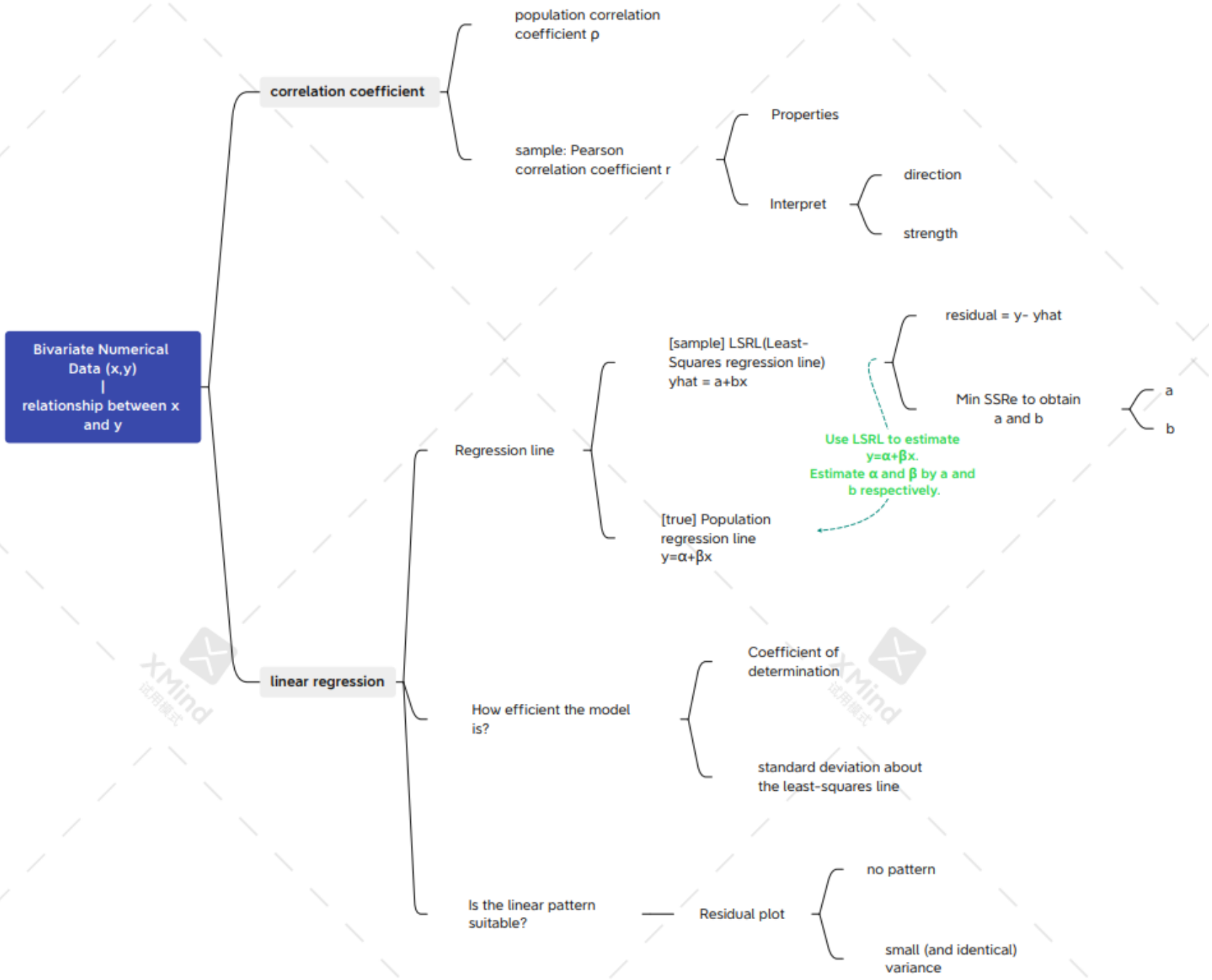
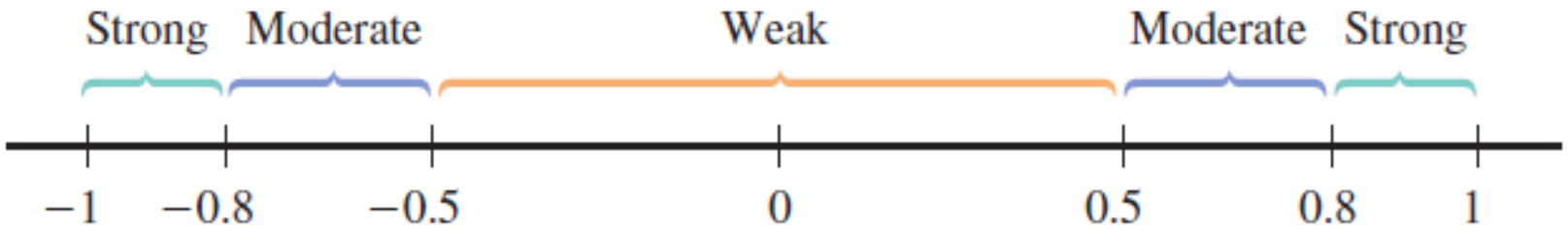
# Review


