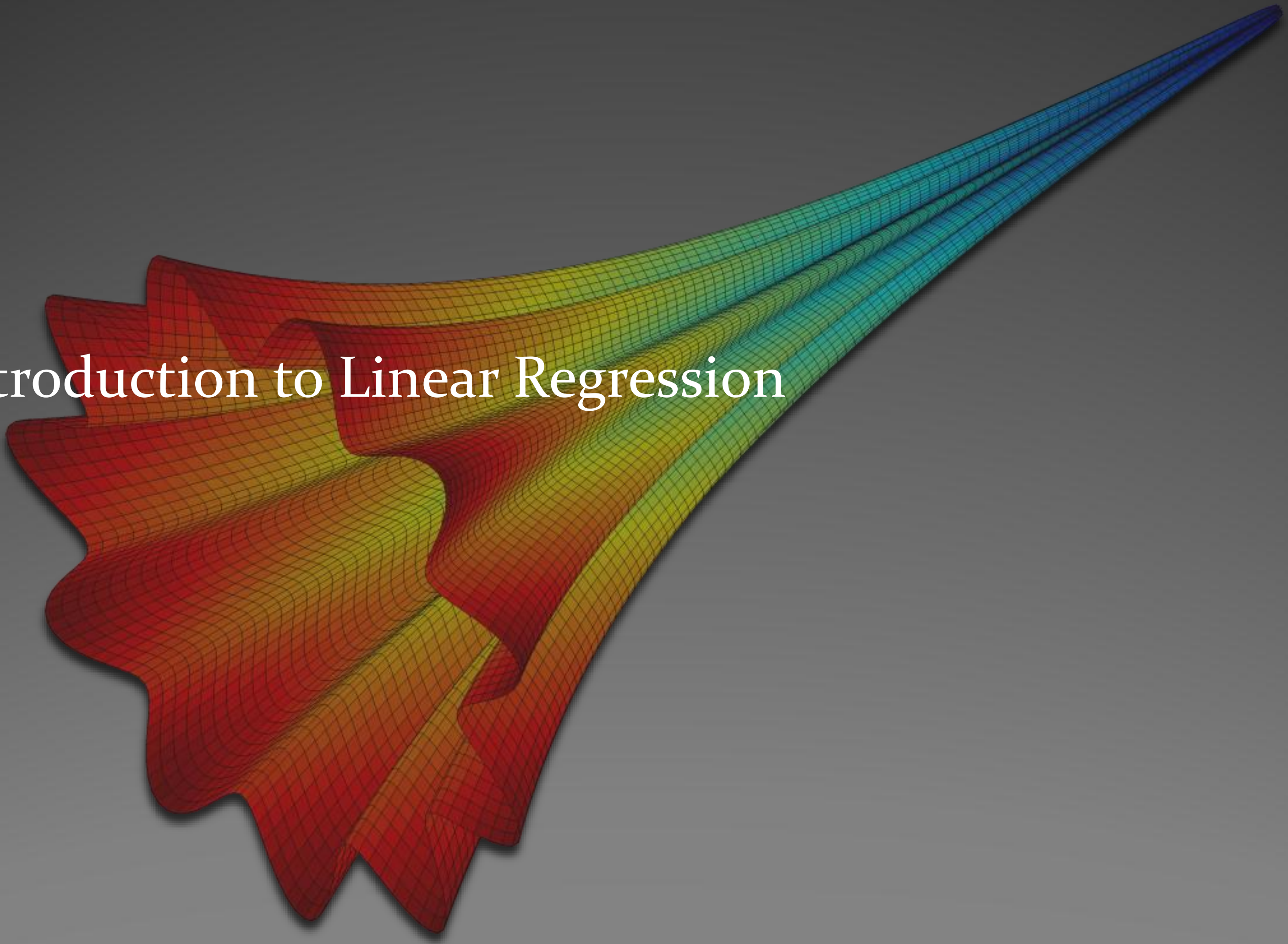
Introduction to Linear Regression



Simple Regression

Definition

A regression model is a mathematical equation that describes the relationship between two or more variables. A ***simple regression*** model includes only two variables: one independent and one dependent. The dependent variable is the one being explained, and the independent variable is the one used to explain the variation in the dependent variable.

What is Regression?

What is regression? Given n data points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ best fit $y = f(x)$ to the data. The best fit is generally based on minimizing the sum of the square of the residuals, S_r.

