

# 내생성 (Endogeneity)

*Foundations • The Valuation Engineer*

William Bert Craytor • October 1, 2026

Version 1.2.0

DOI: [10.67330/z9feps66](https://doi.org/10.67330/z9feps66)

**형식적 정의.** 내생성이란, 회귀 분석에서 가격을 설명하는 데 사용되는 설명변수 (특징량) 가 구조적 오차 (가격에서 인과적 관계에 따라 예상되는 가격을 뺀 값, 즉 적합 모델의 잔차가 아님) 와 함께 변동하여, 그 추정 계수가 관심 있는 인과적 효과를 제대로 반영하지 못하는 상태를 말한다. 흔히 발생하는 원인으로는 변수가 누락된 경우, 가격이 해당 변수에 역으로 영향을 미치는 경우, 판매량이 선별된 집합인 경우, 그리고 변수가 오차와 함께 측정되는 경우가 있습니다.

**직관적 설명.** 특성과 가격이 서로 영향을 미치기 때문에 평가 지표가 부정확할 수 있습니다. 주택이 시장에 머무는 기간은 매도 희망 가격에 따라 달라집니다. 즉, 특성은 부분적으로 가격의 결과이거나, 상승하는 시장과 같이 가격을 높이는 다른 요인의 결과이기도 하며, 이러한 매매 사례를 바탕으로 측정된 평가 지표는 이 두 가지 요인을 혼합하여 반영하게 됩니다. 어떤 회귀 분석도 판매 데이터만으로는 이 두 요소를 분리해 낼 수 없습니다. 특정 특성의 전체 범위에 걸쳐 단일 단가를 적용하든, MARS(다변량 적응 회귀 스플라인, Friedman 1991) 처럼 특성의 구간별로 서로 다른 단가를 적용하든 마찬가지이며, 이를 분리하려면 외부의 변동 요인이나 매매 사례로는 검증할 수 없는 가정이 필요합니다. 이는 어렵지 산술적 계산이 아닌 시장 자체에 있기 때문입니다. 감정평가사가 할 수 있는 일은, 모델이 배제하는 요소들에 따라 특성이 움직일 수 있는 네 가지 방식 (아래에 제시됨) 을 파악하고, 데이터 세트에서 이를 식별한 뒤, 해당 데이터 세트의 효율이 마치 구매자가 그 특성에 대해 지불하는 금액을 측정한 것처럼 사용되기 전에 어떤 방식이 작용하고 있는지 판단하는 것이다.

그 메커니즘을 알고 싶은 독자를 위해. 통계학자들은 어떤 특징이 모델에서 제외된 요소와 함께 변동할 때 이를 ‘내생적’ 이라고 부르며, 이 항목에서는 해당 조건의 정확한 형태를 제시하고 그 네 가지 원인을 구분하여 설명한다. 통계학 문헌에서 제시하는 해결책, 주로 도구변수 (instrumental variables) 는 일반 회귀 분석에 적용되는 것으로, 평가 데이터에서는 거의 활용되지 않습니다. 해당 항목에서도 이를 명시하고 있습니다. 해결책과 달리, 이 진단은 어떤 추정 방법에도 적용됩니다.

이 항목은 통계적 맥락을 다루며, 회귀 계수가 감정사가 원하는 인과적 효과를 제대로 반영하지 못하는 상황을 다룬다는 점에서 인과적 추론도 포함합니다. 이는 계수나 쌍을 이룬 매매 사례에 따른 조정치를 인과 관계의 표현으로 해석하는 감정사들을 대상으로 합니다. 항목 008(생략된 변수 편향) 과 마찬가지로, 이 항목은 전통적인 감정 평가 방식과 마찬가지로 회귀 분석에 ‘상태’ 를 포함시키는 가상 시장인 ‘파시피카 (Pacifica)’ 시장을 사용합니다. 반면, 본 저널이 고안한 방법인 RCA 기반 감정 평가 (항목 007, 조건부 기대값) 는 이를 포함하지 않으며, 이 예시는 따라야 할 모델이 아니라 대조 사례로 제시된 것입니다.

008 번 대상 (‘생략된 변수 편향’) 에서 다룬 생략된 변수 편향은, 별도의 공식이 있는 독립적인 오류 유형으로 보아서는 안 되며, 보다 일반적인 상황인 내생성의 한 예로 이해해야 한다.

