

Espérance conditionnelle (Conditional Expectation)

Foundations • The Valuation Engineer

William Bert Craytor • October 1, 2026

Version 1.2.0

DOI: [10.67330/99hhbc69](https://doi.org/10.67330/99hhbc69)

Définition formelle. L'espérance conditionnelle d'une variable de sortie (ici, le prix de vente) compte tenu d'un ensemble de prédicteurs (les caractéristiques mesurées) correspond à la moyenne de la distribution de la variable de sortie pour chaque valeur fixe des prédicteurs. Considérée comme une fonction des prédicteurs, elle est appelée fonction d'espérance conditionnelle (CEF) ; c'est cette fonction que la régression vise à approximer. La régression par les moindres carrés ordinaires (OLS) fournit la meilleure approximation linéaire de la CEF, et elle reproduit exactement la CEF lorsque celle-ci est elle-même linéaire.

Présentation intuitive. Une réévaluation issue d'une régression part d'un taux : ce que le marché paie par pied carré, par salle de bain, par année d'ancienneté. Cet article explique d'où provient ce taux et ce qu'il représente. En bref, il s'agit de la pente de la courbe des prix moyens du marché lorsqu'une caractéristique varie tandis que les autres restent constantes ; une régression est un moyen de mesurer cette pente à partir des données de vente. La régression OLS mesure une seule pente pour l'ensemble de la fourchette de valeurs de la caractéristique ; la méthode MARS (Multivariate Adaptive Regression Splines, Friedman 1991) mesure une pente pour chaque fourchette dans laquelle le marché paie différemment, et fournit un taux par fourchette. Dans les deux cas, ce taux mesure la même chose, à savoir la façon dont le prix moyen évolue en fonction de cette caractéristique, et la suite de cet article porte précisément sur ce sujet : de quoi s'agit-il, pourquoi la grille de comparaison des ventes le mesure-t-elle également, et que signifie le fait que deux régressions portées sur les mêmes ventes donnent deux taux différents ?

Pour le lecteur qui souhaite en connaître le mécanisme : Le « prix moyen en fonction des caractéristiques » porte un nom, celui de « fonction d'espérance conditionnelle », et l'article l'utilise désormais. Il s'agit d'une surface définie sur l'espace des caractéristiques ; un taux correspond à sa pente dans une direction donnée ; la méthode OLS effectue l'ajustement du plan qui se rapproche le mieux de la surface et ne la reproduit exactement que lorsque la surface est un plan ; la méthode MARS effectue l'ajustement d'une surface composée de morceaux plats qui s'incurvent aux ruptures qu'elle détecte, et se rapproche donc davantage lorsque la surface réelle présente des courbes. Les huit modèles Pacifica biens comparables du numéro 1 constituent l'exemple illustratif.

Cet article porte davantage sur les statistiques que sur le droit : il traite de ce qu'une régression permet d'estimer, et s'adresse aux experts qui réalisent des régressions ou qui utilisent la grille de comparaison des ventes — laquelle permet également d'estimer la même chose.

La régression effectuée dans le numéro 1 pour les biens comparables de la gamme Pacifica a été utilisée pour déduire des prix implicites à partir de ses coefficients (les valeurs que la régression attribue aux caractéristiques) ; un prix implicite correspond au montant en dollars que le marché paie pour une unité d'une caractéristique. Cependant, ce que la régression a estimé n'a pas été clairement précisé, et il convient de le mentionner car la quasi-totalité des questions qui suivent dans ce numéro — concernant le biais, l'endogénéité et l'identification — portent sur la différence entre ce qui a été estimé et ce qui était souhaité.

La régression est généralement présentée comme l'ajustement d'une droite visant à minimiser la somme des carrés des résidus (un résidu correspond à la différence entre le prix de vente et le prix

donné par l'équation ajustée), puis à déterminer la pente. Bien que cette description soit correcte d'un point de vue arithmétique, elle est trompeuse quant à l'objectif visé. Le calcul n'est qu'un moyen ; l'objectif est une grandeur qui décrit l'ensemble du marché (une grandeur *de population*), et non pas uniquement les ventes enregistrées.

Comprendre la notation.

