

# 누락 변수 편향 (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)

**형식적 정의.** 생략 변수 편향이란, 가격에 영향을 미치는 특징 변수가 모델에서 제외되었을 때, 해당 특징 변수가 포함된 다른 특징 변수 중 하나와 동일한 방향으로 움직일 때 (다른 포함된 특징 변수들의 효과가 이미 고려된 후에) 추정된 계수에 발생하는 체계적인 왜곡을 말하며, 이동하는 경우, 이를 보조 회귀 분석 (생략된 변수가 포함된 변수들에 대해 수행하는 두 번째 회귀 분석) 이 측정하는 현상이다. 그러면 생략된 변수의 효과가 포함된 변수들의 계수 (모델이 각 변수에 할당하는 단위당 달러)로 누출된다.

**직관적 설명.** 특정 호가는 다른 특징의 값을 반영하고 있기 때문에 잘못된 값일 수 있습니다. 제 1 호지의 Comp H 는 기록된 네 가지 특징을 바탕으로 한 회귀 분석이 예측한 가격보다 \$46,455 더 높은 가격에 거래되었습니다. 제 1 호에 실린 '잠재 변수 (Latent Variable)' 에서 밝혀진 바에 따르면, 그 이유는 최근에 진행된 주방 리모델링 때문이었는데, 해당 항목에 따르면 이로 인해 주택 가치가 약 \$60,000 증가했다고 합니다. '상태 (Condition)' 는 특징 목록에 포함되어 있었지만, 리모델링에 대한 기록은 없었으며, 리모델링으로 인한 가치는 데이터에서 누락되었습니다. 그 가치가 사라진 것은 아닙니다. 해당 특징들이 판매 시 우연히 리모델링과 함께 나타났기 때문에, 포함된 특징들의 단가를 통해 그 가치가 반영된 것입니다. 이러한 현상이 발생하려면 두 가지 조건이 충족되어야 합니다. 누락된 값이 가격에 영향을 미쳐야 하며, 목록에 포함된 특징과 함께 전달되어야 합니다. 영향을 미치지만 어떤 특징과도 연결되지 않을 경우, 단가는 정확하며 분산만 더 넓어집니다. 이는 특징이 하나의 단가를 적용받든 여러 개의 단가를 적용받든 상관없이 성립합니다: 이 저널의 주요 회귀 분석 방법인 MARS(다변량 적응 회귀 스플라인, Friedman 1991) 는 특성의 값 범위별로 서로 다른 단가를 부여하며, 이 경우 차용된 값은 하나의 숫자에 국한되지 않고 해당 범위 전체에 분산되지만, 동일한 이유로 여전히 존재합니다.

기전을 알고 싶은 독자를 위해. 일반적인 회귀 분석에서 단가는 계수입니다. 즉, 모델이 특성의 단위당 추가로 할당하는 달러 금액이며, 이 차용된 값의 크기는 정확히 다음과 같습니다. 제외된 변수 자체의 계수에, 해당 변수가 포함된 변수와 얼마나 강하게 함께 움직이는지를 나타내는 값을 곱한 것이며, 이 두 번째 요소는 제외된 변수가 포함된 변수들에 대해 수행된 보조 회귀 분석의 계수입니다. 이 글은 선형 회귀의 경우인 해당 공식을 상세히 설명한 뒤, 그리드 분석에서도 동일한 누수 현상이 나타남을 보여줍니다. 여기서 감정사가 제시한 쌍을 이론 매매 사례들이 계수와 마찬가지로 그 누수를 전달합니다. 이 글은 시장 데이터를 바탕으로 조정치를 도출하는 감정사들을 위해 작성되었습니다.

**이번 호의 예시가 보여주는 것과 보여주지 않는 것.** 항목 007(조건부 기대값) 부터 010(매개변수 식별) 까지에서 사용된 가상 파시피카 시장은, 전통적인 감정 평가 방식이나 제 1 호에서와 마찬가지로 상태를 달러 단위의 효율로 표현된 특징으로 취급하며, 이 예시는 그러한 접근 방식의 문제점을 보여줍니다. 전통적인 감정 평가 방법론은 본 저널이 채택한 방법인 RCA(Constraint Approach) 기반 감정 평가—제 1 호 (Craytor 2026) 의 Constraint Approach—와 다르며, 이 방법은 회귀 분석에서 상태와 품질을 제외합니다. 항목 007(조건부 기대값) 에서는 그 방법을 설명하고 있습니다. 독자는 예제의 구체적인 사양보다는 그 교훈을 마음에 새겨야 합니다.

