

Identificação de parâmetros (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)

Definição formal. Um parâmetro é identificado quando é determinado de forma única pela distribuição conjunta dos dados observáveis (a frequência com que cada combinação de características e preço ocorre entre todas essas vendas), dadas as suposições e a especificação do modelo. A identificação é logicamente anterior à estimativa e distinta dela.

Enquadramento intuitivo. Existem algumas taxas que as vendas simplesmente não conseguem indicar, e nenhum método vai mudar isso. Se todas as casas reformadas no mercado estiverem em excelente estado e todas as casas em excelente estado tiverem sido reformadas, as vendas mostram quanto as duas juntas valem e nada sobre cada uma isoladamente; uma regressão que apresenta uma taxa para cada uma delas — como se faz quando as duas quase sempre andam juntas — está relatando uma divisão que o mercado mal conseguiu estabelecer. Se todos os imóveis comparáveis tiverem uma vista melhor do que o imóvel objeto da avaliação, a taxa de vista está sendo interpretada além do que os dados permitem. Se as únicas vendas que mostram uma mudança nas condições de mercado ao longo do tempo forem também as únicas que mostram uma mudança nas taxas de juros hipotecárias, não é possível distinguir uma da outra. Essas são questões sobre o que as vendas contêm, feitas antes de qualquer ajuste, e a resposta é a mesma, quer o método ajuste uma taxa em toda a faixa de uma característica, quer taxas diferentes em trechos distintos dela, como faz o MARS (Splines de Regressão Adaptativa Multivariada, Friedman 1991, o principal método de regressão da Revista): O MARS só consegue identificar onde uma taxa muda quando as vendas contêm o contraste necessário para demonstrá-la. A primeira pergunta do avaliador, portanto, não é “qual é a taxa”, mas “estas vendas podem me dizer isso?”.

Para o leitor que deseja entender o mecanismo. O termo estatístico para isso é identificação: se toda a população de vendas, medida sem erros, determinaria o parâmetro de fato. O método dos mínimos quadrados ordinários torna visível o caso extremo ao descartar uma das duas características que se movem exatamente em sincronia, e oculta o caso perigoso, em que elas se movem quase em sincronia e dois números acabam dependendo das mesmas poucas vendas. O artigo aborda ambos os casos, bem como os casos de extrapolação e de efeitos temporais confundidos com os quais a grade se depara diariamente.

Este artigo trata de estatística: ele analisa o que um conjunto de dados poderia revelar, em princípio, antes da aplicação de qualquer estimador (um método de cálculo das taxas a partir das vendas, como OLS ou MARS), e destina-se tanto a quem está compilando conjuntos de dados comparáveis quanto a quem está realizando o ajuste de modelos de regressão.

Um *parâmetro* aqui é um número que o modelo deve fornecer — uma taxa ou coeficiente, a partir do qual se calcula um ajuste. A questão levantada pela estimativa é a seguinte: com base na amostra, o que devo considerar como minha melhor estimativa do parâmetro? A identificação levanta uma questão prévia: se eu tivesse toda a população — ou seja, todas as transações que algum dia ocorrerão nesse mercado, medidas sem erro —, o parâmetro seria determinado?

Se a resposta for não, então o parâmetro não é identificado e nada do que for feito posteriormente poderá corrigir isso. Um estimador aprimorado não fará diferença, uma amostra maior do mesmo tipo de vendas não será útil e um modelo mais sofisticado também não ajudará, a menos que

acrescente uma suposição que os dados não possam testar (o caso causal abaixo), e pode até mesmo ocultar o problema. A falha em identificar um parâmetro está relacionada à questão e aos dados utilizados — o que o mercado produziu e o que o avaliador coletou —, e não ao método do analista.

