Introductory Statistics

Chapter 14

Descriptive Methods in Regression and Correlation

Simple Regression Analysis

- bivariate (two variables) linear regression -- the most elementary regression model
  - dependent variable, *the variable to be predicted*, usually called Y
  - independent variable, *the predictor or explanatory variable*, usually called X

Table 14.2

Age and price data for a sample of 11 Orions

In this example, we have the predicted price of a particular make and model of car with respect to the age of this particular make of car.

Car	Age (yr) x	Price ($100) y
1	5	85
2	4	103
3	6	70
4	5	82
5	5	89
6	5	98
7	6	66
8	6	95
9	2	169
10	7	70
11	7	48

Two Quantitative Variables

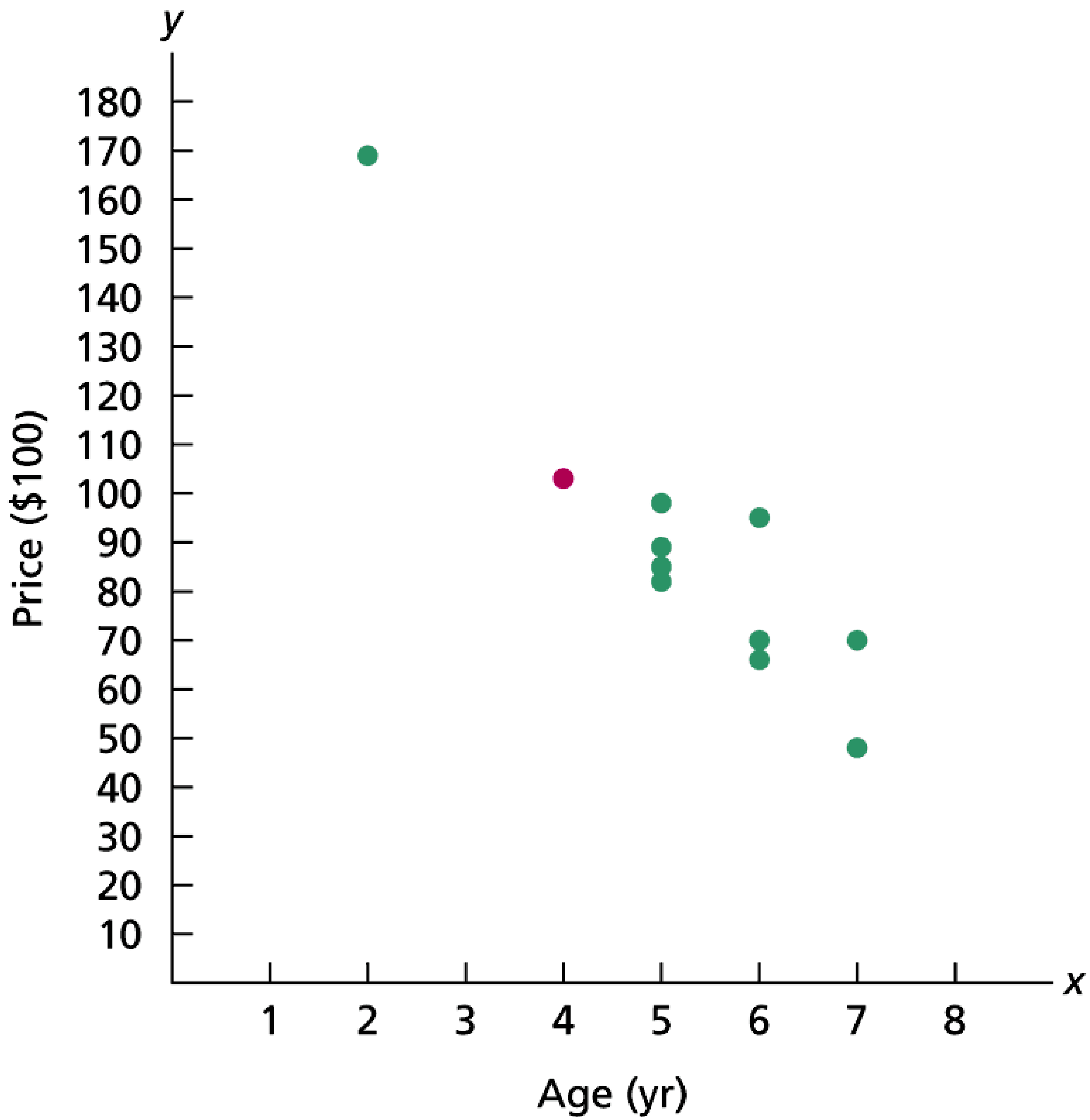
1. Expressed as ordered pairs: (x, y)
2. x: input variable, independent variable
   y: output variable, dependent variable

Scatter Diagram: A plot of all the ordered pairs of bivariate data on a coordinate axis system. The input variable x is plotted on the horizontal axis, and the output variable y is plotted on the vertical axis.

Note: Use scales so that the range of the y-values is equal to or slightly less than the range of the x-values. This creates a window that is approximately square.

Figure 14.7

Scatterplot for the age and price data of Orions from Table 14.2



Questions: what is the line or an equation that best representing the data?

Section 14.1

Linear Equations with One Independent Variable

What does linear mean

- Linear = line (we will see later on that this concept must be extended to higher dimensions)
- Technically, linear refers to the coefficients in the regression model
- What is this regression model we keep referring to?
- In school (hopefully) we learned a number of ways to describe a straight line
- A straight line can be described to points on a graph – *the line passes through the points* (x_1, y_1) *and* (x_2, y_2)

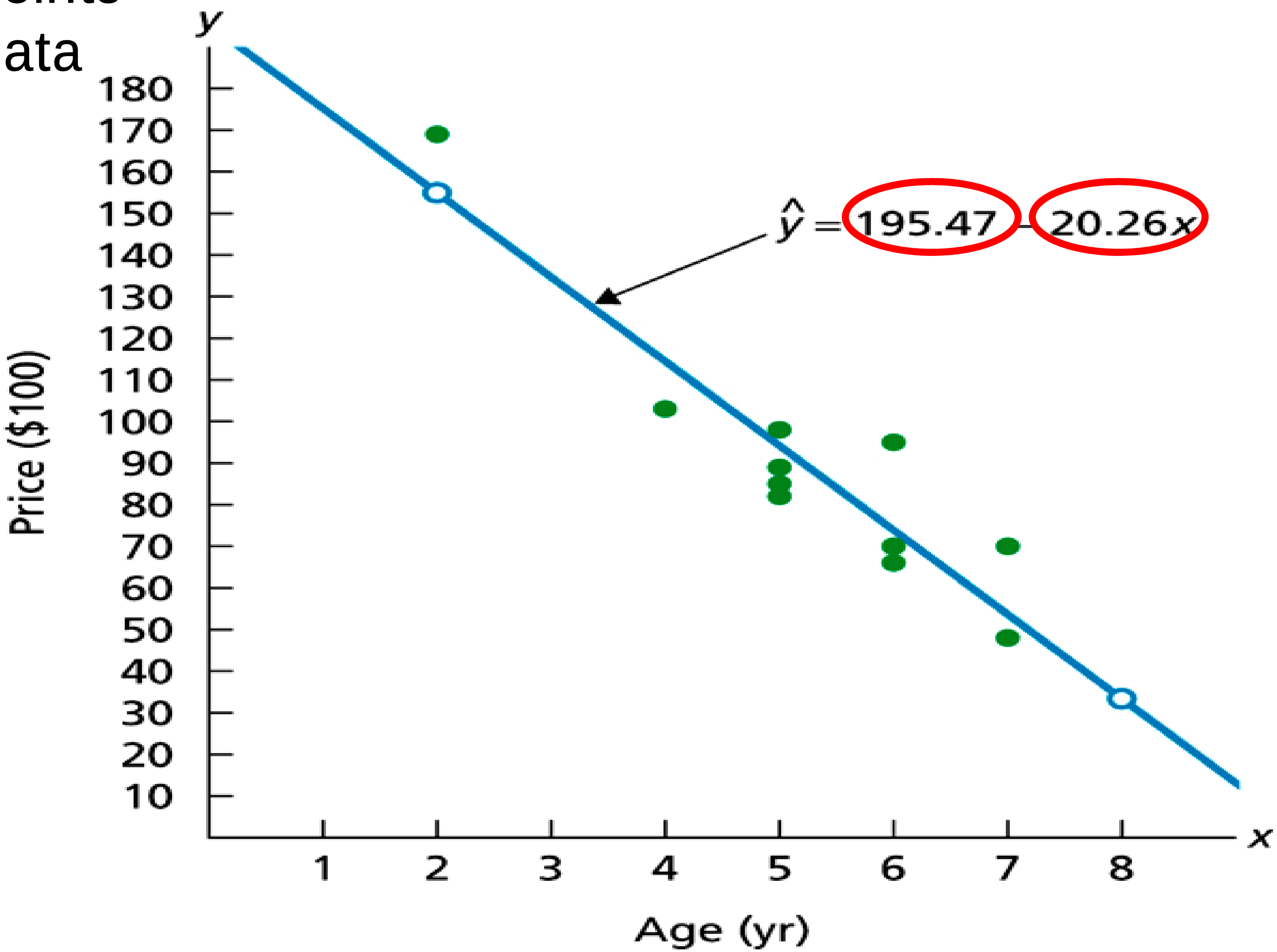
Definition 14.1

y-Intercept and Slope

For a linear equation $y = b_0 + b_1 x$, the number b_0 is called the **y-intercept** and the number b_1 is called the **slope.**

Figure 14.10

Regression line
and data points
for Orion data



Section 14.2

The Regression Equation

Example

Consider the problem of fitting a straight line to the four points in Table 14.3. There are infinitely many lines that can be fitted onto these points, the questions is which one do we want?

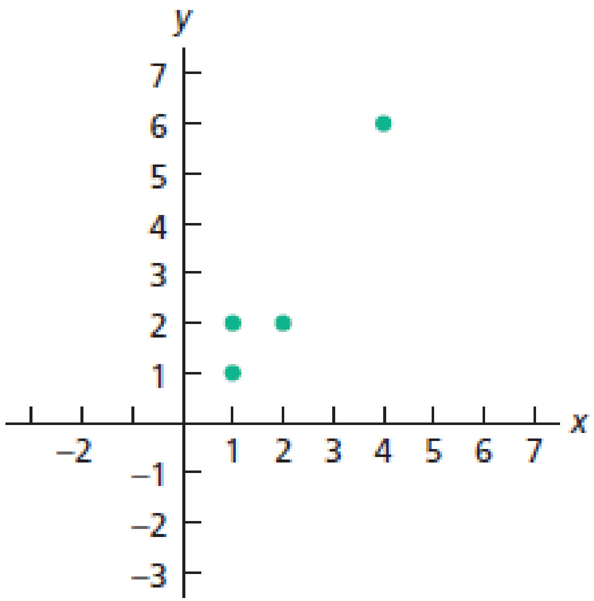
X	Y
1	1
1	2
2	2
4	6

Table 14.3 & Figure 14.8

Four data points

x	y
1	1
1	2
2	2
4	6

Scatterplot for the data points in Table 14.3



Simple linear regression

- Perhaps now we have the tools to begin to write down a model
- Generally we have more than two points to work with
- Ideally we wouldn't fit a regression model to a data set with fewer than thirty points
- We have a number of **responses**, y_i, and an associated measurement (which we assume is taken without measurement error) x_i which we think explains our response.
- However, the points (usually) don't lie exactly on a straight line – there is a bit of "noise" associated with each measurement – does this sound familiar?

Regression Analysis

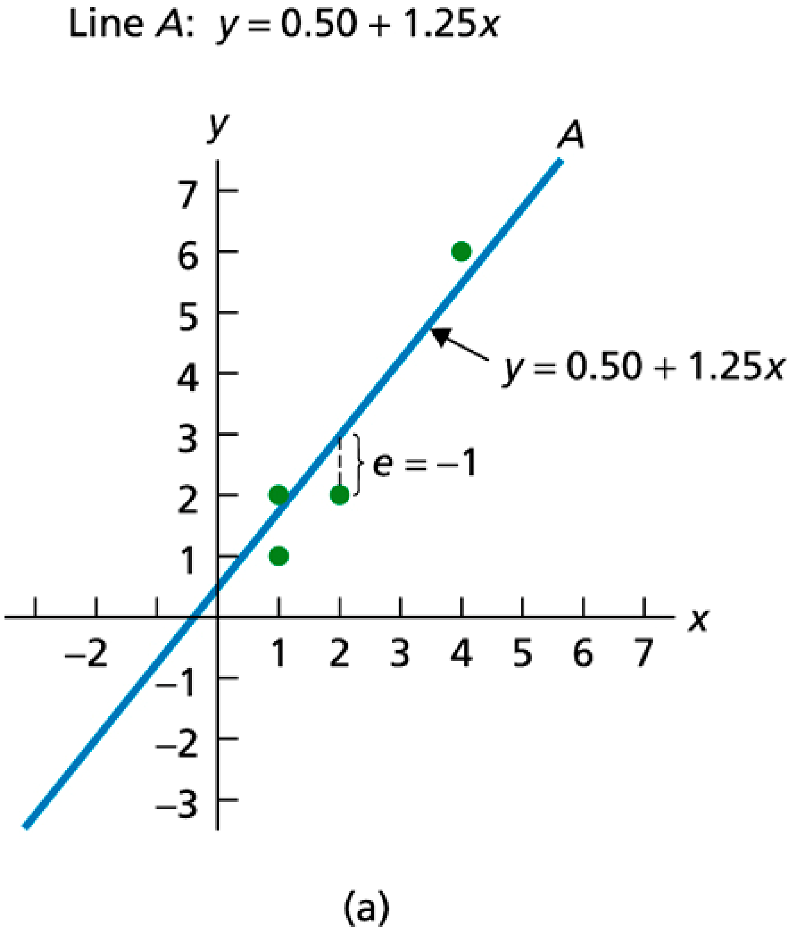
In regression analysis we use the independent variable (X) to estimate the dependent variable (Y).

- The relationship between the variables is linear.
- Both variables must be at least interval scale.
- The least squares criterion is used to determine the equation. That is the term $\sum(Y - \hat{Y})^2$ is minimized (where $\hat{Y}$ is the predicted value of Y).

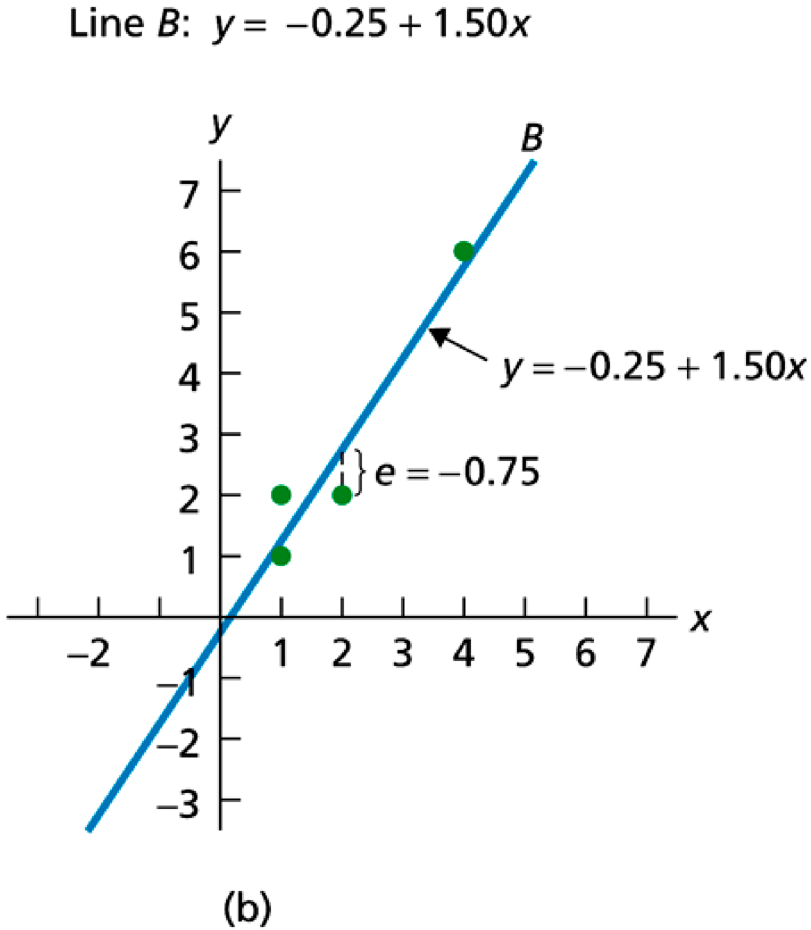
Note: Again, one variable depends on another variable does not imply causation.

Figure 14.9

Two possible lines to fit the data points in Table 14.3



(a)



(b)

Table 14.4

Determining how well the data points in Table 14.3 are fit by (a) Line A and (b) Line B

Line A: $y = 0.50 + 1.25x$

x	y	$\hat{y}$	e	e^2
1	1	1.75	−0.75	0.5625
1	2	1.75	0.25	0.0625
2	2	3.00	−1.00	1.0000
4	6	5.50	0.50	0.2500
				1.8750

(a)

Line B: $y = -0.25 + 1.50x$

x	y	$\hat{y}$	e	e^2
1	1	1.25	−0.25	0.0625
1	2	1.25	0.75	0.5625
2	2	2.75	−0.75	0.5625
4	6	5.75	0.25	0.0625
				1.2500

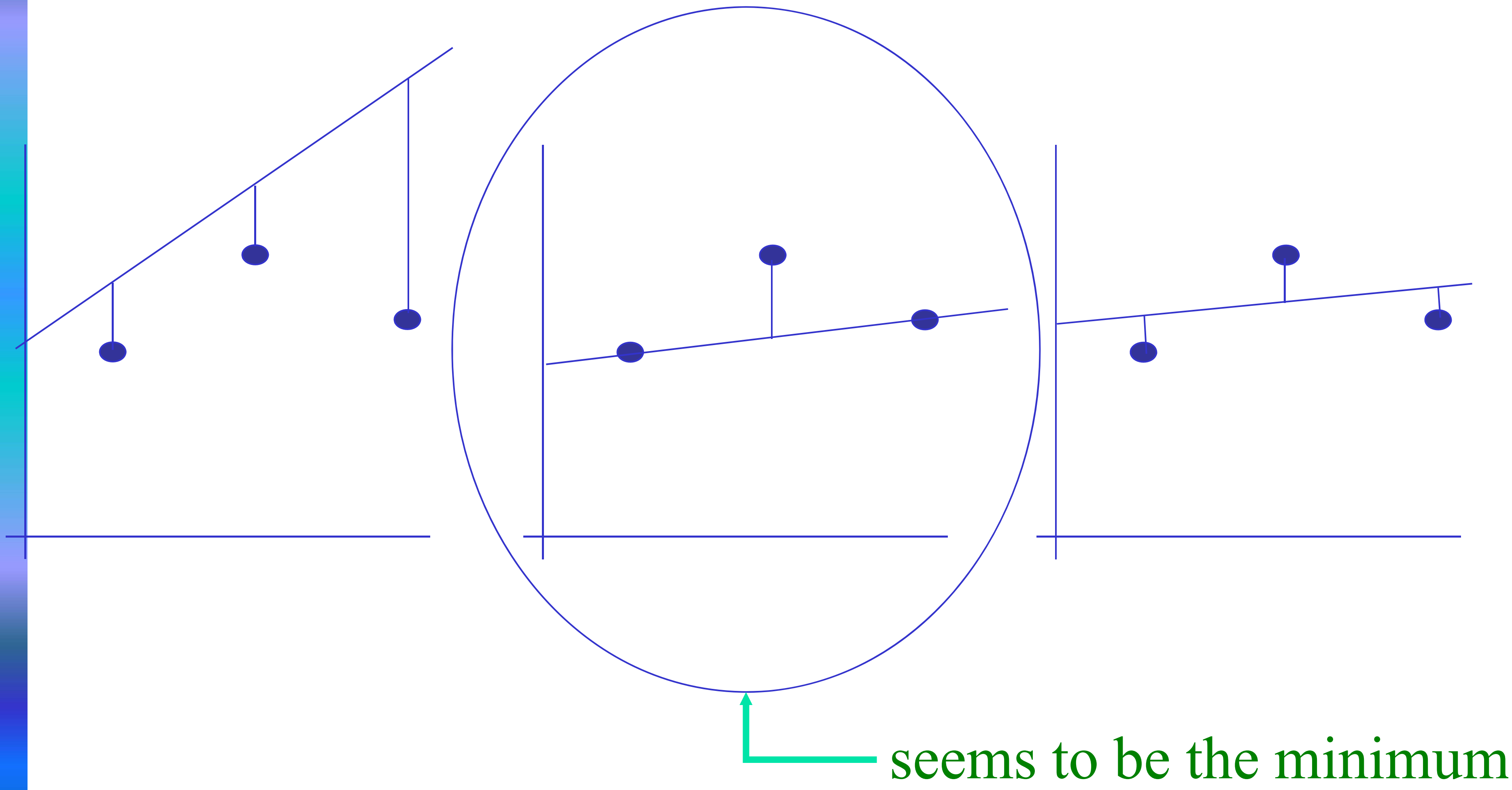
(b)

A probability model for simple linear regression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ \varepsilon_i\ iid\ N\left(0, \sigma^2\right)$$

- *The response variable, Y, is described by the intercept and a coefficient for the predictor X. If we subtract the model we expect to find residuals which are independently and identically normally distributed with mean zero and standard deviation sigma.*
- How do we know we find the intercept and

An illustration of the least squares principle



Key Fact 14.2 & Definition 14.2

Least-Squares Criterion

The **least-squares criterion** is that the line that best fits a set of data points is the one having the smallest possible sum of squared errors.

