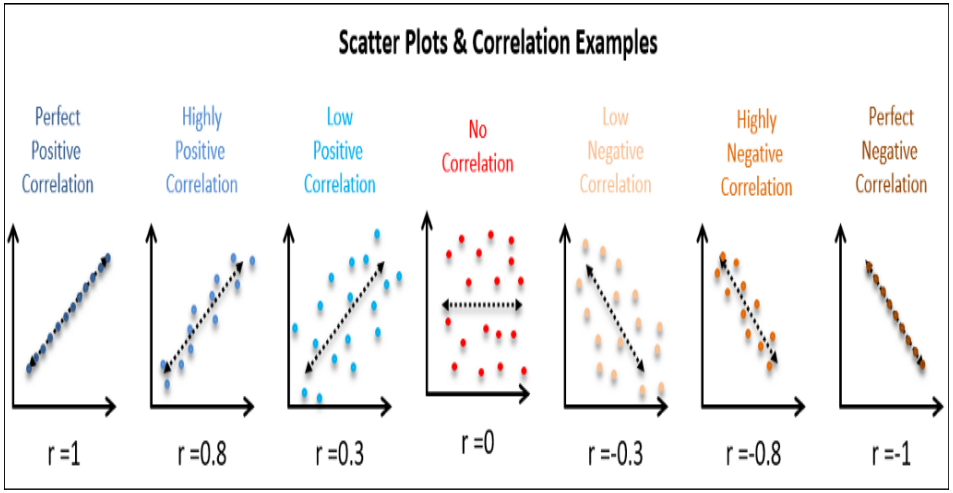
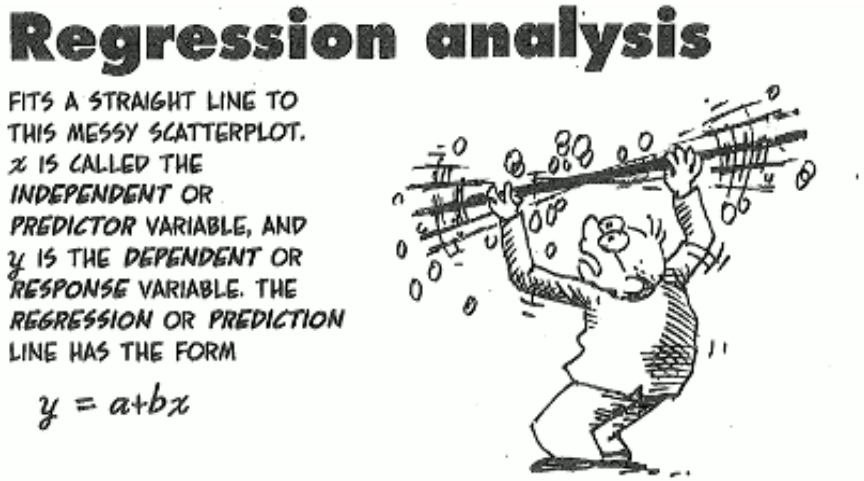
Introductory Statistics: A Problem-Solving Approach

by Stephen Kokoska

Chapter 12

Correlation and Linear Regression





Relationship Models

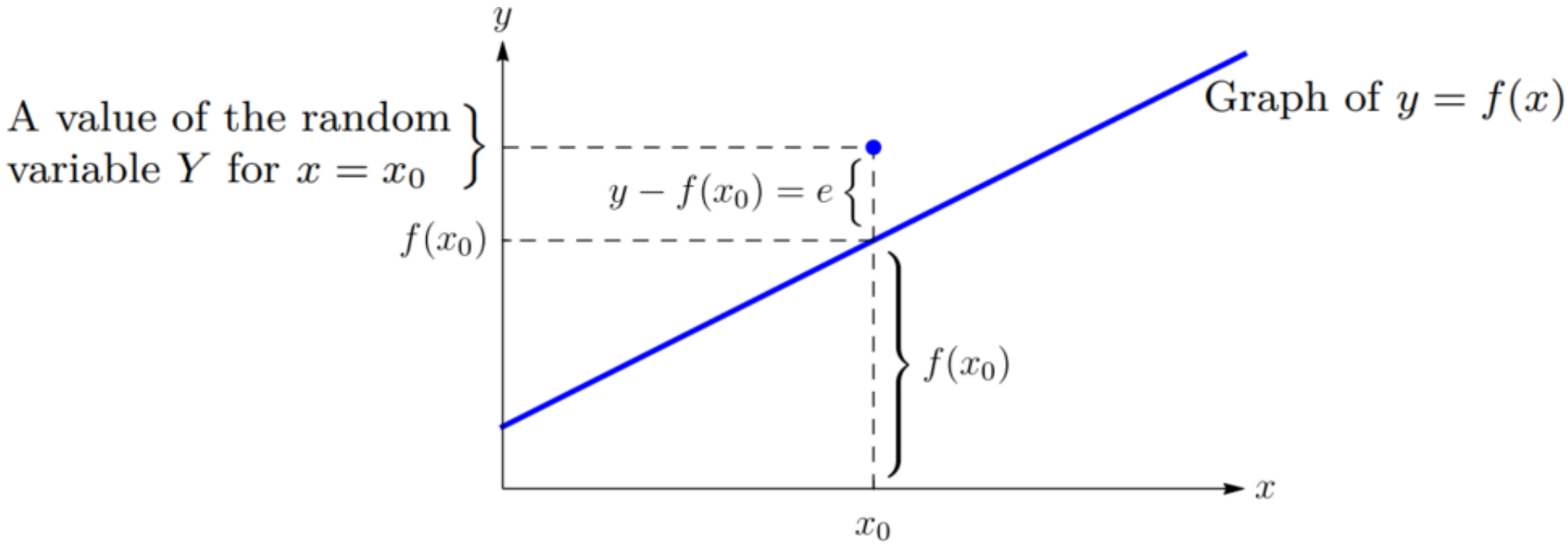
- Two variables may be related in a variety of ways.
- The techniques presented in this chapter can be used to determine whether there is a significant linear relationship between two quantitative variables.
- A **deterministic** relationship between x and y is one in which the value of y is *completely* determined by the value of x.
- In general, y = f(x) is a deterministic relationship between x and y. The independent variable is x and the dependent variable is y.
  - One of the simplest deterministic relationships between *x* and *y* is a linear relationship:

    $$y = b_0 + b_1 x,$$

    where b_0 and b_1 are constants.
  - The set of ordered pairs (*x*, *y*) such that $y = b_0 + b_1 x$ forms a straight line with slope b_1 and *y*-intercept b_0.

Relationship Models

- An extension of the deterministic model is the **probabilistic** model. For a fixed value x, the value of the second variable is randomly distributed.
  - The independent variable, which is fixed by the experimenter, is usually denoted by x. In a probabilistic model, the random variable is the dependent variable and usually denoted by Y.
- In an additive probabilistic model, there is a deterministic part and a random part. The model can be written as:
  Y = (deterministic function of x) + (random deviation) = $f(x) + E$
  where E is a random variable, called the random error.



Simple Linear Regression Model

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be n pairs of observations such that y_i is an observed value of the random variable Y_i. We assume that there exist constants β_0 and β_1 such that

$$Y_i = \beta_0 + \beta_1 x_i + E_i$$

where $E_1, E_2, \ldots, E_n$ are independent, normal random variables (**error terms**) with mean 0 and variance σ^2. that is,

1) The E_i's are normally distributed, which implies that the Y_i's are normally distributed.
2) The expected value of E_i is 0, which implies that $E(Y_i) = \beta_0 + \beta_1 x_i$.
3) $Var(E_i) = \sigma^2$, which implies that $Var(Y_i) = \sigma^2$.
4) The E_i's are independent, which implies that the Y_i's are independent.

