

Идентификация параметров (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)

Формальное определение. Параметр считается идентифицированным, если он однозначно определяется совместным распределением наблюдаемых данных (частотой встречаемости каждой комбинации характеристик и цены среди всех подобных продаж) с учётом допущений и спецификации модели. Идентификация логически предшествует оценке и отличается от неё.

Интуитивное пояснение. Есть показатели, которые данные о продажах вообще не могут предоставить, и никакой метод этого не изменит. Если каждый отремонтированный дом на рынке находится в отличном состоянии, а каждый дом в отличном состоянии был отремонтирован, то отчёт о продажах показывает, сколько стоят эти два фактора вместе, но ничего не говорит ни об одном из них в отдельности; регрессионный анализ, который вычисляет показатель для каждого из них — как это обычно делается, когда эти два фактора почти всегда идут рука об руку — отражает различие, которое рынок едва ли заметил. Если каждый аналог имеет лучший вид, чем объект оценки, то показатель «вид» интерполируется за пределы имеющихся данных. Если единственные сделки, демонстрирующие изменение рыночных условий с течением времени, являются также единственными сделками, отражающими изменение процентных ставок по ипотечным кредитам, то эти два фактора невозможно различить. Это вопросы о том, что содержат сделки, задаваемые до любой подгонки, и ответ на них одинаков независимо от того, подходит ли метод к одному показателю в диапазоне характеристики или к разным показателям на разных участках этого диапазона, как это делает MARS (многомерные адаптивные регрессионные сплайны, Фридман, 1991 г., основной метод регрессии журнала): MARS может определить место изменения ставки только в тех случаях, когда данные о продажах содержат контраст, позволяющий это продемонстрировать. Поэтому первый вопрос оценщика звучит не так: «Какова ставка?», а так: «Могут ли эти данные о продажах дать мне ответ?».

Для читателя, интересующегося механизмом. Статистический термин для этого явления — «идентификация»: определяется ли параметр вообще, если бы вся совокупность продаж измерялась без погрешностей. Метод обычных наименьших квадратов выявляет крайний случай, исключая одну из двух характеристик, которые изменяются точно синхронно, и скрывает опасный случай, когда они изменяются почти синхронно, а результаты зависят от тех же нескольких продаж. В данной статье рассматриваются оба этих случая, а также случаи экстраполяции и смешанных временных эффектов, с которыми ежедневно сталкивается аналитик в таблице.

Эта статья посвящена статистике: в ней рассматривается, что в принципе может показать набор данных до применения какого-либо оценщика (метода расчета коэффициентов на основе данных о продажах, такого как OLS или MARS), и она предназначена как для тех, кто формирует аналогичные наборы данных, так и для тех, кто проводит подгонку регрессионных моделей.

Параметр в данном случае представляет собой число, которое должна предоставить модель — ставку или коэффициент, на основе которого рассчитывается корректировка. Вопрос, возникающий при оценке, заключается в следующем: исходя из выборки, какое значение следует принять в качестве наилучшей оценки параметра? Проблема идентификации ставит предшествующий вопрос: если бы у меня была вся совокупность — то есть каждая сделка, которая когда-либо состоится на этом рынке, измеренная без ошибок, — удалось бы вообще определить этот параметр?

Если ответ «нет», то параметр не идентифицирован, и никакие последующие действия не смогут исправить эту ситуацию. Улучшенная оценка ничего не изменит, более обширная выборка аналогичных продаж не поможет, и более сложная модель также не поможет, если только она не добавит допущение, которое данные не могут проверить (случай причинно-следственной связи, описанный ниже), и может даже скрыть проблему. Неспособность идентифицировать параметр связана с самой задачей и с данными, на которые опирается оценщик — с тем, что дал рынок и что собрал оценщик, — а не с методом аналитика.

