Chapter 11

Simple Linear Regression and Correlation

The Simple Linear Regression (SLR) Model

the estimated or **fitted regression** line is given by

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

The Method of Least Squares

Estimating the Regression Coefficients

Given the sample $\{(x_i, y_i);\ i = 1, 2, \ldots, n\}$, the least squares estimates b_0 and b_1 of the regression coefficients β_0 and β_1 are computed from the formulas

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{and}$$

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - b_1 \bar{x}.$$

11.12 Correlation

Pearson product-moment correlation coefficient

Correlation Coefficient The measure ρ of linear association between two variables X and Y is estimated by the **sample correlation coefficient** r, where

$$r = b_1\sqrt{\frac{S_{xx}}{S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

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}	$\Sigma(y_i - \bar{y})^2$	$\Sigma y_i^2 - (\Sigma y_i)^2/n$

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

Formula for the *t* Test for the Correlation Coefficient

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

with degrees of freedom equal to $n - 2$.

Linear regression and correlation are two commonly used methods

- for examining the relationship between quantitative variables and
- for making predictions regression equation,
- the equation of the line that best fits a set of data points.

Coefficient of determination,

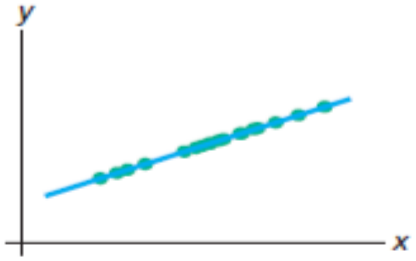
- a descriptive measure of the utility of the regression equation for making predictions linear correlation coefficient,
- it provides a descriptive measure of the strength of the linear relationship between two quantitative variables.

Show that

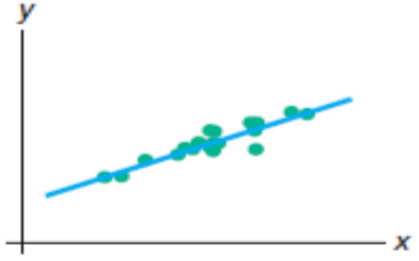
$$S_{xx} = \Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2/n$$

$$S_{xy} = \Sigma(x_i - \bar{x})(y_i - \bar{y}) = \Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n$$

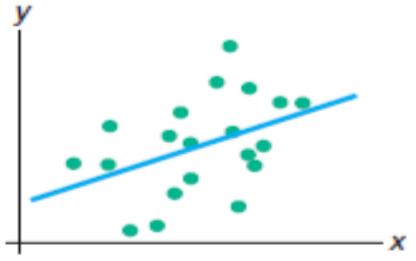
Various degrees of linear correlation



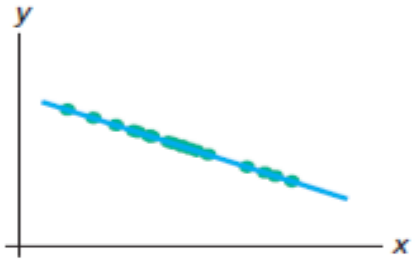
(a) Perfect positive linear correlation
$r = 1$



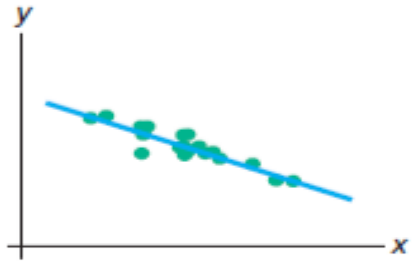
(b) Strong positive linear correlation
$r = 0.9$



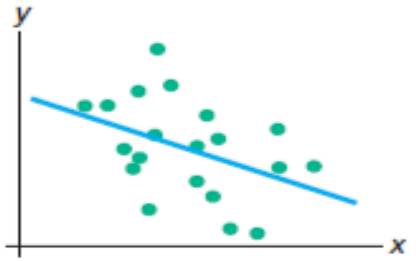
(c) Weak positive linear correlation
$r = 0.4$



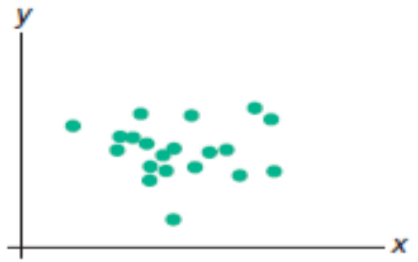
(d) Perfect negative linear correlation
$r = -1$



(e) Strong negative linear correlation
$r = -0.9$



(f) Weak negative linear correlation
$r = -0.4$



(g) No linear correlation (linearly uncorrelated)
$r = 0$

Car Rental Companies

Example:

Construct a scatter plot for the data shown for car rental companies in the United States for a recent year.

