Hedonic Price Model for Houses

Alejandro Lopez Ruiz – s4075315

October 2025

Introduction

This paper analyzes the main determinants of property prices in the Netherlands. It examines how structural, locational, and temporal attributes explain price variation.

The study follows a widely used approach in housing economics, a hedonic price model using Ordinary Least Squares (OLS). The theoretical foundation originates from Rosen (1974), which served as the basis for subsequent studies. Rosen's model views each property as a bundle of attributes whose implicit values determine price together. Subsequent studies (Herath & Maier, 2010; Monson, 2009) have explored the versatility of such framework in capturing structural and spatial determinants.

Accordingly, this study's research question is: "What are the main determinants of house prices in the Netherlands between 1995 and 2000?" The following sections address this question through empirical estimation and discussion of results.

Data Preparation

The dataset consists of 632,790 transactions across the Netherlands between 1995 and 2000. Several cleaning and transformation steps will ensure accuracy and comparability across variables.

The dependent variable *transactionprice* was log-transformed – *lprice* – to correct right-skewness. (Malpezzi, 2003). Extreme outliers were trimmed using the 1st–99th percentile range (Appendix 1). Key structural variables – *usablefloorarea*, *grossvolume*, *plotarea* – were treated similarly, log-transformed, and trimmed for invalid entries. (Koster & Rouwendal, 2020). Urbanization was rescaled categorical variables encoded, and duration log-transformed (Herath & Maier, 2010).

Variable	Description	Transformation / Notes
lprice	Log of transaction price	Dependent variable
larea	Log of usable floor area (m²)	Excludes outliers >1000

lplot	Log of plot area (m^2)	Excludes implausible values
lvolume	Log of gross volume (m^3)	Size control
numberofrooms	Number of rooms	Trimmed at 25
maint_quality	Average of interior/exterior maintenance (1–10)	Combined quality score
price_m2	Transaction price per m^2	Recalculated
land_intensity	Plot area / floor area ratio	Density measure
urban	Urbanization level (1=least urban, 5=most urban)	Reversed scale
lduration	Log of days between listing and sale	+1 adjustment for zeros
prov_id	Province identifier	Encoded categorical variable
constructionperiod	Construction era	9 fixed-effect dummies
number_nr	Home type	10 fixed-effect dummies
year	Transaction year	1995–2000 fixed effects
shed	Type/presence of shed	6-category dummy

Table 1: Summary of variables included in the final models, detailing their definitions, composition, and transformations.

Variable selection follows an incremental complexity approach, detailed in Section 4 and grounded in housing economics literature. Following Rosen (1974) and Monson (2009), the baseline model: *larea*, *lplot*, *lvolume, numberofrooms*, and *number_nr*, captures only core structural attributes, hypothesized to be the primary determinants of housing prices.

Furthermore, Sirmans, Macpherson, & Zietz (2005) and Steegmans & Hassink (2025) highlight that quality features – *maint_quality*, *shed*, *land_intensity* – significantly influence property values.

Koster & Rouwendal (2019) emphasize the role of location attributes and the urban agglomeration effect, while Herath & Maier (2010) highlight the importance of accessibility. Although represented only by the variable *urban*, it remains essential. Regional fixed effects – *prov_id* – follow Forouhar & Van Lierop (2021), who confirm spatial context relevance in the Dutch market.

Finally, temporal attributes—*year* and *constructionperiod*—capture market evolution and structural age effects. *lduration* additionally captures listing dynamics and seller behavior (Monson, 2009).

Descriptive Analysis

The cleaned dataset consists of ≈130,000 complete housing transactions. This section describes main characteristics and price patterns. Each observation in the dataset grasps key determinants of housing value, classified into three broad groups:

- Structural attributes: describe the quality and physical characteristics
- Location attributes: capture spatial variation and regional development differences
- Temporal attributes: reflect market dynamics (inflation, economic growth, and seller behavior…)

Table 2 presents descriptive statistics for key structural variables. The average selling price is ≈€140,000 with a standard deviation of €77,000, reflecting the high variation in prices. The usable floor area dispersion is half its mean, alike plot size and volume, confirming adequate variability for regression analysis.

