

Endogeneidad (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)

Definición formal. La endogeneidad es la situación en la que una variable explicativa —una característica que la regresión utiliza para explicar el precio— varía junto con el error estructural —el precio menos lo que, según la relación causal, debería ser el precio, y no el residuo del modelo ajustado—, de modo que su coeficiente estimado no refleja el efecto causal de interés. Sus causas más comunes son que se haya omitido una característica, que el precio influya a su vez en la característica, que las ventas constituyan un conjunto seleccionado y que la característica se mida con error.

Planteamiento intuitivo. Una tasa puede ser errónea porque las características y el precio se influyen mutuamente. Los días que una vivienda permanece en el mercado dependen del precio de venta: las características son, en parte, consecuencia del precio o de algún otro factor que también lo eleva, como un mercado al alza, y una tasa calculada a partir de esas ventas mezcla ambas variables. Ningún modelo de regresión puede separarlas a partir únicamente de las ventas, ya sea mediante el ajuste de una única tasa a todo el rango de una característica o, como hace el método MARS (Splines de regresión adaptativa multivariante, Friedman 1991), aplique tasas diferentes en distintos tramos de la misma: se necesita una fuente externa de variación, o una hipótesis que las ventas no puedan verificar, porque la confusión está en el mercado, no en la aritmética. Lo que el tasador puede hacer es conocer las cuatro formas en que una característica puede variar en función de lo que el modelo omite (que se exponen a continuación), reconocerlas en un conjunto de datos y determinar cuál de ellas está en juego antes de utilizar una tasa de ese conjunto de datos como si midiera lo que un comprador paga por dicha característica.

Para el lector interesado en el mecanismo. Los estadísticos denominan «endógena» a una variable cuando varía conjuntamente con algo que el modelo omite, y la entrada expone la condición en su forma exacta y distingue sus cuatro fuentes. Las soluciones que ofrece la literatura estadística, principalmente las variables instrumentales, se describen para la regresión ordinaria y rara vez están disponibles en los datos de evaluación; así se indica en la entrada. El diagnóstico, a diferencia de la solución, se aplica a cualquier método de ajuste.

Este artículo tiene un contexto estadístico y también implica un razonamiento causal, ya que aborda situaciones en las que un coeficiente de regresión no refleja el efecto causal que busca un tasador. Está dirigido a tasadores que interpretan los coeficientes o los ajustes por ventas comparables como indicaciones de causalidad. Al igual que la entrada 008 (*Sesgo de variable omitida*), utiliza el mercado Pacífica construido, que incluye el estado del inmueble en la regresión, tal y como hace la tasación tradicional; el método propio de la revista, la tasación basada en el RCA, no lo hace (entrada 007, *Esperanza condicional*), y el ejemplo sirve de contrapunto, no como modelo a seguir.

El sesgo de variable omitida, tema de la entrada 008 (*Sesgo de variable omitida*), no debe considerarse como un tipo de error independiente con su propia fórmula, sino como un ejemplo de una situación más general: la endogeneidad.

Recordemos la descomposición de la entrada 007 (*Esperanza condicional*): $P = E[P \mid \mathbf{Z}] + \varepsilon$, donde $E[P \mid \mathbf{Z}]$ es el precio medio de las viviendas con las características $\mathbf{Z}$ y ε es la distancia, por

encima o por debajo, a la que se sitúa el precio de una vivienda respecto a esa media, de modo que, por definición, $E[\varepsilon | \mathbf{Z}] = 0$. Esto se cumple para la función de esperanza condicional (la CEF, según esa entrada) independientemente de las características que contenga $\mathbf{Z}$. Si, por el contrario, el tasador desea obtener un efecto causal —es decir, cuánto más costaría la vivienda si tuviera vistas, manteniendo realmente constantes todas las demás características de la misma—, se necesita una condición más: las variables explicativas no deben variar junto con ningún otro factor que afecte al precio. Cuando una de ellas lo hace, dicha variable explicativa es *endógena*. Escribamos la relación causal como $P = \beta_0 + \beta_1 z_1 + \dots + \beta_k z_k + u$, donde cada β_j representa el efecto que tendría sobre el precio un cambio de una unidad en la característica z_j . Su término de error es el *error estructural*, que se escribe como u : el precio menos lo que la relación causal indica que debería ser. Es algo distinto de ε , la distancia respecto a la media condicional, que tiene una media nula dada $\mathbf{Z}$ por su propia construcción; u puede variar con las variables explicativas, y la condición de endogeneidad se refiere a ello.

