

Biais de variable omise (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)

Définition formelle. Le biais de variable omise est la distorsion systématique introduite dans un coefficient estimé lorsqu’une caractéristique influant sur le prix est exclue du modèle *et* que cette caractéristique omise évolue de concert avec l’une des caractéristiques incluses — elle évolue avec celle-ci après que les autres caractéristiques incluses ont été prises en compte, ce que mesure la régression auxiliaire (une deuxième régression, de la caractéristique omise sur celles incluses). L’effet de la caractéristique omise se répercute alors sur les coefficients des caractéristiques incluses (le montant en dollars par unité que le modèle attribue à chaque caractéristique).

Présentation intuitive. Un cours peut être erroné s’il intègre la valeur d’une autre caractéristique. Le Comp H du numéro 1 s’est vendu à \$46,455 de plus que ce que prévoyait la régression basée sur ses quatre caractéristiques enregistrées. La raison, révélée dans la rubrique « *Variable latente* » du numéro 1, était une récente rénovation de la cuisine, qui, selon cette entrée, ajoute environ \$60,000 à la valeur de la maison. L’état de la maison figurait dans la liste des caractéristiques, mais rien n’indiquait la rénovation, et la plus-value apportée par celle-ci était absente des données. Cette valeur n’a pas disparu. Elle a été répercutée par le biais des coefficients des caractéristiques incluses, car celles-ci se sont avérées aller de pair avec elle lors des ventes. Deux conditions doivent être remplies pour que cela se produise : la valeur manquante doit avoir une incidence sur le prix, et elle doit être associée à une caractéristique figurant dans la liste. Lorsqu’elle a une incidence mais n’est associée à rien, les taux sont corrects et seule la dispersion est plus large. Cela vaut que la caractéristique se voie attribuer un seul taux ou plusieurs : MARS (Multivariate Adaptive Regression Splines, Friedman 1991), la principale méthode de régression utilisée par le Journal, attribue à une caractéristique des taux différents selon les intervalles de ses valeurs ; la valeur empruntée est alors répartie sur ces intervalles au lieu d’être concentrée dans un seul chiffre, mais elle est bien présente, pour la même raison.

Pour le lecteur qui souhaite en connaître le mécanisme : . Dans une régression classique, un taux correspond à un coefficient — c’est-à-dire la valeur en dollars que le modèle attribue à une unité supplémentaire d’une caractéristique — et la valeur calculée a une valeur précise : le coefficient de la caractéristique exclue multiplié par l’intensité de sa corrélation avec la caractéristique incluse, ce second facteur étant le coefficient d’une régression auxiliaire de la caractéristique exclue sur celles qui sont incluses. Cet article explique cette formule, qui correspond au cas linéaire, puis montre la même fuite dans la grille, où les ventes appariées de l’expert immobilier la transmettent de la même manière que les coefficients. Il s’adresse aux experts immobiliers qui déterminent des ajustements à partir de données de marché.

Ce qu’est et ce que n’est pas l’exemple présenté dans ce numéro. Le marché « Pacifica » fictif utilisé dans les articles 007 (*Espérance conditionnelle*) à 010 (*Identification des paramètres*) traite l’état comme une caractéristique associée à un taux en dollars, à l’instar de l’expertise traditionnelle et du numéro 1, et l’exemple montre en quoi cette approche est erronée. La méthodologie d’évaluation traditionnelle n’est pas celle adoptée par la revue, à savoir l’évaluation basée sur l’approche RCA (Residual Constraint Approach) du numéro 1 (Craytor 2026), qui exclut

l'état et la qualité de la régression ; l'article 007 (*Espérance conditionnelle*) explique comment procéder. Le lecteur doit retenir la leçon de cet exemple, et non sa formulation.

