Chapter 9: Correlation and Regression

- 9.1 Correlation
- 9.2 Linear Regression
- 9.3 Measures of Regression and Prediction Interval

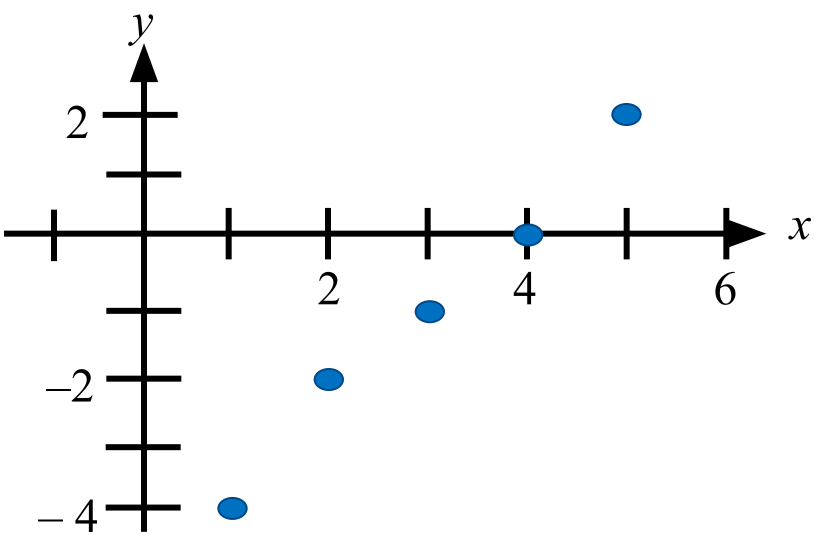
Correlation

Correlation

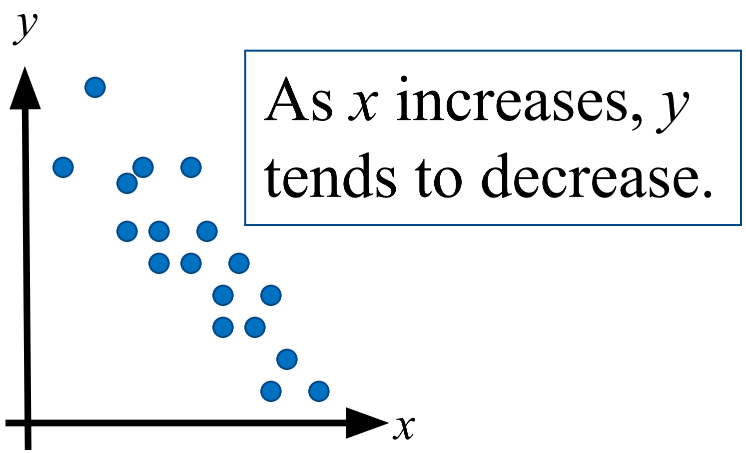
- A relationship between two variables.
- The data can be represented by ordered pairs (x, y)
  - x is the **independent** (or **explanatory**) **variable**
  - y is the **dependent** (or **response**) **variable**

A **scatter plot** can be used to determine whether a linear (straight line) correlation exists between two variables.

