Determinants of Housing Prices in New York State: An Empirical Analysis of Property Characteristics, School Quality, and Socioeconomic Factors *

By:

Sum Yee Chan
1007936272
sumyee.chan@mail.utoronto.ca

Cristina Su Lam
1007928566
cristina.sulam@mail.utoronto.ca

ECO375: Applied Econometrics
University of Toronto
Department of Economics
Date: December 5th, 2024

Abstract:

This paper examines the determinants of housing prices in New York State by integrating property characteristics, public school quality, and socioeconomic factors into a comprehensive hedonic pricing model. Using a cross-sectional dataset of 43,394 properties from ZIP codes with public schools, we first establish a baseline regression, finding that each additional bedroom increases housing prices by 21.3%, though this explains only 10.4% of the variability in housing prices. Building on Sirmans et al. (2006) and Mathur (2016), we expand the model to include bathrooms, house size, lot size, student-teacher ratios, number of public schools, median income, and cost of living, achieving predictive power of 79.2%. Key findings indicate that a one-standard-deviation increase in cost of living is associated with a 71.6% rise in housing prices, while each additional bathroom increases property value by 20.8%. This study contributes to the literature by offering a localized analysis at the ZIP code level, addressing gaps in prior research, which predominantly focuses on broader state or national trends, such as the work of Albouy et al. (2016) and Özmen et al. (2019). By incorporating educational metrics, the findings highlight the role of public school resources in shaping housing markets, consistent with theories of residential sorting based on school quality (Kane et al., 2005). Unlike Yan (2022), which adopts a state-level perspective, our ZIP code-specific approach captures localized interactions among structural, educational, and socioeconomic factors. Limitations include potential omitted variable bias and sample selection bias arising from the focus on ZIP codes with public schools, which may limit generalizability to broader contexts. Future research should consider additional variables, broader geographic contexts, and causal frameworks to enhance applicability. These findings underscore the importance of a multifaceted approach for policymakers addressing housing affordability and regional disparities.

*** Replication Package Link:** Project Files and Code

1 Introduction

Housing prices are a cornerstone of economic stability, influencing both individual financial security and regional prosperity. As affordability challenges intensify and disparities widen, understanding the factors that drive housing markets has become increasingly urgent. New York State, with its economic diversity and stark regional contrasts, provides an ideal setting to explore how localized factors—such as property characteristics, educational quality, and socioeconomic conditions—interact to influence housing demand and prices. By uncovering these interconnections, this research highlights the forces driving housing accessibility and economic disparities within the state.

Motivated by gaps in the existing literature, such as the focus on broader state- or national-level trends in work by Albouy et al. (2016) and Özmen et al. (2019), this study extends prior research by examining housing price determinants at a localized ZIP code level in New York State, focusing on properties within areas that include public schools. While foundational research, including Sirmans et al. (2006), has established the importance of structural characteristics such as bedroom count, our paper incorporates additional dimensions, including public school quality and regional socioeconomic factors. It hypothesizes that housing prices are influenced by a combination of structural attributes and contextual factors, including income distribution and cost of living. A detailed discussion of prior studies is provided in Section 1.1.

The analysis begins with a baseline model using log-transformed housing prices as the dependent variable and bedroom count as the key independent variable, reflecting its consistent role as a robust predictor of property value. The model is then systematically expanded to include additional structural features, educational metrics, and socioeconomic conditions. Significant findings from the final model indicate that factors such as house size, cost of living, and public school resources notably impact housing price variability. By integrating these layers of complexity, the approach demonstrates a substantial improvement in explanatory power, offering nuanced insights into the localized dynamics of housing markets.

The paper is structured as follows: Section 1.1 reviews foundational housing price determinants, Section 2 outlines data sources and methods, Section 3 presents the regression model and results, Section 4 discusses limitations, and Section 5 concludes.

1.1 Literature Review

Property characteristics play a fundamental role in determining housing prices, serving as the foundation for hedonic pricing models that assess property values by isolating the impact of individual attributes. Sirmans et al. (2006) highlight the significance of

structural features like bedroom and bathroom counts, lot size, and square footage, noting their consistent inclusion in hedonic models across regions and time. Among these, bedroom count emerges as a robust predictor, especially in baseline models that capture fundamental property values. While bedroom count is often less sensitive to regional variations, controlling for related variables like bathroom count or square footage enhances the precision of its coefficient.

