

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)

Formal definition. A parameter is identified when it is uniquely determined by the joint distribution of the observable data (how often each combination of features and price occurs among all such sales), given the model's assumptions and specification. Identification is logically prior to, and distinct from, estimation.

Intuitive framing. Some rates the sales cannot give you at all, and no method will change that. If every remodeled house in the market is in excellent condition and every house in excellent condition was remodeled, the sales show what the two together are worth and nothing about either alone; a regression that reports a rate for each — as one will when the two almost always go together — is reporting a split the market barely made. If every comparable has a better view than the subject, the view rate is being read past the end of the data. If the only sales that show a change in market conditions over time are also the only sales that show a change in mortgage interest rates, the two cannot be told apart. These are questions about what the sales contain, asked before any fitting, and the answer is the same whether the method fits one rate across a feature's range or different rates over different stretches of it, as MARS (Multivariate Adaptive Regression Splines, Friedman 1991, the Journal's principal regression method) does: MARS can find where a rate changes only where the sales contain the contrast to show it. The appraiser's first question is therefore not "what is the rate" but "can these sales tell me".

For the reader who wants the mechanism. The statistical name for this is identification: whether the whole population of sales, measured without error, would fix the parameter at all. Ordinary least squares makes the extreme case visible by dropping one of two features that move exactly together, and hides the dangerous case, where they move almost together and two numbers come back resting on the same few sales. The entry works through both, and through the cases of extrapolation and of confounded time effects that the grid meets daily.

This item is about statistics: it looks at what a set of data might show in principle before any estimator (a method of computing the rates from the sales, such as OLS or MARS) has been used, and it is intended for those people who are putting together comparable datasets as well as for those who are fitting regression models.

A *parameter* here is a number the model is to supply — a rate or coefficient, from which an adjustment is computed. The question raised by estimation is this: based on the sample, what should I take as my best estimate of the parameter? Identification poses a prior question: if I had the whole population — that is, every transaction that will ever take place in this market, measured without error — would the parameter be determined at all?

If the answer is no, then the parameter is not identified and nothing done afterward can correct this. An improved estimator will make no difference, a greater sample of the same kind of sales will be of no help, and a more sophisticated model will not help either, unless it adds an assumption the data cannot test (the causal case below), and it might even conceal the problem. Failure to identify a parameter belongs to the question and to the data relied on — what the market produced and what the appraiser gathered — not to the analyst's method.

Identification failure is simplest to see when two characteristics always occur together. Imagine that every property which has been remodeled is also scored excellent (9.00) and that every property scored excellent has also been remodeled; in that case the sets of properties that are remodeled and those that are excellent are identical. The market pays a higher price for that group — for an excellent, remodeled house over a good house that was not remodeled. No matter how many sales we observe, we cannot divide that price difference between the recorded condition and whatever the remodel brought that the other features do not record, because in this market one never varies without the other. We know the total; we do not know the parts.

When least squares makes this clear in the case of exact collinearity (one feature an exact straight-line function of others), it is because the design matrix — that is, the table made up of the columns of the regressors (explanatory variables) which the estimator uses — is rank deficient: one of its columns is an exact combination of the others. The software therefore drops one coefficient or returns a missing value. This is merely a kind of courtesy. The dangerous situations are the near cases — those in which the two variables vary together *almost* always — since in such cases the parameters are technically identifiable, the software provides two numbers, and both of them are based on the small number of observations that break the pattern. They are identified; but with only a few sales breaking the pattern, they are so imprecise that in practice the data cannot tell them apart. That is where an identification problem and a precision problem look alike. This is what multicollinearity (features that move almost exactly together) amounts to from an identification perspective.

Identification can fail even with no collinearity at all, when the parameter in question is causal and the data are observational, unless an assumption the data cannot test is added. The example from entry 009 (*Endogeneity*) in which the same sixteen sales came from an actual market is of this type: the association between remodeling and price can be identified — since it is a feature of the joint distribution and a large sample will estimate it accurately — while the causal effect of remodeling cannot, because the observable distribution contains no information that allows the value which the renovation added to be distinguished from the extra value the houses already had that led their owners to carry out the work. To identify the causal parameter, it is necessary to make an assumption which cannot itself be tested based on the data, and the reliability of the conclusion therefore depends on the credibility of that assumption.

