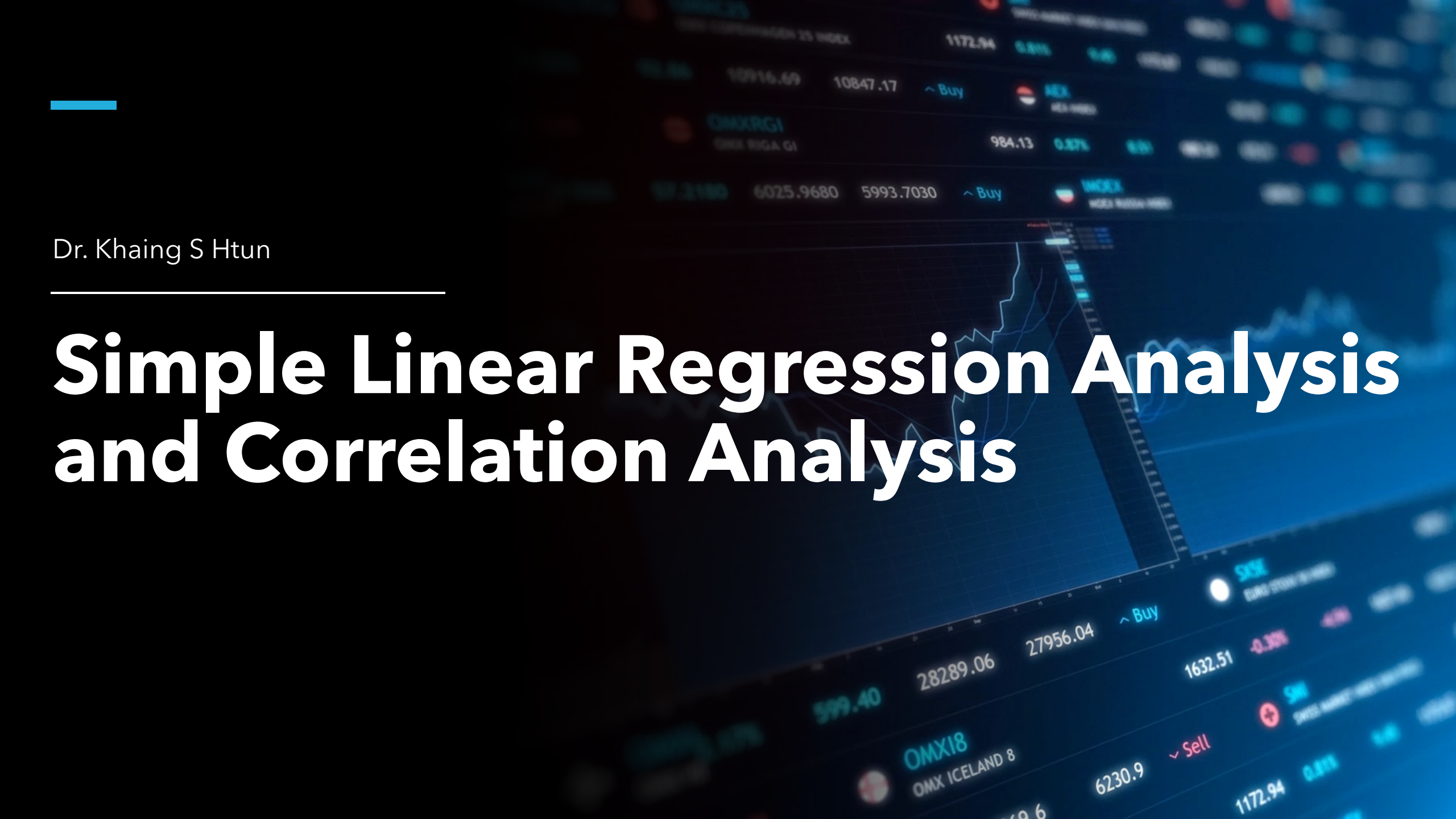
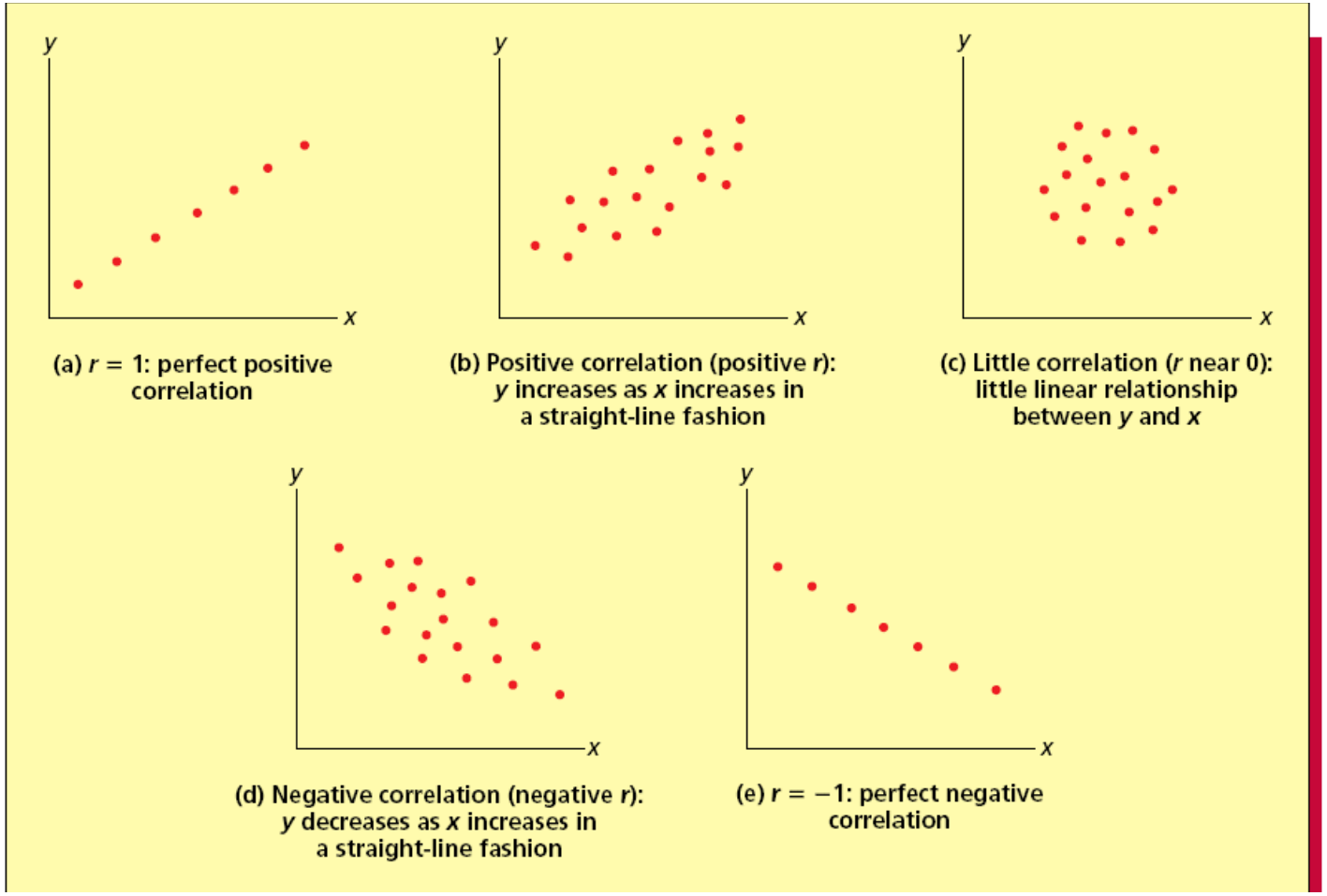


Dr. Khaing S Htun

Simple Linear Regression Analysis and Correlation Analysis

The Simple Correlation Analysis



- A measure of the linear relationship between variables and its strength

$$y = f(x)$$

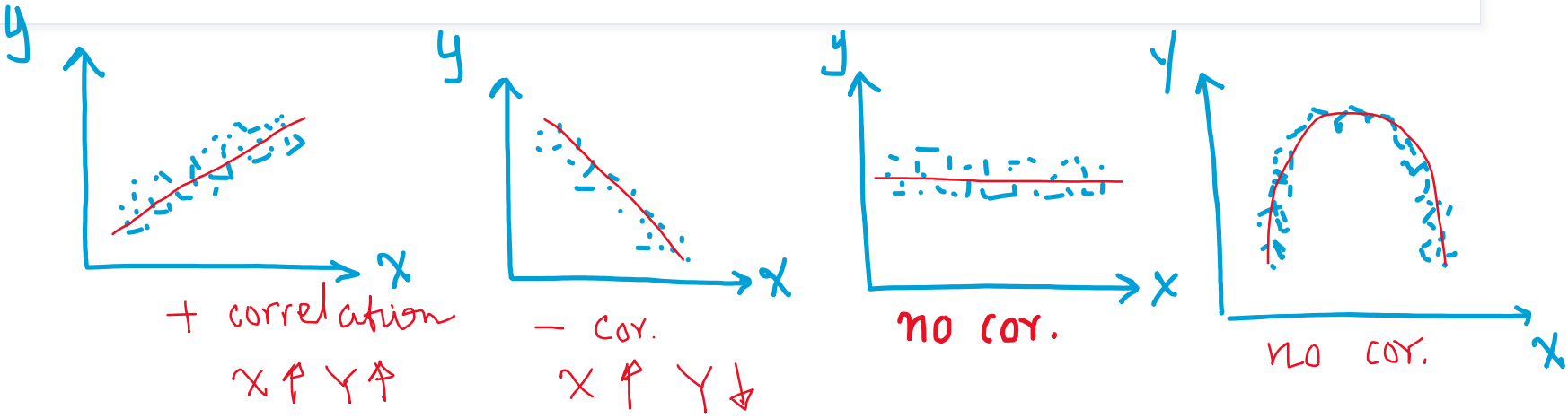
- Scatter diagram
  - Observed values (x_i, y_i),

$$i = 1, 2, 3, 4, \ldots, n$$

The Simple Correlation Analysis

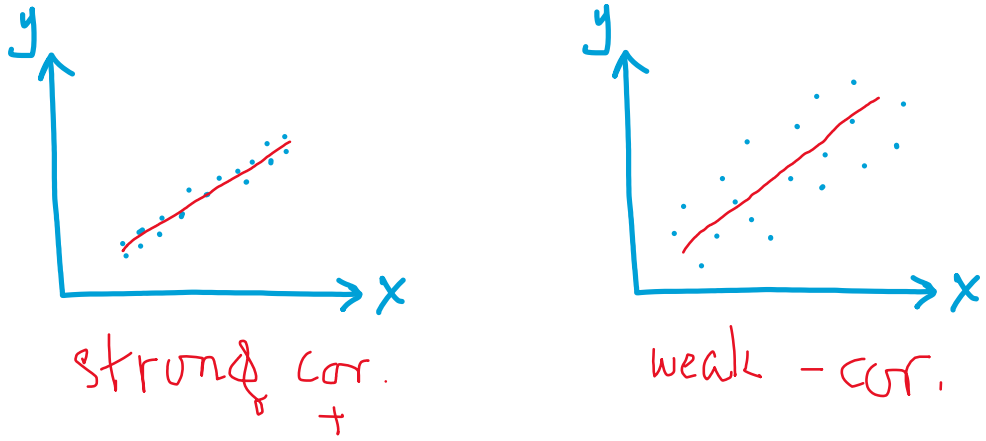
Direction

- Positive
- Negative
- No correlation



Strength

- Strong
- Medium
- Weak



The Simple Correlation Coefficient, r

- Karl Pearson about 1900

- **Correlation Coefficient, r**
  - -1 < **r** < 1
    - +1 = perfect positive linear relationship
    - 0 = no linear relationship
    - -1 = perfect negative linear relationship

Strength

- r= ±0.8 or higher → strong correlation
- r= ±0.5 – 0.8 → medium correlation
- r= ±0.4 or lower → weak correlation

The Simple Correlation Coefficient, r

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

n = the number of paired observations

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

where

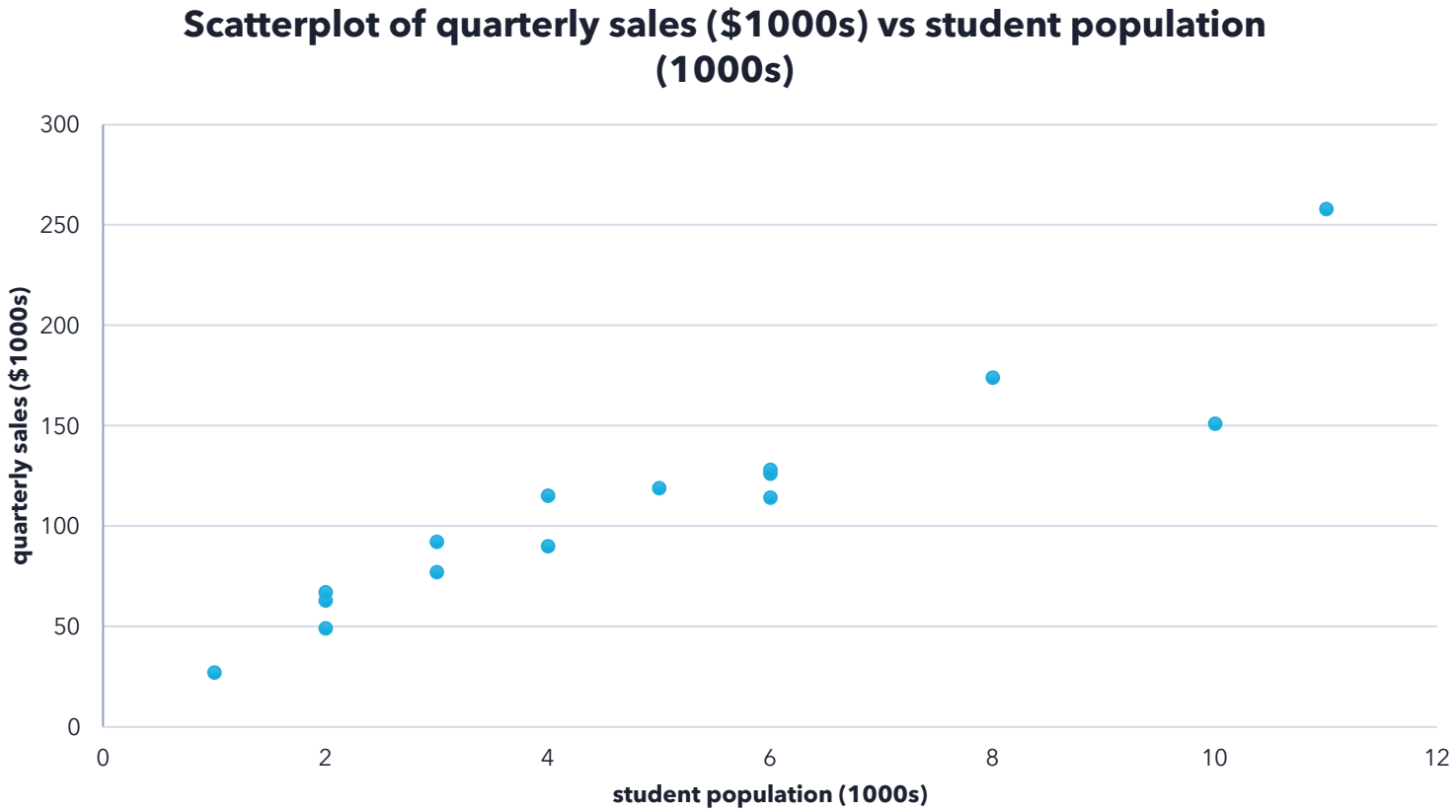
$$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$SS_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

Example: The Simple Correlation Coefficient, r

Restaurant	student population (1000s), x	quarterly sales ($1000s), y
1	3	92
2	2	63
3	6	126
4	8	174
5	2	49
6	4	90
7	5	119
8	6	114
9	2	67
10	4	115
11	6	128
12	11	258
13	3	77
14	10	151
15	1	27



Example: The Simple Correlation Coefficient, r

Restaurant	student population (1000s), x	quarterly sales ($1000s), y	x^2	y^2	xy
1	3	92	9	8464	276
2	2	63	4	3969	126
3	6	126	36	15876	756
4	8	174	64	30276	1392
5	2	49	4	2401	98
6	4	90	16	8100	360
7	5	119	25	14161	595
8	6	114	36	12996	684
9	2	67	4	4489	134
10	4	115	16	13225	460
11	6	128	36	16384	768
12	11	258	121	66564	2838
13	3	77	9	5929	231
14	10	151	100	22801	1510
15	1	27	1	729	27
Sum, Σ	**73**	**1,650**	**481**	**226,364**	**10,255**

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

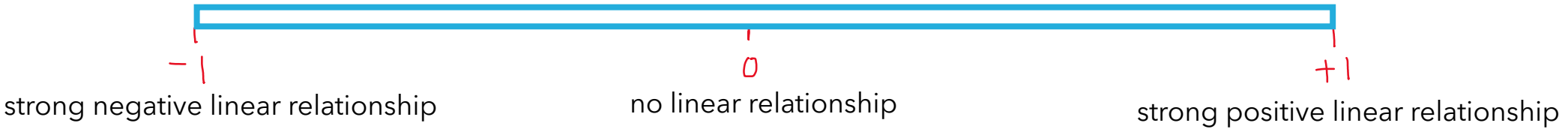
$$= \frac{15(10{,}255) - (73)(1{,}650)}{\sqrt{[15(481) - (73)^2][15(226{,}364) - (1{,}650)^2]}}$$

$$= \frac{153{,}825 - 120{,}450}{\sqrt{[1{,}886]\ [672{,}960]}} = \frac{33{,}375}{35{,}625.87} = +0.9368$$

relationship is **strong**

Testing the Significance of the Population Correlation Coefficient, ρ

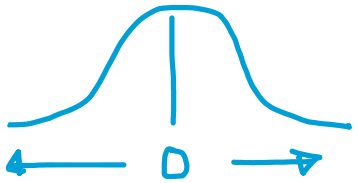
- r – **sample** correlation coefficient *(NOT population)*



$$the\ value\ of\ r\ is\ reliable\ or\ not?$$

Testing the Significance of the Population Correlation Coefficient, ρ

- ρ - **population** correlation coefficient *(NOT sample)*



Hypothesis testing:

1. State the **hypothesis** $H_0: \rho = 0$ $H_1: \rho \neq 0$
2. Find the **critical values** ⟹ from the table (α and d.f.)
3. Compute the **test value** (**t** test) formula
4. Make the **decision** with $\pm$ value
5. Summarize the result

Testing the Significance of the Population Correlation Coefficient, ρ

Hypotheses are:

Null hypothesis - H_0: $\rho = 0$

(There is no linear relationship between *y* and *x*, **or** the population correlation coefficient is zero.)

Alternative hypothesis - H_1: $\rho \neq 0$

(There is a linear relationship between *y* and *x*, **or** the population correlation coefficient is different from zero.)

