

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)

Formal definition. Endogeneity is the condition in which a regressor — a feature the regression uses to explain price — moves with the structural error — the price minus what the causal relationship says the price should be, not the fitted model’s residual — so that its estimated coefficient does not recover the causal effect of interest. Its common causes are a feature left out, the price acting back on the feature, the sales being a selected set, and the feature being measured with error.

Intuitive framing. A rate can be wrong because the feature and the price push on each other. The days a house sits on the market depend on the price asked: the feature is partly a result of the price, or of something else that also raises the price, such as a rising market, and a rate measured from those sales mixes the two directions together. No regression can untangle them from the sales alone, whether it fits one rate across the whole range of a feature or, as MARS (Multivariate Adaptive Regression Splines, Friedman 1991) does, different rates over different stretches of it — it takes an outside source of variation, or an assumption the sales cannot check — because the tangle is in the market, not in the arithmetic. What the appraiser can do is know the four ways a feature can move with what the model leaves out (set out below), recognize them in a dataset, and say which one is in play before a rate from that dataset is used as if it measured what a buyer pays for the feature.

For the reader who wants the mechanism. Statisticians call the feature endogenous when it moves together with something the model leaves out, and the entry gives the condition its exact form and separates its four sources. The remedies the statistical literature offers, chiefly instrumental variables, are stated for ordinary regression and are rarely available in appraisal data; the entry says so. The diagnosis, unlike the remedy, applies to any fitting method.

This item has a statistical context and also involves causal reasoning, since it addresses situations in which a regression coefficient does not capture the causal effect an appraiser wants. It is intended for appraisers who read coefficients or paired-sales adjustments as statements of cause. Like entry 008 (*Omitted Variable Bias*) it uses the constructed Pacifica market, which puts condition into the regression as traditional appraisal does; the Journal’s own method, RCA-based appraisal, does not (entry 007, *Conditional Expectation*), and the example is a foil, not a model to follow.

Omitted variable bias, the subject of entry 008 (*Omitted Variable Bias*), should be seen not as a standalone kind of failure with its own formula, but as an example of a more general situation: endogeneity.

Recall the decomposition of entry 007 (*Conditional Expectation*): $P = E[P \mid \mathbf{Z}] + \varepsilon$, where $E[P \mid \mathbf{Z}]$ is the average price of houses with features $\mathbf{Z}$ and ε is how far, above or below, a house’s price lies from that average, so that by construction $E[\varepsilon \mid \mathbf{Z}] = 0$. This holds for the conditional expectation function (the CEF, from that entry) whatever features $\mathbf{Z}$ contains. If the appraiser wants a causal effect instead — how much more the price would be with a view, all else about the house genuinely held constant — one more condition is needed: the regressors must not move with anything else that affects price. When one of them does, that regressor is *endogenous*. Write the causal relationship as $P = \beta_0 + \beta_1 z_1 + \cdots + \beta_k z_k + u$, where each β_j is what a one-unit change

in feature z_j would do to the price. Its error term is the *structural error*, written u : the price minus what the causal relationship says it should be. It is a different thing from ε , the distance from the conditional average, which has a zero mean given $\mathbf{Z}$ by construction; u may move with the regressors, and the endogeneity condition refers to it.

The term originates in simultaneous-equations modeling, where a variable is called “endogenous” if it is determined within the system rather than provided from outside. In current usage it has a more mechanical meaning: $\text{Cov}(z_j, u) \neq 0$ (Cov, the covariance, measures how two quantities move together) — the regressor and the structural error move together, so the fit (OLS — ordinary least squares regression — or MARS) cannot separate their contributions and assigns part of the error’s movement to the regressor.

Four mechanisms cause this, and it is important to keep them separate because they call for different responses.

Omitted variables. A feature that affects price is left out and moves with an included one; it becomes part of u , and $\text{Cov}(z_j, u) \neq 0$ follows. This is the case of entry 008 (*Omitted Variable Bias*), where the remodel indicator was left out. The remedy is to measure and include the feature, or to work out the largest and smallest values the bias could plausibly take.

Simultaneity. The regressor is not just associated with price but partly caused by it, or settled together with it; causation runs both ways, and since the fit (OLS or MARS) says nothing about direction, it returns a mix of the two. The remedy is an *instrument* — a variable that moves the regressor in question but has no other connection to price (the econometric sense of the word, not the legal sense of a formal document) — or a different design.

Selection. The sample is not a random draw from the houses the conclusion is about, and what put a house in the sample is connected to its price. The remedy is to model how the sample was selected, or to redefine the group the conclusion is about.

