# MATH1. Part II

## Probability and Statistics

# Chapter 12

## Linear Regression and Correlation

# 12.2 The Correlation Coefficient

## Correlation coefficient

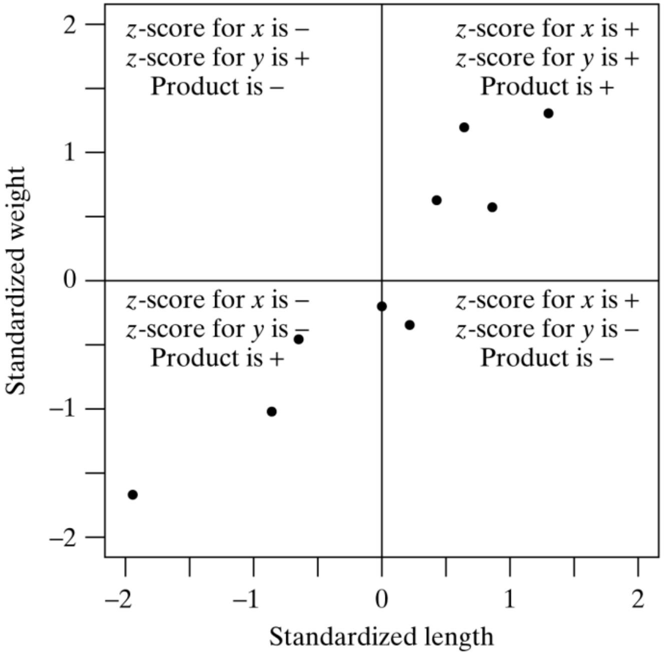
- Suppose we have a sample of n pairs for which each pair represents the measurements of two variables, X and Y.
- **Correlation coefficient** measures the strength of **linear association** between **two** quantitative variables (e.g. X and Y).

The correlation coefficient, $r$

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- Dividing the plot into quadrants based on the sign of the standardized score.
- r positive : points fall into the upper-right and lower-left quadrants.
- r negative: points fall into the upper-left and lower-right quadrants.
- Sum of these products provides a numeric measure of where our points fall.

**Figure 12.2.2** Scatterplot of standardized weight versus standardized length



# 12.2 The Correlation Coefficient

## Correlation coefficient

The correlation coefficient, *r*

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- r measure of the strength of linear association between X and Y.
- Scale-invariant numeric (unit free)
- Sign of r indicates the sign of the relationship: positive (increasing) or negative (decreasing).
- The closer the correlation is to -1 or 1, the stronger the linear relationship between X and Y.
- r = 0, if there were no linear relationship  (the points would fall in evenly in all four quadrants so that the positive and negative products would balance.

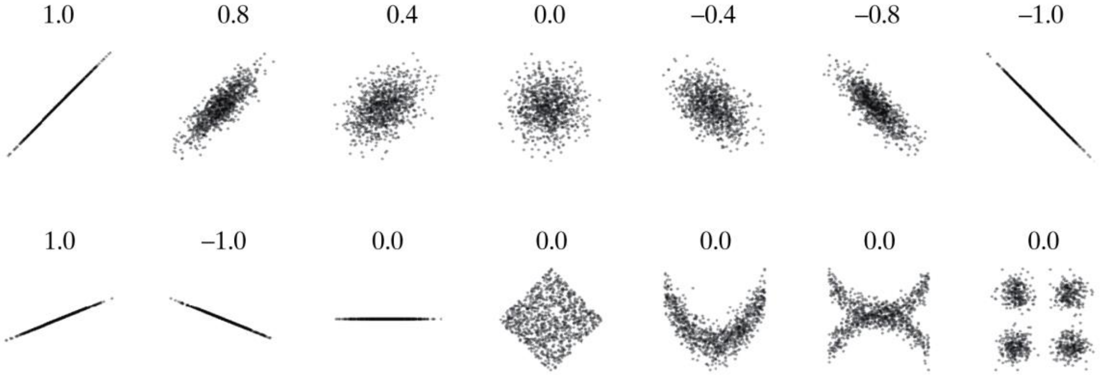


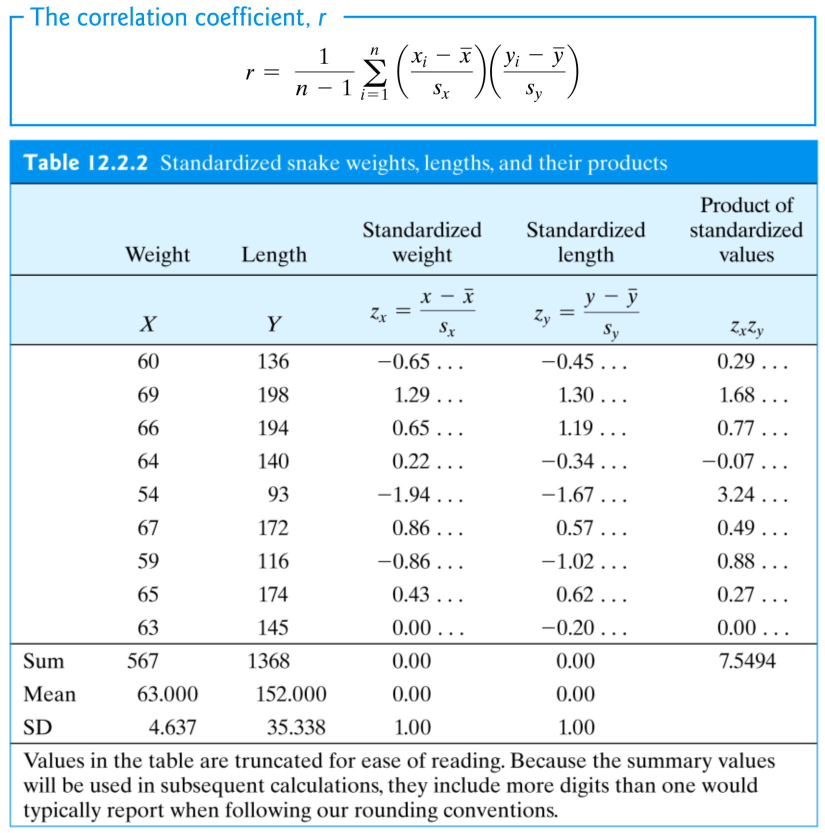
**Figure 12.2.3** Scatterplots of data with a variety of sample correlation values

# 12.2 The Correlation Coefficient

## Correlation coefficient

### Example 12.2.1 Length and Weight of snakes

- study body lengths and weights of a population of the Vipera bertis snake
- researchers caught and measured nine adult females.
- What is the correlation coefficient r?



