

遗漏变量偏误 (Omitted Variable Bias)

Foundations • The Valuation Engineer

William Bert Craytor • October 1, 2026

Version 1.2.0

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

形式定义。 遗漏变量偏差是指当一个影响价格的特征被排除在模型之外，且该被排除的特征与模型中某个已包含特征呈正相关时——在已考虑其他所有已包含特征之后，该被排除特征仍与该已包含特征呈正相关——所引入的系统性偏差，而辅助回归（即对被排除特征与已包含特征之间关系的二次回归）正是用于衡量这一偏差。随后，被排除特征的影响便会渗入到已包含特征的，而辅助回归（即对被排除特征与已包含特征之间关系的第二次回归）正是用来衡量这一现象的。此时，被排除特征的影响便会渗入已包含特征的系数中（即模型赋予每个特征的每单位美元值）。

直观说明。 某笔成交价可能出现偏差，是因为其中包含了另一项特征的数值。第 1 期《Comp H》的成交价比根据其四项记录特征进行的回归分析预测值高出 \$46,455。原因在第 1 期的潜在变量栏目中披露：近期进行的厨房翻新，据该条目所述，这使房屋价值增加了约 \$60,000。“房屋状况”虽在特征列表中，但数据中未记录翻新情况，因此翻新带来的价值在数据中缺失。该价值并未消失，而是通过已包含特征的权重间接体现出来——因为在实际成交中，这些特征恰好与翻新价值相关联。要使这种情况发生，必须满足两个条件：缺失的值必须对价格有影响，并且它必须与列表中的一项特征相关联。当它对价格有影响但未与任何特征相关联时，比率是正确的，只是散点分布范围更广。无论一项特征对应一个比率还是多个比率，情况都是如此：《期刊》的主要回归方法 MARS（多变量自适应回归样条，Friedman 1991）会在特征值的不同区间内赋予其不同的回归率，此时借用的值会被分散到这些区间中，而不是集中在一个数值上，但它确实存在，原因也是一样的。

对于希望了解具体机制的读者，请参阅：。在普通回归中，比率即为系数——即模型赋予特征增加一个单位所对应的美元数额——而借用值的具体大小为：被剔除特征自身的系数乘以其与纳入特征的关联强度，其中后者是被剔除特征对纳入特征进行辅助回归所得的系数。该条目详细阐述了这一公式（即线性情况），随后在网格模型中展示了相同的“泄漏”现象——在此模型中，估价师的配对销售数据与系数一样，同样会传递这种“泄漏”。该条目专为从市场数据中获取调整值的估价师而撰写。

本期示例的要点及其局限性。 在第 007 条（条件期望）至第 010 条（参数识别）中使用的构建性帕西菲卡市场，将车况视为具有美元单价的特征——这与传统评估方法以及《期刊》第 1 期所采用的方法一致——而该示例揭示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，本期刊采用的是基于 RCA（残差约束法）的评估方法——即第 1 期（Craytor 2026）中提出的残差约束法（RCA）——该方法将车况和质量排除在回归分析之外；条目 007（条件期望）阐述了具体方法。读者应汲取该示例的教训，而非照搬其具体规格。

房屋状况、状况评分以及翻新改造。 这三者往往容易被混为一谈。房屋状况指房屋的现状：其表面装饰的磨损程度、各项系统的使用年限以及积压的维护问题。状况评分则是对该现状的评估，其基准是该房屋所在特定市场区域内周边房屋的状况。该评分基于估价师对该市场区域的了解，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实质性冲突。在本期刊中，该评分范围为 0.00 至 10.00：即市场区域内状况更差的房屋所占比例除以十。5.00 分意味着该市场区域内有一半房屋的状况更差；9.00 分则意味着十户中有九户状况更差。该分数仅为估计值，两栋得分相同的房屋实际情况可能存在差异。翻新既不是一种状态，也不是一个分数，而是一项事件：指在近期内完成的、改变了房屋状况的工程。其价值取决于改造所改变的要素——即房屋状况，通常还包括质量、材料和饰面——并按市场对此类要素的定价标准计算。翻新工程的影响通常主要体现在房屋状况上，对质量的影响较小，这主要是由于人工成本所致；但在某些翻新项目中，通过质量提升所增加的价值可能高于状况改善带来的价值。一栋经过翻新的房屋与一栋从未翻新的房屋，若在所有当

