Regression

- **Duration:** 2 hrs
- **Outline:**

  1. What is regression?
  2. Simple linear regression
  3. Example

Regression

- **Duration:** 2 hrs
- **Outline:**

  1. What is regression?

  2. Simple linear regression

  3. Example

Regression

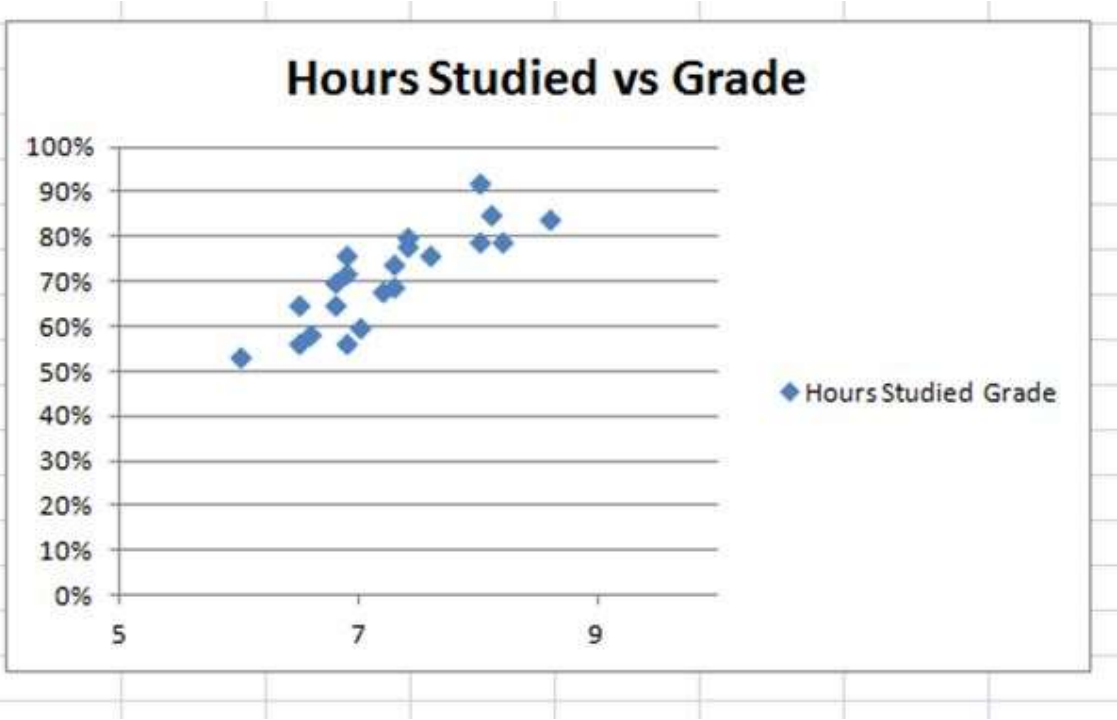
- What is regression analysis?
  - A basic and commonly used predictive analysis (predictive analysis: the use of data, statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data)
- Goals:
  - To establish the possible relationship between variables
  - To predict a real-valued label (target) given an unlabeled example

Regression

	A	B	C	D	E
1	area	bedrooms	balcony	age	price
2	1200	2	0	2	500000
3	2300	3	2	5	620000
4	2500	4	2	1	122500
5	3650	5	3	3	6000000
6	1800	3	1	5	2122000
7	3000	3	1	4	120000
8	1222	1	0	2	450000
9	4600	5	3	1	6500000
10	2050	2	2	2	1530000
11	1450	2	2	3	1563330

Regression

Student Name	Hours Studied	Grade
Jack	6	53%
Anne	7	60%
Harry	6,5	56%
Sharon	8	79%
John	6,6	58%
James	8,1	85%
Jill	6,8	70%
Adam	6,9	56%
Brandon	7,3	69%
Brett	6,9	76%
Brady	8,2	79%
Charles	7,2	68%
Darren	7,3	74%
Dave	6,9	72%
Dawn	8,6	84%
Denise	7,4	78%
Eric	7,6	76%
Emily	6,8	65%
Fred	8	92%
Fran	7,4	80%
Jane	6,5	65%



Regression

- Applications: finance, investing, and other disciplines → to determine the strength and character of the relationship between one dependent variable (y) and a series of other independent variables.
- Ex: estimating house price valuation based on house features [area, the number of bedrooms, location,…]

Types of regression

- Regression describes the relationship between variables by fitting a line to the examples.
- **Linear regression:** line is straight
  - **Simple linear regression**: using one variable to predict the outcome
  - **Multiple linear regression**: using two or more independent variables to predict the outcome
- **Nonlinear regression:** line is curved

Regression

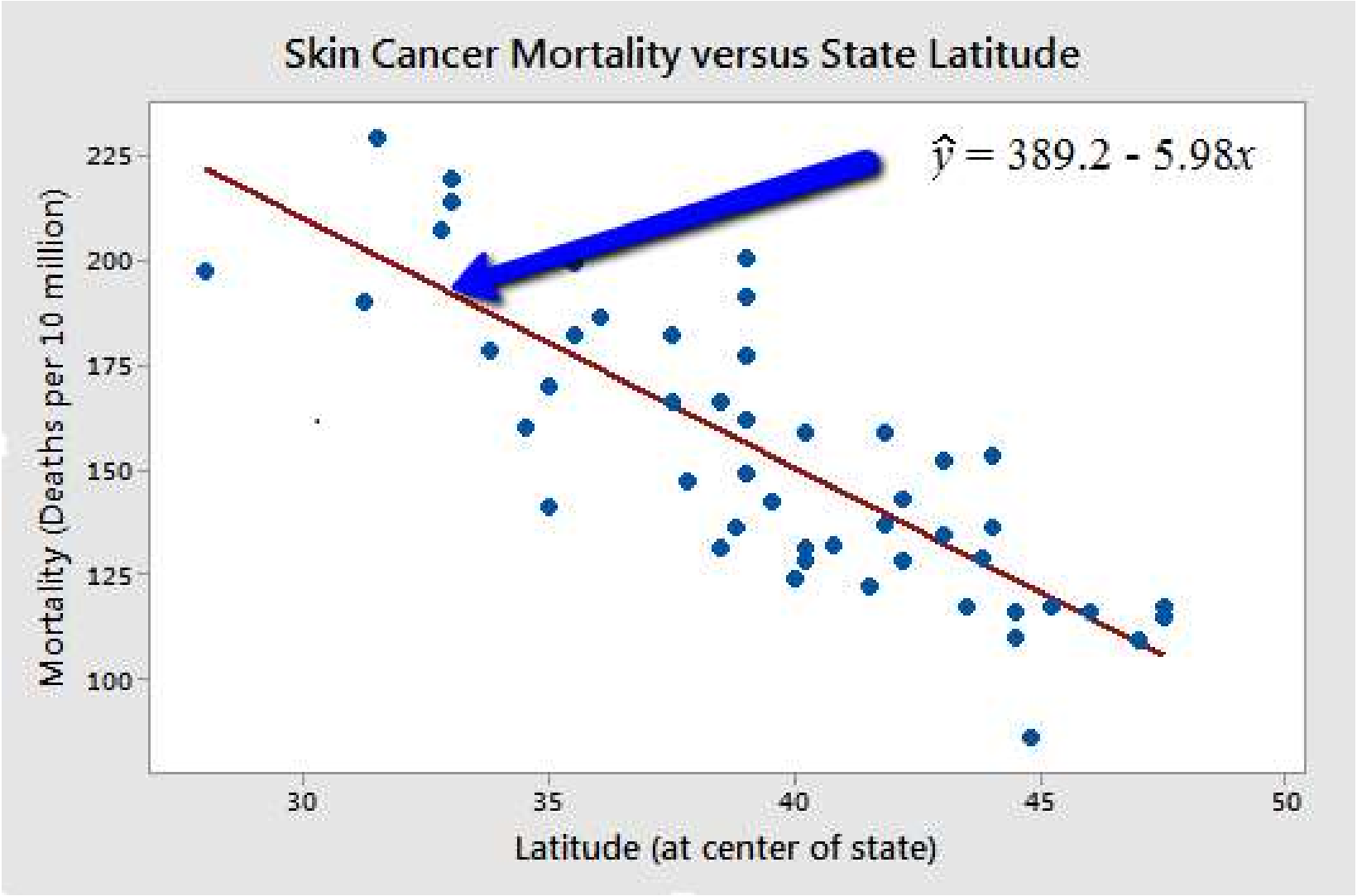
- **Duration:** 2 hrs
- **Outline:**

  1. What is regression?
  2. **Simple linear regression**
  3. Example

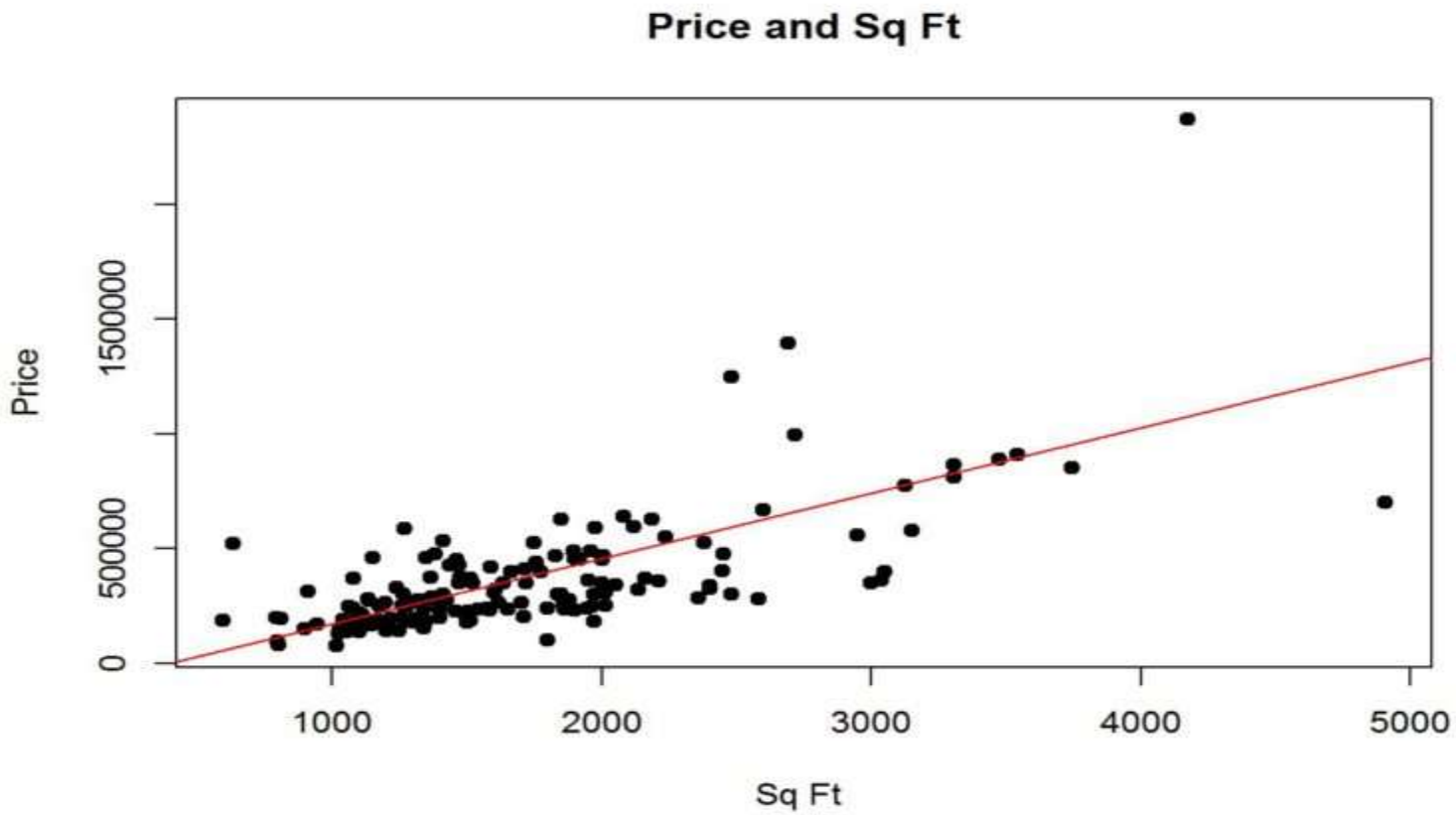
Objectives

- Learning a linear model that summarizes the statistical relationship between two continuous (quantitative) variables
  - Independent variable X, or predictor variable
  - Dependent variable Y, or response
  - Ex: income and spending, student IQ and exam scores
- Linear model: straight line of best fit through the data points
- **Note:** the relationship is statistical, i.e., not perfect, not exact