L'état, le score d'état et une rénovation. Ces trois notions sont souvent confondues. *L'état* correspond à l'état de la maison : l'usure de ses finitions, l'âge de ses installations, son retard d'entretien. Le *score d'état* est une estimation de cet état, par rapport à celui des maisons environnantes dans la zone de marché définie. Il repose sur la connaissance qu'a l'expert de cette zone de marché et ne devrait pas s'écarter de manière significative de celle de ses confrères exerçant dans la même zone. Dans cette revue, il varie de 0,00 à 10,00 : il s'agit du pourcentage de maisons de la zone de marché qui sont en moins bon état, divisé par dix. Un score de 5,00 signifie que la moitié des maisons de la zone de marché sont en moins bon état ; 9,00 signifie que neuf sur dix le sont. Ce score est une estimation, et deux maisons ayant le même score peuvent présenter des différences. Une «*» de rénovation* n'est ni un état ni un score, mais un événement : des travaux réalisés, au cours d'une période récente, qui ont modifié la maison. Sa valeur correspond à la valeur de ce qu'elle a modifié — l'état, et généralement la qualité, les matériaux et les finitions — aux tarifs pratiqués par le marché pour ces éléments. L'effet d'une rénovation porte généralement davantage sur l'état que sur la qualité, en grande partie en raison des coûts de main-d'œuvre, bien que dans certaines rénovations, la valeur ajoutée par la qualité soit la plus élevée des deux. Une maison rénovée et une maison n'ayant jamais fait l'objet de travaux de rénovation, qui présentent toutes les caractéristiques actuelles ayant une incidence sur la valeur, devraient avoir la même valeur ; le marché rémunère l'état et la qualité résultant de la rénovation, et non le simple fait qu'une rénovation ait eu lieu. Une rénovation n'est donc pas une caractéristique ayant un prix qui lui est propre. Lorsqu'un ensemble de données comporte un indicateur de rénovation en plus d'un score d'état, cet indicateur représente tout ce que la rénovation a apporté et que les autres caractéristiques du modèle (surface habitable brute, terrain, vue et score d'état) ne prennent pas en compte : l'état au-delà de ce que le score indique, et la qualité, que l'ensemble de données ne recense pas. C'est ainsi qu'il faut interpréter l'affirmation du numéro 1 dans *Variable latente*, selon laquelle la rénovation de Comp H « ajoute environ \$60,000 à la valeur globale de son ensemble », et c'est sur ce point que repose l'exemple présenté dans cet article.

Pour qu'il y ait biais, deux conditions doivent être réunies : la caractéristique qui n'est pas prise en compte doit avoir une incidence sur le prix, et elle doit évoluer en fonction de l'une des caractéristiques prises en compte. Si l'une ou l'autre de ces conditions n'est pas remplie, il n'y a pas de biais.

Cette exigence en deux parties est importante car la formulation courante de la règle — selon laquelle « un modèle qui omet une variable importante donne des réponses erronées » — est fautive telle qu'elle est énoncée. Si une caractéristique influence le prix mais ne varie pas en fonction des caractéristiques incluses, elle accroît la dispersion des prix autour du modèle et, par conséquent, augmente les erreurs-types (la mesure de l'écart entre un coefficient estimé et la valeur qu'un très grand échantillon donnerait), mais elle ne modifie pas les coefficients eux-mêmes ; l'ordonnée à l'origine (la constante du modèle, exprimée en dollars) absorbe la contribution moyenne de la caractéristique omise. Dans ce cas, c'est la précision qui est perdue, et non l'exactitude. En revanche, une caractéristique qui évolue parallèlement à celles incluses mais qui n'a aucun effet sur le prix ne coûte absolument rien. Ce n'est que lorsque ces deux conditions sont réunies qu'un biais apparaît.

Ce mécanisme repose sur la substitution. Si une caractéristique influant sur le prix est absente de l' des régresseurs (c'est-à-dire les caractéristiques utilisées par la régression pour expliquer le prix), la régression OLS (moindres carrés ordinaires) ne laisse pas simplement sa contribution inexpliquée. Les variables incluses qui évoluent de concert avec celle qui manque servent en quelque sorte à la remplacer partiellement, et leurs coefficients intègrent la part de l'effet de la variable manquante

qu'elles sont capables de prédire. Le coefficient d'une variable incluse ne correspond plus uniquement à l'effet de cette variable ; il s'agit de cet effet, auquel s'ajoute une part de l'effet de la variable manquante qu'elle a « empruntée ».

