Linear Regression

Modeling Reality

Predicting for the Sale Price of a House

Goal

Use available data to predict the sale price of a house (assume a normal distribution)

Available Data

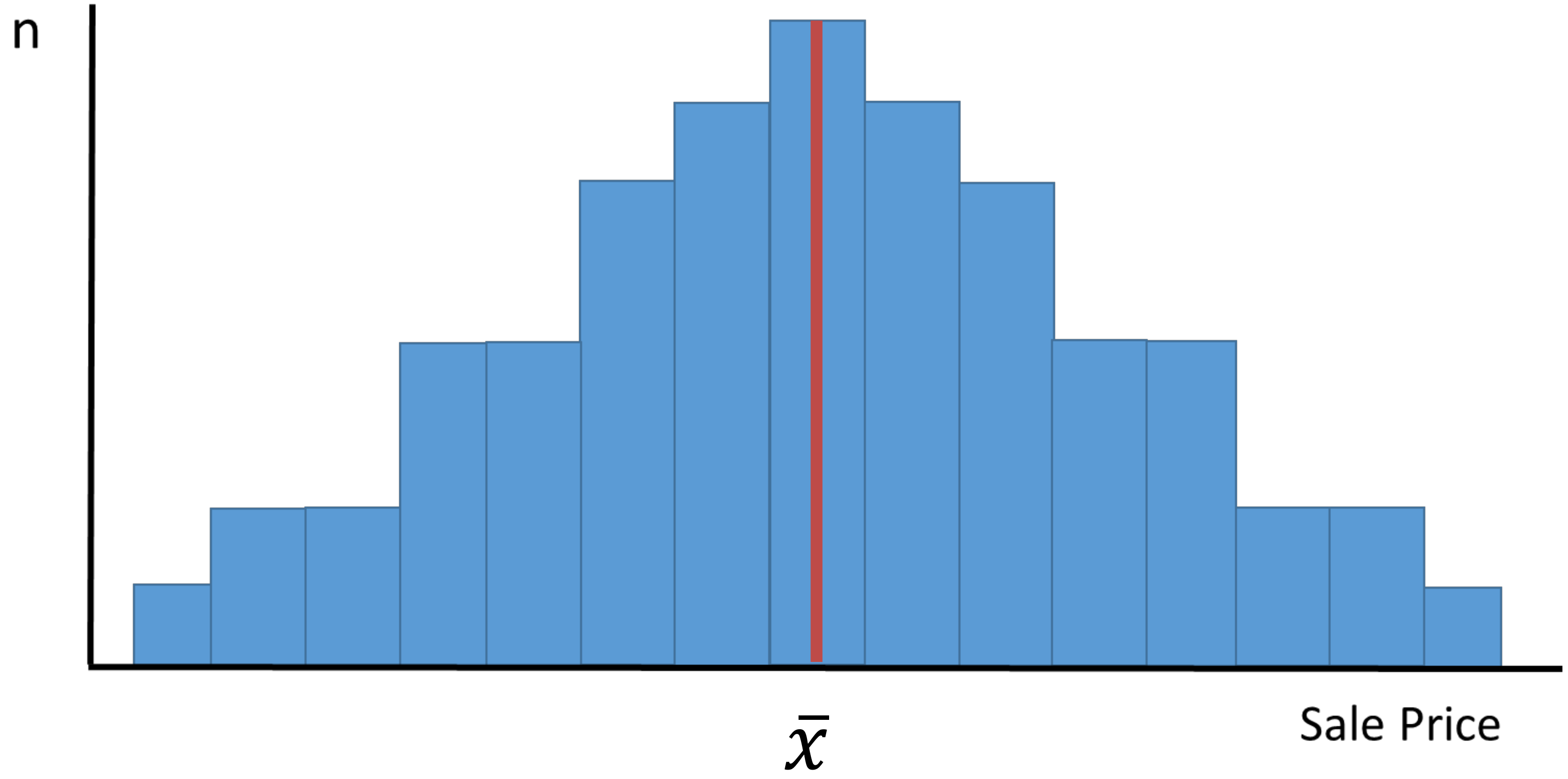
Sale prices of other houses

Unavailable Data

Characteristics about each house (bedrooms, bathrooms, etc.)