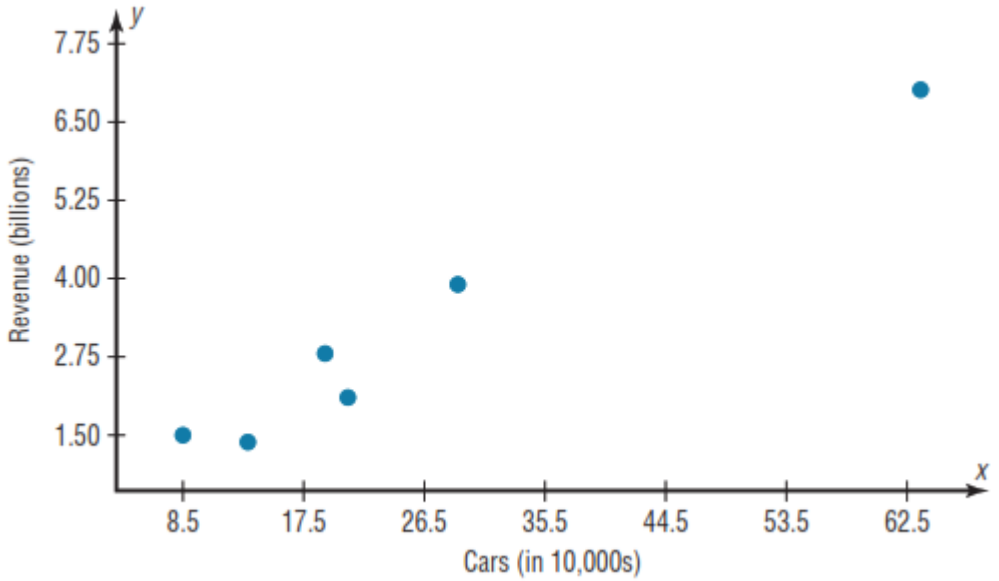
Company	Cars (in ten thousands)	Revenue (in billions)
A	63.0	$7.0
B	29.0	3.9
C	20.8	2.1
D	19.1	2.8
E	13.4	1.4
F	8.5	1.5

Plot a scatter diagram.

Compute the correlation coefficient

Find the equation of the regression line

Test the significance of the correlation coefficient Use $\alpha = 0.05$



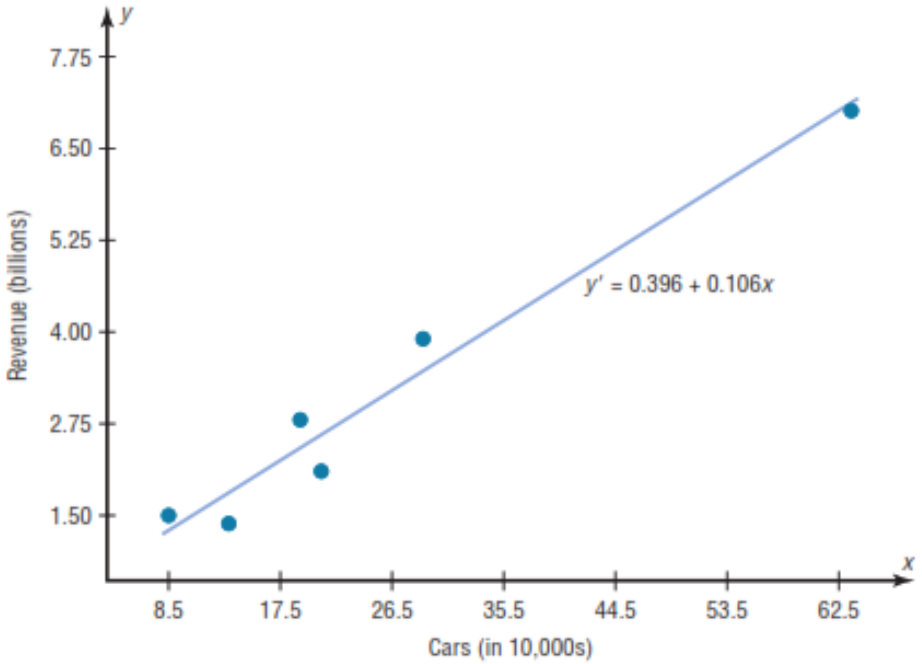
Cars x (in 10,000s)	Revenue y (in billions)	xy	x^2	y^2
63.0	7.0	441.00	3969.00	49.00
29.0	3.9	113.10	841.00	15.21
20.8	2.1	43.68	432.64	4.41
19.1	2.8	53.48	364.81	7.84
13.4	1.4	18.76	179.56	1.96
8.5	1.5	12.75	72.25	2.25
$\Sigma x = 153.8$	$\Sigma y = 18.7$	$\Sigma xy = 682.77$	$\Sigma x^2 = 5859.26$	$\Sigma y^2 = 80.67$

Substitute in the formula and solve for r:

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

$$= \frac{(6)(682.77) - (153.8)(18.7)}{\sqrt{[(6)(5859.26) - (153.8)^2][(6)(80.67) - (18.7)^2]}} = 0.982$$

The correlation coefficient suggests a strong relationship between the number of cars a rental agency has and its annual revenue.



$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{(18.7)(5859.26) - (153.8)(682.77)}{(6)(5859.26) - (153.8)^2} = 0.396$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{6(682.77) - (153.8)(18.7)}{(6)(5859.26) - (153.8)^2} = 0.106$$

$$y' = 0.396 + 0.106x$$

Test the significance of the correlation coefficient Use $\alpha = 0.05$ and $r = 0.982$.

Formula for the *t* Test for the Correlation Coefficient

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

with degrees of freedom equal to $n - 2$.

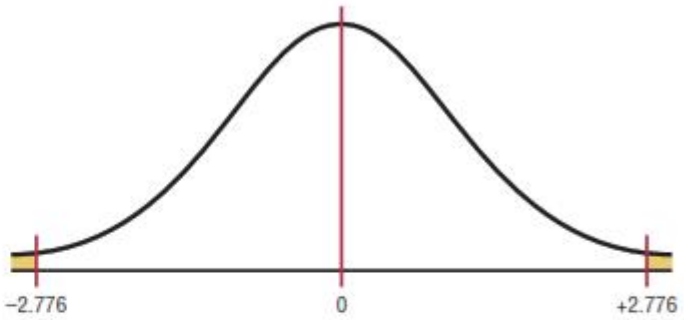
Solution

Step 1 State the hypotheses.

