Estimate the missing value and carry out the analysis of variance to test the effect of the treatments.

SOLUTION

The design is RCDB with one missing observation. The additive model and the assumption are as given for RCBD. Lay out before estimating the missing value.

BLOCKS	A		B	C	D	E	Blocks total
	I	2.62	2.67	1.04	1.88	1.74	9.95
	II	2.58	2.52	M	2.82	2.10	10.02
	III	2.65	1.92	0.91	2.65	2.78	10.91
Treatments totals		7.85	7.11	1.95	7.35	6.62	30.88
Treatments means		2.62	2.37	0.65	2.45	2.21	

To estimate the missing value:

$$M = \frac{rT + bB - S}{(r-1)(b-1)}$$

$$= \frac{5(1.95) + 3(10.02) - 30.88}{(5-1)(3-1)}$$

$$= 1.1163 \text{ or } 1.12 \text{ to 2 d.p}$$

The correction for bias is:

$$Z = \frac{[B - (b-1)M]^2}{t(t-1)}$$

$$= \frac{[10.02 - (5-1)1.12]^2}{5(5-1)}$$

$$= 1.5346$$

The new lay-out (i.e. inserting the estimate of missing value is:

BLOCKS	A		B	C	D	E	Blocks total
	I	2.62	2.67	1.04	1.88	1.74	9.95
	II	2.58	2.52	M	2.82	2.10	10.02
	III	2.65	1.92	0.91	2.65	2.78	10.91
Treatments totals		7.85	7.11	1.95	7.35	6.62	30.88
Treatments means		2.62	2.37	1.02	2.45	2.21	

The hypotheses:

H_O: The treatments effects are not significantly difference

H_A: The treatments effects are not significantly difference

Computations:

$$C.F = \frac{(32.00)^2}{15}$$

$$= 68.2667$$

$$SS_{total} = (2.62)^2 + \ldots\ldots\ldots\ldots + (2.78)^2 - C.F - t$$

$$= 74.5484 - 68.2667 - 1.5346$$

$$= 4.747$$

$$SS_{blocks} = \frac{(9.95)^2 + (11.14)^2 + (10.91)^2}{t=(5)} \quad C.F \quad Z$$

$$= 68.4260 - 68.2667$$

$$= 0.1593$$

$$SS_{trts} = \frac{(7.85)^2 + \ldots\ldots\ldots\ldots + (6.62)^2}{b=3} \quad C.F \quad Z$$

$= 73.1488 - 68.2667 - 1.5346$

$= 3.3475$

$SS_{error} = SS_{total} - SS_{block} - SS_{tts}$

$= 4.7471 - 0.1593 - 3.3475$

$= 1.2403$

Other computations are done in the ANOVA table

ANOVA TABLE

SV	df	SS	MS	F
Blocks	2	0.1593	0.0796	
Treatments	4	3.3475	0.8369	5.3994
Error	7	1.2403	0.1550	
Total	13	4.7471		

Note: The total and df error have been reduced by 1

The F-tabulated for 4 and 7 df at 0.05 level of significance is 4.12. this value is less than the F-calculated, hence, the H_o is not accepted. It is therefore concluded that the effects of the treatments are significantly different.

MISSING VALUE IN LATIN SQUARE DESIGN

$M = \frac{M(R + C + T) - 2S}{(m-1)(m-2)}$

Where M = Estimate of the missing value

m = Number of rows or columns or treatments in the design (e.g. in a 4 x 4 Latin square, m = 4)

R = Sum of observations in the same row as the missing value

C = Sum of observation in the same column as the MV

T = Sum of observation with the same treatment as the MV

S = Sum of all the actual observations

To correct for bias

$$Z = \frac{[S - R - C - (m-1)T]^2 - 2S}{(m-1)^2(m-2)^2}$$

e.g., The following hypothetical data represent the yield (in tones per ha) of four cultivars of Khaya arranged in a 4 x 4 Latin square with one missing value. The four cultivars are A(KH4), B(KH2O), C(KH29) and D(KS15).

COLUMNS

		1	2	3	4
ROWS	I	6.2 (C)	5.1 (D)	3.3 (B)	5.7 (A)
	II	5.1 (B)	3.8 (A)	M (C)	3.9 (D)
	III	5.4 (D)	4.8 (C)	6.9 (A)	4.6 (B)
	IV	3.2 (A)	2.9 (B)	5.2 (D)	5.4 (C)

Where M is the missing value. Estimate this MV and perform the analysis of variance to test for significant differences among the cultivars.

SOLUTION

The design is the Latin square design with one missing observation. The additive model and the assumption are given for Latin square design.

