Univariate Statistical Analysis

Lecture 10

Correlation and Regression
Chapter 13

Today

- (Pearson) Correlation Coefficient → r
- r^2
- Simple Linear Regression Model → $y = \beta_0 + \beta_1 x$
  - How to form the model
  - Prediction
  - Draw the trend line

Correlation Coefficient

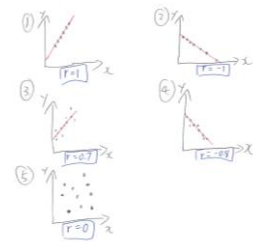
Correlation Coefficient, r, is a way to **measure** the linear **relationship** between two variables.

Scale from -1 to 1

For examples
-1 means perfect negative linear relationship
- 0.8 means strong negative linear relationship
0 means no linear relationship
0.3 means weak positive linear relationship
1 means perfect positive linear relationship

Correlation Coefficient

Describe the relationship between two variables.



Example 1

Given the following set of data

x	y
17	94
13	73
12	59
15	80
16	93
14	85
16	66
16	79
18	77
19	91

i). Calculate and interpret Correlation Coefficient, r
ii). Form a simple regression model
iii). Predict the y value when x = 30
iv). Plot the data and draw a trend line
v). Calculate and interpret r^2

Correlation Coefficient

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

Example 1 Cont'

i). Calculate and interpret Correlation Coefficient, r

$\sum x = 156$, $\sum y = 797$, n = 10

So,

$\bar{x} = \frac{\sum x}{n} = \frac{156}{10} = 15.6$

$\bar{y} = \frac{\sum y}{n} = \frac{797}{10} = 79.7$

x	y
17	94
13	73
12	59
15	80
16	93
14	85
16	66
16	79
18	77
19	91

Example 1 Cont'

i). Calculate and interpret Correlation Coefficient, r

x	y	$x-\bar{x}$	$y-\bar{y}$
17	94	-1.4	-14.3
13	73	2.6	6.7
12	59	3.6	20.7
15	80	0.6	-0.3
16	93	-0.4	-13.3
14	85	1.6	-5.3
16	66	-0.4	13.7
16	79	-0.4	0.7
18	77	-2.4	2.7
19	91	-3.4	-11.3

Example 1 Cont'

i). Calculate and interpret Correlation Coefficient, r

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
17	94	-1.4	-14.3	1.96	204.49
13	73	2.6	6.7	6.76	44.89
12	59	3.6	20.7	12.96	428.49
15	80	0.6	-0.3	0.36	0.09
16	93	-0.4	-13.3	0.16	176.89
14	85	1.6	-5.3	2.56	28.09
16	66	-0.4	13.7	0.16	187.69
16	79	-0.4	0.7	0.16	0.49
18	77	-2.4	2.7	5.76	7.29
19	91	-3.4	-11.3	11.56	127.69

Example 1 Cont'

i). Calculate and interpret Correlation Coefficient, r

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
17	94	-1.4	-14.3	1.96	204.49	20.02
13	73	2.6	6.7	6.76	44.89	17.42
12	59	3.6	20.7	12.96	428.49	74.52
15	80	0.6	-0.3	0.36	0.09	-0.18
16	93	-0.4	-13.3	0.16	176.89	5.32
14	85	1.6	-5.3	2.56	28.09	-8.48
16	66	-0.4	13.7	0.16	187.69	-5.48
16	79	-0.4	0.7	0.16	0.49	-0.28
18	77	-2.4	2.7	5.76	7.29	-6.48
19	91	-3.4	-11.3	11.56	127.69	38.42

Example 1 Cont'

i). Calculate and interpret Correlation Coefficient, r

$\sum((x-\bar{x})(y-\bar{y})) \approx 134.8$ (add the 7th column in the last slide)

$\sum(x-\bar{x})^2 \approx 42.4$ (add the 5th column in the last slide)
$\sqrt{\sum(x-\bar{x})^2} \approx 6.51$

$\sum(y-\bar{y})^2 \approx 1206.1$ (add the 6th column in the last slide)
$\sqrt{\sum(y-\bar{y})^2} \approx 34.72$

So,

$$r = \frac{\sum((x-\bar{x})(y-\bar{y}))}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}} = \frac{134.8}{(6.51)(34.72)} \approx 0.6$$

There is a moderate positive linear relationship between two variables.

Simple Linear Regression Model

$$y = c + mx$$

or

$$y = b + mx$$

Remembered this in high school?

Simple Linear Regression Model

$$y = \beta_0 + \beta_1 x$$

β_1 is the slope of the model. In more technical, slope is the rate of change between y and x, i.e. what is the change in y given 1 unit change in x?

For example
β_1 = 0.5 When 1 unit changes in x-axis, 0.5 units change in y-axis.

β_0 is the y-intercept. It is the value of y when x = 0.

x is the independent variable.

y is the dependent variable. It is because the value of y is subject to the value of x.

