

Identificación de parámetros (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)

Definición formal. Un parámetro se identifica cuando viene determinado de forma única por la distribución conjunta de los datos observables (la frecuencia con la que se da cada combinación de características y precio entre todas esas ventas), dadas las hipótesis y la especificación del modelo. La identificación es lógicamente anterior a la estimación y es distinta de ella.

Planteamiento intuitivo. Hay algunas tasas que las ventas no pueden reflejar en absoluto, y ningún método cambiará eso. Si todas las viviendas reformadas del mercado estuvieran en excelentes condiciones y todas las viviendas en excelentes condiciones hubieran sido reformadas, las ventas reflejarían el valor conjunto de ambas y nada sobre cada una por separado; un informe que indique una tasa para cada una —como es habitual, ya que ambas casi siempre van de la mano— estaría informando una división que el mercado apenas ha marcado. Si todas las viviendas comparables tienen mejores vistas que el inmueble objeto de estudio, la tasa de vistas se está interpretando más allá de los datos disponibles. Si las únicas ventas que muestran un cambio en las condiciones del mercado a lo largo del tiempo son también las únicas que muestran un cambio en los tipos de interés hipotecarios, no es posible distinguir entre ambos factores. Se trata de preguntas sobre el contenido de las ventas, que deben plantearse antes de cualquier ajuste, y la respuesta es la misma tanto si el método ajusta una tasa única a lo largo del rango de una característica como si ajusta tasas diferentes en distintos tramos del mismo, tal y como hace MARS (Multivariate Adaptive Regression Splines, Friedman 1991, el principal método de regresión de la revista): MARS solo puede detectar dónde cambia una tasa cuando las ventas contienen el contraste necesario para mostrarlo. Por lo tanto, la primera pregunta del tasador no es «¿cuál es la tasa?», sino «¿pueden estas ventas decírmelo?».

Para el lector que quiera conocer el mecanismo: . El término estadístico para esto es «identificación»: si la población completa de ventas, medida sin error, permitiría fijar el parámetro en absoluto. El método de los mínimos cuadrados ordinarios hace visible el caso extremo al descartar una de las dos características que varían exactamente al unísono, y oculta el caso peligroso, en el que varían casi al unísono y los dos valores dependen de las mismas pocas ventas. El artículo analiza ambos casos, así como los casos de extrapolación y de efectos temporales confundidos con los que se encuentra a diario la cuadrícula.

Este artículo trata sobre estadística: analiza lo que un conjunto de datos podría mostrar, en principio, antes de que se haya utilizado ningún estimador (un método para calcular las tasas a partir de las ventas, como OLS o MARS), y está dirigido tanto a quienes elaboran conjuntos de datos comparables como a quienes realizan el ajuste de modelos de regresión.

Un *parámetro* es, en este caso, un número que debe proporcionar el modelo —una tasa o un coeficiente— a partir del cual se calculan los ajustes. La cuestión que plantea la estimación es la siguiente: basándome en la muestra, ¿qué valor debo considerar como mi mejor estimación del parámetro? La identificación plantea una cuestión previa: si dispusiera de toda la población —es decir, todas las transacciones que se realizarán jamás en este mercado, medidas sin error—, ¿se determinaría en absoluto el parámetro?

Si la respuesta es no, entonces el parámetro no está identificado y nada de lo que se haga a continuación podrá corregir esto. Un estimador mejorado no supondrá ninguna diferencia, una muestra más amplia del mismo tipo de ventas no servirá de nada, y un modelo más sofisticado tampoco ayudará, a menos que añada una hipótesis que los datos no puedan comprobar (el caso causal que se expone a continuación), y podría incluso ocultar el problema. La falta de identificación de un parámetro se debe a la cuestión en sí y a los datos en los que se basa —lo que generó el mercado y lo que recopiló el tasador—, no al método del analista.

