Owner-occupied dwellings and imputed rent with Hedonic Regression

Asadh 13-14, 2081

Dr. Anil Shrestha
Undersecretary (Account)
Financial Administration Section
National Statistics Office

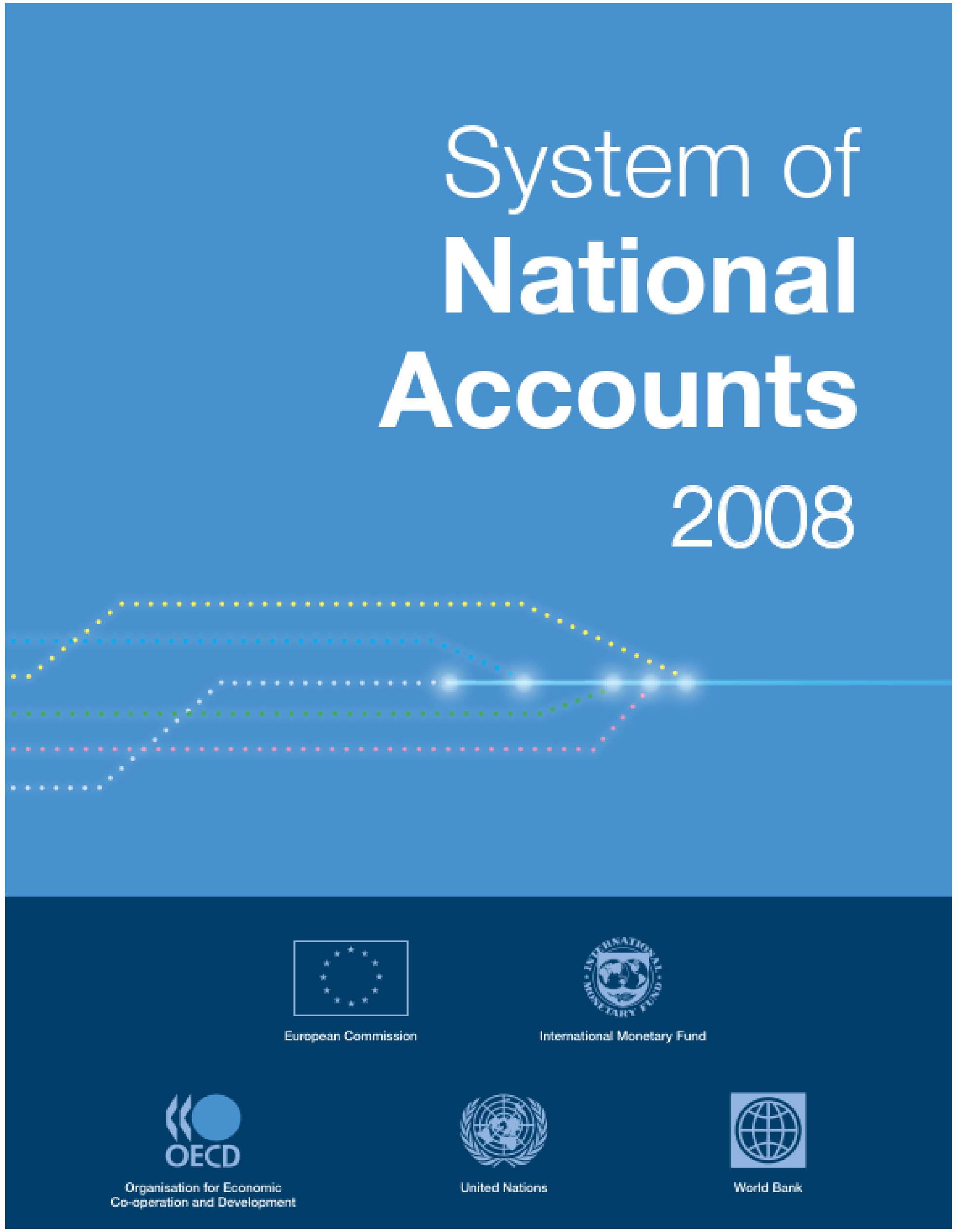
Owner-occupied Dwelling

What is Owner-occupied dwellings?

- Housing units owned and lived in by the occupant.

Treatment in the SNA

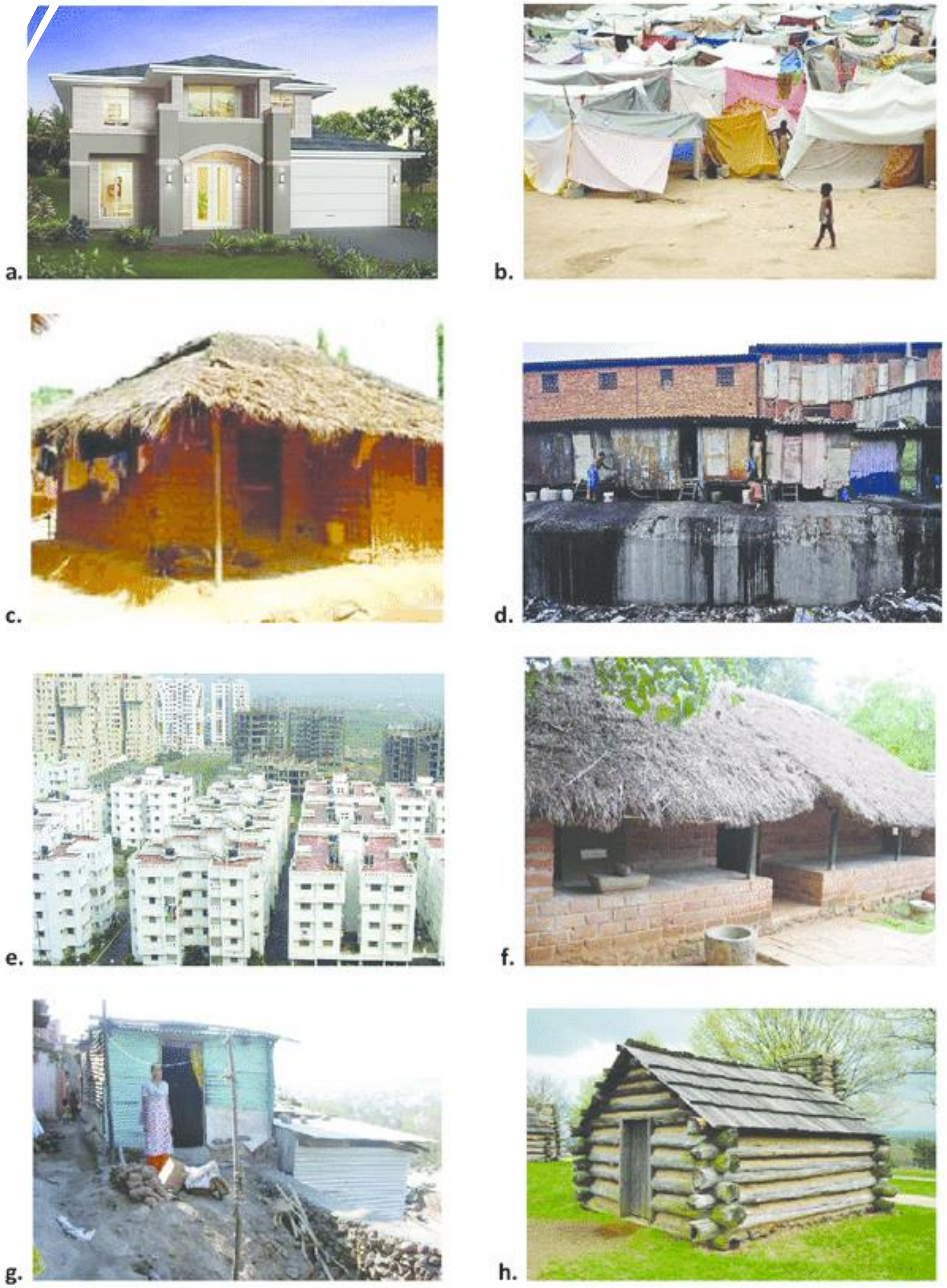
- Not directly included in GDP as it is not considered as part of the market activity.
- Instead, SNA utilizes estimated rental value (imputed rent) that the owner would have to pay if they were to rent a similar property in the market.
- As per SNA, the imputed rent is recorded in both consumption and production of housing services.



Hedonic Regression for imputed rent

What is Hedonic Regression?

It’s a technique to estimate prices of an asset/a product/a service based on its underlying characteristics using an OLS model.



Basic Hedonic regression model

$$P_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_k X_{ik} + \epsilon_i$$

Explanation of Terms:

- P_i: The dependent variable, which is the price or rent of the property.
- β_0: The intercept term, representing the baseline price or rent when all other characteristics are zero.
- $\beta_1, \beta_2, \ldots, \beta_k$: The coefficients that measure the contribution of each characteristic to the price or rent. These are the parameters estimated by the regression.
- $X_{i1}, X_{i2}, \ldots, X_{ik}$: The independent variables representing the characteristics of the property (e.g., size, location, number of bedrooms, age).
- ϵ_i: The error term, capturing the variation in price or rent that is not explained by the model.

Example 1:

For a simplified model, consider three characteristics: solar, city, and garden. The Hedonic regression model would look like this:

$$P_i = \beta_0 + \beta_1(solar_i) + \beta_2(city_i) + \beta_1(garden_i) + \varepsilon_i$$

This equation suggests that the price or rent of a property is determined by whether it has solar, its location (city or rural), and has a garden, with each characteristic contributing a specific amount to the overall price or rent.

Example housing price data

	house_no	price	solar	city	garden
1	1	70000	1	1	0
2	2	60000	0	1	0
3	3	100000	0	0	1
4	4	100000	0	0	1
5	5	50000	0	0	0

All the data, example codes, and do files are provided in https://s.anilz.net/training

Coefficient and its interpretation

```
. reg price solar city garden

      Source |       SS           df       MS      Number of obs   =         5
-------------+----------------------------------   F(3, 1)         =         .
       Model |  2.1200e+09         3   706666667   Prob > F        =         .
    Residual |           0         1           0   R-squared       =    1.0000
-------------+----------------------------------   Adj R-squared   =    1.0000
       Total |  2.1200e+09         4   530000000   Root MSE        =         0

------------------------------------------------------------------------------
       price | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       solar |      10000          .        .       .            .           .
        city |      10000          .        .       .            .           .
      garden |      50000          .        .       .            .           .
       _cons |      50000          .        .       .            .           .
------------------------------------------------------------------------------
```

What if price for the house no 4 changed to 120,000

	house_no	price	solar	city	garden
1	1	70000	1	1	0
2	2	60000	0	1	0
3	3	100000	0	0	1
4	4	120000	0	0	1
5	5	50000	0	0	0

```
. reg price solar city garden

      Source |       SS           df       MS      Number of obs   =         5
-------------+----------------------------------   F(3, 1)         =      5.33
       Model |  3.2000e+09         3  1.0667e+09   Prob > F        =    0.3058
    Residual |   200000000         1   200000000   R-squared       =    0.9412
-------------+----------------------------------   Adj R-squared   =    0.7647
       Total |  3.4000e+09         4   850000000   Root MSE        =     14142

------------------------------------------------------------------------------
       price | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       solar |      10000      20000     0.50   0.705    -244124.1    264124.1
        city |      10000      20000     0.50   0.705    -244124.1    264124.1
      garden |      60000   17320.51     3.46   0.179    -160077.9    280077.9
       _cons |      50000   14142.14     3.54   0.175    -129692.9    229692.9
------------------------------------------------------------------------------
```

Designing Hedonic pricing model

1. Variables selection

- Potential variables are identified based on recent literatures, theories, availability of data.
- Visualizing dependent and independent variables to identify a relationship between them.

Let's use (KIELMC.dta) dataset from the following paper:

Wooldridge Source: K.A. Kiel and K.T. McClain (1995), "House Prices During Siting Decision Stages: The Case of an Incinerator from Rumor Through Operation," Journal of Environmental Economics and Management 28, 241-255.