	count	mean	sd	min	max
Selling price of the property	620784	142874.9	77045.04	37000	533000
Total usable living space (m^2)	319641	104.7828	50.98294	1	350
Plot size (m^2)	469256	310.4276	330.2449	55	2680
Gross volume of the home (m^3)	591899	353.9165	151.8929	10	1100
lprice	620784	11.74815	.4842171	10.51867	13.18628
larea	309759	4.568726	.5365092	.6931472	5.857933
lplot	469256	5.424968	.7136615	4.007333	7.893572
lvolume	591899	5.745195	.6241025	2.302585	7.003066
Number of rooms in the home	632643	4.355317	1.353189	0	25
maint_quality	632790	7.030673	1.102736	1	10
Urbanization level	631377	3.560492	1.304947	1	5

Table 2: Descriptive statistics of key structural and locational variables used in the hedonic model. Logarithmic transformations were applied to mitigate skewness and ensure approximate normality of continuous variables.

Figure 1 contrasts raw and log-transformed distributions for transaction prices. Raw prices show strong right-skewness. Log transformations normalize the distributions (Malpenzzi, 2003), satisfying general assumptions for linear estimation.

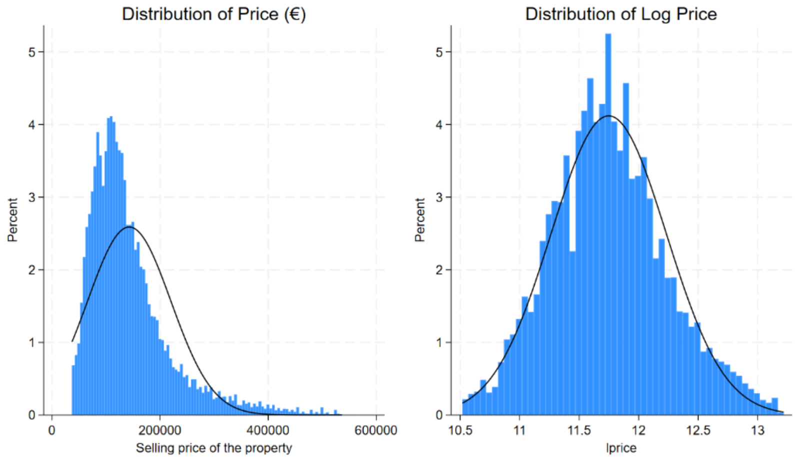


Figure 1: Raw and log-transformed distributions of transaction prices. The log transformation normalizes the data, supporting the linear specification of the hedonic model.

Figure 2 illustrates the relationship between log-price and log-area. The polynomial fit shows a positive but concave trend; larger homes are more expensive, but marginal increases decline with size, consistent with diminishing utility (Herath & Maier, 2010).

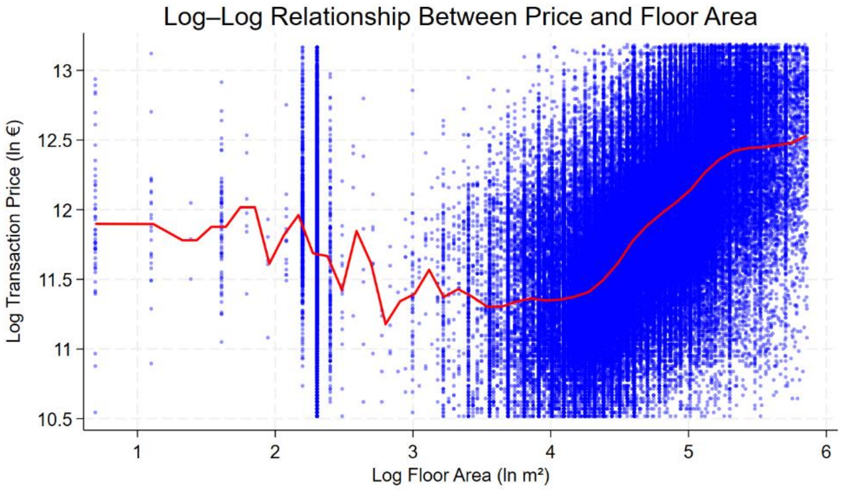


Figure 2: Relationship between log price and log floor area. The concave shape indicates diminishing marginal price increases with larger dwelling size.