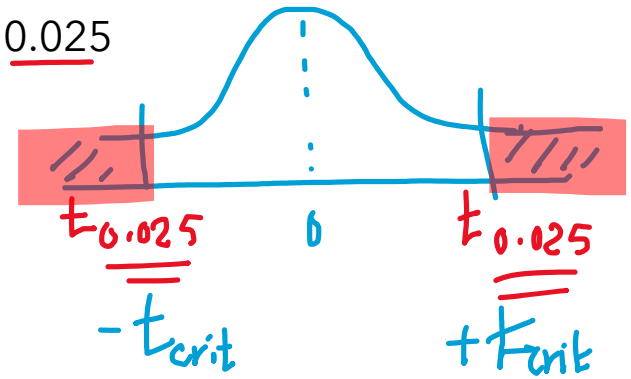
- *Reject* null hypothesis → significant difference between the value of r and 0
- *Fail to reject* → the value of r is **NOT** significantly different from 0

Testing the Significance of the Population Correlation Coefficient, ρ

Critical value

from the table

- degree of freedom, $d.f. = n - 2$
- α is the significant level
  - α = 1 – confidence level *(95%)*
  - α is 0.05 or 5%
    - 0.025 and 0.025



Degrees of Freedom	level of significant, α					
	.20	.10	.05	.025	.01	.005
1	1.376	3.078	6.314	12.706	31.821	63.656
2	1.061	1.886	2.920	4.303	6.965	9.925
3	.978	1.638	2.353	3.182	4.541	5.841
4	.941	1.533	2.132	2.776	3.747	4.604
5	.920	1.476	2.015	2.571	3.365	4.032
6	.906	1.440	1.943	2.447	3.143	3.707
7	.896	1.415	1.895	2.365	2.998	3.499
8	.889	1.397	1.860	2.306	2.896	3.355
9	.883	1.383	1.833	2.262	2.821	3.250
10	.879	1.372	1.812	2.228	2.764	3.169
11	.876	1.363	1.796	2.201	2.718	3.106
12	.873	1.356	1.782	2.179	2.681	3.055
13	.870	1.350	1.771	2.160	2.650	3.012
14	.868	1.345	1.761	2.145	2.624	2.977
15	.866	1.341	1.753	2.131	2.602	2.947
16	.865	1.337	1.746	2.120	2.583	2.921
17	.863	1.333	1.740	2.110	2.567	2.898
18	.862	1.330	1.734	2.101	2.552	2.878
19	.861	1.328	1.729	2.093	2.539	2.861

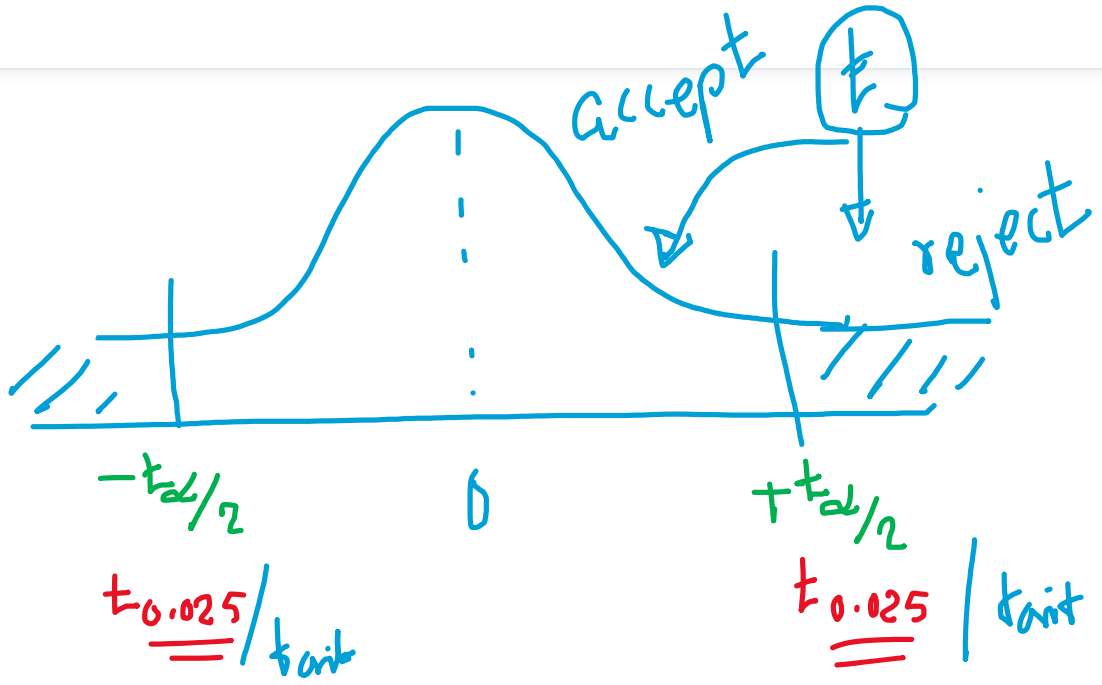
t_{crit} value

Testing the Significance of the Population Correlation Coefficient, ρ

from the formula

Test statistics,

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$



Decision rule: We will **reject** H_0 at α, if the value of $\mathbf{t \leq -t_{\alpha/2}}$ or $\mathbf{t \geq +t_{\alpha/2}}$

Example: Testing the Significance of the Population Correlation Coefficient, ρ

Restaurant	student population (1000s), x	quarterly sales ($1000s), y	x^2	y^2	xy
1	3	92	9	8464	276
2	2	63	4	3969	126
3	6	126	36	15876	756
4	8	174	64	30276	1392
5	2	49	4	2401	98
6	4	90	16	8100	360
7	5	119	25	14161	595
8	6	114	36	12996	684
9	2	67	4	4489	134
10	4	115	16	13225	460
11	6	128	36	16384	768
12	11	258	121	66564	2838
13	3	77	9	5929	231
14	10	151	100	22801	1510
15	1	27	1	729	27
Sum, Σ	**73**	**1,650**	**481**	**226,364**	**10,255**

n

$r = +0.9368$ – strong correlation

with $\alpha = 0.05$ and $d.f. = n - 2 = 15 - 2 = 13$

Test:

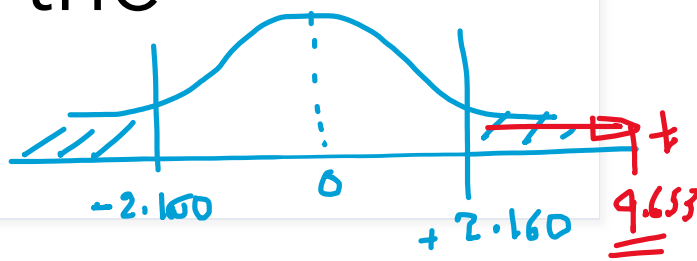
H_0: There is no a linear relationship between the quarterly sales(y) and the student population (x)

H_1: There is a linear relationship between the quarterly sales(y) and the student population(x)

test statistics,

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.9368\sqrt{15-2}}{\sqrt{1-(0.9368)^2}} = \frac{3.3777}{0.3499} = 9.653$$

Example: Testing the Significance of the Population Correlation Coefficient, ρ



Critical value

$\alpha = 0.05$ and $d.f. = n - 2 = 15 - 2 = 13$

- the critical value is $\pm t_{0.025} = \pm 2.160$
- computed t value, $t = 9.653$

Reject H_0 when $t \leq -t_{\alpha/2}$ or $t \geq +t_{\alpha/2}$

Therefore, H_0 is rejected at $\alpha = 0.05$.

level of significant, α

Degrees of Freedom	.20	.10	.05	.025	.01	.005
1	1.376	3.078	6.314	12.706	31.821	63.656
2	1.061	1.886	2.920	4.303	6.965	9.925
3	.978	1.638	2.353	3.182	4.541	5.841
4	.941	1.533	2.132	2.776	3.747	4.604
5	.920	1.476	2.015	2.571	3.365	4.032
6	.906	1.440	1.943	2.447	3.143	3.707
7	.896	1.415	1.895	2.365	2.998	3.499
8	.889	1.397	1.860	2.306	2.896	3.355
9	.883	1.383	1.833	2.262	2.821	3.250
10	.879	1.372	1.812	2.228	2.764	3.169
11	.876	1.363	1.796	2.201	2.718	3.106
12	.873	1.356	1.782	2.179	2.681	3.055
13	.870	1.350	1.771	2.160	2.650	3.012
14	.868	1.345	1.761	2.145	2.624	2.977
15	.866	1.341	1.753	2.131	2.602	2.947
16	.865	1.337	1.746	2.120	2.583	2.921
17	.863	1.333	1.740	2.110	2.567	2.898
18	.862	1.330	1.734	2.101	2.552	2.878
19	.861	1.328	1.729	2.093	2.539	2.861

we can conclude that there is a linear relationship between the student population(x) and the quarterly sales (y).

In-class Assignment

In-class Assignment: Correlation Coefficient

Find the value of the correlation coefficient, **r**, from the following table and test the significant of ρ

SUBJECT	AGE X	GLUCOSE LEVEL Y
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

The Simple Linear Regression Analysis

Correlation vs. Regression

- A scatter diagram can be used to show the relationship between two variables
- Correlation analysis is used to measure strength of the association (linear relationship) between two variables
  - Correlation is only concerned with strength of the relationship
  - **No causal effect** is implied with correlation

The Simple Linear Regression Analysis

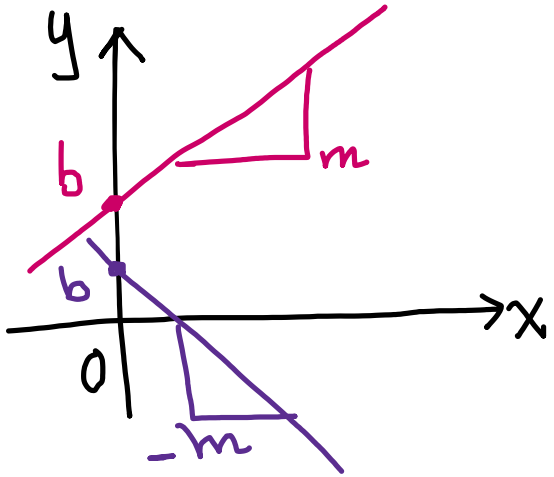
- Predict the value of a dependent variable based on the value of at least one independent variable
- Explain the impact of changes in an independent variable on the dependent variable
  - Dependent variable (y): the variable we wish to predict or explain
  - Independent variable (x): the variable used to predict or explain the dependent variable

The Simple Linear Regression Analysis

Linear equation -

slope y-intercept

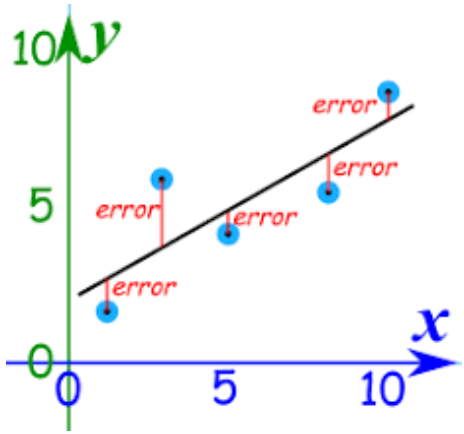
$$y = mx + b$$



Simple Linear regression model

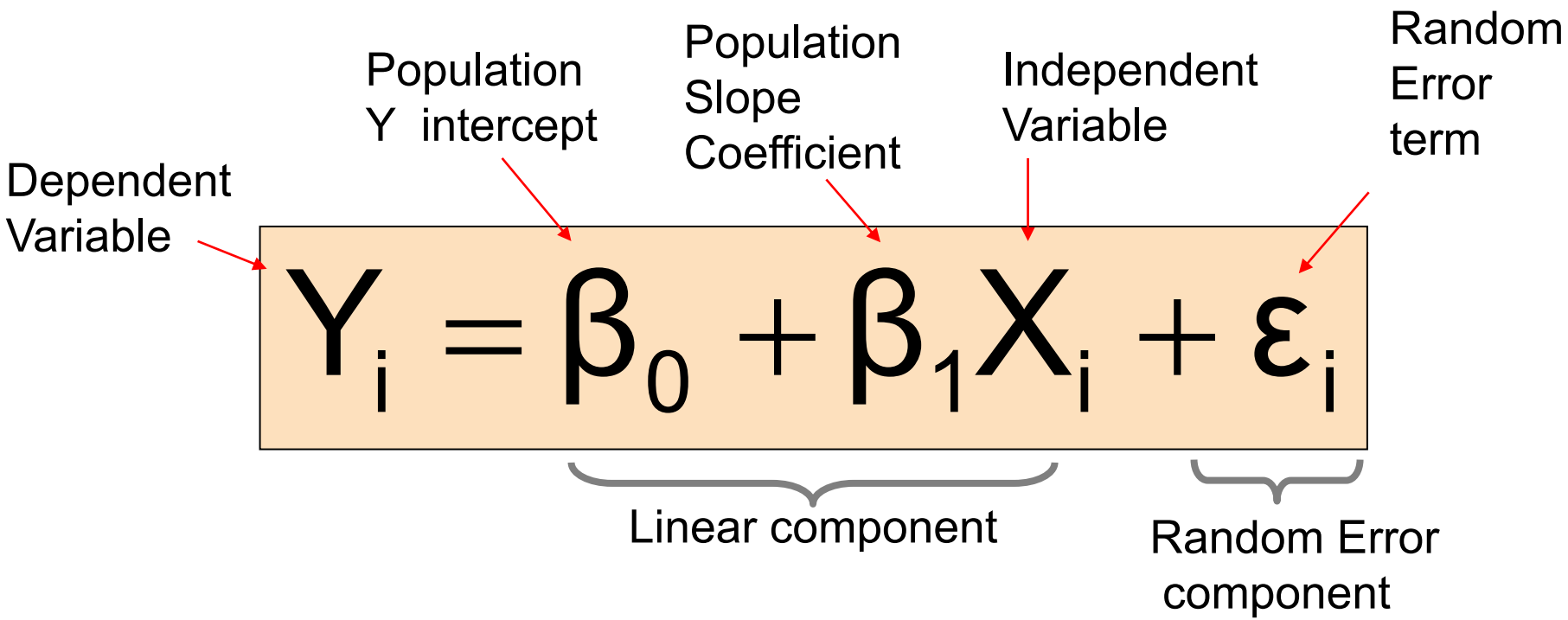
$$y = \beta_0 + \beta_1 x + \varepsilon$$

y-intercept slope error term

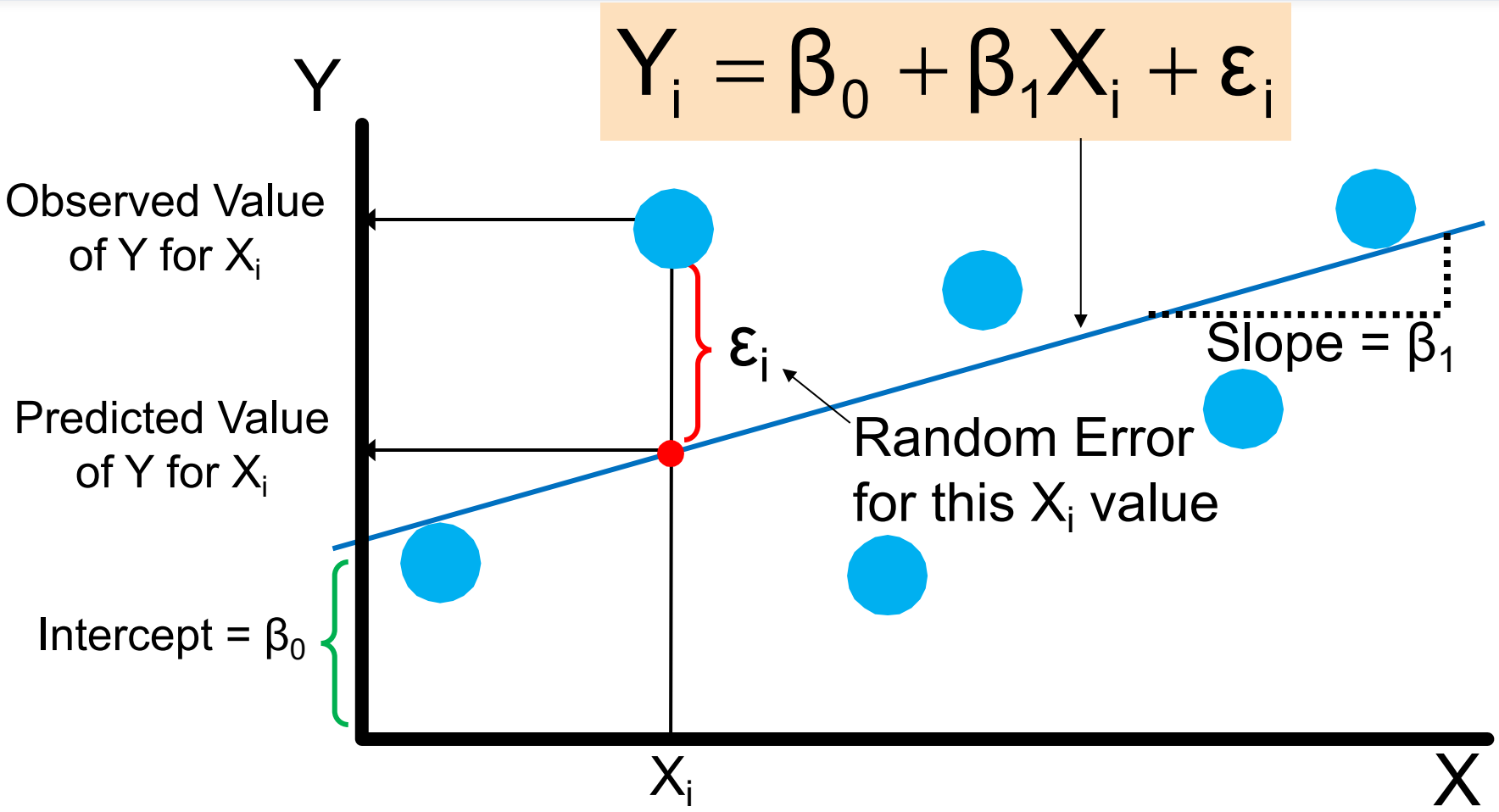


The Simple Linear Regression Analysis

Population Linear regression model



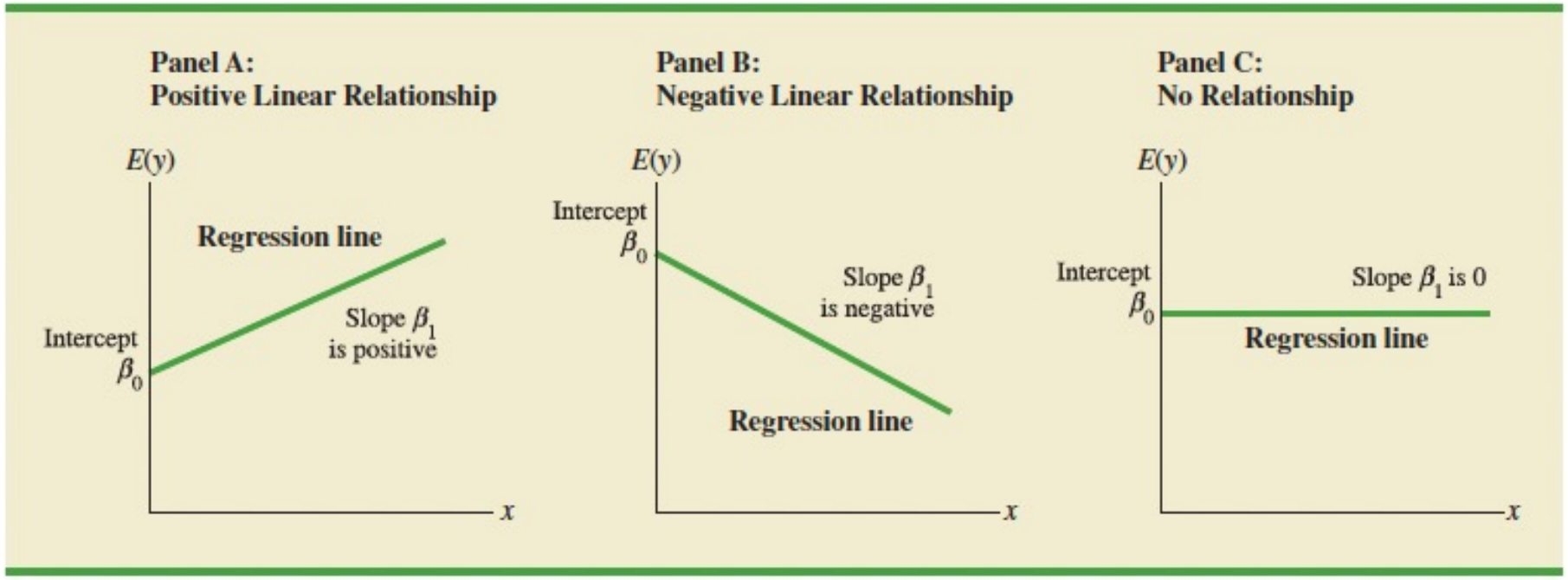
The Simple Linear Regression Analysis



The Simple Linear Regression Analysis

Simple Linear regression equation

$$y = \beta_0 + \beta_1 x$$



The Simple Linear Regression Analysis

The simple linear regression equation provides an **estimate** of the population regression line