Simple linear regression model



Simple linear regression model

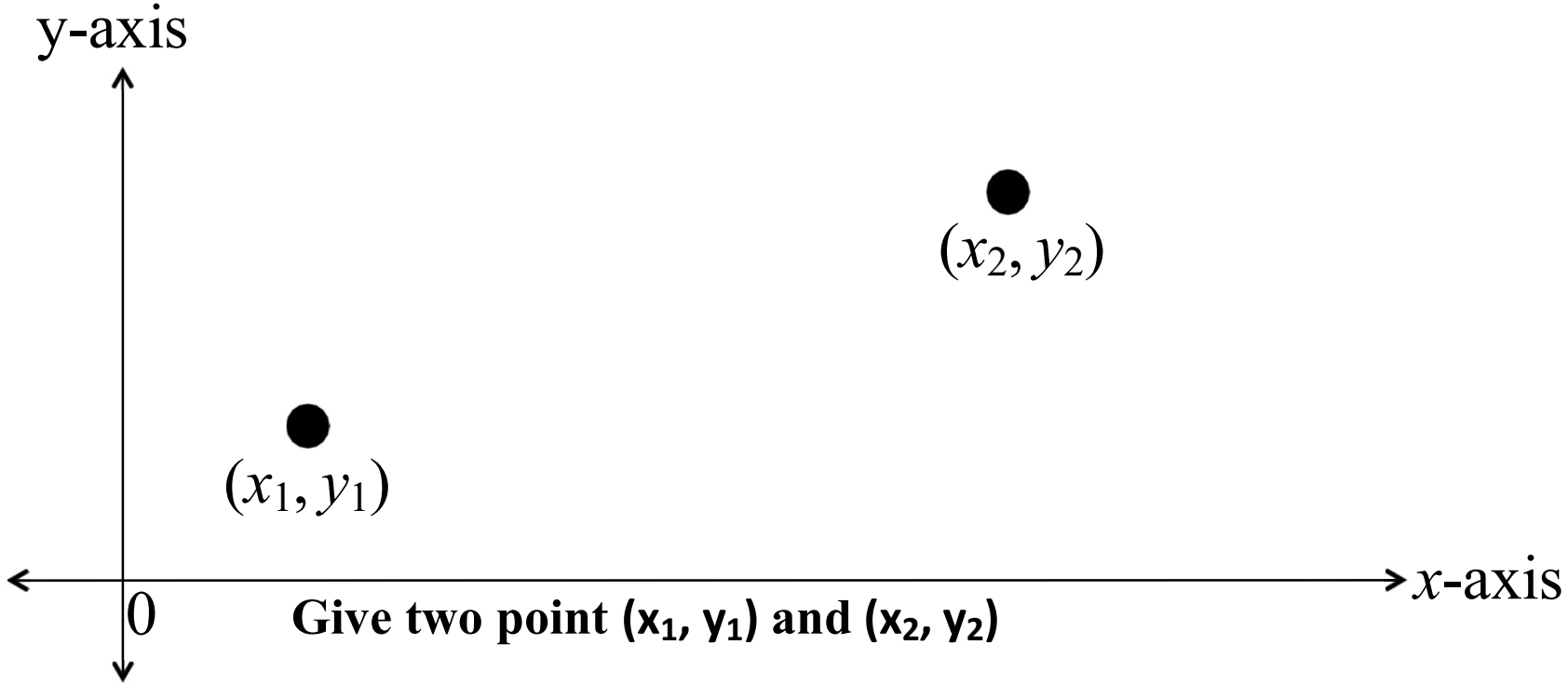


Simple linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

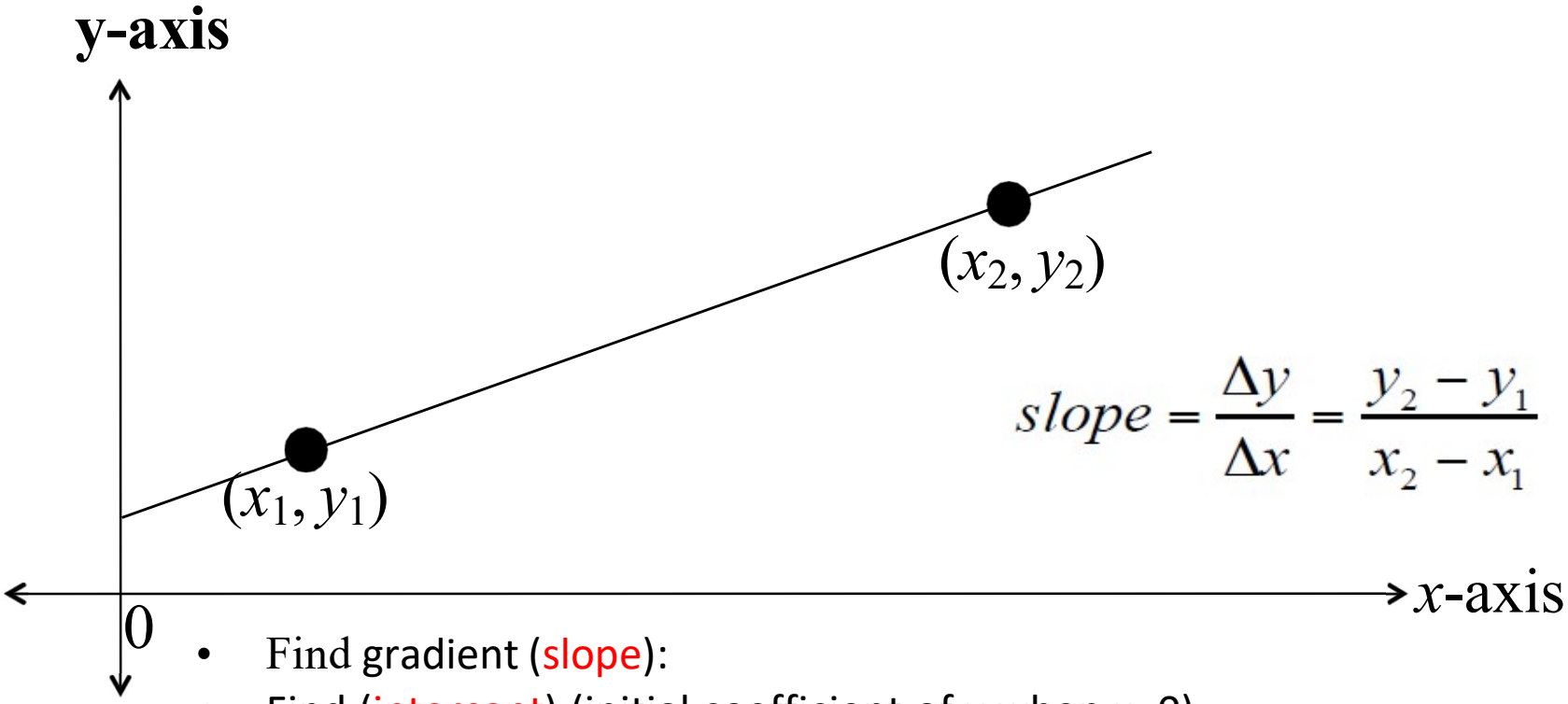
- X: independent variable
- Y: dependent variable
- β_0: *in*tercept
- β_1: slope (gradient)
- ε: random error

Simple linear regression model



How to "develop" an equation connecting these two points?

Simple linear regression model



- Find gradient (slope):
- Find (intercept) (initial coefficient of y when x=0)

Simple linear regression model

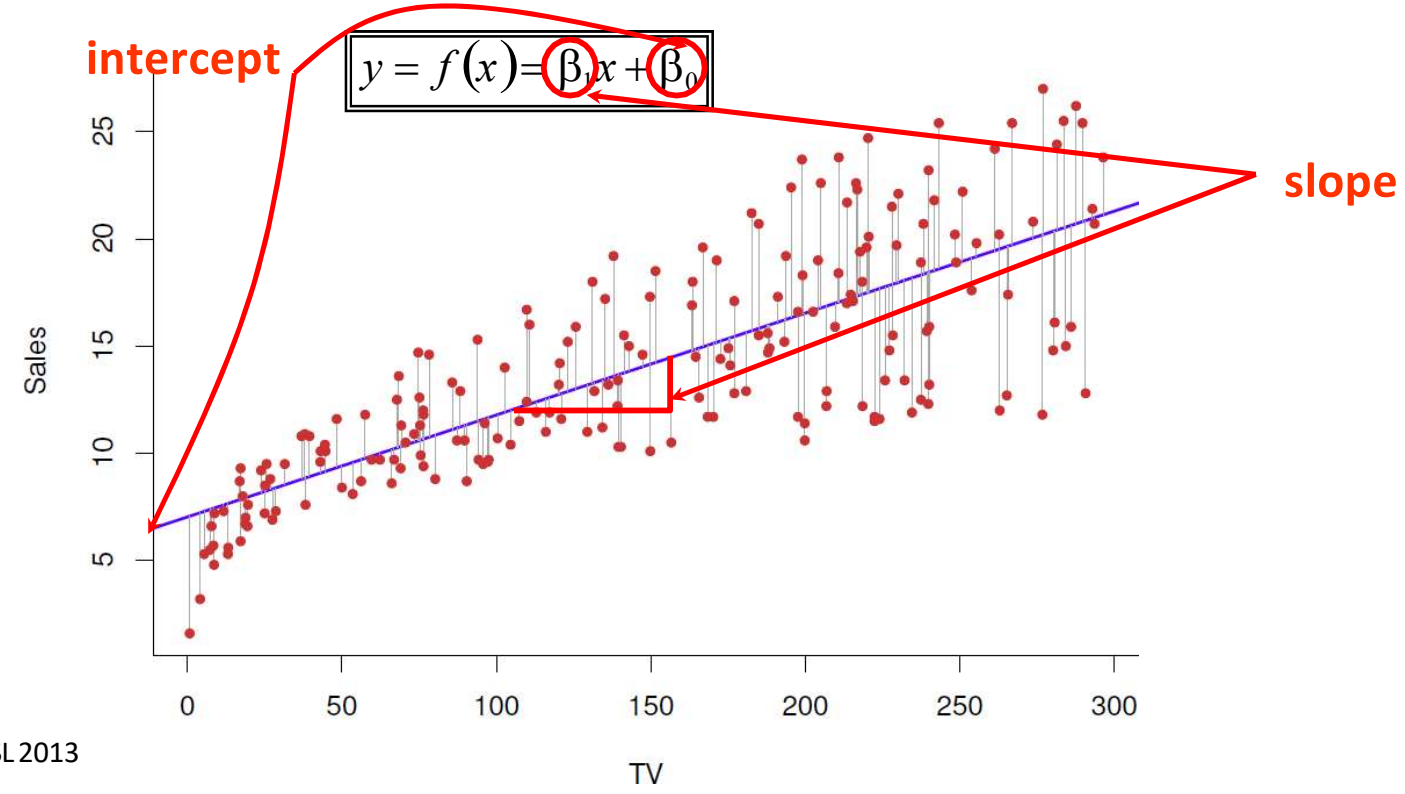
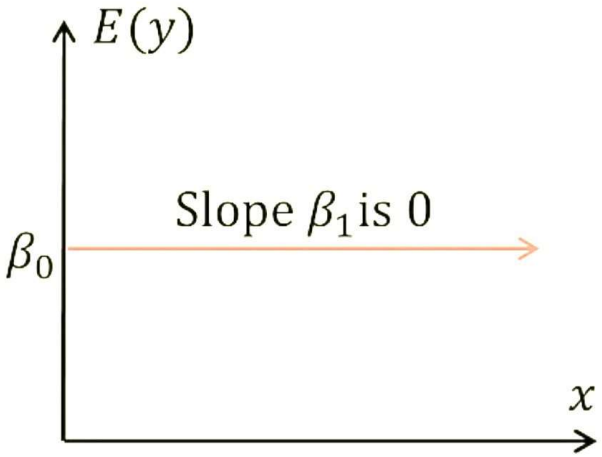
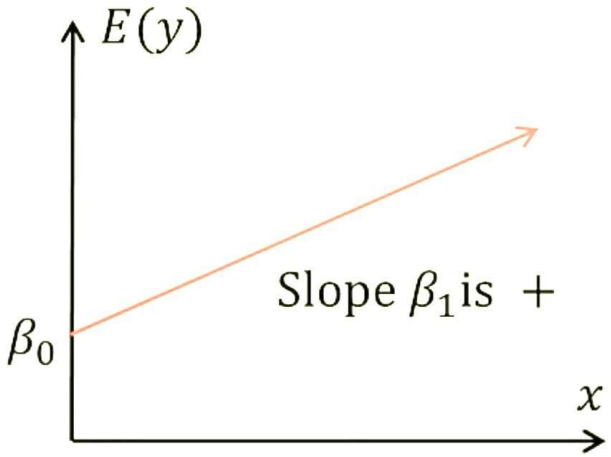