Regression Line and Regression Equation

Regression line: The line that best fits a set of data points according to the least-squares criterion.

Regression equation: The equation of the regression line.

Least squares

- The least squares procedure is a method for fitting regression lines
- It attempts to find the intercept and the slope such that the residual sum of squares is minimised; i.e. Find β_0 and β_1 such that

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

  is minimized
- The minimum value of this function is zero. This is hardly ever achieved.
- The least squares fitted values are denoted b_0 and b_1

How to minimize distance?

We need to look for stationary points:

$$\frac{\partial \sum (Y_i - \hat{Y})^2}{\partial \beta_j} = 0$$

$$i.e. \begin{cases} -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) = 0 \\ -2\sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i) = 0 \end{cases}$$

Definition 14.3

Notation Used in Regression and Correlation

For a set of n data points, the defining and computing formulas for S_{xx}, S_{xy}, and S_{yy} are as follows.

Quantity	Defining formula	Computing formula
S_{xx}	$\Sigma(x_i - \bar{x})^2$	$\Sigma x_i^2 - (\Sigma x_i)^2/n$
S_{xy}	$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n$
S_{yy}	$\Sigma(y_i - \bar{y})^2$	$\Sigma y_i^2 - (\Sigma y_i)^2/n$

Formula 14.1

Regression Equation

The regression equation for a set of n data points is $\hat{y} = b_0 + b_1 x$, where

$$b_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad b_0 = \frac{1}{n}(\Sigma y_i - b_1 \Sigma x_i) = \bar{y} - b_1 \bar{x}.$$

Table 14.5

Some of the calculations that need to be complete, before we find the regression equations for the Orion example:

Age (yr) x	Price ($100) y
5	85
4	103
6	70
5	82
5	89
5	98
6	66
6	95
2	169
7	70
7	48

Table 14.5

Table for computing the regression equation for the Orion data

Age (yr) x	Price ($100) y	xy	x^2
5	85	425	25
4	103	412	16
6	70	420	36
5	82	410	25
5	89	445	25
5	98	490	25
6	66	396	36
6	95	570	36
2	169	338	4
7	70	490	49
7	48	336	49
58	975	4732	326

Figure 14.10

Regression line and data points for Orion data

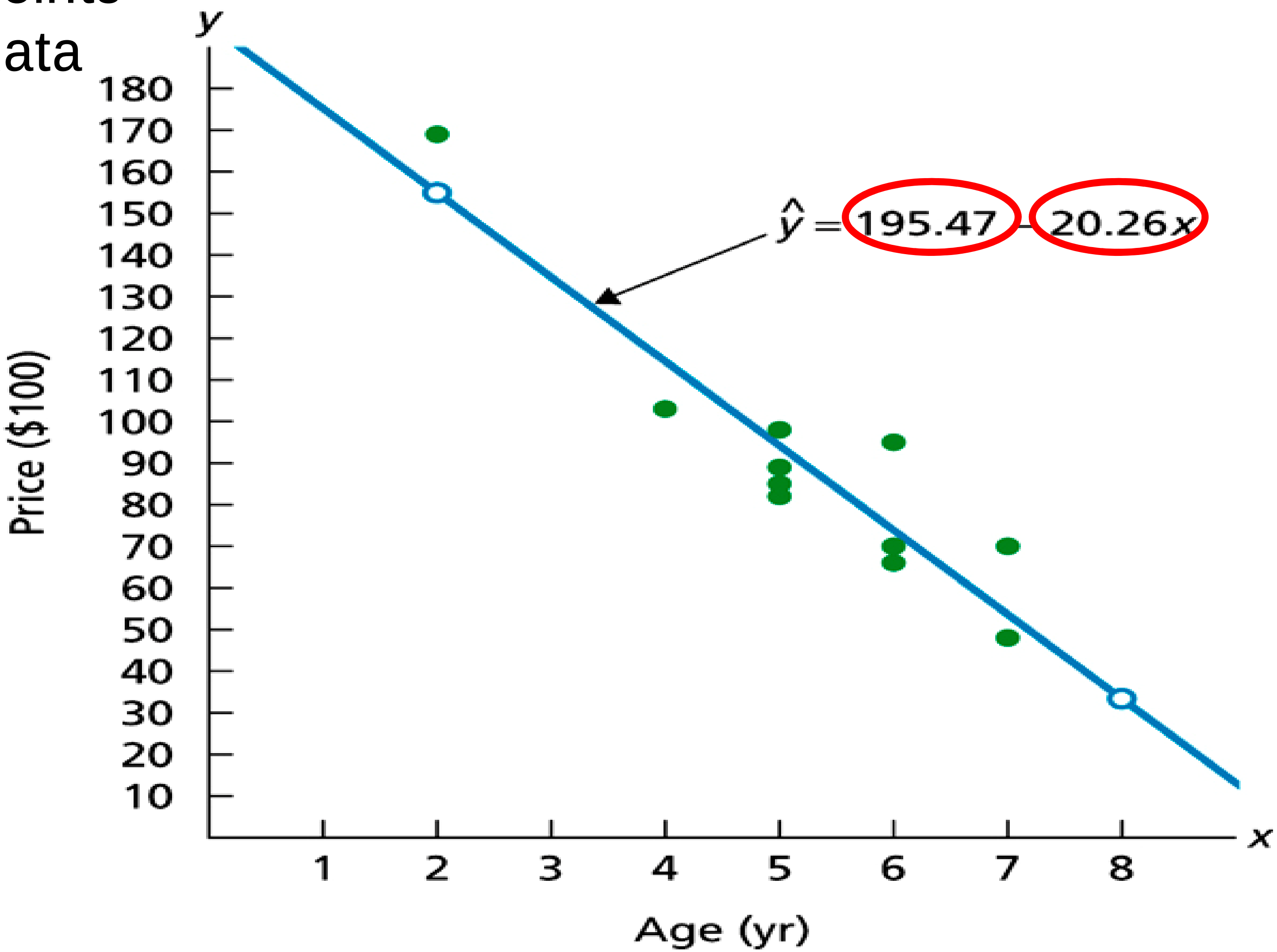


Figure 14.11

Extrapolation in the Orion example

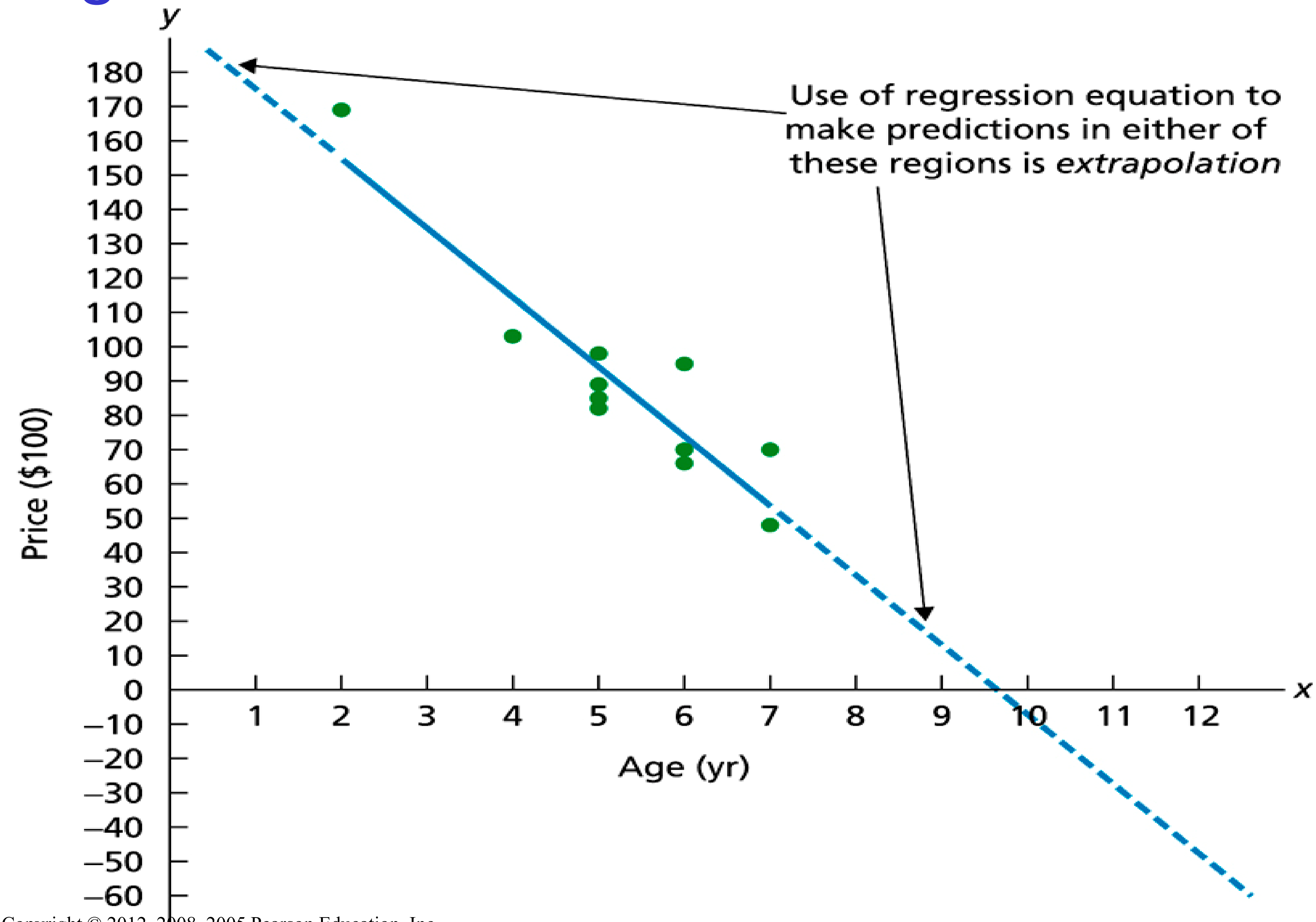
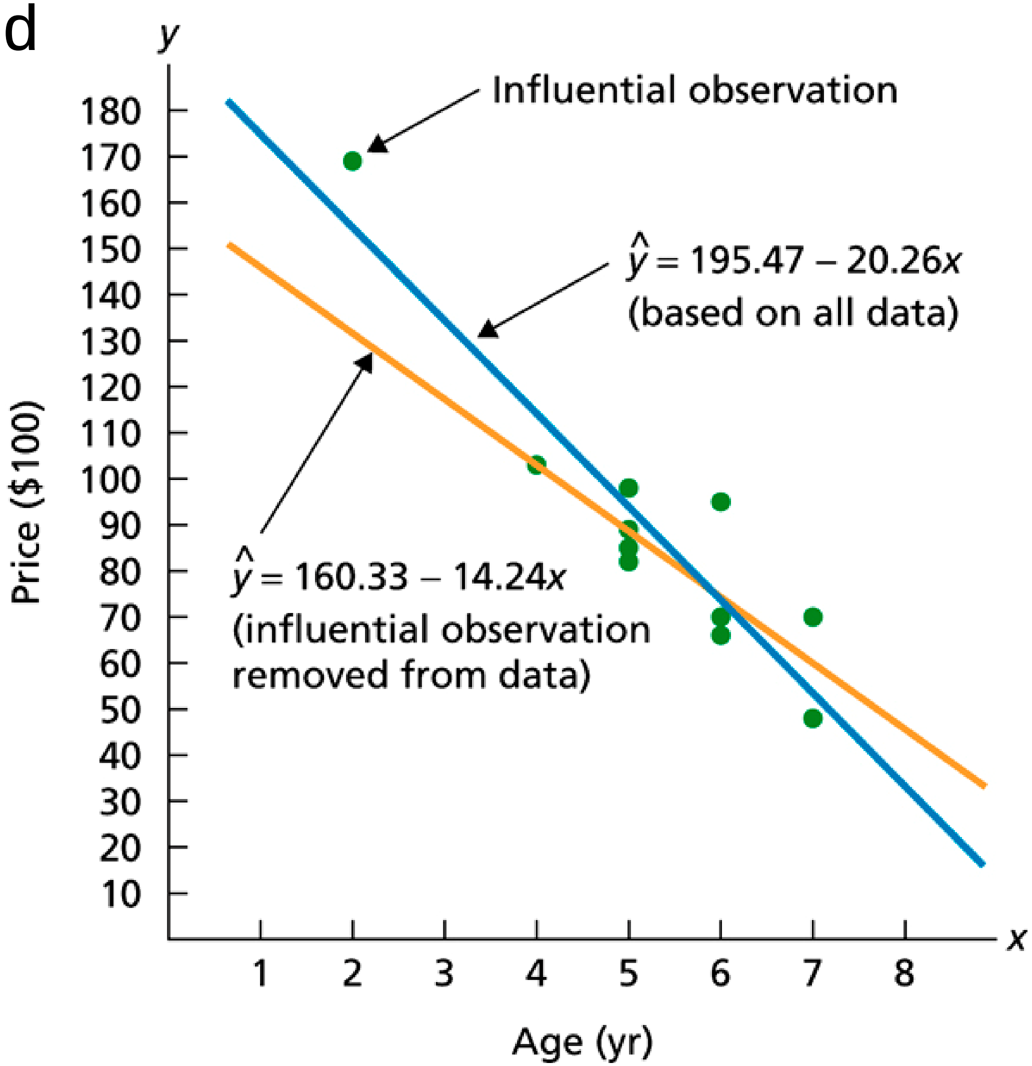


Figure 14.12

Regression lines with and without the influential observation removed



Please Note

- Keep at least three extra decimal places while doing the calculations to ensure an accurate answer
- When rounding off the calculated values of b_0 and b_1, always keep at least two significant digits in the final answer
- The slope b_1 represents the predicted change in y per unit increase in x
- The y-intercept is the value of y where the line of best fit intersects the y-axis
- The line of best fit will always pass through the point $(\bar{x}, \bar{y})$

Making Predictions

1. One of the main purposes for obtaining a regression equation is for making predictions
2. For a given value of x, we can predict a value of $\hat{Y}$
3. The regression equation should be used to make predictions only about the population from which the sample was drawn
4. The regression equation should be used only to cover the sample domain on the input variable. You can estimate values outside the domain interval, but use caution and use values close to the domain interval.
5. Use current data. A sample taken in 1987 should not be used to make predictions in 1999.

Example

✓ **Example:** A recent article measured the job satisfaction of subjects with a 14-question survey. The data below represents the job satisfaction scores, *y*, and the salaries, *x*, for a sample of similar individuals:

x	31	33	22	24	35	29	23	37
y	17	20	13	15	18	17	12	21

1) Draw a scatter diagram for this data
2) Find the equation of the line of best fit

Finding b_1 & b_0

- Preliminary calculations needed to find b_1 and b_0:

x	y	x^2	xy
23	12	529	276
31	17	961	527
33	20	1089	660
22	13	484	286
24	15	576	360
35	18	1225	630
29	17	841	493
37	21	1369	777
234	133	7074	4009
$\sum x$	$\sum y$	$\sum x^2$	$\sum xy$

Line of Best Fit

$$\mathrm{SS}(x) = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 7074 - \left[\frac{234^2}{8}\right] = 229.5$$

$$\mathrm{SS}(xy) = \sum xy - \frac{\sum x \sum y}{n} = 4009 - \left[\frac{(234)(133)}{8}\right] = 118.75$$

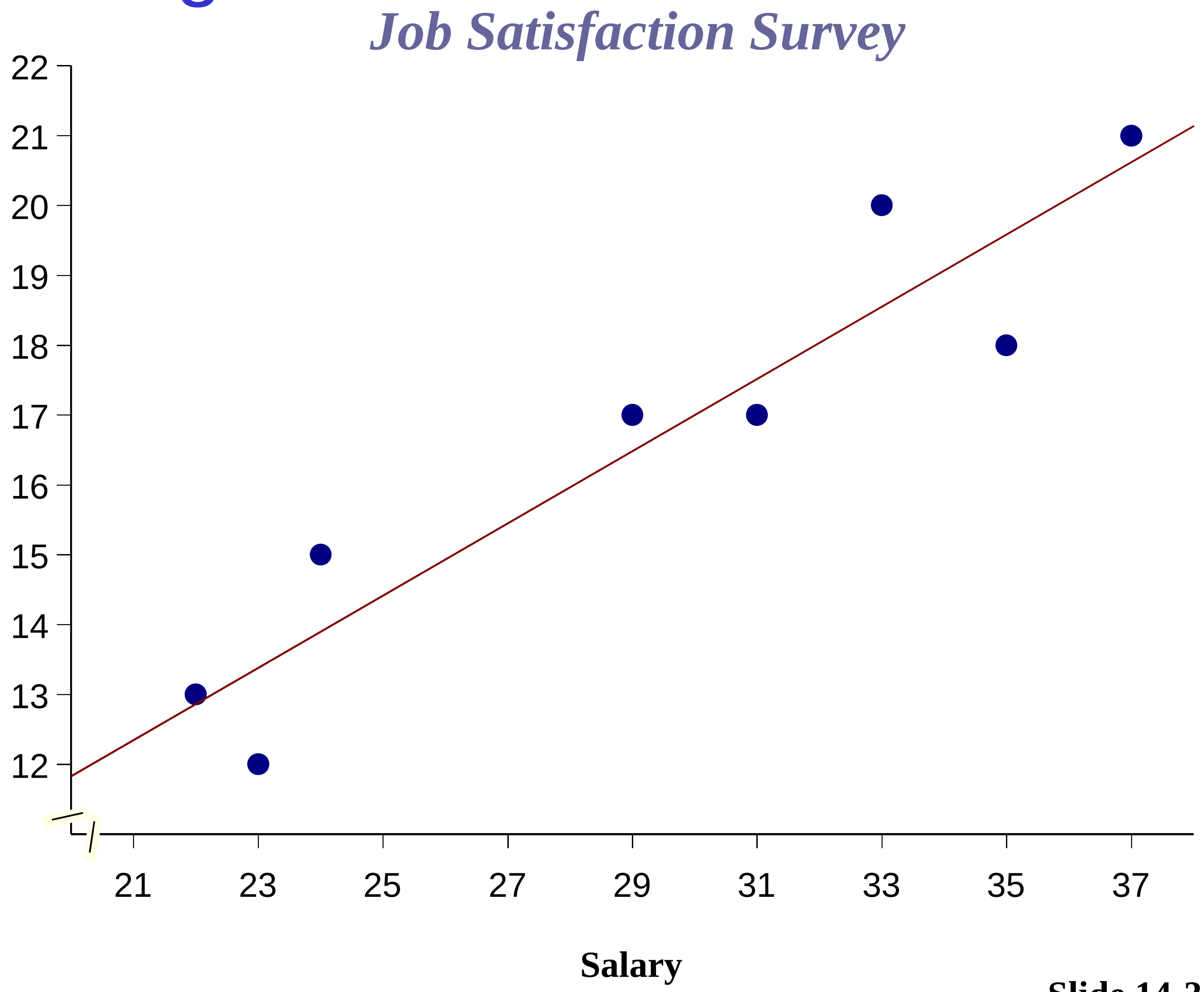
$$b_1 = \frac{\mathrm{SS}(xy)}{\mathrm{SS}(x)} = \frac{118.75}{229.5} = 0.5174$$

$$b_0 = \frac{\sum y - \left(b_1 \times \sum x\right)}{n} = \frac{133 - (0.5174)(234)}{8} = 1.4902$$

Solution 1) Equation of the line of best fit: $\hat{y} = 1.49 + 0.517x$

Scatter Diagram

Solution 2)



Section 14.3

The Coefficient of Determination

Coefficient of Determination

The coefficient of determination (r^2) is the proportion (or %) of the total variation in the dependent variable (Y) that is explained or accounted for by the variation in the independent variable (X).

- It is the square of the coefficient of correlation.
- It ranges from 0 to 1.
- It does not give any information on the direction of the relationship between the variables.

Definition 14.5

Sums of Squares in Regression

Total sum of squares, *SST*: The total variation in the observed values of the response variable: $SST = \Sigma(y_i - \bar{y})^2$.

Regression sum of squares, *SSR*: The variation in the observed values of the response variable explained by the regression: $SSR = \Sigma(\hat{y}_i - \bar{y})^2$.

Error sum of squares, *SSE*: The variation in the observed values of the response variable not explained by the regression: $SSE = \Sigma(y_i - \hat{y}_i)^2$.

