

遺漏變數偏誤 (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)

形式定義。「遺漏變數偏誤」是指當一個影響價格的特徵被排除在模型之外，且該遺漏特徵與其中一個已納入的特徵呈正相關時——在已納入的其他特徵，而輔助迴歸（即針對被排除變數與納入變數間關係所進行的第二個迴歸分析）正是用來衡量此現象。此時，被排除變數的影響便會滲入納入變數的係數中（即模型賦予各變數的每單位美元值）。

直觀說明。某筆成交價之所以不準確，可能是因為其中包含其他特徵的數值。第 1 期的 Comp H 成交價比根據其四項記錄特徵進行迴歸分析所預測的價格高出 \$46,455。原因如第 1 期《潛在變數》所揭露，是近期進行的廚房翻新工程；該條目指出，此工程約為該房產增值 \$60,000。「屋況」雖列於特徵清單中，但並無任何記錄顯示曾進行翻新，而翻新所帶來的價值也未出現在資料中。該價值並未消失，而是透過已納入特徵的比率反映出來，因為這些特徵在實際成交中恰好與其相關聯。要讓這種情況發生，必須滿足兩個條件：缺失的價值必須對價格產生影響，且必須與清單中某項特徵一同出現。當其對價格產生影響卻未與任何特徵相關聯時，係數是正確的，只是散點範圍會更寬。無論一項特徵對應單一係數或多個係數，此現象皆成立：《期刊》的主要迴歸方法 MARS（多變量自適應迴歸樣條，Friedman 1991）會根據特徵數值的不同區間賦予其不同的率值；在此情況下，借用的數值會分散在這些區間中，而非集中於單一數值，但基於相同原因，該數值確實存在。

給想了解運作機制讀者的參考連結：在一般迴歸中，率（rate）即為係數——也就是模型賦予某個特徵增加一個單位所對應的金額——而借用值的數值有精確的計算公式：被剔除特徵本身的係數，乘以其與納入特徵之間變動的強度；而後者即為被剔除特徵對納入特徵進行輔助迴歸所得的係數。該條目詳細闡述了此公式（此為線性情況），並進一步展示在「網格法」（grid）中同樣存在的滲漏現象——在此方法中，估價師的配對成交案例與係數一樣，同樣會傳遞此滲漏。該條目專為從市場數據中獲取調整值的估價師而撰寫。

本期範例的重點與非重點。在第 007 篇（條件期望）至第 010 篇（參數識別）中使用的「帕西菲卡」建模市場，將車況視為具有美元單價的特徵——這與傳統估價方法及《期刊》第 1 期所採用的做法一致——而本例正揭示了此方法的謬誤所在。傳統的估價方法並非本期刊所採用的方法，而是基於 RCA（殘差約束法）的估價法——即第 1 期（Craytor 2026）所提出的「殘差約束法」——該方法將狀況與品質排除在迴歸分析之外；條目 007（條件期望）闡述了具體做法。讀者應汲取該範例的教訓，而非照搬其具體規格。

房屋狀況、狀況評分以及翻新工程。這三者常被混為一談。房屋狀況指的是房屋的現狀：裝潢的磨損程度、設施系統的使用年限，以及積壓的維護需求。狀況評分則是針對該現狀所做的評估，並以該房屋所處的特定市場區域內周邊房屋的狀況為基準。該評分基於估價師對該市場區域的理解，且不應與在同一市場區域執業的同業估價師的理解產生重大衝突。在本期刊中，其範圍為 0.00 至 10.00：即該市場區域內狀況較差的房屋所佔比例，除以十。分數 5.00 表示該市場區域內有一半的房屋狀況較差；9.00 則表示十間中有九間狀況較差。此分數僅為估算值，兩間分數相同的房屋實際狀況仍可能有所不同。翻新工程既非「狀態」亦非「分數」，而是一項「事件」：指在近期某段期間內，為改變房屋而進行的工程。其價值在於被改變的項目——即房屋狀況，以及通常包含的品質、建材與裝修——並以市場對這些項目的報價為基準。翻新工程的影響通常主要體現在房屋狀況上，對品質的影響較小，這主要是由於人工成本所致；不過在某些翻新案例中，透過品質提升所帶來的附加價值反而更高。若兩棟房屋在所有當前與價值相關的特徵上完全相同，其中一棟經過翻新、另一棟從未翻新，兩者的價值應相同；市場所支付的，是翻新所產生的狀況

