

Endogeneidade (Endogeneity)

Foundations • The Valuation Engineer

William Bert Craytor • October 1, 2026

Version 1.2.0

DOI: [10.67330/z9feps66](https://doi.org/10.67330/z9feps66)

Definição formal. A endogeneidade é a condição em que um regressor — uma característica que a regressão utiliza para explicar o preço — varia em conjunto com o erro estrutural — o preço menos o valor que a relação causal indica que o preço deveria ter, e não o resíduo do modelo ajustado — de modo que seu coeficiente estimado não reflete o efeito causal de interesse. Suas causas comuns são: uma característica omitida, o preço exercendo influência sobre a característica, as vendas constituírem um conjunto selecionado e a característica ser medida com erro.

Enquadramento intuitivo. Uma taxa pode estar errada porque as características e o preço influenciam um ao outro. O tempo que uma casa fica no mercado depende do preço pedido: as características são, em parte, resultado do preço ou de algum outro fator que também eleva o preço, como um mercado em alta, e uma taxa calculada a partir dessas vendas mistura essas duas influências. Nenhuma regressão consegue separá-las apenas a partir das vendas, seja realizando um ajuste único para toda a variação de uma característica ou, como faz o MARS (Splines de Regressão Adaptativa Multivariada, Friedman 1991), aplique taxas diferentes em diferentes intervalos dessa característica — é necessária uma fonte externa de variação ou uma suposição que as vendas não possam comprovar —, pois o emaranhado está no mercado, não na aritmética. O que o avaliador pode fazer é conhecer as quatro maneiras pelas quais uma característica pode variar em função do que o modelo deixa de fora (apresentadas abaixo), reconhecê-las em um conjunto de dados e identificar qual delas está em jogo antes que uma taxa desse conjunto de dados seja utilizada como se medisse o que um comprador paga por essa característica.

Para o leitor que deseja conhecer o mecanismo. Os estatísticos chamam de endógena uma variável quando ela varia em conjunto com algo que o modelo deixa de fora, e a entrada apresenta a condição em sua forma exata e distingue suas quatro fontes. As soluções oferecidas pela literatura estatística, principalmente as variáveis instrumentais, são apresentadas para a regressão comum e raramente estão disponíveis em dados de avaliação; o artigo menciona isso. O diagnóstico, ao contrário da solução, aplica-se a qualquer método de ajuste.

Este item tem um contexto estatístico e também envolve raciocínio causal, uma vez que aborda situações em que um coeficiente de regressão não captura o efeito causal desejado pelo avaliador. Ele se destina a avaliadores que interpretam coeficientes ou ajustes de vendas emparelhadas como afirmações de causa. Assim como a entrada 008 (*Viés de Variável Omitida*), ela utiliza o mercado Pacífica construído, que inclui a condição na regressão, como na avaliação tradicional; o método próprio da Revista, a avaliação baseada em RCA, não o faz (entrada 007, *Expectativa Condicional*), e o exemplo serve como contraponto, não como um modelo a ser seguido.

O viés de variável omitida, tema da entrada 008 (*Viés de Variável Omitida*), não deve ser visto como um tipo isolado de erro com sua própria fórmula, mas como um exemplo de uma situação mais geral: a endogeneidade.

Lembre-se da decomposição da entrada 007 (*Expectativa Condicional*): $P = E[P | \mathbf{Z}] + \varepsilon$, onde $E[P | \mathbf{Z}]$ é o preço médio das casas com as características $\mathbf{Z}$ e ε representa o quanto, acima ou abaixo, o preço de uma casa se situa em relação a essa média, de modo que, por definição, $E[\varepsilon | \mathbf{Z}] = 0$. Isso se mantém válido para a função de expectativa condicional (a CEF, conforme essa entrada),

independentemente das características que $\mathbf{Z}$ contenha. Se o avaliador quiser, em vez disso, um efeito causal — quanto mais alto seria o preço se houvesse uma vista, mantendo-se todas as outras características da casa genuinamente constantes —, é necessária mais uma condição: os regressores não devem variar em função de qualquer outro fator que afete o preço. Quando um deles varia, esse regressor é *endógeno*. Escreva a relação causal como $P = \beta_0 + \beta_1 z_1 + \dots + \beta_k z_k + u$, onde cada β_j representa o que uma variação de uma unidade na característica z_j causaria ao preço. Seu termo de erro é o *erro estrutural*, escrito como u : o preço menos o que a relação causal indica que ele deveria ser. Isso é diferente de ε , a distância da média condicional, que tem média zero, dada $\mathbf{Z}$ por construção; u pode variar com os regressores, e a condição de endogeneidade se refere a ela.