- From KIELMC.dta, we consider the following variables

year: 1978 or 1981 (initial discussion and rumors in 1978, actual construction began in 1981)

age: age of house

price: house price

rooms: number of rooms in the house

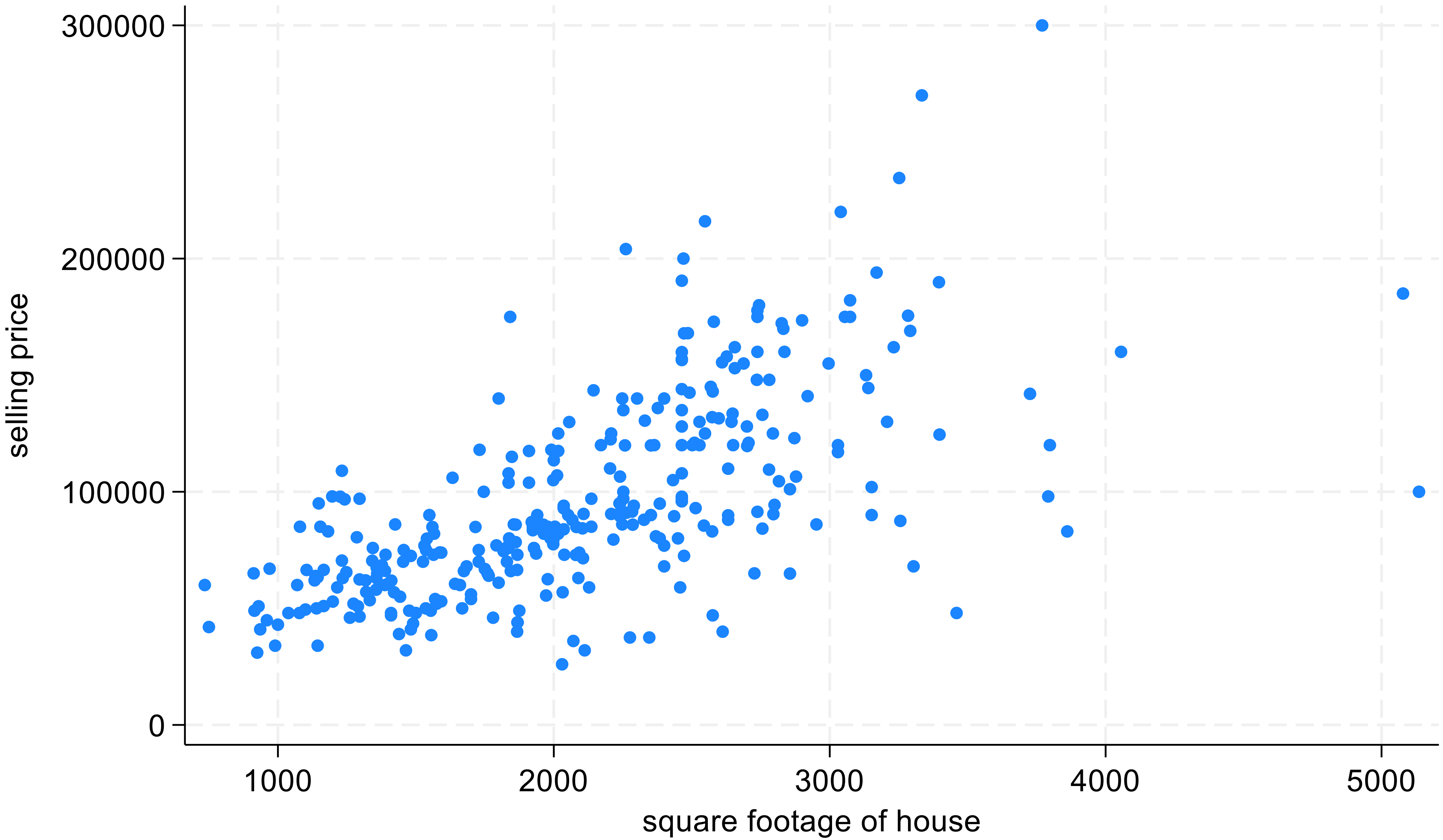
area: square footage of house

land: square footage lot

baths: number of bathrooms

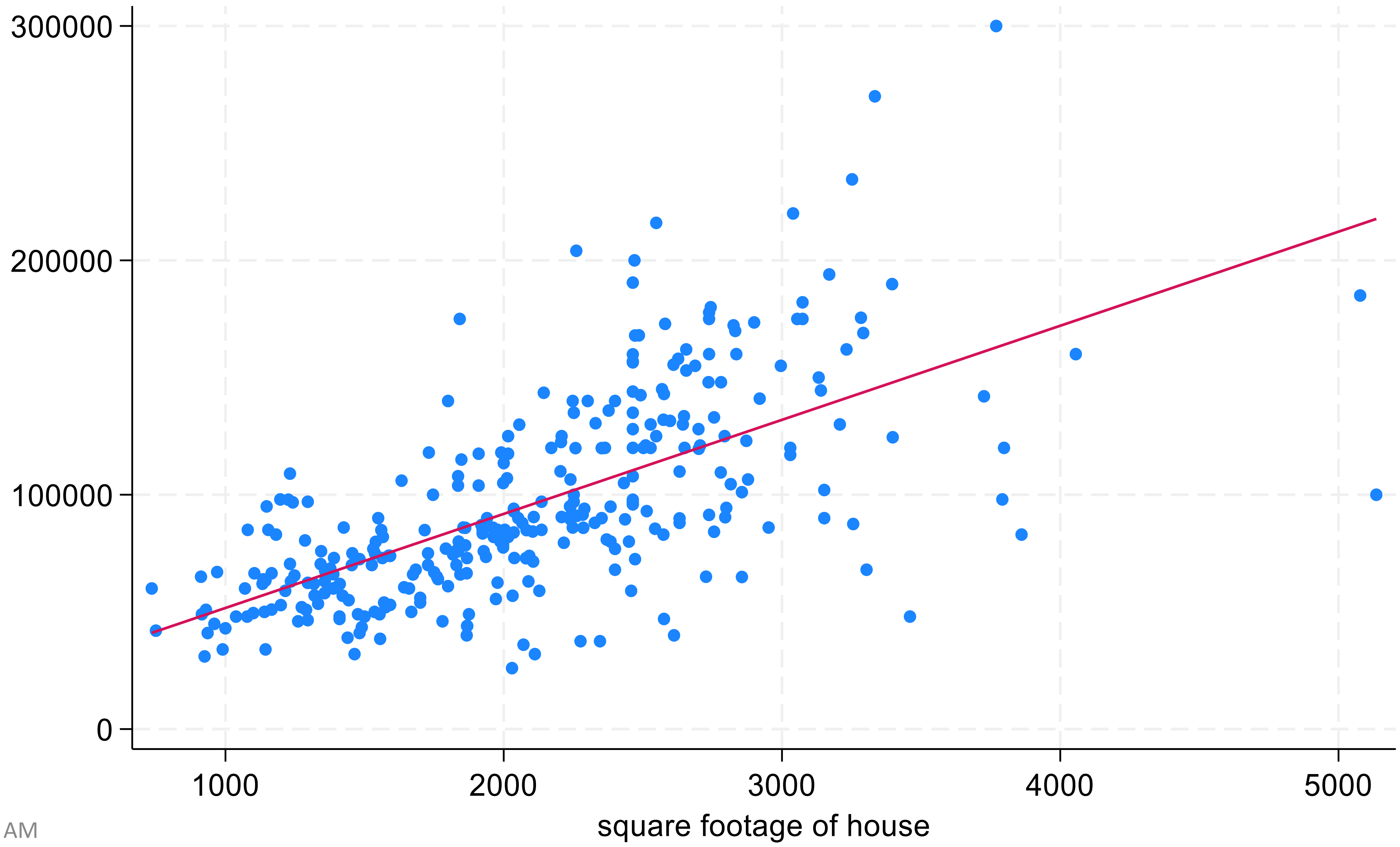
dist: distance from house to incinerator (in feet).

price ~ area

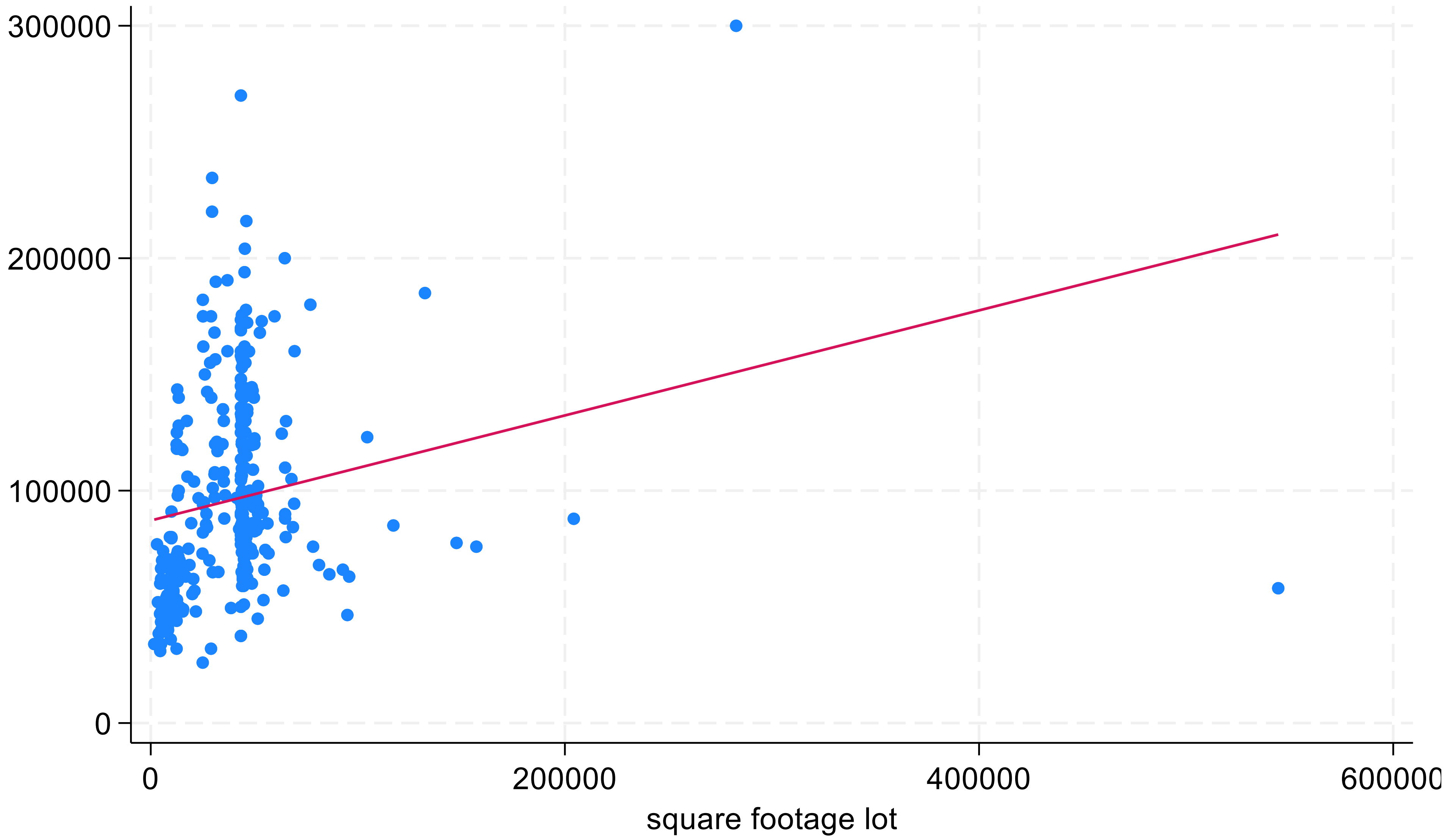


300000
200000
100000
0
selling price
1000
2000
3000
4000
5000
square footage of house