A falha de identificação fica mais fácil de perceber quando duas características sempre ocorrem juntas. Imagine que todos os imóveis que passaram por reforma também recebam a nota “excelente” (9,00) e que todos os imóveis com nota “excelente” também tenham passado por reforma; nesse caso, os conjuntos de imóveis reformados e aqueles com nota “excelente” são idênticos. O mercado paga um preço mais alto por esse grupo — por uma casa excelente e reformada em vez de uma casa boa que não foi reformada. Não importa quantas vendas observemos, não podemos separar essa diferença de preço entre a condição registrada e o que quer que a reforma tenha trazido que as outras características não registram, porque, nesse mercado, uma nunca varia sem a outra. Sabemos o total; não sabemos as partes.

Quando o método dos mínimos quadrados deixa isso claro no caso de colinearidade exata (uma característica é uma função linear exata das outras), é porque a matriz de projeto — ou seja, a tabela composta pelas colunas dos regressores (variáveis explicativas) que o estimador utiliza — é de ra deficiente: uma de suas colunas é uma combinação exata das demais. O software, portanto, descarta um coeficiente ou retorna um valor ausente. Trata-se apenas de uma espécie de cortesia. As situações perigosas são os casos próximos — aqueles em que as duas variáveis variam juntas *quase* sempre —, pois, nesses casos, os parâmetros são tecnicamente identificáveis; o software fornece dois números, e ambos se baseiam no pequeno número de observações que quebram o padrão. Eles são identificados; mas, com apenas algumas vendas quebrando o padrão, são tão imprecisos que, na prática, os dados não conseguem diferenciá-los. É aí que um problema de identificação e um problema de precisão se assemelham. É a isso que se resume a multicolinearidade (características que se movem quase exatamente juntas) do ponto de vista da identificação.

A identificação pode falhar mesmo na ausência total de colinearidade, quando o parâmetro em questão é causal e os dados são observacionais, a menos que seja adicionada uma suposição que os dados não possam testar. O exemplo da entrada 009 (*Endogeneidade*), no qual as mesmas dezesseis vendas provêm de um mercado real, é desse tipo: a associação entre reforma e preço pode ser identificada — já que é uma característica da distribuição conjunta e uma amostra grande a estimará com precisão — enquanto o efeito causal da reforma não pode, porque a distribuição observável não contém informações que permitam distinguir o valor agregado pela reforma do valor extra que as casas já possuíam e que levou seus proprietários a realizar a obra. Para identificar o parâmetro causal, é necessário fazer uma suposição que, por si só, não pode ser testada com base nos dados, e a confiabilidade da conclusão depende, portanto, da credibilidade dessa suposição.

Onde o avaliador a encontra. O requisito de abrangência. A regra que exige que o imóvel objeto da avaliação seja abrangido — ou seja, que sejam incluídos imóveis comparáveis tanto acima quanto abaixo dele para cada característica importante — é apresentada como uma regra prática e responde a duas questões diferentes que vale a pena distinguir. A primeira é a identificação: essas vendas apresentam contraste nas características, variando em relação às outras? Se todos os imóveis comparáveis tiverem a mesma vista, não será possível obter uma taxa para as características a partir deles. A segunda é o suporte: o imóvel objeto da avaliação está dentro do intervalo no qual a taxa foi determinada? Se todos os imóveis comparáveis tiverem uma vista melhor do que o imóvel objeto da avaliação, pode muito bem existir uma taxa — proveniente dessas vendas ou de um conjunto de dados maior —, mas aplicá-la ao imóvel objeto da avaliação significa extrapolar além dos dados. A delimitação protege o segundo critério; somente esse contraste protege o primeiro. Um imóvel objeto da avaliação ainda pode estar localizado em uma área onde duas características nunca variam entre

si, e um imóvel não objeto da avaliação ainda pode receber uma taxa identificada em outro lugar, o que constitui a solução do “conjunto de dados maior” permitida pela seção de defensabilidade.

Um exemplo de características que não variam independentemente dentro de um submercado: um setor em que cada casa no lado da rua com vista possui uma planta baixa maior; a vista e o tamanho não podem ser tratados como fatores separados nesse setor. Um avaliador que faz ajustes separados para ambos está, na prática, apresentando dois valores que os dados da área não conseguem distinguir, e a forma como o valor combinado é dividido entre eles depende do julgamento do avaliador ou de outro mercado. Isso pode ser apropriado, mas é uma afirmação diferente de “derivado do mercado”.

