

Sesgo por variable omitida (Omitted Variable Bias)

Foundations • The Valuation Engineer

William Bert Craytor • October 1, 2026

Version 1.2.0

DOI: [10.67330/460smg10](https://doi.org/10.67330/460smg10)

Definición formal. El sesgo de variable omitida es la distorsión sistemática que se introduce en un coeficiente estimado cuando una característica que afecta al precio se omite del modelo y dicha característica se mueve junto con una de las características incluidas —se mueve con ella después de que se hayan tenido en cuenta las demás características incluidas se hayan tenido en cuenta, que es lo que mide la regresión auxiliar (una segunda regresión, de la característica omitida sobre las incluidas). El efecto de la característica omitida se filtra entonces en los coeficientes de las características incluidas (los dólares por unidad que el modelo asigna a cada característica).

Planteamiento intuitivo. Una cotización puede ser errónea porque refleja el valor de otra característica. El Comp H del número 1 se vendió por \$46,455 más de lo que predecía la regresión basada en sus cuatro características registradas. El motivo, revelado en la «*variable latente*» del número 1, fue una reciente reforma de la cocina, que, según esa entrada, añade unos \$60,000 al valor de la vivienda. El estado de la vivienda figuraba en la lista de características, pero no se había registrado nada sobre la reforma, por lo que el valor que esta aportara no aparecía en los datos. Ese valor no desapareció. Se distribuyó a través de las tasas de las características que sí se incluyeron, ya que estas características coincidieron con él en las ventas. Para que esto ocurra, deben cumplirse dos condiciones: el valor que falta debe influir en el precio y debe ir asociado a una característica que figure en la lista. Cuando influye pero no va asociado a nada, las tasas son correctas y solo la dispersión es mayor. Esto es así tanto si una característica tiene una tasa como si tiene varias: MARS (Multivariate Adaptive Regression Splines, Friedman 1991), el principal método de regresión de la revista, asigna a una característica diferentes tasas en distintos tramos de sus valores, y allí el valor «prestado» se distribuye a lo largo de esos tramos en lugar de concentrarse en un solo número, pero está presente, por la misma razón.

Para el lector que quiera conocer el mecanismo: . En la regresión ordinaria, una tasa es un coeficiente —los dólares que el modelo asigna a una unidad más de una característica— y el valor tomado en préstamo tiene una magnitud exacta: el propio coeficiente de la característica excluida multiplicado por la intensidad con la que varía junto con la característica incluida, siendo este segundo factor el coeficiente de una regresión auxiliar de la característica excluida sobre las incluidas. El artículo analiza esa fórmula, que corresponde al caso lineal, y luego muestra la misma «fuga» en la cuadrícula, donde las ventas emparejadas del tasador las reflejan de la misma manera que lo hacen los coeficientes. Está dirigido a tasadores que obtienen ajustes a partir de datos de mercado.

Qué es y qué no es el ejemplo de este número. El mercado de Pacífica simulado que se utiliza en las entradas 007 (*Esperanza condicional*) a 010 (*Identificación de parámetros*) trata el estado como una característica con un valor en dólares, tal y como lo hace la tasación tradicional y como se hizo en el número 1, y el ejemplo muestra qué falla en este enfoque. La metodología de valoración tradicional no es el método de la revista, sino la valoración basada en el RCA —el Constraint Approach del número 1 (Craytor 2026)—, que mantiene el estado y la calidad fuera de la regresión; la entrada 007 (*Esperanza condicional*) explica cómo. El lector debe quedarse con la lección del ejemplo, no con su especificación.