Ошибка идентификации проще всего заметить, когда две характеристики всегда встречаются вместе. Представьте, что каждому объекту недвижимости, в котором был проведен ремонт, также присваивается балл «отлично» (9,00), и что каждый объект, получивший балл «отлично», также был отремонтирован; в таком случае множества объектов, в которых был проведен ремонт, и объектов, получивших балл «отлично», совпадают. Рынок готов платить более высокую цену за такую группу — за дом в отличном состоянии, прошедший ремонт, по сравнению с хорошим домом, который не ремонтировался. Сколько бы сделок мы ни наблюдали, мы не сможем разделить эту разницу в цене между зарегистрированным состоянием и тем, что привнес ремонт, но не отражено в других характеристиках, потому что на этом рынке одно никогда не варьируется без другого. Мы знаем общую сумму; мы не знаем отдельных составляющих.

Когда метод наименьших квадратов ясно показывает это в случае точной коллинеарности (когда одна переменная является точной линейной функцией остальных), это происходит потому, что матрица дизайна — то есть таблица, состоящая из столбцов регрессоров (объясняющих переменных), которые использует оценщик, — имеет недостаточный ранг: один из её столбцов является точной комбинацией остальных. Поэтому программа исключает один коэффициент или возвращает пропущенное значение. Это всего лишь своего рода любезность. Опасными являются «пограничные» случаи — те, в которых две переменные *почти* всегда изменяются синхронно, — поскольку в таких случаях параметры технически идентифицируемы, программа выдает два числа, и оба они основаны на небольшом количестве наблюдений, выбивающихся из общей закономерности. Они идентифицированы; однако, поскольку отклоняющиеся от закономерности значения продаж составляют лишь небольшую часть, они настолько неточны, что на практике данные не позволяют их различить. Именно в этом случае проблема идентификации и проблема точности выглядят одинаково. Вот к чему сводится мультиколлинеарность (переменные, которые изменяются почти точно синхронно) с точки зрения идентификации.

Идентификация может оказаться невозможной даже при полном отсутствии коллинеарности, если рассматриваемый параметр является причинным, а данные — наблюдательными, за исключением случаев, когда добавляется допущение, которое данные не могут проверить. Пример из записи 009 (*Эндогенность*), в котором те же шестнадцать сделок купли-продажи были взяты с реального рынка, относится к этому типу: связь между ремонтом и ценой можно идентифицировать — поскольку она является характеристикой совместного распределения, и большая выборка позволит точно оценить её — в то время как причинно-следственный эффект ремонта идентифицировать невозможно, поскольку наблюдаемое распределение не содержит информации, позволяющей отличить добавленную ремонтной стоимостью от той дополнительной стоимости, которой дома уже обладали и которая побудила их владельцев провести ремонтные работы. Для идентификации причинно-следственного параметра необходимо сделать допущение, которое само по себе не может быть проверено на основе данных, и поэтому достоверность вывода зависит от обоснованности этого допущения.

Где оценщик с этим сталкивается. Требование о «бракетинге». Правило, согласно которому объект оценки должен быть «обрамлен» — то есть для каждой важной характеристики должны быть включены аналоги как выше, так и ниже него по данной характеристике, — представлено в качестве практического ориентира и дает ответ на два разных вопроса, которые следует рассматривать отдельно. Первый касается идентификации: содержат ли эти сделки контраст по данной характеристике, отличающийся

от остальных? Если все аналоги имеют одинаковый вид из окон, то на их основе невозможно определить ставку с учетом вида. Второй вопрос — подтверждение: находится ли объект оценки в пределах диапазона, по которому был установлен показатель? Если у всех аналогов вид лучше, чем у объекта оценки, показатель вполне может существовать — на основе этих сделок или более обширного набора данных — но его применение к объекту оценки будет означать экстраполяцию за пределы имеющихся данных. Определение диапазона обеспечивает соблюдение второго условия; только такое отличие обеспечивает соблюдение первого. Объект оценки, для которого определен диапазон, все равно может находиться в зоне, где две характеристики никогда не различаются, а объект без диапазона все равно может получить ставку, определенную в другом месте, что является «решением на основе более обширного набора данных», допускаемым разделом об обоснованности.

