

残差约束法 (RCA): 框架与规程 (简体中文版)

本文为 2025 年 3 月 5 日最初发表于 Zenodo 的论文之更新版的简体中文译本。译文借助人工智能辅助完成并经编辑部审校；英文版为权威版本。

William Bert Craytor, SRA

第 1 卷 第 1 期 · 2026 年夏季

发表于 2026 年 7 月 1 日

DOI: [10.67330/kgebf46](https://doi.org/10.67330/kgebf46)

© 2026 William Bert Craytor. Published by Pacific Vista Net.

本开放获取文章依据[知识共享署名 4.0 国际许可协议 \(CC BY 4.0\)](#) 的条款发布。

摘要。传统估价方法的一个重大局限在于其依赖过时的手工技术，例如配对比较分析 (matched pair analysis)。这类方法对房地产特征的价值贡献视而不见：传统的销售网格 (sales grid) 没有价值贡献一栏，只有调整一栏。而事实上，价值贡献正是构建数学约束的必要基础——这些约束能够防止传统估价师的高估与低估。

残差约束法 (Residual Constraint Approach, RCA) 通过整合一个多阶段估价流程来推进传统的市场比较法 (Sales Comparison Approach, SCA)：该流程运用多元自适应回归样条 (Multivariate Adaptive Regression Splines, MARS)，并对特征价值贡献施加严格的数学约束，尤其针对未被测量的特征——诸如美学、设计、状况与质量等潜在变量——其价值在传统估价中通常受制于估价师的主观判断与偏差。

RCA 中的数学约束依赖于对 MARS 回归的专业运用，以估计可测量房地产特征对成交价格的价值贡献。然而，这一估计通常只能解释实际成交价格的约 80%，不包括设计、功能效用、状况和质量等主观评估的特征。在旧金山湾区 (San Francisco Bay Area)，代表这些主观成分的残差平均约占房地产总价值的 20%。80/20 这一数字反映的是作者在旧金山湾区典型数据条件下的经验，不应被解读为跨市场不变的常数。在可实现 R^2 较低的市场中，本框架的应对之道是识别并测量未被测量的价值驱动因素——例如沿海社区中与海洋的距离、山坡社区中的海拔高度，以及当 MLS 数据未提供时从外部来源收集的类似变量——而不是接受这一局限并报告更宽的误差区间。虽然在其他情境下 MARS 残差通常被视为估计误差，但在房地产估价中，残差是潜在变量价值的间接度量。尽管这初看似似乎不切实际，但已有证明表明：只要各描述性成分的贡献之和等于残差，残差就可以被有意义地分解为这些成分——且不影响最终的估价结果。

这一方法论源自作者二十年的实证应用经验，通过在 R 与 Python 环境中对 MARS 的迭代实现逐步发展而成。

关键词：房地产估价，住宅估价，残差约束法，MARS，多元自适应回归样条，市场比较法，残差分析，潜在变量，自动估价

目录

1 引言	3
2 背景	5
3 估价工程师	7
4 假设	8

5 RCA 工作流程	10
5.1 MARS 分析的类型	10
5.2 处理工作流程	10
5.3 数据工作流程	11
5.3.1 MARS 的数据要求	11
5.3.2 规程：已成交交易挂牌信息	12
5.4 项目文件夹结构	14
5.4.1 规程：多次 MARS (earth) 运行	15
5.5 核心 Excel 工作簿	16
5.6 备份	16
6 工作簿记号	17
6.1 描述维度	17
6.1.1 项目维度 (“a”)	18
6.1.2 运行版本维度 (“r”)	18
6.1.3 工作流阶段 (“w”)	19
6.1.4 Excel 工作簿工作表 (“s”)	19
6.1.5 房产挂牌信息 (“p”) ——MLSData 行	20
6.1.6 MLS 变量 (“v”) ——MLSData 列	20
6.1.7 最终记号示例	20
6.1.8 理由	21
6.1.9 矩阵 ²	21
6.1.10 第一次简化	21
6.1.11 第二次简化	22
7 RCA 销售网格	23
8 变量	26
8.1 变量来源本体	26
8.2 变量应用本体	27
8.3 二阶与三阶项	28
8.3.1 调整变量	28
8.3.2 汇总变量	29
8.3.3 特殊变量 & 记号	29
8.3.4 分解残差	30
9 RCA 核心证明	31

10 MARS 模型	32
10.1 模型变量与交互作用的“ g ”函数	32
10.2 g 函数的索引	32
10.3 绘图维度	33
10.4 最终模型方程	33
10.5 绘制一个模型所需的图形数量	34
11 数据特征	36
11.1 残差的用途	36
11.2 作为近似回归变量的状况与质量评级	36
11.3 数据错误	37
11.4 什么是准确性?	38
11.5 作为异常值的估价对象	39
11.6 CVR2 在应对意外数据方面的重要性	39
12 残差	41
12.1 排序与评分	41
12.2 CQA-残差曲线	43
12.3 CQA 曲线的特征	43
12.3.1 CQA 曲线各组成部分的分析	43
12.3.2 市场动态与社会含义	44
12.4 估价对象残差	44
12.5 RCA 价值结论	45
12.5.1 一般考虑	45
12.5.2 复核、审计与投诉	45
13 结论	46

1 引言

本文阐述了一种现代方法的逻辑与数学基础,该方法旨在准确估计住宅及其他房地产的市场价值(market value)。这一方法源自 20 年来将多元自适应回归样条(Multivariate Adaptive Regression Splines, MARS)软件应用于北加州住宅——尤其是旧金山湾区各社区——的实践经验,在估价精度上实现了显著进步。MARS 能够娴熟地捕捉房产特征与价值之间的非线性关系,同时高效地识别并优先处理影响最大的变量,从而提升估价质量。这一能力使估价师(appraiser)能够以远快于传统方法的速度发现新市场区域中的价值动态。MARS 构成了本文提出的残差约束法(Residual Constraint Approach, RCA)的骨干。RCA 方法之所以得名,源于其严谨的框架:

- 它对已测量和未测量房产特征的价值贡献(value contribution)均施加严格约束:

1. 所有由 MARS 导出的价值贡献, 包括残差 (residual) 在内, 其总和必须等于每个可比交易实例 (comparable) 的净售价。
 2. 残差可以分解为具有指定价值的子成分, 前提是这些子成分的总和等于该可比交易实例的残差。
- RCA 通过将主观特征——如美学吸引力、状况和质量——的总价值与残差挂钩, 最大限度地减少了估计此类特征时的偏差; 当由胜任的估价工程师 (valuation engineer) 构建 MARS 模型时, 残差能够被准确量化。

尽管 RCA 主要为住宅房地产设计——在此类物业中主观特征对市场价值影响显著——但它在更广泛的资产估价应用中亦具潜力。鉴于美国住宅房地产市场的估计价值约达 \$45 万亿, 准确的估价至关重要。Fannie Mae 指出, 典型估价与实际售价的偏差约为 5%, 但这一数字掩盖了偏差问题, 因为估价师往往面临与售价保持一致的压力。此外, 由于 60%–70% 的估价用于再融资——缺乏可比售价——RCA 所能提供的精度之必要性便显而易见。美联储 2012 年的分析估计, 2008 年危机前有 15%–20% 的估价与售价的偏差超过 $\pm 10\%$, 这一范围如今很可能仍适用于再融资估价, 并因物业类型和市场而异。准确性为何重要? 对于低贷款价值比 (LTV) 的交易——例如信用良好的买家在 \$1 million 的住宅上支付 \$500,000 首付——贷款机构可能免除估价, 认为止赎风险极小。然而, 跳过估价的买家有多付的风险, 往往在购买之后才发现缺陷。这凸显了 RCA 这类稳健估价方法的价值。量化估价的准确性仍然颇具挑战。研究者缺乏针对“市场价值”偏差的直接度量指标, 只能依赖间接代理。传统估价方法根植于主观判断, 且受估价师胜任程度参差不齐的影响, 使分析更加复杂。相关策略包括:

1. Tzioumis (2016) 研究了 11 个州 4,782 名估价师的 259,436 份估价报告, 发现三分之一的估价与合同价的偏差在 1% 以内, 一半高出合同价 1%–3%, 少数平均高出 5%——表明存在向上偏差。该研究识别出两类估价师: “按市场价值”型 (接近合同价) 和“高于市场价值”型 (持续偏高)。
2. Krivorotov and LaCour-Little (2020) 发现 90% 的购房估价达到或超过合同价, 凸显了“锚定偏差”——向某一期望价值靠拢的倾向——在再融资情境中尤为明显。

这些研究揭示了估价实践中的偏差, 而传统工具的局限性又放大了这种偏差。如果估价师普遍胜任且无偏, 他们的估价结果应当趋同; 某些群体持续的高估 (overvaluation) 表明存在系统性问题。

上述偏差研究及其指向的系统性模式, 反映了传统市场比较法 (Sales Comparison Approach, SCA) 实践的一个更深层特征, 而这正是 RCA 旨在解决的问题。传统方法在两个认识论地位截然不同的渠道上运作。在可测量特征渠道上——GLA、房间数、地块面积、房龄、位置——估价师被期望通过配对分析或对可比交易实例的局部线性回归, 客观地推导调整 (adjustment) 值。胜任的估价师会做这项工作, 由此得到的调整至少大体上能在数据面前站得住脚。但即便在这一渠道上, 传统的网格也只记录调整; 它甚至没有一列用于记录价值贡献。传统的调整也并非作为价值贡献之差来计算——它们是通过配对或类似程序直接估计出来的, 底层的价值贡献从未被计算过。相比之下, 在 RCA 中, 每一项调整都是估价对象 (subject property) 与可比交易实例在给定特征上的价值贡献之差——证明 II 正是精确刻画了这一结构的后果。在主观特征渠道上——状况、施工质量、设计、功能效用、美学——可用的支持程序从单薄到近乎空洞不等。主观特征调整可以通过对直接可比集合之外物业的配对分析、诉诸公认的同行惯例, 或引用估价师在该市场中有记录的经验来支持。这三者在形式上都符合专业标准。但没有一个能达到可测量渠道常规实现的统计严谨性。

其结果是, 标准的要求与方法所能提供的支持之间存在程序上的不对称。标准要求对每一项调整——无论可测量还是主观——都提供支持。而方法在一个渠道上提供稳健的支持, 在另一个渠道上则只有一系列较弱的程序, 从样本量为一的轶事式配对, 到引用其他估价师的习惯做法, 不一而足。诚实而胜任的估价师对景观调整进行细致的配对分析, 其提供的支持显著优于毫无依据地援引同行惯例者; 但在标准表格上比较这两份估价报告的审查者却难以区分二者, 因为

表格记录的是调整值，而非其推导的严谨程度。方法之所以容忍这种不对称，是因为对于主观特征，一直没有更强的程序可供执业估价师普遍使用。

RCA 对这两个渠道都有改进，但方式在范畴上截然不同。在可测量渠道上，MARS 回归能够捕捉配对分析和普通线性回归无法表达的非线性关系、变量交互作用和基函数（basis function）结构。MARS 模型是针对市场区域内全部合格销售数据（通常为数百宗）拟合的，而非传统网格中的三到九个可比交易实例，因此由 MARS 导出的价值贡献既客观，又比其所取代的方法准确得多。在主观渠道上，RCA 所做的事情与任何传统支持程序在性质上都不相同。它根本不试图为单个主观特征估计调整值。所有主观特征——潜在变量（latent variable）——的贡献共同表现为残差：即每个可比交易实例的实际净售价与其 MARS 预测价之差，可直接由数据计算，无需估价师输入。将该残差分解为具名成分——设计、功能效用、景观、状况、质量等——受到其总和必须等于残差总额的约束，并且根据证明 II（第 9 节），任何满足该约束的分解都会得出相同的价值结论。工作流程中剩下的判断，是将估价对象放置在可比交易实例残差排名分布中的位置，这一判断是对照可检查的排名做出的，并在报告中明确记录。一份薄弱的 RCA 估价报告仍然可能出现——缺乏经验的 VE 可能构建出过拟合的 MARS 模型、对估价对象做出不合理的定位，或忽视数据中的异常——但其薄弱之处是审查者通过检查模型和排名即可识别的，而不是标准表格审查所无法揭示的那种薄弱。

除再融资之外，估价师对业主负有提供准确估价的义务，尤其是在高价值购置中，错误可能造成数万乃至数十万的损失。然而，许多人不愿在估价上花费 \$600-\$1,000，怀疑其附加价值。RCA 这类先进方法虽然初期资源投入较大，却有望带来更高的精度。随着方法的改进和经验的积累，其效率——以及采用率——应会不断提高。

2 背景

作者的教育背景为数学、计算机科学与统计学。其职业为软件工程师，兼具估价与会计方面的辅助经验。他自 2002 年起断断续续从事估价工作，并自 2004 年起将 MARS 应用于住宅房地产的调整值确定和价格趋势分析。

使用 MARS 进行估价是估价领域的“深度经验”。MARS 生成的模型能比任何其他统计方法都更细致地反映哪些变量对市场价值有贡献，以及这些变量取值的变化如何影响售价。因此，在使用 MARS 估价的过程中，估价师不断考察由 MARS 模型揭示的变量取值与售价之间的因果关系。

出于透明性的考虑，RCA 方法最适用于具备以下特征的住宅区域：

1. 物业彼此差异很大的复杂社区

- 较老的住宅
- 更新程度参差不齐
- 多种不同的建筑风格与设计
- GLA、地块面积、海拔和景观等物业属性差异较大。

2. 复杂的市场状况

- 快速变化的市场状况
- 类型多样的买家

3. 良好的数据来源：RCA 规程（protocol）基于数据挖掘，需要充足的高质量数据。虽然它也可用于数据稀疏的区域，但可能不值得为此投入所需的时间和成本。

4. 需要准确的市场价值估计。RCA 在多数情况下应能提供 $\pm 1-2\%$ 的精度。估价工程师报告应就最终价值结论的精度条件加以说明。如果数据或信息不足，或未能通过数据挖掘发现

市场中的一致模式，精度就会受损，且应将这一情况告知客户。具体而言，较高的 R2 和 CVR2（交叉验证 R2，通过将 MLS 数据集反复随机划分为训练集和测试集计算得到）、有意义的 MARS 模型，以及排名残差中物业 CQA（状况-质量-吸引力）特征所呈现的强模式，是评估价值结论准确性与可靠性的主要手段。

过去 20 余年间这些 RCA 规程的发展，依托于 MLSListings Inc. (Sunnyvale, CA) 及其他目前可通过 MLSListings 访问的 MLS 数据，这些数据占作者数据分析经验的 95% 以上。作者的专业能力源自在北加州 15 个主要县开展的大量估价工作，并在 10 个次要县积累了额外经验。

RCA 方法在满足以下两个关键条件的市场中最为有效：

1. 高质量估价数据的丰富可得性
2. 复杂、异质的住宅区域，其特征为：
 - 多样的建筑风格
 - 程度不一的物业更新
 - 跨越数十年的开发模式

加利福尼亚州各县经验

本文所作论述基于在以下北加州各县开展物业估价的经验：

San Mateo	Santa Clara	Marin	Napa	Monterey	Santa Rosa
Alameda	San Francisco	Contra Costa	Sacramento	El Dorado	San Joaquin
Stanislaus	Placer	Solano	Mendocino	Lake	Colusa
Merced	Nevada	San Benito	Santa Cruz	Sutter	Yuba