The correlation coefficient, $r$

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

**Table 12.2.2** Standardized snake weights, lengths, and their products

| | Weight | Length | Standardized weight | Standardized length | Product of standardized values |
|---|---|---|---|---|---|
| | $X$ | $Y$ | $z_x = \frac{x - \bar{x}}{s_x}$ | $z_y = \frac{y - \bar{y}}{s_y}$ | $z_x z_y$ |
| | 60 | 136 | −0.65 . . . | −0.45 . . . | 0.29 . . . |
| | 69 | 198 | 1.29 . . . | 1.30 . . . | 1.68 . . . |
| | 66 | 194 | 0.65 . . . | 1.19 . . . | 0.77 . . . |
| | 64 | 140 | 0.22 . . . | −0.34 . . . | −0.07 . . . |
| | 54 | 93 | −1.94 . . . | −1.67 . . . | 3.24 . . . |
| | 67 | 172 | 0.86 . . . | 0.57 . . . | 0.49 . . . |
| | 59 | 116 | −0.86 . . . | −1.02 . . . | 0.88 . . . |
| | 65 | 174 | 0.43 . . . | 0.62 . . . | 0.27 . . . |
| | 63 | 145 | 0.00 . . . | −0.20 . . . | 0.00 . . . |
| Sum | 567 | 1368 | 0.00 | 0.00 | 7.5494 |
| Mean | 63.000 | 152.000 | 0.00 | 0.00 | |
| SD | 4.637 | 35.338 | 1.00 | 1.00 | |

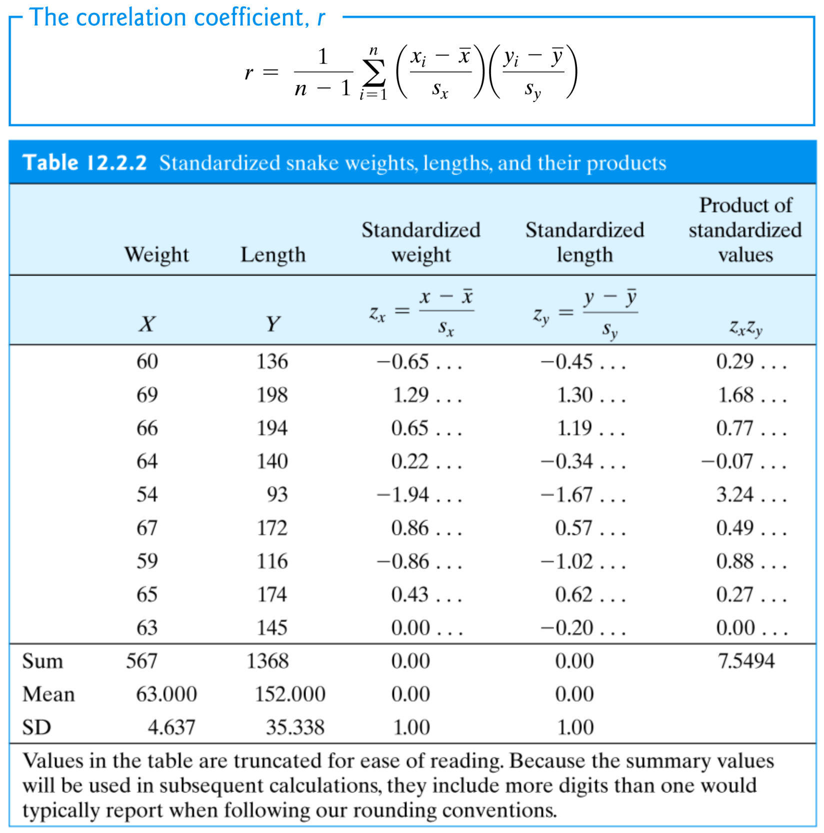
Values in the table are truncated for ease of reading. Because the summary values will be used in subsequent calculations, they include more digits than one would typically report when following our rounding conventions.

# 12.2 The Correlation Coefficient

## Correlation coefficient

### Example 12.2.1 Length and Weight of snakes

- study body lengths and weights of a population of the Vipera bertis snake
- researchers caught and measured nine adult females.
- What is the correlation coefficient r?
  - r = 1/ (9 – 1) x 7.5494 ≈ 0.94
    - the value 0.94 as the **sample correlation**, since the lengths and weights of these nine snakes comprise a sample from a larger population.
    - The sample correlation is an estimate of the **population correlation** (often denoted by the Greek letter “rho,” ρ)
    - In this case, ρ = the entire population of adult female Vipera bertis snakes.