Residual at a point is

$$\varepsilon_i = y_i - f(x_i)$$

Sum of the square of the residuals

$$S_r = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

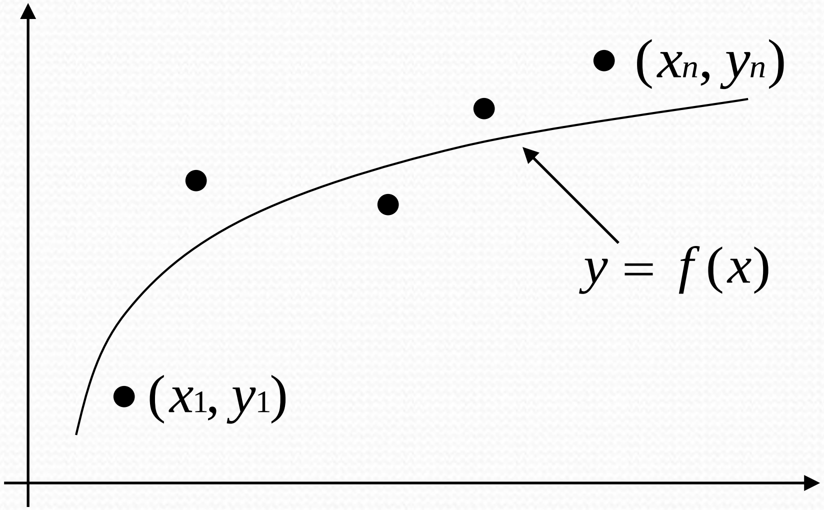


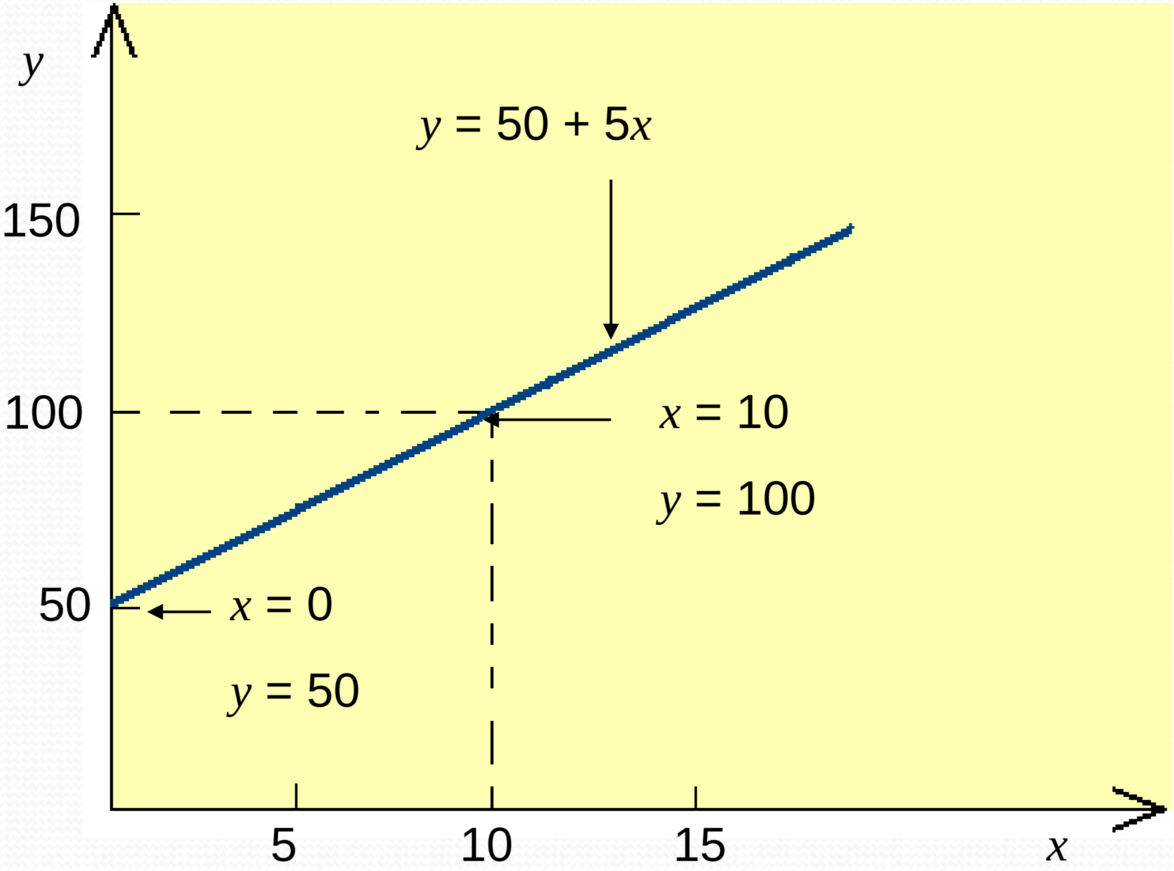
Figure. Basic model for regression

Linear Regression

Definition

A (simple) regression model that gives a straight-line relationship between two variables is called a ***linear regression*** model.

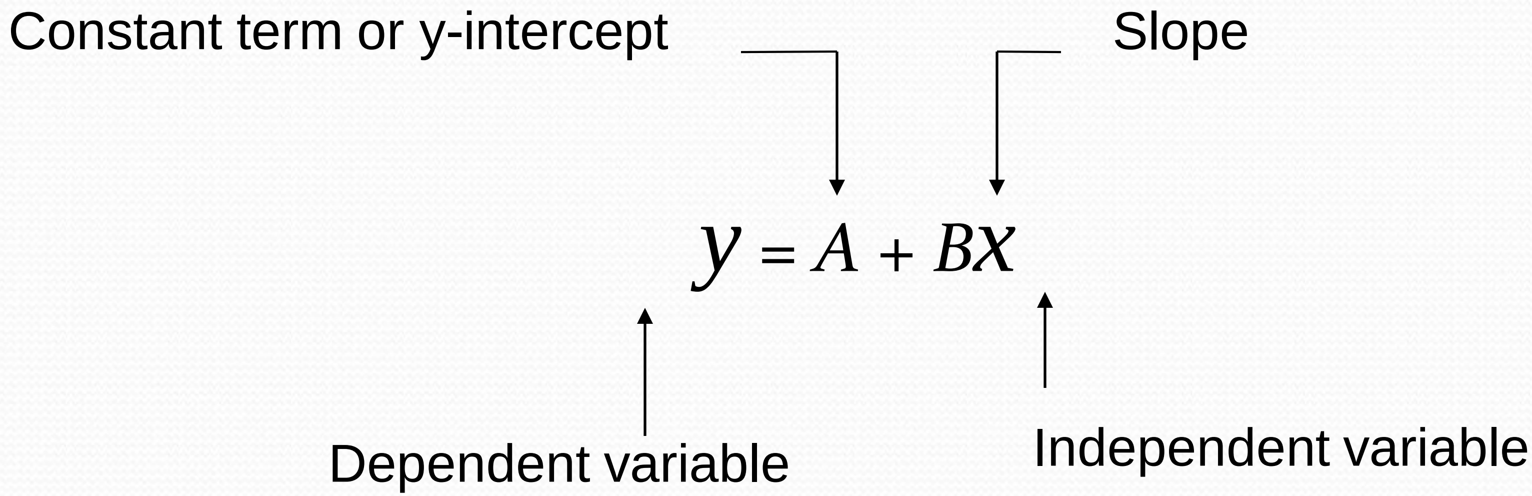
Plotting a linear equation.



SIMPLE LINEAR REGRESSION ANALYSIS

- Scatter Diagram
- Least Square Line
- Interpretation of *a* and *b*
- Assumptions of the Regression Model

SIMPLE LINEAR REGRESSION ANALYSIS cont.



Constant term or y-intercept

Slope

$$y = A + Bx$$

Dependent variable

Independent variable

SIMPLE LINEAR REGRESSION ANALYSIS cont.

Definition

In the ***regression model*** $y = A + Bx +$ Є, A is called the y-intercept or constant term, B is the slope, and Є is the random error term. The dependent and independent variables are y and x, respectively.

SIMPLE LINEAR REGRESSION ANALYSIS

Definition

In the model $\hat{y} = a + bx$, a and b, which are calculated using sample data, are called the ***estimates of A and B***.

Table 1 Incomes (in hundreds of dollars) and Food Expenditures of Seven Households

Income	Food Expenditure
35	9
49	15
21	7
39	11
15	5
28	8
25	9

Scatter Diagram

Definition

A plot of paired observations is called a ***scatter diagram***.

Figure 4 Scatter diagram.

