

条件期望 (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)

形式定义。给定一组预测变量（即测得的特征）时，某个结果（此处为销售价格）的条件期望，即为该结果分布在预测变量每个固定值处的均值。将其视为预测变量的函数时，称为条件期望函数 (CEF)，而回归分析正是旨在对该函数进行逼近。普通最小二乘法 (OLS) 回归可对 CEF 给出最佳的线性逼近，当 CEF 本身为线性时，OLS 回归能精确再现 CEF。

直观说明。基于回归分析得出的调整值始于一个比率：市场对每平方英尺、每个卫生间以及每房龄一年所支付的价格。本文将探讨该比率的来源及其衡量对象。简而言之，该费率是当某一特征发生变化而其他特征保持不变时，市场平均价格的斜率；而回归分析正是通过销售数据来测算该斜率的一种方法。OLS 回归针对该特征的整个范围测算一个斜率；而 MARS（多变量自适应回归样条，Friedman 1991）则针对市场支付价格存在差异的每个区间分别测算斜率，并为每个区间报告一个比率。在两种情况下，该比率都是对同一事物的衡量——即平均价格随该特征变化的趋势，而本文剩余部分将探讨这一问题：它究竟是什么，为什么销售比较网格也会衡量它，以及当针对同一笔销售进行的两次回归得出两个不同的比率时，这又意味着什么。

对于希望了解具体机制的读者，请参阅：“特征变化时的平均价格”有一个专有名词，即“条件期望函数”，本文后续内容将使用该术语。它是在特征空间上的一条曲面；某个方向上的斜率即为该曲面的斜率；OLS 拟合的是最能近似该曲面的平面，且仅当该曲面本身为平面时才能精确还原；MARS 则拟合由平坦片段组成的曲面，这些片段会在其发现的断点处发生弯曲，因此当真实曲面发生弯曲时，MARS 的拟合结果会更接近真实曲面。第 1 期中提到的八辆 Pacifica 可比实例便是贯穿全文的示例。

本文主要涉及统计学而非法律：它探讨了回归分析能够估算的内容，主要面向那些进行回归分析或使用销售比较网格（该网格同样用于估算相同内容）的估价师。

利用第 1 期针对 Pacifica 可比实例进行的回归分析，根据其系数（即回归模型赋予各特征的权重）推导出隐含价格；隐含价格是指市场为该特征的一个单位所支付的美元金额。然而，该回归所估计的内容并未明确说明，这一点值得提及，因为本期后续几乎所有问题——涉及偏倚、内生性和识别性——都与实际估计结果与预期结果之间的差异有关。

回归通常被描述为通过最小化残差平方和（残差即销售价格减去拟合方程给出的价格）来拟合一条直线，然后得出该直线的斜率。虽然从涉及的算术运算来看，这种描述是正确的，但在目的方面却容易引起误解。这些算术运算仅仅是手段；目标是一个描述整个市场（总体量）的量，而不仅仅是手头的销售数据。

解读记谱法。

P	售价。
$\mathbf{Z}$	房产的测量特征列表；这样的列表称为 向量。
$\mathbf{z}$	该列表中的一组具体数值，例如 1,650 平方英尺、7,200 平方英尺的地块、无景观、状况评分 7.00。
$E[\]$	期望值：所有可能值的平均值，每个值有其权重，权重与概率相关。
$ $	“给定”。
$E[P \mid \mathbf{Z} = \mathbf{z}]$	在所有特征与 $\mathbf{z}$ 相等的房产中，其价格的期望值。
ε	销售额与该预期值的偏差。
u	销售中的结构性错误：其价格减去因果关系所决定的合理价格。
$\hat{p}$	拟合方程得出的结果。

