

Endogénéité (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)

Définition formelle. L'endogénéité est la situation dans laquelle une variable explicative — une caractéristique utilisée par la régression pour expliquer le prix — évolue parallèlement à l'erreur structurelle — c'est-à-dire la différence entre le prix et la valeur à laquelle il devrait s'élever selon la relation causale, et non le résidu de l'ajustement du modèle —, de sorte que son coefficient estimé ne reflète pas l'effet causal d'intérêt. Ses causes courantes sont une caractéristique omise, le prix agissant en retour sur la caractéristique, les ventes constituant un ensemble sélectionné, et la caractéristique étant mesurée avec une erreur.

Présentation intuitive. Un taux peut être erroné car les caractéristiques et le prix s'influencent mutuellement. La durée pendant laquelle une maison reste sur le marché dépend du prix demandé : les caractéristiques résultent en partie du prix, ou d'un autre facteur qui fait également grimper le prix, comme un marché en hausse, et un taux calculé à partir de ces ventes mélange ces deux aspects. Aucune régression ne peut les dissocier à partir des seules ventes, lors de l'ajustement d'un taux unique à l'ensemble de la gamme d'une caractéristique ou, comme le fait le modèle MARS (Multivariate Adaptive Regression Splines, Friedman 1991), applique des taux différents selon les segments de cette caractéristique — il faut alors recourir à une source de variation externe, ou à une hypothèse que les ventes ne permettent pas de vérifier — car l'enchevêtrement réside dans le marché, et non dans le calcul. Ce que l'expert immobilier peut faire, c'est connaître les quatre façons dont une caractéristique peut évoluer en fonction de ce que le modèle omet (énoncées ci-dessous), les reconnaître dans un ensemble de données, et déterminer laquelle est en jeu avant qu'un taux issu de cet ensemble de données ne soit utilisé comme s'il mesurait ce qu'un acheteur paie pour cette caractéristique.

Pour le lecteur qui souhaite en connaître le mécanisme : . Les statisticiens qualifient une caractéristique d'« endogène » lorsqu'elle évolue de concert avec un élément que le modèle ne prend pas en compte ; cet article présente la formulation exacte de cette condition et distingue ses quatre sources. Les solutions proposées par la littérature statistique, principalement les variables instrumentales, sont présentées dans le cadre d'une régression ordinaire et sont rarement disponibles dans les données d'évaluation ; l'article le précise. Contrairement à la solution, le diagnostic s'applique à n'importe quelle méthode d'ajustement.

Cet article s'inscrit dans un contexte statistique et fait également appel au raisonnement causal, puisqu'il traite des situations dans lesquelles un coefficient de régression ne rend pas compte de l'effet causal recherché par un expert. Il s'adresse aux experts qui interprètent les coefficients ou les ajustements basés sur des ventes comparables comme des indications de cause. À l'instar de l'entrée 008 (*biais de variable omise*), il utilise le marché Pacifica reconstitué, qui intègre l'état du bien dans la régression, à l'instar de l'expertise traditionnelle ; la méthode propre au Journal, l'expertise basée sur l'analyse RCA, ne le fait pas (entrée 007, *espérance conditionnelle*), et l'exemple sert de contre-exemple, et non de modèle à suivre.

Le biais lié aux variables omises, qui est le bien à évaluer de l'entrée n° 008 (*Biais lié aux variables omises*), ne doit pas être considéré comme un type d'erreur isolé, doté de sa propre formule, mais comme un exemple d'une situation plus générale : l'endogénéité.