항목 007(조건부 기대값)의 분해를 떠올려 보자:  $P = E[P | \mathbf{Z}] + \varepsilon$ . 여기서  $E[P | \mathbf{Z}]$  는  $\mathbf{Z}$  의 특성을 가진 주택들의 평균 가격이며,  $\varepsilon$  는 주택 가격이 그 평균보다 얼마나 높거나 낮은지를 나타내므로, 정의에 따라  $E[\varepsilon | \mathbf{Z}] = 0$  가 성립한다. 이는  $\mathbf{Z}$  에 어떤 특성이 포함되어 있든 상관없이 조건부 기대값 함수 (해당 항목에서 CEF 라고 함) 에 대해 성립한다. 만약 감정평가사가 인과적 효과를 원한다면 — 즉, 주택의 다른 모든 조건이 실제로 일정하게 유지된 상태에서 전망이 있을 때 가격이 얼마나 더 높아지는지 — 한 가지 조건이 더 필요합니다. 즉, 설명변수는 가격에 영향을 미치는 다른 어떤 요인과도 함께 변동해서는 안 됩니다. 설명변수 중 하나가 다른 요인과 함께 변동한다면, 그 설명변수는 내생적입니다. 인과 관계를  $P = \beta_0 + \beta_1 z_1 + \dots + \beta_k z_k + u$  로 표현할 수 있으며, 여기서 각  $\beta_j$  는 특성  $z_j$  이 1 단위 변화했을 때 가격에 미치는 영향을 나타냅니다. 오차항은 구조적 오차이며,  $u$  로 표기됨

니다: 즉, 가격에서 인과 관계가 예측하는 값을 뺀 값입니다. 이는 조건부 평균 ( $\varepsilon$ ) 과의 거리인 와는 다른 개념이며, 후자는 정의상  $\mathbf{Z}$  에 따라 평균이 0 입니다.  $u$  는 설명변수와 함께 변동할 수 있으며, 내생성 조건은 이를 가리킵니다.

이 용어는 연립방정식 모델링에서 유래한 것으로, 시스템 외부에서 주어진 것이 아니라 시스템 내에서 결정되는 변수를 ‘내생적 (endogenous)’ 이라고 부릅니다. 현재 사용에서는 이 용어가 보다 기계적인 의미를 지닙니다:  $\text{Cov}(z_j, u) \neq 0$  (공분산 (Cov) 은 두 수량이 어떻게 함께 움직이는지를 측정합니다) — 설명변수와 구조적 오차가 함께 움직이기 때문에, 회귀 분석 (OLS — 일반 최소제곱 회귀 — 또는 MARS) 은 이들의 기여도를 분리할 수 없으며, 오차의 변동 일부를 설명변수에 귀속시킵니다.

이를 초래하는 네 가지 메커니즘이 있으며, 각각에 따라 다른 대응이 필요하므로 이를 명확히 구분하는 것이 중요합니다.

**생략된 변수.** 가격에 영향을 미치는 한 변수가 제외되었으나, 포함된 변수와 함께 변동하는 경우; 이는 ‘ $u$ ’ 의 일부가 되며, ‘ $\text{Cov}(z_j, u) \neq 0$ ’ 가 뒤따르게 된다. 이는 항목 008(생략된 변수 편향) 의 경우로, 리모델링 지표가 제외된 사례이다. 이를 해결하려면 해당 특징량을 측정하여 포함시키거나, 편향이 타당하게 가질 수 있는 최대값과 최소값을 산출해야 합니다.

**동시성.** 설명변수는 단순히 가격과 연관되어 있을 뿐만 아니라, 부분적으로는 가격에 의해 유발되거나 가격과 함께 결정되기도 합니다. 인과관계는 양방향으로 작용하며, 적합도 (OLS 또는 MARS) 는 방향성에 대해 아무런 정보를 제공하지 않기 때문에 두 가지가 혼합된 결과를 반환합니다. 이를 해결하기 위한 방법은 도구변수를 사용하는 것—즉, 해당 회귀변수를 변화시키지만 가격과는 다른 어떤 연관성도 없는 변수 (법적 의미의 공식 문서가 아닌 계량경제학적 의미에서)—이거나, 다른 분석 설계를 적용하는 것이다.

**표본 추출.** 이 표본은 결론의 대상이 되는 주택들 중에서 무작위로 추출된 것이 아니며, 특정 주택이 표본에 포함되는 여부는 그 주택의 가격과 관련이 있습니다. 이를 해결하기 위해서는 표본이 어떻게 선정되었는지를 모델링하거나, 결론의 대상이 되는 집단을 재정의해야 합니다.