Bivariate Numerical Data (x,y)
|
relationship between x and y
correlation coefficient
population correlation coefficient ρ
sample: Pearson correlation coefficient r
Properties
Interpret
direction
strength
linear regression
Regression line
[sample] LSRL(Least-Squares regression line)
yhat = a+bx
residual = y- yhat
Min SSRe to obtain a and b
a
b
Use LSRL to estimate y=α+βx.
Estimate α and β by a and b respectively.
[true] Population regression line
y=α+βx
How efficient the model is?
Coefficient of determination
standard deviation about the least-squares line
Is the linear pattern suitable?
Residual plot
no pattern
small (and identical) variance

# Properties of r

1. The value of r does not depend on the unit of measurement for either variable
2. The value of r does not depend on which of the variable is considered x
3. the value of $r$ is between $-1$ and $+1$




4. $r = 1$ occurs only when all the points in a scatterplot of data lie exactly on a straight line that slopes upward. Similarly, $r = -1$ only when all the points lie exactly on a downward-sloping line.

5. the value of $r$ is a measure of the extent to which $x$ and $y$ are linearly related.

# Extrapolation

In prediction, we need to pay attention that the least-squares line should not be used to predict the value of y outside the range of x values in the data set, because we do not know whether the linear pattern observed in the scatter plot continues outside the range. This is sometimes referred to as the danger of extrapolation.

**Outlier:** A point with a large residual (but its x value is normal). Outlier fall far away from the regression line in the y direction.

**Influential observation:** A point with x value differs greatly from the rest of the data, and removal of the point has a large impact on the value of the slope or intercept of the regression line. Influential point fall far away from the rest of the data in the x direction

# Coefficient of determination

$$r^2 = 1 - \frac{SSRe}{SSTo}$$

Where

$SSTo = \sum(y - \bar{y})^2$ is the total sum of squares

$SSRe = \sum(y - \hat{y})^2$ is the residual sum of squares

The value of $r^2$ is often converted to a percentage and interpreted as the percentage of variation in y can be explained by an approximate linear relationship between x and y

# Interpretation

**regression:** interpretation, in context, of

1. $r$ – positive or negative, weak or strong linear association between explanatory variable and response variable

2. $r^2$ – x percent of the variation in the response variable can be explained by the approximate linear relationship with the explanatory variable.

3. **slope** – for every 1 unit increase in the explanatory variable, our model predicts an average increase of $y$ units in the response variable.

4. ***y*-intercept** – at an explanatory variable value of 0 units, our model predicts a response variable value of $y$ units. (Does this make any sense?)

