Simple Linear Regression

Objectives

After completing this chapter, you should be able to:

- Explain the simple linear regression model
- Obtain and interpret the simple linear regression equation for a set of data
- Hands on exercise

Correlation vs. Regression

- A scatter plot (or scatter diagram) can be used to show the relationship between two variables
- Correlation analysis is used to measure strength of the association (linear relationship) between two variables
  - Correlation is only concerned with strength of the relationship
  - No causal effect is implied with correlation
  - Correlation was first presented in Chapter 3

Introduction to Regression Analysis

- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

Dependent variable: the variable we wish to explain

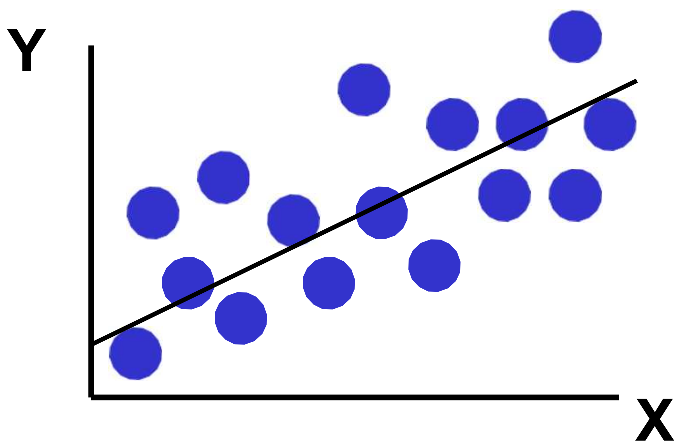
Independent variable: the variable used to explain the dependent variable

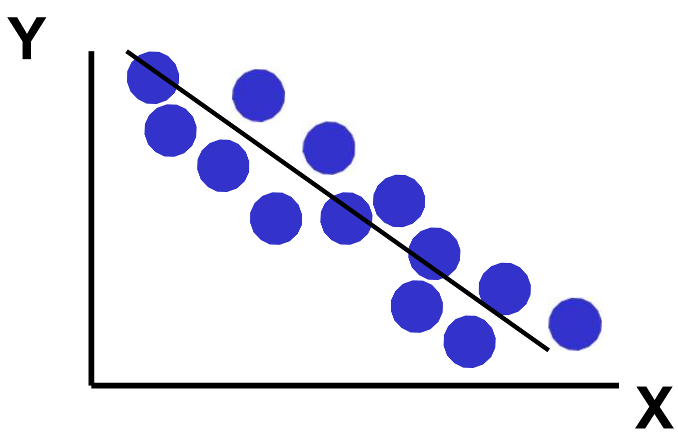
Simple Linear Regression Model

- Only **one** independent variable, X
- Relationship between X and Y is described by a linear function
- Changes in Y are assumed to be caused by changes in X

