

Régression hédonique (version française)

Foundations • The Valuation Engineer

Bert Craytor • 2 août 2026

Version 1.0.0

DOI: [10.67330/ams1t139](https://doi.org/10.67330/ams1t139)

Traduction française de l'entrée « Hedonic Regression » du département Foundations. Traduction préparée avec l'assistance d'une IA et revue par la rédaction ; la version anglaise fait foi.

Définition formelle. La régression hédonique est l'estimation de la fonction de prix hédoniques à partir des transactions observées sur un marché de biens hétérogènes, produisant des estimations empiriques des prix implicites des caractéristiques qui entrent dans le modèle.

Cadrage intuitif. Les entrées précédentes ont décrit la fonction de prix hédoniques et ses prix implicites comme des objets théoriques. La fonction existe en principe, mais en pratique nous n'observons qu'un ensemble fini de transactions : des paniers qui se sont échangés, et les prix auxquels ils se sont échangés. La régression hédonique est le pont empirique entre les deux.

Le dispositif est simple. Chaque transaction fournit une paire $(\mathbf{z}_i, p_i)$: le vecteur de caractéristiques du panier échangé et son prix de vente. La collection de ces paires est un échantillon de la fonction de prix hédoniques (non observée). La régression ajuste une forme fonctionnelle à cet échantillon, en estimant des paramètres qui indexent la fonction. Une fois les paramètres estimés, les dérivées partielles de la fonction ajustée produisent des estimations de prix implicites.

Trois choix de modélisation doivent être faits avant que la régression puisse tourner :

1. *Quelles caractéristiques entrent dans le modèle ?* Le choix détermine les dimensions de la fonction estimée. Omettre une caractéristique pertinente fait fuir son effet dans les coefficients des caractéristiques incluses corrélées (biais de variable omise).
2. *Quelle forme fonctionnelle ?* Linéaire, log-linéaire, log-log, polynomiale, à expansion de base, semi-paramétrique, apprentissage automatique. Le choix contraint la forme que la fonction estimée peut prendre et les motifs de prix implicites qu'elle peut récupérer.
3. *Quel estimateur ?* MCO, moindres carrés généralisés, M-estimateurs robustes, régression quantile, estimateurs d'économétrie spatiale, régression pénalisée. Le choix détermine ce que l'estimateur optimise et quelles hypothèses soutiennent son inférence.

Chaque choix est méthodologiquement substantiel et chacun emporte des conséquences de défendabilité.

Où les évaluateurs rencontrent ce concept. La régression hédonique est le fondement de l'évaluation de masse et le moteur de la plupart des modèles d'évaluation automatisés (AVM). Lorsqu'un assesseur de comté réévalue 100,000 parcelles chaque année, c'est une régression hédonique sur les ventes récentes qui fait le travail. Lorsque Zillow ou Redfin produit une estimation automatisée, c'est un modèle de type hédonique (souvent un ensemble d'arbres plutôt que des MCO, mais conceptuellement toujours un estimateur hédonique) qui la produit. Lorsque l'AVM de marché secondaire d'un prêteur accepte ou rejette un bien pour un refinancement, la même machinerie est en jeu.

L'évaluation d'un bien individuel s'appuie elle aussi de plus en plus sur la régression hédonique, même lorsque l'indication finale de valeur provient de la comparaison des ventes. Des prix implicites estimés à partir d'une régression sur un échantillon de quartier peuvent étayer les amplitudes d'ajustement de manière bien plus défendable que l'intuition de l'évaluateur ou des pourcentages

issus de règles empiriques. Les coefficients de régression deviennent des preuves, et non des opinions, pour les ajustements appliqués dans la grille.

Pourquoi cela importe pour la défendabilité. Les estimations fondées sur la régression présentent un profil de défendabilité distinct de celui de l'analyse par ventes appariées. Les avantages : une régression utilise toutes les ventes disponibles plutôt que deux à la fois ; elle peut isoler l'effet d'une caractéristique tout en contrôlant simultanément beaucoup d'autres ; et elle produit des erreurs types qui quantifient l'incertitude de chaque prix implicite. Les contreparties : une régression impose une forme fonctionnelle qui peut ne pas correspondre à la vraie fonction de prix ; elle exige plus de données qu'une analyse par ventes appariées ; et ses résultats peuvent être sensibles au choix de l'échantillon et de la spécification.