$H_0: \rho = 0$ and $H_1: \rho \neq 0$

Step 2 Find the critical values

. Since $\alpha = 0.05$ and there are $6 - 2 = 4$ degrees of

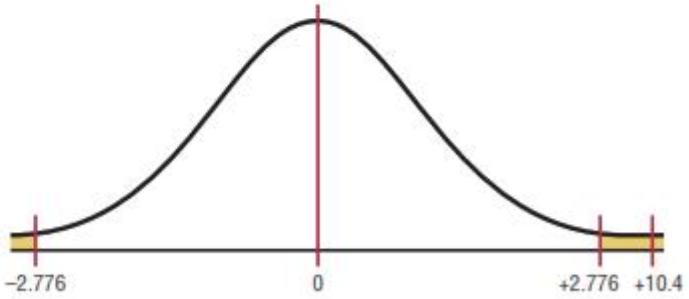


Step 3 Compute the test value.

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.982\sqrt{\frac{6-2}{1-(0.982)^2}} = 10.4$$

Step 4 Make the decision.

Reject the null hypothesis, since the test value falls in the critical region,



Step 5 Summarize the results.

There is a significant relationship between the number of cars a rental agency owns and its annual income.

Example 11.1: $\sum_{i=1}^{33} x_i = 1104, \quad \sum_{i=1}^{33} y_i = 1124, \quad \sum_{i=1}^{33} x_i y_i = 41{,}355, \quad \sum_{i=1}^{33} x_i^2 = 41{,}086$

the estimated regression line is given by

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

Home Activity

formula

$$y = a + bx$$

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

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:

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

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

$$\hat{y} = 37.8 - 6.45x.$$

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

11.2 The grades of a class of 9 students on a midterm report (x) and on the final examination (y) are as follows:

x	77	50	71	72	81	94	96	99	67
y	82	66	78	34	47	85	99	99	68

(a) Estimate the linear regression line.

(b) Estimate the final examination grade of a student who received a grade of 85 on the midterm report.

Solution:

(a) $\sum_i x_i = 707, \sum_i y_i = 658, \sum_i x_i^2 = 57,557, \sum_i x_i y_i = 53,258, n = 9.$

$$b = \frac{(9)(53,258) - (707)(658)}{(9)(57,557) - (707)^2} = 0.7771,$$

$$a = \frac{658 - (0.7771)(707)}{9} = 12.0623.$$

Hence $\hat{y} = 12.0623 + 0.7771x$.

(b) For $x = 85$, $\hat{y} = 12.0623 + (0.7771)(85) = 78$.

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

$$a = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n}$$

11.13 A study of the amount of rainfall and the quantity of air pollution removed produced the following data:

Daily Rainfall, x (0.01 cm)	Particulate Removed, y ($\mu g/m^3$)
4.3	126
4.5	121
5.9	116
5.6	118
6.1	114
5.2	118
3.8	132
2.1	141
7.5	108

(a) Find the equation of the regression line to predict the particulate removed from the amount of daily rainfall.

(b) Estimate the amount of particulate removed when the daily rainfall is $x = 4.8$ units.

Solution:

(a) $\sum_i x_i = 45, \sum_i y_i = 1094, \sum_i x_i^2 = 244.26, \sum_i x_i y_i = 5348.2, n = 9.$

$$b = \frac{(9)(5348.2) - (45)(1094)}{(9)(244.26) - (45)^2} = -6.3240,$$

$$a = \frac{1094 - (-6.3240)(45)}{9} = 153.1755.$$

Hence $\hat{y} = 153.1755 - 6.3240x$.

(b) For $x = 4.8$, $\hat{y} = 153.1755 - (6.3240)(4.8) = 123$.

11.1 A study was conducted at Virginia Tech to determine if certain static arm-strength measures have an influence on the "dynamic lift" characteristics of an individual. Twenty-five individuals were subjected to strength tests and then were asked to perform a weight-lifting test in which weight was dynamically lifted overhead. The data are given here.

Individual	Arm Strength, x	Dynamic Lift, y
1	17.3	71.7
2	19.3	48.3
3	19.5	88.3
4	19.7	75.0
5	22.9	91.7
6	23.1	100.0
7	26.4	73.3
8	26.8	65.0
9	27.6	75.0
10	28.1	88.3
11	28.2	68.3
12	28.7	96.7
13	29.0	76.7
14	29.6	78.3
15	29.9	60.0
16	29.9	71.7
17	30.3	85.0
18	31.3	85.0
19	36.0	88.3
20	39.5	100.0
21	40.4	100.0
22	44.3	100.0
23	44.6	91.7
24	50.4	100.0
25	55.9	71.7

(a) Estimate β_0 and β_1 for the linear regression curve $\mu_{Y|x} = \beta_0 + \beta_1 x$.

(b) Find a point estimate of $\mu_{Y|30}$.

