EE3211 Modelling Techniques

Lecture 6

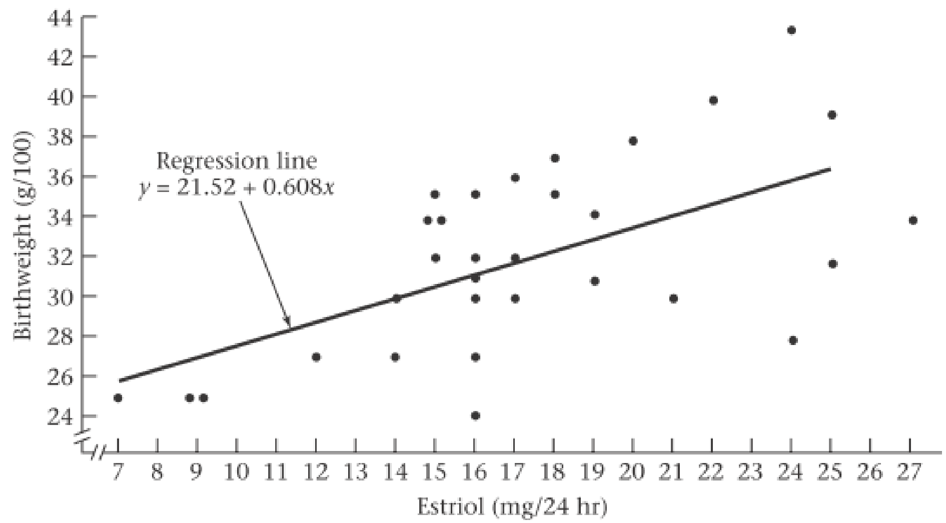
Regression and Correlation Methods

Overview

1. Methods of regression and correlation analysis
2. Multiple-regression analysis
3. Linear regression methods

Example: Some researchers (Greene and Touchstone) conducted a study to relate birthweight and estriol (a female hormone) level in pregnant women

Figure 11.1 Data from the Greene-Touchstone study relating birthweight and estriol level in pregnant women near term



Source: Reprinted with permission of the *American Journal of Obstetrics and Gynecology*, *85*(1), 1–9, 1963.

x = estriol level

y = birthweight

- we can postulate a linear relationship between y and x:

  $E(y|x) = \alpha + \beta x$

- $y = \alpha + \beta x$: regression line; α: intercept; β: slope of the line

Table 11.1 Sample data from the Greene-Touchstone study relating birthweight and estriol level in pregnant women near term

i	Estriol (mg/24 hr) x_i	Birthweight (g/100) y_i	i	Estriol (mg/24 hr) x_i	Birthweight (g/100) y_i
1	7	25	17	17	32
2	9	25	18	25	32
3	9	25	19	27	34
4	12	27	20	15	34
5	14	27	21	15	34
6	16	27	22	15	35
7	16	24	23	16	35
8	14	30	24	19	34
9	16	30	25	18	35
10	16	31	26	17	36
11	17	30	27	18	37
12	19	31	28	20	38
13	21	30	29	22	40
14	24	28	30	25	39
15	15	32	31	24	43
16	16	32			

Source: Reprinted with permission of the *American Journal of Obstetrics and Gynecology, 85*(1), 1–9, 1963.

$y = \alpha + \beta x$: not expected to be true for every woman

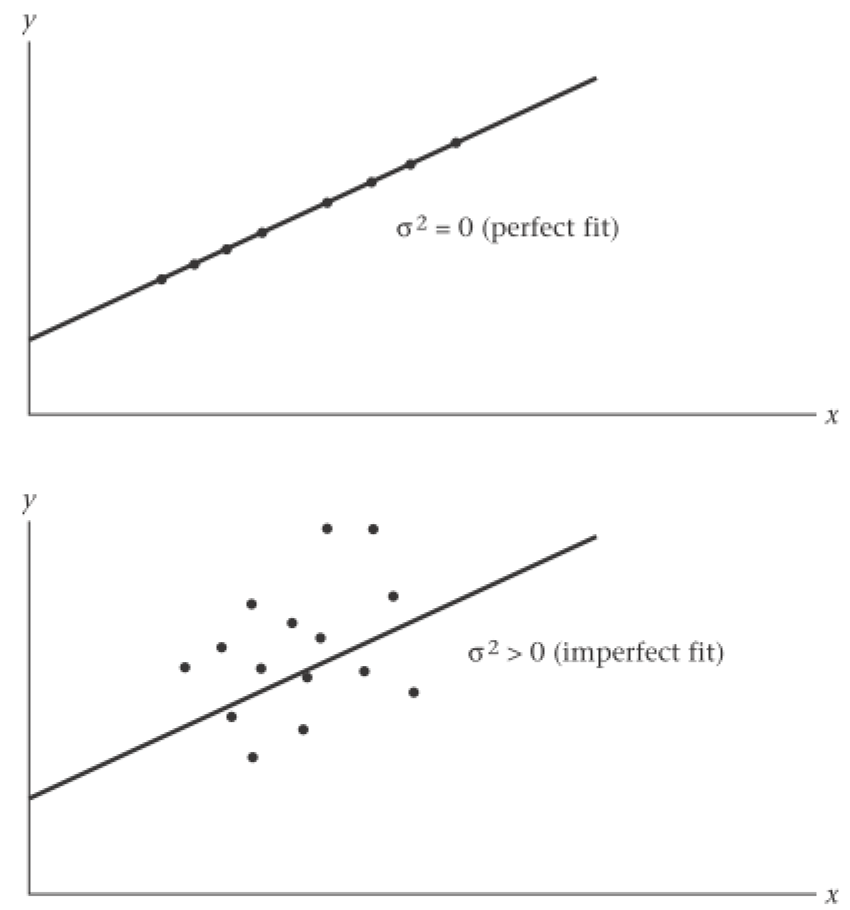
- *e: error term ~N(*$0, \sigma^2$)
- **Full regression line: $y = \alpha + \beta x + e$**

Linear-regression equation: $y = \alpha + \beta x + e$

- y : **dependent variable**
- x : **independent variable** (predict y as a function of x)

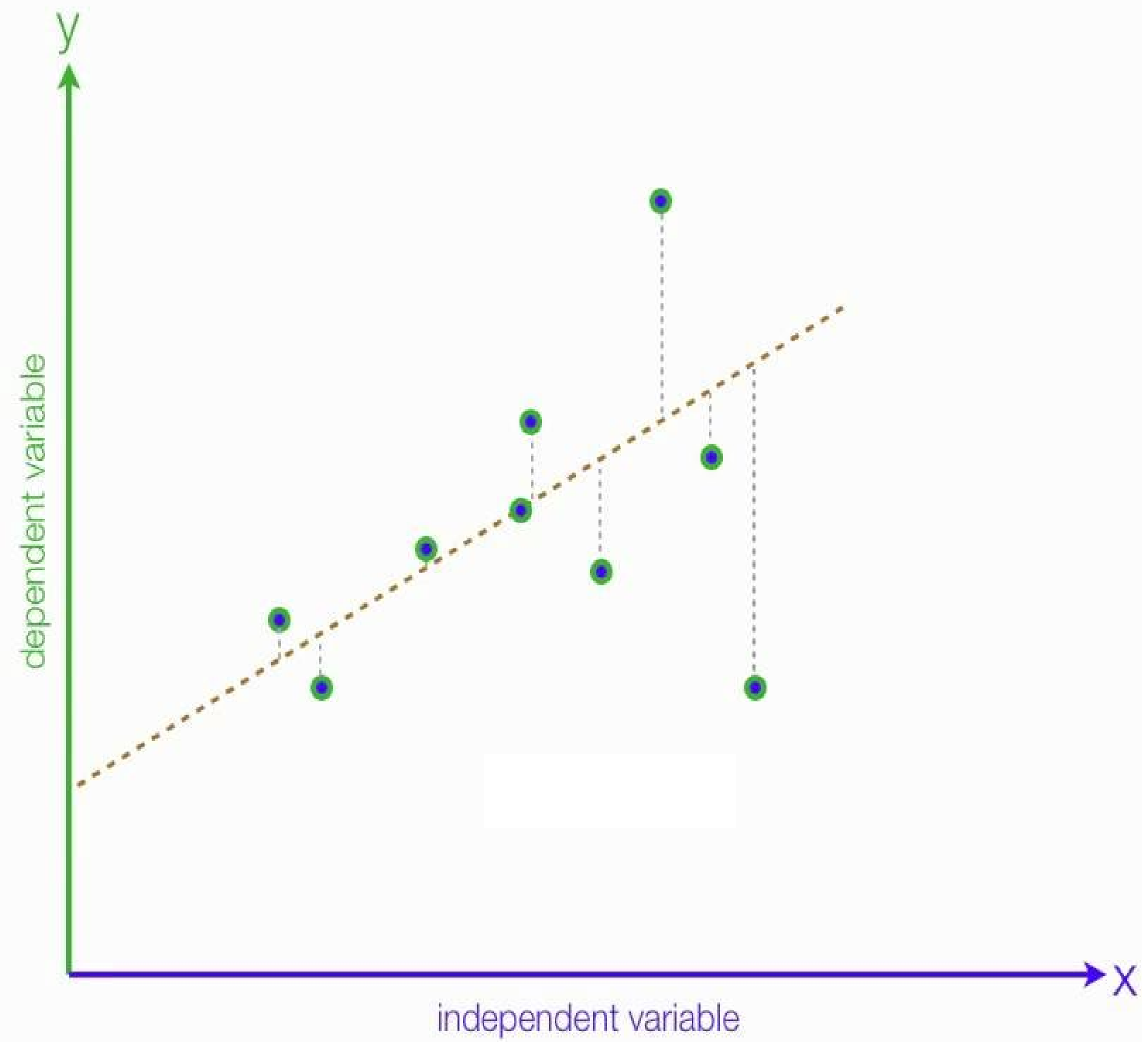
- Birthweight: dependent variable
- Estriol level: independent variable (used to predict birthweight)