The role of structural characteristics is complemented by the influence of public school quality, which has long been recognized as a critical driver of housing prices. Mathur (2016) demonstrates that proximity to high-performing schools significantly increases property values, reflecting the premium placed on educational opportunities by homebuyers. This relationship underscores the importance of including school-related variables—such as student-teacher ratios, school density, and eligibility for free lunch—in housing price models. Kane et al. (2005) extend this analysis by examining how changes in school assignments influence housing prices through residential sorting and demographic shifts. Their study reveals that school quality impacts housing prices indirectly, as families cluster in neighborhoods with access to better schools, further justifying the inclusion of ZIP code fixed effects to control for unobserved location-specific factors.

In addition to structural and educational factors, socioeconomic conditions play a significant role in shaping housing demand. Yan (2022) analyzes housing prices in New York State, emphasizing the importance of economic factors such as cost of living and income in shaping housing demand. While her study adopts a state-level perspective, it underscores the value of integrating these variables to understand affordability dynamics. Extending this analysis, Albouy et al. (2016) highlight how rising housing costs and income disparities disproportionately burden lower-income households, while Özmen et al. (2019) demonstrate that income distribution impacts housing demand, with lower-income shares boosting demand and higher shares dampening it due to supply mismatches.

This study builds on prior research by starting with bedroom count as the primary independent variable in a simple regression model, guided by the findings of Sirmans et al. (2006), which emphasize its consistent role in hedonic pricing models. It then integrates educational metrics, drawing on Mathur (2016) and Kane et al. (2005), to examine the impact of school resources and accessibility on housing markets in New York State. Lastly, ZIP code-specific measures of income, cost of living, and income distribution are added, extending the work of Albouy et al. (2016) and Özmen et al. (2019). This approach offers a comprehensive analysis of how structural, educational, and socioeconomic factors influence housing price variability.

2 The Context and Data

To investigate the determinants of housing prices in New York State, this analysis uses a cross-sectional dataset compiled from three primary sources: USA Real Estate, Public School Characteristics 2021-22, and City/ZIP/County/FIPS - Quality of Life. These datasets collectively provide detailed information on housing, educational, and socioeconomic factors, primarily from 2022, ensuring temporal consistency. Following extensive preprocessing and data cleaning, the final dataset includes 43,394 observations.

The USA Real Estate Dataset (Sakib, 2022), retrieved from Kaggle, contains property-level information collected through web scraping from Realtor.com, including attributes such as property price, number of bedrooms and bathrooms, lot size, and house size. Public school data were sourced from the National Center for Education Statistics (NCES) (2024), capturing educational metrics such as student-teacher ratios and the number of public schools. Socioeconomic variables, including median income and cost of living, were derived from the City/ZIP/County/FIPS - Quality of Life Dataset (Vaughan, 2023), also retrieved from Kaggle.

To ensure consistency and reliability, several preprocessing steps were undertaken. Data from other states were removed to focus on New York State. Observations with missing values in relevant variables, such as property price, house size, or student-teacher ratios, were excluded. Filters were applied to remove outliers, including properties priced below \$60,000 or above \$20 million USD, houses larger than 10,000 ft^2, and those with more than 15 bedrooms or 12 bathrooms. Additional exclusions included student-teacher ratios above 20 and lots exceeding 60 acres. Continuous variables—such as property price, house size, lot size, and the number of students eligible for free lunch—were log-transformed to address skewness and improve interpretability. Median income, cost of living, house size, and student-teacher ratios were standardized for cross-unit comparisons. Finally, the three datasets were merged using ZIP codes as a common key, integrating housing, educational, and socioeconomic attributes. By focusing on properties in areas with public schools, the final dataset provides a robust foundation for analysis.

2.1 Descriptive Statistics

The final dataset highlights significant variability in housing, educational, and socioeconomic attributes across New York State, as summarized in Table 1, underscoring the complexity of the housing market. The average property price is \$615,186, with a substantial standard deviation (SD) of \$1,082,536, reflecting the wide range of housing options, from affordable homes to luxury estates. The original distribution of property

prices was highly skewed due to extreme outliers, particularly at the upper end of the market. To mitigate the influence of outliers, property prices were log-transformed. As shown in Figure 1, the log-transformation brings the distribution closer to normality, although slight left-skewness and potential bimodality remain, indicating regional or structural variations that may influence housing prices.

