

Identification des paramètres (Parameter Identification)

Foundations • The Valuation Engineer

William Bert Craytor • October 1, 2026

Version 1.2.0

DOI: [10.67330/xjszh949](https://doi.org/10.67330/xjszh949)

Définition formelle. Un paramètre est identifié lorsqu'il est déterminé de manière unique par la distribution conjointe des données observables (fréquence à laquelle chaque combinaison de caractéristiques et de prix apparaît parmi l'ensemble de ces ventes), compte tenu des hypothèses et de la spécification du modèle. L'identification précède logiquement l'estimation et s'en distingue.

Présentation intuitive. Il existe certains taux que les ventes ne peuvent tout simplement pas vous fournir, et aucune méthode ne pourra y changer quoi que ce soit. Si toutes les maisons rénovées sur le marché sont en excellent état et que toutes les maisons en excellent état ont été rénovées, les ventes reflètent la valeur de ces deux éléments combinés et ne renseignent en rien sur chacun d'entre eux pris isolément ; une régression qui calcule un taux pour chacun d'entre eux — comme on le ferait lorsque les deux vont presque toujours de pair — rend compte d'une distinction que le marché a à peine opérée. Si chaque bien comparable offre une meilleure vue que le bien à évaluer, le taux lié à la vue est interprété au-delà des limites des données. Si les seules ventes qui montrent une évolution des conditions du marché au fil du temps sont également les seules à refléter une variation des taux d'intérêt hypothécaires, il est impossible de distinguer ces deux facteurs. Il s'agit là de questions sur le contenu des transactions, posées avant tout ajustement, et la réponse est la même, que la méthode ajuste un seul taux sur toute la plage d'une caractéristique ou des taux différents sur différentes portions de celle-ci, comme le fait MARS (Multivariate Adaptive Regression Splines, Friedman 1991, la principale méthode de régression du Journal) : MARS ne peut déterminer où un taux change que lorsque les transactions immobilières contiennent le contraste permettant de le mettre en évidence. La première question de l'expert n'est donc pas « quel est le taux ? », mais « ces transactions peuvent-elles me le dire ? ».

Pour le lecteur qui souhaite comprendre le mécanisme : . En statistique, on appelle cela l'identification : il s'agit de déterminer si l'ensemble de la population des ventes, mesurée sans erreur, permettrait de fixer le paramètre. La méthode des moindres carrés ordinaires met en évidence le cas extrême en supprimant l'une des deux caractéristiques qui évoluent exactement de concert, et masque le cas dangereux où elles évoluent presque de concert et où les deux chiffres reposent sur les mêmes quelques ventes. Cet article traite ces deux cas, ainsi que ceux de l'extrapolation et des effets temporels confondus auxquels la grille est confrontée quotidiennement.

Cet article porte sur les statistiques : il examine ce qu'un ensemble de données pourrait révéler en principe avant même qu'un estimateur (une méthode de calcul des taux à partir des ventes, comme les moindres carrés ordinaires ou MARS) n'ait été utilisé, et s'adresse aussi bien à ceux qui constituent des ensembles de données comparables qu'à ceux qui effectuent l'ajustement de modèles de régression.

Une « » *de paramètre* désigne ici un nombre que le modèle doit fournir — un taux ou un coefficient — à partir duquel un ajustement est calculé. La question soulevée par l'estimation est la suivante : sur la base de l'échantillon, quelle valeur dois-je considérer comme ma meilleure estimation du paramètre ? L'identification pose une question préalable : si je disposais de l'ensemble de la population — c'est-à-dire de toutes les transactions qui auront lieu un jour sur ce marché, mesurées sans erreur —, le paramètre serait-il pour autant déterminé ?