Rappelons la décomposition de l'entrée 007 (*Espérance conditionnelle*) : $P = E[P \mid \mathbf{Z}] + \varepsilon$, où $E[P \mid \mathbf{Z}]$ est le prix moyen des maisons présentant les caractéristiques $\mathbf{Z}$ et ε représente l'écart, à la hausse ou à la baisse, du prix d'une maison par rapport à cette moyenne, de sorte que, par construction, $E[\varepsilon \mid \mathbf{Z}] = 0$. Cela vaut pour la fonction d'espérance conditionnelle (la CEF, d'après cette entrée), quelles que soient les caractéristiques contenues dans $\mathbf{Z}$. Si l'expert souhaite plutôt déterminer un effet causal — à savoir de combien le prix serait plus élevé si la maison disposait d'une vue, toutes les autres caractéristiques de la maison étant véritablement maintenues constantes —, une condition supplémentaire est nécessaire : les régresseurs ne doivent pas varier en fonction d'autres facteurs influençant le prix. Lorsqu'un d'entre eux varie, ce régresseur est *endogène*. On écrit la relation causale sous la forme $P = \beta_0 + \beta_1 z_1 + \dots + \beta_k z_k + u$, où chaque β_j correspond à l'effet sur le prix d'une variation d'une unité de la caractéristique z_j . Son terme d'erreur est l'*erreur structurelle*, notée u : le prix moins la valeur à laquelle la relation causale indique qu'il devrait s'élever. Il s'agit d'une notion différente de ε , qui correspond à la distance par rapport à la moyenne conditionnelle, laquelle a une moyenne nulle étant donné que $\mathbf{Z}$ par construction ; u peut varier en fonction des variables explicatives, et c'est à elle que se réfère la condition d'endogénéité.

Ce terme trouve son origine dans la modélisation par équations simultanées, où une variable est dite « endogène » si elle est déterminée au sein du système plutôt que fournie de l'extérieur. Dans l'usage courant, il revêt une signification plus mécanique : $\text{Cov}(z_j, u) \neq 0$ (Cov, la covariance, mesure la façon dont deux quantités évoluent conjointement) — le régresseur et l'erreur structurelle évoluent de concert, de sorte que l'ajustement (OLS — régression par les moindres carrés ordinaires — ou MARS) ne peut pas distinguer leurs contributions respectives et attribue une partie de la variation de l'erreur au régresseur.

Quatre mécanismes sont à l'origine de ce phénomène, et il est important de les distinguer les uns des autres, car ils nécessitent des réponses différentes.

Variables omises. Une caractéristique qui influe sur le prix est omise et évolue parallèlement à une caractéristique incluse ; elle fait alors partie de u , ce qui entraîne $\text{Cov}(z_j, u) \neq 0$. C'est le cas de l'entrée 008 (*biais de variable omise*), où l'indicateur de rénovation a été omis. Pour y remédier, il faut mesurer et inclure cette caractéristique, ou déterminer les valeurs maximales et minimales que ce biais pourrait raisonnablement prendre.

Simultanéité. La variable explicative n'est pas seulement associée au prix, mais elle est en partie causée par celui-ci, ou évolue de concert avec lui ; la causalité va dans les deux sens, et comme l'ajustement (OLS ou MARS) ne fournit aucune indication sur la direction, il renvoie un mélange des deux. La solution consiste à utiliser une variable *instrumentale* — une variable qui fait varier la variable explicative en question mais qui n'a aucun autre lien avec le prix (au sens économétrique du terme, et non au sens juridique d'un document officiel) — ou à opter pour un autre modèle.

Sélection. L'échantillon n'est pas constitué d'un échantillon aléatoire des maisons sur lesquelles porte la conclusion, et le fait qu'une maison figure dans l'échantillon est lié à son prix. Pour y remédier, il faut modéliser la manière dont l'échantillon a été sélectionné, ou redéfinir le groupe sur lequel porte la conclusion.

Erreur de mesure. La caractéristique est enregistrée de manière imprécise : deux inspecteurs lui attribueraient une note différente, ou la surface habitable est tirée d'un registre plutôt que mesurée. L'erreur d'enregistrement fait partie intégrante de u et varie en fonction de la valeur enregistrée, ce qui entraîne un biais du coefficient. Dans le cas type — l'erreur de mesure *classique*, où l'erreur n'a aucun rapport avec la valeur réelle ni avec aucun autre élément du modèle —, le coefficient de la caractéristique mal mesurée est ramené vers zéro. Un score de condition peut ne pas correspondre à ce cas : son erreur peut dépendre de la qualité réelle de la maison, ainsi que d'autres éléments observés par l'évaluateur. Pour ce type d'erreur, le biais n'a pas de direction générale.