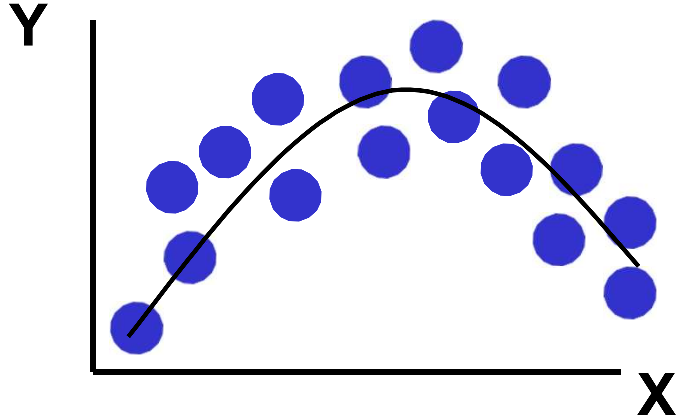
Types of Relationships

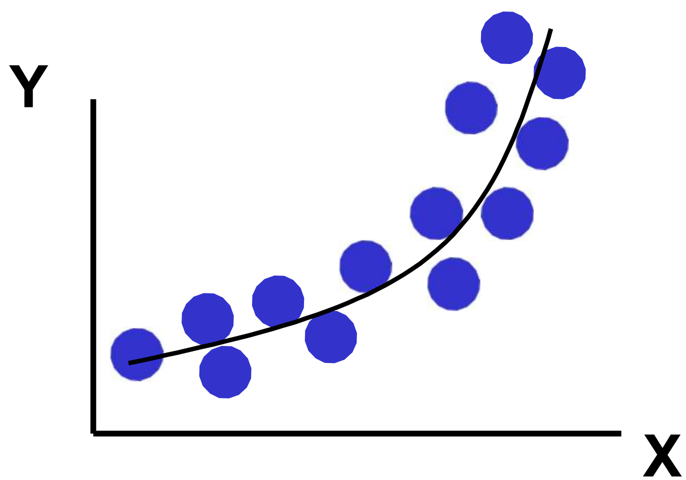
Linear relationships



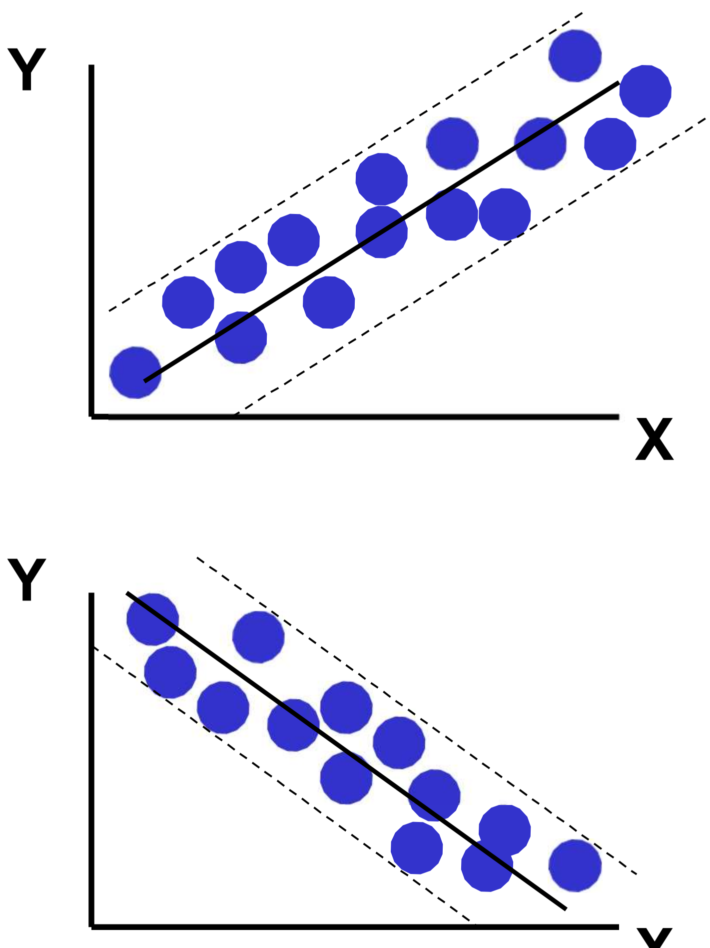


Curvilinear relationships