Plusieurs questions de défendabilité reviennent dans l'évaluation fondée sur la régression :

L'échantillon est-il approprié ? La régression estime la fonction hédonique telle qu'elle opère dans le segment de marché et la période de l'échantillon. Une régression sur des ventes de l'intérieur appliquée à un bien de front de mer est un échec de défendabilité, aussi propres que paraissent les statistiques.

La spécification est-elle appropriée ? Une spécification linéaire est une hypothèse forte. Si la vraie fonction présente des rendements décroissants de la GLA au-delà d'un certain seuil, l'ajustement linéaire estimera mal les prix implicites dans cette plage. Les diagnostics, les graphiques de résidus et les spécifications alternatives font partie du flux de travail défendable.

Que signifie l'erreur type ? Une erreur type capture la variabilité d'échantillonnage sous la spécification supposée. Elle ne capture ni la mauvaise spécification du modèle, ni le biais de variable omise, ni la sélection d'échantillon. Une petite erreur type sur une spécification fautive est plus dangereuse qu'une grande erreur type sur une spécification correcte.

Le résultat se recoupe-t-il avec les autres méthodes ? Un prix implicite issu d'une régression et défendable devrait être cohérent avec ce que suggère l'analyse par ventes appariées, avec ce qu'impliquent les coûts des composants de la méthode du coût, et avec ce que croient les évaluateurs qui connaissent le marché. Un coefficient de régression en désaccord avec les trois est une invitation à investiguer, pas à publier.

Exemple d'évaluation travaillé. Nous pouvons maintenant mettre en œuvre l'ensemble de la machinerie de la régression hédonique sur les huit comps de Pacifica. Le jeu de données est disponible à `shared/data/pacifica_comps.csv`.

Code R :

```
comps <- read.csv("shared/data/pacifica_comps.csv")
fit <- lm(price ~ gla + lot + view + cond, data = comps)
summary(fit)
```

Sortie :

```
Call:
lm(formula = price ~ gla + lot + view + cond, data = comps)

Residuals:
    Min       1Q   Median       3Q      Max
-17995  -8243  -5433  -1409   46455

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  80777.07   164509.35    0.491   0.6571
gla           574.93     109.76    5.238   0.0135 *
lot           22.70      10.38    2.187   0.1165
view         86179.03   31342.97    2.750   0.0708 .
```

```
cond          49824.33      27761.27      1.795      0.1706
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30481 on 3 degrees of freedom
Multiple R-squared:  0.9707, Adjusted R-squared:  0.9317
F-statistic: 24.86 on 4 and 3 DF, p-value: 0.01231
```

Interprétation. Le modèle ajusté estime la fonction de prix hédoniques comme :

$$\hat{p}(\mathbf{z}) = 80,777 + 575 z_{\text{GLA}} + 22.70 z_{\text{lot}} + 86,179 z_{\text{view}} + 49,824 z_{\text{cond}}.$$

Chaque coefficient est le prix implicite estimé de la caractéristique correspondante. La GLA contribue \$575 par pied carré (avec l'erreur type la plus étroite et la significativité la plus forte) ; la vue contribue environ \$86,000 à un bien de Pacifica ; une unité supplémentaire d'état (par exemple, passer de « bon » à « excellent ») vaut à peu près \$50,000.

Plusieurs mises en garde sont immédiatement visibles :

L'échantillon est minuscule. Huit observations et quatre prédicteurs laissent trois degrés de liberté résiduels. Les erreurs types sont en conséquence larges, et les valeurs p pour le terrain, la vue et l'état n'atteignent pas le seuil conventionnel de 0.05 malgré leur plausibilité économique. C'est un artefact de petit échantillon, non une preuve que ces caractéristiques ne comptent pas. Avec 80 comps au lieu de 8, les valeurs p rétréciraient spectaculairement.

