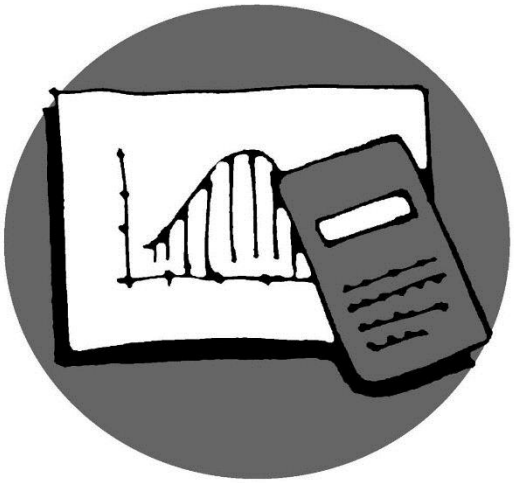
Simple

Linear Regression



Library, Teaching and Learning

2014

Lincoln University
Te Whare Wānaka o Aoraki
CHRISTCHURCH•NEW ZEALAND
New Zealand's specialist land-based university

Simple Linear Regression Analysis is the analysis of the linear relationship between two quantitative continuous variables.

Comment: $y = 3x + 5$ *is a* ***linear*** *relation, ie any one value of y depends on a given value for x and y "is related to x" in that each y value is calculated as "3 times the corresponding x-value, plus 5". The graph of this relation is a straight* ***line****.*

Two techniques are used to study the relationship between the two variables:

Regression Analysis	**Correlation Analysis**
- involves the prediction of values of a dependent variable Y, based on the values of an independent variable X	involves computing a coefficient which measures the strength of the linear relationship between two variables X and Y.
eg If Y is money spent and X is money earned, the value of Y depends on the value of X.	*eg If X is IQ and Y is verbal ability, a correlation coefficient would show the strength of relationship between the two.*

The process involves developing a linear model, which can be used to predict the values of Y (*dependent variable)* from the values of X *(independent variable)*.

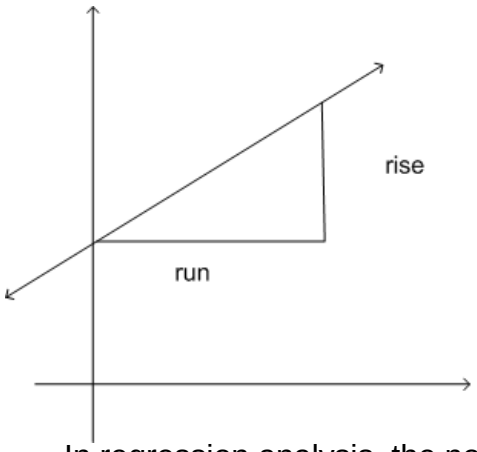
Parameters, symbols

x-values	the observed values of the independent variable
y-values	the observed values of the dependent variable which correspond to the respective x-values
Note: x and y values are always given as "ordered pairs" (x, y) and the order is important:- x-value first.	
Σx	the sum of all the x-values
Σy	the sum of all the y-values
Σx^2	the sum of all the x^2-values. That is, square all values and then add together.
Σy^2	the sum of all the y^2-values. That is, square all values and then add together.
Σxy	the sum of all the xy-values. That is, multiply each pair of x and y values and then add together.
b_1	the slope of the fitted line
b_0	the y-intercept
$\bar{x}$	the mean of all the x-values
$\bar{y}$	the mean of all the y-values
SS_{xy}	sums of squares for cross product xy
$SS_{xx}\,(SS_x), and\ SS_y$	sums of squares for $x\ and\ y$ respectively

Formulae

$y = b_0 + b_1 x$	The equation of the fitted line from the sample data
$b_1 = \dfrac{\left(\Sigma xy - \dfrac{\Sigma x \times \Sigma y}{n}\right)}{\left(\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}\right)} = \dfrac{SS_{xy}}{SS_x}$	The formula to calculate the slope of the line. This tells you the rate at which y is changing as x increases - that is, the gradient.
$b_0 = \bar{y} - b_1 \times \bar{x}$	The formula to calculate the y-intercept – that is, the value of y when x is zero.

We do this by developing a **Simple Linear Regression Equation**
From the scatter diagram, a line is drawn and an equation is developed.
Recall the algebraic equation for straight lines as follows:



$m = \dfrac{rise}{run}$

$y = mx + c$, where m is the gradient and c is the intercept on the y axis.