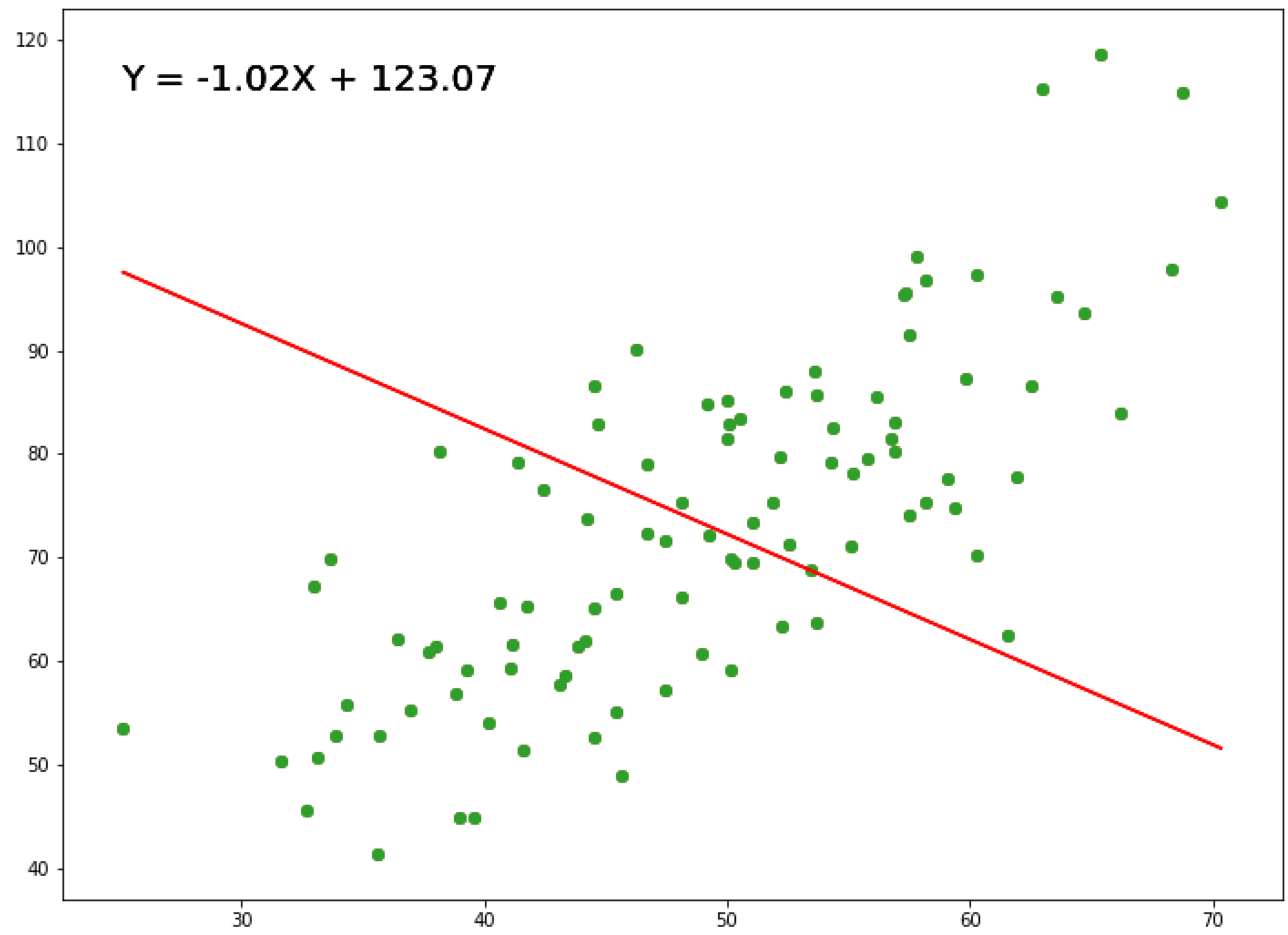
Figure 11.2 The effect of σ^2 on the goodness of fit of a regression line



Interpretation of regression line:

- For a women with estriol level x: her birthweight will be normally distributed with mean α+β*x and variance σ^2
- $\sigma^2 = 0$: every point falls exactly on the regression line
- Larger σ^2 : more scatter occurs about the regression line
- $\beta > 0$: x increases → expected value $y = \alpha + \beta x$ increases


y
dependent variable
x
independent variable


Y = -1.02X + 123.07
120
110
100
90
80
70
60
50
40
30
40
50
60
70

The raw sum of squares for x is defined by

$$\sum_{i=1}^{n} x_i^2$$

The corrected sum of squares for x is denoted by L_{xx} and defined by

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2\Big/n$$

It represents the sum of squares of the deviations of the x_i from the mean. Similarly, the raw sum of squares for y is defined by

$$\sum_{i=1}^{n} y_i^2$$

The corrected sum of squares for y is denoted by L_{yy} and defined by

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2\Big/n$$

Notice that L_{xx} and L_{yy} are simply the numerators of the expressions for the sample variances of x (i.e., s_x^2) and y (i.e., s_y^2), respectively, because

$$s_x^2=\sum_{i=1}^{n}(x_i-\bar{x})^2\Big/(n-1) \text{ and } s_y^2=\sum_{i=1}^{n}(y_i-\bar{y})^2\Big/(\boldsymbol{n}-1)$$

The raw sum of cross products is defined by

$$\sum_{i=1}^{n} x_i y_i$$

The corrected sum of cross products is defined by

$$\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

which is denoted by L_{xy}.

It can be shown that a short form for the corrected sum of cross products is given by

$$\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) \Big/ n$$

Estimation of the Least-Squares Line

The coefficients of the least-squares line $y = a + bx$ are given by

$$b = L_{xy} / L_{xx} \quad \text{and} \quad a = \overline{y} - b\overline{x} = \left(\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i\right) \Big/ n$$

Sometimes, the line $y = a + bx$ is called the *estimated or fitted regression line* or, more briefly, the *regression line*.

Table 11.1 Sample data from the Greene-Touchstone study relating birthweight and estriol level in pregnant women near term

i	Estriol (mg/24 hr) x_i	Birthweight (g/100) y_i	i	Estriol (mg/24 hr) x_i	Birthweight (g/100) y_i
1	7	25	17	17	32
2	9	25	18	25	32
3	9	25	19	27	34
4	12	27	20	15	34
5	14	27	21	15	34
6	16	27	22	15	35
7	16	24	23	16	35
8	14	30	24	19	34
9	16	30	25	18	35
10	16	31	26	17	36
11	17	30	27	18	37
12	19	31	28	20	38
13	21	30	29	22	40
14	24	28	30	25	39
15	15	32	31	24	43
16	16	32			

Source: Based on the *American Journal of Obstetrics and Gynecology, 85*(1), 1–9, 1963.

Example:

Q: Derive the estimated regression line for the data in Table 11.1.