Si la réponse est « non », alors le paramètre n'est pas identifié et aucune mesure prise par la suite ne pourra y remédier. Un estimateur amélioré n'y changera rien, un échantillon plus large du même type de ventes ne sera d'aucune utilité, et un modèle plus sophistiqué n'aidera pas non plus, à moins qu'il n'ajoute une hypothèse que les données ne peuvent pas tester (le cas causal ci-dessous), et il pourrait même masquer le problème. L'incapacité à identifier un paramètre relève de la question posée et des données sur lesquelles on s'appuie — ce que le marché a produit et ce que l'évaluateur a recueilli — et non de la méthode de l'analyste.

L'erreur d'identification est plus facile à repérer lorsque deux caractéristiques vont toujours de pair. Imaginons que chaque bien immobilier ayant fait l'objet de travaux de rénovation obtienne également la note « excellent » (9,00) et que chaque bien noté « excellent » ait également été rénové ; dans ce cas, l'ensemble des biens rénovés et celui des biens notés « excellent » sont identiques. Le marché accorde une plus grande valeur à ce groupe : une maison « excellente » et rénovée vaut plus qu'une maison « bonne » qui n'a pas été rénovée. Quel que soit le nombre de ventes observées, il est impossible de distinguer cette différence de prix entre l'état enregistré et ce que la rénovation a apporté de plus que ce que les autres caractéristiques ne reflètent pas, car sur ce marché, l'une ne va jamais sans l'autre. Nous connaissons le total ; nous ne connaissons pas les éléments qui le composent.

Si la méthode des moindres carrés met clairement cela en évidence en cas de colinéarité exacte (lorsqu'une caractéristique est une fonction linéaire exacte des autres), c'est parce que la matrice de conception — c'est-à-dire le tableau constitué des colonnes des régresseurs (variables explicatives) utilisés par l'estimateur — est de rang déficient : l'une de ses colonnes est une combinaison exacte des autres. Le logiciel supprime donc un coefficient ou renvoie une valeur manquante. Il s'agit simplement d'une mesure de courtoisie. Les situations dangereuses sont les cas limites — ceux où les deux variables varient ensemble *presque* systématiquement — car, dans ces cas-là, les paramètres sont techniquement identifiables : le logiciel fournit deux chiffres, tous deux basés sur le petit nombre d'observations qui rompent la régularité. Elles sont identifiées ; mais comme seules quelques ventes s'écartent de la tendance, elles sont si imprécises que, dans la pratique, les données ne permettent pas de les distinguer. C'est là que se confondent problème d'identification et problème de précision. C'est à cela que revient la multicolinéarité (des caractéristiques qui évoluent presque exactement de concert) du point de vue de l'identification.

L'identification peut échouer même en l'absence totale de colinéarité, lorsque le paramètre en question est causal et que les données sont observationnelles, à moins d'ajouter une hypothèse que les données ne permettent pas de tester. L'exemple tiré de l'entrée 009 (*Endogénéité*), dans lequel les seize ventes provenaient d'un marché réel, est de ce type : l'association entre la rénovation et le prix peut être identifiée — puisqu'il s'agit d'une caractéristique de la distribution conjointe et qu'un échantillon de grande taille permettra de l'estimer avec précision — tandis que l'effet causal de la rénovation ne peut pas l'être, car la distribution observable ne contient aucune information permettant de distinguer la valeur ajoutée par la rénovation de la valeur supplémentaire que possédaient déjà les maisons et qui a conduit leurs propriétaires à réaliser les travaux. Pour identifier le paramètre causal, il est nécessaire de formuler une hypothèse qui ne peut elle-même être testée à partir des données, et la fiabilité de la conclusion dépend donc de la crédibilité de cette hypothèse.