Income	Food Expenditure
35	9
49	15
21	7
39	11
15	5
28	8
25	9

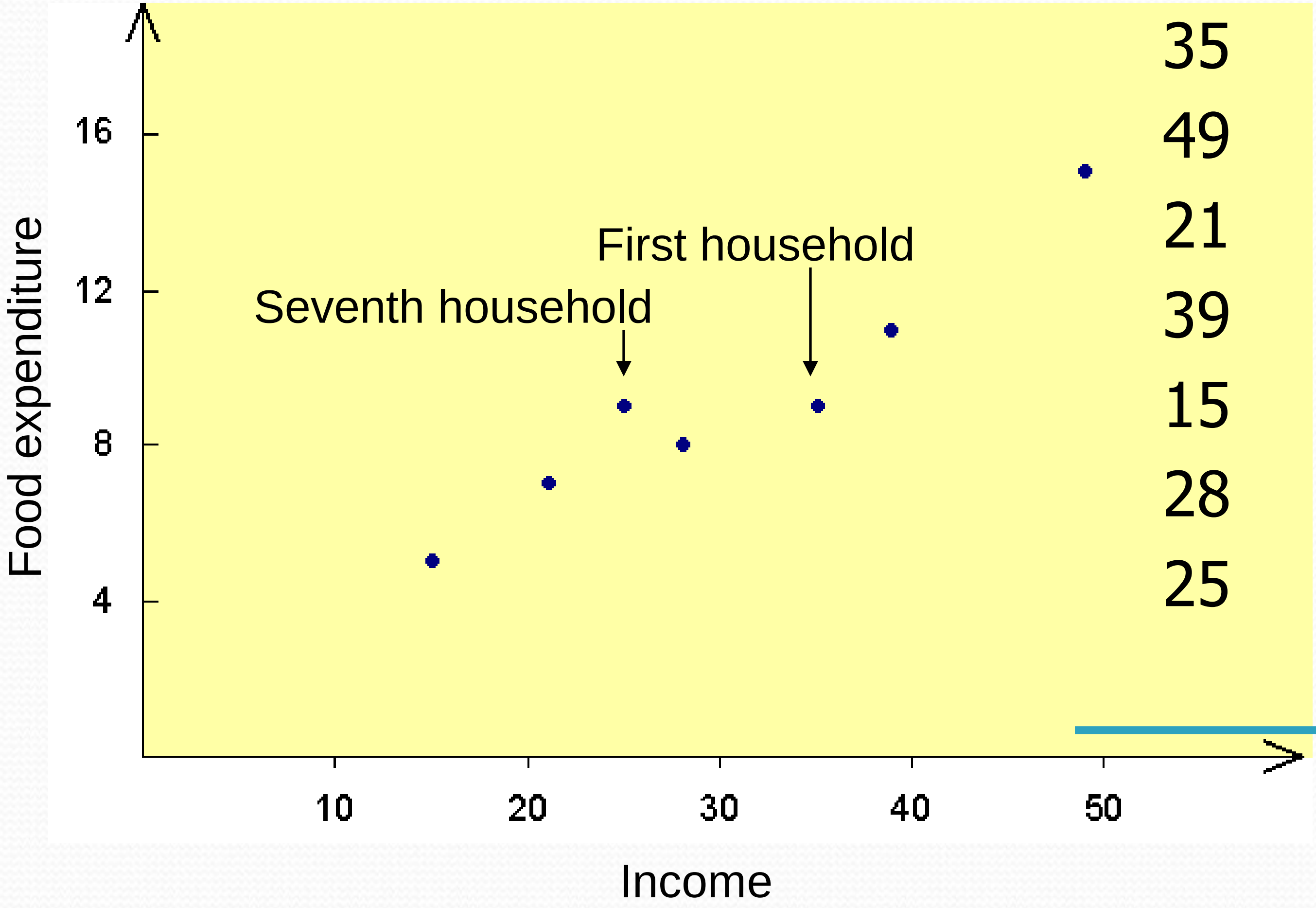
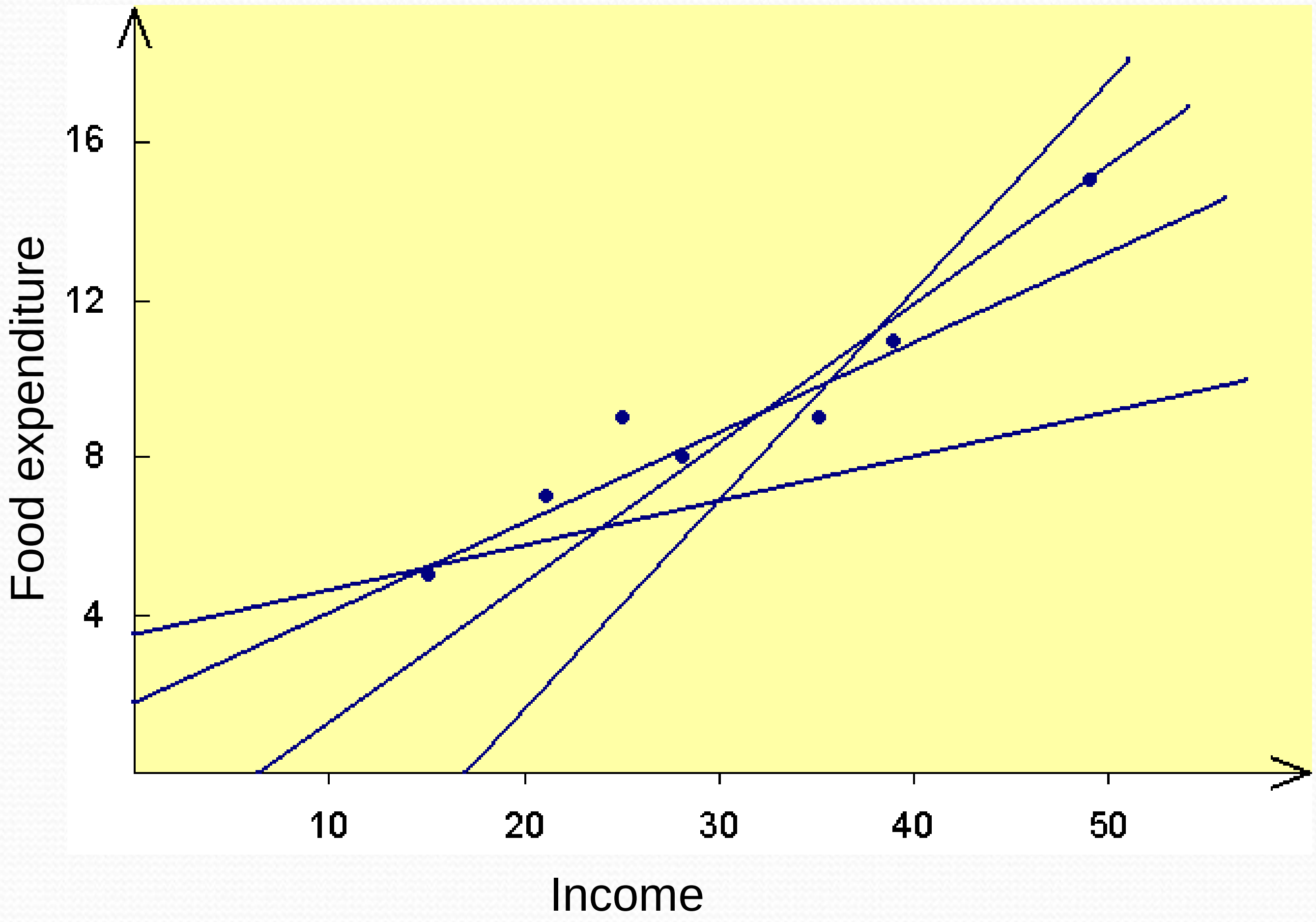
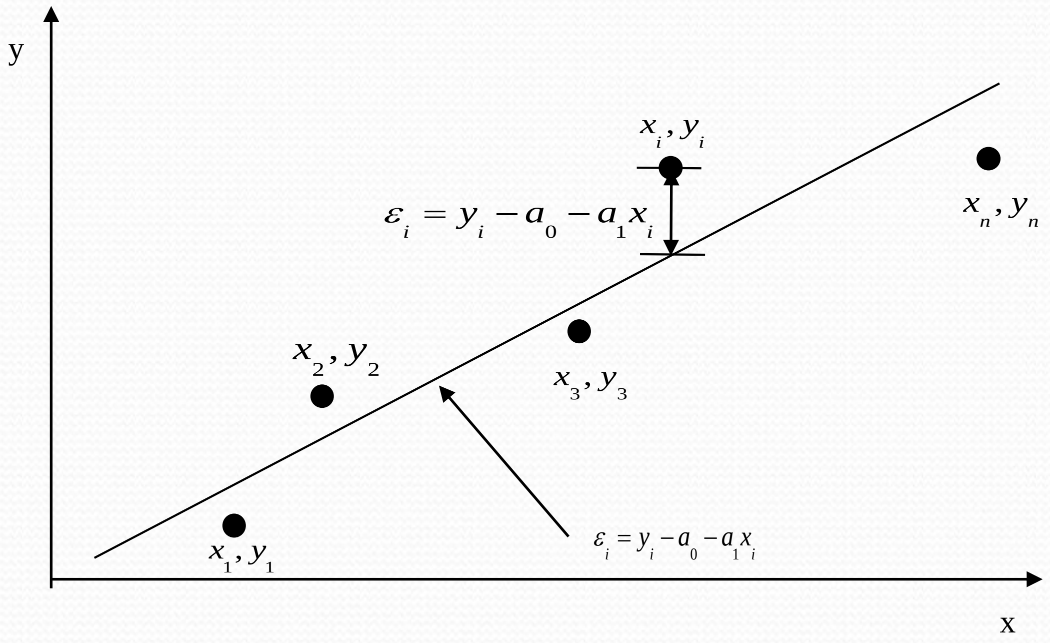


