

Смещение из-за пропущенной переменной (Omitted Variable Bias)

Foundations • The Valuation Engineer

William Bert Craytor • October 1, 2026

Version 1.2.0

DOI: [10.67330/460smg10](https://doi.org/10.67330/460smg10)

Формальное определение. Смещение пропущенной переменной — это систематическое искажение, возникающее в оцененном коэффициенте, когда признак, влияющий на цену, не включен в модель, а этот пропущенный признак изменяется синхронно с одним из включенных признаков — изменяется вместе с ним после того, как другие включенные признаки учтены, что и измеряет вспомогательная регрессия (вторая регрессия, в которой исключённая характеристика зависит от включённых). В таком случае влияние исключённой характеристики проникает в коэффициенты включённых характеристик (количество долларов на единицу, которое модель присваивает каждой характеристике).

Интуитивное пояснение. Стоимость может быть неверной, поскольку в ней учтено значение другого параметра. Дом Comp N из выпуска № 1 был продан за \$46,455 больше, чем предсказывала регрессия на основе четырёх зарегистрированных параметров. Причина, указанная в разделе *Латентные переменные* выпуска № 1, заключалась в недавнем ремонте кухни, который, согласно этой записи, добавляет около \$60,000 к стоимости дома. Состояние дома было в списке характеристик, но о ремонте ничего не было зафиксировано, и то, что принёс этот ремонт, отсутствовало в данных. Это значение никуда не исчезло. Оно было учтено через коэффициенты включённых в список характеристик, поскольку эти характеристики, как оказалось, сопутствовали ему в сделках купли-продажи. Чтобы это произошло, должны выполняться два условия: отсутствующее значение должно влиять на цену и должно сопутствовать характеристике, включенной в список. Когда оно имеет значение, но ни с чем не сопутствует, коэффициенты верны, а разброс лишь шире. Это верно независимо от того, присваивается ли характеристике один коэффициент или несколько: MARS (многомерные адаптивные регрессионные сплайны, Friedman 1991), основной метод регрессии журнала, присваивает признаку разные коэффициенты для разных интервалов его значений, и в этом случае заимствованное значение распределяется по этим интервалам, а не сосредоточено в одном числе, но оно присутствует по той же причине.

Для читателя, интересующегося механизмом. В обычной регрессии коэффициент — это величина, выражающая, сколько долларов модель присваивает одной дополнительной единице признака, — и это значение имеет точный размер: собственный коэффициент исключённой переменной, умноженный на степень её корреляции с включённой переменной, причём второй фактор представляет собой коэффициент вспомогательной регрессии исключённой переменной по включённым. В статье подробно рассматривается эта формула, относящаяся к линейному случаю, а затем демонстрируется аналогичная утечка в таблице, где парные сделки оценщика влияют на результат точно так же, как и коэффициенты. Статья предназначена для оценщиков, которые получают корректировки на основе рыночных данных.

Что представляет собой пример из этого выпуска, а что — нет. Моделируемый рынок Pacifica, использованный в статьях с № 007 (*Условное ожидание*) по № 010 (*Идентификация параметров*), рассматривает состояние как характеристику с долларовой стоимостью — так же, как это делается в традиционной оценке и как было сделано в выпуске № 1, — и данный пример показывает, в чём заключается проблема такого подхода. Традиционная методология оценки не является методом, используемым в журнале; это оценка на основе RCA — Метод остаточных ограничений из выпуска № 1 (Craytor 2026) — который исключает состояние и качество из регрессии; в статье № 007 (*Условное ожидание*) объясняется, как это делается. Читателю следует извлечь урок из примера, а не его конкретную формулировку.