Estimated Simple Linear regression equation

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

Estimated (or predicted) Y value for observation i

Estimate of the regression intercept β_0

Estimate of the regression slope β_1

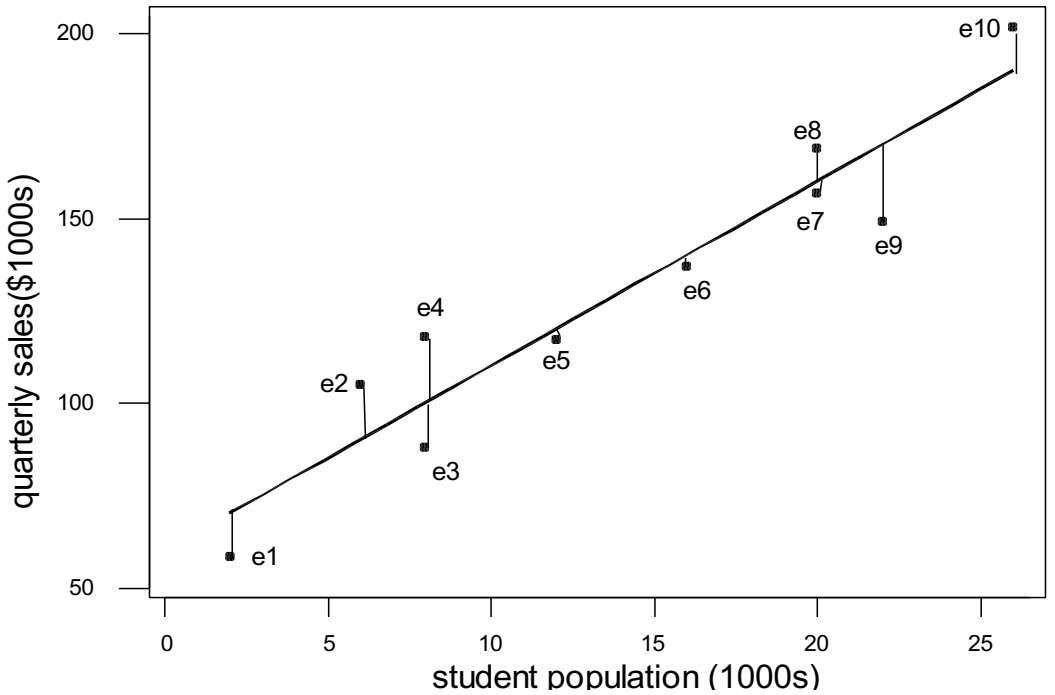
Value of X for observation i

$$y_i = b_0 + b_1 x_i + e_i$$

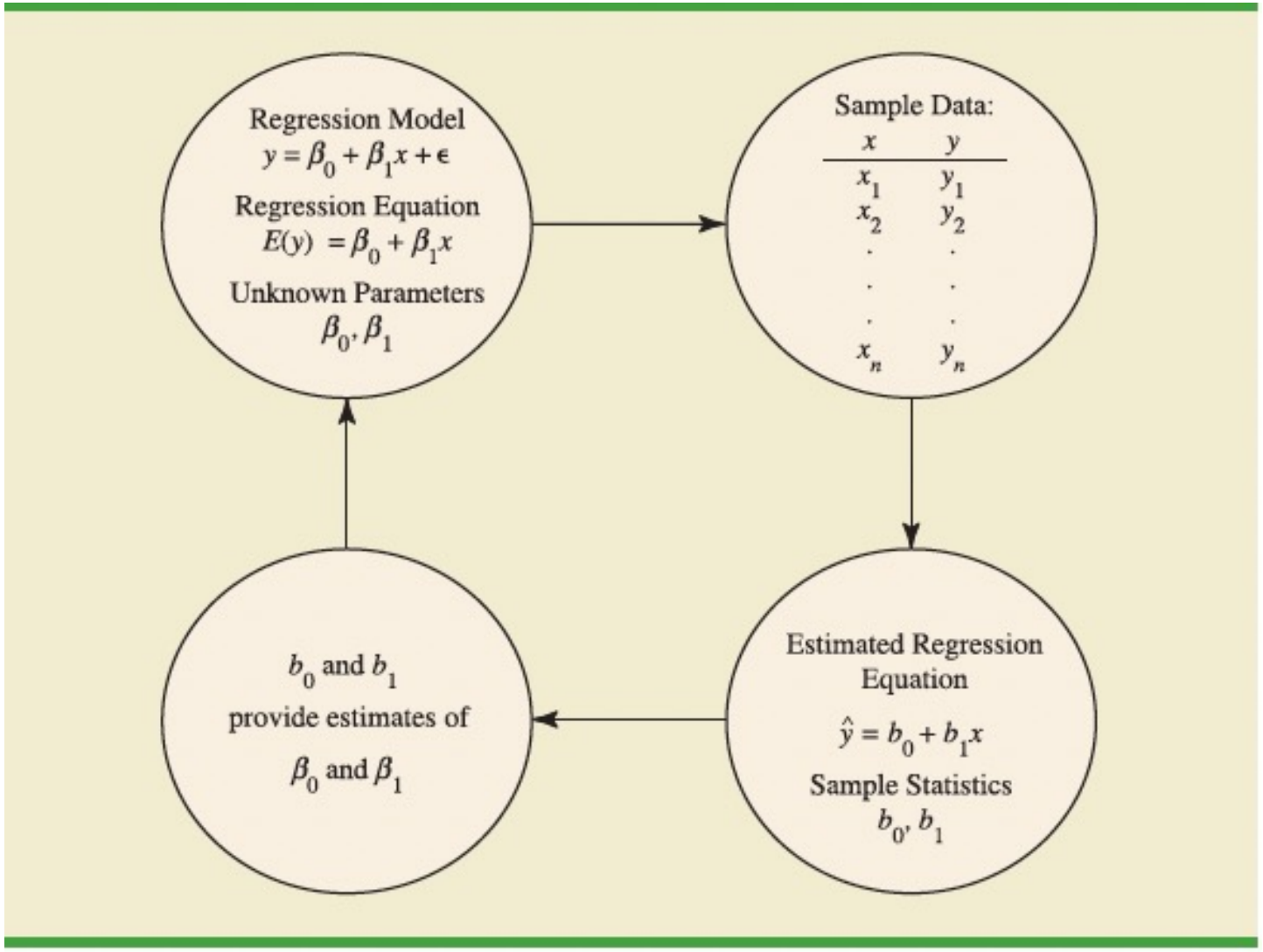
residual or error

The Simple Linear Regression Analysis

residual or error $e_i = y_i - \hat{y}_i$



The estimation process in simple linear regression



The Simple Linear Regression Analysis

The Least squares method

- The $\beta_0 \; and \; \beta_1$ are estimated by the method of least squares.
- b_0 and b_1 are obtained by finding the values that minimize the sum of the squared differences between Y and $\hat{Y}$ (finding the "**best line**"):

The **Sum of Squares of the Errors** about the estimated regression line

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2$$

The Simple Linear Regression Analysis

The estimates for β_0 *and* β_1:

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$$

OR

$$b_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

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

$$\bar{x} = \frac{\Sigma x_i}{n} \qquad \bar{y} = \frac{\Sigma y_i}{n}$$

Formulas and derivations are shown in the teaching materials for those who are interested

The Simple Linear Regression Analysis

Standard error of the estimate, S = $\sqrt{S^2}$

The estimates for $\sigma^2_{b_0}, \sigma^2_{b_1}$ and σ^2-

$$S^2_{b_0} = \frac{S^2 \sum x^2}{n\,SS_{xx}}$$

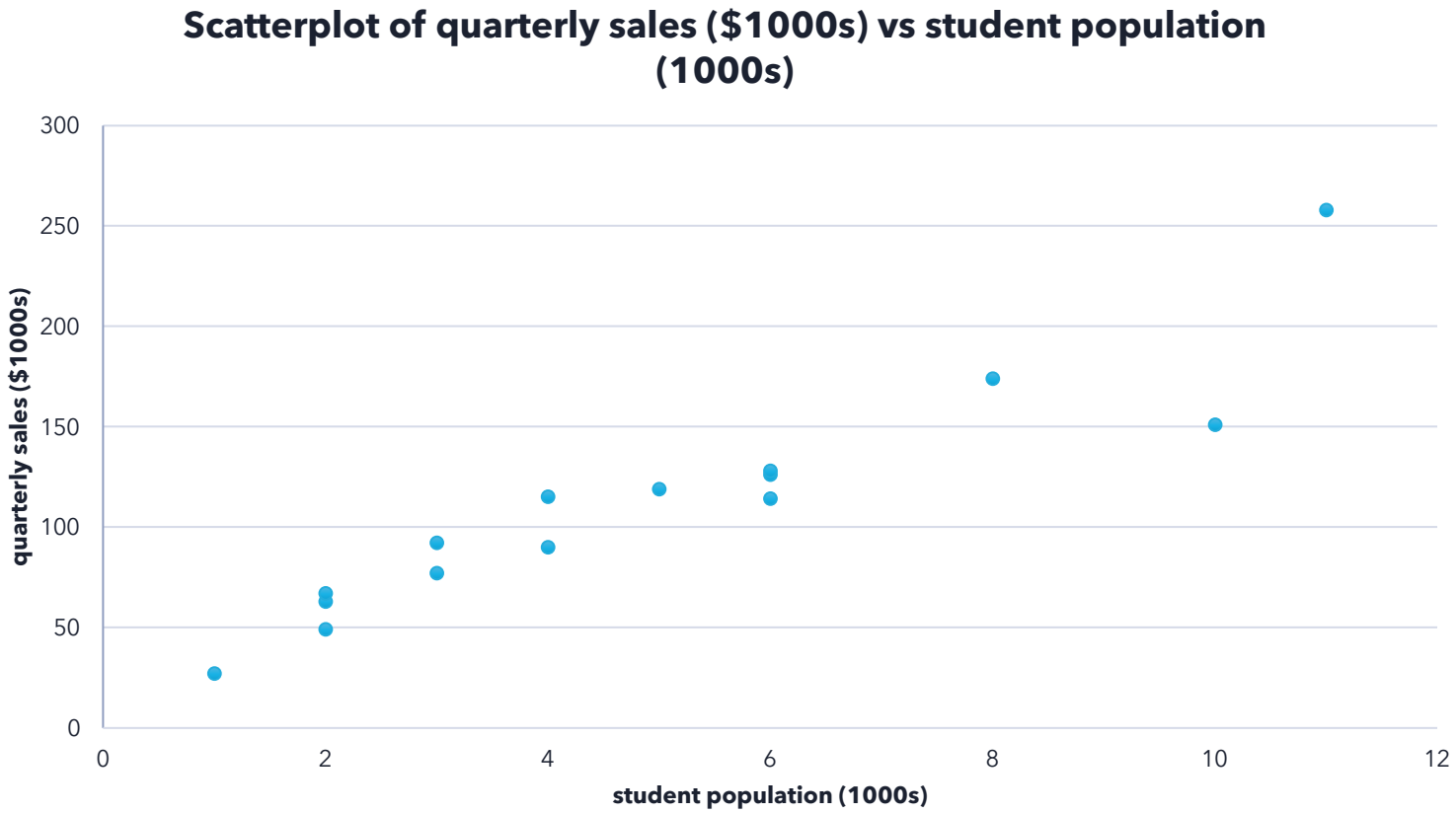
$$S^2_{b_1} = \frac{S^2}{SS_{xx}}$$

$$S^2 = \frac{SSE}{n-2} = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n-2} = \frac{SS_{yy} - b_1 SS_{xy}}{n-2}$$

- If standard error of estimate has a **small** value, this means that the regression line is **good** representative of the data.
- If standard error of estimate has a **large** value, this means that the regression line may **not** be **good** representative of the data.

Example

Restaurant	student population (1000s), x	quarterly sales ($1000s), y
1	3	92
2	2	63
3	6	126
4	8	174
5	2	49
6	4	90
7	5	119
8	6	114
9	2	67
10	4	115
11	6	128
12	11	258
13	3	77
14	10	151
15	1	27



Example: regression coefficients, $\mathbf{b_0}$ and $\mathbf{b_1}$

Restaurant	student population (1000s), x	quarterly sales ($1000s), y	x^2	y^2	xy
1	3	92	9	8464	276
2	2	63	4	3969	126
3	6	126	36	15876	756
4	8	174	64	30276	1392
5	2	49	4	2401	98
6	4	90	16	8100	360
7	5	119	25	14161	595
8	6	114	36	12996	684
9	2	67	4	4489	134
10	4	115	16	13225	460
11	6	128	36	16384	768
12	11	258	121	66564	2838
13	3	77	9	5929	231
14	10	151	100	22801	1510
15	1	27	1	729	27
Sum, Σ	**73**	**1,650**	**481**	**226,364**	**10,255**

r = + 0.9368

Tested the significant, $\boldsymbol{\rho}$

There is a linear relationship between x and y.

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$$

$$b_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

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

$$\bar{x} = \frac{\Sigma x_i}{n} \qquad \bar{y} = \frac{\Sigma y_i}{n}$$

Example: Estimated Simple Linear regression equation

Restaurant	student population (1000s), x	quarterly sales ($1000s), y	x^2	y^2	xy
Sum, Σ	73	1,650	481	226,364	10,255

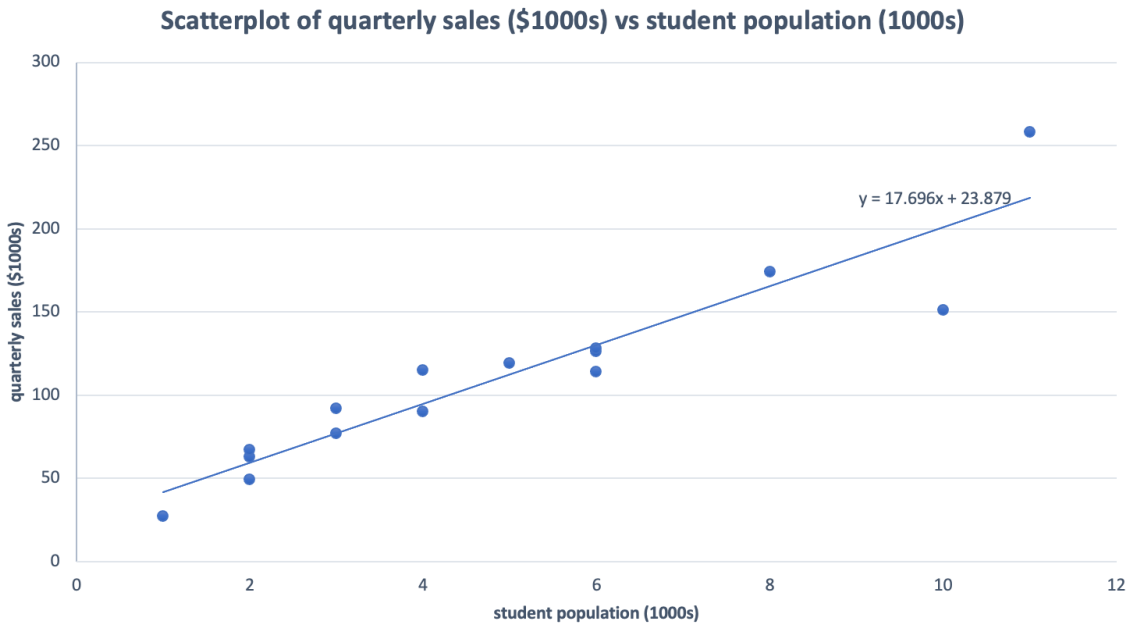
Let's compute $\mathbf{b_0}$ and $\mathbf{b_1}$

$$b_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$= \frac{15(10{,}255) - (73)(1{,}650)}{15(481) - (73)^2} = \frac{33{,}375}{1{,}886} = 17.696$$

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

$$= (1{,}650/15) - 17.696(73/15)$$
$$= 110 - 86.121$$
$$= 23.879$$



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

$$\hat{y} = \mathbf{23.879} + \mathbf{17.696}\, x$$

Estimated Simple Linear regression equation

Testing the Significance of the Population Regression Coefficients

Test!!