El estado, la puntuación de estado y una reforma. Son tres conceptos que a menudo se confunden. *El estado* es el estado en que se encuentra la vivienda: el desgaste de sus acabados, la antigüedad de sus instalaciones y el mantenimiento atrasado. La *puntuación de estado* es una estimación de ese estado, 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 esa zona de mercado y no debería diferir sustancialmente del conocimiento de otros tasadores que operan en la misma zona de mercado. En esta revista, 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. La puntuación es una estimación, y dos viviendas con la misma puntuación pueden diferir entre sí. Una *reforma* no es ni un estado ni una puntuación, sino un acontecimiento: los trabajos realizados, en un periodo reciente, que han modificado la vivienda. Su valor es el valor de lo que ha cambiado —el estado y, por lo general, la calidad, los materiales y los acabados— a los precios que el mercado paga por ellos. La mayor parte del efecto de una reforma suele recaer en el estado y, en menor medida, en la calidad, debido en gran parte a los costes de mano de obra, aunque en algunas reformas el valor añadido a través de la calidad es el mayor de los dos. Una vivienda reformada y una que nunca se ha reformado, que sean idénticas en todas las características actuales que influyen en el valor, deberían tener el mismo valor; el mercado paga por el estado y la calidad que ha aportado la reforma, no por el mero hecho de que se haya llevado a cabo. Por lo tanto, una reforma no es una característica con un precio propio. Cuando un conjunto de datos incluye un indicador de reforma junto a una puntuación de estado, dicho indicador representa todo aquello que la reforma ha aportado y que las demás características del modelo (superficie útil, parcela, vistas y la puntuación de estado) no registran: un estado que va más allá de lo que muestra la puntuación y la calidad, que el conjunto de datos no registra. Así es como debe interpretarse la afirmación del número 1 de *Variable latente*, según la cual la reforma de Comp H «añade aproximadamente \$60,000 a su valor global», y en eso se basa el ejemplo de esta entrada.

Para que se produzca un sesgo, deben cumplirse dos condiciones: la característica que se omite debe influir en el precio y debe variar junto con una de las características incluidas. Si no se cumple alguna de estas condiciones, no hay sesgo.

Este requisito de dos partes es importante porque la versión habitual de la regla —que «un modelo que omite una variable importante da respuestas erróneas»— es falsa tal y como se formula. Si una característica influye en el precio pero no varía junto con ninguna de las características incluidas, amplía la dispersión de los precios en torno al modelo y, por lo tanto, aumenta los errores estándar (la medida de la distancia a la que un coeficiente estimado podría situarse respecto a la cifra que daría una muestra muy grande), pero no modifica los coeficientes en sí mismos; la intersección (el valor constante en dólares del modelo) asume la contribución media de la característica omitida. En este caso, se pierde precisión, no exactitud. Por otro lado, una característica que varía junto con las incluidas pero que no tiene ningún efecto sobre el precio no supone ningún coste. Solo cuando se dan ambas circunstancias se produce un sesgo.

Esto funciona mediante sustitución. Si en el conjunto de variables explicativas (las variables que utiliza una regresión para explicar el precio) falta una variable que influye en el precio, la regresión OLS (mínimos cuadrados ordinarios) no se limita a dejar sin explicar su contribución. Las variables incluidas que varían junto con la que falta actúan como sustitutos parciales de esta, y sus coeficientes incorporan la parte del efecto de la variable que falta que son capaces de predecir. El coeficiente de una variable incluida ya no es solo el efecto de esa variable; es ese efecto junto con una parte del efecto de la variable que falta que ha «tomado prestada».

El sesgo siempre es relativo a una pregunta. Los coeficientes del modelo sin el indicador de

remodelación son la respuesta correcta a una pregunta —cómo varía el precio en función de la puntuación de la condición en estas ventas, teniendo en cuenta todo lo que conlleva dicha puntuación— y la respuesta errónea cuando se interpretan como el valor en dólares por punto de la puntuación de estado del mercado, manteniendo fijos todos los demás factores, que es como los interpreta un ajuste. El modelo no es erróneo porque exista otra característica; sus coeficientes son erróneos para el uso que se les da.

La proporción de influencia tiene un tamaño exacto, determinado por una fórmula. Para una característica excluida w con coeficiente verdadero β_w , el sesgo en el coeficiente de una característica incluida z_j es

$$\text{bias}(\hat{\beta}_j) = \beta_w \cdot \delta_j,$$

, donde $\hat{\beta}_j$ —el sombrero indica una estimación— es el coeficiente que la regresión estima para z_j , y δ_j es el coeficiente de z_j en la regresión auxiliar de w sobre todas las características incluidas. Esta es la fórmula de Theil,¹. Calculada a partir de las regresiones de ajuste, con el coeficiente estimado de la característica omitida, se cumple exactamente en cada muestra; con el coeficiente verdadero, se cumple en promedio.