## EXAMPLE 13.2 Mother's Age and Baby's Birth Weight

$x$ = maternal age (in years)

and

$y$ = birth weight of baby (in grams)

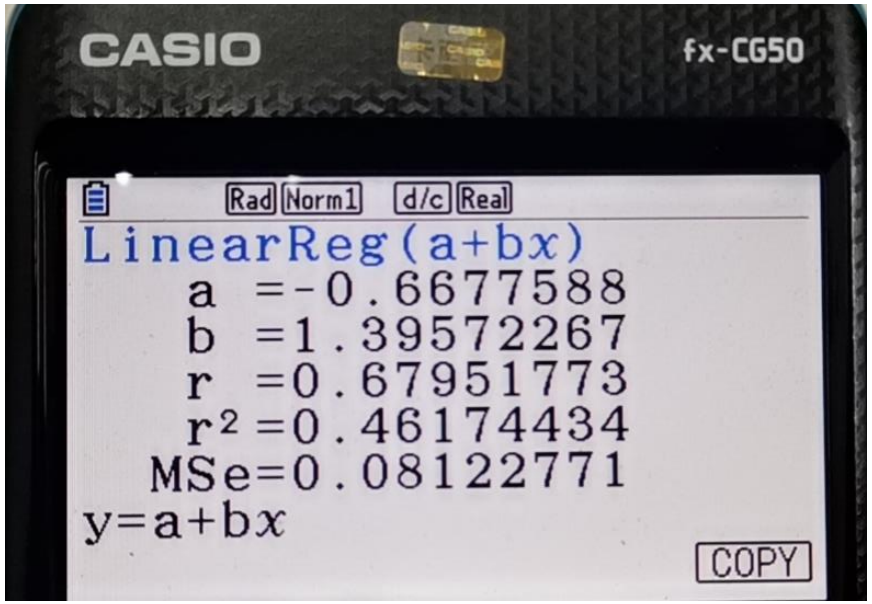
| | OBSERVATION | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| $x$ | 15 | 17 | 18 | 15 | 16 | 19 | 17 | 16 | 18 | 19 |
| $y$ | 2289 | 3393 | 3271 | 2648 | 2897 | 3327 | 2970 | 2535 | 3138 | 3573 |

$$r^2 = 1 - \frac{\text{SSResid}}{\text{SSTo}} = 1 - \frac{0.81228}{1.50909} = 1 - .538 = .462$$