Testing the Significance of

- the *y*-intercept of the regression line, β_0
- the slope of the regression line, β_1

using **test statistics**, $\mathbf{T_0}$ and $\mathbf{T_1}$

$$y = \beta_0 + \beta_1 x + \varepsilon$$

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

Testing the Significance of the y-intercept of the regression line, β_0

The **hypotheses** for testing are

- H_0: $\beta_0 = 0$ (The true regression line passes through the origin)
- H_1: $\beta_0 \neq 0$ (The true regression line does not pass through the origin)

Test statistics

$$T_0 = \frac{b_0 - \beta_0}{S_{b_0}} \qquad \text{where} \quad S_{b_0} = \frac{S\sqrt{\sum x^2}}{\sqrt{n\,SS_{xx}}}$$

We will **reject** H_0, if $T_0 \leq -t_{\alpha/2}$ or $T_0 \geq +t_{\alpha/2}$ with *d.f.=n-2*

A **(1 - α)100%** **confidence interval** **for** β_0:

$$b_0 \pm (t_{\alpha/2,n-2} S_{b_0})$$

where:

b_0 = regression y-intercept coefficient

β_0 = hypothesized y-intercept

S_{b0} = standard error of the y-intercept

Testing the Significance of the slope of the regression line, β_1

The **hypotheses** for testing are

- H_0: $\beta_1 = 0$ (The independent variable *x* has no the effect on the dependent variable *y*)
- H_1: $\beta_1 \neq 0$ (The independent variable *x* has the effect on the dependent variable *y*)

Test statistics

$$T_1 = \frac{b_1 - \beta_1}{S_{b_1}} \qquad \text{where} \quad S_{b_1} = \frac{S}{\sqrt{SS_{xx}}}$$

where:

b_1 = regression slope coefficient

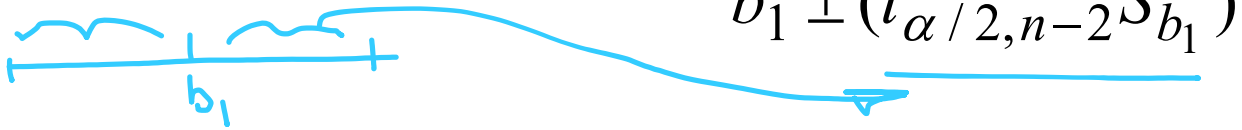
β_1 = hypothesized slope

S_{b1} = standard error of the slope

We will **reject** H_0, if $T_1 \leq -t_{\alpha/2}$ or $T_1 \geq +t_{\alpha/2}$ with *d.f.=n-2*

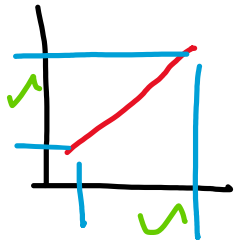
A **(1 - α)100%** **confidence interval** **for** β_1:

$$b_1 \pm (t_{\alpha/2,n-2} S_{b_1})$$



A Confidence Interval and A Prediction Interval (based on $\hat{y} = b_0 + b_1 x$)

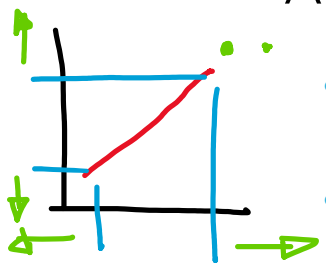
- A (1 - α)100% **confidence interval** for the mean value of *y*, *E(y)*, for a given x_0



  - Predicting for observations in the study sample
  - Uncertainty due to random error

$$\hat{y} \pm t_{\alpha/2,n-2}\, S \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

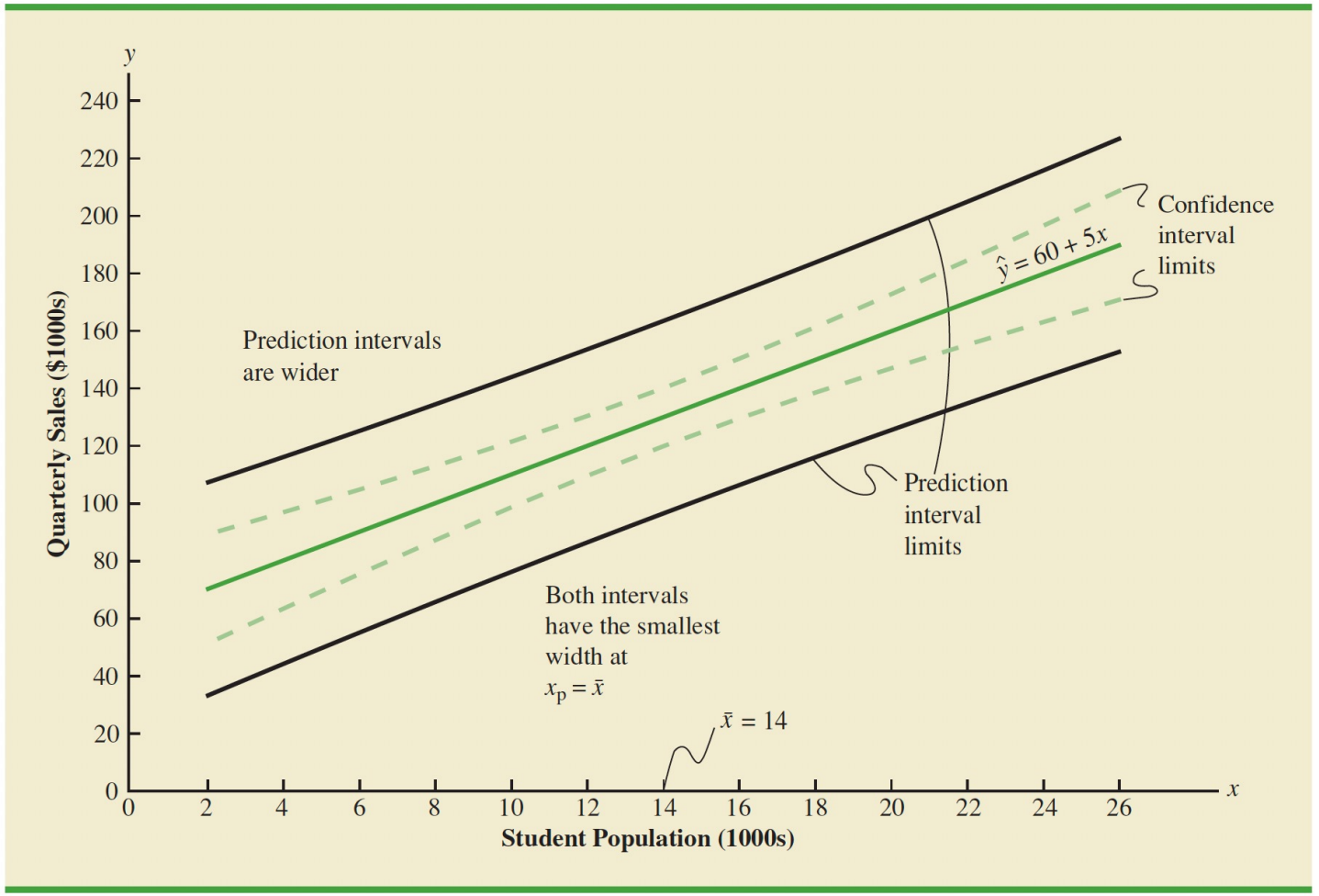
- A (1 - α)100% **prediction interval** for an individual value of *y* for a given x_0



  - Predicting for new observations, outside the study sample
  - Uncertainty due to random + true error (eg. Unmeasured risk factors)

$$\hat{y} \pm t_{\alpha/2,n-2}\, S \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

A Confidence Interval and A Prediction Interval



Constructing the best estimated regression line

H_0: β_0=0 ?	H_0: β_1=0 ?	Result
accept	accept	**can not find** the best estimated regression line
accept	reject	the best estimated regression line is $\hat{y} = b_1 x$
reject	accept	the best estimated regression line is $\hat{y} = b_0$
reject	reject	the best estimated regression line is $\hat{y} = b_0 + b_1 x$

Sums of Squares: Measures of Variation

Measure of **TOTAL** variation

Measure of **EXPLAINED** variation

Measure of **Unexplained** variation

$$SST = SSR + SSE$$

Total Sum of Squares

Regression Sum of Squares

Error Sum of Squares

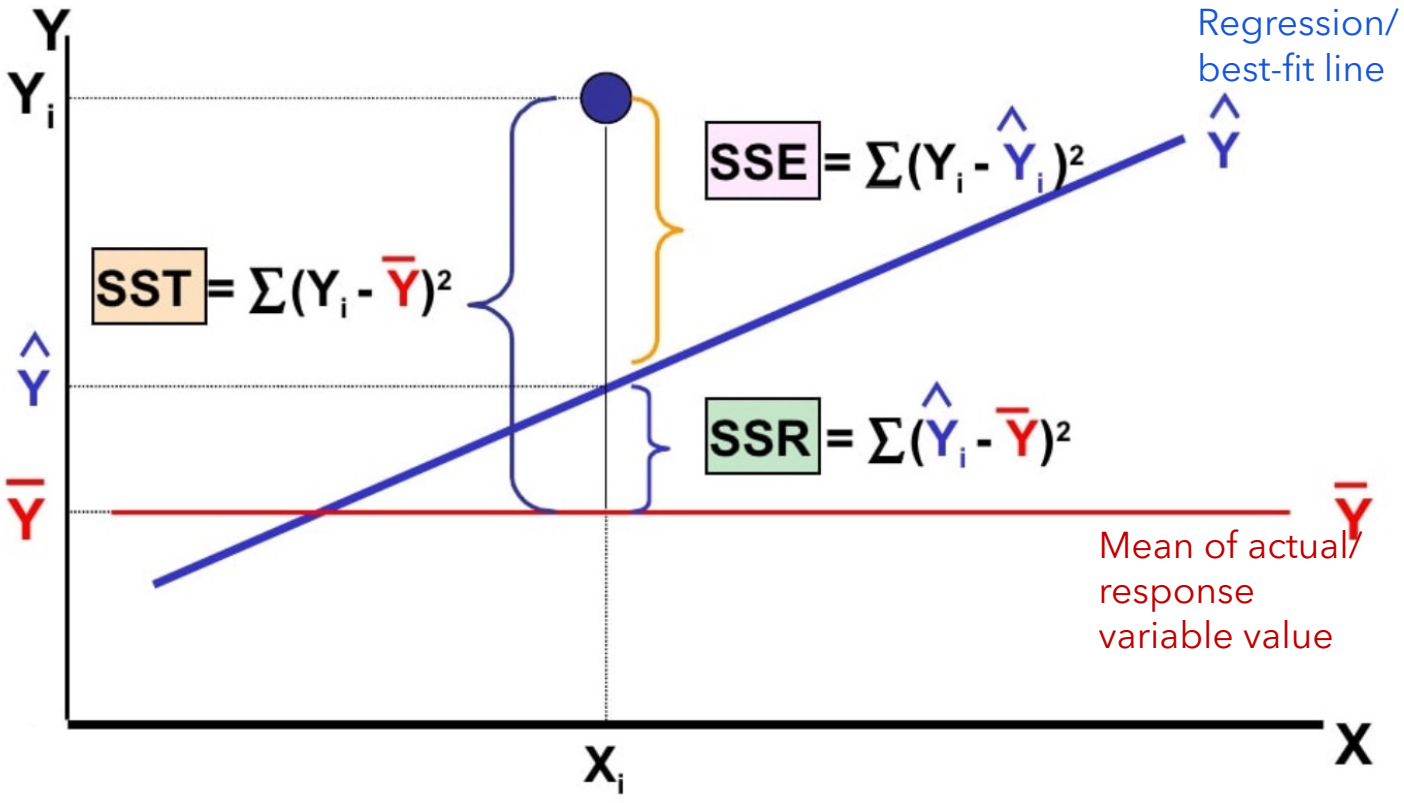
$$SST = \sum (Y_i - \bar{Y})^2 \quad SSR = \sum (\hat{Y}_i - \bar{Y})^2 \quad SSE = \sum (Y_i - \hat{Y}_i)^2$$

$\bar{Y}$ = Average value of the dependent variable

Y_i = Observed values of the dependent variable

$\hat{Y}_i$ = Predicted value of Y for the given X_i value

Measures of Variation



SST = total sum of squares

- Measures the variation of the Y_i values around their mean Y

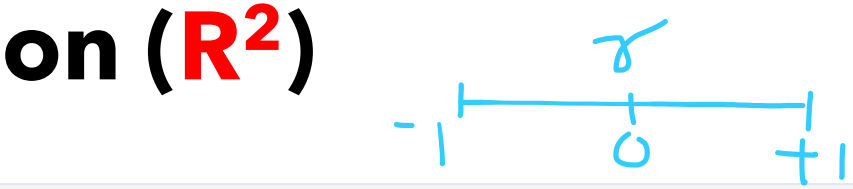
SSR = regression sum of squares

- Explained variation attributable to the relationship between X and Y

SSE = error sum of squares

- Variation attributable to factors other than the relationship between X and Y

A Measure of Quality of Fit: Coefficient of Determination (R^2)



$$-1 \leq r \leq +1$$

$$0 \leq r^2 \leq +1$$

Coefficient of Determination, R-square

the portion of the total variation in the dependent variable that is explained by variation in the independent variable

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}$$

$$0 \leq R^2 \leq 1$$

- If R^2 has a **large** value, this means that the dependent variable *y* is explained by the simple linear regression model quite **well.**
- For the simple linear regression, $R^2 = r^2$

Example: testing the significance of regression coefficients

$\hat{y} = 23.879 + 17.696\, x$

Test the significance of regression coefficients!!!

hypotheses

$$T_0 = \frac{b_0 - \beta_0}{S_{b_0}} \qquad (\beta_0 = 0)$$

$$T_1 = \frac{b_1 - \beta_1}{S_{b_1}} \qquad (\beta_1 = 0)$$

$$S_{b_0}^2 = \frac{S^2 \sum x^2}{n\, SS_{xx}}$$

$$S_{b_1}^2 = \frac{S^2}{SS_{xx}}$$

Example: testing the significance of regression coefficients

$$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 481 - \frac{(73)^2}{15} = 481 - 355.27 = 125.73$$

$$SS_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 226{,}364 - \frac{(1{,}650)^2}{15} = 226{,}364 - 181{,}500 = 44{,}864$$

$$SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 10{,}255 - \frac{(73)(1{,}650)}{15} = 10{,}255 - 8{,}030 = 2{,}225$$

$$S^2 = \frac{SS_{yy} - b_1 SS_{xy}}{n-2} = \frac{44{,}864 - (17.696)(2{,}225)}{15-2} = \frac{5{,}940.4}{13} = 422.338$$

S = $\sqrt{S^2} = \sqrt{422.338^2} = 20.5509$

Example: testing the significance of regression coefficients

$\hat{y} = 23.879 + 17.696\, x$

Test the significance of regression coefficient!!!

$$S_{b_0} = \frac{S\sqrt{\sum x^2}}{\sqrt{n\,SS_{xx}}} = \frac{20.5509\sqrt{481}}{\sqrt{15(125.73)}} = 10.379$$

$$T_0 = \frac{b_0 - \beta_0}{S_{b_0}} = \frac{23.879 - 0}{10.379} = 2.3007$$

$$S_{b_1} = \frac{S}{\sqrt{SS_{xx}}} = \frac{20.5509}{\sqrt{125.73}} = 1.833$$

$$T_1 = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{17.696 - 0}{1.833} = 9.654$$

Example: testing the significance of regression coefficients

$$\hat{y} = 23.879 + 17.696\,x$$

Hypothesis test:

$$y = \beta_0 + \beta_1 x$$

y-intercept

$H_0: \beta_0 = 0$

$H_1: \beta_0 \neq 0$

slope

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

Test statistics

T_0=2.3007

With $\alpha = 0.05$ and $df. = 15 - 2 = 13$

T_1=9.654

$T_0 \geq +t_{\alpha/2}(2.160)$

H_0 is **rejected**

reject	reject	the best estimated regression line is $\hat{y} = b_0 + b_1 x$

$T_1 \geq +t_{\alpha/2}(2.160)$

H_1 is **rejected**

Example: Simple Linear Regression Equation

$$\hat{y} = 23.879 + 17.696\,x$$

Simple Linear Regression Equation

If student population is 16,000, what will be the quarterly sales?

$\hat{y} = 23.879 + 17.696\,x$

$\hat{y} = 23.879 + 17.696\,(16)$

$\hat{y} = 307.015$

Example: measure the Quality of Fit **(R^2)**

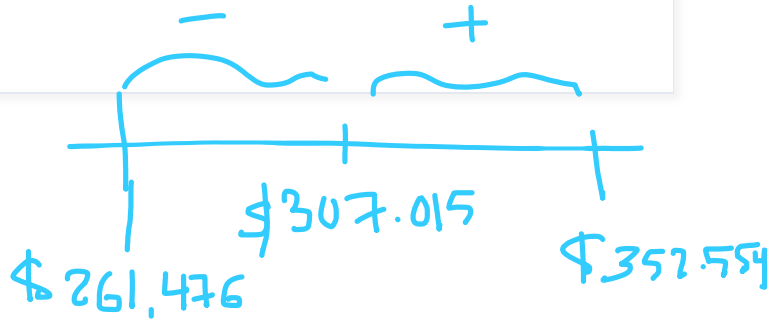
Coefficient of Determination (R^2)

- the simple linear regression, $\mathbf{R^2 = r^2}$ $= (0.9368)^2 =$ **0.8776**

  $r = +0.9368$ or **87.76%**

- 87.76 % of the variation in the quarterly sales is explained by the student
- 12.24 % of the variation in the quarterly sales cannot be explained by the variation in the student population population

Example: confidence interval



confidence interval – 95% for $\boldsymbol{\beta_0}$ *and* $\boldsymbol{\beta_1}$

$$b_0 \pm (t_{\alpha/2,n-2} S_{b_0}) = 23.879 \pm 2.16(10.379) = 23.879 \pm 22.419 = (1.46, 46.298)$$

$$b_1 \pm (t_{\alpha/2,n-2} S_{b_1}) = 17.696 \pm 2.16(1.833) = 17.696 \pm 3.959 = (13.737, 21.655)$$

$$\hat{y} \pm t_{\alpha/2,n-2}\, S \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}} = 307.015 \pm 2.16(20.5509)\sqrt{\frac{1}{15} + \frac{(16 - 4.867)^2}{125.73}}$$

$= 307.015 \pm 2.16(21.083)$
$= 307.015 \pm 45.539$
$= (261.476, 352.554)$

95% confidence interval for the mean quarterly sales when the student population is 16,000 students is $261,476 to $352,554.

Example: prediction interval

prediction interval

$$\hat{y} \pm t_{\alpha/2,n-2}\, S \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}} = 307.015 \pm 2.16(20.5509)\sqrt{1 + \frac{1}{15} + \frac{(16 - 4.867)^2}{125.73}}$$

$$= 307.015 \pm 2.16(29.442)$$
$$= 307.015 \pm 63.595$$
$$= (243.420, 370.610)$$

the 95% prediction interval for the quarterly sales when the student population is 16,000 students is $243,420 to $370,610 .

Recap and warm-up exercise

The advertising expenditure x and sales y in thousands of dollars for a small retail business in its first eight years in operation are shown in the table.

x	1.4	1.6	1.6	2.0	2.0	2.2	2.4	2.6
y	180	184	190	220	186	215	205	240

(a) Compute the linear correlation coefficient, **r** for these sample data and interpret its meaning in the context of the problem.