The lay-out (before estimating the missing value)

COLUMNS

		1	2	3	4	Row Total
ROWS	I	6.2	5.1	3.3	5.7	20.3
	II	5.1	3.8	M	3.9	12.8
	III	5.4	4.8	6.9	4.6	21.7
	IV	3.2	2.9	5.2	5.4	16.7
Column Totals		19.9	16.6	15.4	19.6	71.5

TREATMENTS

	A	B	C	D
	5.7	3.3	6.2	5.1
	3.8	5.1	M	3.9
	6.9	4.6	4.8	5.4
	3.2	2.9	5.4	5.2
Treatment Totals	19.6	15.9	16.4	19.6
Treatment Means	4.90	3.98	5.47	4.90

To estimate the missing value:

$$M = \frac{M(R + C + T) - 2S}{(m-1)(m-2)}$$

$$= \frac{4(12.0 + 15.4 + 16.4) - 2(71.5)}{(4-1)(4-2)}$$

$$= 5.9$$

The correction for bias is:

$$Z = \frac{[S - R - C - (m-1)T]^2}{(m-2)^2(m-2)^2}$$

$$= \frac{[71.5 - 12.8 - 15.4 - (4-1)16.4]^2}{(m-2)^2(m-2)^2}$$

$$= 0.9669$$

The new la-out (after estimating the MV) is:

COLUMN

		1	2	3	4	Row Total
ROWS	I	6.2	5.1	3.3	5.7	20.3
	II	5.1	3.8	M	3.9	12.8
	III	5.4	4.8	6.9	4.6	21.7
	IV	3.2	2.9	5.2	5.4	16.7
Column Totals		19.9	16.6	15.4	19.6	71.5

TREATMENTS

	A	B	C	D
	5.7	3.3	6.2	5.1
	3.8	5.1	M	3.9

	6.9	4.6	4.8	5.4
	3.2	2.9	5.4	5.2
Treatment Totals	19.6	15.9	16.4	19.6
Treatment Means	4.90	3.98	5.47	4.90

The hypotheses:

H_O: There is no significant difference in the yield of the cultivars

H_A: There is no significant difference in the yield of the cultivars

Computations:

$$C.F = \frac{(77.12)^2}{16}$$

$$= 374.4225$$

$$SS_{total} = (6.2)^2 + \ldots\ldots\ldots\ldots + (5.4)^2 - C.F - Z$$

$$= 394.12 - 374.4225 - 0.9669$$

$$= 18.7306$$

$$SS_{rows} = \frac{(20.3)^2 + \ldots\ldots\ldots\ldots + (16.7)^2}{r-1} \quad C.F$$

$$= 3.4675$$

$$SS_{coulmns} = \frac{(19.9)^2 + \ldots\ldots\ldots\ldots + (19.6)^2}{r-1} \quad C.F$$

$$= 2.9325$$

$$SS_{trts} = \frac{(19.6)^2 + \ldots\ldots\ldots\ldots + (19.6)^2}{4} \quad C.F \quad Z$$

$$= 379.605 - 374.4225 - 0.9669$$

$= \quad 4.2156$

$SS_{error} \quad = \quad SS_{total} - SS_{rows} - SS_{coulmns} - SS_{tits}$

$= \quad 4.7471 - 0.1593 - 3.3475$

$= \quad 1.2403$

Other computations are done in the ANOVA table

ANOVA TABLE

SV	df	SS	MS	F
Rows	3	3.4675	1.1558	
Columns	3	2.9325	0.9775	
Treatments	3	4.2156	1.4052	0.8658
Error	7	1.2403	0.1550	
Total	14	18.7306		

Note: The total and df error have been reduced by 1

The F-tabulated for 3 and 5 df at 0.05 level of significance is 5.41. This value is greater than the F-calculated, hence, the H_o is not rejected. The conclusion therefore is that the yields of the four cultivars of khaya are not significantly different from one another.

Regression and correlation

When 2 quantity vary together cannot being functionally related. We may wish to measure degree of association which exist between them e.g. looking at relationship between height (h) and diameter (bh).

Nu inference are made about how they are connected, but are may nearly measure the way in which they move together

X (dbh)	y(Ht)
2	4
3	5
4	9
5	15

Three (3) types of trends are usually common in regression i.e positive, negative, and trend (o). if there is positive association it can be deduce that most if the data will be found in the 1st and 3rd quadrant while a negative association will have its data in the 2nd and 4th quadrant but a number association will always distribute its data in all the 4 quadrant.

