

Regressão hedônica

Tradução para o português do verbete publicado originalmente em inglês. Tradução elaborada com assistência de IA e revisada pela redação; a versão em inglês é a autorizada.

Foundations • The Valuation Engineer

Bert Craytor • 3 de agosto de 2026

Versão 1.0.0

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

Definição formal. A regressão hedônica é a estimação da função de preços hedônicos a partir de transações observadas em um mercado de bens heterogêneos, que produz estimativas empíricas dos preços implícitos das características que entram no modelo.

Enquadramento intuitivo. Os verbetes anteriores descreveram a função de preços hedônicos e seus preços implícitos como objetos teóricos. A função existe em princípio, mas na prática observamos apenas um conjunto finito de transações: as cestas que foram negociadas e os preços pelos quais foram negociadas. A regressão hedônica é a ponte empírica entre as duas coisas.

A montagem é direta. Cada transação fornece um par $(\mathbf{z}_i, p_i)$: o vetor de características da cesta transacionada e seu preço de venda. A coleção desses pares é uma amostra da função de preços hedônicos (não observada). A regressão ajusta uma forma funcional a essa amostra, estimando os parâmetros que indexam a função. Uma vez estimados os parâmetros, as derivadas parciais da função ajustada produzem estimativas dos preços implícitos.

Três decisões de modelagem precisam ser tomadas antes que a regressão possa ser rodada:

1. *Quais características entram no modelo?* A escolha determina as dimensões da função estimada. Omitir uma característica relevante faz com que seu efeito vaze para os coeficientes das características incluídas que sejam correlacionadas (viés de variável omitida).
2. *Qual forma funcional?* Linear, log-linear, log-log, polinomial, com expansão de bases, semi-paramétrica, de aprendizado de máquina. A escolha restringe que formato a função estimada pode assumir e quais padrões de preços implícitos ela pode recuperar.
3. *Qual estimador?* MQO, mínimos quadrados generalizados, M-estimadores robustos, regressão quantílica, estimadores de econometria espacial, regressão penalizada. A escolha determina o que o estimador está otimizando e quais suposições sustentam sua inferência.

Cada escolha é metodologicamente substantiva e cada uma acarreta consequências de defensabilidade.

Onde os avaliadores o encontram. A regressão hedônica é o fundamento da avaliação em massa e o motor da maioria dos modelos de avaliação automatizada (AVM). Quando o órgão fiscal de um condado reavalia 100.000 imóveis por ano, é uma regressão hedônica sobre vendas recentes que está fazendo o trabalho. Quando o Zillow ou o Redfin produz uma estimativa automatizada, quem a produz é um modelo de estilo hedônico (muitas vezes um *ensemble* de árvores em vez de MQO, mas conceitualmente ainda um estimador hedônico). Quando o AVM de mercado secundário de um credor aprova ou rejeita um imóvel para refinanciamento, o mesmo maquinário está envolvido.

A avaliação de imóveis individuais também recorre cada vez mais à regressão hedônica, mesmo quando a indicação final de valor vem do método comparativo direto de dados de mercado. Preços implícitos estimados a partir de uma regressão sobre uma amostra do bairro podem sustentar as magnitudes dos ajustes de forma muito mais defensável do que a intuição do avaliador ou percentuais

de regra prática. Os coeficientes da regressão tornam-se evidência, e não opinião, para os ajustes aplicados na planilha.

Por que importa para a defensabilidade. Estimativas baseadas em regressão têm um perfil de defensabilidade distinto do da análise de pares de vendas. As vantagens: uma regressão usa todas as vendas disponíveis, em vez de duas por vez; consegue isolar o efeito de uma característica controlando simultaneamente muitas outras; e produz erros padrão que quantificam a incerteza de cada preço implícito. As contrapartidas: uma regressão impõe uma forma funcional que pode não corresponder à verdadeira função de preços; exige mais dados do que uma análise de pares de vendas; e seus resultados podem ser sensíveis à escolha da amostra e da especificação.

Várias questões de defensabilidade recorrem na avaliação baseada em regressão:

A amostra é apropriada? A regressão estima a função hedônica tal como ela opera no segmento de mercado e no período da amostra. Uma regressão sobre vendas do interior aplicada a um avaliando de frente para o mar é uma falha de defensabilidade, por mais limpas que pareçam as estatísticas.