**측정 오차.** 해당 특징이 부정확하게 기록된 경우 — 두 명의 검사관이 서로 다른 점수를 매기거나, 실측 대신 기록된 자료를 바탕으로 연면적을 산출한 경우 등. 이러한 기록 오차는 계수 ( $u$ ) 의 일부가 되어 기록된 값과 함께 변동하므로, 계수에 편향이 발생한다. 교과서적인 사례, 즉 오차가 실제 값이나 모델 내의 다른 요소와 전혀 무관한 고전적인 측정 오차의 경우, 잘못 측정된 특성의 계수는 0 으로 수렴합니다. CQA 의 경우, 이러한 경우와 다를 수 있습니다. 즉, 오차는 주택의 실제 상태나 CQA 가 파악하는 다른 요소들에 따라 달라질 수 있습니다. 이러한 오차의 경우, 편향은 일정한 방향을 가지지 않습니다. 해결책은 측정 정확도를 높이거나, 동일한 주택의 동일한 특성에 대해 두 번째 독립적인 측정을 확보하는 것입니다. 그러면 측정 오차를 위해 설계된 방법이 이를 활용하여 추정치를 보정할 수 있습니다.

이 네 가지 메커니즘에는 한 가지 공통점이 있습니다. 바로 같은 종류의 거래량을 더 늘려도 그로 인해 발생하는 오차는 사라지지 않는다는 점입니다. 이것이 내생성과 부정확성 (거래 건수가 제한적 이어서 발생하는 추정치의 분산) 을 구분하는 가장 명확한 실용적 차이점입니다. 표본 크기가 커지면 부정확성은 줄어들지만, 내생성은 줄어들지 않습니다.

**변수를 선택하기 전에, 그 효과의 이름을 정하십시오.** 특정 특징이 모델에 포함되어야 할지는 무엇을 추정하느냐에 따라 달라지며, 두 가지 유형의 특징은 쉽게 혼동되기 쉽습니다. 첫 번째는 회귀변수와 가격 모두에 영향을 미치는 원인입니다. 동네 내의 바람직한 위치 (조용한 막다른 골목, 전망이 좋은 모퉁이 등) 는 고가의 리모델링을 유도하며, 그 자체만으로도 높은 가격을 형성합니다. 통계학자들은 이러한 특징을 교란변수 라고 부릅니다. 이를 고정시켜야만, 회귀변수의 계수가 해당 변수의 효과를 반영하지 않게 됩니다. 두 번째는 회귀변수가 가격에 영향을 미치는 경로입니다. 리모델링은 주택의 상태와 품질을 개선 함으로써 가격을 상승시킵니다. 이러한 특징은 매개변수이며, 이를 고정 함으로써 회귀 변수의 계수에서 해당 경로를 통해 전달되는 효과의 정확한 부분만을 제거하게 됩니다. 따라서 항목 008(생략된 변수 편향) 의 리모델링은 두 가지 서로 다른 질문을 제기합니다. 전통적인 감정 평가에서, 만약 질문이 상태에 대한 시장의 평가율이라면, 상태는 설명변수이며 (RCA 기반 감정 평가는 결코 그러한 질문을 하지 않는데, 항목 007 ‘조건부 기대값’ 에서 그 이유를 설명함), 리모

텔링은 단지 일부 주택이 그 상태에 이르게 된 과정일 뿐입니다; 이러한 요소들이 완전히 기록된 후에는 리모델링이 만들어진 상태와 품질 이상의 가치를 창출하지 못하며, 상태와 품질이 동일한 리모델링된 주택과 잘 관리된 주택은 동일한 가치를 지닙니다. 만약 질문이 리모델링이 특정 주택의 가격에 어떤 영향을 미치는지에 관한 것이라면, 상태와 품질은 매개변수가 되며, 이를 고정시킨 모델은 다른 질문에 답하게 됩니다. 즉, 리모델링이 만들어진 상태와 품질을 제외하고 리모델링 자체가 가지는 가치가 얼마인지에 대한 질문인데, 이는 미미할 수 있으며 어쨌든 리모델링의 총 효과와는 다른 양입니다. “변수를 포함하라”는 것은 교란 요인에 대해서는 올바른 규칙이지만 매개 변수에 대해서는 잘못된 규칙이다. 원하는 효과가 명시되기 전까지는 어느 쪽도 선택할 수 없다.