Один из примеров характеристик, которые не варьируются независимо друг от друга в пределах субрынка: участок, на котором каждый дом, выходящий на сторону улицы с видом, имеет более просторную планировку; вид и размер не могут рассматриваться как отдельные факторы на этом участке. Оценщик, который вносит отдельные корректировки на оба этих фактора, фактически приводит две цифры, которые данные по данному участку не позволяют различить, а то, как совокупная сумма распределяется между ними, зависит от суждения оценщика или от другого рынка. Это может быть уместно, но такое утверждение отличается от формулировки «основанное на рыночных данных».

Возраст, состояние и период постройки. Эти факторы зачастую тесно взаимосвязаны. Старые дома, как правило, находятся в худшем состоянии и были построены по другим стандартам. Использование таблицы корректировок с отдельными строками для возраста и состояния предполагает, что рынок оценивает эти факторы отдельно и что данные позволяют различать их. Можно ли различить их на основе имеющихся данных — это нужно проверять, а не предполагать, и данные часто не выдерживают такой проверки.

Время и рыночные условия. Здесь скрываются два разных вида ошибок. Если в таблице есть строка «рыночные условия», но нет строк для других факторов, изменившихся за этот период — процентные ставки, запасы, новый работодатель — то временный коэффициент определяется, но он является составным и не отражает влияние исключительно рыночных условий: все изменение относится к единственной доступной категории, что является случаем «пропущенной переменной» для записи 008 (*Omitted Variable Bias*). Если обе линии присутствуют, но единственные продажи, отражающие изменение процентных ставок, — это те, которые отражают изменение рыночных условий, то разграничение между ними не идентифицируется, что и является случаем данной записи.

Почему это важно для обоснованности. Задавая идентификационный вопрос, необходимо задать его до вопроса об оценке, поскольку порядок имеет значение. Если оценщик сначала задается вопросом «что позволяют мне различить имеющиеся данные?», а уже потом — «как лучше всего это оценить?», то полученные корректировки будут такими, которые могут быть подтверждены данными. И наоборот, если вопрос об идентификации задается в последнюю очередь, то в результате будут получены достоверные цифры по величинам, которые никто не мог бы измерить.

Параметр, который не идентифицирован, следует указать как таковой. Когда данные позволяют определить совокупный эффект, но не его разбивку, разумным подходом является не опускать корректировку, а честно описать, как была произведена разбивка: а именно: что аналогичная группа не может разделить эти два эффекта, что использованное разделение основано на более обширном наборе данных или на субъективном суждении, а также каков совокупный эффект, поскольку именно совокупный эффект фактически определяется данными. Тогда читатель сможет самостоятельно оценить это разделение. Напротив, читатель не может оценить два показателя, представленных как полученные на основе рыночных данных, если на самом деле рыночными являются только их суммарные значения.

Идентификация — это точка соприкосновения методологии и USPAP. USPAP — это Единые стандарты профессиональной оценочной практики. Они обязывают оценщика собирать, проверять

и анализировать всю информацию, необходимую для получения достоверных результатов оценки (Правило № 1-4 Стандартов). Руководства кредиторов прямо указывают на необходимость корректировок: Руководство по продажам Fannie Mae требует, чтобы корректировки отражали реакцию рынка на различия между объектами недвижимости. Ни одно из правил не гласит, что каждая строка таблицы должна быть отдельно обоснована на основе аналогов, представленных читателю; реакция рынка может быть обусловлена сопоставимыми сделками, не входящими в таблицу, более широким выбором данных или обоснованным суждением, как указано выше. По мнению журнала, правила фактически делают вопрос идентификации обязательным: оценщик, утверждающий, что корректировка подкреплена рынком, должен быть в состоянии указать, какие именно данные ее идентифицируют. В такой формулировке этот вопрос поддается проверке — можно изучить данные, на которые опирается оценщик, и увидеть, содержат ли они это различие, — не превращая математический тест на идентификацию в тест на соответствие требованиям.

Решение кроется в проектировании, а не в вычислениях. Если какой-либо параметр не удастся определить, выход заключается в изменении используемых данных — то есть в расширении географического охвата, удлинении временного интервала, поиске объектов недвижимости, выбивающихся из общей закономерности, — либо в корректировке постановки задачи. Ни одна оценочная функция не способна извлечь информацию, которой нет в данных.