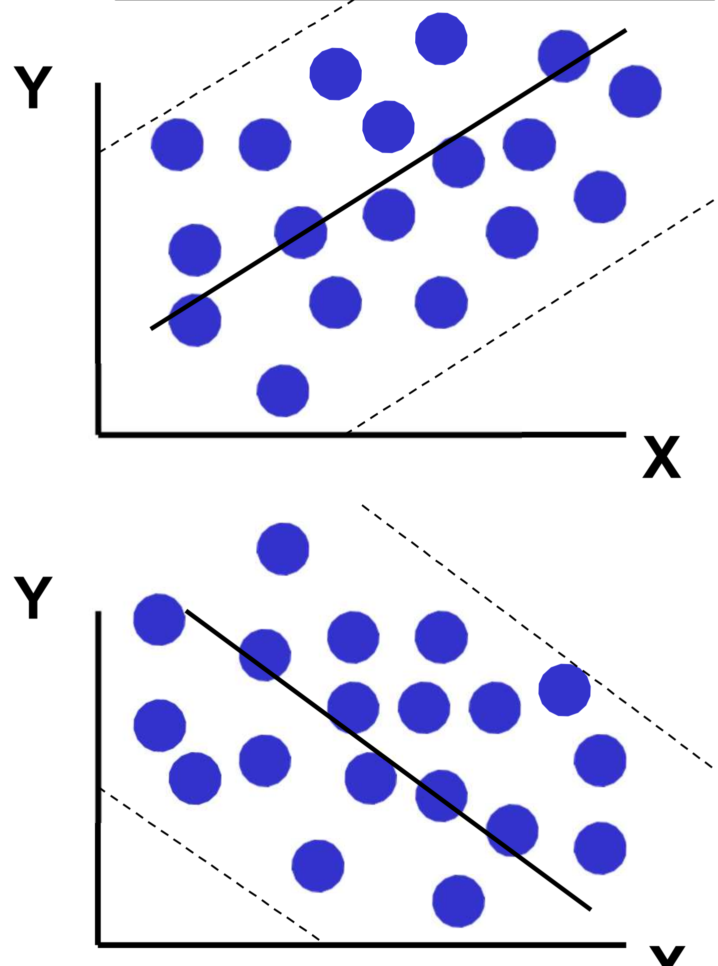




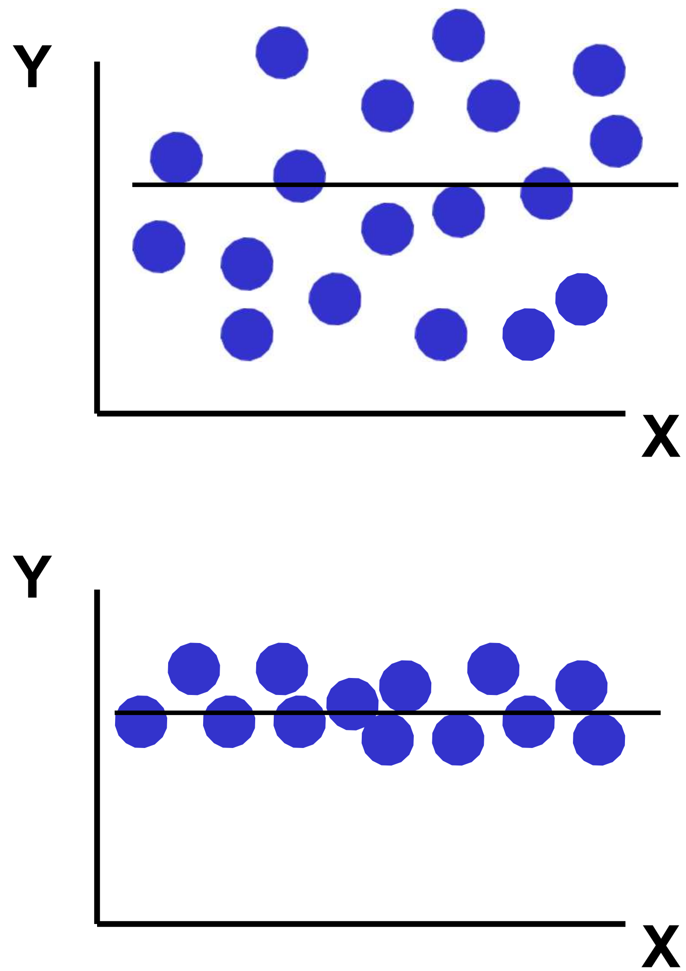
Strong relationships



Weak relationships



No relationship



Simple Linear Regression Model

The population regression model:

Population Y intercept

Population Slope Coefficient

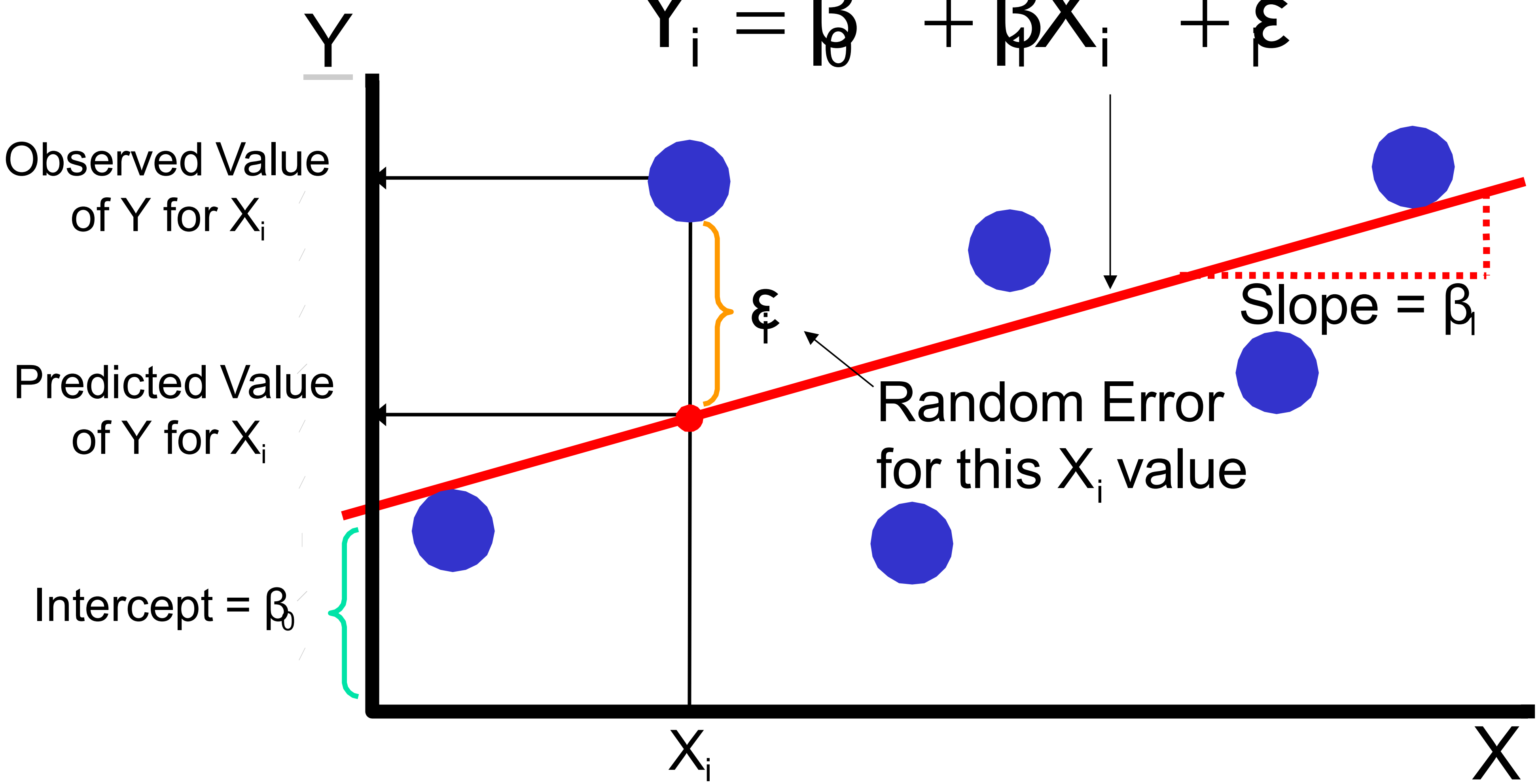
Independent Variable

Random Error term

Dependent Variable

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



Simple Linear Regression Equation

The simple linear regression equation provides an estimate of the population regression line

Estimated (or predicted) Y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

Value of X for observation i

$$\hat{Y}_i = b_0 + b_1 X_i$$

The individual random error terms e_i have a mean of zero

Least Squares Method

- b_0 and b_1 are obtained by finding the values of b_0 and b_1 that minimize the sum of the squared differences between Y and $\hat{Y}$:

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2$$

Interpretation of Slope and Intercept

- b_0 is the estimated average value of Y when the value of X is zero

- b_1 is the estimated change in the average value of Y as a result of a one-unit change in X

Simple Linear Regression Example

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)