**평가 실무에서 만나는 상황.** 역인과관계는 위치에 따라 달라집니다. 동시성으로 인한 내생성은 학교 교육의 질을 통해 잘 설명되는데, 이는 가설적인 문제가 아니라 실제 존재하는 문제입니다. 주택 가격을 학교 평점에 대해 회귀분석하면, 계수는 구매자가 학교를 위해 얼마나 더 지불하는지를 보여주는 것처럼 보입니다. 하지만 인과관계는 반대 방향으로도 작용합니다. 즉, 주택 가격이 더 비싼 지역은 세수 기반이 더 크고, 자녀 교육에 더 적극적으로 참여하며 자원이 풍부한 학부모들이 많아, 교사를 더 잘 유치하고 유지할 수 있습니다. 학교의 질이 주택 가격을 올릴 수 있고, 높은 주택 가격은 학교의 질을 향상시킬 수 있습니다. 계수는 이 두 가지 효과 중 어느 것도 포착하지 못하며, 해당 시장에서 두 경로의 상대적 강도에 따라 달라지는 복합적인 관계를 반영할 뿐입니다. 범죄율, 소매 편의시설 밀도, 수목 덮개율에서도 동일한 양방향 관계가 나타날 수 있다. 각 사례에서 해당 요인은 지역 가격 수준의 원인이자 결과일 수 있다.

리모델링은 ‘처리’에 해당하지만, 무작위로 배정되는 것은 아닙니다. 인과 추론의 용어에서 ‘처리’란 그 효과가 조사 대상이 되는 모든 개입을 의미하며, 주방 리모델링도 그 중 하나입니다. 문제는 누가 그 처리를 받느냐는 점입니다. 많은 주택 소유자는 수익이 비용을 상회할 것으로 예상할 때 리모델링을 진행하는데, 이러한 기대는 데이터셋에 기록되지 않은 자신들이 집에 대해 알고 있는 정보—좋은 평면도, 동네 내에서의 유리한 입지, 투자를 뒷받침해 줄 부지 등—에 기반을 두고 있습니다. 바로 이러한 특성들이 주택 가격을 직접적으로 상승시키는 요인이며, 리모델링의 교란 변수이자 데이터셋이 측정하지 못하는 요소들입니다. 따라서 리모델링 지표가 포함된 모델은 해당 지표의 누락을 해결했을 뿐, 교란 문제는 해결하지 못한 것입니다. 이 지표의 계수는 가격에 미치는 리모델링의 효과와, 기록된 특징들로 설명되는 범위를 넘어 리모델링을 선택한 주택 소유자들의 주택에서 발생하는 추가 가격을 결합한 것입니다. 이 변수를 포함함으로써 누락은 해소되었지만, 회귀 변수가 외생적—즉, 구조적 오차와 무관한—상태가 된 것은 아닙니다.

**매각된 부동산 표본은 무작위적으로 선정된 것이 아닙니다.** 사람들은 상황이 적절하거나 가격이 매력적일 때 매도하기도 하고, 어쩔 수 없이 매도해야 하는 경우도 있으므로, 매도된 부동산은 매도되지 않은 부동산과 모델이 포착하지 못하는 측면에서 차이가 있을 수 있습니다. 이러한 차이가 존재할 때, 매도된 부동산 가격의 특이적 요소는 무작위적이지 않습니다. 즉, 전자의 경우 가격을 한 방향으로 밀어붙이는 반면, 부실매각 및 강제매각은 반대 방향으로 작용합니다. 매매 사례 중에서도 정상 거래와 부실 매물은 각각 무작위적이지 않은 하위 그룹을 형성합니다. 각 그리드의 모든 비교 대상은 이러한 선별된 매매 사례에서 추출됩니다. 어떤 평가 결론도 매도된 매물이 어떻게 선별되었는지에 대한 가정을 전제로만 미매매 매물로 확장될 수 있으며, 보고서는 이러한 가정을 명시해야 합니다.

**상태 평점은 가격에 의해 왜곡될 수 있습니다.** 평점은 주택의 상태만을 측정하기 위한 것이며, 감정평가사는 가치를 산정하기 전에 실사를 통해 대상 주택에 점수를 매깁니다. 비교 대상 주택은 또 다른 문제입니다. 비교 대상 주택의 경우, 매매 가격이 확인된 후에 매물 사진과 설명을 바탕으로 CQA가 매겨질 수 있습니다. 알려진 가격이 CQA에 영향을 미친다면—예를 들어, 높은 가격에 팔린 주택이 관리가 조금 더 잘된 것으로 해석되는 경우—회귀 분석 변수는 부분적으로 결과의 함수가 되어 버립니다. 이는 상태 자체의 특성이 아니라 측정 과정의 결함, 즉 결과에 의존하는 측정 오차로, 가격이 CQA에 영향을 미치는 현상입니다. 이는 방금 설명한 비교전적 유형에 속하지만 한 가지 차이점이 있습니다. 가격 편차의 일부가 CQA에 반영되는 이러한 종류의 영향에 대해서는 그 방향이 알려져 있습니다. 즉, CQA는 매매 가격이 시사하는 CQA 쪽으로 이동하고, 상태 계수는 상승하게 됩니다(CQA에 포함된 다른 오차들은 여전히 반대 방향으로 작용할 수 있습니다). 보호 조치는 절차적인 것

입니다. 즉, 비교 대상 물건이 얼마에 팔렸는지 확인하기 전에 그 상태에 대한 증거를 바탕으로 CQA를 매기고, 각 CQA의 근거를 작업 파일에 기록해야 합니다.

