DECISION ANALYSIS for BIM
LECTURE NOTES

ANOVA
REGRESSION & CORRELATION

ANALYSIS of VARIANCE (ANOVA)

We have seen how to test if mean difference between two populations (two sample mean test). ANOVA is used to test difference of group means for "multiple samples" (more then 2 populations)

The necessary calculations to prepare an ANOVA table is shown below:

(1) Correction term = $CT = \frac{T^2}{n}$ where $T = \sum\sum x_{ij}$

$n = n_1 + n_2 + \dots + n_k$

(2) Sum of Squares Total = $SST = \sum\sum x_{ij}^2 - CT$

(3) Sum of Squares Treatment = $SSTR = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - CT$ where $T_i = \sum_{j=1}^{n_i} x_{ij}$

$i = 1, 2, \dots, k$

(6) Sum of Squares Error = $SSE = SST - SSTR$

We use F statistics to decide if "At least one mean is different"

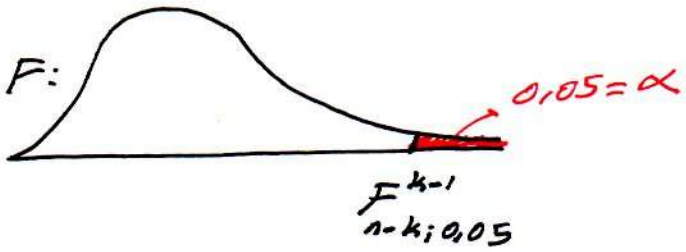
(i) $H_0: \mu_1 = \mu_2 = \dots = \mu_k$

H_A: At least one μ_j is different, $i = 1, 2, \dots, k$

$\alpha = 0.05$

(ii) $F = \dfrac{MSTR}{MSE}$ → $dfTR = k-1$, → $dfE = n-k$

(iii)



Reject H_0 if $F > F^{k-1}_{n-k;0,05}$

(iv) ANOVA TABLE

Source of Variation	Degrees of Freedom df	Sum of Squares SS	Mean Square MS	F
Treatments	$k-1$	$SSTR$	$MSTR = \dfrac{SSTR}{k-1}$	$F = \dfrac{MSTR}{MSE}$
Error	$(n-1)-(k-1) = n-k$	$SST - SSTR = SSE$	$MSE = \dfrac{SSE}{n-k}$	—
TOTAL	$n-1$	SST	—	—

(v) Decide and Conclude

"Do NOT Reject H_0. Means are NOT significantly different at $\alpha = 0,05$"

OR

"Reject H_0. Means are significantly different at $\alpha = 0,05$"

14.17 To study its performance, a newly designed motorboat was timed over a marked course under various wind and water conditions. Use the following data (in minutes) to test, at the 0.05 level of significance, whether the differences among the three sample means are significant:

Calm conditions:	26	19	16	22		
Moderate conditions:	25	27	25	20	18	23
Choppy conditions:	23	25	28	31	26	

14.17)

I Calm	II Moderate	III Choppy
26	25	23
19	27	25
16	25	28
22	20	31
	18	26
	23	
$T_1 = 83$	$T_2 = 138$	$T_3 = 133$

$n_1 = 4$, $n_2 = 6$, $n_3 = 5$

$T = 83 + 138 + 133 = 354$

$n = 4 + 6 + 5 = 15$

$\sum\sum x_{ij}^2 = 26^2 + 19^2 + \dots + 31^2 + 26^2 = 8584$

$\sum_{i=1}^{3} \frac{T_i^2}{n_i} = \frac{83^2}{4} + \frac{138^2}{6} + \frac{133^2}{5} = 8434,05$

(1) $CT = \frac{354^2}{15} = 8354,4$

(2) $SST = 8584 - 8354,4 = 229,6$

(3) $SSTR = 8434,05 - 8354,4 = 79,65$

(6) $SSE = 229,6 - 79,65 = 169,95$

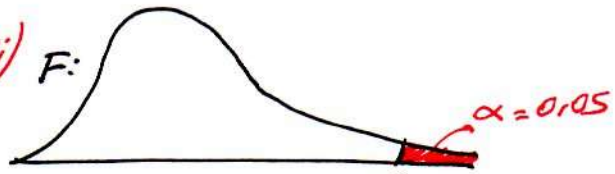
(i) $H_0 : \mu_I = \mu_{II} = \mu_{III}$

H_A: At least one mean is different

$\alpha = 0,05$

(ii) $F = \frac{MSTR}{MSE}$ → $df = 3 - 1 = 2$ (numerator); → $df = 15 - 3 = 12$ (denominator)

(iii) F:



$F^2_{12;0,05} = 3,89$

Reject H_0 if $F > 3,89$

(iv) ANOVA Table

Source	df	SS	MS	F
Treatment	2	79,65	$\frac{79,65}{2} = 39,8$	$\frac{39,8}{12,5} = 3,186$
Error	$16 - 2 = 12$	$229,6 - 79,65 = 149,95$	$\frac{149,95}{12} = 12,5$	—
Total	$n - 1 = 14$	229,6	—	—

(v) Do NOT Reject H_0.