L'ajustement global est fort. Un R^2 ajusté de 0.93 et une valeur p de modèle de 0.012 indiquent que les caractéristiques incluses expliquent conjointement l'essentiel de la variation des prix dans l'échantillon. La combinaison d'un ajustement conjoint fort et d'une significativité individuelle faible est un signal classique de multicollinéarité, qui est ici structurelle : la GLA, le terrain et l'état covarient dans l'échantillon.

Un résidu se détache. Le résidu maximal est de +\$46,455, correspondant au comp H. Les sept autres résidus se situent tous dans une bande de $\pm \$18,000$. Quelque chose du comp H n'est pas capturé par la fonction ajustée — le modèle sous-prédit systématiquement son prix d'un montant près de trois fois supérieur au résidu suivant.

Ce résidu aberrant n'est pas du bruit. C'est le signal qu'une caractéristique non observée du comp H contribue à son prix d'une manière que les variables incluses ne peuvent récupérer. L'entrée sur la variable latente reprend exactement cette question.

Validation par le raisonnement des ventes appariées. L'analyse par ventes appariées de l'entrée 004 prédisait une différence de \$127,039 entre les comps A et B à partir de la vue et du terrain seuls. Les prix implicites de la régression sont identiques (puisque'ils sont la source de ce calcul). La régression généralise la logique des ventes appariées aux huit comps simultanément, mais elle ne produit rien de fondamentalement nouveau à cette taille d'échantillon ; elle confirme l'analyse appariée tout en rapportant des erreurs types que le raisonnement par ventes appariées ne peut fournir.

Références.

- Court, A. T. (1939). Hedonic Price Indexes with Automotive Examples. In *The Dynamics of Automobile Demand*. General Motors Corporation.
- Griliches, Z. (1961). Hedonic Price Indexes for Automobiles : An Econometric Analysis of Quality Change. In *The Price Statistics of the Federal Government*. National Bureau of Economic Research.
- Rosen, S. (1974). Hedonic Prices and Implicit Markets : Product Differentiation in Pure Competition. *Journal of Political Economy*, 82(1), 34–55.
- Tripllett, J. E. (2006). *Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes*. OECD Publishing.

Renvois : bien hétérogène ; espace des caractéristiques ; fonction de prix hédoniques ; prix implicite ; variable latente.