Housing characteristics reveal a mix of property types, with an average house size of 2,025 ft^2 (SD: 1,121 ft^2). The dataset includes properties with an average of 3.60 bedrooms (SD: 1.42) and 2.43 bathrooms (SD: 1.25). Educational attributes also exhibit disparities, with ZIP codes averaging 6.20 public schools (SD: 4.63) and a student-teacher ratio of 11.78 (SD: 1.95), which could shape housing demand in areas with access to better schools.

Socioeconomic conditions further illustrate differences across ZIP codes. The average median household income is $89,361 (SD: $20,724), and the average cost of living is $96,834 (SD: $19,841). Higher-income areas with elevated costs of living often attract wealthier buyers, driving up housing prices in those regions. These descriptive statistics provide essential context for understanding how structural, educational, and socioeconomic factors interact to shape housing prices.

3 Regression Analysis

This section builds on the prior exploration of structural, educational, and socioeconomic factors influencing housing prices by employing regression analysis to quantify these relationships. All regression analyses were conducted using StataCorp LLC (2023), which employs robust standard errors by default, addressing potential heteroscedasticity and ensuring reliable inference even in the presence of non-constant variance in residuals. Starting with a baseline model, the analysis evaluates the association between log-transformed housing prices (*log(Price)*) and the number of bedrooms (*Bed*) as a primary structural characteristic. Successive models incorporate additional predictors to account for housing attributes, educational factors, and socioeconomic conditions. The results of the regression models, summarized in Table 2, provide insights into the relative importance and statistical significance of these factors in explaining variations in housing prices across New York State.

3.1 Simple Linear Regression

$$\log(\text{Price}_i) = \beta_0 + \beta_1 \text{Bed}_i + \epsilon_i \tag{1}$$

The baseline model, Equation (1), shown in Column (1) of Table 2, estimates the relationship between *log(Price)* and *Bed*, capturing the direct effect of a structural feature

on property values. As hypothesized in Section 1, the number of bedrooms is expected to serve as a robust predictor of housing prices in hedonic pricing models. This simple regression provides a foundational framework for subsequent analyses, allowing clearer interpretation of how housing prices vary with changes in this single characteristic.

The regression results indicate that an additional bedroom is associated with a 21.3% increase in housing prices, with statistical significance at the 1% level ($\beta_1 = 0.213$, $p < 0.01$). The model explains approximately 10.4% of the variation in housing prices ($R^2 = 0.104$). While this is a relatively modest proportion, it reflects the influence of bedrooms as a single factor, without accounting for other potential determinants. The residual variation underscores the importance of incorporating additional housing, educational, and socioeconomic variables in subsequent models.

The OLS estimator for this model is unbiased and efficient if the Least Squares Assumptions (LSA)—linearity, random sampling, and homoscedasticity—are satisfied. However, Figure 2, which displays a histogram of the residuals from Equation (1) overlaid with a normal density curve, reveals slight deviations from normality, including a bimodal tendency and asymmetry in the tails. These patterns suggest that the model may be under-specified, likely due to omitted variables, underscoring the need to incorporate additional variables to enhance model reliability and explanatory power.

3.2 Multiple Regression Analysis

$$\begin{aligned}\log(\text{Price}_i) = {} & \beta_0 + \beta_1 \text{Bed}_i + \beta_2 \text{Bath}_i + \beta_3 \log(\text{House Size}_i) \\ & + \beta_4 \log(\text{Acre Lot}_i) + \beta_5 \text{Total Schools}_i + \beta_6 \text{Student-Teacher Ratio}_i \\ & + \beta_7 \log(\text{Total Free Lunch}_i) + \beta_8 \text{Median Income}_i \\ & + \beta_9 \text{Cost of Living}_i + \beta_{10} \left(\log(\text{House Size}_i) \times \log(\text{Acre Lot}_i)\right) + \mu_j \quad (2)\end{aligned}$$

To address the baseline model's limitations, the analysis incorporates structural, educational, and socioeconomic variables to account for additional factors influencing housing prices. The final model, Equation (2), shown in Column (7) of Table 2, builds incrementally by adding predictors, such as structural features like bathroom count and property size (Columns 2-4), followed by educational metrics (Column 5), and finally socioeconomic variables and interaction terms (Columns 6-7), to enhance explanatory power and reduce omitted variable bias (OVB).