$z_{\text{GLA}}, \dots$

某项特征的值：建筑面积（GLA）和地块面积（平方英尺）；景观和翻新情况：是则为 1，否则为 0；残差范围为 0.00–10.00。

确定特征向量 $\mathbf{z}$ ——即面积 1,650 平方英尺，地块面积 7,200 平方英尺，无景观，且房况为“良好”（在《期刊》的房况评分标准中为 7.00 分，具体说明见例题解析）。在市场上，具有完全相同特征向量的房产，其价格各不相同。价格之所以不同，是因为房产在特征向量无法涵盖的方面存在差异，因为买卖双方各不相同，也因为谈判结果并不总是达成相同的价格。在该市场中，存在一个与 $\mathbf{z}$ 相对应的完整价格分布。该分布的平均值即给定 $\mathbf{z}$ 时的价格条件期望，记作 $E[P | \mathbf{Z} = \mathbf{z}]$ ；其一般定义由科尔莫戈洛夫（1933）提出。已成交价格的平均值是该期望值的估计量，而非期望值本身。

由于房产特征在整个特征空间中分布，条件期望描绘出一条曲面：曲线上的每个点都对应一个数值，即具有该特征的房产的平均价格。该曲面被称为条件期望函数，或简称 CEF。它是市场的固有特征，而非属于任何特定数据集的属性——即使从未对任何房产进行过评估，它依然存在。图1绘制了本示例中的该曲面。

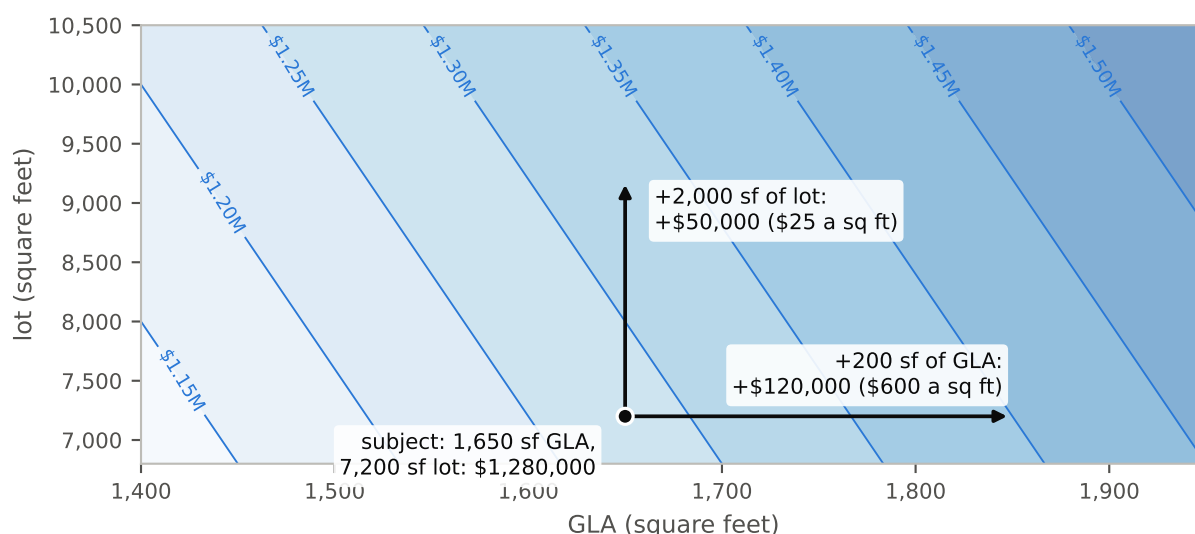


Figure 1: 构建的帕西菲卡（Pacifica）市场的条件期望函数（示例），针对无景观、条件评分 7.00 且未经翻新的住宅，以地形图形式绘制：每条线连接平均价格相同的 GLA 与地块组合，且随着平均价格的上升，阴影逐渐加深。标记的点即为本文所述的房屋，其总建筑面积（GLA）为 1,650 平方英尺，地块面积为 7,200 平方英尺，位于 \$1,280,000。两条箭头代表两个斜率，即价格变动率：总建筑面积每增加 200 平方英尺，平均价格上升 \$120,000，即每平方英尺 \$600；地块面积每增加 2,000 平方英尺，平均价格上升 \$50,000，即每平方英尺 \$25。

有两个事实使得条件期望函数（CEF）成为估计的自然目标。首先，在 $\mathbf{z}$ 的所有函数中，CEF 是将预测误差平方的期望值最小化的那个：当你只能基于 $\mathbf{z}$ 来猜测价格，且以误差平方作为评估标准时， $E[P | \mathbf{Z} = \mathbf{z}]$ 就是你能做出的最佳猜测。其次，任何结果均可表示为