El error de identificación se aprecia más fácilmente cuando dos características siempre van de la mano. Imagina que todos los inmuebles que han sido reformados también obtienen una puntuación de «excelente» (9,00) y que todos los inmuebles con puntuación de «excelente» también han sido reformados; en ese caso, el conjunto de inmuebles reformados y el de inmuebles con calificación de «excelente» son idénticos. El mercado paga un precio más alto por ese grupo: por una vivienda excelente y reformada en lugar de por una vivienda buena que no lo haya sido. Por muchas ventas que observemos, no podemos desglosar esa diferencia de precio entre el estado registrado y lo que haya aportado la reforma que las demás características no reflejen, ya que en este mercado una nunca varía sin la otra. Conocemos el total; no conocemos las partes.

Cuando el método de mínimos cuadrados deja esto claro en el caso de colinealidad exacta (una característica es una función lineal exacta de las demás), se debe a que la matriz de diseño —es decir, la tabla formada por las columnas de los regresores (variables explicativas) que utiliza el estimador— tiene rango deficiente: una de sus columnas es una combinación exacta de las demás. Por lo tanto, el software omite un coeficiente o devuelve un valor en falta. Se trata simplemente de una cuestión de cortesía. Las situaciones peligrosas son los casos límite —aquellos en los que las dos variables varían juntas *casi* siempre—, ya que, en tales casos, los parámetros son técnicamente identificables, el software proporciona dos números y ambos se basan en el reducido número de observaciones que rompen el patrón. Se identifican; pero, al ser tan pocas las ventas que rompen el patrón, resultan tan imprecisas que, en la práctica, los datos no pueden distinguirlas. Ahí es donde un problema de identificación y un problema de precisión se parecen. A esto se reduce la multicolinealidad (características que varían casi exactamente al unísono) desde una perspectiva de identificación.

La identificación puede fallar incluso si no existe colinealidad alguna, cuando el parámetro en cuestión es causal y los datos son observacionales, a menos que se añada una hipótesis que los datos no puedan comprobar. El ejemplo de la entrada 009 (*Endogeneidad*), en el que las mismas dieciséis ventas procedían de un mercado real, es de este tipo: la asociación entre la reforma y el precio puede identificarse —puesto que es una característica de la distribución conjunta y una muestra grande la estimará con precisión— mientras que el efecto causal de la reforma no puede identificarse, porque la distribución observable no contiene información que permita distinguir el valor añadido por la reforma del valor adicional que las viviendas ya tenían y que llevó a sus propietarios a realizar las obras. Para identificar el parámetro causal, es necesario formular una hipótesis que, en sí misma, no puede comprobarse a partir de los datos, por lo que la fiabilidad de la conclusión depende de la credibilidad de dicha hipótesis.

Dónde la encuentra el valorador. El requisito de la «franja de comparación». La norma que exige que se establezca una «franja de comparación» para el inmueble objeto —es decir, que se incluyan elementos comparables tanto por encima como por debajo de él para cada característica importante— se presenta como una regla general y responde a dos cuestiones distintas que conviene diferenciar. La primera es la identificación: ¿presentan estas ventas un contraste en las características, variando con respecto a las demás? Si todos los comparables tienen las mismas vistas, no se puede obtener de ellos ninguna tarifa basada en las vistas. La segunda es el respaldo:

¿se encuentra el inmueble objeto dentro del rango en el que se determinó la tasa? Si todos los comparables tienen mejores vistas que el inmueble objeto, es muy posible que exista una tasa —ya sea a partir de estas ventas o de un conjunto de datos más amplio—, pero aplicarla al inmueble objeto supone una extrapolación más allá de los datos. El uso de intervalos protege el segundo criterio; solo ese contraste protege el primero. Un objeto incluido en un intervalo puede seguir situándose en una zona en la que dos características nunca varían entre sí, y un objeto no incluido en un intervalo puede seguir recibiendo una tarifa identificada en otro lugar, lo cual constituye la solución del «conjunto de datos más amplio» que permite la sección sobre defendibilidad.

Un ejemplo de características que no varían de forma independiente dentro de un submercado: una zona en la que todas las viviendas situadas en el lado de la calle con vistas tienen una superficie útil mayor; las vistas y el tamaño no pueden considerarse factores independientes en esa zona. Un tasador que realiza ajustes por separado para ambos factores está, en la práctica, proporcionando dos cifras que los datos de la zona no pueden distinguir, y la forma en que se reparte entre ellas el importe combinado depende del criterio del tasador o de otro mercado. Puede que eso sea adecuado, pero es una afirmación diferente de «derivado del mercado».