**타당성 확보에 갖는 의미.** 회귀변수가 내생적인지 여부는 일반적인 회귀 분석으로 확인할 수 있는 것이 아니라 논증을 통해 결정된다. 일반적인 적합도 검증 방법—잔차 플롯, 가격 변동의 설명 비율 ( $R^2$ ), 적합 분석에서 제외된 판매량에 대한 모델 검증 (교차 검증), 또는 계수가 0 일 수 있는지 여부에 대한 검정 (유의성 검정) — 이 중 어느 것도 설명변수가 외생적임을 입증할 수 없는데, 이는 구조적 오차가 관측 불가능하며, 따라서 설명변수들과의 연관성도 관측할 수 없기 때문이다. 형식적인 검정 방법도 존재한다 — 하우스만 검정은 일반적인 추정치를 다른 방법으로 얻은 추정치와 비교한다 — 그러나 이러한 방법에는 도구변수나 제 2 추정량이 필요하며, 이는 데이터 자체로는 검증할 수 없는 추가적인 가정을 전제로 한다. 언제나 제시할 수 있는 것은 메커니즘에 대한 논리적인 설명이다: 왜 이 설명변수가 오차와 연결될 수 있는지, 어떤 방향으로, 어느 정도까지 연결되는지. 합리적인 보고서는 이 문제를 언급하지 않음으로써 문제가 존재하지 않는 것처럼 암시하기보다는 이러한 논증을 제시해야 한다.

계수가 ‘어떻게 되기를 바라는지’가 아니라 ‘실제로 얼마인지’를 말해야 합니다. 내생적 회귀변수의 계수는 실제적이고 명확하게 정의된 수치입니다. 즉, 측정된 변수들을 고정시킨 상태에서 해당 특성의 단위당 추정 가격의 차이를 의미합니다. 이는 인과적 효과가 아닙니다. 리모델링 지표가 제외된 16 건의 Pacifica 판매 사례 (항목 008, 누락된 변수 편향) 에서, “측정된 특징들을 고정했을 때, 전망이 있는 이 주택들은 \$79,178 더 비싸게 팔렸다” 라고 말하는 것이 옳다. “전망이 가치에 \$79,178 을 기여한다” 고 말하는 것은 추정치가 뒷받침하지 않는 인과 관계를 주장하는 것입니다. 첫 번째 진술은 반박에 견딜 수 있지만, 두 번째 진술은 그렇지 못합니다.

**설계가 추정치를 능가한다.** 데이터의 구성 방식에 따라 기능과 가격 간의 상호 영향을 차단할 수 있는 경우— 조망이 차단되기 전후의 부동산 특성을 비교하거나, 학교 배정 구역은 바뀌지만 주택 채고는 변하지 않는 경계를 활용하거나, 소유주가 선택한 리모델링이 아닌 화재나 홍수 같은 외부 사건으로 인해 불가피하게 진행된 리모델링을 고려함으로써 — 각 설계 방식은 여전히 입증되어야 할 가정들 (동시에 다른 어떤 것도 변하지 않았다는 가정, 경계가 학교 배정만 구분한다는 가정 등) 에 기반을 두고 있지만, 이러한 가정이 입증된 설계 방식은 관측 데이터를 사용하는 그 어떤 추정기보다 더 방어 가능한 타당성을 확보한다. 이것이 바로 도구변수 (instrumental variables) 문헌이 실질적으로 말하는 바이며, 감정평가와의 관련성은 잘 선정된 비교 대상 집합 자체가 그 자체로 하나의 설계라는 점에 있다.

**평가 계산 예시.** Pacifica 데이터셋은 내생성을 설명해 줄 수는 있지만 이를 해결할 수는 없으며, 이 데이터셋이 왜 해결하지 못하는지가 이 글이 주는 진정한 교훈이다.

먼저, 질문부터 시작하겠습니다. 감정평가사에게 특정 주택의 주방 리모델링 공사비 \$45,000 이 매도 시 회수될 수 있는지 물어본다고 가정해 봅시다. 이는 한 채의 주택과, 동시에 가질 수 없는 두 가지 가격, 즉 리모델링 후 가격과 현 상태의 가격에 관한 질문입니다.  $P_i(1) - P_i(0)$ , 여기서  $P_i(1)$  는 주택  $i$  이 리모델링되었을 때의 가격이고,  $P_i(0)$  는 리모델링되지 않았을 때의 가격입니다. 이는 주택의 상태에 따른 시장 가격에 대한 질문이 아니며, 상태에 대한 변수를 고정시킨 채 이 질문에 답하는 것은 앞서 설명한 중개인의 실수에 해당합니다.