上述区域基本涵盖硅谷及其周边各县，并向外延伸至更多乡村地区，包括中央谷地和内华达山麓。应当指出，与美国其他欠发达地区的许多估价师所能获得的数据相比，作者始终拥有充足且相对准确的数据。

所用的 MLS 数据经过 RESO 认证，符合 RCF (RESO Common Format)。尽管有 RCF 认证，一些物业的、通常属次要重要性的数据有时仍会缺失或有误，其原因或是最初税务评估员的录入错误，或是后续销售经纪人的 MLS 输入错误。这些错误往往遵循某种模式，例如特征不存在时干脆不做输入。举例来说，在没有车棚时，“carport spaces”一栏不输入 0，而是什么都不填。在车棚罕见的地区，分析师可以用 0 替代缺失数据，以便得到有用的车棚调整值。虽然 R/earth 及其他 MARS 实现可以处理缺失数据，但如何处理由估价工程师 (Valuation Engineer, VE) 决定。

MLSListings 的数据也可从其他一些州（从佛罗里达到夏威夷）的 MLS 获得。这对于测试规程和代码在不同州的运行效果很有用。

3 估价工程师

定义 1：估价工程师

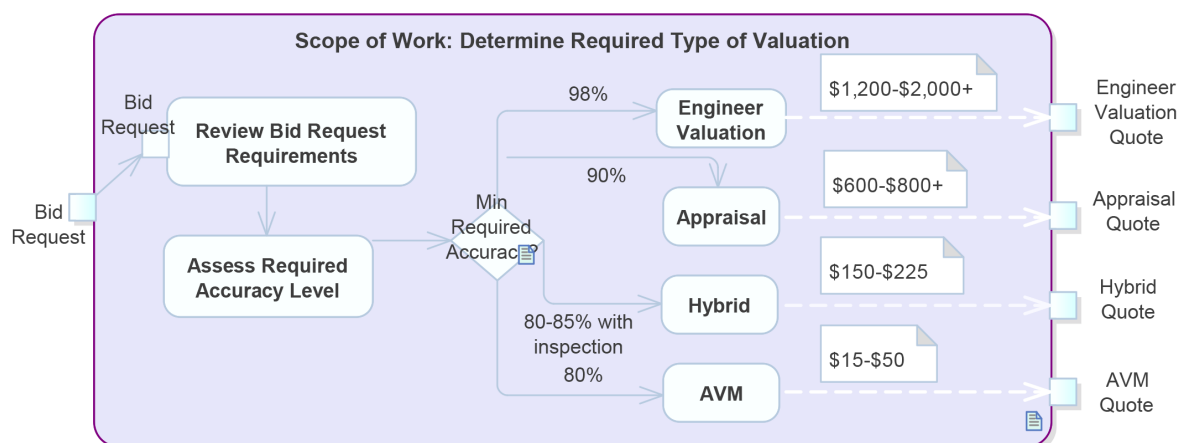
为避免歧义，“估价工程师”（Valuation Engineer，简称“VE”）一词指这样的人员：他既具备持牌估价师的专业能力，又在运用数据挖掘、统计学、编程、详细规程和人工智能来确定复杂物业市场价值方面拥有高级的知识、技能与经验。还应认识到，鉴于既已证实的欺诈可能性，估价是一项需要某种形式执业许可的工作。

贷款机构通常提供以下几种估价选项：

1. 自动估价模型（AVMs）——成本最低的选项
2. 混合估价——由无执照查勘员实地查勘，持牌估价师负责监督
3. 标准估价——由持牌估价师完成

通常不在估价工作考虑范围之内的，是工程师级估价。具备这类技能组合的估价师实在不多，但市场应当存在对他们的需求。借助 RCA 方法（MARS 加约束），他们能够提供显著更高的准确性，并几乎保证相当良好的客观性。

对于较新住宅小区中的合规物业，标准估价通常已经足够。然而，对于翻新历史长达 30 年以上区域中的老宅，则建议采用工程师估价。在这两种方法之间的选择应遵循基于物业特征和市场状况的结构化决策流程。



4 假设

假设 1: 地理区域 & 物业类型

本文的范围主要聚焦于加利福尼亚州大都市及城市地区的独户住宅与共管公寓。虽然市场比较法也适用于其他物业类型和地理区域，但在对农业用地、多户物业和商业房地产进行估价时，收益法和成本法通常发挥更主导的作用。在此类情况下，RCA 方法可能需要修改，或适用性有限。

假设 2: 数据存储

CSV 和 Excel 在 MLS 系统中被大量使用，并被假定为向 RCA 程序馈送数据的主要方式；当然，也可以使用经纪商数据源将数据持续送入 SQL 数据库。备份和中间存储常使用 SQL 数据库，其中扮演重要角色的两个是免费开源的 SQLite 和 PostgreSQL。文本或 RTF 文件也用于报告片段，PNG、JPG、TIFF 文件用于图形。需要注意的是，许多人也使用 Microsoft 的 SQL Server，但其免费版“SQL Server Express”限制很大（例如只能使用 1 个 CPU 核心），而其他版本用于商业用途则需支付高昂的许可费。加利福尼亚州的房地产数据库可能非常庞大，连接（join）操作可能非常耗时，因此对大型连接而言，一台 8 核以上、配备 128G 内存的计算机会有用。当需要查找任意边界区域内的所有物业时，SQL 数据库同样重要。因此，举例来说，如果使用聚类分析来划定社区，就会倾向于使用某种 SQL 数据库。然而，不同数据库处理地理信息的内置函数各不相同，因而互不兼容。在这一系列关于 RCA 的文章中，复杂的 SQL 任务假定使用 PostgreSQL，简单任务使用 SQLite。

假设 3: MARS 处理使用“R/earth”。

2017 年之前，Salford Systems 的 MARS 软件是一个价格亲民的解决方案，每年 \$270。然而，在 Minitab 收购 Salford Systems 之后，MARS 被打包进其 SPM 软件包，年费高达 \$16,000——这一价位令大多数小企业望而却步。幸运的是，由 Stephen Milborrow 维护的 R 语言“earth”包 (Milborrow 2024a) 已发展到与 Minitab 的 SPM MARS 相当乃至更强的水平。earth 包还具有额外优势：它免费可得，与 R 全面的统计生态系统无缝集成，并可通过 Python 借助 Pandas 访问，为不同工作流程提供了出色的灵活性。与 MARS 相关的讨论均假定具体使用 R/earth。

假设 4: 符号约定

来自 CSV 文件和 Excel 电子表格等外部来源的原始数据用标准拉丁字符表示。这包括多重挂牌服务 (MLS) 物业数据、使用 R/earth 包的 MARS (多元自适应回归样条) 配置设置，以及项目特定数据。这一记法同样适用于它们在 R 或 Python 中对应的数据框表示。在讨论物业特征及其各种类型和子集时，记法改用希腊字母，以增强精确性和分析上的清晰度。虽然这种数学记法看起来更为严格，但它有一个关键目的：将抽象的物业特征与其原始数据表示区分开来。按照惯例，希腊字符始终表示物业特征，而拉丁字符表示电子表格数据结构。例如，符号 M 表示 MLS 数据矩阵，其行为各物业，列为相应的特征。

假设 5: 编程语言

在本领域，R 和搭配 Pandas 的 Python 是主导语言，在对速度要求较高时则使用 C 或 C++。R/earth 本身即以 R 和 C 编写。

假设 6: 索引

索引方式在不同编程语言与数学之间可能各不相同，这可能令人困惑：

- 数学和 R 从 1 开始索引。
- C、C#、C++ 和 Python 从 0 开始。

在本文中，为保持一致和简单，所有索引均从 0 开始（如 Python 或 C）。具体规则如下：

- 对于列表、矩阵或数据框：
 - 索引 0 保留给“估价对象”（即被研究的主要物业）。
 - 示例：在以行表示物业的表格中，第 0 行是估价对象，第 0 列存放所有物业的唯一 ID。
 - 实用提示：排列物业时将估价对象放在最前（第 0 行），其余物业按销售日期从新到旧排列，并按顺序编号（0、1、2 等）。
- 当没有估价对象时：
 - 如果数据结构不包含物业（例如仅有变量或函数），则跳过索引 0，索引从 1 开始。
 - 给 R 用户的特别说明：由于 R 使用 1 基索引，在用 R 实现本文的数学内容时，需将所有索引加 1，以从 0 基调整为 1 基。

5 RCA 工作流程

理解 RCA 工作流程是进行有效分析的关键。整个过程遵循一系列结构化的分析任务与决策节点

5.1 MARS 分析的类型

MARS 分析的工作流程取决于仅需进行市场分析, 还是需要对特定房地产进行估值或估价。请注意, 对特定房地产的估价通常需要对其所在市场区域进行分析。

- 市场分析: 侧重于县、市或界定区域等较大范围。这种方式不涉及具体的估价对象, 因此无需得出价值结论。工作流程以模型开发、验证以及生成配套分析结果为核心。一旦确定并记录了最优模型, 流程即告结束。
- 单一房地产分析涵盖市场分析的全部内容, 并进一步确定特定房地产的价值。这与传统估价方法有着根本区别。在传统方法中, 价值通常通过对 3-9 个可比交易实例进行调整并计算加权平均值得出。相比之下, RCA 主要通过两个部分来确定价值:

1. MARS 模型的预测值
2. 估计的房地产特定残差 (连同相应的 CQA 评分)

在 RCA 中, 销售网格的主要作用是通过展示各变量的贡献, 来解释估价对象与可比交易实例之间的价值差异。RCA 的一个关键优势在于其稳定性: 尽管客户可能会对可比交易实例的选择提出异议, 但无论选择哪些可比交易实例用于展示, 价值结论都保持一致, 因为所有成交案例 (无论是 15 个还是 600 个) 都是使用同一模型进行调整的。虽然较早的成交案例会被自动调整至当前市场状况, 但一般而言仍以近期成交案例为佳。随着时间跨度的拉长, 未被测量的市场特征变得愈发重要, 模型捕捉价格因素的效果也随之减弱。

5.2 处理工作流程

整体处理工作流程可以用以下步骤来描述:

1. 工作范围。
2. 建立初始项目文件夹与文件。
3. 下载 MLS 数据。关于所需房地产记录数量以及日期范围限制的设定, 请参阅规程 1.1-1.3。
4. 完成各配置工作表: 回归变量、交互项、聚合项、计算项。
5. 运行 MARS (R/earth) 回归以建立价格模型。
6. 由 MARS 模型生成各房地产的价值贡献与残差 (纯残差以及残差/SF)。
7. 按残差/SF 对房地产进行排序。
8. 为每个可比交易实例创建 CQA 评分。
9. 假设已经生成了一份真实反映各房地产整体吸引力的质量排序, 则通过研究排序中的总体模式趋势, 为估价对象在排序中找到最合适的位置, 并为其赋予 CQA 评分。
10. 将相应的残差/SF 赋予估价对象, 即残差/SF · (估价对象) GLA。
11. 将 MARS 价格估计值与所赋残差相加, 得出估价对象的价值结论。
12. 将估价对象的残差分解为描述性组成部分及其价值贡献。

13. 计算可比交易实例由模型产生的价值贡献与调整额, 以及所有小计和调整后的成交价格。
14. 调整后成交价格将等于估价对象的最终价值结论。
15. 选取 6–12 个最相似的房地产纳入销售网格。
16. 将可比交易实例的残差总额分解为残差组成部分的价值贡献, 分解完成后将残差总值设为 0。
17. 重新计算所有小计以及销售网格中可比交易实例的调整后成交价格。核对数值。
18. 将销售网格上传至报告。

请注意, 采用 RCA 时, 最终价值结论并非直接来源于销售网格中的可比价值。在 RCA 方法下, 销售网格仅仅成为一种解释工具, 用于说明为何那些看似与估价对象相似的房地产, 其成交价格会高于或低于最终价值结论或估价对象的评估价值。必须换一种方式来看待这一点: 估价对象的价值结论取决于 MARS 模型的价格估计值, 以及估价工程师所估计的残差。但这两个量在很大程度上依赖于参与 R/earth 回归的全部房地产, 该回归生成了房地产残差的排序, 而估价工程师通过将估价对象置于该排序中的某一位置, 确定了其 CQA 评分, 进而确定了其残差。销售网格中的调整实际上是由 RCA 所施加的数学约束决定的; 尽管残差分解在一定程度上取决于估价工程师的判断, 但分解出的价值贡献合计必须始终等于相应的残差。我们从证明 2 可知, 虽然残差各组成部分的实际数值可能因估价工程师的判断而异, 但对于给定房地产, 其合计值是完全受约束的, 因此无论估价工程师如何改动残差分解, 都不会改变该房地产的调整后成交价格。

5.3 数据工作流程

请注意, 本文仅对 RCA 方法给出高层级的工作流程。其他文章将深入探讨 RCA 分析的各个阶段与子阶段。

5.3.1 MARS 的数据要求

我们需要指定销售交易的起始日期, 并根据估价或估值的基准日确定截止日期。这通常需要考虑每一交互度 (大于 1 时) 所需回归变量与交互项数量的估计值。

5.3.2 规程：已成交交易挂牌信息

规程 1.1：确定所需 MLS 已成交交易挂牌信息的数量

本规程说明如何计算输入 MARS 回归所建议的已成交交易挂牌信息的最低数量。你需要先做一个粗略估计，随后运行一阶、二阶乃至三阶的 MARS 以生成初始模型，然后按照下述规则审查数据行数，判断是否需要更多记录。本规程假设各自变量之间是 100% 相互独立的，而这在房地产领域几乎不可能成立；例如，GLA 与房间数量通常高度相关，GLA 也很可能与地块面积存在相关性。这就涉及到“自由度 (DOF)”的计算。每一个真正独立的变量占一个自由度。但是，当所谓的自变量之间存在依赖关系时，准确计算 DOF 需要进行一定的分析，这已超出本文的讨论范围。只需记住，由于变量之间很可能存在共线性，下述估计值可能偏高。

1. 对于一阶项（无交互项），每个变量约 ~ 15 行的经验规则效果良好
2. 对于每一个允许的双向交互项，应额外增加约 15-30 处房地产
3. 对于更高阶的交互项（3 个及以上变量），每项应有 30 处以上房地产。

不过，这并非硬性规则。实际所需数量可能取决于：

- 数据中的信噪比
- 所要建模的关系的复杂程度
- 响应变量的变异程度
- 数据的质量

示例：

因此，如果你有：

- 10 个主变量
- 5 个允许的双向交互项
- 2 个允许的三向交互项

则可按如下方式计算所需成交案例的最低数量：

$$\begin{aligned} N &= (10 \text{ 个变量} * 15 \text{ 行/变量}) + \\ &\quad (5 \text{ 个双向交互项} * 22 \text{ 行/双向交互项}) + \\ &\quad (2 \text{ 个三向交互项} * 30 \text{ 行/三向交互项}) \\ &= 320 \text{ 行} \end{aligned}$$

或者，也可以采用以下做法：在对一组数据运行 R/earth 之后，可以用模型中基函数的数量加一作为自由度的估计值，再将该数值乘以 20-30，即可得到所需输入记录的数量。

