

Hedonische Regression

Foundations • The Valuation Engineer

Bert Craytor • 2. August 2026

Version 1.0.0

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

Deutsche Fassung — Übersetzung mit KI-Unterstützung erstellt und redaktionell geprüft; die englische Fassung ist maßgeblich.

Formale Definition. Hedonische Regression ist die Schätzung der hedonischen Preisfunktion aus beobachteten Transaktionen in einem Markt heterogener Güter; sie liefert empirische Schätzungen der impliziten Preise der Merkmale, die in das Modell eingehen.

Intuitive Einordnung. Die vorangehenden Einträge beschrieben die hedonische Preisfunktion und ihre impliziten Preise als theoretische Objekte. Die Funktion existiert im Prinzip, aber in der Praxis beobachten wir nur eine endliche Menge von Transaktionen: Bündel, die gehandelt wurden, und die Preise, zu denen sie gehandelt wurden. Die hedonische Regression ist die empirische Brücke zwischen beidem.

Der Aufbau ist geradlinig. Jede Transaktion liefert ein Paar $(\mathbf{z}_i, p_i)$: den Merkmalsvektor des gehandelten Bündels und seinen Verkaufspreis. Die Sammlung dieser Paare ist eine Stichprobe aus der (unbeobachteten) hedonischen Preisfunktion. Die Regression passt dieser Stichprobe eine funktionale Form an und schätzt Parameter, die die Funktion indizieren. Sind die Parameter geschätzt, liefern partielle Ableitungen der angepassten Funktion Schätzungen der impliziten Preise.

Drei Modellierungsentscheidungen müssen getroffen werden, bevor die Regression laufen kann:

1. *Welche Merkmale gehen in das Modell ein?* Die Wahl bestimmt die Dimensionen der geschätzten Funktion. Wird ein relevantes Merkmal ausgelassen, sickert sein Effekt in die Koeffizienten korrelierter aufgenommener Merkmale (Verzerrung durch ausgelassene Variablen, Omitted-Variable Bias).
2. *Welche funktionale Form?* Linear, log-linear, log-log, polynomial, basiserweitert, semiparametrisch, maschinelles Lernen. Die Wahl beschränkt, welche Gestalt die geschätzte Funktion annehmen kann und welche Muster impliziter Preise sie zurückgewinnen kann.
3. *Welcher Schätzer?* OLS, verallgemeinerte kleinste Quadrate, robuste M-Schätzer, Quantilsregression, räumlich ökonometrische Schätzer, penalisierte Regression. Die Wahl bestimmt, was der Schätzer optimiert und welche Annahmen seine Inferenz stützen.

Jede Entscheidung ist methodisch substanziell, und jede hat Konsequenzen für die Vertretbarkeit.

Wo Gutachter dem Konzept begegnen. Die hedonische Regression ist das Fundament der Massenbewertung und der Motor der meisten automatisierten Bewertungsmodelle (AVMs). Wenn ein County Assessor jährlich 100,000 Parzellen neu bewertet, leistet eine hedonische Regression auf jüngeren Verkäufen die Arbeit. Wenn Zillow oder Redfin eine automatisierte Schätzung produziert, erzeugt sie ein Modell hedonischen Typs (oft ein Baum-Ensemble statt OLS, konzeptionell aber weiterhin ein hedonischer Schätzer). Wenn das Zweitmarkt-AVM eines Kreditgebers eine Immobilie für eine Refinanzierung durchwinkt oder ablehnt, ist dieselbe Maschinerie beteiligt.

Auch das Einzelgutachten stützt sich zunehmend auf hedonische Regression, selbst wenn die endgültige Wertindikation aus dem Vergleichswertverfahren stammt. Aus einer Regression auf einer

Nachbarschaftsstichprobe geschätzte implizite Preise können Anpassungshöhen weit vertretbarer stützen als Gutachterintuition oder Faustregel-Prozentsätze. Die Regressionskoeffizienten werden zu Beweismitteln, nicht Meinungen, für die im Raster angewandten Anpassungen.

Warum es für die Vertretbarkeit wichtig ist. Regressionsbasierte Schätzungen tragen ein Vertretbarkeitsprofil, das sich von der Paarvergleichsanalyse unterscheidet. Die Vorteile: Eine Regression nutzt alle verfügbaren Verkäufe statt jeweils zwei; sie kann den Effekt eines Merkmals isolieren und dabei viele andere gleichzeitig kontrollieren; und sie liefert Standardfehler, die die Unsicherheit jedes impliziten Preises quantifizieren. Die Kompromisse: Eine Regression erlegt eine funktionale Form auf, die zur wahren Preisfunktion möglicherweise nicht passt; sie benötigt mehr Daten als eine Paarvergleichsanalyse; und ihre Ergebnisse können empfindlich auf die Wahl von Stichprobe und Spezifikation reagieren.

Mehrere Vertretbarkeitsfragen kehren in der regressionsbasierten Bewertung wieder:

Ist die Stichprobe angemessen? Die Regression schätzt die hedonische Funktion, wie sie im Marktsegment und Zeitraum der Stichprobe wirkt. Eine Regression auf Binnenland-Verkäufen, angewandt auf ein Bewertungsobjekt in Strandlage, ist ein Vertretbarkeitsversagen, egal wie sauber die Statistik aussieht.