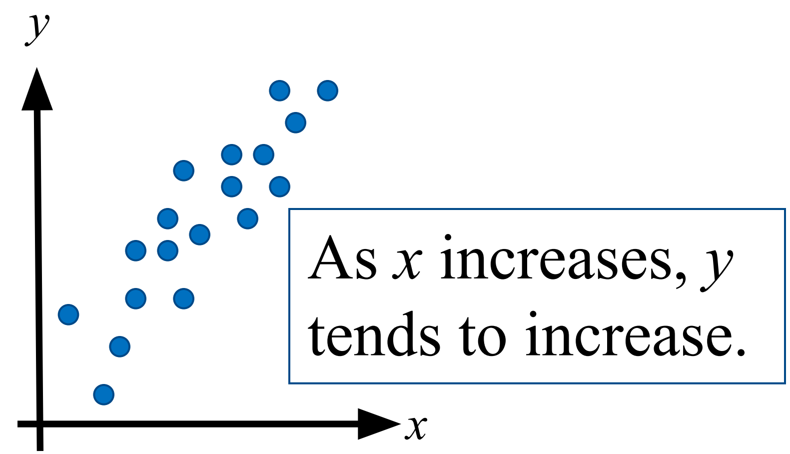
x	1	2	3	4	5
y	– 4	– 2	– 1	0	2



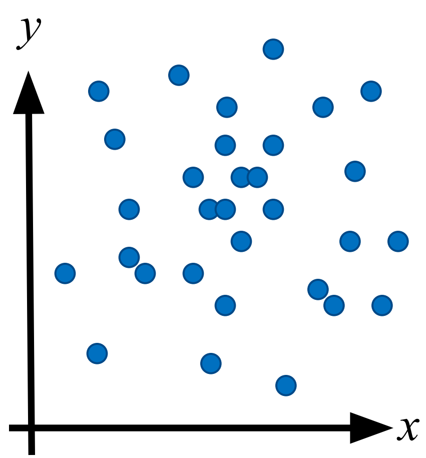
Types of Correlation



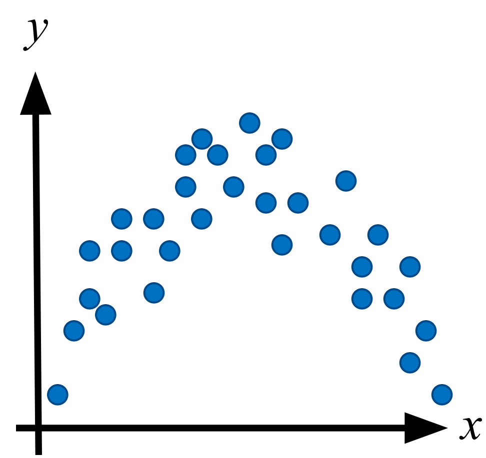
Negative Linear Correlation



Positive Linear Correlation



No Correlation

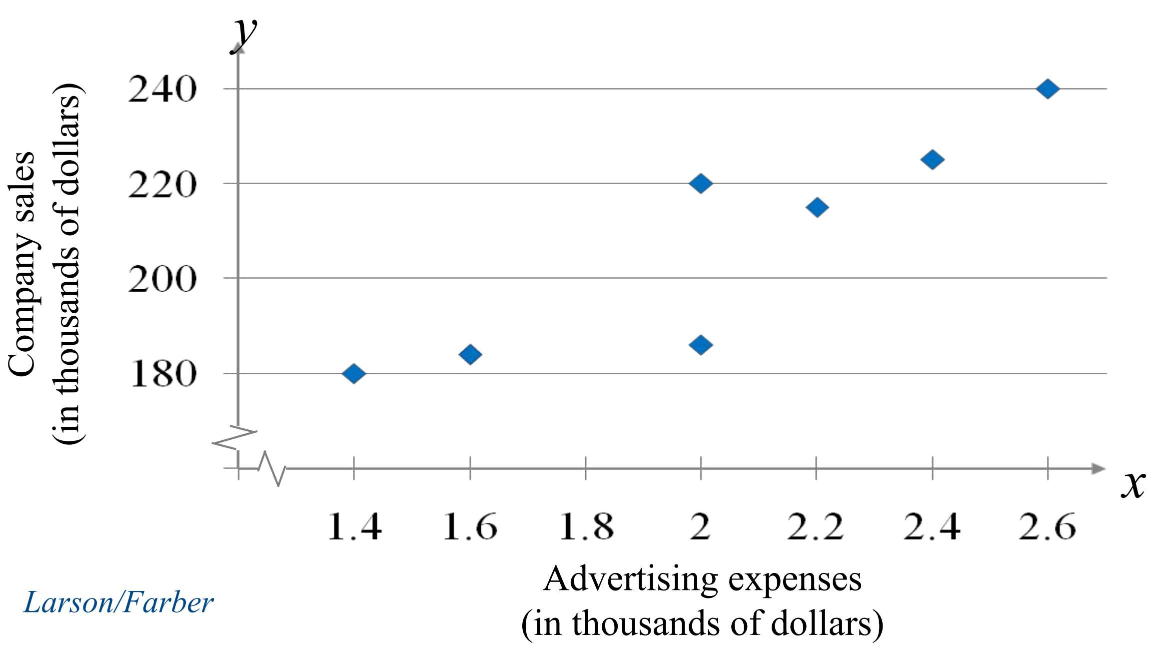


Nonlinear Correlation

Example: Constructing a Scatter Plot

A marketing manager conducted a study to determine whether there is a linear relationship between money spent on advertising and company sales. The data are shown in the table. Display the data in a scatter plot and determine whether there appears to be a positive or negative linear correlation or no linear correlation.

Advertising expenses, ($1000), x	Company sales ($1000), y
2.4	225
1.6	184
2.0	220
2.6	240
1.4	180
1.6	184
2.0	186
2.2	215

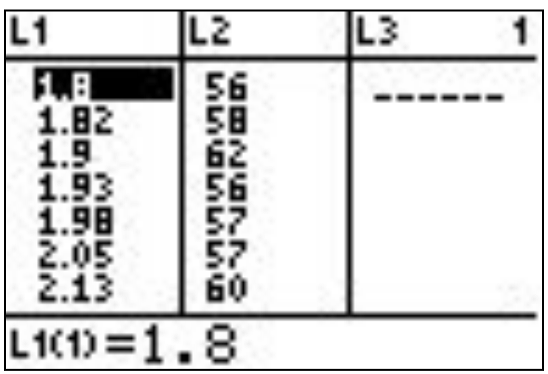


Positive linear correlation. As the advertising expenses increase, the sales tend to increase.