规程 1.2: 销售期间应尽量避免购房者偏好发生变化

在收集历史房地产数据时，如果市场偏好的变化能够通过适当的变量加以捕捉，则对时间跨度没有严格限制。在处理市场中建筑风格演变时，请考虑以下要点：主流建筑风格的变化（例如从加州牧场式向地中海式的转变）可以通过以下方式建模：

- 小区 (subdivision) 标识符
- MLS 区域代码
- 建造日期区间

然而，如果这些风格变化无法通过特定变量得到充分捕捉，你将面临一个重要约束：销售日期通常用于调整市场状况，若同时用它来控制建筑风格偏好，会使这两种截然不同的效应相互混淆。在这种情况下，较为可取的做法是二者择一：

1. 将数据收集范围限制在较新建筑风格出现之后的时期
2. 将可比交易实例的选择限制在具有相同建筑风格的小区之内

这样既能确保分析在房地产特征上保持一致性，又仍然允许对市场状况随时间的变化进行恰当调整。或者，至少在这种情形下，假设你能够从数据中提取出建筑风格，可以考虑这两个变量之间的交互作用。然而，这种提取并非总是可行的。

规程 1.3: 若处于成交稀少的区域，则尽力而为

在使用 MARS 对销售数据建模时，进行有意义的分析大约需要 15 笔交易。如果成交案例更少（例如只有 5 处房地产），可以将每条记录复制两次，以达到 15 处房地产的门槛。采用这种做法时的重要注意事项：

1. 仅使用一阶回归，以降低过拟合
2. 由于数据集过于有限，忽略或跳过交叉验证
3. 在这种情形下，MARS 的表现很可能优于配对比较法或标准线性回归方法，但要注意，复制数据会影响不确定性度量与统计显著性的有效性

最好的长期解决方案是通过补充真实成交案例来扩充数据集。不过，如果必须在数据有限的情况下继续开展工作，数据复制在承认其统计局限性的前提下，不失为一种可行的临时方案。

5.4 项目文件夹结构

有必要讨论一下项目文件夹结构。建议为市场分析和单一房地产估价分别设立单独的文件夹。这两个文件夹具有相似的子文件夹结构：

```

RCA_Projects/
├── MarketAnalysis/
│   ├── Burlingame_20240501/
│   ├── Pacifica_20240310/
│   ├── RedwoodCity_20240615/
│   └── ...
├── SingleFamily/
│   ├── Pacific_Rosewood_123_20240510/
│   │   ├── Code/
│   │   │   ├── R/
│   │   │   │   └── ...
│   │   │   └── Python/
│   │   │       └── ...
│   │   └── Data/
│   │       ├── MLS_Original.xlsx
│   │       ├── MLS/
│   │       │   ├── VER202402510153248/
│   │       │   │   ├── MLS.xlsx
│   │       │   │   ├── MLS_Stage2.xlsx
│   │       │   │   ├── MLS_Stage3.xlsx
│   │       │   │   ├── MLS_Stage4.xlsx
│   │       │   │   ├── MLS_Stage5.xlsx
│   │       │   │   └── Documents/
│   │       │   │       ├── LinearModels.pdf
│   │       │   │       ├── Cum_Dist.pdf
│   │       │   │       ├── Model.txt
│   │       │   │       ├── PartContrib.pdf
│   │       │   │       ├── Pairs.pdf
│   │       │   │       ├── Residual1.png
│   │       │   │       ├── Residual2.png
│   │       │   │       ├── ResidualSF.png
│   │       │   │       ├── Rpart.pdf
│   │       │   │       ├── VariablesImp.pdf
│   │       │   │       ├── VariableVsSP.pdf
│   │       │   │       └── ...
│   │       └── ...
│   └── ...
└── ...

```

第 1 阶段处理将创建版本文件夹及其子文件夹。它会复制 `MLS_Original.xlsx`——尽管该电子表格理应永不更改，而是精确反映在某一日期和时间从 MLS 下载的原始内容；但在周期较长的项目中，仍有可能需要重新从 MLS 下载数据。接着，此阶段会将经过扩展的 `MLS.xlsx` 工作簿连同其配置参数复制到版本文件夹中。版本文件夹可以有很多个，每个都对应一次完整的 MARS 运行。

随后的各个阶段将对工作簿进行列的添加、排序及其他更改，并在估价完成时将更新后的版本以 `MLS_Stage(N)` 的形式写入项目数据文件夹。再次强调，根目录下的 `MLS.xlsx` 处于持续变动之中，因此版本副本是对原始 MLS 工作簿的一种持续演化，包含变量的删除与添加以及其他更改。处理用的 R 或 Python 代码很可能以 Data 文件夹中的 `MLS.xlsx` 工作簿作为输入运

行。这样，我们始终保有每一版 RCA 运行背后全部数据的副本。

版本可以根据需要不断创建，直至无法再做出改进为止，此时选出最佳版本作为最终版本。可以在文件夹名称末尾加上“_Final”，以便浏览目录时一目了然。一些包含有趣模型或竞争性模型的旧版本通常应予保留。每个版本文件夹都包含若干文本、图示和图形文件，可嵌入到报告中。最终的 `MLS_Stage5.xlsx` 工作簿包含报告所需的全部工作表。特别是，它可能包含一张用于转入 Fannie Mae 销售网格的电子表格。它还必然包含聚合项背后的计算或分解过程。这可能是 RCA 方法中最耗时的部分——在一个困难的市场区域中苦苦寻找完美模型的过程。

重建销售网格所需的一切都必须复制到版本文件夹中，包括所使用的 R 或 Python 代码——如果代码经常变动的話。

5.4.1 规程：多次 MARS (earth) 运行

规程 2.1：利用过拟合运行精细剖析市场区域的定价因素

在最初的 earth 运行中，应当有策略地利用过拟合来发现潜藏的价格因素。例如，一个由 800 笔成交案例构成的市场区域中，可能只有 3 处房地产带有工房 (workshop)，而在这些孤立案例中，工房可能对最终成交价格产生重大影响。仅执行前向传递的运行会产生一个过拟合模型，从中可以看出这些工房的近似价值贡献。然而，由于 800 笔成交中只有 3 笔这样的案例，工房相关项将在后向传递中被剔除，在交叉验证下更是必然如此。可以将仅前向传递赋予工房的价值记录下来，转入潜在变量分析，供销售网格中出现的任何带工房的房地产使用。此类过拟合模型还能在许多其他数据稀薄的特征给出详细估计，而这些特征本会被后向传递剪枝或交叉验证运行所剔除。

将每一次这样的探索性运行都保存为独立的版本文件夹：所记录的价值贡献是估价工程师日后分解残差时会想要引用的证据。当然要明白，估价工程师最终必须依靠剪枝和交叉验证，才能使最终模型对回归从未见过的房地产——例如后来出现的估价对象——保持稳健。过拟合运行是用于研究的工具；它们绝不是最终模型。

规程 2.2：在不同 earth() 参数下进行多次交叉验证运行

市场区域在结构和规模上各不相同，估价工程师通常需要通过一系列运行来进行试验：以了解哪些变量对价格影响最大；判断现有的已测变量集合是否足以估计价值，抑或必须补充新数据（如距水域的距离、海拔、距购物中心的距离等）；并发现新型的交互关系，同时将其他交互关系作为低效或具有误导性的项予以排除。最终，估价工程师希望在训练数据上获得良好的 R^2 ，并在测试数据上取得尽可能好的 CVR^2 。要达到这一目标，可能需要收集新数据、提高现有数据的准确性，或根据市场区域的规模与特点（例如针对可用的自由度——参见规程 1.1）调整 earth() 参数。

这正是上述项目结构支持多个运行版本文件夹的原因：每个候选模型都连同其参数及其 R^2/CVR^2 一并保留，从而可以比较各次运行之间的演变轨迹。若某一版本的 R^2 上升而其 CVR^2 停滞甚至下降，则说明该版本拟合的是噪声而非市场信号，而正是版本历史使这一点得以显现。

规程 2.3: 利用 Prolog 展开备注及描述性特征列

尽管这一功能仍处于一定的试验阶段, earthUI 现已内置 Prolog 引擎 (vProlog), 并使用定子句文法 (DCG) 来解析 MLS 的自由文本列——首先是 Public Remarks (公开备注), 此外还有经纪人备注以及以逗号分隔的特征列表, 例如房间特征或既有建筑物。这种展开会将“owned solar”“3 car garage”“40x50 workshop”之类的短语转化为可供回归使用的普通特征列, 而以此方式挖掘出的数据稀薄特征可直接输入潜在变量分析 (参见后文关于将状况与质量评级作为近似回归变量的小节; 该功能本身在本期的 earthUI 配套文章中有详细说明)。

文本展开会成倍扩大候选预测变量集合, 因而必须进行多次 earth/RCA 运行加以测试: 展开出的列必须在模型中反复取舍试验, 在存在共线性时合并为派生的计数或布尔变量, 并对照 CVR2 检验其承载的是市场信号而非噪声。应将每一次此类试验视为独立的运行版本文件夹, 与规程 2.1 和 2.2 的做法完全一致。此外, 由于解析文法本身仍在开发之中, 估价工程师在将提取出的列用于模型之前, 应先将其与源文本进行核对。

5.5 核心 Excel 工作簿

由于大多数 MLS 服务将房地产数据下载为 Excel CVS 工作表, Excel 可能是存储数据及各类配置参数最便捷的方式。每个项目都有一个名为 MLS.xlsx 的核心 Excel 工作簿, 其中包含一张名为“MLSDData”的工作表, 用于存储从 MLS 下载的已成交房地产销售挂牌信息, 此外还有若干配置与参数工作表:

```
Workbook: MLS.xlsx/
├─ Sheet: Project Data
├─ Sheet: MLSDData
├─ Sheet: Regression Variables
├─ Sheet: Calculations
├─ Sheet: Allowed Interactions
├─ Sheet: Aggregations
└─ Sheet: (MARS run parameters)
```

原始数据存储在一个名称类似 MLS_Original.xlsx 的工作簿中。应当以不同名称——比如“MLS.xlsx”——制作一个副本, 由 R 或 Python 的 RCA 程序加载。这两个工作簿都存放在 Data 文件夹下。此后, Data/MLS.xlsx 工作簿很可能会经历多轮迭代运行, 尝试不同的变量组合、MARS 参数以及其他设置的更改。每次迭代都会创建一个以日期和时间命名的新输入/输出文件夹, 格式为 VER_YYYYMMDDHHMMSS。每个文件夹都会保存输入的 MLS.xlsx 工作簿的一个副本, 并包含四个处理阶段 (第 2-5 阶段) 各自生成的 Excel 工作簿, 以及输出的文本文档、图示和图形。请注意, 最后一个阶段即第 5 阶段, 可能会根据客户类型和工作范围 (SOW) 包含若干供报告使用的其他工作表。

5.6 备份

在将新建或修改后的 Excel 电子表格写入项目版本文件夹时, 还可以在项目文件夹中向 SQLite 或 PostgreSQL 数据库进行备份。数据库文件在后续时期的审阅或参考使用过程中, 较不容易被意外损坏。

6 工作簿记号

6.1 描述维度

工作簿“M”大约有六个描述维度：

1. 估价项目“a”
2. 运行版本“r”
3. 工作流阶段“w”
4. 工作表“s”
5. 房产挂牌信息“p”
6. 变量集“v”

因此，特定工作簿 M 中的某个单元格可以描述为：

$$\begin{matrix} r_* \\ w_* \end{matrix} \begin{matrix} a_* \\ W \\ s_* \end{matrix} \begin{matrix} p_* \\ v_* \end{matrix} \quad (1)$$

(注：“*”表示某个有效索引)

或：

$$\begin{matrix} r_k \\ w_m \end{matrix} \begin{matrix} a_g \\ W \\ s_h \end{matrix} \begin{matrix} p_i \\ v_j \end{matrix} \quad \begin{matrix} i = 1, \dots, n_p \\ j = 1, \dots, n_v \\ k = 1, \dots, n_r \\ m = 1, \dots, n_m \\ g = 1, \dots, n_g \\ h = 1, \dots, n_h \end{matrix} \quad (2)$$

其中：

- n_p = 行数（取决于工作表）
- n_v = 列数（取决于工作表）
- n_r = 运行版本数
- n_m = 工作流阶段数
- n_g = 分析项目数
- n_h = 工作簿工作表数

请理解，“W”指的是整个 Excel 工作簿，它包含六个核心数据工作表，每个工作表都有自己的表格；在运行相应的 R 或 Python 程序时，这些表格会被加载到各自独立的内部数据框中。R 和 Python 都有内部函数，可以根据名称查找它属于数据框的哪一行或哪一列。例如，如果在下标 v_j 中， $v = (\text{“GLA”}, \text{“LotSize”}, \text{“BathRms”}, \text{“BedRms”}, \dots)$ ，且 $j=0$ ，则 $v_j = \text{“GLA”}$ 。R 或 Python 都能够在内部将“GLA”解析为其所存储的列，例如，这可能是相应数据框的第 15 列。因此，在本文中我们对名称数组进行索引，把名称传递给运行中的程序，程序再将该名称重新索引为内部数据框中实际的行或列索引。这有点复杂。但为了精确起见，这样做是必要的，也是最方便的。这是分析师角色与开发者角色之间的“职责分离”；尽管在现实中，估价工程师 (VE) 可能同时承担这两个角色。请注意，R 中的“data frame”在 Python 中称为“dataframe”。

另请注意，MARS 运行时参数工作表并不是 W 的一个维度，因为 W 关系到输入 MARS 的数据版本，而与决定 MARS/earth 如何处理这些数据的参数无关。

为清晰起见，本文给出了一套复杂的多维工作簿记号，它与数据工作簿各工作表所加载到的内部 R 或 Python 数据框更直接地对应。这是一套“沉重”的记号。但当我们进入证明等环节时，通过对所处理对象作出假设，就可以抛掉大部分包袱。尽管如此，这套记号有助于我们从一开始就更好地总览工作流和数据。

我们先从将要处理的工作簿集合的记号开始。

6.1.1 项目维度 (“a”)

一旦工作范围 (SOW) 获得批准, 就需要建立一个项目文件夹来存储输入和输出数据, 包括配置参数、输出图形、图表、电子表格和报告片段。VE 应当用简单的顺序 ID (例如整数) 标识所有项目, 这样一旦某份报告丢失, 就会一目了然。如果项目被取消, 应当有一致的方式在相应项目文件夹内注明该项目已被取消, 但绝不应删除该文件夹。我们假设项目 ID 从 0 开始并逐次加 1, 尽管也存在其他可能。我们进一步假设所有有效的项目 ID 都存储在由处理程序创建的有序集合 “a” 中。

估价项目数据位于工作簿的项目工作表中, 但也可以存储在某个中心 SQL 数据库中,

以下是建议的项目字段:

1. Local ID: 为公司顺序生成的短整数 ID。它非常适合作为存储用的主键——并且可以由数据库在每个后续项目中自动生成。
2. Global GUID: 如果需要, 可以使用以下网址创建一个 GUID (全局唯一 ID)

<https://www.uuidgenerator.net/>

3. 房产位置 (市场分析) 或地址
4. 估价生效日期
5. 委托日期
6. 查勘日期
7. 原始报告日期
8. 第一次更新报告日期
9. 第二次更新报告日期
10. 第三次更新报告日期
11. 客户名称
12. 客户地址
13. 贷款机构名称
14. 贷款机构地址
15. 业主名称
16. 业主地址
17.

