

Viés de variável 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)

Definição formal. O viés de variável omitida é a distorção sistemática introduzida em um coeficiente estimado quando uma característica que afeta o preço é deixada de fora do modelo e essa característica se move em conjunto com uma das características incluídas — se move com ela após as outras características incluídas tenham sido consideradas, que é o que a regressão auxiliar (uma segunda regressão, da característica omitida em relação às incluídas) mede. O efeito da característica omitida, então, se reflete nos coeficientes das características incluídas (os dólares por unidade que o modelo atribui a cada característica).

Enquadramento intuitivo. Um valor pode estar incorreto porque incorpora o valor de outra característica. O Comp H da Edição 1 foi vendido por \$46,455 a mais do que a regressão baseada em suas quatro características registradas havia previsto. O motivo, divulgado na seção de *variáveis latentes* da Edição 1, foi uma recente reforma na cozinha, que, segundo essa entrada, agrega cerca de \$60,000 ao valor da casa. O estado de conservação constava na lista de características, mas nada registrava a reforma, e o que quer que a reforma tenha acrescentado estava ausente dos dados. Esse valor não desapareceu. Ele foi distribuído pelas taxas das características que estavam incluídas, porque essas características, por acaso, estavam associadas a ele nas vendas. Duas condições precisam ser verdadeiras para que isso aconteça: o valor ausente deve ser relevante para o preço e deve estar associado a uma característica que conste na lista. Quando ele é relevante, mas não está associado a nada, as taxas estão corretas e apenas a dispersão é maior. Isso se aplica independentemente de uma característica receber uma ou várias taxas: O MARS (Multivariate Adaptive Regression Splines, Friedman 1991), o principal método de regressão da revista, atribui a uma característica taxas diferentes em diferentes intervalos de seus valores; nesse caso, o valor emprestado é distribuído por esses intervalos em vez de ficar concentrado em um único número, mas ele está presente, pela mesma razão.

Para o leitor que deseja conhecer o mecanismo. Na regressão comum, uma taxa é um coeficiente — o valor em dólares que o modelo atribui a mais uma unidade de uma variável — e o valor calculado tem uma magnitude exata: o próprio coeficiente da característica excluída multiplicado pela intensidade com que ela varia em função da característica incluída, sendo esse segundo fator o coeficiente de uma regressão auxiliar da característica excluída em relação às incluídas. O artigo explica essa fórmula, que corresponde ao caso linear, e em seguida mostra o mesmo desvio na grade, onde as vendas emparelhadas dos avaliadores refletem isso da mesma forma que os coeficientes. Ele foi elaborado para avaliadores que obtêm ajustes a partir de dados de mercado.

O que o exemplo desta edição é — e o que não é. O mercado da Pacífica simulado, utilizado nas entradas 007 (*Expectativa Condicional*) a 010 (*Identificação de Parâmetros*), trata a condição como uma característica com um valor em dólares, tal como faz a avaliação tradicional e como fez a Edição 1, e o exemplo mostra o que há de errado nisso. A metodologia tradicional de avaliação não é o método adotado pela Revista, a avaliação baseada na RCA — a Residual Constraint da Edição 1 (Craytor 2026) —, que mantém a condição e a qualidade fora da regressão; a entrada 007 (*Expectativa Condicional*) explica como. O leitor deve reter a lição do exemplo, e não sua especificação.