Idade, estado de conservação e época de construção. Esses fatores costumam estar intimamente ligados. Casas mais antigas tendem a estar em pior estado de conservação e foram construídas de acordo com padrões diferentes. O uso de uma tabela de ajustes com linhas separadas para idade e estado de conservação sugere que o mercado avalia esses fatores separadamente e que os dados podem diferenciá-los. Se os dados disponíveis são capazes de diferenciá-los é algo que deve ser verificado, e não presumido, e os dados muitas vezes não passam nessa verificação.

Tempo e condições de mercado. Há duas falhas diferentes ocultas aqui. Se a tabela tiver uma linha para as condições de mercado e nenhuma linha para os outros fatores que mudaram ao longo do período — taxas de juros, estoque, um novo empregador —, o coeficiente de tempo é identificado, mas é composto e não representa apenas o efeito das condições de mercado: toda a variação é atribuída ao único rótulo disponível, o que corresponde ao caso de variável omitida da entrada 008 (*Viés de Variável Omitida*). Se ambas as linhas estiverem presentes, mas as únicas vendas que apresentam uma variação nas taxas de juros forem aquelas que apresentam uma variação nas condições de mercado, a divisão entre elas não é identificada, o que é o caso desta entrada.

Por que importa para a solidez. Ao fazer a pergunta de identificação, é preciso fazê-la antes da pergunta de estimativa, pois a ordem é importante. Se um avaliador considerar primeiro “o que meus dados me permitem distinguir?” e só depois “qual a melhor forma de estimá-lo?”, os ajustes obtidos serão aqueles que os dados podem sustentar. Por outro lado, se a pergunta de identificação for feita por último, o resultado serão valores confiáveis para quantidades que ninguém poderia ter medido.

Um parâmetro que não seja identificado deve ser relatado como tal. Quando os dados determinam um efeito combinado, mas não sua divisão, a abordagem razoável não é omitir os ajustes, mas sim apresentar uma descrição honesta de como eles foram divididos: ou seja, que o grupo comparável não consegue separar esses dois efeitos, que a divisão utilizada se baseia em um conjunto de dados maior ou em um julgamento, e qual é o efeito combinado, já que é o efeito combinado que os dados realmente determinam. O leitor pode então avaliar a divisão. Em contrapartida, o leitor não pode avaliar dois valores apresentados como derivados do mercado quando, na verdade, apenas a soma deles o é.

A identificação é o ponto de encontro entre a metodologia e a USPAP. USPAP significa Normas Uniformes de Prática Profissional de Avaliação. Ela exige que o avaliador colete, verifique e analise todas as informações necessárias para que os resultados da avaliação sejam confiáveis (Regra 1-4 das Normas). As diretrizes dos credores abordam os ajustes de forma direta: o Guia de Vendas da Fannie Mae exige que os ajustes reflitam a reação do mercado às diferenças entre os imóveis. Nenhuma das regras determina que cada linha da tabela deva ser identificada separadamente a partir dos imóveis comparáveis apresentados ao leitor; a reação do mercado pode provir de vendas emparelhadas fora da tabela, de uma amostra mais ampla ou de um julgamento divulgado — a solução apresentada acima. Na interpretação do Journal, o que as regras fazem é tornar a questão da identificação obrigatória: um avaliador que alega apoio do mercado para os

ajustes deve ser capaz de indicar quais dados o identificam. Enquadrada dessa forma, a questão é verificável — é possível inspecionar os dados utilizados e verificar se eles contêm o contraste — sem transformar o teste matemático de identificação em um teste de conformidade.

A solução está no projeto, não no cálculo. Se um parâmetro não for identificado, a solução é alterar os dados utilizados — ou seja, ampliar o escopo geográfico, prolongar o período de tempo, identificar os imóveis que quebram o padrão — ou modificar a afirmação. Nenhum estimador pode recuperar informações que os dados não contenham.