Solution:

(a) $\sum_i x_i = 778.7$, $\sum_i y_i = 2050.0$, $\sum_i x_i^2 = 26,591.63$,

$\sum_i x_i y_i = 65,164.04$, $n = 25$.

$$b = \frac{(25)(65,164.04) - (778.7)(2050.0)}{(25)(26,591.63) - (778.7)^2} = 0.5609,$$

$$a = \frac{2050 - (0.5609)(778.7)}{25} = 64.53.$$

$$\hat{y} = 64.53 + 0.5609x$$

(b) Using the equation $\hat{y} = 64.53 + 0.5609x$ with $x = 30$,

$$= 81.40.$$

11.44 With reference to Exercise 11.1 on page 398, assume that x and y are random variables with a bivariate normal distribution.

(a) Calculate r.

(b) Test the hypothesis that $\rho = 0$ against the alternative that $\rho \neq 0$ at the 0.05 level of significance.

Solution:

(a)

data of Exercise 11.1 we can calculate

$$S_{xx} = 26,591.63 - (778.7)^2/25 = 2336.6824,$$
$$S_{yy} = 172,891.46 - (2050)^2/25 = 4791.46,$$
$$S_{xy} = 65,164.04 - (778.7)(2050)/25 = 1310.64.$$

$$r = b_1\sqrt{\frac{S_{xx}}{S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

$$r = \frac{1310.64}{\sqrt{(2236.6824)(4791.46)}} = 0.392.$$

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$ V=n-2 degree of freedom

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

formulas

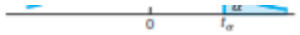
(b) The hypotheses are

$$H_0 : \rho = 0,$$
$$H_1 : \rho \neq 0.$$

$\alpha = 0.05.$

Critical regions: $t < -2.069$ or $t > 2.069$.
Computations: $t = \frac{0.392\sqrt{23}}{\sqrt{1-0.392^2}} = 2.04$.
Decision: Fail to reject H_0 at level 0.05.

Table A.4 Critical Values of the t-Distribution



	α						
v	0.40	0.30	0.20	0.15	0.10	0.05	0.025
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.538	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960

Table A.4 (continued) Critical Values of the t-Distribution

	α						
v	0.02	0.015	0.01	0.0075	0.005	0.0025	0.0005
1	15.894	21.205	31.821	42.433	63.656	127.321	636.578
2	4.849	5.643	6.965	8.073	9.925	14.089	31.600
3	3.482	3.896	4.541	5.047	5.841	7.453	12.924
4	2.999	3.298	3.747	4.088	4.604	5.598	8.610
5	2.757	3.003	3.365	3.634	4.032	4.773	6.869
6	2.612	2.829	3.143	3.372	3.707	4.317	5.959
7	2.517	2.715	2.998	3.203	3.499	4.029	5.408
8	2.449	2.634	2.896	3.085	3.355	3.833	5.041
9	2.398	2.574	2.821	2.998	3.250	3.690	4.781
10	2.359	2.527	2.764	2.932	3.169	3.581	4.587
11	2.328	2.491	2.718	2.879	3.106	3.497	4.437
12	2.303	2.461	2.681	2.836	3.055	3.428	4.318
13	2.282	2.436	2.650	2.801	3.012	3.372	4.221
14	2.264	2.415	2.624	2.771	2.977	3.326	4.140
15	2.249	2.397	2.602	2.746	2.947	3.286	4.073
16	2.235	2.382	2.583	2.724	2.921	3.252	4.015
17	2.224	2.368	2.567	2.706	2.898	3.222	3.965
18	2.214	2.356	2.552	2.689	2.878	3.197	3.922
19	2.205	2.346	2.539	2.674	2.861	3.174	3.883
20	2.197	2.336	2.528	2.661	2.845	3.153	3.850
21	2.189	2.328	2.518	2.649	2.831	3.135	3.819
22	2.183	2.320	2.508	2.639	2.819	3.119	3.792
23	2.177	2.313	2.500	2.629	2.807	3.104	3.768
24	2.172	2.307	2.492	2.620	2.797	3.091	3.745
25	2.167	2.301	2.485	2.612	2.787	3.078	3.725
26	2.162	2.296	2.479	2.605	2.779	3.067	3.707
27	2.158	2.291	2.473	2.598	2.771	3.057	3.689
28	2.154	2.286	2.467	2.592	2.763	3.047	3.674
29	2.150	2.282	2.462	2.586	2.756	3.038	3.660
30	2.147	2.278	2.457	2.581	2.750	3.030	3.646
40	2.123	2.250	2.423	2.542	2.704	2.971	3.551
60	2.099	2.223	2.390	2.504	2.660	2.915	3.460
120	2.076	2.196	2.358	2.468	2.617	2.860	3.373
∞	2.054	2.170	2.326	2.432	2.576	2.807	3.290

Use of z-test

11.45 With reference to Exercise 11.13 on page 400, assume a bivariate normal distribution for x and y.

(a) Calculate r.

(b) Test the null hypothesis that $\rho = -0.5$ against the alternative that $\rho < -0.5$ at the 0.025 level of significance.

(c) Determine the percentage of the variation in the amount of particulate removed that is due to changes in the daily amount of rainfall.

$$z = \boxed{\frac{\sqrt{n-3}}{2} \ln \left[\frac{(1+r)(1-\rho_0)}{(1-r)(1+\rho_0)} \right]}$$

Solution:

(a)

$$S_{xx} = 244.26 - 45^2/9$$

$$S_{yy} = 133,786 - 1094^2/9 = 804.2222,$$

and $S_{xy} = 5348.2 - (45)(1094)/9 = -121.8.$

$$r = \boxed{\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}}.$$

$$r = \frac{-121.8}{\sqrt{(19.26)(804.2222)}} = -0.979.$$

(b) The hypotheses are

$$H_0 : \rho = -0.5,$$
$$H_1 : \rho < -0.5.$$

$\alpha = 0.025.$
Critical regions: $z < -1.96.$
Computations: $z = \frac{\sqrt{6}}{2} \ln \left[\frac{(0.021)(1.5)}{(1.979)(0.5)} \right] = -4.22.$
Decision: Reject H_0; $\rho < -0.5$.