Figure 3.1 , ISL 2013

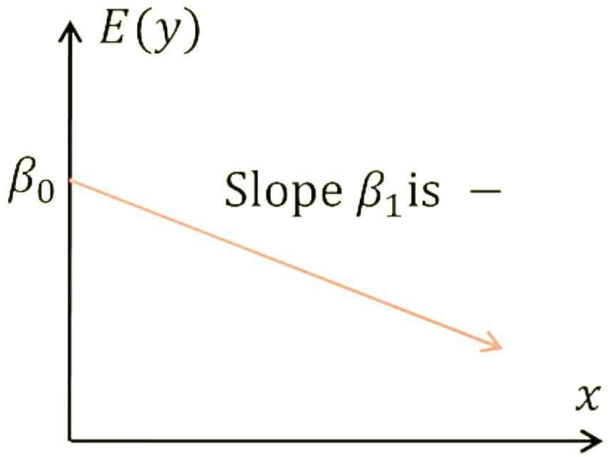
Simple linear regression model



$$E(y) = \beta_0 + 0(x)$$



$$E(y) = \beta_0 + \beta_1 x$$

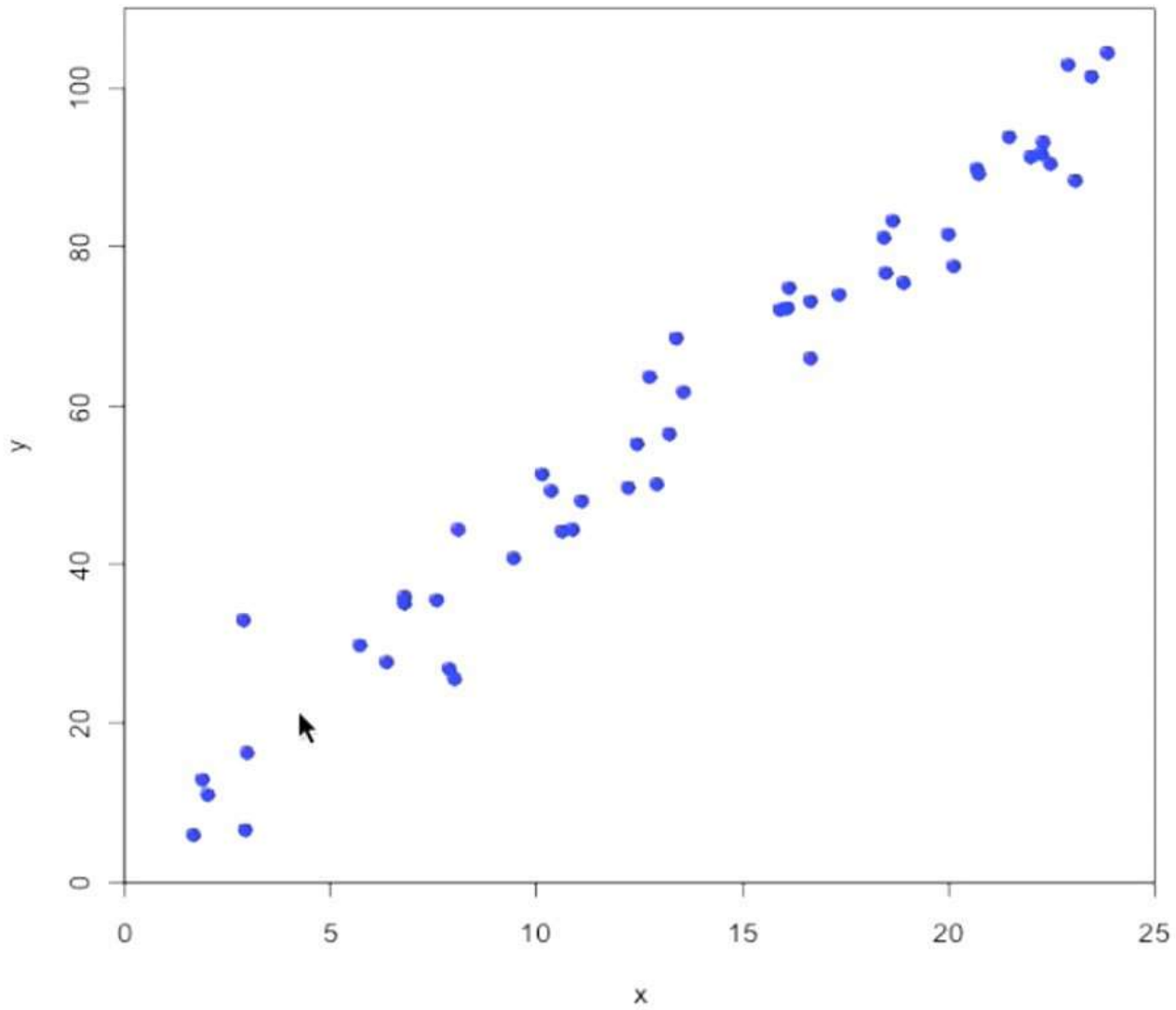


$$E(y) = \beta_0 - \beta_1 x$$

Assumptions of simple linear regression

- **Homogeneity of variance (homoscedasticity)**: The size of the error in the prediction doesn't change significantly across the values of the independent variable.
- **Independence of observations**: the observations in the dataset were collected using statistically valid methods, and there are no hidden relationships among variables.
- **Normality**: the data follows a normal distribution (Gaussian distribution).
- **Linearity**: the line of best fit through the data points is a straight line, rather than a curve or some sort of grouping factor.

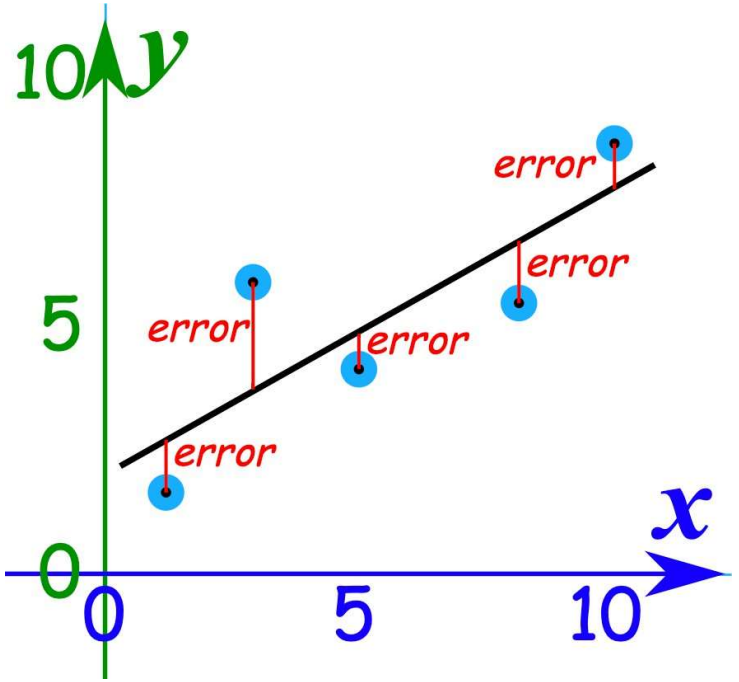
Assumptions of simple linear regression



Simple linear regression problem

- **Goal:** to estimate the linear regression line $\hat{Y} = \beta_0 + \beta_1 X$
- The vertical distance from the data points to the regression line is as small as possible

Y^ = predicted dependent variable (output)
X = independent variable (input)
β0 = **intercept** (the value of Y^ when X=0),
β1 = **slope** (the rate of change of Y^ with respect to X).



Simple linear regression problem

Slope (β_1)

The **slope** β_1 represents how much $\hat{Y}$ changes when X increases by one unit. It is calculated as:

$$\beta_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

where:

- $\bar{X}$ and $\bar{Y}$ are the means of X and Y,
- X_i, Y_i are the individual data points.

Intercept (β_0)

The **intercept** β_0 is the predicted value of $\hat{Y}$ when $X = 0$. It is calculated as:

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

This tells us where the regression line crosses the Y-axis.