Table 14.6

Table for computing *SST* for the Orion price data

Age (yr) x	Price ($100) y	$y - \bar{y}$	$(y - \bar{y})^2$
5	85	−3.64	13.2
4	103	14.36	206.3
6	70	−18.64	347.3
5	82	−6.64	44.0
5	89	0.36	0.1
5	98	9.36	87.7
6	66	−22.64	512.4
6	95	6.36	40.5
2	169	80.36	6458.3
7	70	−18.64	347.3
7	48	−40.64	1651.3
	975		9708.5

An graphical illustration

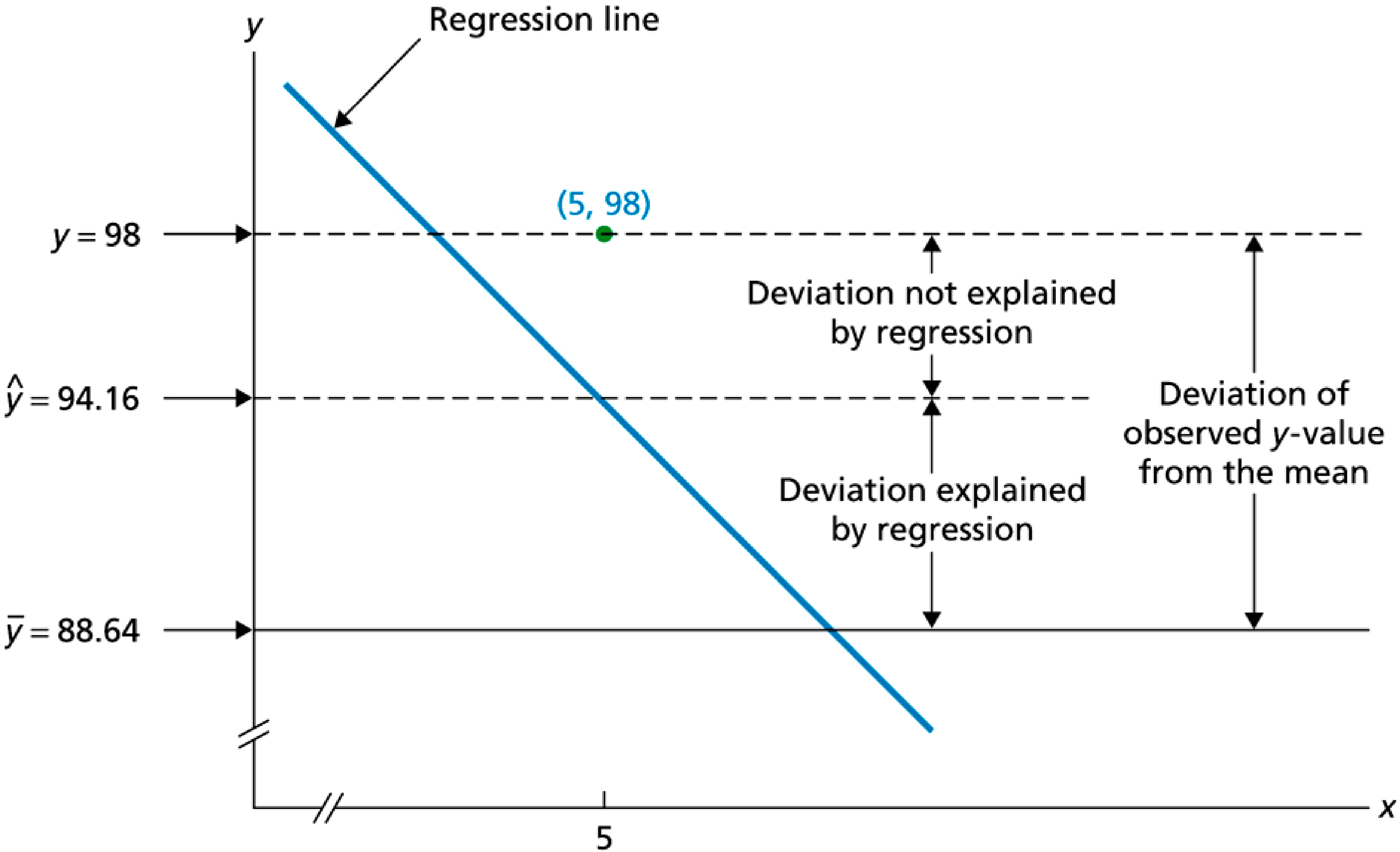


Table 14.7

Table for computing *SSR* for the Orion price data

Age (yr) x	Price ($100) y	$\hat{y}$	$\hat{y} - \bar{y}$	$(\hat{y} - \bar{y})^2$
5	85	94.16	5.53	30.5
4	103	114.42	25.79	665.0
6	70	73.90	−14.74	217.1
5	82	94.16	5.53	30.5
5	89	94.16	5.53	30.5
5	98	94.16	5.53	30.5
6	66	73.90	−14.74	217.1
6	95	73.90	−14.74	217.1
2	169	154.95	66.31	4397.0
7	70	53.64	−35.00	1224.8
7	48	53.64	−35.00	1224.8
				8285.0

Table 14.8

Table for computing *SSE* for the Orion data

Age (yr) x	Price ($100) y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
5	85	94.16	−9.16	83.9
4	103	114.42	−11.42	130.5
6	70	73.90	−3.90	15.2
5	82	94.16	−12.16	147.9
5	89	94.16	−5.16	26.6
5	98	94.16	3.84	14.7
6	66	73.90	−7.90	62.4
6	95	73.90	21.10	445.2
2	169	154.95	14.05	197.5
7	70	53.64	16.36	267.7
7	48	53.64	−5.64	31.8
				1423.5

Squared multiple correlation coefficient

- We define R^2 as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}$$

Taking square roots, we can rewrite this as

$$R = \frac{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)\left(\hat{y}_i - \overline{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2 \sum_{i=1}^{n}\left(\hat{y}_i - \overline{y}\right)^2}}$$

Using this definition we can see that R^2 is the squared sample correlation between the observed and fitted values

Properties of R^2

- It is easily seen that if the fit is very good then

$$\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \approx 0$$

so R^2 will be close to one.

- When the fit is poor, then

$$\hat{y}_i = \hat{\beta}_0 + r_i \quad \text{and} \quad \hat{\beta}_0 \approx \overline{y}$$

so $\quad y_i - \hat{y}_i \approx 0 \quad$ and $\quad \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \approx \sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$

and consequently, R^2 will be close to zero.

Coefficient of Determination for the Airline Cost Example

$$SSE = 0.31434$$

$$SS_{YY} = \sum Y^2 - \frac{\left(\sum Y\right)^2}{n} = 270.9251 - \frac{(56.69)^2}{12} = 3.11209$$

$$r^2 = 1 - \frac{SSE}{SS_{YY}}$$

$$= 1 - \frac{.31434}{3.11209}$$

$$= .899$$

89.9% of the variability of the cost of flying a Boeing 737 is accounted for by the number of passengers.

Section 14.4

Linear Correlation

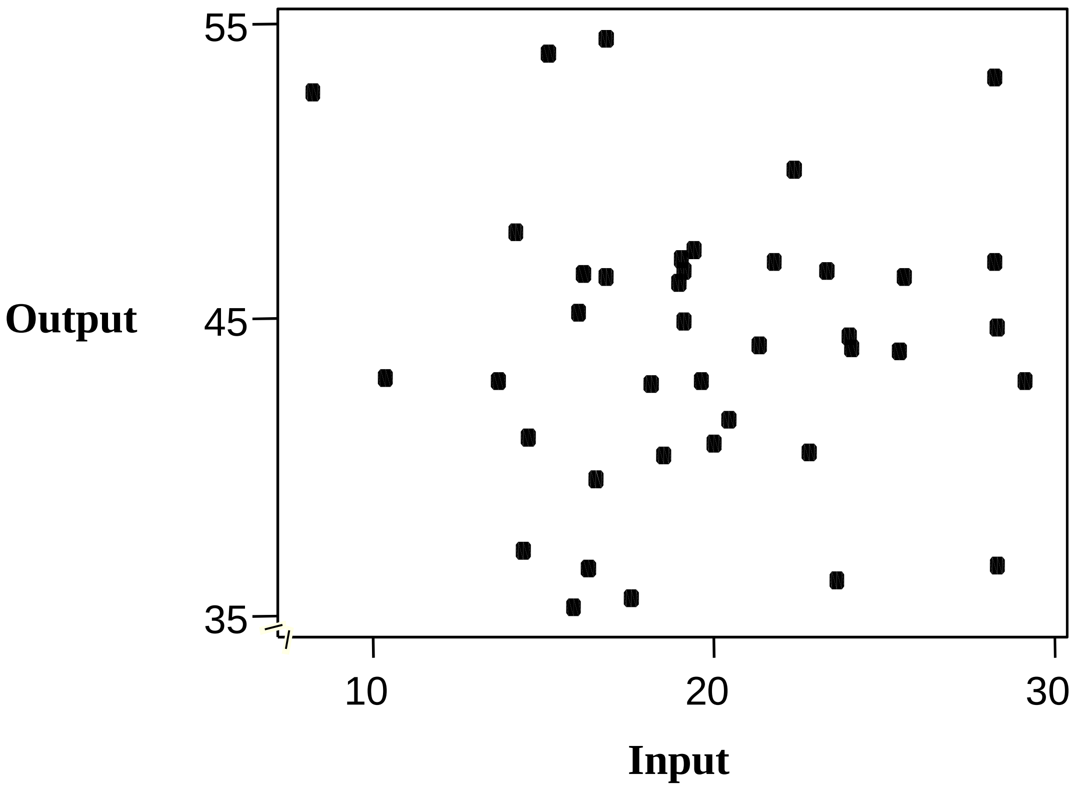
Linear Correlation

Measures the strength of a *linear* relationship between two variables

- As x increases, no definite shift in y: ***no correlation***
- As x increases, a definite shift in y: ***correlation***
- Positive correlation: x increases, y increases
- Negative correlation: x increases, y decreases
- If the ordered pairs follow a straight-line path: *linear correlation*

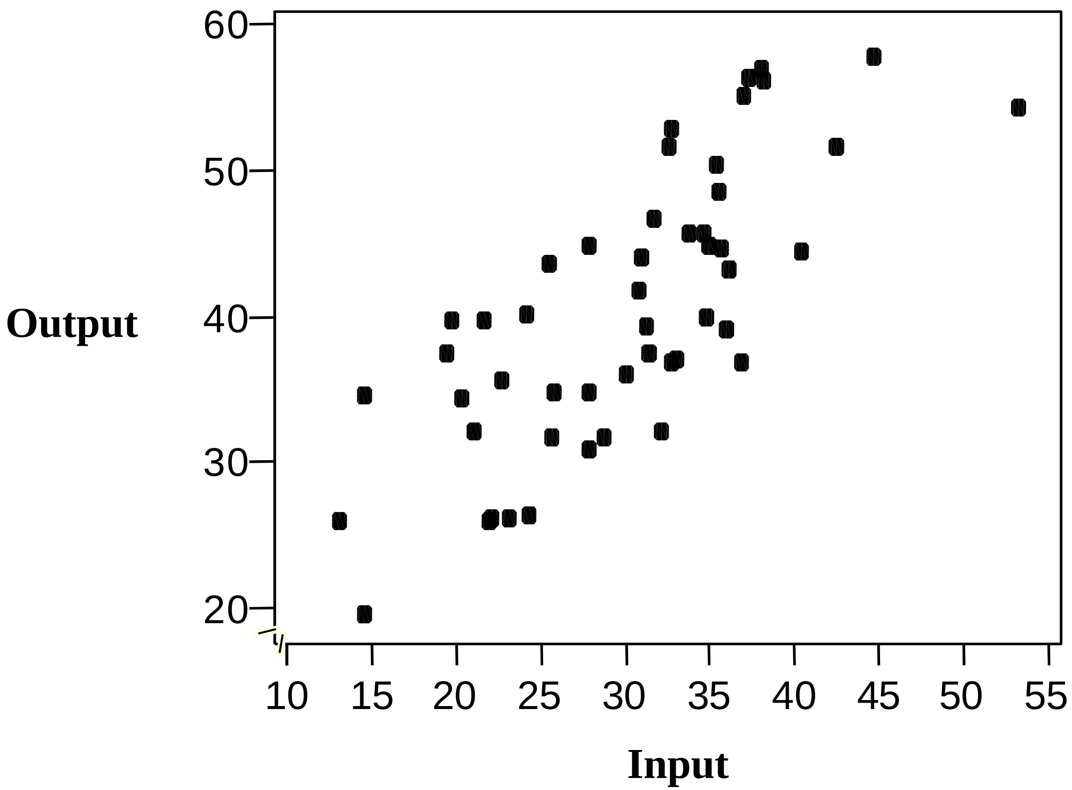
Example: No Correlation

- As *x* increases, there is no definite shift in *y:*



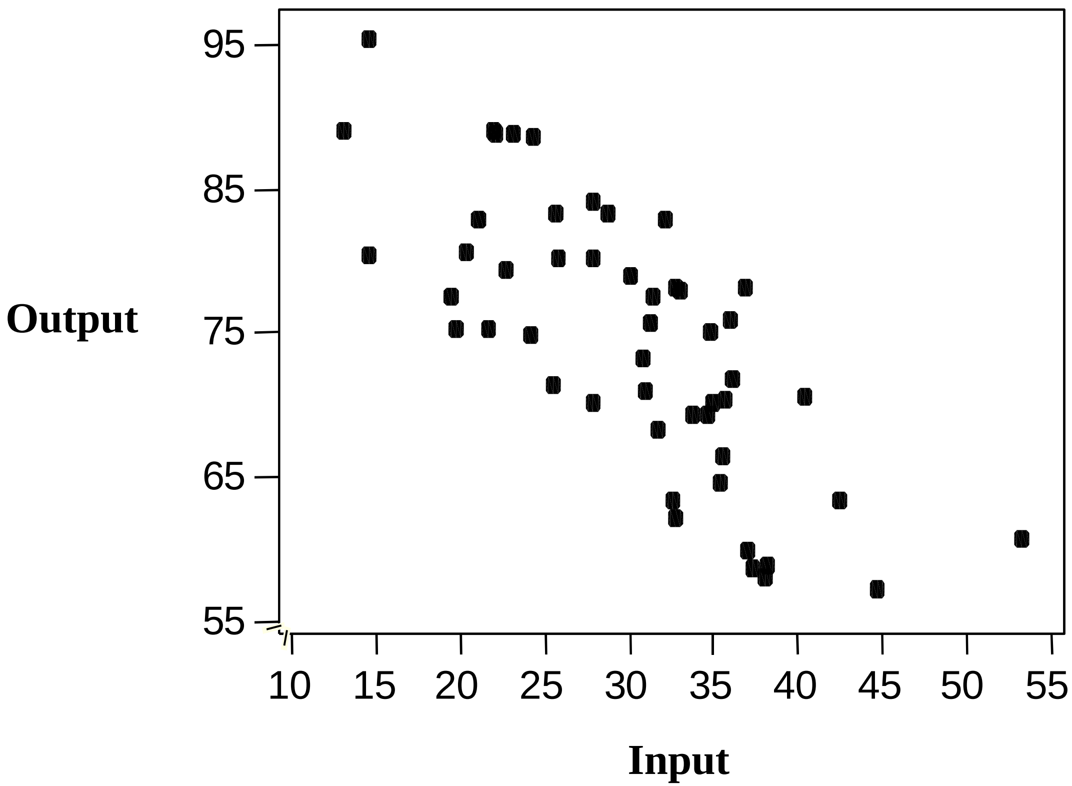
Example: Positive Correlation

- As *x* increases, *y* also increases:



Example: Negative Correlation

- As x increases, y decreases:



Please Note

- *Perfect positive correlation*: all the points lie along a line with positive slope
- *Perfect negative correlation*: all the points lie along a line with negative slope
- If the points lie along a horizontal or vertical line: *no correlation*
- If the points exhibit some other nonlinear pattern: *no linear relationship, no correlation*
- Need some way to measure correlation

Definition 14.7 & Formula 14.3

Linear Correlation Coefficient

For a set of n data points, the **linear correlation coefficient, *r*,** is defined by

$$r = \frac{\frac{1}{n-1}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y},$$

where s_x and s_y denote the sample standard deviations of the x-values and y-values, respectively.

Using algebra, we can show that the linear correlation coefficient can be expressed as $r = S_{xy}/\sqrt{S_{xx}S_{yy}}$, where S_{xx}, S_{xy}, and S_{yy} are given in Definition 14.3 on page 637. Referring again to that definition, we get Formula 14.3.

Computing Formula for a Linear Correlation Coefficient

The computing formula for a linear correlation coefficient is

$$r = \frac{\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n}{\sqrt{[\Sigma x_i^2 - (\Sigma x_i)^2/n][\Sigma y_i^2 - (\Sigma y_i)^2/n]}}.$$

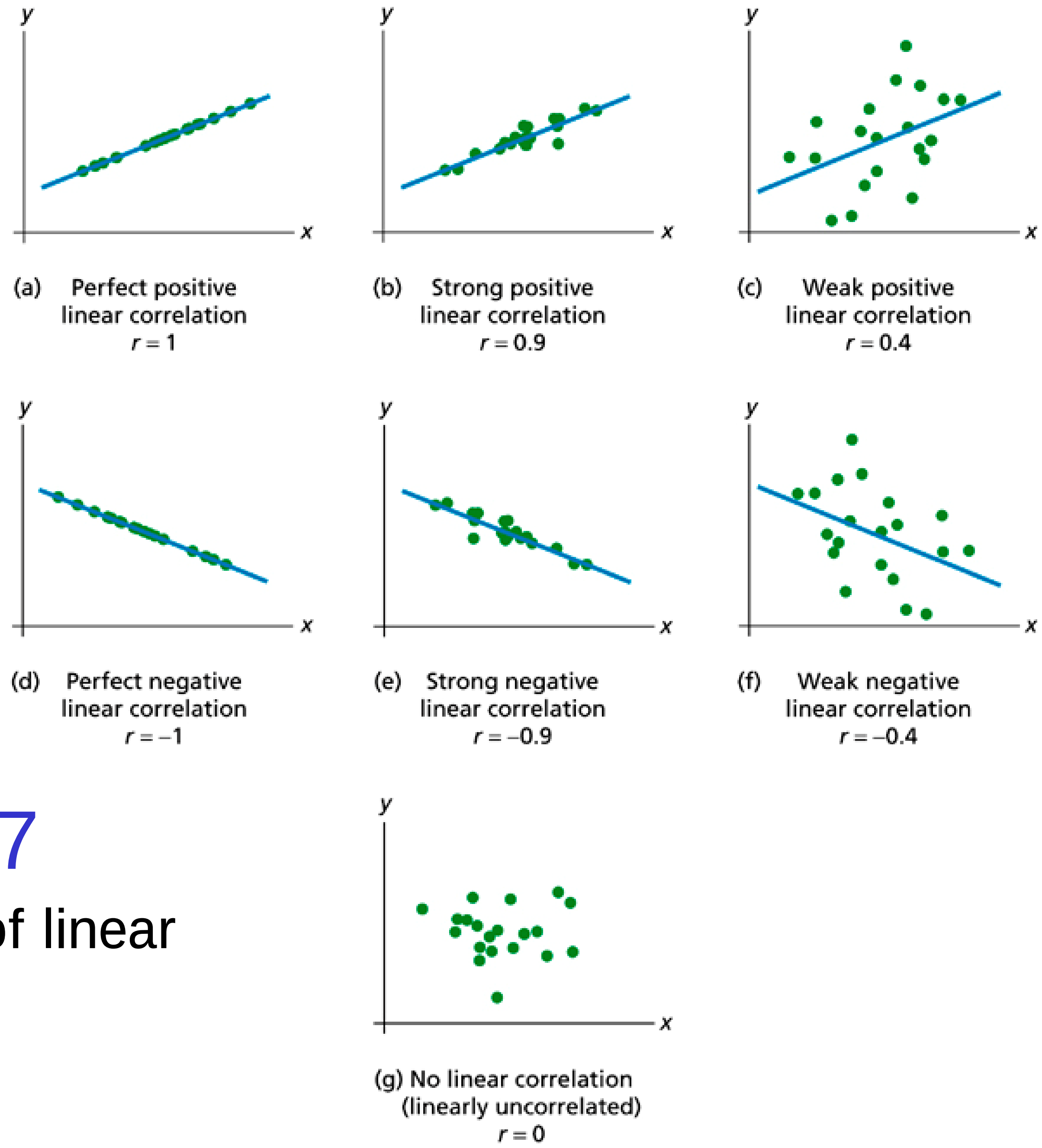


Figure 14.17

Various degrees of linear correlation

Example

- **Example:** The table below presents the weight (in thousands of pounds) x and the gasoline mileage (miles per gallon) y for ten different automobiles. Find the linear correlation coefficient:

	x	y	x^2	y^2	xy
	2.5	40	6.25	1600	100.0
	3.0	43	9.00	1849	129.0
	4.0	30	16.00	900	120.0
	3.5	35	12.25	1225	122.5
	2.7	42	7.29	1764	113.4
	4.5	19	20.25	361	85.5
	3.8	32	14.44	1024	121.6
	2.9	39	8.41	1521	113.1
	5.0	15	25.00	225	75.0
	2.2	14	4.84	196	30.8
Sum	34.1	309	123.73	10665	1010.9
	$\sum x$	$\sum y$	$\sum x^2$	$\sum y^2$	$\sum xy$

模式摘要[b]

模式	R	R平方	調過後的 R平方	估計的標準誤
1	.469[a]	.220	.123	10.435

a. 預測變數：(常數), weight of car (in thousands of pounds)

b. 依變數：gasoline mileage (miles per gallon)

變異數分析[b]

模式		平方和	自由度	平均平方和	F檢定	顯著性
1	迴歸	245.803	1	245.803	2.257	.171[a]
	殘差	871.097	8	108.887		
	總和	1116.900	9			

a. 預測變數：(常數), weight of car (in thousands of pounds)