Où l'évaluateur la rencontre. L'exigence de « bracketing ». La règle qui exige que le sujet fasse l'objet d'un « bracketing » — c'est-à-dire que des biens comparables, tant supérieurs qu'inférieurs, soient inclus pour chaque caractéristique importante — est présentée comme une règle empirique, et elle répond à deux questions distinctes qu'il convient de bien différencier. La première concerne l'identification : ces ventes présentent-elles un contraste au niveau de la caractéristique, se démarquant des autres ? Si tous les biens comparables offrent la même vue, il est impossible

d'en déduire un taux pour cette caractéristique. La seconde concerne la validité : le bien à évaluer se situe-t-il dans la fourchette sur laquelle le taux a été déterminé ? Si tous les biens comparables offrent une meilleure vue que le bien à évaluer, un taux peut très bien exister — issu de ces ventes ou d'un ensemble de données plus large — mais son application au bien à évaluer revient à extrapoler au-delà des données disponibles. La mise entre parenthèses garantit le respect du deuxième critère ; seul un tel contraste garantit le premier. Un bien à évaluer mis entre parenthèses peut tout de même se situer dans une zone où deux caractéristiques ne varient jamais l'une par rapport à l'autre, et un bien à évaluer non mis entre parenthèses peut tout de même se voir attribuer un taux identifié ailleurs, ce qui constitue la solution du « jeu de données plus large » autorisée par la section relative à la défendabilité.

Voici un exemple de caractéristiques qui ne varient pas indépendamment au sein d'un sous-marché : un secteur où chaque maison située du côté de la rue offrant une vue dispose d'une surface habitable plus grande ; la vue et la superficie ne peuvent pas être considérées comme des facteurs distincts dans ce secteur. Un expert qui procède à des ajustements distincts pour ces deux éléments fournit en réalité deux chiffres que les données du secteur ne permettent pas de distinguer, et la manière dont le montant combiné est réparti entre eux repose sur le jugement de l'expert ou sur un autre marché. Cela peut être approprié, mais il s'agit d'une affirmation différente de celle de « dérivé du marché ».

Âge, état et période de construction. Ces facteurs sont souvent étroitement liés. Les maisons plus anciennes sont généralement en moins bon état et ont été construites selon des normes différentes. L'utilisation d'une grille d'ajustement comportant des lignes distinctes pour l'âge et l'état suggère que le marché évalue ces facteurs séparément et que les données permettent de les distinguer. Il convient de vérifier, et non de supposer, si les données disponibles permettent de les distinguer, et celles-ci échouent souvent à cette vérification.

Le temps et les conditions de marché. Deux types d'erreurs distincts se cachent ici. Si la grille comporte une ligne « conditions de marché » et aucune ligne pour les autres éléments ayant évolué au cours de la période — taux d'intérêt, stock, nouvel employeur —, le coefficient temporel est identifié, mais il est composite et ne correspond pas à l'effet des seules conditions de marché : l'ensemble de la variation est imputé à la seule étiquette disponible, ce qui correspond au cas de variable omise de l'entrée 008 (*biais de variable omise*). Si les deux lignes sont présentes mais que les seules ventes présentant une variation des taux d'intérêt sont celles qui montrent une variation des conditions de marché, la distinction entre les deux n'est pas identifiée, ce qui est le cas de cette entrée.

Pourquoi elle compte pour la solidité. Lorsque vous posez la question d'identification, vous devez la poser avant la question d'estimation, car l'ordre a son importance. Si un évaluateur se demande d'abord « que puis-je distinguer à partir de mes données ? » et seulement ensuite « comment puis-je l'estimer au mieux ? », les ajustements obtenus seront alors ceux que les données peuvent étayer. À l'inverse, si la question d'identification est posée en dernier, on obtiendra alors des chiffres fiables pour des quantités que personne n'aurait pu mesurer.

Un paramètre non identifié doit être signalé comme tel. Lorsque les données permettent de déterminer un effet combiné mais pas sa répartition, l'approche raisonnable consiste non pas à omettre l'ajustement, mais plutôt à rendre compte en toute honnêteté de la manière dont il a été réparti : à savoir que le groupe de comparaison ne peut pas distinguer ces deux effets, que la répartition utilisée repose sur un ensemble de données plus vaste ou sur un jugement, et quel est l'effet combiné, puisque c'est cet effet combiné que les données déterminent effectivement. Le lecteur peut alors se faire sa propre opinion sur cette répartition. En revanche, le lecteur ne peut pas se faire une opinion sur deux chiffres présentés comme étant issus du marché alors qu'en réalité,

seule leur somme l'est.