All assumptions can be stated compactly: $E_i \overset{\text{ind}}{\sim} N(0, \sigma^2)$

Example: Get the Lead Out

The Environmental Protection Agency suggests that children are very susceptible to the effects of lead exposure. Children who are exposed to high levels of lead may experience developmental, behavioral, and learning problems. For six-year-old children, suppose IQ level (y) is related to blood lead level (x, measured in μg/dL) and that the true regression line is $y = 96.8 - 0.45x$

a) Find the expected IQ for a six-year-old when the blood lead level is 30 µg/dL.

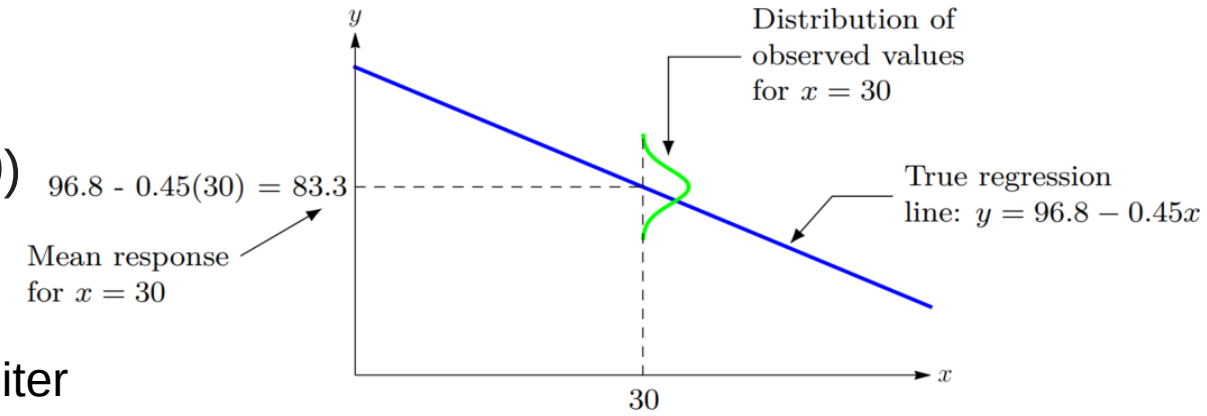
b) How much change in IQ is expected if the blood lead level increases by 10 µg/dL? What if it decreases by 20 µg/dL?

c) Suppose σ = 8 µg/dL. Find the probability that an observed IQ is greater than 100 when the blood lead level is 20 µg/dL.

$E(Y \mid x) = b_0 + b_1 x$

$E(Y \mid 30) = 96.8 - 0.45(30)$

$= \mathbf{83.3}$



µg/dL =micrograms per deciliter

Example: Get the Lead Out

b) How much change in IQ, *y*, is expected if the blood lead level, *x*, increases by 10 µg/dL? What if it decreases by 20 µg/dL?

The slope of the true regression line, $b_1 = -0.45$, is the change in IQ associated with a 1 μg/dL change in blood lead level.

- If the blood lead level increases by 10 μg/dL, the expected change in IQ is $(b_1) \cdot (\text{change in } x) = (-0.45)(10) = \mathbf{-4.5}$
- If the blood lead level decreases by 20 µg/dL, the expected change in IQ is $(b_1) \cdot (\text{change in x}) = (-0.45)(-20) = \mathbf{9.0}$

c) Suppose σ = 8 µg/dL. Find the probability that an observed IQ is greater than 100 when the blood lead level is 20 µg/dL.

$Y \sim N(\mu=87.8, \sigma^2=64)$

$\mu = E(Y \mid 30) = 96.8 - 0.45(20) = \mathbf{87.8}$

$$P(Y > 100) = P\left(\frac{Y - 87.8}{8} > \frac{100 - 87.8}{8}\right)$$

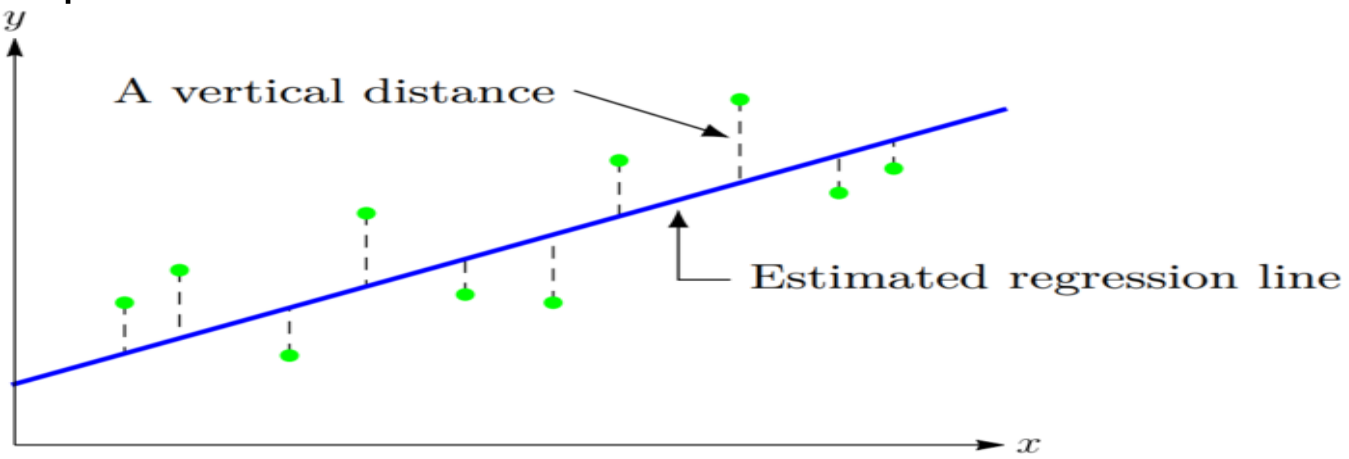
$$= P(Z > 1.53)$$

$$= 1 - P(Z \le 1.53) = 1 - 0.9370 = 0.0630$$

Least Squares Estimates

- Suppose two variables are related via a simple linear regression model. The parameters β_0 and β_1 are usually unknown. However, if we assume that the observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are independent, then these sample data can be used to estimate the model parameters β_0 and β_1.
- The **line of best fit**, or **estimated regression line**, is obtained using the **principle of least squares**: Minimize the sum of the squared deviations, or vertical distances from the observed points to the line. Consider the vertical distances from the points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ to the line. The principle of least squares produces an estimated regression line such that the sum of all squared vertical distances is a minimum.

$\sum e_i^2$ is minimized



Least Squares Estimates

The **least squares estimates** of the *y* intercept (constant term), β_0 and the slope (regression coefficient), β_1 of the true regression line are:

$$\hat{\beta}_1 = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{n\sum x_i^2 - \left(\sum x_i\right)^2} \qquad \hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x}$$

The estimated regression line is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Note:

Before using these equations, always consider constructing a scatter plot and compute the sample correlation coefficient (presented in the next slides) to make sure a linear model is reasonable.

Example: Estimating Coefficients

Climate change and snow depth may affect permafrost and soil composition, especially in Canada. **A** study was conducted to examine the impacts of snow cover on soil temperature across northern latitudes. A random sample of locations was obtained, and the snow depth (in centimeters) and the difference between the soil temperature and the air temperature (in °C; called the surface offset) was measured for each.

a. Find the estimated regression line.

b. Estimate the true mean surface offset for a snow depth of 45 cm.