항목 008(생략된 변수 편향) 의 16 개 속성에는 계수가 \$60,000 인 리모델링 지표가 포함되어 있습니다. 이는 생성된 데이터가 리모델링에 부여하는 가격 차이로, 동일한 GLA, 대지, 전망 및 CQA 점수를 가진 주택보다 \$60,000 더 높은 가격을 의미합니다. 이 \$60,000 은 리모델링 행위 자체에 대한 가격이 아니라, 다른 특징들로는 기록되지 않은 리모델링이 가져온 가치를 나타냅니다. 해당 구성된 데이터셋에서 어떤 주택이 리모델링되었는지는 저자가 결정했습니다. 리모델링 지표는 구조상 외생 변수이며, 오차항에 남은 요소가 없기 때문에 오차항과도 아무런 관련이 없습니다. 따라서 해당 데이터셋에서 \$60,000 은 기록된 CQA 점수를 고정했을 때 리모델링이 가격에 더하는 가치를 나타냅니다. CQA 점수까지 상승시킨 리모델링이라면 더 큰 가치를 더하게 될 것입니다.

이제 동일한 16 개 부동산이 실제 시장에서 나온 것이라고 가정해 보자. 소유주들은 리모델링을 선택하며, 타당하게도 관측되지 않는 우수한 특성을 지닌 주택의 소유주들 (이 가격 요소를  $u$  라고 부

르자) 이 다른 면에서는 비슷한 주택의 소유주들보다 리모델링을 할 가능성이 더 높다고 가정한다. 이제 리모델링 지표가 포함된 회귀 분석은 매매 건수가 누적됨에 따라 \$60,000 가 아니라

$$\hat{\beta}_{\text{remodel}} \xrightarrow{p} \underbrace{60,000}_{\text{what the remodel brought that the other features do not record}} + \underbrace{\frac{\text{Cov}(\tilde{r}, u)}{\text{Var}(\tilde{r})}}_{\text{extra price of the houses selected into remodeling}},$$

로 수렴하게 되는데, 여기서  $\tilde{r}$  는 포함된 특징들로 예측할 수 없는 리모델링 지표의 부분이며,  $\text{Var}$ (분산) 는 그 변동 정도를 측정하고,  $p$  로 표시된 화살표는 “매매 건수가 누적됨에 따라 점점 더 가까워진다” 는 것을 의미합니다. 리모델링된 주택이 평균적으로 \$25,000 만큼 더 많은  $u$  를 가지고 있다면, 계수는 대략 \$85,000 로 수렴하며, 동일한 시장에서 아무리 많은 추가 데이터를 확보하더라도 이 초과분은 줄어들지 않습니다. 이전 008 번 항목 (누락 변수 편향) 에서 연구했던 리모델링 지표가 이제 측정되어 포함되었지만, 계수는 여전히 인과적 효과를 나타내지 않습니다. 여전히 누락된 부분은  $u$  이며, 이는 어떤 데이터셋에도 기록되어 있지 않습니다.

이 구분은 실제 상황에서 실질적인 의미를 지닙니다. 이 계수는 평균적으로 리모델링을 한 주택들을 나타내는데, 이는 집주인들이 그 투자에 대한 수익이 가치 있다고 판단한 주택들입니다. 이 대상은 아직 결정을 내리지 않은 특정 주택에 관한 것이며, 해당 계수는 대상 주택이 소유주들이 알고 있던 정보 측면에서 그 주택들과 유사할 때에만 적용될 수 있는데, 이는 바로 모델이 파악할 수 없는 부분이다. \$85,000 을 바탕으로 투자 회수 기간에 대한 질문에 답하는 감정사는 수익률을 과대평가하는 것이며; 리모델링으로 인해 발생한 모든 조건과 품질을 측정하고 고정시킨 모델을 바탕으로 답변하는 감정사는 수익률을 과소평가하게 되며, 이는 리모델링이 그 조건과 품질 외에 추가하는 가치만큼만 반영되는데, 이는 항목 008(생략된 변수 편향) 의 원칙에 따라 거의 없거나 전무한 수준이다. 두 경우 모두 리모델링과 관련된 조건부 가격 차이를 리모델링 자체의 효과로 해석한 것이다.

