UECS3213 / UECS3453 DATA MINING

SESSION: January 2019

TUTORIAL 4
Chapter 4 - Regression Analysis

1. The time *x* in years that an employee spent at a company and the employee's hourly pay, *y*, for 5 employees are listed in the table below.

x	y
5	25
3	20
4	21
10	35
15	38

   a) Calculate and interpret the correlation coefficient *r*. Include a plot of the data in your discussion.

Hint:

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

$$\mathrm{corr}(\mathrm{x}, \mathrm{y}) = \mathrm{cov}(\mathrm{x}, \mathrm{y})/\sigma_x\sigma_y$$

Answer:

Reference: https://www.kean.edu/~fosborne/bstat/09rc.html

x	y	x^2	y^2	xy
5	25	25	625	125
3	20	9	400	60
4	21	16	441	84
10	35	100	1225	350
15	38	225	1444	570
$\sum x = 37$	$\sum y = 139$	$\sum x^2 = 375$	$\sum y^2 = 4135$	$\sum xy = 1189$

Hint: Calculate the numerator:

$$n\sum(xy) - \left(\sum x\right)\left(\sum y\right) = 5 \cdot 1189 - 37 \cdot 139 = 802$$

Then calculate the denominator:

$$\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2} = \sqrt{5 \cdot 375 - (37)^2}\sqrt{5 \cdot 4135 - (139)^2}$$
$$= \sqrt{506}\sqrt{1354} \approx 827.72$$

Now, divide to get $r \approx \dfrac{802}{827.72} \approx 0.97$.

Interpret this result: There is a **strong positive correlation** between the number of years and employee has worked and the employee's salary, since *r* is very close to 1.

b) Find the equation of the *least square regression line* for the abovementioned relationship.

Answer:

$y = a + bx.$

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

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

First, find the slope m. Start by determining the numerator:

$$n\sum xy - \left(\sum x\right)\left(\sum y\right) 5 \cdot 1189 - 37 \cdot 139 = 802$$

Next, find the denominator:

$$n\sum(x^2) - \left(\sum x\right)^2 = 5 \cdot 375 - (37)^2 = 506$$

Divide to obtain $m = \dfrac{802}{506} \approx 1.58$

Now, find the y-intercept: $b = \dfrac{\sum y}{n} - m\dfrac{\sum x}{n} \approx \dfrac{139}{5} - 1.58 \cdot \dfrac{37}{5} \approx 16.11$

Therefore, the equation of the regression line is $\hat{y} = 1.58x + 16.11$

c) Use the equations in part (b) to predict the hourly pay rate of an employee who has worked for 20 years.

Answer:

For an employee who has worked 20 years, $x = 20$. Plug this into the equation for the regression line: $\hat{y} = 1.58 \cdot 20 + 16.11 = 47.71$ is the predicted salary, based on the data.

2. The table below shows the number of absences, x, in a Calculus course and the final exam grade, y, for 7 students.

   a) Find the correlation coefficient, r and interpret your result.

x	1	0	2	6	4	3	3
y	95	90	90	55	70	80	85

Answer:

$$\sum x = 19, \quad \sum y = 565, \quad \sum x^2 = 75, \quad \sum y^2 = 46,775, \quad \sum xy = 1,380.$$

Calculate the numerator:

$$n\sum(xy) - \left(\sum x\right)\left(\sum y\right) = 7 \cdot 1380 - 19 \cdot 565 = -1075$$

Then calculate the denominator:

$$\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2} = \sqrt{7 \cdot 75 - (19)^2}\sqrt{7 \cdot 46775 - (565)^2}$$
$$= \sqrt{164}\sqrt{8200} \approx 1159.66$$

Now, divide to get $r \approx \dfrac{-1075}{1159.66} \approx -0.93$.

Interpret this result: There is a strong negative correlation between the number of absences and the final exam grade, since r is very close to -1. Thus, as the number of absences increases, the final exam grade tends to decrease.

   b) Find the equation of the *least square regression line* for the abovementioned relationship.