Basic idea

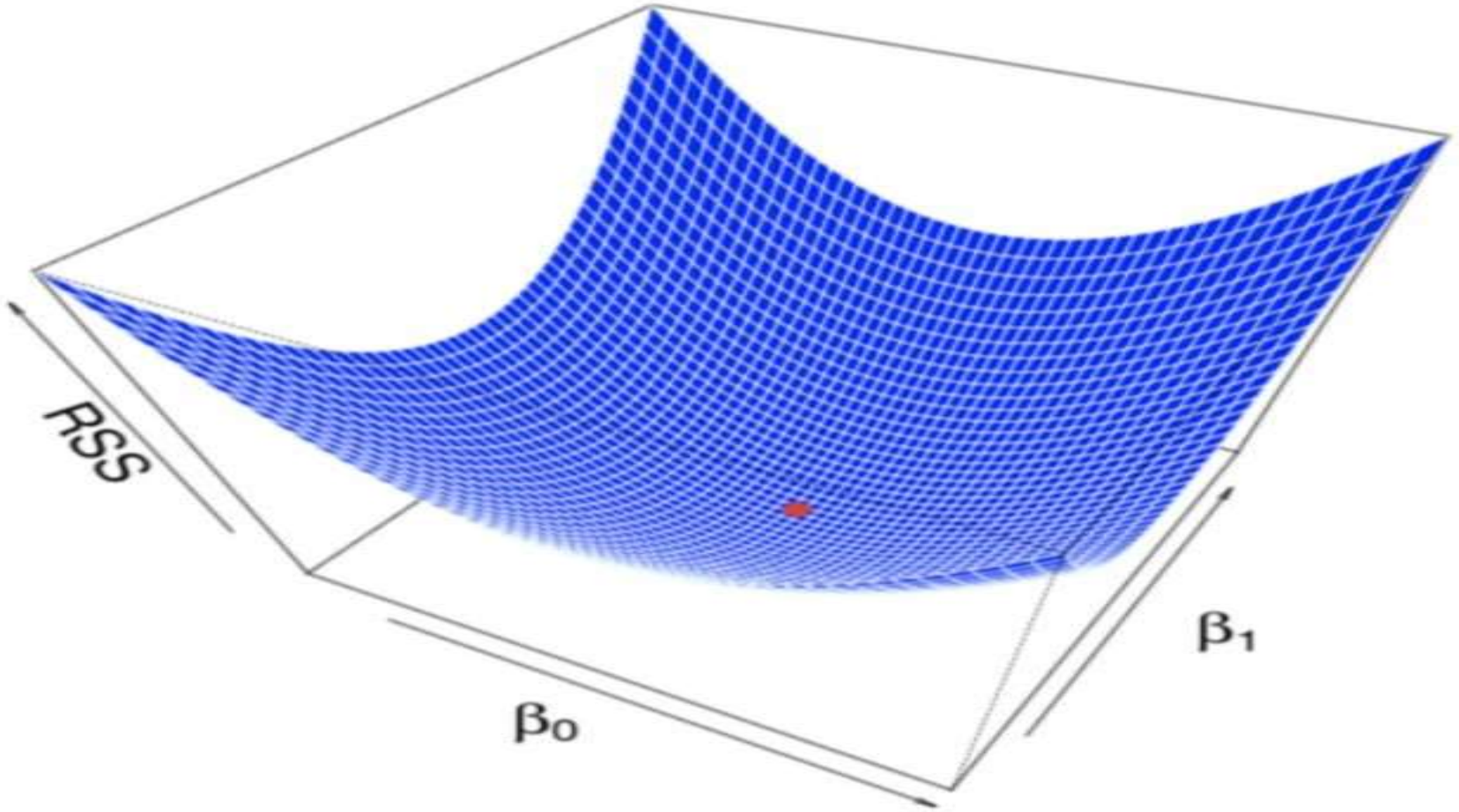
- **Goal:** to find b_0 *and* b_1 to minimize the total of squares of the errors

$$MSE = \frac{1}{n}\sum(Y_i - \hat{Y}_i)^2$$

y_i = observed value of dependent variable

$\hat{y}_i$ = estimated(predicted)value of the dependent variable

Basic idea



Basic idea

- **Goal:** to find β_0 *and* β_1 to minimize the total of squares of the errors

$$MSE = \frac{1}{n}\sum(Y_i - \hat{Y}_i)^2$$

y_i = observed value of dependent variable

$\hat{y}_i$ = estimated(predicted)value of the dependent variable

Regression

- **Duration:** 2 hrs
- **Outline:**

  1. What is regression?
  2. Simple linear regression
  3. **Example**

Restaurant tipping 1

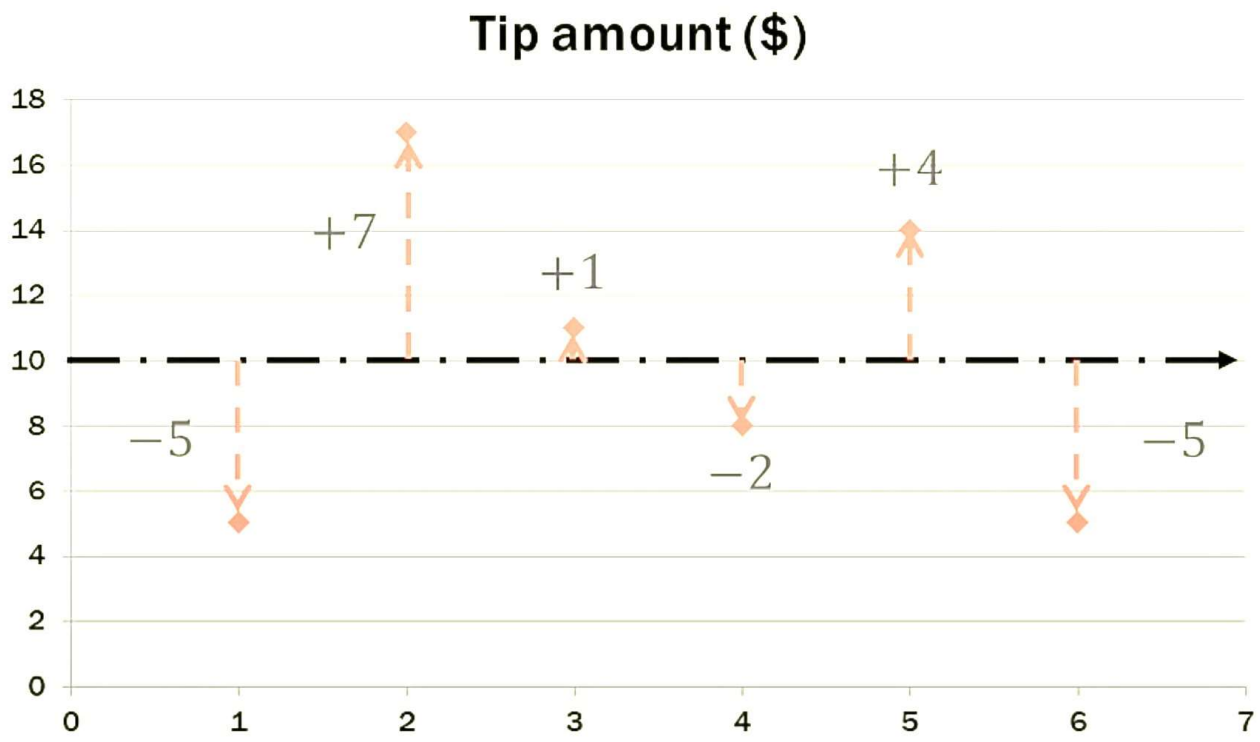
- Predicting the amount of tip a waiter is expected to earn from the restaurant.

Meal#	Tip amount ($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00

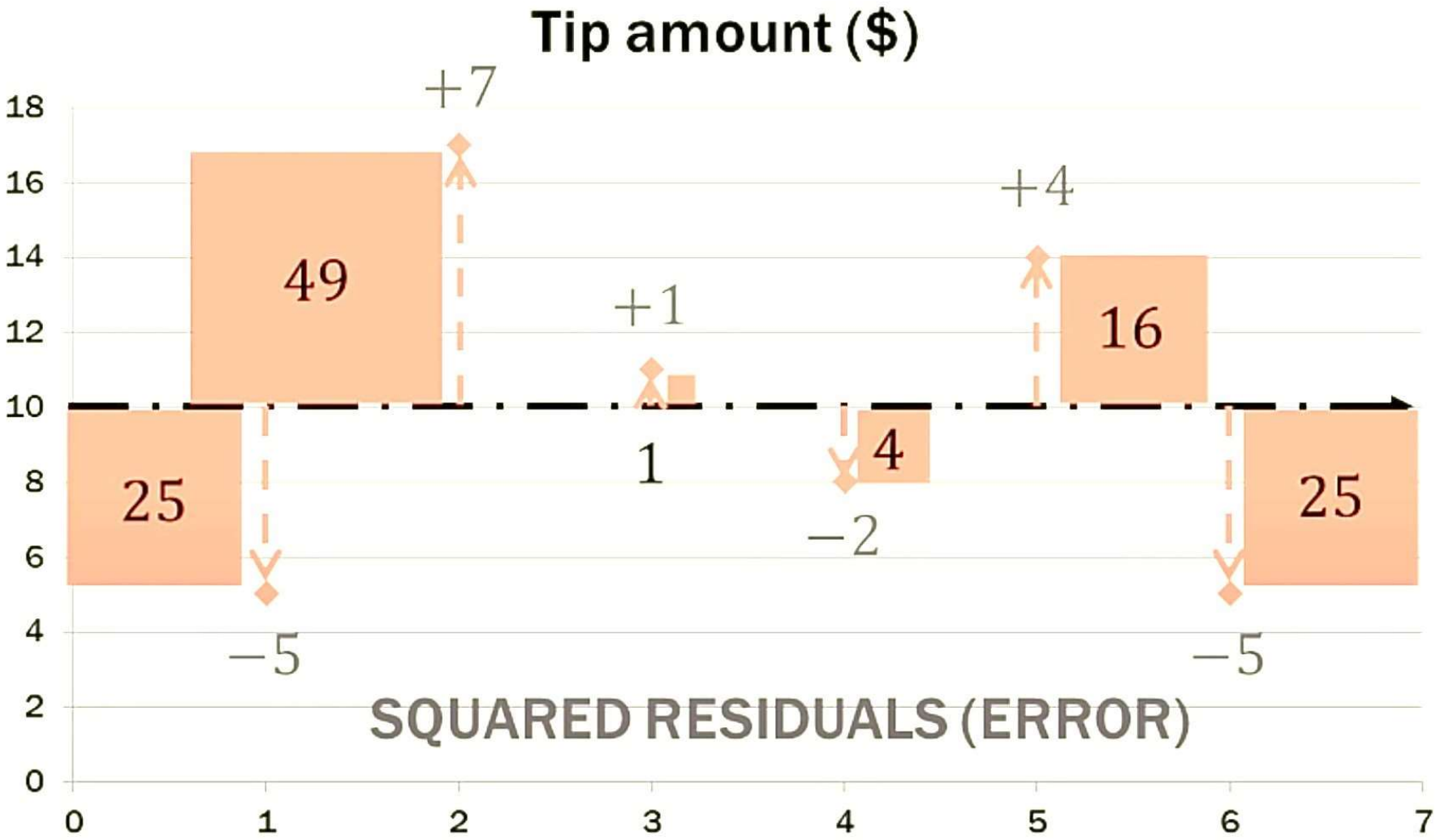
Example (cont)

Meal#	Tip amount ($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00

$$\bar{y} = \$10$$



Example (cont)



