

# 條件期望 (Conditional Expectation)

*Foundations • The Valuation Engineer*

William Bert Craytor • October 1, 2026

Version 1.2.0

DOI: [10.67330/99hhbc69](https://doi.org/10.67330/99hhbc69)

形式定義。「在給定一組預測變數（即已測量的特徵）的情況下，某個結果（此處指售價）的條件期望，即是該結果在預測變數每個固定值處的分布均值。若將其視為預測變數的函數，則稱為條件期望函數（CEF），而迴歸分析正是旨在對此函數進行近似。普通最小二乘法（OLS）迴歸可提供對 CEF 的最佳線性近似，且當 CEF 本身為線性時，該迴歸能精確重現 CEF。

直觀說明。由迴歸分析得出的調整值，是從某個「比率」開始計算的：即市場針對每平方英尺、每個浴室、每房齡一年所支付的價格。這篇文章將探討這個比率的來源，以及它究竟是根據什麼來計算的。簡而言之，它代表當某一項特徵發生變化而其他特徵保持不變時，市場平均價格的斜率；而迴歸分析正是透過銷售數據來測量該斜率的一種方法。OLS 迴歸會針對該特徵的整個範圍計算一個斜率；而 MARS（多變量自適應迴歸樣條，Friedman 1991）則針對市場支付價格不同的每個區間分別計算斜率，並針對每個區間報告一個調整率。在兩種情況下，該比率皆是對同一事物的衡量——即平均價格隨該特徵變動的方式——而本文剩餘部分將探討此事物：它究竟是什麼、為何銷售比較網格也會衡量它，以及當針對同一筆銷售進行的兩次迴歸分析得出兩種不同比率時，這代表什麼意義。

給想了解運作機制讀者的參考連結：。所謂「隨特徵變化而變動的平均價格」有一個專有名詞，即「條件期望函數」，本文後續內容將以此術語指稱。它是在特徵空間上的一種曲面；某個方向上的斜率即為其斜率；OLS 會擬合最能近似該曲面的平面，且僅當該曲面本身為平面時，才能精確重現；MARS 則會擬合由平坦片段組成的曲面，並在偵測到的斷點處進行彎曲，因此當真實曲面出現彎曲時，其擬合結果便更為貼近。第 1 期中提到的八個 Pacifica 可比車款，便是此處的實例。

這篇文章探討的是統計學而非法律：內容聚焦於迴歸分析能夠估算什麼，並旨在供進行迴歸分析或使用銷售比較表（該表同樣用於估算相同內容）的估價師參考。

本刊第 1 期針對 Pacifica 可比車款所進行的迴歸分析，被用來從其係數（即迴歸模型賦予各項特徵的權重）推導出隱含價格；所謂隱含價格，即市場為一單位特徵所支付的金額。然而，該迴歸分析所估計的內容並未明確說明，這一點值得特別提及，因為本期隨後幾乎所有問題——涉及偏誤、內生性與識別問題——皆與「實際估計結果」與「預期結果」之間的差異有關。

迴歸分析通常被描述為透過將殘差平方和最小化（殘差即為實際銷售價格減去擬合方程式所給出的價格）來擬合一條直線，並據此得出斜率。雖然就涉及的算術運算而言，這種描述是正確的，但若論及分析目的，則會產生誤導。算術運算僅是手段；真正的目標是一個能描述整個市場（總體量）的數值，而不僅僅是手頭的銷售數據。

解讀記譜法。

|                                     |                                                        |
|-------------------------------------|--------------------------------------------------------|
| $P$                                 | 售價。                                                    |
| $\mathbf{Z}$                        | 某個物體所測量特徵的清單；這樣的清單稱為 向量。                               |
| $\mathbf{z}$                        | 該清單中的一組特定數值，例如 1,650 平方英尺、7,200 平方英尺的地塊、無景觀、狀況評分 7.00。 |
| $E[\ ]$                             | 期望值：所有可能數值的平均值，每個數值均以其機率為權重。                           |
| $ $                                 | “已給定”。                                                 |
| $E[P \mid \mathbf{Z} = \mathbf{z}]$ | 在所有特徵均等於 $\mathbf{z}$ 的物業中，價格的期望值。                     |
| $\varepsilon$                       | 銷售額與該預期值的偏差。                                           |
| $u$                                 | 一筆交易的結構性錯誤：其價格減去因果關係所指明的合理價格。                          |