Simple Linear Regression Model

β_1 is the slope of the model.

$$\beta_1 = r \frac{S_y}{S_x}$$

r is the Correlation Coefficient

Standard Deviation of y

$$S_y = \sqrt{\frac{\sum(y-\bar{y})^2}{(n-1)}}$$

Standard Deviation of x

$$S_x = \sqrt{\frac{\sum(x-\bar{x})^2}{(n-1)}}$$

Simple Linear Regression Model

β_0 is the y-intercept.

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

β_1 is the slope

$\bar{y}$ is the mean of y

$\bar{x}$ is the mean of x

Example 1 Cont'

ii). Form a simple regression model

From part i).
$\sum(x-\bar{x})^2 \approx 42.4$; $\sum(y-\bar{y})^2 \approx 1206.1$; r ≈ 0.6 ; n=10

$$S_y = \sqrt{\frac{\sum(y-\bar{y})^2}{(n-1)}} = \sqrt{\frac{1206.1}{(10-1)}} \approx 11.58$$

$$S_x = \sqrt{\frac{\sum(x-\bar{x})^2}{(n-1)}} = \sqrt{\frac{42.4}{(10-1)}} \approx 2.17$$

So,

$$\beta_1 = r \frac{S_y}{S_x} = (0.6)\frac{(11.58)}{(2.17)} \approx \frac{6.9}{2.17} \approx 3.18$$

Example 1 Cont'

ii). Form a simple regression model

From part i). and ii).

$\bar{x} = 15.6$; $\bar{y} = 79.7$; $\beta_1 = 3.18$

$\beta_0 = \bar{y} - \beta_1 \bar{x}$
$\beta_0 = 79.7 - (3.18)(15.6)$
$\beta_0 = 79.7 - 49.6$
$\beta_0 = 30.1$

So, the regression model is
$y = \beta_0 + \beta_1 x$
$y = 30.1 + 3.18x$

Example 1 Cont'

iii). Predict the y value when x = 30

From part ii).

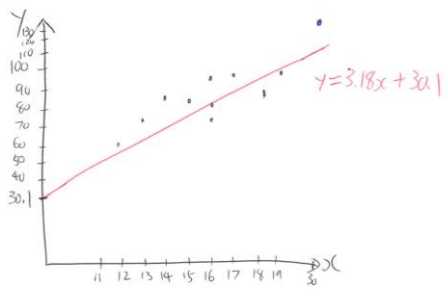
The regression model is
$y = 30.1 + 3.18x$

when x = 30,

$y = 30.1 + (3.18)(30)$
$y \approx 30.1 + 95.38$
$y \approx 125.48$

Example 1 Cont'

iv). Plot the data and draw a trend line



Example 1 Cont'

v). Calculate and interpret r^2

From part i).

$r = 0.6$
$r^2 \approx (0.6)^2$
$r^2 \approx 0.36 \rightarrow 36\%$

So, 36% of the variability in y can be explained by the model (the variability of x).

Example 1 Cont'

For your information
Use statistics software to do the analysis, the following result is shown:

r^2

Regression Statistics	
Multiple R	0.596
R Square	0.355
Adjusted R Square	0.275
Standard Error	9.859
Observations	10.000

ANOVA

	df	SS	MS	F	Significance F
Regression	1	428.5622642	428.5623	4.40943	0.068952083
Residual	8	777.5377358	97.19222		
Total	9	1206.1			

β_0 β_1

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	30.104	23.824	1.264	0.242	-24.834	85.041	-24.834	85.041
x	3.179	1.514	2.100	0.069	-0.312	6.671	-0.312	6.671

Example 2 – Mother's Age and Baby's Birth Weight (Example 13.2, p.695)

Required
i). Calculate and interpret Correlation Coefficient, r
ii). Calculate and interpret r^2
iii). Form a simple regression model
iv). Predict the y value when x = 20 and x = 40
v). Plot the data and draw a trend line

Example 2 Cont'

Solutions
i). Calculate and interpret Correlation Coefficient, r
$r \approx 0.88$ means strong positive linear relationship between maternal age and birth weight of baby.

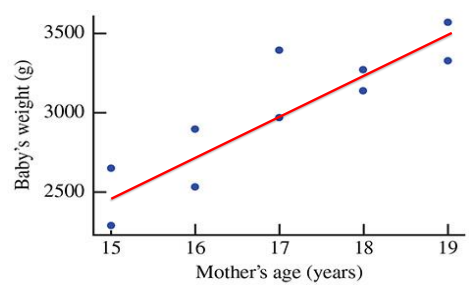
ii). Calculate and interpret r^2
$r^2 \approx 0.78 \rightarrow 78\%$
So, 78% of the variability in birth weight of baby can be explained by the variability of maternal age.

iii). Form a simple regression model
$y = -1{,}163.45 + 245.15x$

iv). Predict the y value when x = 20 and x = 40
$x = 20, y = 3{,}739.55$ grams
$x = 40, y = 8{,}642.55$ grams (Not realistic)

Example 2 Cont'

Solutions
v). Plot the data and draw a trend line



Conclusion

- (Pearson) Correlation Coefficient → r
- r^2
- Simple Linear Regression Model → $y = \beta_0 + \beta_1 x$
  - How to form the model
  - Prediction
  - Draw the trend line