• **Equation 11.3** Estimation of the Least-Squares Line

The coefficients of the least-squares line $y = a + bx$ are given by

$b = L_{xy} / L_{xx}$ and $a = \bar{y} - \bar{x} = \left(\sum_{i=1}^{31} y_i - b\sum_{i=1}^{31} x_i\right)/n$

Sometimes, the line $y = a + bx$ is called the *estimated or fitted regression line* or, more briefly, the *regression line.*

Solution: First,

$\sum_{i=1}^{31} x_i = 534 \quad \sum_{i=1}^{31} x_i^2 = 9876 \quad \sum_{i=1}^{31} y_i = 992 \quad \sum_{i=1}^{31} x_i y_i = 17{,}500$

Then, compute L_{xy} and L_{xx}:

$$L_{xy} = \sum_{i=1}^{31} x_i y_i - \left(\sum_{i=1}^{31} x_i\right)\left(\sum_{i=1}^{31} y_i\right)/31 = 17{,}500 - \frac{(534)(992)}{31} = 412$$

$$L_{xx} = \sum_{i=1}^{31} x_i^2 - \frac{\left(\sum_{i=1}^{31} x_i\right)^2}{31} = 9876 - \frac{(534)^2}{31} = 677.42$$

$\sum_{i=1}^{n}(x_i - \bar{x})^2$

$$L_{xx} = \sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2 \Big/ n$$

$$L_{xy} = \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

Finally, compute the slope of the regression line:

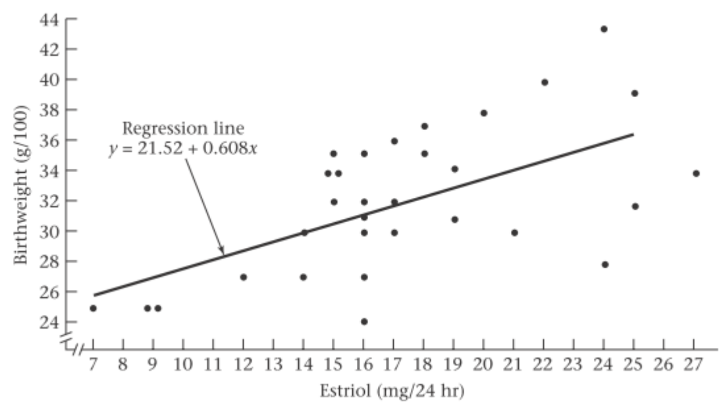
$$b = L_{xy} / L_{xx} = \frac{412}{677.42} = 0.608$$

The intercept of the regression line can also be computed. Note from Equation 11.3 that

$$a = \frac{\left(\sum_{i=1}^{31} y_i - 0.608\sum_{i=1}^{31} x_i\right)}{31} = \frac{[992 - 0.608(534)]}{31} = 21.52$$

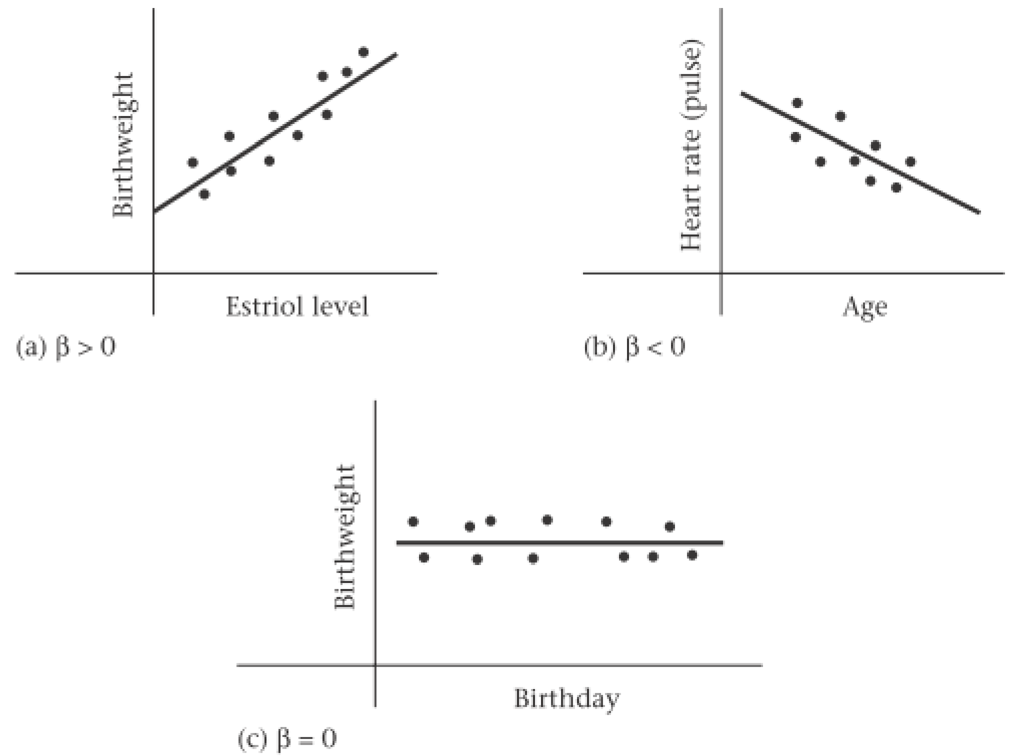
Thus, the regression line is given by $y = 21.52 + 0.608x$. This regression line is shown in Figure 11.1.

Figure 11.1 Data from the Greene-Touchstone study relating birthweight and estriol level in pregnant women near term



Source: Reprinted with permission of the *American Journal of Obstetrics and Gynecology*, *85*(1), 1–9, 1963.

Figure 11.3 Interpretation of the regression line for different values of β



- β = 0: no linear relationship between *x* and *y* (figure c)

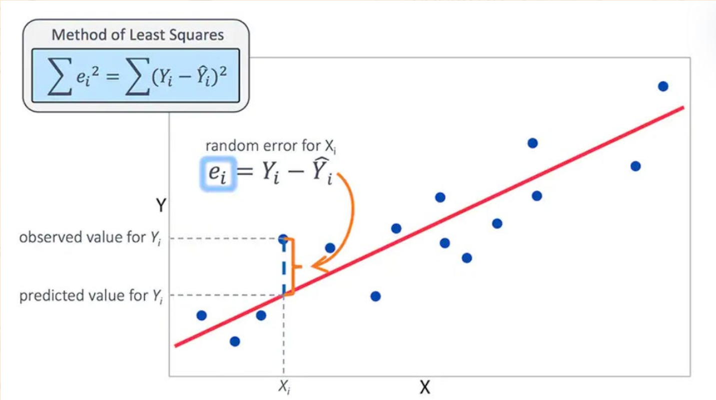
Residuals and fitted regression line

(x_i, y_i): regression line, $y = \alpha + \beta x$.

If $y = a + bx$ is the estimated regression line and

$\hat{e}_i$ = residual for the point (x_i, y_i) about the estimated regression line, then

$\hat{e}_i = y_i - (a + bx_i)$ and $sd(\hat{e}_i) = \sqrt{\hat{\sigma}^2\left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{L_{xx}}\right]}$



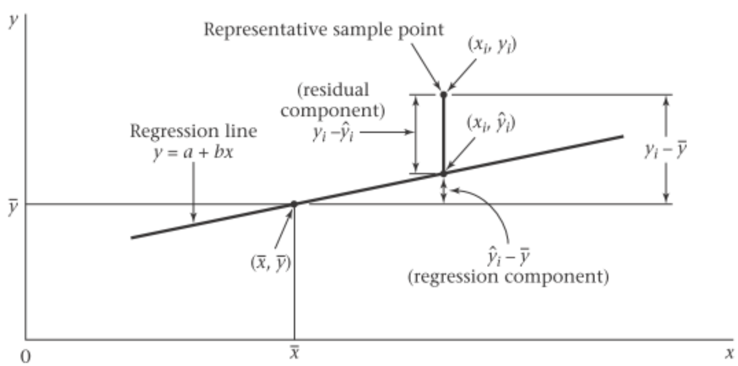
The **Studentized residual** corresponding to the point (x_i, y_i): $\hat{e}_i / sd(\hat{e}_i)$