. La forma de producto corresponde al caso lineal. En un modelo MARS, el efecto de la característica omitida es absorbido por cualquiera de los tramos que la acompañan, por lo que el sesgo se distribuye entre ellos en lugar de concentrarse en un único coeficiente, pero está ahí, y por la misma razón.

La fórmula resulta más útil de lo que parece, ya que a menudo es posible asignar un signo a ambos factores basándose en el conocimiento del mercado, aunque ninguno de ellos pueda medirse. Cuando la característica excluida hace subir el precio ($\beta_w > 0$) y va acompañada de una mayor presencia de una característica incluida, manteniéndose constantes las demás características incluidas ($\delta_j > 0$), el coeficiente de esa característica es demasiado alto. Un tasador que conozca tan solo la dirección de ambas relaciones sabrá en qué sentido se produce el error. Una salvedad: δ_j es una relación en la que las demás características incluidas se mantienen constantes, y cuando hay más de una característica incluida, la correlación simple (una cifra que va de -1 a $+1$ y que mide cómo dos características suben y bajan juntas, ignorando todas las demás) puede ir en sentido contrario. En el ejemplo siguiente, la superficie habitable se correlaciona con el indicador de reforma (correlación $+0.09$), pero su δ es negativo y, por lo tanto, sus coeficientes son demasiado bajos.

Dónde la encuentra el valorador. La cuadrícula comparativa de ventas puede expresarse como un modelo aditivo con la misma estructura que una regresión —unas cuantas características seleccionadas manualmente, con un ajuste en dólares para cada una— y, cuando sus ajustes se extraen de los datos del mercado, está sujeta al sesgo de variables omitidas, al igual que cualquier regresión. Las líneas de la cuadrícula son las características incluidas. Cualquier característica de una vivienda que influya en su precio pero que no figure en una línea queda excluida.

Entre los aspectos que suelen omitirse se incluyen la parte del estado que la puntuación de estado no refleja, la calidad cuando no se registra ninguna puntuación de calidad, la funcionalidad de la distribución, la luz natural, la exposición al ruido, la ubicación concreta dentro de la manzana, la motivación del vendedor y el esfuerzo de comercialización dedicado al anuncio. Cada uno de estos factores puede influir en el precio, y muchos de ellos van de la mano de aspectos que la cuadrícula sí tiene en cuenta.

La reforma es un caso ilustrativo, ya que queda registrada y, aun así, sigue habiendo lagunas. El URAR (Informe Uniforme de Tasación Residencial) recoge si la cocina y los baños se han actualizado

¹Theil, H. (1957). Specification Errors and the Estimation of Economic Relationships. *Revue de l'Institut International de Statistique*, 25(1/3), 41–51.

o reformado, y en qué periodo, y las propias definiciones de estado del UAD (Conjunto de Datos Uniforme de Tasación) dependen del alcance de la actualización. Sin embargo, la cuadrícula incluye líneas para el estado y para la calidad, pero ninguna para la reforma —y con razón, ya que cuando la cuestión es la valoración de mercado del estado, las líneas de estado y calidad ya reflejan lo que ha aportado la reforma, por lo que una tercera línea lo contabilizaría dos veces—. Cuando la cuestión es, en cambio, qué efecto tendría una reforma propuesta en el precio de esta vivienda, mantener fijas el estado y la calidad elimina precisamente el canal a través del cual actúa la reforma; la entrada 009 (*Endogeneidad*) aborda esa cuestión. La fuga se produce a través de lo que las puntuaciones no recogen. En un mercado en el que las puntuaciones más altas de estado corresponden principalmente a reformas recientes, y las reformas aportan un valor que las puntuaciones no registran, un ajuste por estado extraído de esas ventas en la tasación tradicional —mediante ventas emparejadas o mediante una regresión que incluya el estado— lleva consigo ese valor. Según la fórmula de Theil, esos ajustes resultan entonces demasiado elevados: incluyen no solo el efecto propio del estado, sino también una parte de lo que las puntuaciones no registraron. Un tasador que lo tome y lo aplique a una vivienda con la puntuación máxima —que se ha ganado dicha puntuación gracias al mantenimiento— realizará ajustes excesivos. Según el «Constraint Approach» de The Journal, no se extrae ninguna tasa de estado. El tasador divide el residuo de cada inmueble comparable entre el estado, la calidad y las demás características que los datos no miden, según su criterio, bajo una única restricción: la suma de las contribuciones de valor de los componentes residuales debe ser igual al residuo del inmueble comparable. El valor de una reforma se divide entonces según lo considere el tasador, en lugar de figurar en una sola partida.