前影响价值的特征上完全一致，其价值应相同；市场支付的是翻新所产生的状况和质量，而非单纯因为发生了翻新这一事实。因此，翻新本身并非一项具有独立价格的特征。当数据集在房屋状况评分旁包含一个翻新指标时，该指标代表着翻新带来的、而模型中其他特征（建筑面积、地块、景观及状况评分）未记录的内容：即超出评分所反映的房屋状况，以及数据集未记录的装修质量。这就是《潜在变量》第 1 期中关于 Comp H 翻新工程“使其房产组合价值增加了约 \$60,000”这一陈述的解读方式，也是本文例子的核心所在。

要产生偏差，必须同时满足两个条件：被剔除的特征必须对价格产生影响，并且其变化必须与被纳入的特征之一呈正相关。如果这两个条件中任一不成立，则不存在偏差。

这一两部分的要求之所以重要，是因为该规则的常见表述——即“省略了重要变量的模型会得出错误结果”——按其表述本身而言是错误的。如果某个特征影响价格，但与任何已纳入的特征都不呈同步变动关系，它会扩大价格在模型周围的离散程度，从而增大标准误（衡量估计系数与超大样本所得数值之间偏差大小的指标），但不会改变系数本身；截距（模型中的常数金额）会吸收被剔除特征的平均贡献。在这种情况下，丢失的是精度，而非准确度。另一方面，如果某个特征虽与纳入的特征呈正相关，但对价格没有影响，则完全不会造成任何损失。只有当上述两种情况同时成立时，才会产生偏倚。

其工作原理是基于替代法。如果影响价格的某个特征未包含在自变量（即回归模型用于解释价格的特征）中，普通最小二乘法（OLS）不会简单地将其影响留作未解释。那些与缺失特征呈正相关或负相关、且已被纳入模型的特征，会部分替代该缺失特征，其系数会包含它们所能预测的缺失特征效应中的那部分。因此，模型中某个特征的系数不再仅仅代表该特征本身的效应，而是该特征的效应加上它所“借用”的缺失特征效应中的一部分。

偏差总是与特定问题相关的。不包含“重新建模”指示变量的模型系数，是对一个问题的正确答案——即在销售案例中，价格如何随条件评分变化，且与该评分相关的所有因素均随之变化——但若将其解读为“在其他条件不变的情况下，市场对车况评分每增加 1 分所对应的美元价值”（而调整模型正是如此解读的），则这些系数就成了错误的答案。该模型并非因为存在其他特征而错误；而是其系数对于特定用途而言是错误的。

借用部分的大小是确定的，由一个公式给出。对于被排除的特征 w ，其真实系数为 β_w ；此时，被纳入特征 z_j 的系数偏差为

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

，其中 $\hat{\beta}_j$ （帽子符号表示估计值）是回归对 z_j 所估计的系数，而 δ_j 是 z_j 在辅助回归中的系数，该回归以 w 为自变量，所有被纳入特征为因变量。这就是泰尔（Theil）公式¹。根据拟合的回归模型计算得出，若使用被剔除特征的估计系数，该公式在每个样本中都完全成立；若使用真实系数，则在平均意义上成立。

乘积形式对应线性情况。在 MARS 模型中，被剔除特征的影响会被与其相关的任意一段区间所吸收，因此偏差分散在这些区间中，而非集中于一个系数，但偏差确实存在，且成因相同。

该公式比表面看起来更有用，因为即使这两个因素都无法量化，通常也能根据市场知识为它们赋予正负号。当被剔除的特征推高价格（ $\beta_w > 0$ ），且与某个已纳入特征呈正相关时（ $\delta_j > 0$ ），在其他已纳入特征保持不变的情况下，该特征的系数就过高了。如果估价师仅了解这两种关系的趋势方向，就能判断误差的方向。需要注意的是： δ_j 是指在其他纳入特征保持恒定的情况下建立的关系；而当纳入的特征不止一个时，简单相关性（-1 至 +1 中的数值，用于衡量两个特征如何同步涨跌，忽略所有其他因素）可能会呈现相反的趋势。在下面的示例中，居住面积与翻新指标呈正相关（系数 +0.09），但其 δ 为负值，因此系数过低。

估价实务中何处遇到。销售比较网格可以写成一个加法模型，其结构与回归模型相同——即人工挑选的少数几个特征，每个特征对应一个美元调整值——而当其调整值从市场数据中提取出来时，它就会像任何回归模型一样受到遗漏变量偏差的影响。网格中的行即为纳入的特征。任何影响房屋价格但未被纳入网格的特征均会被忽略。

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

通常被忽略的因素包括：房屋状况中未被状况评分所涵盖的部分、未记录质量评分的房屋质量、户型布局的功能性、自然采光、噪音影响、街区内的具体位置、卖家的出售动机，以及房源推广所付出的营销努力。这些因素都可能影响价格，其中许多因素与网格已纳入考量的因素相关。