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

的形式，其中第二个等式并非假设，而是一个恒等式，因为根据定义，偏离条件均值的偏差的条件均值为零（Angrist、&、Pischke 2009，第 3 章）。

本条目位于条目 008（遗漏变量偏倚）至 010（参数识别）之前，因为这些条目均探讨了回归所估计的结果是否正是我们所关注的条件期望（CEF）。恒等式 $E[\varepsilon | \mathbf{Z}] = 0$ 本身始终成立，因为 ε 被定义为与条件均值的距离。可能失效的是，当面对另一种误差时，同一条件（即条件均值为零）不再成立：这种结构性误差在条目 009（内生性）中记作 u ——即价格与估价师希望测量的因果关

系之间的差异。该误差会随特征而变化。例如，当某个影响价格的特征被遗漏时，就会发生这种情况；当价格反作用于某个特征时——例如，随着价格上涨，在市天数（DOM）通常也会增加，因此使用 DOM 来解释价格的回归模型会捕获价格自身对 DOM 的部分影响，这也是为何很少有分析师将 DOM 纳入净销售价格的回归模型中；或者当某个特征的测量存在误差时。遗漏变量偏差（条目 008）和“内生性”（条目 009）涉及上述情况；而“参数识别”（条目 010）则探讨了更根本的问题：销售数据能否真正反映该比率。

估价实务中何处遇到。一位估价师若表示“该地区一套面积为 1,650 平方英尺且无景观的住宅售价约为 \$1.28 百万”，则是在给出一个有条件的预期。此处使用的词是“约”。该陈述并非声称每套此类房屋的售价均为 \$1.28 million；而是指 \$1.28 million 是此类房屋成交价格的平均值。通过使用“约”这一词，估价师承认个别房屋的售价会围绕该数值有所波动。

销售比较网格是一种用于估算估价对象在特征向量处 CEF 值的方法；每个可比实例都是从该可比实例特征向量处的价格分布中提取的一个观测值；每次调整都会将该观测值沿 CEF 曲面朝估价对象的向量方向进行平移；经调整后的结果即可作为 CEF 在估价对象处高度的估计值。从结构上看，当其调整项从销售数据中提取出来时，该网格可被视为条件均值的简单非参数估计量（即不假设曲面存在任何公式的估计量）——它采用局部平均和手动指定的斜率，而非普通最小二乘法（OLS）；从成本或判断中得出的调整项属于结构化意见，而非此类估计值。

以这种方式构建网格，将估价师那些未言明的假设揭示出来。每当使用可比实例时，都默认条件期望函数（CEF）足够光滑，以至于邻近点的值与估价对象所在点的价值相关。线性调整假设该特定特征的 CEF 斜率在所涵盖的范围内大致恒定。赋予三项指标同等权重，则假设每项指标估算的条件均值相同，均无系统偏差，且精度相近。这些假设均不罕见；它们都涉及曲面的形状，且一旦确定了所讨论的曲面，这些假设都会成为讨论的主题。

这种表述方式也消除了人们对区间的常见混淆。不确定性主要有两种：一种是关于条件期望函数（CEF）的位置——即拟合曲面与真实平均值之间的偏差；另一种是单个价格围绕该平均值的离散程度。置信区间仅涵盖前者；预测区间则同时涵盖这两者，因此其范围总是比置信区间更宽。在《期刊》的主要回归方法 MARS 中，earth 包（Milborrow 2026a）通过第二个模型——方差模型（Milborrow 2026b）——为估价对象提供预测区间，该模型用于估计价格波动在整个市场中的变化情况。其前提条件是：成交案例数量充足，且误差大致服从正态分布且彼此独立；该软件包会报告成交案例落在该区间内的频率，这便是验证依据。在基于 RCA 的评估中，根据销售比较法，估价对象的市场价值等于模型估算的销售价格加上估价师估算的残差。在满足上述条件的情况下，方差模型给出了估价对象的销售价格以 95% 的概率落入的区间。当估价师被问及“你有多确定？”时，提问者通常想了解单个价格的分散程度（预测区间），而估价师通常会回答他们对平均值的确定程度（置信区间）。