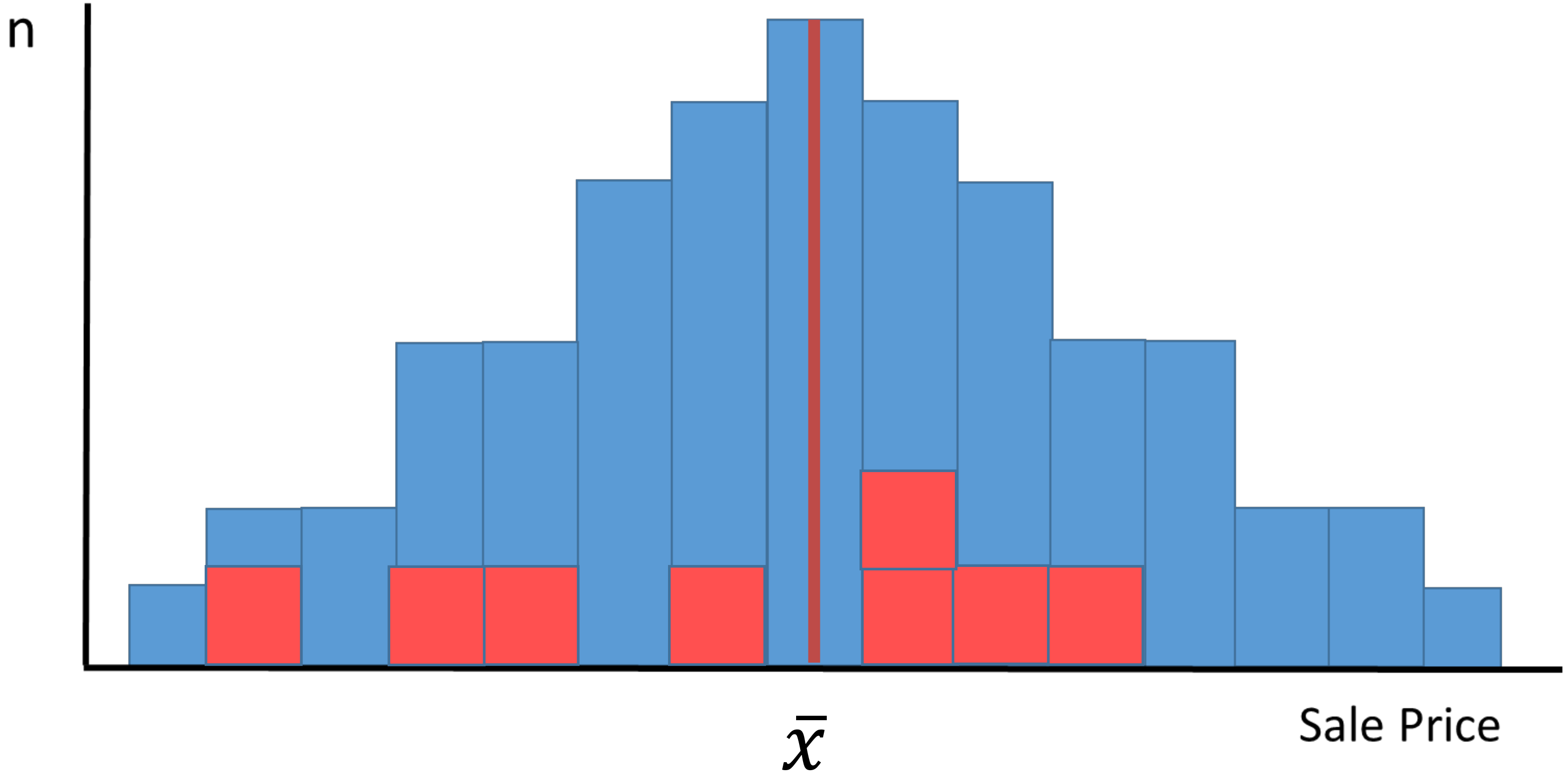
Predicting for the Sale Price of a House

What is our best prediction for the sale price of the house?



Predicting for the Sale Price of a House

What is our best prediction for the sale price of the house?



Predicting for the Sale Price of a House

Given no other information, the mean is our best prediction for each house.

We do not expect our predictions to be perfect.

Prediction Error

(predicted value - actual value)

In the long run, predicting with the mean will result in the lowest total prediction error.