Antigüedad, estado y época de construcción. Estos factores suelen estar estrechamente relacionados. Las viviendas más antiguas suelen estar en peor estado y se construyeron siguiendo normas diferentes. El uso de una cuadrícula de ajustes con líneas separadas para la antigüedad y el estado sugiere que el mercado valora estos factores por separado y que los datos pueden distinguirlos. Es necesario comprobar, y no dar por sentado, si los datos de que se dispone permiten diferenciarlos, y a menudo los datos no superan esa comprobación.

El tiempo y las condiciones del mercado. Aquí se esconden dos fallos distintos. Si la cuadrícula tiene una línea para las condiciones del mercado y ninguna línea para los demás factores que cambiaron durante el periodo —tipos de interés, existencias, un nuevo empleador—, el coeficiente de tiempo queda identificado, pero es compuesto y no representa únicamente el efecto de las condiciones del mercado: todo el cambio se atribuye a la única etiqueta disponible, lo que constituye el caso de variable omitida de la entrada 008 (*sesgo de variable omitida*). Si ambas líneas están presentes, pero las únicas ventas que muestran un cambio en los tipos de interés son las que muestran un cambio en las condiciones de mercado, la distinción entre ambas no se identifica, lo cual es el caso de esta entrada.

Por qué importa para la solidez. Al plantear la pregunta de identificación, debes formularla antes que la pregunta de estimación, ya que el orden es importante. Si un tasador se pregunta primero «¿qué me permiten distinguir mis datos?» y solo después «¿cuál es la mejor forma de estimarlo?», los ajustes obtenidos serán aquellos que los datos puedan respaldar. Por el contrario, si la pregunta de identificación se plantea en último lugar, el resultado serán cifras fiables correspondientes a magnitudes que nadie podría haber medido.

Un parámetro que no se haya identificado debe informarse como tal. Cuando los datos determinan un efecto combinado, pero no su desglose, el enfoque razonable no es omitir los ajustes, sino explicar con sinceridad cómo se han desglosado: a saber, que el grupo comparable no puede separar estos dos efectos, que la división utilizada se basa en un conjunto de datos más amplio o en un criterio de valoración, y cuál es el efecto combinado, ya que es este el que los datos determinan realmente. De este modo, el lector puede juzgar por sí mismo la división. Por el contrario, el lector no puede juzgar dos cifras presentadas como derivadas del mercado cuando, en realidad, solo lo está su suma.

La identificación es el punto de encuentro entre la metodología y las USPAP. Las USPAP son las Normas Uniformes de Práctica Profesional de Tasación. Exigen al tasador que recopile, verifique y analice toda la información necesaria para obtener resultados fiables en el

encargo (Norma 1-4). Las directrices de las entidades crediticias se refieren directamente a los ajustes: la Guía de Venta de Fannie Mae exige que los ajustes reflejen la reacción del mercado ante las diferencias entre inmuebles. Ninguna de las normas establece que cada línea de la cuadrícula deba identificarse por separado a partir de los comparables que tiene ante sí el lector; la reacción del mercado puede provenir de ventas emparejadas fuera de la cuadrícula, de una muestra más amplia o de un criterio justificado, la solución indicada anteriormente. Según la interpretación de la revista, lo que hacen las normas es convertir la cuestión de la identificación en obligatoria: un tasador que alegue que los ajustes cuentan con respaldo del mercado debe poder indicar qué datos lo identifican. Planteada de este modo, la cuestión es verificable —se pueden examinar los datos en los que se basa y comprobar si contienen el contraste— sin convertir la prueba matemática de identificación en una prueba de cumplimiento.

La solución está en el diseño, no en el cálculo. Si no se identifica un parámetro, la solución consiste en modificar los datos utilizados —es decir, ampliar el ámbito geográfico, alargar el periodo de tiempo, buscar los inmuebles que rompen el patrón— o bien modificar la hipótesis. Ningún estimador puede recuperar información que los datos no contengan.