Where appraisers encounter it. The bracketing requirement. The rule which asks that the subject be bracketed — that is, that comparables both above and below it be included for each important characteristic — is presented as a rule of thumb, and it answers two different questions that are worth keeping apart. The first is identification: do these sales contain contrast in the characteristic, varying apart from the others? If every comparable has the same view, no rate for view can be had from them. The second is support: does the subject lie inside the range over which the rate was found? If every comparable has a better view than the subject, a rate may well exist — from these sales or from a larger dataset — but applying it to the subject extrapolates beyond the data. Bracketing guards the second; only such contrast guards the first. A bracketed subject can still sit in a tract where two characteristics never vary apart, and an unbracketed subject can still receive a rate identified elsewhere, which is the “larger dataset” remedy the defensibility section allows.

One example of characteristics that do not vary independently within a submarket: a tract where each house on the view side of the street has the larger floor plan; view and size cannot be treated as separate factors in that tract. An appraiser who makes separate adjustments for both is effectively giving two figures that the tract data cannot distinguish, and how the combined amount is divided between them rests on the appraiser’s judgment or on another market. That may be

appropriate, but it is a different statement from “market-derived”.

Age, condition, and era of construction. These factors are often closely linked. Older houses tend to be in poorer condition and were built to different standards. Using an adjustment grid with separate lines for age and condition suggests that the market prices these factors separately and that the data can distinguish between them. Whether the data at hand can tell them apart is a thing to check, not to assume, and the data often fail that check.

Time and market conditions. Two different failures hide here. If the grid has a market-conditions line and no line for the other things that changed over the period — interest rates, inventory, a new employer — the time coefficient is identified, but it is composite and is not the effect of market conditions alone: the whole change is loaded on the only label available, which is the omitted-variable case of entry 008 (*Omitted Variable Bias*). If both lines are present but the only sales that show a change in interest rates are the sales that show a change in market conditions, the split between them is not identified, which is this entry’s case.

Why it matters for defensibility. When asking the identification question, you must ask it before the estimation question because the order matters. If an appraiser first considers “what can my data allow me to distinguish?” and only afterward “how should I best estimate it?” then the adjustments obtained will be ones that the data can support. Conversely, if the identification question is asked last, then the result will be confident figures for quantities which no one could have measured.

A parameter which is not identified should be reported as such. When the data determine a combined effect but not its division, the reasonable approach is not to omit the adjustment but rather to give an honest account of how it was divided: namely, that the comparable group cannot separate these two effects, that the split used is based on a larger dataset or on judgment, and what the combined effect is, since it is the combined effect that the data actually determine. The reader can then judge the split. In contrast, the reader cannot judge two figures given as being derived from the market when in fact only their sum is.

Identification is where methodology and USPAP meet. USPAP is the Uniform Standards of Professional Appraisal Practice. It requires the appraiser to collect, verify, and analyze all information necessary for credible assignment results (Standards Rule 1-4). Lenders’ guidelines put the point on adjustments directly: Fannie Mae’s Selling Guide requires adjustments to reflect the market’s reaction to differences between properties. Neither rule says that each grid line must be separately identified from the comparables in front of the reader; the market’s reaction may come from paired sales outside the grid, from a broader extraction, or from a disclosed judgment, the remedy given above. In the Journal’s reading, what the rules do is make the identification question obligatory: an appraiser who claims market support for an adjustment has to be able to say which data identify it. Framed this way the question is checkable — one can inspect the data relied upon and see whether they contain the contrast — without turning the mathematical test of identification into a compliance test.

The solution lies in design, not in computation. If a parameter is not identified, the remedy is to alter the data used — that is, to expand the geographical scope, to lengthen the time period, to find the properties that break the pattern — or to modify the claim. No estimator can recover information the data do not contain.

Worked appraisal example. Take the sixteen-property Pacifica dataset of entry 008 (*Omitted Variable Bias*). Condition in it is the Journal’s condition score: an estimate of the house’s condition in relation to the condition of the surrounding homes in its defined market area. It is based on the appraiser’s understanding of that market area, and should not conflict substantially with the understanding of peer appraisers operating in the same market area. The score runs from

0.00 to 10.00: the percentage of houses in the market area that are in worse condition, divided by ten. The dataset records three levels, average (5.00), good (7.00) and excellent (9.00), and the constructed data pay \$20,000 per point. As in entries 008 (*Omitted Variable Bias*) and 009 (*Endogeneity*), putting condition into the regression is what traditional appraisal does, not the Journal’s method, RCA-based appraisal (entry 007, *Conditional Expectation*). Restrict the dataset to the ten properties that lie on one pattern — good and not remodeled, or excellent and remodeled: comps A, B, C, D, G, I, K, M, N, and O. Two cuts make that subsample, and they do different work. Dropping E, H and J (the excellent house never remodeled and the two remodeled houses at good) leaves a realistic pattern: what a market looks like when renovation is the main reason a house reaches the highest condition score in the data (here 9.00), as can happen. Setting aside the three average homes, F, L and P, is artificial, and is done to make the failure exact rather than near; the near case, which is the dangerous one, is taken up after the table. Treat the ten properties as the appraiser’s whole data world — a submarket in which remodeling and a top recorded condition never occur separately. Relative to that population the split is unidentified; the full sixteen show that the failure came from the submarket’s data, not from the market.