Example. Significance of 3 treatment effects is tested with sample size 15. Complete the anova table below. Is treatment significant?

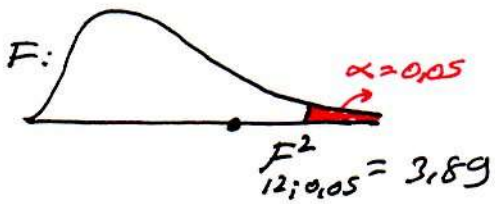
Source	df	SS	MS	F
Treatment		79,65		
Error				
TOTAL		229,6		

Answer:

$k=3$

$n=15$

Source	df	SS	MS	F
Treatment	② $3-1=2$	79,65	⑤ $\frac{79,65}{2}=39,8$	⑦ $\frac{39,8}{12,5}=3,186$
Error	④ $14-2=12$	$229,6-79,65$ ① $=149,95$	⑥ $\frac{149,95}{2}=12,5$	–
TOTAL	③ $15-1=14$	229,6	–	–



Do NOT Reject H_0. Treatment effect is NOT significant.

Block Design - 2 way ANOVA

To reduce variation of Error and obtain more reliable results, we design our experiment by using blocks. In 2-way ANOVA (block design) we have a cross-table. Namely, there's a row variable and column variable, in which effect of both are tested.

The additional calculation here is, we have one more column in ANOVA table whose source is "Block" Given we have b blocks, df for Blocks is $b-1$ and Sum of Squares Block is;

$$(4)\quad SSB = \sum_{j=1}^{b} \frac{T_j^2}{k} - CT$$

ANOVA table:

Source of Variation	Degrees of Freedom df	Sum of Squares SS	Mean Square MS	F
Treatments	$k-1$	$SSTR$	$MSTR = \frac{SSTR}{k-1}$	$F_{TR} = \frac{MSTR}{MSE}$
Blocks	$b-1$	SSB	$MSB = \frac{SSB}{b-1}$	$F_B = \frac{MSB}{MSE}$
Error	$(n-1)-(k-1)-(b-1)$ $=(k-1).(b-1)$	$SST-SSTR-SSB$ $=SSE$	$MSE = \frac{SSE}{(k-1)(b-1)}$	—
TOTAL	$n-1$	SST	—	—

Then, we have,

$$(6)\quad SSE = SST - SSTR - SSB$$

Our Hypothesis are;

(i) $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$

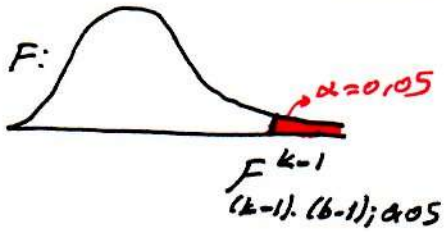
H_A: At least one μ_i is different $i=1,\dots,k$ (Treatment Effect is significant)

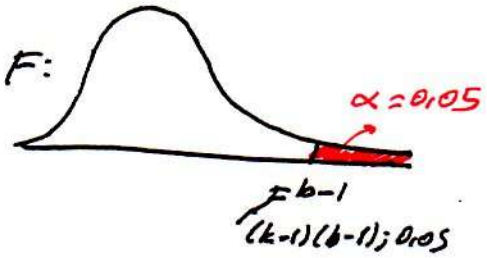
$\alpha = 0.05$

$H_0': \mu_1 = \mu_2 = \cdots = \mu_b$

H_A': At least one μ_j is different $j=1,\dots,b$ (Block effect is significant)

$\alpha = 0.05$

(iii) 



Reject H_0 if $F_{TR} > F^{k-1}_{(k-1)(b-1);0,05}$

Reject H_0' if $F_B > F^{b-1}_{(k-1)(b-1);0,05}$

14.30 The following are the numbers of defectives produced by four workmen operating, in turn, three different machines:

Machine \ Workman	B_1	B_2	B_3	B_4	Total
A_1	35	38	41	32	146
A_2	31	40	38	31	140
A_3	36	35	43	25	139
Total	102	113	122	88	425=T

Perform a two-way analysis of variance, using the 0.05 level of significance for both tests.

$k = 3$

$b = 4$

$n = 3 \cdot 4 = 12$

(1) $CT = \frac{425^2}{12} = 15052,08$

$\sum\sum x_{ij}^2 = 35^2 + 31^2 + \cdots + 25^2 = 15335$

(2) $SST = 15335 - 15052,08 = 282,92$

$\sum \frac{T_i^2}{b} = \frac{146^2}{4} + \frac{140^2}{4} + \frac{139^2}{4} = 15059,25$