與品質，而非單純因為進行了翻新。因此，翻新並非一項具有獨立價格的特徵。當資料集在「屋況分數」旁包含「翻新指標」時，該指標代表翻新所帶來的、而模型中其他特徵（可使用面積、地塊、景觀及屋況分數）未記錄的內容：超出分數所顯示範圍的屋況，以及資料集未記錄的品質。這正是《潛在變數》（*Latent Variable*）第 1 期中關於 Comp H 翻新工程「使其房產總價值增加約 \$60,000」這項陳述的解讀方式，也是本條目範例的核心所在。

要產生偏誤，必須同時滿足以下兩個條件：被排除的特徵必須會影響價格，且其變化必須與其中一個被納入的特徵同步。若任一條件不成立，則不會產生偏誤。

這項兩部分的要求之所以重要，是因為該規則的常見說法——即「省略重要變數的模型會得出錯誤答案」——按其表述本身而言是錯誤的。若某個特徵會影響價格，但與任何已納入的特徵均無同步變動關係，它會擴大價格在模型周圍的分散程度，進而擴大標準誤差（用以衡量估計係數與極大樣本所得數值之間的偏差），但並不會改變係數本身的數值；截距（模型中的常數金額）會吸收該被排除變數的平均貢獻。在此情況下，失去的是精密度，而非準確度。另一方面，若某個變數雖與納入的變數呈正相關，但對價格毫無影響，則完全不會造成任何損失。唯有當上述兩種情況同時成立時，才會產生偏誤。

其運作原理是透過代入法。如果影響價格的某個特徵未包含在迴歸變數（即迴歸分析用來解釋價格的特徵）中，普通最小二乘法（OLS）並不會直接將該特徵的貢獻視為未解釋。那些與缺失變量呈正相關的納入變量，會充當該缺失變量的部分替代變量，而其係數會納入它們所能預測的缺失變量效應之部分。納入變量的係數不再僅代表該變量本身的效應；而是該變量本身的效應，加上它所「借用」的缺失變量效應之部分。

偏誤總是與特定問題相關。未包含「重新建模」指標的模型係數，對於某個問題而言是正確答案——即在這些銷售案例中，價格會隨著條件分數如何變化，且分數所包含的所有相關因素皆一併納入考量——但若將其解讀為「在其他變數均保持不變的情況下，市場對每點車況評分所對應的價格」，這便是調整分析的解讀方式，此時這些係數便成了錯誤的答案。該模型並非因為存在其他特徵而錯誤；而是其係數對於當前的應用方式而言並不正確。

借用份額具有精確的大小，由一個公式給出。對於被排除的特徵 w ，其真實係數為 β_w ；此時，被納入特徵 z_j 的係數偏差為

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

，其中 $\hat{\beta}_j$ —— 帽子符號表示估計值 —— 是迴歸對 z_j 所估計的係數，而 δ_j 則是 z_j 在輔助迴歸中，針對 w 對所有被納入特徵所計算出的係數。這是 Theil 的公式，¹ 根據擬合的迴歸模型計算得出，若使用被排除特徵的估計係數，該公式在每個樣本中均完全成立；若使用真實係數，則在平均層面上成立。

乘積形式即為線性情況。在 MARS 模型中，被排除特徵的影響會被與其相關的那些區段所吸收，因此偏差會分散在這些區段之中，而非集中在單一係數上；但偏差確實存在，且成因相同。