$$s_e^2 = \frac{\text{SSResid}}{n-2} = \frac{0.81228}{10} = .081$$

and

$$s_e = \sqrt{.081} = .285$$



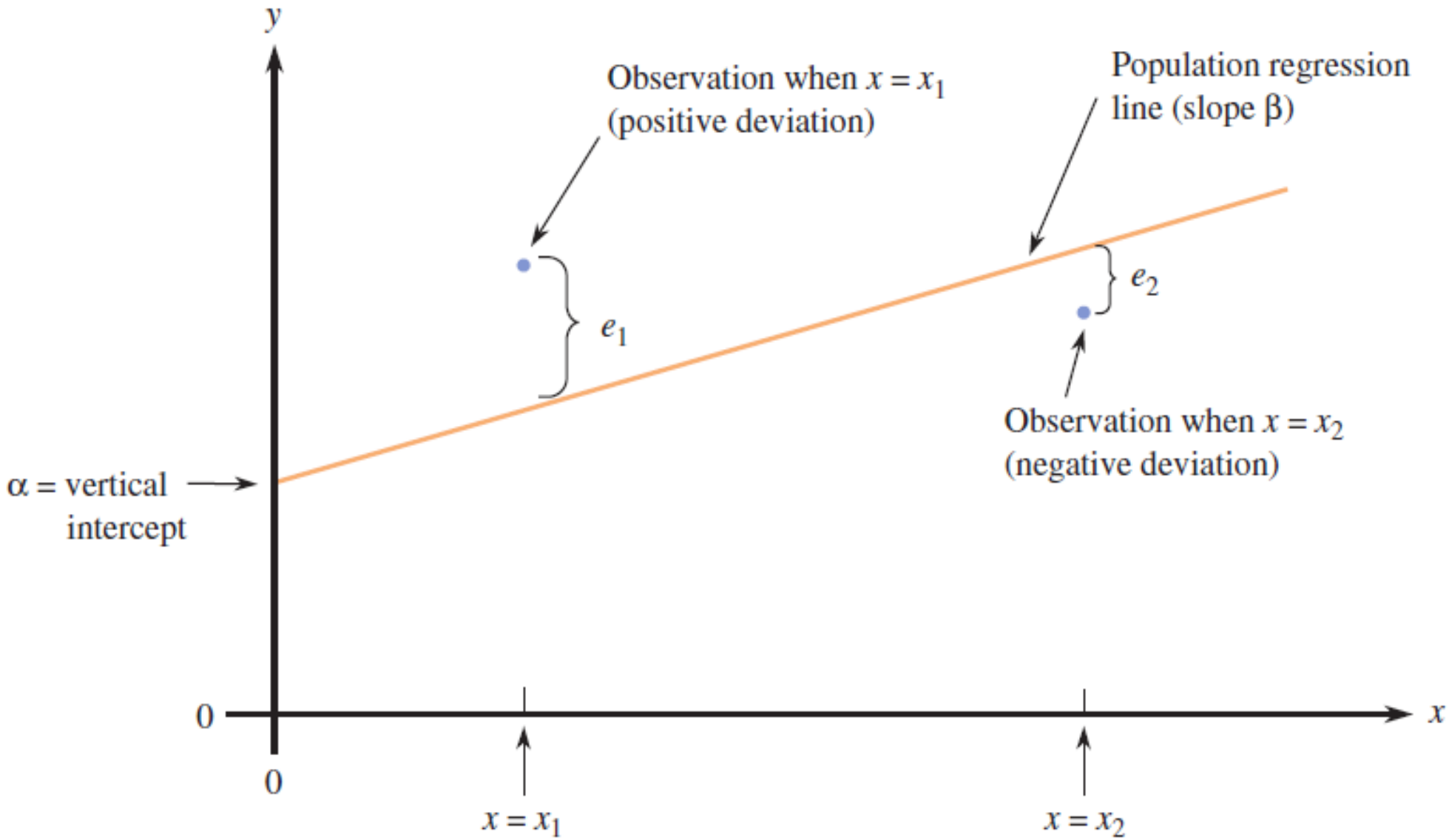
# Simple linear regression model

## DEFINITION

The **simple linear regression model** assumes that there is a line with vertical or $y$ intercept $\alpha$ and slope $\beta$, called the **population regression line**. When a value of the independent variable $x$ is fixed and an observation on the dependent variable $y$ is made,

$$y = \alpha + \beta x + e$$

Without the random deviation $e$, all observed $(x, y)$ points would fall exactly on the population regression line. The inclusion of $e$ in the model equation recognizes that points will deviate from the line by a random amount.


y
Observation when $x = x_1$
(positive deviation)
Population regression
line (slope β)
$e_1$
$e_2$
α = vertical
intercept
Observation when $x = x_2$
(negative deviation)
0
0
x
$x = x_1$
$x = x_2$

# Basic Assumptions of the Simple Linear Regression Model

1. The distribution of $e$ at any particular $x$ value has mean value 0. That is, $\mu_e = 0$.
2. The standard deviation of $e$ (which describes the spread of its distribution) is the same for any particular value of $x$. This standard deviation is denoted by $\sigma$.
3. The distribution of $e$ at any particular $x$ value is normal.
4. The random deviations $e_1, e_2, \ldots, e_n$ associated with different observations are independent of one another.