Depth	Offset
6.7	7.0
41.1	22.8
20.9	8.5
58.9	20.9
18.2	15.1
44.6	14.7
53.8	21.8
23.1	11.5
47.2	17.8
48.2	21.3

Snow depth (x) = independent variable
Surface offset (y) = dependent variable

Summary statistics:

$$\sum x_i = 362.7 \quad \sum y_i = 161.4 \quad \sum x_i y_i = 6628.39$$

$$\sum x_i^2 = 15{,}939.65 \quad \sum y_i^2 = 2900.02$$

Independent = Predictor = Explanatory
Dependent = Response

Example: Estimating Coefficients

$$\hat{\beta}_1 = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

$$= \frac{(10)(6628.39) - (362.7)(161.4)}{10(15{,}939.65) - (362.7)^2}$$

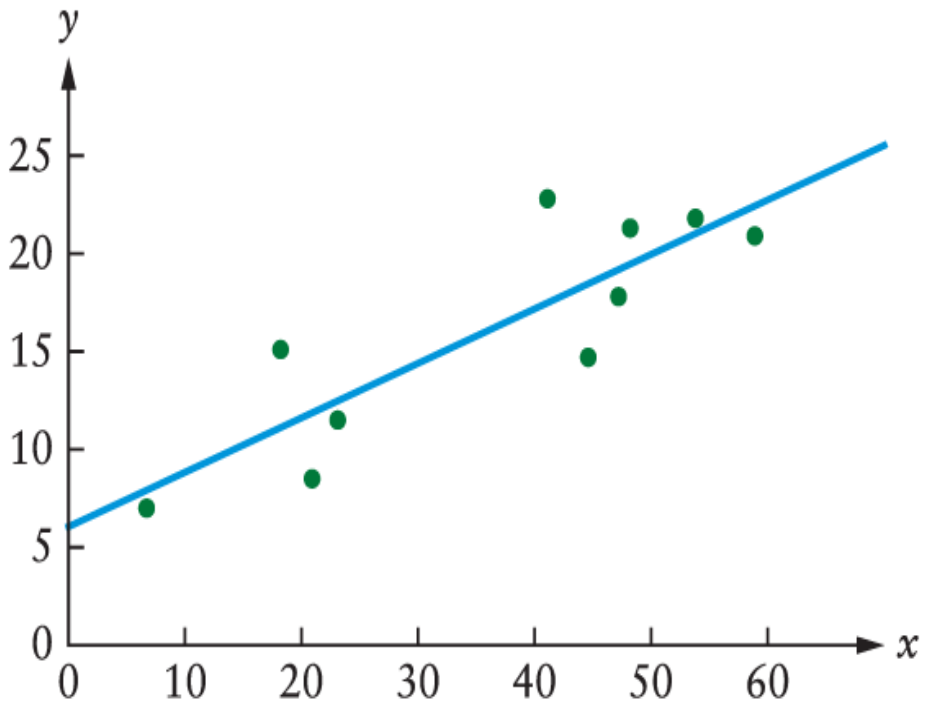
$$= \frac{7744.12}{27{,}845.21} = \mathbf{0.2781}$$

$$\hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n}$$

$$= \frac{161.4 - 0.2781(362.7)}{10}$$

$$= 6.0533$$

$$\hat{y}_i = 6.0533 + 0.2781 x_i$$



Interpretation

$\hat{\beta}_1$ = 0.2781 tells us that the surface offset increases by 0.2781 °C, on average, for each additional one cm in the snow depth.

Example: Estimating Coefficients

b. Estimate the true mean surface offset for a snow depth of 45 cm.

$$\hat{y} = 6.0533 + 0.2781(45)$$

$$= 18.57\ ^{\circ}\text{C}$$

$\hat{y} =$ (Predicted = estimated = fitted value)

Although regression equations can be used for prediction, a few cautions should be considered whenever you are interpreting the predicted values.

1. The predicted value is not perfect (unless r = +1.00 or –1.00).
2. The regression equation should not be used to make predictions for X values that fall outside the range of values covered by the original data.

ANOVA Table and Coefficient of Determination

One method to assess the accuracy of a simple linear regression model involves an analysis of variance table.

The *i*th residual/error = $y_i - \hat{y}_i$. This difference is a measure of how far away the observed value of Y is from its estimated value.

Source of variation	Sum of squares	Degrees of freedom	Mean square	F
Regression	SSR	1	$\text{MSR} = \frac{\text{SSR}}{1}$	$\frac{\text{MSR}}{\text{MSE}}$
Error	SSE	$n-2$	$\text{MSE} = \frac{\text{SSE}}{n-2}$	
Total	SST	$n-1$		

$$\underbrace{\sum(y_i - \bar{y})^2}_{\text{SST}} = \underbrace{\sum(\hat{y}_i - \bar{y})^2}_{\text{SSR}} + \underbrace{\sum(y_i - \hat{y}_i)^2}_{\text{SSE}}$$

$$\text{SST} = S_{yy} \qquad \text{SSR} = \hat{\beta}_1 S_{xy} \qquad \text{SSE} = \text{SST} - \text{SSR}$$

ANOVA Table and Coefficient of Determination

$$S_{yy} = \underbrace{\sum (y_i - \bar{y})^2}_{\text{definition}} = \underbrace{\sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2}_{\text{computational formula}}$$

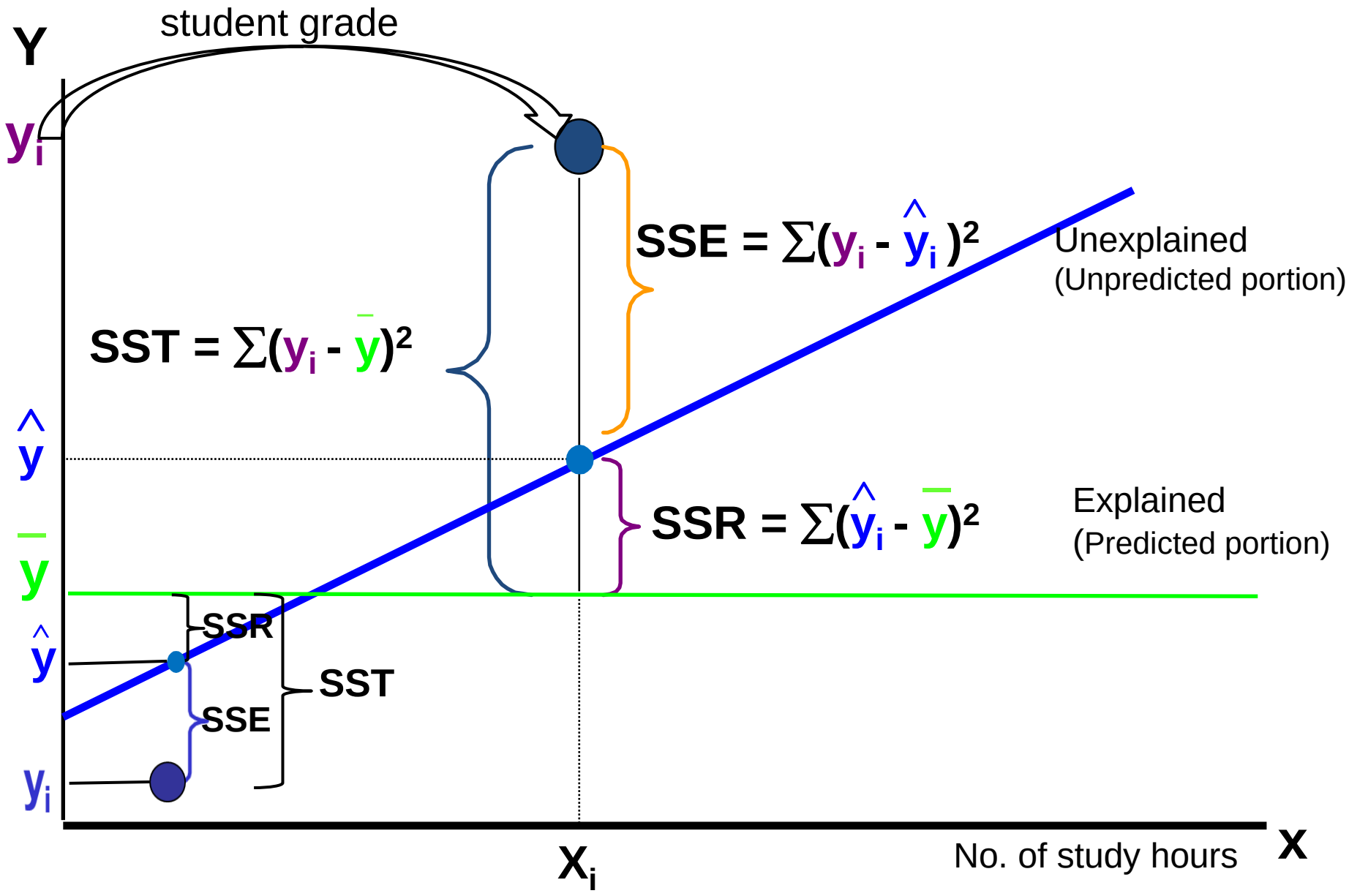
$$S_{xy} = \underbrace{\sum (x_i - \bar{x})(y_i - \bar{y})}_{\text{definition}} = \underbrace{\sum x_i y_i - \frac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right)}_{\text{computational formula}}$$