這個公式比表面上看起來更有用，因為即使這兩項因素都無法量化，通常仍可根據市場知識為其賦予正負號。當被排除的變項會推高價格（ $\beta_w > 0$ ），且與某個已納入變項呈正相關時（ $\delta_j > 0$ ），在其他已納入變項保持不變的情況下，該變項的係數便過高。若估價師僅知曉這兩項關係的方向，便能掌握誤差的方向。需注意一點： δ_j 是指在其他納入特徵保持不變的情況下所建立的關係；當納入特徵超過一項時，簡單相關係數（-1 至 +1 所提供的數值，用以衡量兩項特徵如何同步漲跌，且忽略所有其他變數）可能呈現相反的趨勢。在以下範例中，居住面積與翻新指標呈正相關（相關係數 +0.09），但其 δ 值為負，因此其相關係數過低。

估價實務中何處遇到。銷售比較表可以寫成一種加法模型，其結構與迴歸分析相同——即人工挑選的幾項特徵，每項特徵對價格進行美元調整——而當其調整值從市場數據中提取時，便會像任何迴歸分析一樣受到遺漏變數偏差的影響。該表格中的行即為納入的特徵。任何會影響房價但未被列為行項的房屋特徵，皆被排除在外。

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

常被忽略的因素包括：狀況評分未能涵蓋的狀況細節、未記錄品質評分的房況、平面圖的實用性、自然採光、噪音影響、建築區塊內的具體位置、賣方的出售動機，以及房源刊登所投入的行銷力度。上述各項皆可能影響價格，其中許多因素往往與評分表已納入考量的項目相輔相成。

翻新是一個極具參考價值的案例，因為相關紀錄雖已留存，但仍存在資訊洩漏的情況。統一住宅估價報告（URAR）會記載廚房和浴室是否經過更新或翻新，以及具體的翻新期間；而統一估價資料集（UAD）中的狀況定義本身，則取決於更新的程度。然而，該評估表格中設有「狀況」與「品質」兩欄，卻未設有「翻新」一欄——這在探討市場對房屋狀況的定價時是合理的，因為「狀況」與「品質」兩欄已涵蓋翻新所帶來的效益；若再增設第三欄，便會造成重複計算。倘若問題轉為「擬議中的翻新工程將如何影響該房屋的價格」，在將「狀況」與「品質」視為恆量時，便剝奪了翻新工程發揮作用的途徑；條目 009（內生性）正是探討此問題。這種「洩漏」源於評分系統未能涵蓋的面向。在一個市場中，若最高房況評分主要屬於近期翻修的房產，且翻修帶來的價值未被評分系統記錄，那麼傳統估價中從這些交易中提取的房況調整值——無論是透過配對銷售法，還是包含房況變數的迴歸分析——都會將該價值一併納入。根據泰爾（Theil）的公式，此項調整值便會過高：它不僅包含房況本身的影響，還包含評分系統未記錄的部分價值。若估價師採用此調整值，並將其套用至因維護良好而獲得最高評分的房屋，調整幅度將會過大。根據《期刊》的「殘差約束法」，不會提取任何狀況率。估價師會根據其判斷，將每筆可比物件的殘差，分配至房屋狀況、品質以及數據未測量的其他特徵，且須遵循一項約束條件：各殘差成分價值貢獻的總和必須等於該可比物件的殘差。如此一來，翻新工程的價值將由估價師依其判斷進行分配，而非由單一項目承載。

剛才所述的「洩漏」現象，正是對估價師針對市場導向調整所提出異議的更精確闡述，即源自某一組資料集的調整值難以有效轉移至其他情境。原因在於，所提取的調整值既包含已命名的效應，也包含隨之而來的未命名效應，而其組成方式取決於原始資料集中各項因素之間的關聯性。