将估计量命名为——即被估计总体中的特征——是一项具有合理性的行为，因为若无此命名，便无法对结论进行评判。陈述“调整值为 \$100,000”尚不能被视为正确或错误。只有当我们说“在总居住面积（GLA）、地块和房屋状况保持不变的情况下，带景观和不带景观房产的条件均值售价之差为 \$100,000”时，我们才提出了一项主张，而该主张可能会因“保持恒定”的条件、所涉及的样本群体以及具体房产单位等因素而受到质疑。

对结论可辩护性的意义。由此衍生出以下三点辩护性的后果。

条件集（即保持恒定的特征）的选择是一个方法论问题：它定义了被估量。数据能否估计该被估量则是另一个问题，属于参数识别的估价对象。表达式 $E[P \mid \text{GLA, view}]$ 和 $E[P \mid \text{GLA, view, condition}]$ 代表不同的函数，它们可能具有不同的数值和不同的斜率。当估价师将一个变量纳入评估网格时，他们并非仅仅是在改进估计结果；而是改变了被估量对象。合理的做法是明确说明价值意见所依据的条件集。

“在其他条件不变的情况下”这一表述描述了估价师在分析中的推理过程，且具有明确的含义。在 CEF 框架内，它指的是比较仅在某一特征上存在差异的两个特征向量所对应的条件均值。这是一项与总体相关的陈述，当市场数据充足时，可通过市场数据进行验证。其背后的分析通常会记录在工作文件中——即《美国房地产评估准则》（USPAP）“记录保存规则”要求估价师保存的评估任

务记录——而非写入报告中。

残差分散度是一种信息，而非令人尴尬的来源。价格围绕条件期望函数的分散是市场的特征；在一个价格围绕该函数广泛分散的市场中，任何单一数字的估计值——无论采用何种方法——都会伴随着极大的不确定性。若在陈述数值时未提及这一波动范围，就会使结论显得比市场实际情况更为精确。

估价计算示例。第 1 期中列出的 8 个 Pacifica 可比实例是一个无噪声样本：它们是被构造出来以精确满足一个线性方程的。当将这八个价格作为因变量，以总可租面积（GLA）、地块、景观、房况和翻新指标为自变量进行回归分析时，每个价格都能精确还原到美元——没有任何偏差——且回归结果与构建这些价格所依据的方程完全一致，各项系数完全吻合：

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

此处的“状况”指的是本报告中的状况评分：这是根据该房屋所在的界定市场区域内周边房屋状况，对其状况所作的评估。该评分基于估价师对该市场区域的理解，且不应与在同一市场区域内执业的同行估价师的理解存在重大分歧。该评分范围为 0.00 至 10.00：即该市场区域内状况更差的房屋所占比例除以十。评分 5.00 意味着该市场区域内有一半房屋的状况更差；9.00 则意味着十户中有九户状况更差。第 1 项指标用文字描述房屋状况，并在回归分析中将其编码为 1–3；根据本期刊的评分标准，文字描述对应“一般”5.00、“良好”7.00、“极佳”9.00，因此第 1 项指标的每一步对应 2 分，而市场每分支支付 \$20,000。最后一项是构建数据赋予翻新工程的价格差异：\$60,000，比具有相同 GLA、地块、景观和状况评分的房屋高出这一金额。这并非翻新行为本身的价格；它代表了翻新带来的、而其他特征未记录的任何因素（条目 008，遗漏变量偏差）。

这正是针对构建的 Pacifica 市场得出的 CEF。必须明确刚刚描述的是何种对象，以及为何能够如此精确地进行描述。该数据集是经过构建的，而构建的数据集实际上可以不含噪声；真实数据集则永远不可能不含噪声。构建该数据集的优势在于，它彻底消除了价格中的随机噪声——每个价格都精确符合公式——因此，在第 008 条（遗漏变量偏差）至第 010 条（参数识别）中，估计值与其目标值之间的任何差异都不可能源于噪声：每种情况要么是偏倚，要么（在参数识别中）是识别失败。