L'identification est le point de rencontre entre la méthodologie et l'USPAP. L'USPAP correspond aux « Normes uniformes de pratique professionnelle en matière d'évaluation ». Ces normes exigent que l'évaluateur recueille, vérifie et analyse toutes les informations nécessaires pour obtenir des résultats d'évaluation crédibles (Règle n° 1-4). Les directives des prêteurs mettent directement l'accent sur les ajustements : le guide de vente de Fannie Mae exige que les ajustements reflètent la réaction du marché aux différences entre les biens immobiliers. Aucune de ces règles n'impose que chaque ligne de la grille soit identifiée séparément à partir des biens comparables présentés au lecteur ; la réaction du marché peut provenir de ventes appariées ne figurant pas dans la grille, d'un échantillon plus large ou d'un jugement explicite, comme indiqué ci-dessus. Selon l'interprétation du Journal, ces règles ont pour effet de rendre obligatoire la question de l'identification : un expert qui invoque le soutien du marché pour justifier un ajustement doit être en mesure d'indiquer quelles données le caractérisent. Formulée ainsi, la question est vérifiable — on peut examiner les données sur lesquelles on s'appuie et voir si elles contiennent le contraste — sans transformer le test mathématique d'identification en un test de conformité.

La solution réside dans la conception, et non dans le calcul. Si un paramètre n'est pas identifié, la solution consiste à modifier les données utilisées — c'est-à-dire à élargir le champ géographique, à allonger la période considérée, à rechercher les biens qui s'écartent de la tendance — ou à modifier l'hypothèse. Aucun estimateur ne peut extraire des informations que les données ne contiennent pas.

Exemple d'évaluation chiffré. Prenons l'ensemble de données Pacifica comprenant seize biens immobiliers, à l'entrée n° 008 (*biais de variable omise*). L'état indiqué correspond au score d'état attribué par le Journal : il s'agit d'une estimation de l'état de la maison par rapport à celui des habitations environnantes situées dans sa zone de marché définie. Il repose sur la compréhension qu'a l'expert immobilier de cette zone de marché et ne devrait pas s'écarter de manière significative de celle de ses confrères exerçant dans la même zone. Le score varie de 0,00 à 10,00 : il correspond au pourcentage de maisons de la zone de marché qui sont en moins bon état, divisé par dix. L'ensemble de données enregistre trois niveaux : moyen (5,00), bon (7,00) et excellent (9,00), et les données construites attribuent un \$20,000 e par point. Comme dans les entrées 008 (*Biais de variable omise*) et 009 (*Endogénéité*), l'intégration de l'état dans la régression correspond à la pratique de l'expertise traditionnelle, et non à la méthode du Journal, à savoir l'expertise basée sur l'analyse RCA (entrée 007, *Espérance conditionnelle*). Limiter l'ensemble de données aux dix biens immobiliers qui s'inscrivent dans un même schéma — « bon » et non rénové, ou « excellent » et rénové : les biens comparables A, B, C, D, G, I, K, M, N et O. Deux critères de sélection permettent de constituer ce sous-échantillon, et ils jouent des rôles différents. En écartant E, H et J (la maison « excellente » jamais rénovée et les deux maisons « bonnes » rénovées), on obtient un profil réaliste : à quoi ressemble un marché lorsque la rénovation est la principale raison pour laquelle une maison atteint le score d'état le plus élevé dans les données (ici 9,00), comme cela peut arriver. Mettre de côté les trois maisons moyennes, F, L et P, est artificiel, et vise à rendre l'échec exact plutôt qu'approximatif ; le cas approximatif, qui est le plus dangereux, est abordé après le tableau. Considérons les dix biens immobiliers comme l'ensemble des données dont dispose l'expert — un sous-marché dans lequel la rénovation et un état enregistré au plus haut niveau ne se produisent jamais séparément. Par rapport à cette population, la division n'est pas identifiée ; les seize biens dans leur ensemble montrent que l'erreur provenait des données du sous-marché, et non du marché.