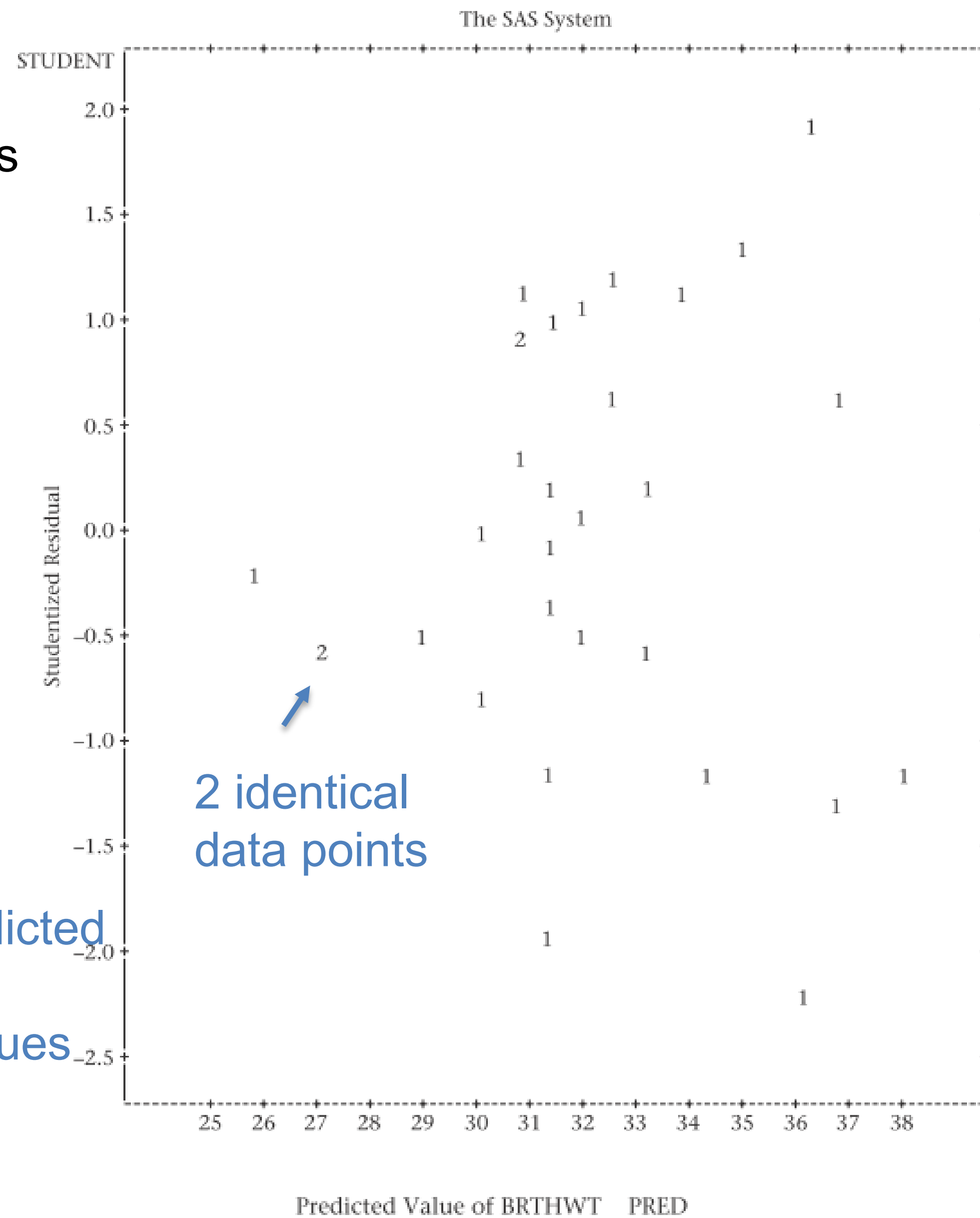
Unequal residual variances:

- **Variance-stabilizing transformation**: transform y to a different scale e.g. ln and square-root

Figure 11.11 Plot of Studentized residuals vs. the predicted value of birthweight for the birthweight–estriol data in Table 11.1

- Compute residuals about fitted regression line
- Construct a scatter plot of residuals vs. estriol values (x) or predicted birthweights ($\hat{y}$)





- Spread increases slightly as the predicted birthweight increases
- Data points with lowest predicted values have residuals close to 0

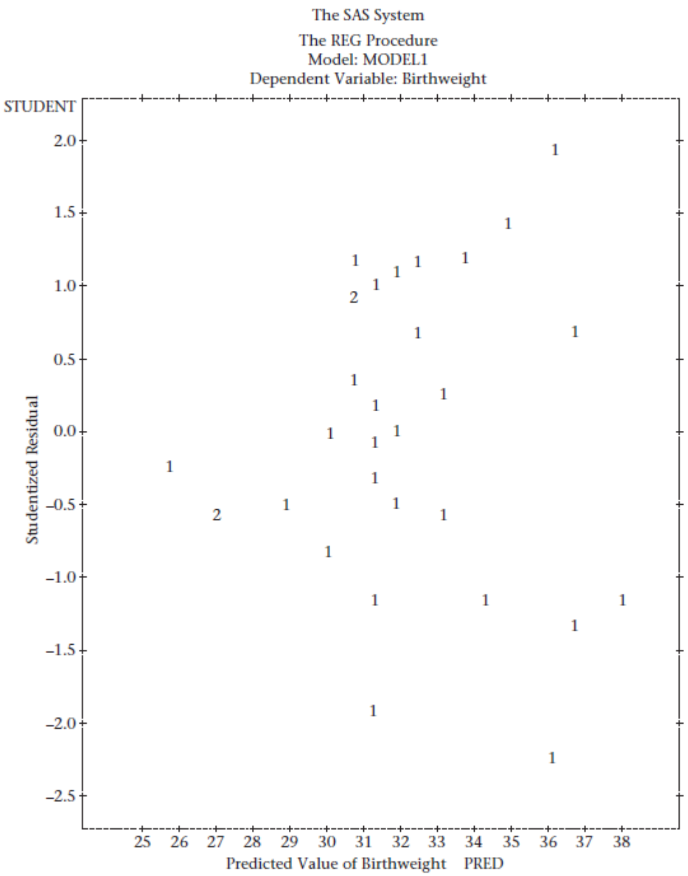


FIGURE 11.11 Plot of Studentized residuals vs. the predicted value of birthweight for the birthweight–estriol data in Table 11.1

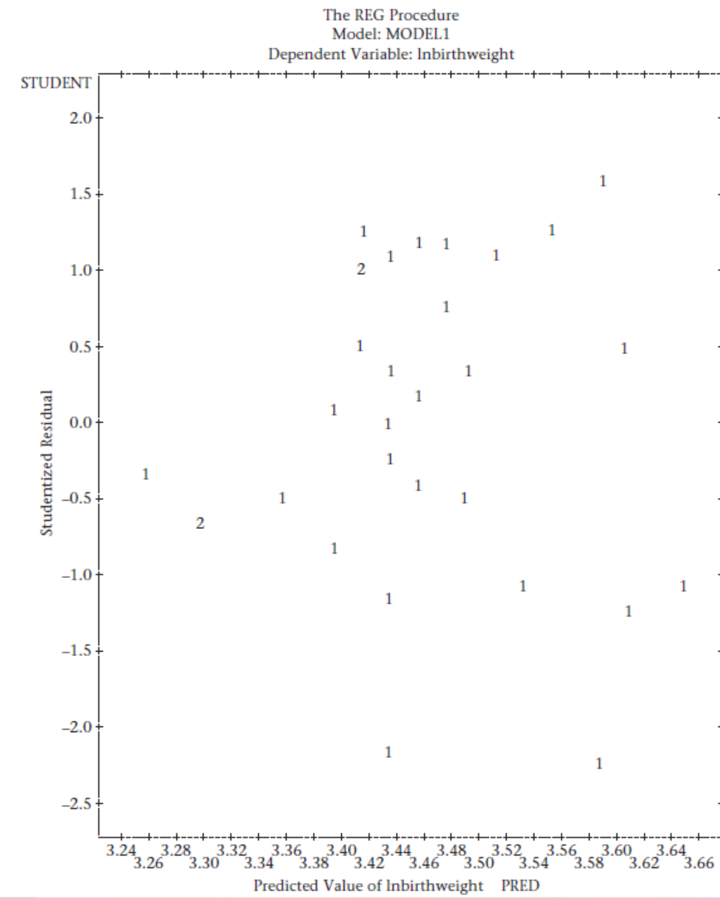


FIGURE 11.12 Plot of Studentized residuals vs. the predicted value of ln(birthweight) for the birthweight–estriol data in Table 11.1

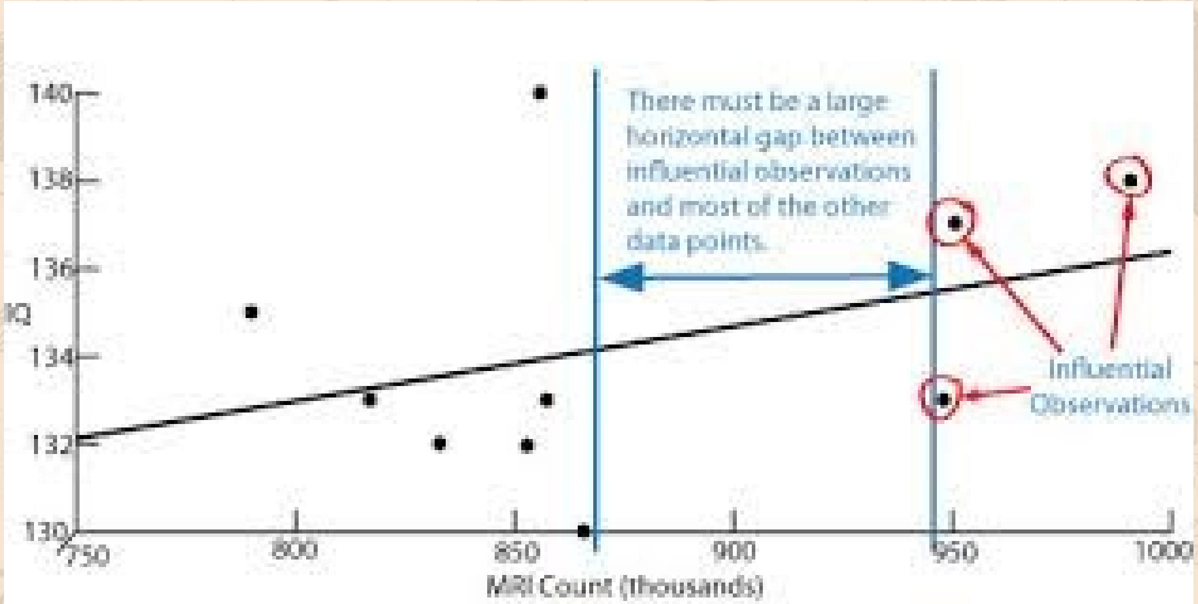
- The two plots are similar
- Simplicity: keep the data in the original scale