项目 ID 列表示例:

```
a = ("1",
      "2",
      "...",
      "141",
      "142")
```

6.1.2 运行版本维度 (“r”)

工作流的第二步是对输入执行 MARS 回归。这很可能需要数次乃至多次执行, 因为我们会调整各种参数以改进模型。每次执行都会为输出创建一个新的版本文件夹, 其名称基于执行的日期和时间。例如,

```
r = ("ver20241210_083124",
      "ver20241209_151130",
      "ver20241208_160944")
```

latest_version : $r_1 = \text{"ver20241210_083124"}$

具有竞争力的模型和输出的有希望的版本应当保留，直到确定最终版本为止；如果它们仍有某些用处，或许还可以保留更久。你可能希望修改版本文件夹名，在名称末尾加上“Final”，以表明它是最终版本。

6.1.3 工作流阶段 (“w”)

RCA 工作流通常分为五个独立的阶段，如下所列。每个阶段必须成功完成后才能开始下一个阶段。VE 应能够在某个阶段结束时停止处理，先审查结果再继续。这是必要的，因为市场区域相当复杂，而且往往颇为独特。可能会遇到需要采用新方法的异常问题。VE 必须能够进入 R 或 Python 代码进行必要的修改，甚至尝试新方法。而且在这种情况下，他可能希望进行试验并尝试不同的方法，这就需要对他正在处理的同一阶段进行多次运行。当然，VE 也应当可以用一条命令运行所有阶段——但有一些前提（例如，三阶聚合必须已通过之前的 Stage 2 运行确定）。这意味着每个阶段完成时，都应把其全部内部数据存入文件或数据库，以便下一个阶段可以通过把上一阶段的数据重新加载到内存中来启动，就好像上一阶段仍驻留在内存中一样。

以下是各个阶段：

```
w = ("stage1", -设置
      "stage2", -MARS 处理
      "stage3", -最终销售网格
      "stage4", -报告片段
      "stage5") -审查/清理
```

stage2 : $w_2 = \text{"stage2"}$

6.1.4 Excel 工作簿工作表 (“s”)

在 RCA 核心 Excel 工作簿中，我们有以下工作表，每个工作表都映射到一个内部数据框：

```
s = ("MLSDData",
      "Regression Variables",
      "Calculations",
      "Interactions",
      "Aggregations 1st & 3rd Degree",
      "Aggregations 2nd Degree")
```

之所以把二阶聚合放在单独的工作表上，是因为它们最容易通过一个对称表来定义：该表展示所有可能的交互对，允许为每一对指定一个聚合变量，将其相应的价值贡献和调整聚合（相加）到该变量上。三阶聚合无法这样处理，因为那需要一个三维对称表，而这在 Excel 中是不可能的。此外，一个 20x20x20 的表将有 8,000 个单元格需要处理。更简单的做法是把任何三阶项的名称直接列在包含一阶聚合的表中，例如：

Variable	Aggregate To
AboveGradeFinArea	GLA
BelowGradeFinArea	BGLA
UnfinArea	UBA
LotSize	LotSize
...	...
GLA___LotSize	GLA
GLA___DateOfSale	GLA
BathRms___DateOfSale	GLA
Lat___Long	Location
Lat___Long___GLA	Location
...	...

6.1.5 房产挂牌信息 (“p”) ——MLSDData 行

房产挂牌信息构成 MLSDData 工作表的行。它们不会出现在其他工作表中，直到 Stage 4——在该阶段，将使用最终选定的可比交易实例生成最终的报告片段。以下是房产 ID 的示例列表：

$$p = (\text{“MLS823211”, “MLS823456”, “MLS823542”, “MLS823721”, “...”})$$

6.1.6 MLS 变量 (“v”) ——MLSDData 列

关于变量的处理，后文还会详述。此处只需指出：MLSDData 工作表通常有 30-40 个类型各异的变量，但通常只有 10-20 个会被选作 MARS 回归的自变量。这些变量也出现在其他工作表中，不过通常是作为可供提交给 MARS 的所有可能变量的列表。变量列表的一个示例是：

$$v = (\text{“GLA”, “SalePrice”, “BedRms”, “BathRms”, “GarageSF”, “...”})$$

6.1.7 最终记号示例

让我们把上述记号汇总成一个最终示例。使用上面的列表，令

$$r_1 \begin{matrix} a_{141} \\ W \\ w_1 \end{matrix} \begin{matrix} p_2 \\ v_0 \end{matrix} \begin{matrix} s_0 \end{matrix} \quad (3)$$

请注意，这里我们使用 0 起始索引，即假定列表的第一个元素的索引为 0。Python 也使用 0 起始索引，而 R 使用 1 起始索引。

因此，上式表示我们所讨论的是以下单元格的值：

- 变量 v_0 ，即列 “GLA”
- 房产 p_2 ，即 ID “MLS823542”
- 工作表 s_0 ，即 “MLSDData”
- workflow 阶段 w_1 ，即 “stage2”
- 版本 r_1 ，即
“ver20241209_151130”

- 项目 a_{141} ，即“142”

上述记号也可以写成：

$$\begin{array}{ccc} \text{"ver20241209_151130"} & \overset{\text{"142"}}{W} & \text{"MLS823542"} \\ \text{"stage2"} & \underset{\text{"MLSDData"} \text{ "GLA"}}{} & \end{array} \quad (4)$$

6.1.8 理由

“为什么不在记号中直接使用索引或数字呢？”我们对记号有两个要求：

1. 子集。例如，分析人员经常选取列表的子集，比如变量的子集。举例来说，MLSDData 工作表中可能有 35-40 个变量，但某次运行只使用其中的 10-20 个。从一次运行到下一次运行，VE 会频繁更改提交给 MARS/earth 的变量。
2. 迭代。例如， $i = 1, 2, \dots, n_1$
我们的算法、规程和证明通常需要能够从头到尾遍历有序列表中的所有项——或挑出某个特定的项。

因此，我们需要一个字母来表示项的子集或列表，并且它必须带有另一个下标，供我们遍历列表中的各项，以给出用于索引数据框行或列的字符串。

6.1.9 矩阵？

那么线性代数中的矩阵呢？矩阵只能处理数字，即整数或实数。R 和 Python 都支持矩阵这种数据结构。然而，我们在 MLS 数据中处理的许多变量是所谓的“字符串”或字符数据。即使有些变量只是数字，它们往往也不是实数，而只是名称。例如，某城市的 MLS 区域代码可能是（“660”、“661”、“662”、“663”、...、“672”）。Earth 回归允许把变量指定为“因子”变量，此时这些数字仅被当作名称处理，不含任何顺序。在这种情况下，生成的模型只会为每个数字提供一个价值贡献，而不是一个在序列上取值的函数。其他例子包括表示状况的“C1”、...、“C6”和表示质量的“Q1”、...、“Q6”。

6.1.10 第一次简化

上述针对核心 Excel 项目工作簿的记号很适合把各部分联系起来，但用于解释算法或数学证明就显得太沉重了。在做繁琐的具体工作时，我们可以假设一切都发生在给定的项目、版本、阶段和工作表之内。如果所处理的对象尚不明确，我们直接说明即可。一旦指明了所处理的对象，我们真正需要的只是：

$$M_{s_h}^{p_i v_j} \quad \begin{array}{l} i = 1, 2, \dots, n_p \\ j = 1, 2, \dots, n_v \\ h = 1, 2, \dots, n_s \end{array} \quad (5)$$

其中：

n_p	=	房产数量
n_v	=	输入变量数量
n_s	=	输入工作表数量

6.1.11 第二次简化

更方便的做法是按工作表所承载的数据框的具体类型来重新命名它们，而不是使用索引。对每个工作表而言，行和列具有不同的功能。因此，W 这一数据框集合被拆分为 M、V、A、I 和 C 数据框，如下所示：

$$\diamond \text{ MLS 数据} \quad M_{v_j}^{p_i} = W_0^{p_i} \quad \begin{matrix} i = 0, 1, \dots, n_p \\ j = 0, 1, \dots, n_v \end{matrix} \quad (6)$$

其中： n_p = 房产数量
 n_v = 输入变量数量

注意：估价对象默认始终位于房产列表首位

在“MLSDData”Excel 电子表格及相应的数据框中，第一行始终保留给估价对象。它绝不会与其下方的其他 MLS 房产一起参与排序。它也从不作为自变量数据输入 MARS 回归，而是作为目标数据，即因变量输入。房产列表同样如此。在所有房产列表中，除非另有说明，索引为 0 的第一个元素始终被假定为估价对象的 ID。

即使任务是为某个没有估价对象的市场区域建立模型，第一行也仍会加载一些虚拟的房产数据。这使得编写程序、给出算法或进行证明都方便得多。因此，如果我们用 c 表示房产列表，则 c_0 是估价对象，而 c_i ($i = 1, \dots, n_p$) 是将输入 MARS/earth 回归的可比挂牌信息。

◇ 回归变量规格说明

$$V_{s_j}^{v_i} = W_1^{v_i} \quad \begin{matrix} i = 1, 2, \dots, n_v \\ j = 1, 2, \dots, n_s \end{matrix} \quad (7)$$

其中： n_v = 可用变量数量
 n_s = 规格说明数量

注意， $n_s \leq n_v$ ，因为我们通常从可用变量中选取一个子集作为回归的输入。

◇ 计算

$$C_{c_j}^{v_i} = W_2^{v_i} \quad \begin{matrix} i = 1, 2, \dots, n_v \\ j = 1, 2, \dots, n_c \end{matrix} \quad (8)$$

其中： n_v = 变量数量
 n_c = 计算规格数量

◇ 交互

$$I_{t_j}^{f_i} = W_3^{f_i} \quad \begin{matrix} i = 1, 2, \dots, n_f \\ j = 1, 2, \dots, n_t \end{matrix} \quad (9)$$

其中： n_f = “from” 变量数量
 n_t = “to” 变量数量

请注意，数据框 I 是一个对称数据框，变量按字母顺序排列于“from”行和“to”列。单元格标记为“1”表示允许相应变量对之间的交互；否则，“0”表示不允许交互。对于二阶交互，这相对简单，只可能存在一种交互。对于三阶回归，三个变量之间共有三个变量对交互，必须允许其中每一对交互，才能生成三变量交互项（或变量）。

◇ 聚合

$$A_{v_j}^{v_i} = W_4^{v_i} \quad \begin{matrix} i = 1, 2, \dots, n_v \\ j = 1, 2, \dots, n_v \end{matrix} \quad (10)$$

其中： n_v = 变量数量

7 RCA 销售网格

请记住，RCA 规程需要一个能够调用 R/earth 并执行所有细节步骤的 R 或 Python 程序。这项工作数据密集，绝不可能以手工方式完成。同时，假设程序没有缺陷，您也需要一个软件程序来确保所需工作精确无误地完成，避免人为差错。

对于从事这项工作的估价工程师来说，能够根据需要修改此类程序也是一种必要能力，尤其是在面对新的、未曾预见的市场区域、社区、房产或其他新需求、新问题时。

在 RCA 处理过程中，我们有三种类型的销售网格：

1. RCA 处理销售网格，在本文中标记为“MLSDData”，它是一个非常庞大的数据框及其关联的电子表格，可能包含许多中间细节，这些细节最终经过聚合处理，得到下文提到的 RCA 销售网格。示例太大，无法在此展示。估价工程师很可能会按照各自的偏好来定制这一电子表格。我打算在将来另撰一篇论文，专门讨论此类网格的开发。
2. RCA 最终销售网格，用于非 GSE 报告，其形式类似于表 2。它确实一定程度上模仿了例如 Fannie Mae 的表格，足以让读者有一定的熟悉感。不过请注意，它为每个可比交易实例提供的列数多于 Fannie Mae 网格：每个可比交易实例除“测量值”和“调整”列之外，还有一个“价值贡献”列，估价对象也同样有一个——这些列是 Fannie Mae 表格所没有的。一个更简单的形式见表 1。
3. GSE 上传销售网格，用于将销售网格上传到诸如 Alamode 之类的软件中。但请注意，并非所有估价表格软件包都提供从电子表格上传销售网格数据的功能。出于必要，此类网格不能包含单独的价值贡献列，而 Fannie Mae 等 GSE 对价值贡献毫无概念。我们的策略是：向它们提供其所要求的表格，但以 RCA 销售网格作为支撑，并辅以其他支持性文档，例如各变量聚合的详细信息。

RCA 处理销售网格看起来不像 Fannie Mae 销售网格，而大致相当于 Fannie Mae 销售网格的转置，行与列互换：每一行包含某一房产的数据，每一列包含某一房产变量的数据。房产行可以轻易达到数百或数千行，列数可达 200 列或更多。第一行保留给变量名，第二行保留给估价对象，即使我们只是在做市场分析也是如此。当电子表格被读入数据框后，工作表的第一行成为 names 属性，而估价对象所在的第二行则成为数据框的第一行。

RCA 销售网格中的列包括标识列，然后是主要变量、外部变量和合成变量，它们合并后作为自变量输入 MARS。MARS 从这些变量中输出它认为对成交价格（也称因变量或目标变量）有显著影响的变量。在二阶回归中，我们可能得到成对的变量，它们作为新的变量发挥作用。在三阶回归中，我们还可能得到变量三元组，它们作为所谓的自变量发挥作用，尽管它们当然与底层变量高度相关。

表 1: RCA 电子表格示例：纵向布局

估价对象			可比交易实例 1			可比交易实例 2		
	测量值	价值贡献	测量值	价值贡献	调整	测量值	价值贡献	调整
成交价格			\$1,150,000			\$1,500,000		
基础价值		\$500,000		\$500,000	\$0.00		\$500,000	\$0
		已测量特征		已测量特征			已测量特征	
居住面积	1,530	\$463,500	1,350	\$382,500	\$81,000	1,700	\$540,000	-\$76,500
地块面积	6,000	\$200,000	6,500	\$220,000	-\$20,000	7,000	\$240,000	-\$40,000
房龄	53	\$63,600	57	\$68,400	-\$4,800	51	\$61,200	\$2,400
浴室数	2	\$15,000	3	\$30,000	-\$15,000	3	\$30,000	-\$15,000
卧室数	3	\$6,000	4	\$9,000	-\$3,000	4	\$9,000	-\$3,000
测量合计		\$1,248,100.00		\$1,209,900.00	\$38,200		\$1,380,200.00	-\$132,100
		残差特征		残差特征			残差特征	
残差	CQA 5.8	\$30,000	CQA 2.5	-\$59,900	\$89,900	CQA 8.2	\$119,800	-\$89,800
设计	残差分解			残差分解			残差分解	
状况		7,000		-\$15,000	\$22,000		\$35,000	-\$28,000
质量		5,000		-\$8,000	\$13,000		\$20,000	-\$15,000
功能效用		8,000		-\$10,000	\$18,000		\$30,000	-\$22,000
配套设施		5,000		-\$12,000	\$17,000		\$15,000	-\$10,000
小计		5,000		-\$14,900	\$19,900		\$19,800	-\$14,800
		30,000		-\$9,900	\$89,900		119,800	-\$89,800
合计			合计			合计		
价值贡献总计		\$1,278,100.00		\$1,150,000.00			\$1,500,000.00	
调整后成交价格					\$1,278,100.00			\$1,278,100.00