price ~ area

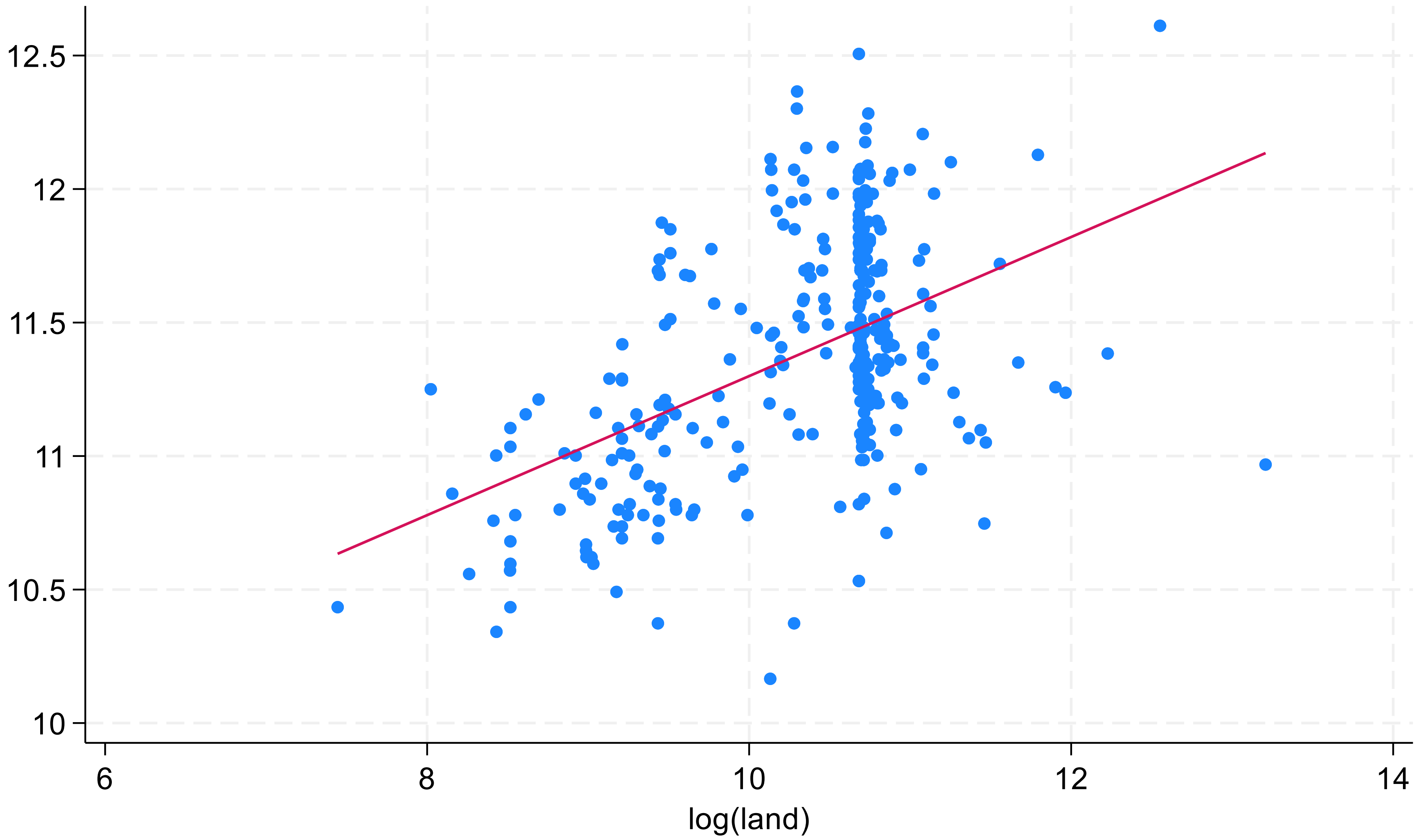

300000
200000
100000
0
1000
2000
3000
4000
5000
square footage of house

price ~ land


300000
200000
100000
0
0
200000
400000
600000
square footage lot

- If a variable has extreme values, it could distort the OLS estimates.
- In such a case, it is better to
  - drop those extreme values.
  - scale down those extreme values.
  - convert the variable in logarithmic form.
  - use quantile regression models instead of OLS.

log(price) ~ log(land)



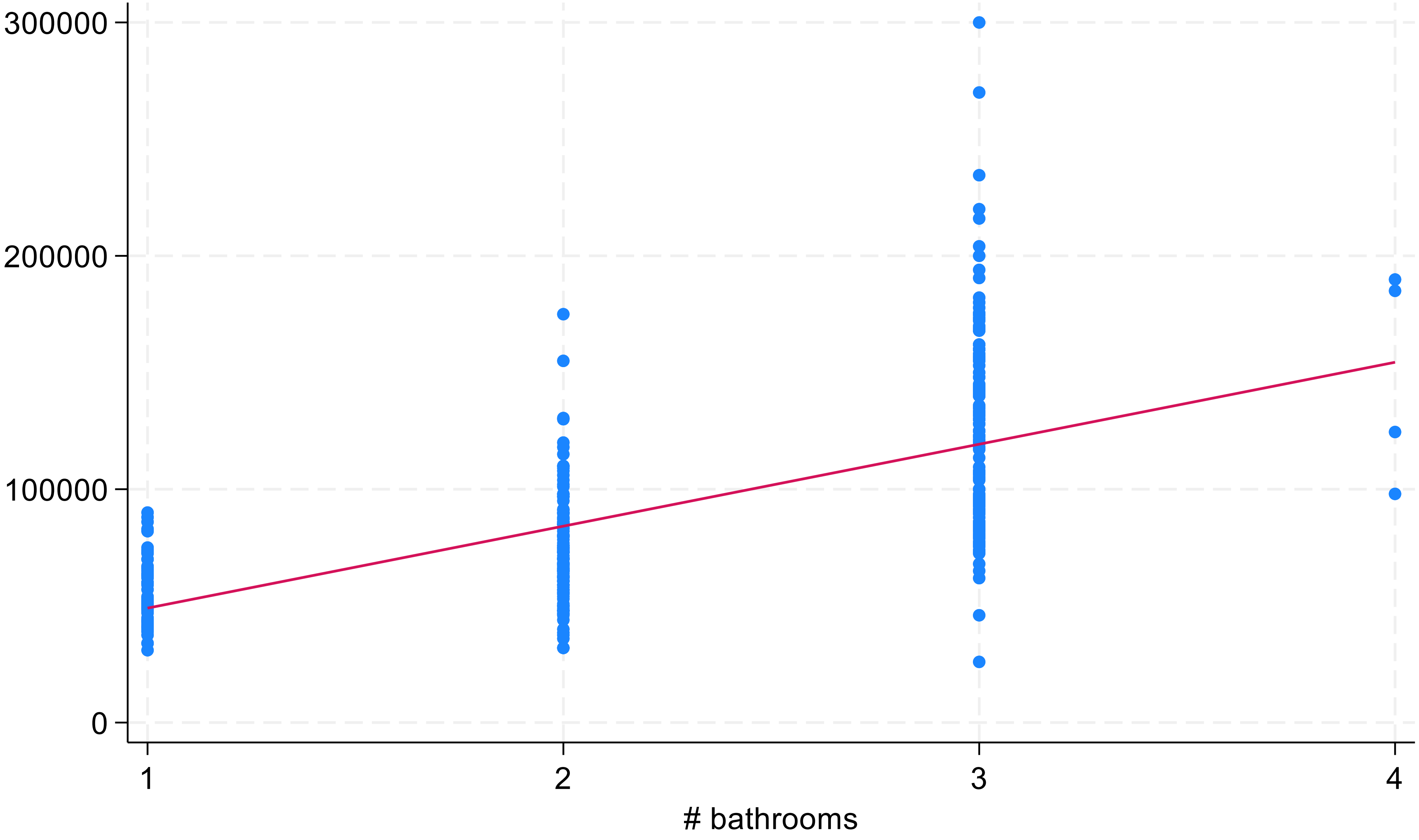
Why we take log?

Generally, we take log of variables with exponential growths such as stock prices, GDP, wages etc.

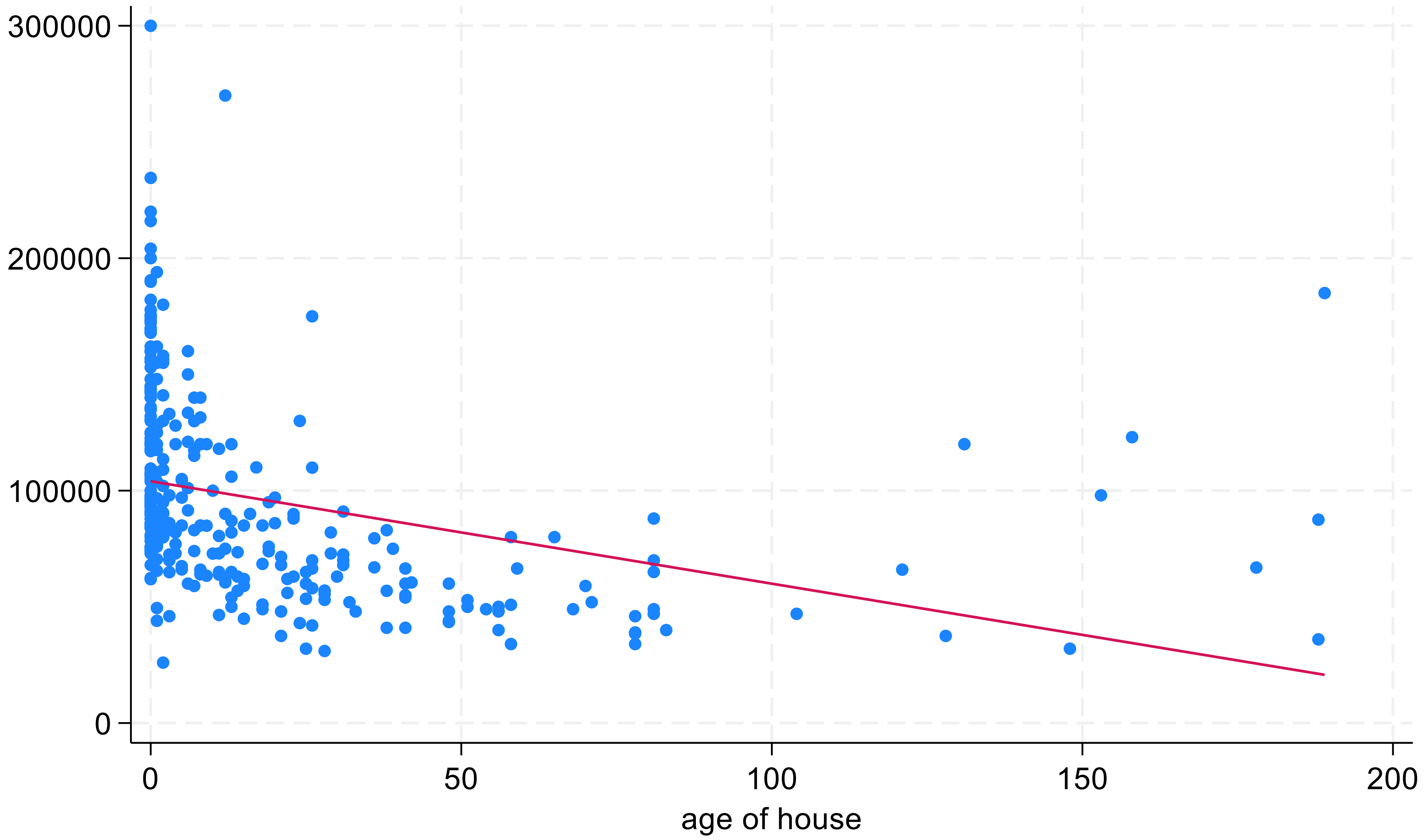
Why we take log?

- To linearize the exponential growth.
- Coefficients represent percentage changes.
- Easier to calculate elasticities.