P	prix de vente.
$\mathbf{Z}$	la liste des caractéristiques mesurées des biens ; une telle liste est appelée « <i>vecteur</i> ».
$\mathbf{z}$	un ensemble précis de valeurs pour cette liste, par exemple 1 650 pieds carrés, un terrain de 7 200 pieds carrés, sans vue, état 7,00.
$E[\]$	la <i>valeur attendue</i> : la moyenne de toutes les valeurs possibles, chacune pondérée par sa probabilité.
$ $	« donné ».
$E[P \mid \mathbf{Z} = \mathbf{z}]$	la valeur attendue du prix parmi les biens immobiliers dont les caractéristiques correspondent à celles d' $\mathbf{z}$.
ε	l'écart entre le montant d'une vente et cette valeur attendue.
u	une erreur structurelle de vente : son prix moins ce qu'il devrait être selon le lien de causalité.
$\hat{p}$	la valeur fournie par l'équation d'ajustement.
$z_{\text{GLA}}, \dots$	valeur d'une caractéristique : surface brute (GLA) et superficie du terrain en pieds carrés ; vue et rénovation : 1 si oui, 0 si non ; note comprise entre 0,00 et 10,00.

Fixons le vecteur de caractéristiques $\mathbf{z}$ — c'est-à-dire une superficie de 1 650 pieds carrés, un terrain de 7 200 pieds carrés, aucune vue et un état « bon » (7,00 sur l'échelle d'état du Journal, expliquée dans l'exemple détaillé). Parmi les biens immobiliers disponibles sur le marché qui présentent exactement ce vecteur de caractéristiques, les prix varient. Les prix ne sont pas identiques car les biens présentent des différences qui ne sont pas prises en compte par le vecteur, parce que les acheteurs et les vendeurs sont différents, et parce que la négociation ne débouche pas toujours sur le même prix. Sur ce marché, il existe une distribution complète des prix correspondant à l' $\mathbf{z}$. La moyenne de cette distribution est l'espérance conditionnelle du prix compte tenu de l' $\mathbf{z}$, notée $E[P \mid \mathbf{Z} = \mathbf{z}]$; la définition générale est due à Kolmogorov (1933). La moyenne des ventes en cours constitue une estimation de cette valeur attendue, et non la valeur attendue elle-même.

Comme les caractéristiques des biens immobiliers couvrent l'ensemble de l'espace des caractéristiques, l'espérance conditionnelle dessine une surface : à chaque point se trouve un nombre, à savoir le prix moyen des biens présentant ces caractéristiques. Cette surface est appelée « *fonction d'espérance conditionnelle* » (ou CEF). Il s'agit d'une caractéristique du marché et non d'un élément propre à un ensemble de données particulier : elle existerait même si aucun bien immobilier n'avait jamais été évalué. La figure 1 la représente pour l'exemple présenté.

Deux faits font de la CEF la cible naturelle de l'estimation. Premièrement, parmi toutes les fonctions d' $\mathbf{z}$ s, la CEF est celle qui minimise la valeur attendue de l'erreur de prédiction au carré : lorsque l'on doit estimer un prix en se basant uniquement sur $\mathbf{z}$ et que l'on est évalué en fonction de l'erreur au carré, $E[P \mid \mathbf{Z} = \mathbf{z}]$ constitue la meilleure estimation possible. Deuxièmement, tout résultat peut s'exprimer sous la forme

$$P = E[P \mid \mathbf{Z}] + \varepsilon, \quad E[\varepsilon \mid \mathbf{Z}] = 0,$$

, la deuxième équation n'étant pas une hypothèse mais une identité, puisque l'écart par rapport à la moyenne conditionnelle a, par définition, une moyenne conditionnelle nulle (Angrist & Pischke 2009, chap. 3).

Cette entrée précède les entrées 008 (*Biais de variable omise*) à 010 (*Identification des paramètres*), car chacune d'entre elles vise à déterminer si ce qu'une régression permet de récupérer correspond