The **coefficient of determination**, denoted r^2, is a measure of the proportion of the variation in the data that is explained by the regression model, and is defined by:

$$r^2 = \text{SSR} / \text{SST}$$

- Because $0 \le \text{SSR} \le \text{SST}$, the coefficient of determination r^2 is always a number between 0 and 1 (inclusive).
- The higher r^2 is, the better the model is.

Explained and Unexplained Variation



Hypothesis Test for a Significant Linear Regression

- The null hypothesis states that the variation in *Y* is completely random, independent of the value of *x*. In this case, a scatter plot would have no discernible linear pattern.
- To test the significance of simple linear regression model, we have two equivalent techniques. F-test ⟺ T-test

H_0: $\beta_1 = 0$ (There is no significant linear relationship)

H_a: $\beta_1 \neq 0$ (There is a significant linear relationship)

$$\text{TS: } F = \frac{\text{MSR}}{\text{MSE}}$$

$$\text{RR: } F > F_{\alpha,1,n-2}$$

The null hypothesis is rejected only for large values of the test statistic.

H_0: $\beta_1 = 0$

H_a: $\beta_1 > 0$, $\beta_1 < 0$, or $\beta_1 \neq 0$

$$S_{xx} = \sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2$$

$$\text{TS: } T = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} \qquad \text{Where, } S_{\hat{\beta}_1} = S/\sqrt{S_{xx}},\ S = \sqrt{MSE}$$

$$\text{RR: } T > t_{\alpha,n-2},\ T < -t_{\alpha,n-2},\ \text{or } |T| > t_{\alpha/2,n-2}$$

Confidence Intervals for β_1

A 100(1 − α)% confidence interval for β_1 has the following values as endpoints:

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot S_{\hat{\beta}_1}$$

EXAMPLE 12.5 Ready to Operate

The Pulsar Corporation sells a large sterilizer with four extendable shelves for medical tools. **Company** engineers believe that the time to reach the operating temperature from a cold start (y, measured in minutes) is linearly related to the thickness of insulation (x , in inches). A random sample of n = 12 thicknesses was selected, and the time to reach operating temperature was recorded for each. The summary statistics are given.

$$\sum x_i = 23.1 \qquad \sum y_i = 84.8 \qquad \sum x_i y_i = 158.5$$
$$\sum x_i^2 = 50.13 \qquad \sum y_i^2 = 607.66$$

a. Find the estimated regression line and interpret the coefficient of regression (slope).
b. Complete the ANOVA table and conduct an *F*-test for a significant regression. Use a significance level of 0.05.
c. Conduct a *t*-test concerning β_1 for a significant regression. Use a significance level of 0.05.
d. Find and interpret the coefficient of determination.

EXAMPLE 12.5 Ready to Operate

$$\hat{\beta}_1 = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2}$$

$$= \frac{(12)(158.5) - (23.1)(84.8)}{(12)(50.13) - (23.1)^2}$$

$$= \frac{-56.88}{67.95} = -0.8371$$

$$\hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n}$$

$$= \frac{84.8 - (-0.8371)(23.1)}{12}$$

$$= 8.6781$$

$$\hat{y}_i = 8.6781 - 0.8371x_i$$

The value $\hat{\beta}_1 = 0.8371$ suggests that an increase of 1 inch in the thickness of insulation leads to a decrease of approximately 0.84 min in the time to reach operating temperature .

EXAMPLE 12.5 Ready to Operate

H_0: $\beta_1 = 0$ (There is no significant linear relationship)

H_a: $\beta_1 \neq 0$ (There is a significant linear relationship)

Source of variation	Sum of squares	Degrees of freedom	Mean square	F	p Value
Regression	3.9678	1	3.9678	8.9387	0.0136
Error	4.4389	10	0.4439		
Total	8.4067	11			

TS: $F = 8.9387$

RR: $F > F_{\alpha,1,n-2} = F_{0.05,1,10} = 4.96$

1-pf(8.9387,1,10)
0.01357909

Because F lies in the rejection region, there is evidence to suggest that insulation thickness is linearly related to time to reach the operating temperature.

OR

Because of *p-value* = 0.0136 is < α = 0.05, we reject the H_0, concluding that there is a significant linear relationship between the time to reach operating temperature and the insulation thickness.

EXAMPLE 12.5 Ready to Operate

H_0: $\beta_1 = 0$

H_a: $\beta_1 \neq 0$

$$\text{TS: } T = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = \frac{-0.8371}{\sqrt{0.0784}} = -2.9898$$

$$S^2_{\hat{\beta}_1} = \frac{S^2}{S_{xx}} = \frac{0.4439}{5.6625} = 0.0784$$

$$S_{xx} = \sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2 = 50.13 - \frac{1}{12}(23.1)^2 = 5.6625$$

$$\text{RR: } |T| > t_{0.05/2,12-2} = t_{0.025,10} = 2.2281$$

S^2=*MSE*

|-2.9898| = 2.9898 > 2.2281

Because |-2.9898| lies in the rejection region, there is evidence to suggest that $\beta_1 \neq 0$, so the regression is significant.

Note: For simple linear regression, $t^2 = f$ in every case, subject to round-off error. Comparing the two tests for a significant regression, notice that $t^2 \approx f$.

