Linear Regression

Learning Objectives

- Regression analysis
- Types of Regression Models
- Linear Regression Model
- Slope and Intercept
- Population Linear Regression
- Estimated Regression Model
- Least Squares Criterion
- The Least Squares Equation

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

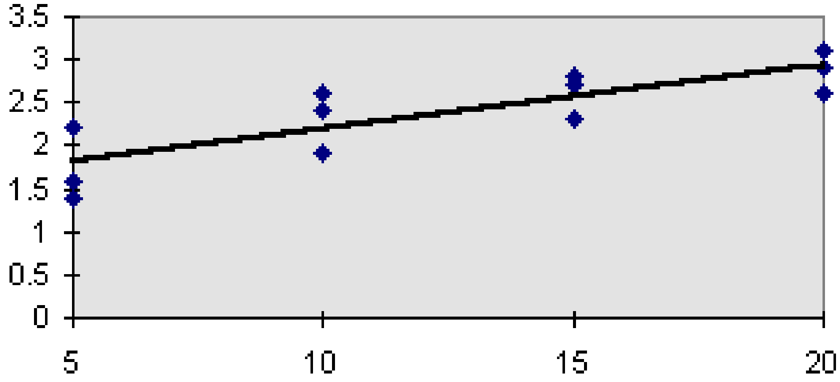
Independent variable: the variable used to explain the dependent variable

Simple Linear Regression Model

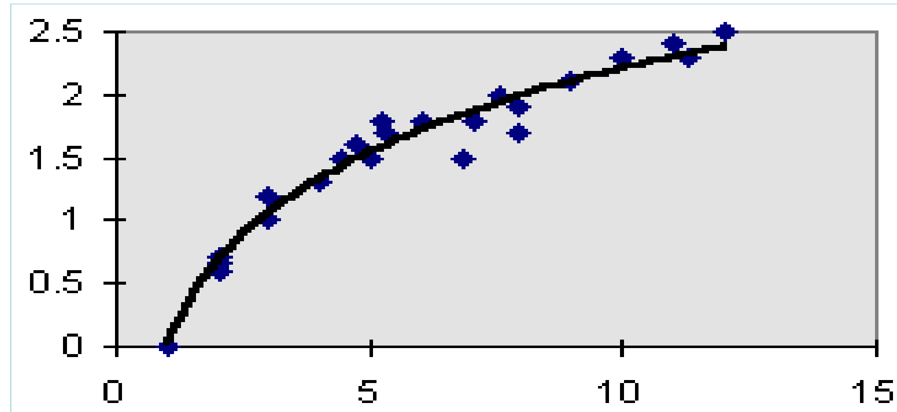
- Only **one independent variable**, x
- Relationship between x and y is described by a linear function
- Changes in y are assumed to be caused by changes in x

