Analysis-of-Variance Approach

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Regression Identity
The total sum of squares equals the regression sum of squares plus the error sum of squares:
SST = SSR + SSE.

Coefficient of Determination
The **coefficient of determination, r^2,** is the proportion of variation in the observed values of the response variable explained by the regression. Thus,

$$r^2 = \frac{SSR}{SST}.$$

$$r^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}.$$

EXAMPLE

The Coefficient of Determination Consider Age and Price of Orions data

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{975}{11} = 88.64$$

a. Compute SST, SSR, and SSE.

b. Compute the coefficient of determination, r^2.

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

$$SST = \Sigma(y_i - \bar{y})^2 = 9708.5.$$

Table for computing SSR for the Orion data

$$\hat{y} = 195.47 - 20.26x$$

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

$$SSR = \Sigma(\hat{y}_i - \bar{y})^2 = 8285.0,$$

$$r^2 = \frac{SSR}{SST} = \frac{8285.0}{9708.5} = 0.853 \quad (85.3\%).$$

Example 11.1:

$$\sum_{i=1}^{33} x_i = 1104, \ \sum_{i=1}^{33} y_i = 1124, \ \sum_{i=1}^{33} x_i y_i = 41{,}355, \ \sum_{i=1}^{33} x_i^2 = 41{,}086$$

the estimated regression line is given by

$$\hat{y} = 3.8296 + 0.9036x.$$

11.14 A professor in the School of Business in a university polled a dozen colleagues about the number of professional meetings they attended in the past five years (x) and the number of papers they submitted to refereed journals (y) during the same period. The summary data are given as follows:

$$n = 12, \quad \bar{x} = 4, \quad \bar{y} = 12,$$

$$\sum_{i=1}^{n} x_i^2 = 232, \quad \sum_{i=1}^{n} x_i y_i = 318.$$

Fit a simple linear regression model between x and y by finding out the estimates of intercept and slope. Com-

11.11 The thrust of an engine (y) is a function of exhaust temperature (x) in °F when other important variables are held constant. Consider the following data.

y	x	y	x
4300	1760	4010	1665
4650	1652	3810	1550
3200	1485	4500	1700
3150	1390	3008	1270
4950	1820		

(a) Plot the data.

(b) Fit a simple linear regression to the data and plot the line through the data.

12.2: Given the data

x	0	1	2	3	4	5	6	7	8	9
y	9.1	7.3	3.2	4.6	4.8	2.9	5.7	7.1	8.8	10.2

fit a regression curve of the form $\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2$ and then estimate $\mu_{Y|2}$.

$$\begin{aligned} 10b_0 + \quad 45\, b_1 + \quad 285\, b_2 &= 63.7, \\ 45b_0 + \quad 285b_1 + \quad 2025\, b_2 &= 307.3, \\ 285b_0 + 2025\, b_1 + 15{,}333b_2 &= 2153.3. \end{aligned}$$

Solving these normal equations, we obtain

$$b_0 = 8.698, \quad b_1 = -2.341, \quad b_2 = 0.288.$$

Therefore,

$$\hat{y} = 8.698 - 2.341x + 0.288x^2.$$

our estimate of $\mu_{Y|2}$ is

$$\hat{y} = 8.698 - (2.341)(2) + (0.288)(2^2) = 5.168.$$

12.4 An experiment was conducted to determine if the weight of an animal can be predicted after a given period of time on the basis of the initial weight of the animal and the amount of feed that was eaten. The following data, measured in kilograms, were recorded:

Final Weight, y	**Initial Weight, x_1**	**Feed Weight, x_2**
95	42	272
77	33	226
80	33	259
100	45	292
97	39	311
70	36	183
50	32	173
80	41	236
92	40	230
84	38	235

(a) Fit a multiple regression equation of the form

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Example:

y	X_1	X_2
140	60	22
155	62	25
159	67	24
179	70	20
192	71	15
200	72	14
212	75	14
215	78	11

$\hat{y} = -6.867 + 3.148x_1 - 1.656x_2$

	y	X_1	X_2
	140	60	22
	155	62	25
	159	67	24
	179	70	20
	192	71	15
	200	72	14
	212	75	14
	215	78	11
Mean	181.5	69.375	18.125
Sum	1452	555	145

	X_1^2	X_2^2	X_1y	X_2y	X_1X_2
	3600	484	8400	3080	1320
	3844	625	9610	3875	1550
	4489	576	10653	3816	1608
	4900	400	12530	3580	1400
	5041	225	13632	2880	1065
	5184	196	14400	2800	1008
	5625	196	15900	2968	1050
	6084	121	16770	2365	858
Sum	38767	2823	101895	25364	9859

12.13 A study was performed on a type of bearing to find the relationship of amount of wear y to x_1 = oil viscosity and x_2 = load. The following data

were obtained. (From *Response Surface Methodology*, Myers, Montgomery, and Anderson-Cook, 2009.)

y	x_1	x_2	y	x_1	x_2
193	1.6	851	230	15.5	816
172	22.0	1058	91	43.0	1201
113	33.0	1357	125	40.0	1115

(a) Estimate the unknown parameters of the multiple linear regression equation

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

(b) Predict wear when oil viscosity is 20 and load is 1200.

Example 12.4:

y (% survival)	x_1 (weight %)	x_2 (weight %)	x_3 (weight %)
25.5	1.74	5.30	10.80
31.2	6.32	5.42	9.40
25.9	6.22	8.41	7.20
38.4	10.52	4.63	8.50
18.4	1.19	11.60	9.40
26.7	1.22	5.85	9.90
26.4	4.10	6.62	8.00
25.9	6.32	8.72	9.10
32.0	4.08	4.42	8.70
25.2	4.15	7.60	9.20
39.7	10.15	4.83	9.40
35.7	1.72	3.12	7.60
26.5	1.70	5.30	8.20

Estimate the multiple linear regression model for the given data.

***Solution*:** The least squares estimating equations,

$$\begin{bmatrix} 13.0 & 59.43 & 81.82 & 115.40 \\ 59.43 & 394.7255 & 360.6621 & 522.0780 \\ 81.82 & 360.6621 & 576.7264 & 728.3100 \\ 115.40 & 522.0780 & 728.3100 & 1035.9600 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 377.5 \\ 1877.567 \\ 2246.661 \\ 3337.780 \end{bmatrix}$$

$$b_0 = 39.1574,\ b_1 = 1.0161,\ b_2 = -1.8616,\ b_3 = -0.3433.$$

Hence, our estimated regression equation is

$$\hat{y} = 39.1574 + 1.0161x_1 - 1.8616x_2 - 0.3433x_3.$$

Analysis-of-Variance Technique (One-Way ANOVA)

Random samples of size n are selected from each of k populations. The k different populations are classified on the basis of a single criterion such as different treatments or groups. Today the term **treatment** is used generally to refer to the various classifications, whether they be different aggregates, different analysts, different fertilizers, or different regions of the country.

Three Important Measures of Variability

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = \text{total sum of squares},$$

$$SSA = n\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2 = \text{treatment sum of squares},$$

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2 = \text{error sum of squares}.$$

$$SST = SSA + SSE.$$

Table 13.3: Analysis of Variance for the One-Way ANOVA

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	**Computed f**
Treatments	SSA	$k-1$	$s_1^2 = \dfrac{SSA}{k-1}$	$\dfrac{s_1^2}{s^2}$
Error	SSE	$k(n-1)$	$s^2 = \dfrac{SSE}{k(n-1)}$	
Total	SST	$kn-1$		

Sum of Squares, Unequal Sample Sizes

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2, \; SSA = \sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2, \; SSE = SST - SSA$$