*Appropriate transformation is critical→ meeting linearity, equal-variance and normality assumptions (conflict between different assumptions e.g. equal-variance and linearity)

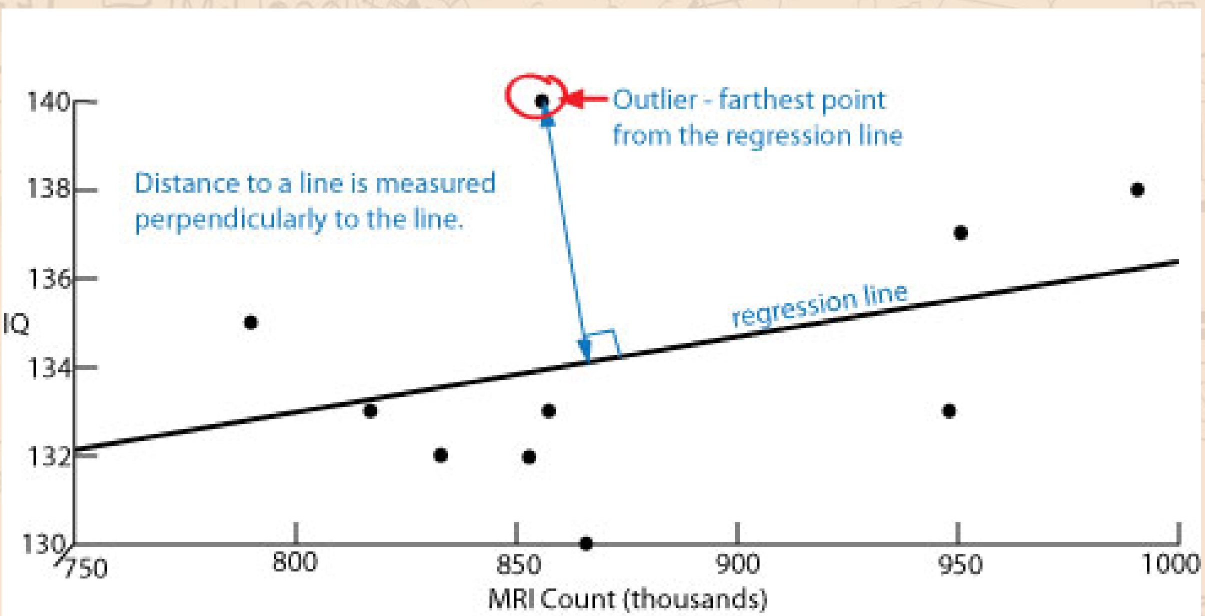
- Residual variance is proportional to average value of y: square-root transformation is useful
- Residual variance is proportional to the square of the average values : log transformation is useful

Outliers and ***influential points*****:** goodness-of-fit of a regression line

- **Influential points**: have an important influence on the coefficients of the fitted regression lines
- **Outlier** (x_i, y_i): may or may not be influential, depends on its location relative to the remaining sample points
- If $|x_i - \bar{x}|$ is small: even a gross outlier will have a relatively small influence on the slope estimate, but important influence on the intercept estimate



Influential points



Outlier

Correlation Coefficient

$$L_{xy} = \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) \Big/ n$$

$$L_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n$$

Sample (Pearson) correlation coefficient (r) : $L_{xy}/\sqrt{L_{xx}L_{yy}}$.

- not affected by changes in location or scale in either variable
- lie between -1 and +1
- useful for quantifying relationship between variables

<u>Interpretation of the sample correlation coefficient</u>

- **Correlation > 0**: variables are positively correlated
- **Correlation < 0**: variables are negatively correlated
- **Correlation = 0**: variables are uncorrelated

R command to obtain a correlation coefficient

```
#use cor.test command
>a=cor.test(Birthweight, Estriol)
#obtain the estimate
>a$estimate
```

Correlation coefficient: a *quantitative* measure of the dependence between two variables

- closer |r| is to 1: more closely related the variables are
- |r| = 1: one variable can be predicted exactly from the other

- Interpretation of the sample correlation coefficient (r): degree of dependence is only correct if the variables *x* and *y* are **normally distributed**

Example on Correlation Coefficient

- FEV is related to both age and height.
- Focus: boys who are ages 10−15 and postulate a regression model of the form FEV = $\alpha + \beta$(height) + e.
- Data were collected on FEV and height for 655 boys in this age group residing in Tecumseh, Michigan.
- Table 11.4 presents the mean FEV in liters for each of twelve 4-cm height groups.

Table 11.4. Mean FEV by height group for boys ages 10−15 in Tecumseh, Michigan

Height (cm)	Mean FEV (L)	Height (cm)	Mean FEV (L)
134[a]	1.7	158	2.7
138	1.9	162	3.0
142	2.0	166	3.1
146	2.1	170	3.4
150	2.2	174	3.8
154	2.5	178	3.9

[a]The middle value of each 4-cm height group is given here.
Source: Based on the *American Review of Respiratory Disease*, *108*, 258−272, 1973.

Question: Compute the correlation coefficient between FEV and height for the pulmonary-function data in Example 11.15 (on p. 471).

Solution: From table 11.4 (previous slide)

$$L_{xy} = 5156.20 - \frac{(1872)(32.3)}{12} = 117.4$$

$$L_{xy} = \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) \Big/ n$$

$$L_{xx} = 294{,}320 - \frac{1872^2}{12} = 2288$$

$$L_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n$$

$$L_{yy} = 93.11 - \frac{32.3^2}{12} = 6.169$$

So, $L_{xy} = 117.4 \quad L_{xx} = 2288 \quad L_{yy} = 6.269$

Therefore, $\mathrm{r} = \frac{117.4}{\sqrt{2288(6.169)}} = \frac{117.4}{118.81} = 0.988$

$\boxed{r = L_{xy}/\sqrt{L_{xx} L_{yy}}}$

- Conclusion: a very strong positive correlation exists between FEV and height

Multiple Regression

- Multiple regression analysis : examine relationship between each of the more than one independent variables ($x_1,\ldots,x_k$) and the dependent variable (y) after taking into account the remaining independent variables

- Estimation of the regression equation: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e$
  e is an error: normally distributed with mean 0 and variance σ^2
  (y=SBP, x_1=birthweight, x_2= age in days)

- If we have k independent variables $x_1,\ldots, x_k$ then a linear-regression model relating y to $x_1,\ldots, x_k$: $y = \alpha + \sum_{j=1}^{k} \beta_j x_j + e$

Example: Pediatrics Hypertension

- Investigate how the relationship between the blood-pressure levels of newborns and infants relate to subsequent adult blood pressure.

- Problem: blood pressure of a newborn is affected by several extraneous factors that make this relationship difficult to study.

- Newborn blood pressures are influenced by:

(1) Birthweight

(2) the day of life on which blood pressure is measured

- Infants: weighed at the time of the blood-pressure measurements
- Expect: infants seen at 5 days of life would on average have a greater weight than those seen at 2 days of life
- Dependent variable: SBP
- Independent variables: age and birthweight

TABLE 11.8 Sample data for infant blood pressure, age, and birthweight for 16 infants

i	Age (days) (x_1)	Birthweight (oz) (x_2)	SBP (mm Hg) (y)
1	3	135	89
2	4	120	90
3	3	100	83
4	2	105	77
5	4	130	92
6	5	125	98
7	2	125	82
8	3	105	85
9	5	120	96
10	4	90	95
11	2	120	80
12	3	95	79
13	3	120	86
14	4	150	97
15	3	160	92
16	3	125	88

Estimates were obtained with SAS PROC REG program

TABLE 11.9 Least-squares estimates of the regression parameters for the newborn blood-pressure data in Table 11.8 using the SAS PROC REG program

```
                              The REG Procedure
Model:  MODEL1
Dependent Variable: sysbp
                     Number of Observations Read     16
                     Number of Observations Used     16
                           Analysis of Variance

                                  Sum of          Mean
      Source              DF     Squares        Square     F Value     Pr > F

      Model                2   591.03564     295.51782       48.08     <.0001
      Error               13    79.90186       6.14630
      Corrected Total     15   670.93750

             Root MSE            2.47917     R-Square     0.8809
             Dependent Mean     88.06250     Adj R-Sq     0.8626
             Coeff Var           2.81524

                           Parameter Estimates
                                                                        Squared
                         Parameter      Standard                        Partial
Variable         DF       Estimate         Error    t Value   Pr > |t|  Corr Type II

Intercept         1       53.45019       4.53189      11.79     <.0001          .
agedys            1        5.88772       0.68021       8.66     <.0001    0.85214
brthwgt           1        0.12558       0.03434       3.66     0.0029    0.50715
```

$y=53.45+5.89x_1+0.126x_2$

For a newborn, the average blood pressure increases by an estimated 5.89 mm Hg per day of age, and 0.126 mm Hg per ounce of birthweight.

