Simple Linear Regression

- A *statistical relation*, unlike a functional relation, is not a perfect one. If X is the *independent variable* and Y the *dependent variable*, a statistical relation is of the form:

$$Y = f(X) + \epsilon.$$

  In such cases, we call X an *explanatory variable* and Y a *response variable.*

- In a *simple linear regression* model, the response variable Y is linearly related to one *explanatory* variable X. That is,

$$Y_i = (\beta_0 + \beta_1 x_i) + \epsilon_i. \qquad i = 1, 2, \ldots, n.$$

  Assumptions:

  1. The mean of ϵ_i is 0 and the variance of ϵ_i is σ^2.

  2. The random errors ϵ_i are uncorrelated.

  3. β_0 and β_1 are parameters.

  4. x_i is a known constant.

- For example, let x denote the distance of a marathon and y the time that it will take a certain runner to finish it. Consider the following 22 practice finish times of our runner.

	1	2	3	4	5	6	7	8	9	10	11
Distance (x)	2	2	3	3	2	2.5	2.5	3	3.5	3.5	4
Time (y)	25	22	35	36	23	30	31	35	41	40	49

	12	13	14	15	16	17	18	19	20	21	22
Distance (x)	4	4	4	4.5	4.5	5	5	5	3.5	3.5	4
Time (y)	47	48	48	56	53	62	60	61	42	41	47

  Determine the regression line and use it to find his expected finish time for a 6-mile marathon.

- **Equation of the Least-Squares Regression Line .** Suppose we have data on an explanatory variable x and a response variable y for n individuals. The means and standard deviations of the sample data are $\bar{x}$ and s_x for x and $\bar{y}$ and s_y for y, and the correlation between x and y is r. The equation of the least-squares regression line of y on x is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

with *slope*

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{(\Sigma xy) - \frac{1}{n}(\Sigma x)(\Sigma y)}{(\Sigma x^2) - \frac{1}{n}(\Sigma x)^2} = r\frac{s_y}{s_x} \tag{1}$$

and *intercept*

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2}$$

Proof:

- The **fitted** (or **predicted**) **values** $\hat{y}_i$'s are obtained by successively substituting the x_i's into the estimated regression line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$. The **residuals** are the vertical deviations, $e_i = y_i - \hat{y}_i$, from the estimated line.

- The **error sum of squares**, (equivalently, **residual sum of squares**) denoted by SSE, is

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \tag{3}$$
$$= SS_{yy} - \hat{\beta}_1 SS_{xy} = \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i \tag{4}$$

and the estimate of σ^2 is

$$\hat{\sigma}^2 = s^2 = \frac{SSE}{n-2} = \frac{(n-1)s_y^2(1-r^2)}{n-2}. \tag{5}$$

- R commands

```
data.jog=read.csv("Jogging.csv",header=T)
attach(data.jog)
plot(Distance,Time)

results=lm(Time~Distance)
coef(results)
abline(results)             # This will plot the regression line in the scatterplot
predict(results,newdata=data.frame(Distance=6))
attributes(results)
results$fitted              # This will give all the predicted values
results$residuals           # This will give all the residuals
```

- **Practice.**

1. **Production Run.** Consider the following data of 10 production runs of a certain manufacturing company.

Production run	1	2	3	4	5	6	7	8	9	10
Lot size (x)	30	20	60	80	40	50	60	30	70	60
Man-Hours (y)	73	50	128	170	87	108	135	69	148	132

a. Determine the correlation coefficient r.

b. What can you say about the linear relationship of x and y? Is it a strong linear relationship.

c. Determine the regression line.

d. Predict the number of man-hours ($\hat{y}$) required to produce a lot size 100.

2. The table below displays data on age (in years) and price (in \$100)for a sample of 11 cars.

Age (x)	5	4	6	6	5	5	6	6	2	7	7
Price (y)	85	102	70	80	89	98	66	90	169	68	50

a. Determine the regression line.

b. Estimate the expected value of a car that is 3 years old.

c. Determine the residual for the first car. Is this value unusual?

3. Tree Circumference and Height. Listed below are the circumferences (in feet) and the heights (in feet) of trees in Marshall, Minnesota (based on data from "Tree Measurements" by Stanley Rice, *American Biology Teacher*.

x (circ)	1.8	1.9	1.8	2.4	5.1	3.1	5.5
y (height)	21.0	33.5	24.6	40.7	73.2	24.9	40.4
x (circ)	5.1	8.3	13.7	5.3	4.9	3.7	3.8
y (height)	45.3	53.5	93.8	64.0	62.7	47.2	44.3

a. Determine the regression line.

b. Estimate the expected height of a tree that has a circumference of 10 feet.

c. Determine the residual for the first tree. Is this value unusual?

- **Recommended Exercises:** Answer the following questions:
  On pages 35-37, #19, 20, 21, 22, 28.