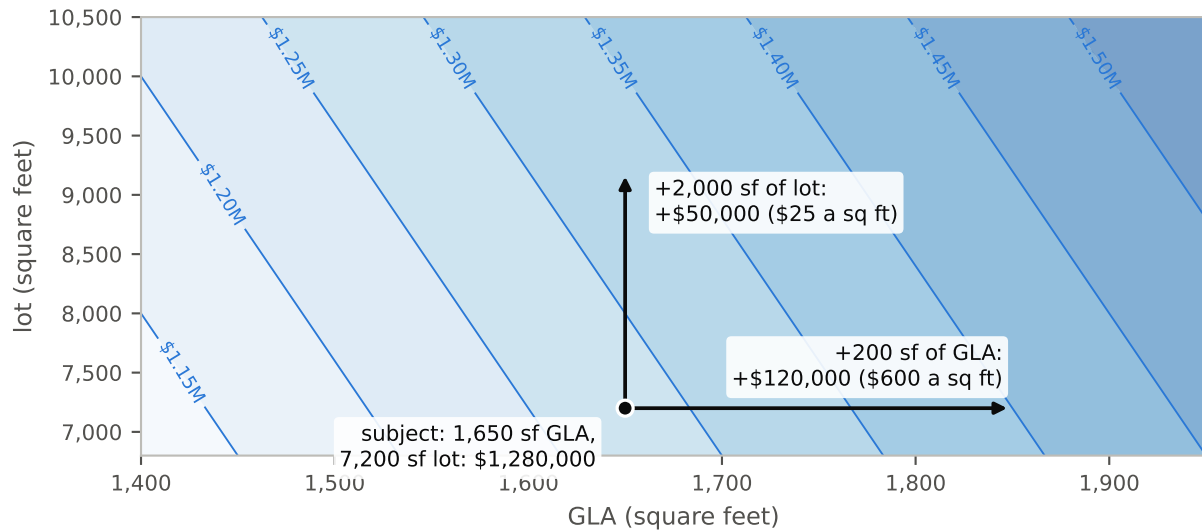


Figure 1: La fonction d'espérance conditionnelle du marché Pacifica modélisé (exemple concret), pour les maisons sans vue, avec un score de condition de 7,00 et sans travaux de rénovation, représentée à la manière d'une carte topographique : chaque ligne relie les combinaisons de surface brute locative (GLA) et de superficie de terrain dont le prix moyen est identique, et l'ombrage s'intensifie à mesure que le prix moyen augmente. Le point marqué correspond à la maison faisant l'objet de cet article, d'une surface brute habitable (GLA) de 1,650 pieds carrés sur un terrain de 7,200 pieds carrés, située à l'adresse \$1,280,000 . Les deux flèches représentent les deux pentes, qui correspondent aux taux : 200 pieds carrés supplémentaires de surface brute habitable (GLA) font augmenter le prix moyen de \$120,000 , soit \$600 par pied carré ; 2,000 pieds carrés supplémentaires de terrain le font augmenter de \$50,000 , soit \$25 par pied carré.

bien à la CEF d'intérêt. $L'E[\varepsilon \mid \mathbf{Z}] = 0$ -identité est toujours vérifiée, car ε est définie comme la distance par rapport à la moyenne conditionnelle. Ce qui peut ne pas être vérifié, c'est cette même condition — une moyenne conditionnelle nulle — pour une erreur différente : l'erreur structurelle, notée u dans l'entrée 009 (*Endogénéité*) — c'est-à-dire la différence entre le prix et la relation causale que l'évaluateur souhaite mesurer. Cette erreur peut varier en fonction des caractéristiques. C'est le cas, par exemple, lorsqu'une caractéristique influant sur le prix est omise ; lorsque le prix agit en retour sur une caractéristique — la durée de mise en vente (DOM) augmente généralement à mesure que le prix augmente ; ainsi, une régression utilisant la DOM pour expliquer le prix capte une partie de l'effet propre du prix sur la DOM, ce qui explique pourquoi peu d'analystes incluent la DOM dans une régression du prix de vente net ; ou lorsqu'une caractéristique est mesurée avec une erreur. *Le biais de variable omise* (entrée 008) et *l'endogénéité* (entrée 009) concernent ces cas de figure ; *l'identification des paramètres* (entrée 010) pose la question préalable de savoir si les ventes permettent réellement de déterminer ce taux.

Où l'évaluateur la rencontre. Un expert immobilier qui affirme qu'« une maison d'une superficie de 1 650 pieds carrés et sans vue, dans ce quartier, se vend environ à \$1.28 million » exprime une estimation conditionnelle. Le mot utilisé ici est « environ ». Cette affirmation ne signifie pas que toutes les maisons de ce type se vendent à \$1.28 million ; elle signifie plutôt que \$1.28 million correspond à la moyenne des prix auxquels se vendent les maisons de ce type. En utilisant le terme « environ », l'expert reconnaît que les prix individuels varient autour de cette valeur.

Une grille de comparaison des ventes est une méthode permettant d'estimer la valeur du CEF au vecteur caractéristique du sujet ; chaque bien comparable correspond à une observation tirée de la distribution des prix au vecteur caractéristique de ce bien ; chaque ajustement déplace cette observation le long de la surface du CEF dans la direction du vecteur du sujet ; l'indication ainsi harmonisée fournit alors une estimation de la hauteur de la CEF au bien à évaluer. D'un point de vue structurel, lorsque ses ajustements sont extraits des ventes, la grille peut être considérée comme un simple estimateur non paramétrique (qui ne suppose aucune formule pour la surface) d'une moyenne conditionnelle — un estimateur qui utilise des moyennes locales et des pentes spécifiées manuellement au lieu de la méthode des moindres carrés ordinaires (OLS) ; les ajustements tirés du coût ou d'un jugement constituent des opinions structurées, et non des estimations de ce type.