Состояние, балл состояния и реконструкция. Эти три понятия часто путают. *Состояние* — это состояние дома: степень износа отделочных материалов, возраст инженерных систем, отложенные ремонтные работы. *Балл состояния* — это оценка этого состояния по сравнению с состоянием соседних домов в определённой рыночной зоне. Он основан на понимании оценщиком данной рыночной зоны и не должен существенно расходиться с мнением других оценщиков, работающих в той же зоне. В данном журнале он варьируется от 0,00 до 10,00: это процент домов в рыночной зоне, находящихся в худшем состоянии, деленный на десять. Балл 5,00 означает, что половина домов в рыночной зоне находится в худшем состоянии; 9,00 — что девять из десяти. Этот балл является оценкой, и два дома с одинаковым баллом могут отличаться друг от друга. *Ремонт* — это не состояние и не балл, а событие: работы, выполненные в течение недавнего периода, которые изменили дом. Её стоимость — это стоимость того, что изменилось: состояние и, как правило, качество, материалы и отделочные работы — по ценам, которые рынок за них платит. В большинстве случаев ремонт в большей степени влияет на состояние, чем на качество, в основном из-за затрат на рабочую силу, хотя в некоторых случаях добавленная стоимость за счёт качества оказывается выше. Дом, в котором был проведен ремонт, и дом, в котором ремонт никогда не проводился, но которые схожи по всем текущим характеристикам, влияющим на стоимость, должны стоить одинаково; рынок оценивает состояние и качество, достигнутые благодаря ремонту, а не сам факт проведения ремонта. Таким образом, ремонт не является характеристикой, имеющей собственную цену. Когда набор данных содержит индикатор ремонта наряду с баллом оценки состояния, этот индикатор обозначает всё то, что принёс ремонт, но что не фиксируется другими характеристиками в модели (общая жилая площадь, участок, вид из окна и балл оценки состояния): состояние, выходящее за рамки показанного баллом, и качество, которое в наборе данных не фиксируется. Именно так следует понимать утверждение из *Латентная переменная* (Latent Variable), приведенное в выпуске № 1, о том, что ремонт Comp H «добавляет примерно \$60,000 к стоимости его набора», и именно на этом основан пример, приведенный в данной статье.

Чтобы возникло смещение, должны выполняться одновременно два условия: исключенная переменная должна влиять на цену и должна изменяться синхронно с одной из включенных переменных. Если хотя бы одно из этих условий не выполняется, смещения нет.

Это требование, состоящее из двух частей, имеет важное значение, поскольку общепринятая формулировка правила — что «модель, в которой отсутствует важная переменная, даёт неверные ответы» — в таком виде является ложной. Если признак влияет на цену, но не изменяется синхронно ни с одним из включённых признаков, он увеличивает разброс цен вокруг модели и, следовательно, увеличивает стандартные ошибки (показатель того, насколько оцененный коэффициент может отличаться от значения, которое дала бы очень большая выборка), но не смещает сами коэффициенты; пересечение с осью (постоянная сумма в долларах модели) компенсирует средний вклад исключенной характеристики. В этом случае теряется точность измерения, а не достоверность. С другой стороны, характеристика, которая изменяется вместе с включенными характеристиками, но не влияет на цену, не вносит никакого вклада. Смещение возникает только в том случае, когда выполняются оба условия.

Этот механизм основан на принципе подстановки. Если в наборе регрессоров (переменных, используемых в регрессии для объяснения цены) отсутствует переменная, влияющая на цену, регрессия OLS (метод обычных наименьших квадратов) не оставляет её вклад необъяснённым. Включённые переменные, которые изменяются вместе с отсутствующей, служат её частичной заменой, и их коэффициенты учитывают ту часть эффекта отсутствующей переменной, которую они способны предсказать. Коэффициент включённой переменной больше не отражает только эффект этой самой переменной; он представляет собой этот эффект вместе с долей эффекта отсутствующей переменной, которую он «заимствовал».

Смещение всегда относительно конкретного вопроса. Коэффициенты модели без индикатора «re-model» являются правильным ответом на один вопрос — как цена изменяется в зависимости от бала, полученного по данным этих продаж, причем все факторы, влияющие на этот балл, учитываются — и являются неправильным ответом, когда их интерпретируют как рыночную стоимость в долларах

за один балл оценки состояния при фиксированных остальных параметрах, что и происходит при корректировке. Модель не является неверной из-за наличия другой характеристики; её коэффициенты являются неверными для того применения, к которому они относятся.

Размер заимствованной доли определяется формулой. Для исключённой характеристики w с истинным коэффициентом β_w смещение коэффициента включённой характеристики z_j равняется

$$\text{bias}(\hat{\beta}_j) = \beta_w \cdot \delta_j,$$

, где $\hat{\beta}_j$ — шляпа обозначает оценку — представляет собой коэффициент, оцененный регрессией для z_j , а δ_j — коэффициент z_j во вспомогательной регрессии, в которой w зависит от всех включённых характеристик. Это формула Тейла,¹. Рассчитанная на основе подгоняемых регрессий с использованием оцененного коэффициента исключённой характеристики, она точно выполняется в каждой выборке; с использованием истинного коэффициента она выполняется в среднем.