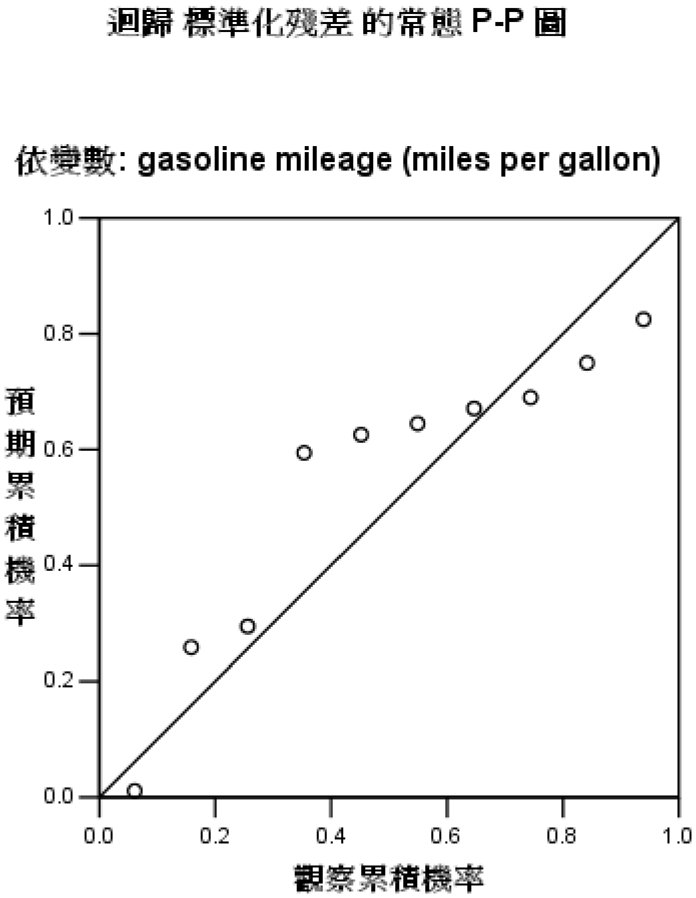
b. 依變數：gasoline mileage (miles per gallon)

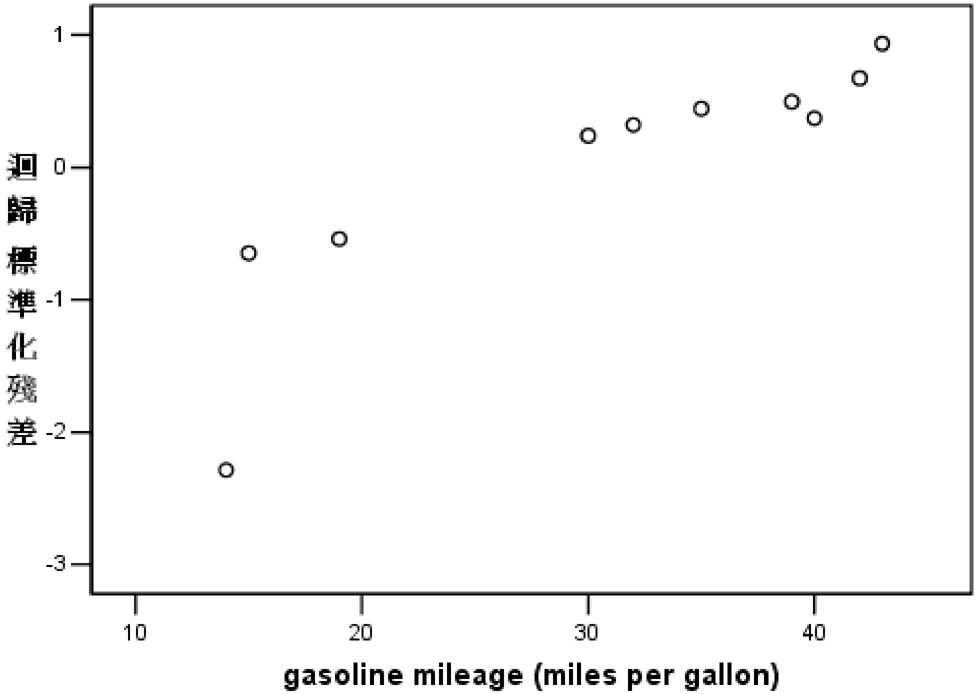
係數[a]

模式		未標準化係數		標準化係數	t	顯著性
		B之估計值	標準誤	Beta分配		
1	(常數)	50.488	13.449		3.754	.006
	weight of car (in thousands of pounds)	-5.744	3.823	-.469	-1.502	.171

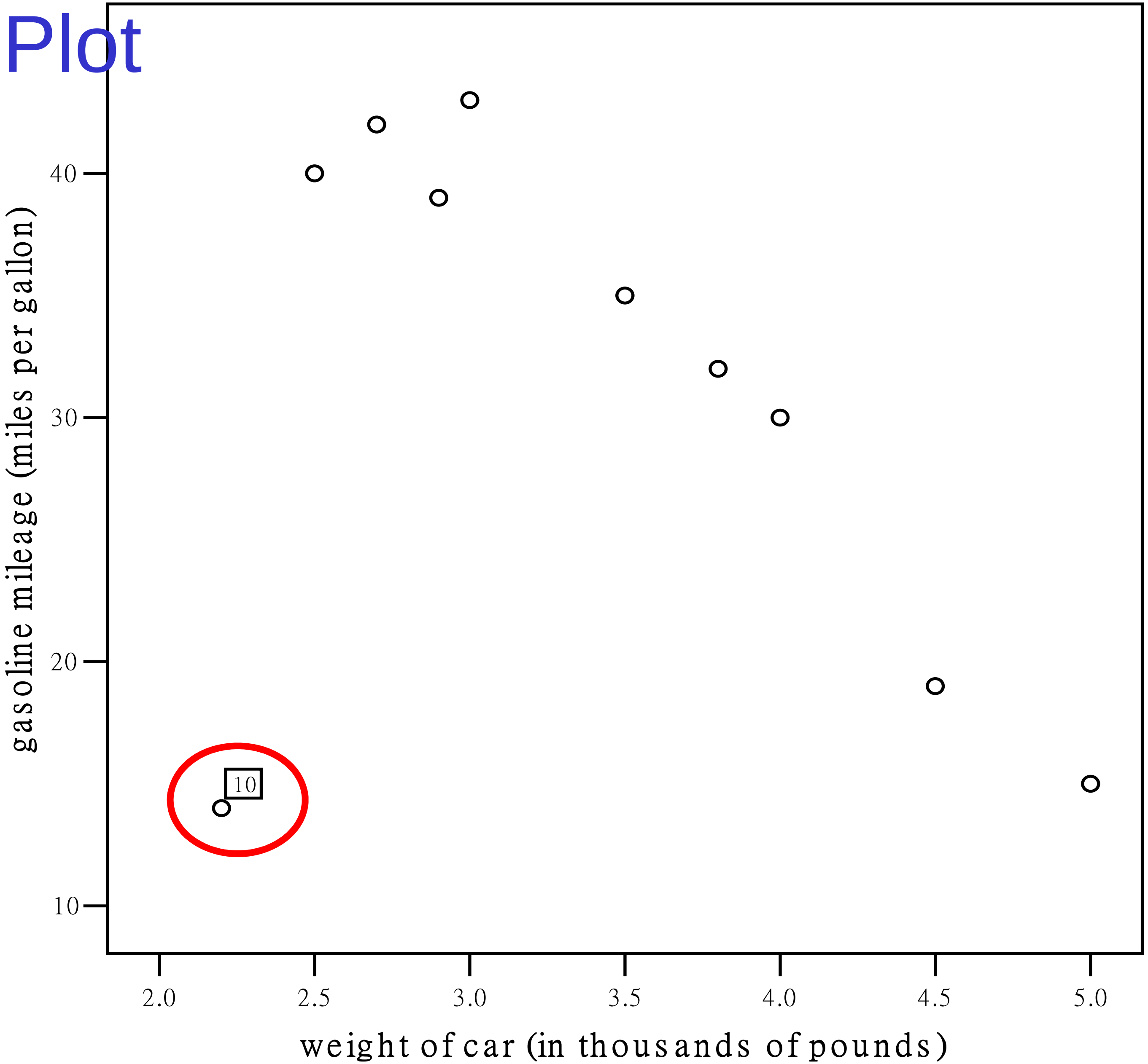
a. 依變數：gasoline mileage (miles per gallon)

迴歸 標準化殘差與 mileage 的散佈圖





A Scatter Plot



What if we remove the point?

模式摘要[b]

模式	R	R平方	調過後的 R平方	估計的標準誤
1	.960[a]	.922	.911	2.986

a. 預測變數：(常數), weight of car (in thousands of pounds)

b. 依變數：gasoline mileage (miles per gallon)

變異數分析[b]

模式		平方和	自由度	平均平方和	F檢定	顯著性
1	迴歸	737.125	1	737.125	82.650	.000[a]
	殘差	62.431	7	8.919		
	總和	799.556	8			

a. 預測變數：(常數), weight of car (in thousands of pounds)

b. 依變數：gasoline mileage (miles per gallon)

係數[a]

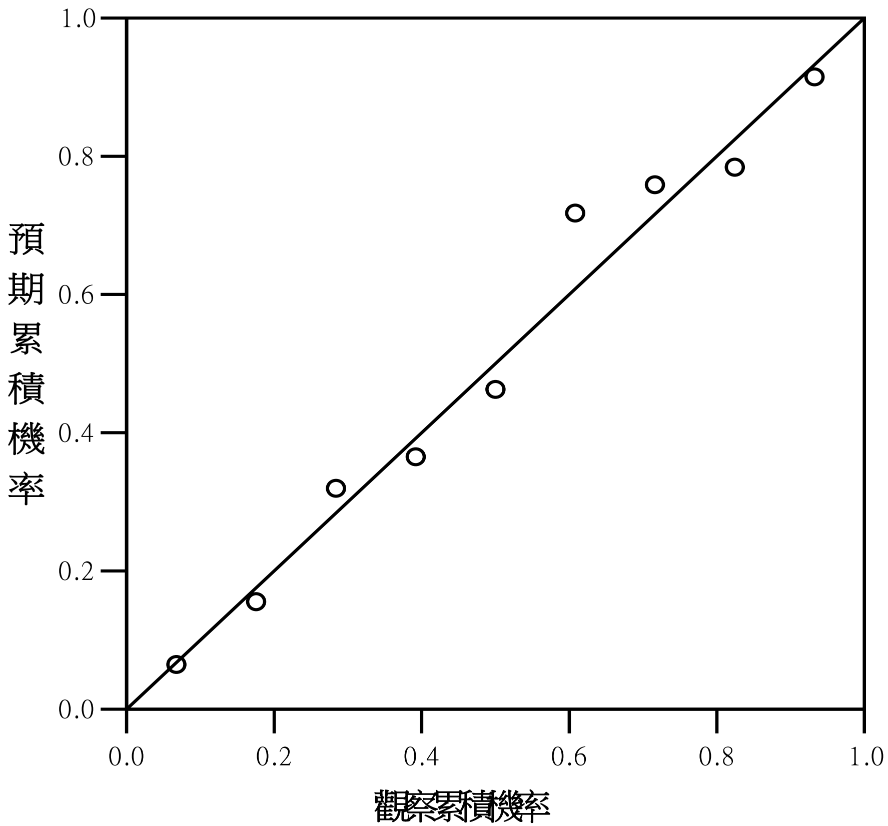
模式		未標準化係數		標準化係數	t	顯著性
		B之估計值	標準誤	Beta分配		
1	(常數)	72.660	4.498		16.152	.000
	weight of car (in thousands of pounds)	-11.252	1.238	-.960	-9.091	.000

a. 依變數：gasoline mileage (miles per gallon)

What if we remove the point?

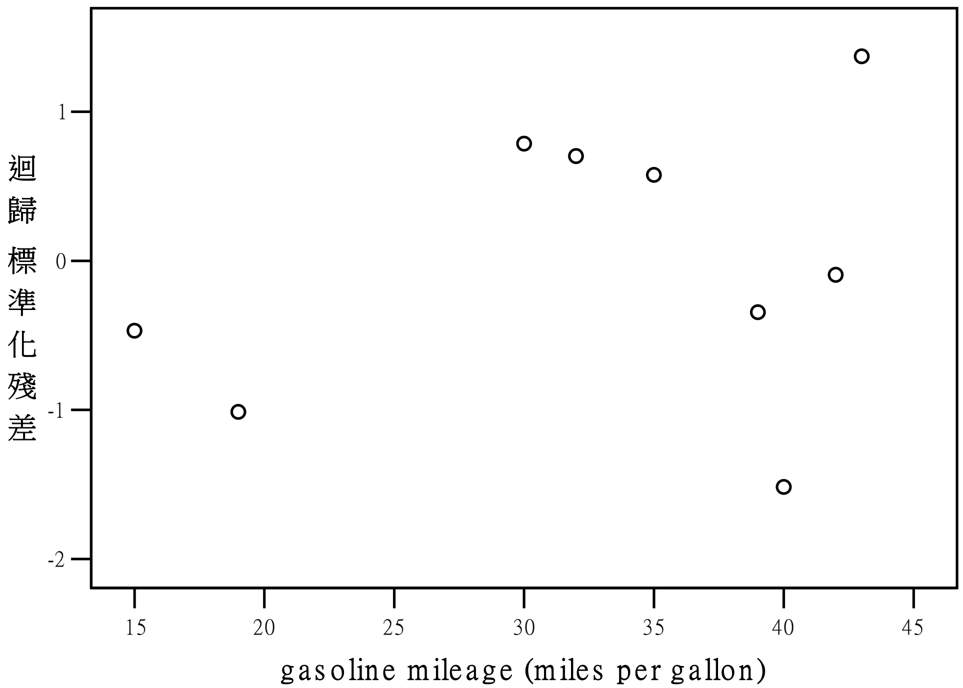
迴歸標準化殘差的常態P-P 圖

依變數 gasoline mileage (miles per gallon)



散佈圖

依變數 gasoline mileage (miles per gallon)



Example: In a study involving children's fear related to being hospitalized, the age and the score each child made on the Child Medical Fear Scale (CMFS) are given in the table below:

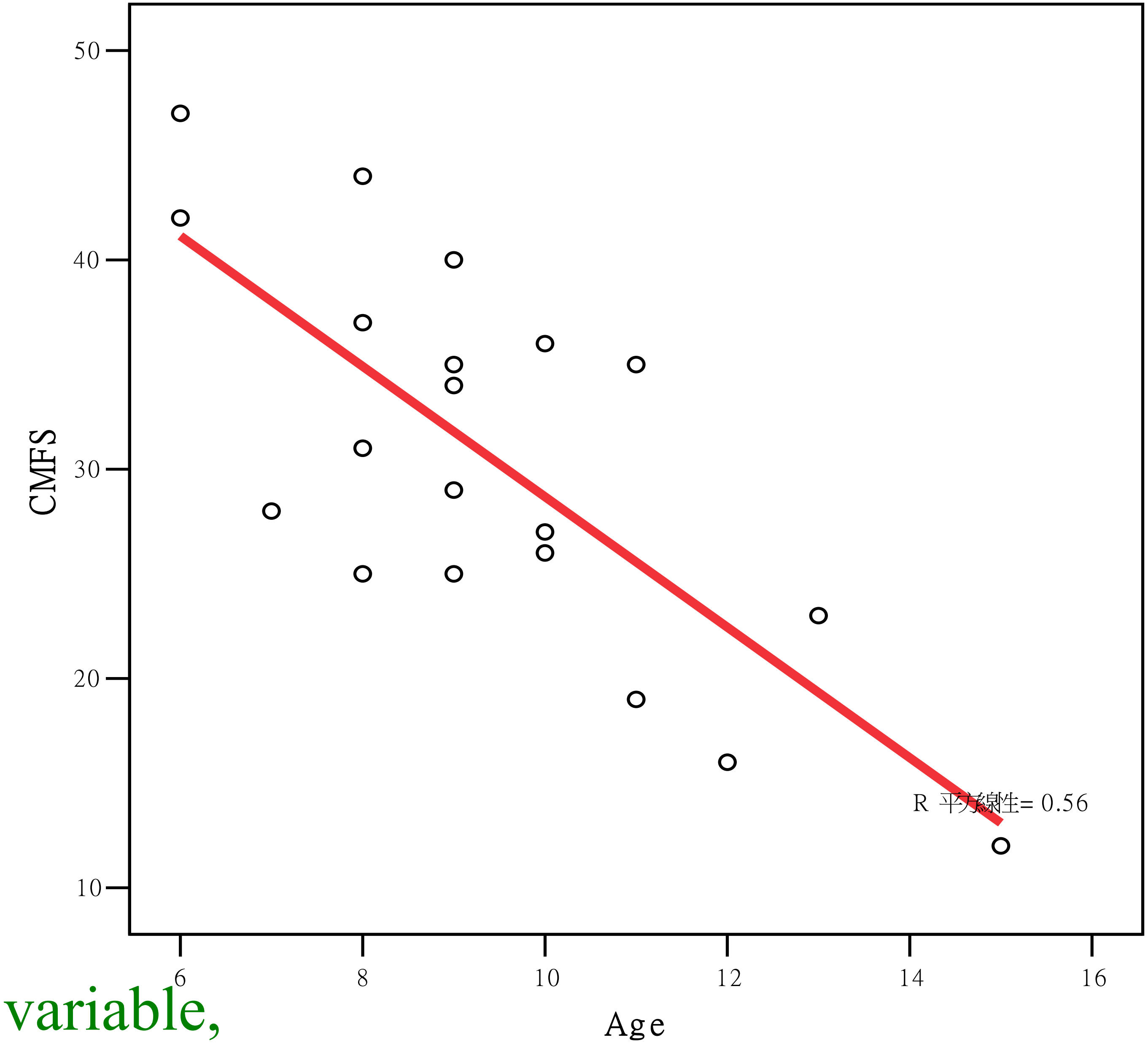
Age (x)	8	9	9	10	11	9	8	9	8	11
CMFS (y)	31	25	40	27	35	29	25	34	44	19

Age (x)	7	6	6	8	9	12	15	13	10	10
CMFS (y)	28	47	42	37	35	16	12	23	26	36

Construct a scatter diagram for this data and build a regression model.

Solution

Child Medical Fear Scale



- age = input variable,
- CMFS = output variable

模式摘要b

模式	R	R平方	調過後的 R平方	估計的標準誤
1	.748[a]	.560	.535	6.344

a. 預測變數：(常數), Age

b. 依變數：CMFS

變異數分析b

模式		平方和	自由度	平均平方和	F檢定	顯著性
1	迴歸	920.476	1	920.476	22.870	.000[a]
	殘差	724.474	18	40.249		
	總和	1644.950	19			

a. 預測變數：(常數), Age

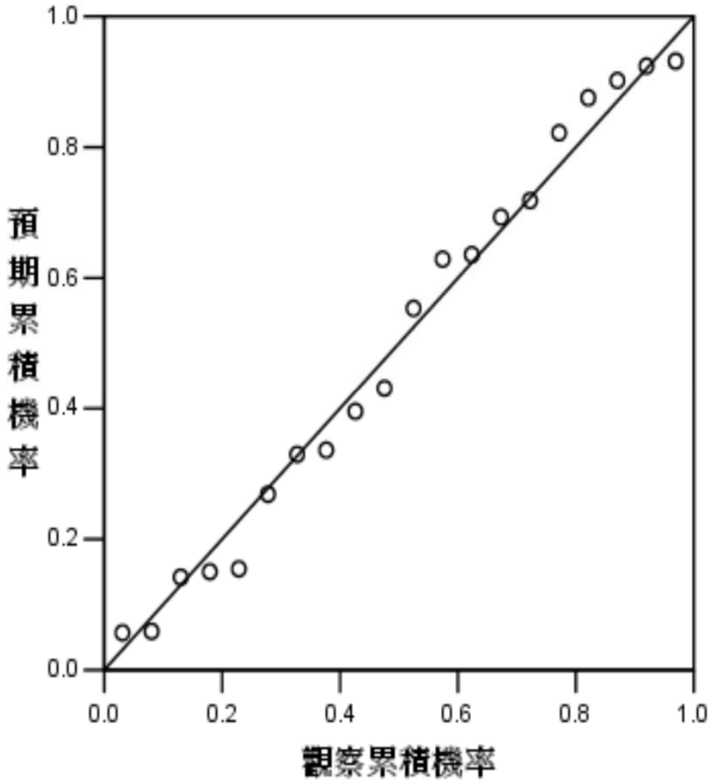
b. 依變數：CMFS

係數a

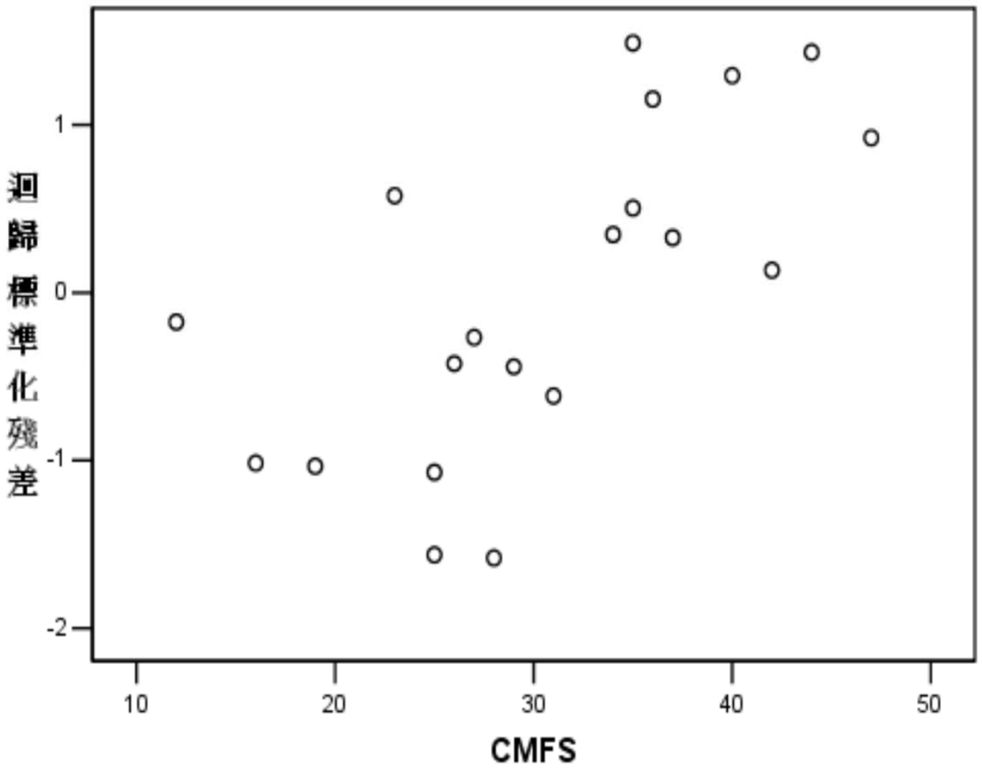
模式		未標準化係數		標準化係數	t	顯著性	迴歸係數B的95% 信賴區間	
		B之估計值	標準誤	Beta 分配			下限	上限
1	(常數)	59.841	6.287		9.518	.000	46.632	73.049
	Age	-3.116	.652	-.748	-4.782	.000	-4.485	-1.747