(c) $(-0.979)^2(100\%) = 95.8\%.$

Use of t-test

Class Activity

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

11.47 The following data were obtained in a study of the relationship between the weight and chest size of infants at birth.

Weight (kg)	Chest Size (cm)
2.75	29.5
2.15	26.3
4.41	32.2
5.52	36.5
3.21	27.2
4.32	27.7
2.31	28.3
4.30	30.3
3.71	28.7

(a) Calculate r.

(b) Test the null hypothesis that $\rho = 0$ against the alternative that $\rho > 0$ at the 0.01 level of significance.

(c) What percentage of the variation in infant chest sizes is explained by difference in weight?

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

Solution:

(a) $S_{xx} = 128.6602 - 32.68^2/9 = 9.9955,$

$S_{yy} = 7980.83 - 266.7^2/9 = 77.62,$

$S_{xy} = 990.268 - (32.68)(266.7)/9 = 21.8507.$

$$r = \frac{21.8507}{\sqrt{(9.9955)(77.62)}} = 0.784.$$

(b) The hypotheses are

$H_0 : \rho = 0,$

$H_1 : \rho > 0.$

$\alpha = 0.01.$

Critical regions: $t > 2.998.$

Computations: $t = \frac{0.784\sqrt{7}}{\sqrt{1-0.784^2}} = 3.34.$

Decision: Reject H_0; $\rho > 0.$

(c) $(0.784)^2(100\%) = 61.5\%.$

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,

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

$$r^2 = \frac{\text{explained variation}}{\text{total variation}}$$

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

Class Activity

EXAMPLE

The Coefficient of Determination Consider Age and Price of Orions data

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

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

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

Regression Identity

SST=SSR + SSE

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

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

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

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

a. Compute SST, SSR, and SSE.
b. Compute the coefficient of determination, r^2.

Table for computing SSR for the Orion data

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

the estimated or **fitted regression** line is

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

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

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

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

Chapter 12

Multiple Linear Regression and Nonlinear Regression Models

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_k x_k,$$

Multiple Regression coefficient

$$\widehat{b_1} = \frac{(\sum x_1 y)(\sum x_2{}^2) - (\sum x_2 y)(\sum x_1 x_2)}{(\sum x_1{}^2)(\sum x_2{}^2) - (\sum x_1 x_2)^2}$$

$$\widehat{b_2} = \frac{(\sum x_2 y)(\sum x_1{}^2) - (\sum x_1 y)(\sum x_1 x_2)}{(\sum x_1{}^2)(\sum x_2{}^2) - (\sum x_1 x_2)^2}$$

$$\widehat{b_0} = \bar{Y} - \hat{b}_1 \bar{X}_1 - \hat{b}_2 \bar{X}_2$$

polynomial regression model

$$\hat{y} = b_0 + b_1 x + b_2 x^2 + \cdots + b_r x^r.$$

$$\boldsymbol{y = a + bx + cx^2} \qquad \textbf{(Quadratic)}$$

$$\sum y = na + b\sum x + c\sum x^2$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$$

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

Normal Estimation Equations for Multiple Linear Regression

$$
\begin{array}{llllll}
nb_0 + b_1 \sum_{i=1}^{n} x_{1i} & + b_2 \sum_{i=1}^{n} x_{2i} & + \cdots + b_k \sum_{i=1}^{n} x_{ki} & = \sum_{i=1}^{n} y_i \\
b_0 \sum_{i=1}^{n} x_{1i} + b_1 \sum_{i=1}^{n} x_{1i}^2 & + b_2 \sum_{i=1}^{n} x_{1i}x_{2i} + & \cdots + b_k \sum_{i=1}^{n} x_{1i}x_{ki} & = \sum_{i=1}^{n} x_{1i}y_i \\
\vdots \quad \vdots & \vdots & \vdots & \vdots \\
b_0 \sum_{i=1}^{n} x_{ki} + b_1 \sum_{i=1}^{n} x_{ki}x_{1i} + & b_2 \sum_{i=1}^{n} x_{ki}x_{2i} + & \cdots + b_k \sum_{i=1}^{n} x_{ki}^2 & = \sum_{i=1}^{n} x_{ki}y_i
\end{array}
$$

These equations can be solved for $b_0, b_1, b_2, \ldots, b_k$ by any appropriate method for solving systems of linear equations. Most statistical software can be used to obtain numerical solutions of the above equations.

Formula for the multiple correlation coefficient:

$$R = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1} \cdot r_{yx_2} \cdot r_{x_1x_2}}{1 - r_{x_1x_2}^2}}$$

Example:

Estimate the multiple linear regression equation

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

Class Activity

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

Multiple Regression coefficient

$$\hat{b}_1 = \frac{(\sum x_1 y)(\sum x_2^2) - (\sum x_2 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$\hat{b}_2 = \frac{(\sum x_2 y)(\sum x_1^2) - (\sum x_1 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}_1 - \hat{b}_2 \bar{X}_2$$

= 181.5 – 3.148(69.375) – (-1.656)(18.125) = **-6.867**

Solution:

	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
38767	2823	101895	25364	9859

Sum

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

For example, the nursing instructor wishes to see whether a student's grade point average and age are related to the student's score on the state board nursing examination. She selects five students and obtains the following data.