En structurant ainsi la grille, on met en lumière les hypothèses tacites de l'expert. Chaque fois que l'on utilise des biens comparables, on part du principe que la courbe d'évaluation conditionnelle (CEF) est suffisamment régulière pour que les points voisins soient pertinents pour la valeur au point à évaluer. Un ajustement linéaire suppose que la pente de la CEF pour cette caractéristique particulière est approximativement constante sur l'ensemble de la plage considérée. Attribuer un poids égal à trois indications suppose que chacune d'entre elles estime la même moyenne conditionnelle, qu'aucune ne présente d'écart systématique et qu'elles ont toutes une précision similaire. Aucune de ces hypothèses n'est inhabituelle ; elles concernent toutes la forme d'une surface, et font toutes l'objet de discussions dès lors que la surface en question a été identifiée.

Cette formulation met également fin à une confusion courante concernant les intervalles. Il existe deux types d'incertitude. L'un concerne la position du CEF : à quelle distance la surface d'ajustement peut se situer de la moyenne réelle. L'autre concerne l'écart des prix individuels par rapport à cette moyenne. Un intervalle de confiance ne couvre que le premier ; un intervalle de prédiction couvre les deux, et est donc toujours le plus large des deux. Avec MARS, la principale méthode de régression utilisée par le Journal, le module « earth » (Milborrow 2026a) fournit un intervalle de prédiction pour un bien à évaluer à partir d'un deuxième modèle, son modèle de variance (Milborrow 2026b), qui estime comment la dispersion des prix varie sur l'ensemble du marché. Les conditions requises sont qu'il y ait suffisamment de ventes et que les erreurs

soient approximativement normales et indépendantes les unes des autres ; le module indique la fréquence à laquelle les ventes se situent à l'intérieur de l'intervalle, ce qui constitue le contrôle. Dans l'évaluation fondée sur l'analyse RCA, la valeur de marché du bien à évaluer, selon l'approche par comparaison des ventes, correspond au prix de vente estimé par le modèle, majoré d'un résidu estimé par l'expert. Le modèle de variance fournit l'intervalle dans lequel, avec une probabilité de 95 %, se situerait une vente du bien à évaluer, à condition que ces conditions soient remplies. Lorsqu'on demande à un expert « dans quelle mesure êtes-vous sûr de votre estimation ? », on lui demande généralement quelle est l'amplitude de la dispersion des prix individuels (l'intervalle de prédiction), et il répond généralement en indiquant son degré de certitude quant à la moyenne (l'intervalle de confiance).

Nommer l'estimande — c'est-à-dire la caractéristique de la population faisant l'objet de l'estimation — est un acte de justification, car aucune conclusion ne peut être évaluée sans cela. L'affirmation « L'ajustement est de \$100,000 » ne peut pas encore être considérée comme vraie ou fausse. Ce n'est que lorsque nous affirmons : « La différence de prix de vente moyen conditionnel entre les biens immobiliers avec vue et ceux sans vue, à surface habitable brute (SHB), superficie du terrain et état constants, est de \$100,000 sur ce segment de marché » que nous formulons une affirmation susceptible d'être contestée en fonction des variables considérées comme constantes, de la population concernée et des logements en question.

Pourquoi elle compte pour la solidité. Il en découle trois conséquences en matière de défendabilité.

Le choix d'un ensemble de conditionnement (les caractéristiques maintenues constantes) relève d'une décision méthodologique : il définit l'estimande. La question de savoir si les données permettent d'estimer cette estimande est distincte et relève de *l'identification des paramètres*. Les expressions $E[P \mid \text{GLA, view}]$ et $E[P \mid \text{GLA, view, condition}]$ représentent des fonctions différentes, qui peuvent avoir des valeurs et des pentes différentes. Lorsqu'un expert inclut une variable dans la grille, il ne se contente pas d'améliorer l'estimation ; il modifie l'estimande. La démarche raisonnable consiste à préciser l'ensemble de conditionnement sur lequel repose l'avis de valeur.

L'expression « toutes choses égales par ailleurs » décrit le raisonnement suivi par l'expert dans son analyse et revêt une signification bien précise. Dans le cadre du CEF, elle désigne la comparaison des moyennes conditionnelles de deux vecteurs caractéristiques qui ne diffèrent que sur un seul critère. Il s'agit d'une affirmation relative à la population, qui peut être vérifiée à l'aide des données de marché lorsque celles-ci sont suffisantes. L'analyse qui la sous-tend figure généralement dans le dossier de travail — c'est-à-dire le registre de la mission que l'expert est tenu de conserver en vertu de la règle de conservation des documents de l'USPAP — plutôt que dans le rapport.