(b) Test the significant of $\boldsymbol{\rho}$

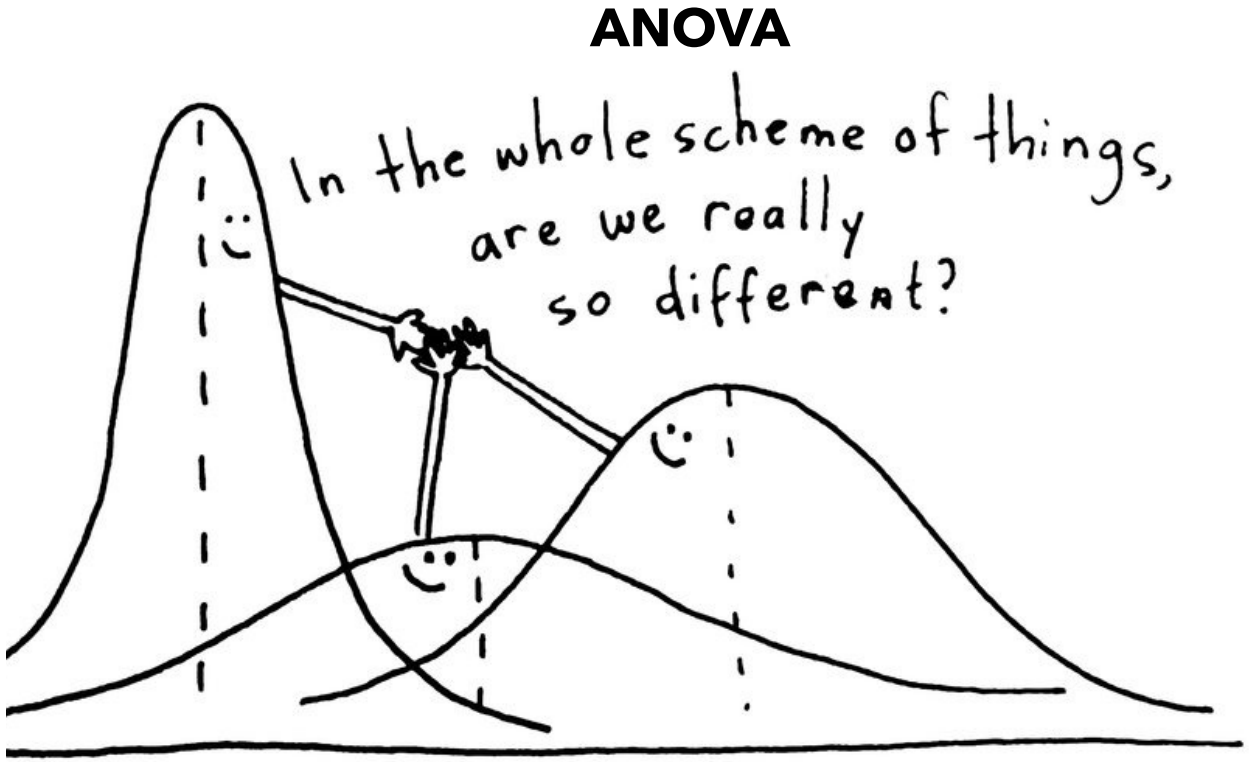
(c) Obtain the estimated linear regression equation (Hence find b_1 and b_0)

(d) Test the significance of regression coefficient. (Hence make an assumption about the appropriate model.)

(e) Compute the confidence interval for β_1 and β_0

(f) Compute the confidence interval for predicted y.

Analysis of Variance F Test for the Model



- Test whether the estimated regression line $\hat{y} = b_0 + b_1 x$ fit better than $\hat{y} = \bar{y}$
- The hypotheses
  - H_0: $\beta_1 = 0$ (The equation $\hat{y} = b_0 + b_1 x$ **cannot** be provided a best fit to the data)
  - H_1: $\beta_1 \neq 0$ (The equation $\hat{y} = b_0 + b_1 x$ **can** be provided a best fit to the data)

F Test for Significance

F Test statistic:

$$F_{STAT} = \frac{MSR}{MSE}$$

where

$$MSR = \frac{SSR}{k}$$

$$MSE = \frac{SSE}{n - k - 1}$$

reject H_0, if $F \geq F_\alpha$ with *d.f.* = *(1, n-2)*

where F_{STAT} follows an F distribution with k numerator and (n – k - 1) denominator degrees of freedom

(k = the number of independent variables in the regression model)

F Test for Significance

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	p-value
Regression	SSR	1	$\text{MSR} = \frac{\text{SSR}}{1}$	$F = \frac{\text{MSR}}{\text{MSE}}$	
Error	SSE	$n - 2$	$\text{MSE} = \frac{\text{SSE}}{n-2}$		
Total	SST	$n - 1$			

If H_0: $\beta_1 = 0$ is rejected, then the conclusion is that x and y are linearly related. In other words, the line $\hat{y} = b_0 + b_1 x$ provides the best fit to the data.

F Test for Significance *(example)*

$H_0: \beta_1 = 0$

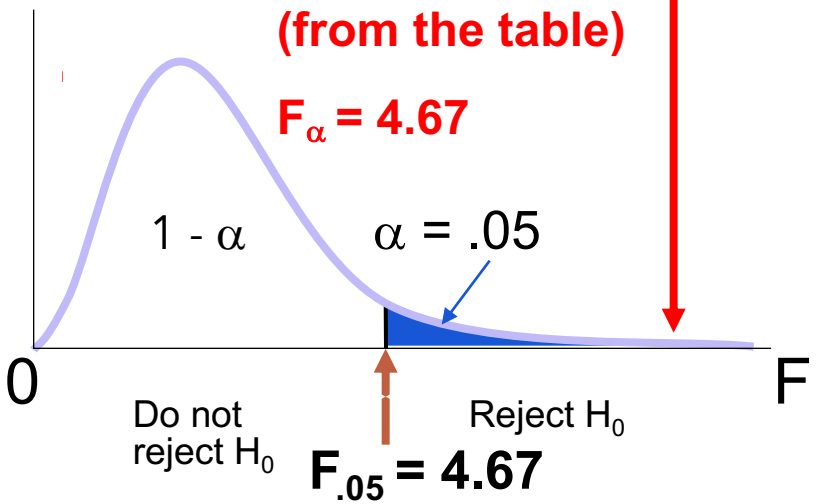
$H_1: \beta_1 \neq 0$

$\alpha = .05$

$df_1 = 1 \quad df_2 = 13$

Critical Value: (from the table)

$F_\alpha = 4.67$



Test Statistic:

$$F_{STAT} = \frac{MSR}{MSE} = 93.24$$

Decision:

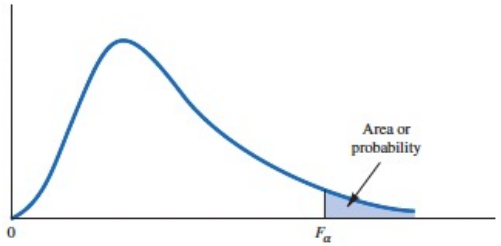
Reject H_0 at $\alpha = 0.05$

Conclusion:

There is sufficient/strong evidence that quarterly sales and the student population are significantly related.

F Distribution Table

TABLE 4 *F* DISTRIBUTION



Entries in the table give F_α values, where α is the area or probability in the upper tail of the F distribution. For example, with 4 numerator degrees of freedom, 8 denominator degrees of freedom, and a .05 area in the upper tail, $F_{.05} = 3.84$.

Denominator Degrees of Freedom	Area in Upper Tail	Numerator Degrees of Freedom																	
		1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	60	100	1000
1	.10	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	61.22	61.74	62.05	62.26	62.53	62.79	63.01	63.30
	.05	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	245.95	248.02	249.26	250.10	251.14	252.20	253.04	254.19
	.025	647.79	799.48	864.15	899.60	921.83	937.11	948.20	956.64	963.28	968.63	984.87	993.08	998.09	1001.40	1005.60	1009.79	1013.16	1017.76
	.01	4052.18	4999.34	5403.53	5624.26	5763.96	5858.95	5928.33	5980.95	6022.40	6055.93	6156.97	6208.66	6239.86	6260.35	6286.43	6312.97	6333.92	6362.80
2	.10	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
	.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.45	19.46	19.46	19.47	19.48	19.49	19.49
	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
	.01	98.50	99.00	99.16	99.25	99.30	99.33	99.36	99.38	99.39	99.40	99.43	99.45	99.46	99.47	99.48	99.48	99.49	99.50
3	.10	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.20	5.18	5.17	5.17	5.16	5.15	5.14	5.13
	.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	8.63	8.62	8.59	8.57	8.55	8.53
	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.25	14.17	14.12	14.08	14.04	13.99	13.96	13.91
	.01	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	26.87	26.69	26.58	26.50	26.41	26.32	26.24	26.14
4	.10	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
	.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
	.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.66	8.56	8.50	8.46	8.41	8.36	8.32	8.26
	.01	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.20	14.02	13.91	13.84	13.75	13.65	13.58	13.47
5	.10	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.324	3.21	3.19	3.17	3.16	3.14	3.13	3.11
	.05	6.61.	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.52	4.50	4.46	4.43	4.41	4.37
	.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.43	6.33	6.27	6.23	6.18	6.12	6.08	6.02
	.01	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.72	9.55	9.45	9.38	9.29	9.20	9.13	9.03

TABLE 4 *F* DISTRIBUTION (*Continued*)

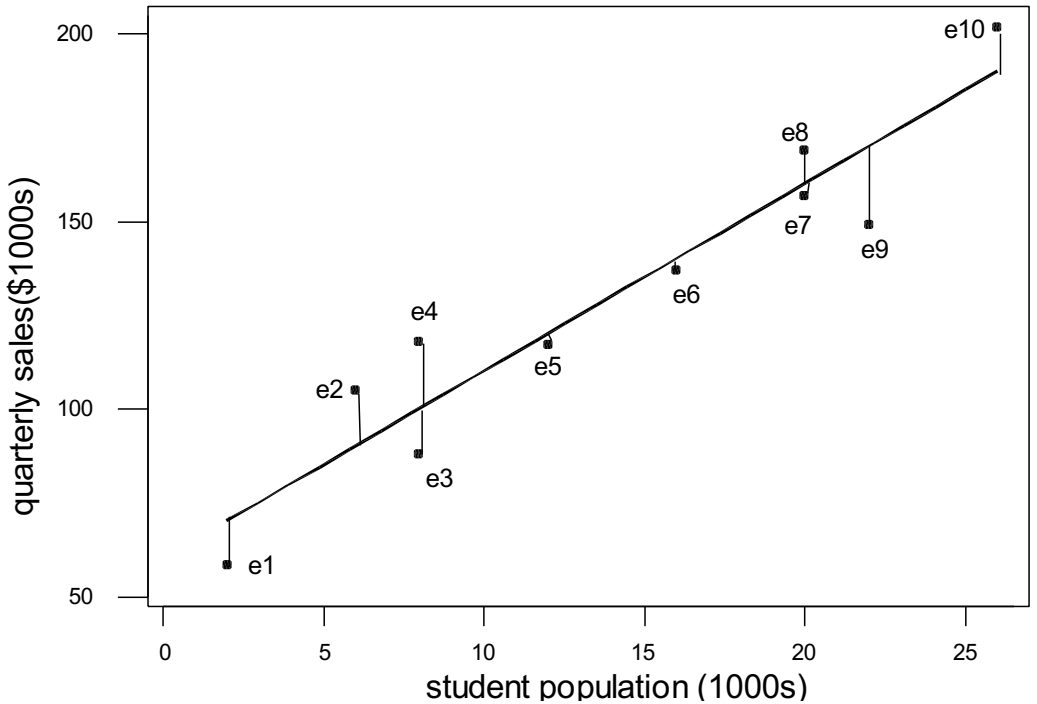
Denominator Degrees of Freedom	Area in Upper Tail	Numerator Degrees of Freedom																	
		1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	60	100	1000
6	.10	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.87	2.84	2.81	2.80	2.78	2.76	2.75	2.72
	.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	3.83	3.81	3.77	3.74	3.71	3.67
	.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.27	5.17	5.11	5.07	5.01	4.96	4.92	4.86
	.01	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.56	7.40	7.30	7.23	7.14	7.06	6.99	6.89
7	.10	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.63	2.59	2.57	2.56	2.54	2.51	2.50	2.47
	.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44	3.40	3.38	3.34	3.30	3.27	3.23
	.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.57	4.47	4.40	4.36	4.31	4.25	4.21	4.15
	.01	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.31	6.16	6.06	5.99	5.91	5.82	5.75	5.66
8	.10	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.30
	.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15	3.11	3.08	3.04	3.01	2.97	2.93
	.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.10	4.00	3.94	3.89	3.84	3.78	3.74	3.68
	.01	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.52	5.36	5.26	5.20	5.12	5.03	4.96	4.87
9	.10	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.34	2.30	2.27	2.25	2.23	2.21	2.19	2.16
	.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.89	2.86	2.83	2.79	2.76	2.71
	.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.77	3.67	3.60	3.56	3.51	3.45	3.40	3.34
	.01	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	4.96	4.81	4.71	4.65	4.57	4.48	4.41	4.32
10	.10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.24	2.20	2.17	2.16	2.13	2.11	2.09	2.06
	.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.73	2.70	2.66	2.62	2.59	2.54
	.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.52	3.42	3.35	3.31	3.26	3.20	3.15	3.09
	.01	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.56	4.41	4.31	4.25	4.17	4.08	4.01	3.92
11	.10	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.17	2.12	2.10	2.08	2.05	2.03	2.01	1.98
	.05	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.65	2.60	2.57	2.53	2.49	2.46	2.41
	.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.33	3.23	3.16	3.12	3.06	3.00	2.96	2.89
	.01	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.25	4.10	4.01	3.94	3.86	3.78	3.71	3.61
12	.10	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.10	2.06	2.03	2.01	1.99	1.96	1.94	1.91
	.05	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.54	2.50	2.47	2.43	2.38	2.35	2.30
	.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.18	3.07	3.01	2.96	2.91	2.85	2.80	2.73
	.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.01	3.86	3.76	3.70	3.62	3.54	3.47	3.37
13	.10	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
	.05	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.46	2.41	2.38	2.34	2.30	2.26	2.21
	.025	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.05	2.95	2.88	2.84	2.78	2.72	2.67	2.60
	.01	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.82	3.66	3.57	3.51	3.43	3.34	3.27	3.18
14	.10	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.01	1.96	1.93	1.99	1.89	1.86	1.83	1.80
	.05	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.39	2.34	2.31	2.27	2.22	2.19	2.14
	.025	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	2.95	2.84	2.78	2.73	2.67	2.61	2.56	2.50
	.01	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.66	3.51	3.41	3.35	3.27	3.18	3.11	3.02
15	.10	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	1.97	1.92	1.89	1.87	1.85	1.82	1.79	1.76
	.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.28	2.25	2.20	2.16	2.12	2.07
	.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.86	2.76	2.69	2.64	2.59	2.52	2.47	2.40
	.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.52	3.37	3.28	3.21	3.13	3.05	2.98	2.88

Residual Analysis: Validating Model Assumptions

$$y = \beta_0 + \beta_1 x + \varepsilon$$

residual or error

$$e_i = y_i - \hat{y}_i$$



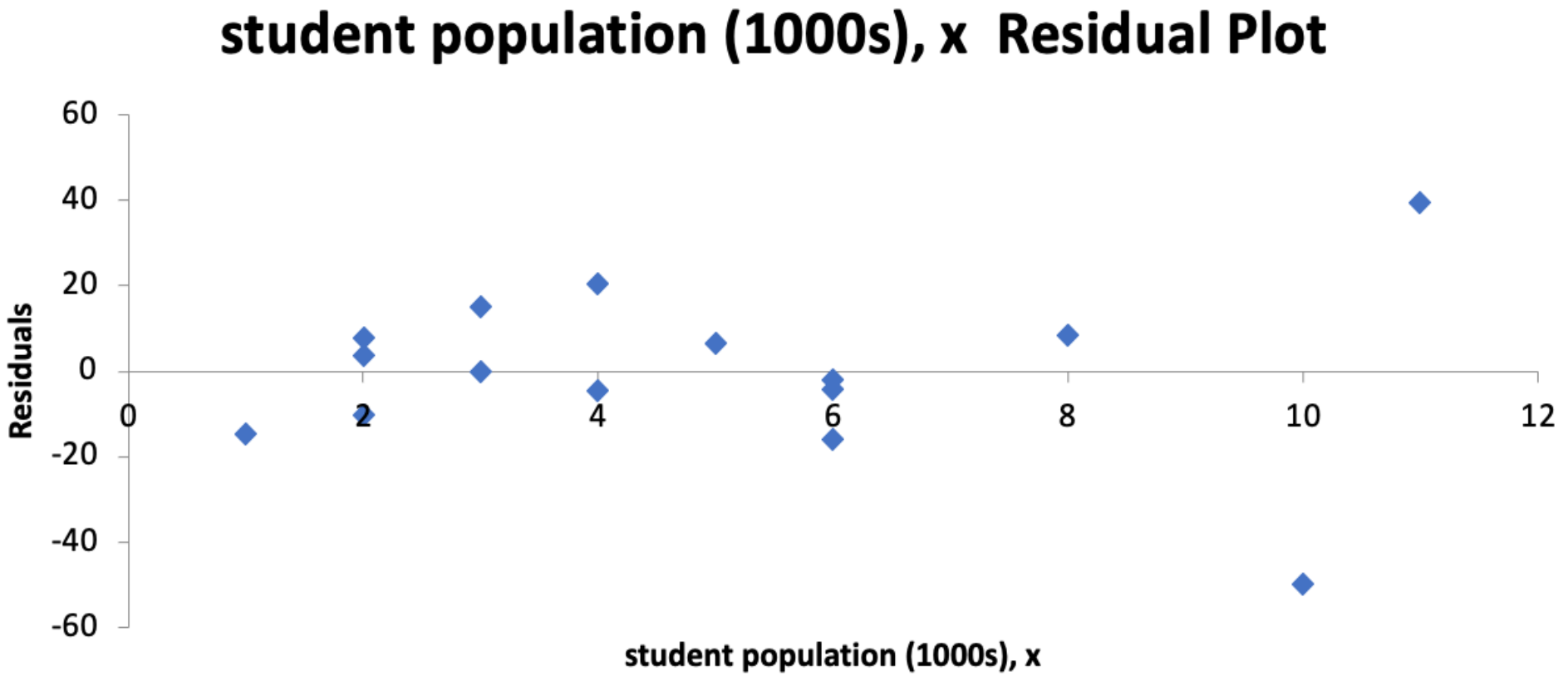
- The error term, ε is a random variable with a **mean of zero**; $\mathrm{E}(\varepsilon) = 0$
- The variance of ε, σ_e^2, is the same for all values of x. It is called that the **constant variance assumption**.
- Each value of ε are **independent** or ε_i and ε_j are independent $i \neq j$
- The error term ε is a **normally distributed random variable**