Temporal patterns are plotted in Figure 3. Median log prices increased steadily between 1995 and 2000, with a cumulative increase of over 40%, consistent with the late-1990s housing boom described by Steegmans & Hassink (2025), supporting the inclusion of year fixed effects.

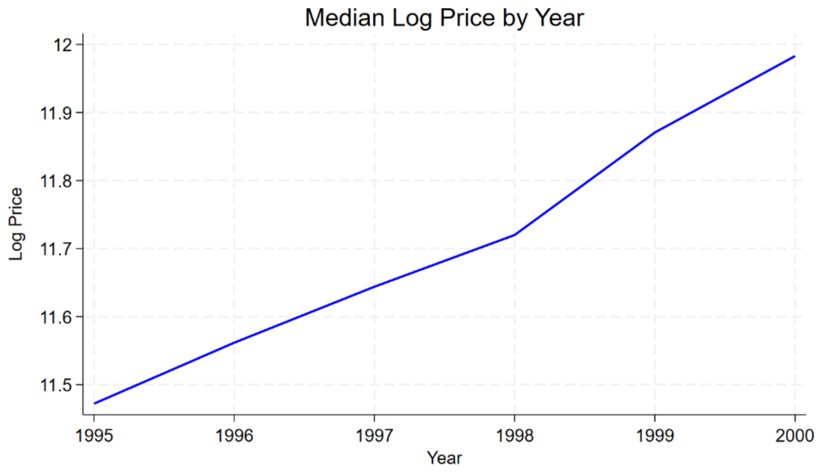


Figure 3: Evolution of median log house prices (1995–2000). The figure highlights steady appreciation consistent with the national housing boom of the late 1990s.

Finally, Figure 4 shows significant regional heterogeneity. Provinces such as Utrecht and Zuid-Holland have highest average log-prices, while Groningen and Fryslân show lower prices. These patterns align with urban concentration effects highlighted by Koster & Rouwendal (2020) and Forouhar & Van Lierop (2021).

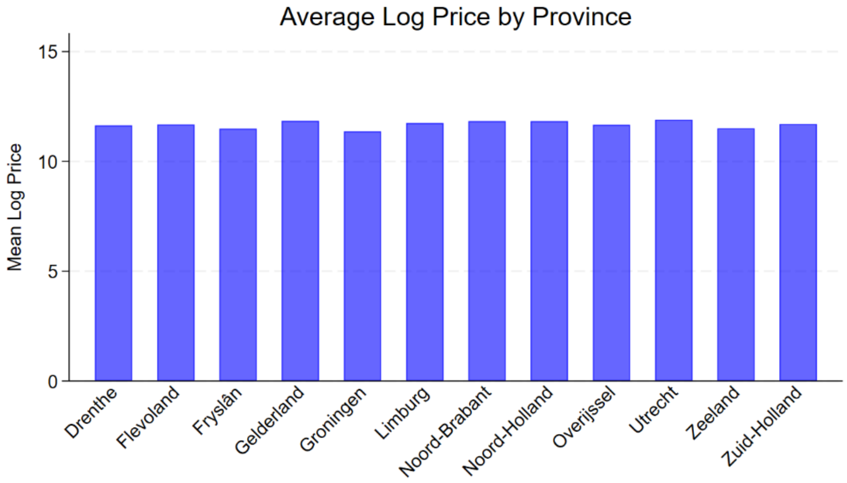


Figure 4: Mean log transaction prices by province, showing pronounced regional disparities consistent with urban concentration effects.

Overall, the descriptive analysis shows that Dutch housing prices between 1995 and 2000 varied systematically with structural, temporal, and spatial characteristics.

Population Model and Estimation Method

The hedonic price model estimates the logarithm of transaction price as a function of structural, qualitative, locational, and temporal attributes. Following Rosen (1974), housing prices reflect the implicit value of their attributes, under competitive conditions, allowing price to be decomposed.