$\hat{p}$  由擬合方程式得出的數值。  
 $z_{GLA}, \dots$  一項特徵的數值：建築總面積（GLA）及地塊面積（以平方英尺為單位）；景觀與翻新狀況：是則為 1，否則為 0；評分範圍為 0.00 至 10.00。

設定以下特徵向量： $\mathbf{z}$ ——也就是說，面積為 1,650 平方英尺，地塊面積為 7,200 平方英尺，無景觀，且房況為「良好」（在《Journal》的房況評分標準中為 7.00，詳見範例說明）。在市場上，具備此特徵向量之房產的價格各不相同。價格之所以不同，是因為房產存在該向量未能涵蓋的差異、買賣雙方各不相同，以及協商結果未必總是達成相同價格。在該市場中，存在一個與  $\mathbf{z}$  相對應的完整價格分佈。此分佈的平均值即為在  $\mathbf{z}$  條件下價格的條件期望，寫作  $E[P | \mathbf{Z} = \mathbf{z}]$ ；其一般定義由科爾莫戈洛夫（Kolmogorov, 1933）提出。手頭銷售價格的平均值是此期望值的估計值，而非期望值本身。

由於房產的特徵涵蓋了整個特徵空間，條件期望便勾勒出一條曲面：在每個點上都有一個數值，即具備該特徵之房產的平均價格。該曲面稱為 條件期望函數，或簡稱 CEF。這是市場的特徵，而非屬於任何特定資料集的屬性——即使從末有任何房產被估價過，它依然存在。圖1 針對本例繪製了該曲面。