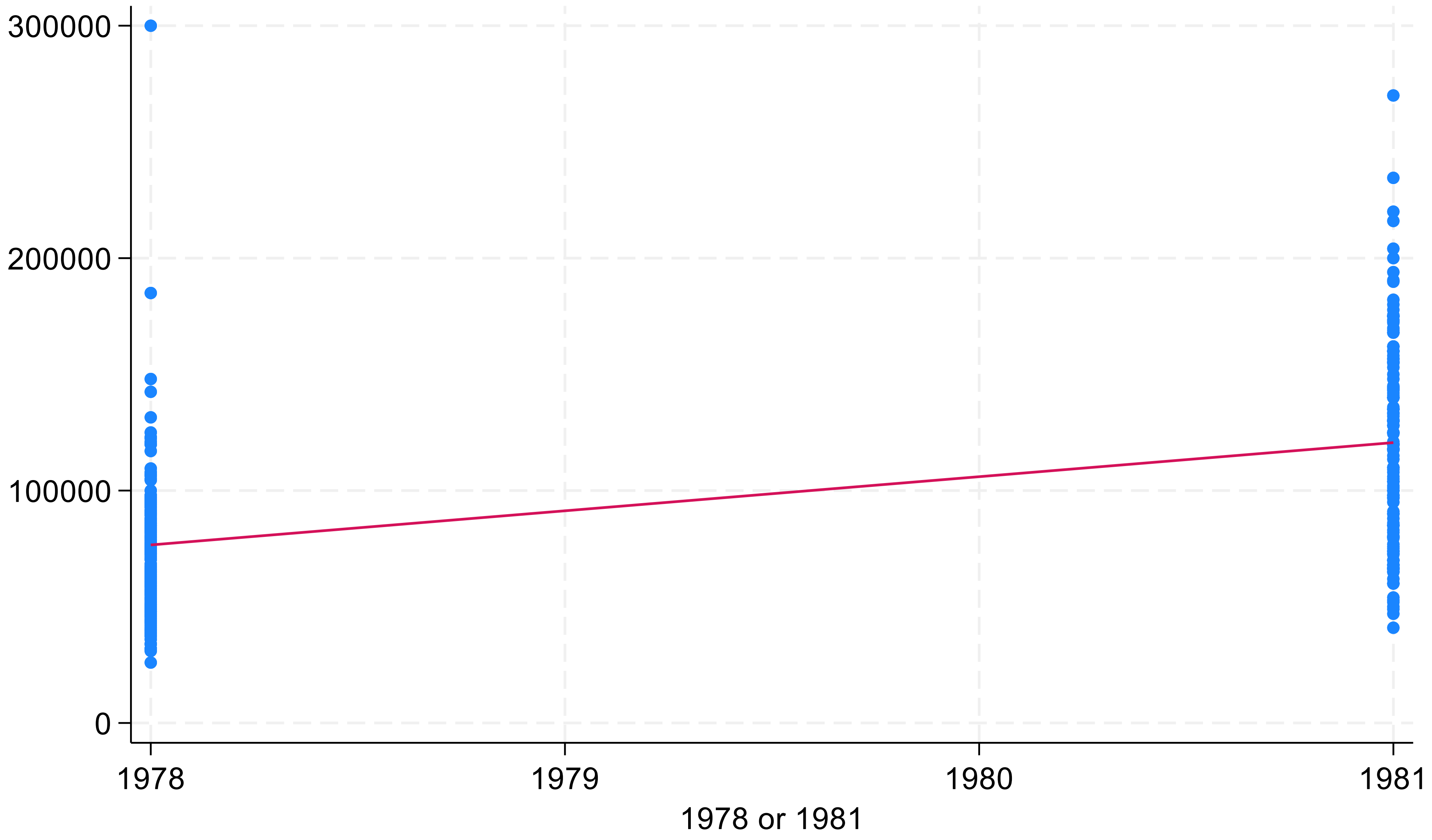
price ~ baths



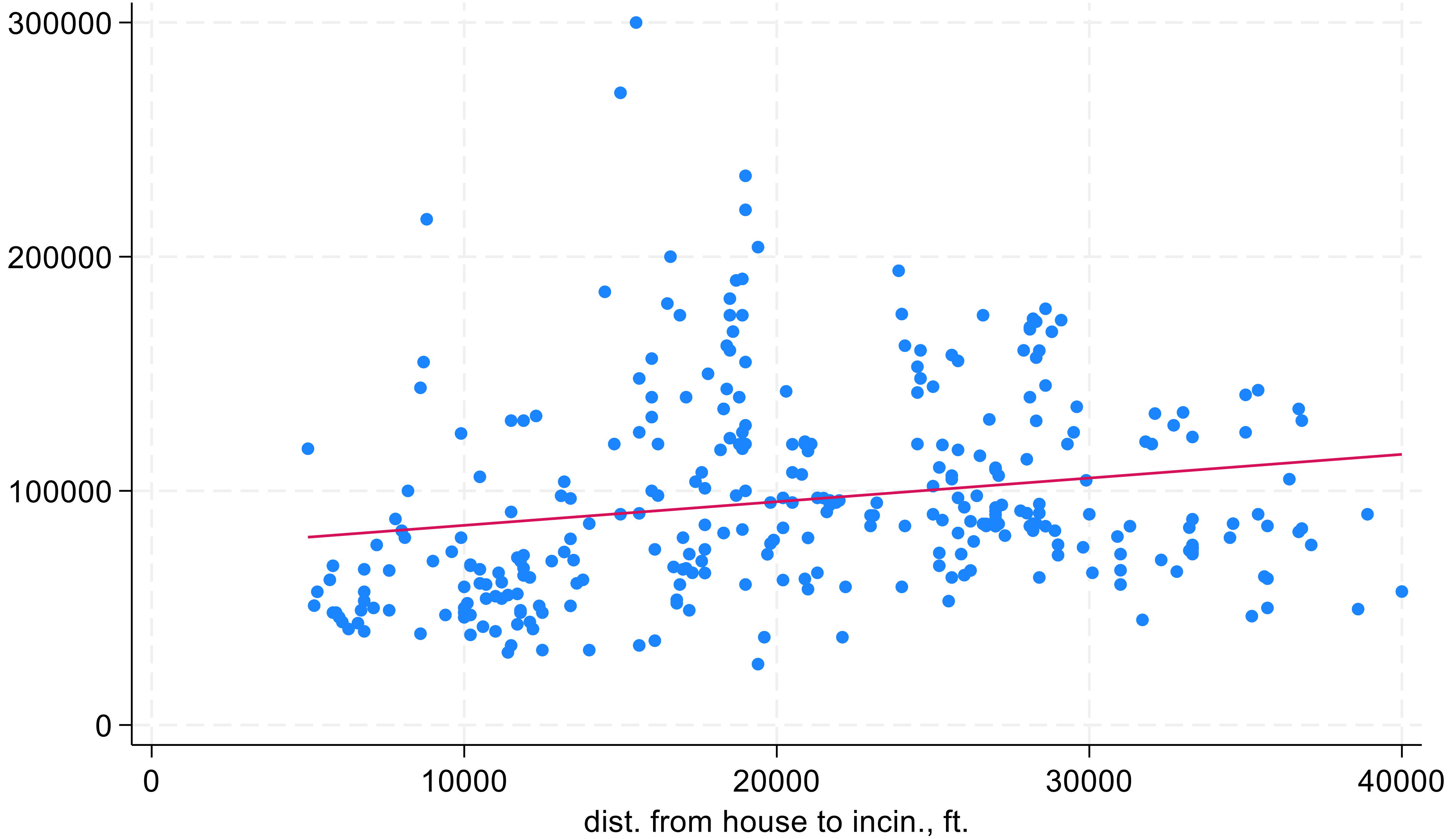
price ~ age



price ~ year

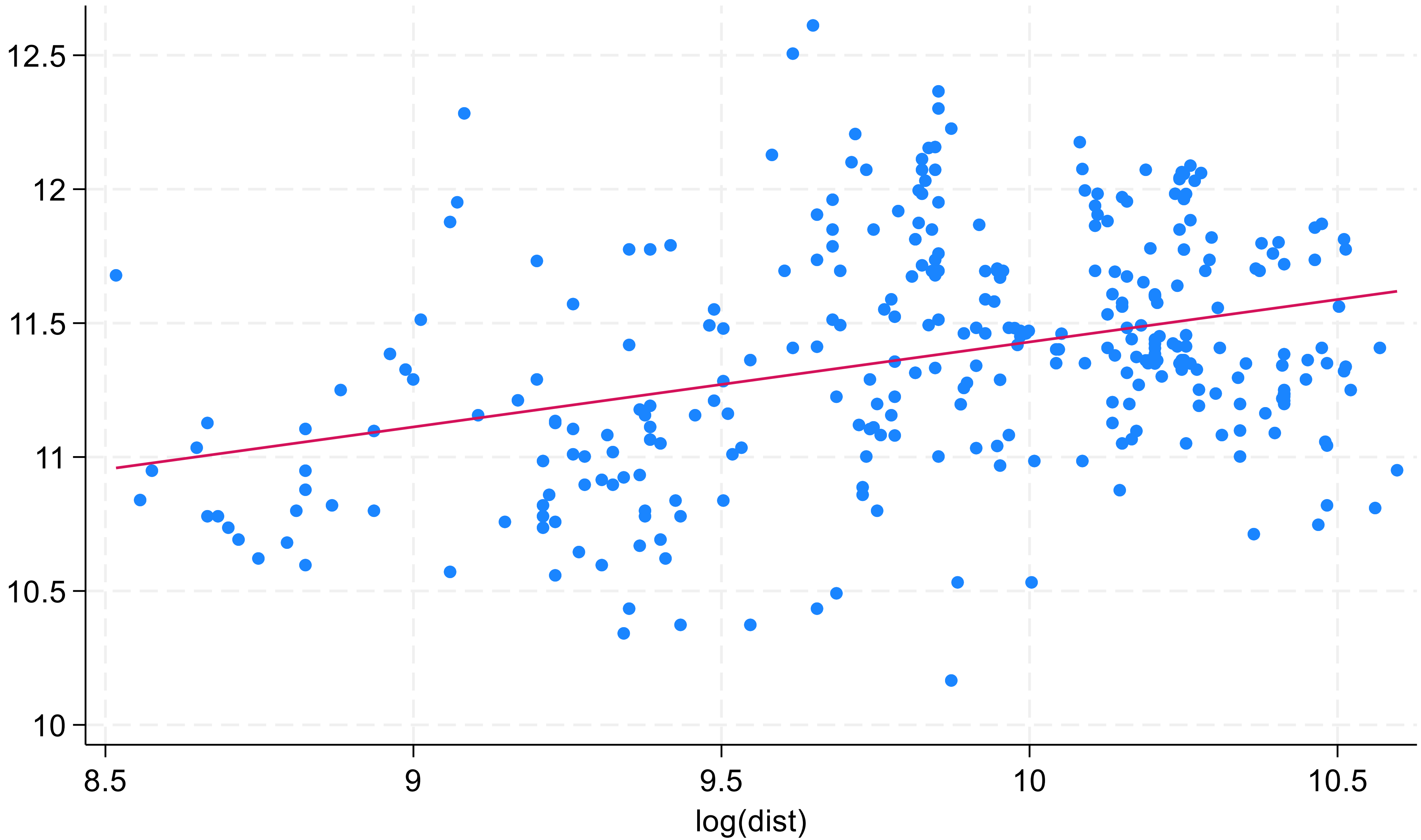


price ~ dist

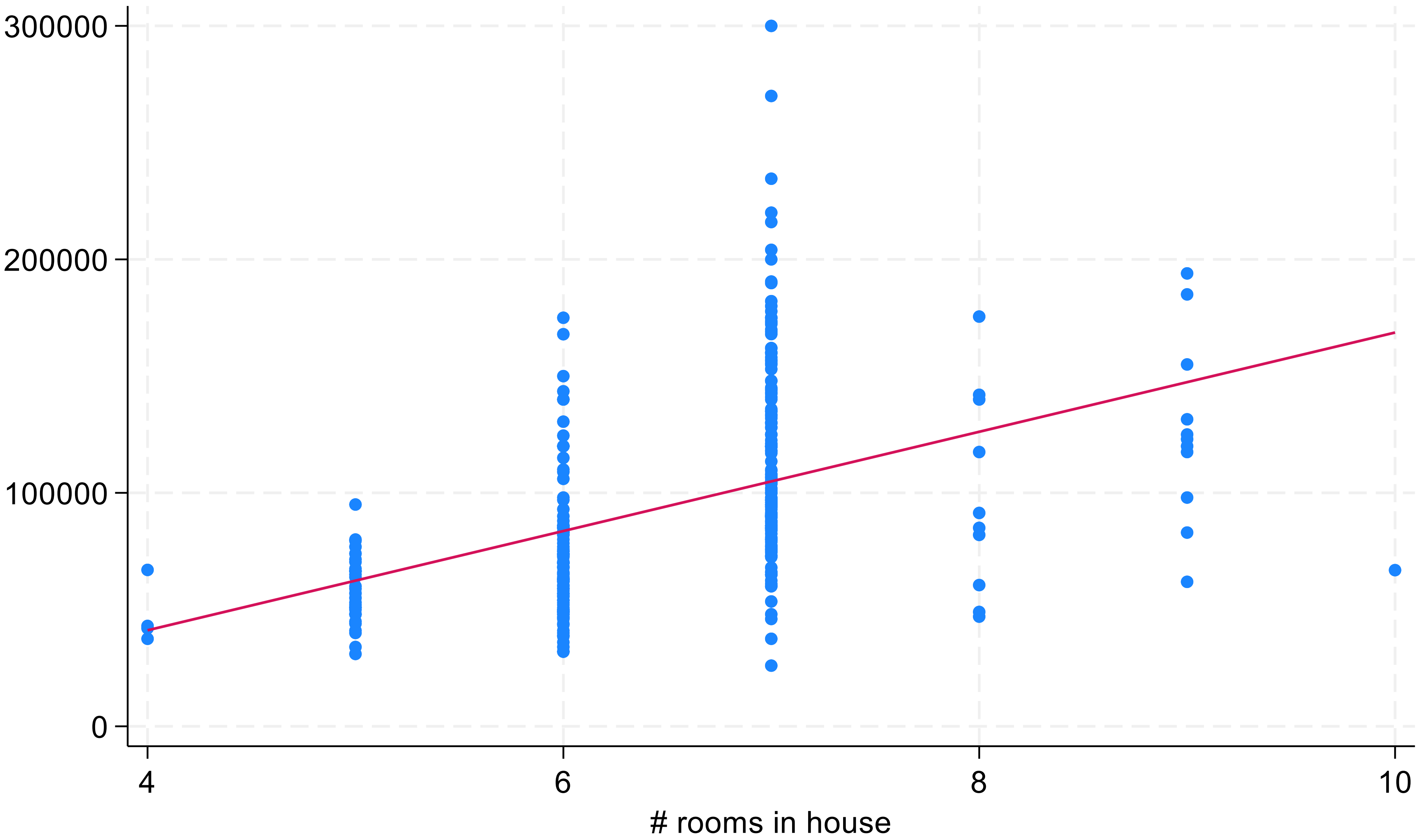


300000
200000
100000
0
0
10000
20000
30000
40000
dist. from house to incin., ft.