El término tiene su origen en la modelización de ecuaciones simultáneas, donde una variable se denomina «endógena» si se determina dentro del sistema en lugar de proporcionarse desde el exterior. En su uso actual, tiene un significado más mecánico: $\text{Cov}(z_j, u) \neq 0$ (Cov, la covarianza, mide cómo se mueven conjuntamente dos magnitudes) — la variable explicativa y el error estructural varían conjuntamente, por lo que el ajuste (OLS —regresión por mínimos cuadrados ordinarios— o MARS) no puede separar sus contribuciones y atribuye parte de la variación del error a la variable explicativa.

Hay cuatro mecanismos que provocan esto, y es importante diferenciarlos, ya que requieren respuestas distintas.

Variables omitidas. Se omite una característica que afecta al precio y que varía junto con una que sí se ha incluido; pasa a formar parte de u , y se produce $\text{Cov}(z_j, u) \neq 0$. Este es el caso de la entrada 008 (*sesgo de variable omitida*), en la que se omitió el indicador de remodelación. La solución consiste en medir e incluir la característica, o bien determinar los valores máximo y mínimo que el sesgo podría adoptar de forma plausible.

Simultaneidad. La variable explicativa no solo está asociada al precio, sino que en parte es causada por él, o se determina conjuntamente con él; la causalidad es bidireccional y, dado que el ajuste (OLS o MARS) no indica nada sobre la dirección, arroja una combinación de ambas. La solución es un *instrumento* —una variable que influye en la variable dependiente en cuestión pero que no tiene ninguna otra conexión con el precio (en el sentido econométrico del término, no en el sentido jurídico de un documento formal)— o un diseño diferente.

Selección. La muestra no es una selección aleatoria de las viviendas a las que se refiere la conclusión, y el hecho de que una vivienda forme parte de la muestra está relacionado con su precio. La solución consiste en modelar cómo se seleccionó la muestra o en redefinir el grupo al que se refiere la conclusión.

Error de medición. La característica se registra de forma inexacta: dos inspectores le asignarían diferentes puntuaciones, o la superficie habitable se toma de un registro en lugar de medirse. El error de registro pasa a formar parte de u y varía junto con el valor registrado, por lo que el coeficiente presenta un sesgo. En el caso típico —el error de medición *clásico*, en el que el error no tiene nada que ver con el valor real ni con ningún otro elemento del modelo—, el coeficiente de la característica mal medida tiende a cero. Una puntuación de estado puede no ser ese caso: su error puede depender de la calidad real de la vivienda y de otros aspectos que observe el evaluador. Para este tipo de error, el sesgo no tiene una dirección general. La solución consiste en medir mejor u obtener una segunda medición independiente de la misma característica de la misma vivienda, que un método diseñado para el error de medición pueda utilizar para corregir la estimación.

Los cuatro mecanismos tienen algo en común: aumentar el número de ventas del mismo tipo

no elimina el error que provocan. Esta es la distinción práctica más clara entre endogeneidad e imprecisión (la dispersión de una estimación que se debe a que solo se dispone de un número limitado de ventas): la imprecisión se reduce a medida que crece la muestra, mientras que la endogeneidad no.