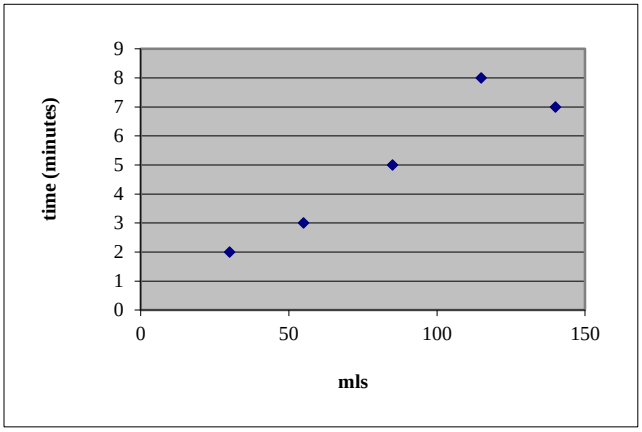
In regression analysis, the notation for a simple linear regression line is as follows:
$y = b_0 + b_1 x$, where b_1 is the gradient and b_0 is the intercept on the y axis.

The following is a simplistic example to show the process.
An industrial psychologist wishes to **predict** the *time* it takes to complete a task (in minutes) based on the *level of alcohol consumed* (in mls) during the previous hours.
The following data are obtained for a sample of $n = 5$ college students.

	Level of alcohol consumed (mls)	Time to complete task (minutes)
Subject	X	Y
Elana	55	3
Irene	30	2
Kathy	85	5
Micki	140	7
Roni	115	8

1. Plot the scattergram.
2. Fit a straight line to the data and plot the line on the scatter diagram.
3. Predict the time needed to complete the task if alcoholic consumption is:
   (a) 30 mls (b) 130 mls.

1. Plot the scatter diagram:



2 Fit the straight line:

- Calculate and complete the table:

	x	y	xy	x^2	y^2
	55	3	165	3025	9
	30	2	60	900	4
	85	5	425	7225	25
	140	7	980	19600	49
	115	8	920	13225	64
Totals	$425 = \Sigma x$	$25 = \Sigma y$	$2550 = \Sigma xy$	43975 = Σx^2	$151 = \Sigma y^2$

- Use the totals above appropriately in the formula to calculate the slope-

$$b_1 = \frac{\left(\Sigma xy - \frac{\Sigma x \times \Sigma y}{n}\right)}{\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right)} = \frac{\left(2550 - \frac{425 \times 25}{5}\right)}{\left(43975 - \frac{425^2}{5}\right)} = 0.054$$

That is the slope, gradient or rate of change = 0.054 which means for each additional ml of alcohol consumed, the time taken to complete the task increases by 0.054 minutes.

- Calculate the constant:

$b_0 = \bar{y} - b_1 \times \bar{x}$, where $\bar{y} = \frac{\Sigma y}{n} = \frac{25}{5} = 5$ and $\bar{x} = \frac{\Sigma x}{n} = \frac{426}{5} = 85$

Hence: $b_0 = 5 - 0.054 \times 85 = 0.398$. That is, the predicted time to complete the task when no alcohol is consumed is 0.4 minutes (ie when $X = 0$).

- Write the equation: $y = b_0 + b_1 x = 0.4 + 0.054x$

Plot the line

First plot any two points. We then use a ruler to draw the line (through the two points) from the *Y* axis to the largest value of *X*.

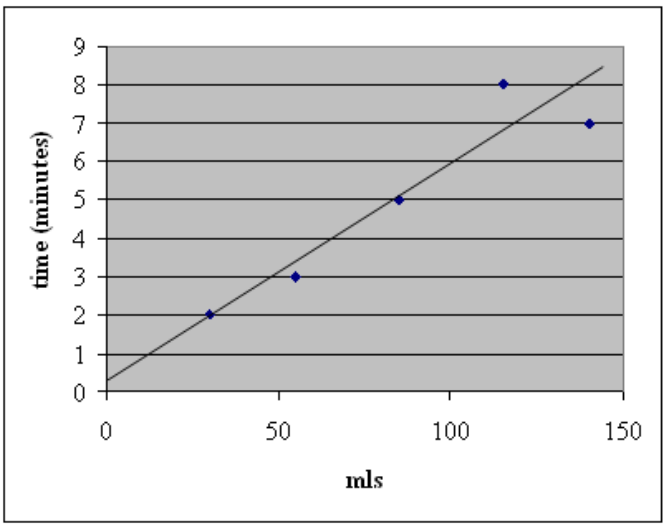
When $x = 0$, $\hat{Y}_i = 0.4 + 0.054 \times 0 = 0.4$

Plot the co-ordinates ($x = 0, y = 0.4$).

When $x = 142$, $\hat{Y}_i = 0.4 + 0.054 \times 140 = 7.96$ (Round to 8.)