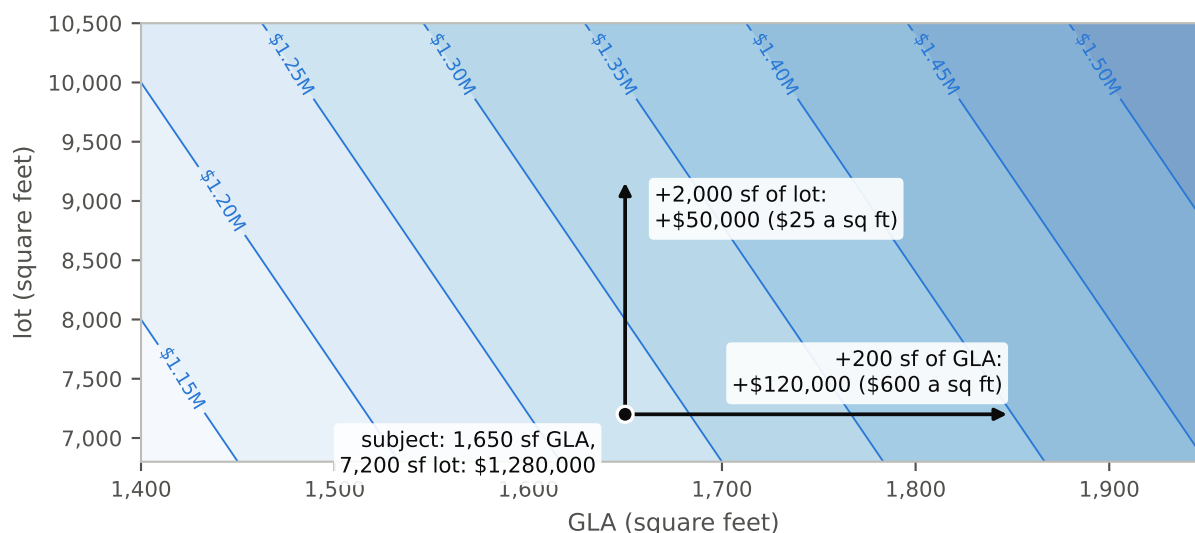


Figure 1: 所建構的帕西菲卡（Pacifica）市場之條件期望函數（範例），適用於無景觀、條件評分 7.00 且未經翻修的住宅，以地形圖形式呈現：每條線連接平均價格相同的總可租賃面積（GLA）與地塊組合，且平均價格越高，陰影越深。標記的點即為本條目所列的房屋，其總可租用面積（GLA）為 1,650 平方英尺，地塊面積為 7,200 平方英尺，位於 \$1,280,000。兩支箭頭代表兩組斜率，即價格變動率：總可租用面積增加 200 平方英尺，平均價格上升 \$120,000，即每平方英尺 \$600；地塊面積增加 2,000 平方英尺，平均價格上升 \$50,000，即每平方英尺 \$25。

有兩項事實使得 CEF 成為估計的自然目標。首先，在  $\mathbf{z}$  中的所有函數中，CEF 是能將預測誤差平方的期望值最小化的那個：當你必須僅根據  $\mathbf{z}$  來推測價格，且以誤差平方作為評斷標準時， $E[P | \mathbf{Z} = \mathbf{z}]$  便是你能做出的最佳推測。其次，任何結果皆可表示為

$$P = E[P | \mathbf{Z}] + \varepsilon, \quad E[\varepsilon | \mathbf{Z}] = 0,$$

的形式，其中第二個方程式並非假設，而是一項恆等式，因為根據定義，偏離條件均值的偏差其條件均值為零（Angrist、&、Pischke 2009，第 3 章）。

本條目位於條目 008（遺漏變數偏差）至 010（參數辨識）之前，因為這些條目均探討迴歸所推估的結果是否為所關注的條件期望值（CEF）。等式  $E[\varepsilon | \mathbf{Z}] = 0$  本身始終成立，因為  $\varepsilon$  被定義為

與條件平均值的距離。可能失效的情況是：對於另一種誤差——即結構性誤差（在條目 009（內生性）中以  $u$  表示）——同一條件（條件均值為零）不再成立；此結構性誤差正是價格與估價師欲測量的因果關係之間的差異。該誤差會隨特徵而變動。例如，當某個影響價格的特徵被遺漏時，就會發生這種情況；當價格反作用於某個特徵時——例如，市場待售天數（DOM）通常隨價格上升而增加，因此若使用 DOM 來解釋價格的迴歸分析，便會汲取部分價格本身對 DOM 的影響，這正是為何少有分析師將 DOM 納入淨售價的迴歸分析中；或是當某個特徵的測量存在誤差時。遺漏變數偏差（條目 008）與內生性（條目 009）即探討此類情形；而「參數識別」（條目 010）則探討更根本的問題：銷售數據究竟能否反映該比率。

估價實務中何處遇到。一位估價師若表示「在該地區，一棟面積為 1,650 平方英尺且無景觀的住宅，售價約為 \$1.28 百萬」，這屬於條件性預期。此處使用的詞彙是「約」。該陳述並非聲稱每棟此類房屋的售價皆為 \$1.28 million；而是指 \$1.28 million 是此類房屋成交價格的平均值。透過使用「約」一詞，該估價師承認個別成交價格會圍繞該數值有所波動。

銷售比較網絡是一種用於估算目標物體在特徵向量處 CEF 值的方法；每個可比物件皆為從該可比物件特徵向量處的價格分佈中擷取的觀測值；每次調整皆會將該觀測值沿著 CEF 曲面朝目標物體向量的方向進行位移；經調和後的指標隨即可提供標本點處 CEF 高度的估計值。從結構上看，當其調整值從銷售數據中提取時，該網絡可視為條件均值的簡單非參數估計量（即不假設曲面存在任何公式）——其採用局部平均與手動指定的斜率，而非普通最小二乘法（OLS）；從成本或判斷中得出的調整值屬於結構化意見，而非此類的估計值。