Condição, escore de condição e remodelação. Três conceitos que costumam ser confundidos. A *condição* é o estado da casa: o desgaste de seus acabamentos, a idade de seus sistemas e a manutenção adiada. O *escore de condição* é uma estimativa desse estado, em relação à condição das casas vizinhas no seu mercado definido. Ela se baseia no entendimento do avaliador sobre esse mercado e não deve entrar em conflito substancial com o entendimento de outros avaliadores que atuam no mesmo mercado. Nesta revista, ela varia de 0,00 a 10,00: a porcentagem de casas no mercado que estão em pior condição, dividida por dez. Uma pontuação de 5,00 significa que metade das casas no mercado está em pior estado; 9,00 significa que nove em cada dez estão. O escore é uma estimativa, e duas casas com o mesmo escore podem apresentar diferenças. Uma *reforma* não é nem um estado nem um escore, mas um evento: obra realizada, em algum período recente, que alterou a casa. Seu valor é o valor do que foi alterado — condição e, geralmente, qualidade, materiais e acabamentos — aos preços que o mercado paga por esses itens. A maior parte do efeito de uma reforma costuma recair sobre a condição e, em menor medida, sobre a qualidade, em grande parte devido aos custos de mão de obra, embora em algumas reformas o valor agregado por meio da qualidade seja o maior dos dois. Uma casa reformada e uma que nunca passou por reforma, que sejam semelhantes em todas as características atuais que influenciam o valor, devem ter o mesmo valor; o mercado paga pelo estado de conservação e pela qualidade que a reforma produziu, não pelo fato de que uma reforma ocorreu. Portanto, uma reforma não é uma característica com um preço próprio. Quando um conjunto de dados apresenta um indicador de reforma ao lado de um escore de estado de conservação, esse indicador representa tudo o que a reforma trouxe e que as outras características do modelo (área útil, terreno, vista e o escore de estado de conservação) não registram: estado de conservação além do que o escore mostra e qualidade, que o conjunto de dados não registra. É assim que deve ser interpretada a afirmação da Edição 1 em *Variáveis Latentes*, de que a reforma de Comp H “acrescenta aproximadamente \$60,000 ao valor do pacote”, e é nisso que se baseia o exemplo desta entrada.

Para que ocorra um viés, duas condições devem ser satisfeitas simultaneamente: a característica que foi deixada de fora deve ter um efeito sobre o preço e deve variar em conjunto com uma das características incluídas. Se qualquer uma dessas condições não for satisfeita, não há viés.

Esse requisito de duas partes é importante porque a versão comum da regra — de que “um modelo que omite uma variável importante fornece respostas erradas” — é falsa tal como está formulada. Se uma característica influencia o preço, mas não varia em conjunto com nenhuma das características incluídas, ela amplia a dispersão dos preços em torno do modelo e, conseqüentemente, amplia os erros-padrão (a medida de quão distante um coeficiente estimado pode estar do valor que uma amostra muito grande forneceria), mas não altera os próprios coeficientes; a interceptação (o valor constante em dólares do modelo) absorve a contribuição média da característica deixada de fora. Nesse caso, perde-se a precisão, não a exatidão. Por outro lado, uma característica que varia em conjunto com as incluídas, mas não tem efeito sobre o preço, não acarreta nenhum custo. É somente quando ambas as condições se verificam que ocorre o viés.

Isso funciona por meio da substituição. Se uma variável que afeta o preço estiver ausente dos regressores (as variáveis que uma regressão utiliza para explicar o preço), a regressão OLS (mínimos quadrados ordinários) não deixa simplesmente sua contribuição sem explicação. As variáveis incluídas que variam em conjunto com a ausente servem como substitutos parciais para ela, e seus coeficientes incorporam a parte do efeito da variável ausente que conseguem prever. O coeficiente de uma variável incluída não representa mais apenas o efeito dessa variável; ele representa esse efeito juntamente com uma parcela do efeito da variável ausente que ela incorporou.

O viés é sempre relativo a uma pergunta. Os coeficientes do modelo sem o indicador de remodelação são a resposta correta para uma pergunta — como o preço se varia em função do escore da condição nessas vendas, considerando todos os fatores associados ao escore — e a resposta errada

quando interpretados como o valor em dólares por ponto do escore de condição, com todos os outros fatores mantidos constantes, que é como os ajustes os interpretam. O modelo não está errado porque existe outra característica; seus coeficientes estão errados para o uso a que se destinam.

A parcela emprestada tem um tamanho exato, determinado por uma fórmula. Para uma característica excluída w com coeficiente verdadeiro β_w , o viés no coeficiente de uma característica incluída z_j é

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

, onde $\hat{\beta}_j$ — o símbolo “hat” indica uma estimativa — é o coeficiente que a regressão estima para z_j , e δ_j é o coeficiente de z_j na regressão auxiliar de w em todas as características incluídas. Essa é a fórmula de Theil,¹. Calculada a partir das regressões de ajuste, com o coeficiente estimado da característica deixada de fora, ela se mantém exatamente em todas as amostras; com o coeficiente verdadeiro, ela se mantém em média.

. A forma de produto corresponde ao caso linear. Em um modelo MARS, o efeito da característica excluída é absorvido por qualquer um dos trechos que a acompanham; assim, o viés é distribuído entre eles, em vez de concentrado em um único coeficiente, mas ele está presente, e pela mesma razão.