Antes de elegir las variables, define el efecto. Que una característica forme parte del modelo depende de lo que se esté estimando, y hay dos tipos de características que se confunden fácilmente. La primera es una causa tanto de la variable explicativa como del precio: las ubicaciones más codiciadas de un vecindario (una calle sin salida tranquila, una esquina con vistas) atraen reformas costosas y alcanzan precios elevados por sí mismas. Los estadísticos denominan a esta característica *factor de confusión*. Debe mantenerse fija, o de lo contrario los coeficientes de la variable explicativa reflejarán su efecto. La segunda es el canal a través del cual la variable explicativa actúa sobre el precio: una reforma aumenta el precio *al* mejorar el estado y la calidad. Dicha característica es un *mediador*, y mantenerla fija elimina del coeficiente de la variable predictora exactamente la parte de su efecto que discurre a través de ese canal. Así pues, la remodelación de la entrada 008 (*sesgo de variable omitida*) plantea dos cuestiones diferentes. En la tasación tradicional, si la cuestión es la valoración de mercado del estado, este es la variable explicativa (la tasación basada en el RCA nunca plantea esa cuestión; la entrada 007, *Expectativa condicional*, explica por qué) y la reforma es simplemente la forma en que algunas viviendas llegaron a ese estado; no aporta ningún valor más allá del estado y la calidad que ha generado, una vez que estos se han registrado por completo, y una vivienda reformada y una bien conservada, si son iguales en estado y calidad, tienen el mismo valor. Si la pregunta es qué efecto tiene una reforma en el precio de una vivienda concreta, entonces el estado y la calidad son mediadores, y un modelo que los mantiene fijos responde a una pregunta diferente: cuánto vale la reforma *al margen del* estado y la calidad que ha generado, lo cual puede ser poco y, en cualquier caso, es una magnitud distinta del efecto total de la reforma. «Incluir la variable» es la regla correcta para un factor de confusión y la incorrecta para un mediador; no se puede elegir ninguna de las dos hasta que se haya identificado el efecto deseado.

Dónde la encuentra el valorador. La causalidad inversa es una cuestión de ubicación. La endogeneidad debida a la simultaneidad queda bien ilustrada por la calidad de los colegios, que es un problema real, no hipotético. Si se realiza una regresión del precio en función de la calificación de un colegio, el coeficiente parece indicar cuánto pagan los compradores por el colegio. Pero la causalidad también funciona a la inversa: los vecindarios más caros cuentan con una base impositiva mayor y con padres más implicados y con más recursos, y pueden atraer y retener mejor a los profesores. La calidad de los colegios puede elevar los precios, y unos precios más altos pueden mejorar la calidad de los colegios. Los coeficientes no captan ninguno de estos efectos; reflejan una combinación que depende de la fuerza relativa de ambos canales en ese mercado. La misma relación bidireccional puede darse con las tasas de delincuencia, la densidad de comercios minoristas y la cobertura arbórea. En cada caso, el factor puede ser tanto una causa como un efecto de los niveles de precios del vecindario.

Una reforma es un tratamiento, pero no se asigna al azar. En el vocabulario de la inferencia causal, un *tratamiento* es cualquier intervención cuyo efecto se analiza, y una reforma de cocina es uno de ellos. La dificultad radica en quién la recibe. Muchos propietarios renuevan cuando esperan que el rendimiento supere el coste, y esa expectativa se basa en lo que saben de su vivienda y que ningún conjunto de datos recoge: una buena distribución, una ubicación privilegiada dentro del vecindario, una parcela que respalda la inversión. Esas mismas cualidades elevan el precio directamente: son factores de confusión de la reforma, y el conjunto de datos no los mide. Por lo tanto, un modelo que incluye el indicador de reforma ha subsanado la omisión del indicador, pero

no la confusión. Los coeficientes del indicador combinan el efecto de la reforma en el precio con el sobreprecio —más allá de lo que explican las características registradas— de las viviendas cuyos propietarios decidieron reformar. La inclusión de la variable subsanó la omisión; no convirtió a la variable explicativa en exógena —es decir, ajena al error estructural—.

La muestra de inmuebles vendidos no es aleatoria. La gente vende cuando las circunstancias son favorables o los precios resultan atractivos, y otros venden porque se ven obligados a ello, por lo que los inmuebles que se venden pueden diferir de los que no se venden en aspectos que el modelo no tiene en cuenta; cuando esto ocurre, el elemento idiosincrásico del precio de un inmueble vendido no es aleatorio: el primer grupo lo empuja en una dirección, mientras que las ventas por dificultades económicas y las ventas forzadas lo empujan en la otra. Dentro de las ventas, las transacciones en condiciones de mercado y las ventas en situación de dificultad constituyen, a su vez, subgrupos no aleatorios. Cada dato comparable de cada cuadrícula se extrae de estas ventas seleccionadas. Cualquier conclusión de valoración solo puede extrapolarse a los inmuebles no vendidos partiendo de supuestos sobre cómo se seleccionaron los vendidos, lo cual debe indicarse en el informe.