따라서 인과적 관점에서 해석하고자 하는 어떤 계수에 대해서도 물어야 할 진단적 질문은 “모델이 적절하게 명시되었는가?” 가 아니라 “어떤 효과를 기대하는가, 그리고 이 변수가 어떻게 그런 값을 가지게 되었는가?” 이다. 만약 그 답이 모델이 파악할 수 없는 사물을 볼 수 있는 누군가가 내린 선택을 포함한다면, 그 계수는 해당 특성만큼이나 그 선택을 묘사하는 것이 된다.

## References.

- Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1), 1–67. <https://doi.org/10.1214/aos/1176347963>
- Haavelmo, T. (1943). The Statistical Implications of a System of Simultaneous Equations. *Econometrica*, 11(1), 1–12.
- Hausman, J. A. (1978). Specification Tests in Econometrics. *Econometrica*, 46(6), 1251–1271.
- Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press.
- Wright, P. G. (1928). *The Tariff on Animal and Vegetable Oils*. Macmillan.

---

*Cross-references: omitted variable bias; conditional expectation; parameter identification.*

---

## Addendum: Prolog Formalization

The following machine-readable companion (009-endogeneity.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>.

%% 009-endogeneity.pl
%%
%% Prolog companion to Foundations Vol. 1, Issue 2, Entry 009
%% Title: Endogeneity
%% Author: Bert Craytor
%% Concept: endogeneity
%% Version: 1.2.0
%%
%% v1.1.0: the estimand is named before the variables are chosen
%% (confounder vs. mediator); a remodel is a treatment that is not randomly
%% assigned; its 60,000 is the price difference the constructed data
%% attach to a remodel; measurement error attenuates only in the classical case; a rating
%% contaminated by the known price is outcome-dependent measurement error.

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

uses_terms([ endogeneity, omitted_variable_bias,
             conditional_expectation, hedonic_regression,
             latent_variable, residual, comp_selection,
             paired_sales_analysis, adjustment,
             sales_comparison_approach ]).
uses_relations([depends_on/2, related/2, introduced_by/2, formalized_by/2]).

article_concept(endogeneity).
article_issue(volume(1), number(2), year(2026)).
article_id('foundations.2026.009').

local_term(instrument).
local_term(simultaneity).
local_term(selection).
local_term(causal_effect).
local_term(conditional_association).
local_term(treatment).
local_term(confounder).
local_term(mediator).
local_term(estimand).

defining_condition(endogeneity,
    "The regressor is correlated with the model's structural error term.").

mechanism(omitted_variables,
    "A price-affecting variable is excluded and correlates with an included regressor; it now sits
    ↪ inside the error.").
mechanism(simultaneity,
    "The regressor is partly caused by, or jointly determined with, the outcome. Least squares
    ↪ returns a blend of both directions.").
mechanism(selection,
    "Sample membership is related to the outcome, so the estimated relationship is conditional on
    ↪ being in the sample.").
mechanism(measurement_error,

```

```

    "The regressor is recorded with error; the recording error enters the structural error and
    ↪ biases the coefficient. Under classical error (independent of the true value and of everything
    ↪ else) the coefficient is attenuated toward zero; under non-classical error no general
    ↪ direction follows.")).

measurement_error_kind(classical,          attenuates_toward_zero).
measurement_error_kind(non_classical,      no_general_direction).
measurement_error_kind(outcome_dependent,  toward_the_outcome).
measurement_error_of(condition_rating, non_classical).
measurement_error_of(condition_rating_after_price, outcome_dependent).

remedy(omitted_variables, "Measure and include the variable, or bound the bias.")).
remedy(simultaneity,      "Find an instrument, or a design that breaks the loop.")).
remedy(selection,        "Model the selection, or redefine the population the conclusion is about
    ↪ .").
remedy(measurement_error, "Measure better, or obtain a second independent measurement of the same
    ↪ characteristic.")).

generalizes(endogeneity, omitted_variable_bias).
special_case_of(omitted_variable_bias, endogeneity).

shrinks_with_sample_size(imprecision).

distinguishable_from_imprecision(endogeneity) :-
    \+ shrinks_with_sample_size(endogeneity).

role(confounder, "A cause of both the regressor and the price.")).
role(mediator,   "The channel through which the regressor acts on the price.")).
include_rule(confounder, include, "Its effect otherwise rides on the regressor's coefficient.")).
include_rule(mediator,   exclude, "Holding it fixed leaves only the direct effect, a different
    ↪ quantity from the total effect.")).