Разобраный пример оценки. Рассмотрим набор данных Rasifica, включающий шестнадцать объектов недвижимости, запись № 008 (*Систематическая ошибка пропущенной переменной*). Показателем состояния в данном случае является балл, присвоенный журналом: это оценка состояния дома по сравнению с состоянием окружающих домов в определённой рыночной зоне. Он основан на понимании оценщиком данной рыночной зоны и не должен существенно противоречить мнению других оценщиков, работающих в той же рыночной зоне. Балл варьируется от 0,00 до 10,00: это процент домов в данной рыночной зоне, находящихся в худшем состоянии, деленный на десять. В наборе данных фиксируются три уровня: среднее (5,00), хорошее (7,00) и отличное (9,00), а в построенных данных за каждый балл выплачивается \$20,000. Как и в разделах 008 (*Смещение пропущенной переменной*) и 009 (*Эндогенность*), включение состояния в регрессию соответствует традиционной оценке, а не методу журнала — оценке на основе RCA (раздел 007, *Условное ожидание*). Ограничим набор данных десятью объектами недвижимости, которые соответствуют одной из двух категорий — «хорошее» и «без ремонта» или «отличное» и «с ремонтом»: объекты A, B, C, D, G, I, K, M, N и O. Эта подвыборка формируется двумя критериями, и они выполняют разные функции. Исключение объектов E, H и J (дом в отличном состоянии, никогда не подвергавшийся реконструкции, и два дома в хорошем состоянии, подвергшиеся реконструкции) оставляет реалистичную картину: как выглядит рынок, когда реконструкция является основной причиной того, что дом достигает наивысшего балла по состоянию в данных (здесь 9,00), как это может произойти. Исключение трёх домов со средним состоянием — F, L и P — является искусственным и делается для того, чтобы неудача была точной, а не приблизительной; приблизительный случай, который является опасным, рассматривается после таблицы. Рассмотрим десять объектов недвижимости как всю совокупность данных оценщика — субрынок, на котором ремонт и максимальный зарегистрированный балл состояния никогда не встречаются отдельно. По отношению к этой совокупности различие не идентифицируется; все шестнадцать объектов показывают, что несоответствие возникло из-за данных субрынка, а не из-за рынка в целом.

В пределах этой подвыборки зарегистрированное состояние и данные о реконструкции связаны точным соотношением: $cond = 7 + 2 \cdot remodel$ для каждого из десяти объектов недвижимости. При подгонке полной спецификации — общая жилая площадь (GLA), участок, вид из окна, состояние, ремонт — программа должна исключить один из двух коллинеарных столбцов, и от того, какой из них будет исключен, зависит, что будет выведено на печать. Показаны оба варианта исключения:

	Подвыборка (10)		Полный набор образцов (16)
	отменена реконструкция	условие снято	
Перехват	–240,000	110,000	–30,000
GLA	600	600	600
Лот	25	25	25
Просмотр	100,000	100,000	100,000
Условие (за каждый балл)	50,000	упал	20,000
Реконструкция	упал	100,000	60,000

В этой таблице следует обратить внимание на три момента.

Во-первых, ни коэффициент состояния, ни коэффициент реконструкции не являются просто неточными; если один из них исключается, он отсутствует, а другой принимает значение, зависящее исключительно от этого исключения. Матрица модели имеет шесть столбцов (столбец единиц для пересечения, GLA, участка, вида, состояния и ремонта), но только пять из них несут отдельную информацию (ранг 5) —, и метод наименьших квадратов не дает единственного ответа: каждая пара $(\beta_{\text{cond}}, \beta_{\text{remodel}})$ на прямой $\beta_{\text{cond}} + \frac{1}{2} \beta_{\text{remodel}} = 50,000$, каждая со своим собственным пересечением, одинаково и точно соответствует десяти сделкам купли-продажи. Два столбца подвыборок в таблице представляют собой две точки на этой прямой. Пересечения с осью абсцисс различаются по той же причине: каждое из них представляет собой «подгоняемую» цену дома без жилой площади, без участка и без вида — при оценке состояния с баллом равным нулю в одной подгонке и без ремонта в другой — каким ни один дом не является, и ни то, ни другое не является рыночным фактом.