Types of Regression Models

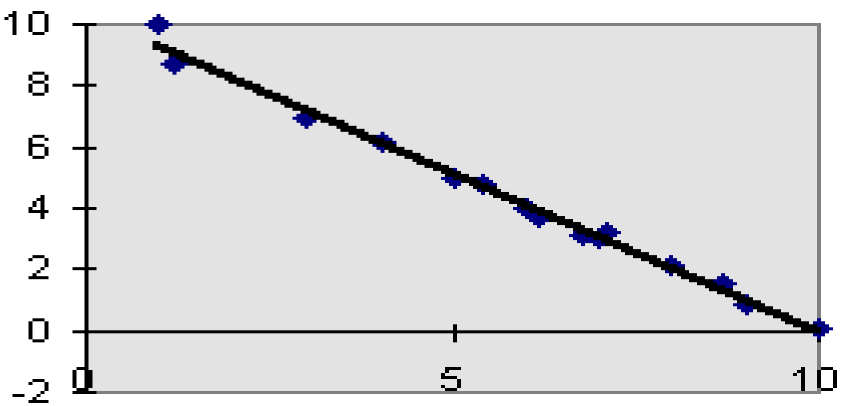
Positive Linear Relationship



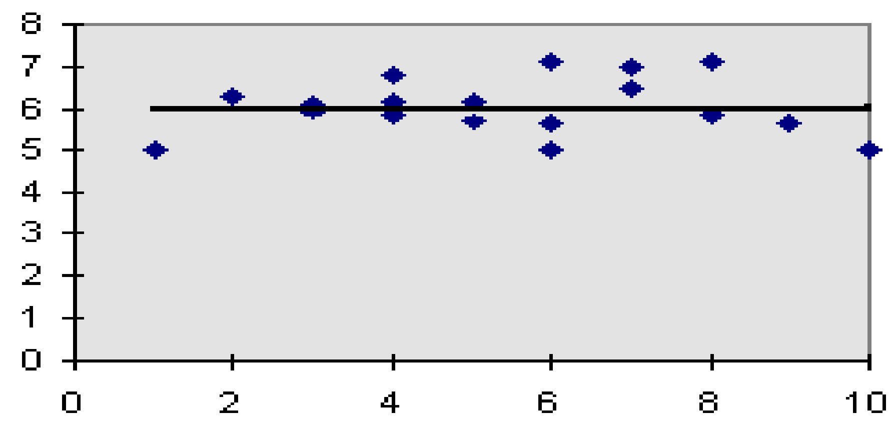
Relationship NOT Linear



Negative Linear Relationship



No Relationship



Population Linear Regression

The population regression model:

Dependent Variable

Population y intercept

Population Slope Coefficient

Independent Variable

Random Error term, or residual

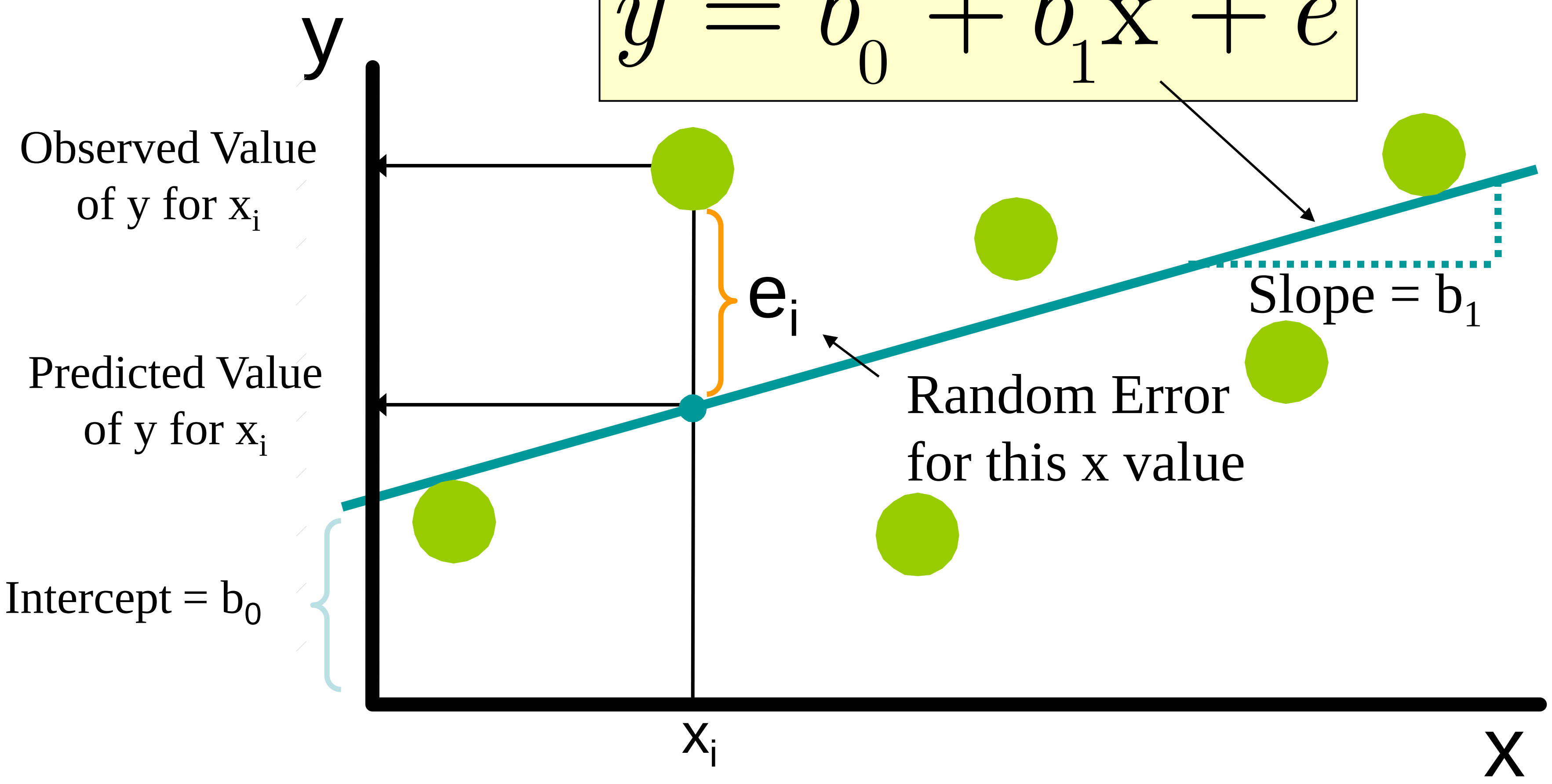
$$y = \underbrace{b_0 + b_1 x}_{\text{Linear component}} + \underbrace{e}_{\text{Random Error component}}$$