以這種方式構建網絡，能將估價師未言明的假設揭露出來。每當使用可比物件時，便假設條件期望值（CEF）足夠平滑，以致鄰近點的數值與標的物所在點的價值具有關聯性。線性調整假設該特定特徵的 CEF 斜率在所涵蓋的範圍內大致恆定。賦予三項指標同等權重，則假設每項指標估算的條件均值相同、均無系統性偏差，且精確度相近。這些假設均不罕見；它們皆涉及曲面的形狀，且一旦將該曲面命名，這些假設便會成為討論的議題。

這種框架也釐清了關於區間的常見混淆。不確定性主要分為兩類：一類是關於 CEF 的位置——即擬合曲面與真實平均值之間的偏差；另一類則是個別價格圍繞該平均值的分散程度。信賴區間僅涵蓋前者；預測區間則涵蓋兩者，因此總是兩者中較寬的那一個。在《期刊》的主要迴歸方法 MARS 中，earth 套件（Milborrow 2026a）會根據第二個模型——其方差模型（Milborrow 2026b）——為某個研究對象提供預測區間；該方差模型用以估計價格波動範圍在市場中的變化情況。其前提條件是：交易案例數量充足，且誤差大致呈正態分佈且彼此獨立；該套件會報告交易價格落在該區間內的頻率，這便是驗證依據。在基於 RCA 的估價中，根據銷售比較法，標的物之市場價值等於模型估算的售價加上估價師估算的殘差。在上述條件成立的情況下，方差模型會給出一個區間，標的物之銷售價格有 95% 的機率會落在該區間內。當估價師被問及「您有多確定？」時，通常是詢問個別價格的分散範圍（預測區間），而他們通常會回答對平均值的確定程度（信賴區間）。

為估量值——也就是正在估計的母體特徵——命名，是一項具有論證價值的行為，因為若無此命名，便無法對結論進行判斷。陳述「調整值為 \$100,000」目前尚無法被視為正確或錯誤。唯有當我們說：「在總居住面積（GLA）、地塊及房況保持不變的情況下，有景觀與無景觀物業之間的條件平均售價差異為 \$100,000」時，我們才提出了一項主張，而這項主張可能因「保持不變的變數」、「所指的樣本群體」以及「涉及的單位數量」等理由而受到質疑。

對結論可辯護性的意義。由此衍生出三項可辯護性的後果。

選擇一組條件集（即保持不變的特徵）屬於方法論層面的考量：它定義了估計量。至於數據能否估計該估計量，則是另一個問題，屬於「參數識別」的範疇。表達式  $E[P \mid \text{GLA}, \text{view}]$  與  $E[P \mid \text{GLA}, \text{view}, \text{condition}]$  代表不同的函數，它們可能具有不同的數值和不同的斜率。當評估師將一個變數納入評估框架時，並非僅是改善估計結果；而是改變了被估量值。合理的做法是明確說明價值意見所依據的條件集。

「在其他條件不變的情況下」這一表述描述了估價師在分析中的推理方式，且具有明確的涵義。在 CEF 框架內，它意指比較僅在單一特徵上有所不同的兩個特徵向量之條件均值。這是一項與總體相關的陳述，當市場數據充足時，可據此進行驗證。其背後的分析通常會記錄於工作檔案

中——即《美國專業評估實務準則》(USPAP)的「記錄保存規則」要求估價師必須保存的委託案紀錄——而非載入報告中。

殘差散佈是資訊，而非令人尷尬的來源。價格圍繞條件期望函數的散佈是市場的一項特徵；在價格圍繞該函數散佈範圍極廣的市場中，任何單一數值的估計——無論採用何種方法——都將伴隨著極大的不確定性。若在陳述數值時未提及此波動範圍，便會使結論顯得比市場實際情況更為精確。

