

헤도닉 회귀분석

(한국어판)

기초 개념 (Foundations) • The Valuation Engineer

Bert Craytor • 2026년 8월 3일

판본 1.0.0

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

한국어 번역본. 인공지능의 도움을 받아 번역하고 편집부가 검토했습니다. 영문판이 정본입니다.

형식적 정의. 헤도닉 회귀분석 (hedonic regression)이란 이질재 시장에서 관측된 거래로부터 헤도닉 가격함수를 추정하는 것으로, 모형에 들어간 특성들의 잠재가격에 대한 경험적 추정치를 산출한다.

직관적 이해. 앞선 항목들은 헤도닉 가격함수와 그 잠재가격을 이론적 대상으로 서술했다. 그 함수는 원리적으로 존재하지만, 실제로 우리가 관측하는 것은 유한한 거래 집합뿐이다. 거래된 묶음과 거래된 가격이 그것이다. 헤도닉 회귀분석은 이 둘을 잇는 경험적 다리이다.

설정은 간명하다. 각 거래는 (z_i, p_i) 쌍을 제공한다. 거래된 묶음의 특성 벡터와 그 거래가격이다. 이 쌍들의 모임은 (관측되지 않은) 헤도닉 가격함수로부터 얻은 하나의 표본이다. 회귀분석 (regression)은 그 표본에 함수형태를 적합시켜 함수를 특징짓는 모수를 추정한다. 모수가 추정되고 나면 적합한 함수의 편도함수가 잠재가격 추정치를 준다.

회귀분석을 돌리기 전에 세 가지 모형화 선택을 해야 한다:

1. 어떤 특성을 모형에 넣을 것인가? 이 선택이 추정되는 함수의 차원을 결정한다. 관련 있는 특성을 누락하면 그 효과가 상관된 포함 특성들의 계수로 스며든다 (누락변수 편).
2. 어떤 함수형태인가? 선형, 로그선형, 이중로그, 다항식, 기저확장, 준모수적, 기계학습 등이 있다. 이 선택은 추정된 함수가 취할 수 있는 모양과 복원 가능한 잠재가격 양상을 제약한다.
3. 어떤 추정량인가? OLS, 일반화최소제곱, 로버스트 M-추정량, 분위수회귀, 공간계량경제 추정량, 벌점회귀 등이 있다. 이 선택은 추정량이 무엇을 최적화하는지, 그리고 어떤 가정이 그 추론을 뒷받침하는지를 결정한다.

각 선택은 방법론적으로 실질적이며, 각각 방어가능성에 결과를 남긴다.

감정평가사가 마주치는 지점. 헤도닉 회귀분석은 대량평가 (mass appraisal)의 기초이며 대부분의 자동가치산정모형 (AVM)의 엔진이다. 카운티 과세평가관이 해마다 10만 필지를 재평가할 때, 그 일을 해내는 것은 최근 거래에 대한 헤도닉 회귀분석이다. Zillow 나 Redfin 이 자동 추정치를 내놓을 때, 그것을 산출하는 것은 헤도닉 계열 모형이다 (OLS가 아니라 트리 앙상블인 경우가 많지만 개념적으로는 여전히 헤도닉 추정량이다). 대출기관의 유통시장 AVM이 어떤 부동산의 재융자 가부를 판정할 때에도 같은 장치가 관여한다.

개별 부동산 감정평가도 점점 더 헤도닉 회귀분석에 기대고 있다. 최종 가치 도출이 거래사례비교에서 나오는 경우에도 그렇다. 근린 표본에 대한 회귀분석에서 추정한 잠재가격은 감정평가사의 직관이나 경험적 백분율보다 훨씬 방어 가능하게 보정 크기를 뒷받침할 수 있다. 회귀계수는 격자에 적용되는 보정에 대해 의견이 아니라 증거가 된다.

왜 방어가능성에 중요한가. 회귀분석에 기반한 추정은 짝지은 매매 분석과는 다른 방어가능성 프로파일을 가진다. 장점은 이렇다. 회귀분석은 한 번에 두 건이 아니라 이용 가능한 모든 거래를 사용한다. 다른 여러 특성을 동시에 통제하면서 한 특성의 효과를 분리할 수 있다. 그리고 각 잠재가격의 불확실성을 정량화하는 표준오차를 산출한다. 대가는 이렇다. 회귀분석은 참된 가격함수와 맞지 않을 수도 있는 함수형태를 부과한다. 짝지은 매매 분석보다 더 많은 자료를 요구한다. 그리고 그 결과는 표본과 모형설정의 선택에 민감할 수 있다.

회귀분석 기반 감정평가에서는 몇 가지 방어가능성 문제가 되풀이된다:

표본은 적절한가? 회귀분석은 표본이 속한 시장 부문과 기간에서 헤도닉 함수가 작동하는 방식을 추정한다. 내륙 거래로 돌린 회귀분석을 해변 대상에 적용하는 것은, 통계 결과가 아무리 깔끔해 보여도 방어가능성의 실패이다.