Answer:

$y = a + bx.$

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

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

First, find the slope m. Start by determining the numerator:

$$n \sum xy - \left(\sum x\right)\left(\sum y\right) = 7 \cdot 1380 - 19 \cdot 565 = -1075$$

Next, find the denominator:

$$n \sum (x^2) - \left(\sum x\right)^2 = 7 \cdot 775 - (19)^2 = 164$$

Divide to obtain $m = \frac{-1075}{164} \approx -6.55$

Now, find the y-intercept: $b = \frac{\sum y}{n} - m\frac{\sum x}{n} \approx \frac{565}{7} - (-6.55) \cdot \frac{19}{7} = 98.49$

Therefore, the equation of the regression line is $\hat{y} = -6.55x + 98.49$

c) Use the equations to part (b) to predict the test score for a student with 5 absences.

For a student with 5 absences, $x = 5$. Plug this into the equation for the regression line: $\hat{y} = -6.55 \cdot 5 + 98.49 = 65.74$ is the predicted score, based on the data.

3. The table below shows the height, x, in inches and the pulse rate, y, per minute, for 9 people. Find the correlation coefficient, r and interpret your result.

x	68	72	65	70	62	75	78	64	68
y	90	85	88	100	105	98	70	65	72

Answer:

$$\sum x = 622, \quad \sum y = 773, \quad \sum x^2 = 43,206, \quad \sum y^2 = 68,007, \quad \sum xy = 53,336.$$

Calculate the numerator:

$$n\sum(xy) - \left(\sum x\right)\left(\sum y\right) = 9 \cdot 53336 - 622 \cdot 773 = -782$$

Then calculate the denominator:

$$\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2} = \sqrt{9 \cdot 43206 - (622)^2}\sqrt{9 \cdot 68007 - (773)^2}$$
$$= \sqrt{1970}\sqrt{14534} \approx 5350.89$$

Now, divide to get $r \approx \dfrac{-782}{5350.89} \approx -0.15$.

Interpret this result: There appears to be an extremely weak, if any, correlation between height and pulse rate, since r is close to 0.

4. Consider the following set of points: {(-2 , -1) , (1 , 1) , (3 , 2)}
   a) Find the *least square regression line* for the given data points.
   b) Plot the given points and the *regression line* in the same rectangular system of axes.

Answer:

Organize the data in a table.

x	y	x y	x^2
-2	-1	2	4
1	1	1	1
3	2	6	9
$\Sigma x = 2$	$\Sigma y = 2$	$\Sigma xy = 9$	$\Sigma x^2 = 14$

We now use the above formula to calculate a and b as follows

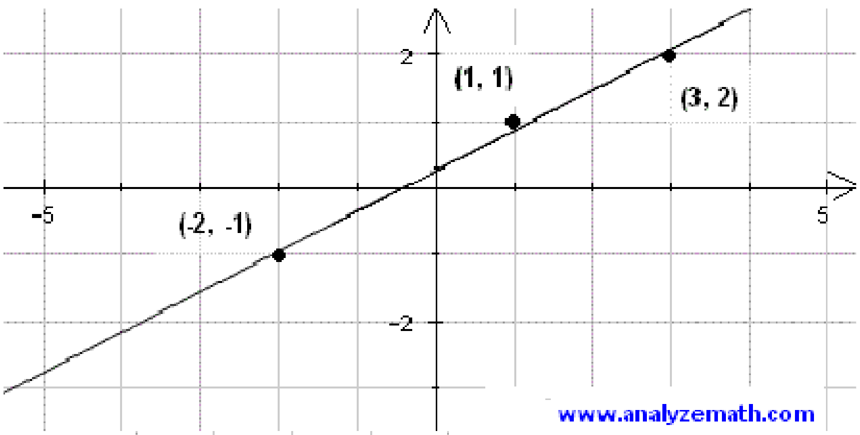
$a = (n\Sigma x\,y - \Sigma x\Sigma y) / (n\Sigma x^2 - (\Sigma x)^2) = (3*9 - 2*2) / (3*14 - 2^2) =$

$23/38$

$b = (1/n)(\Sigma y - a\,\Sigma x) = (1/3)(2 - (23/38)*2) = 5/19$

b) We now graph the regression line given by y = a x + b and

the given points.



Graph of linear regression

5. Given the following data: {(-1 , 0),(0 , 2),(1 , 4),(2 , 5)}
   a) Find the *least square regression line* for the following set of data
   b) Plot the given points and the regression line in the same rectangular system of axes.