Multiple-regression model:

$$y = \alpha + \sum_{j=1}^{k} \beta_j x_j + e \quad \text{where } e \sim N(0, \sigma^2)$$

β_j, j = 1, 2…,k

- partial-regression coefficients:
  - represents the average increase in y per unit increase in x_j, with all other variables held constant (adjusting all other variables)
  - estimated by the parameter b_j

- Partial regression coefficients vs. simple linear-regression coefficients

- Simple linear-regression coefficients: average increase in y per unit increase in x (do not consider any other independent variables)

- Strong relationships among the independent variables in a multiple-regression model: partial-regression coefficients and simple linear-regression coefficients are different considerably

- ranking the independent variables according to their predictive relationship with the dependent variable y
  - hard to rank based on magnitude of partial-regression coefficients (different units)

Standardized regression coefficient (b_s) : $b \times (s_x/s_y)$

- represents the estimated average increase in *y* per standard deviation increase in *x*, after adjusting for all other variables in the model

- useful measure for comparing the predictive value of several independent variables

- can control for differences in the units of measurement for different independent variables by expressing change in standard-deviation units of *x*

Example: Pediatrics Hypertension

Question: Calculate the predicted average SBP of a three-day-old baby with birthweight 8 lb (128 oz).

Solution: The average SBP is estimated by 53.45 + 5.89(3) + 0.126(128) = 87.2 mm Hg
The regression coefficients in Table 11.9 are called *partial-regression coefficients*.

Example: Pediatrics Hypertension

Question: Compute the standardized regression coefficients for age in days and birthweight using the data in Tables 11.8 and 11.9.

Solution: From Table 11.8, s_y = 6.69, s_{x_1} = 0.946, s_{x_2} = 18.75. Therefore,

$$b_s \text{ (age in days)} = 5.888 \times 0.946/6.69 = 0.833$$

$$b_s \text{(birthweight)} = 0.1256 \times 18.75/6.69 = 0.352$$

$$s_x^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 \Big/ (n-1)$$

$$s_y^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 \Big/ (n-1)$$

- The average increase in SBP is 0.833 standard-deviation units of blood pressure per standard-deviation increase in age, holding birthweight constant, and 0.352 standard-deviation units of blood pressure per standard-deviation increase in birthweight, holding age constant

- Age: more important variable after controlling for both variables simultaneously in the multiple- regression model

Hypothesis Testing

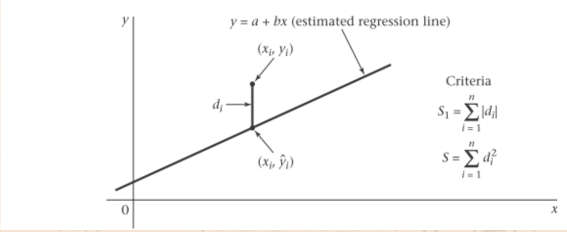
H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$
vs. H_1: at least one of the $\beta_j \neq 0$ in multiple linear regression

1.Estimate the regression parameters and compute Reg SS and Res SS using method of least squares

$$\text{Res SS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$\text{Reg SS} = \text{Total SS} - \text{Res SS}$$

$$\text{Total SS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$
$$\hat{y}_i = a + \sum_{j=1}^{k} b_j x_{ij}$$

Figure 11.4 Possible criteria for judging the fit of a regression line



xij = *j*th independent variable for the *i*th subject, $j = 1,\ldots,k$; $i = 1,\ldots,n$

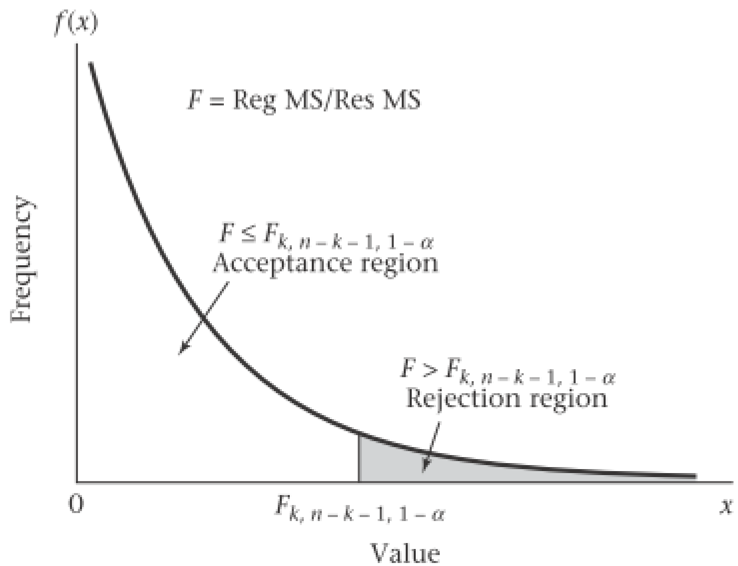
2.Compute Reg MS = Reg SS/k, Res MS = Res SS/(n-k-1)

3.Compute the test statistic F = Reg MS/Res MS following an $F_{k,n-k,k-1}$ distribution under H_0.

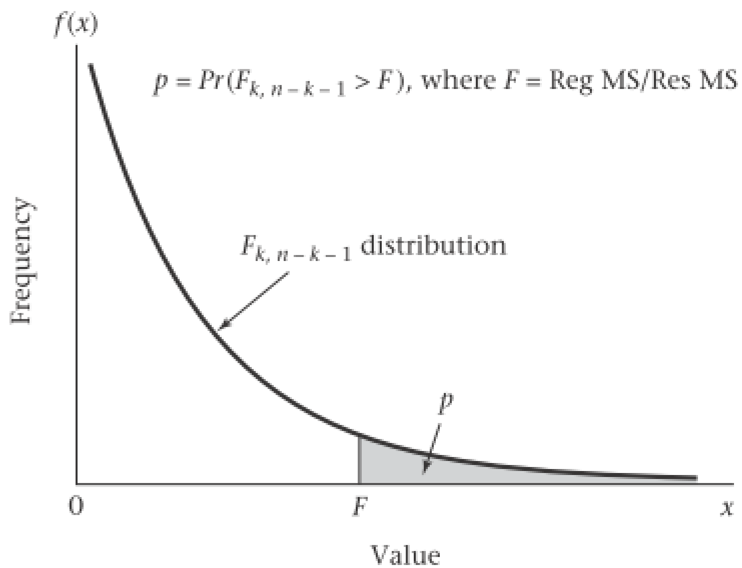
4.For a level α test:
- $F > F_{k,n-k-1,1-\alpha}$ → reject H_0.
- $F \leq F_{k,n-k-1,1-\alpha}$ → accept H_0.

5. Exact p-value : area to the right of F under an F > $F_{k,n-k-1}$ *distribution*
   $= Pr(F_{k,n-k-1} > F)$

Acceptance and rejection regions for testing the hypothesis H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ vs. H_1: at least one of the $\beta_j \neq 0$ in multiple linear regression



Computation of the p-value for testing the hypothesis H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ vs. H_1: at least one of the $\beta_j \neq 0$ in multiple linear regression



t test for testing the hypothesis H_0: $\beta_l = 0$, All other $\beta_j \neq 0$ vs. H_1: $\beta_l \neq 0$, all other $\beta_j \neq 0$ in multiple linear regression

Compute $t = b_l/se(b_l)$ which should follow a t distribution with n – k -1 *df* under H_0.