a. 依變數：CMFS

標準化殘差 的常態 P-P 圖



迴歸 標準化殘差與 CMFS 的散佈圖



Section 15.1

The Regression Model; Analysis of Residuals

Figure 15.1

Population regression line

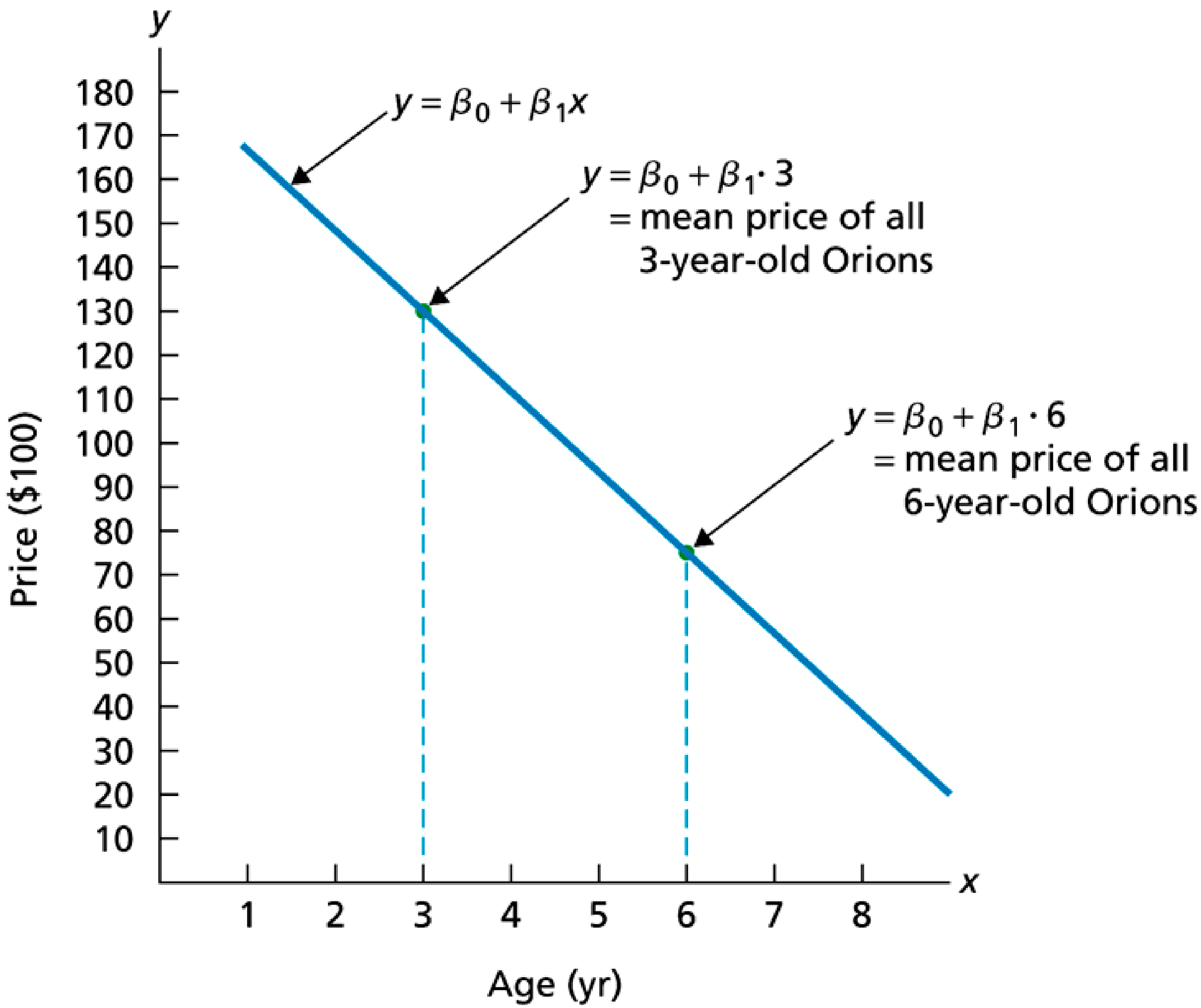


Figure 15.2

Price distributions for 2-, 5-, and 7-year-old Orions under Assumptions 2 and 3 (The means shown for the three normal distributions reflect Assumption 1)

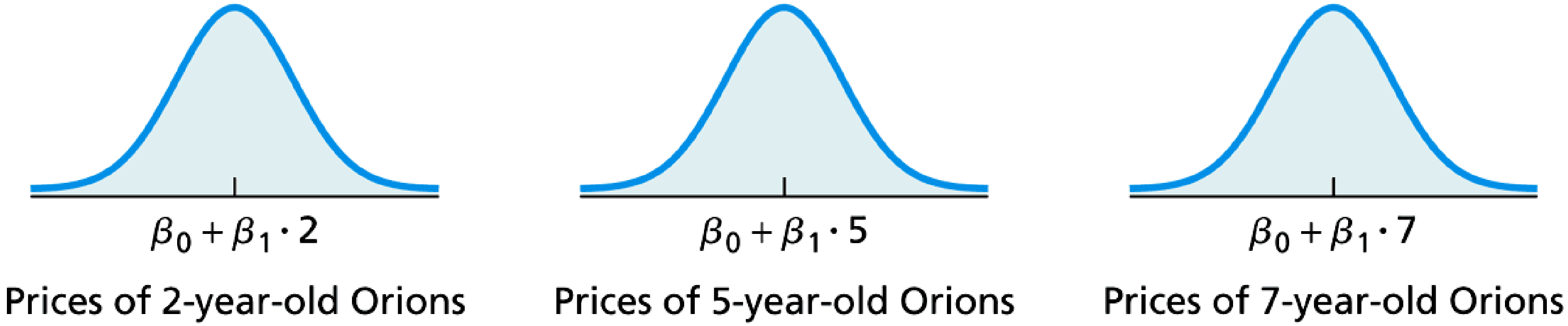


Figure 15.3

Graphical portrayal of Assumptions 1–3 for regression inferences pertaining to age and price of Orions

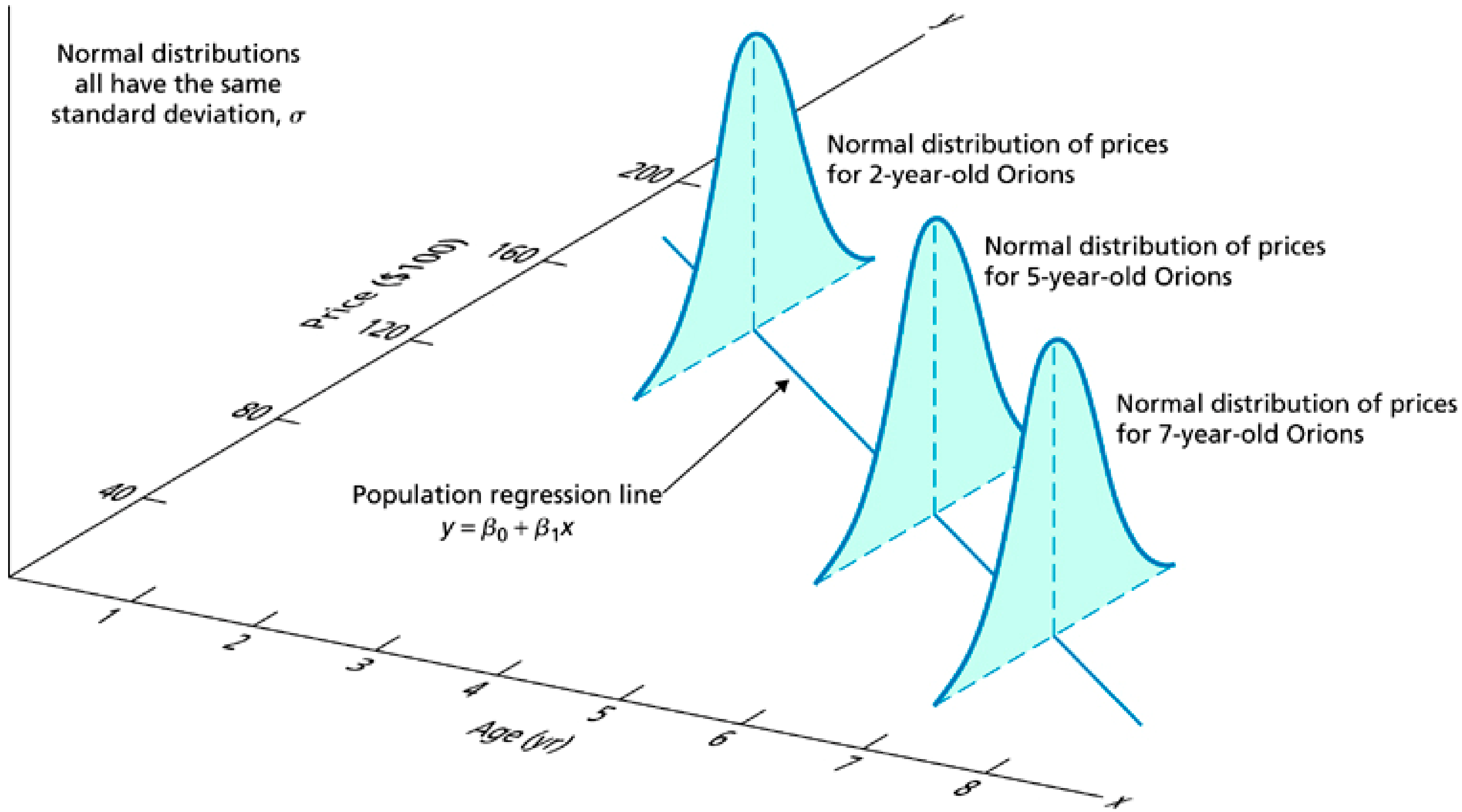
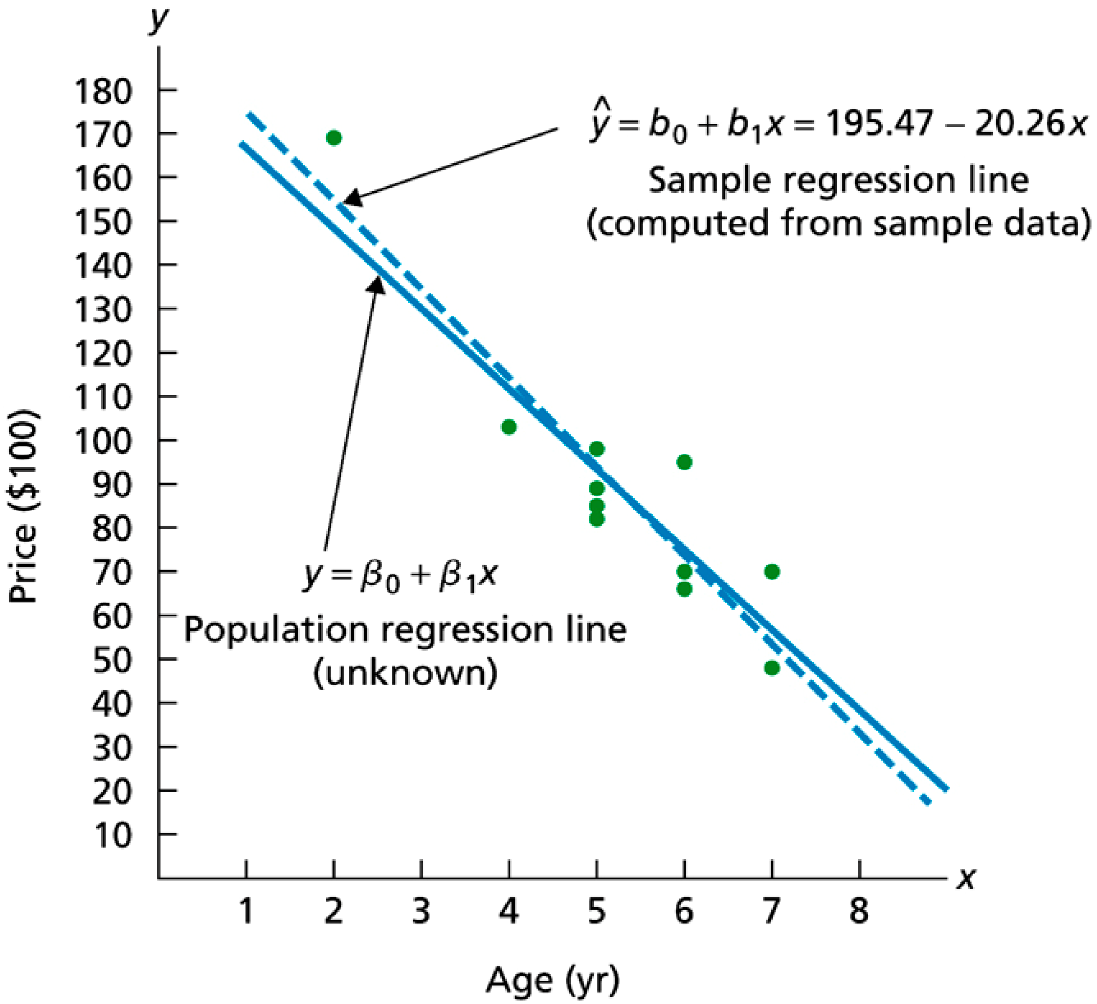


Figure 15.4

Population regression line and sample regression line for age and price of Orions



Definition 15.1

Standard Error of the Estimate

The **standard error of the estimate, s_e,** is defined by

$$s_e = \sqrt{\frac{SSE}{n-2}},$$

where *SSE* is the error sum of squares.

Estimating the Variance of the Experimental Error

Assumption: The distribution of y's is approximately normal and the variances of the distributions of y at all values of x are the same (The standard deviation of the distribution of y about $\hat{y}$ is the same for all values of x)

Consider the sample variance: $s^2 = \dfrac{\sum (x - \bar{x})^2}{n-1}$

1. The variance of y involves an additional complication: there is a different mean for y at each value of x
2. Each "mean" is actually the predicted value, ŷ
3. Variance of the error e estimated by: Degrees of freedom: $n - 2$ $s_e^2 = \dfrac{\sum (y - \hat{y})^2}{n-2}$

Alternative (Computational) Formula for Variance of Experimental Error

Rewriting s_e^2:

$$s_e^2 = \frac{\sum (y-\hat{y})^2}{n-2}$$

$$= \frac{\sum (y-b_0-b_1 x)^2}{n-2}$$

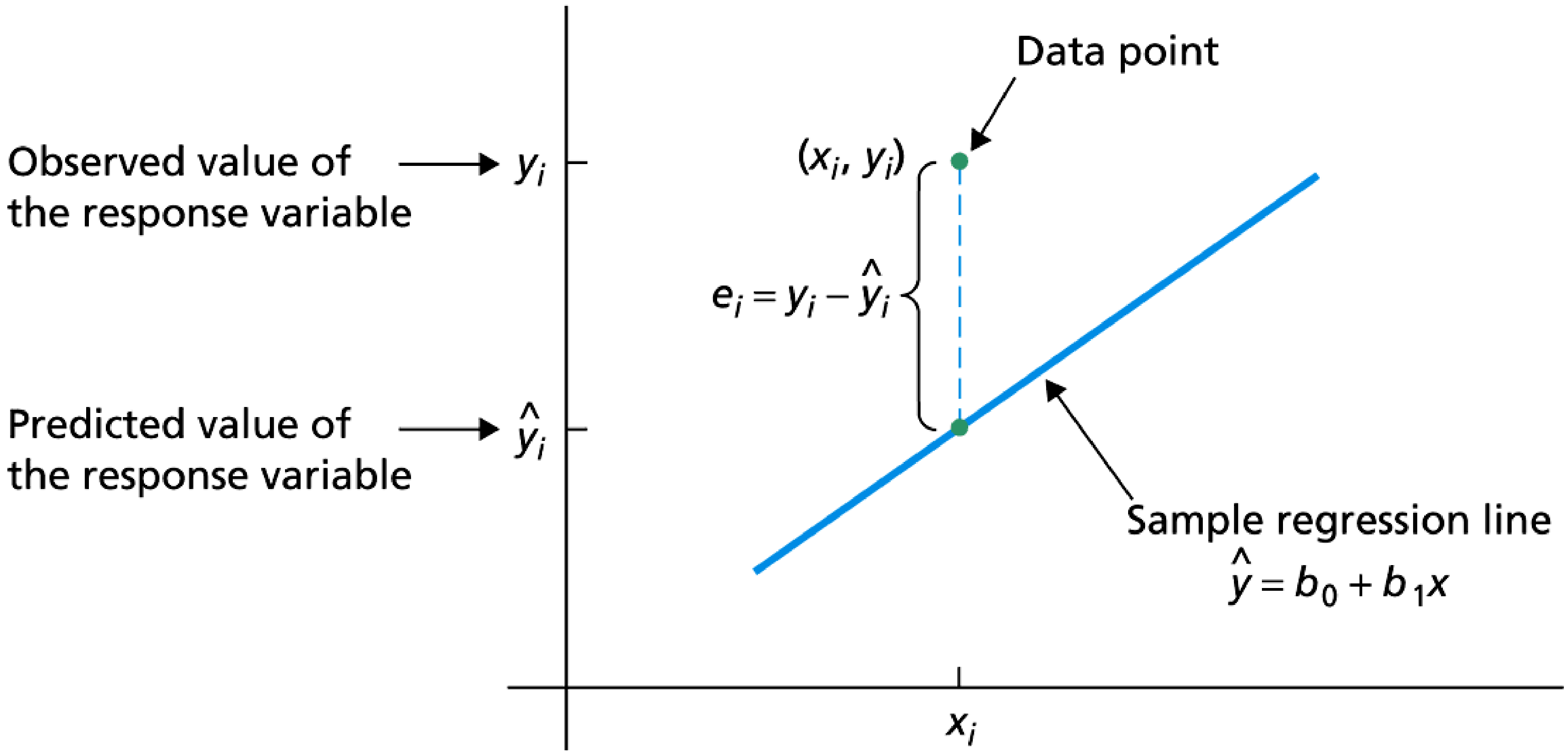
$$= \frac{\left(\sum y^2\right)-(b_0)\left(\sum y\right)-(b_1)\left(\sum xy\right)}{n-2}$$

$$= \frac{\text{SSE}}{n-2}$$

SSE = sum of squares for error

Figure 15.5

Residual of a data point



Standard Error of the Estimate

Sum of Squares Error

$$SSE = \sum\left(Y - \hat{Y}\right)^2$$

$$= \sum Y^2 - b_0 \sum Y - b_1 \sum XY$$

Standard Error of the Estimate

$$S_e = \sqrt{\frac{SSE}{n-2}}$$

Example

✓ **Example:** A recent study was conducted to determine the relation between advertising expenditures and sales of statistics texts (for the first year in print). The data is given below (in thousands). Find the line of best fit and the variance of y about the line of best fit.