Au sein de ce sous-échantillon, l'état enregistré et la rénovation sont liés par une identité exacte : $\text{cond} = 7 + 2 \cdot \text{remodel}$ pour chacune des dix propriétés. À l'ajustement de la spécification

complète — surface habitable brute (GLA), terrain, vue, état, travaux de rénovation —, le logiciel doit supprimer l’une des deux colonnes colinéaires, et c’est celle qu’il supprime qui détermine ce qui s’affiche. Les deux cas de suppression sont présentés :

	Sous-échantillon (10)		Échantillon complet (16)
	projet de rénovation abandonné	condition supprimée	
Interception	−240,000	110,000	−30,000
GLA	600	600	600
Lot	25	25	25
Afficher	100,000	100,000	100,000
Condition (par point)	50,000	<i>laissé tomber</i>	20,000
Rénovation	<i>laissé tomber</i>	100,000	60,000

Trois éléments de ce tableau méritent une attention particulière.

Tout d’abord, ni le coefficient d’état ni le coefficient de rénovation ne sont simplement imprécis ; celui qui est supprimé est absent, et l’autre prend une valeur qui dépend uniquement de cette suppression. La matrice de conception comporte six colonnes (une colonne de 1 pour l’interception, la surface brute, le lot, la vue, l’état et la rénovation), mais seules cinq d’entre elles contiennent des informations distinctes (rang 5), et la méthode des moindres carrés ne fournit pas de réponse unique : chaque paire $(\beta_{\text{cond}}, \beta_{\text{remodel}})$ sur la droite $\beta_{\text{cond}} + \frac{1}{2} \beta_{\text{remodel}} = 50,000$, chacune avec sa propre interception, assure un ajustement aussi bien et exactement de la même manière aux dix ventes. Les deux colonnes de sous-échantillons du tableau correspondent à deux points sur cette droite. Les ordonnées à l’origine diffèrent pour la même raison : chacune correspond au prix ajusté d’une maison sans surface habitable, sans terrain et sans vue — avec un score de condition de zéro dans un ajustement, et non rénovée dans l’autre — ce qui ne correspond à aucune maison et ne constitue pas non plus une réalité du marché.

Deuxièmement, ce qui ne dépend pas de la baisse, c’est la différence de prix entre une excellente maison rénovée et une bonne maison qui ne l’a pas été : \$100,000, qui se lit \$50,000 pour un point sur deux points dans une colonne, et \$100,000 pour la rénovation dans l’autre, et dans les deux cas, précisément $2 \cdot 20,000 + 60,000$ — les deux points propres à l’état enregistré plus tout ce que la rénovation a apporté. Un lecteur qui compare les \$50,000 aux \$20,000 du marché et qualifie cela d’exagération a interprété une manière arbitraire de présenter le résultat (celle où la rénovation n’est pas prise en compte) comme s’il s’agissait d’un taux ; l’en-tête de la colonne indique clairement de quoi il s’agit. Ces dix ventes fixent, au dollar près, la différence de prix entre une excellente maison rénovée et une bonne maison non rénovée, sans rien dire sur la manière dont cette différence se répartit. Notez qu’il ne s’agit pas d’un biais au sens de l’entrée 008 (*biais de variable omise*). L’\$e 100,000 est tout à fait exacte quant à la question à laquelle le sous-échantillon peut répondre ; il s’agit simplement d’une question différente.

