Chapter 6: Relating Continuous Variables

Scatterplots and Relationships between Variables

Positive and Negative Associations

The Correlation Coefficient

Limitations of Correlation

Causation

Scatterplots (page 140)

A **scatterplot** is an *x*-*y* plot of pairs of variables.

Scatterplots permit us to look for relationships between variables.

Is there a tendency for one variable to increase when the other one increases?

Or for one to decrease when the other one increases?

How much variation is there around the general pattern or trend?

Example: Inflation Rates and Central Bank Independence, 1955-1988

Rate	Indepen	Country
7.8	1.00	New Zealand
8.5	1.50	Spain
7.2	1.75	Italy
6.8	2.00	United Kingdom
6.5	2.00	Australia
6.0	2.00	France
6.0	2.00	Norway
6.0	2.00	Sweden
4.1	2.00	Belgium
6.5	2.50	Denmark
4.9	2.50	Japan
4.5	2.50	Canada
4.2	2.50	The Netherlands
4.2	3.50	United States
3.3	4.00	Switzerland
2.9	4.00	Germany

The Correlation Coefficient *r* (page 147)

- Compute the means and standard deviations of both variables
- Standardize each observation, both variables, case by case
- Multiply the standardized observations together, case by case
- Add the resulting products and divide by $n-1$ to get r

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(\frac{x_i-\bar{x}}{s_x}\right)\cdot\left(\frac{y_i-\bar{y}}{s_y}\right)\right]$$

$$= \frac{1}{(n-1)s_x s_{yi}}\sum_{i=1}^{n}[(x_i-\bar{x})(y_i-\bar{y})]$$

Facts about *r* (page 148)

- The correlation coefficient is unitless
- *r* is always between –1 and +1
- If *r* = +1, the pairs lie exactly on a straight line with positive slope; If *r* = –1, the pairs lie exactly on a straight line with negative slope
- An *r* = 0 indicates that there is no straight-line component in the relationship between the two variables

Example:

Row	x	y	Std x	Std y	Prod
1	1	2	–1	–2/2 = –1	+1
2	3	4	+1	0	0
3	2	6	0	2/2 = +1	0
Total	6	12	0	0	+1

$\bar{x} = 2$ and $\bar{y} = 4$. Also $s_x = 1$ and $s_y = 2$.

So $r = +1/2 = 0.5$

Central Bank Example:

Variable	N	Mean	StDev
Rate	16	5.588	1.629
Indepen	16	2.359	0.837

Rate	Indepen	StdRate	StdIndep	Product
7.8	1.00	1.35832	–1.62507	–2.20735
8.5	1.50	1.78807	–1.02734	–1.83695
7.2	1.75	0.98996	–0.72848	–0.72116
6.8	2.00	0.74439	–0.42962	–0.31980
6.5	2.00	0.56021	–0.42962	–0.24067
6.0	2.00	0.25325	–0.42962	–0.10880
6.0	2.00	0.25325	–0.42962	–0.10880
6.0	2.00	0.25325	–0.42962	–0.10880
4.1	2.00	–0.91322	–0.42962	0.39233
6.5	2.50	0.56021	0.16811	0.09418
4.9	2.50	–0.42208	0.16811	–0.07096
4.5	2.50	–0.66765	0.16811	–0.11224
4.2	2.50	–0.85183	0.16811	–0.14320
4.2	3.50	–0.85183	1.36356	–1.16152
3.3	4.00	–1.40436	1.96129	–2.75436
2.9	4.00	–1.64993	1.96129	–3.23599
Total		0	0	–12.644

So $r = –12.644/(16–1) = –0.843$

The Limitations of Correlations (page 151)

The correlation coefficient is sensitive to outliers.

x	y
1	1
2	1.5
3	1
4	2
5	1.5
6	2.5

The correlation coefficient is 0.777

If the last *y* is changed to 1 instead of 2.5 the correlation coefficient changes to 0.131.

Ecological correlations are correlations computed from data that are totals of smaller units of study. (page 152)

Example:

Curved relationships

Example:

Correlation and Cause (page 154)

- **Consistency**

  The relationship must hold in other contexts.

- **Responsiveness**

  Must show that a change in *x* leads to a change in *y*.

- **Mechanism**

  Must construct a model that shows step by step how the presumed cause leads to the presumed effect.

Lurking variables (page 156)

The relationship between *x* and *y* may be caused, wholly or in part, by a third variable *z* that is related to both *x* and *y*.

Examples:

- There is a high positive correlation between the number of fire trucks on the scene and the amount of damage suffered.

- There is a high positive correlation between the percent of the population married and the consumption of candy.

- A study of hospitals finds a high positive correlation between the number of medical specialists on staff and the death rates of patients.