**주택 상태, 상태 점수, 리모델링.** 이 세 가지는 흔히 혼동되기 쉽습니다. 주택 상태는 마감재의 마모 정도, 설비의 노후 정도, 미뤄진 유지보수 사항 등을 포함한 주택의 전반적인 상태를 의미합니다. 상태 점수는 특정 시장 지역 내 주변 주택들의 상태와 비교하여 해당 주택의 상태를 평가한 추정치입니다. 이는 감정사가 해당 시장 지역에 대해 가지고 있는 이해를 바탕으로 하며, 동일한 시장 지역에서 활동하는 동료 감정사들의 이해와 크게 상충되어서는 안 됩니다. 본 저널에서는 이 점수를 0.00 에

서 10.00 까지로 표시합니다. 이는 시장 지역 내 주택 중 상태가 더 나쁜 주택의 비율을 10 으로 나눈 값입니다. 5.00 이라는 CQA 는 시장 지역 내 주택의 절반이 더 나쁜 상태임을 의미하며, 9.00 은 10 채 중 9 채가 더 나쁜 상태임을 의미합니다. 이 CQA 는 추정치이며, 같은 CQA 를 받은 두 주택의 실제 상태는 다를 수 있습니다. 리모델링은 상태나 CQA 가 아니라 하나의 사건입니다. 즉, 최근 일정 기간 내에 주택을 변화시킨 공사를 의미합니다. 리모델링의 가치는 그것이 변화시킨 요소—상태, 그리고 대개 품질, 자재 및 마감재—에 대해 시장이 지불하는 가격에 따라 결정됩니다. 리모델링의 효과 대부분은 주로 인건비 때문에 상태에 더 큰 영향을 미치고 품질에는 상대적으로 적은 영향을 미치지 만, 일부 리모델링의 경우 품질을 통해 추가된 가치가 두 가지 중 더 높은 경우도 있습니다. 가치에 영향을 미치는 모든 현재 특성이 동일한 리모델링된 주택과 리모델링된 적이 없는 주택은 동일한 가치를 가져야 합니다. 시장은 리모델링이 이루어졌다는 사실 자체에 대해 대가를 지불하는 것이 아니라, 리모델링을 통해 창출된 상태와 품질에 대해 대가를 지불합니다. 따라서 리모델링은 그 자체로 가격이 책정되는 특성이 아닙니다. 데이터셋에 상태 CQA 외에 리모델링 지표가 포함되어 있다면, 이 지표는 모델의 다른 특징들 (총연면적, 대지, 전망, CQA) 이 기록하지 않는 리모델링으로 인한 모든 요소를 나타냅니다. 즉, CQA 가 보여주는 것 이상의 상태와 데이터셋에 기록되지 않은 품질이 이에 해당합니다. 이것이 바로 『*Latent Variable*』 제 1 호에서 Comp H 의 리모델링이 “그 부동산의 묶음 가치에 약 \$60,000 을 더한다” 고 한 진술을 해석해야 하는 방식이며, 이 글의 예시가 바로 이를 바탕으로 하고 있습니다.

편향이 발생하려면 다음 두 가지 조건이 모두 충족되어야 합니다. 즉, 제외된 특징이 가격에 영향을 미쳐야 하며, 포함된 특징 중 하나와 함께 변동해야 합니다. 이 중 하나라도 충족되지 않으면 편향은 발생하지 않습니다.

이 두 가지 요건이 중요한 이유는, “중요한 변수를 생략한 모델은 잘못된 답을 내놓는다” 는 이 규칙의 일반적인 해석이, 그대로 받아들이는 경우 거짓이기 때문이다. 어떤 특징 변수가 가격에 영향을 미치지만 포함된 특징 변수 중 어느 것과도 함께 변동하지 않는다면, 이는 모델 주변의 가격 분산을 확대하여 표준오차 (추정된 계수가 매우 큰 표본에서 도출될 수치로부터 얼마나 멀리 떨어져 있을 수 있는지를 나타내는 척도) 를 넓히게 되지만, 계수 자체를 이동시키지는 않습니다. 절편 (모델의 상수 달러 금액) 이 제외된 변수의 평균 기여도를 흡수하기 때문이다. 이 경우 정확도는 떨어지지 않고 정밀도만 저하된다. 반면, 포함된 변수들과 함께 변동하지만 가격에 영향을 미치지 않는 변수는 아무런 비용도 발생시키지 않는다. 두 조건이 모두 충족될 때만 편향이 발생한다.