估價計算示例。《第 1 期》中列出的八項 Pacifica 可比價格，是一個無雜訊的樣本：它們是為了精確滿足單一線性方程式而構建的。當將這八個價格以總可租面積 (GLA)、地塊、景觀、屋況及翻新指標為自變數進行迴歸分析時，每個價格都能精確重現至美元為單位——毫無殘差——且迴歸結果完全還原了構建這些價格所依據的方程式，各項係數完全吻合：

$$E[P | \mathbf{Z}] = -30,000 + 600 \cdot z_{\text{GLA}} + 25 \cdot z_{\text{lot}} + 100,000 \cdot z_{\text{view}} + 20,000 \cdot z_{\text{cond}} + 60,000 \cdot z_{\text{remodel}}$$

此處的「狀況」指的是本《期刊》所採用的狀況評分：這是根據該房屋所在的定義市場區域內周邊房屋狀況，對其狀況所作出的評估。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存在重大衝突。該評分範圍為 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房屋的狀況較差；9.00 則表示十間中有九間狀況較差。第 1 期以文字描述房屋狀況，並為其迴歸分析編碼為 1 至 3；根據本期刊的評分標準，文字描述對應的數值為：一般 5.00、良好 7.00、極佳 9.00；因此第一期中的一級對應兩分，而市場每分價格為 \$20,000。最後一項是建模數據賦予翻新工程的價格差異：\$60,000，高於一棟擁有相同總可租面積 (GLA)、地塊、景觀及狀況評分的房屋。這並非翻新工程本身的價格；它代表翻新所帶來的、而其他特徵未記錄的任何價值（項目 008，遺漏變數偏差）。

這正是針對建構出的 Pacifica 市場所建立的 CEF。必須釐清剛才所描述的究竟是什麼樣的對象，以及為何能如此精確地加以描述。這份資料集是經由建構而成的，而建構出的資料集事實上可以是無雜訊的；真實的資料集則永遠不可能如此。建構此資料集的優勢在於，它能徹底消除價格中的隨機雜訊——每個價格都精確符合公式——因此，在第 008 項（遺漏變數偏差）至第 010 項（參數識別）中，估計值與目標值之間的任何差異都不可能源自雜訊：每種情況要麼是偏誤，要麼（在「參數識別」中）是識別失敗。

此範例將「狀況」視為一項具有美元費率的特徵，這與傳統估價方法及《期刊》第 1 期所採用的做法一致；而條目 008（遺漏變數偏差）、009（內生性）及 010（參數識別）則說明了這種做法會產生哪些問題。傳統的估價方法並非本期刊所採用的「殘差約束法 (RCA) 估價法」。在殘差約束法中，迴歸分析僅採用可測量的特徵；「狀況」與「品質」則是刻意排除在外的。迴歸分析無法解釋的售價部分即為殘差。在編輯所處的加州市場中，可測量的特徵通常能解釋售價變異性的約 80%，其餘約五分之一則歸因於殘差。估價師會依據其評分作為指引，在「殘差成分價值貢獻之總和必須等於可比物件的殘差」這一約束條件下，將殘差分配至狀況、品質以及數據未測量的其他特徵。價值結論並不取決於估價師如何分配殘差——RCA 論文的證明 II (Craytor 2026, 第 36 頁)：在「所有殘差組成部分的價值貢獻總和必須等於可比物件的殘差」這一約束下，改變殘差的分配方式，對最終的價值結論並無影響。因此，狀況係數從未被估算，也從未被主張；估價師賦予狀況的權重，僅是那些已經測量特徵（包括樓齡在內）所已考量範圍之外的狀況因素。

若將第 1 期迴歸與此 CEF 進行比較，則可對該迴歸作出判斷——儘管首先必須釐清一項區別。一般而言，普通最小二乘法 (OLS) 迴歸並不會重建其目標的條件期望函數；相反地，它會找出該函數的最佳線性近似，也就是線性投影，而這兩者僅在條件期望函數為線性時才一致。像 MARS 這樣的非參數方法則不需要條件期望函數是線性的：它以分段方式近似條件期望函數，在曲面彎曲處設有鉸鏈點（比率變化的點），在曲面扭轉處設有交互作用項（取決於另一特徵的比率），因此遵循的是一條非線性條件期望函數，而線性投影會將其撫平。這仍然是對同一對象的近似，而後續條目中關於省略變數、內生性與識別性的問題，也以完全相同的方式適用於此。上述的五變數 CEF 屬於線性關係，這正是為何包含重建指標的迴歸模型能夠精確重現該 CEF 的原因。以四項特徵為條件（排除翻新指標）的第 1 期迴歸，估算了  $E[P | \text{GLA}, \text{lot}, \text{view}, \text{cond}]$  的最佳線性近