翻新是一个极具参考价值的案例，因为相关信息已被记录在案，且至今仍能反映其价值。URAR（统一住宅评估报告）会记录厨房和卫生间是否经过更新或翻新，以及具体是在何时进行的；而UAD（统一评估数据集）中的状况定义本身，正是基于更新程度来确定的。然而，该网格中设有“状况”和“质量”两行，却没有“翻新”这一行——这完全合理，因为当问题在于市场对房屋状况的估值时，“状况”和“质量”两行已涵盖了翻新带来的影响，若再增加第三行，则会导致重复计算。反之，当问题在于拟议的翻新将如何影响该房屋的价格时，若将房屋状况和质量固定不变，便剥夺了翻新发挥作用的途径；条目 009（内生性）探讨了这一问题。这种“泄漏”源于残差评分体系的遗漏。在一个市场中，最高房况评分主要属于近期翻新过的房屋，而翻新带来的价值并未被评分记录下来；在传统评估中，从这些交易中提取的房况调整值——无论是通过配对销售法，还是通过包含房况变量的回归分析——都会将这一价值一并计入。根据泰尔（Theil）公式，这种调整值便过高：它不仅包含了房屋状况本身的效应，还包含了一部分残差未记录的价值。如果估价师采用该调整值，并将其应用于因维护得高分的房屋，调整幅度就会过大。根据《期刊》的“残差约束法”，不会提取任何状况率。估价师将每套可比实例的残差，根据其判断，在房屋状况、质量以及数据未测量的其他特征之间进行分配，且须遵循一项约束条件：各残差成分价值的贡献之和必须等于该可比实例的残差。因此，翻新工程的价值将由估价师根据实际情况进行拆分，而非由单项指标承担。

刚才提到的“泄漏”现象，更准确地反映了估价师对市场基准调整所提出的异议，即基于某一组数据集得出的调整结果难以有效迁移。原因在于，提取出的调整值既包含明确指出的效应，也包含与其相关的未明确指出的效应，而其构成方式取决于原始数据集中各项因素之间的关联关系。

另一个例子是，在传统评估中，所谓的“配对销售”——即两笔交易在除待评估特征外的所有方面都完全相同，其价格差被视为该特征的价值——得出的数值却高得难以置信。原因可能并非测量误差，而是“泄漏”现象：这两笔交易不仅在所指特征上存在差异，还存在与其相关的其他因素，而价格差异正是由这两方面共同造成的。

对结论可辩护性的意义。一份未讨论遗漏变量偏差的报告，其享乐主义调整和网格调整将因此面临这一质疑。

该偏差无法从拟合统计量中检测出来。一个模型可能具有较高的 R^2 （即模型再现的价格变异比例）、较小的残差（即销售价格与模型预测价格之间的差异），且系数通过了常规的统计检验（ t 检验，该检验旨在判断系数是否可能为零），但仍可能存在严重的偏倚，因为偏倚源于特征之间的相关性，而非残差的大小。仅凭 R^2 来证明模型的可靠性并不能解答这个问题。在下面的示例中， R^2 为 0.97，而条件系数却存在 74% 的偏差。

更多的数据并不能解决问题。随着样本量的增加，偏差并不会减小。如果从同一个市场中抽取一千笔销售数据，但仍遗漏了同一项特征，所得的估计值仍会以同一个错误数值为中心，只是估计范围更窄了：标准误差会围绕着这个存在偏差的估计值收敛，这比围绕该估计值建立一个宽广的置信区间更糟糕，因为报告看起来会显得更有把握。对于 MARS 模型也是如此：更多的销售数据会使拟合曲面更平稳，但仍围绕着同一个存在偏差的估计值；地球的 varmod 区间衡量的是价格的分散程度，而非偏差。这纠正了第 1 期中关于“潜在变量”的论点，该论点曾指出随着可比实例的增加，偏差会缩小，且其他系数存在向上偏差。偏差不会随着交易量增加而缩小；在第 1 期报告中仅有的 8 笔交易中，与下文的 16 笔交易一样，视野、GLA 和地块系数均存在向下偏差；只有条件系数存在向上偏差。

即使无法确定偏置的大小，通常也能对其方向进行合理论证。在传统评估中，当通过配对成交案例或包含房屋状况变量的回归分析从成交数据中提取房屋状况调整值时，实际做法并非断言该调整值是无偏的，而是列举可能被忽略的特征，运用市场知识为 β_w 和 δ_j 中的每个特征赋予正负符号，并说明由此产生的偏差方向。“房屋状况调整值可能被高估了。评分最高的交易大多是近期翻新过的房产，而翻新带来了数据中未记录的改善，例如新厨房和新的装修面层。这些改善推高了价格，且主要出现在评分较高的交易中，因此其价值的一部分已被计入房屋状况调整值中。”这是一