La solution consiste à améliorer la mesure ou à obtenir une deuxième mesure indépendante de la même caractéristique de la même maison, qu'une méthode conçue pour traiter les erreurs de mesure pourra ensuite utiliser pour corriger l'estimation.

Ces quatre mécanismes ont un point commun : le fait d'ajouter davantage de ventes du même type ne permet pas d'éliminer l'erreur qu'ils engendrent. C'est là la distinction pratique la plus évidente entre l'endogénéité et l'imprécision (la dispersion d'une estimation due au nombre limité de ventes) : l'imprécision diminue à mesure que l'échantillon s'élargit, contrairement à l'endogénéité.

Avant de choisir les variables, nommez l'effet. L'inclusion d'une caractéristique dans le modèle dépend de ce qui est estimé, et deux types de caractéristiques peuvent facilement prêter à confusion. Le premier est une cause à la fois du régresseur et du prix : les emplacements recherchés au sein d'un quartier (une impasse tranquille, un coin avec vue) attirent des travaux de rénovation coûteux et justifient à eux seuls des prix élevés. Les statisticiens appellent une telle caractéristique un *facteur de confusion*. Elle doit être maintenue fixe, sinon le coefficient du régresseur en reflétera l'effet. La seconde est le canal par lequel le régresseur agit sur le prix : une rénovation augmente le prix *en* améliorant l'état et la qualité du bien. Une telle caractéristique est un *médiateur*, et le fait de la maintenir fixe élimine du coefficient de la variable explicative exactement la partie de son effet qui passe par ce canal. Ainsi, la rénovation de l'entrée 008 (*biais de variable omise*) soulève deux questions distinctes. Dans l'évaluation traditionnelle, si la question porte sur le taux du marché correspondant à l'état, c'est l'état qui est la variable explicative (l'évaluation basée sur l'analyse RCA ne pose jamais cette question ; l'entrée 007, *Espérance conditionnelle*, en explique la raison) et la rénovation n'est que la manière dont certaines maisons sont parvenues à cet état ; elle n'apporte aucune valeur au-delà de l'état et de la qualité qu'elle a générés, une fois ceux-ci pleinement pris en compte, et une maison rénovée et une maison bien entretenue, de même état et de même qualité, ont la même valeur. Si la question porte sur l'impact d'une rénovation sur le prix d'une maison particulière, alors l'état et la qualité sont des médiateurs, et un modèle qui les maintient fixes répond à une question différente : quelle est la valeur de la rénovation *indépendamment de* l'état et de la qualité qu'elle a générés, valeur qui peut être faible et qui, en tout état de cause, est une grandeur différente de l'effet total de la rénovation. « Inclure la variable » est la règle appropriée pour un facteur de confusion, mais erronée pour un médiateur ; aucune des deux ne peut être choisie tant que l'effet recherché n'a pas été défini.

Où l'évaluateur la rencontre. La causalité inverse est une question de localisation. L'endogénéité due à la simultanéité est bien illustrée par la qualité des écoles, qui constitue un véritable problème, et non un problème hypothétique. Si l'on effectue une régression du prix sur la note attribuée à une école, le coefficient semble indiquer le surcoût que les acheteurs sont prêts à payer pour cette école. Mais la causalité fonctionne également dans l'autre sens : les quartiers plus chers disposent d'une assiette fiscale plus importante et de parents plus impliqués et mieux lotis, ce qui leur permet de mieux attirer et retenir les enseignants. La qualité des écoles peut faire grimper les prix, et des prix plus élevés peuvent améliorer la qualité des écoles. Le coefficient ne rend compte d'aucun de ces deux effets ; il reflète un mélange qui dépend de l'importance relative de ces deux canaux sur ce marché. Cette même relation bidirectionnelle peut se manifester avec les taux de criminalité, la densité des commerces de proximité et la couverture arborée. Dans chaque cas, le facteur peut être à la fois une cause et une conséquence des niveaux de prix du quartier.