Measurement error. The feature is recorded inaccurately — two inspectors would score it differently, or the living area is taken from a record rather than measured. The recording error becomes part of u and moves with the recorded value, so the coefficient is biased. In the textbook case — *classical* measurement error, where the error has nothing to do with the true value or with anything else in the model — the mismeasured feature’s coefficient is pulled toward zero. A condition score may not be that case: its error may depend on how good the house really is, and on other things the scorer sees. For such error the bias has no general direction. The remedy is to measure better, or to obtain a second, independent measurement of the same feature of the same house, which a method designed for measurement error can then use to correct the estimate.

The four mechanisms have one thing in common: adding more sales of the same kind does not remove the error they cause. This is the clearest practical distinction between endogeneity and imprecision (the scatter of an estimate that comes from having only a limited number of sales): imprecision shrinks as the sample grows, endogeneity does not.

Before choosing the variables, name the effect. Whether a feature belongs in the model depends on what is being estimated, and two kinds of feature are easily confused. The first is a cause of both the regressor and the price: desirable spots within a neighborhood (a quiet cul-de-sac, a corner with a view) attract expensive renovations and command high prices on their own. Statisticians call such a feature a *confounder*. It must be held fixed, or the regressor’s coefficient carries its effect. The second is the channel through which the regressor acts on the price: a remodel raises the price *by* improving condition and quality. Such a feature is a *mediator*, and holding it fixed removes from the regressor’s coefficient exactly the part of its effect that runs through the channel. So the remodel of entry 008 (*Omitted Variable Bias*) poses two different questions. In traditional appraisal, if the question is the market’s rate for condition, condition is the regressor (RCA-based appraisal never asks that question; entry 007, *Conditional Expectation*, says why) and

the remodel is merely how some houses came to their condition; it earns no value beyond the condition and quality it produced, once those are fully recorded, and a remodeled house and a well-kept house alike in condition and quality are worth the same. If the question is what a remodel does to a particular house's price, then condition and quality are mediators, and a model that holds them fixed answers a different question — what the remodel is worth *apart from* the condition and quality it produced, which may be small and is in any case a different quantity from the total effect of remodeling. “Include the variable” is the right rule for a confounder and the wrong one for a mediator; neither can be chosen until the effect wanted is named.

Where appraisers encounter it. Reverse causation is a matter of location. Endogeneity due to simultaneity is well illustrated by school quality, which is a genuine problem, not a hypothetical one. If you regress price on a school rating, the coefficient appears to show how much buyers pay for the school. But causation also runs the other way: more expensive neighborhoods have a bigger tax base and more involved, better-resourced parents, and can better attract and keep teachers. School quality can raise prices, and higher prices can improve school quality. The coefficient captures neither effect; it reflects a blend that depends on the relative strength of the two channels in that market. The same two-way relationship can appear with crime rates, retail amenity density, and tree canopy. In each case the factor may be both a cause and an effect of neighborhood price levels.

A remodel is a treatment, but it is not randomly assigned. In the vocabulary of causal inference a *treatment* is any intervention whose effect is asked about, and a kitchen remodel is one. The difficulty is who receives it. Many homeowners renovate when they expect the return to exceed the cost, and that expectation rests on what they know about their house that no dataset records — a good floor plan, a strong spot within the neighborhood, a lot that supports the investment. Those same qualities raise the price directly: they are confounders of the remodel, and ones the dataset does not measure. So a model that includes the remodel indicator has cured the omission of the indicator and not the confounding. The indicator's coefficient combines the remodel's effect on the price with the extra price, beyond what their recorded features explain, of the houses whose owners chose to remodel. Including the variable removed the omission; it did not make the regressor exogenous — that is, unconnected with the structural error.

The sample of sold properties is not random. People sell when circumstances are right or prices are attractive, and others sell because they must, so the properties that sell may differ from those that do not in ways the model does not see; when they do, the idiosyncratic element of a sold property's price is not random: the first group pushes it one way, distress and forced sales the other. Within the sales, arm's-length transactions and distressed sales are each further non-random subgroups. Every comparable on every grid is drawn from these selected sales. Any valuation conclusion can be extended to unsold properties only on assumptions about how the sold ones were selected, which a report should state.

A condition score can be contaminated by the price. A score is meant to measure the house's condition and nothing else, and the appraiser scores the subject on inspection, before any value is concluded. The comparables are another matter: they may be scored from listing photographs and remarks, after their sale prices are known. If the known price nudges the score — a house that sold high being read as a little better kept — then the regressor has become partly a function of the outcome. This is not a characteristic of condition itself but a defect of the measurement process: *outcome-dependent measurement error*, the price acting on the score. It is of the non-classical kind just described, with one difference. For a nudge of this kind, which carries part of the price's deviation into the score, the direction is known: the score moves toward the score the sale price would suggest, and the condition coefficient is pushed upward (other errors in

the score may still pull the other way). The protection is procedural: score the comparables from the evidence of their condition before looking at what they sold for, and record the basis of each score in the work file.