이 원리는 대입법을 통해 작동합니다. 가격에 영향을 미치는 특징이 회귀 변수(회귀 분석에서 가격을 설명하기 위해 사용하는 특징들) 에서 누락된 경우, OLS(최소제곱법) 회귀 분석은 해당 특징의 기여도를 단순히 설명하지 않은 채로 두지 않습니다. 누락된 변수와 함께 변동하는 포함된 변수들이 그 변수를 부분적으로 대신하며, 이들의 계수는 예측할 수 있는 범위 내에서 누락된 변수의 효과 중 일부를 반영합니다. 포함된 변수의 계수는 더 이상 해당 변수만의 효과를 나타내는 것이 아니라, 그 효과에 더해 누락된 변수의 효과 중 일부를 차용한 값을 의미합니다.

편향은 항상 특정 질문과 관련이 있습니다. 리모델링 지표가 포함되지 않은 모델의 계수들은 한 가지 질문에 대한 정답입니다. 즉, ‘이 판매 사례들에서 CQA 점수에 따라 가격이 어떻게 변동하는가?’ 라는 질문에 대한 정답이며, 이 CQA 점수와 관련된 모든 요소가 함께 반영된 경우를 말합니다 — 그러나 다른 모든 변수를 고정시킨 상태에서 시장 가격이 CQA 1 점당 얼마인지로 해석될 때, 즉 조정 (adjustment) 이 이를 해석하는 방식에 비추어 보면 잘못된 답이 됩니다. 이 모델이 틀린 것은 다른 특징이 존재하기 때문이 아니라, 그 계수들이 적용된 용도에 비추어 볼 때 잘못된 것입니다.

차용된 비율은 공식에 의해 주어진 정확한 크기를 가집니다. 제외된 특징  $w$  의 실제 계수가  $\beta_w$  일 때, 포함된 특징  $z_j$  의 계수에 대한 편향은

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

로 주어집니다. 여기서  $\hat{\beta}_j$  (모자는 추정치를 나타냄) 는 회귀 분석에서  $z_j$  에 대해 추정된 계수이며,  $\delta_j$  는 모든 포함된 특징을 변수로 한  $w$  에 대한 보조 회귀 분석에서  $z_j$  의 계수입니다. 이것이 바로 테

일 (Theil) 의 공식입니다.<sup>1</sup> 적합된 회귀 분석 결과와 제외된 특징의 추정 계수를 바탕으로 계산했을 때, 이 공식은 모든 표본에서 정확히 성립하며, 실제 계수를 사용했을 때는 평균적으로 성립합니다.

곱의 형태는 선형 경우를 나타냅니다. MARS 모델에서는 제외된 특징의 효과가 그 특징과 함께 고려되는 구간들에 흡수되므로, 편향은 하나의 계수에 집중되기보다는 해당 구간들에 분산되지만, 동일한 이유로 여전히 존재합니다.

이 공식은 겉보이는 것보다 더 유용합니다. 두 요인 모두 측정할 수 없더라도 시장 지식을 바탕으로 두 요인에 부호를 부여할 수 있는 경우가 많기 때문입니다. 제외된 변수가 가격을 상승시키고 ( $\beta_w > 0$ ) 포함된 변수 중 하나와 더 강한 상관관계를 보이며 ( $\delta_j > 0$ ), 다른 포함된 변수들은 일정하게 유지될 때, 해당 변수의 계수는 지나치게 높게 산출됩니다. 두 관계의 방향만 알고 있는 감정평가사는 오차의 방향을 파악할 수 있습니다. 단, 한 가지 유의할 점은  $\delta_j$  에서 제시된 관계는 다른 포함된 변수들이 일정하게 유지된 상태에서의 관계이며, 포함된 변수가 두 개 이상일 경우 단순 상관관계 (-1 에서 +1 까지의 수치로, 다른 모든 변수를 무시하고 두 변수가 함께 상승 및 하락하는 정도를 측정함) 는 반대 방향으로 나타날 수 있습니다. 아래 예시에서, 주거 면적은 리모델링 지표와 상관관계를 보이지만 (상관계수 +0.09), 그  $\delta$  는 음수이므로 상관계수가 지나치게 낮게 나타납니다.