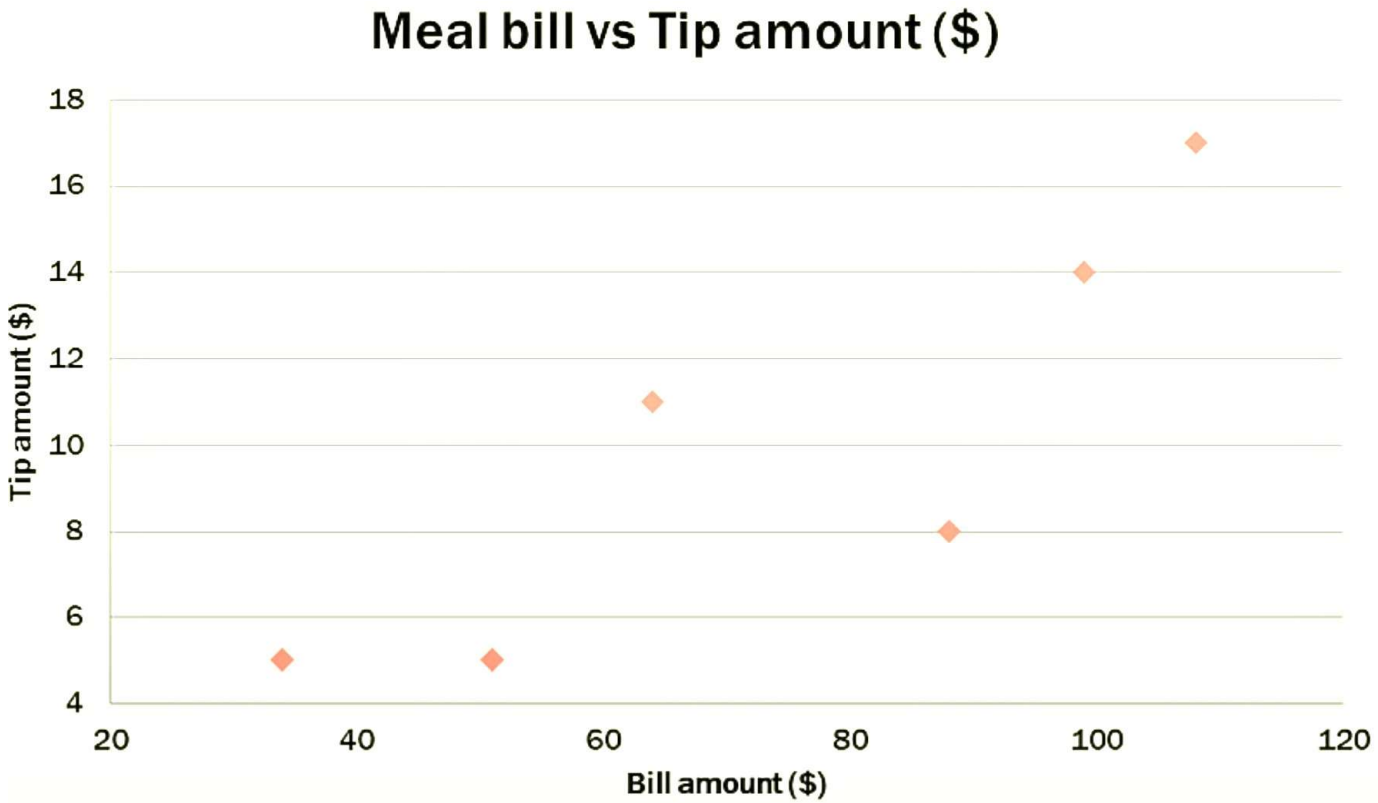
$$Sum\ of\ squared\ errors\ (SSE) = 120$$

Restaurant tipping 2

- Developing a simple linear regression model to predict the amount of tip a waiter can earn given the total bill paid.

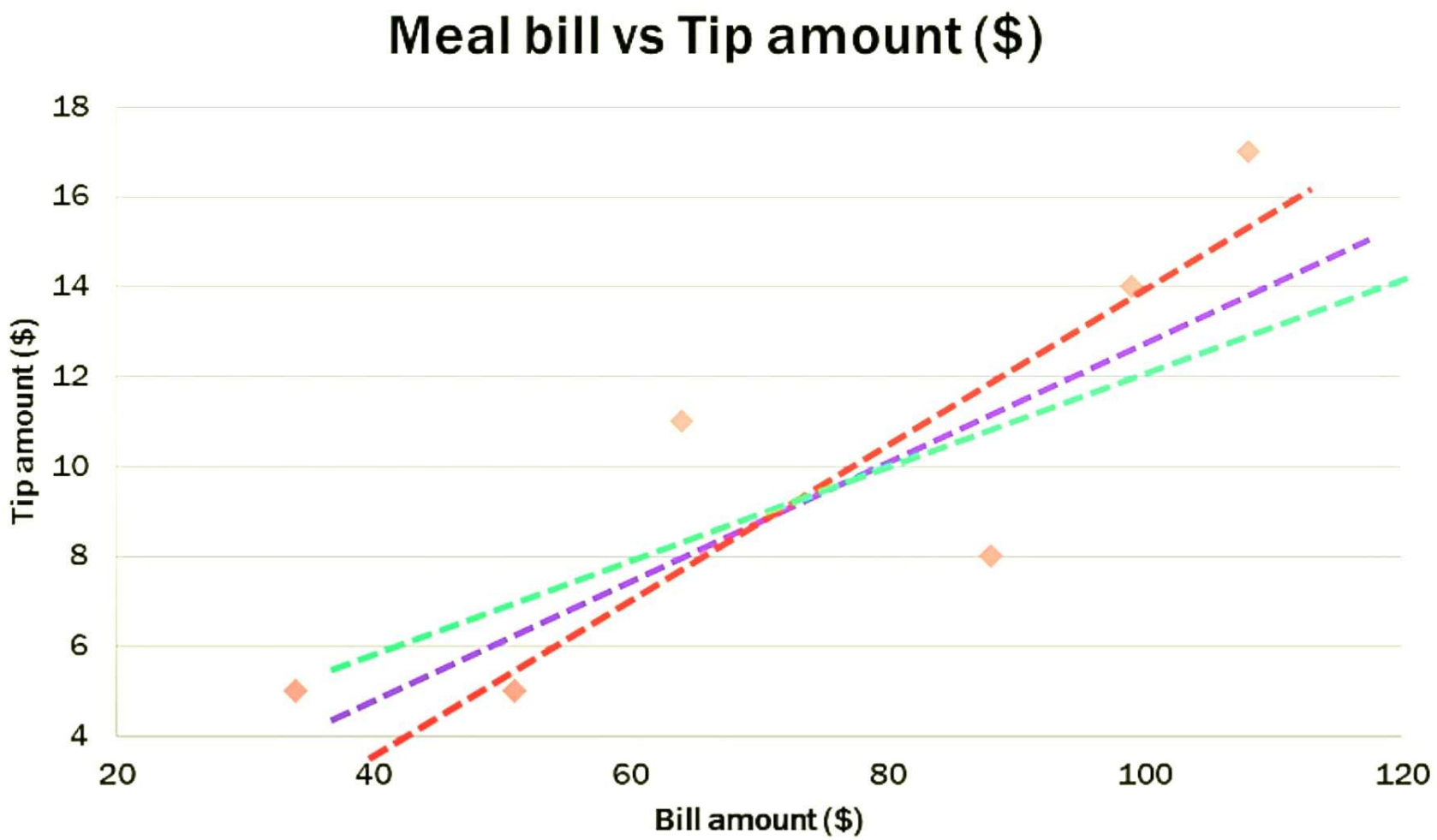
Bill ($)	Tip ($)
34.00	5.00
108.00	17.00
64.00	11.00
88.00	8.00
99.00	14.00
51.00	5.00

Step 1: scatter plot



Bill ($)	Tip ($)
34.00	5.00
108.00	17.00
64.00	11.00
88.00	8.00
99.00	14.00
51.00	5.00

Step 2: look for a visual line



Step 3: correlation check (optional)

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

What is the correlation coefficient, r?

In this case, $r = 0.866$.

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

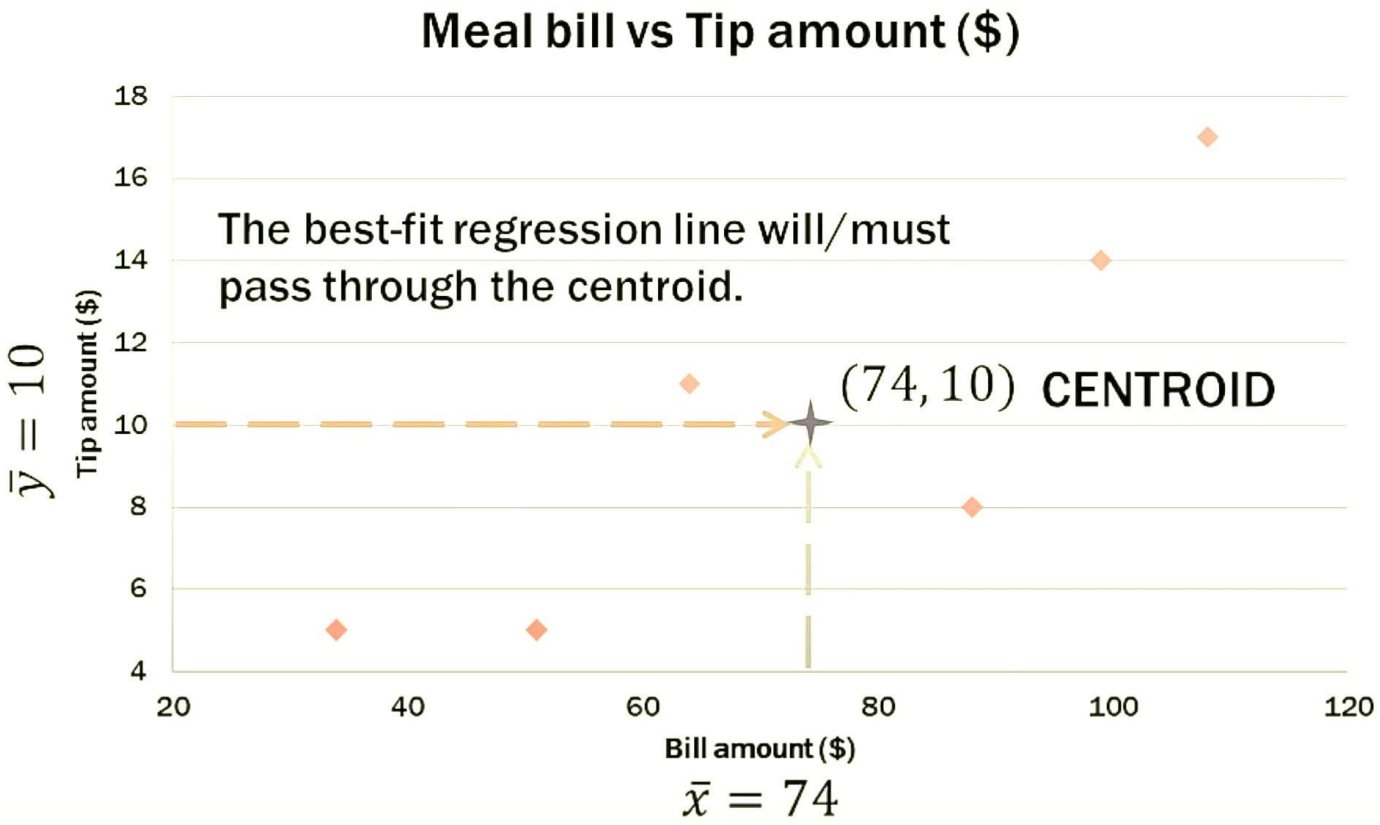
y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Is the relationship strong?

