Multiple linear regression

Inference

Prof. Maria Tackett

Click here for PDF of slides

Topics

- Conduct a hypothesis test for β_j
- Calculate a confidence interval for β_j
- Quick overview of math details for MLR

House prices in Levittown

The data set contains the sales price and characteristics of 85 homes in Levittown, NY that sold between June 2010 and May 2011.

We would like to use the characteristics of a house to understand variability in the sales price.

Variables

Predictors

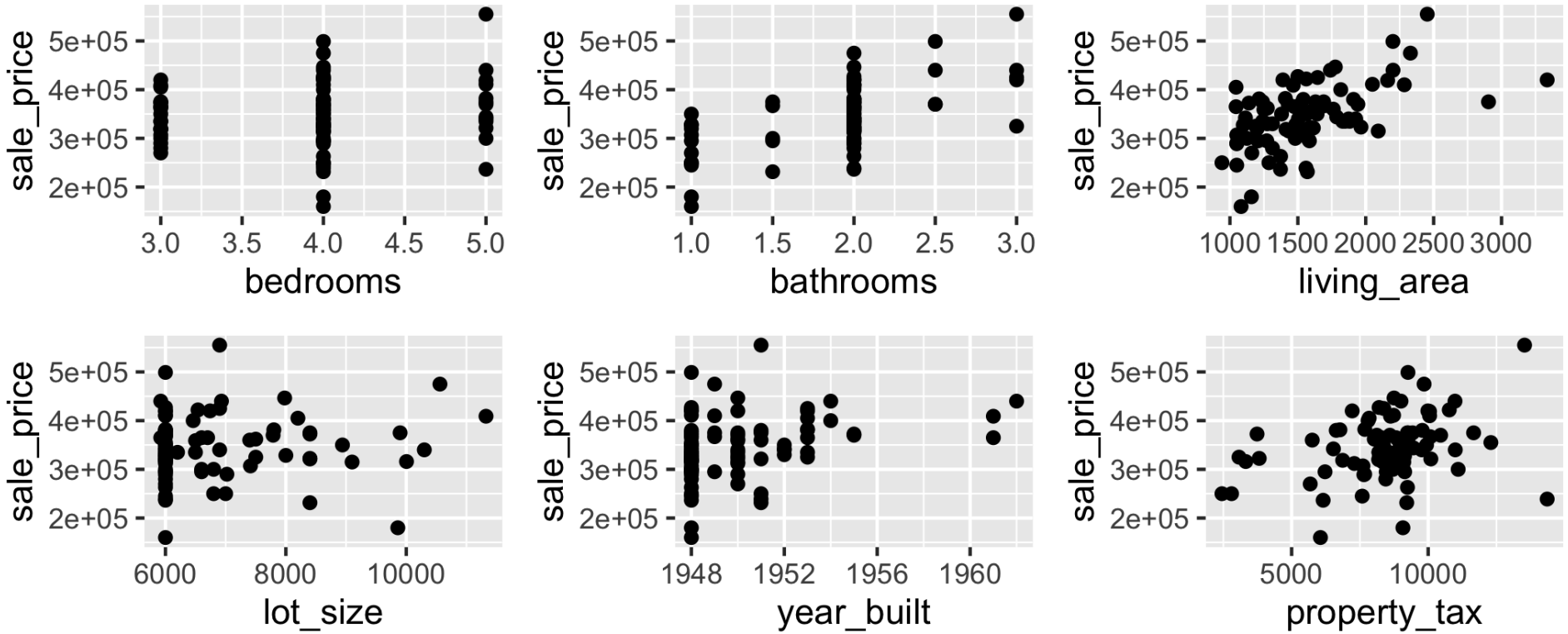
- **`bedrooms`**: Number of bedrooms
- **`bathrooms`**: Number of bathrooms
- **`living_area`**: Total living area of the house (in square feet)
- **`lot_size`**: Total area of the lot (in square feet)
- **`year_built`**: Year the house was built
- **`property_tax`**: Annual property taxes (in U.S. dollars)

Response

- **`sale_price`**: Sales price (in U.S. dollars)

EDA: Response variable

EDA: Response vs. Predictors



Home price model

term	estimate	std.error	statistic	p.value
(Intercept)	-7148818.957	3820093.694	-1.871	0.065
bedrooms	-12291.011	9346.727	-1.315	0.192
bathrooms	51699.236	13094.170	3.948	0.000
living_area	65.903	15.979	4.124	0.000
lot_size	-0.897	4.194	-0.214	0.831
year_built	3760.898	1962.504	1.916	0.059
property_tax	1.476	2.832	0.521	0.604

Hypothesis test for β_j

Outline of a hypothesis test

1️⃣ State the hypotheses.

2️⃣ Calculate the test statistic.

3️⃣ Calculate the p-value.

4️⃣ State the conclusion.