**평가 실무에서 만나는 상황.** 판매 비교 표 ( ) 는 회귀 분석과 동일한 형태를 가진 가산 모델로 표현될 수 있습니다. 즉, 수작업으로 선택된 소수의 특징과 각 특징에 대한 달러 단위 조정값으로 구성되며, 이러한 조정값이 시장 데이터에서 추출될 경우 다른 회귀 분석과 마찬가지로 누락된 변수 편향의 대상이 됩니다. 표의 각 행은 포함된 특징을 나타냅니다. 주택 가격에 영향을 미치지만 행으로 표시되지 않은 특징은 모두 제외됩니다.

일반적으로 간과되는 요소로는 상태 CQA 로는 파악되지 않는 상태의 세부 사항, 품질 CQA 가 기록되지 않은 품질, 평면도의 기능성, 자연 채광, 소음 노출 정도, 블록 내 구체적인 위치, 매도자의 매도 의향, 그리고 매물 등록에 투입된 마케팅 노력 등이 있습니다. 이 모든 요소는 가격에 영향을 미칠 수 있으며, 그중 상당수는 표가 이미 고려하고 있는 항목과 함께 고려되기도 합니다.

리모델링은 기록에 남으면서도 여전히 정보가 누락되기 때문에 좋은 사례입니다. URAR(통일 주거용 감정 보고서) 에는 주방과 욕실이 개조되었는지, 리모델링되었는지, 그리고 그 시기가 언제인지가 기록되며, UAD(통일 감정 데이터 세트) 의 상태 정의 자체도 개조 정도에 따라 달라집니다. 그러나 표에는 ‘상태’ 와 ‘품질’ 에 대한 항목은 있지만 ‘리모델링’ 에 대한 항목은 없습니다. 시장의 상태별 가격을 묻는 질문일 때, ‘상태’ 와 ‘품질’ 항목이 리모델링으로 인한 효과를 이미 반영하고 있으므로, 세 번째 항목을 추가하면 동일한 효과를 두 번 계산하게 되므로 이는 타당한 처리입니다. 반면, 제안된 리모델링이 이 주택의 가격에 어떤 영향을 미칠지가 문제라면, 상태와 품질을 고정시키는 것은 리모델링이 작용하는 바로 그 경로를 제거하는 것이 됩니다. 항목 009(내생성) 가 바로 이 문제를 다룹니다. 누수는 CQA 가 포착하지 못하는 부분을 통해 발생합니다. 최고 CQA 가 대부분 최근 리모델링된 주택에 속하고, 리모델링이 CQA 체계에 기록되지 않은 가치를 창출하는 시장에서, 전통적인 감정 평가에서 해당 매매 사례를 바탕으로 추출한 상태 조정치 (쌍을 이룬 매매 사례 비교나 상태를 포함한 회귀 분석을 통해) 는 그 가치를 함께 반영하게 된다. 테일 (Theil) 의 공식에 따르면, 이러한 조정은 지나치게 높게 산정된 것입니다. 즉, 상태 자체의 효과뿐만 아니라 CQA 가 기록하지 못한 가치의 일부까지 포함하고 있기 때문입니다. 이를 받아들여 유지 관리를 통해 CQA 를 획득한 주택에 적용하는 감정사는 과도하게 조정하게 될 것이다. 『저널』 의 Constraint Approach 에 따르면, 상태 단가는 추출되지 않는다. 감정평가사는 하나의 제약 조건, 즉 잔여 구성 요소 가치 기여도의 합계가 비교 대상 주택의 잔여값과 반드시 일치해야 한다는 전제 하에, 각 비교 대상 주택의 잔여값을 상태, 품질 및 데이터로 측정되지 않는 기타 특징들 사이에 감정평가사의 판단에 따라 분배합니다. 따라서 리모델링의 가치는 단일 항목으로 처리되지 않고, 감정평가사가 적절하다고 판단하는 대로 분할됩니다.

방금 설명한 ‘누수’ 현상은 감정평가사들이 시장 기반 조정 방식에 대해 제기하는 이의, 즉 한 데이터 세트에서 도출된 조정이 다른 데이터 세트로 잘 적용되지 않는다는 점을 더 정확하게 설명해 주는 것입니다. 그 이유는 추출된 조정이 명시된 효과와 그에 수반되는 명시되지 않은 효과로 모두 구성되

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

어 있으며, 이러한 조정의 구성 방식은 해당 조정이 추출된 데이터셋 내에서 어떤 요소가 어떤 요소와 연관되어 있었는지에 따라 달라지기 때문이다.