Форма произведения соответствует линейному случаю. В модели MARS эффект исключённой характеристики поглощается теми участками, которые идут вместе с ней, поэтому смещение распределяется между ними, а не сосредоточено в одном коэффициенте, но оно присутствует, и по той же причине.

Эта формула оказывается более полезной, чем кажется на первый взгляд, поскольку знаки обоих факторов зачастую можно определить на основе знания рынка, даже если ни один из них не поддается измерению. Когда исключённая характеристика приводит к росту цены ($\beta_w > 0$) и сопряжена с более выраженным проявлением какой-либо включённой характеристики при неизменности остальных включённых характеристик ($\delta_j > 0$), коэффициент этой характеристики оказывается завышенным. Оценщик, знающий хотя бы направление обеих зависимостей, знает и направление ошибки. Одно уточнение: « δ_j » — это зависимость при неизменности других включённых характеристик, а при наличии более одной включённой характеристики простая корреляция (число от -1 до $+1$, измеряющее, насколько две характеристики растут и падают вместе, игнорируя все остальные) может иметь противоположное направление. В приведенном ниже примере жилая площадь коррелирует с индикатором ремонта (корреляция $+0.09$), однако её δ отрицательна, и, следовательно, коэффициент слишком низок.

Где оценщик с этим сталкивается. Таблица сравнения продаж может быть представлена в виде аддитивной модели, по форме напоминающей регрессионную — несколько выбранных вручную характеристик с корректировкой в долларах для каждой из них — и когда эти корректировки извлекаются из рыночных данных, она является объектом оценки точно так же, как и любая регрессионная модель. Строки таблицы представляют собой включённые характеристики. Любая характеристика дома, влияющая на его цену, но не являющаяся строкой, остается за пределами модели.

К факторам, которые обычно не учитываются, относятся те аспекты состояния объекта, которые не отражены в баллах состояния, а также качество, если балл качества не зафиксирован; функциональность планировки, естественное освещение, уровень шума, конкретное расположение внутри квартала, мотивация продавца и маркетинговые усилия, затраченные на продвижение объявления. Каждый из этих факторов может повлиять на цену, и многие из них сопутствуют тем факторам, которые учитываются в таблице.

Ремонт является показательным примером, поскольку он фиксируется в документах, но при этом по-прежнему не учитывается в полной мере. В отчёте URAR (Uniform Residential Appraisal Report) указывается, были ли обновлены или отремонтированы кухня и ванные комнаты, а также в течение какого периода; при этом сами определения состояния в наборе данных UAD (Uniform Appraisal Dataset) зависят от степени обновления. Однако в таблице есть строки для «состояния» и «качества», но нет строки для «реконструкции» — и это правильно, когда речь идет о рыночной оценке состояния, а строки «состояние» и «качество» отражают то, что принесла реконструкция, поскольку третья строка

¹Theil, H. (1957). Specification Errors and the Estimation of Economic Relationships. *Revue de l'Institut International de Statistique*, 25(1/3), 41–51.

учитывала бы это дважды. Когда же речь идет о том, как предложенная реконструкция повлияет на цену этого дома, фиксирование состояния и качества устраняет тот самый канал, через который действует реконструкция; этот вопрос рассматривается в пункте 009 (*Эндогенность*). Утечка происходит через то, что упускают из виду баллы. На рынке, где наивысшие балы состояния в основном присущи недавно отремонтированным объектам, а ремонт приносит ценность, которую балы не фиксируют, корректировка по состоянию, извлеченная из этих продаж в традиционной оценке — путем сопоставления пар продаж или с помощью регрессии, включающей состояние — переносит эту ценность вместе с собой. Согласно формуле Тейла такая корректировка оказывается слишком высокой: она включает не только собственный эффект состояния, но и часть того, что не было учтено при подсчете баллов. Оценщик, который принимает эту корректировку и применяет её к дому с максимальным баллом, заслуженным благодаря уходу за ним, скорректирует стоимость слишком сильно. В рамках Метода остаточных ограничений, описанного в журнале, коэффициент состояния не выделяется. Оценщик распределяет остаточную величину каждого аналога между состоянием, качеством и другими характеристиками, которые не измеряются данными, по своему усмотрению, при соблюдении одного ограничения: сумма вкладов остаточных компонентов в стоимость должна равняться остаточной величине аналога. Стоимость реконструкции затем распределяется по усмотрению оценщика, а не отражается одной строкой.