O termo tem origem na modelagem de equações simultâneas, na qual uma variável é chamada de “endógena” se for determinada dentro do sistema, em vez de ser fornecida de fora. No uso atual, ele tem um significado mais mecânico: $\text{Cov}(z_j, u) \neq 0$ (Cov, a covariância, mede como duas grandezas se comportam em conjunto) — o regressor e o erro estrutural se movem em conjunto, de modo que o ajuste (OLS — regressão por mínimos quadrados ordinários — ou MARS) não consegue separar suas contribuições e atribui parte da variação do erro ao regressor.

Existem quatro mecanismos que causam isso, e é importante diferenciá-los, pois exigem respostas diferentes.

Variáveis omitidas. Uma característica que afeta o preço é deixada de fora e varia em conjunto com uma característica incluída; ela passa a fazer parte de u , e $\text{Cov}(z_j, u) \neq 0$ se aplica. Esse é o caso da entrada 008 (*Viés de Variável Omitida*), em que o indicador de reforma foi deixado de fora. A solução é medir e incluir a característica, ou determinar os maiores e menores valores que o viés poderia plausivelmente assumir.

Simultaneidade. O regressor não está apenas associado ao preço, mas é parcialmente causado por ele ou determinado em conjunto com ele; a causalidade ocorre nos dois sentidos e, como o ajuste (OLS ou MARS) não indica nada sobre a direção, ele apresenta uma combinação dos dois. A solução é um *instrumento* — uma variável que influencia a variável explicada em questão, mas não tem nenhuma outra conexão com o preço (no sentido econométrico da palavra, não no sentido jurídico de um documento formal) — ou um modelo diferente.

Seleção. A amostra não é uma seleção aleatória das casas às quais a conclusão se refere, e o que determinou a inclusão de uma casa na amostra está relacionado ao seu preço. A solução é modelar como a amostra foi selecionada ou redefinir o grupo ao qual a conclusão se refere.

Erro de medição. A característica é registrada de forma imprecisa — dois inspetores a escorariam de maneira diferente, ou a área habitável é obtida a partir de um registro, em vez de ser medida. O erro de registro passa a fazer parte de u e acompanha o valor registrado, de modo que os coeficientes ficam enviesados. No caso clássico — o erro de medição *clássico*, em que o erro não tem nada a ver com o valor verdadeiro nem com qualquer outro elemento do modelo —, o coeficiente da característica mal medida é puxado para zero. Um escore de condição pode não ser esse o caso: seu erro pode depender da qualidade real da casa e de outros aspectos observados pelo avaliador. Para esse tipo de erro, o viés não tem uma direção geral. A solução é medir melhor ou obter uma segunda medição independente da mesma característica da mesma casa, que um método projetado para lidar com erros de medição possa então usar para corrigir a estimativa.

Os quatro mecanismos têm uma coisa em comum: aumentar o número de vendas do mesmo tipo não elimina o erro que eles causam. Essa é a distinção prática mais clara entre endogeneidade e imprecisão (a dispersão de uma estimativa decorrente do fato de se dispor apenas de um número limitado de vendas): a imprecisão diminui à medida que a amostra cresce, enquanto a endogeneidade não.

Antes de escolher as variáveis, defina o efeito. A inclusão de uma característica no modelo depende do que está sendo estimado, e há dois tipos de características que são facilmente