A fórmula é mais útil do que parece, já que muitas vezes é possível atribuir um sinal a ambos os fatores com base no conhecimento do mercado, mesmo que nenhum deles possa ser medido. Quando a característica excluída eleva o preço ($\beta_w > 0$) e está associada a uma maior presença de uma característica incluída, mantendo-se as demais características incluídas constantes ($\delta_j > 0$), o coeficiente dessa característica é alto demais. Um avaliador que conheça apenas a direção dessas duas relações já sabe a direção do erro. Uma ressalva: δ_j é uma relação em que as outras características incluídas são mantidas constantes; e, com mais de uma característica incluída, a correlação simples (um número que varia de -1 a $+1$, que mede como duas características aumentam e diminuem juntas, ignorando todas as outras) pode apontar na direção oposta. No exemplo abaixo, a área habitável está relacionada ao indicador de reforma ($+0.09$ de correlação), mas seu δ é negativo e, portanto, seus coeficientes são muito baixos.

Onde o avaliador a encontra. A tabela de comparação de vendas pode ser representada como um modelo aditivo com a mesma estrutura de uma regressão — algumas características selecionadas manualmente, com um ajuste em dólares para cada uma — e, quando seus ajustes são extraídos de dados de mercado, ela está sujeita ao viés de variável omitida, da mesma forma que qualquer regressão. As linhas da tabela representam as características incluídas. Qualquer característica de uma casa que influencie seu preço, mas que não seja uma linha, é deixada de fora.

Entre os aspectos comumente deixados de lado estão a parte do estado do imóvel que o escore de condição não captura, a qualidade quando não há registro de escore de qualidade, a funcionalidade da planta baixa, a iluminação natural, a exposição ao ruído, a localização específica dentro do quarteirão, a motivação do vendedor e o esforço de marketing dedicado ao anúncio. Cada um desses fatores pode afetar o preço, e muitos deles estão associados a aspectos que a tabela de avaliação leva em consideração.

A reforma é um caso ilustrativo, pois é registrada e ainda assim apresenta falhas. O URAR (Relatório Uniforme de Avaliação Residencial) registra se a cozinha e os banheiros foram modernizados ou reformados, e em que período, e as próprias definições de condição do UAD (Conjunto de Dados Uniforme de Avaliação) dependem da extensão da modernização. A tabela, no entanto, possui linhas para condição e para qualidade, mas nenhuma para reforma — e com razão, quando a questão é

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

a taxa de mercado para a condição, e as linhas de condição e qualidade já capturam o que a reforma trouxe, já que uma terceira linha acabaria contabilizando isso duas vezes. Quando a questão é, ao contrário, qual seria o impacto de uma reforma proposta no preço dessa casa, manter a condição e a qualidade fixas remove justamente o canal pelo qual a reforma atua; a entrada 009 (*Endogeneidade*) aborda essa questão. A distorção surge por meio do que os escores deixam de registrar. Em um mercado onde os escores mais altos de estado de conservação pertencem principalmente a reformas recentes, e as reformas agregam valor que os escores não registram, um ajuste de estado de conservação extraído dessas vendas na avaliação tradicional — por meio de vendas emparelhadas ou por uma regressão que inclua o estado de conservação — carrega esse valor consigo. Pela fórmula de Theil, os ajustes ficam então muito altos: eles incluem não apenas o efeito próprio do estado de conservação, mas também uma parcela do que os escores não registraram. Um avaliador que adote esses ajustes e os aplique a uma casa com escore máximo — obtido graças à manutenção — fará um ajuste excessivo. De acordo com a Residual Constraint do Journal, nenhuma taxa de condição é extraída. O avaliador divide o resíduo de cada imóvel comparável entre o estado de conservação, a qualidade e as outras características que os dados não medem, conforme seu julgamento, sob uma única restrição: a soma das contribuições dos valores dos componentes residuais deve ser igual ao resíduo do imóvel comparável. O valor de uma reforma é, então, dividido conforme o avaliador julgar adequado, e não atribuído a uma única linha.

A distorção descrita acima é uma versão mais precisa da objeção que os avaliadores têm em relação aos ajustes baseados no mercado, a saber, que os ajustes derivados de um conjunto de dados não se aplicam bem a outros. A razão é que um ajuste extraído consiste tanto no efeito identificado quanto nos efeitos não identificados que o acompanham, e sua composição depende de quais elementos estavam associados a quais no conjunto de dados do qual foi extraído.