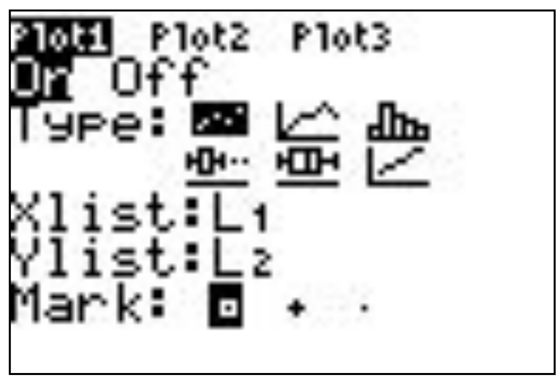
Constructing a Scatter Plot Using Technology

- Enter the *x*-values into list L1 and the *y*-values into list L2.
- Use *Stat Plot* to construct the scatter plot.

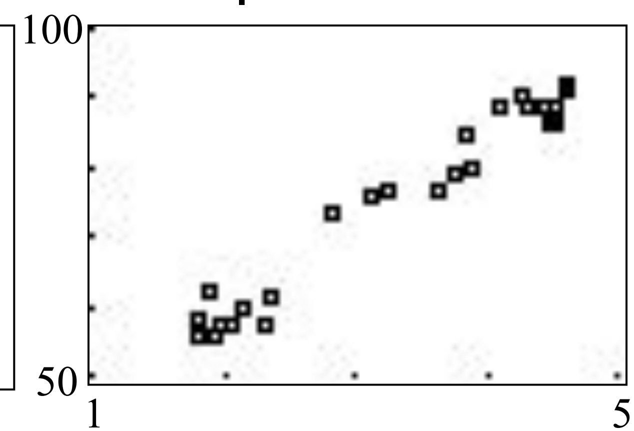
STAT > Edit…



STATPLOT



Graph



Correlation Coefficient

Correlation coefficient

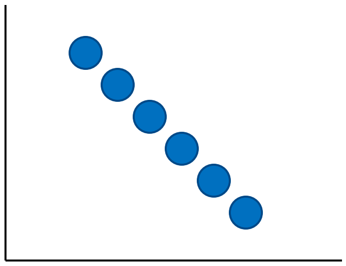
- A measure of the strength and the direction of a linear relationship between two variables.
- ***r*** represents the <u>sample</u> correlation coefficient.
- ***ρ*** (rho) represents the <u>population</u> correlation coefficient

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

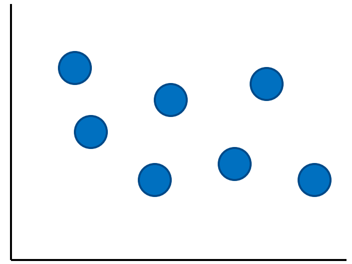
n is the number of data pairs

- The range of the correlation coefficient is -1 to 1.

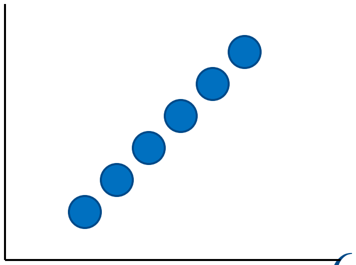
If ***r* = -1** there is a perfect negative correlation



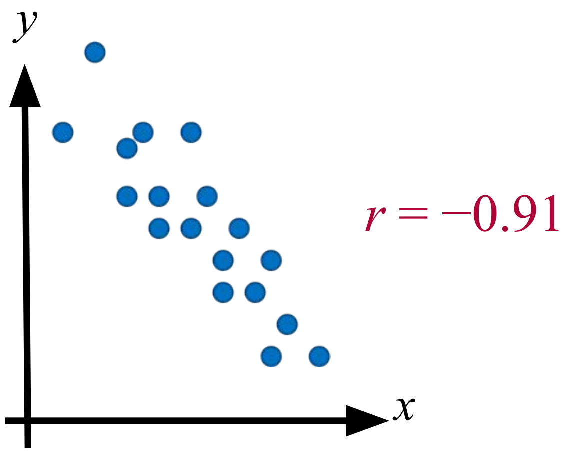
If ***r* is close to 0** there is no linear correlation



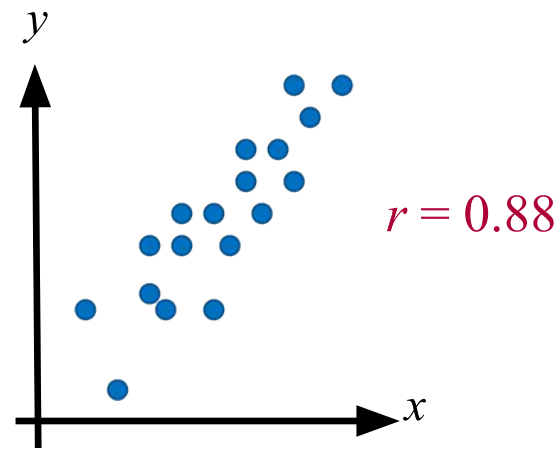
If ***r* = 1** there is a perfect positive correlation