Le biais est toujours relatif à une question. Les coefficients du modèle sans l'indicateur de remodelage constituent la bonne réponse à une question — à savoir comment le prix évolue en fonction du score d'état sur l'ensemble de ces ventes, en tenant compte de tous les éléments associés à ce score — et constituent une réponse erronée lorsqu'ils sont interprétés comme représentant la valeur en dollars par point du score d'état attribué par le marché, toutes les autres variables étant maintenues constantes, ce qui correspond à la manière dont un ajustement les interprète. Le modèle n'est pas erroné en raison de l'existence d'une autre caractéristique ; ses coefficients sont erronés au regard de l'usage qui en est fait.

La part empruntée a une valeur exacte, donnée par une formule. Pour une caractéristique exclue w dont le coefficient réel est β_w , le biais du coefficient d'une caractéristique incluse z_j est

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

où $\hat{\beta}_j$ — le chapeau indique une estimation — est le coefficient estimé par la régression pour z_j , et δ_j est le coefficient de z_j dans la régression auxiliaire de w sur toutes les caractéristiques incluses. Il s'agit de la formule de Theil,¹. Calculée à partir des régressions d'ajustement, avec le coefficient estimé de la caractéristique exclue, elle est exactement vérifiée dans chaque échantillon ; avec le coefficient réel, elle est vérifiée en moyenne.

La forme produit correspond au cas linéaire. Dans un modèle MARS, l'effet de la caractéristique écartée est absorbé par les segments qui lui sont associés ; ainsi, le biais est réparti entre eux plutôt que concentré dans un seul coefficient, mais il est bien présent, et pour la même raison.

Cette formule est plus utile qu'il n'y paraît, car il est souvent possible d'attribuer un signe à ces deux facteurs grâce à la connaissance du marché, même si aucun des deux ne peut être mesuré. Lorsque la caractéristique non prise en compte fait augmenter le prix ($\beta_w > 0$) et est associée à une plus grande présence d'une caractéristique incluse, les autres caractéristiques incluses étant maintenues constantes ($\delta_j > 0$), le coefficient de cette caractéristique est trop élevé. Un expert immobilier qui connaît simplement le sens de ces deux relations connaît le sens de l'erreur. Une précision toutefois : la « δ_j » est une relation où les autres caractéristiques incluses sont maintenues constantes ; or, lorsqu'il y a plus d'une caractéristique incluse, la corrélation simple (un chiffre allant de -1 à $+1$ mesurant la façon dont deux caractéristiques augmentent et diminuent ensemble, en ignorant toutes les autres) peut aller dans le sens inverse. Dans l'exemple ci-dessous, la surface habitable est liée à l'indicateur de rénovation ($+0.09$ de corrélation), mais son « δ » est négatif et son coefficient est donc trop faible.

Où l'évaluateur la rencontre. La grille de comparaison des ventes peut être formulée comme un modèle additif présentant la même structure qu'une régression — une poignée de caractéristiques choisies manuellement, avec un ajustement en dollars pour chacune d'entre elles — et lorsque ses ajustements sont extraits des données du marché, elle est sujette à un biais de variable omise, tout comme n'importe quelle régression. Les lignes du tableau correspondent aux caractéristiques prises en compte. Toute caractéristique d'une maison qui influence son prix mais qui ne figure pas dans une ligne est laissée de côté.

Parmi les éléments souvent omis, on peut citer les aspects de l'état du bien que l'indice d'état ne prend pas en compte, la qualité lorsque aucun indice de qualité n'est enregistré, la fonctionnalité

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

du plan d'étage, la lumière naturelle, l'exposition au bruit, l'emplacement précis au sein du pâté de maisons, la motivation du vendeur et les efforts de promotion consacrés à l'annonce. Chacun de ces éléments peut influencer le prix, et beaucoup d'entre eux s'ajoutent à des critères déjà pris en compte par la grille d'évaluation.