Adv. Costs (x)	Sales (y)	Adv. Costs (x)	Sales (y)
40	289	60	470
55	423	52	408
35	250	39	320
50	400	47	415
43	335	38	389

Solution

$$\mathrm{SS}(x) = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 21677 - \frac{(459)^2}{10} = 608.9$$

$$\mathrm{SS}(xy) = \sum xy - \frac{\sum x \sum y}{n} = 174163 - \frac{(459)(3699)}{10} = 4378.9$$

$$b_1 = \frac{\mathrm{SS}(xy)}{\mathrm{SS}(x)} = \frac{4378.9}{608.9} = 7.1915$$

$$b_0 = \frac{\sum y - \left(b_1 \cdot \sum x\right)}{n} = \frac{3699 - (7.1915)(459)}{10} = 39.8105$$

Solution Continued

- The equation for the line of best fit: $\hat{y} = 39.81 + 7.19x$

- The variance of y about the regression line:

$$s_e^2 = \frac{\left(\sum y^2\right) - (b_0)\left(\sum y\right) - (b_1)\left(\sum xy\right)}{n-2}$$

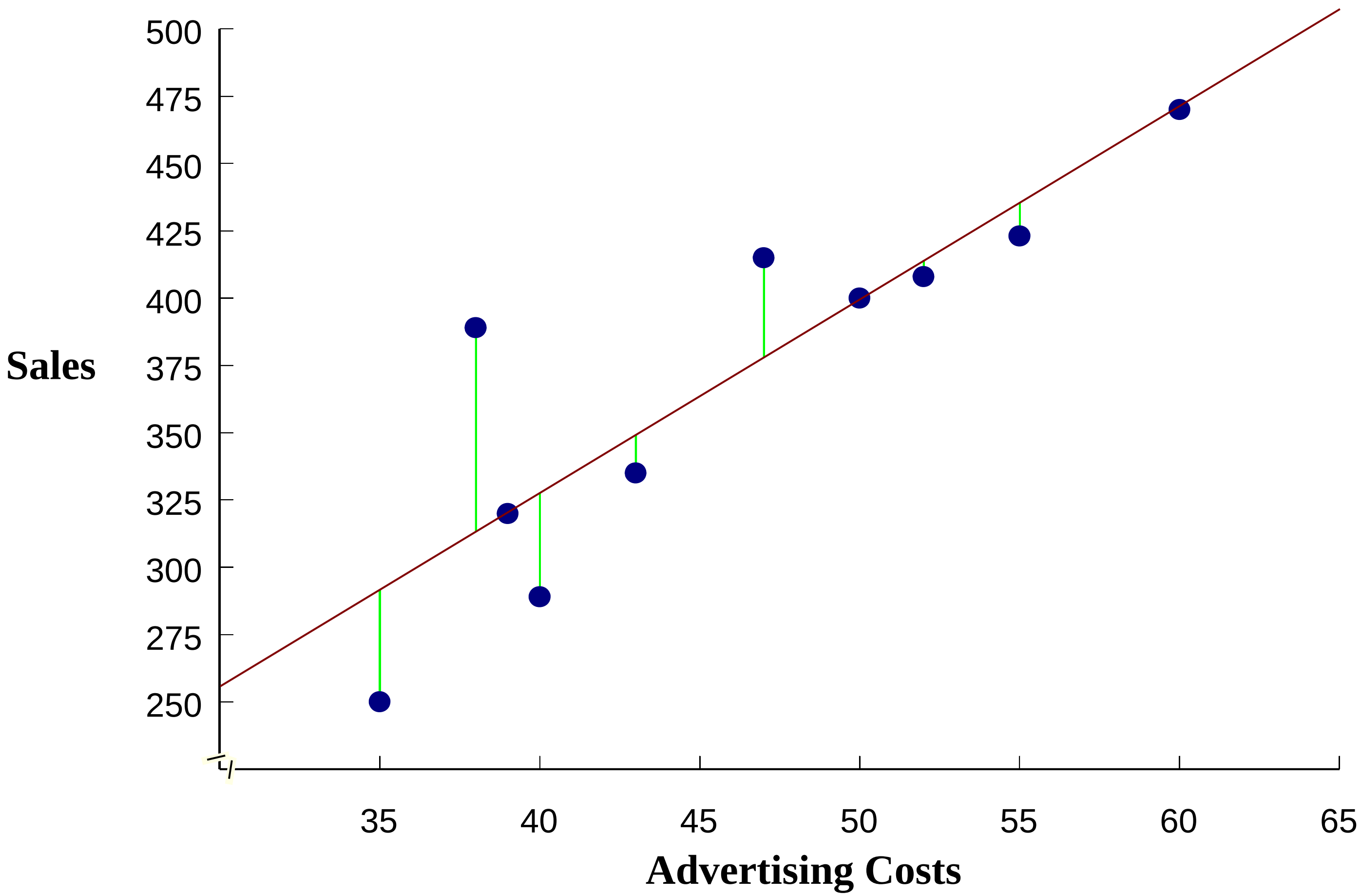
$$= \frac{(1410485) - (39.81)(3699) - (7.1915)(174163)}{8}$$

$$= \frac{10734.5955}{8} = 1341.8244$$

Note: Extra decimal places are often needed for this type of calculation

Illustration

- Scatter diagram, regression line, and random errors as line segments:



Residual Analysis

In the regression model, we made some assumptions. Now, we will investigate the cases when the assumptions do not hold. Here are some cases that we will investigate:

1. The regression function is not linear.
2. The distribution of Y do not have constant variances at all level of X (i.e. the error terms do not have constant variances).
3. The distributions of Y are not normal (i.e. the error terms are not normally distributed).
4. The error terms are not independent.

The linear model:

Let's look at the linear regression model used:

$$Y_i = a + b X_i + \varepsilon_i$$

where ε_i is the error term.

A residual plot is a very useful method to indicate a solution - plot the residual against the fitted value as scatter plot :

$$\varepsilon_i (= Y_i - E[Y_i]) \text{ v.s. } \hat{Y}_i.$$

Residual Plot

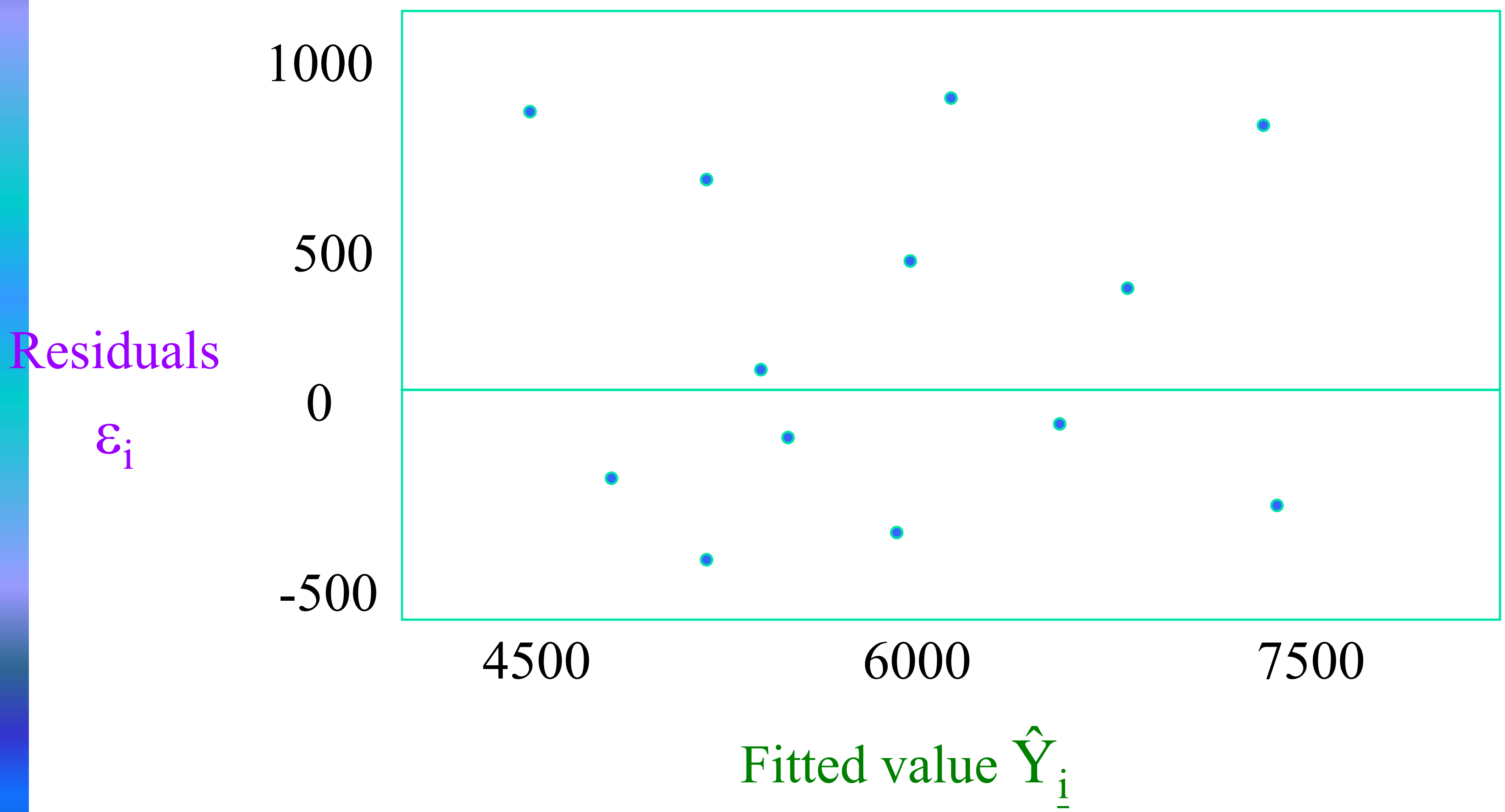


Figure 15.6

Residual plots suggesting (a) no violation of linearity or constant standard deviation, (b) violation of linearity, and (c) violation of constant standard deviation

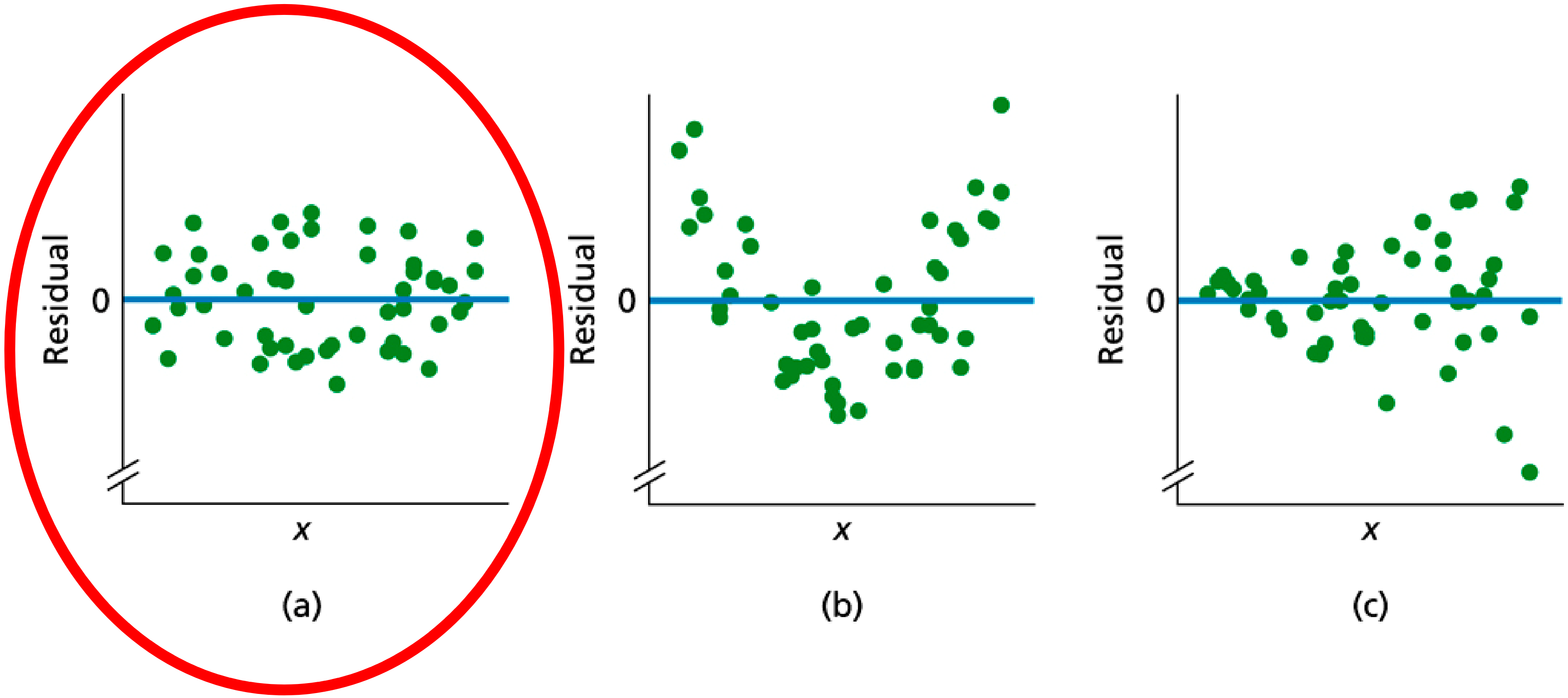
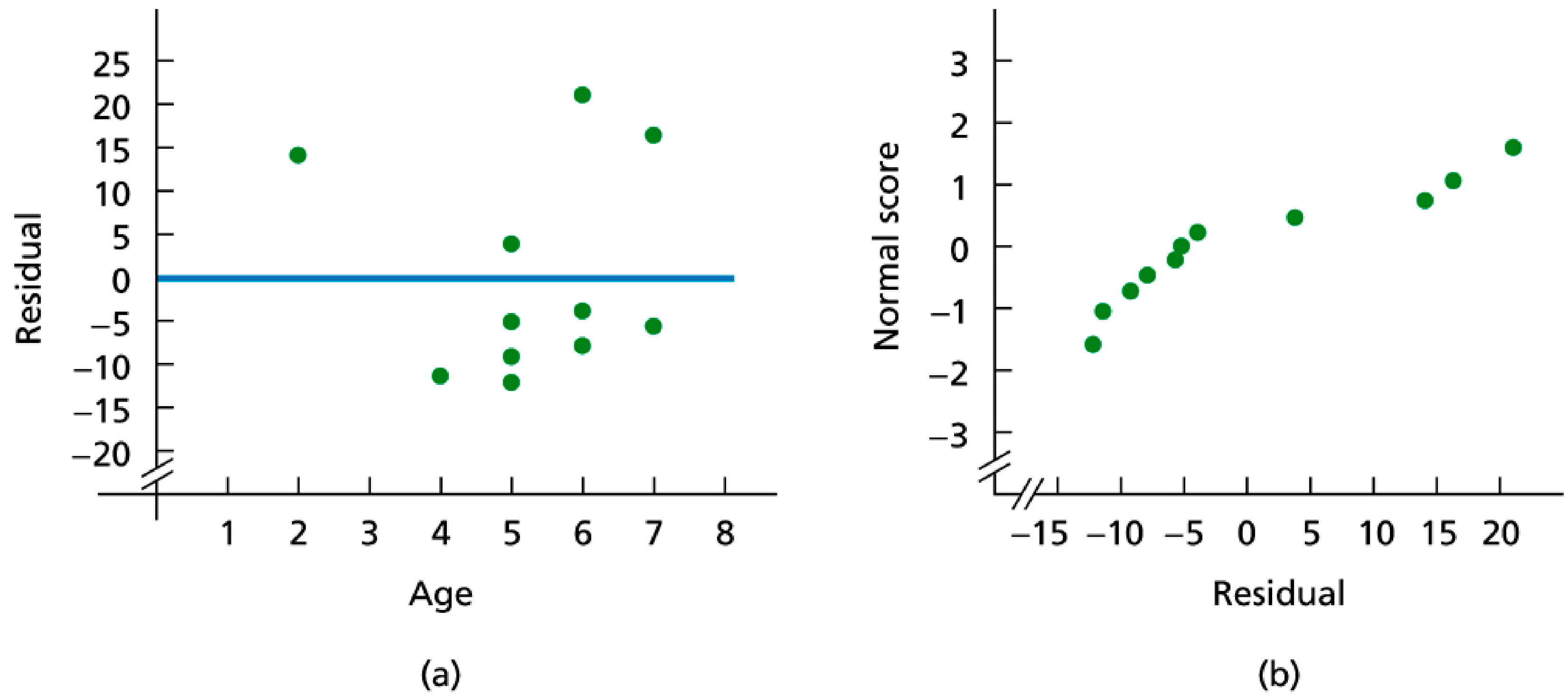


Figure 15.7

(a) Residual plot; (b) normal probability plot for residuals



Equality of variance

- It is possible to have normality without having equality of variance, i.e. in some situations we fit the model

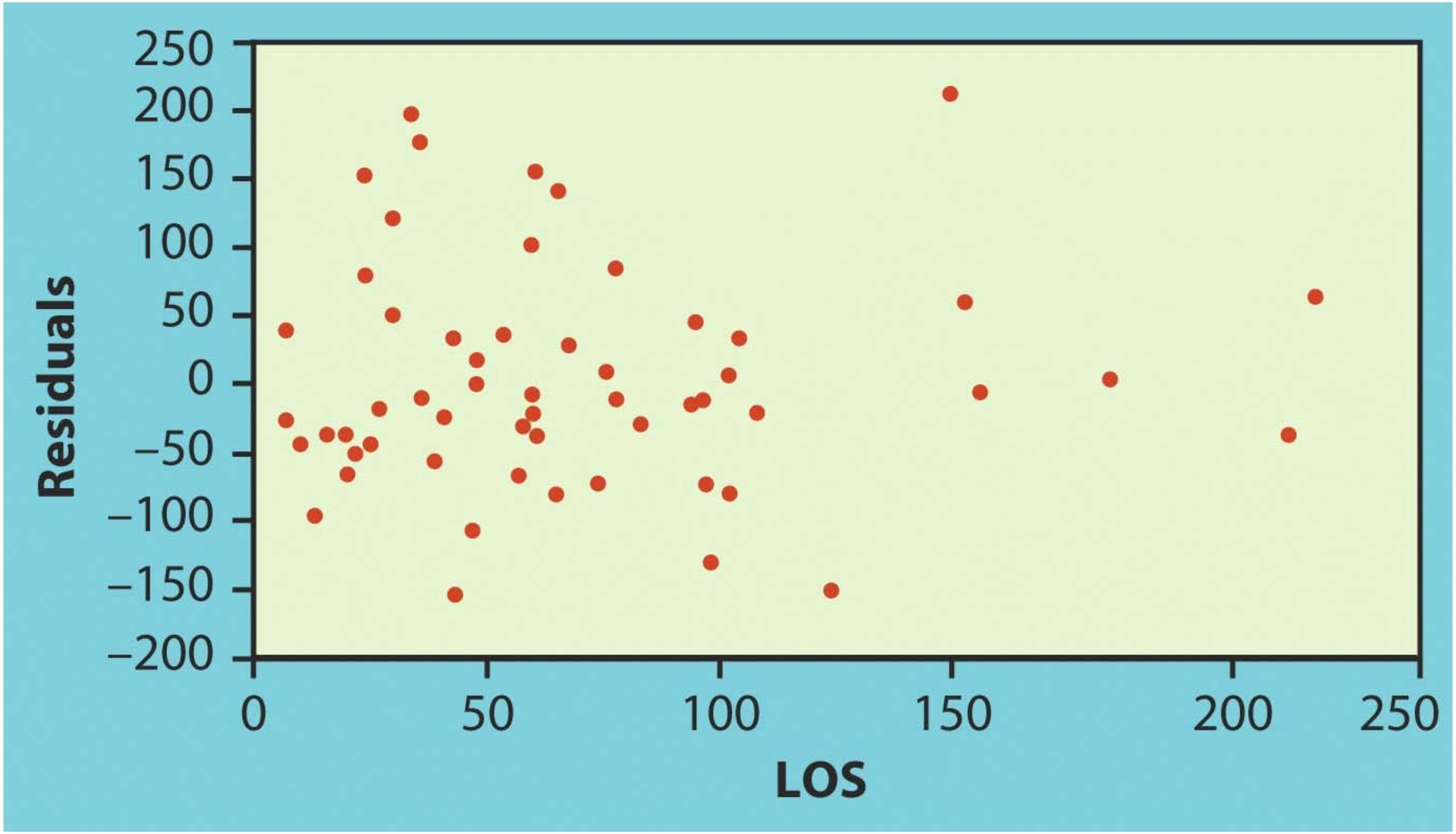
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \text{ iid. } N(0, \sigma_i^2)$$

- However, we did not fit this model to this set of data, we assumed that we had equal variances for every error, i.e.

$$\sigma_i = \sigma, \forall i$$

- We check this assumption, as before, with a residuals plot
- This time however, we will generally see less patterning, because the data are not grouped

An Example of a random residuals plot



Interpreting residuals plots

- If we have strong patterns in the residuals plot then this can mean a number of things

1. The equality of variance assumption has been violated – this is usually shown by a funnel shape in the plot
2. The simple linear model did not explain the trend in the data, i.e. there is some trend that still exists in the data which might require the addition of extra model terms – this is more likely in multivariate regression
3. The data require transformation before a linear model is appropriate

Models or Prediction Equations

- Some examples of various possible relationships:

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

Quadratic: $\hat{y} = (a + bx)^2$

Exponential: $\hat{y} = a(b^x)$

Logarithmic: $\hat{y} = a \log_b x$

Reciprocal: $\hat{y} = \dfrac{1}{a + bx}$

Note: What would a scatter diagram look like to suggest each relationship?