In this case, ***YES!***

Step 4: centroid determination



Bill ($)	Tip ($)
34.00	5.00
108.00	17.00
64.00	11.00
88.00	8.00
99.00	14.00
51.00	5.00
$\bar{x} = 74$	$\bar{y} = 10$

Step 5: calculation

$$\hat{y}_i = b_0 + b_1 x_i$$

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$\bar{x}$ = mean of the independent variable

$\bar{y}$ = mean of the dependent variable

x_i = value of independent variable

y_i = value of dependent variable

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

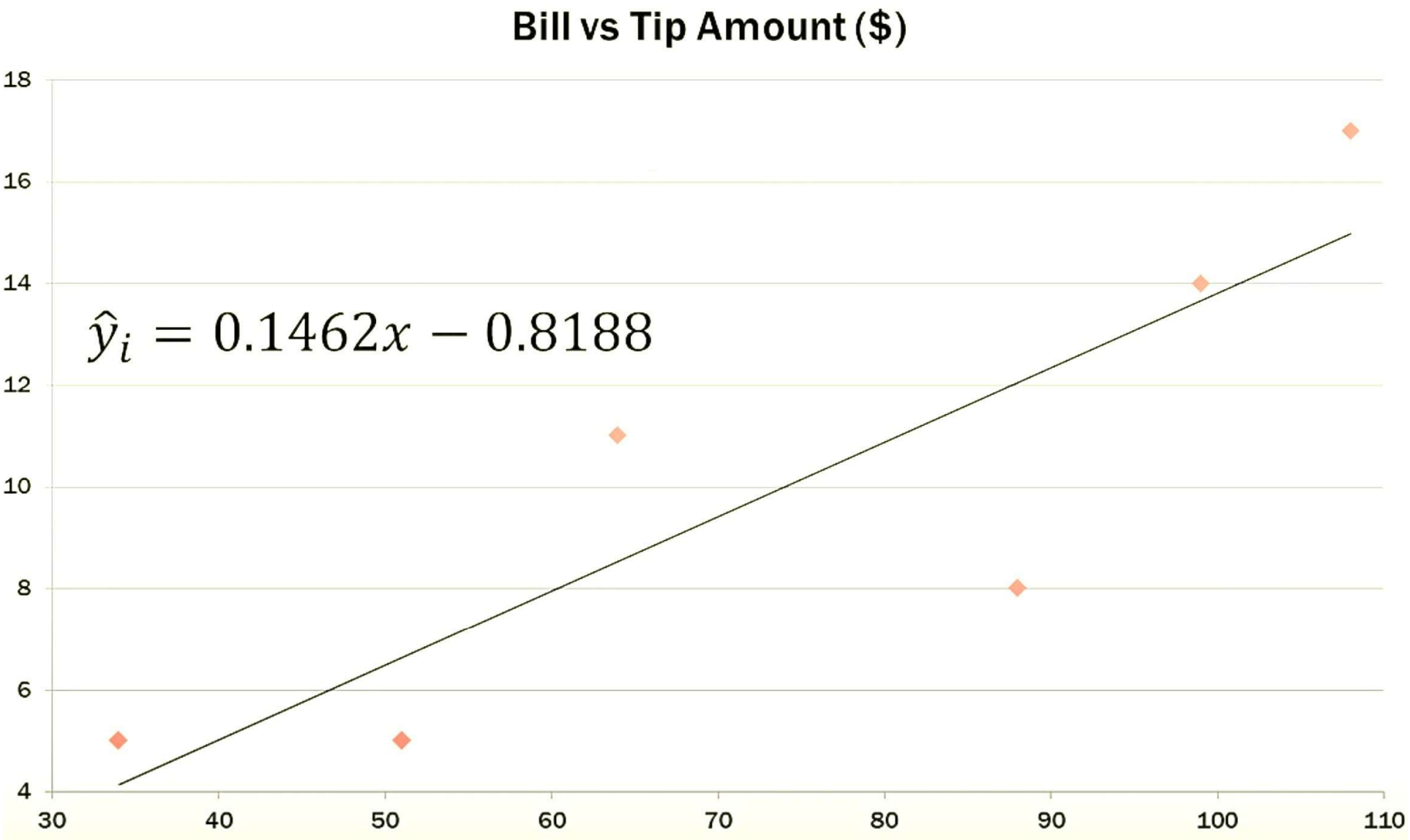
Meal	Total bill ($)	Tip amount ($)	Bill deviation	Tip Deviations	Deviation Products	Bill Deviations Squared
	x	y	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})(y_i - \overline{y})$	$(x_i - \overline{x})^2$
1	34	5				
2	108	17				
3	64	11				
4	88	8				
5	99	14				
6	51	5				
	$\overline{x} = 74$	$\bar{y} = 10$				

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \qquad b_0 = \bar{y} - b_1\bar{x}$$

Meal	Total bill ($)	Tip amount ($)	Bill deviation	Tip Deviations	Deviation Products	Bill Deviations Squared
	x	y	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})(y_i - \overline{y})$	$(x_i - \overline{x})^2$
1	34	5				
2	108	17				
3	64	11				
4	88	8				
5	99	14				
6	51	5				
	$\overline{x} = 74$	$\bar{y} = 10$			$\sum = 615$	$\sum = 4206$

$b_1 = 0.1462$ $\qquad$ $b_0 = 0.8188$

Estimated simple linear regression model



Estimated error

Meal	Total bill ($)	Tip amount ($)	$\hat{y}_i = 0.1462x - 0.8188$	$\hat{y}_i$ (predicted tip amount)
	x	y		
1	34	5	$\hat{y}_i = 0.1462(34) - 0.8188$	4.1505
2	108	17	$\hat{y}_i = 0.1462(108) - 0.8188$	14.9693
3	64	11	$\hat{y}_i = 0.1462(64) - 0.8188$	8.5365
4	88	8	$\hat{y}_i = 0.1462(88) - 0.8188$	12.0453
5	99	14	$\hat{y}_i = 0.1462(99) - 0.8188$	13.6535
6	51	5	$\hat{y}_i = 0.1462(51) - 0.8188$	6.6359
	$\bar{x} = 74$	$\bar{y} = 10$		

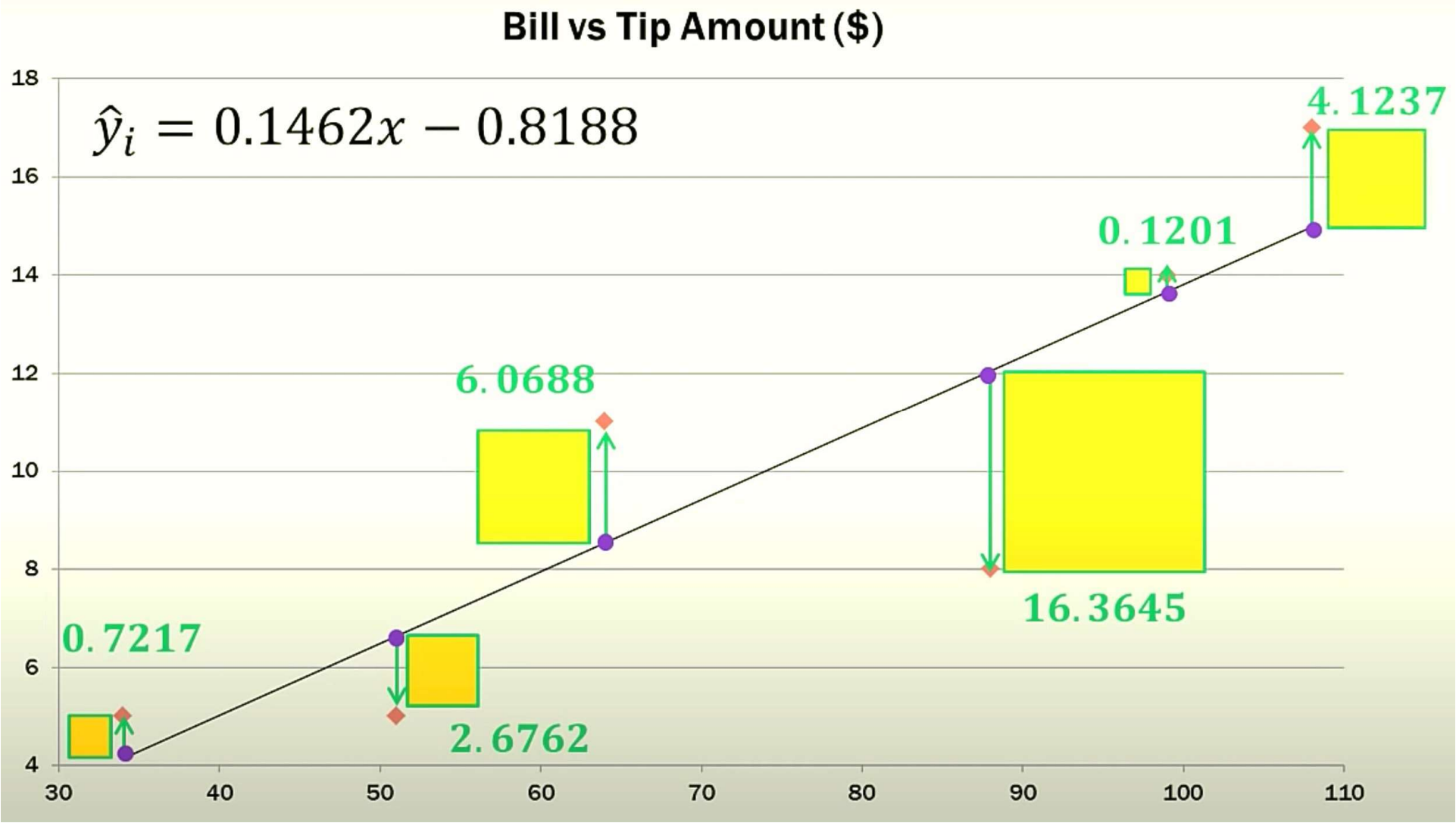
Observed vs. Predicted

Estimated error

Meal	Total bill ($)	Observed tip amount ($)	$\hat{y}_i$ (predicted tip amount)	Error $(y - \hat{y}_i)$	Squared Error $(y - \hat{y}_i)^2$
	x	y			
1	34	5	4.1505	0.8495	0.7217
2	108	17	14.9693	2.0307	4.1237
3	64	11	8.5365	2.4635	6.0688
4	88	8	12.0453	−4.0453	16.3645
5	99	14	13.6535	0.3465	0.1201
6	51	5	6.6359	−1.6359	2.6762

$$SSE = 30.075$$

Estimated error



SSE comparison

Tip amount only

SSE = 120 = SST

Tip amount as a function of bill amount

SSE = 30.075

$$SST = SSR + SSE$$

$$120 = 89.925 + 30.075$$

Solving Linear Regression Using Gradient Descent

Gradient Descent is an optimization algorithm used to find the best values of β0 (intercept) and β1 (slope) by minimizing the Mean Squared Error (MSE).

The goal is to minimize MSE, which is equivalent to minimizing the **Sum of Squared Errors (SSE)**:

$$J(\beta_0, \beta_1) = \frac{1}{2n} \sum (Y_i - (\beta_0 + \beta_1 X_i))^2$$

It can be used gradient descent to iteratively adjust β0 and β1 to minimize this cost.

Basic idea

Compute the Gradients

The gradients (partial derivatives) of the cost function with respect to β0 and β1 are.

$$\frac{\partial J}{\partial \beta_0} = -\frac{1}{n}\sum(Y_i - (\beta_0 + \beta_1 X_i))$$

$$\frac{\partial J}{\partial \beta_1} = -\frac{1}{n}\sum(X_i(Y_i - (\beta_0 + \beta_1 X_i)))$$

These gradients tell us the direction in which we should update β0 and β1.