Figure 5 Scatter diagram and straight lines (which line to choose ?)



Linear Regression-Criterion#1

Given *n* data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ best fit $y = a_0 + a_1 x$ to the data.



y
x_i, y_i
$\varepsilon_i = y_i - a_0 - a_1 x_i$
x_n, y_n
x_2, y_2
x_3, y_3
x_1, y_1
$\varepsilon_i = y_i - a_0 - a_1 x_i$
x

Figure. Linear regression of y vs. x data showing residuals at a typical point, x_i.

Does minimizing $\sum_{i=1}^{n} \varepsilon_i$ work as a criterion, where $\varepsilon_i = y_i - (a_0 + a_1 x_i)$

Example for Criterion#1

Example: Given the data points (2,4), (3,6), (2,6) and (3,8), best fit the data to a straight line using Criterion#1

Table. Data Points

x	y
2.0	4.0
3.0	6.0
2.0	6.0
3.0	8.0

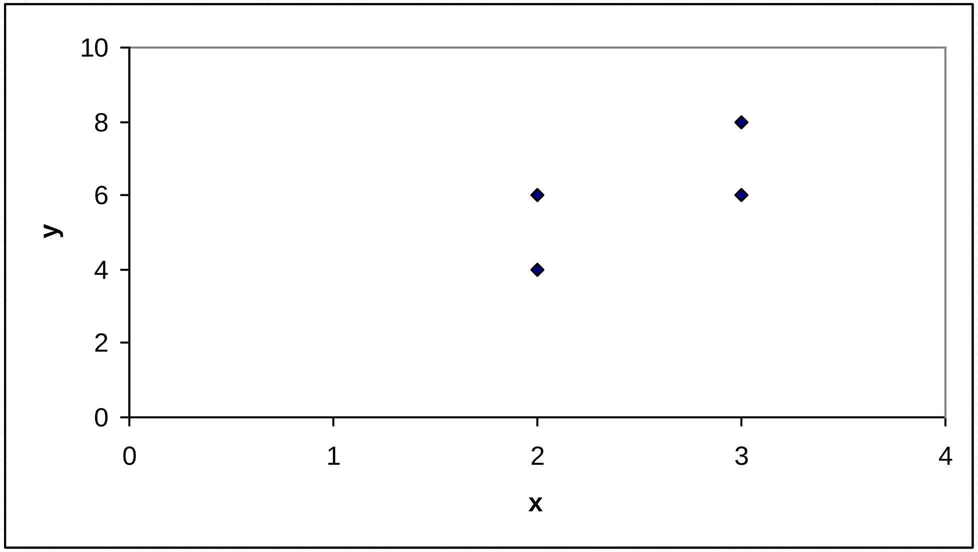


Figure. Data points for y vs. x data.

Linear Regression-Criteria#1

Using *y=4x-4* as the regression curve

Table. Residuals at each point for regression model y = 4x – 4.

x	y	$y_{predicted}$	$\varepsilon = y - y_{predicted}$
2.0	4.0	4.0	0.0
3.0	6.0	8.0	-2.0
2.0	6.0	4.0	2.0
3.0	8.0	8.0	0.0
			$\sum_{i=1}^{4} \varepsilon_i = 0$

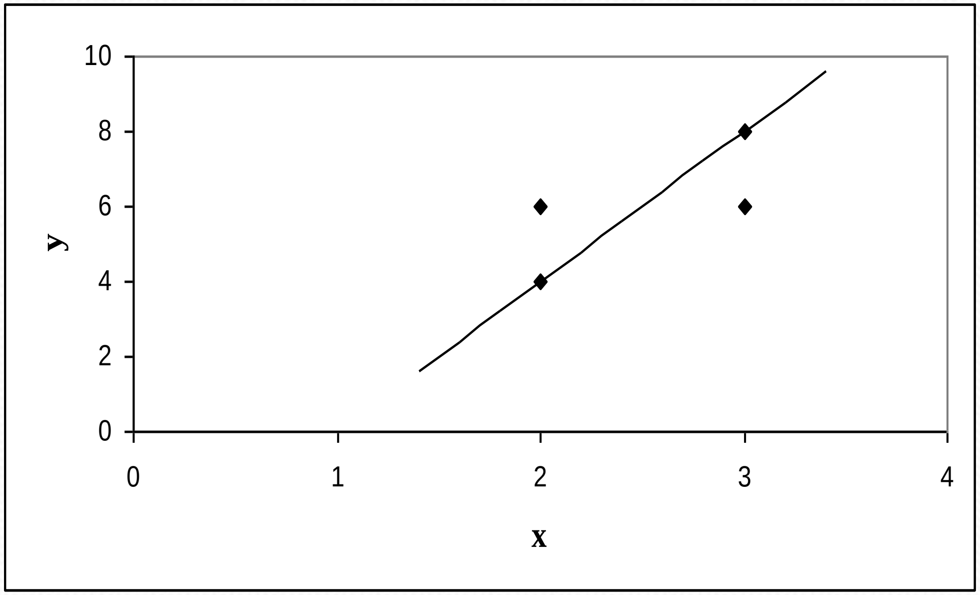


Figure. Regression curve for y=4x-4, y vs. x data

Linear Regression-Criteria#1

Using *y=6* as a regression curve

Table. Residuals at each point for y=6

x	y	$y_{predicted}$	$\varepsilon = y - y_{predicted}$
2.0	4.0	6.0	-2.0
3.0	6.0	6.0	0.0
2.0	6.0	6.0	0.0
3.0	8.0	6.0	2.0
			$\sum_{i=1}^{4} \varepsilon_i = 0$