The correlation coefficient, *r*

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

**Table 12.2.2** Standardized snake weights, lengths, and their products

| | Weight | Length | Standardized weight | Standardized length | Product of standardized values |
|---|---|---|---|---|---|
| | $X$ | $Y$ | $z_x = \frac{x-\bar{x}}{s_x}$ | $z_y = \frac{y-\bar{y}}{s_y}$ | $z_xz_y$ |
| | 60 | 136 | −0.65 . . . | −0.45 . . . | 0.29 . . . |
| | 69 | 198 | 1.29 . . . | 1.30 . . . | 1.68 . . . |
| | 66 | 194 | 0.65 . . . | 1.19 . . . | 0.77 . . . |
| | 64 | 140 | 0.22 . . . | −0.34 . . . | −0.07 . . . |
| | 54 | 93 | −1.94 . . . | −1.67 . . . | 3.24 . . . |
| | 67 | 172 | 0.86 . . . | 0.57 . . . | 0.49 . . . |
| | 59 | 116 | −0.86 . . . | −1.02 . . . | 0.88 . . . |
| | 65 | 174 | 0.43 . . . | 0.62 . . . | 0.27 . . . |
| | 63 | 145 | 0.00 . . . | −0.20 . . . | 0.00 . . . |
| Sum | 567 | 1368 | 0.00 | 0.00 | 7.5494 |
| Mean | 63.000 | 152.000 | 0.00 | 0.00 | |
| SD | 4.637 | 35.338 | 1.00 | 1.00 | |

Values in the table are truncated for ease of reading. Because the summary values will be used in subsequent calculations, they include more digits than one would typically report when following our rounding conventions.

# 12.2 The Correlation Coefficient

## Correlation coefficient

The correlation coefficient, *r*

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- In order to regard the sample correlation coefficient r as an estimate of a population parameter, it must be reasonable to assume that both the X and the Y values were selected at random, as in the following bivariate random sampling model:

Bivariate Random Sampling Model:

We regard each pair $(x_i, y_i)$ as having been sampled at random from a population of $(x, y)$ pairs.

# 12.2 The Correlation Coefficient

## R testing the hypothesis $H_0$: $\rho = 0$

Now we shall consider statistical inference based on r for data from bivariate random sampling model.

- Null hypothesis $H_0$: There is no linear relationship between X and Y.
- A t test
  - the test statistic: $t_s = r\sqrt{\frac{n-2}{1-r^2}}$
  - Critical values are obtained from Student's t-distribution with df = n – 2

### Example 12.2.3-4 blood Pressure and Platelet calcium

- It is suspected that calcium in blood platelets may be related to blood pressure.
- The sample size is n = 38
- The sample correlation is r = 0.5832.
- Is there evidence that blood pressure and platelet calcium are linearly related?

# 12.2 The Correlation Coefficient

## R testing the hypothesis $H_0$: ρ = 0

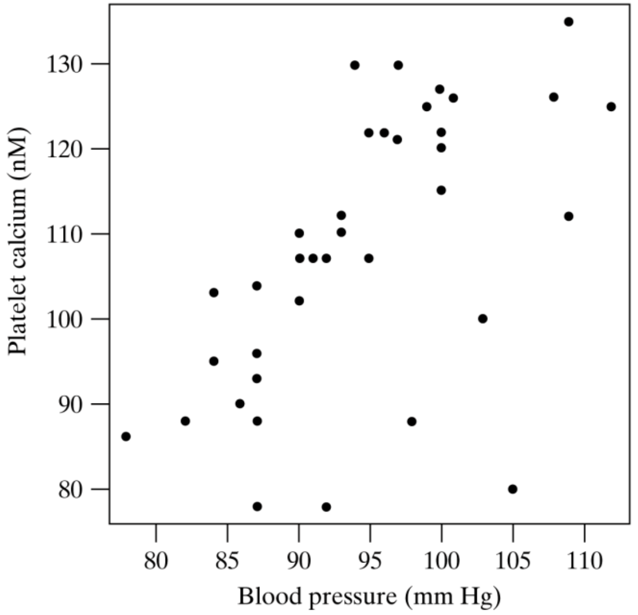
### Example 12.2.3-4 blood Pressure and Platelet calcium

- The sample size is n = 38; the sample correlation is r = 0.5832.
- Is there evidence that blood pressure and platelet calcium are linearly related?
  - $H_0$: ρ = 0. Platelet calcium is not linearly related to blood pressure.
  - $H_A$: ρ ≠ 0. Platelet calcium is linearly related blood pressure (nondirectional alternative).
  - Let us choose α = 0.05. The test statistic is

$$t_s = r\sqrt{\frac{n-2}{1-r^2}} = 0.5832\sqrt{\frac{38-2}{1-0.58322}} = 4.308$$

  - From Table 4 with df = n - 2 = 36 ≈ 40, we find $t_{40,\ 0.0005}$ = 3.551.
  - Thus, we find P-value < 0.05, and we reject $H_0$ .
  - The data provide strong evidence that platelet calcium is linearly related blood pressure.

**Figure 12.2.4** Blood pressure and platelet calcium for 38 persons with normal blood pressure



# 12.2 The Correlation Coefficient

## Correlation and Causation

- We have noted earlier that an observed association between two variables does NOT necessarily indicate any causal connection between them.
- One way to establish causality is to conduct a controlled experiment in which the putative causal factor is varied and all other factors are either held constant or controlled by randomization.

# 12.3 The Fitted Regression Line

## Fitted Regression Line

- In this section, we will learn how to find and interpret the line that best summarizes their linear relationship.
- The equation of a straight line can be written as $Y = b_0 + b_1 X$
  - where $b_0$ is the y-intercept and $b_1$ is the slope of the line.
  - The slope $b_1$ is the rate of change of Y with respect to X.
- The **least-squares regression line** of Y on X is written $\hat{y} = b_0 + b_1 x$

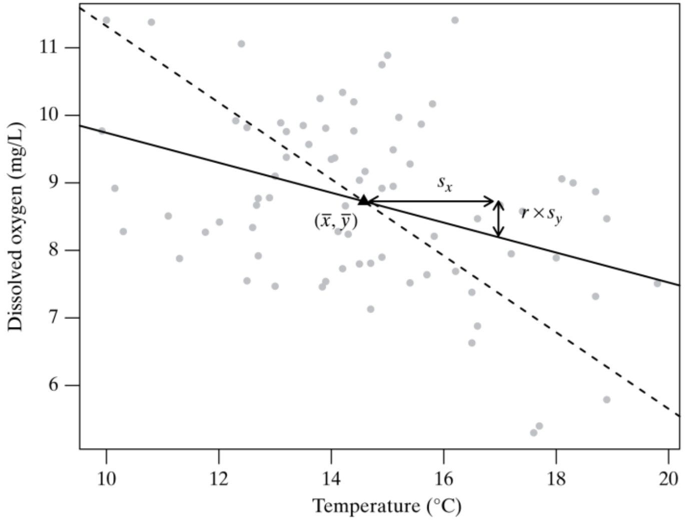
Least-Squares Regression Line of *Y* on *X*

$$\text{Slope: } b_1 = r\left(\frac{s_y}{s_x}\right)$$

$$\text{Intercept: } b_0 = \bar{y} - b_1\bar{x}$$

  - $\hat{y}$ (read "Y-hat") is only the estimated or predicted Y values
  - regression line passes through the joint mean $(\bar{x}, \bar{y})$ of our data
  - the SD line: r = ±1 (i.e., perfect linear relationships ), slope $\pm s_y/s_x$