# The properties of y at a fixed x value

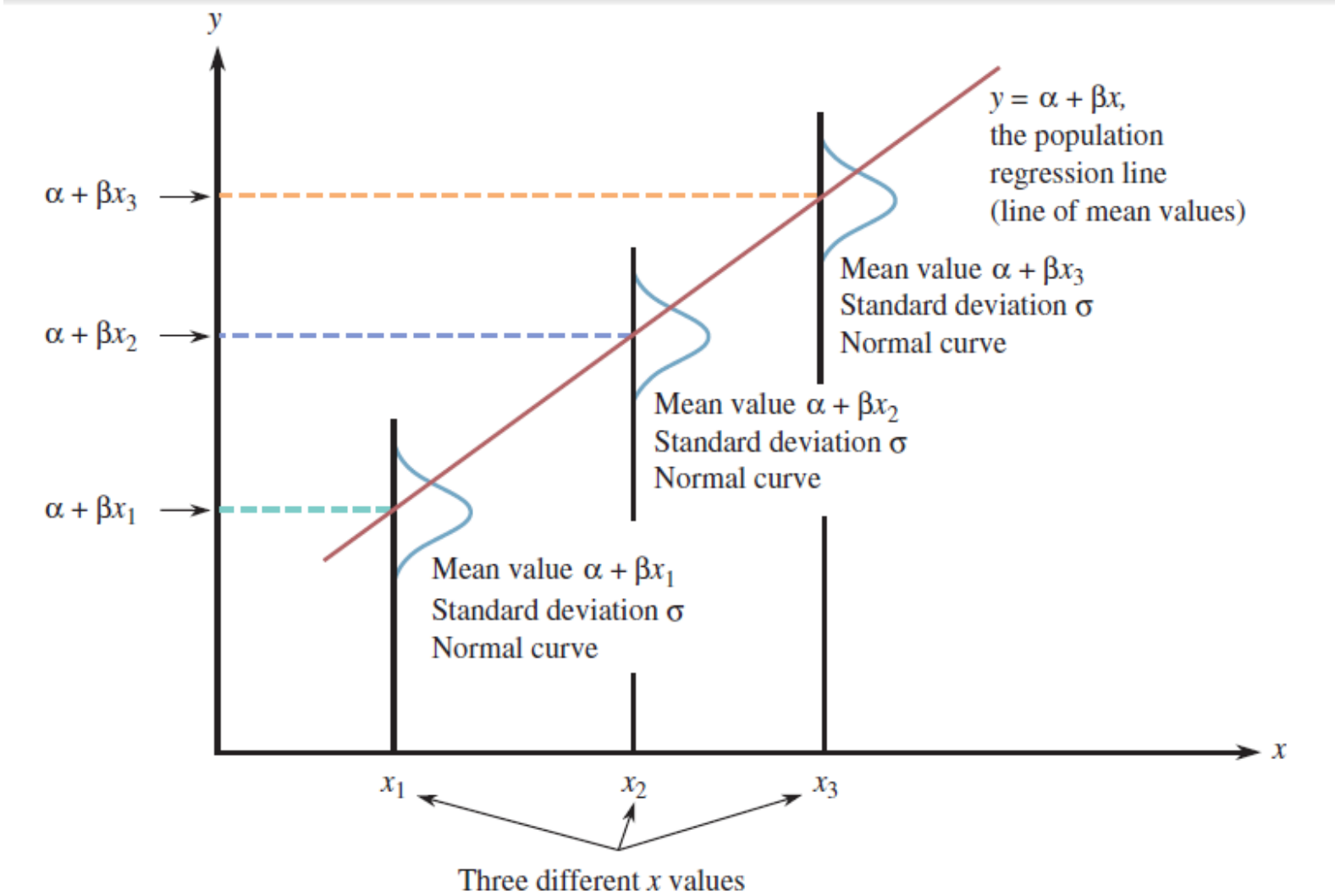
For any fixed x value, what is the distribution of y?

$$y \sim N(\alpha + \beta x, \sigma^2)$$

- slope $\beta$ of the population regression line:
  the average change in y associated with a 1-unit increase in x.