Только что описанная «утечка» представляет собой более точную формулировку возражения, которое высказывают оценщики в отношении корректировок, основанных на рыночных данных, а именно: корректировки, полученные на основе одного набора данных, плохо переносятся на другие наборы. Причина заключается в том, что извлечённая корректировка состоит как из названного эффекта, так и из сопутствующих ему неназванных эффектов, а её состав зависит от того, какие элементы сопутствовали какому в наборе данных, из которого она была взята.

Еще одним примером является ситуация, когда «парные сделки» в традиционной оценке — две сделки, одинаковые во всем, кроме рассматриваемого признака, причем разница в цене интерпретируется как стоимость этого признака — дают неправдоподобно высокую цифру. Причиной может быть не погрешность измерения, а «утечка»: две сделки различаются как по указанной характеристике, так и по чему-то, что с ней связано, и разница в цене обусловлена обоими факторами.

Почему это важно для обоснованности. Отчёт, в котором не рассматривается проблема смещения, связанного с пропущенными переменными, делает его гедонические и табличные корректировки уязвимыми для этой критики.

Смещение невозможно определить по статистике подгонки. Модель может иметь высокий коэффициент объяснения (R^2) (долю вариации цен, которую воспроизводит модель), небольшие остатки (разности между ценами продажи и ценами, предсказанными моделью) и коэффициенты, проходящие обычный статистический тест (t -тест, который проверяет, может ли коэффициент с разумной вероятностью равняться нулю), и при этом иметь сильное смещение, поскольку смещение обусловлено взаимосвязью между признаками, а не величиной остатков. Использование коэффициента R^2 в качестве доказательства надежности модели не решает проблему. В приведенном ниже примере коэффициент R^2 равен 0,97, а коэффициент условия ошибочен на 74 процента.

Большой объем данных не решит проблему. Смещение не уменьшается по мере увеличения выборки. Тысяча продаж с того же рынка, при которых не учитывается та же самая характеристика, даёт оценку, центрированную вокруг того же неверного значения, но с более узким доверительным интервалом: стандартная ошибка уменьшается вокруг смещённой оценки, что хуже, чем широкий интервал вокруг неё, поскольку отчёт будет выглядеть более убедительным. То же самое верно и для MARS: большее количество сделок делает аппроксимирующую поверхность более стабильной, но вокруг того же самого смещенного результата; интервал «*varmod*» () измеряет разброс цен, а не смещение. Это исправляет утверждение о «латентной переменной» из выпуска № 1, в котором говорилось, что смещение уменьшится с увеличением количества аналогов и что остальные коэффициенты

были смещены в сторону завышения. Смещение не уменьшается с увеличением количества продаж, и как в восьми сделках из выпуска № 1, так и в шестнадцати приведенных ниже, коэффициенты «вид», «GLA» и «участок» имеют смещение в сторону занижения; только коэффициент «состояние» имеет смещение в сторону завышения.

Часто можно обосновать направление смещения, даже если его величина неизвестна. В традиционной оценке, где корректировка на состояние извлекается из данных о продажах — путем сопоставления пар продаж или с помощью регрессии, включающей состояние объекта, — практический подход заключается не в том, чтобы утверждать, что корректировка является несмещенной, а в том, чтобы назвать вероятные упущенные характеристики, использовать знание рынка для присвоения каждой из них знака для β_w и для δ_j , а также указать результирующее направление смещения. «Корректировка состояния, вероятно, завышена. Большинство сделок с наивысшим баллом — это недавно отремонтированные объекты, а ремонт привнес улучшения, не отраженные в данных, такие как новые кухни и отделочные материалы. Эти улучшения повышают цену и встречаются преимущественно в сделках с высокой оценкой, поэтому часть их стоимости была учтена как состояние» — это разумное утверждение. Оба требования выполнены, и оба указывают в одном направлении: не учтенные улучшения повышают цену (коэффициент β_w положителен) и сопутствуют сделкам с высокими баллами (коэффициент δ_j положителен), поэтому смещение направлено в сторону увеличения. Такое утверждение превращает неизмеренную проблему в проблему, которая признана и получила определенное направление, что именно и нужно читателю для того, чтобы оценить вывод. В рамках Метода остаточных ограничений коэффициент состояния не извлекается, и вопрос о состоянии не возникает.