Exemplo de avaliação resolvido. Considere o conjunto de dados Pacifica, com dezesseis imóveis, disponível em, na entrada 008 (*Viés de Variável Omitida*). O parâmetro “Condição” nesse conjunto de dados corresponde à pontuação de condição do Journal: uma estimativa da condição da casa em relação à condição das residências vizinhas em seu mercado definido. Ela se baseia na compreensão do avaliador sobre esse mercado e não deve entrar em conflito substancial com a compreensão de outros avaliadores que atuam no mesmo mercado. O escore varia de 0,00 a 10,00: a porcentagem de casas no mercado que estão em pior estado, dividida por dez. O conjunto de dados registra três níveis: média (5,00), boa (7,00) e excelente (9,00), e os dados construídos pagam \$20,000 por ponto. Assim como nas entradas 008 (*Viés de Variável Omitida*) e 009 (*Endogeneidade*), incluir a condição na regressão é o que a avaliação tradicional faz, e não o método da Revista, a avaliação baseada em RCA (entrada 007, *Expectativa Condicional*). Restrinja o conjunto de dados às dez propriedades que se enquadram em um único padrão — em bom estado e não reformadas, ou em excelente estado e reformadas: os imóveis comparáveis A, B, C, D, G, I, K, M, N e O. Dois critérios de seleção formam essa subamostra, e eles têm funções diferentes. Excluir E, H e J (a casa excelente que nunca foi reformada e as duas casas reformadas classificadas como boas) deixa um padrão realista: como é um mercado quando a reforma é a principal razão pela qual uma casa atinge o maior escore de condição nos dados (aqui, 9,00), como pode acontecer. Deixar de lado as três casas medianas, F, L e P, é artificial e é feito para tornar a falha exata, em vez de aproximada; o caso aproximado, que é o perigoso, é abordado após a tabela. Considere os dezimóveis como o conjunto completo de dados do avaliador — um submercado no qual a reforma e a melhor condição registrada nunca ocorrem separadamente. Em relação a essa população, a divisão não é identificada; as dezesseis propriedades completas mostram que a falha veio dos dados do submercado, não do mercado.

Nessa subamostra, a condição registrada e a reforma estão relacionadas por uma identidade exata: $\text{cond} = 7 + 2 \cdot \text{reforma}$ em cada uma das dez propriedades. Ao realizar o ajuste da especificação completa — área habitável bruta (GLA), lote, vista, condição, reforma —, o software deve descartar uma das duas colunas colineares, e a escolha de qual delas será descartada determina o que será impresso. Ambos os resultados são mostrados:

	Subamostra (10)		Amostra completa (16)
	a reforma foi cancelada	condição retirada	
Interceptação	−240,000	110,000	−30,000
GLA	600	600	600
Lote	25	25	25
Visualizar	100,000	100,000	100,000
Condição (por ponto)	50,000	<i>deixou cair</i>	20,000
Reforma	<i>deixou cair</i>	100,000	60,000

Três aspectos desta tabela merecem atenção.

Em primeiro lugar, nem o coeficiente de condição nem o coeficiente de reforma são meramente imprecisos; qualquer que seja o que for omitido, ele fica ausente, e o outro assume um valor que

depende apenas dessa omissão. A matriz de projeto tem seis colunas (uma coluna de uns para a interceptação, GLA, lote, vista, condição e reforma), mas apenas cinco que contêm informações distintas (rajo 5) e o método dos mínimos quadrados não oferece uma resposta única: cada par $(\beta_{\text{cond}}, \beta_{\text{remodel}})$ na reta $\beta_{\text{cond}} + \frac{1}{2} \beta_{\text{remodel}} = 50,000$, cada um com sua própria interceptação, realiza o ajuste às dez vendas de maneira igualmente e exatamente adequada. As duas colunas de subamostras da tabela são dois pontos nessa reta. As interceptações diferem pela mesma razão: cada uma é o preço ajustado de uma casa sem área habitável, sem terreno e sem vista — com um escore de condição igual a zero em um ajuste e sem reforma no outro —, o que nenhuma casa é, e nem isso é um fato de mercado.