条合理的陈述。这两个要求均已满足，且指向同一方向：被忽略的改进推高了价格（ β_w 为正值），且与高残差相关（ δ_j 为正值），因此偏差是向上的。这样的陈述将一个无法量化的问题转化为一个已被承认并明确了方向的问题，这正是读者在权衡结论时所需要的。在“残差约束法”下，不会提取任何状况率，因此关于状况的问题也就不会出现。

估价计算示例。Pacifica 数据集的 16 处房产扩展数据 记录了每处房产的总居住面积（GLA）、地块面积、景观、状况评分、观察到的翻新指标以及成交价格。由于每个价格都是根据已知的公式计算得出的，且不包含随机成分（数据无噪声），因此使用正确特征的回归模型能够精确还原该公式，而任何与该公式的偏差均属于偏倚，完全不包含噪声成分。

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

可比实例 A 至 H 均为第 1 期，其已公布的属性及价格均未发生变化；第 1 期用文字描述了品相，并用 1–3 的代码进行回归分析；根据本期刊的评分标准，这些文字对应“一般”5.00、“良好”7.00、“极佳”9.00，因此第 1 期的一个等级对应两分。Comp H 的翻新——在第 1 期中被披露为其残差的来源，但未作为变量记录——现已记录在案。市场对每一点状况评分支付 \$20,000。在构建的数据集中，一栋经过翻新的房屋售价比拥有相同 GLA、地块、景观和状况评分的房屋高出 \$60,000。这 \$60,000 并非翻新行为本身的价格；它代表着翻新带来的、而其他特征未记录的附加价值。翻新房屋的平均残差也更高（8.20 对比 6.64），因此翻新指标会随其中包含的房屋状况评分而变化，而房屋状况评分正是偏见所要求的第二个条件：corr(remodel, cond) = 0.55。

针对同一组十六处房产进行的两次回归分析：

	重构部分已省略	翻新工程已包含在内	偏见
截获	−83,923.57	−30,000	−53,923.57
GLA	586.64	600	−13.36
Lot	24.32	25	−0.68
查看	79,178.46	100,000	−20,821.54
条件（每项）	34,815.56	20,000	+14,815.56
翻新	—	60,000	
R^2	0.9733	1.0000	

翻新线上的 \$60,000 并非翻新价格。在这个构建的数据集中，这是在总可租赁面积（GLA）、地块、景观和评分均相等的情况下，与翻新指标相关联的价格差；这就是“恢复生成函数”所恢复的内容。

条件系数高出 \$14,816 一个点——其报告的数值为 \$34,816 (每点条件), 而市场数据为 \$20,000, 高估了 74%。视图系数低了 \$20,822。这两个误差在拟合中均不可见: 不含翻新指标的模型具有 $R^2 = 0.9733$, 其残差除三笔接近 \$35,000 的交易外, 看起来都很普通。无论是 R^2 中的内容, 还是对系数的检验, 均未显示任何异常迹象 (如果这些残差是噪声, 那么在通常的 5% 显著性水平下, 每个系数都会通过该检验)。

泰尔公式能精确解释这些数值。将 “remodel” 指示变量以所选的四个特征为自变量进行回归, 得到 $\delta_{\text{cond}} = 0.24692602$ 和 $\delta_{\text{view}} = -0.34702564$ 。于是

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

与观测到的偏移量完全吻合。该偏差并非近似由该公式解释; 它本身就是该公式。噪声的影响可能小于预期: 在存在噪声的数据中, 该公式在两个拟合回归 (无重塑指标和含重塑指标) 之间仍然完全成立, 只需用指标的估计系数 $\hat{\beta}_w$ 代替 β_w ; 与真实系数相比, 该公式在许多样本的平均值上仍然成立。

视角系数值得再仔细研究一番, 因为它的偏向性恰好相反, 而背后的原因颇具启发性。在其他特征保持不变的情况下, 本样本中拥有景观的房产被翻新的可能性较低 ($\delta_{\text{view}} < 0$)。因此, 景观系数被压低而非抬高: 它被 \$20,822 低估了, 而房况系数则被 \$14,816 高估了。这两种误差是同一遗漏所产生的独立影响, 而非彼此之间的转移。仅就状况进行推断的估价师根本不会预见到景观调整中存在误差——然而, 景观调整被低估了 21%, 正是这一同一遗漏的直接后果。