Разобраный пример оценки. Расширение набора данных Pacifica, включающее шестнадцать объектов недвижимости, содержит для каждого объекта данные о общей жилой площади (GLA), площади участка, виде из окна, баллах по состоянию, индикаторе проведенного ремонта и цене продажи. Поскольку каждая цена рассчитывалась по известной формуле без случайной составляющей (данные не содержат шума), регрессия, использующая правильные признаки, точно воспроизводит эту формулу, и любое отклонение от неё является систематической ошибкой, полностью лишенной шумовой составляющей.

Comp	GLA (sf)	Участок (кв. футов)	Просмотр	Условия	Отремонтированный	Цена (\$)
A	1,650	9,000	Да	7,00	Нет	1,425,000
B	1,650	7,200	Нет	7,00	Нет	1,280,000
C	1,850	7,500	Нет	7,00	Нет	1,407,500
D	1,450	7,000	Нет	7,00	Нет	1,155,000
E	1,700	8,000	Да	9,00	Нет	1,470,000
F	1,550	7,100	Нет	5,00	Нет	1,177,500
G	1,700	10,500	Нет	7,00	Нет	1,392,500
H	1,600	7,300	Нет	7,00	Да	1,312,500
I	1,750	8,200	Нет	9,00	Да	1,465,000
J	1,600	7,600	Нет	7,00	Да	1,320,000
K	1,900	9,500	Да	9,00	Да	1,687,500
L	1,500	6,800	Нет	5,00	Нет	1,140,000
M	1,800	8,800	Нет	7,00	Нет	1,410,000
N	1,650	7,400	Да	7,00	Нет	1,385,000
O	1,550	7,900	Нет	9,00	Да	1,337,500
P	1,720	9,200	Нет	5,00	Нет	1,332,000

Выпуски А–Н являются аналогами выпуска № 1, при этом их опубликованные характеристики и цены остались без изменений; в выпуске № 1 состояние описывалось словами и кодировалось по шкале 1–3 для регрессионного анализа; по шкале данного журнала слова соотносятся со средним состоянием 5,00, хорошим — 7,00, отличным — 9,00, поэтому один шаг в выпуске № 1 соответствует

двум баллам. Реконструкция объекта Comp H — указанная в выпуске № 1 как источник его остаточной стоимости, но не зарегистрированная в качестве переменной — теперь учтена. Рынок оценивает \$20,000 за каждый балл состояния. В сгенерированных данных отремонтированный дом продаётся на \$60,000 дороже, чем дом с такой же общей арендуемой площадью (GLA), участком, видом и баллом оценки состояния. Эти \$60,000 не являются ценой за сам факт ремонта; они отражают всё то, что принёс ремонт, но что не учитывается другими характеристиками. Дома после ремонта также имеют в среднем более высокие баллы (8,20 против 6,64), поэтому показатель ремонта изменяется вместе с включённой оценкой состояния, что является вторым условием, требуемым для устранения систематической ошибки: $\text{corr}(\text{remodel}, \text{cond}) = 0.55$.

Две регрессии по одним и тем же шестнадцати объектам недвижимости:

	Переработка опущена	Включено в ремонт	Смещение
Перехват	−83,923.57	−30,000	−53,923.57
GLA	586.64	600	−13.36
Лот	24.32	25	−0.68
Просмотр	79,178.46	100,000	−20,821.54
Условие (за каждый балл)	34,815.56	20,000	+14,815.56
Реконструкция	—	60,000	
R^2	0.9733	1.0000	

Показатель «\$60,000» в строке «remodel» не отражает цену реконструкции. В данном наборе данных он представляет собой разницу в цене, связанную с индикатором «remodel» при одинаковых показателях GLA, участка, вида из окна и балла; именно это и «восстанавливает» «генерирующая функция».