Troisièmement, les valeurs de GLA, lot et view sont récupérées à leurs valeurs d’origine — 600, 25 et 100,000 — inchangées par rapport à l’échantillon complet. L’identification s’effectue paramètre par paramètre. Un modèle peut identifier la plupart de ses coefficients et échouer sur trois d’entre eux — ici « condition », « rénovation » et, avec eux, l’ordonnée à l’origine, dont la valeur affichée dépend uniquement de la colonne parmi les deux colonnes colinéaires que le logiciel élimine — et cet échec est invisible dans les coefficients qui ont été identifiés avec succès. Ce que les dix biens déterminent, c’est la différence de prix combinée de deux points de « condition » et de « remodel » enregistrés : \$100,000.

Le contraste avec l’entrée n° 008 (*biais de variable omise*) est justement l’objet de cette entrée. Dans ce cas-là, les informations nécessaires étaient présentes dans les données, mais le modèle n’a

pas su les exploiter : l'ajout de l'indicateur « remodel » a résolu le problème. Ici, l'inclusion de l'indicateur de rénovation ne change rien, car cette information est totalement absente du sous-échantillon . Même spécification, même estimateur, même logiciel... et un échec d'un tout autre ordre, qui ne peut être résolu que par des biens hors de l' $\text{cond} = 7 + 2 \cdot \text{remodel}$ de la ligne. Les six biens mis de côté procèdent de deux manières différentes. E, une excellente maison jamais rénovée, ainsi que H et J, des maisons rénovées en bon état, séparent chacune directement les deux colonnes : chaque contraste identifie la distinction à partir des ventes elles-mêmes, en se basant uniquement sur la forme du modèle, sans aucune hypothèse concernant les taux en dehors de la plage des données. F, L et P, les maisons moyennes, rompent également cette identité, mais elles ne séparent les colonnes qu'à travers une hypothèse : celle selon laquelle le taux de condition compris entre 5,00 et 7,00, qu'elles fournissent, s'applique également entre 7,00 et 9,00, où se situent toutes les maisons rénovées. Si l'on ne remet que F, L et P, le logiciel renvoie à nouveau deux chiffres — ici, les valeurs générées exactement, car les données ne comportent aucun bruit — mais tous deux reposent sur trois ventes et sur cette hypothèse. Cela s'apparente étroitement au cas limite que l'entrée qualifie de « dangereux » : formellement identifié, mais reposant sur les trois ventes qui rompent la tendance et sur l'hypothèse selon laquelle le taux compris entre 5,00 et 7,00 s'applique également entre 7,00 et 9,00. En présence d'un bruit réel, les deux chiffres varieraient ensemble le long de la ligne $\beta_{\text{cond}} + \frac{1}{2}\beta_{\text{remodel}} = 50,000$. C'est en rajoutant E, H et J que l'on distingue cette division de celle basée uniquement sur les ventes. La correction n'a jamais été statistique.

References.

- Fisher, F. M. (1966). *The Identification Problem in Econometrics*. McGraw-Hill.
- Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1), 1–67.
<https://doi.org/10.1214/aos/1176347963>
- Koopmans, T. C. (1949). Identification Problems in Economic Model Construction. *Econometrica*, 17(2), 125–144.
- Manski, C. F. (1995). *Identification Problems in the Social Sciences*. Harvard University Press.

Cross-references: endogeneity; omitted variable bias; conditional expectation.

Addendum: Prolog Formalization

The following machine-readable companion (010-identification.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>.

%% 010-identification.pl
%%
%% Prolog companion to Foundations Vol. 1, Issue 2, Entry 010
%% Title: Parameter Identification
%% Author: Bert Craytor
```

```

%% Concept: identification
%% Version: 1.2.0
%%
%% v1.1.0: condition on the Journal's 0.00-10.00 score (average 5.00, good
%% 7.00, excellent 9.00; 20,000 a point). The knowledge base carries the
%% score and the generating function in those units.
%%
%% The galley reconstructs the entry's identification failure from the
%% dataset itself: not_identified_in_subsample/0 proves the collinearity
%% rather than asserting it.

% 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').