- If $t < t_{n-k-1,\alpha/2}$ or $t > t_{n-k-1,\alpha/2}$ → reject H_0
- If $t_{n-k-1,\alpha/2} \leq t \leq t_{n-k-1,1-\alpha/2}$ → accept H_0

The exact *p*-value :

- $t \geq 0$: $2 \times Pr(t_{n-k-1} > t)$
- $t < 0$: $2 \times Pr(t_{n-k-1} \leq t)$

FIGURE 11.22 Acceptance and rejection regions for the *t* test for multiple linear regression

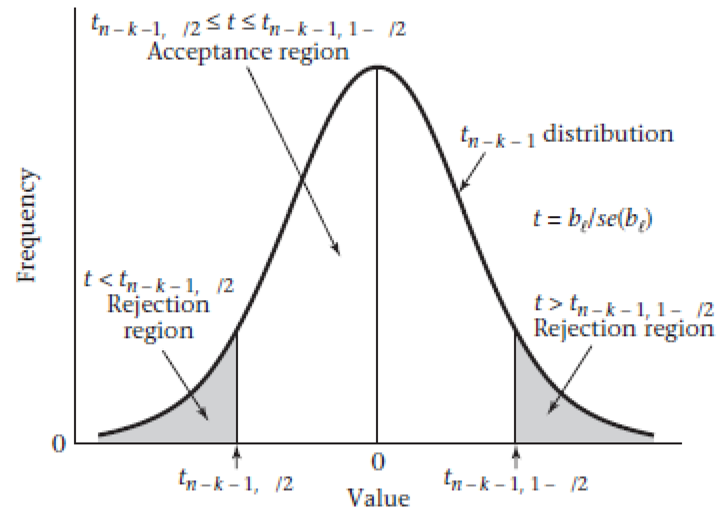
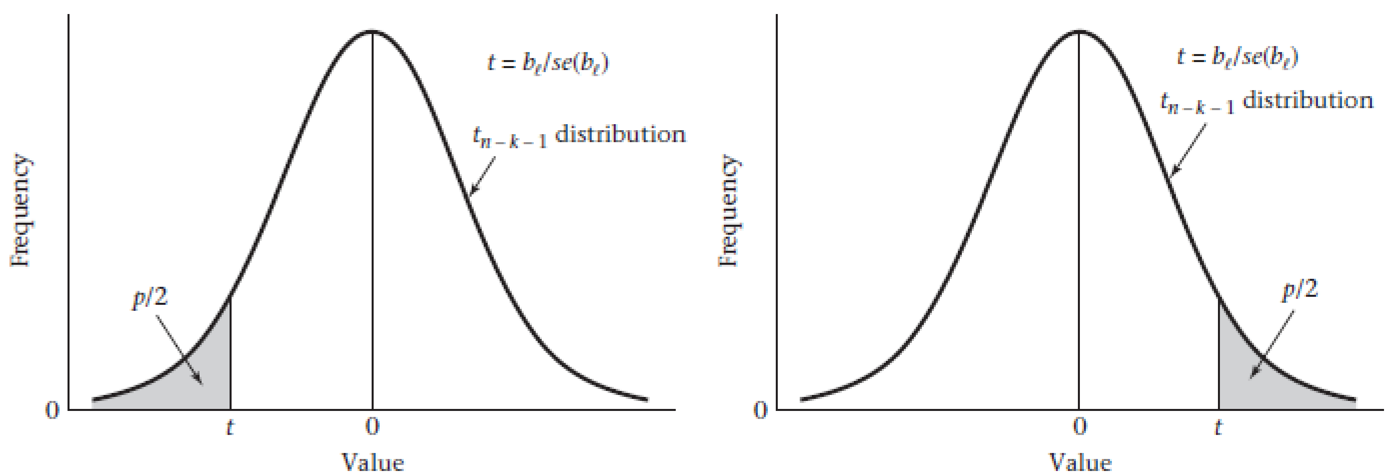


FIGURE 11.23 Computation of the exact *p*-value for the *t* test for multiple linear regression



Example: Pediatrics Hypertension

Question: Test the hypothesis H_0: $\boldsymbol{\beta_1} = \boldsymbol{\beta_2} = \mathbf{0}\ \boldsymbol{vs}.\ \boldsymbol{H_1}$: either $\boldsymbol{\beta_1} \neq \mathbf{0}$ or $\boldsymbol{\beta_2} \neq$ 0 using the data in Tables 11.8 and 11.9.

$$\text{Res SS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\text{Reg SS} = \text{Total SS} - \text{Res SS}$$

Solution: Refer to Table 11.9 and note that

Reg SS = 590.98 (called Model SS)

Reg MS = 590.98/2 = 295.49 (called Model MS)

Res SS = 79.9558 (called Error SS)

Res MS = 79.9558/13 = 6.150 (called Error MS)

F = Reg MS/Res MS = 48.05 $\sim F_{2,13}$ under H_0

Reg MS = Reg SS/k
Res MS = Res SS/(n-k-1)

Using R, the $p - value = \Pr(F_{2,13} > 48.05) = 1 - pf(48.05, 2, 13) < 0.01$

- Conclusion: the two variables, when considered together, are significant predictors of blood pressure.

TABLE 11.8 Sample data for infant blood pressure, age, and birthweight for 16 infants

i	Age (days) (x_1)	Birthweight (oz) (x_2)	SBP (mm Hg) (y)
1	3	135	89
2	4	120	90
3	3	100	83
4	2	105	77
5	4	130	92
6	5	125	98
7	2	125	82
8	3	105	85
9	5	120	96
10	4	90	95
11	2	120	80
12	3	95	79
13	3	120	86
14	4	150	97
15	3	160	92
16	3	125	88

$$\hat{y}_i = a + \sum_{j=1}^{k} b_j x_{ij}$$

$$Y = 53.45 + 5.89\,\text{age} + 0.126\,\text{birthweight}$$

$$\hat{y}_1 = 53.45 + 5.89(3) + 0.126(135) = 88.13$$

$$\hat{y}_2 = 53.45 + 5.89(4) + 0.126(120) = 92.13$$

$$\hat{y}_3 = 53.45 + 5.89(3) + 0.126(100) = 83.72$$

$$\hat{y}_4 = 53.45 + 5.89(2) + 0.126(105) = 78.46$$

$$\hat{y}_5 = 53.45 + 5.89(4) + 0.126(130) = 93.39$$

$$\hat{y}_6 = 53.45 + 5.89(5) + 0.126(125) = 98.65$$

$$\hat{y}_7 = 53.45 + 5.89(2) + 0.126(125) = 80.98$$

$$\hat{y}_8 = 53.45 + 5.89(3) + 0.126(105) = 84.35$$

$$\hat{y}_9 = 53.45 + 5.89(5) + 0.126(120) = 98.02$$

$$\hat{y}_{10} = 53.45 + 5.89(4) + 0.126(90) = 88.35$$

$$\hat{y}_{11} = 53.45 + 5.89(2) + 0.126(120) = 80.35$$

$$\hat{y}_{12} = 53.45 + 5.89(3) + 0.126(95) = 83.09$$

$$\hat{y}_{13} = 53.45 + 5.89(3) + 0.126(120) = 86.24$$

$$\hat{y}_{14} = 53.45 + 5.89(4) + 0.126(150) = 95.91$$

$$\hat{y}_{15} = 53.45 + 5.89(3) + 0.126(160) = 91.28$$

$$\hat{y}_{16} = 53.45 + 5.89(3) + 0.126(125) = 86.87$$

$$
\begin{aligned}
\text{Res SS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= (89-88.13)^2 && = 0.7569 && = 79.9558 \\
&+ (90-92.13)^2 && + 4.5369 \\
&+ (83-83.72)^2 && + 0.5184 \\
&+ (77-78.46)^2 && + 2.1316 \\
&+ (92-93.39)^2 && + 1.9321 \\
&+ (98-98.65)^2 && + 0.4225 \\
&+ (82-80.98)^2 && + 1.0404 \\
&+ (85-84.35)^2 && + 0.4225 \\
&+ (96-98.02)^2 && + 4.0804 \\
&+ (95-88.35)^2 && + 44.2225 \\
&+ (80-80.35)^2 && + 0.1225 \\
&+ (79-83.09)^2 && + 16.7281 \\
&+ (86-86.24)^2 && + 0.0576 \\
&+ (97-95.91)^2 && + 1.1881 \\
&+ (92-91.28)^2 && + 0.5184 \\
&+ (88-86.87)^2 && + 1.2769
\end{aligned}
$$

$$\text{Total SS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$\bar{y} = \frac{89+90+83+77+92+98+82+85+96+95+80+79+86+97+92+88}{16}$$

$$= 88.0625$$

$$\begin{aligned}
\text{Total SS} = {} & (89-88.0625)^2 & = 0.8789 & \quad = 670.9375 \\
& + (90-88.0625)^2 & + 3.7539 & \\
& + (83-88.0625)^2 & + 25.6289 & \\
& + (77-88.0625)^2 & + 122.3789 & \\
& + (92-88.0625)^2 & + 15.5039 & \\
& + (98-88.0625)^2 & + 98.7539 & \\
& + (82-88.0625)^2 & + 36.7539 & \\
& + (85-88.0625)^2 & + 9.3789 & \\
& + (96-88.0625)^2 & + 63.0039 & \\
& + (95-88.0625)^2 & + 48.1289 & \\
& + (80-88.0625)^2 & + 65.0039 & \\
& + (79-88.0625)^2 & + 82.1289 & \\
& + (86-88.0625)^2 & + 4.2539 & \\
& + (97-88.0625)^2 & + 79.8789 & \\
& + (92-88.0625)^2 & + 15.5039 & \\
& + (88-88.0625)^2 & + 0.00391 &
\end{aligned}$$

Reg SS = Total SS − Res SS

$= 670.9375 - 79.9558.$

~~591~~

$= 590.98.$

Reg MS = Reg SS / k $= \cancel{591}^{590.98} / 2 = 295.49$

Res SS = 79.9558.

$$\text{Res MS} = \frac{\text{Res SS}}{16-2-1} = \frac{79.9558}{13} = 6.150.$$

$$F = \frac{\text{Reg MS}}{\text{Res MS}} = \frac{295.49}{6.15} = 48.05.$$

Example: Pediatrics Hypertension

Question: Test for the independent contributions of age and birthweight in predicting SBP in infants, using the output in Table 11.9 .