Why it matters for defensibility. Whether a regressor is endogenous is settled by argument, not by anything the ordinary regression can show. None of the usual checks of a fit — a plot of the residuals, the share of price variation explained (R^2), testing the model on sales held out of the fit (cross-validation), or a test of whether a coefficient could be zero (a significance test) — can establish that a regressor is exogenous, since the structural error is unobservable and so is its connection with the regressors. Formal tests exist — the Hausman test compares the ordinary estimate with one obtained another way — but they need an instrument or a second estimator, that is, additional assumptions which the data cannot themselves check. What can always be given is a reasoned explanation of the mechanism: why this regressor might be connected with the error, in what direction, and to what extent. A reasonable report makes this argument rather than suggesting that the problem does not exist by failing to mention it.

Say what the coefficient is, not what one wishes it were. The coefficient of an endogenous regressor is a real, well-defined figure: the difference in fitted price per unit of the feature, the measured features held fixed. It is not the causal effect. In the sixteen Pacifica sales with the remodel indicator left out (entry 008, *Omitted Variable Bias*), it is correct to say “these houses with views sold for \$79,178 more, holding the measured features fixed”. Saying “the view contributes \$79,178 to value” asserts a cause the estimate does not support. The first statement will withstand cross-examination; the second will not.

Design beats estimation. Where the two-way influence between a feature and the price can be broken by the way the data are organized — by comparing properties before and after a view has been blocked, by using a boundary where school assignment changes but the housing stock does not, or by taking a renovation forced by an external event such as a fire or flood rather than one chosen by the owner — each design still rests on assumptions that must be defended (that nothing else changed at the same time, that the boundary separates only the school assignment), but a design whose assumptions are defended achieves more defensibility than any estimator working with observational data. This is what the instrumental-variables literature has to say in practical terms, and its relevance to appraisal is that a well-chosen set of comparables is in itself a design.

Worked appraisal example. The Pacifica dataset can illustrate endogeneity but cannot cure it, and why the dataset cannot is the entry’s real lesson.

First, the question. Suppose an appraiser is asked whether a proposed \$45,000 kitchen renovation of a particular house would be recovered at sale. That is a question about one house and two prices it cannot have at once: its price remodeled and its price as it stands, $P_i(1) - P_i(0)$, where $P_i(1)$ is house i ’s price if remodeled and $P_i(0)$ its price if not. It is not a question about the market’s rate for condition, and holding condition fixed while answering it would be the mediator mistake described above.

The sixteen properties of entry 008 (*Omitted Variable Bias*) carry a remodel indicator whose coefficient is \$60,000: the price difference the constructed data attach to a remodel — \$60,000 more than a house with the same GLA, lot, view and condition score. That \$60,000 is not a price for the act of remodeling; it stands for whatever the remodel brought that the other features do not record. In that constructed dataset, which houses were remodeled was decided by the author. The remodel indicator is exogenous by construction, connected with nothing left in the error because nothing is left in the error, and so \$60,000 is, in that dataset, what a remodel adds to a price with the recorded condition score held fixed; a remodel that also raised the score would add more.

Now suppose the same sixteen properties came from an actual market. Owners choose to

remodel, and suppose — as is plausible — that owners of houses with strong unobserved qualities (call that component of price u) are more likely to remodel than owners of otherwise similar houses. The regression that includes the remodel indicator now converges, as sales accumulate, not to \$60,000 but to

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

where $\tilde{r}$ is the part of the remodel indicator that the included features do not predict, Var (the variance) measures how much it varies, and the arrow marked p means “gets ever closer to, as sales accumulate”. If remodeled houses carry, on average, \$25,000 more of u than observably similar unremodeled ones, the coefficient converges to roughly \$85,000, and no amount of additional data from the same market reduces the excess. The remodel indicator, whose omission entry 008 (*Omitted Variable Bias*) studied, is now measured and included, and the coefficient is still not the causal effect: what is still left out is u , which no dataset records.

The distinction has teeth in practice. The coefficient describes, on average, the houses that *were* remodeled, which are the houses whose owners judged the return worthwhile. The subject is a particular house whose owner has not yet decided, and the coefficient carries over to it only if the subject resembles those houses in what their owners knew — which is exactly what the model cannot see. An appraiser who answers the payback question from the \$85,000 overstates the return; one who answers it from a model that also measured, and held fixed, all the condition and quality the remodel produced understates it, down to whatever the remodel adds apart from them, which under the principle of entry 008 (*Omitted Variable Bias*) is little or nothing. Both have read the conditional price difference associated with remodeling as the effect of the remodel itself.

The diagnostic question to ask of any coefficient one intends to read causally is therefore not “is the model well specified?” but “what effect is wanted, and how did this variable come to take the values it took?” If the answer involves a choice made by someone who could see things the model cannot, the coefficient describes that choice as much as the feature.

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.>").

% 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)).

```