Nonlinear Regression Models: Model Transformation

$$\hat{y} = ab^x$$

$$\Rightarrow \log(\hat{y}) = \log(a) + x\log(b)$$

$$\Rightarrow \hat{y}' = a' + b'x$$

where:

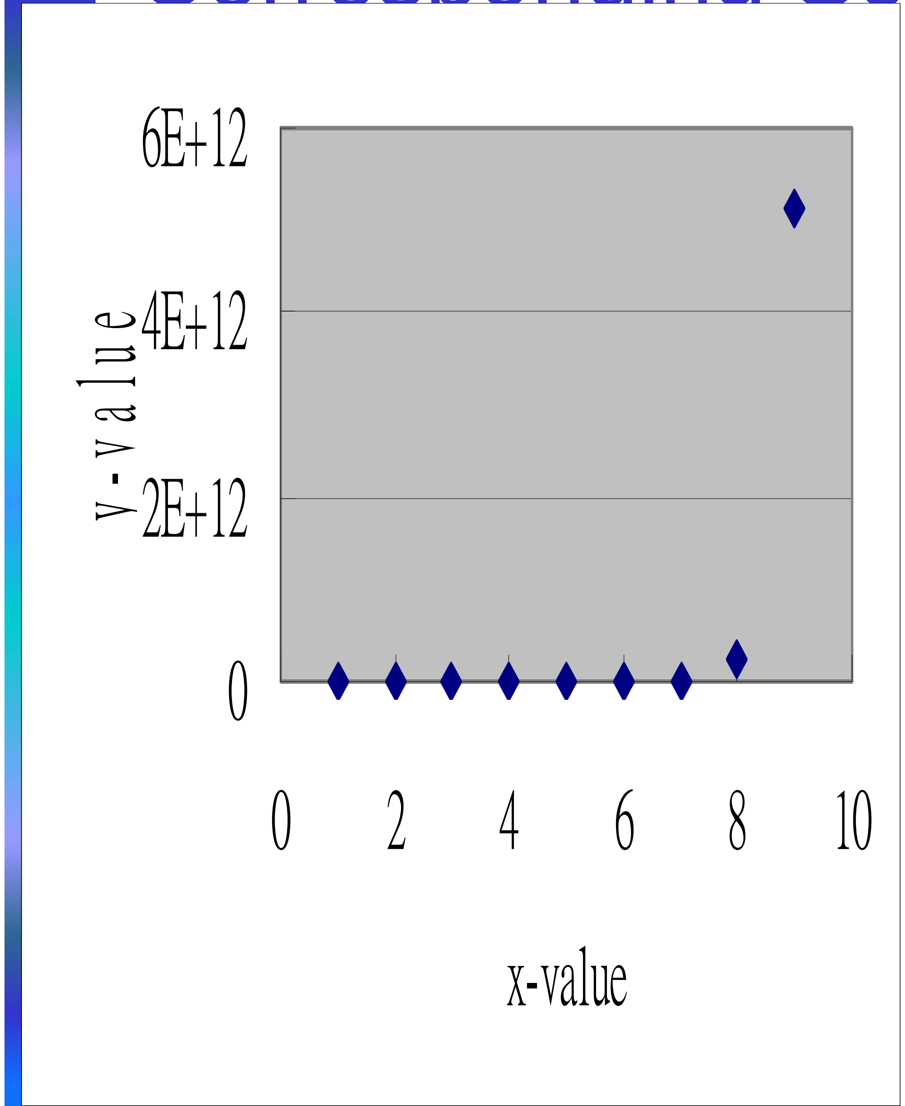
$$\hat{y}' = \log(\hat{y})$$

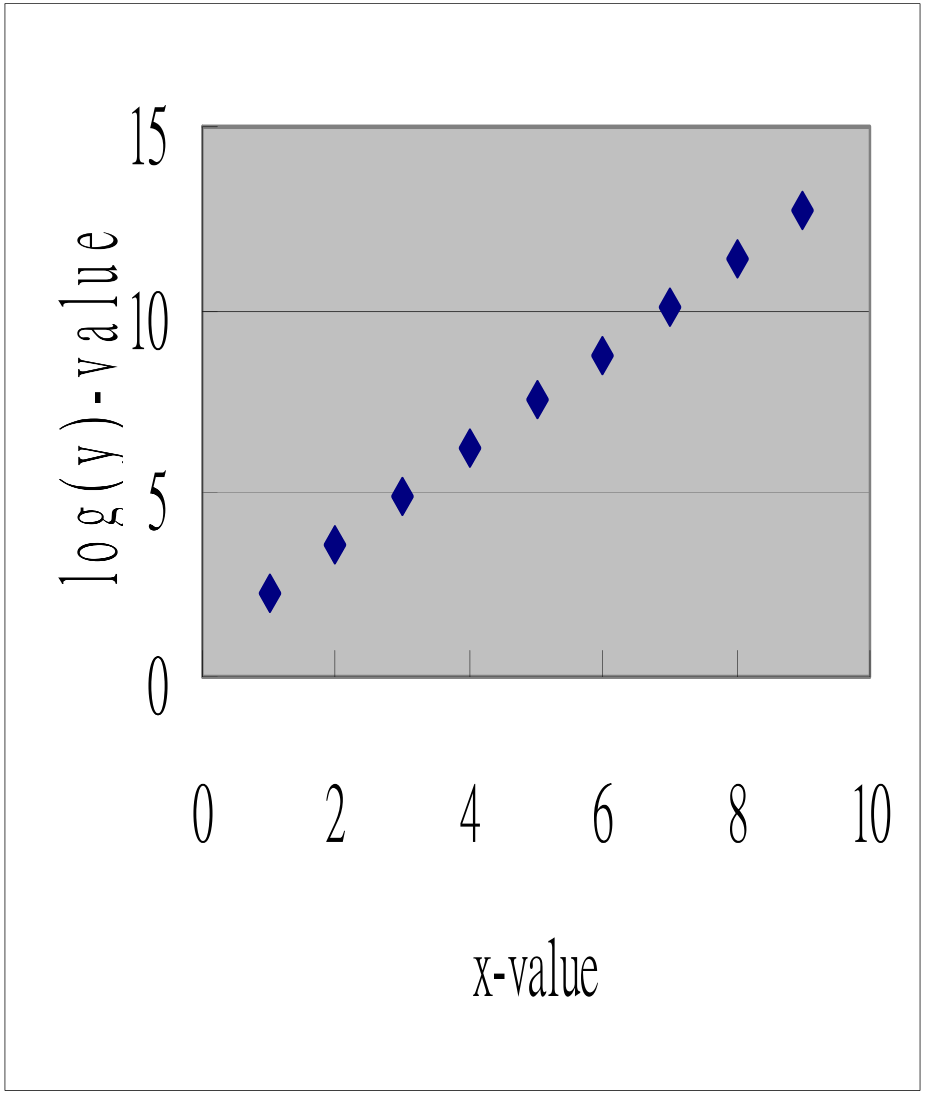
$$a' = \log(a)$$

$$b' = \log(b)$$

Hence we map $\hat{y}'$ vs. x

Corresponding Scatter Plot





Nonlinear Regression Models: Model Transformation

$$\hat{y} = (a + bx)^2$$

$$\Rightarrow \sqrt{\hat{y}} = a + bx$$

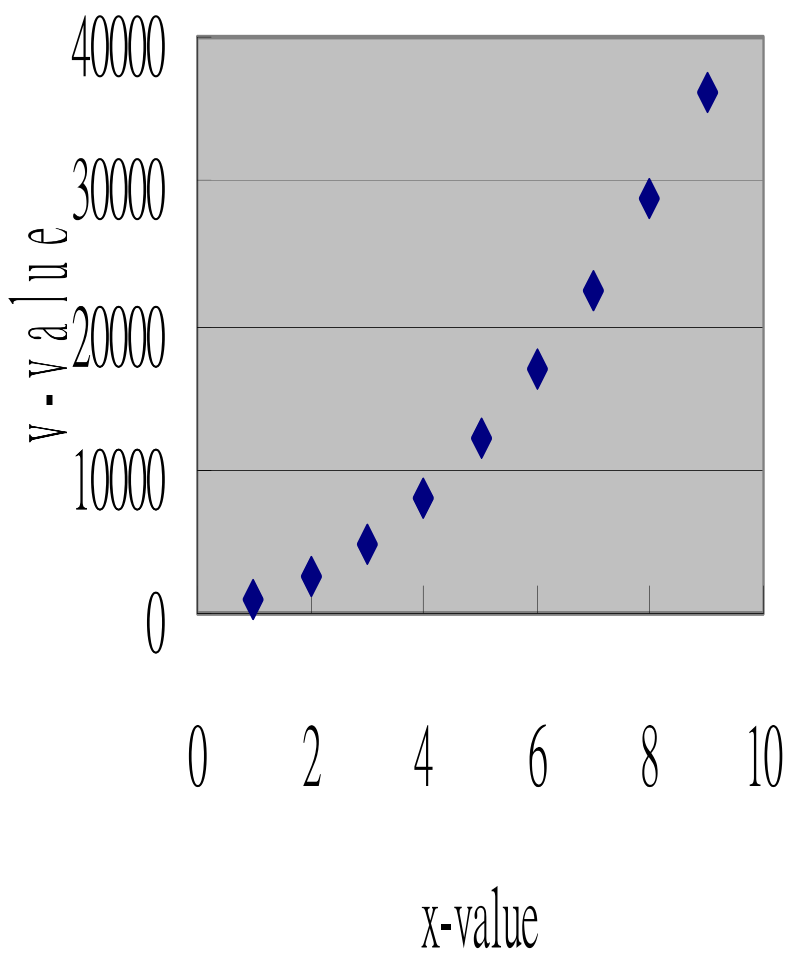
$$\Rightarrow \hat{y}' = a + bx$$

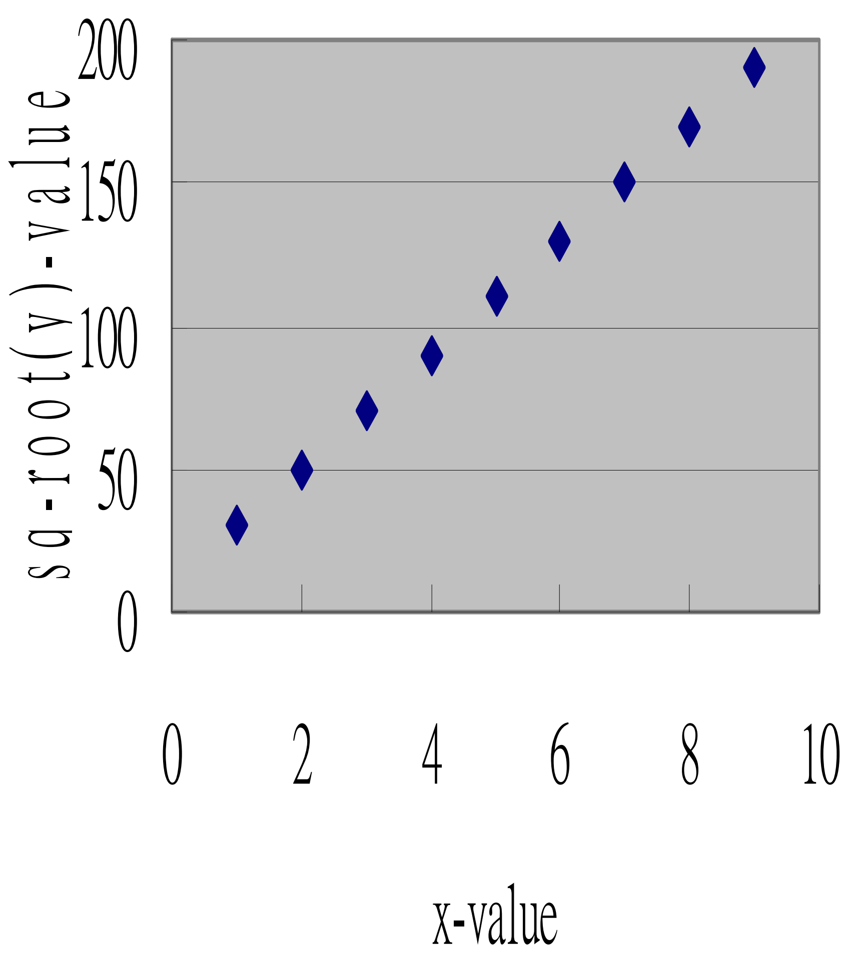
where:

$$\hat{y}' = \sqrt{\hat{y}}$$

Hence we map $\hat{y}'$ vs. x

Corresponding Scatter Plot





Nonlinear Regression Models: Model Transformation

$$\hat{y} = \frac{1}{a + bx}$$

$$\Rightarrow \frac{1}{\hat{y}} = a + bx$$

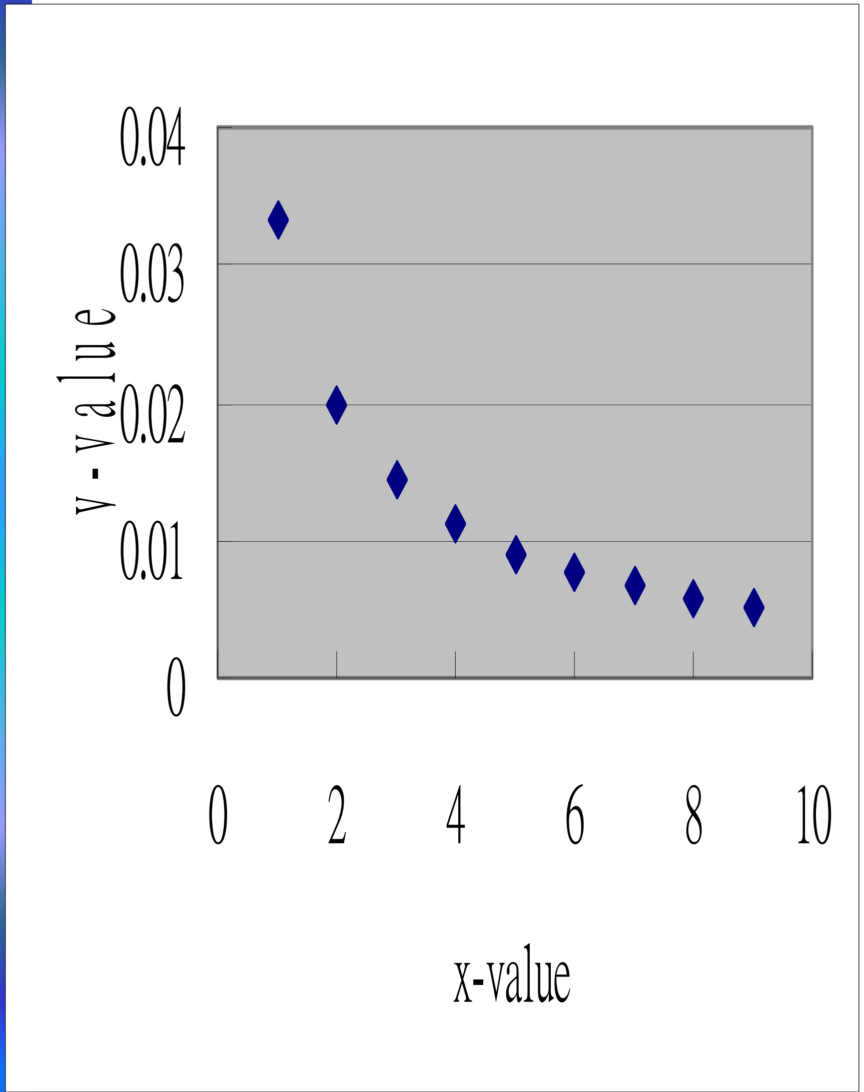
$$\Rightarrow \hat{y}' = a + bx$$

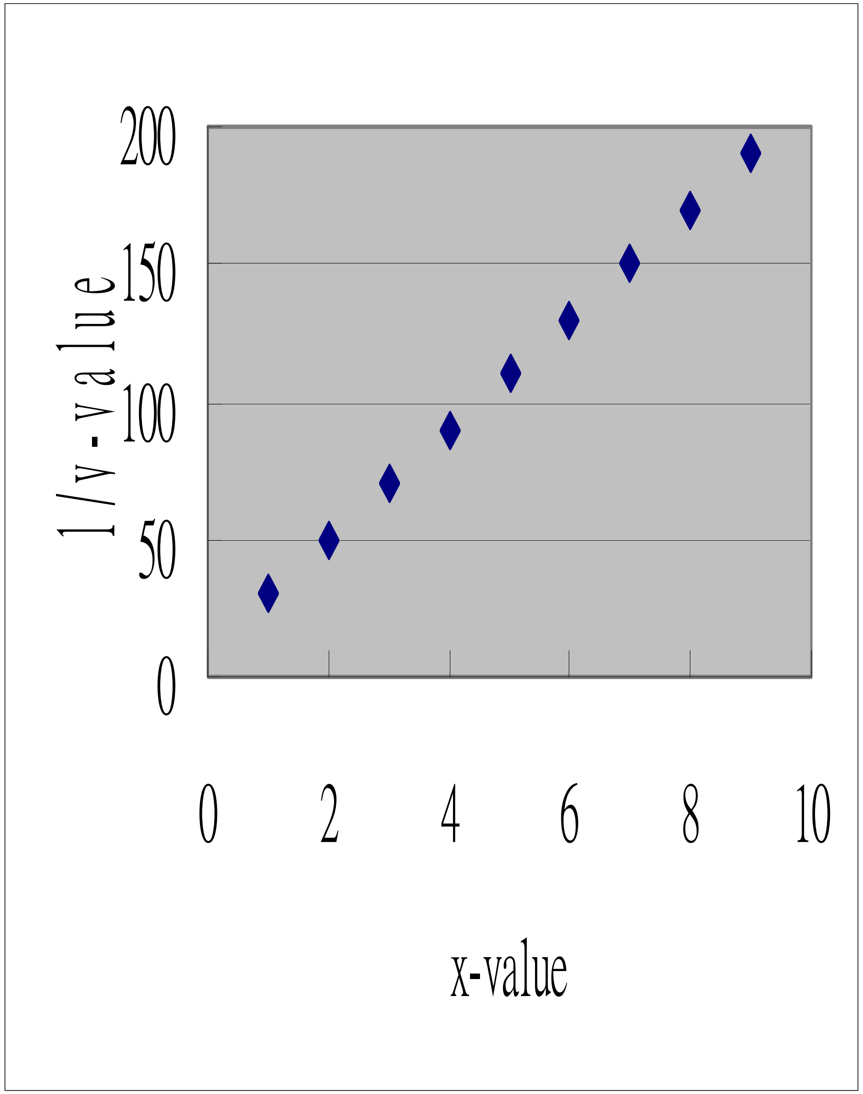
where :

$$\hat{y}' = \frac{1}{\hat{y}}$$

Hence we map $\hat{y}'$ vs. x

Corresponding Scatter Plot





Section 15.2

Inferences for the Slope of the Population Regression Line

Inferences Concerning the Slope of the Regression Line

- **Hypothesis Test for β_1:** Tests the null hypothesis, $\beta_1 = 0$, the slope of the line of best fit is equal to 0, that is, the line is of no use in predicting y for a given value of x

- **Confidence Interval for β_1:** $1-\alpha$ confidence interval estimate for the population slope of the line of best fit

Page 760, Procedure 15.1

PROCEDURE 15.1 **Regression *t*-Test**

Purpose To perform a hypothesis test to decide whether a predictor variable is useful for making predictions

Assumptions

The four assumptions for regression inferences

STEP 1 The null and alternative hypotheses are

H_0: $\beta_1 = 0$ (predictor variable is not useful for making predictions)
H_a: $\beta_1 \neq 0$ (predictor variable is useful for making predictions)

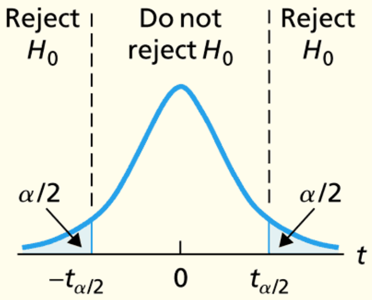
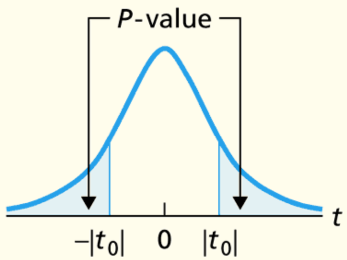
STEP 2 Decide on the significance level, α.

STEP 3 Compute the value of the test statistic

$$t = \frac{b_1}{s_e/\sqrt{S_{xx}}}$$

and denote that value t_0.

Page 760, Procedure 15.1 (cont.)

CRITICAL-VALUE APPROACH	*P*-VALUE APPROACH
STEP 4 The critical value(s) are $\pm t_{\alpha/2}$ with df $= n - 2$. Use Table IV to find the critical values. 	**STEP 4** The t-statistic has df $= n - 2$. Use Table IV to estimate the P-value, or obtain it exactly by using technology. 
STEP 5 If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.	**STEP 5** If $P \leq \alpha$, reject H_0; otherwise, do not reject H_0.

STEP 6 Interpret the results of the hypothesis test.

Minitab Output

Regression Analysis

```
The regression equation is
C2 = 39.8 + 7.19 C1

Predictor        Coef      StDev         T       P
Constant        39.81      69.11      0.58   0.580
C1              7.191      1.484  ?? 4.84   0.001

S = 36.63       R-Sq = 74.6%     R-Sq(adj) = 71.4%

Analysis of Variance

Source           DF          SS          MS         F        P
Regression        1       31491       31491     23.47    0.001
Residual Error    8       10734        1342
Total             9       42225
```

Section 15.3

Estimation and Prediction

Confidence Interval Estimates for Regression

- Use the line of best fit to make predictions

- Predict the population mean *y*-value at a given *x*

- Predict the individual *y*-value selected at random that will occur at a given value of *x*

- The best point estimate, or prediction, for both is $\hat{y}$.

Notation & Background

Notation:

1. Mean of the population y-values at a given value of x: $\mu_{y|x_0}$
2. The individual y-value selected at random for a given value of x: y_{x_0}

Background:

1. Recall: the development of confidence intervals for the population mean μ when the variance was known and when the variance was estimated
2. The confidence interval for $\mu_{y|x_0}$ and the prediction interval for y_{x_0} are constructed in a similar fashion
3. $\hat{y}$ replaces $\bar{x}$ as the point estimate
4. The sampling distribution of $\hat{y}$ is normal

Background Continued

5. The standard deviation in both cases is computed by multiplying the square root of the variance of the error by an appropriate correction factor

6. The line of best fit passes through the centroid: $(\bar{x}, \bar{y})$

Consider a confidence interval for the slope β_1

If we draw lines with slopes equal to the extremes of that confidence interval through the centroid, the value for y fluctuates considerably for different values of x.