Basic idea

Gradient Descent Update Rules

Using a learning rate α\alphaα, the update rules for β0 and β1 are.

$$\beta_0^{new} = \beta_0^{old} - \alpha \cdot \frac{\partial J}{\partial \beta_0}$$

$$\beta_1^{new} = \beta_1^{old} - \alpha \cdot \frac{\partial J}{\partial \beta_1}$$

which expands to:

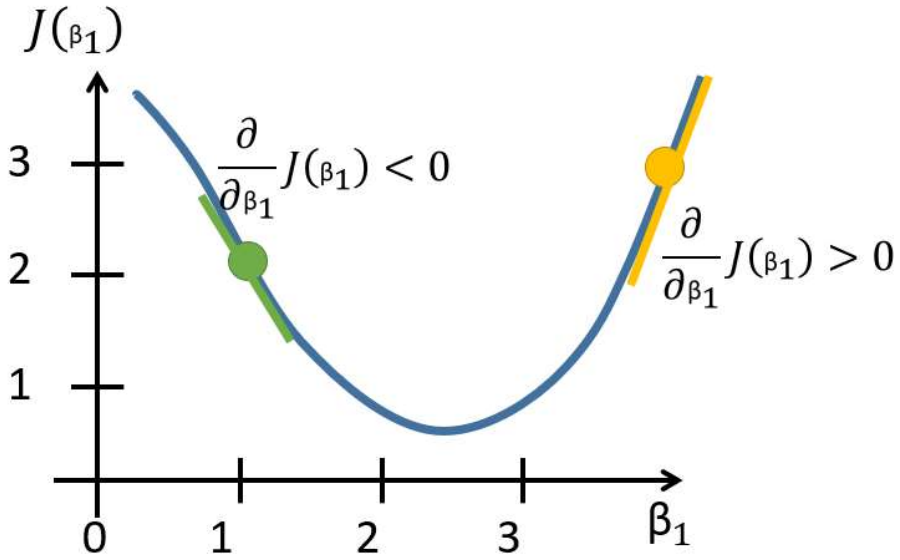
$$\beta_0 = \beta_0 - \alpha \cdot \left(-\frac{1}{n} \sum (Y_i - (\beta_0 + \beta_1 X_i)) \right)$$

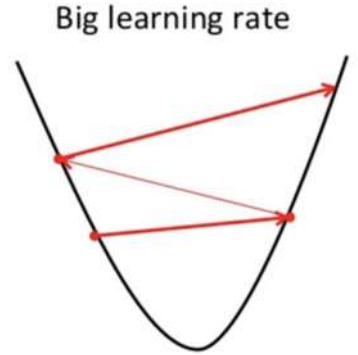
$$\beta_1 = \beta_1 - \alpha \cdot \left(-\frac{1}{n} \sum X_i (Y_i - (\beta_0 + \beta_1 X_i)) \right)$$

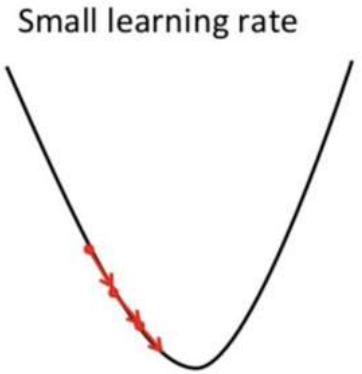
Basic idea

- Initialize β1
- Repeat until convergence

$$\beta_1 := \beta_1 - \alpha \frac{\partial J}{\partial \beta_1}$$







- **Gradient Direction:**
  - If $\frac{\partial J}{\partial \beta_1} < 0$ (left side of the minimum), β_1 increases.
  - If $\frac{\partial J}{\partial \beta_1} > 0$ (right side of the minimum), β_1 decreases.

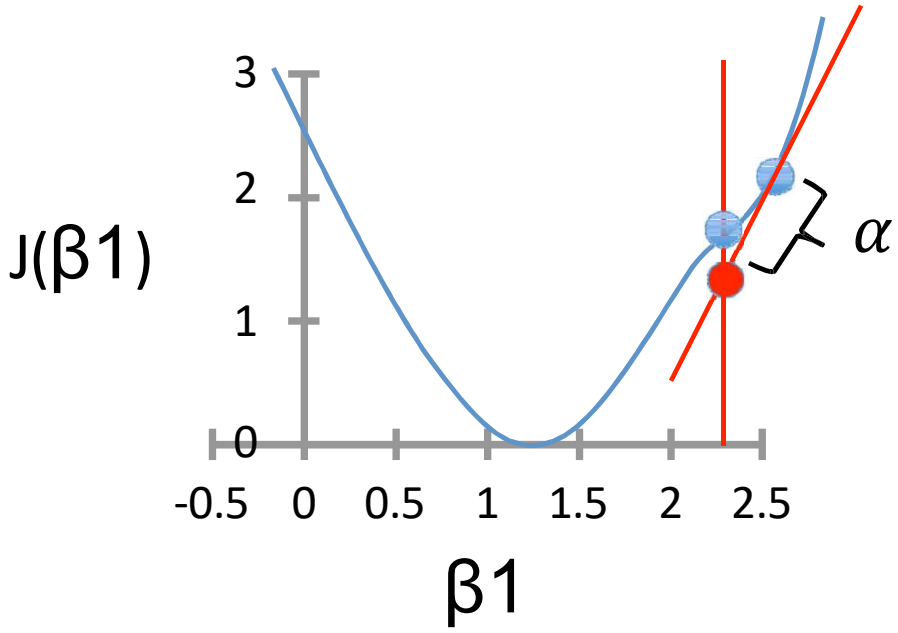
Gradient Descent

- Initialize $\beta 1$
- Repeat until convergence

$$\beta 1 \rightarrow \beta 1 - \alpha \frac{\partial}{\partial \beta 1} J(\beta 1)$$

simultaneous update
for j = 0 ... d

learning rate (small)
e.g., α = 0.05



How to improve the model?

- Including more predictors
  - The bills were paid by male or female?
  - The customers are smokers or non-smokers?
  - The meals are on Monday, Tuesday,… or Sunday?
  - The meals took place during lunchtime or dinner?

→ **Multiple linear regression**

Multiple Linear Regression

$$f(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_p x_p$$

$$\mathbf{w} = \left[w_0, w_1, w_2, \ldots, w_p \right]^T$$

$$\bar{\mathbf{x}} = \left[1, x_1, x_2, \ldots, x_p \right]$$

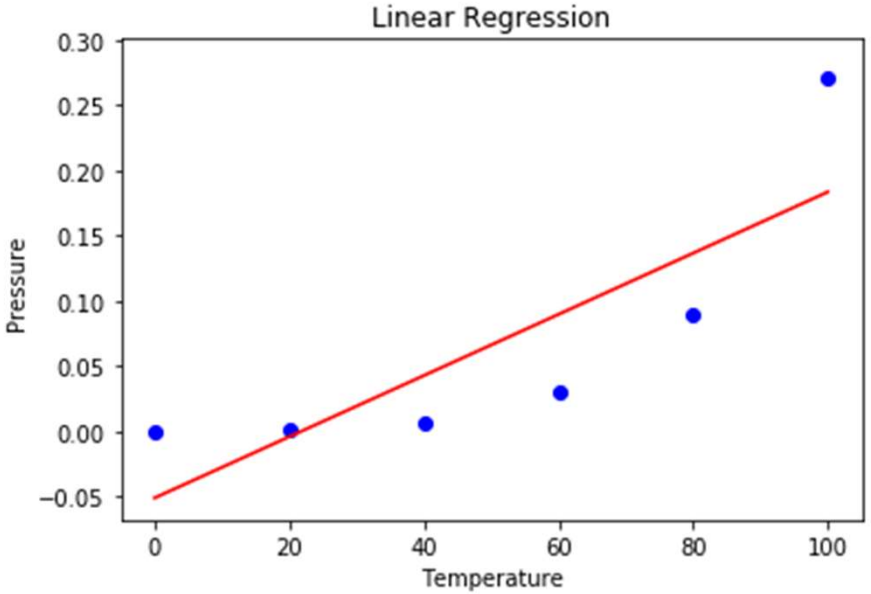
$$f(\mathbf{x}) = \bar{\mathbf{x}}\mathbf{w}$$

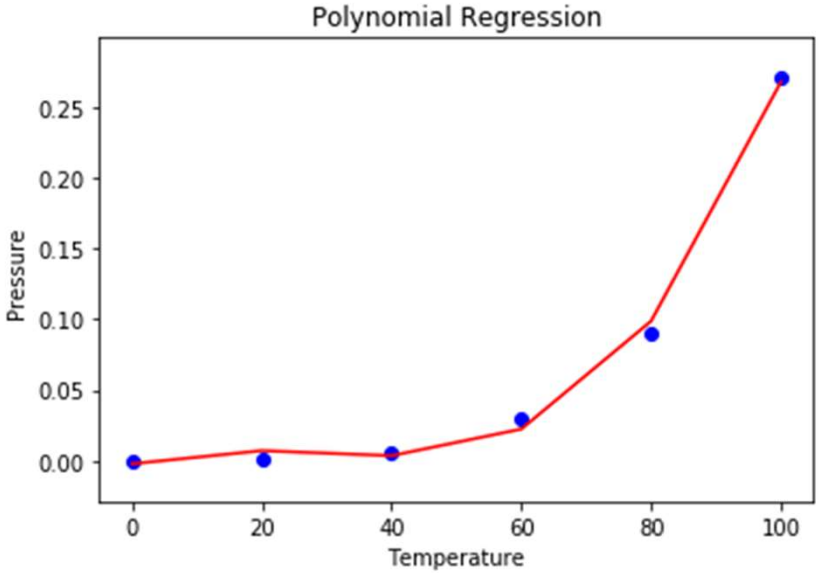
Loss function

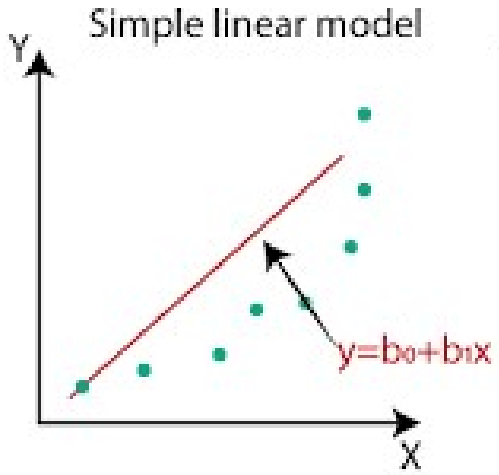
$$\mathcal{L}(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{N} (y_i - \bar{\mathbf{x}}_\mathbf{i}\mathbf{w})^2$$

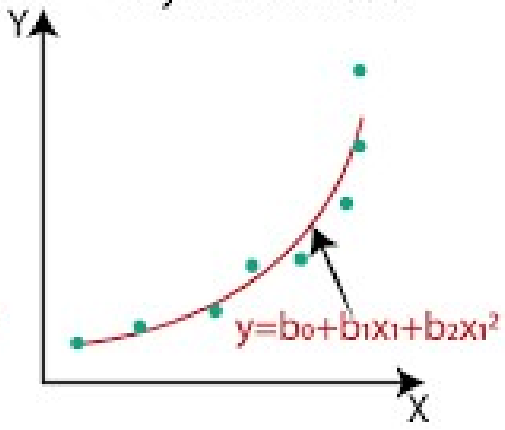
$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w})$$