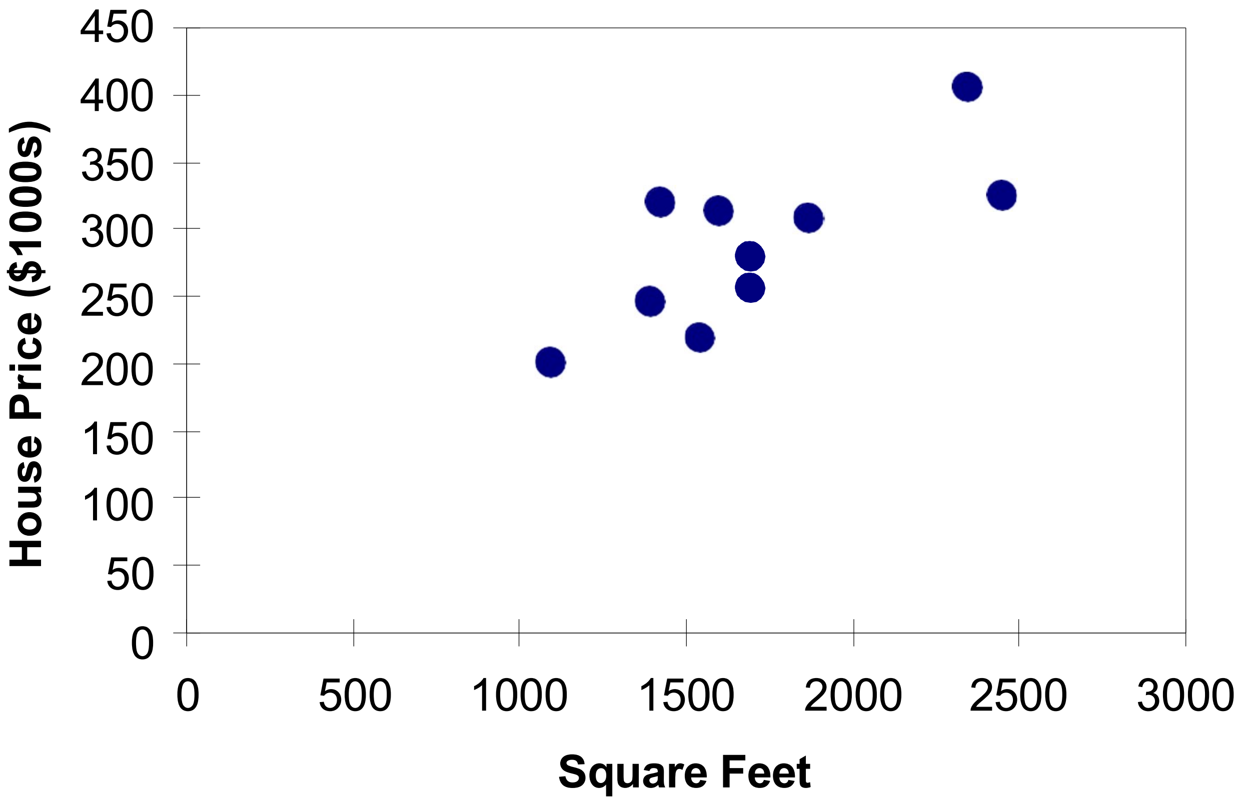
- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in $1000s
  - Independent variable (X) = square feet