Formally, the population model is estimated as:

$$\ln(P_i) = \beta_0 + \beta_1 X_i^S + \beta_2 X_i^Q + \beta_3 X_i^L + \beta_4 X_i^T + \epsilon_i$$

W

An incremental model-complexity approach is applied to evaluate the explanatory power. Model 1 includes only the core structural variables – larea, lplot, lvolume, numberofrooms, number_nr – forming the empirical foundation of the model (Rosen, 1974; Monson, 2009). Model 2 extends that baseline, incorporating quality variables – land_intensity, maint_quality, shed – consistent with Sirmans, Macpherson, and Zietz (2005) and supported by Steegmans and Hassink (2025), who highlight the value effect of maintenance and minor features.

Model 3 introduces the "urban" and provincial effects – prov_id – to account for locational effects (Herath & Maier, 2010; Koster & Rouwendal, 2020). Model 4 adds temporal dynamics – year, constructionperiod – and temporal market behavior – lduration –, capturing time-on-market effects. Finally, model 5 uses spatial insights and integrates regional effects by replacing prov_id for pc2_region, capturing more granular spatial heterogeneity.

Larger properties, better maintained, and higher urbanization are expected to increase prices, while longer selling durations or older construction periods are associated with lower values (Malpezzi, 2003; Monson, 2009). Spatial ispecification should enhance model performance, and prices are expected to rise over time, consistent with macroeconomics.

OLS estimation is applied with robust standard errors, mitigating possible heteroskedasticity. OLS is the appropriate choice as it estimates the marginal contribution of each attribute. This progressive specification further allows for direct comparison of models. Overall, this modeling framework allows systematic testing of how attributes jointly explain housing price variation across the Netherlands.

Results

Table 3 presents the progression of the five hedonic price models, showing how the inclusion of additional explanatory variables improves overall performance. The baseline explains approximately 45% of the variation in log-prices, relying on core structural characteristics. As additional quality, locational and temporal attributes are added, the explanatory power increases greatly, reaching R2=0.771 (see Appendix 2 for Stata model specifications).

Statistics	Model 1	Model 2	Model 3	Model 4	Model 5
R-squared	0.449	0.460	0.639	0.703	0.771
Observations	175,530	129,644	129,425	128,139	128,139

Table 3: Progression of the hedonic price models (Models 1–5). The R-squared increases from 0.449 to 0.771, confirming that successive inclusion of quality, locational, and temporal variables markedly enhances model fit.

The increase in explanatory power demonstrates that housing prices depend not only on physical structure, but also on quality, spatial context and temporal market dynamics. This pattern is consistent with Sirmans (2005) and Steegmans & Hassink (2025), who emphasize the role of qualitative and spatial factors in hedonic models

For the sake of interpretation, Model 4 is selected as the final hedonic price model. Although Model 5 improves explanatory power, it does so by incorporating 90 regional dummies, which limits interpretability. Model 4 provides clearer, economically meaningful coefficients with strong explanatory power.

Table 4 summarizes the main coefficients from the final hedonic model; complete regression output can be found in Appendix 3. Coefficients are interpreted as semi-elasticities due to the log specification.

Variable	**Coefficient (Std. Error)**	**Significance**
Log Area (larea)	0.057 (0.003)	***
Log Plot (lplot)	0.118 (0.003)	***
Log Volume (lvolume)	0.480 (0.012)	***
Number of Rooms (incl. living room)	0.035 (0.001)	***
Maintenance Quality	0.058 (0.001)	***
Land Intensity	0.003 (0.000)	***
Urbanization Level (5 = most urban)	0.010 (0.001)	***
Log Duration (lduration)	–0.021 (0.001)	***
Constant	6.990 (0.055)	***

*Table 4: Main coefficients from the final hedonic regression model. All variables are statistically significant at the 1% level, confirming the robustness of the estimates. Standard errors are shown in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.*

The positive coefficients for larea (0.057), lplot (0.118), and lvolume (0.480) indicate that, as expected, larger properties mean higher prices (Monson, 2009). Specifically, a

1% increase in area is a 0.057% increase in price, while volume corresponds to 0.48%, highlighting the dominant role of spatial capacity in value formation.