La rénovation constitue un cas édifiant, car elle est consignée et continue d'avoir des répercussions. L'URAR (Uniform Residential Appraisal Report) indique si la cuisine et les salles de bains ont été modernisées ou rénovées, et au cours de quelle période, tandis que les définitions de l'état figurant dans l'UAD (Uniform Appraisal Dataset) dépendent elles-mêmes de l'ampleur des travaux de modernisation. Le tableau comporte toutefois des lignes pour l'état et pour la qualité, mais aucune pour la rénovation — et à juste titre, puisque la question porte sur la valeur marchande liée à l'état, et que les lignes « état » et « qualité » reflètent déjà l'apport de la rénovation ; une troisième ligne reviendrait donc à la comptabiliser deux fois. Lorsque la question porte plutôt sur l'impact d'une rénovation envisagée sur le prix de cette maison, le fait de maintenir l'état et la qualité constants supprime le canal même par lequel la rénovation agit ; l'entrée 009 (*Endogénéité*) aborde cette question. La fuite provient de ce que les scores ne prennent pas en compte. Sur un marché où les meilleures notes d'état appartiennent principalement à des rénovations récentes, et où ces rénovations apportent une valeur que les notes ne prennent pas en compte, un ajustement lié à l'état extrait de ces ventes dans le cadre d'une expertise traditionnelle — par comparaison avec des ventes similaires ou par une régression incluant l'état — intègre cette valeur. Selon la formule de Theil, cet ajustement est alors trop élevé : il inclut non seulement l'effet propre à l'état de la maison, mais aussi une partie de ce que les scores n'ont pas pris en compte. Un expert immobilier qui reprend cet ajustement et l'applique à une maison ayant obtenu le meilleur score grâce à son entretien procédera à un ajustement trop important. Dans le cadre de l'approche par contrainte résiduelle du Journal, aucun taux lié à l'état n'est extrait. L'expert répartit le résidu de chaque bien comparable entre l'état, la qualité et les autres caractéristiques non mesurées par les données, selon son appréciation, sous une seule contrainte : la somme des contributions des composantes de valeur au résidu doit être égale au résidu du bien comparable. La valeur d'une rénovation est alors répartie selon l'appréciation de l'expert, et non comptabilisée sur une seule ligne.

La « fuite » que nous venons de décrire correspond à une formulation plus précise de l'objection soulevée par les experts en évaluation concernant les ajustements fondés sur le marché, à savoir que les ajustements dérivés d'un ensemble de données ne sont pas facilement transposables. La raison en est qu'un ajustement extrait comprend à la fois l'effet nommé et les effets non nommés qui l'accompagnent, et que sa composition dépend des associations qui existaient dans l'ensemble de données dont il est issu.

Un autre exemple est celui où les « ventes appariées » dans l'évaluation traditionnelle — deux ventes identiques en tout point sauf sur la caractéristique en question, la différence de prix étant interprétée comme la valeur de cette caractéristique — aboutissent à un chiffre invraisemblablement élevé. La cause n'est peut-être pas une erreur de mesure, mais une « fuite » : les deux ventes diffèrent *non seulement* sur la caractéristique mentionnée, *mais aussi* sur un élément qui y est lié, et la différence de prix est due à ces deux facteurs.

Pourquoi elle compte pour la solidité. Un rapport qui n'aborde pas la question du biais lié aux variables omises expose ses ajustements hédoniques et par grille à cette critique.

Le biais ne peut pas être détecté à partir des statistiques d'ajustement. Un modèle peut présenter une forte « R^2 » (la part de la variation des prix que le modèle reproduit), de faibles résidus (les écarts entre les prix de vente et les prix du modèle) et des coefficients qui satisfont au test statistique habituel (un test de « t », qui consiste à vérifier si un coefficient pourrait raisonnablement être égal à zéro), tout en présentant un biais important, car le biais provient des corrélations entre

les caractéristiques, et non de l'ampleur des résidus. Utiliser l' R^2 comme preuve de la fiabilité du modèle ne répond pas à la question. Dans l'exemple ci-dessous, l' R^2 est de 0,97 et le coefficient conditionnel présente une erreur de 74 %.