It is reasonable to expect a wider confidence interval as we consider values of x further from $\bar{x}$

We need a correction factor to adjust for the distance between x_0 and $\bar{x}$. This factor must also adjust for the variation of the y-values about y

Confidence Interval

Confidence interval for the mean value of y at a given value of x, $\mu_{y|x_0}$

$$\hat{y} \pm t_{(n\text{-}2,\ \alpha/2)} \times \overbrace{s_e \times \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x - \bar{x})^2}}}^{\text{standard error of } \hat{y}}$$

$$= \hat{y} \pm t_{(n\text{-}2,\ \alpha/2)} \times s_e \times \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SS}(x)}}$$

Notes:

1. The numerator of the second term under the radical sign is the square of the distance of x_0 from $\bar{x}$
2. The denominator is closely related to the variance of x and has a standardizing effect on this term

Procedure 15.3

Conditional Mean t-Interval Procedure

Purpose To find a confidence interval for the conditional mean of the response variable corresponding to a particular value of the predictor variable, x_p

Assumptions The four assumptions for regression inferences

Step 1 **For a confidence level of $1-\alpha$, use Table IV to find $t_{\alpha/2}$ with df $= n - 2$.**

Step 2 **Compute the point estimate, $\hat{y}_p = b_0 + b_1 x_p$.**

Step 3 **The endpoints of the confidence interval for the conditional mean of the response variable are**

$$\hat{y}_p \pm t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x_p - \Sigma x_i/n)^2}{S_{xx}}}.$$

Step 4 **Interpret the confidence interval.**

Example

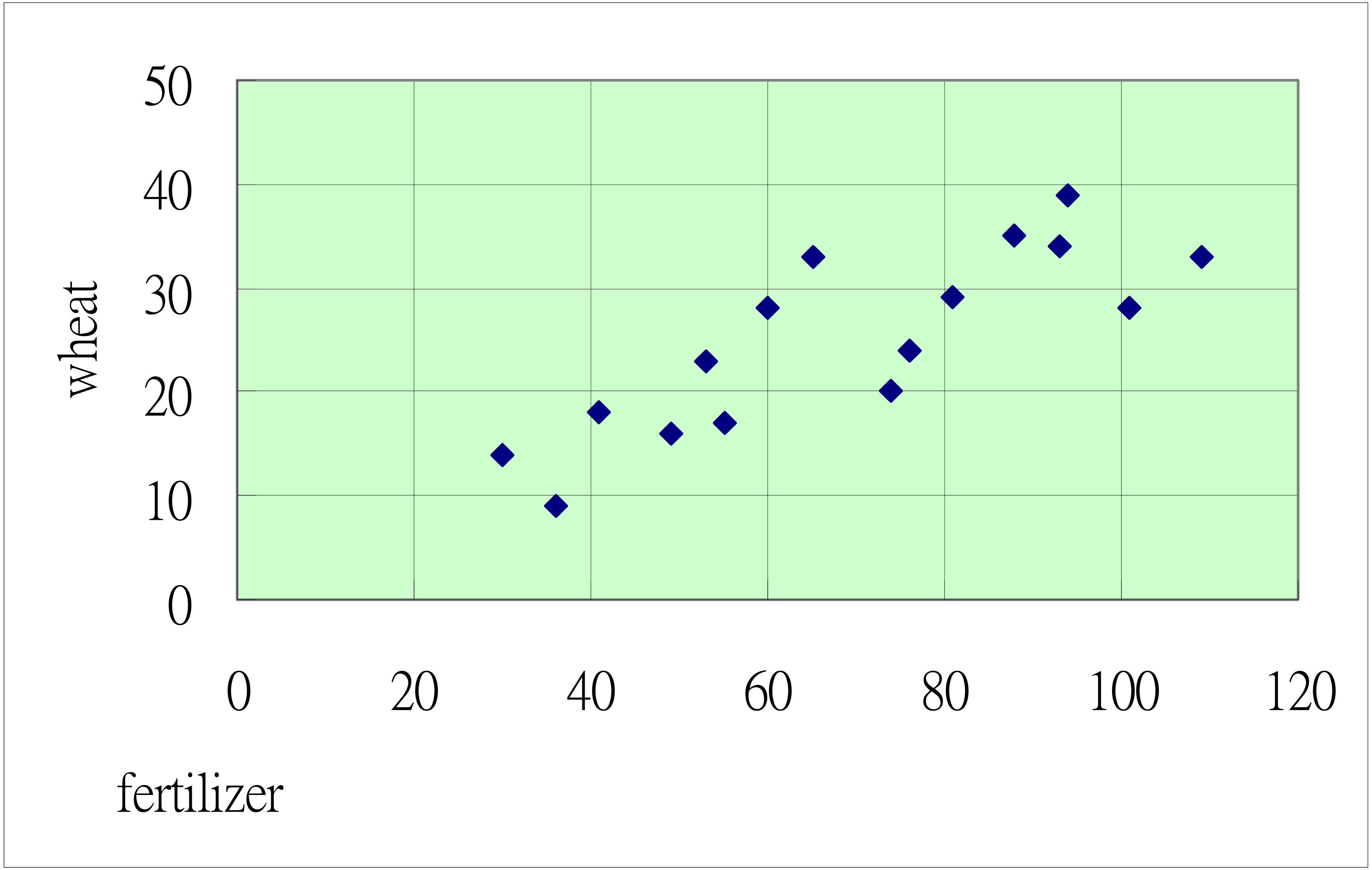
✓ **Example:** It is believed that the amount of nitrogen fertilizer used per acre has a direct effect on the amount of wheat produced. The data below shows the amount of nitrogen fertilizer used per test plot and the amount of wheat harvested per test plot.

a. Find the line of best fit

b. Construct a 95% confidence interval for the mean amount of wheat harvested for 45 pounds of fertilizer

Pounds of Fertilizer (x)	100 Pounds of Wheat (y)	Pounds of Fertilizer (x)	100 Pounds of Wheat (y)
30	14	74	20
36	9	76	24
41	18	81	29
49	16	88	35
53	23	93	34
55	17	94	39
60	28	101	28
65	33	109	33

A scatter plot



Solution:

Pounds of Fertilizer (x)	100 Pounds of Wheat (y)	Pounds of Fertilizer (x)	100 Pounds of Wheat (y)
30	14	74	20
36	9	76	24
41	18	81	29
49	16	88	35
53	23	93	34
55	17	94	39
60	28	101	28
65	33	109	33

$$\sum x = 1105$$
$$\sum y = 400$$
$$\sum x^2 = 85061$$
$$\sum y^2 = 11140$$
$$\sum xy = 30231$$
$$\bar{x} = 1105/16 = 69.0625$$
$$\bar{y} = 400/16 = 25$$

Solution

- the line of best fit: $\hat{y} = 4.42 + 0.298x$

Confidence Interval:

1. *Population Parameter of Interest*
   The mean amount of wheat produced for 45 pounds of fertilizer, $\mu_{y|x=45}$

2. *The Confidence Interval Criteria*
   a. Assumptions: The ordered pairs form a random sample and the *y*-values at each *x* have a mounded distribution
   b. Test statistic: *t* with df = 16 – 2 = 14
   c. Confidence level: 1 – α = 0.95

3. *Sample Information*:

$$s_e^2 = 25.97 \qquad s_e = \sqrt{25.97} = 5.096$$

$$y_{x=45}: \qquad \hat{y} = 4.42 + 0.298(45) = 17.83$$

Solution Continued

4. The Confidence Interval:

$$\hat{y} \pm t_{(n\text{-}2,\,\alpha/2)} \times s_e \times \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SS}(x)}}$$

$$= 17.83 \pm (2.14)(5.096)\sqrt{\frac{1}{16} + \frac{(45 - 69.06)^2}{8746.94}}$$

$$= 17.83 \pm (2.14)(5.096)\sqrt{0.0625 + 0.0662}$$

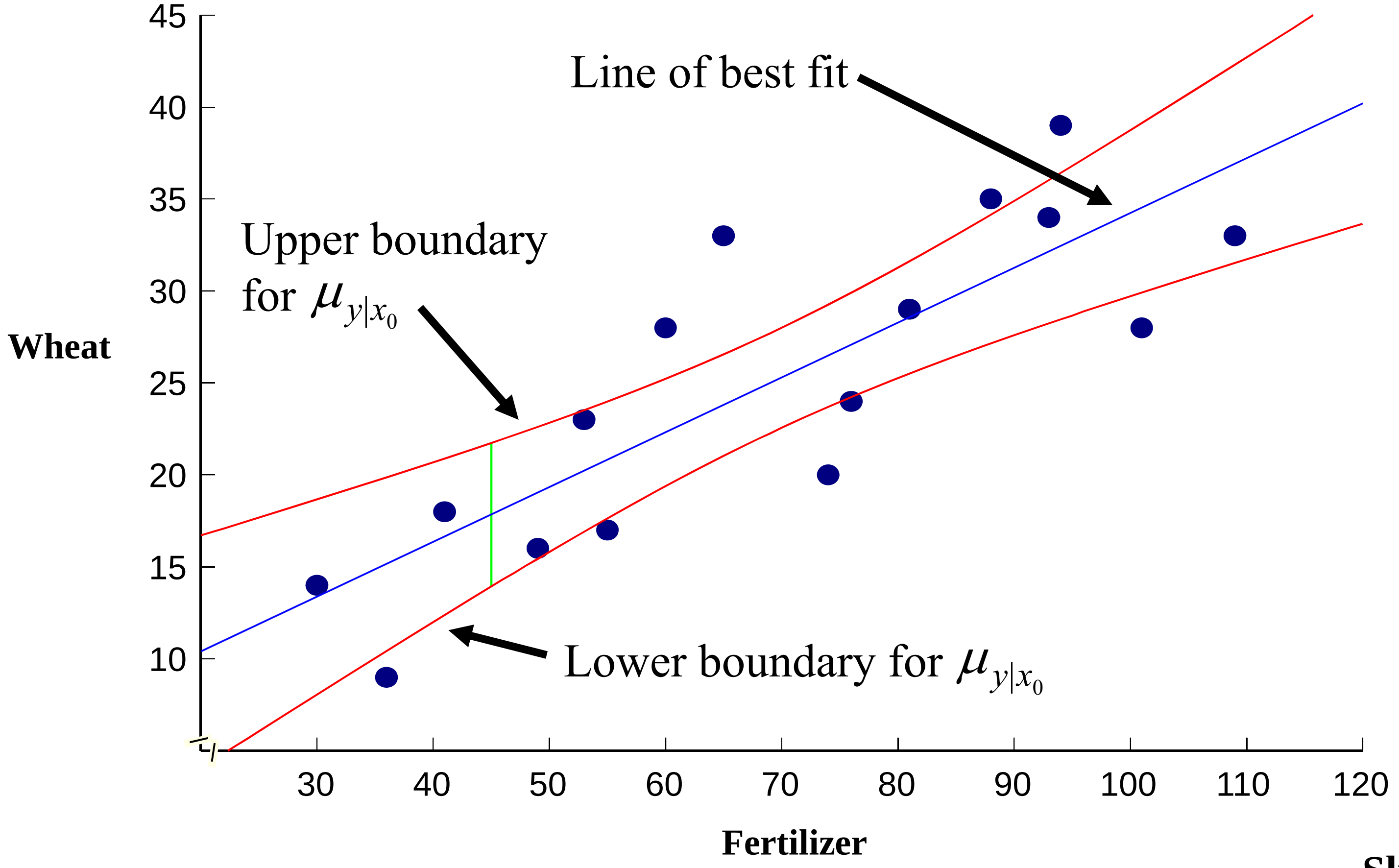
$$= 17.83 \pm (2.14)(5.096)(0.3587)$$

$$= 17.83 \pm 3.91$$

13.92 to 21.74, 95% confidence interval for $\mu_{y|x=45}$

Confidence Belts for $\mu_{y|x_0}$

- *Confidence interval*: green vertical line
- *Confidence interval belt*: upper and lower boundaries of all 95% confidence intervals



Procedure 15.4

Predicted Value *t*-Interval Procedure

Purpose To find a prediction interval for the value of the response variable corresponding to a particular value of the predictor variable, x_p

Assumptions The four assumptions for regression inferences

Step 1 **For a prediction level of $1 - \alpha$, use Table IV to find $t_{\alpha/2}$ with df $= n - 2$.**

Step 2 **Compute the predicted value, $\hat{y}_p = b_0 + b_1 x_p$.**

Step 3 **The endpoints of the prediction interval for the value of the response variable are**

$$\hat{y}_p \pm t_{\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x_p - \Sigma x_i / n)^2}{S_{xx}}}.$$

Step 4 **Interpret the prediction interval.**

Prediction Interval

Prediction interval of the value of a single randomly selected y:

$$\hat{y} \pm t_{(n\text{-}2,\ \alpha/2)} \times s_e \times \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SS}(x)}}$$

✓ **Example:** Find the 95% prediction interval for the amount of wheat harvested for 45 pounds of fertilizer

Solution:

1. *Population Parameter of Interest*

   $y_{x=45}$, the amount of wheat harvested for 45 pounds of fertilizer

Solution Continued

2. *The Confidence Interval Criteria*

   a. Assumptions: The ordered pairs form a random sample and the *y*-values at each *x* have a mounded distribution

   b. Test statistic: *t* with df = 16 – 2 = 14

   c. Confidence level: 1 – α = 0.95

3. *Sample Information*

$$s_e^2 = 25.97 \qquad s_e = \sqrt{25.97} = 5.096$$

$$y_{x=45}: \quad \hat{y} = 4.42 + 0.298(45) = 17.83$$

Solution Continued

4. *The Confidence Interval*

$$\hat{y} \pm t_{(n\text{-}2,\ \alpha/2)} \times s_e \times \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SS}(x)}}$$

$$= 17.83 \pm (2.14)(5.096)\sqrt{1 + \frac{1}{16} + \frac{(45 - 69.06)^2}{8746.94}}$$

$$= 17.83 \pm (2.14)(5.096)\sqrt{1 + 0.0625 + 0.0662}$$

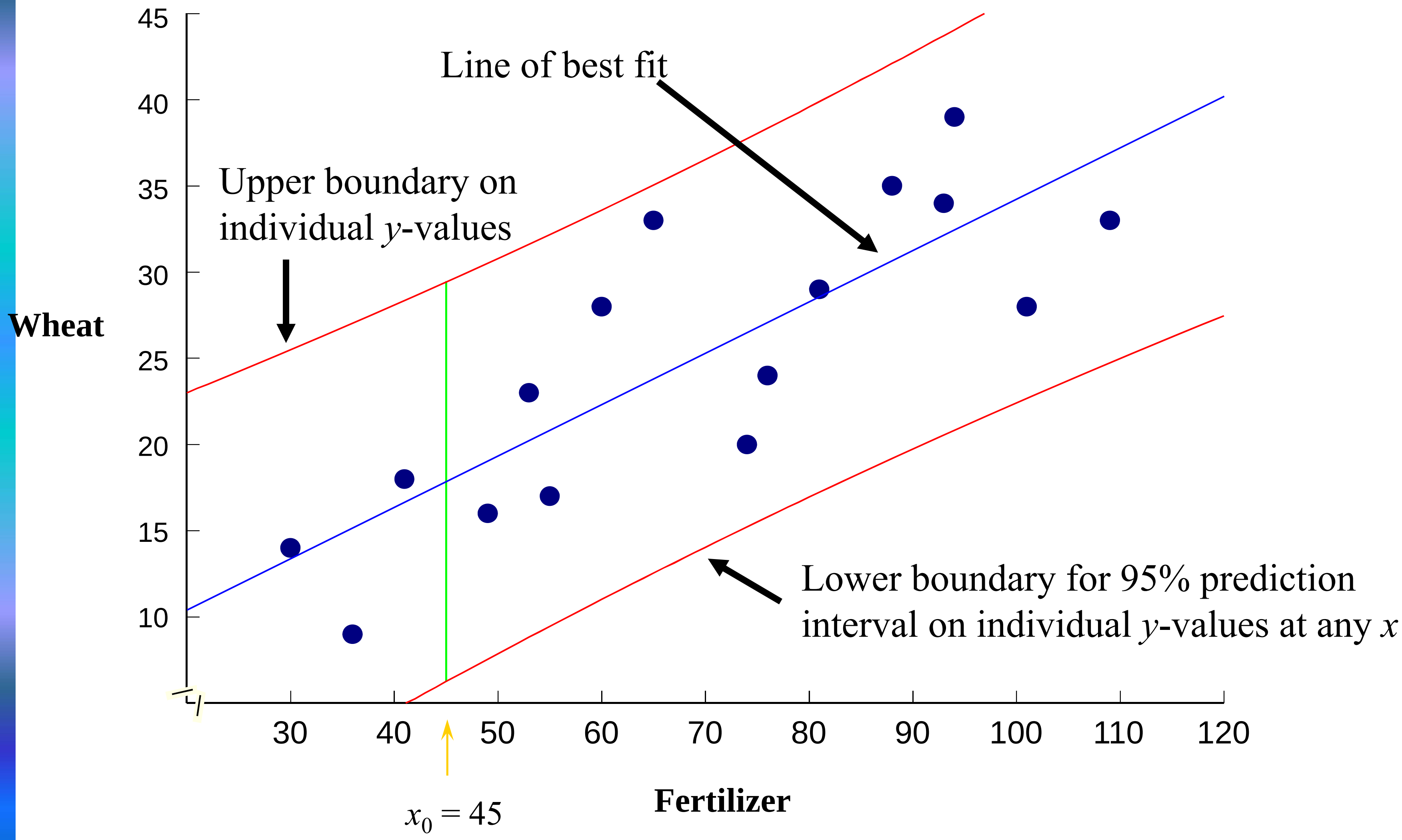
$$= 17.83 \pm (2.14)(5.096)\sqrt{1.1287}$$

$$= 17.83 \pm (2.14)(5.096)(1.0624)$$

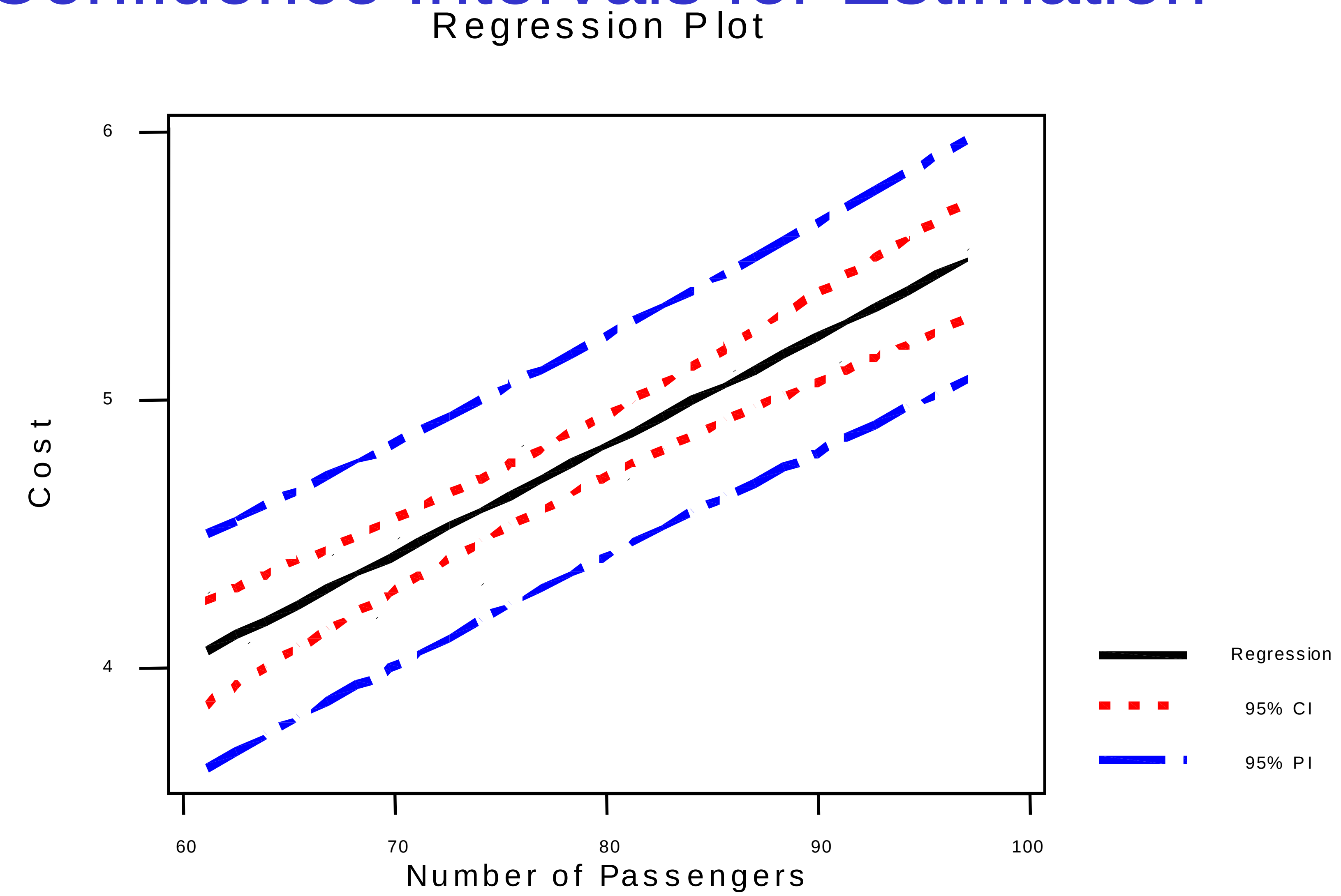
$$= 17.83 \pm 11.5859$$

6.24 to 29.41, 95% prediction interval for $y_{x=45}$

Prediction belts for y_{x_0}



Confidence Intervals for Estimation



Precautions

1. The regression equation is meaningful *only* in the domain of the x variable studied. Estimation outside this domain is risky; it assumes the relationship between x and y is the same outside the domain of the sample data.

2. The results of one sample should not be used to make inferences about a population other than the one from which the sample was drawn

3. Correlation (or association) does *not* imply causation. A significant regression does not imply x causes y to change. Most common problem: missing, or third, variable effect.