Коэффициент состояния слишком высок на \$14,816 пункта — он представляет отчёт о значении \$34,816 на пункт состояния, тогда как рыночный показатель составляет \$20,000, что превышает рыночное значение на 74 процента. Коэффициент view занижен на \$20,822. Обе ошибки незаметны в подгонке: модель без индикатора remodel имеет $R^2 = 0.9733$ и остатки, которые, за исключением трёх продаж вблизи \$35,000, выглядят обычными. Ничто в R^2 или в тестах коэффициентов не указывает на проблемы (если бы эти остатки были шумом, каждый коэффициент прошел бы этот тест на обычном уровне значимости 5 процентов).

Формула Тейла точно описывает эти цифры. Регрессионный анализ индикатора «remodel» по четырём включённым признакам даёт $\delta_{\text{cond}} = 0.24692602$ и $\delta_{\text{view}} = -0.34702564$. Тогда

$$\beta_w \cdot \delta_{\text{cond}} = 60,000 \times 0.24692602 = 14,815.56, \quad \beta_w \cdot \delta_{\text{view}} = 60,000 \times (-0.34702564) = -20,821.54,$$

что точно соответствует наблюдаемым сдвигам до цента. Смещение не объясняется приблизительно этой формулой; оно *и есть* эта формула. Шум изменил бы результат меньше, чем можно было бы ожидать: при наличии шума в данных формула по-прежнему точно выполняется между двумя подгоняемыми регрессиями (без и с индикатором ремоделирования), если использовать оцененный коэффициент индикатора $\hat{\beta}_w$ вместо β_w ; по сравнению с истинным коэффициентом она выполняется в среднем по многим выборкам.

Коэффициент вида заслуживает более пристального внимания, поскольку его смещение направлено в обратную сторону, и причина этого весьма поучительна. При прочих равных условиях объекты недвижимости с хорошим видом в данной выборке *реже* подвергаются реконструкции ($\delta_{\text{view}} < 0$). Таким образом, коэффициент вида смещается вниз, а не вверх: он \$20,822 занижен, в то время как коэффициент состояния \$14,816 завышен. Эти две ошибки являются отдельными последствиями одного и того же упущения, а не результатом их взаимного влияния. Оценщик, руководствующийся

исключительно состоянием объекта, вообще не предвидел бы ошибку в корректировке вида — однако занижение корректировки вида на 21 процент является прямым следствием того же самого упущения.

На практике ущерб сосредоточен именно там, где оценщики вряд ли заметят его. Рассмотрим сначала, как работает эта смещённая модель в качестве *предиктора*. Для объекта оценки без ремонта, идентичного Comp N — 1,650 квадратных футов общей арендуемой площади (GLA) на участке площадью 7,400 квадратных футов с видом и баллом состояния 7,00 — она указывает на \$1,386,914 при истинной средней цене \$1,385,000 для домов с точно такими же характеристиками — по формуле из пяти характеристик из записи 007 (*Условное ожидание*), с учетом ремонта = 0 — погрешность составляет \$1,914, или 0,14 процента. По всем шестнадцати сделкам типичная погрешность составляет около \$21,000, так что модель сильно смещена, но всё же даёт приемлемые прогнозы, что не является совпадением: OLS подгоняется под предоставленную выборку, и смещённые коэффициенты — это именно те, которые наиболее близки к ценам продажи в выборке. Точность прогнозирования не является доказательством отсутствия смещения, связанного с пропущенными переменными, и перекрёстная проверка — тестирование модели на сделках, не включённых в подгонку — также не исключает его: хорошие результаты перекрёстной проверки не означают, что коэффициенты свободны от смещения, поскольку некорректно специфицированная модель может давать точные прогнозы везде, где признаки сочетаются так же, как и при подгонке. Если поочередно исключать каждую характеристику, две наибольшие погрешности — около \$56,000 для K и \$48,000 для G — приходятся на объекты недвижимости, которые соответствуют закономерности данной выборки, в которой высокие баллы состояния в основном присущи домам, прошедшим ремонт; E, который нарушает эту закономерность, занимает третье место с показателем около \$45,000, а H и J, которые также нарушают её, занимают четвертое и пятое места с показателями около \$41,000. Данные здесь не содержат шума, поэтому каждая из этих ошибок обусловлена пропущенным индикатором реконструкции, а не шумом; в реальных сделках купли-продажи, где также присутствует шум, размер ошибки сам по себе не позволяет определить, обусловлена ли она пропущенной характеристикой или обычным разбросом остатков.