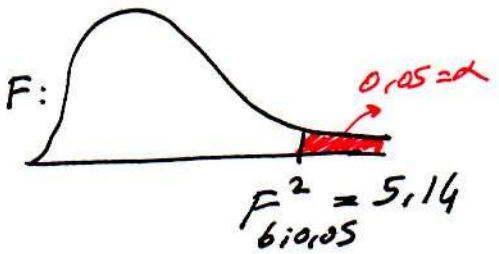
(3) $SSTR = 15059,25 - 15052,08 = 7,17$

$\sum \frac{T_j^2}{k} = \frac{102^2}{3} + \frac{113^2}{3} + \frac{122^2}{3} + \frac{88^2}{3} = 15267$

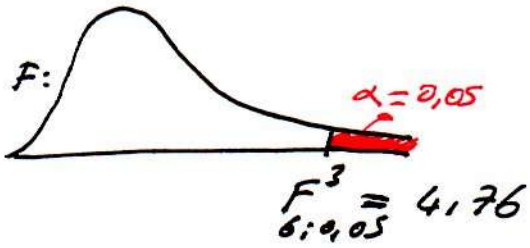
(4) $SSB = 15267 - 15052,08 = 214,92$

(6) $SSE = 282,92 - 7,17 - 214,92 = 60,83$

Source	df	SS	MS	F
Treatments	$3-1=2$	7,17	$\frac{7,17}{2}=3,59$	$\frac{3,59}{10,14}<1$
Blocks	$4-1=3$	214,92	$\frac{214,92}{3}=71,64$	$\frac{71,64}{10,14}=7,065$
Error	$11-2-3=6$	$282,92-7,17-214,92=60,83$	$\frac{60,83}{6}=10,14$	—
TOTAL	$12-1=11$	282,92	—	—



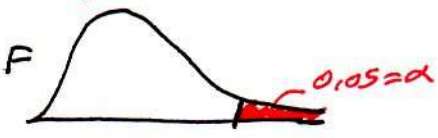
Treatment (Machine) Effect is NOT significant



Block (Workmas) Effect is significant

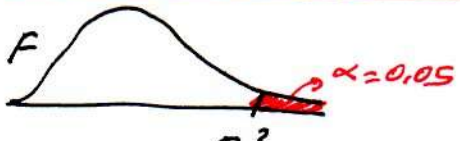
Example. Complete the ANOVA table below. Which effects are significant?

Source	df	SS	MS	F
Treatments	2	② $282,92-214,92-60,83=7,17$	③ $\frac{7,17}{2}=3,59$	⑥ $\frac{3,59}{10,14}<1$
Blocks	① $11-6-2=3$	214,92	④ $\frac{214,92}{3}=71,64$	⑦ $\frac{71,64}{10,14}=7,065$
Error	6	60,83	⑤ $\frac{60,83}{6}=10,14$	
TOTAL	11	282,92		



$F^2_{6;0,05}=5,14$

Treatment Effect is NOT significant



$F^3_{6;0,05}=4,76$

Block Effect is Significant

Latin Square Design

In Latin Square design, Number of Rows and Columns are Equal. The treatment effect is in the square such that NO rows or columns have a level more then once (like SUDOKU)

* 14.34 Making use of the fact that each of the letters must occur once and only once in each row and each column, complete the following Latin squares:

(a)

		A
	B	

(b)

	A		
			B
A	C		
		C	

(c)

	A	E		
		B		E
C			A	
D				
				D

14.34)

a)

B(1)	C(2)	A
C(4)	A(3)	B(7)
A(5)	B	C(6)

b)

B(8)	A	D(9)	C(7)
C(11)	D(2)	A(10)	B
A	C	B(3)	D(4)
D(6)	B(1)	C	A(5)

c)

B(7)	A	E	D(6)	C(5)
A(9)	D(10)	B	C(11)	E
C	E(2)	D(1)	A	B(3)
D	B(16)	C(14)	E(12)	A(4)
E(8)	C(15)	A(14)	B(13)	D

Note that, since this is a square, we have $k = b$ (also $= t$) and df's are the same for both rows, columns ant treatments. The additional calculation here is;

(5) $$SS_{TR} = \sum_{A,B,\dots} \frac{T_t^2}{n_t} - CT$$

and

(6) $$SSE = SST - SS_{TR} - SS_{Rows} - SS_{Columns}$$

14.35 The sample data in the following 3 × 3 Latin square are the scores in an American history test obtained by nine college students of various ethnic backgrounds and of various professional interests, who were taught by instructors *A*, *B*, and *C*:

	Ethnic background: Mexican	German	Polish
Law	A 75	B 86	C 69
Medicine	B 95	C 79	A 86
Engineering	C 70	A 83	B 93

Analyze this Latin square, using the 0.05 level of

14.35)

	Mex.	German	Polish	TOTAL
Law	A 75	B 86	C 69	230
Medicine	B 95	C 79	A 86	260
Eng.	C 70	A 83	B 93	246
TOTAL	240	248	248	736

$n = 3.3 = 9 \quad \Sigma X_A = 244 \quad \Sigma X_B = 274 \quad \Sigma X_C = 218$