另一個例子是，在傳統估價中，所謂的「配對成交案例」——即兩筆成交案例在除待評估特徵外的所有方面均完全相同，而價格差額被解讀為該特徵的價值——所得出的數值往往高得難以置信。原因可能並非測量誤差，而是「洩漏」：這兩筆交易不僅在所述特徵上有所不同，還包含與該特徵相關的其他因素，而價格差異正是由這兩者共同造成的。

對結論可辯護性的意義。一份未探討「遺漏變數偏誤」的報告，其享樂調整與網格調整將因此面臨此項質疑。

此偏差無法從擬合統計量中察覺。一個模型可能具有較高的 R^2 （模型重現價格變異性的比例）、較小的殘差（售價與模型預測價格之間的差異），且其係數通過了常規的統計檢定（ t -檢定，用以判斷係數是否合理地等於零），但仍可能存在嚴重的偏誤，因為偏誤源自特徵間的相關動態，而非殘差的大小。將 R^2 用作模型可靠性的證明並不能解答這個問題。在下面的例子中， R^2 為 0.97，而條件係數的誤差高達 74%。

更多的數據並不能解決問題。隨著樣本規模擴大，偏差並不會隨之縮小。若從同一個市場中，排除同一項特徵後取得一千筆銷售資料，所得到的估計值仍會以同一個錯誤數值為中心，只是估計範圍變得更窄：標準誤差會圍繞著這個帶有偏差的估計值縮小，這比圍繞該值形成寬廣的區間更為糟糕，因為報告看起來會顯得更為篤定。MARS 亦是如此：更多的銷售資料會使擬合曲面更平穩，但仍圍繞著相同的偏誤結果；地球的 varmod 區間衡量的是價格的分散程度，而非偏誤。這糾正了《第一期》中關於潛在變量的論點，該論點曾指出偏誤會隨著更多比較物件而縮小，且其他係數存在向上偏誤。偏差並不會隨著交易量增加而縮小；在《第一期》自身所列的八筆交易中，如同下文的十六筆交易一樣，景觀、GLA 和地塊係數均存在向下偏差；僅有條件係數存在向上偏差。

即使無法得知偏誤的幅度，通常仍可對其方向進行合理化說明。在傳統估價中，當從成交案例中提取狀況調整值——無論是透過配對成交案例，還是透過包含狀況變數的迴歸分析——實務上的做法並非斷言該調整值是無偏的，而是列出可能被忽略的特徵，運用市場知識為 β_w 及 δ_j 中的各項特徵賦予正負號，並說明由此產生的偏誤方向。「房況調整值可能被高估了。多數評分最高的成交案例都是近期翻新過的房產，而這些翻新帶來了數據中未記錄的改善，例如新的廚房和裝潢。

這些改善推高了價格，且主要出現在評分較高的成交案例中，因此其價值的一部分已被計入房況調整值中」是一項合理的陳述。這兩項要求均已滿足，且指向相同結論：被排除在外的改善措施推高了價格（ β_w 為正值），並與高分交易相關（ δ_j 為正值），因此偏差呈上升趨勢。這樣的陳述將一個無法量化的問題轉化為一個已被承認且具有明確方向的問題，這正是讀者用以權衡結論所需要的。在「殘差約束法」下，並未提取任何狀況率，因此關於「狀況」的問題也不存在。

估價計算示例。Pacifica 資料集的十六處房產擴充資料，針對每處房產，記錄了總居住面積 (GLA)、地塊面積、景觀、狀況評分、觀察到的翻新指標，以及售價。由於每個價格皆是根據已知公式計算得出，且不含隨機成分（資料無雜訊），因此使用正確特徵的迴歸模型能精確還原該公式，而任何與該公式的偏差皆屬偏誤，完全不包含雜訊成分。