Linear Regression Assumptions

- Error values (e) are statistically independent
- Error values are normally distributed for any given value of x
- The probability distribution of the errors is normal
- The probability distribution of the errors has constant variance
- The underlying relationship between the x variable and the y variable is linear

Population Linear Regression

$$y = b_0 + b_1 x + e$$



Estimated Regression Model

The sample regression line provides an estimate of the population regression line

Estimated (or predicted) y value

Estimate of the regression intercept

Estimate of the regression slope

Independent variable

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

The individual random error terms e_i have a mean of zero

Least Squares Criterion

- b_0 and b_1 are obtained by finding the values of b_0 and b_1 that **minimize the sum of the squared residuals**

$$\sum e^2 = \sum (y - \hat{y})^2$$
$$= \sum (y - (b_0 + b_1 x))^2$$

The Least Squares Equation

The formulas for b_1 and b_0 are:

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

algebraic equivalent:

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

and

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

Interpretation of the Slope and the 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

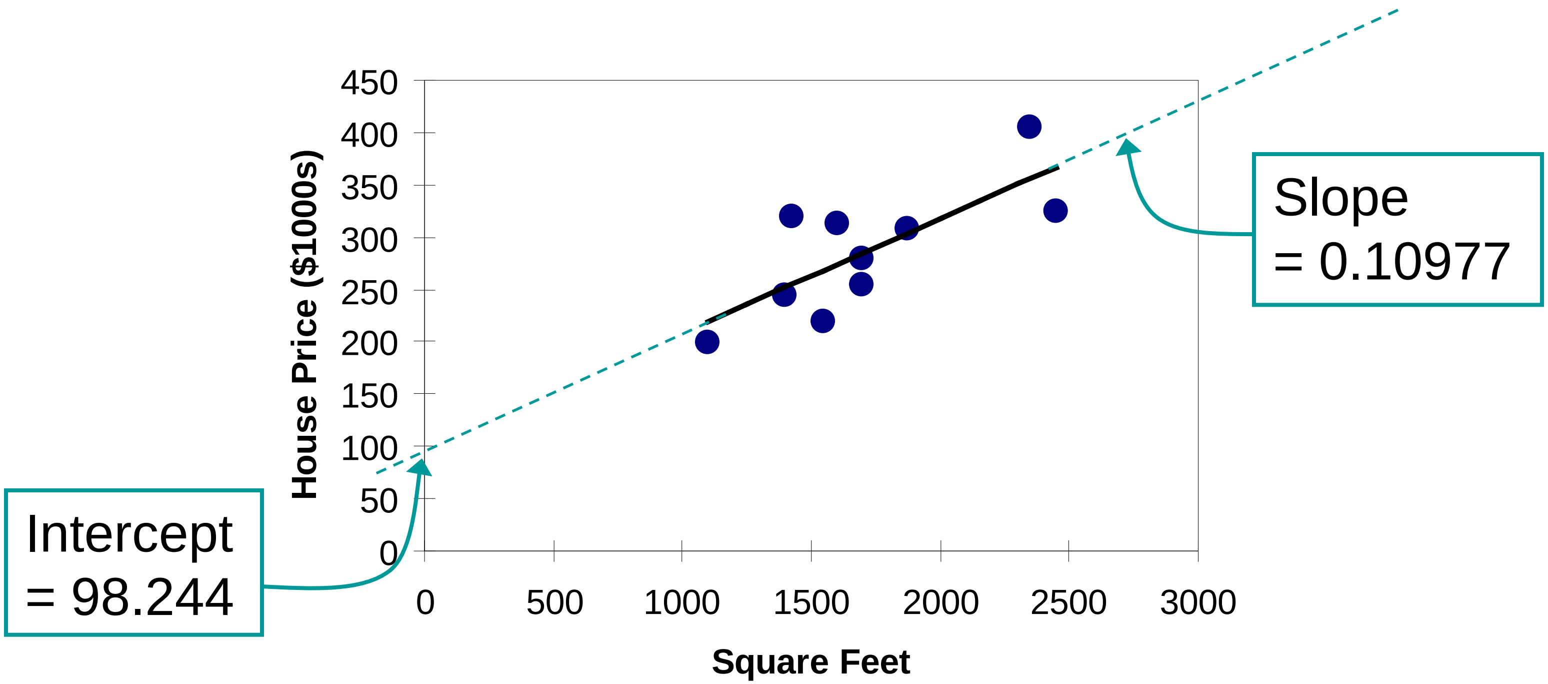
House Price in $1000s	Square Feet			
y	x	xy	y^2	x^2
245	1400	343000	60025	1960000
312	1600	499200	97344	2560000
279	1700	474300	77841	2890000
308	1875	577500	94864	3515625
199	1100	218900	39601	1210000
219	1550	339450	47961	2402500
405	2350	951750	164025	5522500
324	2450	793800	104976	6002500
319	1425	454575	101761	2030625
255	1700	433500	65025	2890000
Σ=2865	Σ=17150	Σ=5085975	Σ=853423	Σ=30983750

$$b_1 = \frac{5085975 - \dfrac{2865 \cdot 17150}{10}}{30983750 - \dfrac{17150^2}{10}} = \frac{172500}{1571500} = 0.10977$$

$$b_0 = \frac{2865}{10} - 0.10977 \cdot \frac{17150}{10} =$$
$$= 286.5 - 188.25555 = 98.24445$$

Graphical Presentation

House price model: scatter plot and regression line



$$\widehat{\text{house price}} = 98.24445 + 0.10977\ (\text{square feet})$$

Interpretation of the Intercept, b_0

$$\widehat{\text{house price}} = \boxed{98.24445} + 0.10977\,(\text{square feet})$$

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 $\boxed{b_0 = 98.24445}$ just indicates that, for houses within the range of sizes observed, \$98,244.45 is the portion of the house price not explained by square feet

Interpretation of the Slope Coefficient, b_1

$$\boxed{\widehat{\text{house price}} = 98.24445 + \boxed{0.10977}\,(\text{square feet})}$$

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

Here, $\boxed{b_1 = 0.10977}$ tells us that the average value of a house increases by 0.10977($1000) = $109.77, on average, for each additional one square foot of size

Least Squares Regression Properties

- The sum of the residuals from the least squares regression line is 0 ($\sum (y - \hat{y}) = 0$)
- The sum of the squared residuals is a minimum (minimized $\sum (y - \hat{y})^2$)
- The simple regression line always passes through the mean of the y variable and the mean of the x variable
- The least squares coefficients are unbiased estimates of b_0 and b_1

Example: House Prices

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

Estimated Regression Equation:

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

Predict the price for a house with 2000 square feet

Example: House Prices

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

The following data represents the age, in years, of California Barracuda and their respective weights, in pounds.