1 State the hypotheses

term	estimate	std.error	statistic	p.value
(Intercept)	-7148818.957	3820093.694	-1.871	0.065
bedrooms	-12291.011	9346.727	-1.315	0.192
bathrooms	51699.236	13094.170	3.948	0.000
living_area	65.903	15.979	4.124	0.000
lot_size	-0.897	4.194	-0.214	0.831
year_built	3760.898	1962.504	1.916	0.059
property_tax	1.476	2.832	0.521	0.604

$$H_0 : \beta_{living\_area} = 0$$
$$H_a : \beta_{living\_area} \neq 0$$

2 Calculate the test statistic

term	estimate	std.error	statistic	p.value
(Intercept)	-7148818.957	3820093.694	-1.871	0.065
bedrooms	-12291.011	9346.727	-1.315	0.192
bathrooms	51699.236	13094.170	3.948	0.000
living_area	65.903	15.979	4.124	0.000
lot_size	-0.897	4.194	-0.214	0.831
year_built	3760.898	1962.504	1.916	0.059
property_tax	1.476	2.832	0.521	0.604

$$t = \frac{65.903 - 0}{15.979} = 4.124$$

2️⃣ Calculate the test statistic

The estimated slope, 65.903, is 4.124 standard errors above the hypothesized mean, 0.

3 Calculate the p-value

term	estimate	std.error	statistic	p.value
(Intercept)	-7148818.957	3820093.694	-1.871	0.065
bedrooms	-12291.011	9346.727	-1.315	0.192
bathrooms	51699.236	13094.170	3.948	0.000
living_area	65.903	15.979	4.124	0.000
lot_size	-0.897	4.194	-0.214	0.831
year_built	3760.898	1962.504	1.916	0.059
property_tax	1.476	2.832	0.521	0.604

$$\text{P-value} = P(|t| \geq |4.124|) = 0.00009$$

3 Calculate the p-value

The p-value is calculated using a t distribution with $n - p - 1$ degrees of freedom, where p is the number of coefficients in the model.

In this example, the p-value is calculated using a t distribution with $85 - 6 - 1 = 78$ degrees of freedom.

Given $\beta_{living\_area} = 0$ the probability of observing a coefficient at least as extreme as the one we've observed, 65.903, is 0.00009.

4 State the conclusion

term	estimate	std.error	statistic	p.value
(Intercept)	-7148818.957	3820093.694	-1.871	0.065
bedrooms	-12291.011	9346.727	-1.315	0.192
bathrooms	51699.236	13094.170	3.948	0.000
living_area	65.903	15.979	4.124	0.000
lot_size	-0.897	4.194	-0.214	0.831
year_built	3760.898	1962.504	1.916	0.059
property_tax	1.476	2.832	0.521	0.604

The p-value is very small, so we reject H_0. The data provide sufficient evidence that the living area is a helpful predictor in the model explaining some of the variability in price.

Confidence interval for β_j

Confidence Interval for β_j

The C confidence interval for β_j

$$\hat{\beta}_j \pm t^* SE(\hat{\beta}_j)$$

where t^* follows a t distribution with $n - p - 1$ degrees of freedom

General Interpretation: We are C confident that the interval LB to UB contains the population coefficient of x_j. Therefore, for every one unit increase in x_j, we expect y to change by LB to UB units, holding all else constant.

Confidence interval for `living_area`

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-7148818.957	3820093.694	-1.871	0.065	-14754041.291	456403.376
bedrooms	-12291.011	9346.727	-1.315	0.192	-30898.915	6316.893
bathrooms	51699.236	13094.170	3.948	0.000	25630.746	77767.726
living_area	65.903	15.979	4.124	0.000	34.091	97.715
lot_size	-0.897	4.194	-0.214	0.831	-9.247	7.453
year_built	3760.898	1962.504	1.916	0.059	-146.148	7667.944
property_tax	1.476	2.832	0.521	0.604	-4.163	7.115

We are 95% confident that for every one additional square foot in living area, we expect the price to increase by $34.09 to $97.71, holding all other characteristics constant.

🛑 Caution: Large sample sizes

If the sample size is large enough, the test will likely result in rejecting $H_0 : \beta_j = 0$ even x_j has a very small effect on y

- Consider the **practical significance** of the result not just the statistical significance
- Use the confidence interval to draw conclusions instead of relying only p-values

🛑 Caution: Small sample sizes

If the sample size is small, there may not be enough evidence to reject $H_0 : \beta_j = 0$

- When you fail to reject the null hypothesis, **DON'T** immediately conclude that the variable has no association with the response.
- There may be a linear association that is just not strong enough to detect given your data, or there may be a non-linear association.

Math details

Regression Model

The multiple linear regression model assumes

$$Y|X_1, X_2, \ldots, X_p \sim N(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p, \sigma^2_\epsilon)$$

For a given observation $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$, we can rewrite the previous statement as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2)$$

Estimating σ_ϵ^2

For a given observation $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$ the residual is

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip})$$

The estimated value of the regression variance , σ_ϵ^2, is

$$\hat{\sigma}_\epsilon^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - p - 1}$$

Estimating Coefficients

One way to estimate the coefficients is by taking partial derivatives of the formula

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip})]^2$$

This produces messy formulas, so instead we can use matrix notation for multiple linear regression and estimate the coefficients using rules from linear algebra. For more details, see A Matrix Formulation of the Multiple Regression Model.

Recap

- Conduct a hypothesis test for β_j
- Calculate a confidence interval for β_j
- Quick overview of math details for MLR