(1) $CT = \frac{736^2}{9} = 60188,4$

$\Sigma\Sigma X_{ij}^2 = 75^2 + 95^2 + \dots + 93^2$
$= 60882$

(2) $SST = 60882 - 60188,4 = 693,6$

(3) $SS_{Rows} = \frac{1}{3}(230^2 + 260^2 + 266^2) - CT$
$= 150,3$

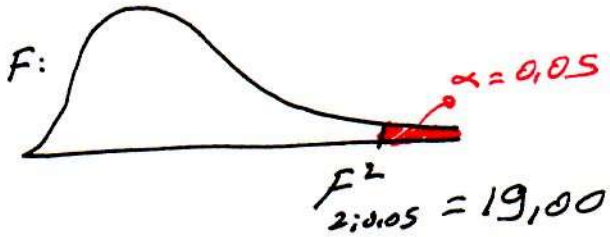
(4) $SS_{Columns} = \frac{1}{3}(260^2 + 268^2 + 268^2) - CT$
$= 16,3$

(5) $SS_{TR} = \frac{1}{3}(246^2 + 276^2 + 218^2) - CT$
$= 523,6$

(6) $SSE = 693,6 - 150,3 - 16,3 - 523,6$
$= 5,4$

ANOVA Table

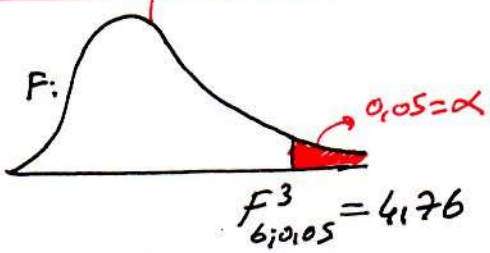
Source	df	SS	MS	F
Rows	$3-1=2$	150,3	$\frac{150,3}{2} = 75,15$	$\frac{75,15}{2,7} = 27,83$
Columns	$3-1=2$	16,3	$\frac{16,3}{2} = 7,15$	$\frac{7,15}{2,7} = 2,648$
Treatment	$3-1=2$	523,6	$\frac{523,6}{2} = 261,8$	$\frac{261,8}{2,7} = 96,96$
Error	$8-(2+2+2) = 2$ (by chance)	$693,6-150,3-16,3-523,6 = 5,4$	$\frac{5,4}{2} = 2,7$	—
TOTAL	$n-1=9-1=8$	693,6	—	—



Rows (department) effect is significant
Columns (Ethnic Background) effect is NOT significant
Treatment (Instructor) effect is significant.

Example. Calculate the missing values in the following ANOVA table. Which effects are significant?

Source	df	SS	MS	F
Rows	3	(9) $5,85 \cdot 3 = 17,55$	(8) $6,5 \cdot 0,9 = 5,85$	0,9
Columns	(2) 3	116,75	(10) $\frac{116,75}{3} = 38,25$	(12) $\frac{38,25}{6,5} = 5,88$
Treatment	(3) 3	(7) $3 \cdot 58,25 = 174,75$	58,25	(13) $\frac{58,25}{6,5} = 8,96$
Error	(5) $15 - 9 = 6$	(6) $6 \cdot 6,5 = 39$	6,5	—
TOTAL	(4) $4 \cdot 4 - 1 = 15$	(11) $17,55 + 116,75 + 174,75 + 39 = 346,85$	—	—



Rows effect is NOT significant.
Columns effect is significant.
Treatment effect is significant.

REGRESSION & CORRELATION

Simple Regression

X: Explanatory Variable
Y: Dependent Variable

We want to estimate (or explain) Y using X via the model $Y = a + bX$. This is called "Simple Linear Regression"

To find a and b, we follow these steps;

(i) $\Sigma x_i \quad \Sigma y_i \quad \Sigma x_i^2 \quad \Sigma x_i y_i$ (also Σy_i^2 for correlation)

(ii) $SS_{xx} = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} = \Sigma (x_i - \bar{x})^2$

$SS_{xy} = \Sigma x_i y_i - \frac{\Sigma x_i \cdot \Sigma y_i}{n} = \Sigma (x_i - \bar{x}) \cdot (y_i - \bar{y})$

also; $SS_{yy} = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n} = \Sigma (y_i - \bar{y})^2$ for correlation

(iii) $b = \frac{SS_{xy}}{SS_{xx}}$; $a = \frac{\Sigma y_i}{n} - b \cdot \frac{\Sigma x_i}{n} = \bar{y} - b \cdot \bar{x}$

15.10 The following show the improvement (gain in reading speed) of eight students in a speed-reading program, and the number of weeks they have been in the program:

Number of weeks x	*Speed gain (words per minute)* y
3	86
5	118
2	49
8	193
6	164
9	232
3	73
4	109

(a) Plot the eight data points to verify that it is reasonable to assume that the relationship between average speed gain and time is linear.
(b) Find the equation of the least-squares line which will enable us to predict speed gain from the number of weeks that a student has been in the program.
(c) Use the results of part (b) to predict the speed gain of a student after he or she has been in the program for seven weeks.

