Distance to Schools and House Prices

Hypotheses & Regression Strategy

Your Name

Distance to Schools and House Prices: Hypotheses & Regression Strategy

H1: House prices are higher closer to schools *(negative coefficients on distance variables)*

Empirical framework

- **Outcome:** $\ln(\text{Price})$
- **Key variables:** distance to nearest **primary** school; distance to nearest **secondary** school
- **Strategy:** estimate **continuous**, **quadratic**, and **binned** distance gradients

Baseline OLS controls

Standard housing attributes

- Living area
- Plot area
- Number of rooms
- House age

Distance designs

Alternative specifications

- **Continuous:**
  - include quadratic term (km^2)
- **Binned**
  - 0–3 km, 3–6 km, 6–9 km, $>$ 9 km

Regression Specifications

1. Naive (Unconditional)

$$\ln(P) = \alpha + \beta_1 D^{prim} + \beta_2 D^{sec} + \varepsilon$$

- **Goal:** Establish raw spatial correlation.
- **Risk:** High Omitted Variable Bias.

2. Baseline (Preferred)

$$\ln(P) = \alpha + \mathbf{D} + \mathbf{X} + \varepsilon$$

- **Controls ($\mathbf{X}$):** Area, Plot, Rooms, Age.
- **Hypothesis:** $\beta < 0$ (Negative semi-elasticity).

3. Non-Linearity (Polynomial)

$$\ln(P) = \text{Base} \cdots + \delta(D)^2 + ...$$

- **Logic:** Test for convex/concave decay.
- **Test:** $\delta \neq 0$ implies varying MWTP.

4&5. Robustness (Single-Type)

$$\ln(P) = \alpha + \beta_k D^k + \mathbf{X} + \varepsilon$$

- **Method:** Estimate Primary/Secondary separately.
- **Check:** Stability against multicollinearity.

Descriptive Evidence: Price-Distance Gradient

Negative Slope (H1)
Clear decay in prices as distance increases across specifications.

Non-Linearity
Steeper drop in the first 3-6 km, flattening afterwards.

Heterogeneity
Primary schools show a visibly steeper gradient.

Fig 1. Continuous Trend (Raw Scatter)

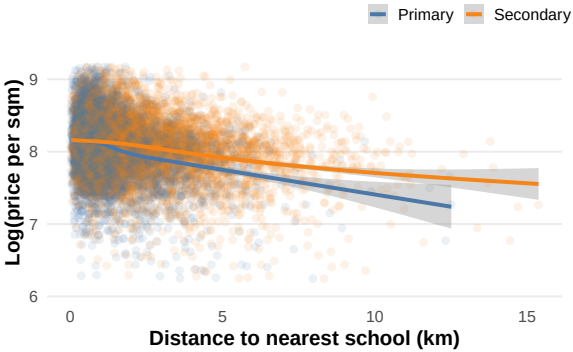
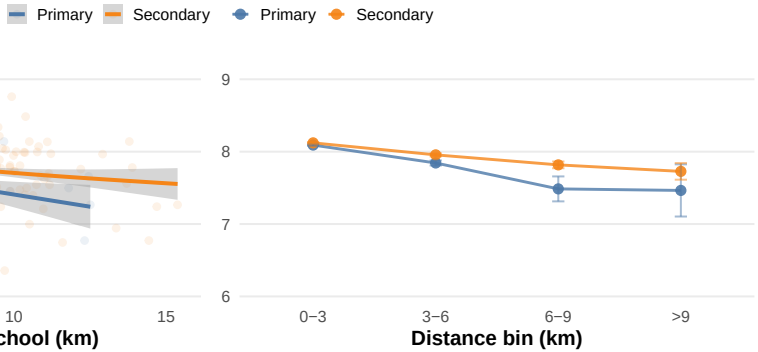


Fig 2. Discrete Gradient (Binned Means)



Main Results: Continuous Model

	Naive (Both)	Baseline (Both)	Polynomial (Both)	Baseline (Primary)	Baseline (Secondary)
Distance to primary school (km)	−0.070***	−0.071***	−0.107***	−0.103***	
Distance to primary school (km)^2			0.007***		
Distance to secondary school (km)	−0.031***	−0.034***	−0.023***		−0.057***
Distance to secondary school (km)^2			−0.001		
Log living area		0.880***	0.879***	0.886***	0.888***
Log plot area		0.053***	0.054***	0.046***	0.042***
Rooms		−0.012***	−0.012***	−0.012***	−0.012***
House age		−0.002***	−0.002***	−0.002***	−0.002***
Intercept	13.327***	8.677***	8.688***	8.649***	8.661***
Num.Obs.	5794	5794	5794	5794	5794
R2	0.060	0.436	0.438	0.425	0.419
R2 Adj.	0.060	0.436	0.437	0.425	0.418