La puntuación del estado de una vivienda puede verse sesgada por el precio. La puntuación tiene como objetivo medir el estado de la vivienda y nada más, y el tasador la asigna tras la inspección, antes de determinar su valor. Los inmuebles comparables son otro asunto: pueden puntuarse a partir de las fotografías y los comentarios de los anuncios, una vez que se conocen sus precios de venta. Si el precio conocido influye en la puntuación —una vivienda que se ha vendido a un precio elevado se interpreta como que está un poco mejor conservada—, entonces el modelo de regresión se ha convertido, en parte, en una función del resultado. Esto no es una característica del estado en sí mismo, sino un defecto del proceso de medición: *un error de medición dependiente del resultado*, en el que el precio influye en la puntuación. Se trata del tipo no clásico que acabamos de describir, con una diferencia. En el caso de un sesgo de este tipo, que traslada parte de la desviación del precio a la puntuación, la dirección es conocida: la puntuación se desplaza hacia el valor que sugeriría el precio de venta, y el coeficiente de estado se ve impulsado al alza (otros errores en la puntuación pueden seguir ejerciendo una influencia en sentido contrario). La protección es de carácter procedimental: realizar las puntuaciones de los inmuebles comparables a partir de las pruebas de su estado antes de fijarse en el precio por el que se vendieron, y registrar el fundamento de cada puntuación en el expediente de trabajo.

Por qué importa para la solidez. El carácter endógeno de una variable explicativa se determina mediante razonamiento, y no a partir de lo que pueda mostrar una regresión ordinaria. Ninguna de las comprobaciones habituales del ajuste —un gráfico de los residuos, la proporción de variación de precios explicada (R^2), la comprobación del modelo con ventas excluidas del ajuste (validación cruzada) o una prueba de si un coeficiente podría ser cero (una prueba de significación)— puede establecer que una variable explicativa sea exógena, ya que el error estructural es inobservable, al igual que su relación con las variables explicativas. Existen pruebas formales —la prueba de Hausman compara la estimación ordinaria con otra obtenida de otra manera—, pero requieren un instrumento o un segundo estimador, es decir, supuestos adicionales que los propios datos no pueden verificar. Lo que siempre se puede ofrecer es una explicación razonada del mecanismo: por qué esta variable explicativa podría estar relacionada con el error, en qué dirección y en qué medida. Un informe razonable expone este argumento en lugar de sugerir que el problema no existe al omitir mencionarlo.

Hay que indicar cuál es el coeficiente, no cuál se desearía que fuera. El coeficiente de una variable explicativa endógena es una cifra real y bien definida: la diferencia en el precio ajustado por unidad de la característica, manteniendo fijas las características medidas. No se trata del efecto

causal. En las dieciséis ventas de Pacifica en las que se omitió el indicador de reforma (entrada 008, *Sesgo de variable omitida*), es correcto afirmar que «estas viviendas con vistas se vendieron por \$79,178 más, manteniendo fijas las características medidas». Decir «las vistas contribuyen \$79,178 al valor» afirma una causa que la estimación no respalda. La primera afirmación resistirá un contrainterrogatorio; la segunda, no.

El diseño prevalece sobre la estimación. En los casos en que la influencia recíproca entre una característica y el precio puede romperse en función de cómo se organicen los datos:— comparando los inmuebles antes y después de que se haya bloqueado una vista, utilizando un límite en el que cambie la asignación escolar pero no el parque inmobiliario, o tomando como referencia una reforma impuesta por un acontecimiento externo, como un incendio o una inundación, en lugar de una elegida por el propietario— cada diseño sigue basándose en supuestos que deben defenderse (que nada más cambió al mismo tiempo, que el límite separa únicamente la asignación escolar), pero un diseño cuyos supuestos se defienden alcanza una mayor defensibilidad que cualquier estimador que trabaje con datos observacionales. Esto es lo que la bibliografía sobre variables instrumentales tiene que decir en términos prácticos, y su relevancia para la tasación radica en que un conjunto de inmuebles comparables bien elegido constituye en sí mismo un diseño.

Ejemplo de valoración resuelto. El conjunto de datos de Pacifica puede ilustrar la endogeneidad, pero no puede solucionarla, y la verdadera lección que nos enseña este artículo es precisamente por qué no puede hacerlo.