La «fuga» que acabamos de describir es una versión más precisa de la objeción que plantean los tasadores respecto a los ajustes basados en el mercado, a saber, que los ajustes derivados de un conjunto de datos no se pueden extrapolar fácilmente. La razón es que un ajuste extraído consta tanto del efecto identificado como de los efectos no identificados que lo acompañan, y su composición depende de la relación entre los distintos elementos en el conjunto de datos del que se ha extraído.

Otro ejemplo es la situación en la que las «ventas emparejadas» en la tasación tradicional —dos ventas idénticas en todo salvo en la característica en cuestión, interpretándose la diferencia de precio como el valor de dicha característica— arrojan una cifra inverosímilmente alta. La causa puede no ser un error de medición, sino una «fuga»: las dos ventas difieren en la característica mencionada y en algo que va asociado a ella, y la diferencia de precio se debe a ambos factores.

Por qué importa para la solidez. Un informe que no aborde el sesgo por variables omitidas deja sus ajustes hedónicos y de cuadrícula expuestos a esa objeción.

El sesgo no se puede detectar a partir de las estadísticas de ajuste. Un modelo puede tener un alto índice de explicación de la variación (R^2 , es decir, la proporción de la variación en los precios que reproduce el modelo), residuos pequeños (las diferencias entre los precios de venta y los precios del modelo) y coeficientes que superen la prueba estadística habitual (una prueba de t , que comprueba si es plausible que un coeficiente sea cero), y, aun así, presentar un sesgo considerable, ya que el sesgo proviene de cómo se relacionan las características entre sí, y no del tamaño de los residuos. Utilizar el « R^2 » como prueba de que el modelo es fiable no responde a la pregunta. En el ejemplo siguiente, el « R^2 » es de 0,97 y los coeficientes condicionales tienen un error del 74 por ciento.

Más datos no resolverán el problema. El sesgo no disminuye a medida que crece la muestra. Mil ventas del mismo mercado en las que se omite la misma característica dan lugar a una estimación centrada en la misma cifra errónea, pero estimada con mayor precisión: el error estándar se reduce en torno a la estimación sesgada, lo cual es peor que un intervalo amplio a su

alrededor, ya que el informe parecerá más fiable. Lo mismo ocurre con MARS: un mayor número de ventas hace que la superficie ajustada sea más estable, pero en torno a la misma respuesta sesgada; el intervalo de «`varmod`» de la Tierra mide la dispersión de los precios, no el sesgo. Esto corrige la «*variable latente*» del Número 1, que afirmaba que el sesgo se reduciría con más comparables y que los demás coeficientes presentaban un sesgo al alza. El sesgo no se reduce con un mayor número de ventas y, tanto en las ocho ventas del propio número 1 como en las dieciséis que se muestran a continuación, los coeficientes de vista, superficie bruta alquilable (GLA) y parcela presentan un sesgo a la baja; solo el coeficiente de condición presenta un sesgo al alza.

A menudo es posible justificar la dirección de un sesgo incluso cuando no se conoce su magnitud. En la tasación tradicional, donde el ajuste por estado se extrae de las ventas —mediante ventas emparejadas o mediante una regresión que incluye el estado—, el enfoque práctico no consiste en afirmar que el ajuste es insesgado, sino en identificar las características plausibles que se han omitido, utilizar el conocimiento del mercado para asignar a cada una un signo para β_w y para δ_j , y indicar la dirección resultante del sesgo. «Probablemente, los ajustes por estado se hayan sobrevalorado. La mayoría de las ventas con puntuaciones más altas corresponden a reformas recientes, y dichas reformas han aportado mejoras que los datos no registran, como cocinas nuevas y acabados. Esas mejoras elevan el precio y se dan principalmente en las ventas con puntuaciones altas, por lo que parte de su valor se ha contabilizado como estado». Esta es una afirmación razonable. Se cumplen ambos requisitos, y ambos apuntan en la misma dirección: las mejoras no tenidas en cuenta elevan el precio (el β_w es positivo) y van acompañadas de puntuaciones altas (el δ_j es positivo), por lo que el sesgo es al alza. Una afirmación de este tipo convierte un problema no cuantificado en uno que se reconoce y al que se le da una orientación, que es exactamente lo que necesita el lector para sopesar la conclusión. En el Constraint Approach no se extrae ninguna tasa de estado, por lo que no se plantea la cuestión del estado.