L'écart résiduel est une information, et non une source d'embarras. La dispersion des prix autour de la fonction d'espérance conditionnelle est une caractéristique du marché ; sur un marché où les prix s'écartent largement de cette fonction, toute estimation sous forme de chiffre unique — quelle que soit la méthode utilisée — s'accompagnera d'une grande incertitude. Indiquer une valeur sans faire référence à cette dispersion donne l'impression que la conclusion est plus précise que ne le justifie le marché.

Exemple d'évaluation chiffré. Les huit prix comparables de Pacifica présentés dans le numéro 1 constituent un échantillon « sans bruit » : ils ont été construits pour satisfaire exactement une seule équation linéaire. Lorsque ces huit prix sont soumis à une régression sur la surface brute locative (GLA), le lot, la vue, l'état et l'indicateur de rénovation, chaque prix est reproduit au dollar près — il ne reste aucun écart — et la régression restitue l'équation à partir de laquelle les prix ont été calculés, taux par taux :

$$E[P \mid \mathbf{Z}] = -30,000 + 600 \cdot z_{\text{GLA}} + 25 \cdot z_{\text{lot}} + 100,000 \cdot z_{\text{view}} + 20,000 \cdot z_{\text{cond}} + 60,000 \cdot z_{\text{remodel}}.$$

L'état mentionné ici correspond à l'indice d'état de ce Journal : il s'agit d'une estimation de l'état de la maison par rapport à celui des habitations environnantes situées dans sa zone de marché définie. Cet indice repose sur la connaissance qu'a l'expert de cette zone de marché et ne devrait pas s'écarter de manière significative de celle d'autres experts exerçant dans la même zone de marché. L'indice varie de 0,00 à 10,00 : il correspond au 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. Le numéro 1 décrivait l'état en mots et l'avait codé de 1 à 3 pour sa régression ; selon le score de cette revue, les descriptions correspondent à « moyen » (5,00), « bon » (7,00) et « excellent » (9,00) ; ainsi, un cran du numéro 1 équivaut à deux points, et le marché paie \$20,000 par point. Le dernier terme correspond à la différence de prix que les données construites attribuent à une rénovation : \$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. Il ne s'agit pas d'un prix lié à l'acte de rénovation en soi ; il représente tout ce que la rénovation a apporté et que les autres caractéristiques ne prennent pas en compte (entrée 008, *biais de variable omise*).

Il s'agit précisément du CEF pour le marché « Pacifica » construit. Il est important de préciser quel type d'objet vient d'être décrit et pourquoi il peut être décrit avec une telle précision. L'ensemble de données a été construit, et un ensemble de données construit peut en effet être exempt de bruit ; un ensemble réel ne l'est jamais. L'avantage de sa construction est qu'elle élimine complètement le bruit aléatoire des prix — chaque prix correspond exactement à la formule — ; ainsi, dans les rubriques 008 (*biais de variable omise*) à 010 (*identification des paramètres*), aucune différence entre une estimation et sa valeur cible ne peut provenir du bruit : chacune correspond à un biais ou, dans le cas de *l'identification des paramètres*, à un échec d'identification.

Cet exemple traite l'état comme une caractéristique associée à un taux en dollars, à l'instar de l'évaluation traditionnelle et de ce qui a été fait dans le numéro 1 ; les entrées 008 (*biais de variable omise*), 009 (*endogénéité*) et 010 (*identification des paramètres*) montrent les problèmes que cela pose. La méthodologie d'évaluation traditionnelle ne correspond pas à la méthode utilisée par la revue, à savoir l'évaluation basée sur l'approche des contraintes résiduelles (RCA). Dans l'approche des contraintes résiduelles, la régression ne prend en compte que les caractéristiques mesurables ; l'état et la qualité en sont délibérément exclus. Ce que la régression ne parvient pas à expliquer dans un prix de vente correspond au résidu. Sur les marchés californiens dont parle le rédacteur en chef, les caractéristiques mesurées expliquent généralement environ 80 % de la variation des prix de vente, laissant environ un cinquième au résidu. L'expert répartit le résidu entre l'état, la qualité et les autres caractéristiques que les données ne mesurent pas, en s'appuyant sur ses propres notes, sous une seule contrainte : la somme des contributions des composantes résiduelles doit être égale au résidu du comparable. La conclusion de valeur ne dépend pas de la manière dont l'expert la répartit — Preuve II de l'article RCA (Craytor 2026, p. 36) : modifier la répartition du résidu sous la contrainte que la somme des contributions des composantes de valeur résiduelles soit égale au résidu du bien comparable n'a aucun impact sur la conclusion finale de valeur. Le taux attribué à l'état n'est donc jamais estimé ni revendiqué ; ce que l'expert attribue à l'état correspond à la part de l'état qui dépasse ce que les caractéristiques mesurées, dont l'âge, prennent déjà en compte.