En primer lugar, la pregunta. Supongamos que se le pregunta a un tasador si la inversión propuesta de \$45,000 en la reforma de la cocina de una vivienda concreta se recuperaría en el momento de la venta. Se trata de una pregunta sobre una vivienda y dos precios que no puede tener a la vez: su precio tras la reforma y su precio tal y como está, $P_i(1) - P_i(0)$, donde $P_i(1)$ es el precio de la vivienda i si se reforma y $P_i(0)$ su precio si no se reforma. No se trata de una pregunta sobre el precio de mercado en función del estado de la vivienda, y mantener fijo el estado de la vivienda al responderla constituiría el error del mediador descrito anteriormente.

Las dieciséis características de la entrada 008 (*sesgo de variable omitida*) incluyen un indicador de reforma cuyo coeficiente es \$60,000: la diferencia de precio que los datos construidos atribuyen a una reforma —\$60,000 más que una vivienda con la misma superficie bruta alquilable (GLA), parcela, vistas y puntuación de estado. Ese \$60,000 no es un precio por el acto de reformar; representa todo aquello que la reforma aportó y que las demás características no registran. En ese conjunto de datos construido, el autor decidió qué viviendas habían sido reformadas. El indicador de reforma es exógeno por definición, no está relacionado con nada de lo que queda en el error porque no queda nada en el error, por lo que \$60,000 representa, en ese conjunto de datos, lo que una reforma añade al precio manteniendo fija la puntuación de estado registrada; una reforma que también aumentara la puntuación añadiría más.

Supongamos ahora que esos mismos dieciséis inmuebles procedieran de un mercado real. Los propietarios deciden reformar y supongamos —como es plausible— que los propietarios de inmuebles con cualidades no observadas destacadas (denominemos a ese componente del precio « u ») son más propensos a reformar que los propietarios de inmuebles por lo demás similares. La regresión que incluye el indicador de reforma ahora converge, a medida que se acumulan las ventas, no hacia \$60,000, sino a

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

donde $\tilde{r}$ es la parte del indicador de reforma que las características incluidas no predicen; Var

(la varianza) mide cuánto varía, y la flecha marcada como p significa «se acerca cada vez más, a medida que se acumulan las ventas». Si las viviendas reformadas tienen, de media, \$25,000 más de u que las viviendas no reformadas observablemente similares, los coeficientes convergen aproximadamente hacia \$85,000, y ninguna cantidad adicional de datos del mismo mercado reduce ese exceso. El indicador de remodelación, cuya entrada de omisión 008 (*sesgo de variable omitida*) se estudió, ahora se mide y se incluye, y el coeficiente sigue sin ser el efecto causal: lo que aún se omite es u , algo que ningún conjunto de datos registra.

En la práctica, esta distinción tiene su importancia. Los coeficientes describen, en promedio, las viviendas que *se* reformaron, es decir, aquellas cuyos propietarios consideraron que la inversión merecía la pena. El inmueble objeto del caso es una vivienda concreta cuyo propietario aún no se ha decidido, y los coeficientes solo son aplicables a ella si se asemeja a aquellas viviendas en lo que sus propietarios sabían —que es precisamente lo que el modelo no puede ver—. Un tasador que responda a la pregunta sobre la amortización a partir de la \$85,000 sobrevalora la rentabilidad; quien la responde a partir de un modelo que también midió, y mantuvo fijas, todas las condiciones y la calidad que la reforma generó, la subestima, reduciéndola a lo que la reforma aporta al margen de ellas, lo cual, según el principio de la entrada 008 (*sesgo de variable omitida*), es poco o nada. Ambos han interpretado la diferencia de precio condicional asociada a la reforma como el efecto de la propia reforma.

La pregunta de diagnóstico que hay que plantearse respecto a cualquier coeficiente que se pretenda interpretar de forma causal no es, por lo tanto, «¿está bien especificado el modelo?», sino «¿qué efecto se busca y cómo llegó esta variable a tomar los valores que tomó?». Si la respuesta implica una elección realizada por alguien capaz de ver cosas que el modelo no puede ver, el coeficiente describe esa elección tanto como la característica en sí.

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 Scryer. 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)).

```