Ejemplo de valoración resuelto. La ampliación del conjunto de datos Pacífica, que incluye dieciséis inmuebles, recoge, para cada inmueble, la superficie habitable bruta (GLA), el tamaño de la parcela, las vistas, la puntuación de estado, un indicador de reforma observada y el precio de venta. Dado que cada precio se calculó a partir de una fórmula conocida sin ningún componente aleatorio (los datos están libres de ruido), la regresión que utiliza las características adecuadas reproduce esa fórmula con exactitud, y cualquier desviación respecto a ella es un sesgo sin ningún componente de ruido.

Comp	GLA (sf)	Parcela (sf)	Ver	Cond	Reformado	Precio (\$)
A	1,650	9,000	Sí	7,00	No	1,425,000
B	1,650	7,200	No	7,00	No	1,280,000
C	1,850	7,500	No	7,00	No	1,407,500
D	1,450	7,000	No	7,00	No	1,155,000
E	1,700	8,000	Sí	9,00	No	1,470,000
F	1,550	7,100	No	5,00	No	1,177,500
G	1,700	10,500	No	7,00	No	1,392,500
H	1,600	7,300	No	7,00	Sí	1,312,500
I	1,750	8,200	No	9,00	Sí	1,465,000
J	1,600	7,600	No	7,00	Sí	1,320,000
K	1,900	9,500	Sí	9,00	Sí	1,687,500
L	1,500	6,800	No	5,00	No	1,140,000
M	1,800	8,800	No	7,00	No	1,410,000
N	1,650	7,400	Sí	7,00	No	1,385,000
O	1,550	7,900	No	9,00	Sí	1,337,500
P	1,720	9,200	No	5,00	No	1,332,000

Los productos A-H son comparables al Número 1, y sus atributos y precios publicados no han variado; La Edición 1 describía el estado con palabras y lo codificaba del 1 al 3 para su regresión; según la puntuación de esta revista, las palabras se corresponden con «normal» (5,00), «bueno» (7,00) y «excelente» (9,00), por lo que un paso de la Edición 1 equivale a dos puntos. La reforma de Comp H —revelada en el Número 1 como el origen de su residuo, pero no registrada como variable— ahora sí se ha registrado. El mercado paga \$20,000 por cada punto de estado. En los datos contruidos, una vivienda reformada se vende por \$60,000 más que una vivienda con la misma superficie bruta alquilable (GLA), parcela, vistas y puntuación de estado. Esos \$60,000 no son el precio de la reforma en sí; representan lo que la reforma ha aportado y que las demás características no reflejan. Las viviendas reformadas también obtienen puntuaciones medias más altas (8,20 frente a 6,64), por lo que el indicador de reforma varía en función de la puntuación de estado incluida, que es el segundo requisito del sesgo: $\text{corr}(\text{remodel}, \text{cond}) = 0.55$.

Dos regresiones sobre los mismos dieciséis inmuebles:

	Se ha omitido la remodelación	La reforma estaba incluida	Sesgo
Interceptación	-83,923.57	-30,000	-53,923.57
GLA	586,64	600	-13.36
Lote	24,32	25	-0.68
Ver	79,178,46	100,000	-20,821.54
Condición (por punto)	34,815,56	20,000	+14,815.56
Reforma	—	60,000	
R^2	0,9733	1,0000	

El valor de \$60,000 en la línea de «remodelación» no es el precio de la remodelación. En este conjunto de datos construido, se trata de la diferencia de precio asociada al indicador de «remodelación» cuando la superficie bruta alquilable (GLA), la parcela, las vistas y la puntuación son iguales; eso es lo que «recupere la función generadora».