似值，並在將條件重新表述為期刊評分後，得出結果（第 1 期刊載的截距為 80,777，以及根據其 1–3 編碼的每一步所對應的 49,824；兩者為相同的擬合結果），

$$\hat{p}(\mathbf{z}) = 6,041 + 575 \cdot z_{\text{GLA}} + 22.70 \cdot z_{\text{lot}} + 86,179 \cdot z_{\text{view}} + 24,912 \cdot z_{\text{cond}}.$$

每個係數都與其上方的係數不同。一旦將迴歸描述為直線擬合，便無需再多做闡述：我們有兩個模型和兩組數值。然而，在 CEF 方法下，可以得出兩項精確的論述。首先，這兩組迴歸是以不同的變數集為條件的，因此近似了不同的條件期望函數：

$$E[P \mid \text{GLA, lot, view, cond}] \quad \text{and} \quad E[P \mid \text{GLA, lot, view, cond, remodel}].$$

在一個市場中，當某些經翻新與未經翻新的房產共享相同的四項特徵時，這兩項函數在這些點上的取值便會有所不同。在此小型的人為市場中，沒有兩處房產具備完全相同的四項特徵，因此這兩項函數在每處房產上雖結果一致，但形式卻有所不同：第二項函數在特徵上呈直線關係，而第一項則非如此。在每種情況下，迴歸分析僅提供其目標變數的最佳線性近似值。這兩項目標變數均為有效的分析對象。其次，只有那個包含五個變數的函數，其係數可被解讀為隱含價格——其中「翻新係數」即為具有相同總可租面積（GLA）、地塊、景觀及狀況評分之翻新房屋與未翻新房屋之間的價格差——而條目 008（遺漏變數偏差）則闡明了箇中原因。

第 1 期中 Comp H 的殘差，恰恰說明了這種區別有多麼重要。若將第 1 期的八項特性視為代表整個市場，則每個四變數特徵向量皆與這八項特性中的一項完全吻合，因此四變數 CEF 會精確地穿過所有八個價格——其中之一即是 Comp H。Comp H 與四變數 CEF 之間的偏差為零；其第 1 期殘差為 \$46,455，即其與擬合的四變數線性迴歸（即該八項屬性市場的線性投影）之間的差異量——而翻新工程的影響，恰恰隱藏於該投影與 CEF 之間的差距之中。採用五變數規格時，投影與 CEF 便完全一致，偏差也隨之消失。該殘差並非源自 Comp H 本身，而是源自規格設定。

## References.

- Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
- 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>
- Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Julius Springer.
- Milborrow, S. (2026a). *earth: Multivariate Adaptive Regression Splines* (R package version 5.3.6; derived from mda:mars by T. Hastie and R. Tibshirani). <https://doi.org/10.32614/CRAN.package.earth>
- Milborrow, S. (2026b, August 7). *Variance Models in earth*. <https://www.milbo.org/doc/earth-varmod.pdf>

---

*Cross-references: hedonic regression; implicit price; omitted variable bias; parameter identification.*

---

## Addendum: Prolog Formalization

The following machine-readable companion (007-conditional-expectation.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>.
```

```
%% 007-conditional-expectation.pl
```

```
%%
```

```
%% Prolog companion to Foundations Vol. 1, Issue 2, Entry 007
```

```
%% Title: Conditional Expectation
```

```
%% Author: Bert Craytor
```

```
%% Concept: conditional_expectation
```

```
%% Version: 1.2.0
```

```
%%
```

```
%% v1.2.0: entry text revised for readers who are not statisticians;
```

```
%% no Prolog clause changes.
```

```
%%
```

```
%% 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 generating function lives in
%% kb/pacifica_comps_extended.pl, which is restated in those units with this
%% release; nothing here changes but the units it reads.
```