Solution: From Table 11.9,

$\boxed{t = b_l / se(b_l)}$

se: standard error

$b_1 = 5.888$

$se(b_1) = 0.6802$

$t\ (age) = b_1 / se\ (b_1) = 8.66$

$p = 2 \times \Pr(t_{13} > 8.66) < 0.001$

$b_2 = 0.1256$

$se(b_2) = 0.0343$

$t\ (birthweight) = b_2 / se\ (b_2) = 3.66$

$p = 2 \times \Pr(t_{13} > 3.66) = 0.003$

$\boxed{\texttt{2*(1-pt(3.66, 13))}}$

- Conclusion: both age and birthweight have highly significant associations with SBP, even after controlling for the other variable.

TABLE 11.9 Least-squares estimates of the regression parameters for the newborn blood-pressure data in Table 11.8 using the SAS PROC REG program

The REG Procedure

Model: MODEL1

Dependent Variable: sysbp

Number of Observations Read	16
Number of Observations Used	16

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	591.03564	295.51782	48.08	<.0001
Error	13	79.90186	6.14630		
Corrected Total	15	670.93750			

Root MSE	2.47917	R-Square	0.8809
Dependent Mean	88.06250	Adj R-Sq	0.8626
Coeff Var	2.81524		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Squared Partial Corr Type II
Intercept	1	53.45019	4.53189	11.79	<.0001	.
agedys	1	5.88772	0.68021	8.66	<.0001	0.85214
brthwgt	1	0.12558	0.03434	3.66	0.0029	0.50715

TABLE 5 Percentage points of the t distribution ($t_{d,u}$)[a]

Degrees of freedom, d	u = .75	.80	.85	.90	.95	.975	.99	.995	.9995
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	0.679	0.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

Spearman rank-correlation coefficient (r_s): ordinary correlation coefficient based on ranks

- $r_s = L_{xy} / \sqrt{L_{xx} \times L_{xy}}$
- L's are computed from the rank (not actual score)

<u>t test for Spearman Rank Correlation</u>

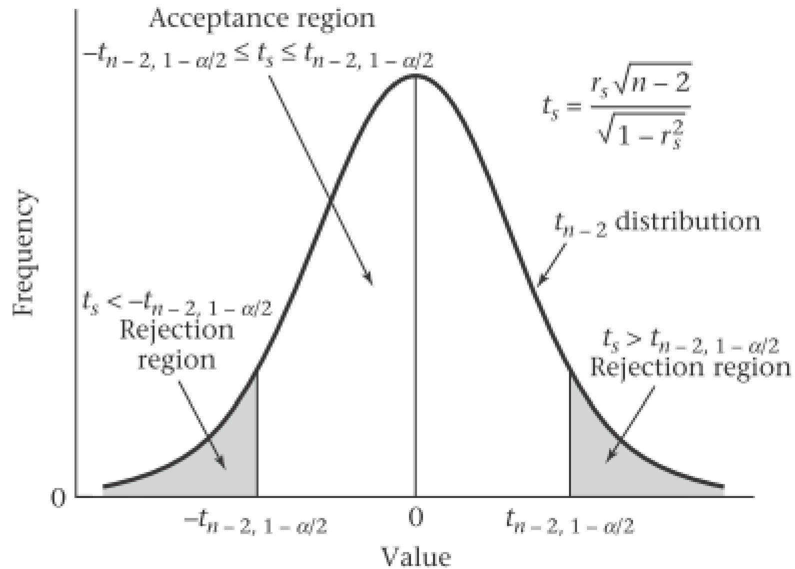
- H_0: no correlation
  - test statistic $t_s = r_s(\sqrt{n-2}) / \sqrt{1-r_s^2}$ follows a t distribution with n-2 degrees of freedom.
- For a two-sided level α test:
  - if $t_s > t_{n-2,1-\alpha/2}$ or $t_s < t_{n-2,\alpha/2} = -t_{n-2,1-\alpha/2}$ then reject H_0
  - otherwise, accept H_0.
- The exact p-value :

  $t_s < 0$: p = 2 × (area to the left of t_s under a t_{n-2} distribution)

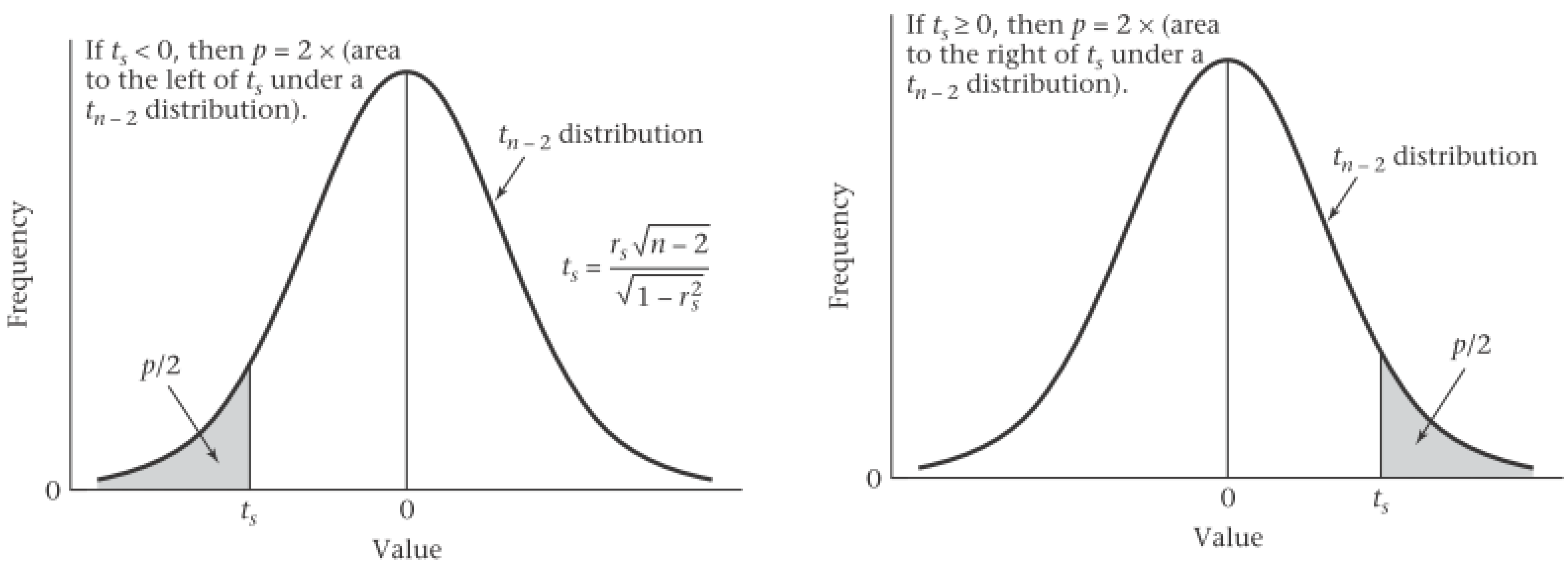
  $t_s \geq 0$: p = 2 × (area to the right of t_s under a t_{n-2} distribution)

*This test is valid only if **n $\geq$ 10**

Acceptance and rejection regions for the t test for a Spearman rank-correlation coefficient



Computation of the exact p-value for the t test for a Spearman rank-correlation coefficient



Summary

1.Statistical inference methods for investigating the relationship between two or more variables

2.If only two variables, both of which are continuous, are being studied, and we wish to predict one variable (the dependent variable) as a function of the other variable (the independent variable) then **simple linear regression analysis** is used.

3.Pearson correlation methods are used to determine the association between two normally distributed variables without distinguishing between dependent and independent variables.

4.Rank correlation may be used if both variables are continuous but not normally distributed or are ordinal variables.

5.Multiple regression methods may be used to predict the value of one variable (the dependent variable which is normally distributed) as a function of several independent variables.