A especificação é apropriada? Uma especificação linear é uma suposição forte. Se a função verdadeira exhibe retornos decrescentes da GLA acima de algum limiar, o ajuste linear estimará mal os preços implícitos nessa faixa. Diagnósticos, gráficos de resíduos e especificações alternativas fazem parte do fluxo de trabalho defensável.

O que significa o erro padrão? Um erro padrão capta a variabilidade amostral sob a especificação suposta. Ele não capta especificação incorreta do modelo, viés de variável omitida nem seleção amostral. Um erro padrão pequeno sobre uma especificação errada é mais perigoso do que um erro padrão grande sobre uma correta.

O resultado triangula com os demais métodos? Um preço implícito defensável baseado em regressão deve ser coerente com o que sugere a análise de pares de vendas, com o que implicam os custos dos componentes no método da quantificação do custo e com o que acreditam os avaliadores que conhecem o mercado. Um coeficiente de regressão em desacordo com os três é um convite a investigar, não a publicar.

Exemplo de avaliação resolvido. Podemos agora exercitar todo o maquinário da regressão hedônica sobre os oito elementos comparativos de Pacífica. O conjunto de dados está disponível em `shared/data/pacifica_comps.csv`.

Código em R:

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

Saída:

```
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

```

Interpretação. O modelo ajustado estima a função de preços hedônicos como:

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

Cada coeficiente é o preço implícito estimado da característica correspondente. A GLA contribui com \$575 por pé quadrado (com o erro padrão mais estreito e a significância mais forte); a vista contribui com cerca de \$86.000 para um imóvel de Pacifica; uma unidade adicional de estado de conservação (por exemplo, passar de “bom” para “excelente”) vale aproximadamente \$50.000.

Várias ressalvas são imediatamente visíveis:

A amostra é minúscula. Oito observações e quatro preditores deixam três graus de liberdade residuais. Os erros padrão são correspondentemente amplos, e os valores p de terreno, vista e estado de conservação não alcançam o limiar convencional de 0,05, apesar de sua plausibilidade econômica. Trata-se de um artefato de amostra pequena, e não de evidência de que essas características não importam. Com 80 elementos comparativos em vez de 8, os valores p encolheriam drasticamente.

O ajuste global é forte. Um R^2 ajustado = 0,93 e um valor p do modelo de 0,012 indicam que as características incluídas explicam conjuntamente a maior parte da variação de preços na amostra. A combinação de um ajuste conjunto forte com significância individual fraca é um sinal de manual de multicolinearidade, que aqui é estrutural: GLA, terreno e estado de conservação covariam na amostra.

Um resíduo se destaca. O resíduo máximo é de +\$46.455 e corresponde ao elemento H. Os outros sete resíduos caem todos dentro de uma faixa de $\pm$ \$18.000. Algo do elemento H não é captado pela função ajustada — o modelo está subprevedendo sistematicamente seu preço por um valor quase três vezes maior que o segundo maior resíduo.

Esse resíduo atípico não é ruído. É o sinal de que uma característica não observada do elemento H está contribuindo para seu preço de uma forma que as variáveis incluídas não conseguem recuperar. O verbete sobre variável latente trata exatamente dessa questão.

Validação frente ao raciocínio por pares de vendas. A análise de pares de vendas do verbete 004 previu uma diferença de \$127.039 entre os elementos A e B a partir da vista e do terreno apenas. Os preços implícitos da regressão são idênticos (uma vez que são a fonte daquele cálculo). A regressão generaliza a lógica dos pares de vendas aos oito elementos comparativos simultaneamente, mas não produz nada fundamentalmente novo com este tamanho de amostra; ela confirma a análise por pares ao mesmo tempo em que reporta erros padrão que o raciocínio por pares não é capaz de oferecer.

Referências.

- 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.
- Triplet, J. E. (2006). *Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes*. OECD Publishing.

Referências cruzadas: bem heterogêneo; espaço de características; função de preços hedônicos; preço implícito; variável latente.

Adendo: formalização em Prolog

O material complementar legível por máquina apresentado a seguir (005-hedonic-regression.pl) formaliza este verbete na base de conhecimento Prolog da revista. Ele também é publicado como composição final navegável junto com este artigo. O código não é traduzido.

```
%% 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 R^2 = 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).

```