Outro exemplo é a situação em que as “vendas emparelhadas” na avaliação tradicional — duas vendas idênticas em tudo, exceto na característica em questão, sendo que a diferença de preço é interpretada como o valor dessa característica — resultam em um valor absurdamente alto. A causa pode não ser um erro de medição, mas sim um “vazamento”: as duas vendas diferem na característica mencionada e em algo que a acompanha, e a diferença de preço se deve a ambos os fatores.

Por que importa para a solidez. Um relatório que não aborde o viés de variável omitida deixa seus ajustes hedônicos e de grade suscetíveis a essa objeção.

O viés não pode ser identificado a partir das estatísticas de ajuste. Um modelo pode apresentar um alto índice de R^2 (a parcela da variação nos preços que o modelo reproduz), pequenos resíduos (as diferenças entre os preços de venda e os preços do modelo) e coeficientes que passam no teste estatístico usual (um teste de t , que verifica se um coeficiente poderia plausivelmente ser zero), e ainda assim apresentar um viés significativo, pois o viés decorre da relação de dependência entre as variáveis, e não do tamanho dos resíduos. Utilizar o R^2 como prova de que o modelo é confiável não responde à questão. No exemplo abaixo, o R^2 é 0,97 e os coeficientes condicionais apresentam um erro de 74 por cento.

Mais dados não resolverão o problema. O viés não diminui à medida que a amostra cresce. Mil vendas do mesmo mercado, com a mesma característica omitida, resultam em uma estimativa centrada no mesmo valor errado, mas estimada com maior precisão: o erro-padrão diminui em torno da estimativa viciada, o que é pior do que um intervalo amplo em torno dela, já que o relatório parecerá mais confiável. O mesmo vale para o MARS: mais vendas tornam a superfície ajustada mais estável, mas em torno da mesma resposta enviesada; o intervalo **varmod** da Terra mede a dispersão dos preços, não o viés. Isso corrige as *variáveis latentes* da Edição 1, que afirmavam que o viés diminuiria com mais comparações comparáveis e que os outros coeficientes estavam enviesados

para cima. O viés não diminui com mais vendas e, nas próprias oito vendas da Edição 1, assim como nas dezesseis abaixo, os coeficientes de vista, área bruta locável (GLA) e lote apresentam viés para baixo; apenas o coeficiente de condição apresenta viés para cima.

Muitas vezes, é possível justificar a direção de um viés, mesmo quando sua magnitude não pode ser determinada. Na avaliação tradicional, em que um ajuste de condição é extraído das vendas — por meio de vendas emparelhadas ou de uma regressão que inclui a condição —, a abordagem prática não é afirmar que o ajuste é insesgado, mas sim identificar as características plausíveis que foram deixadas de fora, usar o conhecimento do mercado para atribuir a cada uma delas um sinal para β_w e para δ_j , e indicar a direção resultante do viés. “Os ajustes de condição provavelmente estão superestimados. A maioria das vendas com escore mais alto são reformas recentes, e essas reformas trouxeram melhorias que os dados não registram, como cozinhas novas e acabamentos. Essas melhorias elevam o preço e ocorrem principalmente nas vendas com escore alto; portanto, parte de seu valor foi contabilizada como condição” é uma afirmação razoável. Os dois requisitos são atendidos e ambos apontam na mesma direção: as melhorias não consideradas elevam o preço (o β_w é positivo) e estão associadas a escores altos (o δ_j é positivo); portanto, o viés é ascendente. Tal afirmação transforma um problema não mensurado em um problema reconhecido e com uma direção definida, que é exatamente o que um leitor precisa para avaliar a conclusão. Na residual Constraint, nenhuma taxa de condição é extraída, e a questão não se coloca para a condição.

Exemplo de avaliação resolvido. A ampliação do conjunto de dados Pacífica, com dezesseis imóveis, registra, para cada imóvel, a área habitável bruta (GLA), o tamanho do lote, a vista, o escore de condição, um indicador de reforma observada e o preço de venda. Como cada preço foi calculado a partir de uma fórmula conhecida, sem nenhuma componente aleatória (os dados não apresentam ruído), a regressão que utiliza as características corretas recupera essa fórmula exatamente, e qualquer desvio em relação a ela é um viés, sem qualquer componente de ruído.