log(price) ~ log(dist)



price ~ rooms

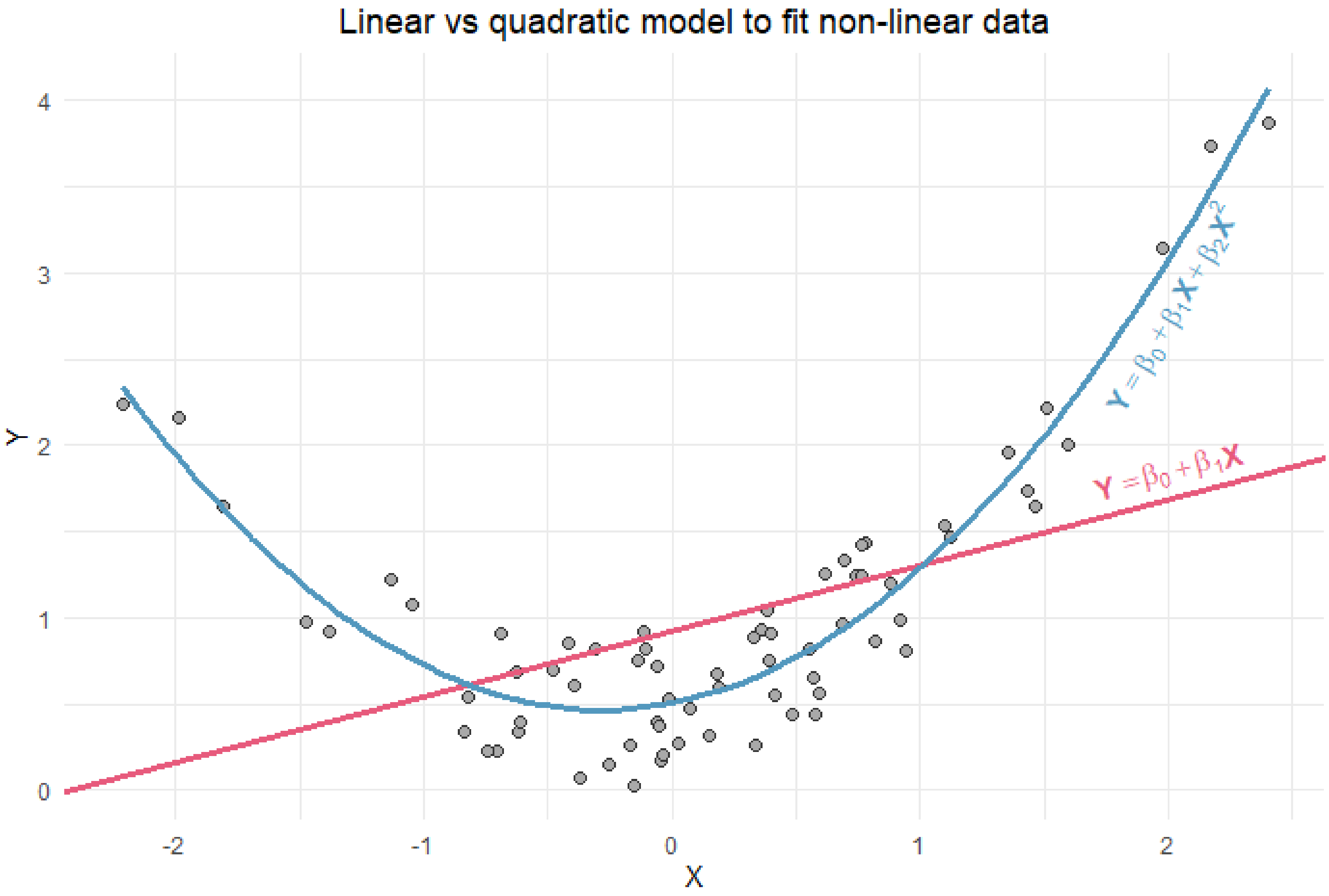


2. Handling non-linear relationship

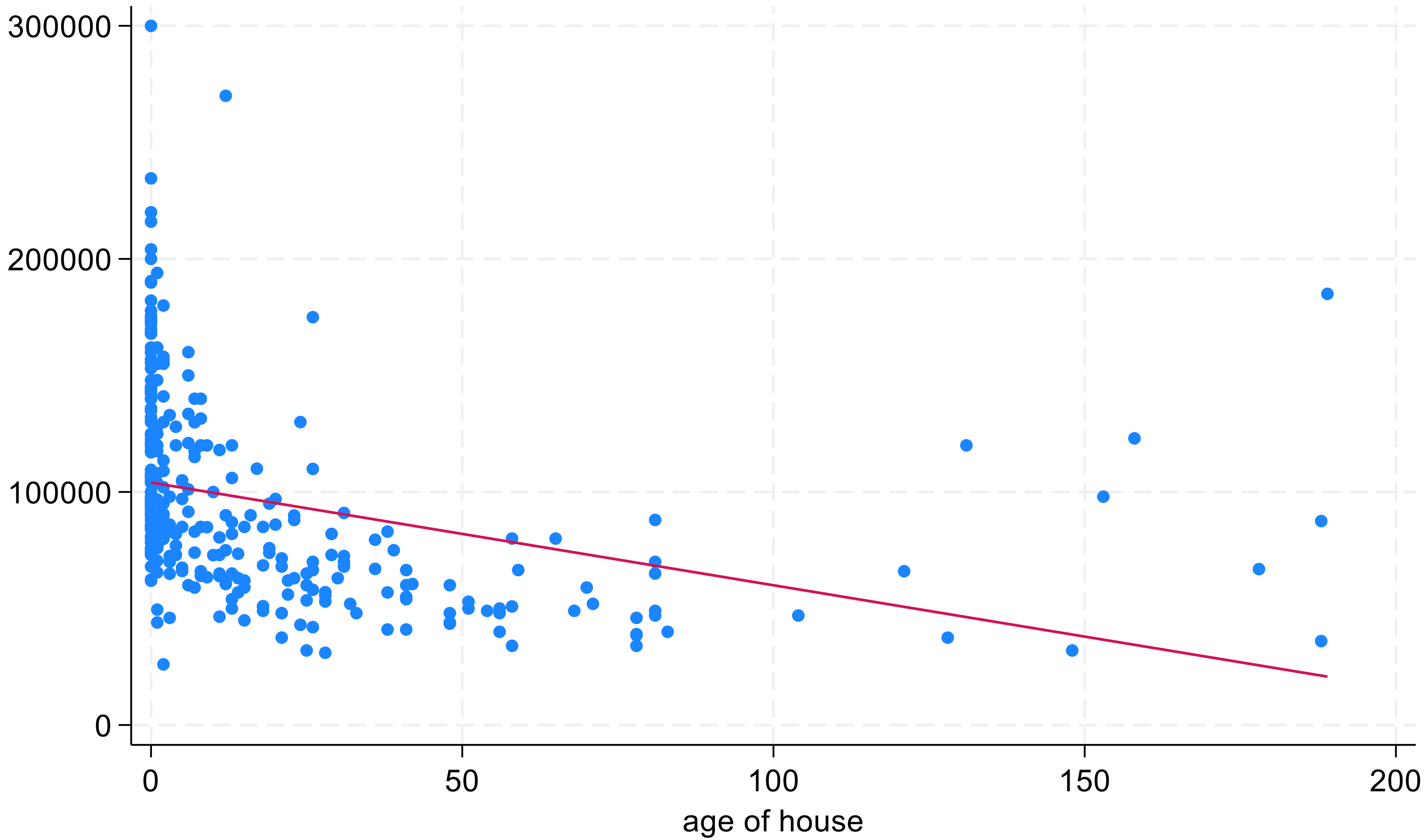
- OLS assumes that outcome Y and predictor X hold a linear relationship.
- If this assumption is invalid, OLS will be a poor fit to data.
- Adding a quadratic term to the regression may help.

Examples of non-linear relationships:

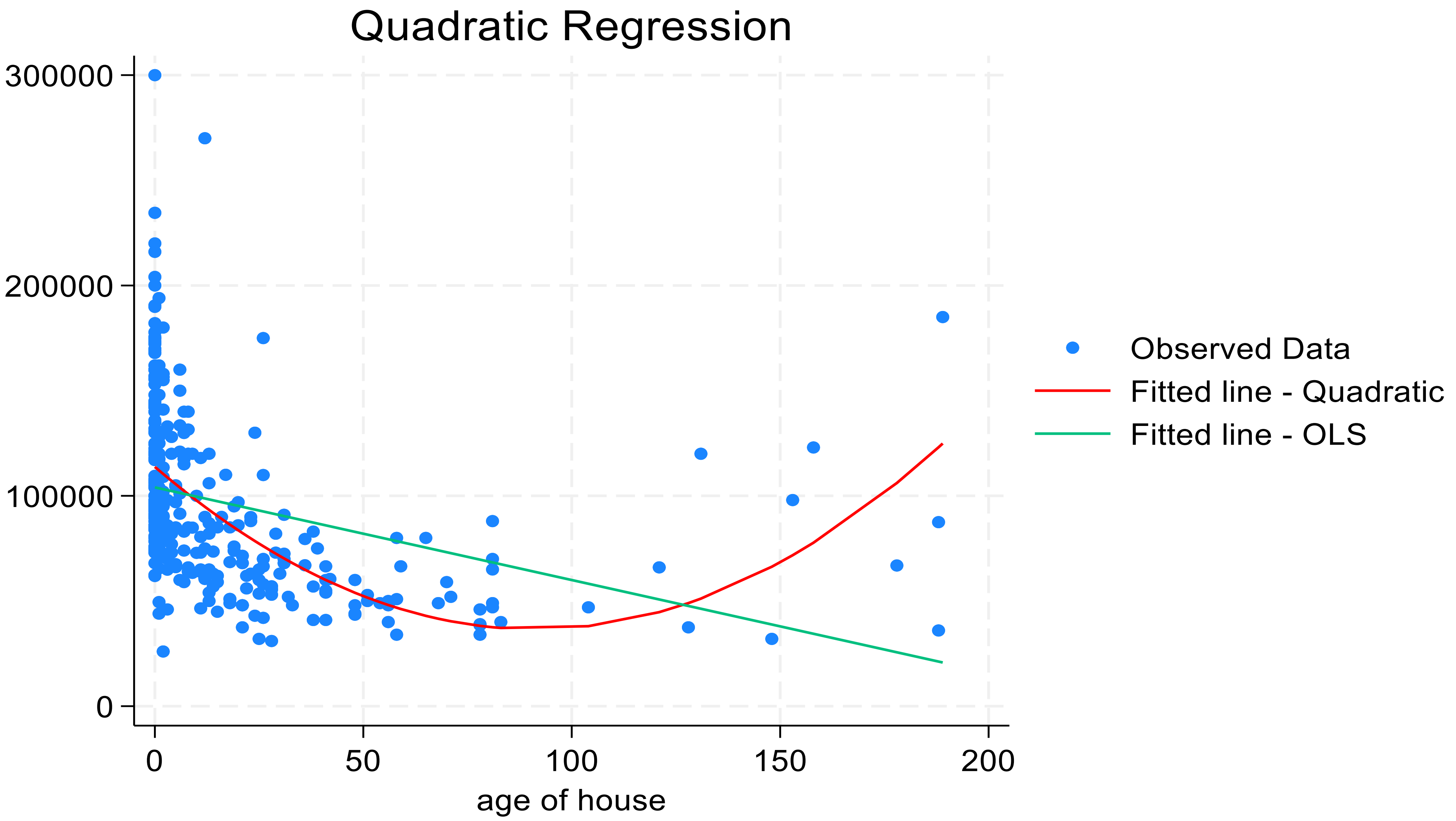
- CO_2 emission ~ GDP per capita (EKC hypothesis)
- Car price ~ life of the car
- Worker productivity ~ worker's age