- y intercept $\alpha$:
  the height of the population line when x=0

- $\sigma$: determines the extent to which (x,y) observations deviate from the population line. When $\sigma$ is small, most observations will be quite close to the line, but when $\sigma$ is large, there are likely to be some large deviations


y
$\alpha + \beta x_3$
$\alpha + \beta x_2$
$\alpha + \beta x_1$
$y = \alpha + \beta x$,
the population
regression line
(line of mean values)
Mean value $\alpha + \beta x_3$
Standard deviation $\sigma$
Normal curve
Mean value $\alpha + \beta x_2$
Standard deviation $\sigma$
Normal curve
Mean value $\alpha + \beta x_1$
Standard deviation $\sigma$
Normal curve
x
$x_1$
$x_2$
$x_3$
Three different $x$ values

# Properties of the sampling distribution of b

When the four basic assumptions of the simple linear regression model are satisfied, the following statements are true:

1. The mean value of $b$ is $\beta$. That is, $\mu_b = \beta$, so the sampling distribution of $b$ is always centered at the value of $\beta$. This means that $b$ is an unbiased statistic for estimating $\beta$.
2. The standard deviation of the statistic $b$ is

$$\sigma_b = \frac{\sigma}{\sqrt{S_{xx}}}$$

3. The statistic $b$ has a normal distribution (a consequence of the model assumption that the random deviation $e$ is normally distributed).

1. The true relationship between x and y is Linear
2. The data points must be Independent
3. For a particular value of the explanatory variable, the responses are Normally distributed.
    - The boxplot of the residuals appears to be roughly symmetric and free of outliers.
4. Equal Variance: The standard deviation of errors must be constant across values of x
5. Random
    - ✓ Experiment: treatments are randomly assigned
    - ✓ Observations: random sampling method

If all conditions are met and $\sigma$ is known,

$$\mathrm{b} \sim \mathrm{N}(\beta, \frac{\sigma^2}{S_{xx}})$$

If all conditions are met but $\sigma$ unknown:

- use **Se** to estimate $\sigma$, then the standard deviation of the statistic b is $s_b = \frac{s_e}{\sqrt{S_{xx}}}$

standardized variable: $t = \frac{b - \beta}{s_b}$ , $\mathbf{t \sim t(n-2)}$

Confidence interval: statistic $\pm$ (critical value) $\cdot$ (standard deviation of statistic)

## Confidence Interval for $\beta$

When the four basic assumptions of the simple linear regression model are satisfied, a **confidence interval for $\beta$**, the slope of the population regression line, has the form

$$b \pm (t \text{ critical value}) \cdot s_b$$

where the $t$ critical value is based on df $= n - 2$. Appendix Table 3 gives critical values corresponding to the most frequently used confidence levels.

# Read the regression table

| Predictor | Coef | SE Coef | T | P |
|---|---|---|---|---|
| Constant | –1163.4 | 783.1 | –1.49 | 0.176 |
| Maternal Age | 245.15 | 45.91 | 5.34 | 0.001 |

S = 205.308 R-Sq = 78.1% R-Sq(adj) = 75.4%

Amos collected data about p[illegible]re kilometer) and average rent for 1-bedroom apartments in [illegible]er output from a least-squares regression analysis on his san[illegible]

| Predictor | |
|---|---|
| Constant | |
| Density | 22 |

$S = 262.1 \quad R\text{-sq} = 54.$[illegible]

Assume that all conditions fo[illegible]