Equation (2) demonstrates a marked improvement in model fit, explaining 79.2% ($R^2 = 0.792$) of the variation in *log(Price)*, compared to 10.4% in the baseline model ($R^2 = 0.104$). Key predictors with high economic and statistical significance include standardized cost of living ($\beta_9 = 0.716$, $p < 0.01$), suggesting that a one-standard-

deviation increase in cost of living corresponds to a 71.6% increase in housing prices, and bathroom count ($\beta_2 = 0.208$, $p < 0.01$), indicating that each additional bathroom is associated with an approximate 20.8% increase in housing prices. The interaction term, *log(House Size)* $\times$ (*Acre Lot*) ($\beta_1 0 = 0.032$, $p < 0.01$), reveals that the relationship between house size and property value is amplified for larger lots, emphasizing the complementary effect of land area on housing price appreciation. Additionally, the inclusion of ZIP code fixed effects helps control for unobserved regional heterogeneity, such as neighborhood amenities or local policy differences. Although the coefficients lack consistent economic significance, as shown in Table 2, their inclusion enhances internal validity by reducing OVB and mitigating potential endogeneity arising from unobserved factors.

Figure 3 illustrates the histogram of residuals for Equation (2), which closely aligns with a normal distribution, addressing concerns raised in the baseline model's residuals as seen in Figure 2. The improved normality and reduced asymmetry indicate the model's ability to account for previously omitted variables, thereby satisfying the LSA and supporting the validity of hypothesis testing. Nonetheless, after controlling for structural, educational, and socioeconomic variables, the treatment can be considered as good as randomly assigned, further enhancing the model's reliability in capturing the determinants of housing prices in New York State. Despite its robustness and strong theoretical foundation, Equation (2) is not without limitations, such as potential functional form misspecification and multicollinearity among covariates, which are discussed further in section 4.

4 Limitations of Results

While this study provides valuable insights into the determinants of housing prices, several limitations warrant attention. External validity refers to the extent to which the conclusions of a study can be generalized to other contexts, populations, or time periods. By focusing solely on properties in New York State ZIP codes with public schools, this ensures consistency in educational metrics, but it excludes properties from ZIP codes without public schools and from regions with differing educational and socioeconomic contexts. As a result, the findings may lack external validity, limiting their generalizability to broader housing markets beyond the study region.

Although Equation (2) incorporated ZIP code fixed effects and additional predictors to address OVB, some limitations persist. In Table 2, the inconsistent economic significance of ZIP code fixed effects across models suggests that unobserved heterogeneity—differences in factors like neighborhood safety or local amenities that vary across ZIP codes but are not captured in the model—remains, and certain omitted factors are

inadequately controlled. This shortfall in accounting for such contextual differences contributes to residual variation, which also raise the possibility of simultaneous causality, as variables like local public school quality and housing prices may influence one another, further complicating causal interpretation. While robust standard errors address heteroscedasticity, they do not resolve challenges related to internal validity, including feedback loops between predictors like public school quality and housing prices, which complicate causal interpretation. These concerns underscore the need to interpret the results as descriptive rather than strictly causal.

The reliability of the results could also be affected by functional form misspecification. While linear and log-linear transformations improve interpretability, they may not fully capture nonlinear relationships or complex interactions. For instance, the interaction term between house size and lot size addresses some subtleties, but other relationships may remain unexplored. Additionally, multicollinearity among predictors, particularly between house size and lot size, could compromise the precision of coefficient estimates. The histogram of residuals for Model 7 (Figure 3) shows improved normality, but slight asymmetry and deviations persist, indicating the model may still be under-specified.

In summary, while the models adhere to key econometric principles and achieve high explanatory power, limitations such as potential OVB, functional form misspecification, and multicollinearity highlight the need for cautious interpretation. Future research could address these issues by employing instrumental variables, expanding datasets to additional regions, or incorporating richer variables to enhance internal and external validity.

5 Conclusion

This study examined the determinants of housing prices in New York State, starting with a baseline model that identified the number of bedrooms as a key predictor of property values. By progressively incorporating additional structural characteristics, educational metrics, and socioeconomic variables, the analysis revealed that factors such as house size, cost of living, and public school resources significantly influence housing price variability. The final model demonstrated a substantial improvement in explanatory power, highlighting the complex interplay of localized factors in shaping housing markets. These findings underscore the importance of considering a multifaceted set of determinants when analyzing housing prices, offering nuanced insights into the dynamics of housing accessibility and economic disparities within the state. Future research could build upon this work by addressing the limitations identified and exploring these relationships in other regions to enhance generalizability.