Age (Years)	Weight (Pounds)
1	0.5
2	1.5
3	2.0
4	3.0
5	4.0
6	4.75
7	5.5
8	6.0
9	6.5

Calculating the Correlation Coefficient

Sample correlation coefficient:

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

or the algebraic equivalent:

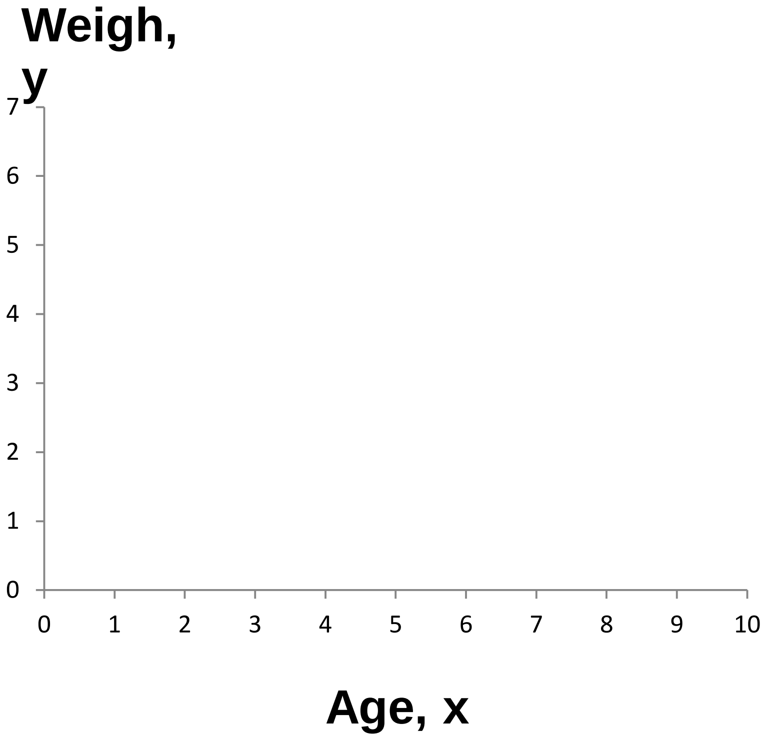
$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

where:

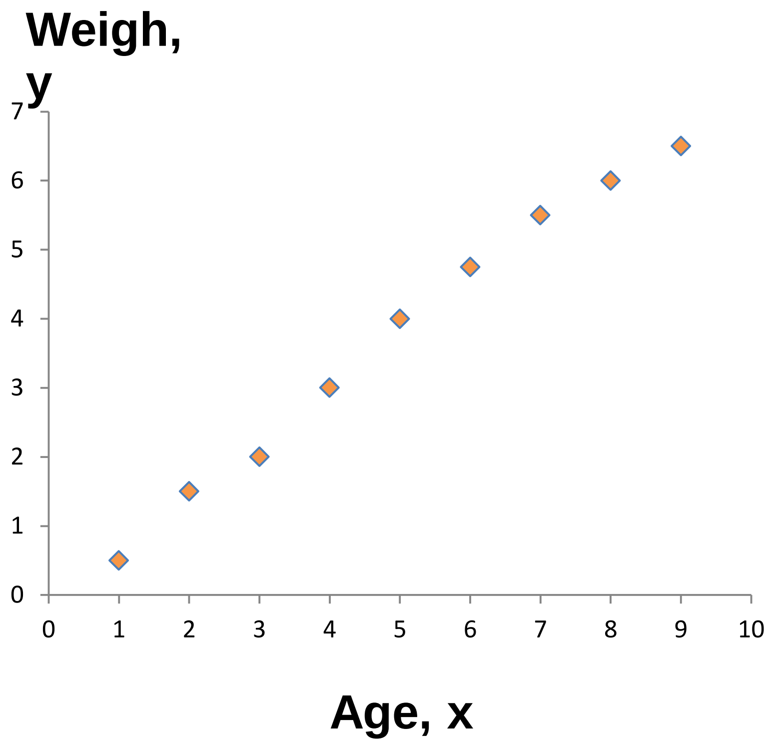
r = Sample correlation coefficient
n = Sample size
x = Value of the independent variable
y = Value of the dependent variable