Si l'on compare la régression du numéro 1 à cette CEF, on peut estimer que la régression est valable — à condition toutefois de faire au préalable une distinction. D'une manière générale, une régression par les moindres carrés ordinaires (OLS) ne reconstitue pas la fonction d'espérance conditionnelle (CEF) qu'elle cible ; elle trouve plutôt la meilleure approximation linéaire de celle-ci, c'est-à-dire la projection linéaire, et ces deux fonctions ne coïncident que dans le cas où la CEF est linéaire. Une méthode non paramétrique telle que MARS n'a pas besoin que la CEF soit linéaire : elle approxime la CEF par morceaux, avec une charnière (un point où le taux change) là où la

surface se courbe et un terme d'interaction (un taux qui dépend d'une autre caractéristique) là où elle se tord, et suit ainsi une CEF non linéaire qu'une projection linéaire aplatiserait. Il s'agit toujours d'une approximation du même objet, et les questions relatives aux variables omises, à l'endogénéité et à l'identification abordées dans les articles suivants s'y appliquent exactement de la même manière. La CEF à cinq variables mentionnée ci-dessus est linéaire, ce qui explique pourquoi la régression incluant l'indicateur de rénovation permet de la restituer exactement. La régression du numéro 1, conditionnée sur quatre caractéristiques à l'exclusion de l'indicateur de rénovation, estime la meilleure approximation linéaire de $E[P \mid \text{GLA}, \text{lot}, \text{view}, \text{cond}]$ et renvoie, avec la condition reformulée sur le score du Journal (le numéro 1 indiquait une ordonnée à l'origine de 80,777 et 49,824 par pas de son codage de 1 à 3 ; il s'agit du même ajustement),

$$\hat{p}(\mathbf{z}) = 6,041 + 575 \cdot z_{\text{GLA}} + 22.70 \cdot z_{\text{lot}} + 86,179 \cdot z_{\text{view}} + 24,912 \cdot z_{\text{cond}}.$$

Chaque coefficient est différent de celui qui le précède. Une fois la régression décrite en termes d'ajustement de droite, il n'y a plus rien à ajouter : nous disposons de deux modèles et de deux ensembles de chiffres. Cependant, dans le cadre de l'approche CEF, deux affirmations précises peuvent être formulées. Premièrement, les deux régressions s'appliquent à des ensembles de variables différents et, par conséquent, fournissent des approximations de fonctions d'espérance conditionnelle différentes :

$$E[P \mid \text{GLA}, \text{lot}, \text{view}, \text{cond}] \quad \text{and} \quad E[P \mid \text{GLA}, \text{lot}, \text{view}, \text{cond}, \text{remodel}].$$

Sur un marché où certains biens immobiliers, qu'ils aient été rénovés ou non, partagent les mêmes quatre caractéristiques, les deux fonctions prennent des valeurs différentes pour ces points. Sur ce petit marché artificiel, il n'existe pas deux biens présentant les mêmes quatre caractéristiques ; ainsi, les deux fonctions coïncident pour chaque bien mais diffèrent dans leur forme : la seconde est une formule linéaire dans ses caractéristiques, contrairement à la première. Dans chaque cas, la régression ne fournit que la meilleure approximation linéaire de sa cible. Les deux cibles sont des objets valides. Deuxièmement, seule la fonction comportant cinq variables présente des coefficients pouvant être interprétés comme des prix implicites — le coefficient de rénovation correspondant à la différence de prix entre une maison rénovée et une maison non rénovée présentant la même surface brute locative (GLA), le même terrain, la même vue et le même score d'état — et l'entrée 008 (*biais de variable omise*) en explique la raison.