Plot the co-ordinates ($x = 142, y = 8$).

Draw the line through these points:



3. Use the equation to predict

To predict y when $x = 30mls$: $\boldsymbol{y} = 0.4 + 0.054 \times 30 = 2.02$ ie 2 minutes.

To predict y when $x = 130mls$: $y = 0.4 + 0.054 \times 130 = 7.42$ ie 7 minutes.

Note

1 The population predictor equation is formally written as: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

2 It is wise to make predictions only *within* the range of the X values observed. As a rule, you do not **extrapolate** your values *outside* this range.

Practice Questions
Multi-choice Questions

4. The intercept (b_0) of a regression line represents the
   a) predicted average value of Y when $X = 0$.
   b) estimated average change in Y per unit change in X.
   c) predicted value of Y.
   d) variation around the line of regression.

5. "Score received on an exam (measured in percentage points)" (Y) is regressed on "percentage attendance" (X) for 22 students in a Statistics for Business and Economics course. If the Y intercept b_0 = 39.39 and the slope b_1 = 0.341, which of the following statement is correct?
   a) If attendance increases by 0.341%, the estimated average score received will increase by 1 percentage point.
   b) If attendance increases by 1%, the estimated average score received will increase by 39.39 percentage points.
   c) If attendance increases by 1%, the estimated average score received will increase by 0.341 percentage points.
   d) If the score received increases by 39.39%, the estimated average attendance will go up by 1%.

Use the following TABLE and the information given to answer the next 5 questions. Tim's Hardware store manager believes that there is a linear relationship between the income from sales of goods (Y, in thousands of dollars) and the amount spent on advertising (X, in thousands of dollars).

Advertising spending	Income from sales
5.3	21
3.8	16
3.1	13
2.9	12
4.4	23
4.9	20
5.1	23
5.4	24
3.2	14
5.1	19

This gives: $\Sigma x = 43.2$; $\Sigma x^2 = 195.34$; $\Sigma y = 185$; $\Sigma y^2 = 3601$; $\Sigma xy = 835$; $n = 10$.

$$SSx = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 195.34 - \frac{(43.2)^2}{10} = 8.716;$$

$$SSxy = \Sigma xy - \frac{\Sigma x \times \Sigma y}{n} = 835 - \frac{43.2 \times 185}{10} = 35.8$$

6. The slope of the regression is:

a 0.7561 b 4.1074 c 1.7014 d 0.1657

7. The regression equation is

a $\hat{Y} = 0.1657 + 4.1074x$ b $\hat{Y} = 4.1074 + 0.1657x$

c $\hat{Y} = 4.1074 + 0.7561x$ d $\hat{Y} = 0.7561 + 4.1074x$

8. Predicted income from the sales of goods when $4,400 is spent on advertising is

a $23.6452 (000's) b $16.9743 (000's)
c $18.8286 (000's) d $13.0864 (000's)

---

9. The slope (b_1) in a regression model represents the

a) predicted value of *Y* when *X* = 0.
b) estimated average change in *Y* per unit change in *X*.
c) predicted value of *Y*.
d) variation around the line of regression.

10. The least-squares method minimizes which of the following in regression?

a) SSR b) SSE c) SST d) All of the above.

Questions 11-15 are related to the following problem:
In a recent study researchers wanted to determine the strength of the relationship between the speed limit (X) and death rate (Y). A sample of 10 countries produced the following data:

Country	Speed limit (miles/hour) X	Death rate (per 100 million miles) Y
Norway	55	3.0
United States	55	3.3
Finland	55	3.4
Britain	70	3.5
Denmark	55	4.1
Canada	60	4.3
Japan	55	4.7
Australia	60	4.9
Netherlands	60	5.1
Italy	75	6.1

This gives: $\Sigma X = 600,\ \Sigma Y = 42.4,\ \Sigma X^2 = 36450,\ \Sigma Y^2 = 188.32,\ \Sigma XY = 2578$

11. Assuming a linear relationship, the slope b_1 of the regression model is:

a 0.076 b 0.81 c 1.16 d 1.47

12. The intercept b_0 of the regression model is

a 0.17 b -0.32 c 0.21 d 0.43

13 The death rate (per 100 million miles) for a country whose speed limit is 65 miles per hour is

a 5.0 b 5.5 c 6.1 d 4.62

More questions

Use the following information to answer the next 2 questions.

The director of cooperative education at a state college wants to examine the effect of cooperative education job experience on marketability in the work place. She takes a random sample of 4 students. For these 4, she finds out how many times each had a cooperative education job and how many job offers they received upon graduation. These data are presented in the table below.