모형설정은 적절한가? 선형 설정은 강한 가정이다. 참된 함수가 어떤 문턱 위에서 GLA 에 대해 수확체감을 보인다면, 선형 적합은 그 구간에서 잠재가격을 잘못 추정할 것이다. 진단, 잔차도, 대안적 설정은 방어 가능한 작업흐름의 일부이다.

표준오차는 무엇을 뜻하는가? 표준오차는 가정된 모형설정 아래에서의 표본변동을 포착한다. 그것은 모형설정 오류, 누락변수 편倚, 표본선택 문제를 포착하지 않는다. 잘못된 설정에서의 작은 표준오차는 올바른 설정에서의 큰 표준오차보다 더 위험하다.

결과가 다른 방식과 삼각검증되는가? 방어 가능한 회귀분석 기반 잠재가격이라면 짝지은 매매 분석이 시사하는 바, 원가법의 구성요소 원가가 함의하는 바, 그리고 해당 시장에 밝은 감정평가사들이 믿는 바와 일관되어야 한다. 이 셋 모두와 어긋나는 회귀계수는 발표하라는 신호가 아니라 조사하라는 초대장이다.

감정평가 실무 예제. 이제 여덟 개의 퍼시픽카 비교사례에 대해 헤도닉 회귀분석 장치 전체를 돌려 볼 수 있다. 데이터셋은 `shared/data/pacifica_comps.csv` 에 있다.

R 코드:

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

출력:

```
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
```

해석. 적합한 모형은 헤도닉 가격함수를 다음과 같이 추정한다:

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

각 계수는 해당 특성의 추정된 잠재가격이다. GLA 는 제곱피트당 \$575 를 기여하며 (표준오차가 가장 좁고 유의성이 가장 강하다), 조망은 퍼시픽카 부동산에 약 \$86,000 을 기여한다. 상태가 한 단계 오르면 (예컨대 “양호” 에서 “우수” 로) 대략 \$50,000 의 가치가 있다.

몇 가지 유의점이 곧바로 눈에 띈다:

표본이 매우 작다. 관측치 여덟 개에 예측변수 네 개이므로 잔차 자유도는 셋뿐이다. 표준오차는 그만큼 넓고, 대자·조망·상태의 p -값은 경제적으로 그럴듯함에도 통상적인 0.05 문턱에 이르지 못한다.

이는 소표본의 인공물이지만, 이 특성들이 중요하지 않다는 증거가 아니다. 비교사례가 8 건이 아니라 80 건이라면 p -값은 극적으로 줄어들 것이다.

전체 적합도는 강하다. 조정 $R^2 = 0.93$ 과 모형 p -값 0.012 는 포함된 특성들이 표본 내 가격 변동의 대부분을 공동으로 설명함을 나타낸다. 강한 결합 적합도와 약한 개별 유의성의 조합은 다중공선성의 교과서적 신호이며, 여기서는 구조적이다. GLA, 대지, 상태가 표본 안에서 함께 움직이기 때문이다.

잔차 하나가 유독 두드러진다. 최대 잔차는 +\$46,455 이며 비교사례 H 에 해당한다. 나머지 일곱 개 잔차는 모두 $\pm \$18,000$ 대역 안에 든다. 비교사례 H 의 무언가가 적합된 함수에 포착되지 않은 것이다. 모형은 그 가격을 두 번째로 큰 잔차의 거의 세 배에 달하는 크기로 체계적으로 과소 예측하고 있다.

이 이상치 잔차는 잡음이 아니다. 그것은 비교사례 H 의 어떤 미관측 특성이, 포함된 변수들로는 복원할 수 없는 방식으로 가격에 기여하고 있다는 신호이다. 잠재변수 항목이 다루는 것이 바로 이 문제이다.

짜지은 매매 추론과의 검증. 항목 004 의 짜지은 매매 분석은 조망과 대지만으로 비교사례 A 와 B 사이에 \$127,039 의 차이를 예측했다. 회귀분석의 잠재가격은 그것과 동일하다 (그 계산의 출처가 바로 이 잠재가격이기 때문이다). 회귀분석은 짜지은 매매의 논리를 여덟 개 비교사례 전체로 동시에 일반화하지만, 이 표본 크기에서는 본질적으로 새로운 것을 낳지 않는다. 짜지은 분석을 확인해 주면서, 짜지은 매매 추론으로는 얻을 수 없는 표준오차를 보고할 뿐이다.

참고문헌.

- 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.

상호 참조: 이질재 (*heterogeneous good*); 특성 공간 (*characteristics space*); 헤도닉 가격함수 (*hedonic price function*); 잠재가격 (*implicit price*); 잠재변수 (*latent variable*).

부록: Prolog 형식화

아래의 기계 판독 가능한 부속 파일 (005-hedonic-regression.pl) 은 본지의 Prolog 지식베이스에서 이 항목을 형식화한 것이다. 이 파일은 본문과 함께 열람 가능한 Prolog 갤리로도 발행된다.

```
%% 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, Bl, _, _, _, _),
    fitted_coefficient(view, Bv, _, _, _, _),
    fitted_coefficient(cond, Bc, _, _, _, _),
    Predicted is B0 + Bg * GLA + Bl * 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).
```