Comp	GLA (sf)	地塊 (平方英尺)	檢視	Cond	翻新過	Price (\$)
A	1,650	9,000	是的	7.00	不	1,425,000
B	1,650	7,200	不	7.00	不	1,280,000
C	1,850	7,500	不	7.00	不	1,407,500
D	1,450	7,000	不	7.00	不	1,155,000
E	1,700	8,000	是的	9.00	不	1,470,000
F	1,550	7,100	不	5.00	不	1,177,500
G	1,700	10,500	不	7.00	不	1,392,500
H	1,600	7,300	不	7.00	是的	1,312,500
我	1,750	8,200	不	9.00	是的	1,465,000
J	1,600	7,600	不	7.00	是的	1,320,000
K	1,900	9,500	是的	9.00	是的	1,687,500
L	1,500	6,800	不	5.00	不	1,140,000
M	1,800	8,800	不	7.00	不	1,410,000
N	1,650	7,400	是的	7.00	不	1,385,000
O	1,550	7,900	不	9.00	是的	1,337,500
P	1,720	9,200	不	5.00	不	1,332,000

比較樣本 A 至 H 均為第 1 期，其已公布的屬性與價格均未變更；第 1 期以文字描述品相，並以 1 至 3 為代碼進行迴歸分析；根據本期刊的評分標準，文字描述對應「平均」5.00、「良好」7.00、「極佳」9.00，因此第 1 期的每一個等級變化相當於兩分。Comp H 的翻新工程——在第 1 期中已揭露為其殘差來源，但未被記錄為變數——現已納入記錄。市場對每點狀況評分支付 \$20,000。在建模數據中，一棟經翻修的房屋售價比擁有相同總可租用面積 (GLA)、地塊、景觀及狀況評分的房屋高出 \$60,000。這 \$60,000 並非翻修行為本身的價格；它代表著翻修所帶來的、而其他特徵未能記錄的價值。翻新過的房屋平均狀況評分也較高 (8.20 對比 6.64)，因此翻新指標會隨其包含的狀況評分波動，而這正是偏誤所要求的第二項條件： $\text{corr}(\text{remodel}, \text{cond}) = 0.55$ 。

針對同一組十六項房產特徵進行的兩項迴歸分析：

	翻新部分已省略	翻新工程已包含在內	偏見
攔截	-83,923.57	-30,000	-53,923.57
GLA	586.64	600	-13.36
洛特	24.32	25	-0.68
檢視	79,178.46	100,000	-20,821.54
條件 (每項)	34,815.56	20,000	+14,815.56
翻新	—	60,000	
R^2	0.9733	1.0000	

翻新線上的 \$60,000 並非翻新工程的價格。在此建構的資料集中，這是指在總可租賃面積 (GLA)、地塊、景觀及評分均相同時，附帶於翻新指標上的價格差異；這正是「恢復生成函數」所恢復的內容。

條件係數高出 \$14,816 一個點——其報告顯示每點條件值為 \$34,816，而市場數據為 \$20,000，高估了 74%。觀點係數低了 \$20,822。這兩項誤差在擬合結果中均無法察覺：未包含「翻新」指標的模型，其 $R^2 = 0.9733$ 及殘差——除了三筆接近 \$35,000 的交易外——看起來都很普通。無論是 R^2 中的結果，還是對係數的檢定，均未顯示任何異常跡象（若這些殘差為隨機誤差，則所有係數在慣常的 5% 顯著水準下均會通過該檢定）。

泰爾 (Theil) 公式 能精確解釋這些數值。將「remodel」指標對所包含的四個特徵進行迴歸分析，可得 $\delta_{\text{cond}} = 0.24692602$ 及 $\delta_{\text{view}} = -0.34702564$ 。接著

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

與觀察到的偏移量完全吻合。該偏差並非僅能由公式近似解釋；它本身就是該公式。雜訊的影響比預期中要小：即使在存在雜訊的數據下，該公式在兩組擬合迴歸（不含與包含「remodel」指標）之間仍完全成立，只要將指標的估計係數 $\hat{\beta}_w$ 取代 β_w ；相較於真實係數，該公式在多組樣本的平均值上依然成立。