**Figure 12.3.4** Levels of dissolved oxygen versus water temperature for 75 days with SD line (dashed) and fitted regression line (solid)

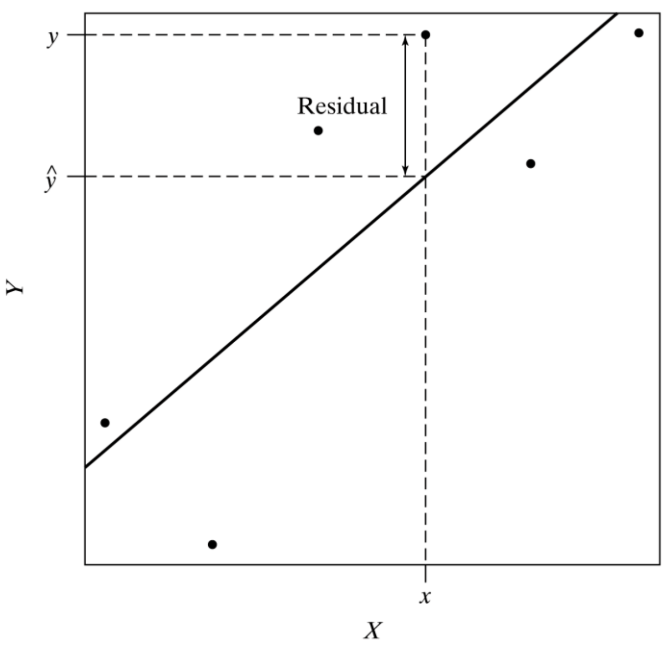


# 12.3 The Fitted Regression Line

## The residual, $e_i$

- We now consider a statistic that describes the scatter of the points about the fitted regression line.
- For each observed $x_i$ in our data there is a predicted Y value of $\hat{y}_i = b_0 + b_1 x_i$
- A **residual** is defined as $\mathbf{e_i = y_i - \hat{y}_i}$
  - Note that a residual is calculated in terms of vertical distance.
  - It can be shown that the sum of the residuals, taking into account their signs, is always zero, because of “balancing” of data points above and below the fitted regression line.
  - The magnitude (absolute value) of each residual is the vertical distance of the data point from the fitted line.

**Figure 12.3.6** $\hat{y}$ and the residual for a typical data point $(x, y)$



# 12.3 The Fitted Regression Line

## The residual sum of squares, SS(resid)

- A **residual** is defined as **$e_i = y_i - \hat{y}_i$**
- A summary measure of the distances of the data points from the regression line is the **residual sum of squares**, or SS(resid),

Residual Sum of Squares

$$\text{SS(resid)} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

  – the residual sum of squares will be small if the data points all lie very close to the line.

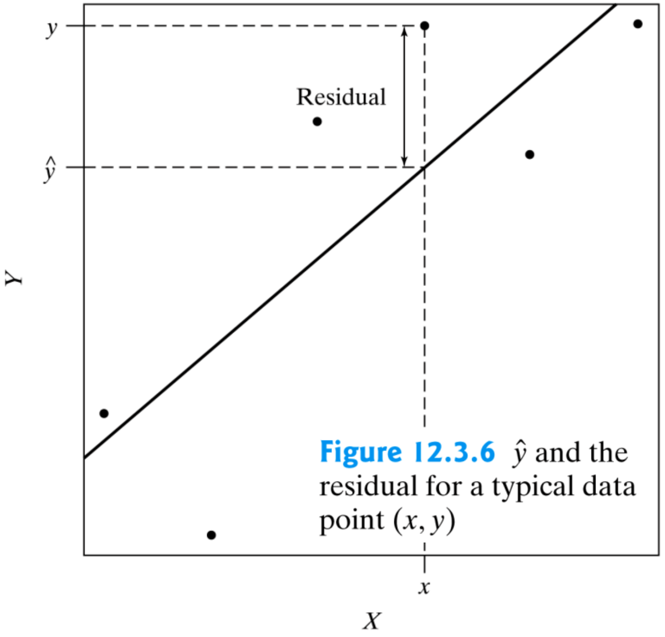


Figure 12.3.6 $\hat{y}$ and the residual for a typical data point $(x, y)$

## The least-squares criterion

- The classical criterion, which define the straight line that “best” fits a set of data point, is the least-squares criterion:

Least-Squares Criterion

The “best” straight line is the one that minimizes the residual sum of squares.

# 12.3 The Fitted Regression Line

## The residual standard deviation, $s_e$

- A **residual** is defined as $\mathbf{e_i = y_i - \hat{y}_i}$
- The **residual standard deviation**, denoted $s_e$, specifies how far off predictions made using the regression model tend to be.

Residual Standard Deviation

$$s_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}e_i^2}{n-2}} = \sqrt{\frac{\text{SS(resid)}}{n-2}}$$

- residual SD measures variability around the regression line
- the ordinary SD measures variability around the mean, $\bar{y}$.

