Multiple Linear Regression

Lecture 16, 17

Ravleen Bajaj

Today:

- "Adjustment" in multiple linear regression model.
- Examples.

Why "Adjustment"?

- In simple regression, we study the *total* association between x and y.
- In multiple regression, predictors can be correlated.
- We want to know the effect of given predictor while accounting for the other predictors.

The Model

Multiple Regression Model for p predictors

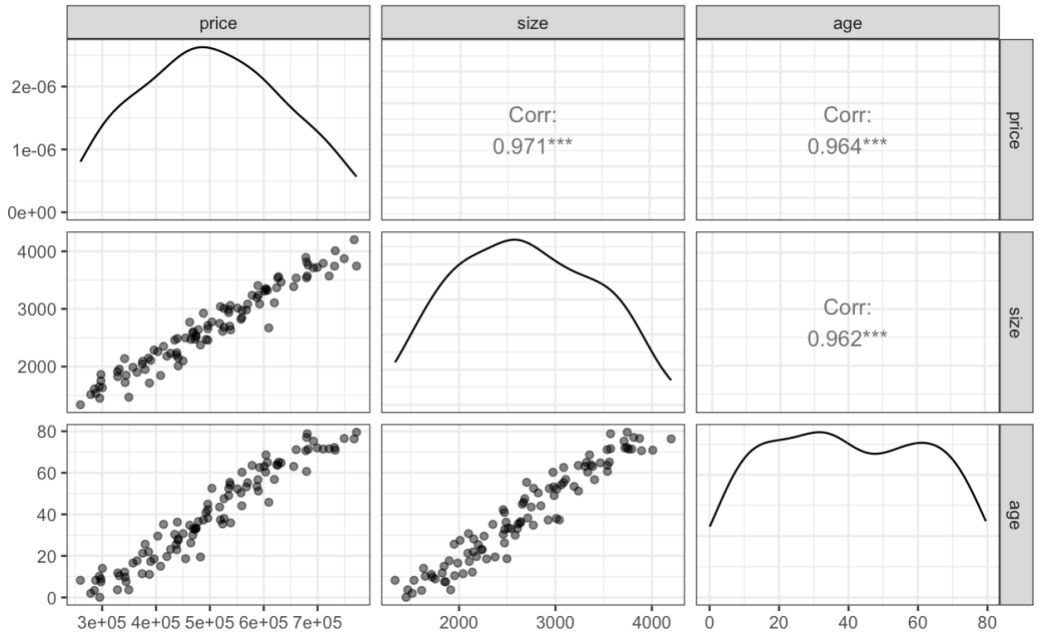
$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \epsilon_i$$

where ϵ_i is the error associated with i^{th} response, ϵ_i's are independent and $\epsilon_i \sim N(0, \sigma^2)$.

Interpretation: β_j is the expected change in y for a one–unit increase in x_j, **holding all other predictors constant**.

Example

- We are interested to predict the price of a house using multiple predictors, i.e., size, and age (how old the house is).



It was observed that:

- Positive relationship between price and size.
- Positive relationship between price and age.

- But are older homes really expensive?

IMPORTANT: These pairwise plots may mask relationships.

Why Pairwise Plots Can Mislead

- These relationships discussed do **not account for other predictors**.
- Example:
  - Older homes tend to be **larger** and have **higher** prices.
  - So price and age appear positively related, even if older homes for a fixed price are actually worth less.

Key idea: Pairwise plots compare houses that differ in multiple ways, not just in age or size.

Discussion: The Apparent Effect of Age

- Based on pairwise plots alone, we might think:

$$\text{Older homes} \rightarrow \text{Higher prices}$$

- The positive effect of size is masking the negative effect of age.
- But perhaps:

$$\text{Older homes} \rightarrow \text{Larger size} \rightarrow \text{Higher price.}$$

- Is the relationship between age and price **direct** or **mediated by age**?

DISCUSSION: What would happen if we compared houses of the same size?