Predictions are just mean values

Like deviation, prediction errors need to be squared to be useful.

If we square the correlation coefficient (r) we have R-Squared, a common regression modeling metric.

R-Squared is more formally known as the coefficient of determination.

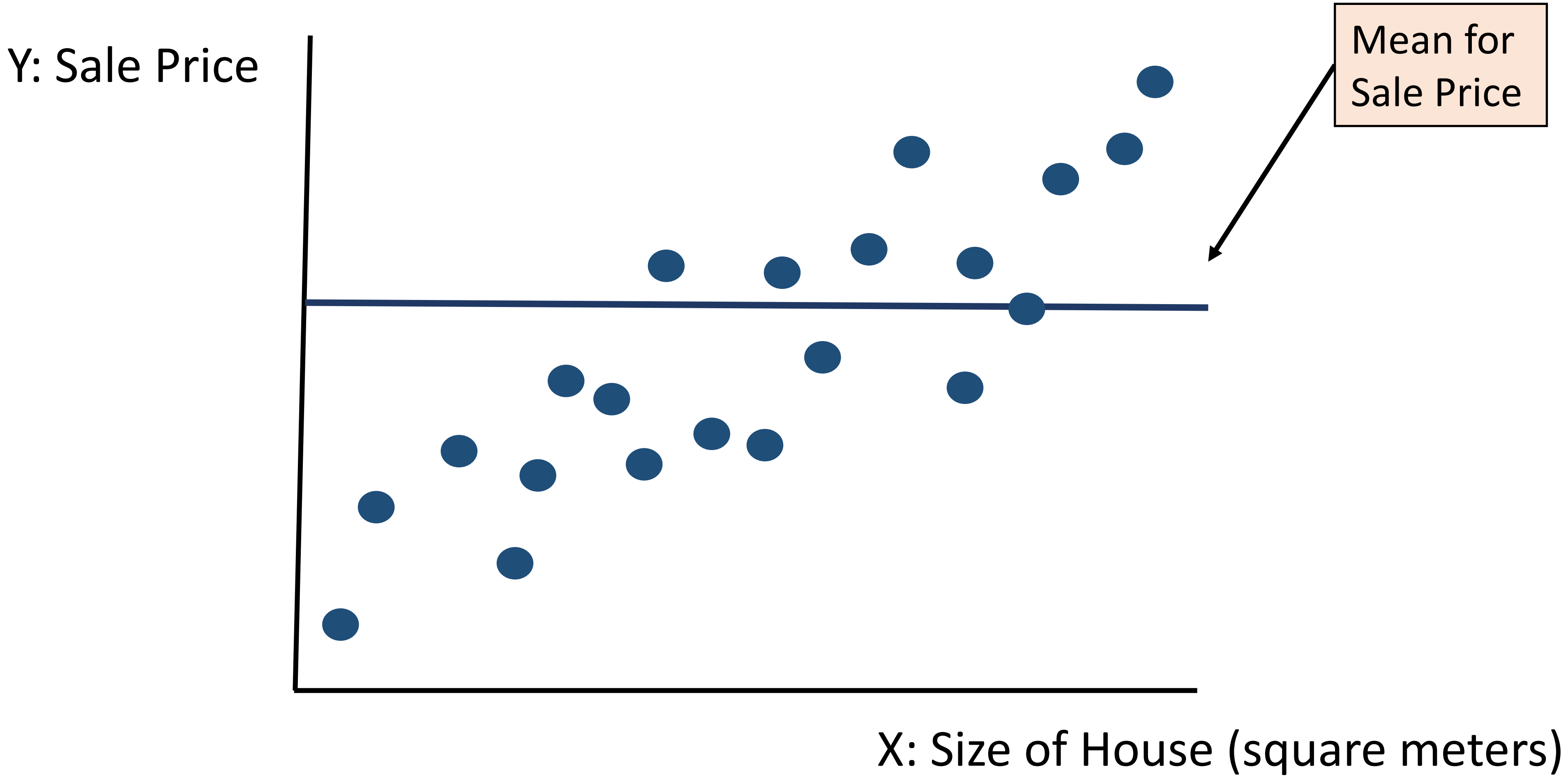
Coefficient of Determination

$$R^2 = 1 - \frac{\text{Squared Errors Using Regression}}{\text{Squared Errors using Mean}}$$