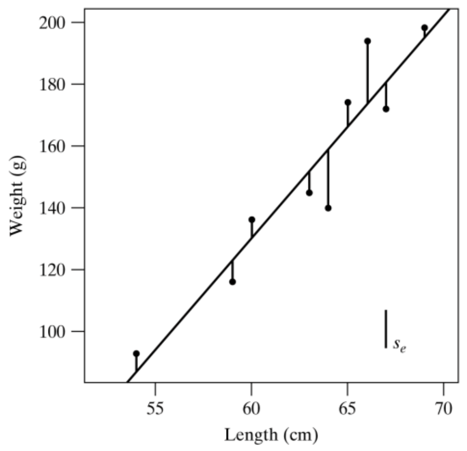
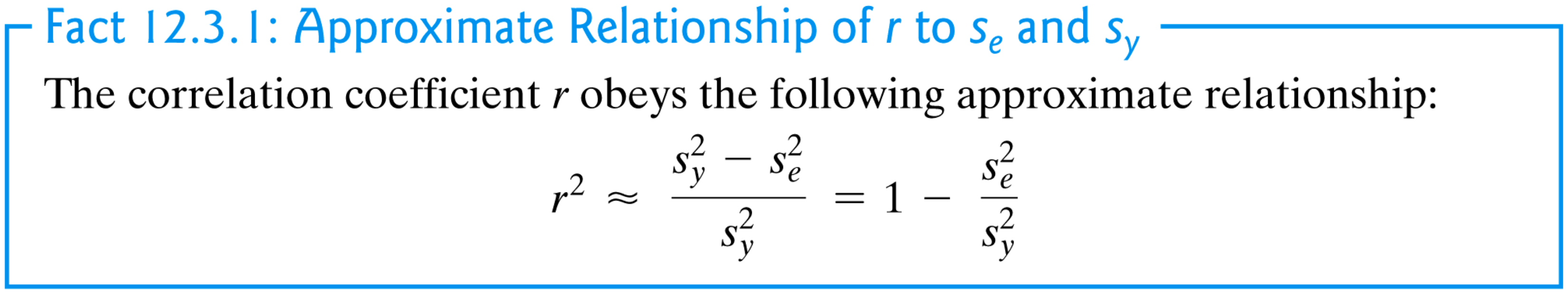


**Figure 12.3.7** Weight versus length of nine snakes showing the residuals and a line segment denoting the magnitude of the residual SD

# 12.3 The Fitted Regression Line

## The coefficient of determination, r$^2$



Fact 12.3.1: Approximate Relationship of $r$ to $s_e$ and $s_y$

The correlation coefficient $r$ obeys the following approximate relationship:

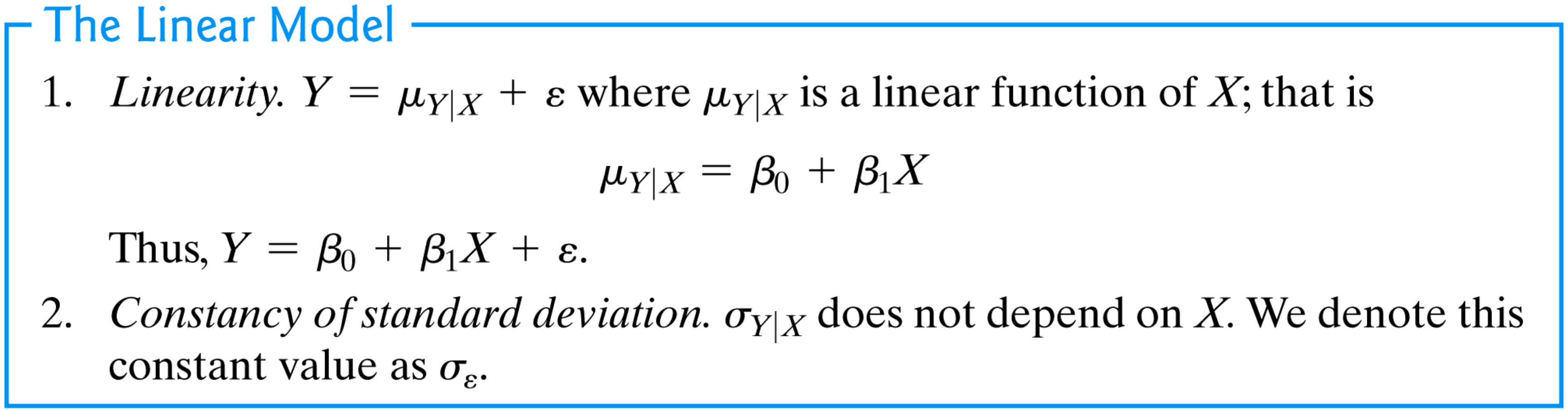
$$r^2 \approx \frac{s_y^2 - s_e^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

- r describes the tightness of the linear relationship between X and Y
- The **coefficient of determination, r$^2$,** describes the proportion of the variance in Y that is explained by the linear relationship between Y and X. This interpretation follows from the above fact (proved in Appendix 12.2).

# 12.4 The Linear Model

## The Linear Model

- These conditions, which constitute the linear model, are given in the following box.



The Linear Model

1. *Linearity.* $Y = \mu_{Y|X} + \varepsilon$ where $\mu_{Y|X}$ is a linear function of $X$; that is

$$\mu_{Y|X} = \beta_0 + \beta_1 X$$

   Thus, $Y = \beta_0 + \beta_1 X + \varepsilon$.

2. *Constancy of standard deviation.* $\sigma_{Y|X}$ does not depend on $X$. We denote this constant value as $\sigma_\varepsilon$.