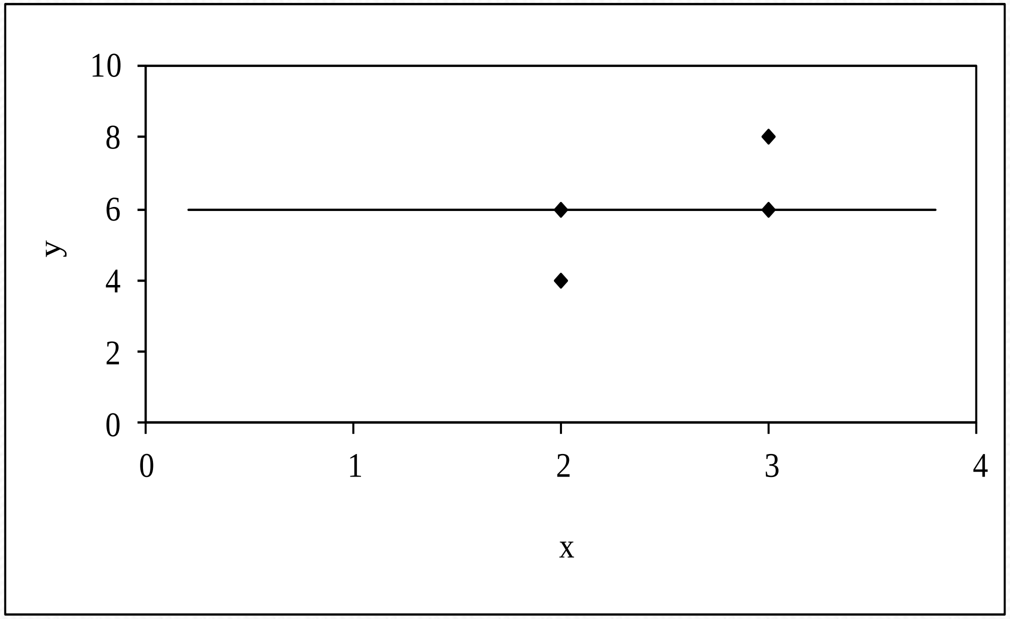


Figure. Regression curve for y=6, y vs. x data

Linear Regression – Criterion #1

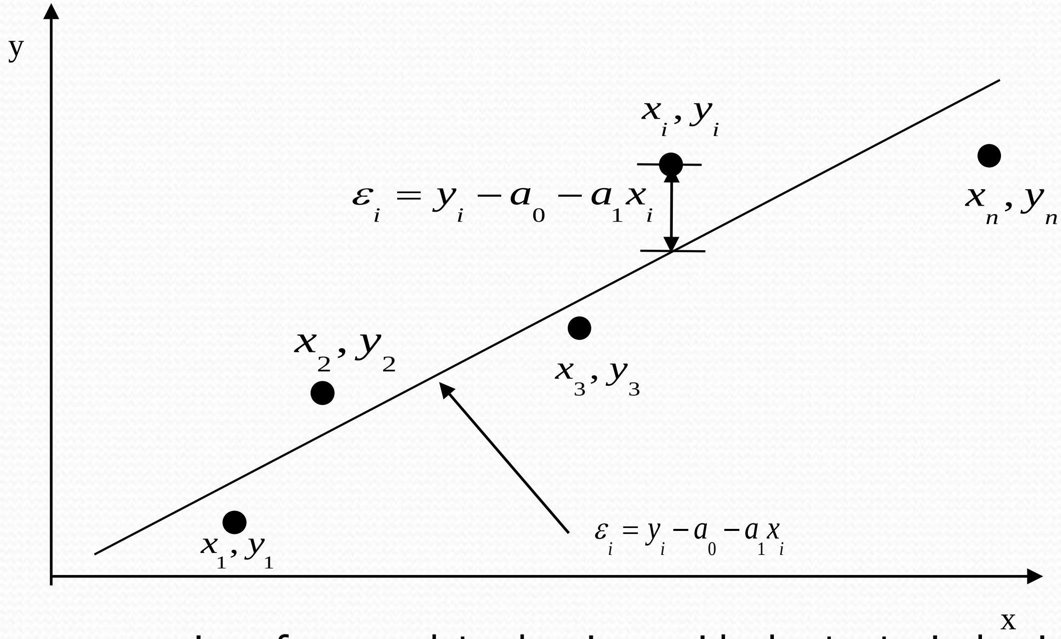
$\sum_{i=1}^{4} \varepsilon_i = 0$ for both regression models of y=4x-4 and y=6.

The sum of the residuals is as small as possible, that is zero, but the regression model is not unique.

Hence the above criterion of minimizing the sum of the residuals is a bad criterion.

Linear Regression-Criterion#2

Will minimizing $\sum_{i=1}^{n} \left| \varepsilon_i \right|$ work any better?



y
x_i, y_i
$\varepsilon_i = y_i - a_0 - a_1 x_i$
x_n, y_n
x_2, y_2
x_3, y_3
x_1, y_1
$\varepsilon_i = y_i - a_0 - a_1 x_i$
x

Figure. Linear regression of y vs. x data showing residuals at a typical point, x_i.

Linear Regression-Criteria 2

Using *y=4x-4* as the regression curve

Table. The absolute residuals employing the y=4x-4 regression model

x	y	$y_{predicted}$	$\|\varepsilon\| = \|y - y_{predicted}\|$
2.0	4.0	4.0	0.0
3.0	6.0	8.0	2.0
2.0	6.0	4.0	2.0
3.0	8.0	8.0	0.0
			$\sum_{i=1}^{4} \left\|\varepsilon_i\right\| = 4$

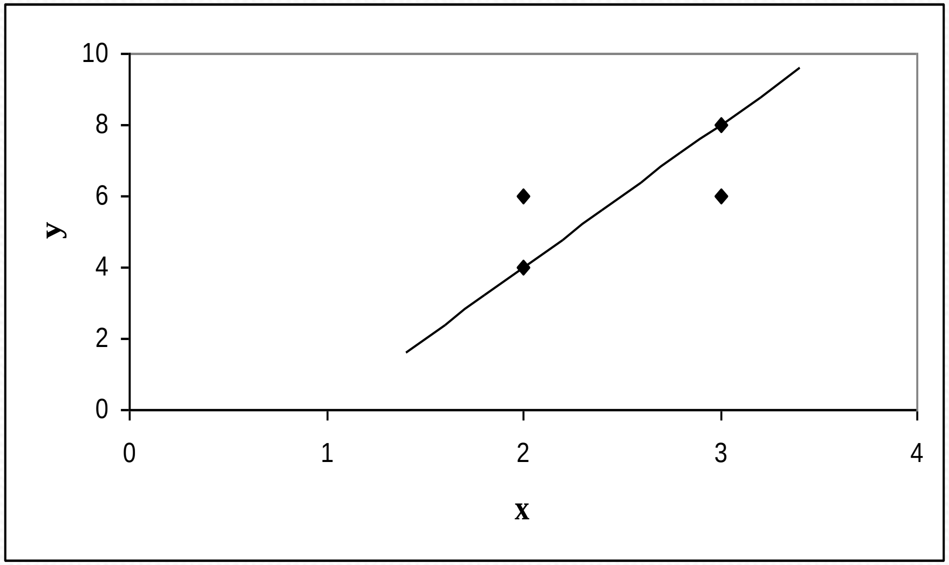


Figure. Regression curve for y=4x-4, y vs. x data

Linear Regression-Criteria#2

Using *y=6* as a regression curve

Table. Absolute residuals employing the y=6 model

x	y	$y_{predicted}$	$\|\varepsilon\| = \|y - y_{predicted}\|$
2.0	4.0	6.0	2.0
3.0	6.0	6.0	0.0
2.0	6.0	6.0	0.0
3.0	8.0	6.0	2.0
			$\sum_{i=1}^{4} \|\varepsilon_i\| = 4$