confundidos. O primeiro é uma causa tanto do regressor quanto do preço: locais atraentes dentro de um bairro (uma rua sem saída tranquila, uma esquina com vista) atraem reformas caras e, por si só, determinam preços elevados. Os estatísticos chamam essa característica de *fator de confusão*. Ela deve ser mantida fixa; caso contrário, o coeficiente do regressor refletirá seu efeito. O segundo é o canal pelo qual o regressor atua sobre o preço: uma reforma eleva o preço *ao* melhorar as condições e a qualidade. Tal característica é um *mediador*, e mantê-la fixa remove dos coeficientes da variável regressora exatamente a parte de seu efeito que passa pelo canal. Portanto, a reforma da entrada 008 (*Viés de Variável Omitida*) levanta duas questões diferentes. Na avaliação tradicional, se a questão for a taxa de mercado para o estado de conservação, este é o regressor (a avaliação baseada em RCA nunca faz essa pergunta; a entrada 007, *Expectativa Condicional*, explica o motivo) e a reforma é meramente a forma como algumas casas chegaram a esse estado; ela não agrega nenhum valor além da condição e da qualidade que produziu, uma vez que estas estejam totalmente registradas, e uma casa reformada e uma casa bem conservada, semelhantes em condição e qualidade, valem o mesmo. Se a questão for o que uma reforma causa no preço de uma casa específica, então a condição e a qualidade são mediadores, e um modelo que as mantém fixas responde a uma questão diferente — quanto vale a reforma *além da* condição e da qualidade que ela produziu, o que pode ser pequeno e, de qualquer forma, é uma grandeza diferente do efeito total da reforma. “Incluir a variável” é a regra correta para um fator de confusão e a errada para um mediador; nenhuma das duas pode ser escolhida até que o efeito desejado seja identificado.

Onde o avaliador a encontra. A causalidade reversa é uma questão de localização. A endogeneidade decorrente da simultaneidade é bem ilustrada pela qualidade das escolas, que é um problema real, e não hipotético. Se você realizar uma regressão do preço em relação à classificação da escola, os coeficientes parecem indicar quanto os compradores pagam pela escola. Mas a causalidade também funciona no sentido inverso: bairros mais caros têm uma base tributária maior e pais mais engajados e com melhores recursos, podendo atrair e reter professores com mais facilidade. A qualidade da escola pode elevar os preços, e preços mais altos podem melhorar a qualidade da escola. Os coeficientes não capturam nenhum desses efeitos; eles refletem uma combinação que depende da força relativa dos dois canais nesse mercado. A mesma relação bidirecional pode ocorrer com as taxas de criminalidade, a densidade de estabelecimentos comerciais e a cobertura arbórea. Em cada caso, o fator pode ser tanto uma causa quanto um efeito dos níveis de preço do bairro.

Uma reforma é um tratamento, mas não é atribuída aleatoriamente. No vocabulário da inferência causal, um *tratamento* é qualquer intervenção cujo efeito se busca avaliar, e uma reforma na cozinha é um exemplo disso. A dificuldade está em determinar quem a recebe. Muitos proprietários reformam quando esperam que o retorno supere o custo, e essa expectativa se baseia no que eles sabem sobre sua casa — algo que nenhum conjunto de dados registra —: uma boa planta baixa, uma localização privilegiada no bairro, um terreno que justifica o investimento. Essas mesmas qualidades elevam o preço diretamente: são fatores de confusão da reforma, e que o conjunto de dados não mede. Portanto, um modelo que inclui o indicador de reforma corrigiu a omissão do indicador, e não a confusão. Os coeficientes do indicador combinam o efeito da reforma sobre o preço com o preço adicional — além do que as características registradas explicam — das casas cujos proprietários optaram por reformar. A inclusão da variável eliminou a omissão; ela não tornou o regressor exógeno — ou seja, desvinculado do erro estrutural.

A amostra de imóveis vendidos não é aleatória. As pessoas vendem quando as circunstâncias são favoráveis ou os preços são atraentes, e outras vendem porque precisam; assim, os imóveis que são vendidos podem diferir daqueles que não são de maneiras que o modelo não percebe; quando isso ocorre, o elemento idiossincrático do preço de um imóvel vendido não é aleatório: o primeiro grupo o influencia em uma direção, enquanto as vendas em situação de dificuldade e as vendas

forçadas o influenciam na outra. Dentro das vendas, as transações em condições de mercado e as vendas em situação de dificuldade financeira constituem, cada uma, subgrupos adicionais não aleatórios. Todos os imóveis comparáveis em cada tabela são extraídos dessas vendas selecionadas. Qualquer conclusão de avaliação só pode ser estendida a imóveis não vendidos com base em suposições sobre como os vendidos foram selecionados, o que deve ser declarado no relatório.