Residual Plots

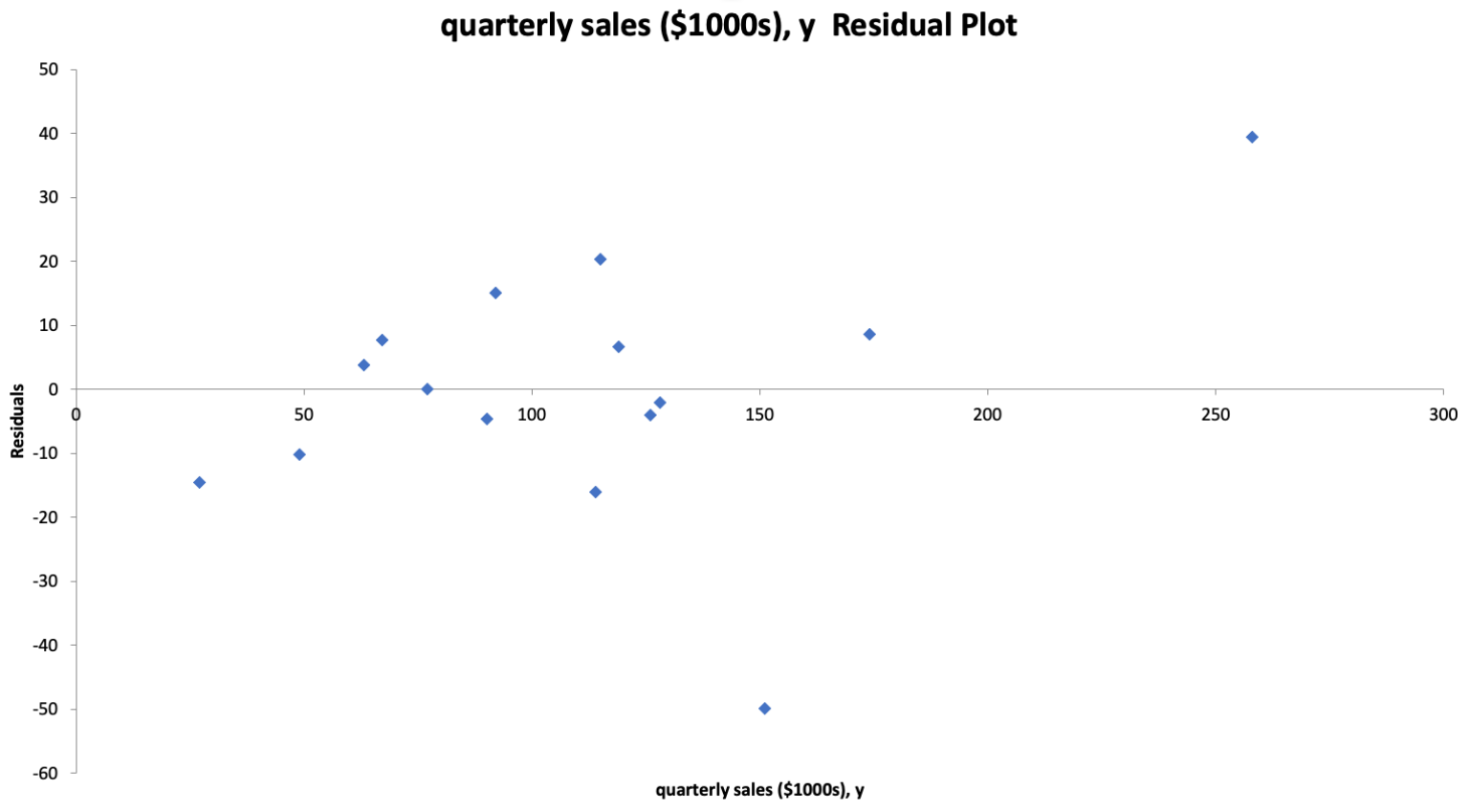
1. Residual Plot Against $\mathbf{x}$
2. Residual Plot Against $\hat{y}$
3. Standardized Residuals
4. Normal Probability Plot

} validity of the constant variance assumption

Residual Plots: against **x**

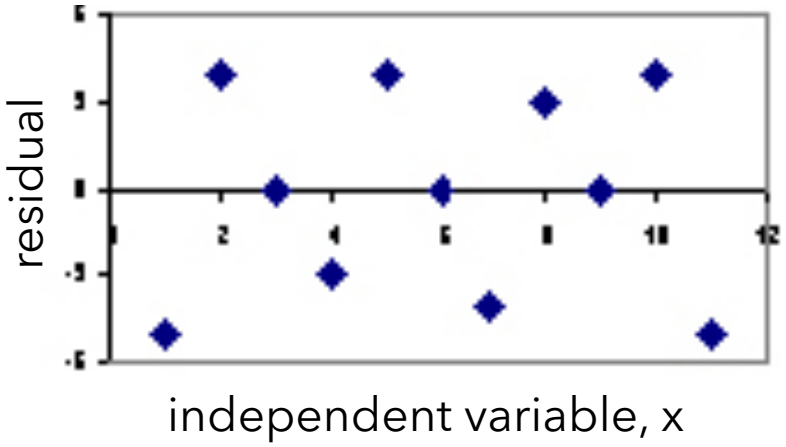


Residual Plots: against $\hat{y}$

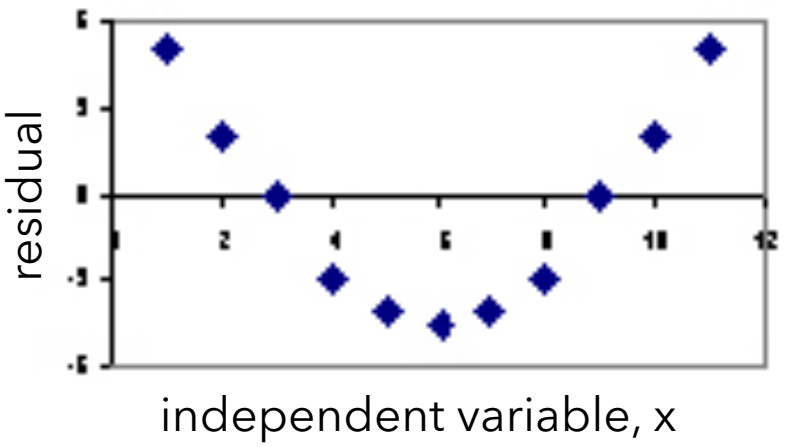


Residual Plot

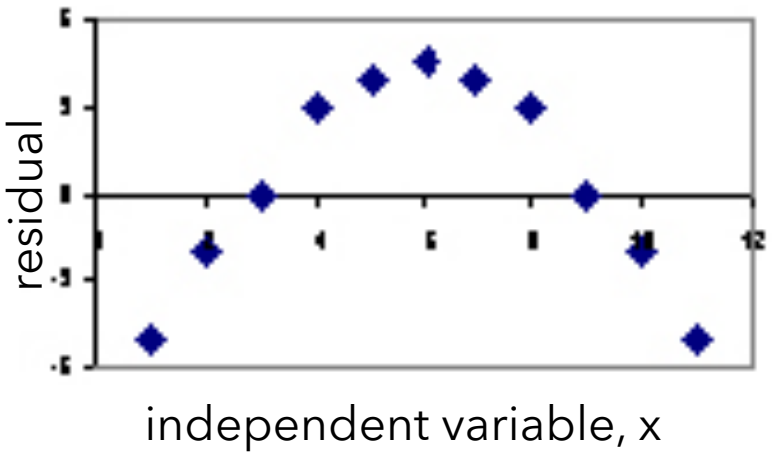
Random pattern



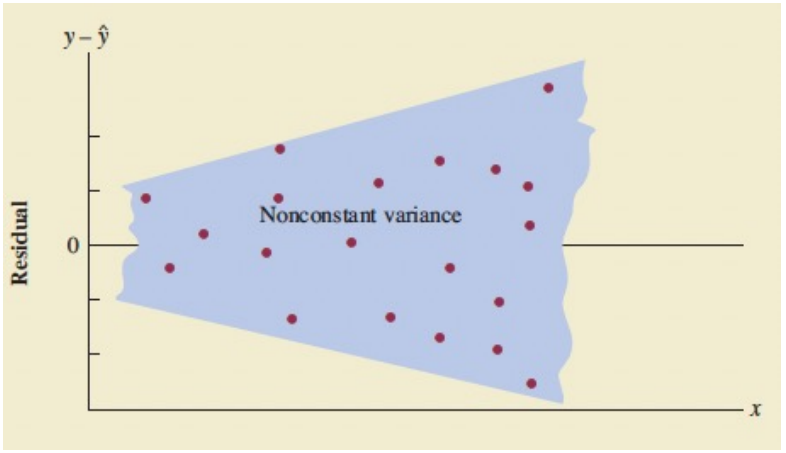
Non-random: U-shaped curve

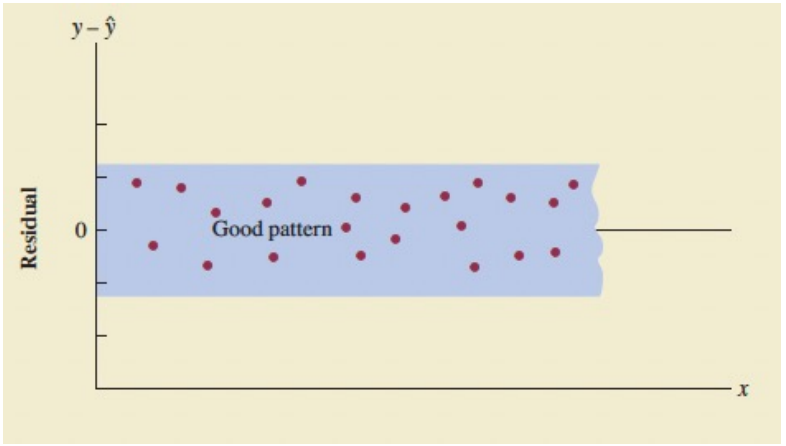


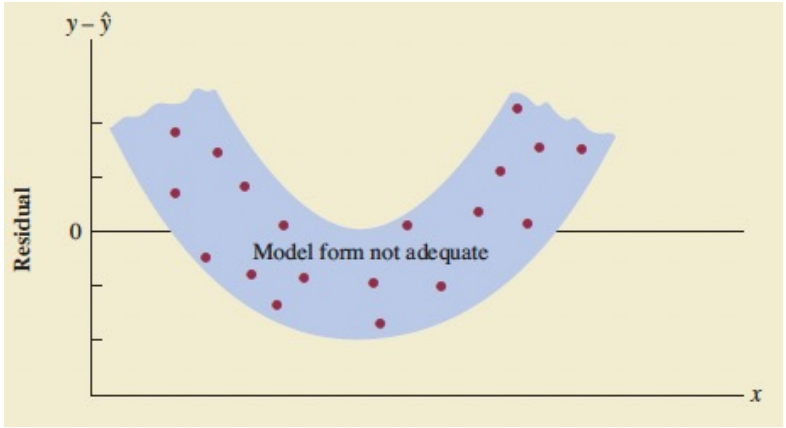
Non-random: Inverted U



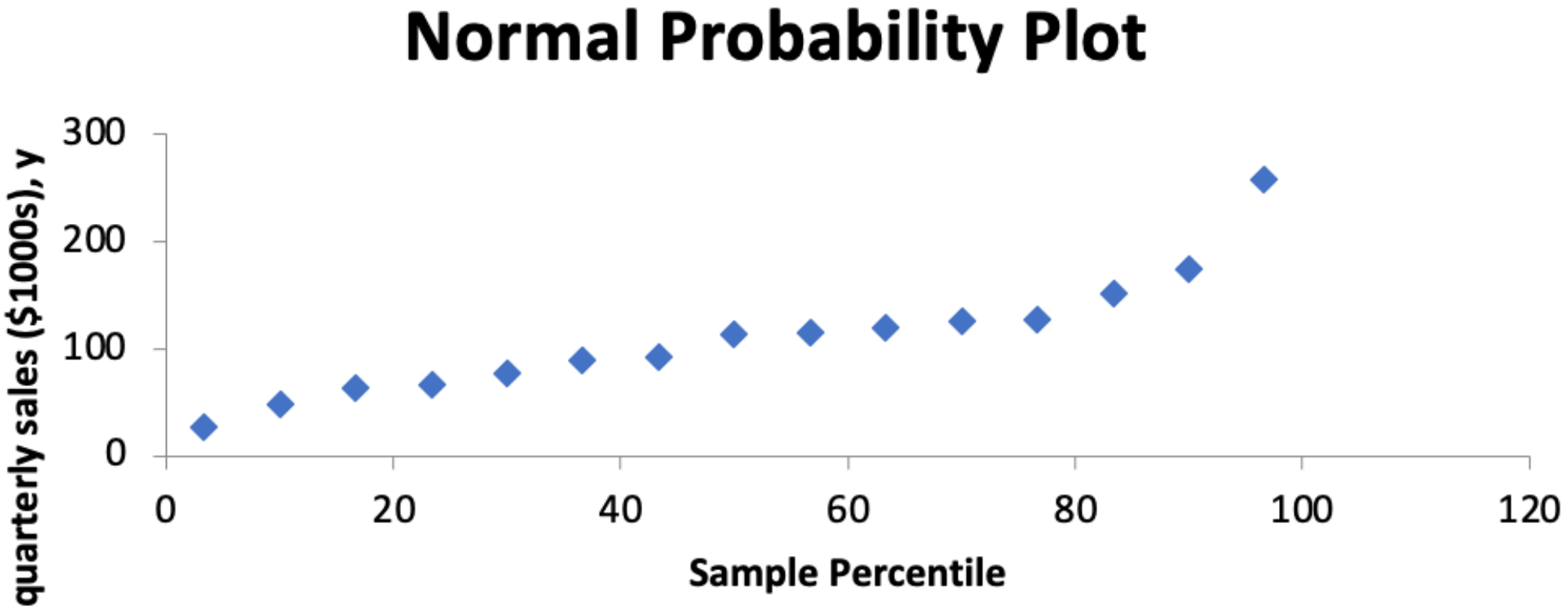
Residual Plots: standardized residuals







Residual Plots: Normal Probability Plot



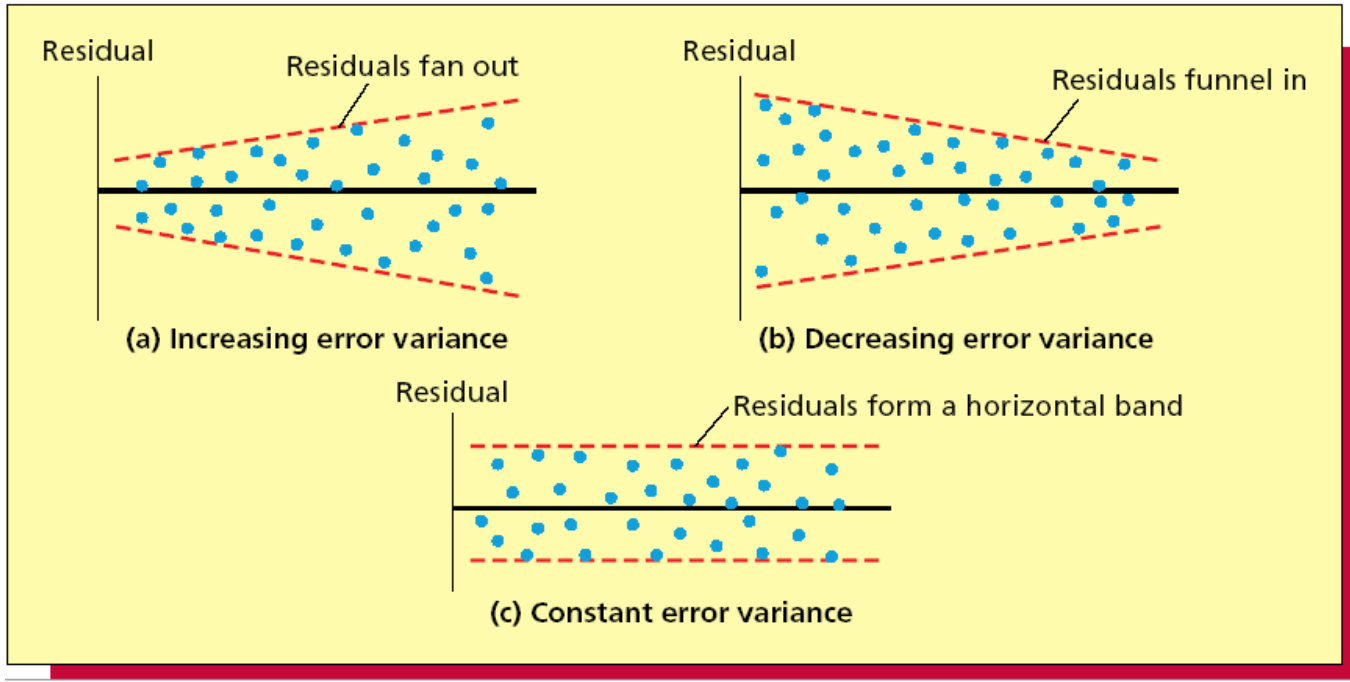
Residual Analysis: Validating Model Assumptions

Constant Variance Assumptions

- The *x* values
- The predicted *y* values
- Time (when data is time series)

- Constant error variance
  - Residuals fluctuates around their mean of zero
  - No pattern
- Increasing error variance
  - "fans out" - spread out
  - Constant variance assumption is violated
- Decreasing error variance
  - "funnels in" - spread out
  - Constant variance assumption is violated

Assumption of Correct Functional Form



Residual Analysis: Validating Model Assumptions

Normal Probability Plot of the residuals

- Bell-shaped and symmetric
- Kolmogorov-Smirnov test (KS)
  - The **hypotheses** are

    H_0: The distribution of the residuals are normally distributed.

    H_1: The distribution of the residuals are not normally distributed.

    Reject H_0 if the ρ-value is $\leq \alpha$

Residual Analysis: Validating Model Assumptions

Independence Assumption

- **Durbin-Watson** test statistic,

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e^2}$$

where

d = Durbin-Watson statistic
e = residual
t = time period counter

- The **hypotheses** are

  H_0: There is no autocorrelation or the values of errors are independent.

  H_1: There is autocorrelation or the values of errors are not independent.

Residual Analysis: Validating Model Assumptions

Independence Assumption

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e^2}$$

- The statistic (d) has a range of from 0 to 4, with a midpoint of 2

Regions of Acceptance and Rejection of the Null Hypothesis at α				
If $d < d_L$, H_0 can be rejected. It means that the error terms are positively autocorrelated.	If $d_L \le d \le d_U$, the test is inconclusive.	If $d_U < d < 4 - d_U$, H_0 can be accepted. It means that the error terms are not autocorrelated.	If $4 - d_U \le d \le 4 - d_L$, the test is inconclusive.	If $4 - d_L < d < 4$, H_0 can be rejected. It means that the error terms are negatively autocorrelated.