Polynomial Regression

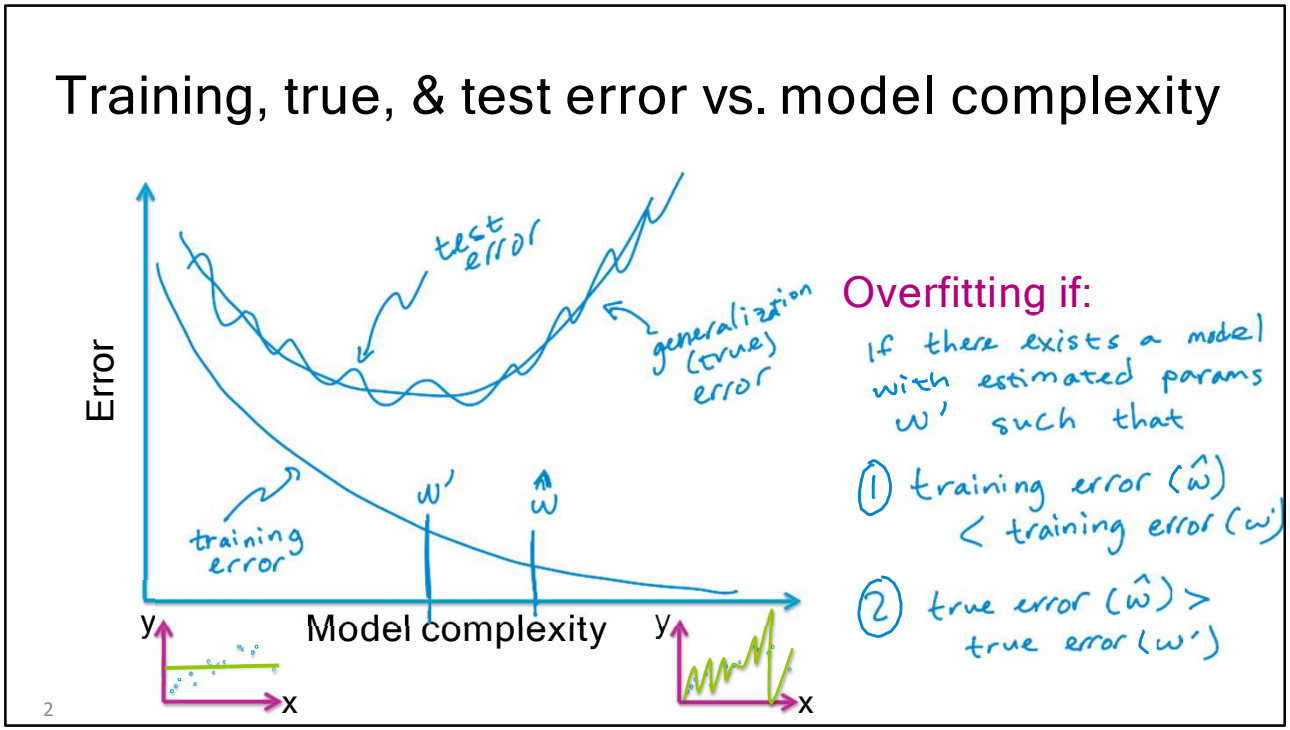




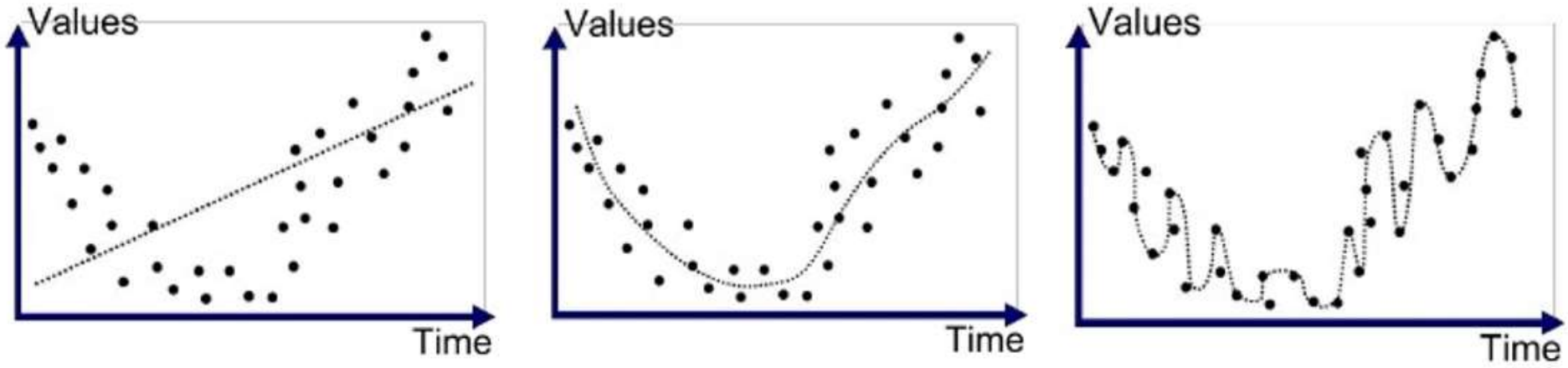




Overfitting



Underfitting and overfitting illustrated

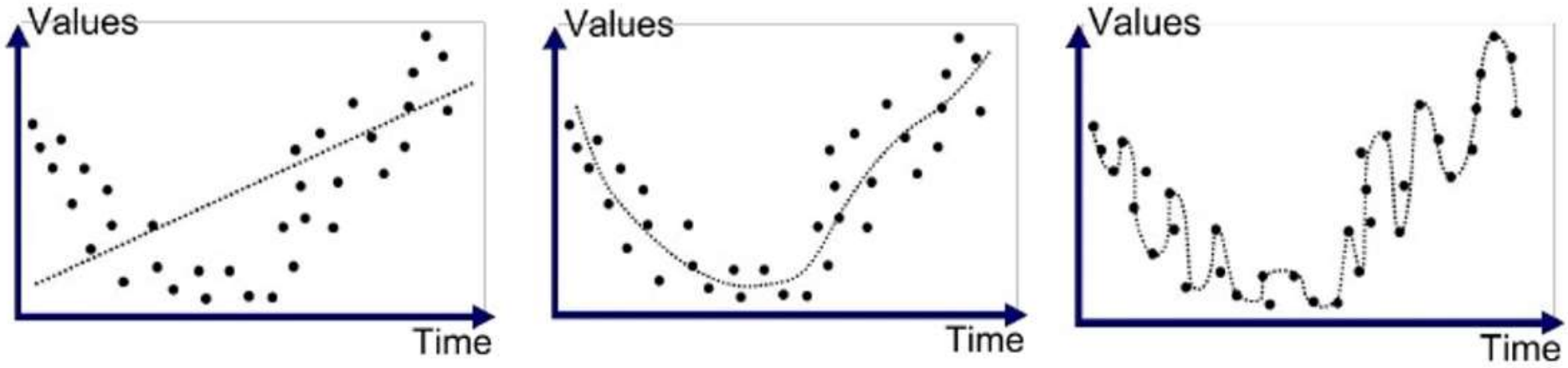


Underfitted

Good Fit/Robust

Overfitted

Underfitting and overfitting illustrated



Underfitted

Good Fit/Robust

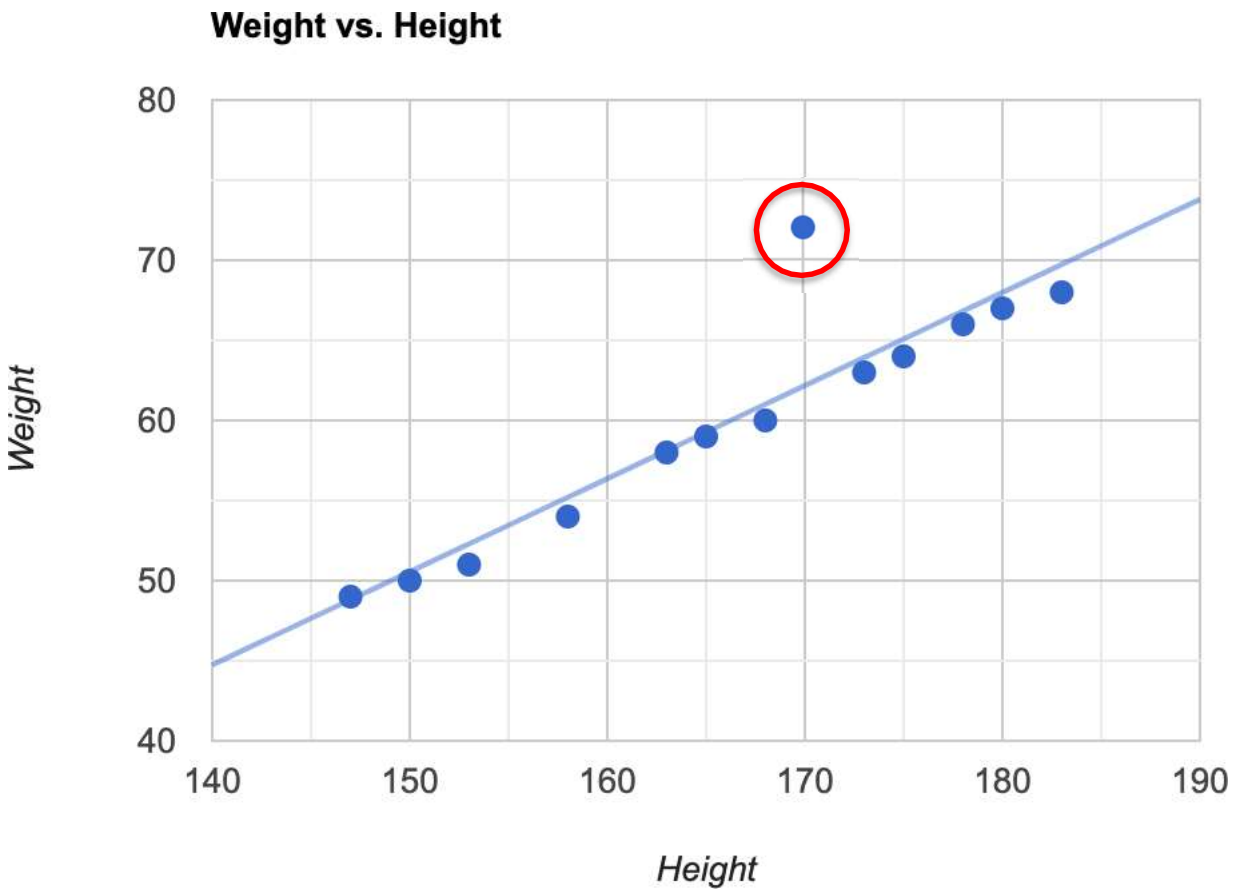
Overfitted

Exercise 1

Predicting weight from height.

Note: remove outlier

Height (cm)	Weight (kg)	Height (cm)	Weight (kg)
147	49	168	60
150	50	170	72
153	51	173	63
155	52	175	64
158	54	178	66
160	56	180	67
163	58	183	68
165	59		


Weight vs. Height
80
70
60
50
40
Weight
140
150
160
170
180
190
Height

Exercise 2

Predicting job performance from IQ scores.

If an applicant's IQ score = 100 → estimate his performance?

id	fname	iq	performance
1	Kevin	106	115
2	Ayden	97	104
3	Madelyn	108	98
4	Madelyn	96	101
5	Tristan	112	98
6	Isaac	111	106
7	Victoria	97	95
8	Christopher	99	98
9	Caroline	87	87
10	Daniel	91	86