El coeficiente de condición es demasiado alto en \$14,816 por punto —informa \$34,816 por punto de condición, mientras que la cifra del mercado es \$20,000, lo que supone una sobreestimación del 74 por ciento. El coeficiente de «view» es demasiado bajo en \$20,822. Ambos errores son invisibles en el ajuste: el modelo sin el indicador «remodel» tiene $R^2 = 0.9733$ y unos residuos que, salvo tres ventas cercanas a \$35,000, parecen normales. Nada en R^2 ni en las pruebas de los coeficientes indica ningún problema (si estos residuos hubieran sido ruido, todos los coeficientes habrían superado esa prueba al nivel habitual del 5 por ciento).

La fórmula de Theil explica los números con exactitud. Al realizar una regresión del indicador «remodel» sobre las cuatro características incluidas, se obtienen $\delta_{\text{cond}} = 0.24692602$ y $\delta_{\text{view}} = -0.34702564$. Entonces,

$$\beta_w \cdot \delta_{\text{cond}} = 60,000 \times 0.24692602 = 14,815.56, \quad \beta_w \cdot \delta_{\text{view}} = 60,000 \times (-0.34702564) = -20,821.54,$$

lo que coincide con los desplazamientos observados al céntimo. El sesgo no se explica aproximadamente mediante la fórmula; es la fórmula misma. El ruido cambiaría menos de lo que cabría esperar: con datos ruidosos, la fórmula sigue siendo válida exactamente entre las dos regresiones de ajuste (sin y con el indicador de remodelación), utilizando el coeficiente estimado del indicador $\hat{\beta}_w$ en lugar de β_w ; en comparación con el coeficiente real, se cumple de media en muchas muestras.

Merece la pena volver a examinar el coeficiente de vista, ya que su sesgo va en la dirección opuesta y el motivo resulta instructivo. Manteniendo constantes las demás características, los inmuebles con vistas en esta muestra tienen *menos* probabilidades de ser remodeladas ($\delta_{\text{view}} < 0$).

Por lo tanto, el coeficiente de vistas se ve reducido en lugar de aumentado: es \$20,822 demasiado bajo, mientras que el coeficiente de estado es \$14,816 demasiado alto. Los dos errores son efectos independientes de la misma omisión, no una transferencia entre ellos. Un tasador que se basara únicamente en el estado no habría previsto en absoluto un error en el ajuste por vistas; sin embargo, un ajuste por vistas subestimado en un 21 por ciento es la consecuencia directa de la misma omisión.

En la práctica, los daños se concentran precisamente donde los tasadores son menos propensos a buscarlos. Consideremos primero qué hace el modelo sesgado como *predictor*. En un inmueble objeto sin reformar idéntico a Comp N —1,650 pies cuadrados de superficie bruta alquilable (GLA) en una parcela de 7,400 pies cuadrados con vistas y una puntuación de estado de 7,00 — indica un valor de \$1,386,914 frente a un precio medio real de \$1,385,000 para viviendas con exactamente estas características — según la fórmula de cinco características de la entrada 007 (*Esperanza condicional*), con la remodelación = 0 — lo que supone un error de \$1,914, o un 0,14 por ciento. En las dieciséis ventas, el error típico es de aproximadamente \$21,000, por lo que el modelo presenta un sesgo considerable y, aun así, ofrece predicciones aceptables, lo cual no es una coincidencia: el método OLS realiza el ajuste a la muestra que se le proporciona, y los coeficientes sesgados son precisamente los que más se acercan a los precios de la muestra. La precisión de la predicción no es una prueba en contra del sesgo por variables omitidas, y la validación cruzada —que consiste en probar el modelo con ventas excluidas del ajuste— tampoco lo descarta: unos buenos resultados de validación cruzada no demuestran que los coeficientes estén libres de sesgo, ya que un modelo mal especificado puede predecir bien siempre que las características se combinen tal y como lo hicieron en el ajuste. Al excluir cada característica por turnos, los dos errores más grandes —alrededor de \$56,000 en K y \$48,000 en G— se producen en inmuebles que siguen el patrón de esta muestra, en los que las puntuaciones de estado elevadas pertenecen en su mayoría a inmuebles reformados; E, que rompe este patrón, ocupa el tercer lugar con unos \$45,000, y H y J, que también lo rompen, ocupan el cuarto y quinto lugar con unos \$41,000. Los datos aquí presentados no contienen ruido, por lo que cada uno de estos errores proviene del indicador de remodelación omitido, y ninguno del ruido; en las ventas reales, donde también hay ruido, el tamaño de un error no indica en absoluto si proviene de una característica omitida o del residuo habitual.