또 다른 예로는, 전통적인 감정 평가에서 ‘쌍을 이룬 매매 사례’—즉, 문제의 특성을 제외하고는 모든 면에서 동일한 두 건의 매매 사례에서 발생하는 가격 차이를 해당 특성의 가치로 해석하는 경우—가 터무니없이 높은 수치를 산출하는 상황이 있습니다. 그 원인은 측정 오류가 아니라 ‘누수’ 때문일 수 있다. 즉, 두 매매 사례는 명시된 특징뿐만 아니라 그와 관련된 다른 요소에서도 차이가 있으며, 가격 차이는 이 두 가지 모두에 기인하는 것이다.

**타당성 확보에 갖는 의미.** 생략된 변수 편향을 다루지 않은 보고서는 그 헤도닉 조정 및 그리드 조정에 대한 반론의 여지를 남기게 된다.

이 편향은 적합 통계량만으로는 파악할 수 없습니다. 모델이 높은 설명력 ( $R^2$ , 모델이 재현하는 가격 변동의 비율), 작은 잔차 (판매 가격과 모델이 예측한 가격 간의 차이), 그리고 일반적인 통계적 검정 (계수가 0 일 가능성이 있는지 여부를 묻는  $t$ -검정) 을 통과하는 계수를 가질지라도, 그럼에도 불구하고 심각한 편향을 보일 수 있습니다. 편향은 잔차의 크기가 아니라, 특징 변수들 간의 상호 관계에서 비롯되기 때문입니다.  $R^2$  를 모델의 신뢰성을 입증하는 근거로 삼는 것은 문제의 해답이 되지 못합니다. 아래 예시에서  $R^2$  는 0.97 이지만, 조건 계수는 74% 나 잘못되었습니다.

데이터가 더 많아진다고 해서 문제가 해결되는 것은 아닙니다. 표본 크기가 커져도 편향은 줄어들지 않습니다. 동일한 시장에서 동일한 특징을 제외한 1,000 건의 판매 데이터를 분석하더라도, 추정치는 여전히 동일한 잘못된 수치를 중심으로 하며, 단지 추정 범위가 더 좁아질 뿐입니다. 즉, 편향된 추정치 주변의 표준 오차가 줄어들게 되는데, 이는 추정치 주변에 넓은 신뢰 구간이 있는 것보다 더 나은 결과입니다. 왜냐하면 보고서가 더 확신에 찬 것처럼 보일 것이기 때문입니다. MARS 의 경우도 마찬가지입니다. 판매 건수가 늘어나면 적합 곡면은 더 안정적이 되지만, 여전히 동일한 편향된 결과 주위를 맴돌게 됩니다. 지구의 **varmod** 구간은 가격의 분산을 측정할 뿐, 편향을 측정하는 것이 아닙니다. 이는 제 1 호에서 ‘비교 대상 건수가 늘어나면 편향이 줄어들며, 다른 계수들은 상향 편향을 띤다’ 고 주장했던 내용을 바로잡는 것입니다. 편향은 판매 건수가 늘어나도 줄어들지 않으며, 제 1 호의 8 건 사례와 아래의 16 건 사례 모두에서 전망, GLA 및 부지 계수는 하향 편향을 보이며, 오직 조건 계수만이 상향 편향을 보입니다.

편향의 크기를 알 수 없는 경우에도, 그 방향을 정당화할 수 있는 경우가 종종 있습니다. 판매 사례 (쌍을 이룬 판매 사례나 상태를 포함한 회귀 분석을 통해) 에서 상태 조정을 도출하는 전통적인 감정 평가에서는, 해당 조정이 편향되지 않았다고 단정하기보다는, 제외되었을 가능성이 있는 특징들을 명시하고, 시장 지식을 바탕으로 각 특징에 대해  $\beta_w$  및  $\delta_j$  의 부호를 부여한 뒤, 그 결과로 나타나는 편향의 방향을 밝히는 것이 실용적인 접근 방식이다. “상태 조정치는 아마도 과대평가되었을 것이다. 최고 CQA 점수를 받은 매매 사례의 대부분은 최근 리모델링된 주택이며, 리모델링을 통해 새로운 주방이나 마감재와 같이 데이터에 기록되지 않은 개선 사항이 추가되었다. 이러한 개선 사항은 가격을 상승시키며 주로 높은 CQA 점수를 받은 매매 사례에서 나타나므로, 그 가치의 일부가 ‘상태’ 항목으로 계산된 것이다” 라는 설명은 타당한 주장이다. 두 가지 요건이 모두 충족되며, 둘 다 동일한 방향을 가리킵니다. 즉, 데이터에서 제외된 개선 사항들이 가격을 상승시키고 ( $\beta_w$  가 양수), 높은 CQA 와 연관되어 있으므로 ( $\delta_j$  가 양수), 편향은 상향 편향입니다. 이러한 주장은 측정되지 않은 문제를 인정되고 방향성이 부여된 문제로 전환시키며, 이는 독자가 결론을 평가하기 위해 정확히 필요로 하는 것입니다. Constraint Approach 에서는 상태 단가를 추출하지 않으므로, 상태에 대한 의문은 제기되지 않습니다.