该示例将“状况”视为一个带有美元费率的特征，这与传统评估方法以及第 1 期中的做法一致，而条目 008（遗漏变量偏差）、009（内生性）和 010（参数识别）则指出了这种做法存在的问题。传统评估方法并非本期刊采用的方法——基于残差约束法（RCA）的评估。在残差约束法中，回归分析仅采用可测量的特征；“车况”和“质量”则被有意排除在外。回归分析无法解释的销售价格部分即为残差。在编辑所关注的加利福尼亚市场中，可测特征通常能解释销售价格变异性的约 80%，剩余约五分之一归因于残差。估价师将残差分配给状况、质量以及数据中未测量的其他特征，以估价师的评分作为参考，同时受制于一个约束条件：各残差成分价值的贡献之和必须等于可比实例的残差。价值结论并不取决于估价师如何进行分配——RCA 论文的证明 II（Craytor 2026，第 36 页）：在“所有残差成分价值的贡献之和等于可比实例的残差”这一约束条件下，改变残差的分解方式不会影响最终的价值结论。因此，状况因素的权重既从未被估算，也从未被主张；估价师赋予状况因素的权重，是指在已测量的特征（包括房龄）所反映的价值之外，状况因素所贡献的额外价值。

如果我们将第 1 期回归与该条件期望函数（CEF）进行比较，则可以对该回归作出判断——不过首先必须明确一点区别。一般而言，普通最小二乘法（OLS）回归并不会重建其目标的条件期望函数；相反，它会找到该函数的最佳线性近似，即线性投影，而这两者仅在条件期望函数（CEF）为线性时才一致。像 MARS 这样的非参数方法并不要求条件期望函数（CEF）为线性：它通过分段方式对条件期望函数进行逼近，在曲面弯曲处设有铰链（比率发生变化的点），在曲面扭转处设有交互项（取决于另一特征的比率），从而遵循一条非线性条件期望函数，而线性投影会将其展平。这仍然是对同一对象的近似，后续条目中提到的变量遗漏、内生性和识别问题也以完全相同的方式与之相关。上述五变量 CEF 是线性的，这就是为什么包含翻新指标的回归能够精确恢复该 CEF 的原因。以四项特征为条件（不包括翻新指标）的第 1 期回归，估算了 $E[P | \text{GLA, lot, view, cond}]$

的最佳线性近似值，并返回了基于《期刊》残差重新表述的条件结果（第 1 期印出的截距为 80,777 和 49,824；拟合结果相同），

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

每个系数都与它上方的系数不同。一旦用直线拟合来描述回归，就无需再多做说明：我们有两个模型和两组数据。然而，在 CEF 方法下，可以得出两个精确的结论。首先，这两个回归是基于不同的变量集进行的，因此近似了不同的条件期望函数：

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

在一个市场上，如果某些已翻新和未翻新的房产具有相同的四个特征，那么在这几个点上，这两个函数的取值就不同。在这个小型的人工市场中，没有两处房产具有完全相同的四项特征，因此这两项函数在每处房产上结果一致，但在形式上有所不同：第二项函数在其特征上呈直线关系，而第一项则不是。在每种情况下，回归分析仅提供对其目标变量的最佳线性近似。这两个目标变量都是有效的分析对象。其次，只有那个包含五个变量的模型，其系数可以被解读为隐含价格——其中“翻新系数”即指在总可租赁面积（GLA）、地块、景观和状况评分均相同的条件下，翻新房屋与未翻新房屋之间的价格差——而条目 008（遗漏变量偏差）解释了其中的原因。

第 1 期中 Comp H 的残差恰恰说明了这种区别是多么显著。如果我们将第 1 期的八个属性视为整个市场的代表，那么每个四变量特征向量都与这八个属性中的一个完全对应，因此四变量 CEF 恰好通过所有八个价格——其中就包括 Comp H。Comp H 与四变量 CEF 之间的偏差为零；其第 1 期中的残差为 \$46,455，即其与拟合的四变量线性回归（即该八个房产市场的线性投影）之间的差异量——而翻新效果恰恰隐藏在该投影与 CEF 之间的差距中。采用五变量规格时，投影与 CEF 变得完全一致，偏差随之消失。该残差并非源自 Comp H 本身，而是源于模型规格。

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 Scryer. 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).

```