8 变量

在运行 MARS 的输入端，我们只有简单变量，它们是：

1. 从 MLS 电子表格加载的物业特征。
2. 由 VE 创建的派生变量，用于细化 MLS 变量或添加现有变量的变换。例如，VE 可能倾向于将销售日期变换为“DaysOffMarket”，其值等于估价生效日期减去销售日期。
3. VE 可以从第三方来源导入数据，例如按揭利率。或者，如果 MLS 数据中缺少纬度、经度或海拔数据，他可以从其他来源导入。这些变量可称为“外部变量”。

在 MARS 的输出端，我们处理的价格模型是若干项的组合，这些项：

1. 仅包含一个输入变量，并给出该变量价值贡献的方程。这些称为一阶项或（输出）变量。该项本身被视为一个价值贡献值并被赋予名称。
2. 包含两个输入变量，是给定两个变量的交互作用之价值贡献的方程。这些项可称为二阶项或（输出）变量。
3. 包含三个输入变量，是三个变量之间交互作用之价值贡献的方程。在本文中，这些项可称为三阶项或（输出）变量。

于是，MARS 的输出变量即为价值贡献，包括残差价值贡献。它们又在销售网格计算中被作为变量使用，例如计算调整或绘制图形。

我们可以有落入两种基本本体的不同类型变量：（1）变量来源；（2）变量应用，即非残差与残差

8.1 变量来源本体

1. 一阶变量

- 1.1 MLS 回归变量
- 1.2 外部回归变量
- 1.3 派生回归变量

2. 因子变量：

为了处理因子变量，需要一个能够识别它们的函数。由于它们是在 Excel 项目配置工作簿中指定的，因此创建了一个函数：当传入的变量已被归类为因子类型时，该函数返回 true。

3. 交互变量：如果阶数设置为 2 或更高，这些变量由 MARS 在生成模型时创建。然而，由于解释上的困难、过拟合的可能性、对自由度的影响以及为确保模型稳健而随之增加的物业记录数量，最多只建议 3 个变量交互。

3.1 二阶变量：这些很常见，好的例子有 GLA:BathRms、GLA:LotSize、DateOfSale:BathRms、Latitude _Longitude。

3.2 三阶变量：较不常见的交互，提示需要谨慎并可能存在过拟合。

4. 残差分解变量

8.2 变量应用本体

1. MARS 生成的价值贡献

1.1 一阶变量

1.1.1. MLS 回归变量

1.1.2. 外部回归变量

1.1.3. 派生回归变量

1.2 交互变量

1.2.1. 二阶变量

1.2.2. 三阶变量

1.3 残差分解变量

1.4 汇总变量

2. 工程师生成的变量 & 数值

2.1 残差分解变量

2.1.1. 设计/美学

2.1.2. 景观

2.1.3. 功能效用

2.1.4. 视野

2.1.5. 状况

2.1.6. 质量

2.1.7. ...

这些变量组将在下一节中给出其记号并作更详细的描述

规程 3.1: 因子变量

因子变量, 或简称因子, 可出现在所有项中, 包括交互项。它们之所以重要, 是因为在 MARS 回归和绘图中必须对其进行特殊处理。

虽然取字符值的变量必然是因子, 但数值变量必须被标记为因子变量, MARS 才能将其识别为因子。例如, MLS 区域常以数字标识。特别是, 旧金山湾区的邻里区域常以唯一的整数标识。如果这些值作为因子传给 MARS, 那么 MARS 将尝试发现某些区域对其边界内的物业交易是否具有平均贡献价值, 而不考虑其与其他区域的任何关系。相反, 对于普通变量, 它会假设随着变量值的变化, 价值贡献存在某种连续性。这是一个重要的区别。此外, 这些因子变量可能与非因子变量一起出现在模型项中, 这使绘图变得困难。具体而言, 在绘制含因子的单变量项时, 我们将所有此类因子归并在一起, 简单地绘制一张显示其正负贡献的条形图。

如果因子与其他变量一起出现在某个项中, 它们总是以一个可能的取值出现, 因为因子变量对给定物业只能有一个值: 如果某区域的值为 662, 根据定义, 该区域不可能有任何其他值。如果你觉得这令人困惑, 那么更糟的是, 在 MARS 生成的模型中, 因子并没有单一的变量名; MARS 会取输入时给它的因子变量的词根名, 然后通过附加各可能取值来构造新的变量名并创建新列。于是, Excel 电子表格中的单个列 (例如 MlsArea) 会为每个可能的取值创建多个数据框列, 例如 MlsArea660、MlsArea661 等。诸如 MlsArea661 这样的变量, 对位于 661 区域的物业取值为 1, 否则为 0。然后, MARS 将把这些列当作布尔变量进行回归。MARS 提供的模型中将使用这些修改后的名称。

你很可能希望将这些模型名称转换回其原始名称, 使图形对喜欢简单明了 (至少尽可能简单) 的客户更易理解。我们看到的将不是 MlsArea661=1, 而是 MlsArea=661。然后, 我们只需在图形标题中注明诸如“对于 MLS 区域 = 662”之类的文字, 即可绘制各项。这样, 含 M 个因子变量的 N 阶项将作为 (N-M) 阶项绘制, 因子维度则消失于标题之中。注意, 单个项中可能存在两个或更多因子变量, 因此你很可能需要为每一对可能的取值组合绘制一张图。如果具有三个不同取值的变量 x 与具有两个不同取值的变量 y 同非因子变量 z 一起出现在同一项中, 那么你将需要 $2 \cdot 3 = 6$ 张图。一个例子是“MlsArea”+“Architectural Style”, 例如旧金山的维多利亚式住宅, 其价值可能因其所处的旧金山区域不同而或高或低。幸运的是, 在大多数区域, 通常只有 1-3 种建筑风格作为价值贡献者重要到足以用于模型 (尽管可能存在 6 种以上不同的建筑风格)。

8.3 二阶与三阶项

在二阶或三阶回归中, 输出模型中的某些项可能分别包含两个或三个变量。例如:

$$\text{BathRms} ___ \text{GLA} = -52 * \max(0, 1200 - \text{GLA}) * \max(0, \text{BathRms} - 1)$$

注意, 我们通过按字母顺序对变量名排序并将其连接 (或许以下划线相连) 来命名这些项。对于模型中产生的变量名三元组, 我们对三阶项也如此处理。例如“BathRms ___ GLA ___ LotSz”。

在第 3 阶段, 当 RCA 程序为每个物业计算所有变量 (包括交互变量) 的贡献时, 这些变量会被加入到主要输入变量之中。

8.3.1 调整变量

RCA 程序还将在第 3 阶段, 即价值贡献计算完成之后, 通过用估价对象某一特征的贡献减去相应可比物业该特征的价值贡献, 来计算所有变量和所有物业的调整。

8.3.2 汇总变量

汇总变量被定义用来取代变量的某些子集，将它们的所有价值贡献和调整加总到一个公共变量中，以便在报告中使用。例如，我们可能已将已完工和未完工居住面积都拆分为更细的变量：

1. 地上合法已完工面积
2. 地下合法已完工面积
3. 地上非法已完工面积
4. 地下非法已完工面积
5. 地上未完工面积
6. 地下未完工面积

是的，确实存在这样做有意义的地区。

MARS 回归将对所有这些变量进行回归，并生成为它们提供贡献和调整的模型。但对于表格式报告而言，这些不同的贡献可能必须合并或汇总为更少的变量。汇总的细节应在附录中提供。一个更常见的例子是将交互调整汇总到某个主要变量的调整中。例如，我们可以将 GLA_Lotsize 的调整汇总到 GLA 或 LotSize，甚至按 50:50 在两者之间拆分，具体由汇总工作表决定。

8.3.3 特殊变量 & 记号

π_i	物业 $i = 1, 2, \dots, n_p$ 的实际净售价；对于估价对象， π_0 表示价值结论（见证明 I）
α_i	物业 $i = 1, 2, \dots, n_p$ 的调整后净售价
ρ_i	物业 $i = 0, 1, \dots, n_p$ 的 MARS 估计净售价
β	模型的基准值
$\varphi_{i,j}$	物业 $i = 0, 1, \dots, n_p$ 的回归变量 $j = 1, 2, \dots, n_r$ 的价值贡献
$\delta_{i,j}$	物业 $i = 1, 2, \dots, n_p$ 的回归变量 $j = 1, 2, \dots, n_r$ 的调整
ϵ_i	物业 $i = 0, 1, \dots, n_p$ 的总残差
$\epsilon_{i,c}$	物业 $i = 0, 1, \dots, n_p$ 的残差分量价值贡献 c ， $c = 1, 2, \dots, n_c$
$\xi_{i,c}$	物业 $i = 0, 1, \dots, n_p$ 的残差分量 c 的调整， $c = 1, 2, \dots, n_c$

其中

n_p = 可比物业的数量

n_r = 回归变量的数量

n_c = 残差分解变量的数量

8.3.4 分解残差

一旦 MARS 回归创建了模型，它便会根据 MARS 选定的输入变量生成估计售价。它用相应的净售价减去该值，得到每个物业的残差值。此时，RCA 规程规定：一旦 VE 确定了最终的可比物业——通常为 6-12 条物业挂牌信息——他需要逐一查看每条物业挂牌信息，并在忽略随机数据误差可能性的前提下，将计算出的残差分解为他认为最可能影响价值的特征（潜在变量）。例如，对于某一物业，他可能有以下分解变量：

1. 设计
2. 功能效用
3. 高于/低于市场的成交
4. 状况
5. 质量

不同物业很可能会有一些不同的残差分解变量，需要将它们收集到一个有序集合中。

估计稀有特征的贡献。有些承载价值的特征对回归通道而言过于稀有：在 780 笔交易中仅出现 3 次的工作间无法获得站得住脚的系数，前向过程通常会完全跳过这样的变量。这些就是残差特征，其贡献可以在残差空间中估计：先拟合不含该特征的模型，然后用普通最小二乘法将模型残差对被排除的列进行回归。对于二元标志，系数即为一次性总额贡献（等价地，即被标记交易的平均残差偏移）；对于数值列，则是单位贡献。当若干相关设施被联合估计时（例如附属建筑面积与工作间标志一起估计），回归会在它们之间分配贡献，而不是在两行分解中重复计算同一笔金额。由不足约十个观测支持的估计仅具指示性，应通过折旧成本或扩大后的样本（例如从公开备注中提取的标志）予以佐证；残差约束本身会对最终赋予的价值设置上限。earthUI (v0.11+) 以 *Residual Contribution* 选项卡的形式直接实现了这一方法，包括联合估计、逐笔交易的残差证据以及用于工作档案的记录说明。

记号：我们使用上表中的记号，得到：

$$\epsilon_i = \epsilon_{i,1} + \epsilon_{i,2} + \dots + \epsilon_{i,n_c} \quad (11)$$

其中 n_c 是残差分解分量的数量。

VE 必须运用其判断将残差分解为给定的分量。有人可能认为这会让偏见进入最终的价值结论。然而事实并非如此，因为我们有数学证明：只要各分解调整的总和等于残差，它们就不会影响价值结论。分解的目的不是确立价值，而是解释残差相对于其他物业交易的残差为何是这个数值。工程师的偏见唯一可能影响最终价值结论的地方，是对估价对象残差的估计，因为估价对象当然没有实际售价可用来为其计算残差。

每个物业都有其残差价值贡献的集合，但这些集合需要合并为覆盖所有物业的一个集合 C：

$$\begin{aligned} \text{设: } A_i &= \{\{\epsilon_{i,c}\}_{c=1}^{n_c}\} \\ C &= \{x \mid x \in \bigcup_{i=0}^{n_p} A_i\} \end{aligned}$$

其中 $[\epsilon_{i,c}]$ 是值 $\epsilon_{i,c}$ 的变量名， A_i 是此类变量的集合， n_c 是每个 ϵ_i 中值的数量。

因此，当估价对象与可比物业的数据分列于不同列中时，在每个物业的所有回归模型变量价值贡献之下，我们应看到由 n_c 个相应残差价值贡献构成的集合 C。

9 RCA 核心证明

设：

$$\begin{aligned} n_p &= \# \text{ of properties} \\ n_r &= \# \text{ of regression variables} \\ n_c &= \# \text{ of residual component variables} \end{aligned}$$

证明 I： 每一个可比交易实例的调整后售价均等于估价对象的估计售价。

对于任意可比交易实例 i ，调整后售价 α_i 等于净售价 π_i 加上各项调整：即估价对象与可比交易实例在基础价值、回归变量 $j = 1, 2, \dots, n_r$ 以及残差上的价值贡献之差。对于没有实际成交的估价对象，其净售价 π_0 定义为价值结论 $\beta + \varphi_{0,1} + \varphi_{0,2} + \dots + \varphi_{0,n_r} + \epsilon_0$ 。于是：

$$\begin{aligned} \alpha_i &= \pi_i + (\beta - \beta) + (\varphi_{0,1} - \varphi_{i,1}) \\ &\quad + (\varphi_{0,2} - \varphi_{i,2}) + \dots + \\ &\quad (\varphi_{0,n_r} - \varphi_{i,n_r}) \\ &\quad + (\epsilon_0 - \epsilon_i) \\ &= \pi_i + (\beta + \varphi_{0,1} + \varphi_{0,2} + \dots \\ &\quad + \varphi_{0,n_r} + \epsilon_0) \\ &\quad - (\beta + \varphi_{i,1} + \varphi_{i,2} + \dots + \\ &\quad \varphi_{i,n_r} + \epsilon_i) \\ &= \pi_i + \pi_0 - \pi_i \\ &= \pi_0 \end{aligned} \tag{12}$$

证明 II： 在所有残差分量价值贡献之和等于可比交易实例残差这一约束下，改变残差的分解方式对 RCA 规程的最终价值结论没有任何影响。

这几乎是显而易见的。然而若不认真思考其中的数学，就很容易忽视这一点，因为传统估价师关注的是调整而非价值贡献。在这里相互抵消的正是价值贡献——这是一个持续了数十年的疏漏。对于某个可比交易实例 k ($k > 0$)，总残差调整等于 n_c 个分解残差分量之和，而后者又等于各价值贡献之差的和，即等于估价对象总残差减去可比交易实例总残差。这与所创建的残差分量的数量或其所代表的内容无关。只要残差分量的价值之和对估价对象和所选可比交易实例都等于总残差， ξ_k 就不会改变。

$$\begin{aligned} \xi_k &= \xi_{k,1} + \xi_{k,2} + \dots + \xi_{k,n_c} \\ &= (\epsilon_{0,1} - \epsilon_{k,1}) + (\epsilon_{0,2} - \epsilon_{k,2}) + \dots + \\ &\quad (\epsilon_{0,n_c} - \epsilon_{k,n_c}) \\ &= (\epsilon_{0,1} + \epsilon_{0,2} + \dots + \epsilon_{0,n_c}) - \\ &\quad (\epsilon_{k,1} + \epsilon_{k,2} + \dots + \epsilon_{k,n_c}) \\ &= \epsilon_0 - \epsilon_k \end{aligned} \tag{13}$$

10 MARS 模型

10.1 模型变量与交互作用的“ g ”函数

RCA 方法的一项核心原则是“解释性推理”至关重要。遗憾的是，为各个价值贡献方程提供有意义的图形——尤其是在使用交互作用和因子变量时——可能颇具挑战。因此，为绘制这些图形制定一套清晰而严格的规程被视为势在必行。本文在此给出这一规程，但请理解，该规程面向的是编写 R 或 Python 代码以生成图形的开发者，对于不参与编码的读者而言，可能仅有粗浅的参考意义。