El error se produce en el momento en que los coeficientes se utilizan como *ajustes* en lugar de como regla de predicción. Un tasador que extraiga un ajuste por estado del inmueble a partir de este modelo aplica \$34,816 por punto, mientras que la cifra del mercado es \$20,000 — \$29,631 de más en cada inmueble comparable ajustado en los dos puntos, de «bueno» a «excelente» — y un ajuste por vistas de \$79,178, cuando la cifra del mercado es \$100,000. La cuadrícula es donde se leen los coeficientes de una regresión uno a uno, y esa es precisamente la lectura que el sesgo destruye.

La predicción también falla, y de forma estrepitosa, en los inmuebles que rompen el patrón en el que se basa el sesgo. Un inmueble bien conservado como el objeto (1,650 pies cuadrados de superficie bruta alquilable, parcela de 7,400 pies cuadrados), pero sin vistas, con una puntuación excelente de 9,00 y que nunca se ha reformado, aparece indicado en \$1,377,367 frente a un valor real de \$1,325,000, sobrevalorada en \$52,367 (3,95 %), ya que, en estas ventas, una puntuación de 9,00 solía ir acompañada de una reforma, y el índice de estado del modelo incluye el valor de la reforma \$60,000 junto con la puntuación. Una vivienda reformada, por lo demás idéntica al inmueble objeto de estudio (con vistas, puntuación 7,00), tiene un precio indicado de \$1,386,914 frente a \$1,445,000, infravalorada en \$58,086 (4,02 %). Se trata de inmuebles atípicos, es decir, aquellos para cuya valoración se suele contratar específicamente a un tasador, y aquellos para los que el informe se leerá con mayor atención.

References.

- 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>
- Theil, H. (1957). Specification Errors and the Estimation of Economic Relationships. *Revue de l'Institut International de Statistique*, 25(1/3), 41–51.

Cross-references: conditional expectation; hedonic regression; implicit price; latent variable; endogeneity.

Addendum: Prolog Formalization

The following machine-readable companion (008-omitted-variable-bias.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>.

%% 008-omitted-variable-bias.pl -- companion to Foundations Vol. 1, Issue 2,
%% Entry 008, Omitted Variable Bias (omitted_variable_bias), v1.2.0.
%% Checks: theil_formula_holds/0 (each bias = beta_omitted * delta_j to the
%% cent); remodel_reading_holds/0 (the remodel indicator's 60,000).

% 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').

uses_terms([ omitted_variable_bias, conditional_expectation,
             hedonic_regression, implicit_price, latent_variable,
             multicollinearity, ordinary_least_squares, residual,
             adjustment, comp_selection, paired_sales_analysis,
             sales_comparison_approach, uspap ]).
uses_relations([depends_on/2, related/2, introduced_by/2]).

article_concept(omitted_variable_bias).
article_issue(volume(1), number(2), year(2026)).
article_id('foundations.2026.008').

:- 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)).

kind(condition,          state).
kind(condition_score,    measurement_of(condition)).
```

```

kind(quality,          state).
kind(remodel,          event).
changes(remodel, condition).
changes(remodel, quality).

stand_in(remodel_numeric, unrecorded_condition).
stand_in(remodel_numeric, unrecorded_quality).

ovb_condition(affects_outcome,
  "The omitted variable must have a nonzero effect on the outcome.").
ovb_condition(correlated_with_included,
  "The omitted variable must have a nonzero coefficient on at least one included regressor in the
  ↪ auxiliary regression (partial, not merely marginal, correlation).").

biased(BetaOmitted, Delta) :-
  BetaOmitted =\= 0,
  Delta =\= 0.

consequence(affects_but_uncorrelated,
  "Inflates residual variance and widens standard errors; slope coefficients remain unbiased (the
  ↪ intercept absorbs the omitted variable's mean).").
consequence(correlated_but_no_effect,
  "No cost at all.").

theil_bias(BetaOmitted, Delta, Bias) :-
  Bias is BetaOmitted * Delta.