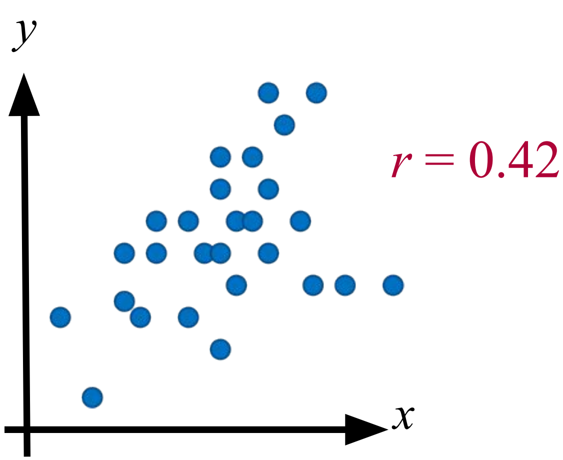
Linear Correlation



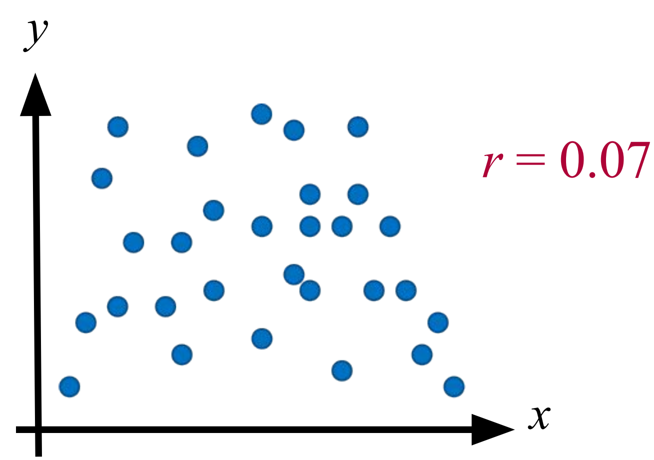
Strong negative correlation



Strong positive correlation



Weak positive correlation



Nonlinear Correlation

Calculating a Correlation Coefficient

	In Words	In Symbols
1.	Find the sum of the x-values.	$\sum x$
2.	Find the sum of the y-values.	$\sum y$
3.	Multiply each x-value by its corresponding y-value and find the sum.	$\sum xy$
4.	Square each x-value and find the sum.	$\sum x^2$
5.	Square each y-value and find the sum.	$\sum y^2$
6.	Use these five sums to calculate the correlation coefficient.	$r = \dfrac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}$

Example: Finding the Correlation Coefficient

Advertising expenses, (\$1000), x	Company sales (\$1000), y
2.4	225
1.6	184
2.0	220
2.6	240
1.4	180
1.6	184
2.0	186
2.2	215

Calculate the correlation coefficient for the advertising expenditures and company sales data. What can you conclude?

x	y	xy	x^2	y^2
2.4	225	540	5.76	50,625
1.6	184	294.4	2.56	33,856
2.0	220	440	4	48,400
2.6	240	624	6.76	57,600
1.4	180	252	1.96	32,400
1.6	184	294.4	2.56	33,856
2.0	186	372	4	34,596
2.2	215	473	4.84	46,225
$\Sigma x = 15.8$	$\Sigma y = 1634$	$\Sigma xy = 3289.8$	$\Sigma x^2 = 32.44$	$\Sigma y^2 = 337,558$

Finding the Correlation Coefficient Example Continued…

$\Sigma x = 15.8 \quad \Sigma y = 1634 \quad \Sigma xy = 3289.8 \quad \Sigma x^2 = 32.44 \quad \Sigma y^2 = 337{,}558$

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

$$= \frac{8(3289.8) - (15.8)(1634)}{\sqrt{8(32.44) - 15.8^2}\sqrt{8(337,558) - 1634^2}}$$

$$= \frac{501.2}{\sqrt{9.88}\sqrt{30,508}} \approx 0.9129$$

r ≈ 0.913 suggests a strong positive linear correlation. As the amount spent on advertising increases, the company sales also increase.

<u>Ti83/84</u>
Catalog – Diagnostic ON
Stat-Calc-4:LinReg(ax+b) L1, L2

Using a Table to Test a Population Correlation Coefficient ρ

- Once the sample correlation coefficient r has been calculated, we need to determine whether there is enough evidence to decide that the population correlation coefficient ρ is significant at a specified level of significance.
- Use Table 11 in Appendix B.
- **If $|r|$ is greater than the critical value, there is enough evidence to decide that the correlation coefficient ρ is significant.**

For Example: To determine whether ρ is significant for five pairs of data ($n = 5$) at a level of significance of $\alpha = 0.01$

Reject H_0: $\rho = 0$ if the absolute value of r is greater than the value given in the table.

n	$\alpha = 0.05$	$\alpha = 0.01$
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875

If $|r| > 0.959$, the correlation is significant. Otherwise, there is not enough evidence to conclude that the correlation is significant.

Hypothesis Testing for a Population Correlation Coefficient ρ

- A hypothesis test (one or two tailed) can also be used to determine whether the sample correlation coefficient r provides enough evidence to conclude that the population correlation coefficient ρ is significant at a specified level of significance.