Linear vs quadratic model to fit non-linear data
4
3
2
1
0
Y
-2
-1
0
1
2
X
$Y = \beta_0 + \beta_1 X + \beta_2 X^2$
$Y = \beta_0 + \beta_1 X$

price ~ age



price ~ age


Quadratic Regression
300000
200000
100000
0
0
50
100
150
200
age of house
Observed Data
Fitted line - Quadratic
Fitted line - OLS

3. Interpreting regression estimates

```
. reg lprice rooms larea lland baths age agesq y81 ldist
```

Source	SS	df	MS
Model	48.1475892	8	6.01844866
Residual	13.2913961	312	.042600629
Total	61.4389853	320	.191996829

Number of obs	=	321
F(8, 312)	=	141.28
Prob > F	=	0.0000
R-squared	=	0.7837
Adj R-squared	=	0.7781
Root MSE	=	.2064

lprice	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
rooms	.0456153	.0174171	2.62	0.009	.0113455	.079885
larea	.3554203	.052142	6.82	0.000	.252826	.4580147
lland	.0664187	.0204743	3.24	0.001	.0261336	.1067039
baths	.1074252	.0277303	3.87	0.000	.0528633	.1619872
age	-.0070673	.0013569	-5.21	0.000	-.0097372	-.0043975
agesq	.0000307	8.45e-06	3.64	0.000	.0000141	.0000474
y81	.3897487	.0239645	16.26	0.000	.3425962	.4369012
ldist	-.0091126	.0329457	-0.28	0.782	-.0739365	.0557112
_cons	7.443987	.4524415	16.45	0.000	6.553765	8.33421

```
. reg lprice rooms larea lland baths age agesq y81 ldist
```

Source	SS	df	MS
Model	48.1475892	8	6.01844866
Residual	13.2913961	312	.042600629
Total	61.4389853	320	.191996829

Number of obs	=	321
F(8, 312)	=	141.28
Prob > F	=	0.0000
R-squared	=	0.7837
Adj R-squared	=	0.7781
Root MSE	=	.2064

lprice	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
rooms	.0456153	.0174171	2.62	0.009	.0113455	.079885
larea	.3554203	.052142	6.82	0.000	.252826	.4580147
lland	.0664187	.0204743	3.24	0.001	.0261336	.1067039
baths	.1074252	.0277303	3.87	0.000	.0528633	.1619872
age	-.0070673	.0013569	-5.21	0.000	-.0097372	-.0043975
agesq	.0000307	8.45e-06	3.64	0.000	.0000141	.0000474
y81	.3897487	.0239645	16.26	0.000	.3425962	.4369012
ldist	-.0091126	.0329457	-0.28	0.782	-.0739365	.0557112
_cons	7.443987	.4524415	16.45	0.000	6.553765	8.33421

Significance of the model:

The F-statistic with a p-value less than 1% indicates that the model is statistically significant at the 1% level. This means the joint estimates of the coefficients are statistically not equal to zero.

```
. reg lprice rooms larea lland baths age agesq y81 ldist
```

Source	SS	df	MS
Model	48.1475892	8	6.01844866
Residual	13.2913961	312	.042600629
Total	61.4389853	320	.191996829

Number of obs	=	321
F(8, 312)	=	141.28
Prob > F	=	0.0000
R-squared	=	0.7837
Adj R-squared	=	0.7781
Root MSE	=	.2064

lprice	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
rooms	.0456153	.0174171	2.62	0.009	.0113455	.079885
larea	.3554203	.052142	6.82	0.000	.252826	.4580147
lland	.0664187	.0204743	3.24	0.001	.0261336	.1067039
baths	.1074252	.0277303	3.87	0.000	.0528633	.1619872
age	-.0070673	.0013569	-5.21	0.000	-.0097372	-.0043975
agesq	.0000307	8.45e-06	3.64	0.000	.0000141	.0000474
y81	.3897487	.0239645	16.26	0.000	.3425962	.4369012
ldist	-.0091126	.0329457	-0.28	0.782	-.0739365	.0557112
_cons	7.443987	.4524415	16.45	0.000	6.553765	8.33421

Model fit:

Independent variables explains 78% of the variation in the log prices.

```
. reg lprice rooms larea lland baths age agesq y81 ldist
```

Source	SS	df	MS			
Model	48.1475892	8	6.01844866	Number of obs	=	321
Residual	13.2913961	312	.042600629	F(8, 312)	=	141.28
				Prob > F	=	0.0000
Total	61.4389853	320	.191996829	R-squared	=	0.7837
				Adj R-squared	=	0.7781
				Root MSE	=	.2064

lprice	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
rooms	.0456153	.0174171	2.62	0.009	.0113455	.079885
larea	.3554203	.052142	6.82	0.000	.252826	.4580147
lland	.0664187	.0204743	3.24	0.001	.0261336	.1067039
baths	.1074252	.0277303	3.87	0.000	.0528633	.1619872
age	-.0070673	.0013569	-5.21	0.000	-.0097372	-.0043975
agesq	.0000307	8.45e-06	3.64	0.000	.0000141	.0000474
y81	.3897487	.0239645	16.26	0.000	.3425962	.4369012
ldist	-.0091126	.0329457	-0.28	0.782	-.0739365	.0557112
_cons	7.443987	.4524415	16.45	0.000	6.553765	8.33421

Root Mean Squared Error (RMSE):

The Root MSE measures the standard deviation of the residuals. Lower values indicate better fit.

```
. reg lprice rooms larea lland baths age agesq y81 ldist

      Source |       SS           df       MS      Number of obs   =       321
-------------+----------------------------------   F(8, 312)       =    141.28
       Model |  48.1475892         8  6.01844866   Prob > F        =    0.0000
    Residual |  13.2913961       312  .042600629   R-squared       =    0.7837
-------------+----------------------------------   Adj R-squared   =    0.7781
       Total |  61.4389853       320  .191996829   Root MSE        =     .2064

------------------------------------------------------------------------------
      lprice | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       rooms |   .0456153   .0174171     2.62   0.009     .0113455     .079885
       larea |   .3554203    .052142     6.82   0.000      .252826    .4580147
       lland |   .0664187   .0204743     3.24   0.001     .0261336    .1067039
       baths |   .1074252   .0277303     3.87   0.000     .0528633    .1619872
         age |  -.0070673   .0013569    -5.21   0.000    -.0097372   -.0043975
       agesq |   .0000307   8.45e-06     3.64   0.000     .0000141    .0000474
         y81 |   .3897487   .0239645    16.26   0.000     .3425962    .4369012
       ldist |  -.0091126   .0329457    -0.28   0.782    -.0739365    .0557112
       _cons |   7.443987   .4524415    16.45   0.000     6.553765     8.33421
------------------------------------------------------------------------------
```

Coefficient interpretation:

- For each additional room, price increases by 0.046 (4.6%).
- 1% increase in area is associated with a 0.36% increase in price.
- agesq +ve coefficient suggests a non-linear relationship between price and age.

```
. reg lprice rooms larea lland baths age agesq y81 ldist
```

Source	SS	df	MS
Model	48.1475892	8	6.01844866
Residual	13.2913961	312	.042600629
Total	61.4389853	320	.191996829

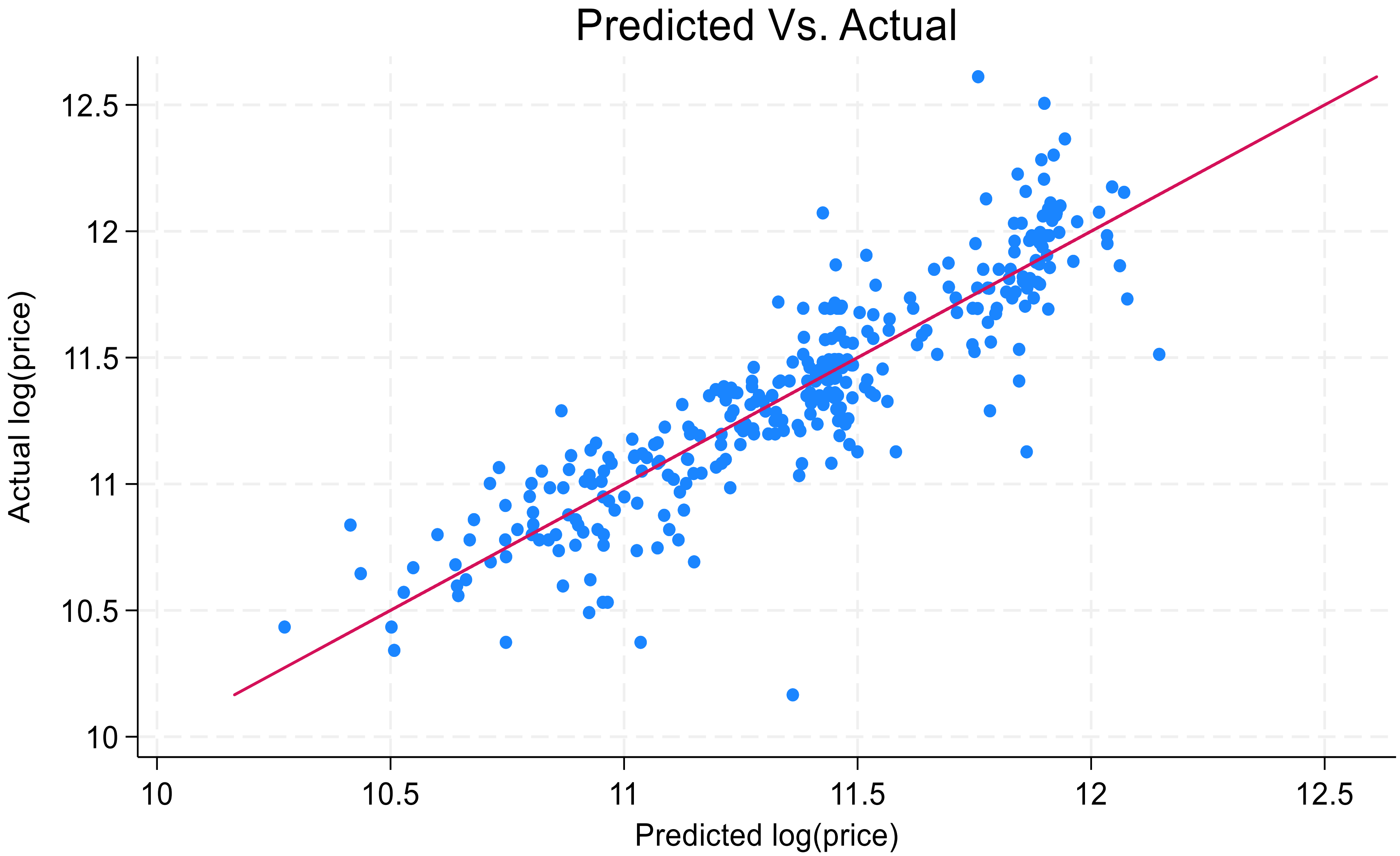
Number of obs	=	321
F(8, 312)	=	141.28
Prob > F	=	0.0000
R-squared	=	0.7837
Adj R-squared	=	0.7781
Root MSE	=	.2064

lprice	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
rooms	.0456153	.0174171	2.62	0.009	.0113455	.079885
larea	.3554203	.052142	6.82	0.000	.252826	.4580147
lland	.0664187	.0204743	3.24	0.001	.0261336	.1067039
baths	.1074252	.0277303	3.87	0.000	.0528633	.1619872
age	-.0070673	.0013569	-5.21	0.000	-.0097372	-.0043975
agesq	.0000307	8.45e-06	3.64	0.000	.0000141	.0000474
y81	.3897487	.0239645	16.26	0.000	.3425962	.4369012
ldist	-.0091126	.0329457	-0.28	0.782	-.0739365	.0557112
_cons	7.443987	.4524415	16.45	0.000	6.553765	8.33421