MARS 生成的价格模型是一个基础常数加上若干附加项之和，每一项涉及 1 到 3 个变量。为了对其进行分析，我们首先将模型拆分为单个的项。然后，将共享同一变量集合、代表单一交互作用的多个项合并为单个交互作用表达式。假设回归的最大阶数为 3。考虑一个三阶回归模型，它由一个基础常数、 n_1 个一阶表达式、 n_2 个二阶交互作用表达式和 n_3 个三阶交互作用表达式组成。这样我们便得到四组不同的表达式，它们或为单变量表达式，或为多变量交互作用，其绘图方式取决于两个因素：（1）它们是否包含因子变量（即分类变量），以及（2）它们所含唯一变量的数量。具体而言：

- 一阶表达式（1 个变量）以二维图形绘制。
- 二阶表达式（2 个变量）以三维图形绘制。
- 三阶表达式（3 个变量）以一组三维图形绘制。
- 若表达式包含因子变量，则每个唯一因子会使图形的维度降低一维（例如，若其中一个变量是因子，则三维图形变为二维图形）。

每个函数 g 的索引提供了确定其绘图方式所需的信息，反映了唯一变量的数量以及是否存在因子变量。

10.2 g 函数的索引

每个函数 g 的索引形式为 $f_{j,k}g_k^j$ ，其中：

- $f_{j,k}$ 是第一个索引（置于 g 之前的上标）：
 - j 的取值范围为 0 到 3，用于标识组别：
 - * $j = 0$ ：基础常数（是通常索引规则的一个例外）。
 - * $j = 1$ ：一阶表达式。
 - * $j = 2$ ：二阶表达式。
 - * $j = 3$ ：三阶表达式。
 - k 是某一具体函数在其所属组内的索引（例如， $k = 1$ 表示组 j 中的第一个函数）。
 - $f_{j,k}$ 的值表示该函数中唯一因子变量的数量：
 - * $f_{j,k} = 0$ ：不含因子变量。
 - * $f_{j,k} > 0$ ：因子变量的数量（例如 1、2 或 3）。
- g 的后置索引：
 - g 之后的上标 j （例如 g^j ）重复标注组号，以求清晰。
 - g 之后的下标 k （例如 g_k ）与 $f_{j,k}$ 中的 k 一致，标识该函数在其组内的位置。

虽然 f 与 g 的索引在数值上一致，但它们的作用不同： $f_{j,k}$ 计数因子变量，这对绘图决策至关重要；而 g 上的 j 和 k 则定义其组别和位置。

10.3 绘图维度

函数 $f_{j,k} g_k^j$ 的图形维度 d 按下式计算:

$$d = j - f_{j,k} \quad (14)$$

其中:

- j 是唯一变量的数量 (即组号, 不考虑常数的情形)。
- $f_{j,k}$ 是因子变量的数量。
- d 是需要绘图的连续 (非因子) 变量的数量, 最大为 3 (因为最大阶数为 3)。含有 d 个连续变量的函数需在 $d+1$ 维中绘制: $d=1$ 对应二维图形, $d=2$ 对应三维图形 (或二维热力图), $d=3$ 则对应一组三维图形。

例如:

- 若 $j=3$ (3 个变量) 且 $f_{j,k}=0$ (无因子变量), 则 $d=3-0=3$, 需要一组三维图形, 每张图固定其中一个变量。
- 若 $j=3$ 且 $f_{j,k}=1$ (一个因子变量), 则 $d=3-1=2$, 需要针对模型中出现的每个因子变量取值绘制一张三维图形 (或二维热力图)。
- 若 $j=0$ (常数), 通常无需绘图, 因为它只是一个单一数值。

当 $d=3$ 时 (例如不含因子变量的三阶表达式), 要将完整函数可视化, 可能需要多张三维图形, 每张图将某一变量固定在不同取值上, 以展示函数行为的一个子集。

10.4 最终模型方程

MARS 模型被拆分为四组表达式 (常数、一阶、二阶和三阶)。索引 $f_{j,k} g_k^j$ 告诉我们:

- j : 组别及变量数量。
- $f_{j,k}$: 因子变量的数量。
- $d = j - f_{j,k}$: 图形维度。

这一结构指导每个 g 函数的可视化方式, 确保绘图方法与表达式的复杂度和变量类型相匹配。

MARS 模型给出的售价估计值 $\hat{P}$ 可以表示为:

$$\begin{aligned} \hat{P} = & f_{0,1} g_1^0 + \sum_{q=1}^{n_1} f_{1,q} g_q^1(x_*) \\ & + \sum_{r=1}^{n_2} f_{2,r} g_r^2(x_*, y_*) \\ & + \sum_{s=1}^{n_3} f_{3,s} g_s^3(x_*, y_*, z_*) \end{aligned} \quad (15)$$

$n_1 = \#$ of first degree terms

$n_2 = \#$ of second degree terms

$n_3 = \#$ of third degree terms

10.5 绘制一个模型所需的图形数量

要将模型中的所有函数全部可视化，需要多少张图？答案取决于变量本身、它们之间的交互作用，以及它们是因子变量还是非因子变量。

估价工程师可以在供 R/earth 使用的 Excel 配置文件中限制变量和交互作用，从而影响哪些交互作用被允许。一种务实的做法是让 R/earth 评估所有可能的交互作用，并选出对售价估计有显著贡献的那些，同时检查过拟合或共线性等问题。某些交互作用可能被禁用，这会影响更高阶的交互作用（例如，若三阶交互作用中的某一变量对被排除，则整个组合都被排除）。

对于因子变量（例如具有 12 个取值的 MLS Area 与 GLA 交互），绘图会变得更为复杂。这可能需要 12 张图——每个区域一张。若另一个因子（如具有 6 个取值的 Architectural Style）再与 MLS Area 和 GLA 交互，则可能需要多达 $12 \cdot 6 = 60$ 张图。不过，R/earth 很少会认定所有组合都显著；通常只有少数关键交互作用（例如特定的区域和风格）值得绘图，而模型会明确指出这些组合。

以下是所需图形数量的简明指南：

1. 单变量函数（非因子）：1 张图。
2. 单变量函数（因子）：1 张将所有因子取值合并展示的条形图。
3. 基础函数：无需绘图。
4. 二阶交互作用（无因子）：1 张图（例如以颜色表示售价的二维热力图）。
5. 二阶交互作用（1 个因子）：模型中每个因子取值 1 张图（例如 2 个取值 = 2 张图）。
6. 二阶交互作用（2 个因子）：1 张以因子组合构成矩阵的图（例如 3 个区域 $\times$ 2 种风格 = 6 个单元格）。
7. 三阶交互作用（无因子）：多张图（例如按 Bath 数量 = 1、2、3、4 绘制 4 张图）。
8. 三阶交互作用（1 个因子）：模型中每个因子取值 1 张图。
9. 三阶交互作用（2 个因子）：模型中每种因子取值组合 1 张图。
10. 三阶交互作用（3 个因子）：以不同取值最少的那个因子的每个取值各绘 1 张图，其余两个因子用颜色矩阵表示。

更复杂的情形可能需要更高级的绘图技术。

此外还存在其他可能性，通常需要更为精细的图形。

由 MARS 生成的完整模型示例见 [图 1](#)。

11 数据特征

以下一节讨论 MARS。MARS 完全依赖于它所接收的数据，而这些数据大多来自某个 MLS（多重挂牌服务）或税务评估机构。不过，数据也可能经由其他数据提供商中转，这些提供商维护并改进数据，然后以一定利润出售。在某些地区，数据相当可靠；在另一些地区则不那么可靠。在关键情形下，房产查勘人员必须亲自进行观察和测量，而其准确性也可能受到质疑。

估价工程师需要理解数据代表什么，以及数据在多大程度上如实地代表了它本应代表的内容。这引出了一系列问题：

- 如何将一个变量归类为测量值？
- 如果一个变量不是测量值，它还能是什么？
- 如果回归所接收的数据不准确，回归又如何能有用？

我们从 MLS 挂牌服务和税务评估机构获得的用于回归的数据，大多是隐含测量值或计数的数值型数据。其中一部分是逻辑型数据，即真或假。

11.1 残差的用途

残差是对未纳入回归模型的变量所增添价值的一种间接量度。用特征价格文献的语言来说，这些未测量的特征就是潜在变量：它们影响成交价格但未被直接测量，多数情况下是因为没有切实可行的方法对其进行客观测量。它们的取值只能通过推断得到，而在实践中，潜在变量的取值往往通过主观判断来估计——在 RCA 中，这种判断以残差排序为锚。通过按残差或每平方英尺残差对房产进行排序，我们得到了一个从吸引力最低到吸引力最高的客观排名。只要由一位对当地市场有深入理解的熟练分析师构建多元自适应回归样条（MARS）模型，这种排序方法就是有效的。以下是一些附加要求：

- 数据与模型的质量：构建良好的回归模型和高质量的数据是获得可靠残差的前提。在旧金山湾区，我的目标是建立一个 R^2 约为 75-80%、交叉验证 R^2 (CV R^2) 为 55-70% 的模型。
- 避免过拟合、限制 R^2 ：在像旧金山湾区这样房价较高地区的多数成熟市场中，估价工程师应以 R^2 最高约 80% 为目标，因为住宅房产竞争性价值的很大一部分在于其主观吸引力，这种吸引力无法量化，因而不能直接纳入回归模型。 R^2 显著高于 80% 可能意味着模型过拟合，需要进一步审查。但需要提醒的是：有些市场区域的 R^2 值高于 90% 而并未过拟合；也有一些难以估价的地区，能达到的最好水平可能只有 50-60% 甚至更低。
- 特定地区的高 R^2 ：不过，在某些区域，例如分块开发的小区或其他同质化的住宅区，由于房产的同质性，较高的 R^2 值出人意料地常见。

11.2 作为近似回归变量的状况与质量评级

房利美的状况与质量评级量表 (C1-C6 和 Q1-Q6) 值得特别说明。这些量表过于粗糙，不能被当作真正的测量特征；而且由于它们由估价师给出，实践中属于主观评级——本身就是潜在变量。明确区分两个层次是有益的：估价师用作回归变量的粗糙评级，即粗粒度状况 (condition-1) 和粗粒度质量 (quality-1)；相对地，从残差中恢复的细粒度潜在变量，即细粒度状况 (condition-2) 和细粒度质量 (quality-2)。VE 可以将估价师的粗粒度评级作为近似值接受并对其进行回归，从而得到近似的价值贡献，这些价值贡献可作为细粒度潜在变量的初始赋值。在采用这种做法时，规程是从模型估计值中减去这两个粗粒度变量经回归估计的价值贡献——把它们从销售网格回归部分下的已测量属性中移出——并将其加回到残差之上。此后，再从残差中移除这部分并将其分配给相应的细粒度潜在变量 `res_condition1` 和 `res_quality1`。随后利用从 MLS 公开

备注 (public_remarks) 及相关字符串列表中提取的附加信息, 对这些潜在变量的取值进行微调。最终结果是, 细粒度价值贡献取代粗粒度价值贡献: 粗糙评级只是近似模型的脚手架, 最终报告的状况与质量取值是从残差推导出的细粒度取值。

11.3 数据错误

根据社交媒体上的评论, 加利福尼亚州的多重挂牌服务 (MLS) 数据总体上比美国许多其他地区质量更高。通常, 在大多数成熟的大都市地区, MLS 数据的质量应当是可以预期的良好, 而在农村地区则较差。

尽管数据中因各种原因偶有错误, 但只要数据量足够, 这些错误一般应相互抵消, 对价格模型所用平均值的影响很小。

然而, 特定地区可能出现系统性偏差。例如, 考虑这样一个社区: 许多住宅最初建有完整的二层, 但部分买家可以选择拆除二楼房间, 为一层的一个或多个房间打造挑高拱顶天花板, 从而显著减少总居住面积 (GLA)。由于种种原因, 这些改动后来没有在评估机构的记录中更新。结果, MLS 可能显示 2,400 平方英尺, 而实测值只有 2,000 平方英尺。在这类情形下, 应在执行多元自适应回归样条 (MARS) 回归之前先更正 MLS 数据。需要注意的是, 为可比交易实例获取这类信息可能需要一些调查工作。发现的差异应予以记录以备将来参考。

综上所述, 可以合理假设, 在 MLS 数据质量较高的地区, 溢入残差的错误与设计、状况、质量等主观且未测量特征的影响相比是微不足道的。在估计 RCA 方法的准确性时, 这是一个重要的考虑因素。例如, 如果某个市场区域的 R^2 值为 80%, 那么价格中约有 20% 归因于 CQA 残差。因此, 在残差排序中定位估价对象时 $\pm 10\%$ 的误差, 传导到最终价值结论上就是 $\pm 10\% \cdot 20\% = \pm 2\%$ 的误差。这就是提高准确性的途径! 在许多情况下, 可以得到误差远低于 $\pm 1\%$ 的准确价值估计。

因此, 在 MARS 回归中把残差当作未建模变量集体价值的间接指标是合理的, 前提是估价工程师具备足够的当地市场经验, 能够识别系统性测量偏差存在于何处以及为何存在。这种方法确保分析准确反映实际房产价值, 同时也承认房地产数据的局限性与细微之处。

实例演算: 在 Sea Ranch 测量未被测量的价值驱动因素

当某个市场区域可实现的 R^2 低于该市场表面同质性所应达到的水平时, 本框架的应对方式是识别并测量缺失的变量, 而不是接受这一局限。索诺马县的 Sea Ranch 社区直接说明了这一点。

这是“测量而非接受”式应对的一个具体实例。索诺马县的 Sea Ranch 社区是一个沿海开发区, 到海洋的距离是主要的价值驱动因素: 在其他特征相同的情况下, 距离悬崖几百英尺的住宅与向内陆四分之一英里的住宅, 价格差异巨大。该社区的 MLS 数据并未把到海距离记录为变量。基于现有 MLS 变量拟合的初始 MARS 模型, 其解释力低于该市场表面同质性所应支持的水平, 且残差结构未能按数据中任何可见特征充分地对房产排序——这一信号表明模型缺失了一个重要的驱动因素, 而不是该市场本身噪声很大。

补救办法是直接测量。作者获取了一张显示房产相对海岸线布局的社区地图, 并为数百套住宅测量了到海距离。将测得的距离作为变量加入 MLS 数据集, 并重新运行 MARS 回归。新变量被模型选为显著变量, 解释力大幅提升, 随后的残差排序按与其余主观特征——景观、状况、质量——相一致的模式对房产排序, 而不再受此前污染残差的潜在到海距离信号的影响。在 MLS 变量缺失或不可靠的山坡社区, 高程问题也以类似方式得到了处理。

这一方法论原则具有普遍性。当可实现的 R^2 低于该市场表面同质性所应支持的水平时, 首要假设是模型缺失了一个重要变量, 恰当的应对是弄清它是什么并直接收集测量数据。数据来源包括社区地图、GIS 数据、税务评估记录、建筑管理部门记录、卫星影像以及实地查勘。这类测量工作的成本是实实在在的, 但有限度; 一旦为某个市场区域收集了该变量, 就可以在该区域的后续估价任务中沿用。RCA 能够顺畅地吸纳新变量, 因为该框架并不在意哪些变量进入