- **Left-tailed test**

  H_0: $\rho \geq 0$ (no significant negative correlation)
  H_a: $\rho < 0$ (significant negative correlation)

- **Right-tailed test**

  H_0: $\rho \leq 0$ (no significant positive correlation)
  H_a: $\rho > 0$ (significant positive correlation)

- **Two-tailed test**

  H_0: $\rho = 0$ (no significant correlation)
  H_a: $\rho \neq 0$ (significant correlation)

Using the t-Test for ρ

In Words	*In Symbols*
1. State the null and alternative hypothesis.	State H_0 and H_a.
2. Specify the level of significance.	Identify α.
3. Identify the degrees of freedom.	d.f. = $n - 2$.
4. Determine the critical value(s) and rejection region(s).	Use Table 5 in Appendix B.
5. Find the standardized test statistic.	$t = \dfrac{r}{\sqrt{\dfrac{1-r^2}{n-2}}}$
6. Make a decision to reject or fail to reject the null hypothesis and interpret the decision in terms of the original claim.	If t is in the rejection region, reject H_0. Otherwise fail to reject H_0.

Example: *t*-Test for a Correlation Coefficient

For the advertising data, we previously calculated **r ≈ 0.9129**. Test the significance of this correlation coefficient. Use α = 0.05.

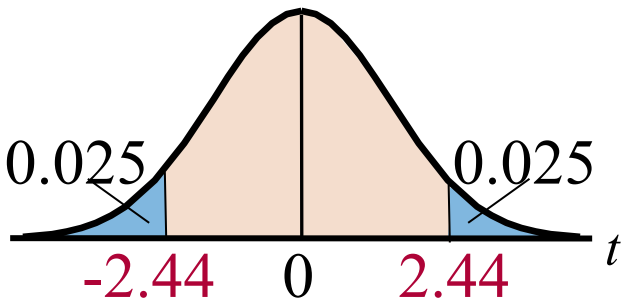
Advertising expenses, (\$1000), x	Company sales (\$1000), y
2.4	225
1.6	184
2.0	220
2.6	240
1.4	180
1.6	184
2.0	186
2.2	215

H_0 $\rho = 0$

H_a $\rho \neq 0$

α **0.05**

d.f. **8 – 2 = 6**



Test Statistic: $t = \dfrac{0.9129}{\sqrt{\dfrac{1-(0.9129)^2}{8-2}}} \approx 5.478$

Stat-Tests
LinRegTTest

Decision: Reject H_0

At the 5% level of significance, there is enough evidence to conclude that there is a significant linear correlation between advertising expenses and company sales.

Correlation and Causation

- The fact that two variables are strongly correlated does not in itself imply a cause-and-effect relationship between the variables.
- If there is a significant correlation between two variables, you should consider the following possibilities:

1. Is there a direct cause-and-effect relationship between the variables?
   - **Does *x* cause *y*?**

2. Is there a reverse cause-and-effect relationship between the variables?
   - **Does *y* cause *x*?**

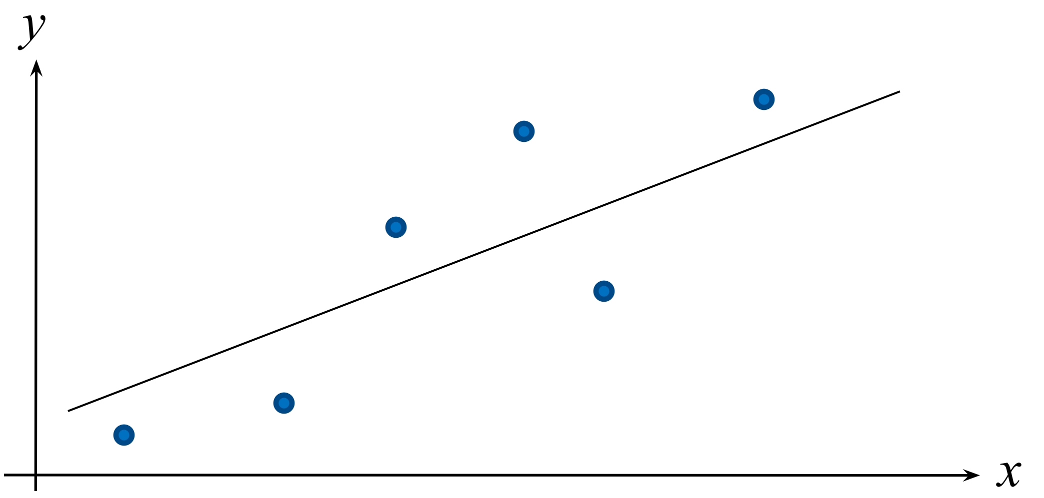
3. Is it possible that the relationship between the variables can be **caused by a third variable** or by a combination of several other variables?

4. Is it possible that the relationship between two variables may be a **coincidence?**

9.2 Objectives

- Find the equation of a regression line
- Predict *y*-values using a regression equation

After verifying that the linear correlation between two variables is significant, we determine the equation of the line that best models the data (**regression line**) - used to predict the value of *y* for a given value of *x*.