Comp	GLA (sf)	Lote (sf)	Visualizar	Cond	Reformado	Preço (\$)
A	1,650	9,000	Sim	7,00	Não	1,425,000
B	1,650	7,200	Não	7,00	Não	1,280,000
C	1,850	7,500	Não	7,00	Não	1,407,500
D	1,450	7,000	Não	7,00	Não	1,155,000
E	1,700	8,000	Sim	9h00	Não	1,470,000
F	1,550	7,100	Não	5,00	Não	1,177,500
G	1,700	10,500	Não	7,00	Não	1,392,500
H	1,600	7,300	Não	7,00	Sim	1,312,500
I	1,750	8,200	Não	9h00	Sim	1,465,000
J	1,600	7,600	Não	7,00	Sim	1,320,000
K	1,900	9,500	Sim	9h00	Sim	1,687,500
L	1,500	6,800	Não	5,00	Não	1,140,000
M	1,800	8,800	Não	7,00	Não	1,410,000
N	1,650	7,400	Sim	7,00	Não	1,385,000
O	1,550	7,900	Não	9h00	Sim	1,337,500
P	1,720	9,200	Não	5,00	Não	1,332,000

Os produtos comparáveis A a H são da Edição 1, com seus atributos e preços publicados inalterados; A Edição 1 descrevia o estado de conservação por meio de palavras e o codificava de 1 a 3 para fins de regressão; no escore deste Jornal, as palavras correspondem a: médio 5,00, bom 7,00, excelente 9,00, de modo que um nível da Edição 1 equivale a dois pontos. A reforma do Comp H — divulgada na Edição 1 como a origem de seu resíduo, mas não registrada como variável — agora está registrada. O mercado paga \$20,000 por ponto de condição. Nos dados construídos, uma casa

reformada é vendida por \$60,000 a mais do que uma casa com a mesma área bruta locável (GLA), terreno, vista e escore de condição. Esses \$60,000 não representam o preço do ato de reformar; eles representam o que quer que a reforma tenha trazido que as outras características não registram. Casas reformadas também apresentam escores médias mais altos (8,20 contra 6,64), de modo que o indicador de reforma varia em conjunto com o escore de condição incluído, o que é o segundo requisito para a existência de viés: $\text{corr}(\text{remodel}, \text{cond}) = 0.55$.

Duas regressões realizadas com base nos mesmos dezesseis imóveis:

	Remodelação omitida	Reforma incluída	Viés
Interceptação	-83,923.57	-30,000	-53,923.57
GLA	586,64	600	-13.36
Lote	24,32	25	-0.68
Visualizar	79,178,46	100,000	-20,821.54
Condição (por ponto)	34,815,56	20,000	+14,815.56
Reforma	—	60,000	
R^2	0,9733	1,0000	

O “\$60,000” na linha de reforma não é um preço referente à reforma. Neste conjunto de dados construído, trata-se da diferença de preço associada ao indicador de reforma, considerando GLA, lote, vista e escore iguais; é isso que a “função geradora” recupera.

O coeficiente de condição está \$14,816 um ponto acima do normal — ele indica \$34,816 por ponto de condição, enquanto o valor de mercado é \$20,000, o que representa um exagero de 74 por cento. O coeficiente de visão está \$20,822 abaixo do normal. Ambos os erros são imperceptíveis no ajuste: o modelo sem o indicador de remodelação apresenta $R^2 = 0.9733$ e resíduos que, com exceção de três vendas próximas a \$35,000, parecem normais. Nada em R^2 ou nos testes dos coeficientes indica problemas (se esses resíduos fossem ruído, todos os coeficientes teriam sido aprovados nesse teste no nível usual de 5%).

A fórmula de Theil explica os números com exatidão. Ao realizar uma regressão do indicador “remodel” nas quatro características incluídas, obtém-se $\delta_{\text{cond}} = 0.24692602$ e $\delta_{\text{view}} = -0.34702564$. Então,

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

correspondendo às variações observadas com precisão de centésimos. O viés não é aproximadamente explicado pela fórmula; ele é a própria fórmula. O ruído mudaria menos do que se poderia esperar: com dados ruidosos, a fórmula ainda se mantém exatamente válida entre as duas regressões de ajuste (sem e com o indicador de remodelação), usando o coeficiente estimado do indicador $\hat{\beta}_w$ no lugar de β_w ; em comparação com o coeficiente verdadeiro, ela se mantém válida, em média, ao longo de muitas amostras.