Avoir plus de données ne résoudra pas le problème. Le biais ne diminue pas à mesure que l'échantillon s'élargit. Mille ventes issues du même marché, pour lesquelles la même caractéristique a été omise, donnent une estimation centrée sur le même chiffre erroné, mais estimée avec plus de précision : l'erreur-type se resserre autour de l'estimation biaisée, ce qui est pire qu'un intervalle large autour de celle-ci, car le rapport paraîtra plus fiable. Il en va de même pour MARS : un plus grand nombre de ventes rend la surface d'ajustement plus stable, mais toujours autour de la même réponse biaisée ; l'intervalle `varmod` de la Terre mesure la dispersion des prix, et non le biais. Cela corrige l'article n° 1 sur *les variables latentes*, qui affirmait que le biais diminuerait avec un plus grand nombre de comparables et que les autres coefficients étaient biaisés à la hausse. Le biais ne diminue pas avec un plus grand nombre de ventes, et dans les huit ventes du numéro 1 lui-même, comme dans les seize ci-dessous, les coefficients « vue », « surface brute locative » et « terrain » présentent un biais à la baisse ; seul le coefficient « état » présente un biais à la hausse.

Il est souvent possible de justifier le sens d'un biais même lorsque son ampleur ne peut être déterminée. Dans l'évaluation traditionnelle, où un ajustement lié à l'état du bien est déduit des ventes — par comparaison de ventes appariées ou par une régression incluant l'état du bien —, l'approche pratique consiste non pas à affirmer que l'ajustement est sans biais, mais à identifier les caractéristiques plausibles qui ont été omises, à utiliser la connaissance du marché pour attribuer à chacune un signe de β_w et de δ_j , et à indiquer le sens du biais qui en résulte. « L'ajustement lié à l'état du bien est probablement surestimé. La plupart des ventes ayant obtenu les meilleures notes concernent des rénovations récentes, et ces rénovations ont apporté des améliorations que les données ne prennent pas en compte, telles que de nouvelles cuisines et de nouvelles finitions. Ces améliorations font augmenter le prix et concernent principalement les ventes ayant obtenu les meilleures notes ; une partie de leur valeur a donc été comptabilisée dans l'état du bien » : il s'agit là d'une affirmation raisonnable. Les deux exigences sont satisfaites, et toutes deux vont dans le même sens : les améliorations non prises en compte font augmenter le prix (le β_w est positif) et vont de pair avec des scores élevés (le δ_j est positif) ; le biais est donc à la hausse. Une telle affirmation transforme un problème non mesuré en un problème reconnu et auquel on attribue une orientation, ce qui est exactement ce dont le lecteur a besoin pour évaluer la conclusion. Dans le cadre de l'approche par contrainte résiduelle, aucun taux d'état n'est extrait, et la question ne se pose pas pour l'état.

Exemple d'évaluation chiffré. L'extension de 16 biens immobiliers de l'ensemble de données Pacifica recense, pour chaque bien, la surface habitable brute (GLA), la superficie du terrain, la vue, le score d'état, un indicateur de rénovation observé et le prix de vente. Comme chaque prix a été calculé à partir d'une formule connue ne comportant aucune composante aléatoire (les données sont exemptes de bruit), la régression utilisant les caractéristiques appropriées reproduit exactement cette formule, et tout écart par rapport à celle-ci correspond à un biais ne comportant aucune composante de bruit.

Comp	GLA (sf)	Terrain (pieds carrés)	Afficher	Cond	Rénové	Prix (\$)
A	1,650	9 000	Oui	7,00	Non	1,425,000
B	1,650	7,200	Non	7,00	Non	1,280,000
C	1,850	7,500	Non	7,00	Non	1,407,500
D	1,450	7 000	Non	7,00	Non	1,155,000
E	1,700	8 000	Oui	9 h 00	Non	1,470,000
F	1,550	7,100	Non	5,00	Non	1,177,500
G	1,700	10,500	Non	7,00	Non	1,392,500
H	1,600	7,300	Non	7,00	Oui	1,312,500
I	1,750	8,200	Non	9 h 00	Oui	1,465,000
J	1,600	7,600	Non	7,00	Oui	1,320,000
K	1,900	9,500	Oui	9 h 00	Oui	1,687,500
L	1,500	6,800	Non	5,00	Non	1,140,000
M	1,800	8,800	Non	7,00	Non	1,410,000
N	1,650	7,400	Oui	7,00	Non	1,385,000
O	1,550	7,900	Non	9 h 00	Oui	1,337,500
P	1,720	9,200	Non	5,00	Non	1,332,000