Une rénovation est un traitement, mais elle n'est pas attribuée de manière aléatoire. Dans le vocabulaire de l'inférence causale, un *traitement* désigne toute intervention dont on cherche à évaluer l'effet, et la rénovation d'une cuisine en fait partie. La difficulté réside dans le choix des bénéficiaires. De nombreux propriétaires rénovent lorsqu'ils s'attendent à ce que le retour sur investissement dépasse le coût, et cette attente repose sur ce qu'ils savent de leur maison et qui

n'est consigné dans aucun ensemble de données : un bon agencement, un emplacement privilégié dans le quartier, un terrain propice à l'investissement. Ces mêmes qualités font directement grimper le prix : ce sont des facteurs de confusion liés à la rénovation, que l'ensemble de données ne mesure pas. Ainsi, un modèle incluant l'indicateur de rénovation a corrigé l'omission de cet indicateur, mais pas la confusion. Le coefficient de l'indicateur combine l'effet de la rénovation sur le prix avec la plus-value, au-delà de ce qu'expliquent les caractéristiques enregistrées, des maisons dont les propriétaires ont choisi de rénover. L'inclusion de la variable a corrigé l'omission ; elle n'a pas rendu la variable explicative exogène — c'est-à-dire sans lien avec l'erreur structurelle.

L'échantillon des biens immobiliers vendus n'est pas aléatoire. Certains vendent lorsque les circonstances s'y prêtent ou que les prix sont attractifs, tandis que d'autres vendent par nécessité ; ainsi, les biens qui se vendent peuvent différer de ceux qui ne se vendent pas selon des critères que le modèle ne prend pas en compte ; lorsque c'est le cas, la composante idiosyncratique du prix d'un bien vendu n'est pas aléatoire : le premier groupe l'oriente dans un sens, tandis que les ventes en situation de détresse et les ventes forcées l'orientent dans l'autre. Au sein des ventes, les transactions entre parties indépendantes et les ventes en situation de détresse constituent chacune des sous-groupes supplémentaires non aléatoires. Chaque bien comparable figurant sur chaque grille est issu de ces ventes sélectionnées. Toute conclusion d'évaluation ne peut être étendue aux biens non vendus qu'en se fondant sur des hypothèses concernant la manière dont les biens vendus ont été sélectionnés, ce qu'un rapport doit préciser.

Le score d'état peut être faussé par le prix. Ce score est censé mesurer l'état de la maison et rien d'autre ; l'expert attribue la note au bien à évaluer lors de l'inspection, avant même d'en déterminer la valeur. Les biens comparables constituent un autre cas de figure : ils peuvent être notés à partir des photos et des commentaires figurant dans les annonces, une fois leurs prix de vente connus. Si le prix connu influence la note — une maison vendue à un prix élevé étant considérée comme un peu mieux entretenue —, alors le modèle de régression devient en partie fonction du résultat. Il ne s'agit pas d'une caractéristique de l'état en soi, mais d'un défaut du processus de mesure : *une erreur de mesure dépendante du résultat*, le prix agissant sur la note. Il s'agit du type non classique que nous venons de décrire, à une différence près. Pour une influence de ce type, qui répercute une partie de l'écart de prix sur la note, la direction est connue : la note tend vers celle que le prix de vente suggérerait, et le coefficient d'état est poussé à la hausse (d'autres erreurs dans la note peuvent toutefois exercer une influence inverse). La protection est d'ordre procédural : attribuer une note aux biens comparables en fonction des indices de leur état avant d'examiner leur prix de vente, et consigner le fondement de chaque note dans le dossier de travail.