实际上, 损坏往往集中在估价师最不可能察觉的地方。首先考虑一下, 这个存在偏差的模型作为预测变量会产生什么结果。以一处与 Comp N 完全相同的未翻新估价对象为例——总建筑面积 (GLA) 为 1,650 平方英尺, 占地 7,400 平方英尺, 拥有景观视野, 状况残差 7.00——该模型给出的预测值为 \$1,386,914, 而具有完全相同特征的房屋真实平均价格为 \$1,385,000——根据条目 007 (条件期望) 的五项特征公式, 且包含翻新成本 = 0——误差为 \$1,914, 即 0.14 而交叉验证——即在未纳入拟合的销售数据上测试模型——也无法排除这种偏差: 良好的交叉验证结果并不能证明系数没有偏倚, 因为当特征的组合方式与拟合结果一致时, 即使模型规格有误, 也能做出良好的预测。依次剔除每个房产后, 两个最大的误差——K 处的约 \$56,000 和 G 处的 \$48,000——出现在符合本样本模式的房产上, 即高状况残差主要属于翻新过的房屋; E (其数据不符合该模式) 以约 \$45,000 的数值位列第三, 而 H 和 J (其数据同样不符合该模式) 则分别以约 \$41,000 的数值位列第四和第五。此处数据不含噪声, 因此这些误差均源于被遗漏的翻新指标, 而非噪声; 而在实际销售中, 由于存在噪声, 仅凭误差大小无法判断其源于被遗漏的特征还是普通的残差波动。

一旦将系数用作调整项而非预测规则, 问题便会显现。估价师若从该模型中提取状况调整值, 则每分将应用 \$34,816, 而市场数据为 \$20,000——对于每处从 “良好” 调整至 “优秀” 的可比实例, 该模型的调整值过高了 \$29,631——此外, 视野调整值为 \$79,178, 而市场标准值为 \$100,000。网格是逐个读取回归系数的地方, 而偏置所破坏的正是这一读数。

在那些打破了偏见所依据的模式模式的房产上, 预测同样会失准, 而且失准得很严重。像本估价对象这样维护良好的住宅 (总建筑面积 1,650 平方英尺, 占地面积 7,400 平方英尺), 虽然没有景观, 但评分极佳 (9.00), 且从未翻新, 其标价为 \$1,377,367, 而实际成交价为 \$1,325,000, 高估幅度达 \$52,367 (3.95%), 因为在这些成交案例中, 评分 9.00 的房屋大多经过翻新, 而该模型的房屋状况评分会将翻新后的 \$60,000 一并计入评分。一处翻新后除景观外与估价对象完全相同的住宅 (景观评分 7.00) 标价为 \$1,386,914, 而 \$1,445,000, 低估了 \$58,086 (4.02%)。这些就是非典型房产——估价师最常被专门聘请来评估的正是这类房产, 而报告也将针对这类房产被最仔细地研读。

References.

- Craytor, W. B. (2026). Residual Constraint Approach (RCA): Framework & Protocol. *The Valuation Engineer*, 1(1). <https://doi.org/10.67330/kgebx46>
- Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1), 1–67. <https://doi.org/10.1214/aos/1176347963>

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

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

Addendum: Prolog Formalization

The following machine-readable companion (008-omitted-variable-bias.pl) formalizes this entry in the journal's Prolog knowledge base. It is also published as a browsable galley alongside this article.

```
% SPDX-License-Identifier: AGPL-3.0-or-later
% Copyright (C) 2026 William Bert Craytor. Published by The Valuation Engineer
% Journal (Pacific Vista Net). This program is free software: you can
% redistribute it and/or modify it under the terms of the GNU Affero General
% Public License as published by the Free Software Foundation, either version 3
% of the License, or (at your option) any later version. It is distributed
% WITHOUT ANY WARRANTY, without even the implied warranty of MERCHANTABILITY or
% FITNESS FOR A PARTICULAR PURPOSE; see the COPYING file beside it or
% <https://www.gnu.org/licenses/agpl-3.0.html>.

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

% The dictionary (Issues/tools/dictionary) is not loaded at run time, so
% that the file runs unchanged in Trealla, SWI-Prolog and Scryer. The
% declarations below are plain facts, read by the dictionary linter.
dictionary_release_pinned('dictionary-2026.1').

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

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

:- initialization(consult('../..kb/pacifica_comps_extended')).
% 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)).

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

```

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

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

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

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

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

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

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

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

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

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

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

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

theil_formula_holds :-

```

```

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

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

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

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

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

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

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

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

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

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

```