HE ^{y}	dbh ^{x}
5	10
6	8
8	15
12	12
9	6
8	9
10	12
13	5
6	10
7	10
72	97

Statistically correlation r = $\frac{spxy}{\sqrt{ssx \cdot ssy}}$

Spxy = $\Sigma xy - \frac{\Sigma x \Sigma y}{n}$

Ssx = $\Sigma x^2 - \frac{[\Sigma x]}{n}$

Ssy = $\Sigma y^2 - \frac{[\Sigma y]}{n}$

Y	X	Xy	Y^2	X^2
5	10	50	25	100
6	8	48	36	64
8	15	120	64	225
12	12	144	144	144
7	6	42	49	36
8	9	72	64	81
10	12	120	100	144
13	5	65	169	25
6	10	60	36	100
7	10	70	49	100
Σ = 72	97	741		

r = $\frac{\quad}{\quad \times \quad}$

= 0.63

r usually ranges from 0-1 it can neither be positive or negative or neutral i.e 0

r^2 = coefficient of variation. Determination

for the data

n = 10

$\Sigma y = 72$

$\bar{Y} = 7.2$

$\Sigma x = 97$

X = 9.7

$\Sigma x^2 = 1019$

$\Sigma y^2 = 576$

$\Sigma xy = 741$

$r = \frac{spxy}{\sqrt{ssx \cdot ssy}}$

$spxy = \Sigma xy - \frac{\Sigma y \Sigma x}{n} = 741 - \frac{72 \times 97}{10}$

= 42.6

$Ssx = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$

$= 1019 - \frac{97}{10}$

= 78.1

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

$576 - \frac{72}{10}$

= 57.6

Value of r cannot be said to be significant or not until the value are checked up in the stat table also the level of threshold should be known. i.e. r(0.1)10 0.1 is the 1% probability level and 10 stands for the two of observation.

Threshold determine time significant level of the result.

The probability level can either be 99% (1%) or 95% (5%).

= r (0.1)10

r (0.5)10

99% means the result will continue to be correct 99 times out of 100.

The 10 outside the bracket is the number of observation

r (0.1) 10 = .794

r (0.5) 10 = .648

This is effect means the result is not significant at 95% probability level.

Inference

If calculate is higher than Tab value then the statistic is significant and the reverse i.e if calculated vale is lower than tabulated value than the statistic is not significant

If null hypothesis says there is correlation between Ht and dbh and the result is not significant than the hull hypothesis rejected but if the statistic is significant then we say the null hypothesis is not rejected.

Coefficient of determination r^2

$R^2 \rightarrow$ this measure the proportion of variation of eh is associated Σ variation in X. if ranges from 0-1 and it must always be positive.

r = 0.63

$r^2 = (0.63)^2$

= 0.3969

= 0.397

What this one means is that 39% of variation in Y is attributable to variation in X

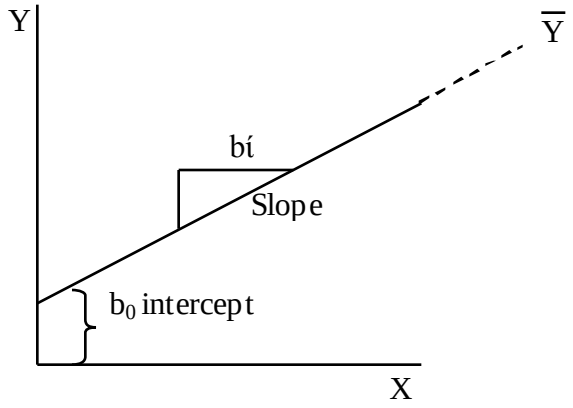
REGRESSION

If there are 2 variable X and Y and the live of relationship is needed then regression is always calculated. If there are only 2 variables then are talk of simple regression also there is always a dependent variable of an independent variable.

In case of Ht and dbh, dbh is easier to get and the Ht is the one we are interred in is the dependent variable while the dbh from which are want to evaluate Ht is the independent variable Hence Ht = Y and dbh = X.

The known variable = X, the unknown is the Y

Single linear regression = $y = b_o \pm bíx$



Intercept is always present in a line that does not urgency from O, if a line start from O then the intercept is O or non exixting.

$Y = B_o + BíX + Eí$

Eí = Error term

Y = estimated value of y for projection.

$bí = \frac{SPXY}{SSX}$

$b_o = \check{y} – bíx$

from the previous data

bí = $\frac{42.6}{78.1} - 0.545$

b_o = 7.2 – 0.545 x 9.7

= 7.2

= 1.91

The equation then become y = 1.91 + 0.545x

If X = 20 then Y = 1.91 + 0.545 X 20

Using computer to do this type of work an ANOVA table will be generated.

ANOVA

Source of variation	df	Ss	ms	F
Regression	1	23.22	23.22	5.4
Error	8	34.38	4.30	
Total	9	57.6		

Regress df = 1 because there are 2 variables

Ss regression = bí SPXY

Ss total = SSY

bíSPXY = 0.545 x 42.6

= 23.21

SSY = 57.6

Ms regression = $\frac{SSr}{df_r}$

MSE = $\frac{SSE}{df}$

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

The essence of going to calculate F is to know whether the reg. line is significant or not.