omitted_variable(remodel_numeric).
omitted_variable_effect(remodel_numeric, 60000).

condition_scale(0.0, 10.0).
condition_score(average, 5.0).
condition_score(good, 7.0).
condition_score(excellent, 9.0).
condition_rate(20000).

coefficient(remodel_omitted, intercept, -83923.571).
coefficient(remodel_omitted, gla_sqft, 586.645).
coefficient(remodel_omitted, lot_sqft, 24.323).
coefficient(remodel_omitted, view_numeric, 79178.461).
coefficient(remodel_omitted, cond_score, 34815.561).

coefficient(remodel_included, intercept, -30000).
coefficient(remodel_included, gla_sqft, 600).
coefficient(remodel_included, lot_sqft, 25).
coefficient(remodel_included, view_numeric, 100000).
coefficient(remodel_included, cond_score, 20000).
coefficient(remodel_included, remodel_numeric, 60000).

aux_slope(intercept, -0.89872625).
aux_slope(gla_sqft, -0.00022259).
aux_slope(lot_sqft, -0.00001129).
aux_slope(view_numeric, -0.34702564).
aux_slope(cond_score, 0.24692602).

model_fit(remodel_omitted, r_squared, 0.9733).
model_fit(remodel_included, r_squared, 1.0000).

observed_bias(R, Bias) :-

```

```

coefficient(remodel_omitted, R, Omitted),
coefficient(remodel_included, R, True),
Bias is Omitted - True.

theil_formula_holds :-
  omitted_variable_effect(_, BetaOmitted),
  forall(( observed_bias(R, Observed), aux_slope(R, Delta) ),
    ( theil_bias(BetaOmitted, Delta, Predicted),
      abs(Observed - Predicted) < 0.01 )).

remodel_reading_holds :-
  omitted_variable_effect(remodel_numeric, Beta),
  coefficient(remodel_included, remodel_numeric, Coef),
  abs(Coef - Beta) < 0.01.

reported_bias(cond_score, 14815.56).
reported_bias(view_numeric, -20821.54).

relative_bias(R, Pct) :-
  observed_bias(R, Bias),
  coefficient(remodel_included, R, True),
  True =\= 0,
  Pct is 100 * Bias / True.

use_of_coefficients(prediction,
  "Evaluate the whole fitted surface at a point; biases in individual coefficients can offset.").
use_of_coefficients(adjustment,
  "Read one coefficient at a time as an implicit price; bias enters directly and does not offset
  ↪ .").

prediction_check(subject_typical, remodel_omitted, 1386914).
prediction_check(subject_typical, remodel_included, 1385000).

prediction_check(excellent_unremodeled, remodel_omitted, 1377367).
prediction_check(excellent_unremodeled, remodel_included, 1325000).
prediction_check(remodeled_with_view, remodel_omitted, 1386914).
prediction_check(remodeled_with_view, remodel_included, 1445000).

atypical_property_fails_worse(S) :-
  prediction_check(S, remodel_omitted, A),
  prediction_check(S, remodel_included, B),
  abs(A - B) > 50000.

defensibility_consequence(fit_statistics,
  "Bias is undetectable from R-squared or residual size; it lives in the correlation structure,
  ↪ not the residuals.").
defensibility_consequence(sample_size,
  "Bias does not shrink with n; more data gives a tightly estimated wrong number.").
defensibility_consequence(direction_disclosure,
  "Signing beta_omitted and delta_j from market knowledge yields a defensible directional
  ↪ statement even when magnitude is unknown.").

practice_encounter(what_the_scores_miss,
  "Traditional appraisal: a condition adjustment extracted where top scores go with remodels
  ↪ carries their unrecorded value. RCA extracts no condition rate.").
practice_encounter(no_remodel_line,
  "The URAR records kitchen and bathroom updates; the grid has condition and quality lines and no
  ↪ remodel line, rightly: a remodel line would count the same value twice.").
practice_encounter(adjustments_do_not_travel,

```

```
"An extracted adjustment is a composite of the named effect and unnamed effects correlated with  
↪ it, so it is dataset-specific.").  
practice_encounter(implausible_paired_sale,  
  "A paired-sales adjustment far larger than cost can reflect leakage from a correlated omitted  
↪ attribute.").
```