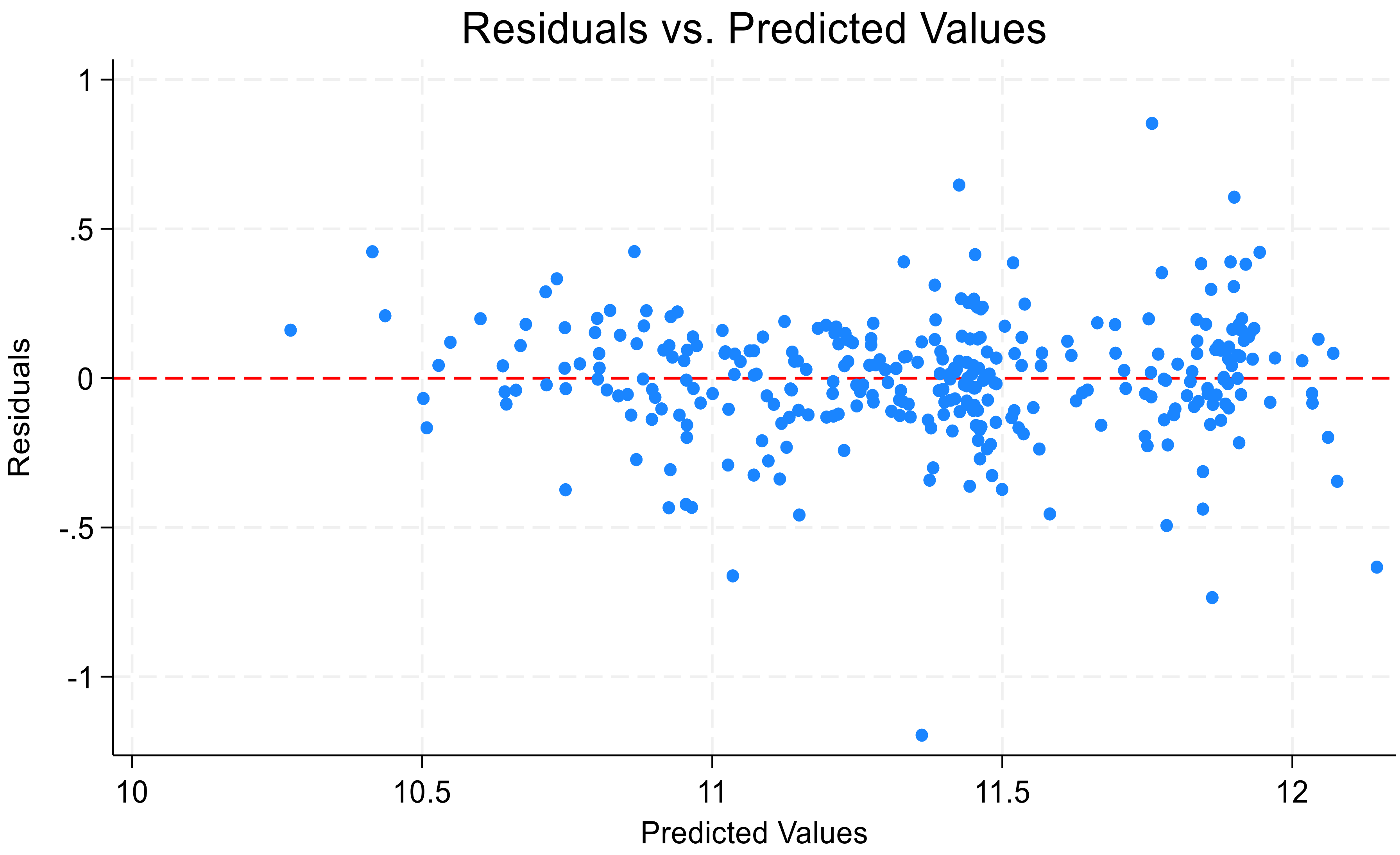
Coefficient interpretation:

- The coefficient of 0.3897 for the dummy variable y81 suggests that, on average, houses are priced 38.97% higher compared to house prices in 1978.
- All the coefficient estimates are statistically significant at 1% except for ***ldist***.

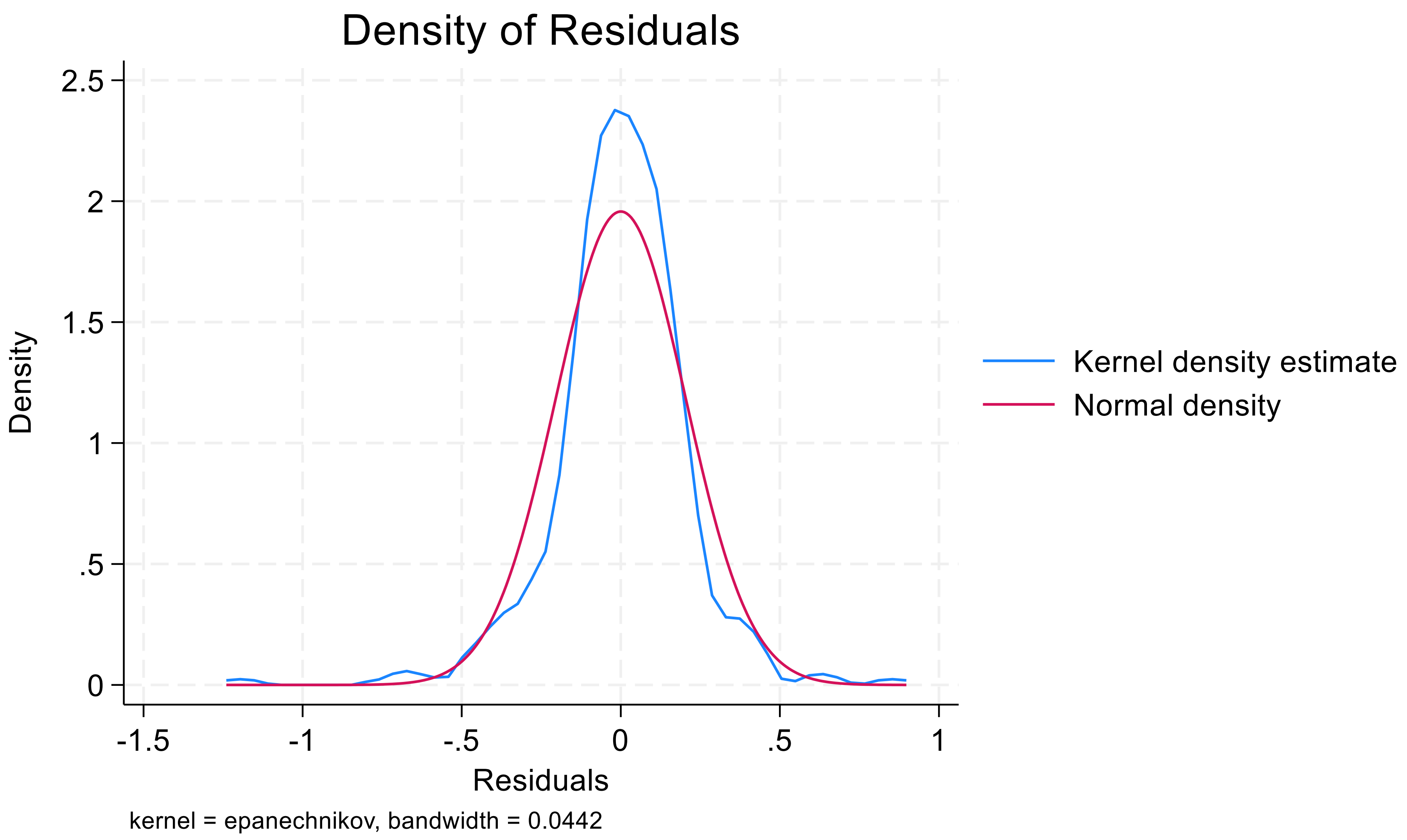
4. Model estimate visualization



In a well-fitted model, the predicted values and the actual values of the dependent variable should closely align along a 45-degree line (i.e. x=y).



In a residual plot, residuals should not exhibit any pattern and should randomly scattered around y=0 line.



A well-fitted model should have a normally distributed residuals.

Formal test of residual normality

```
*Shapiro-Wilk test is generally preferred for smaller samples (n < 2000).
swilk res
```

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
res	321	0.93993	13.590	6.145	0.00000

```
*Shapiro-Francia test, which is an alternative to Shapiro-Wilk, often used for larger samples.
sfrancia res
```

Shapiro-Francia W' test for normal data

Variable	Obs	W'	V'	z	Prob>z
res	321	0.93390	16.181	5.935	0.00001

However, in real world, it is challenging to obtain residuals that are normally distributed because including all the relevant factors that influence the dependent variable might not be possible.

5. The problem of multicollinearity

```
. cor lprice rooms larea lland baths age agesq y81 ldist
(obs=321)
```

	lprice	rooms	larea	lland	baths	age	agesq	y81	ldist
lprice	1.0000								
rooms	0.4933	1.0000							
larea	0.6558	0.5606	1.0000						
lland	0.4765	0.3993	0.3877	1.0000					
baths	0.6746	0.6038	0.7093	0.4949	1.0000				
age	-0.4013	-0.0512	-0.0988	-0.3265	-0.3569	1.0000			
agesq	-0.1858	0.1651	0.0843	-0.1100	-0.0938	0.9159	1.0000		
y81	0.5108	0.0058	0.1528	-0.0256	0.0471	-0.1104	-0.1147	1.0000	
ldist	0.3463	0.3113	0.2168	0.6314	0.3875	-0.3561	-0.1565	-0.0387	1.0000

*According to the above correlation matrix, it is better to drop **baths** and **ldist** from the model specification.*

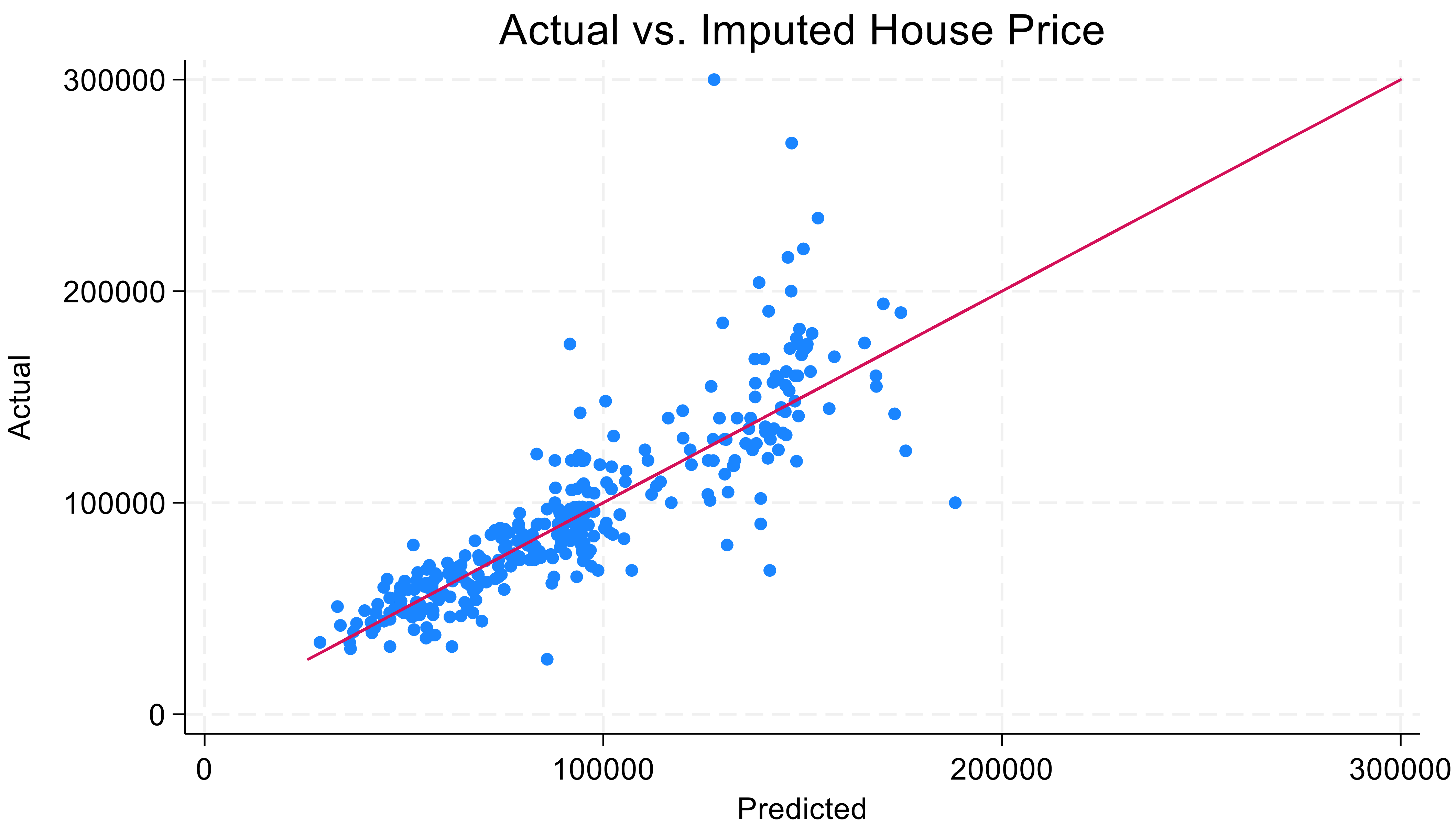
We should not use independent variables that are highly correlated to each other as this can lead to a multicollinearity problem, which can result in biased and unreliable estimates.