Недостаток проявляется в тот момент, когда коэффициенты используются в качестве *корректировок*, а не в качестве правила прогнозирования. Оценщик, использующий корректировку на состояние объекта из этой модели, применяет \$34,816 за каждый балл, тогда как рыночный показатель составляет \$20,000 — на \$29,631 слишком много для каждого аналога, скорректированного по двум пунктам от «хорошего» до «отличного» — а также корректировка вида в размере \$79,178, тогда как рыночный показатель составляет \$100,000. Таблица — это место, где коэффициенты регрессии считываются по одному, и именно это считывание разрушает смещение.

Прогнозирование также даёт сбой — причём серьёзный — в случае объектов недвижимости, которые выбиваются из шаблона, на котором основано данное смещение. Хорошо ухоженный дом, подобный объекту оценки (1,650 квадратных футов общей арендуемой площади, участок площадью 7,400 квадратных футов), но без вида из окна, получивший отличный балл 9,00 и никогда не подвергавшийся реконструкции, указан по адресу \$1,377,367, тогда как фактическая цена составляет \$1,325,000, что на \$52,367 (3,95 процента), поскольку в этих сделках балл 9,00 в основном сопутствовал ремонту, а показатель состояния модели учитывает стоимость ремонта \$60,000 наряду с самим баллом. Цена отремонтированного дома, в остальном идентичного объекту оценки (вид из окна, балл 7,00), указана на сайте \$1,386,914 по сравнению с \$1,445,000, что на \$58,086 (4,02 процента) ниже рыночной стоимости. Это нетипичные объекты недвижимости — те, для оценки которых чаще всего специально привлекают оценщиков, и те, отчёт по которым будет изучаться наиболее внимательно.

References.

- Craytor, W. B. (2026). Residual Constraint Approach (RCA): Framework & Protocol. *The Valuation Engineer*, 1(1). <https://doi.org/10.67330/kgebx46>
- Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1), 1–67. <https://doi.org/10.1214/aos/1176347636>

[//doi.org/10.1214/aos/1176347963](https://doi.org/10.1214/aos/1176347963)

Theil, H. (1957). Specification Errors and the Estimation of Economic Relationships. *Revue de l'Institut International de Statistique*, 25(1/3), 41–51.

Cross-references: conditional expectation; hedonic regression; implicit price; latent variable; endogeneity.

Addendum: Prolog Formalization

The following machine-readable companion (008-omitted-variable-bias.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>.

%% 008-omitted-variable-bias.pl -- companion to Foundations Vol. 1, Issue 2,
%% Entry 008, Omitted Variable Bias (omitted_variable_bias), v1.2.0.
%% Checks: theil_formula_holds/0 (each bias = beta_omitted * delta_j to the
%% cent); remodel_reading_holds/0 (the remodel indicator's 60,000).

% 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').

uses_terms([ omitted_variable_bias, conditional_expectation,
             hedonic_regression, implicit_price, latent_variable,
             multicollinearity, ordinary_least_squares, residual,
             adjustment, comp_selection, paired_sales_analysis,
             sales_comparison_approach, uspap ]).
uses_relations([depends_on/2, related/2, introduced_by/2]).

article_concept(omitted_variable_bias).
article_issue(volume(1), number(2), year(2026)).
article_id('foundations.2026.008').

:- 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)).

kind(condition, state).
kind(condition_score, measurement_of(condition)).
kind(quality, state).
kind(remodel, event).
changes(remodel, condition).
changes(remodel, quality).
```

```

stand_in(remodel_numeric, unrecorded_condition).
stand_in(remodel_numeric, unrecorded_quality).

ovb_condition(affects_outcome,
  "The omitted variable must have a nonzero effect on the outcome.").
ovb_condition(correlated_with_included,
  "The omitted variable must have a nonzero coefficient on at least one included regressor in the
  ↪ auxiliary regression (partial, not merely marginal, correlation).").

biased(BetaOmitted, Delta) :-
  BetaOmitted =\= 0,
  Delta =\= 0.

consequence(affects_but_uncorrelated,
  "Inflates residual variance and widens standard errors; slope coefficients remain unbiased (the
  ↪ intercept absorbs the omitted variable's mean).").
consequence(correlated_but_no_effect,
  "No cost at all.").