Adjusting for Predictors

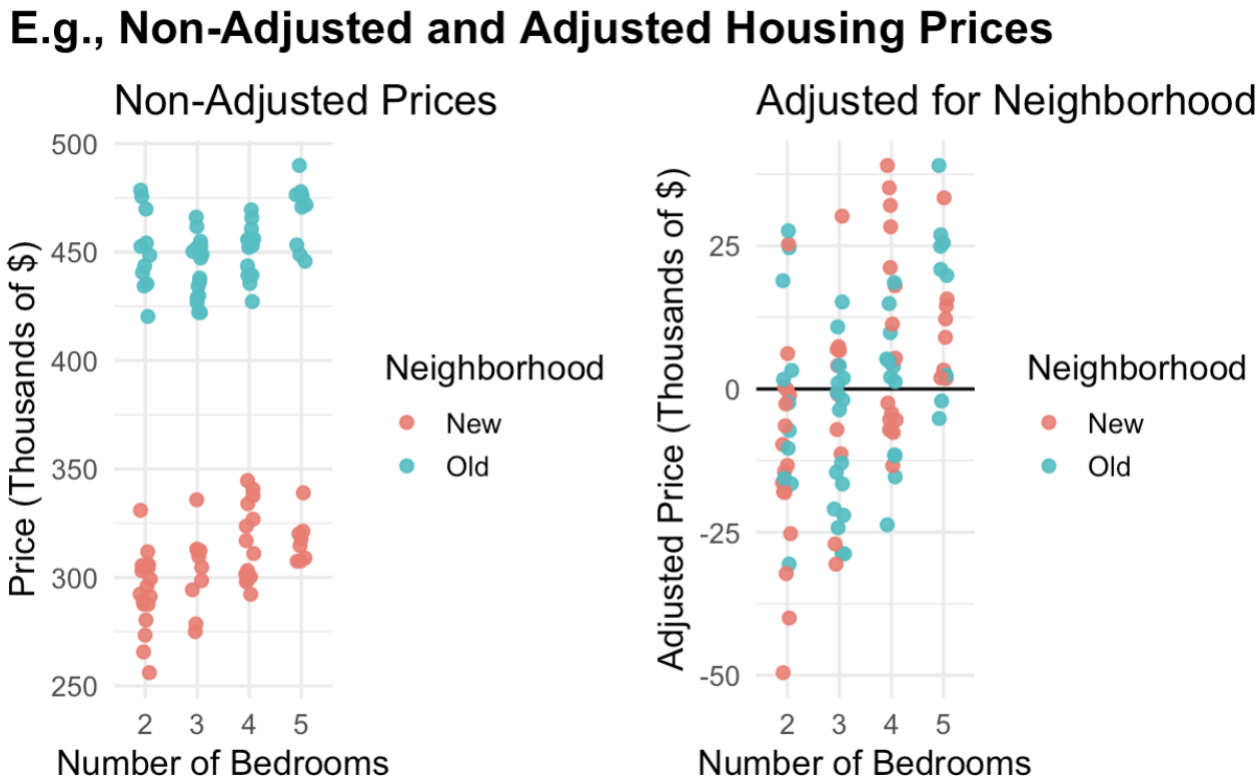
The multiple regression model adjusts automatically:

$$\text{Price}_i = \beta_0 + \beta_1 \text{Size}_i + \beta_2 \text{Age}_i + \varepsilon_i$$

Interpretation: β_2: effect of **Age** on Price, *holding Size constant.*

- In practice, β_2 often is **negative**, i.e., older homes sell for less when other factors are held fixed.
- This is the **adjusted effect** of age.

Non-Adjusted vs Adjusted Comparison



- Theoretically, the inherent age effect has been adjusted by subtracting the mean price for each neighborhood.
- After adjusting, we observe that the price increases with increase in the number of bedrooms.

Takeaways

- "Adjusting (responses) for a predictor" means that we can treat the responses as if they came from individuals who all have the same value of this predictor.
- Adjustment isolates each variable's **unique contribution** to the variability in the response.
- Multiple Linear Regression automatically performs these adjustments.
- Without adjustment, we risk attributing changes in Y to the wrong x.
- Good to think about: What variables are we adjusting for?

Thank You!

Questions?