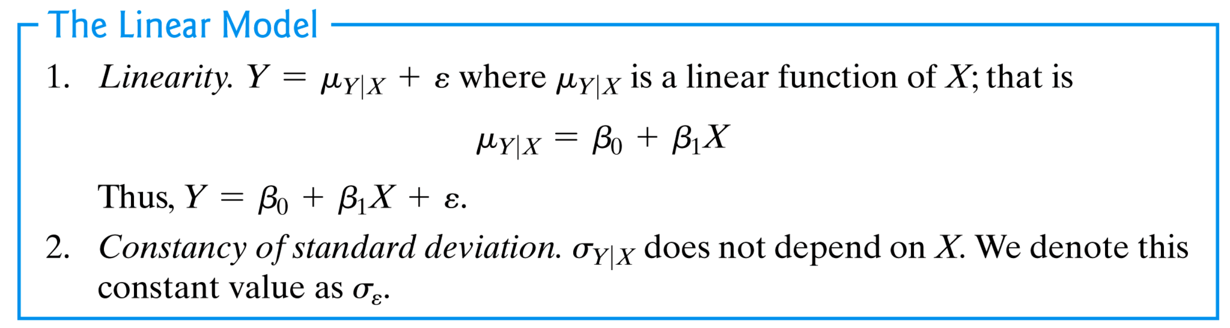
   - $\mu_{Y|X}$ = Population mean Y value for a given X
   - $\sigma_{Y|X}$ = Population SD of Y values for a given X
   - ε term represents random error.

# 12.4 The Linear Model

## The Linear Model

### Example 12.4.4 Height and Weight of Young Men

- We consider an idealized fictitious population of young men whose joint height and weight distribution fits the linear model exactly.
- $\mu_{Y|X} = -145 + 4.25X$, $\sigma_\varepsilon = 20$
- What is the linear model Y?



The Linear Model

1. *Linearity.* $Y = \mu_{Y|X} + \varepsilon$ where $\mu_{Y|X}$ is a linear function of $X$; that is

$$\mu_{Y|X} = \beta_0 + \beta_1 X$$

Thus, $Y = \beta_0 + \beta_1 X + \varepsilon$.

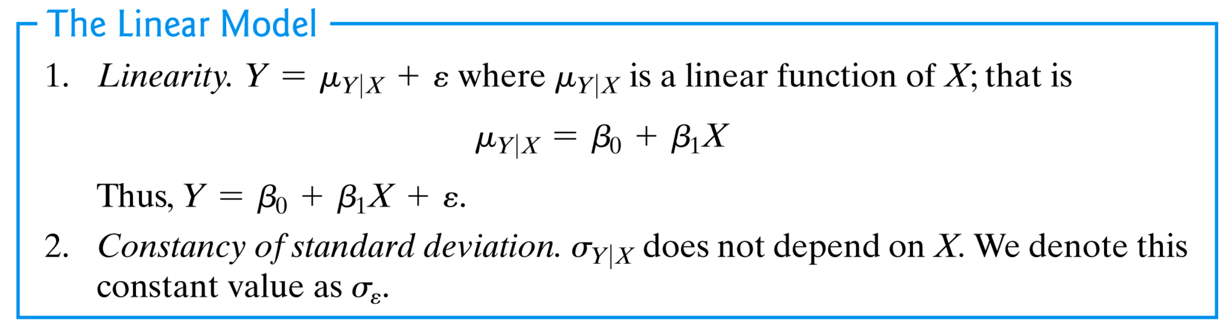
2. *Constancy of standard deviation.* $\sigma_{Y|X}$ does not depend on $X$. We denote this constant value as $\sigma_\varepsilon$.

# 12.4 The Linear Model

## The Linear Model

### Example 12.4.4 Height and Weight of Young Men

- We consider an idealized fictitious population of young men whose joint height and weight distribution fits the linear model exactly.
- $\mu_{Y|X}$ = -145 + 4.25X, $\sigma_\varepsilon$ = 20
- What is the linear model Y?



The Linear Model

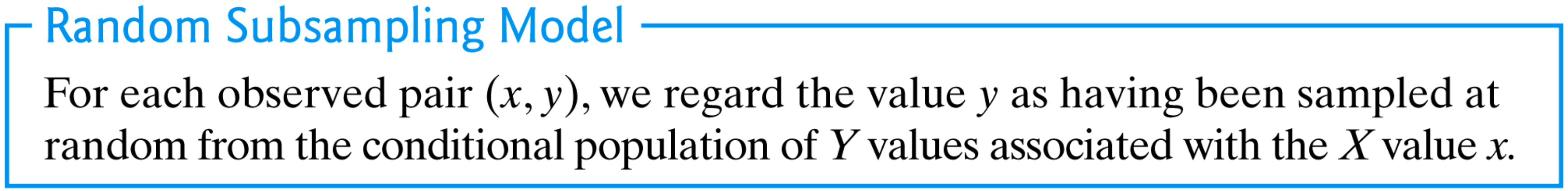
1. *Linearity.* $Y = \mu_{Y|X} + \varepsilon$ where $\mu_{Y|X}$ is a linear function of $X$; that is
$$\mu_{Y|X} = \beta_0 + \beta_1 X$$
Thus, $Y = \beta_0 + \beta_1 X + \varepsilon$.
2. *Constancy of standard deviation.* $\sigma_{Y|X}$ does not depend on $X$. We denote this constant value as $\sigma_\varepsilon$.

  - $\beta_0$ = -145; $\beta_1$ = 4.25
  - The model is Y = -145 + 4.25X + ε.

# 12.4 The Linear Model

## Estimation in the Linear Model

- Population distribution fits the linear model exactly. $Y = \beta_0 + \beta_1 X + \varepsilon$
- Suppose further that we are willing to adopt the following random subsampling model



### Random Subsampling Model

For each observed pair $(x, y)$, we regard the value $y$ as having been sampled at random from the conditional population of $Y$ values associated with the $X$ value $x$.