Answer:

x	y	x y	x^2
-1	0	0	1
0	2	0	0
1	4	4	1
2	5	10	4
$\Sigma x = 2$	$\Sigma y = 11$	$\Sigma x\,y = 14$	$\Sigma x^2 = 6$

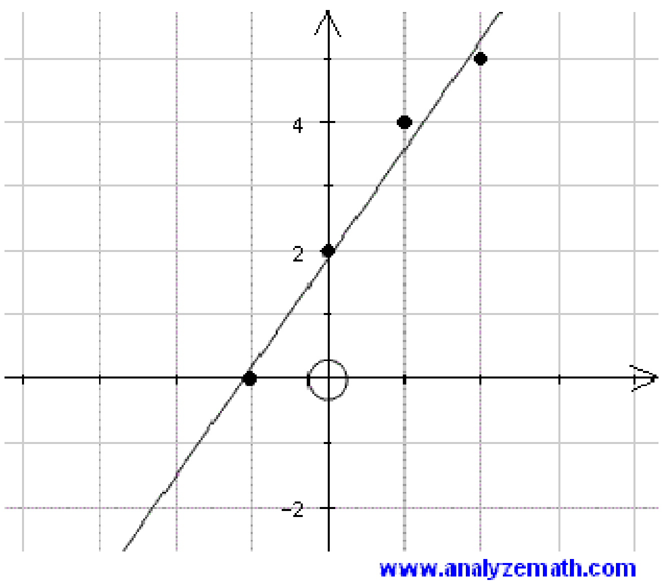
We now use the above formula to calculate a and b as follows

$a = (n\Sigma x\, y - \Sigma x\Sigma y) / (n\Sigma x^2 - (\Sigma x)^2) = (4*14 - 2*11) / (4*6 - 2^2) =$

$17/10 = 1.7$

$b = (1/n)(\Sigma y - a\, \Sigma x) = (1/4)(11 - 1.7*2) = 1.9$

b) We now graph the regression line given by y = ax + b and the given points.



Graph of linear regression.

6. The values of y and their corresponding values of y are shown in the table below

x	0	1	2	3	4
y	2	3	5	4	6

a) Find the *least square regression line,* y = a x + b.
b) Estimate the value of y when x = 10.

Answer

a) We use a table to calculate a and b.

x	y	x y	x^2
0	2	0	0
1	3	3	1
2	5	10	4
3	4	12	9
4	6	24	16
Σx = 10	Σy = 20	Σx y = 49	$\Sigma x^2 = 30$

We now calculate a and b using the least square regression

formulas for a and b.

$a = (n\Sigma x\,y - \Sigma x\Sigma y) / (n\Sigma x^2 - (\Sigma x)^2) = (5*49 - 10*20) / (5*30 - 10^2)$

$= 0.9$

$b = (1/n)(\Sigma y - a\,\Sigma x) = (1/5)(20 - 0.9*10) = 2.2$

b) Now that we have the least square regression line y = 0.9 x + 2.2, substitute x by 10 to find the value of the corresponding y.

$y = 0.9 * 10 + 2.2 = 11.2$

7. The sales of a company (in million dollars) for each year are shown in the table below.

x (year)	2005	2006	2007	2008	2009
y (sales)	12	19	29	37	45

a) Find the *least square regression line* y = a x + b.
b) Use the least squares regression line as a model to estimate the sales of the company in 2012.

Answer

a) We first change the variable x into t such that t = x - 2005 and therefore t represents the number of years after 2005. Using t instead of x makes the numbers smaller and therefore manageable. The table of values becomes.

t (years after 2005)	0	1	2	3	4
y (sales)	12	19	29	37	45

We now use the table to calculate a and b included in the least regression line formula.

t	y	t y	t^2
0	12	0	0
1	19	19	1
2	29	58	4
3	37	111	9
4	45	180	16
$\Sigma x = 10$	$\Sigma y = 142$	$\Sigma xy = 368$	$\Sigma x^2 = 30$

We now calculate a and b using the least square regression formulas for a and b.

$a = (n\Sigma t\,y - \Sigma t\Sigma y) / (n\Sigma t^2 - (\Sigma t)^2) = (5*368 - 10*142) / (5*30 - 10^2)$

$= 8.4$

$b = (1/n)(\Sigma y - a\,\Sigma x) = (1/5)(142 - 8.4*10) = 11.6$

b) In 2012, t = 2012 - 2005 = 7

The estimated sales in 2012 are: y = 8.4 * 7 + 11.6 = 70.4 million dollars.

The End