Sample Data for House Price Model

House Price in $1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

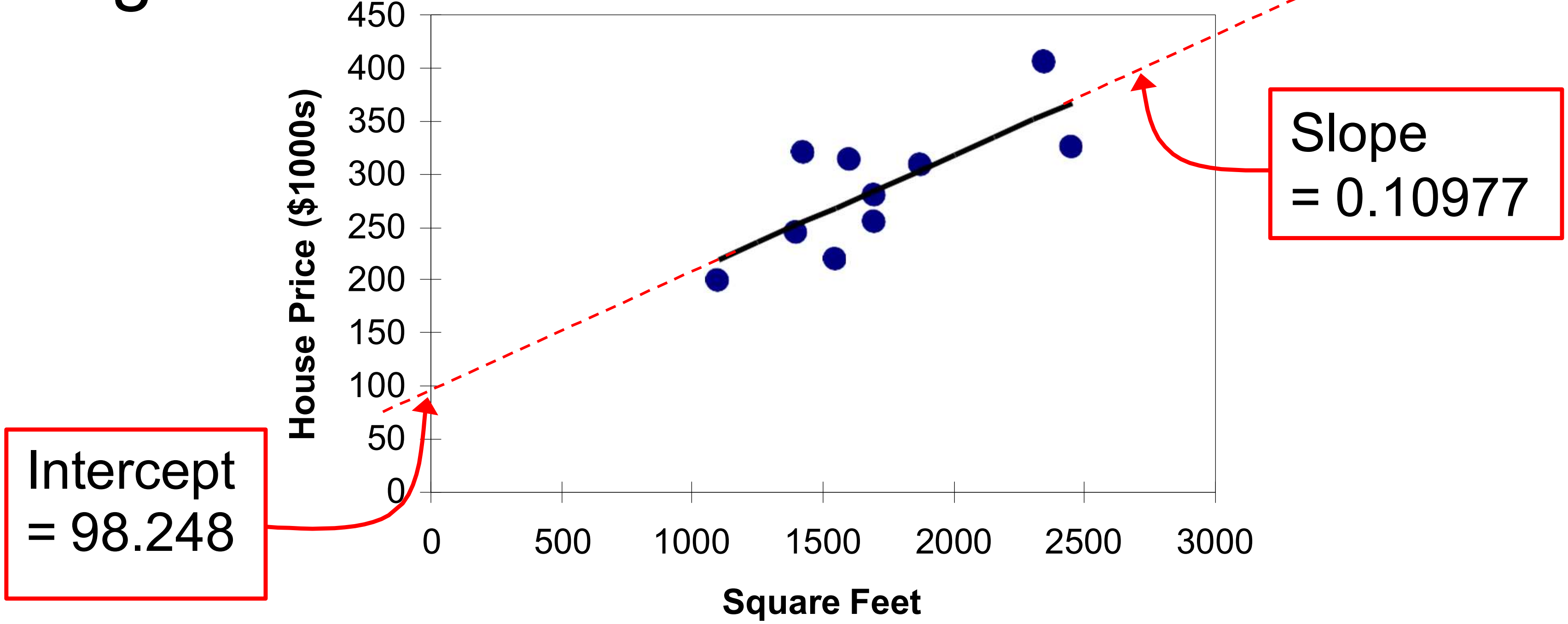
Graphical Presentation

- House price model: scatter plot



Graphical Presentation

- House price model: scatter plot and regression line



$$\widehat{\text{house price}} = 98.24833 + 0.10977 \text{ (square feet)}$$

Interpretation of the Intercept b_0

$$\widehat{\text{house price}} = \boxed{98.24833} + 0.10977\,(\text{square feet})$$

- b_0 is the estimated average value of Y when the value of X is zero (if X = 0 is in the range of observed X values)
  - Here, no houses had 0 square feet, so $\boxed{b_0 = 98.24833}$ just indicates that, for houses within the range of sizes observed, $98,248.33 is the portion of the house price not explained by square feet

Interpretation of the Slope Coefficient b_1

$$\widehat{\text{house price}} = 98.24833 + \boxed{0.10977}\,(\text{square feet})$$

- b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X
  - Here, $\boxed{b_1 = .10977}$ tells us that the average value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

Predictions using Regression Analysis

Predict the price for a house with 2000 square feet:

$$\widehat{\text{house price}} = 98.25 + 0.1098\,(\text{sq.ft.})$$

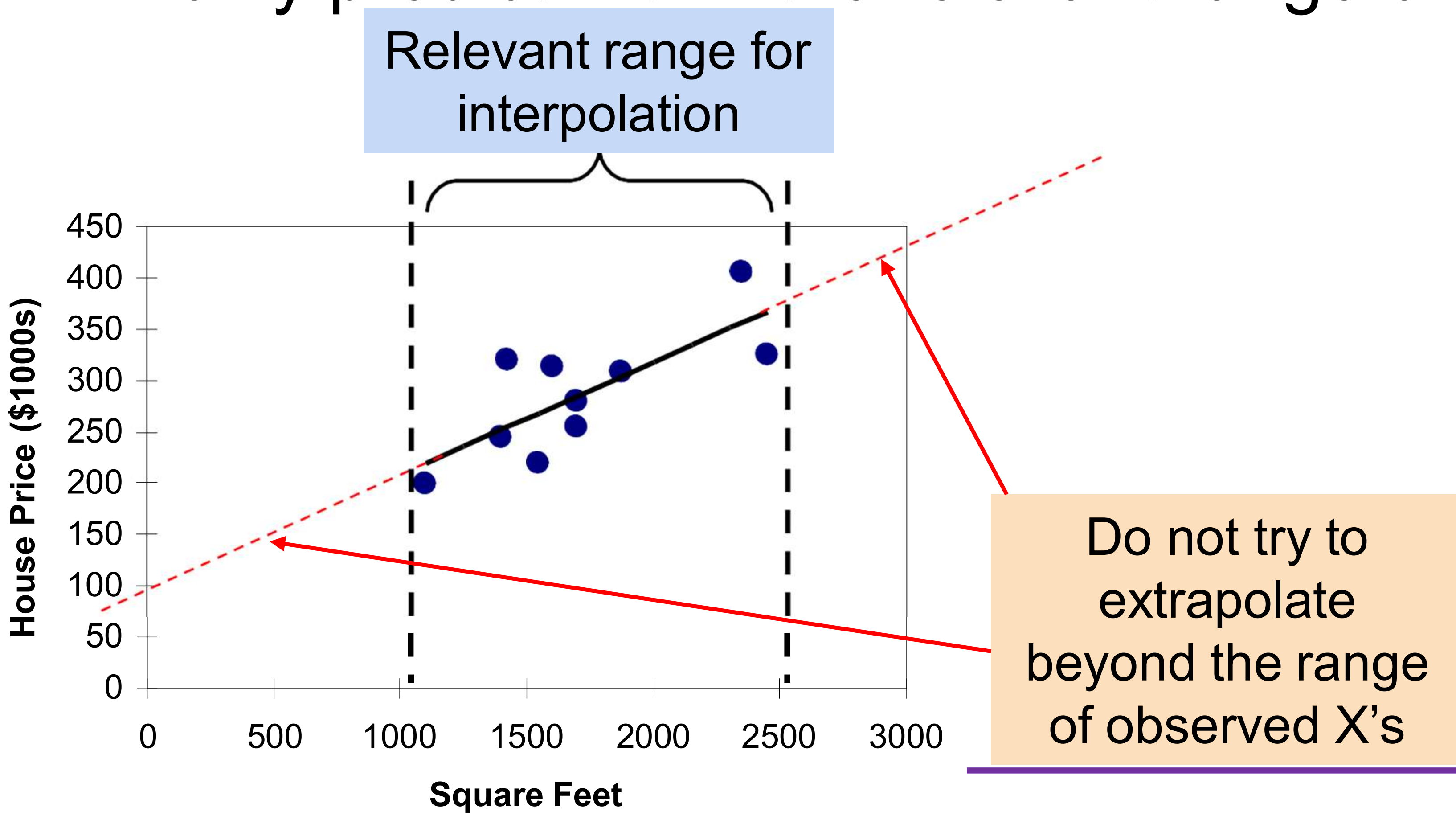
$$= 98.25 + 0.1098(2000)$$

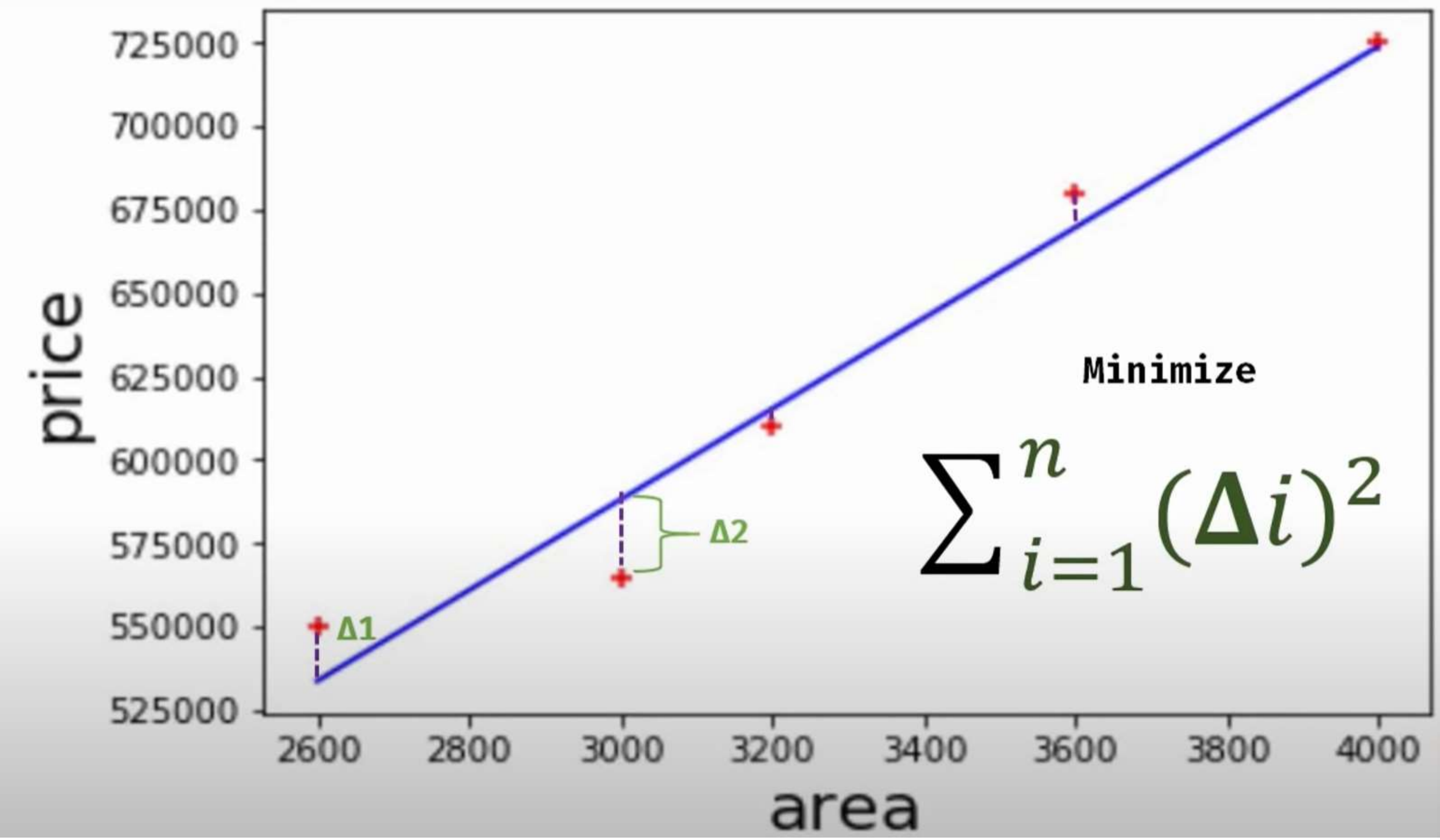
$$= 317.85$$

The predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,850

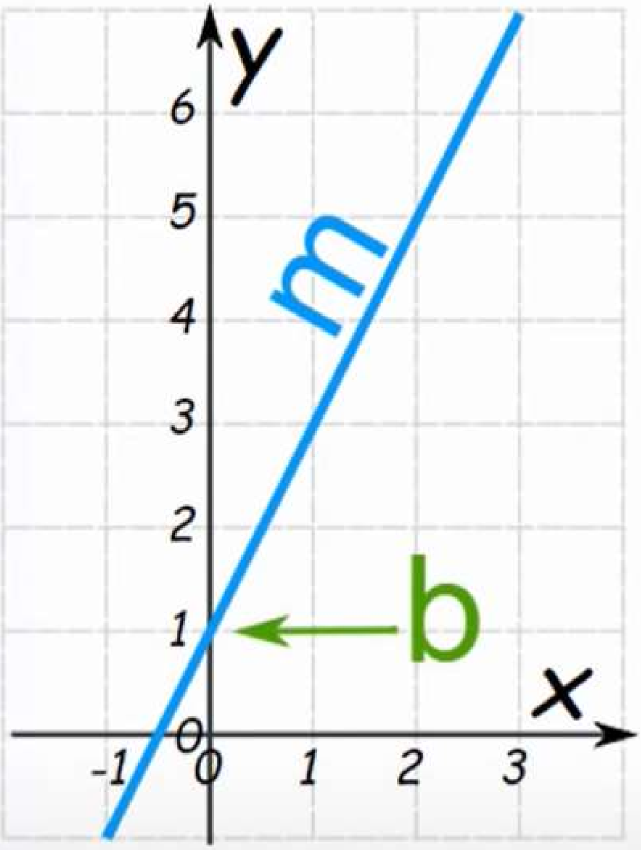
Interpolation vs. Extrapolation

- When using a regression model for prediction, only predict within the relevant range of data





725000
700000
675000
650000
625000
600000
575000
550000
525000
price
2600
2800
3000
3200
3400
3600
3800
4000
area
Δ1
Δ2
Minimize
$\sum_{i=1}^{n} (\Delta i)^2$



price = m * area + b

$$y = mx + b$$

Slope (or Gradient) Y Intercept

price = m * area + b

Dependent variable Independent variable

Scikit-learn

- Scikit-learn (formerly scikits.learn and also known as sklearn) is a free software machine learning library for the Python programming language
- It features various classification, regression and clustering algorithms including support-vector machines, random forests, gradient boosting, k-means and DBSCAN
- is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy

Summary

- *Introduced linear regression model*
- *Build linear regression model using jyputer notebook*

Himanshu Patel, Instructor
Saskatchewan Polytechnic
email: patelh@saskpolytech.ca
Mining Building, Saskatoon