15.10)

	x	y	x*x	x*y
	3	86	9	258
	5	118	25	590
	2	49	4	98
	8	193	64	1544
	6	164	36	984
	9	232	81	2088
	3	73	9	219
	4	109	16	436
TOTAL	40	1024	244	6217

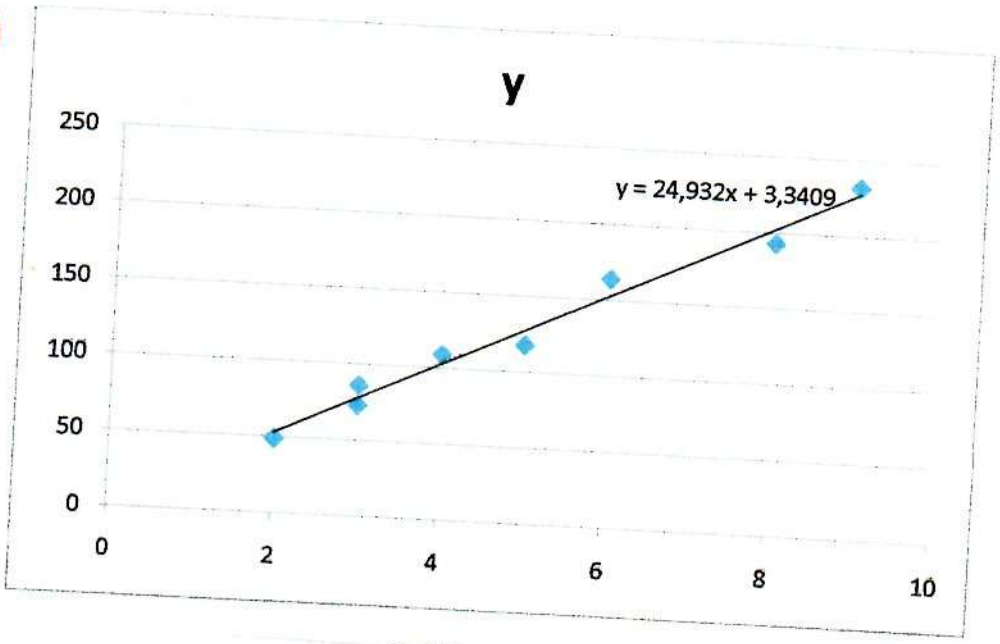
b) (i) $\Sigma x_i = 40$; $\Sigma y_i = 1024$;

$\Sigma x_i^2 = 246$; $\Sigma y_i x_i = 6217$; $n = 8$

(ii) $SS_{xx} = 244 - \frac{40^2}{8} = 44$

$SS_{xy} = 6217 - \frac{40 \cdot 1024}{8} = 1098$

a)



(iii) $b = \frac{1098}{44} = 24,932$

$a = \frac{1024}{8} - 24,932 \cdot \frac{40}{8}$

$= 3,34$

Then;

$y = 3,34 + 24,932 \cdot X$

c) $(\hat{y}/X = 7) = 3,34 + 24,932 \cdot 7 = 178$ words/min.

d) Estimate the speed gain of a student after 20 weeks.

Answer We (mathematically) have,

$(\hat{y}/X = 20) = 3,34 + 24,932 \cdot 20 = 502$

However, this result is NOT reliable because $X = 20$ is far away from the range of X $(2;9)$. We use regression model to estimate y for values of X that are "at least" close to range or in the range.

Multiple Regression

If we have more than one explanatory variables, the regression model is called, "Multiple Regression". For example, one may construct a model that explains

y: Cum. GPA using X_1: Weekly studying hours

X_2 = Cum. GPA

X_3: ÖSS - Score.

Then, we have,

$$y = a + bX_1 + cX_2 + dX_3$$

a, b, c, d are (for example) found using "Normal Equations". We'll only learn how to construct Normal Equations, we won't solve them and find a, b, c, d (parameter estimates)

To find Normal equations, (we have 4 unknowns and so 4 equations), put Σ and variable (if any) near the unknown estimator. Namely, put Σ to both sides of the equation for a; put ΣX_1 for b; ΣX_2 for c; and ΣX_3 for d. Then;

$$(1)\ \Sigma y = \Sigma a + \Sigma bX_1 + \Sigma c X_2 + \Sigma d X_3$$

$$(2)\ \Sigma X_1 y = \Sigma X_1 a + \Sigma X_1 b X_1 + \Sigma X_1 c X_2 + \Sigma X_1 d X_3$$

$$(3)\ \Sigma X_2 y = \Sigma X_2 a + \Sigma X_2 b X_1 + \Sigma X_2 c X_2 + \Sigma X_2 d X_3$$

$$(4)\ \Sigma X_3 y = \Sigma X_3 a + \Sigma X_3 b X_1 + \Sigma X_3 c X_2 + \Sigma X_3 d X_3$$