Pourquoi elle compte pour la solidité. La question de savoir si une variable explicative est endogène se tranche par l'argumentation, et non par ce que peut révéler une régression classique. Aucun des contrôles habituels d'ajustement — représentation graphique des résidus, R^2 , validation croisée du modèle sur des ventes non utilisées pour l'ajustement, ou test visant à déterminer si un coefficient pourrait être nul (un test de significativité) — ne permet d'établir qu'une variable explicative est exogène, car l'erreur structurelle est inobservable, tout comme son lien avec les variables explicatives. Il existe des tests formels — le test de Hausman compare l'estimation ordinaire à une estimation obtenue par une autre méthode — mais ceux-ci nécessitent un instrument ou un deuxième estimateur, c'est-à-dire des hypothèses supplémentaires que les données ne peuvent pas vérifier elles-mêmes. Ce qu'il est toujours possible de fournir, c'est une explication raisonnée du mécanisme : pourquoi cette variable explicative pourrait être liée à l'erreur, dans quel sens et dans quelle mesure. Un rapport raisonnable présente cet argument plutôt que de suggérer que le problème n'existe pas en omettant de le mentionner.

Indiquez la valeur réelle du coefficient, et non celle que l'on souhaiterait. Le coeffi-

cient d'une variable endogène est une valeur réelle et bien définie : il s'agit de la différence de prix ajusté par unité de la caractéristique, toutes les autres caractéristiques mesurées étant maintenues constantes. Il ne s'agit pas de l'effet causal. Dans les seize ventes de Pacifica où l'indicateur de rénovation a été omis (entrée 008, *biais de variable omise*), il est correct de dire « ces maisons avec vue se sont vendues \$79,178 de plus, toutes les caractéristiques mesurées étant maintenues constantes ». Affirmer que « la vue contribue à hauteur de \$79,178 à la valeur » revient à postuler une cause que l'estimation ne corrobore pas. La première affirmation résistera à un contre-interrogatoire ; la seconde, non.

La conception l'emporte sur l'estimation. Lorsque l'influence réciproque entre une caractéristique et le prix peut être isolée grâce à la manière dont les données sont organisées — en comparant les biens avant et après le blocage d'une vue, en utilisant une limite où l'affectation scolaire change mais pas le parc immobilier, ou en prenant en compte une rénovation imposée par un événement extérieur tel qu'un incendie ou une inondation plutôt qu'une rénovation choisie par le propriétaire — chaque modèle repose néanmoins sur des hypothèses qui doivent être justifiées (qu'aucun autre élément n'a changé au même moment, que la limite ne sépare que l'affectation scolaire), mais un modèle dont les hypothèses sont justifiées offre une meilleure défendabilité que n'importe quel estimateur fonctionnant avec des données d'observation. C'est ce que la littérature sur les variables instrumentales a à dire en termes pratiques, et sa pertinence pour l'évaluation réside dans le fait qu'un ensemble de biens comparables bien choisi constitue en soi un modèle.

Exemple d'évaluation chiffré. L'ensemble de données Pacifica permet d'illustrer l'endogénéité, mais ne peut pas y remédier ; et c'est précisément la raison pour laquelle cet ensemble de données n'y parvient pas qui constitue la véritable leçon à retenir de cet article.

Tout d'abord, la question. Supposons qu'on demande à un expert immobilier si les travaux de rénovation de la cuisine d'une maison donnée, proposés sur \$45,000, seraient rentabilisés lors de la vente. Il s'agit d'une question portant sur une maison et deux prix qu'elle ne peut avoir simultanément : son prix après rénovation et son prix en l'état, $P_i(1) - P_i(0)$, où $P_i(1)$ correspond au prix de la maison i si elle est rénovée et $P_i(0)$ à son prix si elle ne l'est pas. Il ne s'agit pas d'une question relative au cours du marché en fonction de l'état de la maison, et maintenir cet état fixe tout en y répondant constituerait l'erreur de médiation décrite ci-dessus.

