo, sino de la especificación.

### 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/kgebx46>
- 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 Scyer. 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
   $\hookrightarrow$  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).

```