Student	GPA x_1	Age x_2	State board score y
A	3.2	22	550
B	2.7	27	570
C	2.5	24	525
D	3.4	28	670
E	2.2	23	490

Estimate the multiple linear regression equation

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

OR

Estimate the multiple linear regression equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2.$$

For the data regarding state board scores, find the value of R.

The multiple regression equation obtained from the data is

$$y' = -44.81 + 87.64x_1 + 14.533x_2$$

Home work

Formula for the multiple correlation coefficient:

$$R = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1} \cdot r_{yx_2} \cdot r_{x_1x_2}}{1 - r_{x_1x_2}^2}}$$

For the data regarding state board scores, find the value of R.

Solution

The values of the correlation coefficients are

$$r_{yx_1} = 0.845$$
$$r_{yx_2} = 0.791$$
$$r_{x_1x_2} = 0.371$$

Substituting in the formula, you get

$$R = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1} \cdot r_{yx_2} \cdot r_{x_1x_2}}{1 - r_{x_1x_2}^2}}$$

$$= \sqrt{\frac{(0.845)^2 + (0.791)^2 - 2(0.845)(0.791)(0.371)}{1 - 0.371^2}}$$

$$= \sqrt{\frac{0.8437569}{0.862359}} = \sqrt{0.9784288} = 0.989$$

Normal Estimation Equation for linear , polynomial and Multiple linear regression

$$\boldsymbol{y = a + bx}$$

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

$$\boldsymbol{y = a + bx + cx^2}$$

$$\sum y = na + b\sum x + c\sum x^2$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$$

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

$$\boldsymbol{y = a + bx_1 + cx_2}$$

$$\sum y = na + b\sum x_1 + c\sum x_2$$

$$\sum x_1 y = a\sum x_1 + b\sum x_1^2 + c\sum x_1 x_2$$

$$\sum x_2 y = a\sum x_2 + b\sum x_1 x_2 + c\sum x_2^2$$

Class Activity with Calculator

Polynomial Regression

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 $\mu_{Y|2}$.

Solution:

$$y = a + bx + cx^2$$

$$\sum y = na + b\sum x + c\sum x^2$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$$

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

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

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

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

(a) $\hat{y} = -22.99316 + 1.39567x_1 + 0.21761x_2.$

Solution 12.4

	x1	x2	y		x1 y	x2 y	Sq(x1)	sq (x2)	x1 x2
	42	272	95		3990	25840	1764	73984	11424
	33	226	77		2541	17402	1089	51076	7458
	33	259	80		2640	20720	1089	67081	8547
	45	292	100		4500	29200	2025	85264	13140
	39	311	97		3783	30167	1521	96721	12129
	36	183	70		2520	12810	1296	33489	6588
	32	173	50		1600	8650	1024	29929	5536
	41	236	80		3280	18880	1681	55696	9676
	40	230	92		3680	21160	1600	52900	9200
	38	235	84		3192	19740	1444	55225	8930
sum	**379**	**2417**	**825**		**31726**	**204569**	**14533**	**601365**	**92628**

Normal equation of Multiple Regression:

$$y = a + bx_1 + cx_2$$

$$\sum y = na + b\sum x_1 + c\sum x_2$$

$$\sum x_1 y = a\sum x_1 + b\sum {x_1}^2 + c\sum x_1 x_2$$

$$\sum x_2 y = a\sum x_2 + b\sum x_1 x_2 + c\sum {x_2}^2$$

Substitute values from table in normal equation and Form the linear equation and solve via calculator for multiple linear coefficient

$$\text{(a)}\ \ \hat{y} = -22.99316 + 1.39567x_1 + 0.21761x_2.$$

Chapter 13

One-Factor Experiments: General

13.1 Analysis-of-Variance Technique

ANOVA.

13.3 One-Way Analysis of Variance: Completely Randomized Design (One-Way ANOVA)

Assumptions and Hypotheses in One-Way ANOVA

H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$,

H_1: At least two of the means are not equal.

Use of *F*-Test in ANOVA

Source	df	*SS*	*MS* = *SS*/df	*F*-statistic
Residual	$k-1$	*SSR*	$MSR = \frac{SSR}{k-1}$	$F = \frac{MSR}{MSE}$
Error	$n-k$	*SSE*	$MSE = \frac{SSE}{n-k}$	
Total	$n-1$	*SST*		

The null hypothesis H_0 is rejected at the α-level of significance when

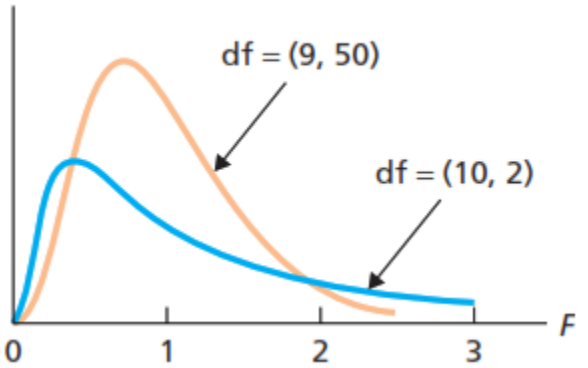
$$f > f_\alpha[k-1, k(n-1)].$$

The *F*-Distribution

Analysis-of-variance procedures rely on a distribution called the *F-distribution*, named in honor of Sir Ronald Fisher.