treatment(remodel).
confounder_of(unobserved_quality_u, remodel, price).    % floor plan, micro-location, suitability
mediator_of(condition, remodel, price).
mediator_of(quality,   remodel, price).

estimand(condition_rate,
    "Traditional appraisal: the market's rate for condition, with condition as the regressor; the
    ↪ remodel is how some houses came to their condition and earns no value of its own. RCA-based
    ↪ appraisal never asks this.")).
estimand(remodel_total_effect,
    "What a remodel does to a particular house's price: condition and quality are mediators and
    ↪ must not be held fixed.")).
estimand(remodel_direct_effect,
    "The remodel's effect apart from the condition and quality it produced: what a model holding
    ↪ those fixed estimates; may be small; not the total effect.")).

variable_choice(remodel_total_effect, V, exclude) :- mediator_of(V, remodel, price).
variable_choice(remodel_total_effect, V, include)  :- confounder_of(V, remodel, price).
variable_choice(condition_rate, condition, regressor).

encounter(school_quality, simultaneity,
    "School quality can raise prices and prices can raise school quality via tax base and parental
    ↪ resources; the coefficient then recovers neither effect.")).
encounter(renovation_choice, omitted_variables,
    "A remodel is a treatment that is not randomly assigned: owners remodel when they expect the
    ↪ return to justify it, on qualities u that also raise price. u is a confounder the dataset does
    ↪ not measure, so including the remodel indicator cures the omission of the indicator, not the

```

```

    ↪ confounding." ).
encounter(comparable_set, selection,
    "Sold properties are selected: sellers who choose their moment push the price one way,
    ↪ distressed and forced sellers the other; this matters when the selection is related to what
    ↪ the model does not see." ).
encounter(condition_rating, measurement_error,
    "Comparables may be rated from photographs after their sale prices are known; if the known
    ↪ price nudges the rating, the regressor is partly a function of the outcome. For a nudge that
    ↪ carries part of the price deviation into the score, the condition coefficient is pushed upward
    ↪ ." ).
protection(condition_rating,
    "Rate the comparables from the evidence of their condition before looking at the price; record
    ↪ the basis of each rating." ).

bidirectional_with_price(school_quality).
bidirectional_with_price(crime_rate).
bidirectional_with_price(retail_amenity_density).
bidirectional_with_price(tree_canopy).

cures(measuring_the_variable, omitted_variables).

omitted_component(renovation_choice, unobserved_quality_u).
measurable(remodel_numeric).

remains_endogenous_after_inclusion(remodel_numeric) :-
    encounter(renovation_choice, omitted_variables, _),
    omitted_component(renovation_choice, U),
    \+ measurable(U).

coefficient_component(remodel_numeric, what_the_work_produced,
    "60,000: the price difference the constructed data attach to a remodel, holding GLA, lot, view
    ↪ and the condition score fixed; it stands for whatever the remodel brought that the other
    ↪ features do not record." ).
coefficient_component(remodel_numeric, value_of_selected_properties,
    "The unobserved quality of properties whose owners chose to remodel; positive under positive
    ↪ selection, negative under negative selection." ).

coefficient_limit(SelectionDifference, Limit) :-
    Limit is 60000 + SelectionDifference.

consequence(renovation_payback_question,
    "A coefficient from observational data is the conditional association among the properties that
    ↪ were remodeled; it need not equal the causal effect of remodeling the particular subject, P_i
    ↪ (1) - P_i(0)." ).
consequence(two_ways_wrong,
    "Reading the selected coefficient overstates the return under positive selection; reading a
    ↪ model that holds condition and quality fixed understates it, to the direct effect alone." ).

defensibility_consequence(not_from_the_regression_itself,
    "Residual plots, R-squared, cross-validation and significance tests cannot establish exogeneity
    ↪ ; formal tests (Hausman) need an instrument or a second estimator, i.e. assumptions the data
    ↪ cannot check." ).
defensibility_consequence(state_what_it_is,
    "A coefficient on an endogenous regressor is a well-defined conditional association; reporting
    ↪ it as a causal contribution asserts more than the estimate supports." ).
defensibility_consequence(design_beats_estimation,
    "Breaking the loop in how data are assembled does more than any estimator applied afterward; a
    ↪ well-chosen comparable set is itself a design." ).

```

```

design(before_after_view_obstruction).
design(boundary_discontinuity_school_assignment).
design(externally_forced_renovation).

diagnostic_question(
    "What effect is wanted, and how did this variable come to take the values it took? If a chooser
    ↪ could observe what the model cannot, the coefficient describes the choice as much as the
    ↪ characteristic.").

% 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)).

```