Les comparables A à H correspondent au numéro 1, dont les caractéristiques et les prix publiés restent inchangés ; le numéro 1 décrivait l'état en mots et l'attribuait un code de 1 à 3 pour sa régression ; selon le barème de ce journal, les mots correspondent à « moyen » (5,00), « bon » (7,00) et « excellent » (9,00), de sorte qu'un cran du numéro 1 équivaut à deux points. La rénovation de Comp H — mentionnée dans le numéro 1 comme source de son résidu, mais non enregistrée en tant que variable — est désormais prise en compte. Le marché paie \$20,000 par point d'état. Dans les données construites, une maison rénovée se vend \$60,000 de plus qu'une maison présentant la même surface brute locative (GLA), le même terrain, la même vue et le même score d'état. Ces \$60,000 ne correspondent pas au prix de la rénovation en tant que telle ; ils représentent tout ce que la rénovation a apporté et que les autres caractéristiques ne prennent pas en compte. Les maisons rénovées obtiennent également des notes moyennes plus élevées (8,20 contre 6,64) ; l'indicateur de rénovation évolue donc en fonction de la note d'état incluse, ce qui correspond au deuxième critère requis par le principe de biais : $\text{corr}(\text{remodel}, \text{cond}) = 0.55$.

Deux régressions portant sur les seize mêmes biens immobiliers :

	Remodel omis	Rénovation comprise	Partialité
Interception	-83,923.57	-30,000	-53,923.57
GLA	586,64	600	-13.36
Lot	24,32	25	-0.68
Afficher	79,178,46	100,000	-20,821.54
Condition (par point)	34,815,56	20,000	+14,815.56
Rénovation	—	60,000	
R^2	0,9733	1,0000	

L'\$60,000e de la ligne « rénovation » ne correspond pas au prix de la rénovation. Dans cet ensemble de données reconstitué, il s'agit de la différence de prix associée à l'indicateur « rénovation » pour une surface brute locative (GLA), un terrain, une vue et un score identiques ; c'est ce que « récupère » la fonction génératrice.

Le coefficient de condition est trop élevé de \$14,816 par point — il indique \$34,816 par point de condition, alors que le chiffre du marché est de \$20,000, soit une surestimation de 74 %. Le coefficient « view » est sous-estimé de \$20,822. Ces deux erreurs sont invisibles dans l'ajustement :

le modèle sans l'indicateur « remodel » présente des $R^2 = 0.9733$ et des résidus qui, à l'exception de trois ventes proches de \$35,000, semblent normaux. Rien dans l' R^2 e ni dans les tests sur les coefficients ne laisse présager de problème (si ces résidus avaient été du bruit, tous les coefficients auraient satisfait à ce test au niveau habituel de 5 %).

La formule de Theil rend compte exactement des chiffres. La régression de l'indicateur « remodel » sur les quatre caractéristiques incluses donne $\delta_{\text{cond}} = 0.24692602$ et $\delta_{\text{view}} = -0.34702564$. On obtient alors

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

qui correspond aux décalages observés au centime près. Le biais n'est pas approximativement expliqué par la formule ; il *est* la formule elle-même. Le bruit modifierait moins que ce à quoi on pourrait s'attendre : avec des données bruitées, la formule reste valable exactement entre les deux régressions d'ajustement (sans et avec l'indicateur « remodel »), en utilisant le coefficient estimé de l'indicateur $\hat{\beta}_w$ à la place de β_w ; par rapport au coefficient réel, elle reste valable en moyenne sur de nombreux échantillons.

Le coefficient de vue mérite qu'on s'y attarde, car son biais va dans le sens inverse et la raison en est instructive. Toutes les autres caractéristiques étant constantes, les biens offrant une vue dans cet échantillon sont *moins* susceptibles d'être rénovés ($\delta_{\text{view}} < 0$). Le coefficient de vue est donc tiré vers le bas plutôt que vers le haut : il est \$20,822 trop bas, tandis que le coefficient d'état est \$14,816 trop élevé. Ces deux erreurs sont des effets distincts d'une même omission, et non un transfert entre elles. Un expert qui se baserait uniquement sur l'état du bien n'aurait pas du tout anticipé d'erreur dans l'ajustement lié à la vue — pourtant, un ajustement lié à la vue sous-évalué de 21 % est la conséquence directe de cette même omission.