Vale a pena dar uma segunda olhada no coeficiente de visualização, pois seu viés segue na direção oposta, e a razão para isso é esclarecedora. Mantendo-se as demais características constantes, os imóveis com vista nesta amostra têm *menor* probabilidade de serem reformados ($\delta_{\text{view}} < 0$). Assim, o coeficiente de vista é empurrado para baixo, em vez de para cima: ele está \$20,822 abaixo do valor correto, enquanto o coeficiente de condição está \$14,816 acima do valor correto. Os dois erros são efeitos distintos da mesma omissão, e não uma transferência entre eles. Um avaliador que se baseasse apenas na condição não teria previsto, de forma alguma, um erro no ajuste de vista — no entanto, um ajuste de vista subestimado em 21 por cento é a consequência direta da mesma omissão.

Na prática, os danos concentram-se exatamente onde os avaliadores estão menos propensos a procurá-los. Considere, primeiro, o que o modelo tendencioso faz como *preditor*. Em um imóvel objeto da avaliação idêntico ao Comp N — 1,650 pés quadrados de área bruta locável (GLA) em um terreno de 7,400 pés quadrados com vista e um escore de condição de 7,00 — ele indica \$1,386,914 contra um preço médio real de \$1,385,000 para casas com exatamente essas características — a partir da fórmula de cinco características da entrada 007 (*Expectativa Condicional*), com reforma = 0 — um erro de \$1,914, ou 0,14 por cento. Em todas as dezesseis vendas, o erro típico é de cerca de \$21,000, portanto, o modelo apresenta um viés significativo e ainda assim faz previsões aceitáveis, o que não é coincidência: o OLS realiza o ajuste à amostra que lhe é fornecida, e os coeficientes com viés são precisamente aqueles que mais se aproximam dos preços de venda da amostra. A precisão da previsão não é evidência contra o viés de variável omitida, e a validação cruzada — que testa o modelo em vendas excluídas do ajuste — também não o descarta: bons resultados de validação cruzada não indicam que os coeficientes estejam livres de viés, pois um modelo mal especificado pode prever bem sempre que as características se combinam da mesma forma que no ajuste. Ao excluir cada imóvel sucessivamente, os dois maiores erros — cerca de \$56,000 em K e \$48,000 em G — ocorrem em imóveis que seguem o padrão desta amostra, na qual os escores de condição elevadas pertencem principalmente a casas reformadas; E, que rompe esse padrão, ocupa o terceiro lugar com cerca de \$45,000, e H e J, que também o rompem, ocupam o quarto e o quinto lugares com cerca de \$41,000. Os dados aqui apresentados não contêm ruído; portanto, cada um desses erros provém do indicador de reforma omitido, e nenhum deles é resultado de ruído; em vendas reais, onde o ruído também está presente, nada no tamanho de um erro indica se ele se deve a uma característica omitida ou à dispersão de resíduo comum.

O prejuízo surge no momento em que os coeficientes são utilizados como *ajustes*, em vez de como uma regra de previsão. Um avaliador que aplique um ajuste de condição com base nesse modelo aplica \$34,816 por ponto, enquanto o valor de mercado é \$20,000 — \$29,631 a mais em cada imóvel comparável ajustado nos dois pontos, de “bom” para “excelente” — e ajustes de vista de \$79,178, enquanto o valor de mercado é \$100,000. A grade é onde os coeficientes de uma regressão são lidos um por um, e essa é exatamente a leitura que o viés destrói.

A previsão também falha, e de forma grave, nos imóveis que quebram o padrão no qual o viés se baseia. Uma casa bem conservada como o imóvel objeto da avaliação (1.650 pés quadrados de área bruta, terreno de 7.400 pés quadrados), mas sem vista, com escore excelente (9,00) e que nunca passou por reforma, está avaliada em \$1.377,367, contra um valor real de \$1,325,000, supervalorizada em \$52,367 (3,95%), pois, nessas vendas, um escore de 9,00 geralmente vinha acompanhado de uma reforma, e a taxa de condição do modelo inclui o valor da reforma \$60,000 junto com o escore. Uma casa reformada, de resto idêntica ao imóvel objeto da avaliação (com vista, escore 7,00), está avaliada em \$1,386,914, contra \$1,445,000, subvalorizada em \$58,086 (4,02%). Esses são os imóveis atípicos — aqueles para os quais um avaliador é, na maioria das vezes, contratado especificamente para avaliar, e aqueles cujos laudos serão lidos com maior atenção.

References.

- 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>
- 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.").

```