Write $\Sigma a = n \cdot a$ and make multiplications, write a, b, c, d before Σ;

$$(1)\ \Sigma y = na + b\Sigma X_1 + c\Sigma X_2 + d\Sigma X_3$$

$$(2)\ \Sigma X_1 y = a\Sigma X_1 + b\Sigma X_1^2 + c\Sigma X_1 X_2 + d\Sigma X_1 X_3$$

$$(3)\ \Sigma X_2 y = a\Sigma X_2 + b\Sigma X_1 X_2 + c\Sigma X_2^2 + d\Sigma X_2 X_3$$

$$(4)\ \Sigma X_3 y = a\Sigma X_3 + b\Sigma X_1 X_3 + c\Sigma X_2 X_3 + d\Sigma X_3^2$$

} Normal Equations

For example, let we are given

$y_i = 1,3 + 0,0762 X_1 + 0,00417 \cdot X_2 + 0,00068 \cdot X_3$

Then, the estimated Cum. GPA of a student who studies 17 hours a week, IQ level is 130 and ÖSS-Score is 321 is found by;

$(\hat{y}_i \mid X_1 = 17,\ X_2 = 130,\ X_3 = 321)$

$= 1,3 + 0,0762 \cdot 17 + 0,00417 \cdot 130 + 0,00068 \cdot 321 = 3,32$

15.39) The following are sample data provided by a moving company on the weights of six shipments, the distances they were moved, and the damage that was incurred:

Weight (thousands of pounds) x_1	Distance (thousands of miles) x_2	Damage (dollars) y
4.0	1.5	160
3.0	2.2	112
1.6	1.0	69
1.2	2.0	90
3.4	0.8	123
4.8	1.6	186

The necessary sums for calculations are;

$\Sigma y_i = 760 \quad \Sigma X_{1i} = 18 \quad \Sigma X_{2i} = 9,1$

$\Sigma X_{1i} y_i = 2505,4 \quad \Sigma X_{1i}^2 = 63,6 \quad \Sigma X_{1i} X_{2i} = 27$

$\Sigma X_{2i} y_i = 1131,4 \quad \Sigma X_{1i} X_{2i}^2 = 44,38$

$\Sigma X_{2i}^2 y_i = 1885,96 \quad \Sigma X_{2i}^2 = 15,29 \quad \Sigma X_{2i}^3 = 27,63$

$n = 6 \quad ; \quad \Sigma X_{2i}^4 = 52,45$

a) Form the normal equations for a multiple regression model of the form $y = a + bX_1 + cX_2 + dX_2^2$

b) Given the following equation, estimate (or predict) the damage when a shipment weighting 2400 pounds is moved 1500 miles

$y_i = 20,1 + 17,03 X_1 + 16,02 X_2 - 0,98 X_2^2$

Answer

b) $(\hat{y} \mid X_1 = 2,4,\ X_2 = 1,5)$

$= 20,1 + 17,03 \cdot 2,4 + 16,02 \cdot 1,5 - 0,98 \cdot 1,5^2 = 82,8\ \$$

a) $y = a + bX_1 + cX_2 + dX_2^2$

$\Sigma \quad \Sigma X_1 \quad \Sigma X_2 \quad \Sigma X_2^2$

(1) $a \rightarrow \Sigma y = \Sigma a + \Sigma b X_1 + \Sigma c X_2 + \Sigma d X_2^2$

(2) $b \rightarrow \Sigma X_1 y = \Sigma X_1 a + \Sigma X_1 b X_1 + \Sigma X_1 c X_2 + \Sigma X_1 d X_2^2$

(3) $c \rightarrow \Sigma X_2 y = \Sigma X_2 a + \Sigma X_2 b X_1 + \Sigma X_2 c X_2 + \Sigma X_2 d X_2^2$

(4) $d \rightarrow \Sigma X_2^2 y = \Sigma X_2^2 a + \Sigma X_2^2 b X_1 + \Sigma X_2^2 c X_2 + \Sigma X_2^2 d X_2$

$$\Sigma y = na + b\Sigma X_1 + c\Sigma X_2 + d\Sigma X_2^2$$

$$\Sigma X_1 y = a\Sigma X_1 + b\Sigma X_1^2 + c\Sigma X_1 X_2 + d\Sigma X_2 X_2^2$$

$$\Sigma X_2 y = a\Sigma X_2 + b\Sigma X_1 X_2 + c\Sigma X_2^2 + d\Sigma X_2^3$$

$$\Sigma X_2^2 y = a\Sigma X_2^2 + b\Sigma X_1 X_2^2 + c\Sigma X_2^3 + d\Sigma X_2^4$$

$$740 = 6a + 18b + 9{,}1c + 15{,}29d$$

$$2505{,}4 = 18a + 63{,}6b + 27c + 44{,}38d$$

$$1131{,}4 = 9{,}1a + 27b + 15{,}29c + 27{,}63d$$

$$1885{,}96 = 15{,}29a + 44{,}38b + 27{,}63c + 52{,}45d$$

* To apply regression model, the model should be "linear" in terms of unknowns. If the model is "multiplicative", we take the log of both sides and the model becomes "linear" by properties of log.

$$\text{Log } a \cdot b = \log a + \log b \quad ; \quad \log a^x = x \cdot \log a$$

Then, if for example the model is; $y = A \cdot X^b \cdot y^c$,

we have; $\log y = \log (A X^b y^c)$

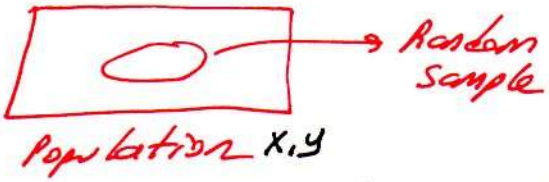
$$\log y = \underbrace{\log A}_{a} + \log X^b + \log y^c$$

$$\log y = a + b \log X + c \log y \rightarrow \text{linear Model.}$$

Correlation

The correlation coefficient shows the direction and strength of the "linear Relationship" between two variables, let's say X and Y.