theil_bias(BetaOmitted, Delta, Bias) :-
  Bias is BetaOmitted * Delta.

omitted_variable(remodel_numeric).
omitted_variable_effect(remodel_numeric, 60000).

condition_scale(0.0, 10.0).
condition_score(average, 5.0).
condition_score(good, 7.0).
condition_score(excellent, 9.0).
condition_rate(20000).

coefficient(remodel_omitted, intercept, -83923.571).
coefficient(remodel_omitted, gla_sqft, 586.645).
coefficient(remodel_omitted, lot_sqft, 24.323).
coefficient(remodel_omitted, view_numeric, 79178.461).
coefficient(remodel_omitted, cond_score, 34815.561).

coefficient(remodel_included, intercept, -30000).
coefficient(remodel_included, gla_sqft, 600).
coefficient(remodel_included, lot_sqft, 25).
coefficient(remodel_included, view_numeric, 100000).
coefficient(remodel_included, cond_score, 20000).
coefficient(remodel_included, remodel_numeric, 60000).

aux_slope(intercept, -0.89872625).
aux_slope(gla_sqft, -0.00022259).
aux_slope(lot_sqft, -0.00001129).
aux_slope(view_numeric, -0.34702564).
aux_slope(cond_score, 0.24692602).

model_fit(remodel_omitted, r_squared, 0.9733).
model_fit(remodel_included, r_squared, 1.0000).

observed_bias(R, Bias) :-
  coefficient(remodel_omitted, R, Omitted),
  coefficient(remodel_included, R, True),
  Bias is Omitted - True.

theil_formula_holds :-

```

```

omitted_variable_effect(_, BetaOmitted),
forall(( observed_bias(R, Observed), aux_slope(R, Delta) ),
      ( theil_bias(BetaOmitted, Delta, Predicted),
        abs(Observed - Predicted) < 0.01 )).

remodel_reading_holds :-
  omitted_variable_effect(remodel_numeric, Beta),
  coefficient(remodel_included, remodel_numeric, Coef),
  abs(Coef - Beta) < 0.01.

reported_bias(cond_score, 14815.56).
reported_bias(view_numeric, -20821.54).

relative_bias(R, Pct) :-
  observed_bias(R, Bias),
  coefficient(remodel_included, R, True),
  True \= 0,
  Pct is 100 * Bias / True.

use_of_coefficients(prediction,
  "Evaluate the whole fitted surface at a point; biases in individual coefficients can offset.").
use_of_coefficients(adjustment,
  "Read one coefficient at a time as an implicit price; bias enters directly and does not offset
  ↪.").

prediction_check(subject_typical, remodel_omitted, 1386914).
prediction_check(subject_typical, remodel_included, 1385000).

prediction_check(excellent_unremodeled, remodel_omitted, 1377367).
prediction_check(excellent_unremodeled, remodel_included, 1325000).
prediction_check(remodeled_with_view, remodel_omitted, 1386914).
prediction_check(remodeled_with_view, remodel_included, 1445000).

atypical_property_fails_worse(S) :-
  prediction_check(S, remodel_omitted, A),
  prediction_check(S, remodel_included, B),
  abs(A - B) > 50000.

defensibility_consequence(fit_statistics,
  "Bias is undetectable from R-squared or residual size; it lives in the correlation structure,
  ↪ not the residuals.").
defensibility_consequence(sample_size,
  "Bias does not shrink with n; more data gives a tightly estimated wrong number.").
defensibility_consequence(direction_disclosure,
  "Signing beta_omitted and delta_j from market knowledge yields a defensible directional
  ↪ statement even when magnitude is unknown.").

practice_encounter(what_the_scores_miss,
  "Traditional appraisal: a condition adjustment extracted where top scores go with remodels
  ↪ carries their unrecorded value. RCA extracts no condition rate.").
practice_encounter(no_remodel_line,
  "The URAR records kitchen and bathroom updates; the grid has condition and quality lines and no
  ↪ remodel line, rightly: a remodel line would count the same value twice.").
practice_encounter(adjustments_do_not_travel,
  "An extracted adjustment is a composite of the named effect and unnamed effects correlated with
  ↪ it, so it is dataset-specific.").
practice_encounter(implausible_paired_sale,
  "A paired-sales adjustment far larger than cost can reflect leakage from a correlated omitted
  ↪ attribute.").

```