Residuals & Equation of Line of Regression

Residual

- The difference between the observed *y*-value and the predicted *y*-value for a given *x*-value on the line.

Regression line

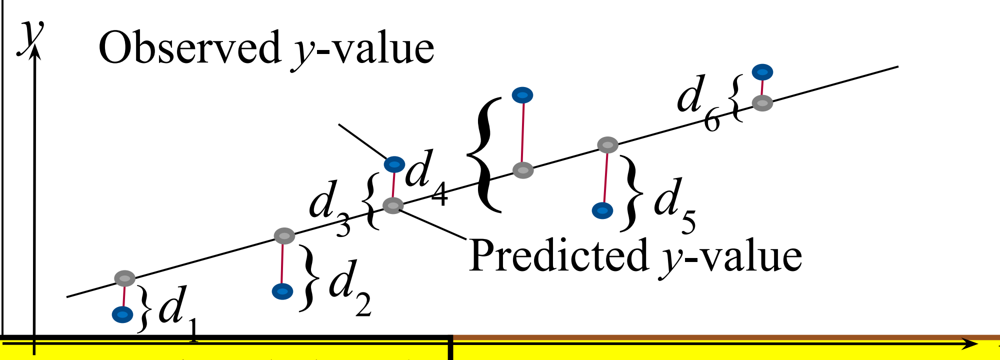
☐ **Line of best fit**

- The line for which the sum of the squares of the residuals is a minimum.
- Equation of Regression

  $\hat{y} = mx + b$

For a given *x*-value,

d_i = **(observed *y*-value) – (predicted *y*-value)**



$\hat{y}$ - predicted y-value
m – slope
b – y-intercept

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

$$b=\bar{y}-m\bar{x}=\frac{\sum y}{n}-m\frac{\sum x}{n}$$

$\bar{y}$ - mean of y-values in the data
$\bar{x}$ - mean of x-values in the data
The regression line always passes through $\left(\bar{x},\bar{y}\right)$

Finding Equation for Line of Regression

Recall the data from section 9.1

x	y	xy	x^2	y^2
2.4	225	540	5.76	50,625
1.6	184	294.4	2.56	33,856
2.0	220	440	4	48,400
2.6	240	624	6.76	57,600
1.4	180	252	1.96	32,400
1.6	184	294.4	2.56	33,856
2.0	186	372	4	34,596
2.2	215	473	4.84	46,225
$\Sigma x = 15.8$	$\Sigma y = 1634$	$\Sigma xy = 3289.8$	$\Sigma x^2 = 32.44$	$\Sigma y^2 = 337{,}558$

Advertising expenses, (\$1000), x	Company sales (\$1000), y
2.4	225
1.6	184
2.0	220
2.6	240
1.4	180
1.6	184
2.0	186
2.2	215

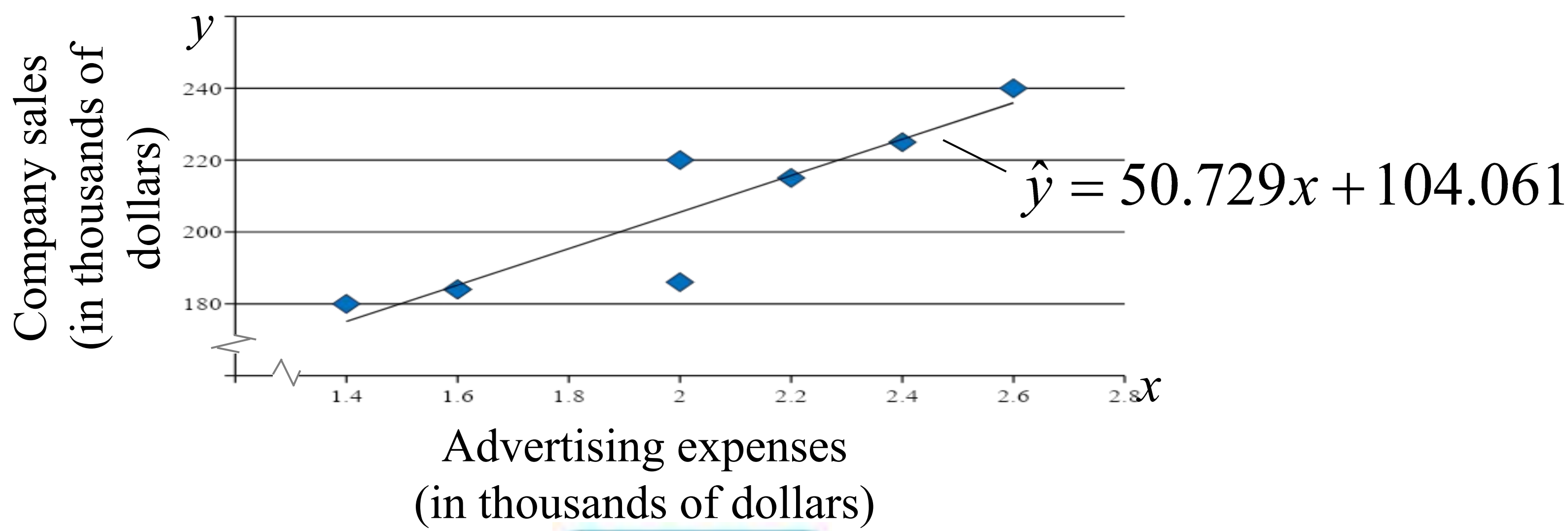
$$m = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{8(3289.8) - (15.8)(1634)}{8(32.44) - 15.8^2} = \frac{501.2}{9.88} \approx 50.72874$$