Le résidu de Comp H du numéro 1 illustre à quel point cette distinction est importante. Si l'on considère que les huit biens du numéro 1 représentent l'ensemble du marché, alors chaque vecteur caractéristique à quatre variables correspond exactement à l'un des huit biens, et ainsi la CEF à quatre variables passe exactement par les huit prix — Comp H étant l'un d'entre eux. Comp H présente un écart nul par rapport au CEF à quatre variables ; son résidu du numéro 1, égal à \$46,455 correspond à la différence par rapport à la régression linéaire à quatre variables avec ajustement — la projection linéaire de ce marché à huit biens — et c'est précisément dans l'écart entre la projection et la CEF que se cache l'effet de la refonte. Avec la spécification à cinq variables, la projection et le CEF deviennent identiques, et l'écart disparaît. Le résidu ne provenait pas de Comp H lui-même ; il provenait de la spécification.

References.

- Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
- 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>
- Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Julius Springer.
- Milborrow, S. (2026a). *earth: Multivariate Adaptive Regression Splines* (R package version 5.3.6; derived from mda:mars by T. Hastie and R. Tibshirani). <https://doi.org/10.32614/CRAN.package.earth>
- Milborrow, S. (2026b, August 7). *Variance Models in earth*. <https://www.milbo.org/doc/earth-varmod.pdf>

Cross-references: hedonic regression; implicit price; omitted variable bias; parameter identification.

Addendum: Prolog Formalization

The following machine-readable companion (007-conditional-expectation.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>.

%% 007-conditional-expectation.pl
%%
%% Prolog companion to Foundations Vol. 1, Issue 2, Entry 007
%% Title: Conditional Expectation
%% Author: Bert Craytor
%% Concept: conditional_expectation
%% Version: 1.2.0
%%
%% v1.2.0: entry text revised for readers who are not statisticians;
%% no Prolog clause changes.
%%
%% v1.1.0: condition on the Journal's 0.00-10.00 score (average 5.00, good
%% 7.00, excellent 9.00; 20,000 a point). The generating function lives in
%% kb/pacifica_comps_extended.pl, which is restated in those units with this
%% release; nothing here changes but the units it reads.
%%
%% This file is the Prolog galley for the entry: it encodes the entry's
%% main claims, its worked example, and the generating function of the
%% running dataset in a form other Prolog code in the journal can
%% consult, query, and verify. Read alongside entry.tex / entry.pdf.

% =====
% DICTIONARY IMPORT
% =====

% 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').
```

```

% --- From the dictionary (quoted for reference) -----
% term(conditional_expectation, "Conditional Expectation",
%     econometric_method,
%     "The conditional expectation of an outcome given a set of
%     predictors is the mean of the outcome's distribution at each
%     fixed value of the predictors. Viewed as a function of the
%     predictors, it is called the conditional expectation function
%     (CEF), and it is the function that regression aims to
%     approximate. Ordinary least squares (OLS) regression gives the
%     best linear approximation to the CEF, and it reproduces the CEF
%     exactly when the CEF itself is linear.", published,
%     'dictionary-2026.1').
%
% depends_on(conditional_expectation, hedonic_regression).
% -----

uses_terms([ conditional_expectation, hedonic_regression,
             implicit_price, hedonic_price_function,
             characteristics_space, residual,
             ordinary_least_squares,
             sales_comparison_approach, adjustment,
             comp_selection, reconciliation ]).
uses_relations([depends_on/2, related/2]).

% =====
% ENTRY SUBJECT
% =====

article_concept(conditional_expectation).
article_issue(volume(1), number(2), year(2026)).
article_id('foundations.2026.007').

% =====
% ENTRY-LOCAL VOCABULARY
% =====
%
% Field names for the worked-example records and for the two properties
% of the CEF the entry relies on. Too narrow to promote to the
% dictionary.

local_term(cef).                % conditional expectation function
local_term(estimand).
local_term(conditioning_set).
local_term(prediction_interval).
local_term(confidence_interval).
local_term(linear_projection).   % best linear approximation to the CEF

% =====
% THE TWO PROPERTIES OF THE CEF THE ENTRY USES
% =====

%! cef_property(?Name, ?Statement) is nondet.

cef_property(best_predictor,
    "Among all functions of the conditioning variables, the CEF minimizes expected squared
     $\hookrightarrow$  prediction error.").
cef_property(orthogonal_deviation,
    "P = E[P|Z] + eps with E[eps|Z] = 0 holds by construction, not by assumption.").

```

```

% The second property is the hinge for entries 008-010: each asks when
% the recovered object is the CEF of interest.
hinge_for(orthogonal_deviation, omitted_variable_bias).
hinge_for(orthogonal_deviation, endogeneity).
hinge_for(orthogonal_deviation, identification).