6. The imputed house prices

```
* imputed housing price
reg lprice rooms larea lland baths age agesq y81 ldist

* Predict fitted values
capture noisily drop fitted_y price_predicted
predict fitted_y, xb

gen price_predicted = exp(fitted_y)
twoway (scatter price price_predicted) (line price price), title(
"Actual vs. Imputed House Price") legend(off) ///
    xtitle(Predicted) ytitle(Actual)
```

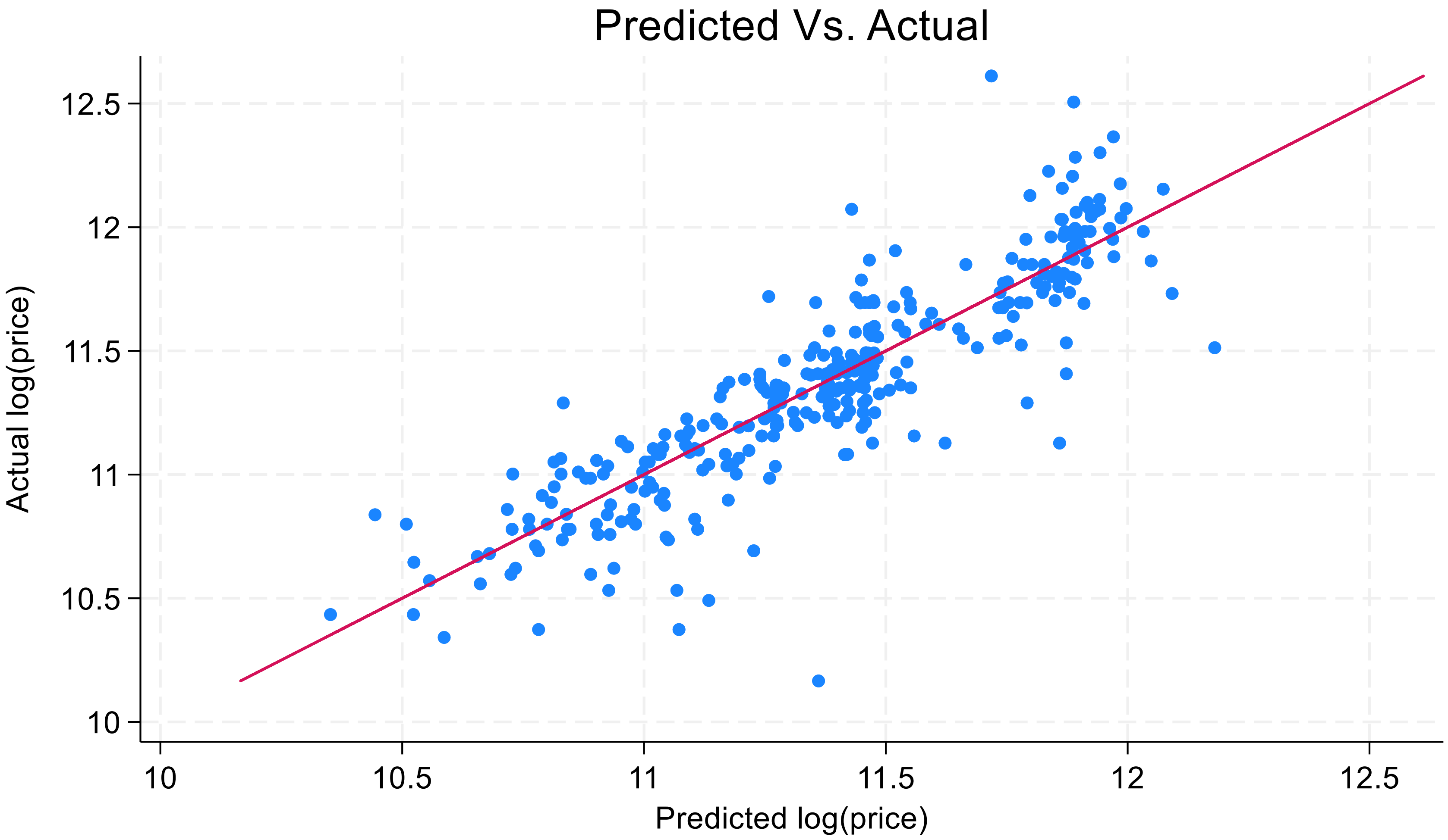

Actual vs. Imputed House Price
300000
200000
100000
0
Actual
0
100000
200000
300000
Predicted

Quantile Regression as an alternative to Hedonic Pricing

Benefits of quantile regression

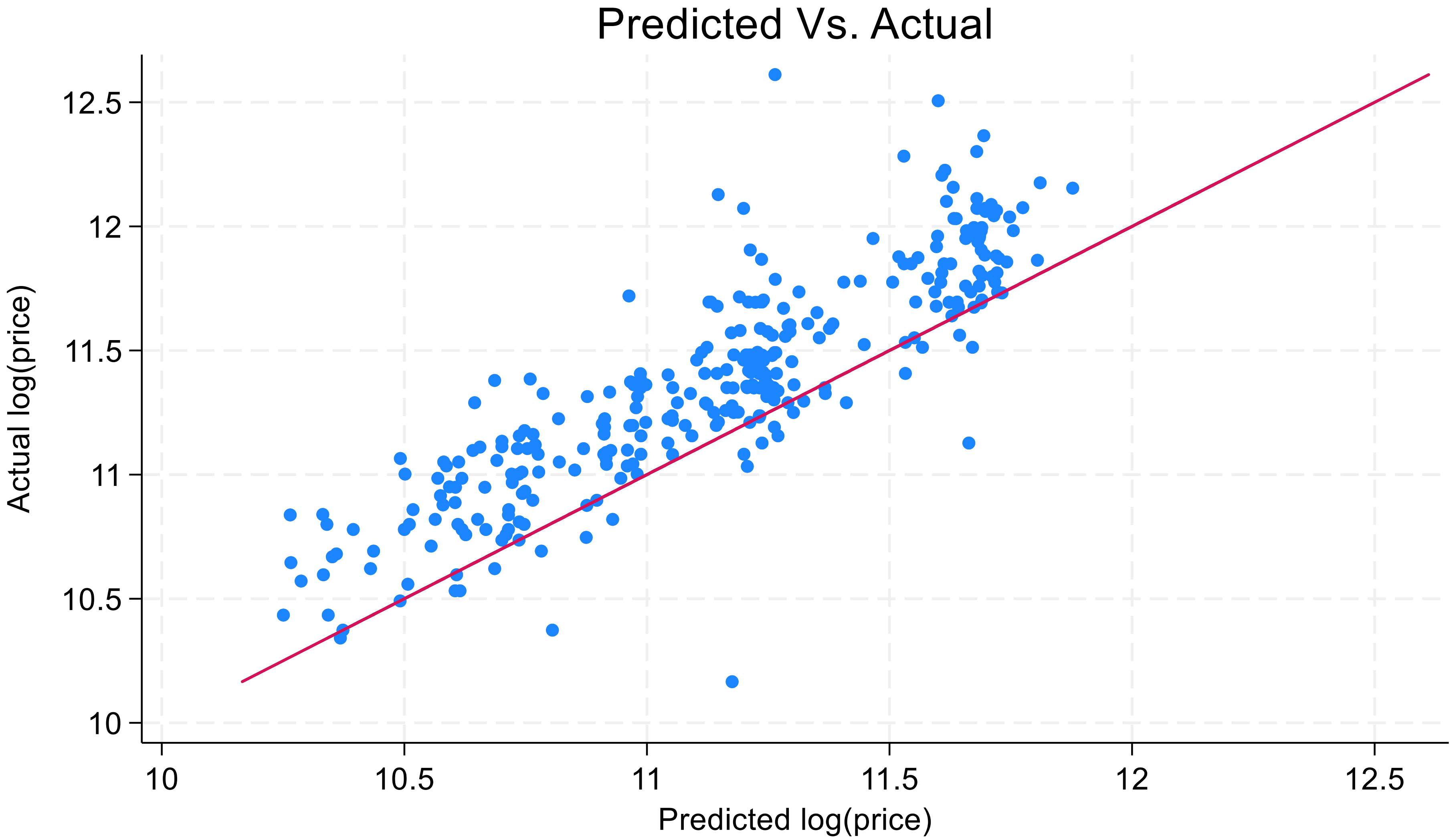
- **Varying effects on different quantiles:** OLS calculates conditional means only. QR shows how relationships change across different quantiles.
- **Outlier Influence:** Mitigates the influence of outliers. Useful for skewed or heavy-tailed data
- **Non-Normal Errors:** OLS assumes normally distributed errors. Quantile regression does not.

Quantile regression (q=0.5)

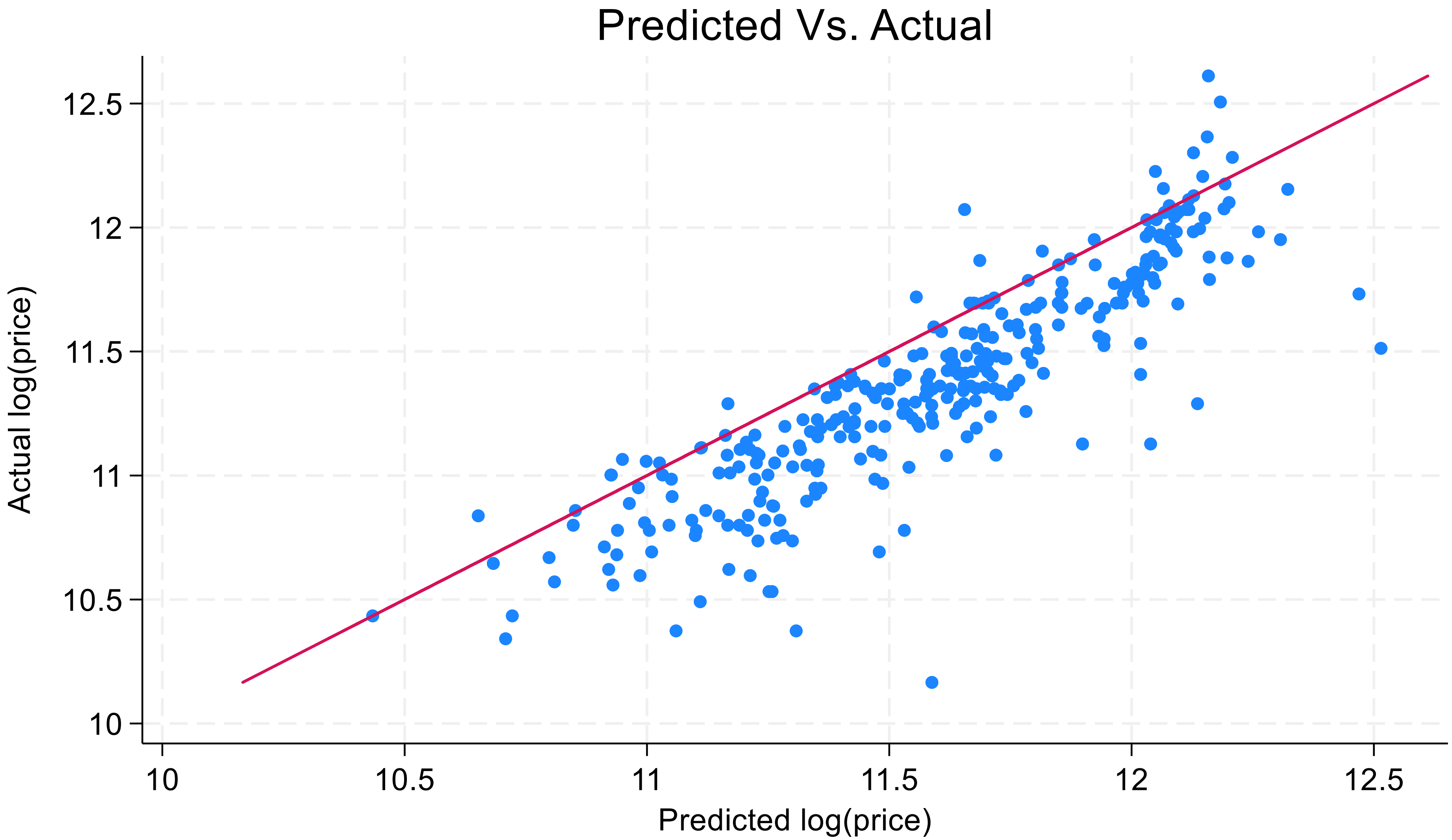


RMSE = 0.2072

Quantile regression (q=0.1)



Quantile regression (q=0.9)



Thank You