The threshold = Fα0.5 (1, 8) = 5.32

1 = df of regression

8 = df of error

Fα0.1 (1,8) = 11.3

Since E calculate (5.4) is > Tab 5.32 at 95% than the statistic is significant.

But at 99% Tab 11.3 is >5.4 and thus the result is not significant.

REGRESSION MODELS

If there is no significant or else then transformation into 1st order polynomial $Y = b_0 + bí$ will be done and if it is still not significant we move to he 2nd order polynomials = $Y = b_0 + bí \times b_2x^2$

3rd = $Y = b_0 + bíX + b_2X^2 + b_3X^3$

2nd order = quadrantic

3rd order = cubic

4th order = quatric

Log transformation

$Y = b_0 + bí \log X$ ------ single log

$\log Y = b_0 + bí \log X$ -------- double log

1st order

$Y = a + bíx$

2nd $Y = a + bíXí \pm b_2X^2$

3rd $Y = a + bíXí + b_2X^2 + b_3X^3$

At every stage ANOVA tale will be constructed and difficult F and r^2 will be found.

4th $Y = a + bíXí + b_2X^2 + b_3X^3 + b_4X^4$

IQ f age NL of a child

Y X Z

Model = $Y = a + bíX \pm b_2Z_2$ ---- multiple regression but not transformation.

Pr	Qw	L	Ec	Mc	U
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-

Predictor	coefficient	standard deviation	t-ratio	Probability
Constraint	-106	938	-1.14	0.299
QW	0.62	0.38	1.61	0.159
L	0.74	3.38	0.22	.835
EC	3.81	4.68	0.81	.447
Mc	-12.55	33.09	-0.38	.718
U	-43.92	51.56	-0.85	.427

$r^2 = 55.1$

ANOVA TABLE

Source	df	Ss	ms	F	P
Regression	5	3.208×10^{12}	7.617×10^{11}	1.77	0.323
Error	6	3.104×10^{12}	5.174×10^{11}		

Total 11 6.91×10^{12}

Initially there are 1 dependent and 5 independent variable = 6 and n = 12

The most important things for the model is the coefficient column

Hence the model for Pr

= Pr = -106 + 0.62Qw + 0.74L + 3.81EC – 12.55 mc – 43.92u.

The other important thing is the P each of taken from 1 will give the level significant

= 1 – p

= 1 – 0.299 = 0.71

= 1 – 0.159 = 0.841

= 1 – 0.835 = 0.165

= 1 – 0.447 = 0.553

= 1- 0.718 = 0.212

= 1- 0.427 = 0.573

Conversing them to percentage are will then get the percentage significant level it will give them help to know whether anyone is significant at 99% or 95%.

i.e 0.71 = 71%

0.841 = 841%

0.165 = 16.5%

0.553 = 55.3%

0.212 = 21.2%

0.573 = 57.3%.

Thus none of this is significant at 99/95% level.

If r^2 is converted so r it become

$r^2 = 55\% = 0.55$

r = 0.742

Then check whether the regression is significant in the stat table

How to check multiple regression stat table check n = 12 under k = 5 (independent variable).

i.e $r_{0.5}$ (5, 12) = 0.886 at r 0.1 (5, 12) = 0.938. our r is not significant or either 95% and 99% level since r calculate = 0.742 while r tab = 0.8 and 0.9. if the tried since a model need to be developed.

P should be ranked in order of importance based on their figure. Neglecting the constraint then, step wise analysis should be done; this will be eliminating 5 variables

STEP WISE REGRESSION

Variable	1	2	3	4
Const	-106	-53	-33	-29
Qw	.65	0.62	0.67	0.72
L	1.9	1.5	1.9	-
Ec	0.9	1.0	-	-
Mc	-7	-	$r^2 = 48.61$	$r^2 = 43.06$
U	$r^2 = 49.66$	$r^2 = 49.32$	-	-

1st step equation

Pr = -106 + 0.65QW + 1.9L + 0.9EC – 7mc

2nd

Pr = -55 + 0.62Qw + 1.5L + 1.0Ec

3rd

$Pr = -33 + 0.67\ Qw + 1.9L$

4th

$Pr = -29 + 0.72\ Qw$

Back to Anova table

$P = 0.323$

$1 - p = 0.677 = 67.7\%$

i.e it is not significant of either 95/99%

Considering

From table $F0.05\ (5, 6) = 4.95$

df error = 6 t stat df error

Trqmt = 5 $F0.1\ (5, 6) = 10.7$

F tab at 95% shows 4.95 while focal shows 1.47 i.e it is not significant at 95% not to talk of 99%

For transformation

r^2 for semilog = 52.1%

r^2 for Normal = 55%

r^2 for Double log = 33.1%.

semilog means conveying dependent variables as it is.

2nd order polynomials

$Pr = a + Qw + Qw^2 + L + L^2 + Ec + Ec^2 + Mc + Mc^2 + U + U^2$