% =====
% CENTRAL CLAIM: CONDITIONING SET DETERMINES THE ESTIMAND
% =====
%
% Two regressions over the same data with different conditioning sets
% target DIFFERENT population objects, and each recovers only the best
% linear approximation to its target (the linear projection); OLS
% recovers the CEF itself exactly when the CEF is linear. Neither is
% wrong; they answer different questions.

conditioning_set(issue1_model,
    [gla_sqft, lot_sqft, view_numeric, cond_score]).
conditioning_set(issue2_model,
    [gla_sqft, lot_sqft, view_numeric, cond_score, remodel_numeric]).

%! different_estimand(?ModelA, ?ModelB) is nondet.
%
% True when two models condition on different variable sets and
% therefore approximate different conditional expectation functions.

different_estimand(A, B) :-
    conditioning_set(A, SA),
    conditioning_set(B, SB),
    A \== B,
    sort(SA, X), sort(SB, Y),
    X \== Y.

% What OLS recovers, relative to the CEF it conditions on.
ols_recovers(best_linear_approximation_to_cef).
ols_recovers_cef_exactly_when(cef_is_linear).

% =====
% WORKED EXAMPLE - THE GENERATING FUNCTION
% =====
%
% The running dataset is defined in ../../kb/pacifica_comps_extended.pl.
% Its generating function IS the conditional expectation function for
% the constructed market, exactly: the data carry no noise.

:- 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)).

%! cef_exact(+Property, -Price) is det.
%
% The CEF evaluated at a property's characteristic vector. Identical to
% true_price/2 in the dataset file; named here to make the entry's claim
% ("the generating function IS the CEF") explicit in the galley.

cef_exact(P, Price) :- true_price(P, Price).

```

```

%! noiseless_dataset is semidet.
%
% Verifies the entry's central factual claim about the data: every
% observed sale price equals the CEF at that property, so the residual
% is zero everywhere.

noiseless_dataset :-
    forall(property(P),
        ( cef_exact(P, Price), sale_price(P, Price) )).

% =====
% THE COMP H POINT
% =====
%
% Issue 1 reported a $46,455 residual for Comp H under the four-variable
% model. Taking issue 1's eight properties as the market, every
% four-variable characteristic vector belongs to exactly one property, so
% the four-variable CEF interpolates all eight prices and Comp H deviates
% from it by ZERO. The 46,455 is Comp H's deviation from the fitted
% four-variable LINEAR regression --- the linear projection --- and the
% projection-CEF gap is where the remodel effect hides. Under the
% five-variable specification, projection and CEF coincide and the
% deviation vanishes.

issue1_residual(comp_h, 46455).

% deviation_from(Property, Object, Dollars).
deviation_from(comp_h, four_variable_projection, 46455).
deviation_from(comp_h, four_variable_cef, 0).
deviation_from(comp_h, five_variable_cef, 0).

residual_under(issue2_model, comp_h, 0) :-
    cef_exact(comp_h, Price),
    sale_price(comp_h, Price).

%! residual_is_artifact_of_specification(?Property) is nondet.
%
% The issue-1 residual is a property of the specification (a linear
% projection under a smaller conditioning set), not of the property.

residual_is_artifact_of_specification(P) :-
    issue1_residual(P, R), R =\= 0,
    deviation_from(P, four_variable_cef, 0),
    residual_under(issue2_model, P, 0).

% =====
% APPRAISER ENCOUNTERS
% =====

grid_assumption(smoothness,
    "Using comparables at all assumes the CEF is smooth enough that nearby points inform the
    ↪ subject's point.").
grid_assumption(local_linearity,
    "Adjusting linearly assumes the CEF's slope in that characteristic is roughly constant over the
    ↪ range crossed.").
grid_assumption(unbiased_indications,
    "Averaging indications assumes each is an unbiased draw around the same conditional mean.").

```

```

% The grid is structurally a nonparametric estimator of a conditional
% mean, using local averaging and hand-specified slopes.
estimator_role(sales_comparison_approach, conditional_mean_estimator).

% =====
% DEFENSIBILITY
% =====

defensibility_consequence(conditioning_set,
    "Adding a variable to a grid changes the estimand, not merely the precision of an estimate.").
defensibility_consequence(ceteris_paribus,
    "'Holding other things equal' means comparing conditional means at vectors differing in one
    ↪ coordinate.").
defensibility_consequence(residual_spread,
    "Conditional dispersion is a market property; a point estimate reported without it overstates
    ↪ precision.").

% The confidence/prediction interval confusion the entry names.
interval_answers(confidence_interval, "uncertainty about where the CEF is").
interval_answers(prediction_interval, "spread of individual prices around the CEF").
commonly_confused(confidence_interval, prediction_interval).

```