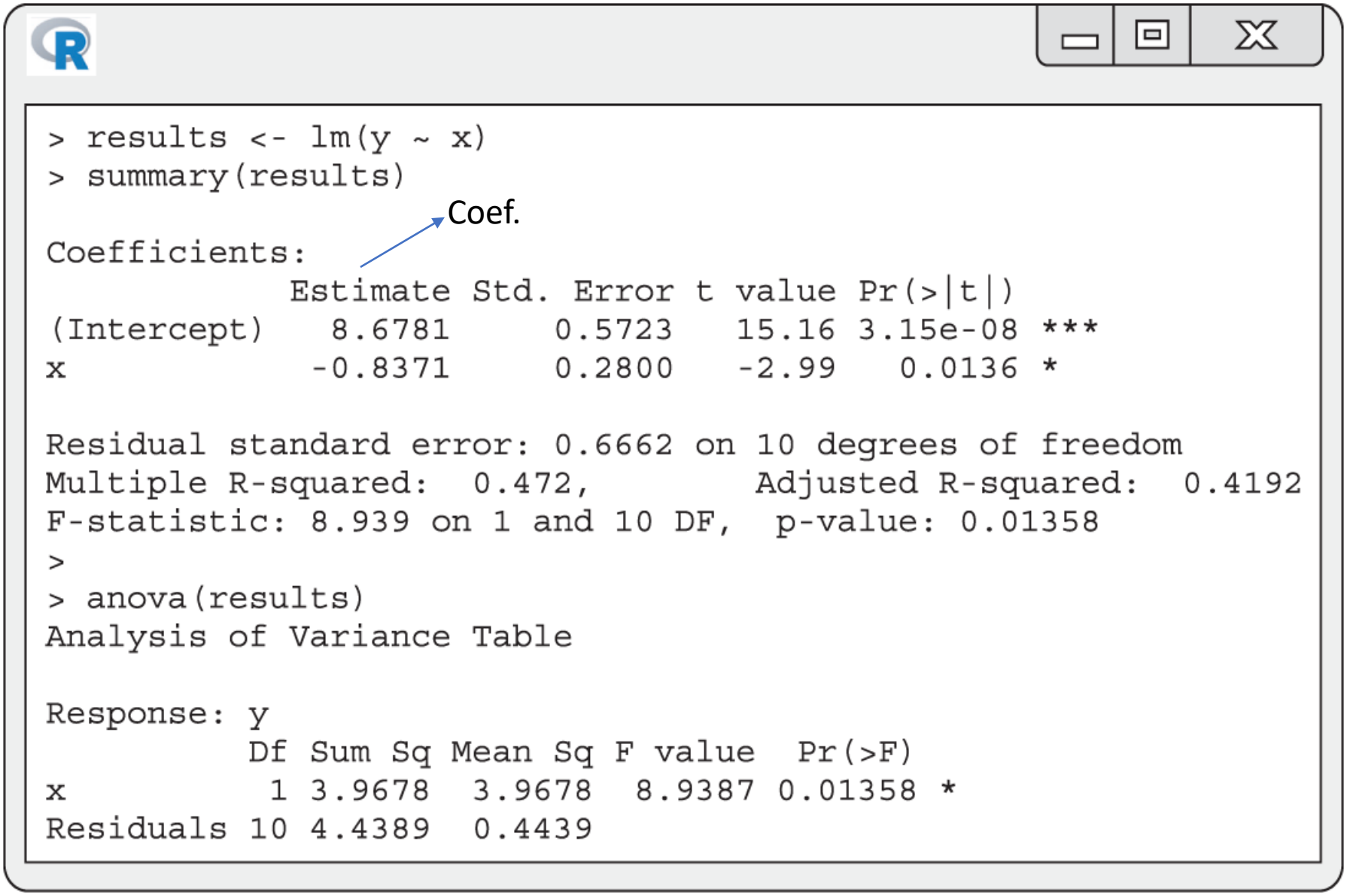
$(-2.9898)^2 \approx 8.9387$

r^2 = SSR / SST = 3.9678 / 8.4067

≈ 0.4720

Approximately 0.4720, or 47.20%, of the variation in the time to reach operating temperature is explained by the regression model (the thickness of insulation)

EXAMPLE 12.5 Ready to Operate




```
> results <- lm(y ~ x)
> summary(results)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.6781     0.5723   15.16 3.15e-08 ***
x             -0.8371     0.2800   -2.99   0.0136 *

Residual standard error: 0.6662 on 10 degrees of freedom
Multiple R-squared:  0.472,        Adjusted R-squared:  0.4192
F-statistic: 8.939 on 1 and 10 DF,  p-value: 0.01358
>
> anova(results)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
x          1 3.9678  3.9678  8.9387 0.01358 *
Residuals 10 4.4389  0.4439
```

Correlation

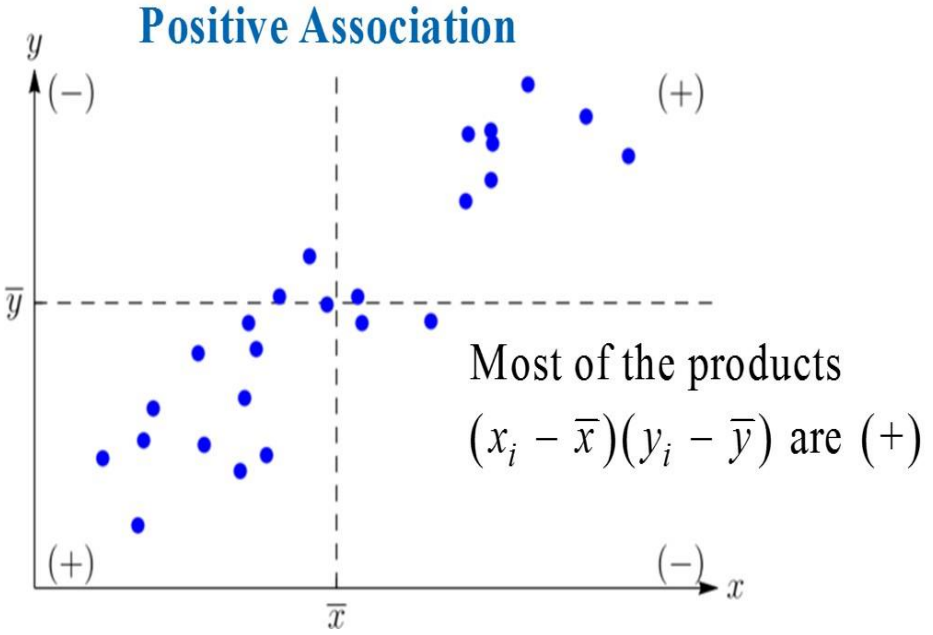
- Correlation is a statistical term indicating a relationship between two variables. For example, the temperature is correlated with the number of cars that will not start in the morning. As the temperature decreases, the number of cars that will not start in the morning increases.
- The **sample correlation coefficient,** denoted *r*, is a measure of the strength of a linear relationship between two quantitative variables *x* and *y*.
- If large values of *x* are associated with large values of *y,* or if as *x* increases, the corresponding value of *y* tends to increase, then *x* and *y* are **positively** related.
- If small values of *x* are associated with large values of *y,* or if as *x* increases, the corresponding value of *y* tends to decrease, then *x* and *y* are **negatively** related.
- **Definition:** Suppose there are *n* pairs of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. The **sample correlation coefficient** for these *n* pairs is

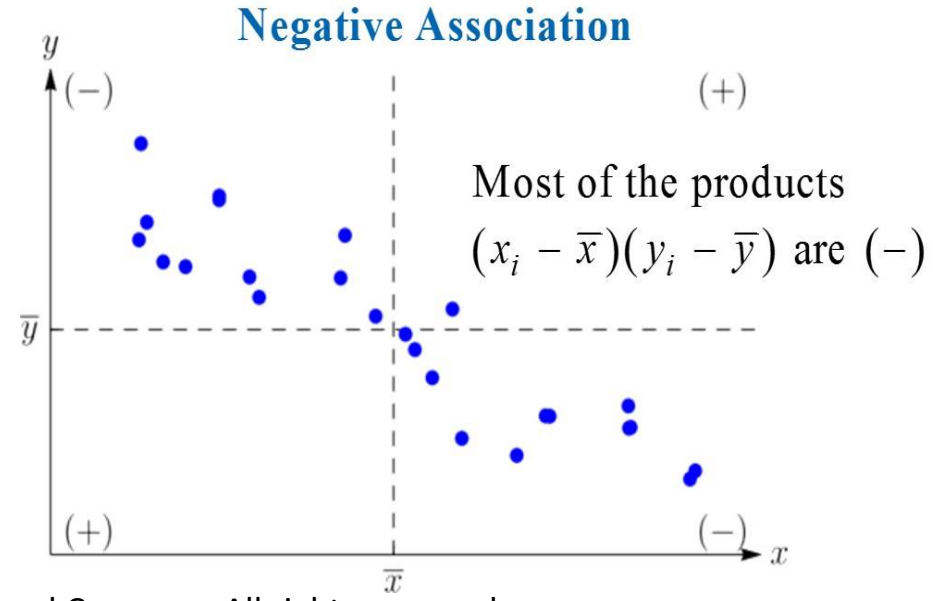
$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

$$= \frac{\sum x_i y_i - \frac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{\left[\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2\right]\left[\sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2\right]}}$$