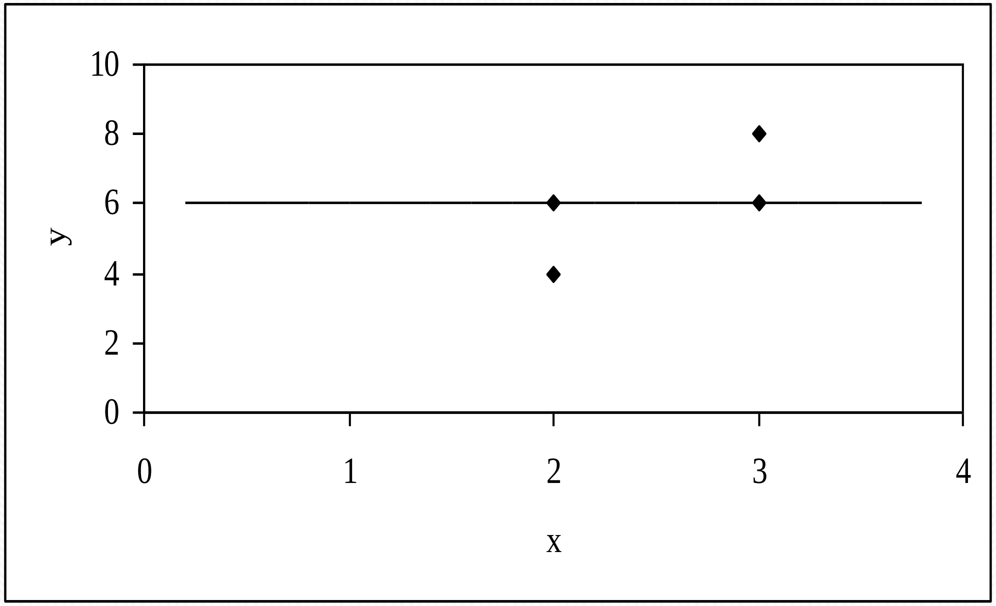


Figure. Regression curve for y=6, y vs. x data

Linear Regression-Criterion#2

$\sum_{i=1}^{4} \left| \varepsilon_i \right| = 4$ for both regression models of y=4x-4 and y=6.

The sum of the absolute residuals has been made as small as possible, that is 4, but the regression model is not unique.

Hence the above criterion of minimizing the sum of the absolute value of the residuals is also a bad criterion.

Can you find a regression line for which $\sum_{i=1}^{4} \left| \varepsilon_i \right| < 4$

Least Squares Criterion

The least squares criterion minimizes the sum of the square of the residuals in the model, and also produces a unique line.

$$S_r = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i\right)^2$$

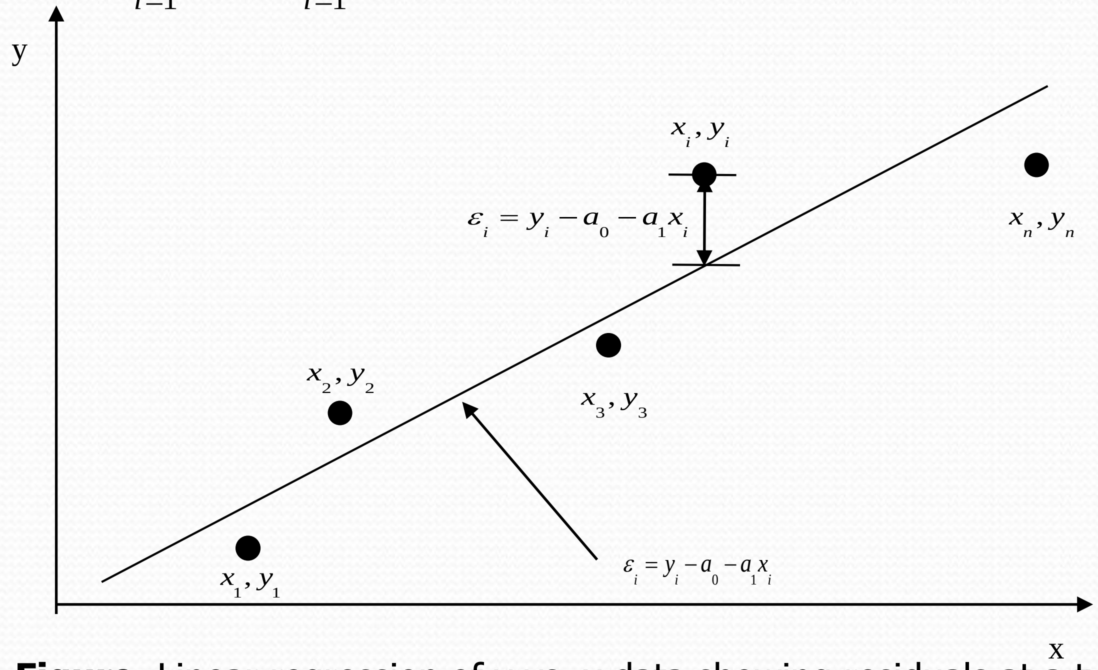
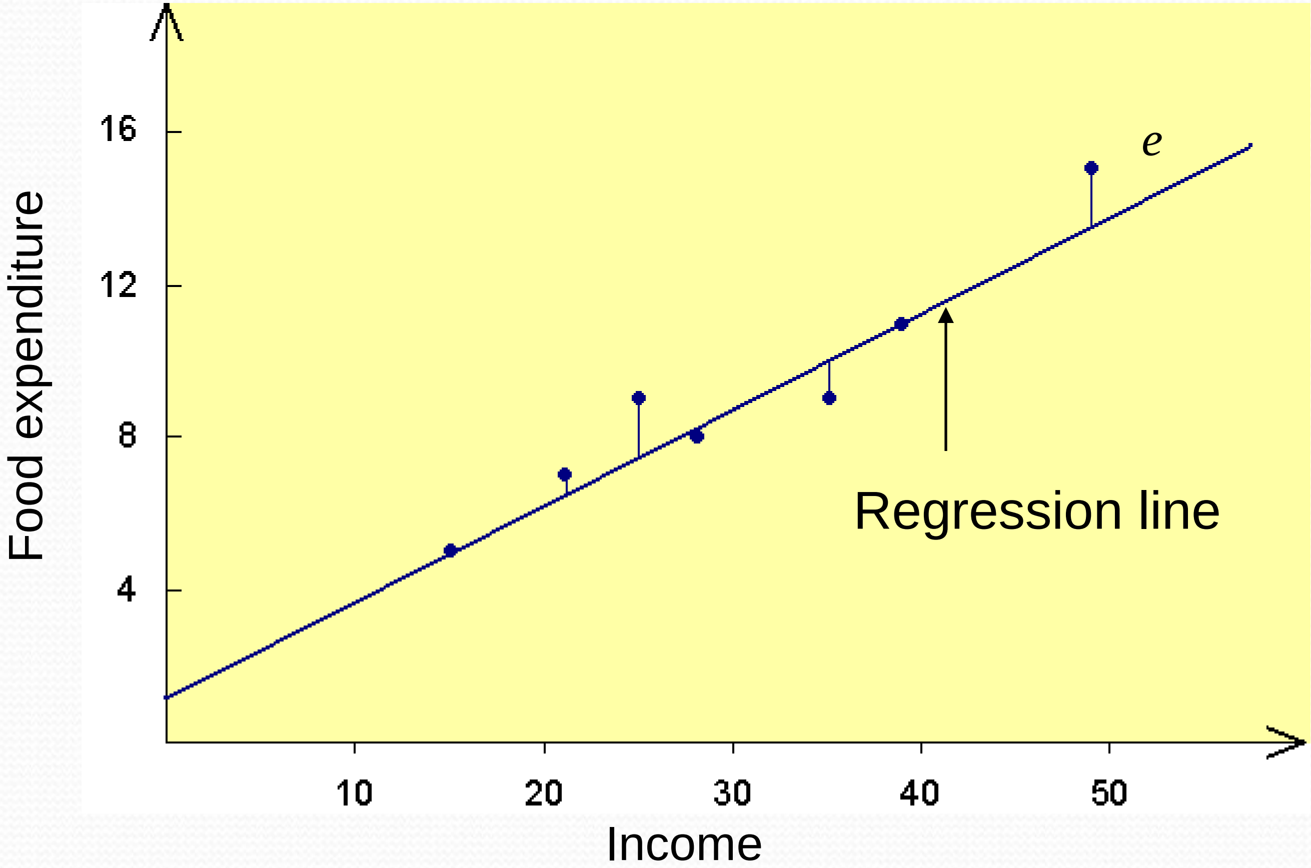