References

Albouy, D., G. Ehrlich, and Y. Liu (2016, November). Housing demand, cost-of-living inequality, and the affordability crisis. *National Bureau of Economic Research*.

Kane, T. J., D. O. Staiger, and S. K. Riegg (2005, May). School quality, neighborhoods and housing prices: The impacts of school desegregation. *National Bureau of Economic Research*.

Mathur, S. (2016). The myth of 'free' public education: Impact of school quality on house prices in the fremont unified school district. *Journal of Planning Education and Research 37*(2), 176–194.

National Center for Education Statistics (NCES) (2024). Education Demographic and Geographic Estimate (EDGE) Program - Public School Characteristics 2021-22. Last modified October 21, 2024.

Sakib, A. S. (2022). Usa real estate dataset. Accessed April 1, 2022.

Sirmans, G. S., L. MacDonald, D. A. Macpherson, and E. N. Zietz (2006). The value of housing characteristics: A meta analysis. *The Journal of Real Estate Finance and Economics 33*(3), 215–240.

StataCorp LLC (2023). *Stata Statistical Software: Release 18*. College Station, TX. Available at https://www.stata.com.

Vaughan, Z. (2023). City/zip/county/fips - quality of life (us). Accessed December 1, 2024.

Yan, Y. (2022, January). Influencing factors of housing price in new york-analysis: Based on excel multi-regression model. In *Proceedings of the International Conference on Big Data Economy and Digital Management*.

Özmen, M. U., C. Yalçın, and E. Yücel (2019). Income distribution and house prices: Evidence from turkey. *Borsa Istanbul Review 19*(3), 258–271.