Les seize biens de l'entrée 008 (*biais de variable omise*) comportent un indicateur de rénovation dont le coefficient est de \$60,000 : il s'agit de la différence de prix que les données construites attribuent à une rénovation — soit \$60,000 de plus qu'une maison présentant la même surface brute locative (GLA), le même terrain, la même vue et le même score d'état. Ce montant de \$60,000 ne correspond pas au coût de la rénovation elle-même ; il représente tout ce que la rénovation a apporté et que les autres caractéristiques ne prennent pas en compte. Dans cet ensemble de données construit, c'est l'auteur qui a décidé quelles maisons avaient été rénovées. L'indicateur de rénovation est exogène par construction ; il n'est lié à aucun élément restant dans l'erreur, car il ne reste rien dans l'erreur. Ainsi, \$60,000 représente, dans cet ensemble de données, la plus-value qu'apporte une rénovation au prix, en supposant que la note d'état enregistrée reste fixe ; une rénovation qui améliorerait également cette note ajouterait une plus-value supplémentaire.

Supposons maintenant que ces seize biens proviennent d'un marché réel. Les propriétaires choisissent de rénover, et supposons — ce qui est plausible — que les propriétaires de maisons présentant des qualités non observables importantes (appelons cette composante du prix « u ») soient plus enclins à rénover que les propriétaires de maisons par ailleurs similaires. La régression qui inclut l'indicateur de rénovation converge désormais, à mesure que les ventes s'accumulent, non

pas vers \$60,000 mais à

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

où $\tilde{r}$ est la partie de l'indicateur de rénovation que les caractéristiques incluses ne permettent pas de prédire, Var (la variance) mesure son ampleur de variation, et la flèche marquée p signifie « se rapproche de plus en plus, à mesure que les ventes s'accumulent ». Si les maisons rénovées affichent, en moyenne, \$25,000 de plus de u que les maisons non rénovées observables similaires, le coefficient converge vers environ \$85,000, et aucune quantité supplémentaire de données provenant du même marché ne réduit cet excès. L'indicateur de rénovation, dont l'entrée d'omission 008 (*biais de variable omise*) a été étudiée, est désormais mesuré et pris en compte, mais le coefficient ne correspond toujours pas à l'effet causal : ce qui manque encore, c'est u , que nul ensemble de données ne recense.

Cette distinction a une réelle incidence dans la pratique. Le coefficient décrit, en moyenne, les maisons qui *ont été* rénovées, c'est-à-dire celles dont les propriétaires ont estimé que le retour sur investissement en valait la peine. Le bien à évaluer est une maison particulière dont le propriétaire n'a pas encore pris de décision, et le coefficient ne s'applique à celle-ci que si elle ressemble à ces maisons en ce que leurs propriétaires savaient — ce qui est précisément ce que le modèle ne peut pas prendre en compte. Un expert qui répond à la question du retour sur investissement à partir de l'\$85,000 surestime le rendement ; celui qui y répond à partir d'un modèle qui a également mesuré, et maintenu fixes, toutes les conditions et la qualité générées par la rénovation le sous-estime, le ramenant à ce que la rénovation apporte en dehors de ces éléments, ce qui, selon le principe n° 008 (*biais de variable omise*), est peu ou rien. Les deux ont interprété la différence de prix conditionnelle liée à la rénovation comme étant l'effet de la rénovation elle-même.

La question diagnostique à se poser concernant tout coefficient que l'on entend interpréter de manière causale n'est donc pas « le modèle est-il bien spécifié ? », mais « quel effet recherche-t-on, et comment cette variable en est-elle venue à prendre les valeurs qu'elle a prises ? » Si la réponse implique un choix effectué par quelqu'un capable de percevoir des éléments que le modèle ne peut pas saisir, le coefficient décrit autant ce choix que la caractéristique elle-même.

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)).

```