Age (Years)	Weight (Pounds)			
x	y	xy	y^2	x^2
1	0.5			
2	1.5			
3	2.0			
4	3.0			
5	4.0			
6	4.75			
7	5.5			
8	6.0			
9	6.5			
Σ=	Σ=	Σ=	Σ=	Σ=

Age (Years)	Weight (Pounds)			
x	y	xy	y^2	x^2
1	0.5	0.5	0.25	1
2	1.5	3	2.25	4
3	2.0	6	4	9
4	3.0	12	9	16
5	4.0	20	16	25
6	4.75	28.5	22.56	36
7	5.5	38.5	30.25	49
8	6.0	48	36	64
9	6.5	58.5	42.25	81
Σ=45	Σ=33.75	Σ=215	Σ=162,56	Σ=285



$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$



$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

$$= \frac{9 \cdot 215 - 45 \cdot 33.75}{\sqrt{[9 \cdot 285 - (45)^2][9 \cdot 162.56 - (33.75)^2]}}$$

$$= 0.99517$$

r = 0.99517 → relatively strong positive linear association between x and y

Estimated Regression Model

The sample regression line provides an estimate of the population regression line

Estimated (or predicted) y value

Estimate of the regression intercept

Estimate of the regression slope

Independent variable

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

The individual random error terms e_i have a mean of zero

$$\boxed{b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}}$$

and

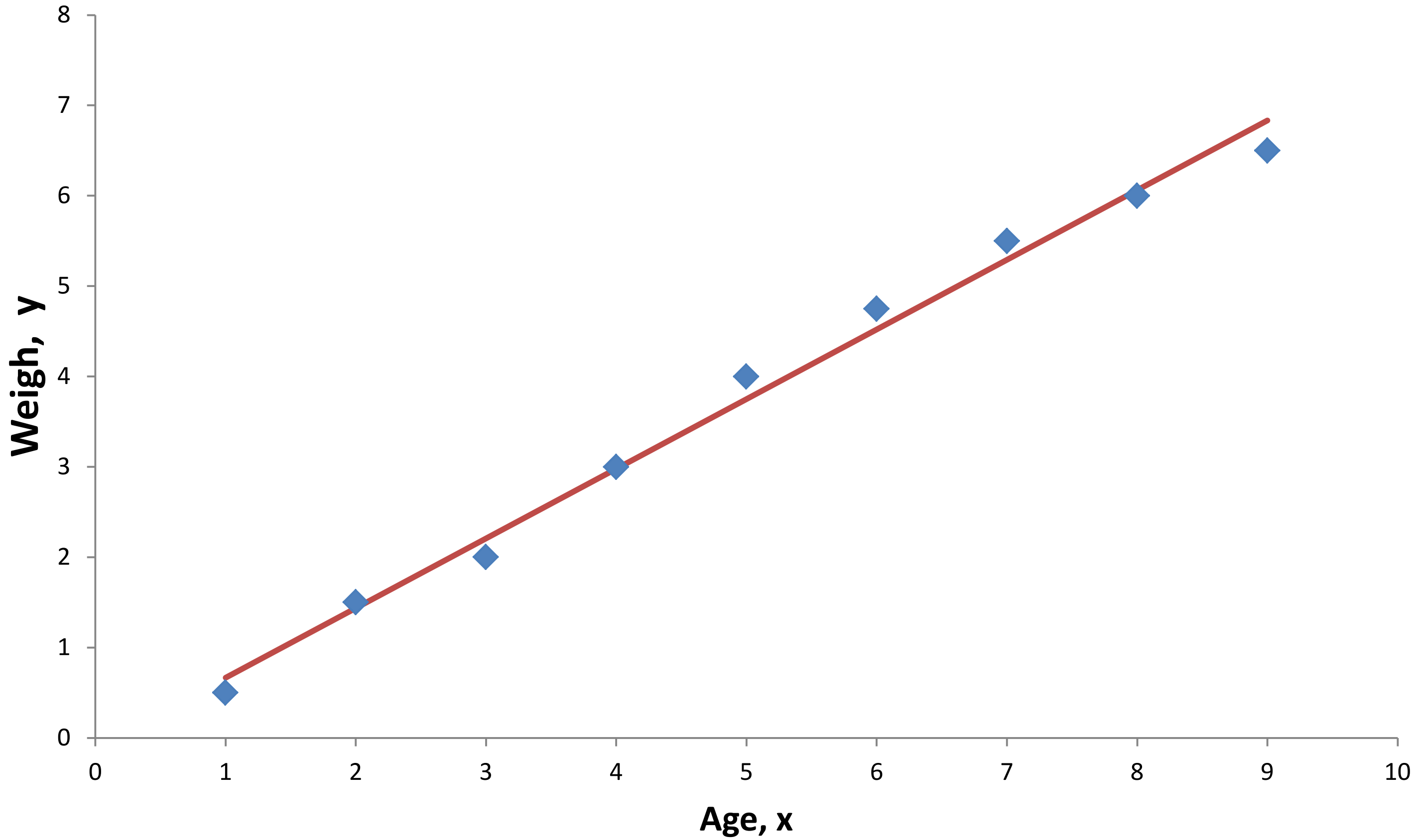
$$\boxed{b_0 = \overline{y} - b_1 \overline{x}}$$