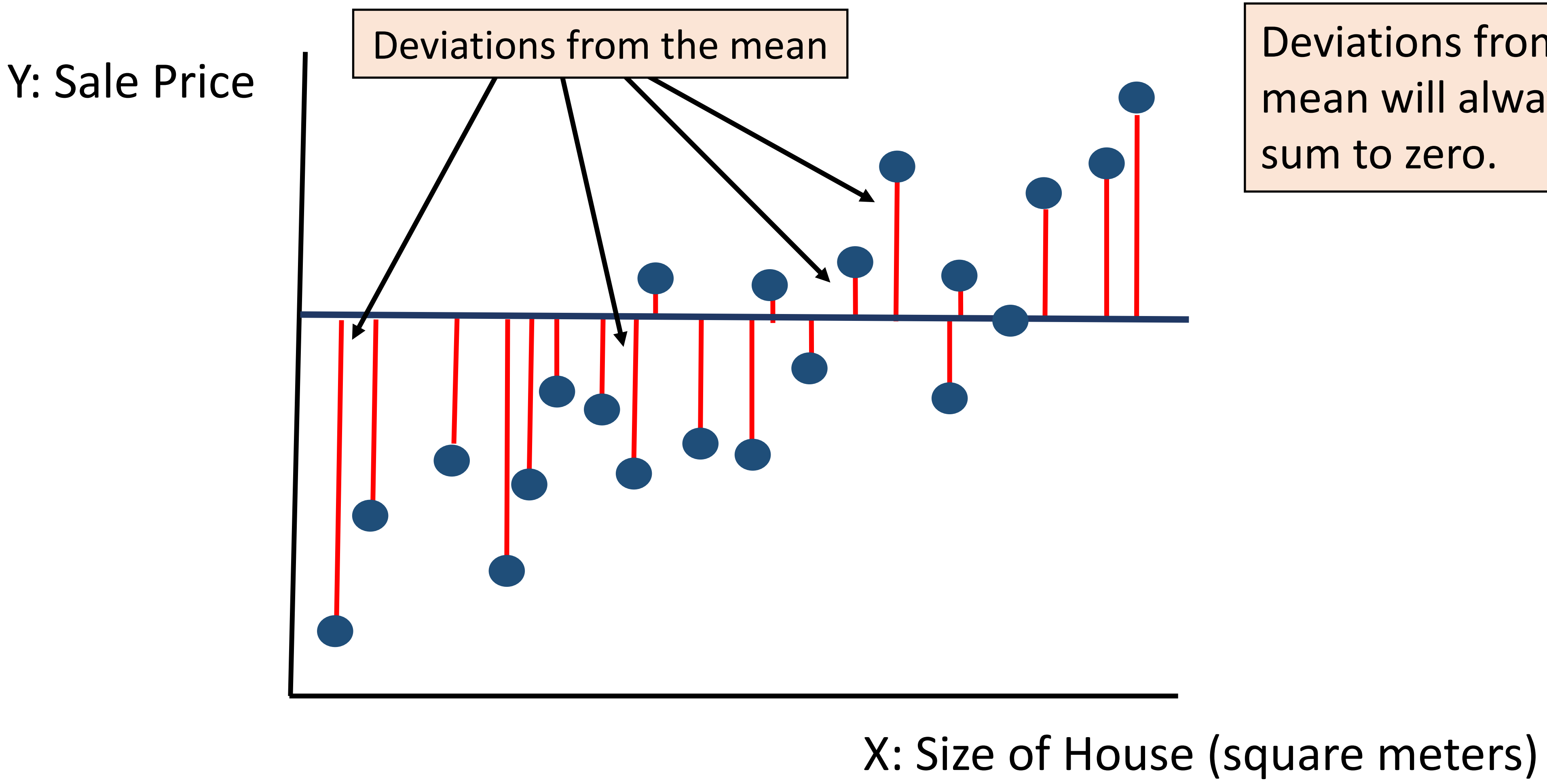
Answers the question:

How much better is our model compared to using the mean of Y to predict?