Addendum : formalisation en Prolog

Le compagnon lisible par machine ci-dessous (005-hedonic-regression.pl) formalise cette entrée dans la base de connaissances Prolog de la revue. Il est également publié sous forme de galée consultable aux côtés de cet article.

```
%% 005-hedonic-regression.pl
%%
%% Prolog companion to Foundations Vol. 1, Issue 1, Article 005
%% Title: Hedonic Regression
%% Author: Bert Craytor
%% Concept: hedonic_regression
%% Version: 0.1.0-draft

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

:- use_module('.././../tools/dictionary/dictionary').
:- dictionary_release('dictionary-2026.1').

% --- From the dictionary (loaded via :- use_module above) -----
% term(hedonic_regression, "Hedonic Regression", econometric_method,
%      "Hedonic regression is the estimation of the hedonic price
%      function from observed transactions in a heterogeneous-goods
%      market, yielding empirical estimates of the implicit prices of
%      the characteristics that enter the model.",
%      published, 'dictionary-2026.1').
%
% term(mass_appraisal, ...).
% term(automated_valuation_model, ...). % synonym: avm
% term(omitted_variable_bias, ...).
% term(multicollinearity, ...).
% term(residual, ...).
% term(ordinary_least_squares, ...).
%
% introduced_by(hedonic_regression, court_1939).
% introduced_by(hedonic_regression, griliches_1961).
% formalized_by(hedonic_regression, rosen_1974).
% depends_on(hedonic_regression, hedonic_price_function).
% depends_on(hedonic_regression, implicit_price).
% -----

:- use_terms([ hedonic_regression, hedonic_price_function, implicit_price,
               characteristics_space, heterogeneous_good, latent_variable,
               mass_appraisal, automated_valuation_model,
               omitted_variable_bias, multicollinearity, residual,
               ordinary_least_squares,
               sales_comparison_approach, paired_sales_analysis, adjustment ]).
:- use_relations([related/2, depends_on/2,
                  introduced_by/2, formalized_by/2]).
```

```

% =====
% ARTICLE SUBJECT
% =====

article_concept(hedonic_regression).
article_issue(volume(1), number(1), year(2026)).
article_id('foundations.2026.005').

% =====
% LOCAL VOCABULARY
% =====

local_term(specification).
local_term(coefficient).
local_term(standard_error).
local_term(p_value).
local_term(r_squared).
local_term(adjusted_r_squared).
local_term(residual_standard_error).
local_term(degrees_of_freedom).
local_term(f_statistic).
local_term(significance_level).
local_term(diagnostic).

% =====
% THE EMPIRICAL BRIDGE
% =====
%
% Hedonic regression is the bridge from theoretical  $p(z)$  to a sample of
% observed  $(z_i, p_i)$  pairs.

setup_of_regression(
    each_transaction_supplies_pair("(z_i, p_i): characteristics vector and sale price"),
    sample_from("the unobserved hedonic price function"),
    fits("a functional form to that sample, estimating parameters")).

% =====
% THREE MODELING CHOICES (per the article)
% =====

modeling_choice(which_characteristics,
    "Determines the dimensions of the estimated function. Omission causes omitted_variable_bias.").
modeling_choice(functional_form,
    "Constrains what shape the estimated function can take. Linear, log-linear, polynomial, basis-
    ↪ expanded, semiparametric, ML.").
modeling_choice(estimator,
    "OLS, GLS, robust, quantile, spatial, penalized. Determines optimization criterion and
    ↪ inference assumptions.").

% =====
% PRACTICE CONTEXTS WHERE REGRESSION IS THE ENGINE
% =====

regression_engine_context(mass_appraisal,
    "Assessor revaluation of 100,000 parcels annually.").
regression_engine_context(automated_valuation_model,
    "Zillow/Redfin/lender AVMs; tree ensembles are conceptually hedonic estimators.").
regression_engine_context(single_property_appraisal,
    "Regression coefficients support adjustments more defensibly than appraiser intuition.").

```

```

% =====
% DEFENSIBILITY PROFILE: REGRESSION vs PAIRED-SALES
% =====

advantage_over_paired_sales(uses_all_data,
    "Regression uses all available sales rather than two at a time.").
advantage_over_paired_sales(simultaneous_control,
    "Isolates one characteristic while controlling for many others.").
advantage_over_paired_sales(quantified_uncertainty,
    "Produces standard errors quantifying uncertainty in each implicit price.").

tradeoff(functional_form_constraint,
    "Regression imposes a form that may not match the true price function.").
tradeoff(data_requirement,
    "Requires more observations than paired-sales analysis.").
tradeoff(sample_specification_sensitivity,
    "Results can be sensitive to the choice of sample and specification.").

% =====
% RECURRING DEFENSIBILITY QUESTIONS
% =====

defensibility_question(sample_appropriateness,
    "The regression estimates the function for the sample's market segment and time period;
    ↪ mismatch is a defensibility failure regardless of statistical fit.").
defensibility_question(specification_appropriateness,
    "Linear is a strong assumption; diagnostics, residual plots, and alternative specifications
    ↪ belong in the defensible workflow.").
defensibility_question(standard_error_interpretation,
    "Standard errors capture sampling variability under the assumed specification, NOT
    ↪ misspecification, omitted-variable bias, or sample selection.").
defensibility_question(triangulation,
    "Regression coefficients should be consistent with paired-sales, cost components, and local
    ↪ appraiser opinion; isolated coefficients are an invitation to investigate.").

% =====
% WORKED EXAMPLE: lm(price ~ gla + lot + view + cond)
% =====

:- consult('../..'/kb/pacifica_comps').

% The fitted model.
fit_call("lm(formula = price ~ gla + lot + view + cond, data = comps)").
fit_estimator(ordinary_least_squares).
fit_software(r_lm).

% Fitted coefficients, with standard errors and significance from
% the R output reproduced in the article.
fitted_coefficient(intercept, 80777.07, 164509.35, 0.491, 0.6571, ' ').
fitted_coefficient(gla, 574.93, 109.76, 5.238, 0.0135, '*').
fitted_coefficient(lot, 22.70, 10.38, 2.187, 0.1165, ' ').
fitted_coefficient(view, 86179.03, 31342.97, 2.750, 0.0708, '.').
fitted_coefficient(cond, 49824.33, 27761.27, 1.795, 0.1706, ' ').

% Diagnostics from the R output.
residual_summary(min, -17995).
residual_summary(q1, -8243).
residual_summary(median, -5433).