$$b_1 = \frac{\sum xy - \dfrac{\sum x \sum y}{n}}{\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}} = \frac{215 - \dfrac{45 \cdot 33.75}{9}}{285 - \dfrac{(45)^2}{9}} = \frac{46.25}{60} = 0.770833$$

and

$$b_0 = \overline{y} - b_1\overline{x} = \frac{33.75}{9} - 0.770833 \cdot \frac{45}{9} = -0.10417$$

$$\boxed{\widehat{\text{weight}} = -0.10417 + 0.770833\ (\text{age})}$$

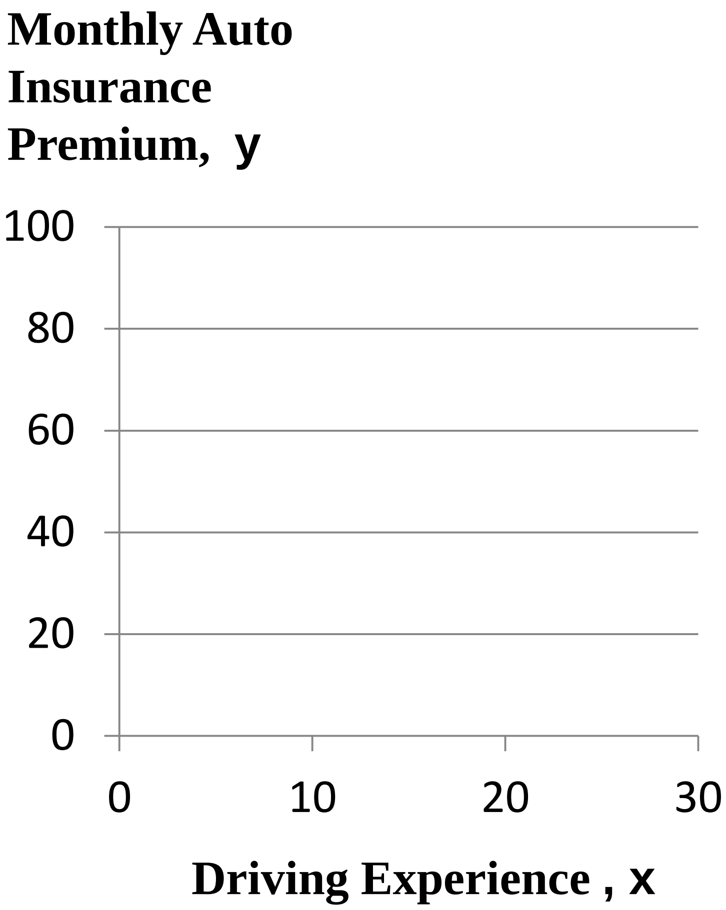


A random sample of eight drivers insured with a company and having similar auto insurance policies was selected. The following table lists their driving experiences (in years) and monthly auto insurance premiums.