MARS 回归——残差结构会告诉 VE 新变量是否捕捉到了真实信号，各项价值贡献也会相应调整。

另一种应对方式——接受较低的 R^2 并报告更宽的误差区间——在缺失的驱动因素确实无法测量（买方特有的偏好、短暂的市场情绪、交易的特殊情形）时是站得住脚的，但当这些因素只是尚未收集时则不然。面对一个弱模型，VE 的第一步应当是追问缺失的信号是否属于后者。

11.4 什么是准确性？

完美准确的估价要求：

1. 符合市场价值：估价对象的假设性销售交易 100% 符合给定的市场价值定义。
2. 准确而稳健的模型开发：建立一个高质量的 MARS 模型，视地区而定，其 R^2 约为 75-80%， CVR^2 约为 55-70%。
3. 良好的数据错误分析：已测量变量所有可能的数据错误都得到理解和控制，几乎不溢入残差。
4. 系统性偏差核算：数据中任何系统性偏差都应予以核算和调整。
5. 残差排名分析：确保按残差排名的房产从最低到最高排名呈现出合理的特征模式。遗产出售、亏本急售和拍卖出售等异常值应予以调查、解释，并在必要时剔除。
6. 估价对象定位：在透彻理解残差排序的基础上，应准确定位估价对象。找出一处吸引力明显更高的房产和一处吸引力明显更低的房产，将估价对象置于两者之间并打分。然后尽量收窄较高与较低房产之间的排名区间。
7. 解释性推理：绝不追求准确性而牺牲解释性推理。估价师必须能够解释估价对象与可比房产之间的价值差异，至少要让一个具有正常理解能力的人明白。例如，一个 R^2 和 CVR^2 达到最优的良好 MARS 模型可用于解释已测量特征的价值贡献，而估价工程师对残差的分解则可用于解释其他特征的价值贡献。
8. 过拟合：避免模型过拟合，不使用那些单独或组合起来能够识别一处或数处房产的房产特征。尤其要避免使用其取值能够锁定少数房产的因子变量。在规模足够大的数据集中，用于标识社区的因子变量是合适的，但在游泳池极少的地区标识带游泳池房产的因子变量则必须慎用。标识唯一房产的特征，例如精确的 GIS 坐标、宗地号或地址，是绝对不允许的。

如果回归的 R^2 为 80%，我们预计估价对象价值中约有 20% 归因于残差。当然，这只是近似值。从 CQA—残差曲线可知，残差分布的大部分集中在 CQA 的低端和高端两个极值区。基于 CQA—残差曲线以及低、高 CQA 值处的极端斜率，我们可能预期在这些区域对残差的估计不那么准确。然而，这一点被这些区域房产的相对稀少以及相邻房产之间吸引力差异更大这两个因素所抵消。因此，尽管在中间区段我们谈论的影响范围可能是 $\pm 5\%$ ，在两端我们更应关注的则是略窄的 $\pm 2 - \pm 3\%$ 范围。

一般而言，由于 CQA—残差曲线的非线性，计算 RCA 的准确性有一定难度。但只要你有一个合理的残差排序，就可以逐案计算。CQA 分值的范围是 0.0 到 10.0。有了良好的 MARS 模型，你的残差排序应呈现出从吸引力最低到最高的清晰模式，你应该不难把评分控制在总范围的 $\pm 5\%$ 之内，即 CQA 分值的 ± 0.5 之内。如果回归中有 100 处房产，这相当于把估价对象准确地排入一个 10 处房产的组内。那么，参照表 3 中的伯灵格姆销售数据，让我们分别针对 VE 估计的 CQA 为 7.5、5.1 和 1.9 的情形计算近似准确性（或误差），其中他对估价对象的价值结论为 \$3,000,000。注意，由于表 3 只包含所用数百处房产中的一个样本，必要时你需要（假设线性）对残差进行插值。需要明确的是，我们假定 VE 在每种情形下都非常确信，他对估价对象的排名在仅含 100 处房产的残差排序中处于 ± 5 处房产的范围之内。

CQA= 7.5 ±0.5					
CQA	残差	绝对差	预期误差	假设成交价	误差百分比
8.00	\$257,112				
7.50	\$197,113	\$59,999			
7.00	\$154,615	\$42,498	±\$51, 249	\$3,000,000	±1.71%

CQA=5.1 ±0.5					
CQA	残差	绝对差	预期误差	假设成交价	误差百分比
5.60	\$50,911				
5.10	\$19,559	\$31,352			
4.60	-\$28,207	\$47,766	±\$39, 559	\$3,000,000	±1.32%

CQA= 1.9 ±0.5					
CQA	残差	绝对差	预期误差	假设成交价	误差百分比
2.40	-\$228,411				
1.90	-\$276,381	\$47,970			
1.40	-\$326,938	\$50,557	±\$49, 264	\$3,000,000	±1.64%

最后，验证准确性的唯一客观方式，是建立一套严格的规程（本文只是粗线条地部分提及）并核查其遵守情况。之所以必要，是因为受市场不完善的影响，实际成交价格可能偏离市场价值。将估价与其后的成交进行比较，可以在一定程度上反映准确性。然而，房地产市场的复杂性和不稳定性，要求由估价师或高度合格的估价工程师依据既定标准和规程给出价值估计。这些标准和规程应当足够严格，以确保胜任的估价工程师能够各自独立地得出几乎相同的价值结论。在大多数能够获得良好数据的情况下，这应当是一个可以实现的目标。

11.5 作为异常值的估价对象

既然估计一处房产的市场价值需要分析近期的销售交易，那么就有理由追问：对于一处缺乏现代化改造和维修的较旧的估价对象，近期销售交易能否很好地代表其各项特征。在旧金山湾区尤其如此，那里近期成交的住宅通常都经过翻新，以提高其售价。因此，可比销售实例的平均质量和状况通常优于那些很久没有进入市场的老旧住宅。这一价值差异应当由对估价对象残差的估计——换言之，由其 CQA 分值——来捕捉。就估价对象估值中的数据质量与准确性而言，这是一个重要的考虑因素。

11.6 CVR2 在应对意外数据方面的重要性

MARS 回归基于对随机选取的房产子集多次迭代运行 R/earth，该子集通常占 MLS 数据集的 80% 到 90%，称为训练集。其余 10% - 20% 的房产作为测试集。在这一阶段，先由训练数据集建立 MARS 模型，再用它预测测试数据集中房产的成交价格，然后将预测值与实际成交价格进行比较。这一方法已被数据挖掘者公认为构建稳健模型、预测尚未观察到的房产成交价格的最有效策略。

然而，CVR2 指标的主要用途并不在于反映回归模型对实际市场区域成交的事实准确性，而在于优化模型对以指定生效日期为准的假设性房产交易的预测能力。这些房产不在训练数据集中，可能表现出训练集和测试集房产数据中都缺失的新的或意外的特征组合。

在这一背景下，假设一处过去 15 年从未翻新的房产在生效估价日期前一个月挂牌，我们如何估计其成交价格？一个 CVR2 表现强劲的模型，比一个未经过这种反复交叉验证的模型更有可能给出最准确的估计。同样至关重要的是，数据集中要包含足够数量、翻新、维修和现代化改造程度各异的可比房产。

归根结底，如果我们拥有一个稳健的模型，能够准确预测可比房产数据集中各项房产特征的价值贡献，我们就能有效地把新的估价对象放入模型按残差对这些房产的排序之中。这一定位使我们能够为该房产赋予恰当的 CQA 分值和残差。

12 残差

正如“残差约束法”这一名称所示，残差是本方法论的核心。此处的残差是指房产的净成交价格与回归分析所预测的成交价格之间的差额。

正残差表明，基于建筑面积（GLA）、地块面积等典型可测量属性，该房产的成交价高于其估计价格，暗示其吸引力可能高于平均水平。负残差则表明该房产的成交价偏低，可能意味着吸引力较低或存在其他未被捕捉到的负面属性。

考虑到较大的住宅通常具有较大的残差，我们可以使用基于 GLA 的每平方英尺残差进行更规范化的比较。关于这一问题的更深入讨论将留待另一篇论文。

12.1 排序与评分

在为已测量变量建立 MARS 模型之后，为每一处房产计算残差。这些残差表示观测到的净成交价格与模型预测的成交价格之间的差额。还应额外建立一系列“残差/平方英尺”，即残差除以住宅的居住面积——这是基于这样的认识：一般而言，具体房产的残差会随居住面积的大小而增大。其他可能的统计量在某些社区中或许效果更好，但那是另一个话题。

假设 7：“按残差排序”

在本文中，凡提及“按残差排序”，均应理解为按残差排序、按残差/平方英尺排序，或按估价工程师认为最适合该市场区域的其他残差函数排序。作者在旧金山湾区的经验表明，残差/平方英尺通常最为可靠，尤其是在处理 GLA 异常小或异常大的估价对象时。

接下来，将各房产按其残差从最负到最正进行排序。然后为每一处房产生成一个“状况—质量—吸引力”（CQA）评分，取值范围为 0.0 至 10.0。该评分的计算方法如下：

CQA 评分对应于某一房产的残差在所有房产中的百分位排名。具体而言，评分的计算方法是：确定残差低于该房产的房产所占的百分比，再将该百分比除以 10。例如，若某一房产的残差高于 50% 的房产的残差，则其评分为 5.0。

由于四舍五入的原因，可能有一处房产获得 10.0 分，即残差最高的那处房产。然而，10.0 的满分在理论上是不可能的，因为从逻辑上讲，没有任何房产的残差能够高于 100% 的房产（包括其自身）。另一方面，最低分 0.0 是可能出现的，因为至少有一处房产不存在残差更低的其他房产。请注意，当比较的房产数量很多时，由于四舍五入，可能不止一处房产得到 0.0 分。

因此，CQA 评分提供了一个直观的度量，用以衡量某一房产的属性与模型预期的吻合程度——评分越高，表明与模型预测的吻合度越好。

必须承认，只有当销售条件与其市场价值相称时，CQA 评分才可被视为可比房产吸引力的间接指标。估价工程师有责任通过对销售交易进行全面解释和彻底调查，来查明残差价值中是否包含高于或低于市场的因素。虽然显著高于或低于市场的销售并不常见，但在强制出售、遗嘱认证、当事人无行为能力、串通等情形下仍可能发生。

表 3 展示了 MLS 数据的一个半真实节选，连同残差和 CQA 评分（包括总额和每平方英尺两种形式）。本研究分析的房产总数约为 600 处。值得注意的是，若干参与回归的变量并未显示。换言之，所展示的残差是在 MARS 回归已考虑了成交日期等其他变量的影响之后得出的。

表 3: 按残差排序的可比交易实例

成交价格	估计成交价	残差	CQA	残差/平方英尺	CQA SF	GLA	成交日期
\$3,450,000	\$2,318,865	\$1,131,135	10	\$420.50	9.98	2,690	2021-04
\$4,360,757	\$3,399,922	\$960,835	9.9	\$290.28	9.82	3,310	2021-09
\$4,000,000	\$3,134,444	\$865,556	9.8	\$234.57	9.60	3,690	2020-11
\$5,850,000	\$5,150,966	\$699,034	9.7	\$109.91	8.35	6,360	2021-12
\$3,600,000	\$3,023,366	\$576,634	9.5	\$269.46	9.79	2,140	2022-06
\$4,425,000	\$3,987,873	\$437,127	9.2	\$127.44	8.68	3,430	2021-08
\$2,530,000	\$2,164,564	\$365,436	8.9	\$180.02	9.28	2,030	2019-12
\$4,250,000	\$3,884,719	\$365,281	8.8	\$87.18	7.91	4,190	2021-08
\$3,350,000	\$3,034,613	\$315,387	8.5	\$130.87	8.75	2,410	2021-09
\$3,800,000	\$3,522,028	\$277,972	8.2	\$77.43	7.72	3,590	2020-09
\$4,200,000	\$3,963,748	\$236,252	7.8	\$56.12	6.89	4,210	2021-01
\$2,900,000	\$2,702,887	\$197,113	7.5	\$89.19	7.92	2,210	2021-07
\$4,250,000	\$4,103,885	\$146,115	6.9	\$39.38	6.29	3,710	2022-08
\$5,000,000	\$4,864,958	\$135,042	6.8	\$28.25	5.85	4,780	2021-07
\$4,320,000	\$4,199,672	\$120,328	6.5	\$33.90	6.04	3,550	2022-07
\$2,600,000	\$2,479,904	\$120,096	6.5	\$60.96	7.00	1,970	2021-10
\$3,400,000	\$3,301,146	\$98,854	6.2	\$30.32	5.90	3,260	2021-06
\$2,000,000	\$1,901,987	\$98,013	6.2	\$73.14	7.54	1,340	2019-11
\$3,000,000	\$2,919,718	\$80,282	5.9	\$26.50	5.79	3,030	2020-05
\$6,750,000	\$6,699,089	\$50,911	5.6	\$7.75	5.20	6,570	2020-08
\$5,652,000	\$5,601,570	\$50,430	5.5	\$12.83	5.37	3,930	2022-05
\$2,600,000	\$2,579,082	\$20,918	5.2	\$10.51	5.33	1,990	2022-03
\$1,925,000	\$1,905,441	\$19,559	5.1	\$11.51	5.36	1,700	2021-01
\$3,200,000	\$3,209,500	-\$9,500	4.8	-\$3.21	4.88	2,960	2020-11
\$4,150,000	\$4,187,560	-\$37,560	4.5	-\$10.49	4.68	3,580	2021-04
\$2,700,000	\$2,776,510	-\$76,510	3.9	-\$23.61	4.14	3,240	2021-07
\$3,450,000	\$3,567,668	-\$117,668	3.4	-\$31.05	3.90	3,790	2020-09
\$1,975,000	\$2,109,255	-\$134,255	3.3	-\$55.48	3.24	2,420	2020-02
\$1,950,000	\$2,085,431	-\$135,431	3.2	-\$68.40	2.86	1,980	2021-02
\$3,695,000	\$3,859,154	-\$164,154	3.1	-\$47.04	3.56	3,490	2021-05
\$2,288,000	\$2,455,747	-\$167,747	2.9	-\$56.86	3.19	2,950	2020-06
\$2,635,000	\$2,802,781	-\$167,781	2.9	-\$56.87	3.18	2,950	2020-09
\$2,050,000	\$2,217,794	-\$167,794	2.9	-\$82.66	2.46	2,030	2021-07
\$1,550,000	\$1,731,804	-\$181,804	2.7	-\$130.79	1.32	1,390	2021-05
\$2,800,000	\$3,028,411	-\$228,411	2.4	-\$81.58	2.51	2,800	2021-02
\$3,700,000	\$3,976,381	-\$276,381	1.9	-\$76.99	2.62	3,590	2021-08
\$2,517,500	\$2,834,974	-\$317,474	1.5	-\$120.71	1.48	2,630	2019-11
\$2,550,000	\$2,905,330	-\$355,330	1.1	-\$115.74	1.64	3,070	2019-12
\$1,815,000	\$2,263,944	-\$448,944	0.7	-\$273.75	0.15	1,640	2021-11
\$3,711,000	\$4,683,454	-\$972,454	0	-\$226.68	0.37	4,290	2021-12
\$2,225,000	\$3,332,960	-\$1,107,960	0	-\$370.56	0.03	2,990	2021-08
\$3,550,000	\$4,746,489	-\$1,196,489	0	-\$274.42	0.14	4,360	2022-06

