Section 1.6 Worksheet – Linear Regression by Hand

MDM4U
Jensen

1) Sand driven by wind creates large dunes at the Great Sand Dunes National Monument in Colorado. Is there a linear relationship correlation between wind velocity and sand drift rate? A test site at the Great Sand Dunes National Monument gave the following information about x, wind velocity in cm/sec, and y, drift rate of sand in g/cm/sec.

a) Complete the chart

Wind Speed $[x]$	**Drift Rate** $[y]$	x^2	y^2	xy
70	3	4 900	9	210
115	45	13 225	2 025	5 175
105	21	11 025	441	2 205
82	7	6 724	49	574
93	16	8 649	256	1 488
125	62	15 625	3 844	7 750
88	12	7 744	144	1 056
$\sum x = 678$	$\sum y = 166$	$\sum x^2 = 67\,892$	$\sum y^2 = 6\,768$	$\sum xy = 18\,458$

b) Determine the equation of the least squares regression line ($\hat{y} = a + bx$). Interpret the slope and y-intercept in context.

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

Slope:

Slope = $b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} = \frac{7(18\,458) - (678)(166)}{7(67\,892) - (678)^2} = \frac{16\,658}{15\,560} = 1.07$

This indicates that for every 1 cm/sec increase in wind velocity, the model predicts a 1.07 g/cm/sec increase in drift rate of sand.

y-intercept:

$$\bar{x} = \frac{\sum x}{n} = \frac{678}{7} = 96.857$$

$$\bar{y} = \frac{\sum y}{n} = \frac{166}{7} = 23.714$$

$$y - intercept = a = \bar{y} - b\bar{x} = 23.714 - 1.07(96.857) = -79.92$$

y-intercept = $a = \bar{y} - b\bar{x}$

This tells us that at a wind speed of 0, the model predicts a sand drift rate of -79.92 g/cm/sec.

Linear Regression Equation:

$\hat{y} = a + bx$ → $predicted\ drift\ rate = -79.92 + 1.07(wind\ velocity)$

c) Compute the correlation coefficient using the formula. Interpret r and r^2 in context.

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

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} = \frac{7(18\,458) - (678)(166)}{\sqrt{[7(67\,892) - (678)^2][7(6\,768) - (166)^2]}} = \frac{16\,658}{17\,561.29836} = 0.94856$$

$r = 0.94856$; This indicates that there is a strong, positive, linear correlation between wind speed and drift rate.

$r^2 = 0.8998$; This tells us that about 89.98% of the variation in drift rate can be explained by the approximate linear correlation with wind speed.

2) A study was conducted to determine if larger universities tend to have more property crime. Let x represent student enrollment (in thousands) and let y represent the number of burglaries in a year on the campus. A random sample of 8 universities in California gave the following information:

a) Complete the chart

Student Enrollment [x]	**Burglaries [y]**	x^2	y^2	xy
12.5	26	156.25	676	325
30	73	900	5 329	2 190
24.5	39	600.25	1 521	955.5
14.3	23	204.49	529	328.9
7.5	15	56.25	225	112.5
27.7	30	767.29	900	831
16.2	15	262.44	225	243
20.1	25	404.01	625	502.5
$\sum x = 152.8$	$\sum y = 246$	$\sum x^2 = 3\,350.98$	$\sum y^2 = 10\,030$	$\sum xy = 5\,488.4$

b) Determine the equation of the least squares regression line ($\hat{y} = a + bx$) by hand. Interpret the slope and y-intercept in context.

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

Slope:

Slope = $b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} = \frac{8(5488.4) - (152.8)(246)}{8(3350.98) - (152.8)^2} = \frac{6318.4}{3460} = 1.826$

The slope tells us that for every 1000 more students enrolled, the model predicts 1.826 more burglaries a year.

y-intercept:

y-intercept = $a = \bar{y} - b\bar{x}$

$$\bar{x} = \frac{\sum x}{n} = \frac{152.8}{8} = 19.1$$

$$\bar{y} = \frac{\sum y}{n} = \frac{246}{8} = 30.75$$

$$y - intercept = a = \bar{y} - b\bar{x} = 30.75 - 1.826(19.1) = -4.1266$$

The y-intercept tells us that if 0 students were enrolled, the model predicts -4.1266 crimes a year.

Linear Regression Equation:

$\hat{y} = a + bx$ → $predicted\ burglaries = -4.1266 + 1.826(student\ enrollment)$

c) Compute the correlation coefficient using the formula. Interpret r and r^2 in context.

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

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} = \frac{8(5488.4) - (152.8)(246)}{\sqrt{[8(3350.98) - (152.8)^2][8(10030) - (246)^2]}} = \frac{6318.4}{8261.055623} = 0.7648$$

$r = 0.7648$; this tells us there is a moderate, positive, linear correlation between student enrollment and burglaries.

$r^2 = 0.5849$; this tells us that about 58.49% of the variation in burglaries can be explained by the approximate linear correlation with student enrollment.