y
x_i, y_i
$\varepsilon_i = y_i - a_0 - a_1 x_i$
x_n, y_n
x_2, y_2
x_3, y_3
x_1, y_1
$\varepsilon_i = y_i - a_0 - a_1 x_i$
x

Figure. Linear regression of y vs. x data showing residuals at a typical point, x_i.

Least Squares Line

Figure 6 Regression line and random errors.



Finding Constants of Linear Model

Minimize the sum of the square of the residuals: $S_r = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i\right)^2$

To find a_0 and a_1 we minimize S_r with respect to a_1 and a_0.

$$\frac{\partial S_r}{\partial a_0} = -2\sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)(-1) = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2\sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)(-x_i) = 0$$

giving

$$\sum_{i=1}^{n} a_0 + \sum_{i=1}^{n} a_1 x_i = \sum_{i=1}^{n} y_i$$

$$\sum_{i=1}^{n} a_0 x_i + \sum_{i=1}^{n} a_1 x_i^{\,2} = \sum_{i=1}^{n} y_i x_i$$

Finding Constants of Linear Model

Solving for a_0 and a_1 directly yields,

$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

and

$$a_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$a_0 = \bar{y} - a_1 \bar{x}$$

Curve Fitting

- In many cases, data in the form of (x,y) points is ***sampled*** in time (e.g., every second or every hour) or in space (e.g., every inch or every kilometer)
- ***Curve fitting*** is finding a curve that "best fits" the data points in order to fill in the "missing" data from the sampling
- MATLAB has several curve-fitting functions (and also Curve Fitting Toolbox)
- the function **polyfit(x,y,d)** fits a polynomial of the degree d specified to the data points represented by x and y and returns a polynomial row vector of coefficients

- **Example 1:**
- You have to study the relationship between the monthly e-commerce sales and the online advertising costs. You have the survey results for 7 online stores for the last year. Your task is to find the equation of the straight line that fits the data best.
- The following table represents the survey results from the 7 online stores.

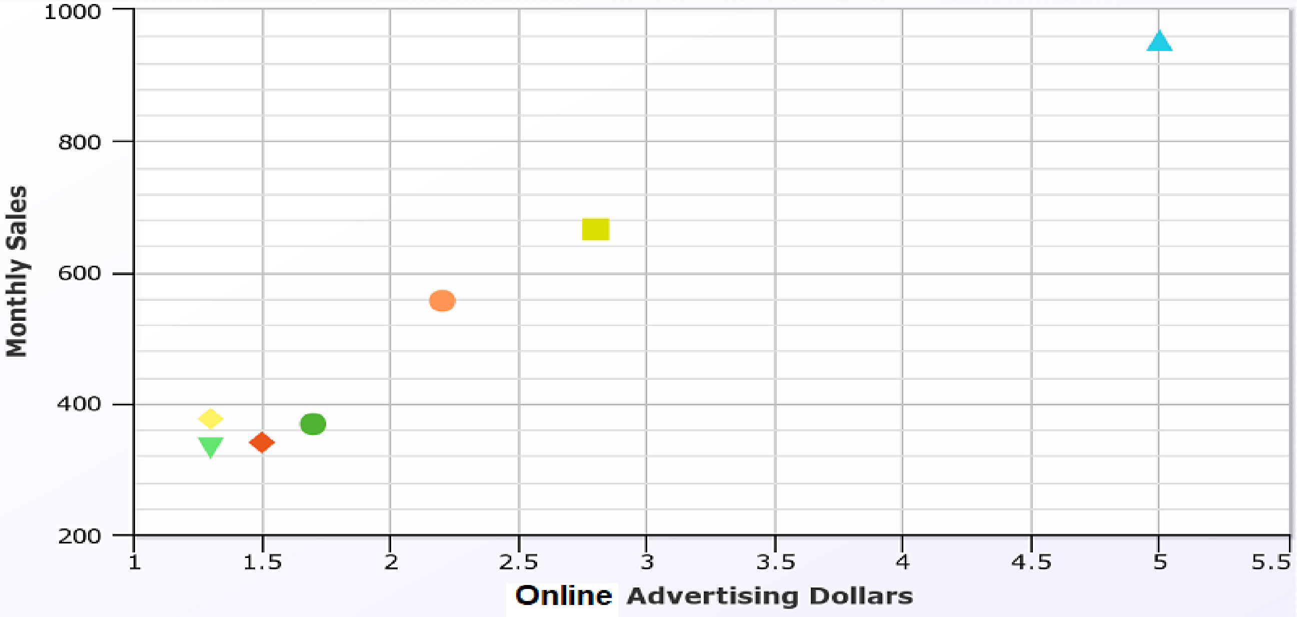
Online Store	Monthly E-commerce Sales (in 1000 s)	Online Advertising Dollars (1000 s)
1	368	1.7
2	340	1.5
3	665	2.8
4	954	5
5	331	1.3
6	556	2.2
7	376	1.3

We can see that there is a **positive relationship** between the monthly e-commerce sales (Y) and online advertising costs (X).