A variable is said to have an ***F*-distribution** if its distribution has the shape of a special type of right-skewed curve, called an ***F*-curve.** There are infinitely many *F*-distributions, and we identify an *F*-distribution (and *F*-curve) by its number of degrees of freedom, just as we did for *t*-distributions

Two different *F*-curves



An *F*-distribution, however, has two numbers of degrees of freedom instead of one. Figure depicts two different *F*-curves;

one has df = (10, 2), and the other has df = (9, 50).

The first number of degrees of freedom for an *F*-curve is called the **degrees of freedom for the numerator,** and the second is called the **degrees of freedom for the denominator.**

Thus, for the *F*-curve in Fig. with df = (10, 2), we have

df = (10, 2)

Degrees of freedom for the numerator

Degrees of freedom for the denominator

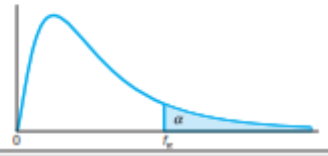


Table A.6 Critical Values of the *F*-Distribution

$f_{0.05}(v_1, v_2)$

	v_1								
v_2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

Table A.6 (continued) Critical Values of the *F*-Distribution

$f_{0.05}(v_1, v_2)$

	v_1									
v_2	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
1	241.88	243.91	245.95	248.01	249.05	250.10	251.14	252.20	253.25	254.31
2	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

PROCEDURE

One-Way ANOVA Test

Purpose To perform a hypothesis test to compare k population means, $\mu_1, \mu_2, \ldots, \mu_k$

Assumptions

1. Simple random samples
2. Independent samples
3. Normal populations
4. Equal population standard deviations

Step 1 **The null and alternative hypotheses are, respectively,**

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

H_a: **Not all the means are equal.**

Step 2 **Decide on the significance level, α.**

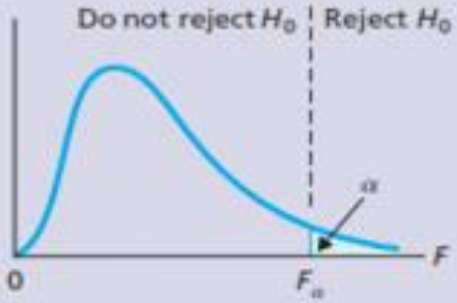
Step 3 **Compute the value of the test statistic**

$$F = \frac{MSR}{MSE}$$

and denote that value F_0. To do so, construct a one-way ANOVA table:

Source	df	*SS*	$MS = SS/df$	F-statistic
Regression	$k-1$	*SSR*	$MSR = \frac{SSR}{k-1}$	$F = \frac{MSR}{MSE}$
Error	$n-k$	*SSE*	$MSE = \frac{SSE}{n-k}$	
Total	$n-1$	*SST*		

Step 4 **The critical value is F_α with df $= (k-1, n-k)$. Use Table VIII to find the critical value.**



Step 5 **If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.**

Step 6 **Interpret the results of the hypothesis test.**

Example:

$$SSR = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \cdots + n_k(\bar{x}_k - \bar{x})^2.$$

$$SSE = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2.$$

Energy Consumption The Energy Information Administration gathers data on residential energy consumption and expenditures and publishes its findings in *Residential Energy Consumption Survey: Consumption and Expenditures*. Suppose that we want to decide whether a difference exists in mean annual energy consumption by households among the four U.S. regions.

Let μ_1, μ_2, μ_3, and μ_4 denote last year's mean energy consumptions by households in the Northeast, Midwest, South, and West, respectively.

Northeast	Midwest	South	West
15	17	11	10
10	12	7	12
13	18	9	8
14	13	13	7
13	15		9
	12		
13.0	14.5	10.0	9.2

← *Means*

a) Construct one-way ANOVA table

b) Test the hypothesis at 0.05 level of significance

Solution:

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{15 + 10 + 13 + \cdots + 7 + 9}{20} = \frac{238}{20} = 11.9.$$

$$SSR = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + n_3(\bar{x}_3 - \bar{x})^2 + n_4(\bar{x}_4 - \bar{x})^2$$

$$= 5(13.0 - 11.9)^2 + 6(14.5 - 11.9)^2 + 4(10.0 - 11.9)^2 + 5(9.2 - 11.9)^2$$

$$= 97.5.$$

$$MSR = \frac{SSR}{k - 1} = \frac{97.5}{4 - 1} = 32.5.$$

, we determine *MSE*. We have $k = 4, n_1 = 5, n_2 = 6, n_3 = 4, n_4 = 5,$ $n = 20$. Computing the variance of each sample gives

$s_1^2 = 3.5, s_2^2 = 6.7, s_3^2 = 6.\overline{6},$ and $s_4^2 = 3.7.$

Use calculator

$$MSE = \frac{SSE}{n - k} = \frac{82.3}{20 - 4} = 5.144.$$

Finally, $$F = \frac{MSR}{MSE} = \frac{32.5}{5.144} = 6.32.$$

Source	df	*SS*	*MS* = *SS*/df	*F*-statistic
Resudual	3	97.5	32.500	6.32
Error	16	82.3	5.144	
Total	19	179.8		

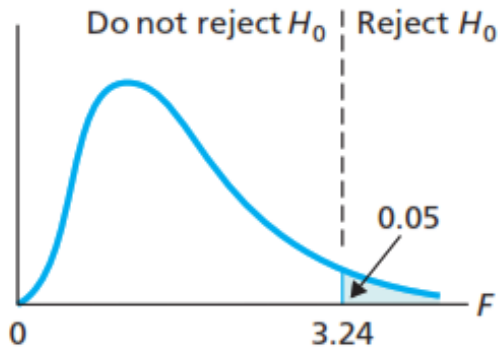
H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ (mean energy consumptions are all equal) **Step 1 State the null and alternative hypotheses.**

H_a: Not all the means are equal.