12.2 CQA-残差曲线

将某一社区的 CQA 与残差/平方英尺作图，必然得到图 2 所示的 CQA-残差特征曲线。该图所示为 2022 年前后加州 Burlingame 的数据。将 CQA 评分与残差值作图通常也会观察到类似的曲线；不过，我更倾向于使用残差/平方英尺，因为它能够结合估价对象的建筑面积（GLA）来解读 CQA。将残差/平方英尺乘以 GLA，即可把某一房产的 CQA 转换为相应的残差价值。

估价工程师必须仔细评估房产的规模——具体而言是建筑面积（GLA）——是否显著影响残差的大小。这一评估至关重要，因为它可以决定计算房产价值的最准确方法。例如，高端厨房电器和豪华卫浴设施可以大幅提升住宅的残差价值。然而，在厨房和卫生间的尺寸与布局普遍统一的社区中，这些升级可能与 GLA 的相关性并不强。它们对残差价值的影响，可能更多在于它们所代表的质量，而非它们所占据的空间。

相反，昂贵的外部特征和材料能够显著提升房产的门面吸引力，可能与房产的规模有更直接的关系。较大的住宅可以容纳更精致的外部设施，例如大规模的园林绿化、更优质的屋顶材料或宽敞的露台，这些自然会提升房产的整体吸引力，进而提升其残差价值。

因此，估价工程师在分析残差时必须考虑这些动态因素。他们不仅应评估高品质设施和特征所带来的内在价值，还应理解这些改进如何与房产的物理尺寸相互作用。这种全面的方法可确保估价准确反映房产的定性与定量属性。

因此，建议同时绘制 CQA-残差曲线和 CQA-残差/平方英尺曲线并研究其差异，这可以揭示某一市场区域中的模式。估价工程师可以让程序同时计算两种残差，并取某种加权平均来计算分配给估价对象的残差。其影响可能大，也可能不大。

12.3 CQA 曲线的特征

“状况—质量—吸引力”（CQA）曲线提供了一个精细的模型，用以理解房产价值如何随给定区域内的吸引力、状况和市场动态而波动。该模型捕捉了影响房地产市场的经济因素与物理因素之间的相互作用，反映了不同细分市场中房产估价的微妙现实。

12.3.1 CQA 曲线各组成部分的分析

1. 曲线左侧的指数衰减：

当 CQA 评分从约 2 降至 0 时，房产表现出快速恶化的特征。曲线的这一部分可以很好地用指数衰减函数来表示，突显了缺乏适当维护和更新的房产会以多快的速度丧失其吸引力和价值。木材腐烂、颜色褪去、金属锈蚀等物理退化会加速这种衰退，尤其是在未定期维护的住宅中。

2. 中间区段的线性增长：

市场的主体通常属于这一类别，其吸引力在评分约 2 至 8 之间近乎线性地增长。这一区段反映了那些总体上维护达标、但不一定具备突出品质或区位的房产。它代表了住房市场中广泛的中间层级，涵盖了住房市场的大多数——这些住宅基于其标准特征和整体状况进行有竞争力的定价。

3. 曲线右侧的指数增长：

在此，曲线在评分 8 至 10 之间呈指数上升，映照出市场中吸引力较高的一端。这一细分市场中的房产往往属于较富裕的人群，他们不仅投入于更优质的维护，还投入于能显著提升房产吸引力和价值的改进。指数增长反映了此类高品质房产的稀缺性以及富裕买家的旺盛需求。而且要明白，一套价格较低的住宅在吸引力方面可能评分很高，而一套面积大、价格昂贵的住宅评分却可能很低。这些动态可通过对 MARS 的专业运用而揭示出来，而大多数估价师通常不会注意到它们。

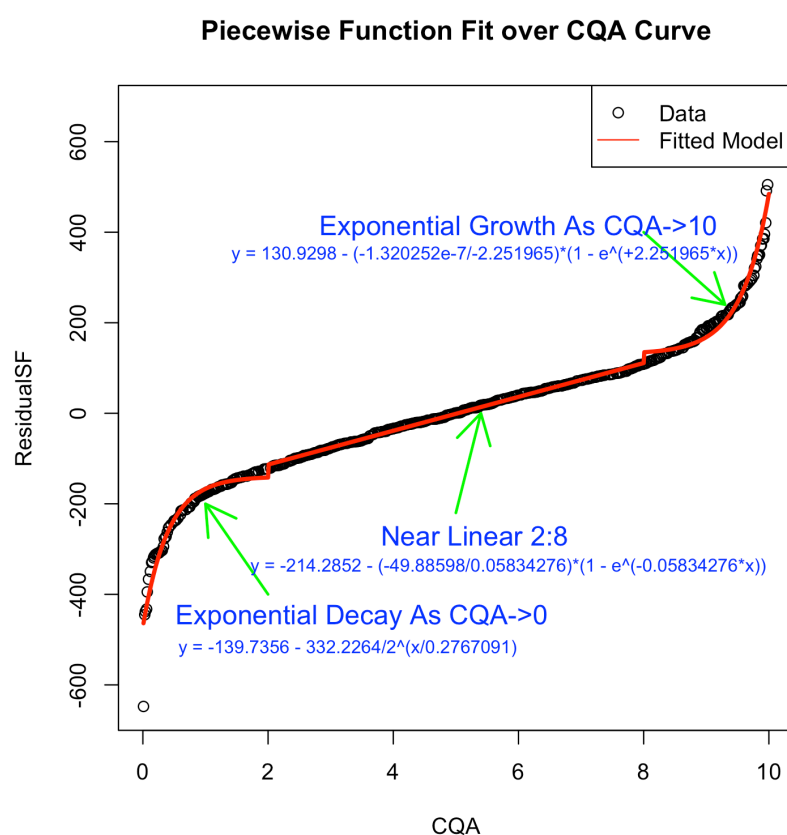
12.3.2 市场动态与社会含义

CQA 曲线隐含地讲述了不同区域内相对财富与经济差距的故事。在圣马特奥县这样的地方，Burlingame 等区域与 Hillsborough、Woodside 或 Atherton 等更富裕社区之间的房产价值差异，展示了显著的社会经济分化。这种差距影响着从房产维护到市场估值的方方面面，并最终影响买家群体构成和房地产市场趋势。

此外，还有这样一个例子：一位相对富有的人士，在自己感到舒适、且出于种种原因更愿意居住的中档社区中投资建造高品质住宅——这凸显了房地产经济学的另一个关键方面——“过度建造”房产的概念。此类房产虽然极具吸引力，但可能在当地找不到能够支付其在更高价位市场区域本可卖出的价格的买家。

CQA-残差曲线是每个市场区域所特有的，在进行分析时不应被忽视。

图 2: CQA-残差特征函数



12.4 估价对象残差

与可比交易实例不同，估价对象并不具有客观确定的残差；它必须由估价工程师（VE）估计。这一估计为残差约束法（RCA）规程引入了一定程度的主观性，标志着个人判断发挥关键作用的一个重要环节。

为了估计“状况—质量—吸引力”（CQA）评分，VE 会细致地分析已排序的房产，在排序中寻找可辨识的模式。这一分析通常涉及对每处房产相关 MLS 照片的详细审阅。对于大型数据集（如 600 处房产）而言，逐一详尽审阅并不现实。相反，VE 运用实务推理来近似判断估价对象在整体排序中的可能位置。需要大量修缮的房产（即“待修房”）通常位于排序底部附近，而更新程度高、吸引力强的房产则往往排在顶部附近。

随后, VE 采用更聚焦的方法, 在排序中“跳跃”, 以识别与估价对象吸引力最接近的可比交易实例。随着范围的缩小, VE 会更加关注各房产之间的相似之处, 尤其是在质量、状况、设计美感(包括油漆、木作、石作和窗户等元素)、功能效用及其他相关特征等方面

必须理解的关键一点是, VE 的判断从根本上受市场所确立的排序的引导。这确保了该过程中的主观因素被锚定于市场现实, 力求对估价对象作出公平而准确的估价。

12.5 RCA 价值结论

12.5.1 一般考虑

确定 CQA 评分对于确立房产的估计残差价值至关重要, 因为这两个要素本质上是相互关联的。虽然不使用 CQA 也可以对房产进行评估, 但这一指标通过展示该房产相对于市场区域内其他房产所处的位置, 提供了有价值的背景信息。例如, CQA 评分为 3.2 表明该房产的吸引力高于周边 32% 的可比房产。如果对所赋评分存在异议, 异议方必须查阅排序, 并基于照片证据或其他实质性数据来论证其提出的任何调整。如果排序体系是稳健的, 且估价工程师(VE)已在此框架内准确定位了该房产, 则不太可能出现重大分歧。

在报告阶段, 通常会选择 6 至 12 个近期且相关的可比交易实例来支持估价结论。对这些可比交易实例与估价对象之间的差异进行分析, 并对每个变量作相应调整。然后将这些调整应用于可比交易实例的净成交价格, 得出调整后成交价格。正如数学上已证明的那样, 这些调整后价格应收敛于估价对象的估计成交价格。

12.5.2 复核、审计与投诉

一旦确定了估价对象的估计残差, 就可以为所有可比交易实例计算残差调整。这些调整对 VE 针对其认为相关的特定 CQA 变量所作的进一步手工分解构成约束。必须强调的是, 正如证明 II 所确认的, 这些细化分解仅具解释性, 并不影响最终估值。

对于不满意估价价值结论的用户所提出的投诉, 传统估价师通常不得不浪费时间解释为什么某处价格更高或更低的房产并不具有用户所期望的价值影响。而使用 RCA, 所有可能的可比交易实例(通常为 100-600 个)通常都被评估为与价格结论完全相同的价值, 因此展示任意某处房产的调整几乎不费力气。除估价对象残差的评分外, 整个 RCA 过程都是自动化且客观的。因此, RCA 的另一个优势在于, 针对 VE 主观判断不完善之处的批评主要局限于这一个环节——并且可以得到有效而高效的管理。此外, 假设 R^2 约为 80%, 估计估价对象残差时 $\pm 10\%$ 的误差对最终价值估计的影响仅为 $\pm 2\%$ 的误差。

在许多情况下, 估价对象在房产残差排序中的契合度如此之高, 以至于可以预期其准确度在 $\pm 1\%$ 以内。

估计残差与估价对象的 MARS 估计成交价格之和, 即得出最终的价值结论, 称为残差约束法(RCA)价值结论。这一过程确保了房产估价方法的全面性与透明度, 增强了人们对评估结果准确性与公正性的信心。

13 结论

残差约束法是一个具有三项实质性贡献的框架：其一，源于在整个市场而非小规模可比交易实例集合上拟合模型的结构抗锚定性（证明 I）；其二，一项不变性结果，使残差分解得以作为解释层而不扰动价值结论（证明 II 及其后的注记）；其三，一项经验规律——CQA-残差特征曲线——它似乎以特征性的“指数—线性—指数”形态在各市场区域中反复出现。该框架的局限性同样具有实质性：残差分解具有不变性但不具识别性；CQA 定位步骤依赖尚未经正式评分者间一致性研究验证的 VE 判断；其准确性主张尚待与替代方法进行基准比较。

本文所呈现的是一个框架和一套规程。本文尚未呈现、但目前正在进行的更大研究计划所提供的，是一条可执行、可审计追溯的流水线，它以具有可辩护默认设置的软件将该规程付诸操作。这条流水线——用于 MARS 基函数阶段的 CRAN 上的 earthUI 包、用于在这些基函数上构建正则化结构模型的 glmnetUI 包，以及用于空间平滑的 mgcvUI 包——是本期姊妹文章的主题，也是从“作为方法的 RCA”通往“作为可交付成果的 RCA”的桥梁。估价监管中的理论与实践鸿沟（Craytor, 2026, 本期）说明了这座桥梁为何重要：估价产出在经济上被当作前瞻性风险输入来使用，而一个使现值意见可审计、可复现的框架，是对这种前瞻性用途进行任何诚实探讨的前提条件。RCA 是这一更大计划中的一步棋。它并非整个计划，本文也并未如此宣称。

参考文献

- Ambrose, Brent W. et al. (2025). “Do Appraiser and Borrower Race Affect Mortgage Collateral Valuation?” In: *Review of Finance*. DOI: [10.1093/rof/rfaf046](https://doi.org/10.1093/rof/rfaf046).
- Craytor, William Bert (2026). “The Doctrine–Practice Gap in Real Estate Appraisal: A Structured Account of Functions, Boundaries, and Tensions”. In: *Valuation Engineer Journal* 1.1.
- Friedman, Jerome H. (1991). “Multivariate Adaptive Regression Splines”. In: *The Annals of Statistics* 19.1, pp. 1–141.
- Hastie, Trevor, Robert Tibshirani, and Jerome H. Friedman (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second. Springer Series in Statistics. New York, NY: Springer. ISBN: 978-0-387-84857-0.
- James, Gareth et al. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics 103. New York: Springer. ISBN: 978-1-4614-7137-0.
- Krivorotov, George and Michael LaCour-Little (2020). “AVM versus Appraisal-Based Underwriting in Refinance Mortgages: The Trade-off Between Noise and Bias”. In: *SSRN Electronic Journal*. DOI: [10.2139/ssrn.3603327](https://doi.org/10.2139/ssrn.3603327).
- Kuhn, Max and Kjell Johnson (2016). *Applied Predictive Modeling*. Corrected at 5th printing. New York: Springer. ISBN: 978-1-4614-6848-6.
- Milborrow, Stephen (2024a). *Earth: Multivariate Adaptive Regression Splines*. R Package. URL: <https://CRAN.R-project.org/package=earth>.
- (2024b). *Notes on the Earth Package*. Tech. rep. URL: <http://www.milbo.org/doc/earth-notes.pdf>.
- (2024c). *Variance Models in Earth*. Tech. rep. URL: <http://www.milbo.org/doc/earth-varmod.pdf>.
- Perry, Andre M., Jonathan Rothwell, and David Harshbarger (2018). *The Devaluation of Assets in Black Neighborhoods*. Tech. rep. Brookings Institution.
- Perry, Andre M., Jonathan Rothwell, J. Williamson, et al. (2024). *How Racial Bias in Appraisals Affects the Devaluation of Homes in Majority-Black Neighborhoods*. Tech. rep. Brookings Institution.
- Property Appraisal and Valuation Equity Task Force (2022). *Action Plan to Advance Property Appraisal and Valuation Equity*. Tech. rep. U.S. Department of Housing and Urban Development.

- Tzioumis, Konstantinos (2016). “Appraisers and Valuation Bias: An Empirical Analysis”. In: *Real Estate Economics* 45.3, pp. 679–712. DOI: [10.1111/1540-6229.12133](https://doi.org/10.1111/1540-6229.12133).
- Williamson, Jake and Mark Palim (2022). *Appraising the Appraisal: A Closer Look at Divergent Appraisal Values for Black and White Borrowers Refinancing Their Home*. Tech. rep. Fannie Mae.
- Zhang, Wengang (2020). *MARS Applications in Geotechnical Engineering Systems: Multi-Dimension with Big Data*. Singapore: Springer. ISBN: 978-981-13-7421-0.