Section 1.6 – Linear Regression by Hand

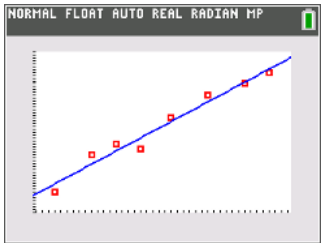
MDM4U
David Chen

Part 1: Linear Regression Using Technology Practice

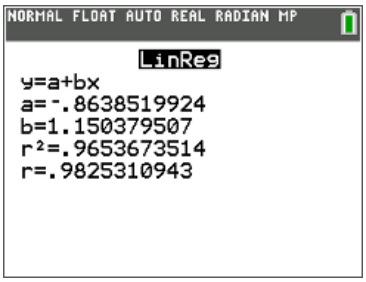
This table shows data for the full-time employees of a small company.

Age (years)	Annual Income ($000)
33	33
25	31
19	18
44	52
50	56
54	60
38	44
29	35

a) Generate a scatterplot of the data.



b) Perform a linear regression and state the equation of the line of best fit. Explain what the slope and y-intercept mean in context.



Equation: $predicted\ income = -0.86 + 1.15(age)$

y-intercept: when age is zero, the expected income is -0.86 thousand dollars

slope: for every one year increase in age, the model predicts an average increase of $1150.

c) What is the correlation coefficient? What does this tell you about the relationship between age and annual income?

$r = 0.98$; this indicates a strong, positive, linear correlation between age and income.

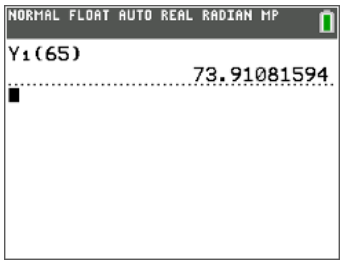
d) What is the coefficient of determination? What does it mean?

$r^2 = 0.965$; approximately 96.5% of the variation in income can be explained by the approximate linear correlation with age.

e) Use the line of best fit to predict the income for a 65 year old employee.

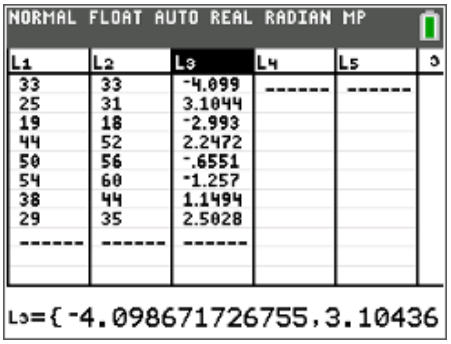
$predicted\ income = -0.86 + 1.15(65)$
$= 73.89$

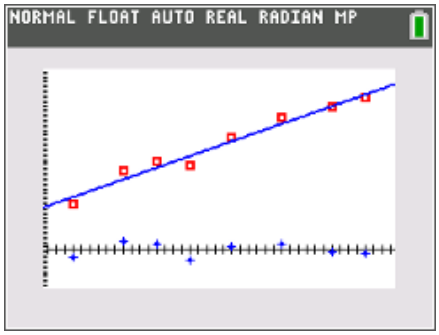
The model predicts an approximate income of $73 890 for a 65 year old employee.



f) Find the residual values. What do they tell you about the correlation between the two variables?

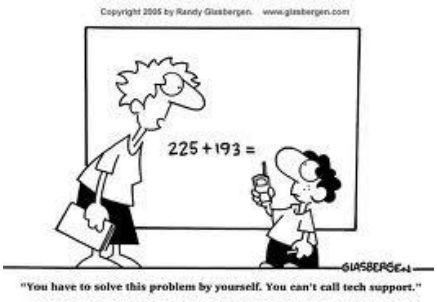
The residual values are relatively small and there is no distinguishable pattern on the residual plot. This indicates that the linear regression is an appropriate model for the relationship between age and income.





Part 2: Linear Regression by Hand

Example: The following table lists the mathematics of data management marks and grade 12 averages for a small group of students. Start by completing filling in the missing cells. You will need these values to calculate the correlation coefficient and equation of the line of best fit.



MDM4U Mark (x)	Grade 12 Average (y)	x^2	y^2	xy
74	77	5476	5929	5698
81	87	6561	7569	7047
66	68	4356	4624	4488
53	67	2809	4489	3551
92	85	8464	7225	7820
45	55	2025	3025	2475
80	76	6400	5776	6080
$\sum x = 491$	$\sum y = 515$	$\sum x^2 = 36091$	$\sum y^2 = 38637$	$\sum xy = 37159$

a) Determine the equation of the least squares regression line (line of best fit)

$$\text{Slope} = b = \frac{(n\sum xy(\,!\,\sum(x\ \sum y}{n(\sum x^2)\,!\,(\sum x)^2} = \frac{7(371\$\%)\,!\,(491)(515)}{7(36\#\$1)\,!\,(491)^2} = \frac{72\#\$}{1155\#} = 0.6272$$

This indicates that for every one percent increase in MDM4U average, the model predicts a 0.6272 percent increase in overall grade 12 average.

To calculate the y-intercept, we will need to find the average of the x values ($\bar{x}$) and y values ($\bar{y}$)

$$\bar{x} = \frac{\sum x}{n} = \frac{491}{7} = 70.14$$

$$\bar{y} = \frac{\sum y}{n} = \frac{515}{7} = 73.57$$

y-intercept = $a = \bar{y} - b\bar{x} = 73.57 - 0.6272(70.14) = 29.58$

When a students MDM4U mark is 0, we would expect a grade 12 average of approximately 29.58%.

The equation of the regression line is:

$$\hat{y} = a + bx \quad \rightarrow \quad predicted\ grade\ 12\ average = 29.58 + 0.6272(MDM4U\ grade)$$

b) Calculate the correlation coefficient by hand

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

$$= \frac{7(37159) - (491)(515)}{\sqrt{[7(36091) - (491)^2][7(38637) - (515)^2]}}$$

$$= \frac{7248}{7777.152692}$$

$$= 0.93196$$

This indicates that there is a strong positive linear correlation between Data grades and overall grade 12 average.

Approximately 86.86% of the variation in grade 12 average can be explained by the linear correlation with Data grades.