How Linear Regression Works

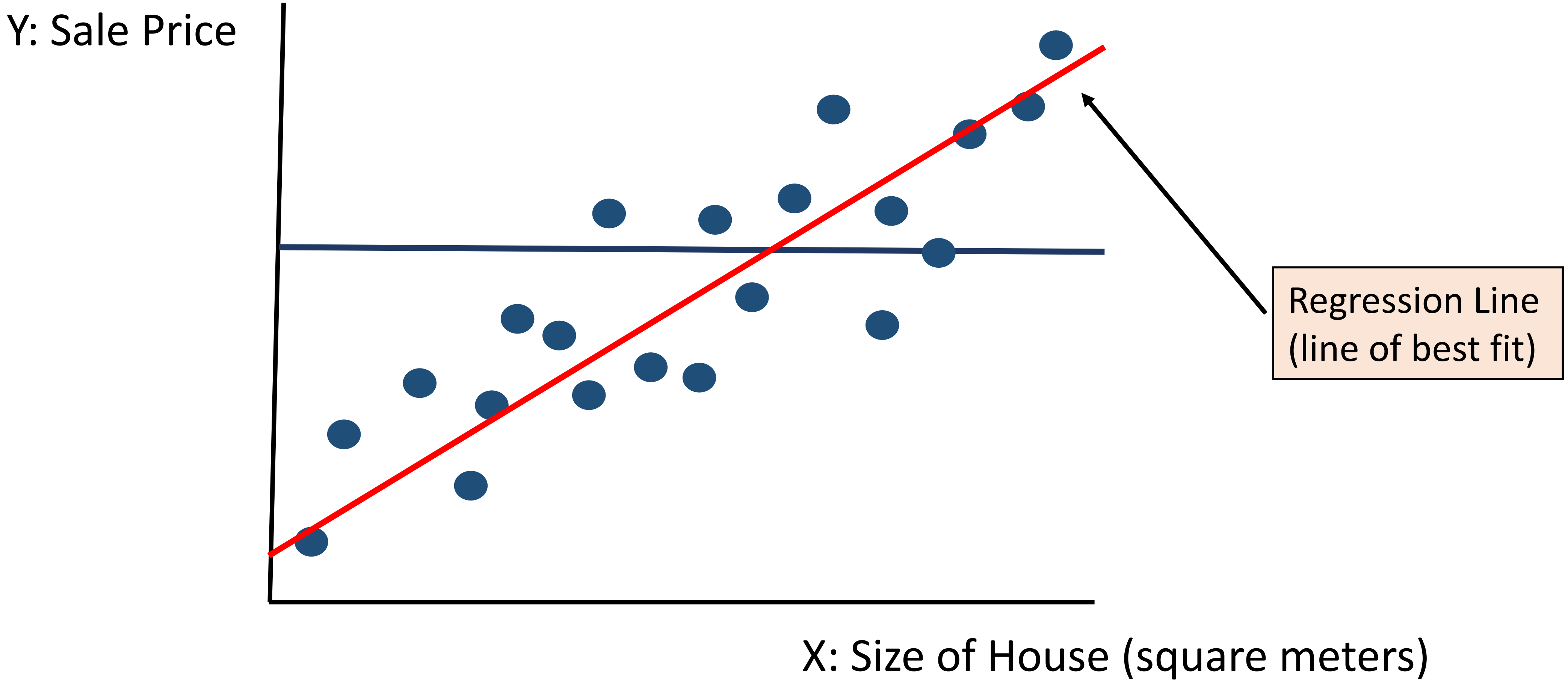


How Linear Regression Works



Deviations from the mean will always sum to zero.

How Linear Regression Works



How Linear Regression Works

Linear regression simply adds slope to the mean line

The regression line minimizes the squared deviations from the data

The quality of this line is measured with R-Squared

$$R^2 = 1 - \frac{\text{Squared Errors Using Regression}}{\text{Squared Errors using Mean}}$$

$$R^2 = \text{Percentage reduction in squared errors}$$

How Linear Regression Works

Linear regression simply adds slope to the mean line

The regression line minimizes the squared deviations from the data

The quality of this line is measured with R-Squared

$$0 \leq R^2 \leq 1$$

$R^2 \approx 0$	Little variation explained
$R^2 \approx 1$	A lot of variation explained

Adjusted R-Squared

R-squared can never decrease with the addition of new variables to a model.

Adjusted R-square penalizes models when they contain variables that do not add value.

Adjusted R-square is suitable when comparing models of different sizes.