- Within the framework of the linear model and the random subsampling model, the quantities $b_0$ , $b_1$ , and $s_e$ calculated from a regression analysis can be interpreted as estimates of population parameters:
  - The y-intercept of least-squares regression line, $b_0$ , is an estimate of $\beta_0$
  - The slope of least-squares regression line, $b_1$ , is an estimate of $\beta_1$
  - The residual standard deviation, $s_e$ , is an estimate of $\sigma_\varepsilon$

# 12.4 The Linear Model

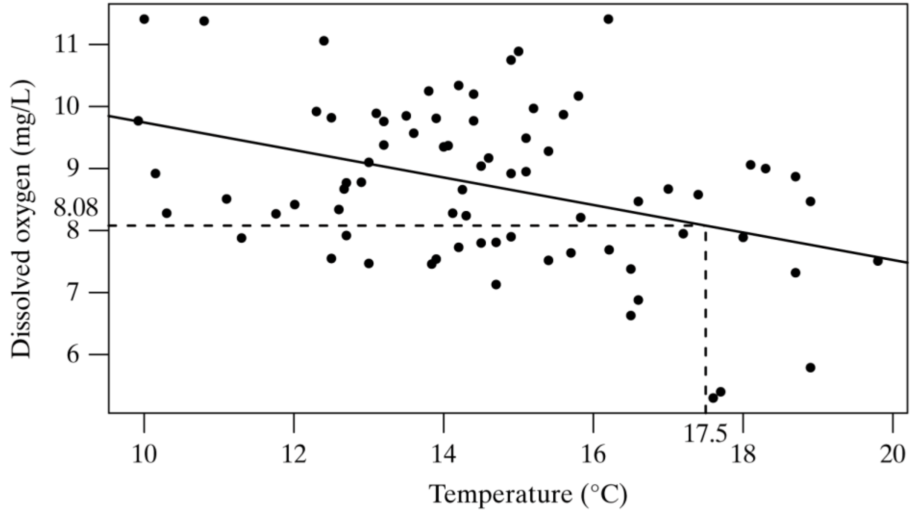
## Estimation in the Linear Model

### Example 12.4.6 Dissolved Oxygen

- regression equation to be $\hat{y} = 11.94 - 0.22x$ and $s_e = 1.21$.
- $x = 17.5$, $\hat{y}$?

- **Interpolation**: predict Y for values of X within the range of observed values of X.
- **Extrapolation**: predict Y for values of X that are outside the range of the data.
  - avoided whenever possible.

**Figure 12.4.2** Levels of dissolved oxygen versus water temperature for 75 days



# 12.4 The Linear Model

## Estimation in the Linear Model

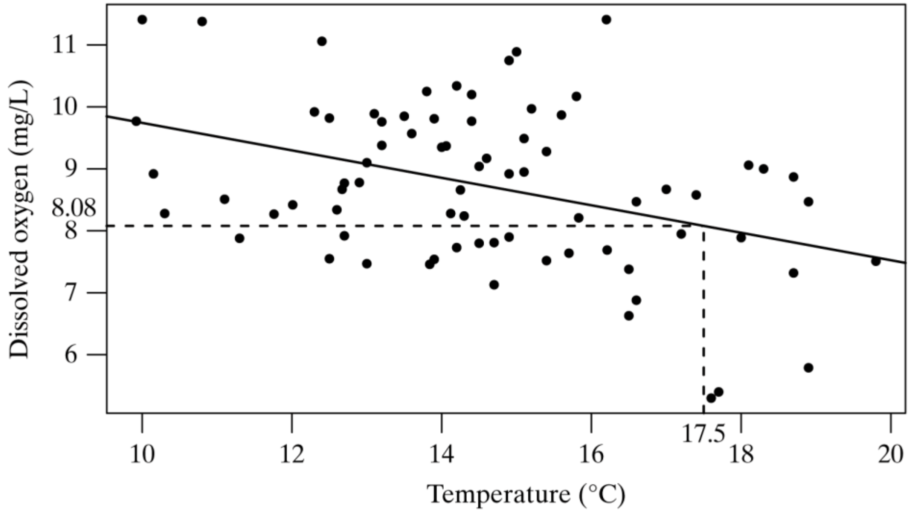
### Example 12.4.6 Dissolved Oxygen

- regression equation to be ŷ = 11.94 - 0.22x and $s_e$ = 1.21.
- x = 17.5, ŷ?
  - 11.94 - 0.22 x 17.5 = 8.1 with $s_e$ = 1.21.

- **Interpolation**: predict Y for values of X within the range of observed values of X.

- **Extrapolation**: predict Y for values of X that are outside the range of the data.
  - avoided whenever possible.

**Figure 12.4.2** Levels of dissolved oxygen versus water temperature for 75 days



# 12.5 Statistical Inference Concerning $\beta_1$

## The standard Error of $b_1$

- Within the framework of the linear model: $Y = \beta_0 + \beta_1 X + \varepsilon$ , and the random subsampling model,

Standard Error of $b_1$

$$\mathrm{SE}_{b_1} = \frac{s_e}{s_x\sqrt{n-1}}$$

  - The slope of least-squares regression line, $b_1$ , is an estimate of $\beta_1$

## Confidence interval for $\beta_1$

- 95% confidence interval: b1 ± $t_{0.025}$ $\mathrm{SE}_{b1}$
  where the critical value $t_{0.025}$ is determined from Student’s t distribution with **df = n - 2**

# 12.5 Statistical Inference Concerning $\beta_1$

## Confidence interval for $\beta_1$

### Example 12.5.2 Length and Weight of snakes

- 9 observations, $b_1 = 7.19186$, $SE_{b1} = 0.9531$.
- Construct 95% confident interval for $\beta_1$.

# 12.5 Statistical Inference Concerning β$_1$

## Confidence interval for β$_1$

### Example 12.5.2 Length and Weight of snakes

- 9 observations, $b_1 = 7.19186$, $SE_{b1} = 0.9531$.
- Construct 95% confident interval for $\beta_1$.
  - df = 9 - 2 = 7, $t_{7,\,0.025} = 2.365$
  - The 95% confidence interval is
    $7.19186 \pm 2.365 \times 0.9531 \rightarrow 4.94 \text{ gm/cm} < \beta_1 < 9.45 \text{ gm/cm}$
  - We are 95% confident that the true slope of the regression of weight on length for this snake population is between 4.94 gm/cm and 9.45 gm/cm.