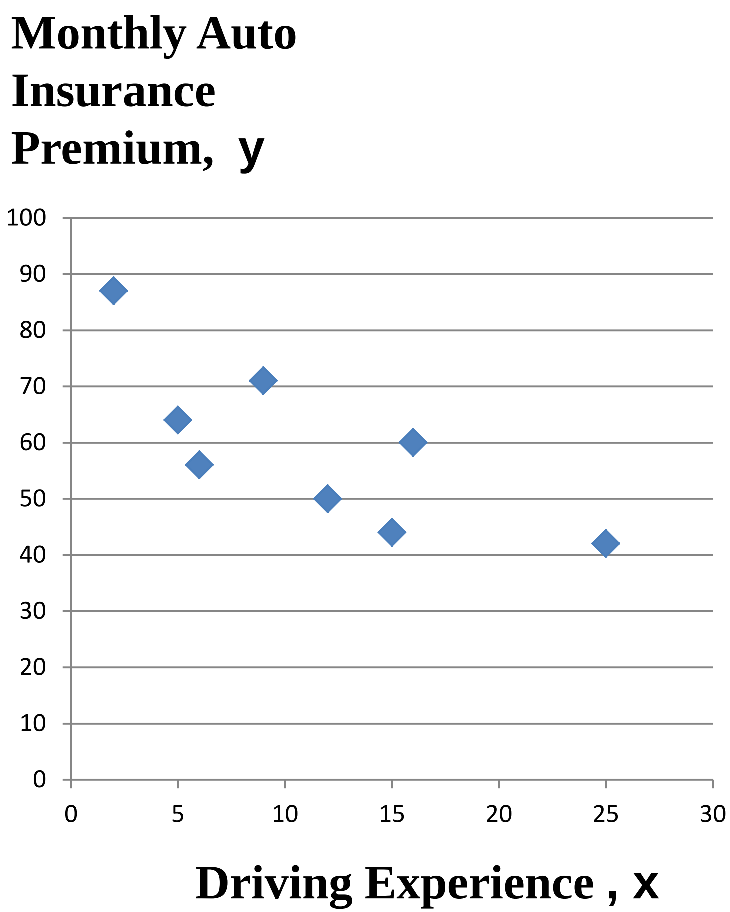
Driving Experience (years)	Monthly Auto Insurance Premium($)
5	64
2	87
12	50
9	71
15	44
6	56
25	42
16	60

Driving Experience (years)	Monthly Auto Insurance Premium			
x	y	xy	y^2	x^2
2	87			
5	64			
6	56			
9	71			
12	50			
15	44			
16	60			
25	42			
Σ=	Σ=	Σ=	Σ=	Σ=

Driving Experience (years)	Monthly Auto Insurance Premium			
x	y	xy	y^2	x^2
2	87	174	7569	4
5	64	320	4096	25
6	56	336	3136	36
9	71	639	5041	81
12	50	600	2500	144
15	44	660	1936	225
16	60	960	3600	256
25	42	1050	1764	625
$\Sigma=90$	$\Sigma=474$	$\Sigma=4739$	$\Sigma=29642$	$\Sigma=1396$



$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$



$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

$$= \frac{8 \cdot 4739 - 90 \cdot 474}{\sqrt{[8 \cdot 1396 - (90)^2][8 \cdot 29642 - (474)^2]}} =$$

$$= -0.76793$$

r = -0,76793 → relatively negative linear association between x and y

$$\boxed{b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}}$$

and

$$\boxed{b_0 = \overline{y} - b_1 \overline{x}}$$

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{4739 - \frac{90 \cdot 474}{8}}{1396 - \frac{(90)^2}{8}} = -1.547588$$

and

$$b_0 = \overline{y} - b_1 \overline{x} = \frac{474}{8} + 1.547588 \cdot \frac{90}{8} = 76.66$$

$$\hat{y} = 76.66 - 1.547588x$$