Concrètement, les dommages se concentrent là où les experts sont le moins susceptibles de les repérer. Examinons d'abord ce que donne ce modèle biaisé en tant que *prédicteur*. Prenons un bien à évaluer non rénové identique à Comp N — 1,650 pieds carrés de surface brute locative (GLA) sur un terrain de 7,400 pieds carrés avec vue et un score d'état de 7,00 — il indique \$1,386,914 contre un prix moyen réel de \$1,385,000 pour les maisons présentant exactement ces caractéristiques — d'après la formule à cinq critères de l'entrée 007 (*Espérance conditionnelle*), avec une valeur de rénovation de = 0 — soit une erreur de \$1,914, soit 0,14 pour cent. Sur l'ensemble des seize ventes, l'erreur typique est d'environ \$21,000 ; le modèle présente donc un fort biais mais fournit tout de même des prédictions acceptables, ce qui n'est pas une coïncidence : l'ajustement par les moindres carrés ordinaires (OLS) s'adapte à l'échantillon qui lui est fourni, et les coefficients biaisés sont précisément ceux qui se rapprochent le plus des prix de l'échantillon. La précision des prédictions ne constitue pas une preuve contre le biais de variable omise, et les résultats de la validation croisée — qui consiste à tester le modèle sur des ventes non utilisées pour l'ajustement — ne l'excluent pas non plus : de bons résultats de validation croisée ne prouvent pas que les coefficients sont exempts de biais, car un modèle mal spécifié peut bien prédire dès lors que les caractéristiques se combinent comme elles l'ont fait lors de l'ajustement. En excluant tour à tour chaque bien, les deux erreurs les plus importantes — environ \$56,000 pour K et \$48,000 pour G — concernent des biens qui suivent le schéma de cet échantillon, dans lequel les scores d'état élevés appartiennent principalement à des maisons rénovées ; E, qui rompt ce schéma, arrive en troisième position avec environ \$45,000, tandis que H et J, qui le rompent également, se classent quatrième et cinquième avec environ \$41,000. Les données présentées ici ne comportent aucun bruit ; ainsi, chacune de ces erreurs provient de l'indicateur de rénovation omis, et aucune n'est due au bruit. Dans les ventes réelles, où le bruit est également présent, l'amplitude d'une erreur ne permet pas de déterminer si elle provient d'une caractéristique omise ou d'une dispersion résiduelle ordinaire.

Le problème apparaît dès lors que les coefficients sont utilisés comme *ajustements* plutôt que comme règle de prévision. Un expert qui applique un ajustement lié à l'état du bien à partir de ce modèle applique \$34,816 par point, alors que le chiffre du marché est de \$20,000 — soit \$29,631 de trop sur chaque comparable ajusté entre les deux niveaux « bon » et « excellent » — et un ajustement lié à la vue de \$79,178 alors que la valeur du marché est de \$100,000. La grille est l'endroit où les coefficients d'une régression sont lus un par un, et c' est exactement cette lecture que le biais détruit.

Les prévisions échouent également, et de manière flagrante, lorsqu'il s'agit de biens immobiliers qui ne correspondent pas au modèle sur lequel repose ce biais. Une maison bien entretenue comme le bien à évaluer (1,650 pieds carrés de surface brute, terrain de 7,400 pieds carrés), mais sans vue, ayant obtenu une note excellente de 9,00 et n'ayant jamais fait l'objet de travaux de rénovation, est répertoriée à l'adresse \$1,377,367, alors que le prix réel s'élève à \$1,325,000, soit une surévaluation de \$52,367 (3,95 %), car, dans ces ventes, une note de 9,00 s'accompagnait généralement d'une rénovation, et le taux d'état du modèle inclut l'\$60,000 de la rénovation en plus de la note. Une maison rénovée, par ailleurs identique au bien à évaluer (vue, note de 7,00), est indiquée à \$1,386,914, contre \$1,445,000, soit une sous-évaluation de \$58,086 (4,02 %). Ce sont là des biens atypiques — ceux pour lesquels un expert est le plus souvent spécifiquement mandaté pour en déterminer la valeur, et ceux pour lesquels le rapport sera lu avec le plus d'attention.

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 Scyer. 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')).
% Scyer 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_unrelated,
    "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.").

```