Residual Analysis: Validating Model Assumptions

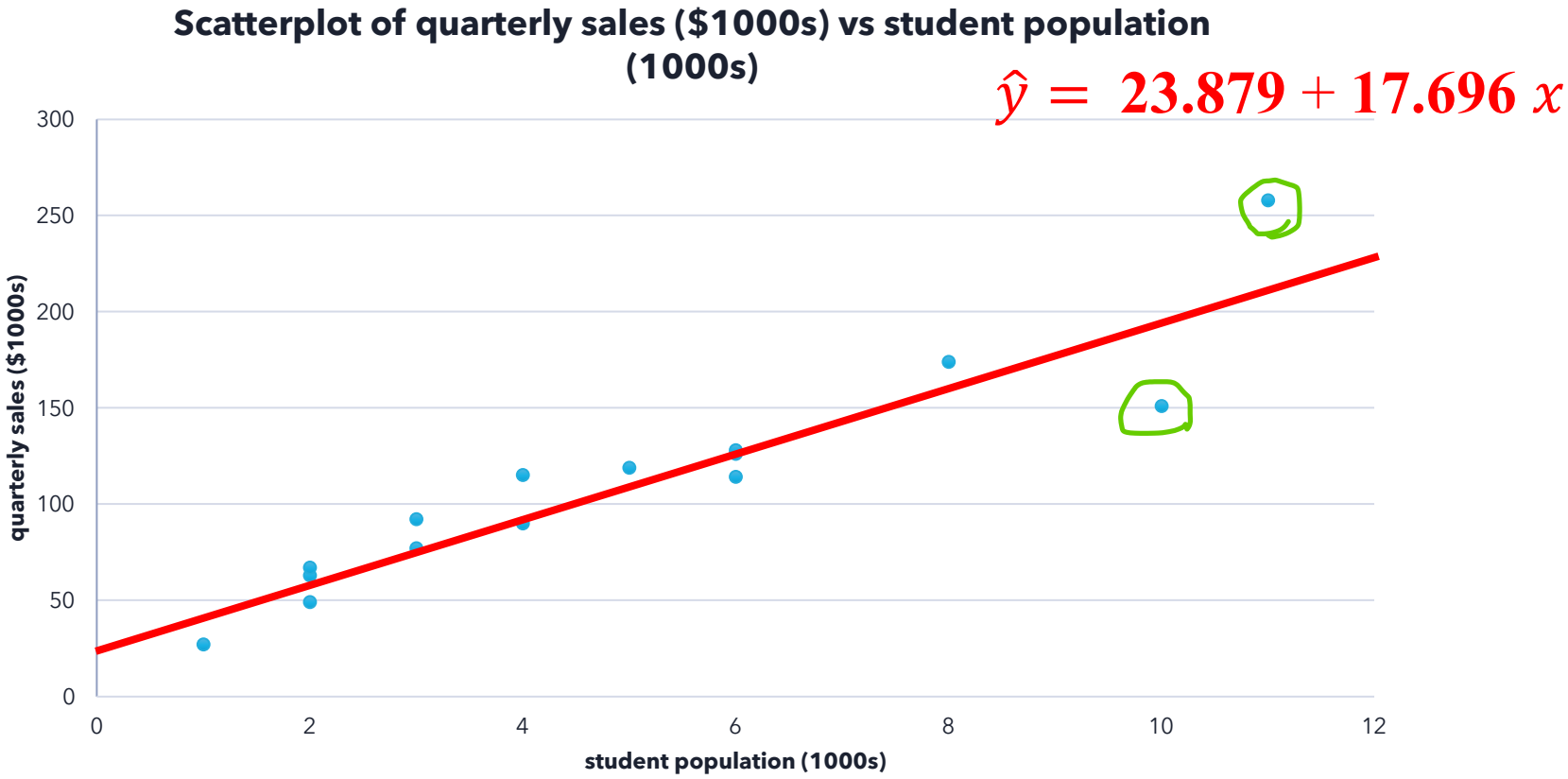
The Critical Values for the Durbin-Watson Statistic ($\alpha = 0.05$)

n	$k=1$		$k=2$		$k=3$		$k=4$		$k=5$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	0.95	1.53	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.53	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.54	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.54	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

Remarks: n is the total observations and k is the number of independent variable(s)

Example: Residual Analysis

Restaurant	student population (1000s), x	quarterly sales ($1000s), y
1	3	92
2	2	63
3	6	126
4	8	174
5	2	49
6	4	90
7	5	119
8	6	114
9	2	67
10	4	115
11	6	128
12	11	258
13	3	77
14	10	151
15	1	27



Example: Residual Analysis

$$y = \beta_0 + \beta_1 x + \varepsilon$$

$$\hat{y} = 23.879 + 17.696\, x$$

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	39374	39374	93.24	0.000
Residual Error	13	5490	422		
Total	14	44864			

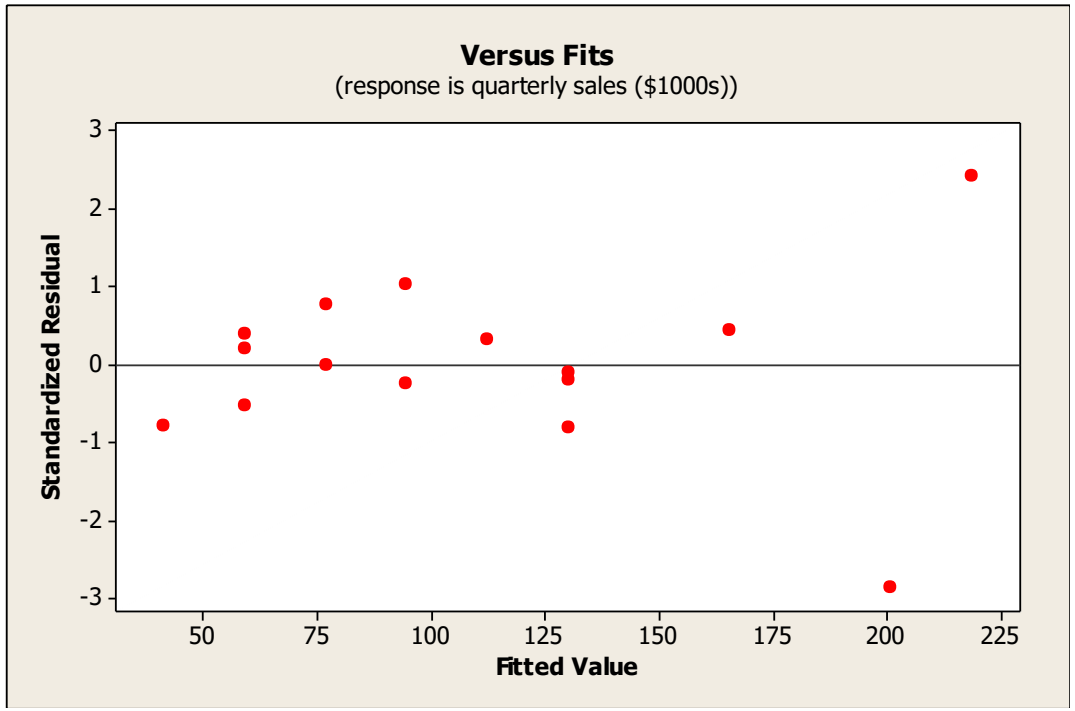
Unusual Observations

Obs	student population (1000s)	quarterly sales ($1000s)	Fit	SE Fit	Residual	St Resid
12	11.0	258.00	218.54	12.43	39.46	2.41R
14	10.0	151.00	200.84	10.80	-49.84	-2.85R

R denotes an observation with a large standardized residual.

Durbin-Watson statistic = 1.75076

Example: Residual Analysis

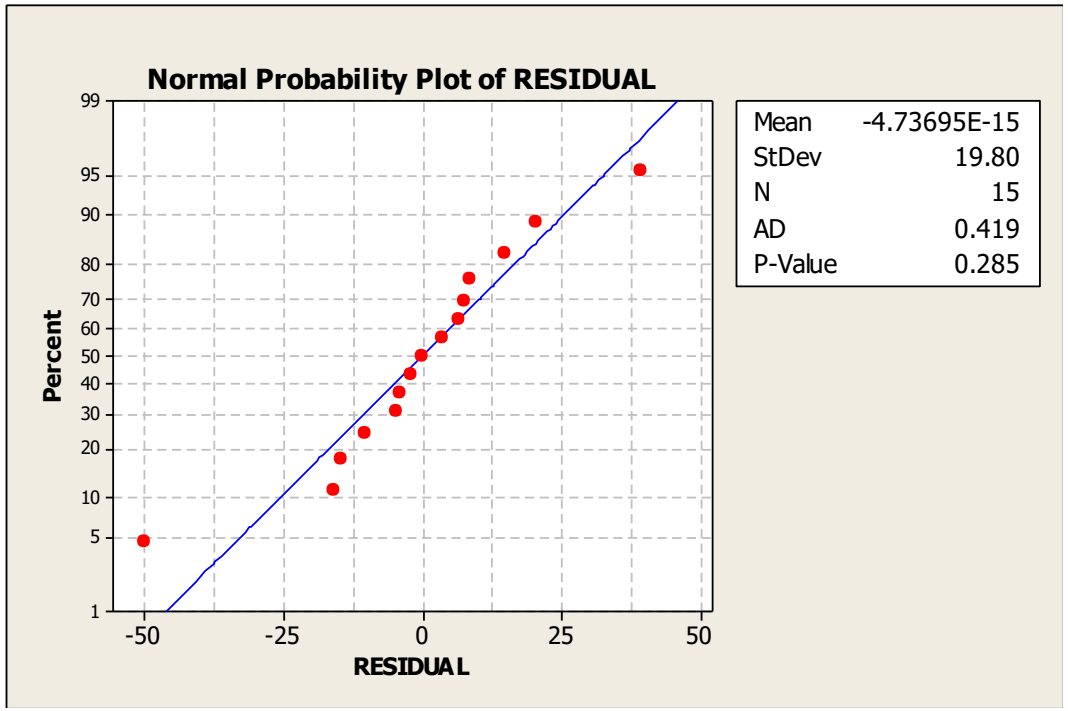


Constant Variance Assumptions

do not fluctuate around their mean of zero

- **constant variance assumption is violated**

Example: Residual Analysis



Normal Probability Plot of the residuals

Mean of ε = -4.73695 x 10^{-15}

- $\mathbf{E(\varepsilon) = 0}$

p-value = 0.285 >α = 0.05

- **Fail to reject H_0**
- ***residuals distribution are normally distributed***

Example: Residual Analysis

n	$k=1$		$k=2$		$k=3$		$k=4$		$k=5$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	0.95	1.53	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.53	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.54	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.54	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

Remarks: n is the total observations and k is the number of independent variable(s)

Independence Assumption

- Durbin-Watson statistic, d = 1.75076

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e^2}$$

Regions of Acceptance and Rejection of the Null Hypothesis				
Reject H_0, It has the positive autocorrelation.	The test is inconclusive.	Accept H_0: There is no autocorrelation.	The test is inconclusive.	**Reject H_0, It has the negative autocorrelation.**

0 $d_L = 1.08$ $d_U = 1.36$ $4 - d_U = 2.64$ $4 - d_L = 2.92$ 4

Independence assumption is valid.

Example: Validating Model Assumptions

$$y = \beta_0 + \beta_1 x + \varepsilon$$

$$\hat{y} = 23.879 + 17.696\, x$$

1. **Constant Variance Assumptions:** violated
2. **Normal Probability Plot of the residuals:** *normally distributed*
3. **Independence Assumption:** *valid*

the estimated regression equation $\hat{y} = 23.879 + 17.696\, x$ should not be used to predict

Data Transformations

- **Correct violation** of the assumptioms in linear regression model
  - Linear function and equall variance assumptions
- **Residual plot** suggests transformation of x and/or y
  - **important predictor variables (x) are omitted -** adding the omitted predictors
  - mean of the response (y) is **not a linear function of the predictors (x)** - different function
  - **unequal error variances** – transform y and/or x or use "**weighted least squares regression**."
  - **outlier** exists - "**robust estimation procedure**."
  - **error terms are not independent** - fit "**time series model**."

Data Transformations

first consider a simple linear regression model in which:

- transform the predictor (x) values only.
- transform the response (y) values only.
- transform both the predictor (x) values and response (y) values.

Data Transformations

- **Model building**
  - Model formulation
  - Model estimation
  - Model evaluation
- **Model use**

Data Transformations

check the appropriateness of a simple linear regression model.

Assumptions of Regression - **L.I.N.E**:

- **Linear Function:** The mean of the response, $E(Y_i)$, at each value of the predictor, x_i, is a Linear function of the x_i.
- **Independent:** The errors, ϵ_i, are Independent.
- **Normally Distributed:** The errors, ϵ_i, at each value of the predictor, x_i, are Normally distributed.
- **Equal variances:** The errors, ϵ_i, at each value of the predictor, x_i, have Equal variances (denoted σ^2).

Assumptions of Regression - L.I.N.E

Linearity

- The relationship between X and Y is linear

Independence of Errors

- Error values are statistically independent
- Particularly important when data are collected over a period of time

Normality of Error

- Error values are normally distributed for any given value of X

Equal Variance (also called homoscedasticity)

- The probability distribution of the errors has constant variance

Data Transformations

Your model

- is not overly complicated
- meets the four conditions of the linear regression model, and
- allows you to answer your research question of interest.

Data Transformations

Methods of Transforming Variables to Achieve Linearity

Method	Transform	Regression equation	Predicted value ($\hat{y}$)
Standard linear regression model	None	$y = b_0 + b_1x$	$\hat{y} = b_0 + b_1x$
Exponential model	DV = log(y)	$\log(y) = b_0 + b_1x$	$\hat{y} = 10^{b_0 + b_1x}$
Quadratic model	DV = sqrt(y)	$\text{sqrt}(y) = b_0 + b_1x$	$\hat{y} = (b_0 + b_1x)^2$
Reciprocal model	DV = 1/y	$1/y = b_0 + b_1x$	$\hat{y} = 1 / (b_0 + b_1x)$
Logarithmic model	IV = log(x)	$Y = b_0 + b_1\log(x)$	$\hat{y} = b_0 + b_1\log(x)$
Power model	DV = log(y) IV = log(x)	$\log(y) = b_0 + b_1\log(x)$	$\hat{y} = 10^{b_0 + b_1\log(x)}$

Note: DV is dependent variable and IV is independent variable

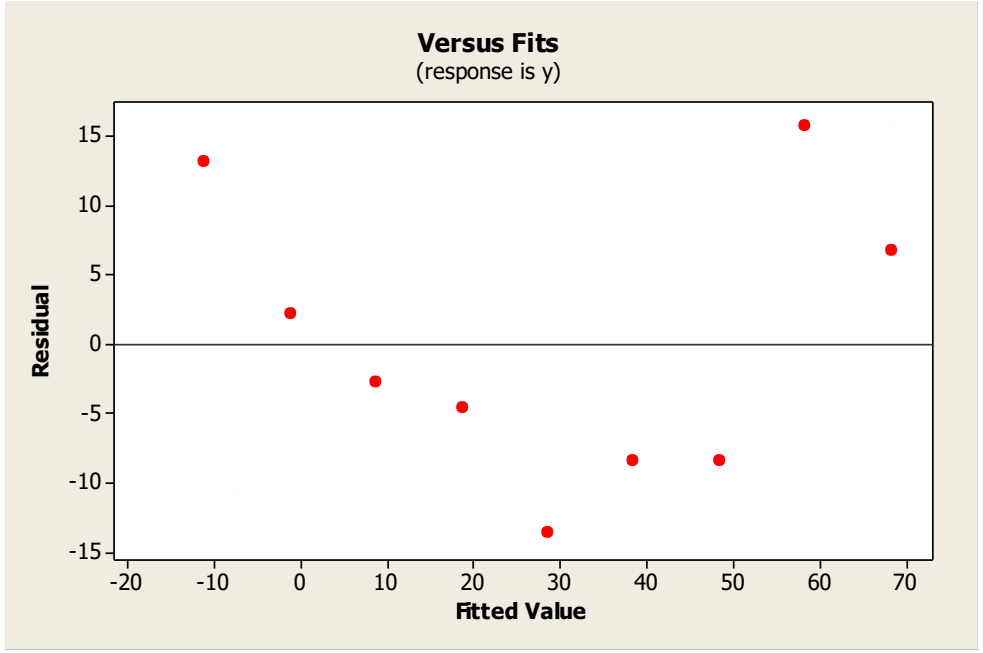
Data Transformations

How to Perform a Transformation to Achieve Linearity

- Conduct a standard regression analysis on the raw data.
- Construct a residual plot.
  - If the plot pattern is random, do not transform data.
  - If the plot pattern is not random, continue.
- Compute the coefficient of determination (R^2).
- Choose a transformation method (see above table).
- Transform the independent variable, dependent variable, or both.
- Conduct a regression analysis, using the transformed variables.
- Compute the coefficient of determination (R^2), based on the transformed variables.
  - If the transformed R^2 is greater than the raw-score R^2, the transformation was successful. Congratulations!
  - If not, try a different transformation method.

Transformation example

x	1	2	3	4	5	6	7	8	9
y	2	1	6	14	15	30	40	74	75

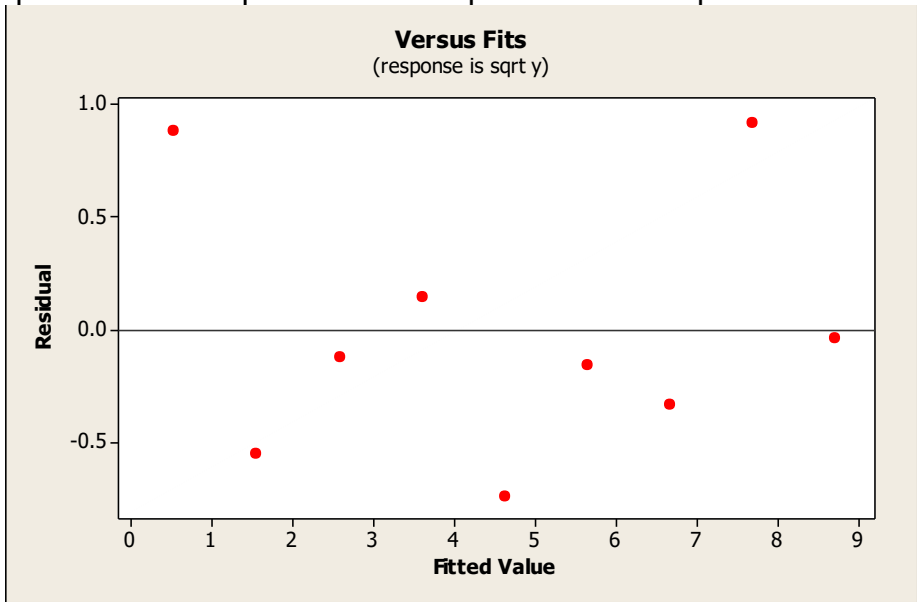


Transformation example

$$y'_t = b_0 + b_1 x \quad \text{or} \quad \sqrt{y} = b_0 + b_1 x$$

Quadratic model **DV = sqrt(y)** **$\sqrt{y} = b_0 + b_1 x$** **$\hat{y} = (b_0 + b_1 x)^2$**

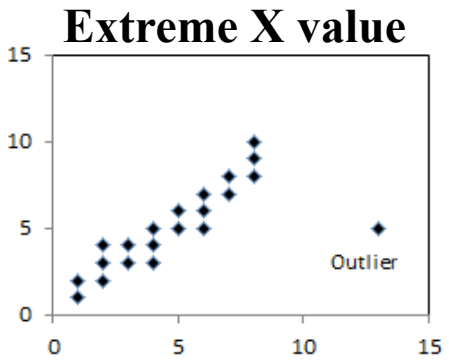
x	1	2	3	4	5	6	7	8	9
$\sqrt{y}$	1.41	1.00	2.45	3.74	3.87	5.48	6.32	8.60	8.66

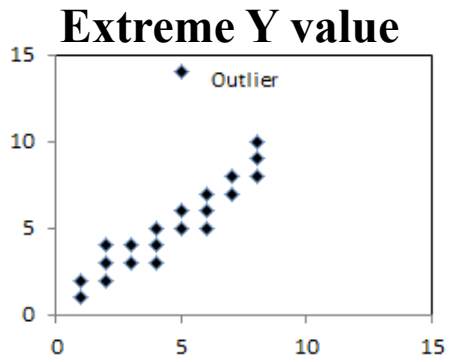


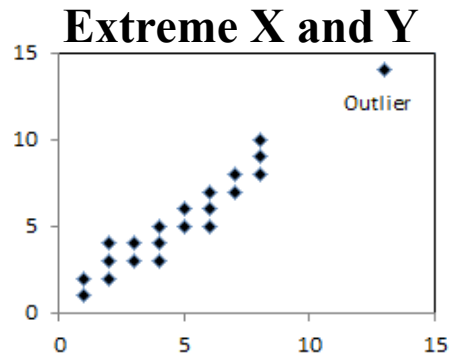
Influential Points and Outliers

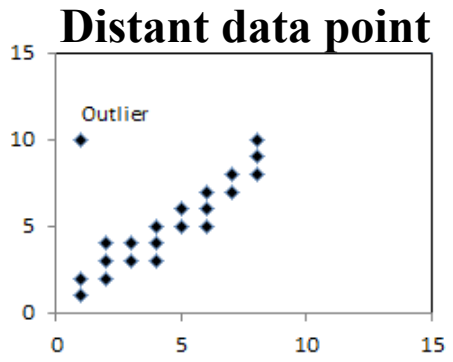
Outliers

- a value that is inconsistent with the rest of data (**extreme values**)
- Data points that diverge in a big way from the overall pattern





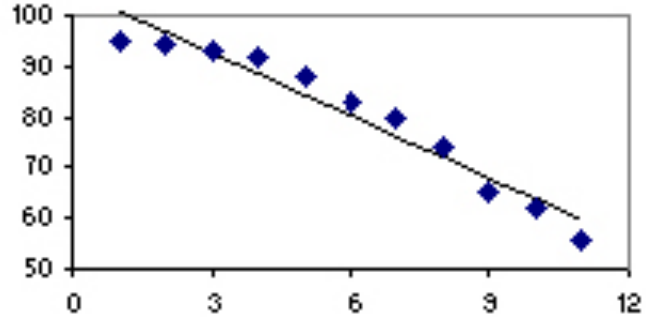




Influential Points and Outliers

An **influential point** is an outlier that greatly affects the slope of the regression line.