Ejemplo de valoración resuelto. Tomemos el conjunto de datos «Pacifica», compuesto por dieciséis inmuebles,, de la entrada 008 (*sesgo de variable omitida*). El «estado» que figura en él es la puntuación de estado de la revista: una estimación del estado de la vivienda en relación con el estado de las viviendas circundantes en su zona de mercado definida. Se basa en el conocimiento que tiene el tasador de esa zona de mercado y no debería diferir sustancialmente del conocimiento de otros tasadores que operan en la misma zona de mercado. La puntuación oscila entre 0,00 y 10,00: el porcentaje de viviendas de la zona de mercado que se encuentran en peor estado, dividido entre diez. El conjunto de datos registra tres niveles: medio (5,00), bueno (7,00) y excelente (9,00), y los datos contruidos pagan \$20,000 por punto. Al igual que en las entradas 008 (*Sesgo de variable omitida*) y 009 (*Endogeneidad*), incluir el estado en la regresión es lo que hace la tasación tradicional, no el método de la revista, la tasación basada en el RCA (entrada 007, *Esperanza condicional*). Restringir el conjunto de datos a los diez inmuebles que siguen un mismo patrón: «bueno» y sin reformar, o «excelente» y reformado: las comparables A, B, C, D, G, I, K, M, N y O. Dos criterios de selección conforman esa submuestra, y cada uno cumple una función diferente. Al descartar E, H y J (la vivienda excelente que nunca se ha reformado y las dos viviendas reformadas calificadas como «buenas») queda un patrón realista: cómo se presenta un mercado cuando la renovación es la razón principal por la que una vivienda alcanza la puntuación de estado más alta de los datos (en este caso, 9,00), tal y como puede ocurrir. Dejar de lado las tres viviendas de nivel medio, F, L y P, es artificial, y se hace para que el fallo sea exacto en lugar de aproximado; el caso aproximado, que es el peligroso, se aborda tras la tabla. Consideremos los diez inmuebles como el conjunto completo de datos del tasador: un submercado en el que la reforma y el mejor estado registrado nunca se dan por separado. En relación con esa población, la división no está identificada; las dieciséis propiedades completas muestran que el error procedía de los datos del submercado, no del mercado.

Dentro de esta submuestra, la condición registrada y la reforma están relacionadas mediante una identidad exacta: $\text{cond} = 7 + 2 \cdot \text{remodel}$ en cada una de las diez propiedades. Al realizar el ajuste de la especificación completa —superficie habitable bruta (GLA), parcela, vistas, estado, reforma—, el software debe eliminar una de las dos columnas colineales, y la que elimine determinará lo que se imprima. Se muestran ambas opciones:

	Submuestra (10)		Muestra completa (16)
	se ha cancelado la reforma	condición eliminada	
Interceptación	-240,000	110,000	-30,000
GLA	600	600	600
Lote	25	25	25
Ver	100,000	100,000	100,000
Condición (por punto)	50,000	<i>se cayó</i>	20,000
Reforma	<i>se cayó</i>	100,000	60,000

Hay tres aspectos de esta tabla que merecen especial atención.

En primer lugar, ni el coeficiente de condición ni el coeficiente de remodelación son simplemente imprecisos; el que se omite está ausente, y el otro toma un valor que depende únicamente de la omisión. La matriz de diseño tiene seis columnas (una columna de unos para la intersección, la superficie bruta de venta, la parcela, las vistas, el estado y la reforma), pero solo cinco contienen información independiente (rango 5), y el método de los mínimos cuadrados no ofrece una única respuesta: cada par $(\beta_{\text{cond}}, \beta_{\text{remodel}})$ en la recta $\beta_{\text{cond}} + \frac{1}{2} \beta_{\text{remodel}} = 50,000$, cada uno con su propia intersección, tiene un ajuste igual y exactamente igual a las diez ventas. Las dos columnas de submuestras de la tabla son dos puntos en esa recta. Las intersecciones difieren por la misma razón: cada una es el precio ajustado de una vivienda sin superficie habitable, sin parcela y sin vistas —con una puntuación de estado de cero en un ajuste, y sin reformar en el otro—, lo cual no corresponde a ninguna vivienda real ni constituye un hecho del mercado.