Em segundo lugar, o que não depende da queda é a diferença de preço entre uma casa excelente, reformada, e uma boa, que não foi reformada: \$100,000, que se lê como \$50,000 por ponto em dois pontos de uma coluna e como \$100,000 na reforma na outra; e, em ambos os casos, precisamente $2 \cdot 20,000 + 60,000$ — os dois pontos da condição registrada mais tudo o que a reforma trouxe consigo. Um leitor que coloca o \$50,000 ao lado do \$20,000 do mercado e considera isso um exagero leu uma forma arbitrária de apresentar o resultado (aquela em que a reforma foi omitida) como se fosse uma taxa; o título da coluna indica o que realmente é. As dez vendas determinam, ao dólar, a diferença de preço entre uma casa excelente, reformada, e uma boa que não foi reformada, e não dizem nada sobre como essa diferença se distribui. Observe que isso não é viés no sentido da entrada 008 (*Viés de Variável Omitida*). O \$100,000 está absolutamente correto quanto à pergunta que a subamostra pode responder; trata-se simplesmente de uma pergunta diferente.

Em terceiro lugar, GLA, lote e vista são recuperados com seus valores de geração — 600, 25 e 100,000 — inalterados em relação à amostra completa. A identificação é feita parâmetro por parâmetro. Um modelo pode identificar a maioria de seus coeficientes e falhar em três — neste caso, condição, reforma e, junto com eles, a interceptação, cujo valor exibido depende apenas de qual das duas colunas colineares o software descarta — e a falha é invisível nos coeficientes que foram bem-sucedidos. O que os dez imóveis determinam é a diferença combinada de preço entre dois pontos registrados de “condição” e “remodelação”: \$100,000.

O contraste com a entrada 008 (*Viés de Variável Omitida*) é o ponto central desta entrada. Nesse caso, as informações necessárias estavam presentes nos dados e o modelo não conseguiu utilizá-las: a inclusão do indicador de remodelação resolveu tudo. Aqui, incluir o indicador de reforma não muda nada, porque a informação simplesmente não está presente na subamostra. Mesma especificação, mesmo estimador, mesmo software — e uma falha de um tipo totalmente diferente, que só pode ser corrigida por imóveis fora da linha $\text{cond} = 7 + 2 \cdot \text{remodel}$. As seis propriedades separadas fazem isso de duas maneiras diferentes. E, uma excelente casa que nunca foi reformada, e H e J, casas reformadas em bom estado, separam as duas colunas diretamente: cada um desses contrastes identifica a divisão a partir das próprias vendas, considerando apenas a forma do próprio modelo, sem nenhuma suposição sobre taxas fora do intervalo dos dados. F, L e P, como casas médias, também quebram a identidade, mas separam as colunas apenas por meio de uma suposição: que a taxa de condição entre 5,00 e 7,00, que elas fornecem, também se aplica entre 7,00 e 9,00, onde se situam todas as casas reformadas. Coloque de volta apenas F, L e P e o software retorna dois números novamente — aqui, exatamente os valores gerados, pois os dados não contêm ruído —, mas ambos se baseiam em três vendas e nessa suposição. Isso é muito semelhante ao caso próximo que a entrada chamou de perigoso: formalmente identificado, mas baseado nas três vendas que quebram o padrão e na suposição de que a taxa entre 5,00 e 7,00 se mantém entre 7,00 e 9,00. Com ruído real, os dois números oscilariam juntos ao longo da linha $\beta_{\text{cond}} + \frac{1}{2} \beta_{\text{remodel}} = 50,000$. Adicionar E, H e J de volta é o que identifica a diferença em relação às vendas isoladas. A correção nunca foi estatística.

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')).
% 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.").

```