The number of rooms also contributes positively (0.035); holding other factors constant, each additional room increases prices by 3.5%. Quality-related attributes, maintenance quality (0.058) and land intensity (0.003), also show significant effects, emphasizing the influence of maintenance and land use efficiency.

Locational and temporal variables further improve R^2. Urbanization level (0.010) confirms that urbanized areas have higher values, in line with the premium associated with accessibility and agglomeration discussed by Koster and Rouwendaal (2020). Longer selling durations (–0.021) reduce prices, consistent with market efficiency theory.

The inclusion of provincial effects captures geographical information. Properties in Noord-Holland or Utrecht exhibit higher coefficients (0.518, 0.554) respective to the reference category, Drenthe. Consistent with Forouhar & Van Lierop (2021), who highlight spatial dynamics in Dutch prices. Temporal effects confirm steady yearly price increases (Figure 3), reflecting late-1990s macroeconomic expansion (Herath & Maier, 2010).

Overall, the final model demonstrates strong explanatory power (R^2=0.703) and robust statistical significance. The findings support theoretical premises that housing value is jointly determined by structural, locational and temporal attributes. These results align with existing literature and confirm the effectiveness of the incremental hedonic framework.

Discussion

The results confirm that Dutch housing prices in the late 1990s were determined by structural, locational, and temporal attributes. Despite the model’s high explanatory power, several methodological considerations remain.

Concerns arise from omitted variable bias. The dataset contains structural and contextual information but omits neighborhood and environmental variables. Examples include energy efficiency or proximity to schools and transport, which strongly influence prices (Herath & Maier, 2010). Likewise, unobserved property features, such as design or renovation status, could also affect prices. This may explain some of the residual variation in the model.

Another potential issue is sample bias. The dataset covers exclusively executed transactions, excluding unsold or withdrawn properties. If unobservable factors influence both the probability of sale and the price, the model might reflect selection effects rather than price determinants. Although a large sample size mitigates some concern, future work should correct selection bias.

Regarding causal interpretation, coefficients represent associations, not direct causal effects. The cross-sectional design limits ability to control unobservable phenomena or reverse causality; for instance, higher prices may induce better maintenance rather than result from it. To partially address this, regional fixed effects were incorporated to control for time-invariant spatial characteristics (Koster & Rouwendal, 2020). Nevertheless, further robustness could be achieved using panel data to better identify causal mechanisms (Malpezzi, 2003).

The chosen log-linear functional form follows standard hedonic practice. It allows interpretation of coefficients as elasticities and ensures comparability with previous studies. Future research could test alternative specifications to capture potential non-linearities, particularly for land area.

Nonetheless, the analysis remains limited by the absence of micro-spatial variables, potential endogeneity between maintenance quality and price, and the cross-sectional nature of the data. As two examples, incorporating neighborhood-level indicators would enhance the modeling of urbanization and accessibility effects, while panel data capture temporal dynamics and long-term price adjustments more effectively. The results describe a market of strong structural fundamentals and steady regional differences, consistent with the late-1990s economic expansion.

In conclusion, the developed hedonic model provides a robust foundation for understanding Dutch housing price determinants but remains constrained by the limits of cross-sectional OLS estimation. Future research should integrate finer spatial and temporal data to capture micro-scale heterogeneity in price. In addition, examining environmental features also offers a promising direction. These considerations motivate two follow-up questions:

- How do neighborhood-level amenities and accessibility factors influence housing prices across Dutch regions?

- What is the effect of environmental and energy-efficient attributes on Dutch property values?

Bibliography

Forouhar, A., & Van Lierop, D. (2021). If you build it, they will change. *The Journal of Transport and Land Use Vol. 14 No.1*, 949–973.

Herath, S., & Maier, G. (2010). *The hedonic price method in real estate and housing market research: a review of the literature.*

Koster, H., & Rouwendal, J. (2020). Household preferences. In N. Verloo, & L. Bertolini, *Seeing the City* (pp. 124-144).

Malpenzzi, S. (2003). Hedonic pricing models: a selective and applied review. In *Housing Economics: Essays in Honor of Duncan Maclennan* (pp. 67-89).