**Which of these is a $90\%$ con[illegible]res regression line?**

| df | .25 | .20 | .15 | .10 | .05 |
|---|---|---|---|---|---|
| 1 | 1.000 | 1.376 | 1.963 | 3.078 | 6.314 |
| 2 | .816 | 1.061 | 1.386 | 1.886 | 2.920 |
| 3 | .765 | .978 | 1.250 | 1.638 | 2.353 |
| 4 | .741 | .941 | 1.190 | 1.533 | 2.132 |
| 5 | .727 | .920 | 1.156 | 1.476 | 2.015 |
| 6 | .718 | .906 | 1.134 | 1.440 | 1.943 |
| 7 | .711 | .896 | 1.119 | 1.415 | 1.895 |
| 8 | .706 | .889 | 1.108 | 1.397 | 1.860 |
| 9 | .703 | .883 | 1.100 | 1.383 | 1.833 |
| 10 | .700 | .879 | 1.093 | 1.372 | 1.812 |
| 11 | .697 | .876 | 1.088 | 1.363 | 1.796 |
| 12 | .695 | .873 | 1.083 | 1.356 | 1.782 |
| 13 | .694 | .870 | 1.079 | 1.350 | 1.771 |
| 14 | .692 | .868 | 1.076 | 1.345 | 1.761 |
| 15 | .691 | .866 | 1.074 | 1.341 | 1.753 |
| 16 | .690 | .865 | 1.071 | 1.337 | 1.746 |
| 17 | .689 | .863 | 1.069 | 1.333 | 1.740 |
| 18 | .688 | .862 | 1.067 | 1.330 | 1.734 |
| 19 | .688 | .861 | 1.066 | 1.328 | 1.729 |
| 20 | .687 | .860 | 1.064 | 1.325 | 1.725 |
| 21 | .686 | .859 | 1.063 | 1.323 | 1.721 |
| 22 | .686 | .858 | 1.061 | 1.321 | 1.717 |
| 23 | .685 | .858 | 1.060 | 1.319 | 1.714 |

(A) $22.615 \pm 1.714(4.179)$

(B) $22.615 \pm 1.645(4.179)$

(C) $22.615 \pm 1.319(4.179)$

(D) $812 \pm 1.645(109.8)$

(E) $812 \pm 1.319(109.8)$

# Read the regression table

Analysis of Variance

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Regression | 1 | 74.630 | 74.630 | 15.58 | 0.003 |
| Residual Error | 9 | 43.097 | 4.789 | | |
| Total | 10 | 117.727 | | | |

## EXAMPLE 13.4 Athletic Performance and Cardiovascular Fitness

● Is cardiovascular fitness (as measured by time to exhaustion from running on a treadmill) related to an athlete's performance in a 20-km ski race? The following data on

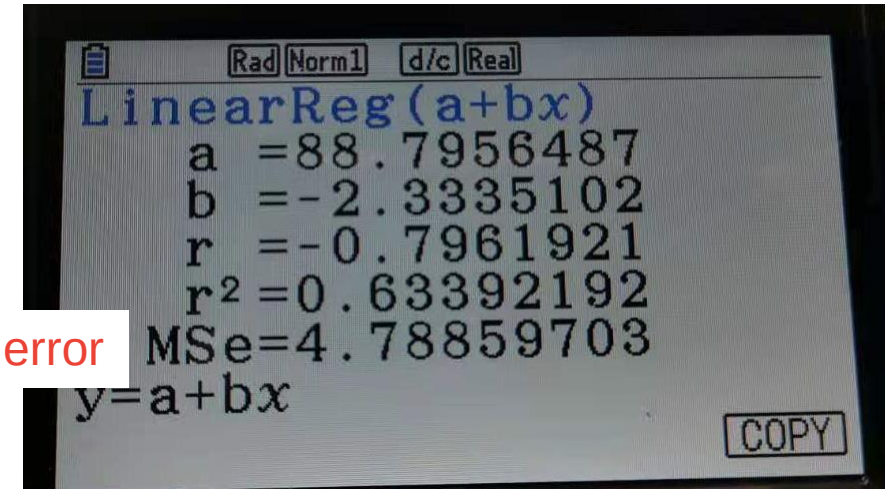
$x$ = treadmill time to exhaustion (in minutes)

and

$y$ = 20-km ski time (in minutes)

were taken from the article "Physiological Characteristics and Performance of Top U.S. Biathletes" (*Medicine and Science in Sports and Exercise* [1995]: 1302–1310):