Within this subsample, recorded condition and remodel are related by an exact identity: $\text{cond} = 7 + 2 \cdot \text{remodel}$ at every one of the ten properties. Fitting the full specification — gross living area (GLA), lot, view, condition, remodel — the software must drop one of the two collinear columns, and which one it drops decides what is printed. Both drops are shown:

	Subsample (10)		Full sample (16)
	remodel dropped	condition dropped	
Intercept	−240,000	110,000	−30,000
GLA	600	600	600
Lot	25	25	25
View	100,000	100,000	100,000
Condition (per point)	50,000	<i>dropped</i>	20,000
Remodel	<i>dropped</i>	100,000	60,000

Three things in this table deserve attention.

First, neither the condition coefficient nor the remodel coefficient is merely imprecise; whichever is dropped is absent, and the other takes a value that depends only on the drop. The design matrix has six columns (a column of ones for the intercept, GLA, lot, view, condition and remodel) but only five that carry separate information (rank 5), and least squares has no single answer to give: every pair $(\beta_{\text{cond}}, \beta_{\text{remodel}})$ on the line $\beta_{\text{cond}} + \frac{1}{2} \beta_{\text{remodel}} = 50,000$ each with its own intercept, fits the ten sales equally and exactly well. The two subsample columns of the table are two points on that line. The intercepts differ for the same reason: each is the fitted price of a house with no living area, no lot and no view — at a condition score of zero in one fit, and unremodeled in the other — which no house is, and neither is a market fact.

Second, what does not depend on the drop is the price difference between an excellent, remodeled house and a good one that was not remodeled: \$100,000, read as \$50,000 a point across two points in one column and as \$100,000 on the remodel in the other, and in either case precisely $2 \cdot 20,000 + 60,000$ — the recorded condition’s own two points plus everything the remodel brought with it. A reader who sets the \$50,000 beside the market’s \$20,000 and calls it an overstatement has read one arbitrary way of printing the result (the one with remodel dropped) as if it were a rate; the column head says what it is. The ten sales fix, to the dollar, the price difference between an excellent, remodeled house and a good one that was not remodeled, and say nothing about how that difference divides. Note that this is not bias in the sense of entry 008 (*Omitted Variable Bias*).

The \$100,000 is exactly right about the question the subsample can answer; it is simply a different question.

Third, GLA, lot, and view are recovered at their generating values — 600, 25, and 100,000 — unchanged from the full sample. Identification is parameter-by-parameter. A model can identify most of its coefficients and fail on three — here condition, remodel and, with them, the intercept, whose printed value depends only on which of the two collinear columns the software drops — and the failure is invisible in the coefficients that succeeded. What the ten properties do determine is the combined price difference of two points of recorded condition and the remodel: \$100,000.

The contrast with entry 008 (*Omitted Variable Bias*) is the point of the entry. There, the information needed was present in the data and the model failed to use it: including the remodel indicator fixed everything. Here, including the remodel indicator changes nothing, because the information is not in the subsample at all. Same specification, same estimator, same software — and a failure of an entirely different kind, curable only by properties off the line $\text{cond} = 7 + 2 \cdot \text{remodel}$. The six properties set aside do this in two different ways. E, an excellent house never remodeled, and H and J, remodeled houses at good, each separate the two columns directly: either contrast identifies the split from the sales themselves, given only the model's own form, with no assumption about rates outside the range of the data. F, L and P, the average homes, break the identity too, but they separate the columns only through an assumption: that the condition rate between 5.00 and 7.00, which they supply, also holds between 7.00 and 9.00, where every remodeled house sits. Put only F, L and P back and the software returns two numbers again — here the generating values exactly, because the data carry no noise — but both rest on three sales and on that assumption. That is close kin to the near case the entry called dangerous: formally identified, but resting on the three sales that break the pattern and on the assumption that the rate between 5.00 and 7.00 holds between 7.00 and 9.00. With real noise the two numbers would wander together along the line $\beta_{\text{cond}} + \frac{1}{2}\beta_{\text{remodel}} = 50,000$. Adding E, H and J back is what identifies the split from the sales alone. The fix was never statistical.

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.").

```