Hedonic Regression

Foundations • The Valuation Engineer

Bert Craytor• May 29, 2026
Version 0.1.0-draft

Formal definition. 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.

Intuitive framing. The previous entries described the hedonic price function and its implicit prices as theoretical objects. The function exists in principle, but in practice we observe only a finite set of transactions: bundles that traded, and the prices at which they traded. Hedonic regression is the empirical bridge between the two.

The setup is straightforward. Each transaction supplies a $(\mathbf{z}_i, p_i)$ pair: the characteristics vector of the transacted bundle and its sale price. The collection of these pairs is a sample from the (unobserved) hedonic price function. Regression fits a functional form to that sample, estimating parameters that index the function. Once the parameters are estimated, partial derivatives of the fitted function yield implicit-price estimates.

Three modeling choices have to be made before regression can run:

1. *Which characteristics enter the model?* The choice determines the dimensions of the estimated function. Omitting a relevant characteristic causes its effect to leak into the coefficients of correlated included characteristics (omitted-variable bias).
2. *What functional form?* Linear, log-linear, log-log, polynomial, basis-expanded, semiparametric, machine-learning. The choice constrains what shape the estimated function can take and which implicit-price patterns it can recover.
3. *What estimator?* OLS, generalized least squares, robust M-estimators, quantile regression, spatial econometric estimators, penalized regression. The choice determines what the estimator is optimizing and what assumptions support its inference.

Each choice is methodologically substantive and each carries defensibility consequences.

Where appraisers encounter it. Hedonic regression is the foundation of mass appraisal and the engine of most automated valuation models (AVMs). When a county assessor revalues 100,000 parcels annually, a hedonic regression on recent sales is doing the work. When Zillow or Redfin produces an automated estimate, a hedonic-style model (often a tree ensemble rather than OLS, but conceptually still a hedonic estimator) is producing it. When a lender's secondary-market AVM passes or fails a property for a refi, the same machinery is involved.

Single-property appraisal increasingly draws on hedonic regression as well, even when the final value indication comes from sales comparison. Implicit prices estimated from a regression on a neighborhood sample can support adjustment magnitudes far more defensibly than appraiser intuition or rule-of-thumb percentages. The regression coefficients become evidence, not opinion, for the adjustments applied in the grid.

Why it matters for defensibility. Regression-based estimates carry a defensibility profile distinct from paired-sales analysis. The advantages: a regression uses all available sales rather than two at a time; it can isolate the effect of one characteristic while controlling for many others simultaneously; and it produces standard errors that quantify the uncertainty in each implicit price.

The trade-offs: a regression imposes a functional form that may not match the true price function; it requires more data than a paired-sales analysis; and its results can be sensitive to the choice of sample and specification.

Several defensibility questions recur in regression-based appraisal:

Is the sample appropriate? The regression estimates the hedonic function as it operates in the sample's market segment and time period. A regression on inland sales applied to a beachfront subject is a defensibility failure regardless of how clean the statistics look.

Is the specification appropriate? A linear specification is a strong assumption. If the true function exhibits diminishing returns to GLA above some threshold, the linear fit will misestimate implicit prices in that range. Diagnostics, residual plots, and alternative specifications are part of the defensible workflow.

What does the standard error mean? A standard error captures sampling variability under the assumed specification. It does not capture model misspecification, omitted-variable bias, or sample selection. A small standard error on a wrong specification is more dangerous than a large standard error on a correct one.

Does the result triangulate with other approaches? A defensible regression-based implicit price should be consistent with what paired-sales analysis suggests, with what cost-approach component costs imply, and with what appraisers familiar with the market believe. A regression coefficient at odds with all three is an invitation to investigate, not to publish.

Worked appraisal example. We can now exercise the full hedonic regression machinery on the eight Pacifica comps. The dataset is available at `shared/data/pacifica_comps.csv`.

R code:

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

Output:

```
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. The fitted model estimates the hedonic price function as:

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

Each coefficient is the estimated implicit price of the corresponding characteristic. GLA contributes \$575 per square foot (with the narrowest standard error and the strongest significance);

view contributes about $86,000 to a Pacifica property; an additional unit of condition (e.g., moving from "good" to "excellent") is worth roughly $50,000.

Several caveats are immediately visible:

The sample is tiny. Eight observations and four predictors leaves three residual degrees of freedom. The standard errors are correspondingly wide, and the p-values for lot, view, and condition fail to reach the conventional 0.05 threshold despite their economic plausibility. This is a small-sample artifact, not evidence that these characteristics do not matter. With 80 comps rather than 8, the p-values would shrink dramatically.

The overall fit is strong. Adjusted $R^2 = 0.93$ and a model p-value of 0.012 indicate that the included characteristics jointly explain most of the price variation in the sample. The combination of strong joint fit with weak individual significance is a textbook signal of multicollinearity, which here is structural: GLA, lot, and condition covary in the sample.

One residual stands out. The maximum residual is +$46,455, corresponding to Comp H. The other seven residuals all fall within a ±$18,000 band. Something about Comp H is not captured by the fitted function — the model is systematically under-predicting its price by an amount nearly three times the next largest residual.

This outlier residual is not noise. It is the signal that an unobserved characteristic of Comp H is contributing to its price in a way the included variables cannot recover. The latent variable entry takes up exactly this question.

Validation against paired-sales reasoning. The paired-sales analysis from entry 004 predicted a $127,039 difference between Comps A and B from view and lot alone. The regression's implicit prices are identical (since they are the source of that calculation). The regression generalizes the paired-sales logic to all eight comps simultaneously, but it produces nothing fundamentally new at this sample size; it confirms the paired analysis while reporting standard errors that paired-sales reasoning cannot.

Cross-references: heterogeneous good; *characteristics space*; *hedonic price function*; *implicit price*; *latent variable.*