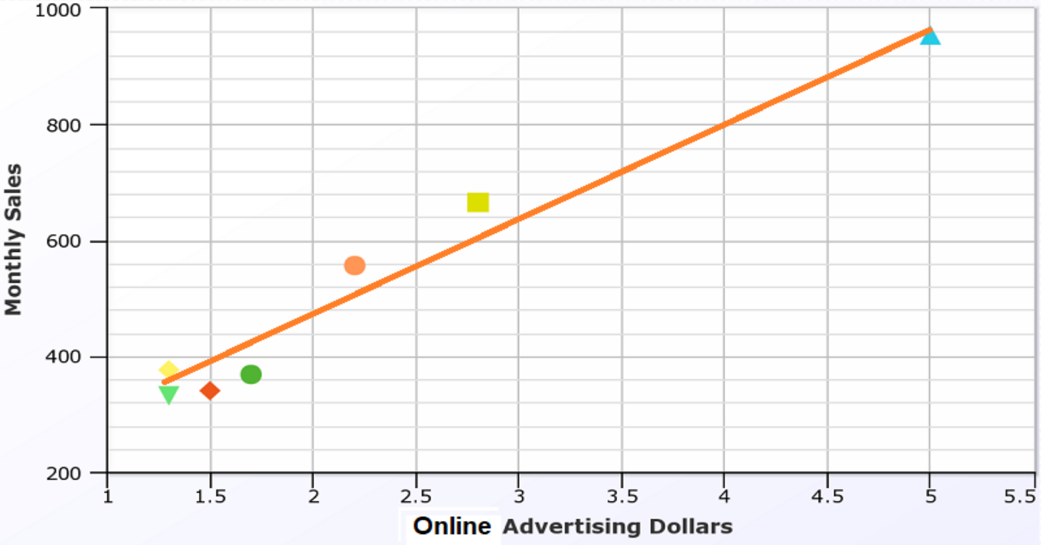
The positive correlation means that the values of the dependent variable (y) increase when the values of the independent variable (x) rise.

So, if we want to predict the monthly e-commerce sales from the online advertising costs, the higher the value of advertising costs, the higher our prediction of sales.

We will use the above data to build our Scatter diagram.



- The Scatter plot shows how much one variable affects another. In our example, above Scatter plot shows how much online advertising costs affect the monthly e-commerce sales. It shows their correlation.
- Let's see the simple linear regression equation.
- $\mathbf{Y = B_0 + B_1X}$
- **Y**= 125.8 + 171.5*X (Matlab Calculation)



- Linear regression aims to find the best-fitting straight line through the points. The best-fitting line is known as the regression line.
- If data points are closer when plotted to making a straight line, it means the correlation between the two variables is higher. In our example, the relationship is strong.
- The orange diagonal line in diagram 2 is the regression line and shows the predicted score on e-commerce sales for each possible value of the online advertising costs.
- **Interpretation of the results:**
- The slope of 171.5 shows that each increase of one unit in X, we predict the average of Y to increase by an estimated 171.5 units.
- The formula estimates that for each increase of 1 dollar in online advertising costs, the expected monthly e-commerce sales are predicted to increase by $171.5.
- This was a simple linear regression example for a positive relationship in business.

Example 2:

You have to examine the relationship between the age and price for used cars sold in the last year by a car dealership company. Here is the table of the data:

Car Age (in years)	Price (in dollars)
4	6300
4	5800
5	5700
5	4500
7	4500
7	4200
8	4100
9	3100
10	2100
11	2500
12	2200

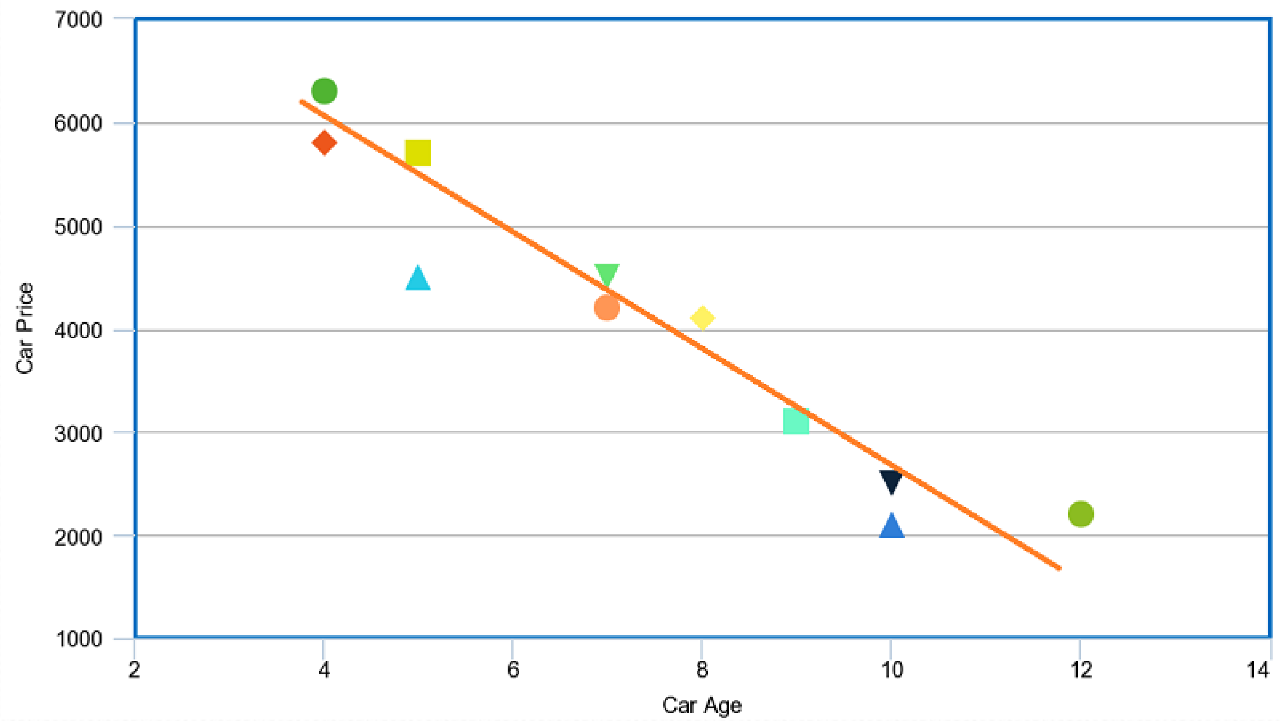
Now, we see that we have a negative relationship between the car price (Y) and car age(X) – as car age increases, price decreases.

When we use the simple linear regression equation, we have the following results:

$\mathbf{Y = B_0 + B_1X}$

Y = 7836 – 502.4*X

Let’s use the data from the table and create our Scatter plot and linear regression line:



- **Result Interpretation:**
- With an estimated slope of – 502.4, we can conclude that the average car price decreases $502.2 for each year a car increases in age.
- The above simple linear regression examples and problems aim to help you understand better the whole idea behind simple linear regression equation.

  Problem-solving using linear regression has so many applications in engineering, business, social, biological, and many many other areas.