6 Tables and Figures

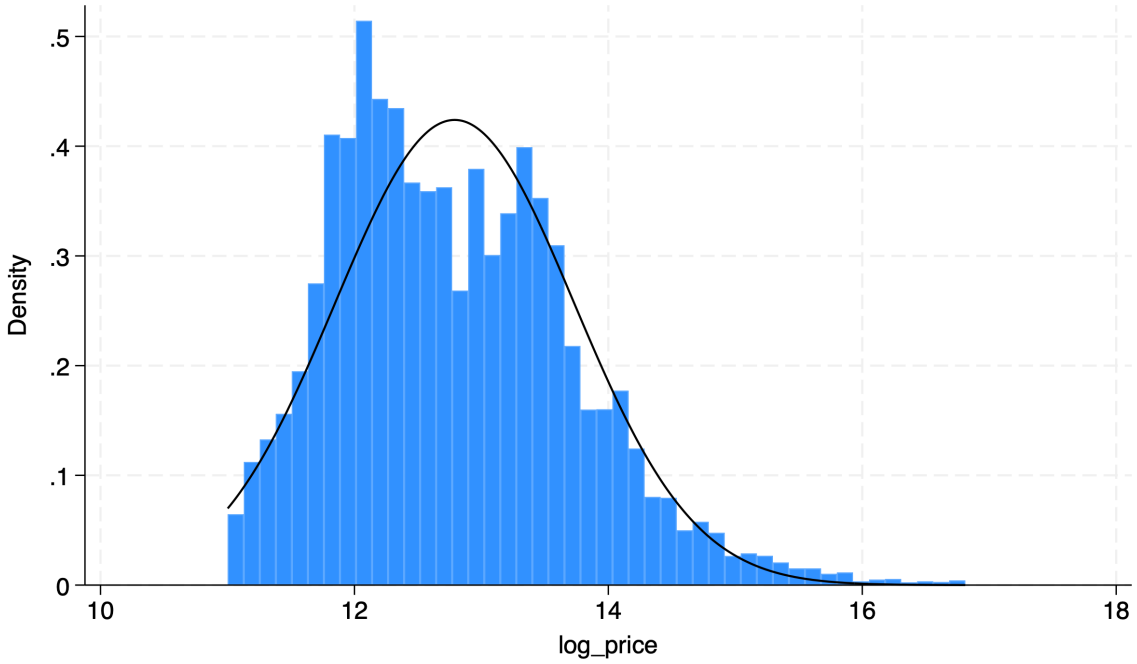

Figure 1: Distribution of Log Price

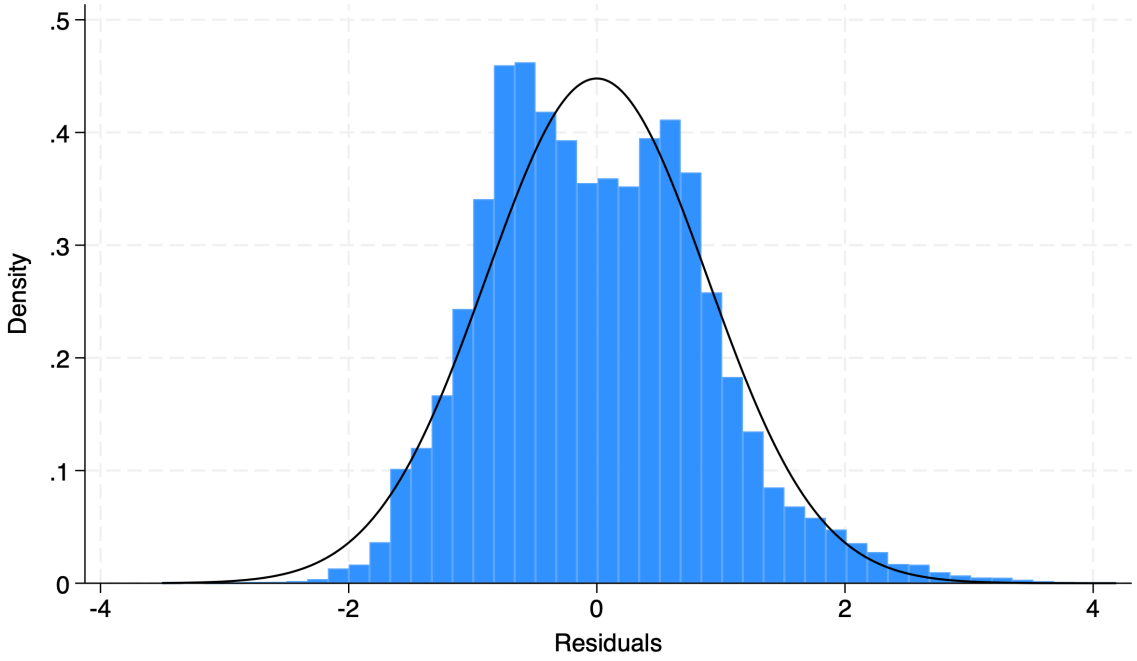

Figure 2: Histogram of Residuals (Model 1) with Normal Curve

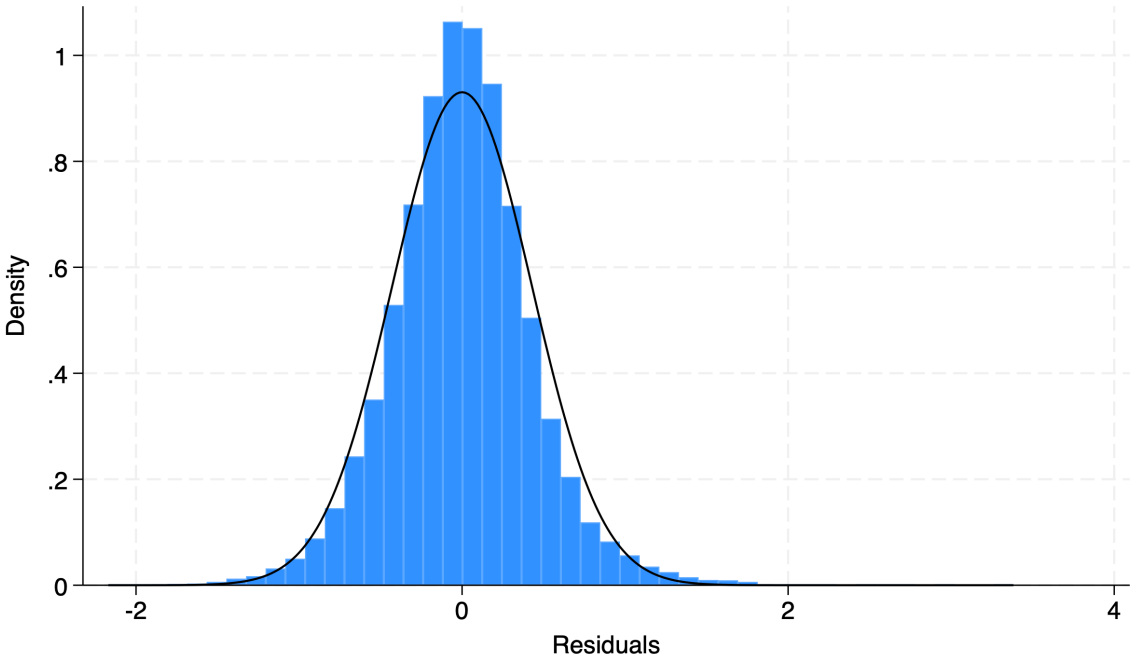

Figure 3: Histogram of Residuals (Model 7) with Normal Curve

Table 1: Summary Statistics of Numerical Variables

Variable	Obs	Mean	Std. Dev.	Min	Max
Price	43,394	615,186.1	1,082,536	60,000	20,000,000
Bed	43,394	3.59	1.42	1	15
Bath	43,394	2.43	1.25	1	12
Acre Lot	43,394	1.17	4.10	0.01	60
House Size	43,394	2,025.35	1,120.87	122	10,000
Total Free Lunch	43,394	1,780.41	2,024.24	3	11,548
Total Schools	43,394	6.20	4.63	1	34
Cost of Living	43,394	96,834.18	19,841.24	67,463.92	137,874.7
Median Income	43,394	89,360.55	20,723.90	48,566.82	139,281.8
Avg Student-Teacher Ratio	43,394	11.78	1.95	3.29	19.63

Note: Summary statistics include the mean, standard deviation, minimum, and maximum values for key variables. Monetary values are in USD, and log transformations were applied to mitigate skewness where indicated. Observations reflect properties in ZIP codes with public schools in New York State, following data cleaning to exclude extreme or missing values in key variables (e.g., properties with prices below $60,000 or above $20 million).

Table 2: Results of Regression Models for Housing Prices

	(1) Log Price	(2) Log Price	(3) Log Price	(4) Log Price	(5) Log Price	(6) Log Price	(7) Log Price
Bed	0.213***	-0.072***	-0.096***	-0.111***	-0.106***	-0.067***	-0.060***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)
Bath		0.415***	0.425***	0.406***	0.404***	0.355***	0.208***