```
%%
```

```
%% This file is the Prolog galley for the entry: it encodes the entry's
%% main claims, its worked example, and the generating function of the
%% running dataset in a form other Prolog code in the journal can
%% consult, query, and verify. Read alongside entry.tex / entry.pdf.
```

```
% =====
```

```
% DICTIONARY IMPORT
```

```
% =====
```

```
% 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 (quoted for reference) -----
```

```
% term(conditional_expectation, "Conditional Expectation",
%     econometric_method,
%     "The conditional expectation of an outcome given a set of
%     predictors is the mean of the outcome's distribution at each
%     fixed value of the predictors. Viewed as a function of the
%     predictors, it is called the conditional expectation function
%     (CEF), and it is the function that regression aims to
%     approximate. Ordinary least squares (OLS) regression gives the
%     best linear approximation to the CEF, and it reproduces the CEF
%     exactly when the CEF itself is linear.", published,
%     'dictionary-2026.1').
```

```
%
```

```
% depends_on(conditional_expectation, hedonic_regression).
```

```
% -----
```

```
uses_terms([ conditional_expectation, hedonic_regression,
              implicit_price, hedonic_price_function,
              characteristics_space, residual,
              ordinary_least_squares,
              sales_comparison_approach, adjustment,
              comp_selection, reconciliation ]).
```

```
uses_relations([depends_on/2, related/2]).
```

```

% =====
% ENTRY SUBJECT
% =====

article_concept(conditional_expectation).
article_issue(volume(1), number(2), year(2026)).
article_id('foundations.2026.007').

% =====
% ENTRY-LOCAL VOCABULARY
% =====
%
% Field names for the worked-example records and for the two properties
% of the CEF the entry relies on. Too narrow to promote to the
% dictionary.

local_term(cef).                % conditional expectation function
local_term(estimand).
local_term(conditioning_set).
local_term(prediction_interval).
local_term(confidence_interval).
local_term(linear_projection).   % best linear approximation to the CEF

% =====
% THE TWO PROPERTIES OF THE CEF THE ENTRY USES
% =====

%! cef_property(?Name, ?Statement) is nondet.

cef_property(best_predictor,
  "Among all functions of the conditioning variables, the CEF minimizes expected squared
  ↪ prediction error.").
cef_property(orthogonal_deviation,
  "P = E[P|Z] + eps with E[eps|Z] = 0 holds by construction, not by assumption.").

% The second property is the hinge for entries 008-010: each asks when
% the recovered object is the CEF of interest.
hinge_for(orthogonal_deviation, omitted_variable_bias).
hinge_for(orthogonal_deviation, endogeneity).
hinge_for(orthogonal_deviation, identification).

% =====
% CENTRAL CLAIM: CONDITIONING SET DETERMINES THE ESTIMAND
% =====
%
% Two regressions over the same data with different conditioning sets
% target DIFFERENT population objects, and each recovers only the best
% linear approximation to its target (the linear projection); OLS
% recovers the CEF itself exactly when the CEF is linear. Neither is
% wrong; they answer different questions.

conditioning_set(issue1_model,
  [gla_sqft, lot_sqft, view_numeric, cond_score]).
conditioning_set(issue2_model,
  [gla_sqft, lot_sqft, view_numeric, cond_score, remodel_numeric]).

%! different_estimand(?ModelA, ?ModelB) is nondet.
%

```

```

% True when two models condition on different variable sets and
% therefore approximate different conditional expectation functions.

different_estimand(A, B) :-
    conditioning_set(A, SA),
    conditioning_set(B, SB),
    A \== B,
    sort(SA, X), sort(SB, Y),
    X \== Y.

% What OLS recovers, relative to the CEF it conditions on.
ols_recovers(best_linear_approximation_to_cef).
ols_recovers_cef_exactly_when(cef_is_linear).

