4.2 Linear Regression

Linear Regression – A technique in which a straight line is fitted to a set of points.

Interpolation – When you estimate within the range of data values.

Extrapolation – When you estimate beyond the range of data values

Least Squares Method – Determines the residuals (vertical deviation from the line of best fit)

- residuals are positive for points above the line, and negative for points below the line
- the sum of the squares of the residuals has the least possible value
- the sum of the residuals is zero

[We can calculate the equation of the line of best fit using the following formula

We can calculate the equation of the line of best fit using the following formulae:

$y = ax + b$	
$a = \dfrac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$	$b = \dfrac{\sum y}{n} - a\left(\dfrac{\sum x}{n}\right)$

Example 1: Researchers monitoring the number of wolves and rabbits in a wildlife reserve think that the wolf population depends on the rabbit population since wolves prey on rabbits. Over a period of 8 years, researchers have collected the following data.

Year	1994	1995	1996	1997	1998	1999	2000	2001
Rabbit Population	61	72	78	76	65	54	39	43
Wolf Population	26	33	42	49	37	30	24	19

a) Determine the line of best fit and the correlation coefficient for this data.

Rabbit (x)	Wolf (y)	x^2	y^2	xy
61	26	3721	676	1586
72	33	5184	1089	2376
78	42	6084	1764	3276
76	49	5776	2401	3724
65	37	4223	1369	2405
54	30	2916	900	1620
39	24	1521	576	936
43	19	1849	361	817
$\Sigma x = 488$	$\Sigma y = 260$	$\Sigma x^2 = 31,276$	$\Sigma y^2 = 9136$	$\Sigma xy = 16,740$

$$r = \frac{(8)(16740) - (488)(260)}{\sqrt{[(8)(31,276) - (488)^2][(8)(9136) - (260)^2]}}$$

$$= \frac{7040}{\sqrt{(12064)(5488)}}$$

$$\doteq 0.87$$

(same as r)

$$a = \frac{7040}{12064}$$

$$= 0.58$$

$$b = \frac{260}{8} - 0.58\left(\frac{488}{8}\right) = -2.88$$

$$\therefore y = 0.58x - 2.88$$

as the rabbit popⁿ increases, ↑ (x) more rabbit means 0.58 more wolves

when there are no more rabbits there are "negative" amounts of wolves

b) Does this data support the researchers' theory?

Yes, the data supports the theory

Example 2: To evaluate the performance of one of its instructors, a driving school tabulates the number of hours of instruction and the drive-test scores for the instructor's students.

a) Analyze this data to determine whether the instructor is an effective teacher or not.

Hours x	Score y	x^2	y^2	xy
10	78	100	6084	780
15	85	225	7225	1275
21	96	441	9216	2016
6	75	36	5625	450
18	84	324	7056	1512
20	45	400	2025	900
12	82	144	6724	984
$\sum x = 102$	$\sum y = 545$	$\sum x^2 = 1670$	$\sum y^2 = 43,955$	$\sum xy = 7917$

$n = 7$

$$r = \frac{7(7917) - (102)(545)}{\sqrt{[7(1670) - (102)^2][7(43955) - (545)^2]}}$$

$$= \frac{-171}{\sqrt{(1286)(10,660)}} \doteq -0.05$$

∴ This is a weak negative correlation

b) Comment on any data that appears to be unusual.

Check y-values for outliers:

45 75