BACS3013 Data Science

Tutorial 5 (Data Collection, Sampling and Pre-processing - part 3)

Q1. Given a set of paired data (X, Y)

a. if Y is independent of X, then what value of a correlation coefficient would you expect?

b. if Y is linearly dependent on X, then what value of a correlation coefficient would you expect?

c. How could Y be closely dependent upon X yet $r \approx 0$?

Ans:

a. r = 0.
b. $r \approx 1$ or $r \approx -1$ (these two are same as $|r| \approx 1$).
c. Y could be a quadratic function of X, as an example. The correlation coefficient is a measure of the scatter about a straight line — a linear function of X.

Q2. Some data are given as:

X	Y
1	16
2	23
4	35
3	28
5	44
6	40
3	22
8	61
9	82

a. Sketch a scatterplot.

b. Compute the correlation coefficient, r.

c. Compute the coefficients of the linear regression line, $y = b_1x + b_0$.

d. What is the estimated value, y_p for $x = 7$?

Ans:

b.

$$r = \frac{1}{(n-1)} \sum \left[\frac{(X-\bar{X})}{s_x} \cdot \frac{(Y-\bar{Y})}{s_y} \right]$$

BACS3013 Data Science

n = 9, $\Sigma X = 41$, $\Sigma Y = 351$, $\Sigma X^2 = 245$, $\Sigma Y^2 = 17259$, $\Sigma XY = 2038$.
$\bar{X} = 4.5556$, $\bar{Y} = 39.0000$, $s_x = 2.6977$, $s_y = 21.1246$.

$r = 2038 - 9 \bullet 4.5556 = 0.9629$.

c.

$b_1 = r(s_y/s_x) = 0.9629(21.1246/2.6977) = 7.5401$, and
$b_0 = \bar{Y} - b_1\bar{X} = 39.0000 - (0.9629) \bullet 4.5556 = 4.6508$.

Which gives: $y_p = b_0 + b_1 x = 4.6508 + 7.5401x$.

d. Evaluating for $x = 7$ gives: $y_p = 4.6508 + 0.9629 \bullet 7 = 57.43$.

Q3.

Interesting data are given as:

X	Y
72	45
73	38
75	41
76	35
77	31
78	40
79	25
80	32
80	36
81	29
82	34
83	38
84	26
85	32
86	28
88	27

a. Sketch a scatterplot.
b. Compute the correlation coefficient, r.
c. Compute the coefficients of the linear regression line, $y = b_1 x + b_0$.
d. What is the estimated value for $X = 7$?

Ans:

b.

$n = 16,\ \Sigma X = 1279,\ \Sigma Y = 537,\ \Sigma X^2 = 102563,\ \Sigma Y^2 = 18535,\ \Sigma XY = 42650.$
$\overline{X} = 79.9375,\ \overline{Y} = 33.5625,\ s_x = 4.6400,\ s_y = 5.8420.$

$r = -0.6799.$

c.

$b_1 = r(s_y/s_x) = -0.6799(5.4842/4.6400) = -0.8560$, and
$b_0 = \overline{Y} - b_1\overline{X} = 33.5625 - (-0.8560) \cdot 79.9375 = 101.9879.$

Which gives: $y_p = b_0 + b_1x = 101.9897 - 0.8560x.$

d. Evaluating for x = 80 gives: $y_p = 101.9897 - 0.8560 \cdot 80 = 33.51.$

Q4. There are 110 houses in a particular neighborhood. Liberals live in 25 of them, moderates in 55 of them, and conservatives in the remaining 30. An airplane carrying 65 lb. sacks of flour passes over the neighborhood. For some reason, 20 sacks fall from the plane, each miraculously slamming through the roof of a different house. None hit the yards or the street, or land in trees, or anything like that. Each one slams through a roof. Anyway, 2 slam through a liberal roof, 15 slam through a moderate roof, and 3 slam through a conservative roof. Should we reject the hypothesis that the sacks of flour hit houses at random?

Given the numbers of liberals, moderates and conservative households, we can calculate the expected number of sacks of flour to crash through each category of house:

20 sacks x 25/110 = 4.55 liberal roofs smashed
20 sacks x 55/110 = 10.00 moderate roofs smashed
20 sacks x 30/110 = 5.45 conservative roofs smashed
Set up the table for the goodness-of-fit test:

Category	Observed	Expected	Obs-Exp	(Obs-Exp)^2 / Exp
Liberal	2	4.55	-2.55	1.43
Moderate	15	10.00	5.00	2.50
Conservative	3	5.45	-2.45	1.10
Total	20	20.00	0	5.03

In a simple test like this, where there are three categories and where the expected values are not influenced by the observed values, there are two degrees of freedom. Checking the table of critical values of the chi-square distribution for 2 d.f., we find that $0.05 < p < 0.10$. That is, there is greater than a 5% probability, but less than a 10% probability, of getting at least this much departure between observed and expected results by chance. Therefore,

while it appears that moderates have had worse luck than liberals and conservatives, we cannot reject the hypothesis that the sacks of flour struck houses at random.

Q5. The Acme Battery Company has developed a new cell phone battery. On average, the battery lasts 60 minutes on a single charge. The standard deviation is 4 minutes.

Suppose the manufacturing department runs a quality control test. They randomly select 7 batteries. The standard deviation of the selected batteries is 6 minutes. What would be the chi-square statistic represented by this test?

Ans:

The standard deviation of the population is 4 minutes.
The standard deviation of the sample is 6 minutes.
The number of sample observations is 7.
To compute the chi-square statistic, we plug these data in the chi-square equation, as shown below.

$$X^2 = [(n - 1) * s^2] / \sigma^2$$

$$X^2 = [(7 - 1) * 6^2] / 4^2 = 13.5$$

where X^2 is the chi-square statistic, n is the sample size, s is the standard deviation of the sample, and σ is the standard deviation of the population.

Q6. A bank had disbursed 60816 auto loans with around 2.5% of the bad rate in the quarter between April–June 2012. The summary of data is shown in the table below. By using information value filter, deduct the level of predictive power of this dataset.

Age Group	Total Number of Loans	Number of Bad Loans	Number of Good Loans
21 - 30	4821	206	4615
30 - 36	10266	357	9909
36 - 48	32926	776	32150
48 - 60	12788	183	12605
Total	**60801**	**1522**	**59279**

BACS3013 Data Science

Ans:

Age Group	% Bad Loans	Name of Coarse Group	Distibution Bad (DB)	Distibution Good (DG)	WOE	DG - DB	(DG - DB)* WOE
21 - 30	4.3%	G1	0.135	0.078	-0.553	-0.057	0.0318
30 - 36	3.5%	G2	0.235	0.167	-0.339	-0.067	0.0228
36 - 48	2.4%	G3	0.510	0.542	0.062	0.032	0.0020
48 - 60	1.4%	G4	0.120	0.213	0.570	0.092	0.0527
Total						Information Value -->	0.1093

The information value is 0.1093. It is barely falling into medium predictors' range