# 12.5 Statistical Inference Concerning $\beta_1$

## Testing the hypothesis $H_0 : \beta_1 = 0$

- null hypothesis $H_0$ : $\mu_{Y|X}$ does not depend on X
- A t test of $H_0$ is based on the test statistic $\mathbf{t_s = (b_1 - 0) / SE_{b1}}$
- Critical values are obtained from Student's t distribution with **df = n – 2**

### Example 12.5.3 Blood Pressure and Platelet calcium

- The blood pressure and platelet calcium data are shown in Figure 12.5.2.
- $b_0 = -2.2009$ and $b_1 = 1.16475$; $SE_{b1} = 0.2704$; $n = 38$
- Is the calcium level linearly related to blood pressure?

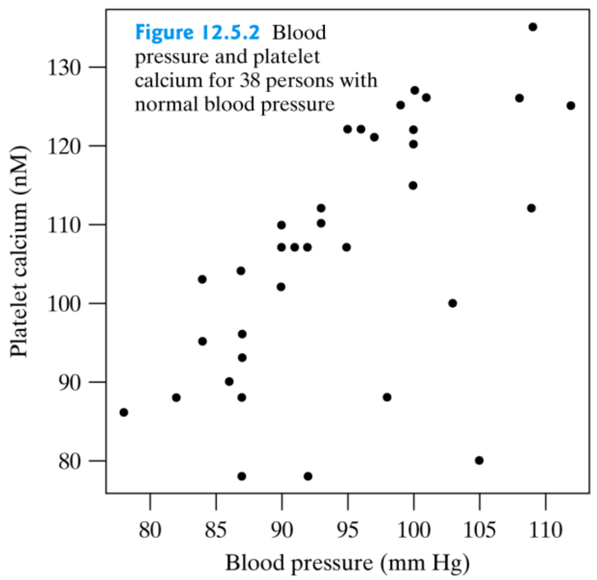


**Figure 12.5.2** Blood pressure and platelet calcium for 38 persons with normal blood pressure

# 12.5 Statistical Inference Concerning $\beta_1$

## Testing the hypothesis $H_0 : \beta_1 = 0$

### Example 12.5.3 Blood Pressure and Platelet calcium

- $b_0 = -2.2009$ and $b_1 = 1.16475$; $SE_{b1} = 0.2704$; $n = 38$
- Is the calcium level linearly related to blood pressure?
  - $H_0$: β1 = 0. Mean platelet calcium is not linearly related to blood pressure
  - $H_A$: β1 ≠ 0. Mean platelet calcium is linearly related to blood pressure
  - Let us choose α= 0.05. The test statistic is

    $$t_s = 1.16475/\ 0.2704 = 4.308$$
  - From Table 4 with df = n - 2 = 36 ≈ 40, we find $t_{40,0.0005} = 3.551$.
  - Thus, we find P-value < 0.001 and we reject $H_0$
  - The data provide sufficient (and very strong) evidence to conclude that the true slope of the regression of platelet calcium on blood pressure in this population is positive (i.e., β1 > 0).

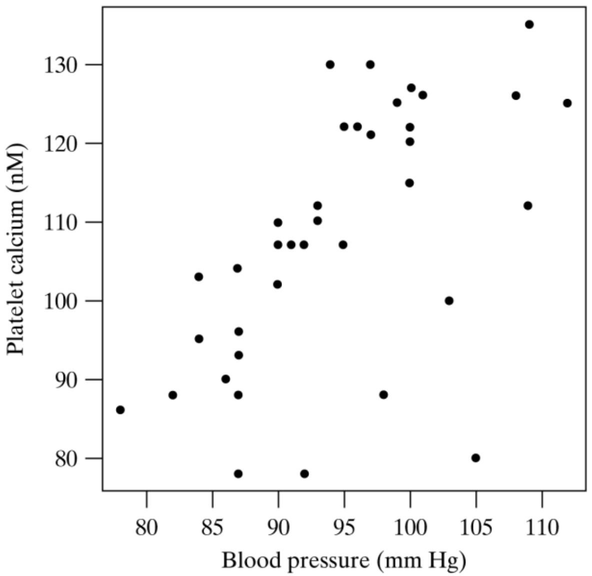


**Figure 12.5.2** Blood pressure and platelet calcium for 38 persons with normal blood pressure

# 12.9 Summary of Formulas

For convenient reference, we summarize the formulas presented in Chapter 12.

### Correlation Coefficient

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Fact 12.3.1: $$r^2 \approx \frac{s_y^2 - s_e^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

# 12.9 Summary of Formulas

## Fitted Regression Line

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

where

$$b_1 = r \times \left( \frac{s_y}{s_x} \right)$$

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

Residuals:

$$y_i - \hat{y}_i \quad \text{where} \quad \hat{y}_i = b_0 + b_1 x_i$$

Residual Sum of Squares:

$$\text{SS(resid)} = \sum (y_i - \hat{y}_i)^2$$

Residual Standard Deviation:

$$s_e = \sqrt{\frac{\text{SS(resid)}}{n - 2}}$$

# 12.9 Summary of Formulas

## Inference

Standard Error of $b_1$:

$$\mathrm{SE}_{b_1} = \frac{s_e}{s_x\sqrt{n-1}}$$

95% confidence interval for $\beta_1$:

$$b_1 \pm t_{0.025}\mathrm{SE}_{b_1}$$

Test of $H_0$: $\beta_1 = 0$ or $H_0$: $\rho = 0$:

$$t_s = \frac{b_1}{\mathrm{SE}_{b_1}} = r\sqrt{\frac{n-2}{1-r^2}}$$

Critical values for the test and confidence interval are determined from Student's $t$ distribution with df $= n - 2$.

# Summary

**Chapter 12. Linear Regression and Correlation**

- 12.2 The Correlation Coefficient
- 12.3 The Fitted Regression Line
- 12.4 Parametric Interpretation of Regression: The Linear Model
- 12.5 Statistical Inference Concerning β1
- 12.9 Summary of Formulas

# Homework

## Chapter 12

- 12.2.2;
- 12.3.1;
- 12.4.5;
- 12.5.1.