% --- From the dictionary -----
% term(identification, "Identification", econometric_method,
%      "A parameter is identified when it is uniquely determined by the
%      joint distribution of the observable data given the assumptions
%      of the model; identification concerns what the data could reveal
%      in principle and is logically prior to, and separate from,
%      estimation.", published, 'dictionary-2026.1').
%
% depends_on(identification, conditional_expectation).
% depends_on(identification, endogeneity).
% -----

uses_terms([ identification, endogeneity, omitted_variable_bias,
               conditional_expectation, multicollinearity,
               hedonic_regression, comp_selection,
               paired_sales_analysis, adjustment, uspap,
               standards_rule ]).

uses_relations([depends_on/2, related/2, introduced_by/2,
               formalized_by/2, generalized_by/2])).

article_concept(identification).
article_issue(volume(1), number(2), year(2026)).
article_id('foundations.2026.010').

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

local_term(rank_deficiency).
local_term(design_matrix).
local_term(bracketing).
local_term(estimand).

% =====
% THE DISTINCTION
% =====

asks(estimation,      "Given this sample, what is the best guess at the parameter?").
asks(identification, "Would the entire population pin the parameter down at all?").

```

```

prior_to(identification, estimation).

% What does NOT repair an identification failure.
does_not_repair(better_estimator).
does_not_repair(larger_sample).
does_not_repair(more_sophisticated_model).

% =====
% THE SUBSAMPLE ON WHICH REMODEL IS NOT IDENTIFIED
% =====
%
% Condition score from the dataset's three levels: average 5.00, good 7.00,
% excellent 9.00.
cond_score(P, S) :- attribute_value(P, cond_score, S).

% The generating function, from the knowledge base.
true_coefficient_score(K, V) :- true_coefficient(K, V).

% Restrict to good (7.00) or excellent (9.00) properties, in a submarket
% where every excellent house was remodeled and no good one was (the three
% average homes are set aside to keep the illustration exact). Within the
% restriction, cond = 7 + 2*remodel identically, so the two columns are
% perfectly collinear given an intercept.

in_subsample(P) :-
    property(P),
    cond_score(P, C),
    attribute_value(P, remodel_numeric, R),
    ( C =:= 9.0 -> R =:= 1 ; C =:= 7.0, R =:= 0 ).

subsample(Ps) :- findall(P, in_subsample(P), Ps).

%! collinear_in_subsample is semidet.
%
% Proves cond = 7 + 2*remodel at every property of the subsample. This is
% the identification failure, derived from the data rather than asserted.

collinear_in_subsample :-
    forall(in_subsample(P),
        ( cond_score(P, C),
          attribute_value(P, remodel_numeric, R),
          C =:= 7 + 2*R ) ).

%! not_identified_in_subsample(?ParamA, ?ParamB) is semidet.

not_identified_in_subsample(cond_score, remodel_numeric) :-
    collinear_in_subsample.

% The six properties that BREAK the pattern. Two kinds: E, H and J separate
% the columns directly (an excellent house never remodeled; remodeled houses
% at good); F, L and P separate them only under the assumption that the
% condition rate from 5.00 to 7.00 holds from 7.00 to 9.00.
breaks_pattern(P) :- property(P), \+ in_subsample(P).
direct_contrast(P) :- breaks_pattern(P), cond_score(P, C), C >= 7.0.
linearity_contrast(P) :- breaks_pattern(P), cond_score(P, C), C < 7.0.

% =====
% WHAT THE SUBSAMPLE DOES AND DOES NOT DETERMINE
% =====

```

```

% Fitted on the subsample (n = 10) with the remodel column dropped; see
% entry.tex. With the condition column dropped instead: intercept 110,000,
% remodel 100,000, the same three rates. Both lie on
%  $\beta_{\text{cond}} + \beta_{\text{remodel}}/2 = 50,000$ .
subsample_coefficient(intercept,      -240000).
subsample_coefficient(gla_sqft,       600).
subsample_coefficient(lot_sqft,       25).
subsample_coefficient(view_numeric,   100000).
subsample_coefficient(cond_score,     50000). % per point
% remodel_numeric: not estimable (rank deficiency).

not_estimable(remodel_numeric, subsample).
alternative_normalization(condition_dropped, intercept, 110000).
alternative_normalization(condition_dropped, remodel_numeric, 100000).