Monson, M. (2009). *Valuation using hedonic pricing models.*

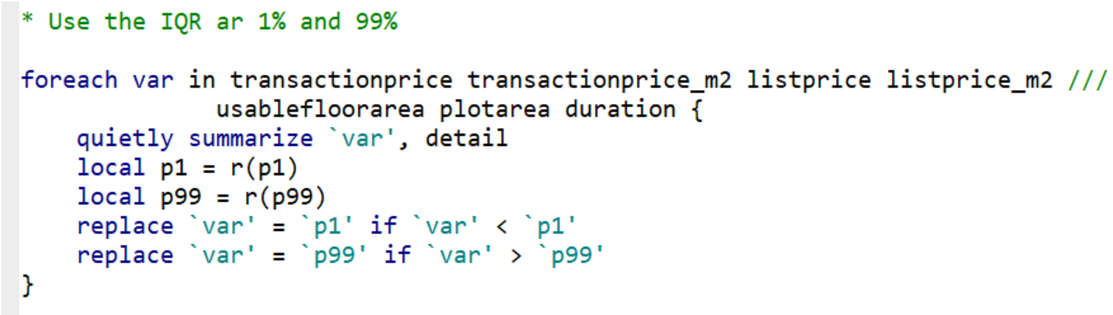
Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy, Vol. 82, No. 1*, 34-55.

Sirmans, S., Macpherson, D., & Zietz, E. (2005). The Composition of Hedonic Pricing Models. *Journal of Real Estate Literature , 2005, Vol. 13, No. 1*, 3-43.

Steegmans, J., & Hassink, W. (2025). Nominal loss aversion and equity constraints in house price determination: Empirical evidence in the absence of down-payment constraints. *Journal of Housing Economics Vol. 69 102084*.

Appendix

Appendix 1: Outlier Remover Loop

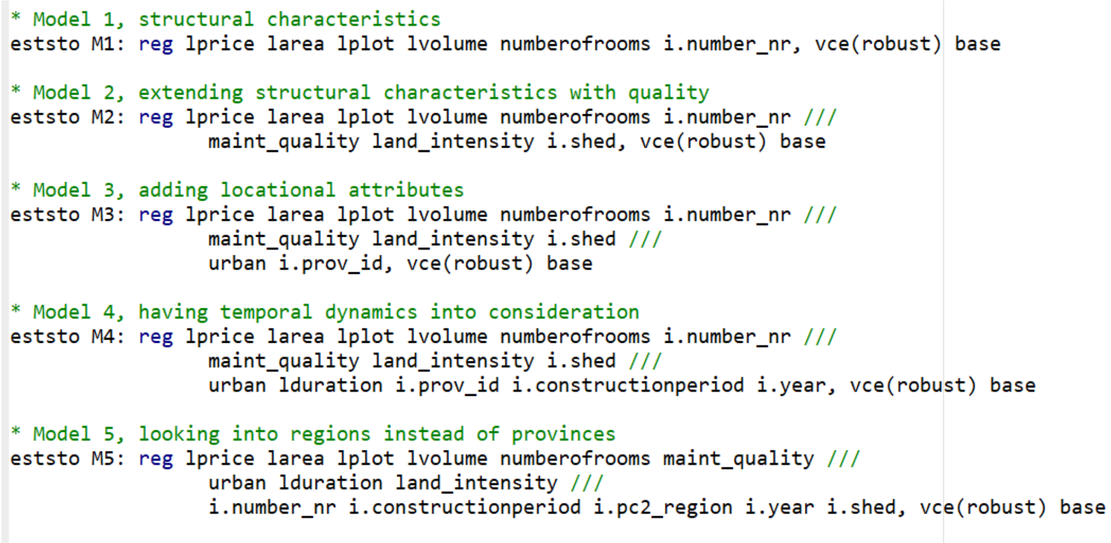


```
* Use the IQR ar 1% and 99%

foreach var in transactionprice transactionprice_m2 listprice listprice_m2 ///
               usablefloorarea plotarea duration {
    quietly summarize `var', detail
    local p1 = r(p1)
    local p99 = r(p99)
    replace `var' = `p1' if `var' < `p1'
    replace `var' = `p99' if `var' > `p99'
}
```

Appendix 1: For Loop Code from Stata aimed to eliminate outliers outside of the 1st and 99th percentiles. Copied from a question's answer on a Stata help forum.