**평가 계산 예시.** Pacifica 데이터셋 의 16 개 부동산으로 구성된 확장 데이터에는 각 부동산별로 총 주거 면적 (GLA), 대지 면적, 전망, CQA 점수, 확인된 리모델링 여부, 그리고 판매 가격이 기록되어 있습니다. 모든 가격은 무작위 요소가 없는 알려진 공식에 따라 계산되었기 때문에 (데이터에 잡음이 없음), 올바른 특징을 사용하는 회귀 분석은 해당 공식을 정확히 재현하며, 그로부터의 이탈은 잡음 성분이 전혀 없는 편향에 해당합니다.

| Comp | GLA (sf) | 부지 (sf) | 보기  | 조건   | 리모델링된 | 가격 (\$)   |
|------|----------|---------|-----|------|-------|-----------|
| A    | 1,650    | 9,000   | 네   | 7.00 | 아니요   | 1,425,000 |
| B    | 1,650    | 7,200   | 아니요 | 7.00 | 아니요   | 1,280,000 |
| C    | 1,850    | 7,500   | 아니요 | 7.00 | 아니요   | 1,407,500 |
| D    | 1,450    | 7,000   | 아니요 | 7.00 | 아니요   | 1,155,000 |
| E    | 1,700    | 8,000   | 네   | 9.00 | 아니요   | 1,470,000 |
| F    | 1,550    | 7,100   | 아니요 | 5.00 | 아니요   | 1,177,500 |
| G    | 1,700    | 10,500  | 아니요 | 7.00 | 아니요   | 1,392,500 |
| H    | 1,600    | 7,300   | 아니요 | 7.00 | 네     | 1,312,500 |
| I    | 1,750    | 8,200   | 아니요 | 9.00 | 네     | 1,465,000 |
| J    | 1,600    | 7,600   | 아니요 | 7.00 | 네     | 1,320,000 |
| K    | 1,900    | 9,500   | 네   | 9.00 | 네     | 1,687,500 |
| L    | 1,500    | 6,800   | 아니요 | 5.00 | 아니요   | 1,140,000 |
| M    | 1,800    | 8,800   | 아니요 | 7.00 | 아니요   | 1,410,000 |
| N    | 1,650    | 7,400   | 네   | 7.00 | 아니요   | 1,385,000 |
| O    | 1,550    | 7,900   | 아니요 | 9.00 | 네     | 1,337,500 |
| P    | 1,720    | 9,200   | 아니요 | 5.00 | 아니요   | 1,332,000 |

비교 대상 A H는 제 1 호이며, 게재된 속성과 가격은 변경되지 않았습니다; 제 1 호에서는 상태를 문구로 기술하고 회귀 분석을 위해 1-3 의 코드로 표기했습니다. 본 저널의 CQA 평가 기준에 따르면, 해당 문구는 평균 5.00, 양호 7.00, 우수 9.00 에 해당하므로, 제 1 호의 한 단계는 2 점에 해당합니다. Comp H 의 리모델링 — 제 1 호에서 잔차의 원인으로 공개되었으나 변수로 기록되지는 않았던 — 이 이제 기록되었습니다. 시장은 상태 점수 1 점당 \$20,000 을 지불합니다. 구성된 데이터에서 리모델링 된 주택은 동일한 GLA, 대지, 전망 및 CQA 점수를 가진 주택보다 \$60,000 더 비싸게 판매됩니다. 이 \$60,000 은 리모델링 행위 자체에 대한 가격이 아니라, 다른 특징들로는 기록되지 않는 리모델링이 가져온 가치를 나타냅니다. 리모델링된 주택은 또한 평균적으로 더 높은 CQA(8.20 대 6.64) 를 기록하므로, 리모델링 지표는 포함된 상태 CQA 와 함께 변동하는데, 이는 편향 (bias) 이 요구하는 두 번째 조건인 ‘ $\text{corr}(\text{remodel}, \text{cond}) = 0.55$ ’ 에 해당합니다.