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Main Results: Binned Model

	Naive (Both)	Baseline (Both)	Baseline (Primary)	Baseline (Secondary)
(Intercept)	13.226***	8.594***	8.571***	8.600***
dist_primary_bin3-6	−0.147***	−0.146***	−0.268***	
dist_primary_bin6-9	−0.361***	−0.369***	−0.577***	
dist_primary_bin>9	−0.833***	−0.538***	−0.793***	
dist_secondary_bin3-6	−0.142***	−0.146***		−0.179***
dist_secondary_bin6-9	−0.241***	−0.256***		−0.331***
dist_secondary_bin>9	−0.209***	−0.294***		−0.435***
log_area		0.887***	0.899***	0.891***
log_plot_area		0.042***	0.031***	0.038***
zimmeranzahl		−0.011***	−0.012***	−0.011***
house_age		−0.002***	−0.002***	−0.002***
Num.Obs.	5794	5794	5794	5794
R2	0.046	0.420	0.401	0.413
R2 Adj.	0.045	0.419	0.400	0.412

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Compare: Continuous vs Binned

	Baseline (Continuous)	Baseline (Binned)
Distance to primary school (km)	−0.071***	
Distance to secondary school (km)	−0.034***	
Primary: 3–6 km (ref: 0–3)		−0.146***
Primary: 6–9 km (ref: 0–3)		−0.369***
Primary: >9 km (ref: 0–3)		−0.538***
Secondary: 3–6 km (ref: 0–3)		−0.146***
Secondary: 6–9 km (ref: 0–3)		−0.256***
Secondary: >9 km (ref: 0–3)		−0.294***
Num.Obs.	5794	5794
R2	0.436	0.420
R2 Adj.	0.436	0.419

Note:

Binned coefficients are relative to the reference group: 0-3 km. Controls included but omitted from display.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Model comparison & robustness (main models)

Continuous fits slightly better;

binned is slightly more robust to influential points.

Model comparison

Table 4. Model comparison
In-sample fit and 5-fold out-of-sample performance

Model	N	R^2	Adj. R^2	AIC	BIC	CV RMSE	CV MAE	CV R^2
Baseline (Continuous)	5794	0.436	0.436	4973.8	5027.1	0.371	0.288	0.434
Baseline (Binned)	5794	0.420	0.419	5143.6	5223.6	0.377	0.293	0.416

In-sample: higher R^2/Adj. R^2 is better; lower AIC/BIC is better.
Out-of-sample (5-fold CV): lower RMSE/MAE is better; higher R^2 is better.

Model comparison & robustness (main models)

Continuous fits slightly better;

binned is slightly more robust to influential points.

Robustness checks

Table 5. Robustness checks (main models)

Diagnostics summary (heteroskedasticity, multicollinearity, influence)

Model	BP p-value	Max VIF	Influential (D>4/n)	Max Cook's D
Good Reference	<0.001	27.57	315	0.033
Average Reference	<0.001	109.13	315	0.033
Bad Reference	<0.001	578.70	315	0.033

Inference uses **HC1 robust SE**.

Lower BP p-values indicate heteroskedasticity; higher VIF indicates collinearity; larger Cook's D indicates influential points.

Economic Mechanisms & Interpretations

1. Capitalization

Why are houses closer to schools more expensive?

- **Commuting Costs:** Time saved is directly capitalized into property values
- **Residential Sorting:** Families cluster here, increasing local demand

2. Primary vs Secondary

Why is the effect stronger for primary schools?

- **Age Constraints:** Young children need escorting
- **Community Ties:** Primary schools define local neighborhoods

3. Non-Linearity

Why does the gradient flatten beyond 3km?

- **Walkability Premium:** Only exists within $<$3km distance
- **Substitution:** Public transit replaces driving at longer distances

4. Limitation

What is missing from this analysis?

- **Distance $\neq$ Quality:** Proximity captures accessibility, not desirability
- **Next Step:** Incorporate **Social Index** for school quality