A pontuação de estado de conservação pode ser influenciada pelo preço. A pontuação tem como objetivo medir o estado de conservação da casa e nada mais, e o avaliador atribui a pontuação ao imóvel objeto da avaliação durante a inspeção, antes de se chegar a qualquer conclusão sobre o valor. Os imóveis comparáveis são outra questão: eles podem ser escorados com base em fotos e comentários dos anúncios, após o conhecimento de seus preços de venda. Se o preço conhecido influenciar o escore — uma casa vendida por um valor alto sendo interpretada como um pouco mais bem conservada —, então o modelo de regressão passa a ser, em parte, uma função do resultado. Essa não é uma característica do estado de conservação em si, mas um defeito do processo de medição: *erro de medição dependente do resultado*, com o preço influenciando o escore. Trata-se do tipo não clássico descrito acima, com uma diferença. Para um desvio desse tipo, que transporta parte do desvio do preço para o escore, a direção é conhecida: o escore se move em direção ao escore que o preço de venda sugeriria, e o coeficiente de condição é empurrado para cima (outros erros no escore ainda podem puxar na direção oposta). A proteção é de natureza processual: escore os imóveis comparáveis com base nas evidências de seu estado antes de verificar por quanto foram vendidos e registre a justificativa de cada pontuação no arquivo de trabalho.

Por que importa para a solidez. A questão de saber se uma variável regressora é endógena é decidida por meio de argumentação, e não por qualquer resultado que a regressão comum possa mostrar. Nenhuma das verificações habituais de ajuste — um gráfico dos resíduos, a proporção da variação de preço explicada (R^2), o teste do modelo com vendas excluídas do ajuste (validação cruzada) ou um teste para verificar se um coeficiente poderia ser zero (um teste de significância) — pode estabelecer que uma variável explicativa é exógena, uma vez que o erro estrutural é inobservável, assim como sua relação com as variáveis explicativas. Existem testes formais — o teste de Hausman compara a estimativa comum com outra obtida de maneira diferente —, mas eles exigem um instrumento ou um segundo estimador, ou seja, suposições adicionais que os próprios dados não podem verificar. O que sempre pode ser fornecido é uma explicação fundamentada do mecanismo: por que esse regressor pode estar relacionado ao erro, em que direção e em que medida. Um relatório razoável apresenta esse argumento, em vez de sugerir que o problema não existe ao deixar de mencioná-lo.

Diga qual é o coeficiente, não qual se gostaria que fosse. O coeficiente de um regressor endógeno é um valor real e bem definido: a diferença no preço ajustado por unidade da característica, mantendo-se fixas as características medidas. Não se trata do efeito causal. Nas dezesseis vendas de Pacífica em que o indicador de reforma foi omitido (entrada 008, *Viés de Variável Omitida*), é correto afirmar que “essas casas com vista foram vendidas por \$79,178 a mais, mantendo as características medidas fixas”. Afirmar que “a vista contribui com \$79,178 para o valor” sugere uma causa que a estimativa não sustenta. A primeira afirmação resistirá a um questionamento; a segunda, não.

O projeto supera a estimativa. É possível romper a influência mútua entre uma característica e o preço por meio da forma como os dados são organizados — comparando imóveis antes e depois do bloqueio de uma vista, utilizando um limite onde a designação escolar muda, mas o parque habitacional permanece o mesmo, ou considerando uma reforma imposta por um evento externo, como um incêndio ou uma enchente, em vez de uma escolhida pelo proprietário — cada modelo ainda se baseia em suposições que devem ser defendidas (de que nada mais mudou ao mesmo

tempo, de que o limite separa apenas a designação escolar), mas um modelo cujas suposições são defendidas alcança maior defensabilidade do que qualquer estimador que trabalhe com dados observacionais. É isso que a literatura sobre variáveis instrumentais tem a dizer em termos práticos, e sua relevância para a avaliação imobiliária é que um conjunto bem escolhido de imóveis comparáveis é, em si mesmo, um projeto.

Exemplo de avaliação resolvido. O conjunto de dados Pacifica pode ilustrar a endogeneidade, mas não pode resolvê-la, e a verdadeira lição do artigo é justamente explicar por que o conjunto de dados não consegue fazer isso.