%! normalizations_agree_on_the_difference is semidet.
% 2 points * 50,000 = 100,000 on the remodel = 2*20,000 + 60,000.
normalizations_agree_on_the_difference :-
    subsample_coefficient(cond_score, S),
    alternative_normalization(condition_dropped, remodel_numeric, M),
    2 * S =:= M,
    true_coefficient_score(cond_score, A), true_coefficient_score(remodel_numeric, B),
    M =:= 2 * A + B.

%! identified_sum_is_correct is semidet.
%
% The subsample's condition coefficient, per point, equals the recorded
% condition's rate plus the remodel's 60,000 spread over the two points
% that separate good from excellent: 20,000 + 60,000/2 = 50,000. Across
% those two points, 100,000 = 2*20,000 + 60,000, the price difference the
% data determine and cannot divide.

identified_sum_is_correct :-
    subsample_coefficient(cond_score, S),
    true_coefficient_score(cond_score, A),
    true_coefficient_score(remodel_numeric, B),
    S =:= A + B / 2,
    2 * S =:= 2 * A + B.

%! identification_is_parameter_by_parameter is semidet.
%
% GLA, lot and view are recovered at their generating values in the
% subsample even though two other parameters are not identified.

identification_is_parameter_by_parameter :-
    forall(member(R, [gla_sqft, lot_sqft, view_numeric]),
        ( subsample_coefficient(R, X), true_coefficient_score(R, X) ) ).

% =====
% CONTRAST WITH ENTRY 008
% =====

failure_kind(omitted_variable_bias,
    "The information is present in the data and the model fails to use it. Including the variable
    ↪ fixes it.").
failure_kind(identification,
    "The information is absent from the data relied upon. Including the variable fixes nothing.").

```

```

remedy(omitted_variable_bias, include_the_variable).
remedy(identification,      change_the_data_or_the_claim).

% =====
% APPRAISAL ENCOUNTERS
% =====

encounter(bracketing,
    "Bracketing answers two questions: contrast varying apart from the other characteristics (is
    ↪ the rate identified at all?) and support (does the subject lie inside the range where it was
    ↪ found?). Bracketing guards support; only such contrast guards identification." ).
encounter(tract_subdivision,
    "Where every view-side home is also the larger plan, view and size cannot be separated within
    ↪ that tract." ).
encounter(age_and_condition,
    "Age, condition and era of construction often move together; separate grid lines assert a
    ↪ separation the data may not support." ).
encounter(time_adjustment,
    "A market-conditions coefficient is identified but composite: it attributes to time everything
    ↪ else that changed over the period, and is not the effect of market conditions alone." ).

% Near-misses are the dangerous case: formally identified, practically not.
dangerous_case(near_collinearity,
    "Two coefficients technically identified but determined by the handful of observations that
    ↪ break the pattern; this is multicollinearity seen from the identification side." ).

% =====
% DEFENSIBILITY
% =====

defensibility_consequence(order_of_questions,
    "Ask what the data can distinguish before asking how best to estimate it; the reverse order
    ↪ produces confident numbers for unmeasurable quantities." ).
defensibility_consequence(report_the_sum,
    "Where two effects are not separately identified, report the combined effect the data do
    ↪ determine and disclose the basis of any split applied." ).
defensibility_consequence(uspap_link,
    "In the Journal's reading, USPAP SR 1-4 (credible results) and the Fannie Mae Selling Guide (
    ↪ market-based adjustments) make the identification question obligatory -- which data identify
    ↪ this adjustment? -- without requiring each grid line to be identified from the comparables in
    ↪ front of the reader." ).
defensibility_consequence(remedy_is_design,
    "Widen the geography, extend the period, or find the properties that break the correlation; no
    ↪ estimator recovers absent information." ).

```