$$b = \bar{y} - m\bar{x} = \frac{1634}{8} - (50.72874)\frac{15.8}{8} = 204.25 - (50.72874)(1.975) \approx 104.0607$$

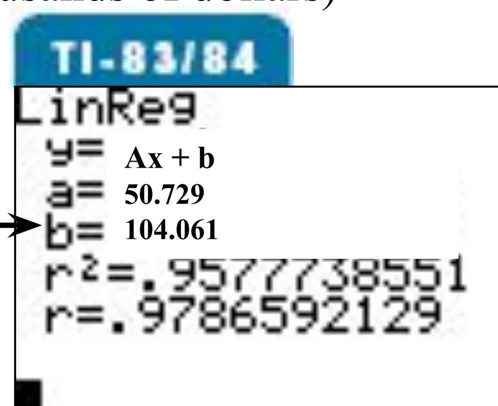
Equation of Line of Regression : $\hat{y} = 50.729x + 104.061$

Solution: Finding the Equation of a Regression Line

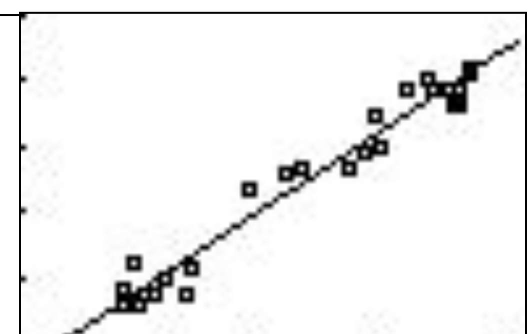
- To sketch the regression line, use any two *x*-values within the range of the data and calculate the corresponding *y*-values from the regression line.



Ti83/84
Catalog – Diagnostic ON
Stat-Calc-4:LinReg(ax+b) L1, L2



StatPlot and Graph



Example: Predicting y-Values Using Regression Equations

The regression equation for the advertising expenses (in thousands of dollars) and company sales (in thousands of dollars) data is $\hat{y} = 50.729x + 104.061$. Use this equation to predict the *expected* company sales for the advertising expenses below:

1. **1.5 thousand dollars** : $\hat{y} = 50.729(1.5) + 104.061 \approx 180.155$

   When advertising expenses are \$1500, company sales are about \$180,155.

2. **1.8 thousand dollars** $\hat{y} = 50.729(1.8) + 104.061 \approx 195.373$

   When advertising expenses are \$1800, company sales are about \$195,373.

3. **2.5 thousand dollars** $\hat{y} = 50.729(2.5) + 104.061 \approx 230.884$

   When advertising expenses are \$2500, company sales are about \$230,884.

Prediction values are meaningful only for *x*-values in (or close to) the range of the data. *X*-values in the original data set range from 1.4 to 2.6. It is not appropriate to use the regression line to predict company sales for advertising expenditures such as 0.5 (\$500) or 5.0 (\$5000).

9.3 Measures of Regression and Prediction Intervals (Objectives)

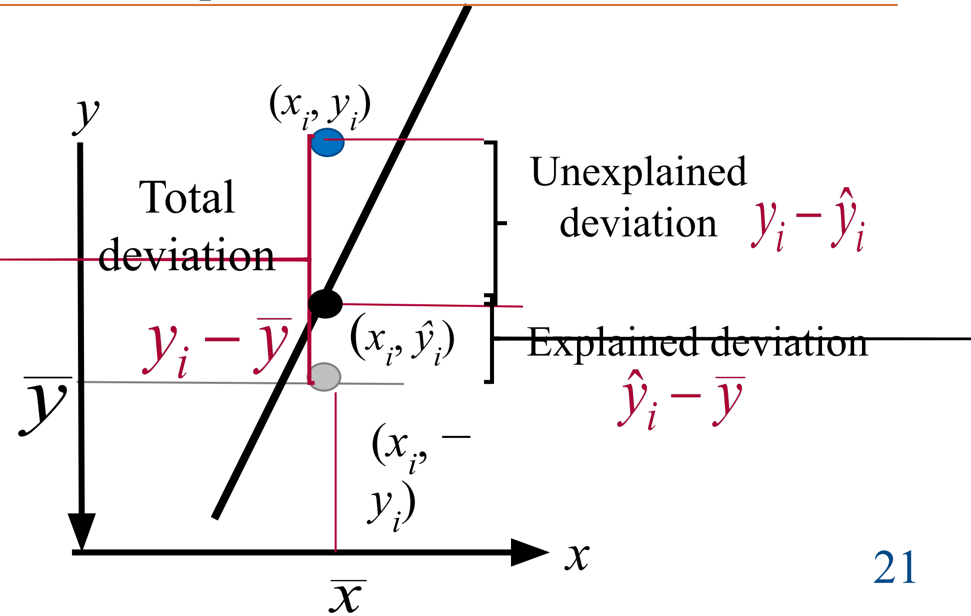
- Interpret the <u>three types of variation</u> about a regression line
- Find and interpret the <u>coefficient of determination</u>
- Find and interpret the <u>standard error of the estimate</u> for a regression line
- Construct and interpret a <u>prediction interval</u> for *y*