Primeiro, a pergunta. Suponha que perguntem a um avaliador se uma proposta de reforma da cozinha de uma determinada casa, no valor de \$45,000, seria recuperada na venda. Essa é uma pergunta sobre uma casa e dois preços que ela não pode ter ao mesmo tempo: seu preço após a reforma e seu preço no estado atual, $P_i(1) - P_i(0)$, onde $P_i(1)$ é o preço da casa i caso seja reformada e $P_i(0)$ é o preço caso não seja. Não se trata de uma questão sobre a cotação de mercado para o estado atual, e manter esse estado fixo ao responder seria o erro do mediador descrito acima.

As dezesseis características do registro 008 (*Viés de Variável Omitida*) apresentam um indicador de reforma cujo coeficiente é \$60,000: a diferença de preço que os dados construídos atribuem a uma reforma — \$60,000 a mais do que uma casa com a mesma área bruta locável (GLA), terreno, vista e escore de condição. Esse valor de \$60,000 não representa o preço do ato de reforma em si; ele representa tudo o que a reforma trouxe de mais que as outras características não registram. Nesse conjunto de dados construído, a decisão sobre quais casas foram reformadas foi tomada pelo autor. O indicador de reforma é exógeno por definição, não está relacionado a nada que reste no erro, pois nada resta no erro; portanto, \$60,000 representa, nesse conjunto de dados, o que uma reforma agrega ao preço, mantendo-se fixa a pontuação de condição registrada; uma reforma que também aumentasse o escore agregaria ainda mais.

Agora, suponhamos que esses mesmos dezesseis imóveis tenham vindo de um mercado real. Os proprietários optam por reformar e, suponhamos — o que é plausível — que os proprietários de casas com qualidades não observadas significativas (chamemos esse componente do preço u) sejam mais propensos a reformar do que os proprietários de casas, de resto, semelhantes. A regressão que inclui o indicador de reforma agora converge, à medida que as vendas se acumulam, não para \$60,000, mas a

$$\hat{\beta}_{\text{remodel}} \xrightarrow{p} \underbrace{60,000}_{\text{what the remodel brought that the other features do not record}} + \underbrace{\frac{\text{Cov}(\tilde{r}, u)}{\text{Var}(\tilde{r})}}_{\text{extra price of the houses selected into remodeling}},$$

onde $\tilde{r}$ é a parte do indicador de reforma que as características incluídas não predizem, Var (a variância) mede o quanto ela varia, e a seta marcada com p significa “aproxima-se cada vez mais, à medida que as vendas se acumulam”. Se as casas reformadas apresentam, em média, \$25,000 a mais de u do que as não reformadas observavelmente semelhantes, o coeficiente converge para aproximadamente \$85,000, e nenhuma quantidade adicional de dados do mesmo mercado reduz esse excesso. O indicador de reforma, cuja entrada de omissão 008 (*viés de variável omitida*) foi estudada, agora é medido e incluído, e o coeficiente ainda não representa o efeito causal: o que ainda está sendo deixado de fora é u , que nenhum conjunto de dados registra.

Essa distinção tem peso na prática. Os coeficientes descrevem, em média, as casas que *foram* reformadas, ou seja, aquelas cujos proprietários consideraram que o retorno valia a pena. O imóvel objeto da avaliação é uma casa específica cujo proprietário ainda não se decidiu, e os coeficientes se aplicam a ela apenas se o imóvel se assemelhar àquelas casas no que diz respeito ao que seus proprietários sabiam — o que é exatamente o que o modelo não consegue perceber. Um avaliador

que responda à questão do retorno com base no valor de \$85,000 superestima o retorno; aquele que a responde com base em um modelo que também mediu e manteve fixas todas as condições e qualidades resultantes da reforma subestima o retorno, reduzindo-o ao que a reforma acrescenta além dessas condições, o que, segundo o princípio da entrada 008 (*Viés de Variável Omitida*), é pouco ou nada. Ambos interpretaram a diferença de preço condicional associada à reforma como o efeito da própria reforma.

A pergunta diagnóstica a ser feita sobre qualquer coeficiente que se pretenda interpretar causalmente não é, portanto, “o modelo está bem especificado?”, mas “qual é o efeito desejado e como essa variável chegou a assumir os valores que assumiu?” Se a resposta envolver uma escolha feita por alguém capaz de enxergar coisas que o modelo não consegue, os coeficientes descrevem essa escolha tanto quanto a característica em si.

References.

- Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1), 1–67. <https://doi.org/10.1214/aos/1176347963>
- Haavelmo, T. (1943). The Statistical Implications of a System of Simultaneous Equations. *Econometrica*, 11(1), 1–12.
- Hausman, J. A. (1978). Specification Tests in Econometrics. *Econometrica*, 46(6), 1251–1271.
- Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press.
- Wright, P. G. (1928). *The Tariff on Animal and Vegetable Oils*. Macmillan.

Cross-references: omitted variable bias; conditional expectation; parameter identification.

Addendum: Prolog Formalization

The following machine-readable companion (009-endogeneity.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>.

%% 009-endogeneity.pl
%%
%% Prolog companion to Foundations Vol. 1, Issue 2, Entry 009
%% Title: Endogeneity
%% Author: Bert Craytor
%% Concept: endogeneity
%% Version: 1.2.0
%%
%% v1.1.0: the estimand is named before the variables are chosen
%% (confounder vs. mediator); a remodel is a treatment that is not randomly
%% assigned; its 60,000 is the price difference the constructed data
%% attach to a remodel; measurement error attenuates only in the classical case; a rating
%% contaminated by the known price is outcome-dependent measurement error.
```

```

% 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([ endogeneity, omitted_variable_bias,
             conditional_expectation, hedonic_regression,
             latent_variable, residual, comp_selection,
             paired_sales_analysis, adjustment,
             sales_comparison_approach ]).
uses_relations([depends_on/2, related/2, introduced_by/2, formalized_by/2])).

article_concept(endogeneity).
article_issue(volume(1), number(2), year(2026)).
article_id('foundations.2026.009').

local_term(instrument).
local_term(simultaneity).
local_term(selection).
local_term(causal_effect).
local_term(conditional_association).
local_term(treatment).
local_term(confounder).
local_term(mediator).
local_term(estimand).

defining_condition(endogeneity,
    "The regressor is correlated with the model's structural error term.").

mechanism(omitted_variables,
    "A price-affecting variable is excluded and correlates with an included regressor; it now sits
    ↪ inside the error.").
mechanism(simultaneity,
    "The regressor is partly caused by, or jointly determined with, the outcome. Least squares
    ↪ returns a blend of both directions.").
mechanism(selection,
    "Sample membership is related to the outcome, so the estimated relationship is conditional on
    ↪ being in the sample.").
mechanism(measurement_error,
    "The regressor is recorded with error; the recording error enters the structural error and
    ↪ biases the coefficient. Under classical error (independent of the true value and of everything
    ↪ else) the coefficient is attenuated toward zero; under non-classical error no general
    ↪ direction follows.").

measurement_error_kind(classical,          attenuates_toward_zero).
measurement_error_kind(non_classical,      no_general_direction).
measurement_error_kind(outcome_dependent,  toward_the_outcome).
measurement_error_of(condition_rating, non_classical).
measurement_error_of(condition_rating_after_price, outcome_dependent).

remedy(omitted_variables, "Measure and include the variable, or bound the bias.").
remedy(simultaneity,      "Find an instrument, or a design that breaks the loop.").
remedy(selection,         "Model the selection, or redefine the population the conclusion is about
    ↪ .").
remedy(measurement_error, "Measure better, or obtain a second independent measurement of the same
    ↪ characteristic.").

generalizes(endogeneity, omitted_variable_bias).
special_case_of(omitted_variable_bias, endogeneity).

```

```

shrinks_with_sample_size(imprecision).

distinguishable_from_imprecision(endogeneity) :-
    \+ shrinks_with_sample_size(endogeneity).

role(confounder, "A cause of both the regressor and the price.").
role(mediator, "The channel through which the regressor acts on the price.").
include_rule(confounder, include, "Its effect otherwise rides on the regressor's coefficient.").
include_rule(mediator, exclude, "Holding it fixed leaves only the direct effect, a different
    ↪ quantity from the total effect.").

treatment(remodel).
confounder_of(unobserved_quality_u, remodel, price).    % floor plan, micro-location, suitability
mediator_of(condition, remodel, price).
mediator_of(quality, remodel, price).

estimand(condition_rate,
    "Traditional appraisal: the market's rate for condition, with condition as the regressor; the
    ↪ remodel is how some houses came to their condition and earns no value of its own. RCA-based
    ↪ appraisal never asks this.").
estimand(remodel_total_effect,
    "What a remodel does to a particular house's price: condition and quality are mediators and
    ↪ must not be held fixed.").
estimand(remodel_direct_effect,
    "The remodel's effect apart from the condition and quality it produced: what a model holding
    ↪ those fixed estimates; may be small; not the total effect.").