「觀點係數」值得我們再細細端詳一番，因為它的偏誤方向恰好相反，而箇中原因頗具啟發性。在其他特徵保持不變的情況下，本樣本中擁有景觀的房產較少進行翻新 ($\delta_{\text{view}} < 0$)。因此，景觀係數被壓低而非抬高：它被 \$20,822 低估了，而房況係數則被 \$14,816 高估了。這兩種誤差是同一項遺漏所產生的獨立影響，而非彼此之間的轉移。若估價師僅針對狀況進行推論，根本不會預料到景觀調整會出現誤差——然而，景觀調整值被低估 21 百分比，正是該同一項遺漏所直接導致的後果。

實際上，損壞往往集中在估價師最不可能察覺的地方。首先考慮一下，這個存在偏誤的模型作為預測變數會產生什麼結果。以一處與比較物件 N 完全相同、未經翻修的標的物為例——總樓面面積 1,650 平方英尺，佔地 7,400 平方英尺，擁有景觀，且狀況評分為 7.00——該模型預測的價格為 \$1,386,914，而實際平均價格為 \$1,385,000 美元的實際平均價格——根據條目 007（條件期望）的五項特徵公式，並考量翻新因素 = 0——誤差為 \$1,914，即 0.14 百分比。在全部十六筆交易中，典型誤差約為 \$21,000，因此該模型雖存在嚴重偏誤，但預測結果仍尚可接受，這並非巧合：OLS 會針對所給的樣本進行擬合，而這些帶有偏誤的係數恰恰是最接近樣本價格的那些。預測準確度並不能作為排除遺漏變數偏誤的證據，而交叉驗證——即在未納入擬合的銷售資料上測試模型——亦無法排除此偏誤：良好的交叉驗證結果並不能證明係數沒有偏誤，因為當特徵的組合方式與擬合結果一致時，即使模型規格錯誤，仍能產生良好的預測結果。依序排除各房產特徵後，兩項最大的誤差——K 處的約 \$56,000 以及 G 處的 \$48,000——皆出現在符合本樣本模式的房產上；在此模式中，高狀況評分多屬經翻修的房屋；E 雖打破此模式，但以約 \$45,000 排名第三；而 H 和 J，同樣打破此模式，則分別以約 \$41,000 排名第四和第五。此處的數據不含雜訊，因此這些誤差全都源自被遺漏的「翻新」指標，而非雜訊；但在實際銷售中，由於也存在雜訊，僅憑誤差的大小無法判斷其源自被遺漏的特徵，還是源自普通的殘差波動。

一旦將係數用作調整而非預測規則，損害便隨即顯現。一位從此模型中提取狀況調整值的估價師，每點適用 \$34,816，而市場數據則為 \$20,000——每點多計了 \$29,631，而市場數據為 \$20,000——且景觀調整值為 \$79,178，而市場數據為 \$100,000。該網絡是逐一讀取迴歸係數的位置，而這正是偏誤所破壞的讀數。

在那些打破偏見所依據之模式的物業上，預測同樣會失準，而且失準得很嚴重。像本案這般維護良好的住宅（總可租賃面積 1,650 平方英尺，地塊面積 7,400 平方英尺），儘管沒有景觀、評分為 9.00（極佳），且從未進行翻新，其估價顯示為 \$1,377,367，而實際成交價為 \$1,325,000，高估幅度達 \$52,367 (3.95%)，因為在這些成交案例中，評分達 9.00 的房屋多半經過翻新，而該模型的房屋狀況評分會將翻新工程的 \$60,000 納入評分考量。一棟與本案完全相同（視野，評分 7.00）但經翻修的住宅，其價格標示為 \$1,386,914，而對照物件的評分則為 \$1,445,000，低估金額為 \$58,086 (4.02%)。這些屬於非典型物業——正是估價師最常受聘專門進行估價的物件，也是估價報告將被最仔細審閱的對象。

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/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.").
```