<u>Three types of variation about a regression line</u>

● Total variation ● Explained variation ● Unexplained variation

First calculate

- The **total deviation** $y_i - \bar{y}$
- The **explained deviation** $\hat{y}_i - \bar{y}$
- The **unexplained deviation** $y_i - \hat{y}_i$



Variation About a Regression Line

Total variation $= \sum (y_i - \overline{y})^2$

- The sum of the squares of the differences between the *y*-value of each ordered pair and the mean of *y*.

Total variation = Explained variation + Unexplained variation

---

Explained variation $\sum (\hat{y}_i - \overline{y})^2$

- The sum of the squares of the differences between each predicted *y*-value and the mean of *y*.

Unexplained variation $\sum (y_i - \hat{y}_i)^2$

- The sum of the squares of the differences between the *y*-value of each ordered pair and each corresponding predicted *y*-value.

Coefficient of determination (r^2)

- Ratio of the explained variation to the total variation.

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

For the advertising data, correlation coefficient $r \approx 0.913 \Rightarrow r^2 = (.913)^2 = .834$

About **83.4%** of the variation in company sales can be explained by variation in advertising expenditures. About **16.9%** of the variation is unexplained.

The Standard Error of Estimate

Standard error of estimate

- The standard deviation (s_e) of the observed y_i-values about the predicted $\hat{y}$-value for a given x_i-value.

$$s_e = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n-2}}$$

***n* = number of ordered data pairs.**

- The closer the observed *y*-values are to the predicted *y*-values, the smaller the standard error of estimate will be.

The regression equation for the advertising expenses and company sales data as calculated in section 9.2 is : $\hat{y} = 50.729x + 104.061$

x	y	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
2.4	225	225.81	$(225 - 225.81)^2 = 0.6561$
1.6	184	185.23	$(184 - 185.23)^2 = 1.5129$
2.0	220	205.52	$(220 - 205.52)^2 = 209.6704$
2.6	240	235.96	$(240 - 235.96)^2 = 16.3216$
1.4	180	175.08	$(180 - 175.08)^2 = 24.2064$
1.6	184	185.23	$(184 - 185.23)^2 = 1.5129$
2.0	186	205.52	$(186 - 205.52)^2 = 381.0304$
2.2	215	215.66	$(215 - 215.66)^2 = 0.4356$
			$\Sigma = 635.3463$

Unexplained variation

$$\sqrt{\frac{635.3463}{8-2}} \approx 10.290$$

The standard error of estimate of the company sales for a specific advertising expense is about $10.29.

Stat-Tests
LinRegTTest

Prediction Intervals

- Two variables have a **bivariate normal distribution** if for any fixed value of x, the corresponding values of y are normally distributed and for any fixed values of y, the corresponding x-values are normally distributed.

> Given a linear regression equation $\hat{y}_i = mx_i + b$ and x_0(a specific value of x), d.f. = n-2, a ***c*-prediction interval** for y is:
>
> $\hat{y} - E < y < \hat{y} + E$, ***where,*** $E = t_c s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n\sum x^2 - (\sum x)^2}}$
>
> The point estimate is $\hat{y}$ and the margin of error is E. The probability that the prediction interval contains y is c.

Point estimate:

$\hat{y} = 50.729(2.1) + 104.061 \approx \mathbf{210.592}$

Critical value:

d.f. = $n - 2 = 8 - 2 = 6$

$t_c = 2.447$

Example: Construct a 95% prediction interval for the company sales when the advertising expenses are $2100. What can you conclude?

Recall, $n = 8$, $\hat{y} = 50.729x + 104.061$, $s_e = 10.290$ $\Sigma x = 15.8,\quad \Sigma x^2 = 32.44,\quad \bar{x} = 1.975$

$$E = t_c s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n\sum x^2 - (\sum x)^2}}$$

$$= (2.447)(10.290)\sqrt{1 + \frac{1}{8} + \frac{8(2.1 - 1.975)^2}{8(32.44) - (15.8)^2}} \approx 26.857$$

Prediction Interval:

210.592 – 26.857 to 210.592 + 26.857

$\mathbf{183.735 < y < 237.449}$

You can be 95% confident that when advertising expenses are $2100, sales will be between $183,735 and $237,449.