Во-вторых, от падения цен не зависит разница в цене между отличным отремонтированным домом и хорошим домом, который не подвергался ремонту: \$100,000, что читается как \$50,000 за один пункт по двум пунктам в одном столбце и как \$100,000 на реконструкцию в другой, и в любом случае именно $2 \cdot 20,000 + 60,000$ — две точки, характерные для зафиксированного состояния, плюс всё, что принесла с собой реконструкция. Читатель, который ставит \$50,000 рядом с рыночным показателем \$20,000 и называет это завышением, прочитал один произвольный способ представления результата (тот, где не учтён ремонт) так, как будто это был показатель; заголовок столбца говорит, что это такое. Эти десять продаж фиксируют, с точностью до доллара, разницу в цене между отличным, отремонтированным домом и хорошим домом, который не был отремонтирован, и ничего не говорят о том, как распределяется эта разница. Обратите внимание, что это не является систематической ошибкой в смысле записи 008 (*Систематическая ошибка пропущенной переменной*). \$100,000 абсолютно верно отражает вопрос, на который может ответить подвыборка; просто это другой вопрос.

В-третьих, значения GLA, lot и view восстанавливаются в тех же значениях, что и в исходной выборке, — 600, 25 и 100,000 — без изменений. Идентификация проводится по каждому параметру в отдельности. Модель может определить большинство своих коэффициентов, но не справиться с тремя — в данном случае это «состояние», «ремонт» и, вместе с ними, пересечение с осью, значение которого зависит только от того, какой из двух коллинеарных столбцов программа отсечет — и эта неудача незаметна в коэффициентах, которые были определены успешно. Десять объектов недвижимости определяют совокупную разницу в цене между двумя значениями «состояние» и «ремонт»: \$100,000.

Суть данной записи заключается в контрасте с записью № 008 (*Систематическая ошибка из-за пропущенной переменной*). В том случае необходимая информация присутствовала в данных, но модель не смогла её использовать: включение индикатора «remodel» решило все проблемы. Здесь добавление индикатора ремонта ничего не меняет, поскольку этой информации в подвыборке вообще нет. Та же спецификация, тот же оценщик, то же программное обеспечение — и сбой совершенно иного рода, устранимый только с помощью объектов недвижимости, не входящих в линию $\text{cond} = 7 + 2 \cdot \text{remodel}$. Шесть выделенных объектов недвижимости решают эту задачу двумя разными способами. Е — превосходный дом, никогда не подвергавшийся реконструкции, а также Н и J —

реконструированные дома в хорошем состоянии — каждый из них напрямую разделяет два столбца: любой из этих контрастов определяет различие на основе самих данных о продажах, исходя исключительно из формы модели, без каких-либо допущений о показателях за пределами диапазона данных. F, L и P — дома среднего уровня — также нарушают тождество, но разделяют столбцы лишь на основе предположения: что показатель состояния в диапазоне от 5,00 до 7,00, который они предоставляют, также справедлив в диапазоне от 7,00 до 9,00, где находятся все отремонтированные дома. Если вернуть только F, L и P, программа снова выдаст два числа — в данном случае именно генерирующие значения, поскольку данные не содержат шума, — но оба они основаны на трех сделках и на этом допущении. Это очень похоже на тот случай, который в записи назван «опасным»: формально идентифицированный, но основанный на трёх сделках, нарушающих закономерность, и на предположении, что коэффициент в диапазоне от 5,00 до 7,00 сохраняется и в диапазоне от 7,00 до 9,00. При наличии реального шума эти два числа будут вместе колебаться вдоль линии $\beta_{\text{cond}} + \frac{1}{2}\beta_{\text{remodel}} = 50,000$. Именно добавление E, H и J позволяет отличить этот случай от ситуации, основанной исключительно на данных продаж. Исправление никогда не было статистическим.

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_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.>").

```