Remember, population parameters are unknown constants which are estimated by sample statistics (known variables)



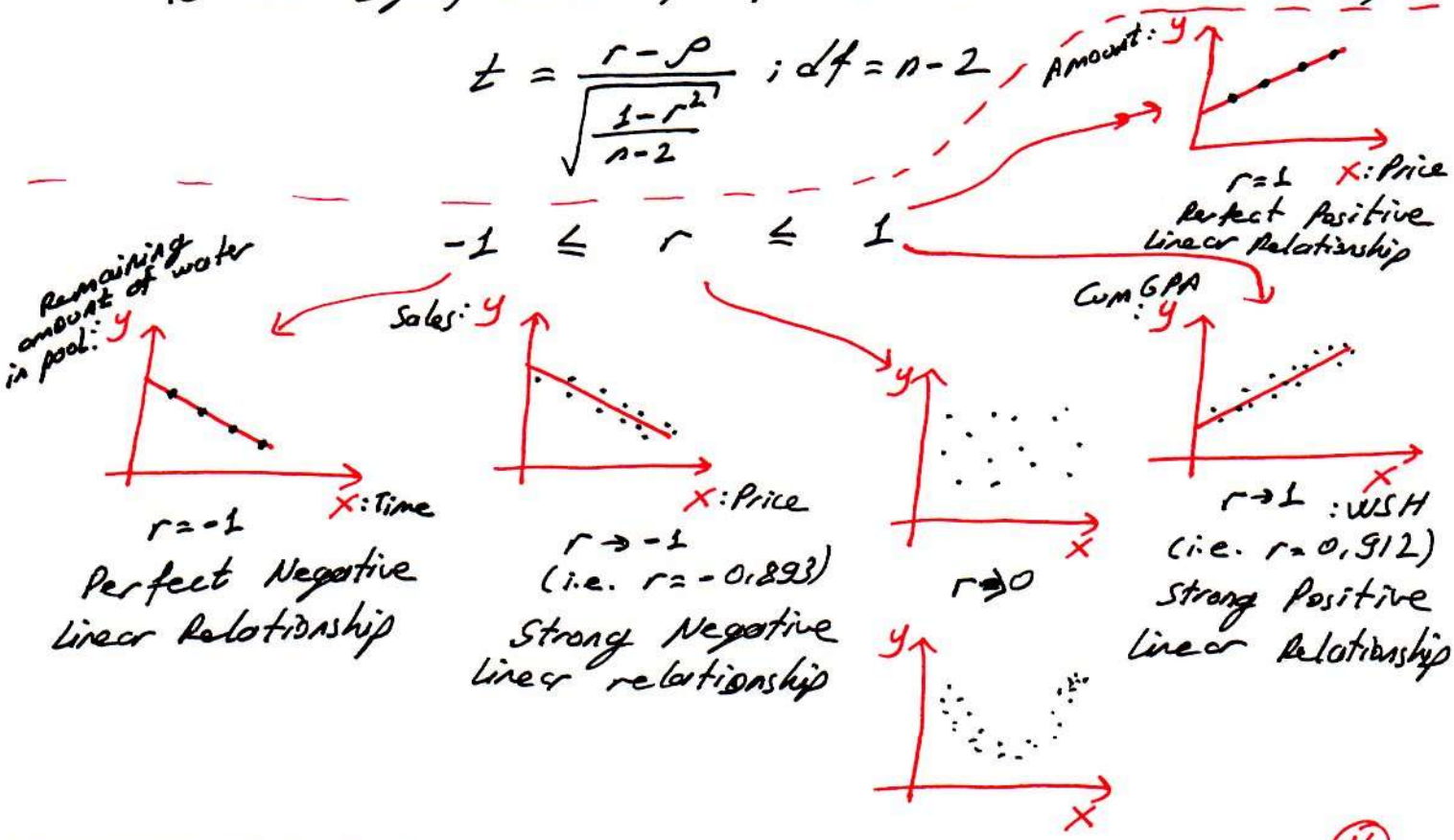
Population Correlation Coefficient: ρ

Sample Correlation Coefficient: $r = \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}}$

To test significance of r, the test statistics is;

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

$$-1 \leq r \leq 1$$



16.11 State in each case whether you would expect a positive correlation, a negative correlation, or no correlation:
(a) the ages of husbands and wives;
(b) the amount of rubber on tires and the number of miles they have been driven;
(c) the number of hours that golfers practice and their scores;
(d) shoe size and IQ;
(e) the weight of the load of trucks and their gasoline consumption.