variable_choice(remodel_total_effect, V, exclude) :- mediator_of(V, remodel, price).
variable_choice(remodel_total_effect, V, include) :- confounder_of(V, remodel, price).
variable_choice(condition_rate, condition, regressor).

encounter(school_quality, simultaneity,
    "School quality can raise prices and prices can raise school quality via tax base and parental
    ↪ resources; the coefficient then recovers neither effect.").
encounter(renovation_choice, omitted_variables,
    "A remodel is a treatment that is not randomly assigned: owners remodel when they expect the
    ↪ return to justify it, on qualities u that also raise price. u is a confounder the dataset does
    ↪ not measure, so including the remodel indicator cures the omission of the indicator, not the
    ↪ confounding.").
encounter(comparable_set, selection,
    "Sold properties are selected: sellers who choose their moment push the price one way,
    ↪ distressed and forced sellers the other; this matters when the selection is related to what
    ↪ the model does not see.").
encounter(condition_rating, measurement_error,
    "Comparables may be rated from photographs after their sale prices are known; if the known
    ↪ price nudges the rating, the regressor is partly a function of the outcome. For a nudge that
    ↪ carries part of the price deviation into the score, the condition coefficient is pushed upward
    ↪ .").
protection(condition_rating,
    "Rate the comparables from the evidence of their condition before looking at the price; record
    ↪ the basis of each rating.").

bidirectional_with_price(school_quality).
bidirectional_with_price(crime_rate).
bidirectional_with_price(retail_amenity_density).
bidirectional_with_price(tree_canopy).

cures(measuring_the_variable, omitted_variables).

```

```

omitted_component(renovation_choice, unobserved_quality_u).
measurable(remodel_numeric).

remains_endogenous_after_inclusion(remodel_numeric) :-
    encounter(renovation_choice, omitted_variables, _),
    omitted_component(renovation_choice, U),
    \+ measurable(U).

coefficient_component(remodel_numeric, what_the_work_produced,
    "60,000: the price difference the constructed data attach to a remodel, holding GLA, lot, view
    ↪ and the condition score fixed; it stands for whatever the remodel brought that the other
    ↪ features do not record.").
coefficient_component(remodel_numeric, value_of_selected_properties,
    "The unobserved quality of properties whose owners chose to remodel; positive under positive
    ↪ selection, negative under negative selection.").

coefficient_limit(SelectionDifference, Limit) :-
    Limit is 60000 + SelectionDifference.

consequence(renovation_payback_question,
    "A coefficient from observational data is the conditional association among the properties that
    ↪ were remodeled; it need not equal the causal effect of remodeling the particular subject, P_i
    ↪ (1) - P_i(0).").
consequence(two_ways_wrong,
    "Reading the selected coefficient overstates the return under positive selection; reading a
    ↪ model that holds condition and quality fixed understates it, to the direct effect alone.").

defensibility_consequence(not_from_the_regression_itself,
    "Residual plots, R-squared, cross-validation and significance tests cannot establish exogeneity
    ↪; formal tests (Hausman) need an instrument or a second estimator, i.e. assumptions the data
    ↪ cannot check.").
defensibility_consequence(state_what_it_is,
    "A coefficient on an endogenous regressor is a well-defined conditional association; reporting
    ↪ it as a causal contribution asserts more than the estimate supports.").
defensibility_consequence(design_beats_estimation,
    "Breaking the loop in how data are assembled does more than any estimator applied afterward; a
    ↪ well-chosen comparable set is itself a design.").

design(before_after_view_obstruction).
design(boundary_discontinuity_school_assignment).
design(externally_forced_renovation).

diagnostic_question(
    "What effect is wanted, and how did this variable come to take the values it took? If a chooser
    ↪ could observe what the model cannot, the coefficient describes the choice as much as the
    ↪ characteristic.").

% Scryer Prolog loads forall/2 and the list predicates on request;
% SWI-Prolog and Trealla have them built in. The guarded loads below
% let the file run unchanged in all three.
:- initialization(catch(use_module(library(lists)), _, true)).
:- initialization(catch(use_module(library(iso_ext)), _, true)).

```