Correlation

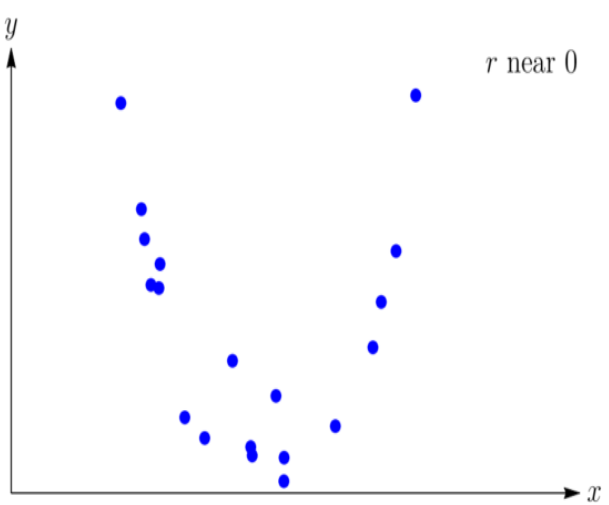
1. The value of *r* does not depend on the order of the variables and is independent of units.
2. $-1 \le r \le +1$. *r* is exactly +1 if and only if all of the ordered pairs lie on a straight line with positive slope. *r* is exactly −1 if and only if all of the ordered pairs lie on a straight line with negative slope.
3. r^2= coefficient of determination
4. If *r* is near 0, there is no evidence of a linear relationship, but x and y may be related in another way.
5. Suppose there is a horizontal line ($y = \beta_0$) with zero slope, and all the data points lie very close to this line. So, $r \approx 0$
6. Correlation between two variables does not imply causation.

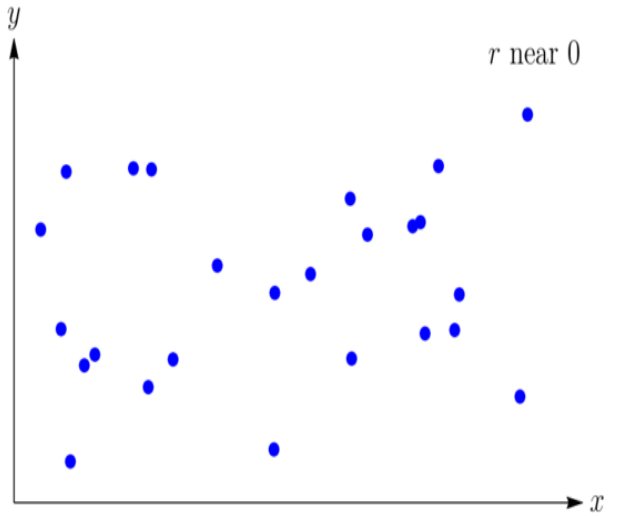


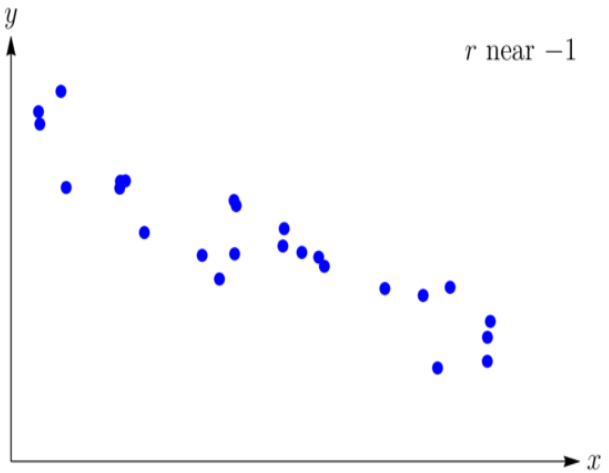


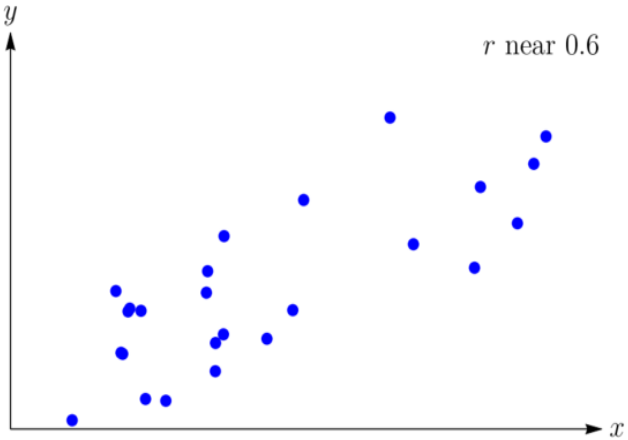
General Guidelines

1. If $0 \le |r| \le 0.5$, then there is a **weak linear relationship**.
2. If $0.5 \le |r| \le 0.8$, then there is a **moderate linear relationship**.
3. If $0.8 \le |r| \le 1$, then there is a **strong linear relationship**.





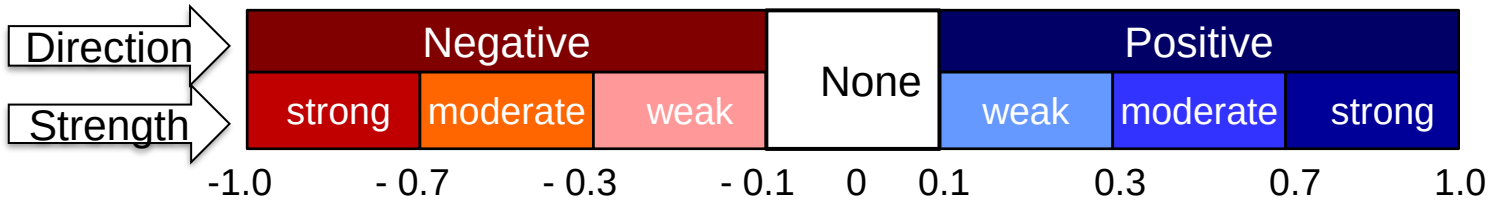




r = + 0.85

Direction: positive correlation

Strength: a strong relationship



EXAMPLE 12.6 Credit Scores and Income

Although income is not used in calculating a credit score, some evidence suggests that income is related to credit score. A random sample of adult consumers was obtained, and their credit score and yearly income level (in hundreds of thousands of dollars) were recorded. The data are given in the table. Find the sample correlation coefficient between credit score and yearly income, and interpret this value.

Income, x	1.3	1.1	0.8	1.2	1.4	0.9	0.9	1.4	1.2	1.0
Credit score, y	756	728	635	599	760	722	743	726	694	726

$$S_{xx} = \sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2 = 12.96 - \frac{1}{10}(11.2)^2 = 0.416$$

$$S_{yy} = \sum y_i^2 - \frac{1}{10}\left(\sum y_i\right)^2 = 5{,}050{,}267 - \frac{1}{10}(7089)^2 = 24{,}874.9$$

$$S_{xy} = \sum x_i y_i - \frac{1}{10}\left(\sum x_i\right)\left(\sum y_i\right) = 7968.1 - \frac{1}{10}(11.2)(7089) = 28.42$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\,S_{yy}}} = \frac{28.42}{\sqrt{(0.416)(24{,}874.9)}} = 0.2794$$

Because $r = 0.2794 < 0.5$, there is a weak positive linear relationship between income level and credit score (as a person's income level increases, so does the credit score).