```

```

residual_summary(q3, -1409).
residual_summary(max, 46455).

model_summary(residual_standard_error, 30481).
model_summary(degrees_of_freedom, 3).
model_summary(r_squared, 0.9707).
model_summary(adjusted_r_squared, 0.9317).
model_summary(f_statistic, 24.86).
model_summary(model_p_value, 0.01231).

% =====
% INTERPRETIVE CAVEATS FROM THE ARTICLE
% =====

caveat(tiny_sample,
  "n=8, four predictors, three residual df. Standard errors are correspondingly wide; lot/view/
  ↳ cond fail to reach the 0.05 threshold despite economic plausibility. Small-sample artifact.").
  ↳

caveat(structural_multicollinearity,
  "Adj R2 = 0.93 with model p = 0.012 indicates strong joint fit; weak individual significance
  ↳ is a textbook multicollinearity signature. Here GLA, lot, and condition covary in the sample
  ↳ .").

caveat(comp_h_residual,
  "Comp H's residual is +$46,455; the other seven fall within +/- $18,000. The model
  ↳ systematically under-predicts Comp H by ~3x the next-largest residual. NOT noise. Signal of
  ↳ an unobserved characteristic.").

% =====
% RELATING TO PAIRED-SALES (entry 004)
% =====
%
% The regression's implicit prices ARE the values used in entry 004's
% paired-sales decomposition of A vs B. The regression generalizes that
% logic to all 8 comps simultaneously while reporting standard errors
% that paired-sales reasoning cannot produce.

confirms_paired_sales_from(entry_004,
  "A vs B differential of $127,039 from view and lot alone - coefficients are the same;
  ↳ regression confirms the paired analysis while adding standard errors.").

% =====
% RESIDUAL OF COMP H POINTS FORWARD TO ENTRY 006
% =====

points_forward_to(latent_variable,
  "Comp H's anomalous +$46,455 residual is the signal that an unobserved characteristic is
  ↳ contributing to price. Entry 006 takes up the question.").

% =====
% PREDICTED PRICES FROM THE FITTED MODEL
% =====
%
% Computes predicted price from the fitted coefficients. The residual
% for each comp is sale_price - predicted_price.

predicted_price(P, Predicted) :-
  coord(P, gla_sqft, GLA),

```

```

coord(P, lot_sqft, Lot),
coord(P, view, View),
coord(P, cond_numeric, Cond),
fitted_coefficient(intercept, B0, _, _, _, _),
fitted_coefficient(gla,      Bg, _, _, _, _),
fitted_coefficient(lot,      B1, _, _, _, _),
fitted_coefficient(view,     Bv, _, _, _, _),
fitted_coefficient(cond,     Bc, _, _, _, _),
Predicted is B0 + Bg * GLA + B1 * Lot + Bv * View + Bc * Cond.

residual_of(P, R) :-
    sale_price(P, Observed),
    predicted_price(P, Predicted),
    R is Observed - Predicted.

coord(P, A, N) :-
    attribute_value(P, A, V),
    ( number(V) -> N = V
    ; A == view, V == yes -> N = 1
    ; A == view, V == no  -> N = 0
    ; V = N
    ).

% =====
% CROSS-REFERENCES
% =====

cross_reference(heterogeneous_good).
cross_reference(characteristics_space).
cross_reference(hedonic_price_function).
cross_reference(implicit_price).
cross_reference(latent_variable).

```