Ist die Spezifikation angemessen? Eine lineare Spezifikation ist eine starke Annahme. Zeigt die wahre Funktion oberhalb einer Schwelle abnehmende Erträge der GLA, wird die lineare Anpassung die impliziten Preise in diesem Bereich falsch schätzen. Diagnostik, Residuenplots und alternative Spezifikationen gehören zum vertretbaren Arbeitsablauf.

Was bedeutet der Standardfehler? Ein Standardfehler erfasst die Stichprobenvariabilität unter der angenommenen Spezifikation. Er erfasst nicht Modellfehlspezifikation, Verzerrung durch ausgelassene Variablen oder Stichprobenselektion. Ein kleiner Standardfehler bei falscher Spezifikation ist gefährlicher als ein großer Standardfehler bei richtiger.

Triangulieren die Ergebnisse mit anderen Ansätzen? Ein vertretbarer regressionsbasierter impliziter Preis sollte mit dem konsistent sein, was die Paarvergleichsanalyse nahelegt, was die Komponentenkosten des Sachwertverfahrens implizieren und was marktkundige Gutachter glauben. Ein Regressionskoeffizient im Widerspruch zu allen dreien ist eine Einladung zur Untersuchung, nicht zur Veröffentlichung.

Durchgerechnetes Bewertungsbeispiel. Wir können nun die volle Maschinerie der hedonischen Regression auf die acht Pacifica-Comps anwenden. Der Datensatz ist verfügbar unter `shared/data/pacifica_comps`

R-Code:

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

Ausgabe:

```
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
```

Interpretation. Das angepasste Modell schätzt die hedonische Preisfunktion als:

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

Jeder Koeffizient ist der geschätzte implizite Preis des entsprechenden Merkmals. GLA trägt \$575 pro Quadratfuß bei (mit dem engsten Standardfehler und der stärksten Signifikanz); Aussicht trägt etwa \$86,000 zu einer Pacifica-Immobilie bei; eine zusätzliche Zustandsstufe (z.B. von „gut“ zu „ausgezeichnet“) ist rund \$50,000 wert.

Mehrere Vorbehalte sind unmittelbar sichtbar:

Die Stichprobe ist winzig. Acht Beobachtungen und vier Prädiktoren lassen drei Freiheitsgrade für die Residuen. Die Standardfehler sind entsprechend breit, und die p -Werte für Grundstück, Aussicht und Zustand verfehlen die konventionelle 0.05-Schwelle trotz ihrer ökonomischen Plausibilität. Das ist ein Kleinstichproben-Artefakt, kein Beleg dafür, dass diese Merkmale nicht zählen. Mit 80 statt 8 Comps würden die p -Werte drastisch schrumpfen.

Die Gesamtanpassung ist stark. Ein adjustiertes $R^2 = 0.93$ und ein Modell- p -Wert von 0.012 zeigen an, dass die aufgenommenen Merkmale gemeinsam den Großteil der Preisvariation in der Stichprobe erklären. Die Kombination aus starker gemeinsamer Anpassung und schwacher individueller Signifikanz ist ein Lehrbuchsignal für Multikollinearität, die hier strukturell ist: GLA, Grundstück und Zustand kovariieren in der Stichprobe.

Ein Residuum sticht heraus. Das maximale Residuum beträgt +\$46,455 und entspricht Comp H. Die übrigen sieben Residuen liegen alle innerhalb eines Bandes von $\pm \$18,000$. Irgendetwas an Comp H wird von der angepassten Funktion nicht erfasst — das Modell unterschätzt seinen Preis systematisch, um einen Betrag, der fast dreimal so groß ist wie das nächstgrößte Residuum.

Dieses Ausreißer-Residuum ist kein Rauschen. Es ist das Signal, dass ein unbeobachtetes Merkmal von Comp H auf eine Weise zu dessen Preis beiträgt, die die aufgenommenen Variablen nicht zurückgewinnen können. Der Eintrag zur latenten Variable greift genau diese Frage auf.

Validierung gegen das Paarvergleichsdenken. Die Paarvergleichsanalyse aus Eintrag 004 prognostizierte eine Differenz von \$127,039 zwischen Comp A und Comp B allein aus Aussicht und Grundstück. Die impliziten Preise der Regression sind identisch (da sie die Quelle jener Berechnung sind). Die Regression verallgemeinert die Paarvergleichslogik auf alle acht Comps gleichzeitig, bringt aber bei dieser Stichprobengröße nichts grundsätzlich Neues hervor; sie bestätigt die Paaranalyse und berichtet dabei Standardfehler, die das Paarvergleichsdenken nicht liefern kann.

Literatur.

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

Querverweise: heterogenes Gut; Merkmalsraum; hedonische Preisfunktion; impliziter Preis; latente Variable.

Addendum: Prolog-Formalisierung

Die folgende maschinenlesbare Begleitdatei (005-hedonic-regression.pl) formalisiert diesen Eintrag in der Prolog-Wissensbasis der Zeitschrift. Sie wird zudem als browsbare Fahne (Galley) neben diesem Artikel veröffentlicht.

```
%% 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).

```