동일한 16 개 부동산에 대한 두 가지 회귀 분석:

|          | 리모델링 내용 생략 | 리모델링 포함 | 편향         |
|----------|------------|---------|------------|
| 인터셉트     | -83,923.57 | -30,000 | -53,923.57 |
| GLA      | 586.64     | 600     | -13.36     |
| 롯        | 24.32      | 25      | -0.68      |
| 보기       | 79,178.46  | 100,000 | -20,821.54 |
| 조건 (점수당) | 34,815.56  | 20,000  | +14,815.56 |
| 리모델링     | —          | 60,000  |            |
| $R^2$    | 0.9733     | 1.0000  |            |

리모델링 라인에 표시된 \$60,000 은 리모델링 비용이 아닙니다. 이 구축된 데이터셋에서 이 수치는 총 임대 가능 면적 (GLA), 대지, 전망 및 CQA 가 동일한 조건에서 리모델링 지표에 부여된 가격 차이를 나타내며, 이것이 바로 “생성 함수를 복원” 하는 과정에서 복원되는 값입니다.

조건 계수는 \$14,816 포인트나 너무 높게 산출되었습니다. — 이 모델은 조건 1 포인트당 \$34,816 을 보고하는 반면, 시장의 수치는 \$20,000 이며, 이는 74% 과대평가된 것입니다. 뷰 계수는 \$20,822 만큼 너무 낮게 산출되었다. 두 오류 모두 적합도 분석에서는 드러나지 않는다: 리모델 지표가 포함되지 않은 모델의  $R^2 = 0.9733$  와 잔차들은, \$35,000 근처의 세 건의 매매를 제외하면 평범해 보인다.  $R^2$  의 결과나 계수 검정 결과 어디에서도 문제가 될 만한 징후는 나타나지 않습니다 (만약 이 잔차들이 잡음이었다면, 모든 계수가 통상적인 5% 유의수준에서 해당 검정을 통과했을 것입니다).

Theil 의 공식은 수치를 정확히 설명합니다. 리모델링 지표에 포함된 네 가지 특징을 회귀분석한 결과,  $\delta_{\text{cond}} = 0.24692602$  및  $\delta_{\text{view}} = -0.34702564$  가 도출됩니다. 그러면

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

관측된 변화폭이 백분을 단위와 일치합니다. 편향은 이 공식으로 대략적으로 설명되는 것이 아니라, 바로 그 공식 그 자체입니다. 잡음의 영향은 예상보다 적을 것입니다: 잡음이 있는 데이터에서도 이 공식은 두 회귀 분석 결과 (리모델링 지표 포함 및 미포함) 사이에서 여전히 정확히 성립하며, 이때  $\beta_w$  대신 지표의 추정 계수  $\hat{\beta}_w$  을 사용합니다; 실제 계수에 비추어 볼 때, 이 공식은 다수의 표본에 걸쳐 평균적으로 성립합니다.

뷰 계수는 다시 한 번 살펴볼 가치가 있는데, 그 편향이 반대 방향으로 나타나며 그 이유가 시사하는 바가 크기 때문이다. 다른 요인들을 일정하게 유지했을 때, 이 표본에서 전망이 좋은 주택은 리모델링될 가능성이 더 낮습니다 ( $\delta_{\text{view}} < 0$ ). 따라서 전망 계수는 상승하기보다는 하락하게 됩니다. 즉, 이 계수는 \$20,822 만큼 너무 낮게 산정된 반면, 상태 계수는 \$14,816 만큼 너무 높게 산정되었습니다. 이 두 가지 오류는 동일한 누락에서 비롯된 별개의 효과일 뿐, 서로 간에 전이된 것이 아닙니다. 상태만 고려하는 감정평가사는 전망 조정에서 오류가 발생할 것이라고 전혀 예상하지 못했을 것입니다. 그러나 21% 나 과소평가된 전망 조정은 바로 그 동일한 누락의 직접적인 결과입니다.