Regression Diagnostics

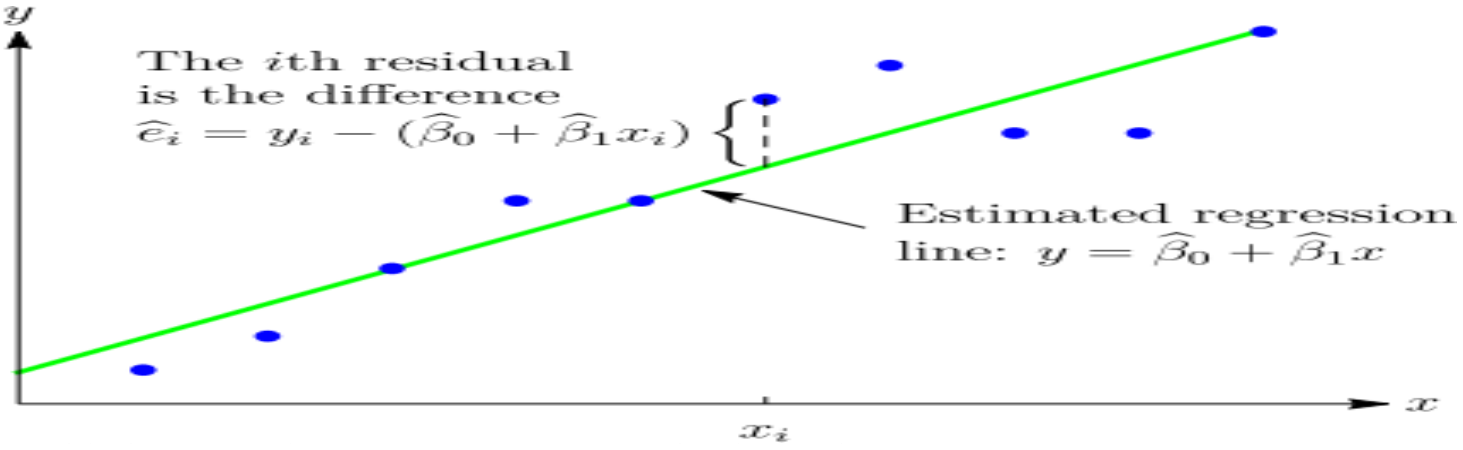
Recall that the assumptions in a simple linear regression model are stated in terms of the random deviations E_i, $i = 1, 2, \ldots, n$. It is assumed that the E_i's are:

- independent,
- normal random variables
- with mean 0 and constant variance σ^2.

$$E_i \overset{\text{ind}}{\sim} \mathrm{N}\left(0, \sigma^2\right)$$

If any of these assumptions is violated, then the results and subsequent inferences are in doubt.

- If the true regression line were known, the set of actual random errors could be computed and used to check the assumptions.
  $e_i = y_i - (\beta_0 + \beta_1 x_i)$
- We usually do not know the values for β_0 and β_1, so we must use the **residuals** (deviations from the estimated regression line) to check assumptions.

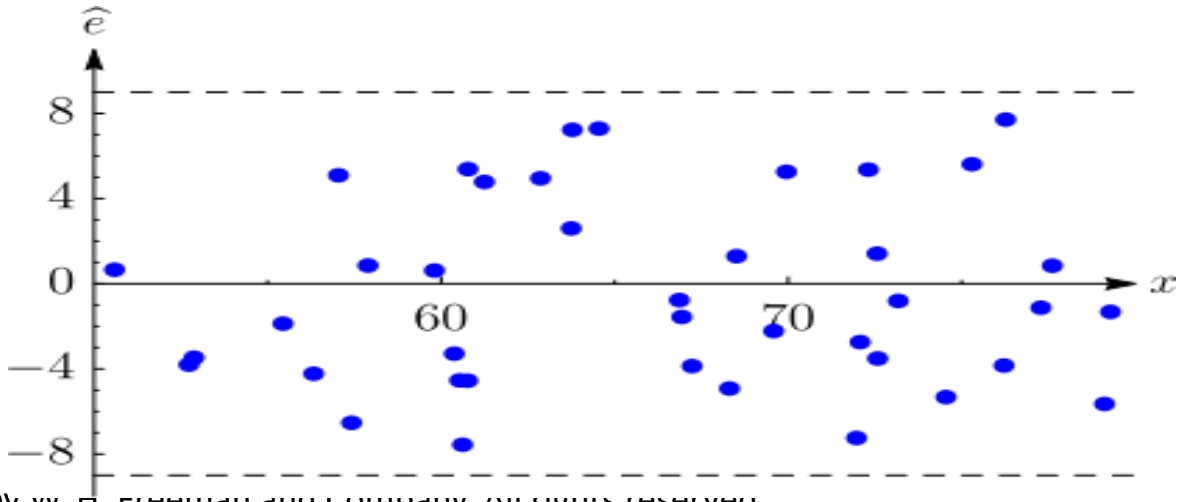


Regression Diagnostics

- Residuals (estimates of the random errors) are used in a variety of diagnostic checks.
- Several preliminary *graphical procedures* are used to reveal assumption violations.

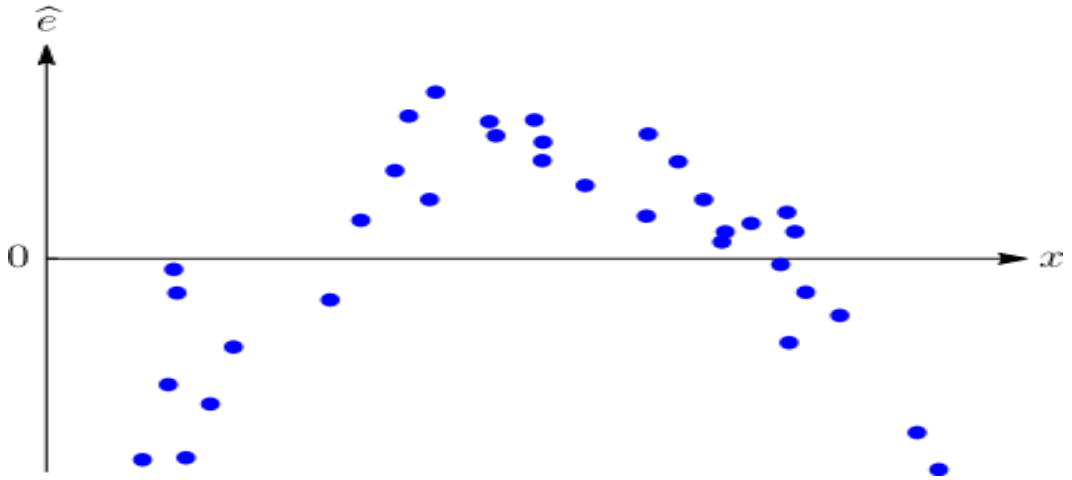
1. Use a **normal probability plot** of the residuals to check the **normality assumption.**
2. Use a **histogram** or stem-and-leaf plot of residuals to check the **normality assumption.**
3. Use a **scatter plot of residuals versus the independent variable** values to check the assumptions. If there are no violations of assumptions, a scatter plot of the residuals versus the independent variable should look like a horizontal band around zero with randomly distributed points and no discernible pattern.

Example1:
No assumption violations (a random pattern in the residual plot).