% =====
% WORKED EXAMPLE — THE GENERATING FUNCTION
% =====
%
% The running dataset is defined in ../../kb/pacifica_comps_extended.pl.
% Its generating function IS the conditional expectation function for
% the constructed market, exactly: the data carry no noise.

:- initialization(consult('../../kb/pacifica_comps_extended')).
% Scryer Prolog loads forall/2 and the list predicates on request;
% SWI-Prolog and Trealla have them built in. The guarded loads below
% let the file run unchanged in all three.
:- initialization(catch(use_module(library(lists)), _, true)).
:- initialization(catch(use_module(library(iso_ext)), _, true)).

%! cef_exact(+Property, -Price) is det.
%
% The CEF evaluated at a property's characteristic vector. Identical to
% true_price/2 in the dataset file; named here to make the entry's claim
% ("the generating function IS the CEF") explicit in the galley.

cef_exact(P, Price) :- true_price(P, Price).

%! noiseless_dataset is semidet.
%
% Verifies the entry's central factual claim about the data: every
% observed sale price equals the CEF at that property, so the residual
% is zero everywhere.

noiseless_dataset :-
    forall(property(P),
        ( cef_exact(P, Price), sale_price(P, Price) ) ).

% =====
% THE COMP H POINT
% =====
%
% Issue 1 reported a $46,455 residual for Comp H under the four-variable
% model. Taking issue 1's eight properties as the market, every
% four-variable characteristic vector belongs to exactly one property, so
% the four-variable CEF interpolates all eight prices and Comp H deviates
% from it by ZERO. The 46,455 is Comp H's deviation from the fitted
% four-variable LINEAR regression --- the linear projection --- and the
% projection-CEF gap is where the remodel effect hides. Under the
% five-variable specification, projection and CEF coincide and the

```

```

% deviation vanishes.

issue1_residual(comp_h, 46455).

% deviation_from(Property, Object, Dollars).
deviation_from(comp_h, four_variable_projection, 46455).
deviation_from(comp_h, four_variable_cef, 0).
deviation_from(comp_h, five_variable_cef, 0).

residual_under(issue2_model, comp_h, 0) :-
    cef_exact(comp_h, Price),
    sale_price(comp_h, Price).

%! residual_is_artifact_of_specification(?Property) is nondet.
%
% The issue-1 residual is a property of the specification (a linear
% projection under a smaller conditioning set), not of the property.

residual_is_artifact_of_specification(P) :-
    issue1_residual(P, R), R \= 0,
    deviation_from(P, four_variable_cef, 0),
    residual_under(issue2_model, P, 0).

% =====
% APPRAISER ENCOUNTERS
% =====

grid_assumption(smoothness,
    "Using comparables at all assumes the CEF is smooth enough that nearby points inform the
    ↪ subject's point.").
grid_assumption(local_linearity,
    "Adjusting linearly assumes the CEF's slope in that characteristic is roughly constant over the
    ↪ range crossed.").
grid_assumption(unbiased_indications,
    "Averaging indications assumes each is an unbiased draw around the same conditional mean.").

% The grid is structurally a nonparametric estimator of a conditional
% mean, using local averaging and hand-specified slopes.
estimator_role(sales_comparison_approach, conditional_mean_estimator).

% =====
% DEFENSIBILITY
% =====

defensibility_consequence(conditioning_set,
    "Adding a variable to a grid changes the estimand, not merely the precision of an estimate.").
defensibility_consequence(ceteris_paribus,
    "'Holding other things equal' means comparing conditional means at vectors differing in one
    ↪ coordinate.").
defensibility_consequence(residual_spread,
    "Conditional dispersion is a market property; a point estimate reported without it overstates
    ↪ precision.").

% The confidence/prediction interval confusion the entry names.
interval_answers(confidence_interval, "uncertainty about where the CEF is").
interval_answers(prediction_interval, "spread of individual prices around the CEF").
commonly_confused(confidence_interval, prediction_interval).

```