16.11) a) Positive correlation
b) Negative correlation
c) Positive correlation
d) No correlation
e) Positive correlation

16.5 The following table shows the percentages of the vote predicted by a poll for eight candidates for the U.S. Senate in different states, x, and the percentages of the vote which they actually received, y:

x	y
43	50
46	42
51	57
59	55
41	46
53	48
52	53
62	56

Calculate r. Is it significant?

16.5)

	x	y	x^2	y^2	$x.y$
(i)	43	50	43^2	50^2	43.50
$n=8$	46	42	46^2	42^2	46.52
	!	!	!	!	!
	62	56	62^2	56^2	62.56
+	$\Sigma x_i=407$	$\Sigma y_i=410$	$\Sigma x_i^2=21085$	$\Sigma y_i^2=20903$	$\Sigma x_i y_i=20892$

(ii) $SS_{xx} = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} = 21085 - \frac{407^2}{8} = 378,9 = \Sigma(x_i-\bar{x})^2$

$SS_{yy} = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n} = 20903 - \frac{410^2}{8} = 196,9 = \Sigma(y_i-\bar{y})^2$

$SS_{xy} = \Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n} = 20892 - \frac{407.410}{8} = 185,9$

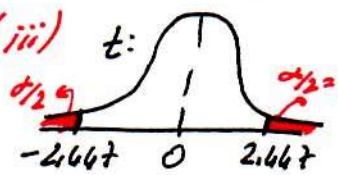
(iii) $r = \frac{SS_{xy}}{\sqrt{SS_{xx}.SS_{yy}}} = \frac{185,9}{\sqrt{378,9.196,9}} = 0,68$

Is it significant? : Is ρ significantly different from 0.

HYPOTHESIS TESTING

(i) $H_0: \rho = 0$
$H_A: \rho \neq 0$
$\alpha = 0,05$

(ii) $t = \frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}}$ idf = 6

(iii) t:


Reject H_0 if $|t| > 2,447$

(iv) $t = \frac{0,68-0}{\sqrt{\frac{1-0,68^2}{6}}} = 2,275$

Do NOT reject H_0. Correlation is NOT significant at $\alpha=0,05$

Partial Correlation.

Let, X_1 : weekly hot chocolade sold

X_2 : weekly number of visitors

If on the basis of suitable data, we get $r = -0{,}30$ for these variables, this should come as a surprise — after all, we would expect more sales of hot chocolade when there are more visitors.

However; the fact is that, these two variables are bot related to a third variable,

X_3 : Average weekly temperature

Let us suppose the data yield $r = -0{,}70$ for X_1 and X_3, $r = 0{,}80$ for X_2 and X_3. Then, to see the "true" linear relationship between X_1 and X_2, we must hold X_3 constant. This is the "partial correlation coefficient" which is given by,

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{1 - r_{13}^2} \cdot \sqrt{1 - r_{23}^2}}$$

Substituting $r_{12} = -0{,}30$; $r_{13} = -0{,}70$; $r_{23} = 0{,}80$,

we get

$$r_{12.3} = \frac{(-0{,}30) - (-0{,}70 \cdot 0{,}80)}{\sqrt{1 - (-0{,}70)^2} \cdot \sqrt{1 - 0{,}80^2}} = 0{,}61$$

When the effect of differences in temperature is eliminated, there is a positive relationship between the sales of hot chocolade and number of visitors.

Correlation Matrix

A variable has correlation 1 with itself.

Summarizing many variables in terms of correlation, matrix notation is useful. An example follows.

Example. Given the following correlation matrix, find and interpret the partial correlation between C_2 and C_7 keeping C_9 constant.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
C1	1.000									
C2	0.274	1.000								
C3	-0.134	-0.269	1.000							
C4	0.201	-0.153	0.075	1.000						
C5	-0.129	-0.166	0.278	-0.011	1.000					
C6	-0.095	0.280	-0.348	-0.378	-0.009	1.000				
C7	0.171	-0.122	0.288	0.086	0.193	0.002	1.000			
C8	0.219	0.242	-0.380	-0.227	-0.551	0.324	-0.082	1.000		
C9	0.518	0.238	0.002	0.082	-0.015	0.304	0.347	-0.013	1.000	
C10	0.299	0.568	0.165	-0.122	-0.106	-0.169	0.243	0.014	0.352	1.000

Answer. $r_{27} = -0,122$; $r_{29} = 0,238$; $r_{79} = 0,347$

$$r_{27.9} = \frac{r_{27} - r_{29} \cdot r_{79}}{\sqrt{(1 - r_{29}^2)(1 - r_{79}^2)}} = \frac{-0,122 - 0,238 \cdot 0,347}{\sqrt{(1 - 0,238^2) \cdot (1 - 0,347^2)}} = -0,225$$

The true relationship between C_2 and C_7 becomes slightly more powerful (in negative way) when we hold C_9 constant. (Remove the effect of C_9)