Step 2 Decide on the significance level, α.

$$\text{df} = (k - 1, n - k) = (4 - 1, 20 - 4) = (3, 16).$$

Step 3 Compute the value of the test statistic



Step 4 The critical value is F_α with df = $(k - 1, n - k)$. Use Table to find the critical value.

Step 5 If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.

From Step 3, the value of the test statistic is $F = 6.32$, falls in the rejection re gion. Thus we reject H_0.

The test results are statistically significant at the 5% level.

Example

Compute SST ,SSR ,SSE ?

use of the computing formulas.

Northeast	Midwest	South	West
15	17	11	10
10	12	7	12
13	18	9	8
14	13	13	7
13	15		9
	12		
13.0	14.5	10.0	9.2

← *Means*

$$k = 4$$

$$n_1 = 5 \qquad n_2 = 6 \qquad n_3 = 4 \qquad n_4 = 5$$

$$T_1 = 65 \qquad T_2 = 87 \qquad T_3 = 40 \qquad T_4 = 46$$

and

$$n = \Sigma n_j = 5 + 6 + 4 + 5 = 20$$

$$\Sigma x_i = \Sigma T_j = 65 + 87 + 40 + 46 = 238.$$

Summing the squares of all the data

$$\Sigma x_i^2 = (15)^2 + (10)^2 + (13)^2 + \cdots + (7)^2 + (9)^2 = 3012.$$

$$SST = \Sigma x_i^2 - (\Sigma x_i)^2/n$$

$$= 3012 - (238)^2/20 = 3012 - 2832.2 = 179.8,$$

$$SSR = \Sigma(T_j^2/n_j) - (\Sigma x_i)^2/n$$

$$= (65)^2/5 + (87)^2/6 + (40)^2/4 + (46)^2/5 - (238)^2/20$$

$$= 2929.7 - 2832.2 = 97.5,$$

$$SSE = SST - SSR = 179.8 - 97.5 = 82.3.$$

Lowering Blood Pressure

A researcher wishes to try three different techniques to lower the blood pressure of individuals diagnosed with high blood pressure. The subjects are randomly assigned to three groups; the first group takes medication, the second group exercises, and the third group follows a special diet. After four weeks, the reduction in each person's blood pressure is recorded. At $\alpha = 0.05$, test the claim that there is no difference among the means. The data are shown.

Medication	Exercise	Diet
10	6	5
12	8	9
9	3	12
15	0	8
13	2	4

$$SSR = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \cdots + n_k(\bar{x}_k - \bar{x})^2.$$

$$SSE = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2.$$

Solution

Step 1 State the hypotheses and identify the claim.

H_0: $\mu_1 = \mu_2 = \mu_3$ (claim)

H_1: At least one mean is different from the others.

Step 2 Find the critical value. Since $k = 3$ and $N = 15$,

d.f.N. $= k - 1 = 3 - 1 = 2$

d.f.D. $= N - k = 15 - 3 = 12$

The critical value is 3.89, $\alpha = 0.05$.

Step 3 Compute the test value,

$\bar{X}_1 = 11.8$ $\bar{X}_2 = 3.8$ $\bar{X}_3 = 7.6$

$s_1^2 = 5.7$ $s_2^2 = 10.2$ $s_3^2 = 10.3$

Step 4 Make the decision. The decision is to reject the null hypothesis, since $9.17 > 3.89$.

Step 5 Summarize the results. There is enough evidence to reject the claim and conclude that at least one mean is different from the others.

13.2 The data in the following table represent the number of hours of relief provided by five different brands of headache tablets administered to 25 subjects experiencing fevers of 38°C or more. Perform the analysis of variance and test the hypothesis at the 0.05 level of significance that the mean number of hours of relief provided by the tablets is the same for all five brands. Discuss the results.

Tablet				
A	***B***	***C***	***D***	***E***
5.2	9.1	3.2	2.4	7.1
4.7	7.1	5.8	3.4	6.6
8.1	8.2	2.2	4.1	9.3
6.2	6.0	3.1	1.0	4.2
3.0	9.1	7.2	4.0	7.6

a. State the hypotheses and identify the claim.
b. Find the critical value.
c. Compute the test value.
d. Make the decision.
e. Summarize the results, and explain where the differences in the means are.

Solution:

The hypotheses are

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_5,$$
$$H_1 : \text{At least two of the means are not equal.}$$

$\alpha = 0.05$.
Critical region: $f > 2.87$ with $v_1 = 4$ and $v_2 = 20$ degrees of freedom.
Computation:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Computed f
Tablets	78.422	4	19.605	6.59
Error	59.532	20	2.977	
Total	137.954	24		

Decision: Reject H_0.

The mean number of hours of relief differ significantly.


ANY
Questions?