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $x$ | 7.7 | 8.4 | 8.7 | 9.0 | 9.6 | 9.6 | 10.0 | 10.2 | 10.4 | 11.0 | 11.7 |
| $y$ | 71.0 | 71.4 | 65.0 | 68.7 | 64.4 | 69.4 | 63.0 | 64.6 | 66.9 | 62.6 | 61.7 |




mean squared error

$r^2 = .634$ (63.4% of the observed variation in ski time can be explained by the simple linear regression model)

$s_e^2 = 4.789 \qquad s_e = 2.188$

$$s_b = \frac{s_e}{\sqrt{S_{xx}}} = \frac{2.188}{3.702} = .591$$

## Regression Analysis

The regression equation is
ski time = 88.8 − 2.33 treadmill time

| Predictor | Coef | StDev | T | P |
|---|---|---|---|---|
| Constant | 88.796 | 5.750 | 15.44 | 0.000 |
| treadmill | −2.3335 | 0.5911 | −3.95 | 0.003 |

S = 2.188  R-Sq = 63.4%  R-Sq (adj) = 59.3%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Regression | 1 | 74.630 | 74.630 | 15.58 | 0.003 |
| Residual Error | 9 | 43.097 | 4.789 | | |
| Total | 10 | 117.727 | | | |

**Regression Analysis**

The regression equation is

ski time = 88.8 − 2.33 treadmill time

*Equation of the estimated regression line* $\hat{y} = a + bx$

*Estimated y intercept a*

*Estimated slope b*

| Predictor | Coef | StDev | T | P |
|---|---|---|---|---|
| Constant | 88.796 | 5.750 | 15.44 | 0.000 |
| treadmill | −2.3335 | 0.5911 | −3.95 | 0.003 |

$s_b$ = *estimated standard deviation of b*

S = 2.188 R-Sq = 63.4% R-Sq (adj) = 59.3%

$s_e$

$r^2$

Analysis of Variance

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Regression | 1 | 74.630 | 74.630 | 15.58 | 0.003 |
| Residual Error | 9 | 43.097 | 4.789 | | |
| Total | 10 | 117.727 | | | |

$n - 2$ = *residual df*

*SSResid*

*SSTo*

$s_e^2$

# Hypothesis test concerning $\beta$

## Summary of Hypothesis Tests Concerning $\beta$

**Null hypothesis:** $H_0$: $\beta$ = hypothesized value

**Test statistic:** $t = \dfrac{b - \text{hypothesized value}}{s_b}$

The test is based on df $= n - 2$.

| **Alternative hypothesis:** | ***P*-value:** |
|---|---|
| $H_a$: $\beta >$ hypothesized value | Area to the right of the computed $t$ under the appropriate $t$ curve |
| $H_a$: $\beta <$ hypothesized value | Area to the left of the computed $t$ under the appropriate $t$ curve |
| $H_a$: $\beta \neq$ hypothesized value | (1) 2(area to the right of $t$) if $t$ is positive<br>or<br>(2) 2(area to the left of the $t$) if $t$ is negative |

The **model utility test for simple linear regression** is the test of

$$H_0: \beta = 0$$

versus

$$H_a: \beta \neq 0$$

$H_0$ specifies that there is no useful linear relationship between $x$ and $y$,

$H_a$ specifies that there is a useful linear relationship between $x$ and $y$.

- If $H_0$ is rejected, we conclude that the simple linear regression model is useful for predicting y.