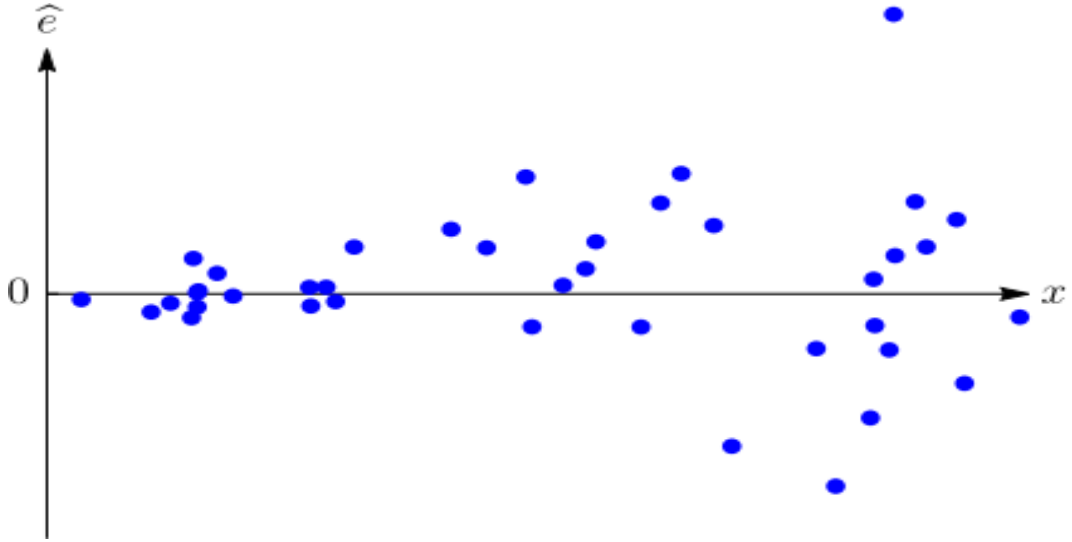


Regression Diagnostics

Example2: A distinct curve in the plot, either mound- or bowl-shaped (parabolic) suggests that an additional predictor variable may be necessary. A linear model is not appropriate.

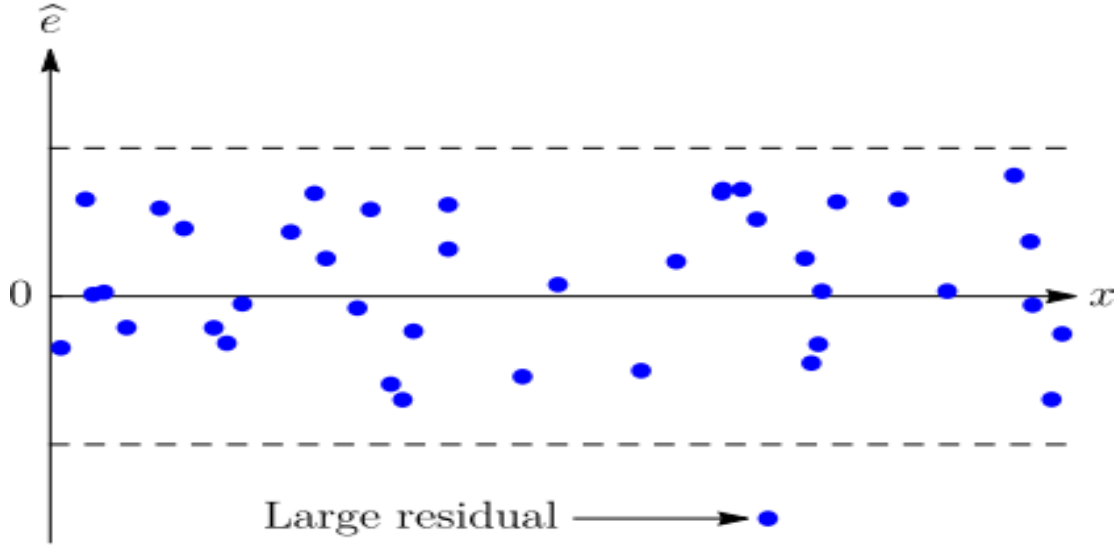


Example3: A nonconstant spread. If there is not a uniform horizontal band, or if the spread of the residuals varies outside this band, this suggests that the variance is not constant/same for each value of *x*.

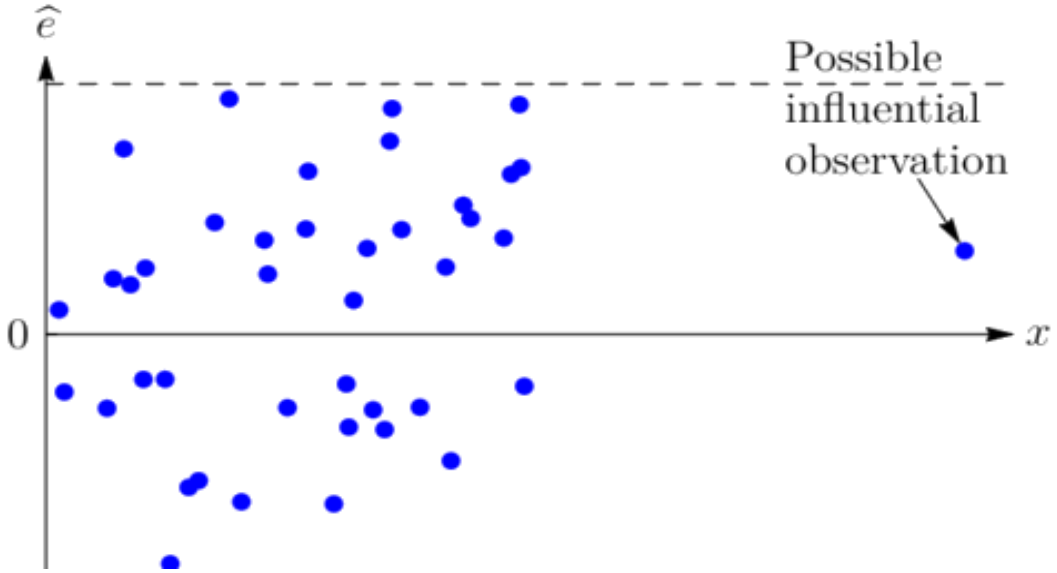


Regression Diagnostics

Example4: An unusually large (in magnitude) residual. This suggests that one observation is very far from the rest. The data may have been recorded or entered incorrectly.



Example5: Any outliers? If the observation is correct, an outlying residual suggests that one observation has an unusually large influence on the estimated regression line.

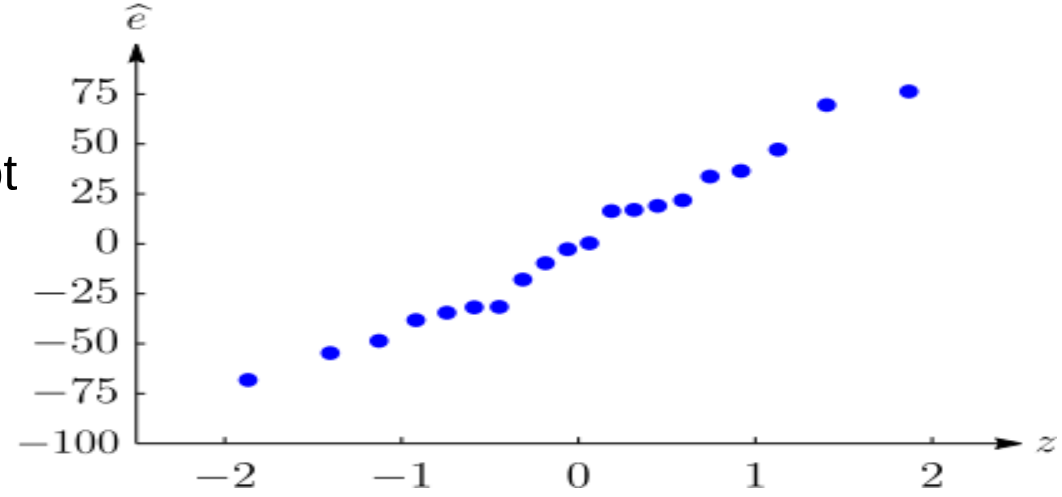


Example: Pillbugs and Red Clover

Some scientists believe that their preferred natural shelter is red clover, and that the density of pillbugs can be predicted from the density of red clover. The number of red clover plants per square meter (x) and the number of pillbugs per square meter (y) were measured for each 20 random field.

y = −12.360 + 14.213x.

- After computing the residuals, construct a normal probability plot
- The points lie along an approximate straight line.
- There is no evidence to suggest that the normality assumption is violated.



- After sketching the scatter plot of the residuals versus the predictor variable values, there is no discernible/recognizable (a distinct curve, nonconstant spread, an unusually large residual, or an outlier) pattern in this plot.