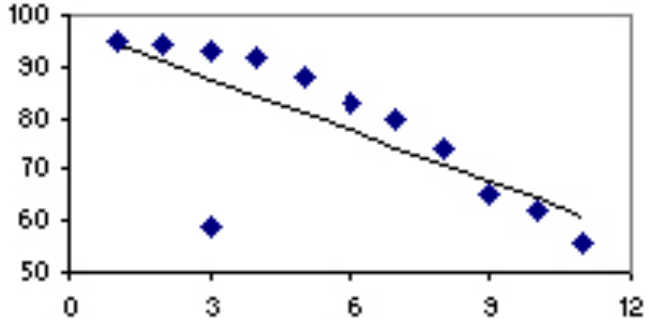
Without Outlier



Regression equation: $\hat{y} = 104.78 - 4.10x$

Coefficient of determination: $R^2 = 0.94$

With Outlier



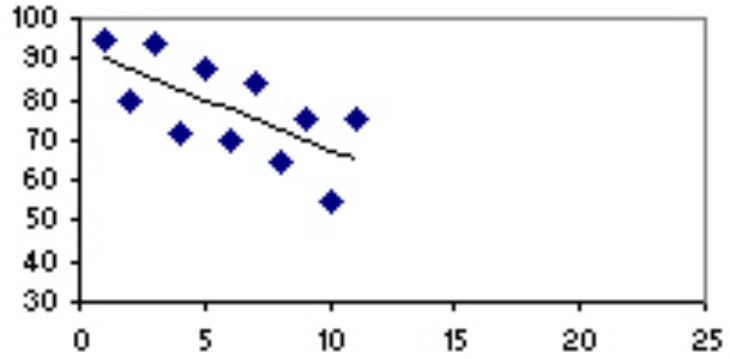
Regression equation: $\hat{y} = 97.51 - 3.32x$

Coefficient of determination: $R^2 = 0.55$

Influential Points and Outliers

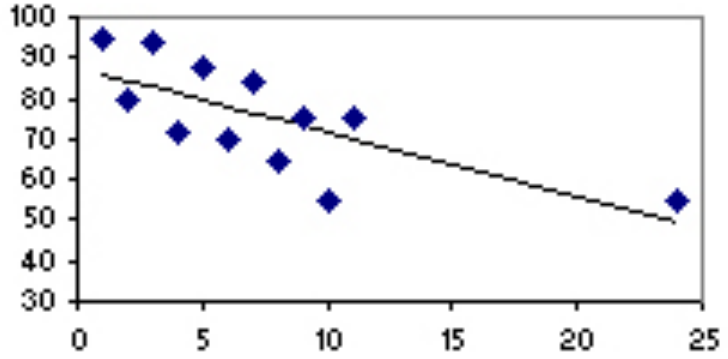
influential point will cause the R^2 sometimes bigger or smaller

Without Outlier



Regression equation: $\hat{y} = 92.54 - 2.5x$
Slope: $b_0 = -2.5$
Coefficient of determination: $R^2 = 0.46$

With Outlier



Regression equation: $\hat{y} = 87.59 - 1.6x$
Slope: $b_0 = -1.6$
Coefficient of determination: $R^2 = 0.52$

Influential Points and Outliers

Consider following if data set has influential point

- An influential point may represent bad data, possibly the result of measurement error. If possible, check the validity of the data point.
- Compare the decisions that would be made based on regression equations defined with and without the influential point. If the equations lead to contrary decisions, use caution.

Detecting Influential Observations

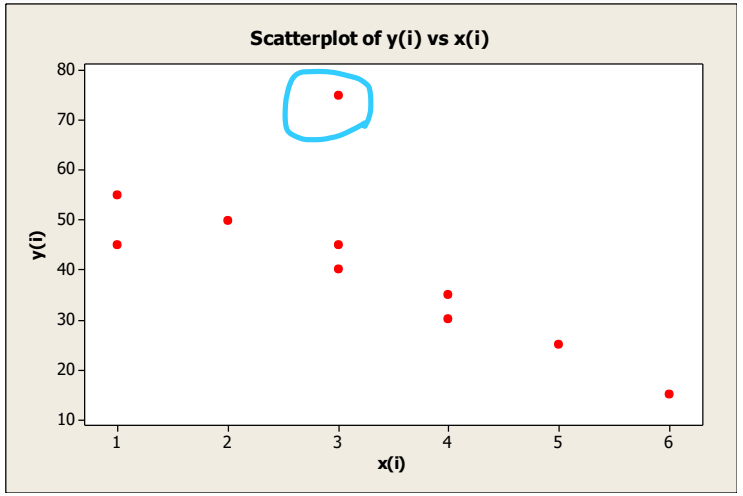
Scatterplot

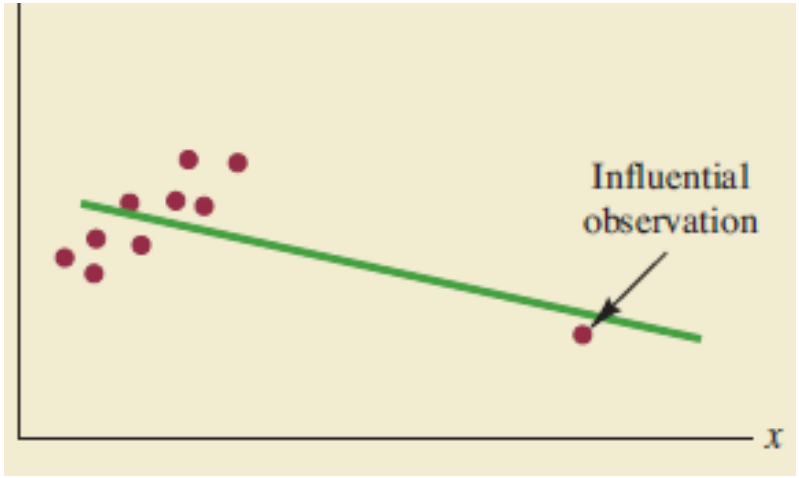
Standardized residual

- Unsual observations

Influential observation

- High leverage points - *how far they are from mean values*
- Cook's Distance Measure – *to identify influential observations*





What to do about Outliers?

- Data was recorded correctly?
  - If not, discard the observation and rerun.
  - If correct, search for a reason for the observation. It might be caused by a situation we do not wish to model. If so, drop the observation.
- Missing an important independent variable in the model.

Key Formula

Key Formulas

Simple Linear Regression Model

$$y = \beta_0 + \beta_1 x + \epsilon \quad (14.1)$$

Simple Linear Regression Equation

$$E(y) = \beta_0 + \beta_1 x \quad (14.2)$$

Estimated Simple Linear Regression Equation

$$\hat{y} = b_0 + b_1 x \quad (14.3)$$

Least Squares Criterion

$$\min \Sigma(y_i - \hat{y}_i)^2 \quad (14.5)$$

Slope and *y*-Intercept for the Estimated Regression Equation

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \quad (14.6)$$

$$b_0 = \bar{y} - b_1 \bar{x} \quad (14.7)$$

Sum of Squares Due to Error

$$\text{SSE} = \Sigma(y_i - \hat{y}_i)^2 \quad (14.8)$$

Total Sum of Squares

$$\text{SST} = \Sigma(y_i - \bar{y})^2 \quad (14.9)$$

Sum of Squares Due to Regression

$$\text{SSR} = \Sigma(\hat{y}_i - \bar{y})^2 \quad (14.10)$$

Relationship Among SST, SSR, and SSE

$$\text{SST} = \text{SSR} + \text{SSE} \quad (14.11)$$

Coefficient of Determination

$$r^2 = \frac{\text{SSR}}{\text{SST}} \quad (14.12)$$

Key Formula

Sample Correlation Coefficient

$$r_{xy} = (\text{sign of } b_1)\sqrt{\text{Coefficient of determination}} = (\text{sign of } b_1)\sqrt{r^2} \quad (14.13)$$

Mean Square Error (Estimate of σ^2)

$$s^2 = \text{MSE} = \frac{\text{SSE}}{n-2} \quad (14.15)$$

Standard Error of the Estimate

$$s = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{n-2}} \quad (14.16)$$

Standard Deviation of b_1

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\Sigma(x_i - \bar{x})^2}} \quad (14.17)$$

Estimated Standard Deviation of b_1

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}} \quad (14.18)$$

***t* Test Statistic**

$$t = \frac{b_1}{s_{b_1}} \quad (14.19)$$

Mean Square Regression

$$\text{MSR} = \frac{\text{SSR}}{\text{Number of independent variables}} \quad (14.20)$$

***F* Test Statistic**

$$F = \frac{\text{MSR}}{\text{MSE}} \quad (14.21)$$

Estimated Standard Deviation of $\hat{y}_p$

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} \quad (14.23)$$

Key Formula

Confidence Interval for $E(y_p)$

$$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p} \qquad (14.24)$$

Estimated Standard Deviation of an Individual Value

$$s_{ind} = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} \qquad (14.26)$$

Prediction Interval for y_p

$$\hat{y}_p \pm t_{\alpha/2} s_{ind} \qquad (14.27)$$

Regression Analysis with Excel

STUDENT POPULATION AND QUARTERLY SALES DATA FOR 10 ARMAND'S PIZZA PARLORS

Restaurant i	Student Population (1000s) x_i	Quarterly Sales ($1000s) y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

Steps to produce the regression results

Step 1. Click the Data tab on the Ribbon

Step 2. In the Analysis group, click Data Analysis

Step 3. Choose Regression from the list of Analysis Tools

Step 4. Click OK

Step 5. When the Regression dialog box appears:

Enter C1:C11 in the Input Y Range box

Enter B1:B11 in the Input X Range box

Select Labels

Select Confidence Level

Enter 99 in the Confidence Level box

Select Output Range

Enter A13 in the Output Range box

Click OK

EXCEL Output

	A	B	C	D	E	F	G	H	I	J
1	**Restaurant**	**Population**	**Sales**							
2	1	2	58							
3	2	6	105							
4	3	8	88							
5	4	8	118							
6	5	12	117							
7	6	16	137							
8	7	20	157							
9	8	20	169							
10	9	22	149							
11	10	26	202							
12										
13	SUMMARY OUTPUT									
14										
15	*Regression Statistics*									
16	Multiple R	0.9501								
17	R Square	0.9027								
18	Adjusted R Square	0.8906								
19	Standard Error	13.8293								
20	Observations	10								
21										
22	ANOVA									
23		*df*	*SS*	*MS*	*F*	*Significance F*				
24	Regression	1	14200	14200	74.2484	2.55E-05				
25	Residual	8	1530	191.25						
26	Total	9	15730							
27										
28		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*	
29	Intercept	60	9.2260	6.5033	0.0002	38.7247	81.2753	29.0431	90.9569	
30	Population	5	0.5803	8.6167	2.55E-05	3.6619	6.3381	3.0530	6.9470	
31										
32										
33										
34										

Interpretation Estimated Regression Equation Output

27									
28		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
29	Intercept	60	9.2260	6.5033	0.0002	38.7247	81.2753	29.0431	90.9569
30	Population	5	0.5803	8.6167	2.55E-05	3.6619	6.3381	3.0530	6.9470

Interpretation of ANOVA Output

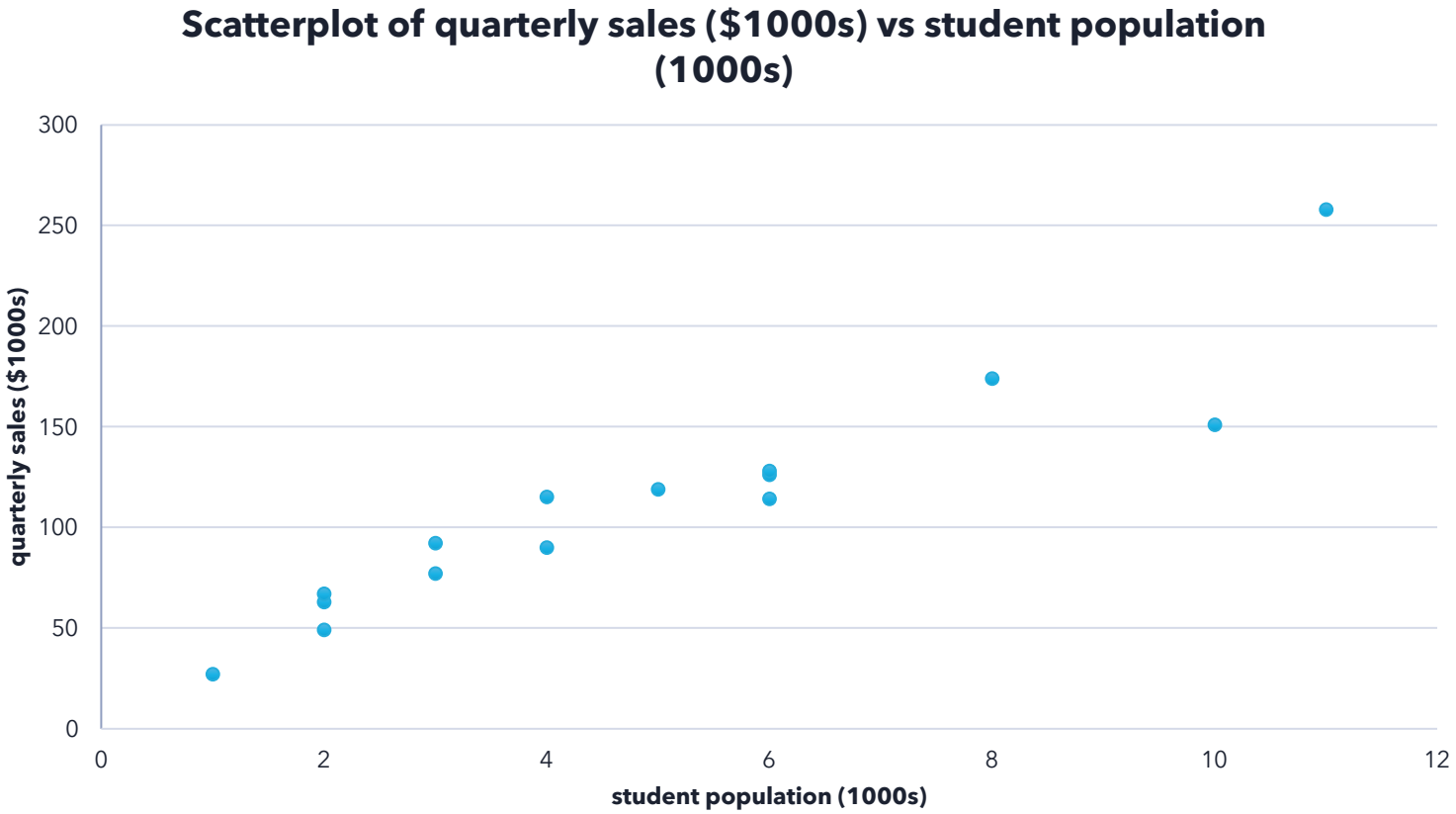
21						
22	ANOVA					
23		*df*	*SS*	*MS*	*F*	*Significance F*
24	Regression	1	14200	14200	74.2484	2.55E-05
25	Residual	8	1530	191.25		
26	Total	9	15730			

Interpretation of Regression Statistics Output

12		
13	SUMMARY OUTPUT	
14		
15	*Regression Statistics*	
16	Multiple R	0.9501
17	R Square	0.9027
18	Adjusted R Square	0.8906
19	Standard Error	13.8293
20	Observations	10

Let’s try this example with Excel

Restaurant	student population (1000s), x	quarterly sales ($1000s), y
1	3	92
2	2	63
3	6	126
4	8	174
5	2	49
6	4	90
7	5	119
8	6	114
9	2	67
10	4	115
11	6	128
12	11	258
13	3	77
14	10	151
15	1	27



Assignment #1(a)

Assignment #1: Simple Linear Regression

- use **Excel** (or Minitab or any other software)

House Price in $1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

- Develop a **scatter diagram** for these data
- Determine the **sample correlation coefficient** and interpret its value, and test the significance of the correlation coefficient at α = 0.05 and d.f = n-2.
- **Test the significance** of correlation coefficient
- Develop an **estimated regression equation** that can be used to predict the house price given the square feet.
- Use the estimated regression equation to **predict** house price for a square feet of 2000.
- Is there a **relationship** between the square footage of the house and its sales price? Hence perform
  - **Test the significance of b_1** (T-test), and
  - **F-Test** for Significance
  - **Confidence Interval Estimate** for the Slope

Assignment #1 (b)

"Best Fitting Line"

- heights (x) and weights (y) of 10 students
- which line – the red line or the green line – do you think best summarizes the trend between height and weight?