		(0.005)	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)
Std House Size		0.154***					
		(0.006)					
Std Acre Lot		0.000					
		(0.004)					
Log House Size			0.394***	0.439***	0.426***	0.349***	0.505***
			(0.013)	(0.012)	(0.012)	(0.011)	(0.008)
Log Acre Lot			-0.063***	-0.018***	-0.028***	-0.005*	-0.172***
			(0.003)	(0.003)	(0.003)	(0.002)	(0.021)
Total Schools				0.046***	0.057***	0.052***	0.021***
				(0.001)	(0.001)	(0.001)	(0.001)
Std Avg Student-Teacher					0.024***	0.002	-0.023***
					(0.004)	(0.003)	(0.002)
Log Total Free Lunch					-0.066***	-0.012**	-0.072***
					(0.005)	(0.004)	(0.003)
Std Median Income						0.300***	-0.234***
						(0.003)	(0.003)
Log House Size × Log Acre Lot							0.032***
							(0.003)
Std Cost of Living							0.716***
							(0.005)
Zip Code							-0.000***
							(0.000)
Constant	12.022***	12.041***	9.066***	8.599***	9.059***	9.301***	9.921***
	(0.012)	(0.014)	(0.085)	(0.082)	(0.088)	(0.080)	(0.068)
R^2	0.104	0.386	0.393	0.439	0.442	0.536	0.792
N	43394	43394	43394	43394	43394	43394	43394

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Dependent variable: Log-transformed property price. Standard errors are reported in parentheses, with statistical significance denoted as follows: ***p ¡ 0.01, **p ¡ 0.05, *p ¡ 0.1. Models sequentially incorporate variables, beginning with structural characteristics in Model 1 and adding educational (Model 4) and socioeconomic factors (Model 6-7). Interaction terms and ZIP code fixed effects (Model 7) capture regional heterogeneity. Preprocessing steps ensured the reliability of key variables, with outliers and missing values excluded as detailed in the text.