En segundo lugar, lo que no depende de la caída es la diferencia de precio entre una casa excelente que ha sido reformada y una buena que no lo ha sido: \$100,000, que se lee como \$50,000 un punto a lo largo de dos puntos en una columna y como \$100,000 en la reforma en la otra, y en ambos casos exactamente $2 \cdot 20,000 + 60,000$ — los dos puntos propios de la condición registrada más todo lo que la reforma ha aportado. Un lector que compare el \$50,000 con el \$20,000 del mercado y lo califique de exageración ha interpretado una forma arbitraria de presentar el resultado (en la que se omite la reforma) como si fuera una tasa; el encabezado de la columna indica lo que es. Las diez ventas fijan, al dólar, la diferencia de precio entre una casa excelente y reformada y una buena que no lo ha sido, y no dicen nada sobre cómo se distribuye esa diferencia. Téngase en cuenta que esto no es un sesgo en el sentido de la entrada 008 (*sesgo de variable omitida*). La «\$100,000» es totalmente correcta en cuanto a la pregunta que la submuestra puede responder; simplemente se trata de una pregunta diferente.

En tercer lugar, GLA, «lot» y «view» se recuperan con sus valores de generación — 600, 25 y 100,000 — sin cambios respecto a la muestra completa. La identificación se realiza parámetro por parámetro. Un modelo puede identificar la mayoría de sus coeficientes y fallar en tres —en este caso, «condición», «remodelación» y, con ellos, la intersección, cuyo valor impreso depende únicamente de cuál de las dos columnas colineales el software descarta—; y el fallo es invisible en los coeficientes que se han identificado correctamente. Lo que sí determinan las diez propiedades es la diferencia de precio combinada de dos puntos de «condición» y «remodelación» registrados: \$100,000.

El contraste con la entrada 008 (*sesgo de variable omitida*) es el punto central de esta entrada. En aquel caso, la información necesaria estaba presente en los datos y el modelo no supo aprovecharla: la inclusión del indicador «remodel» lo solucionó todo. En este caso, incluir el indicador de reforma no cambia nada, porque la información no se encuentra en la submuestra en absoluto. Misma especificación, mismo estimador, mismo software... y un fallo de un tipo totalmente diferente, que solo puede subsanarse mediante inmuebles fuera de la línea $\text{cond} = 7 + 2 \cdot \text{remodel}$. Los seis inmuebles apartados lo hacen de dos maneras diferentes. E, una casa excelente que nunca se ha reformado,

y H y J, casas reformadas en buen estado, separan cada una de las dos columnas directamente: cualquiera de estos contrastes identifica la división a partir de las propias ventas, dada únicamente la forma del modelo, sin ningún supuesto sobre tasas fuera del rango de los datos. F, L y P, las viviendas medias, también rompen la identidad, pero separan las columnas únicamente mediante una suposición: que la tasa de condiciones entre 5,00 y 7,00, que ellas aportan, también se mantiene entre 7,00 y 9,00, donde se sitúan todas las viviendas reformadas. Si solo se vuelven a introducir F, L y P, el software vuelve a arrojar dos cifras —en este caso, los valores generados exactamente, ya que los datos no contienen ruido—, pero ambas se basan en tres ventas y en esa suposición. Esto se asemeja mucho al caso cercano que la entrada calificó de peligroso: identificado formalmente, pero basado en las tres ventas que rompen el patrón y en la suposición de que la tasa entre 5,00 y 7,00 se mantiene entre 7,00 y 9,00. Con ruido real, los dos números oscilarían juntos a lo largo de la línea $\beta_{\text{cond}} + \frac{1}{2}\beta_{\text{remodel}} = 50,000$. Añadir de nuevo E, H y J es lo que permite distinguir la separación de las ventas por sí solas. La corrección nunca fue estadística.

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 Scryer. 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')).
% 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)).

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_cond + beta_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.>").

```