Appendix 2: Code for the OLS regression models



```
* Model 1, structural characteristics
eststo M1: reg lprice larea lplot lvolume numberofrooms i.number_nr, vce(robust) base

* Model 2, extending structural characteristics with quality
eststo M2: reg lprice larea lplot lvolume numberofrooms i.number_nr ///
                 maint_quality land_intensity i.shed, vce(robust) base

* Model 3, adding locational attributes
eststo M3: reg lprice larea lplot lvolume numberofrooms i.number_nr ///
                 maint_quality land_intensity i.shed ///
                 urban i.prov_id, vce(robust) base

* Model 4, having temporal dynamics into consideration
eststo M4: reg lprice larea lplot lvolume numberofrooms i.number_nr ///
                 maint_quality land_intensity i.shed ///
                 urban lduration i.prov_id i.constructionperiod i.year, vce(robust) base

* Model 5, looking into regions instead of provinces
eststo M5: reg lprice larea lplot lvolume numberofrooms maint_quality ///
                 urban lduration land_intensity ///
                 i.number_nr i.constructionperiod i.pc2_region i.year i.shed, vce(robust) base
```

Appendix 2: Stata regression specifications for Models 1–5.

Appendix 3: Table 4 Extended

Variable	**Coefficient (Std. Error)**	**Variable**	**Coefficient (Std. Error)**
Structural Attributes		**Amenities & Quality**	
Log Area (larea)	**0.057***** (0.003)	Maintenance Quality	**0.058***** (0.001)
Log Plot (lplot)	**0.118***** (0.003)	Land Intensity	**0.003***** (0.000)
Log Volume (lvolume)	**0.480***** (0.012)	Shed = 1 (ref.)	0.000 (.)
Number of Rooms (incl. living room)	**0.035***** (0.001)	Shed = 2	**0.013***** (0.002)
Home Type (Terraced, ref.)	0.000 (.)	Shed = 3	**0.008***** (0.004)

Home Type = Linked	**0.105***** (0.004)	Shed = 4	**0.029***** (0.002)
Home Type = Corner	**0.040***** (0.002)	Shed = 5	**0.051***** (0.003)
Home Type = Semi (5)	**0.163***** (0.003)	Shed = 6	**–0.291***** (0.008)
Home Type = 6	**0.265***** (0.004)		
Home Type = 8	–0.005*** (0.009)	**Spatial Variables**	
Home Type = 9	**–0.068***** (0.006)	Urbanization Level (5 = most urban)	**0.010***** (0.001)
Home Type = 10	**0.107***** (0.006)	Log Duration (lduration)	**–0.021***** (0.001)
Provincial Fixed Effects		**Temporal Variables**	
Drenthe (ref.)	0.000 (.)	Year 1996 (ref.)	0.000 (.)
Flevoland	**0.231***** (0.004)	Year 1997	**0.070***** (0.008)
Fryslân	**–0.062***** (0.005)	Year 1998	**0.186***** (0.007)
Gelderland	**0.373***** (0.004)	Year 1999	**0.314***** (0.007)
Groningen	**–0.163***** (0.005)	Year 2000	**0.439***** (0.007)
Limburg	**0.139***** (0.007)		
Noord-Brabant	**0.307***** (0.004)	**Construction Periods**	
Noord-Holland	**0.518***** (0.004)	Period 1 (ref.)	0.000 (.)
Overijssel	**0.180***** (0.004)	Period 2	**–0.059***** (0.005)
Utrecht	**0.554***** (0.003)	Period 3	**–0.023***** (0.005)
Zeeland	**–0.058***** (0.008)	Period 4	**–0.040***** (0.005)
Zuid-Holland	**0.431***** (0.003)	Period 5	**–0.061***** (0.005)
		Period 6	**–0.082***** (0.005)
		Period 7	**–0.032***** (0.005)
		Period 8	**0.042***** (0.005)
Constant	**6.990***** (0.055)	**Observations**	**128,139**

Appendix 3: Extended regression output for the final hedonic model.