실제로 손상은 감정평가사가 가장 살피지 않을 법한 곳에 집중되어 있습니다. 먼저, 편향된 모델이 예측 변수로서 어떤 역할을 하는지 살펴보겠습니다. Comp N 과 동일한 리모델링되지 않은 대상 물건의 경우 — 7,400 제곱피트 대지에 1,650 제곱피트의 총 임대 가능 면적 (GLA) 을 갖추고, 전망이 있으며 CQA 점수는 7.00 인 경우 —인 경우, 이 모델은 \$1,386,914 를 예측하지만, 정확히 이러한 특징을 가진 주택의 실제 평균 가격은 \$1,385,000 인 주택의 실제 평균 가격과 대비됩니다 — 항목 007(조건부 기대값) 의 5 가지 특징 공식에 따라, 리모델링 비용은 = 0 — 오차는 \$1,914, 즉 0.14 퍼센트입니다. 총 16 건의 매매 사례를 통틀어 전형적인 오차는 약 \$21,000 이며, 따라서 모델은 심각한 편향을 보이지만 여전히 용납할 만한 수준의 예측을 제공합니다. 이는 우연이 아닙니다. OLS 는 주어진 표본에 맞춰 추정되며, 편향된 계수들은 바로 표본 가격에 가장 근접한 값들입니다. 예측 정확도는 누락된 변수 편향을 반박하는 증거가 되지 않으며, 교차 검증 — 적합 과정에서 제외된 거래 데이터를 대상으로 모델을 테스트하는 것 — 역시 이를 배제하지 못합니다: 좋은 교차 검증 결과가 계수에 편향이 없다는 것을 의미하지는 않습니다. 왜냐하면 잘못 지정된 모델이라도, 설명변수들이 적합 과정에서와 같이 함께 나타나는 곳이라면 어디서든 잘 예측할 수 있기 때문입니다. 각 속성을 차례로 제외해 보면, 가장 큰 두 오차 — K 에서 약 \$56,000, G 에서 \$48,000 — 는 이 표본의 패턴을 따르는 속성들에서 발생하는데, 이 표본에서는 높은 CQA 점수가 주로 리모델링된 주택에 속합니다; 이 패턴을 깨는 E 는 약 \$45,000 으로 3 위를 차지하고, 마찬가지로 패턴을 깨는 H 와 J 는 약 \$41,000 으로 각각 4 위와 5 위를 기록합니다. 이 데이터에는 잡음이 없으므로, 이러한 오차들은 모두 누락된 리모델링 지표에서 비롯된 것이며, 잡음에서 비롯된 것은 하나도 없습니다. 잡음이 존재하는 실제 매매의 경우, 오차의 크기만으로는 그것이 누락된 특징에서 비롯된 것인지, 아니면 일반적인 잔차 분포에서 비롯된 것인지 알 수 없습니다.

이 계수들을 예측 규칙이 아닌 조정 요소로 사용하는 순간 문제가 발생합니다. 이 모델에서 상태 조정치를 도출하는 감정평가사는 포인트당 \$34,816 을 적용하지만, 시장 가격은 \$20,000 입니다. — ‘양호’ 에서 ‘우수’ 로 등급이 조정된 모든 비교 대상에 대해 \$29,631 을 과다 적용하고 — 전망 조정치는 \$79,178 인 반면, 시장 기준치는 \$100,000 입니다. 그리드란 회귀 분석의 계수를 하나씩 읽어들이는 것이며, 바로 그 판독값을 편향이 무너뜨리는 것입니다.

또한, 편향이 형성된 패턴에서 벗어난 부동산의 경우 예측도 크게 빗나갑니다. 대상과 같이 잘 관리된 주택 (총 임대 가능 면적 1,650 제곱피트, 대지 면적 7,400 제곱피트) 이지만 전망이 없고, CQA 9.00 이라는 ‘우수’ 등급을 받았으며, 리모델링 이력이 전혀 없는 주택은 \$1,377,367 로 표시된 반면, 실제 거래가인 \$1,325,000 에 비해 \$52,367 (3.95%) 만큼 과대평가된 것으로 나타났습니다. 이는 해당 매매 사례에서 9.00 CQA 를 받은 주택은 대부분 리모델링이 이루어졌으며, 모델의 상태 평가 CQA 에는 리모델링 비용인 \$60,000 이 CQA 와 함께 반영되었기 때문입니다. 대상 주택 (전망, CQA 7.00)

과 그 외 모든 면에서 동일한 리모델링된 주택은 \$1,386,914 로 표시되어 있으며, \$1,445,000 과 비교했을 때 \$58,086(4.02%) 만큼 저평가되어 있습니다. 이것들은 비정형 부동산들입니다. 즉, 감정평가사가 가치를 산정하기 위해 가장 자주 특별히 위임받는 대상이며, 보고서가 가장 면밀히 검토되는 대상이기도 합니다.

### References.

- Craytor, W. B. (2026). Residual Constraint Approach (RCA): Framework & Protocol. *The Valuation Engineer*, 1(1). <https://doi.org/10.67330/kgebf46>
- 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.").

```