Student	Co-Operative Jobs	Job Offer
1	1	4
2	2	6
3	1	3
4	0	1

1. What is the independent variable X?

a) Co Operative Jobs
b) Job Offers
c) Marketability in the workplace
d) None of the Above

2. Referring to the above data, the estimate of the slope is

a) 2 b) 2.50 c) 5 d) 0.4

Use the following information to answer the next 2 questions.

It is believed that the average number of hours spent studying per day (HOURS) during undergraduate education should have a positive linear relationship with the starting salary (SALARY, measured in thousands of dollars per month) after graduation. Given below is the Excel output from regressing starting salary on number of hours spent studying per day for a sample of 51 students.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	Lower 95%	Upper 95%
Intercept	-1.8940	0.4018	-4.7134	2.051E-05	-2.7015	-1.0865
Hours	0.9795	0.0733	13.3561	5.944E-18	0.8321	1.1269

3 Referring to the table, the estimated average change in salary (in thousands of dollars) as a result of spending an extra hour per day studying is

A) 0.9795 B) 0.7845 C) 335.0473 D) -1.8940

4. The 90% confidence interval for the average change in SALARY (in thousands of dollars) as a result of spending an extra hour per day studying is

A) narrower than [-2.7015, -1.0865] B) wider than [-2.7015, -1.0865]
C) narrower than [0.83219, 1.1269] D) wider than [0.83219, 1.1269].

Use the following information to answer the next 3 questions.

The owner of an intra-city moving company is interested in determining the number of labour hours needed for a move. He has collected data for 36 moves in which the volume of goods (X) in cubic metres and corresponding labour hours (Y) were recorded. This data was summarized and an Analysis of Variance table was produced. The results are shown below:

$\sum X = 6864$; $\sum Y = 1042.5$; $\sum XY = 240833$; $\sum X^2 = 1564761$; $\sum Y^2 = 37960.5$

5. Calculate the linear regression coefficients b_0 and b_1 for this data.

6. Interpret the meaning of b_1 in this problem.

7. Estimate the average labour hours needed to move 200 cubic metres.

8. During the 1950s, radioactive material leaked from a storage area near Hanford, Washington, into the Columbia River nearby. For nine counties downstream in Oregon, an index of exposure *X* was calculated (based on the distance from Hanford). Also the cancer mortality *Y* was calculated (deaths per 100,000 person-years, 1959 – 1964). This data is summarised as follows:

County	**Radioactive Exposure *X***	**Cancer Mortality *Y***
Clatsop	8.3	210
Columbia	6.4	180
Gilliam	3.4	130
Hood River	3.8	170
Morrow	2.6	130
Portland	11.6	210
Sherman	1.2	120
Umatilla	2.5	150
Wasco	1.6	140

$\Sigma X = 41.4$; $\Sigma X^2 = 287.42$; $\Sigma Y = 1440$; $\Sigma Y^2 = 239800$; $\Sigma XY = 7500$

a) Calculate the linear regression coefficients b_0 and b_1 for this data.

b) Estimate the cancer mortality (*Y*) associated with a radioactive exposure index value of 8.0. (2 marks)

Solutions

Multi-choice Questions

4	**a**	**5**	**c**	**6**	**b**	**7**	**d**	**8**	**c**
9	**b**	**10**	**b**	**11**	**a**	**12**	**b**	**13**	**d**

More Questions

1	**a**	**2**	**b**	**3**	**a**	**4**	**c**

5. $$b_1 = \frac{\left(\Sigma xy - \frac{\Sigma x \times \Sigma y}{n}\right)}{\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right)} = \frac{\left(240833 - \frac{6864 \times 1042.5}{36}\right)}{\left(1564761 - \frac{6864^2}{36}\right)} = 0.1643$$

$$b_0 = \bar{y} - b_1 \times \bar{x} = \frac{1042.5}{36} - 0.1643 \times \frac{6864}{36} = -2.37$$

6. For each cubic meter 0.164 labours hours are required

7. $x = 200m^3 \Rightarrow y = -2.37 + 0.1643 \times 200 = 30.49hrs$

8.

a) $$b_1 = \frac{\left(\Sigma xy - \frac{\Sigma x \times \Sigma y}{n}\right)}{\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right)} = \frac{\left(7500 - \frac{41.4 \times 1440}{9}\right)}{\left(287.42 - \frac{41.4^2}{9}\right)} = 9.033$$

$$b_0 = \bar{y} - b_1 \times \bar{x} = \frac{41.4}{9} - 9.033 \times \frac{1440}{9} = 118.45$$

b) $x = 8 \Rightarrow y = 118.45 + 9.033 \times 8 = 190.71$