Public Policy 529
Correlation Analysis

Jonathan Hanson

Gerald R. Ford School of Public Policy
University of Michigan

November 13, 2023

Outline

1. Preliminaries

2. Calculating Pearson's *r*

3. Using Software

Outline

1. Preliminaries

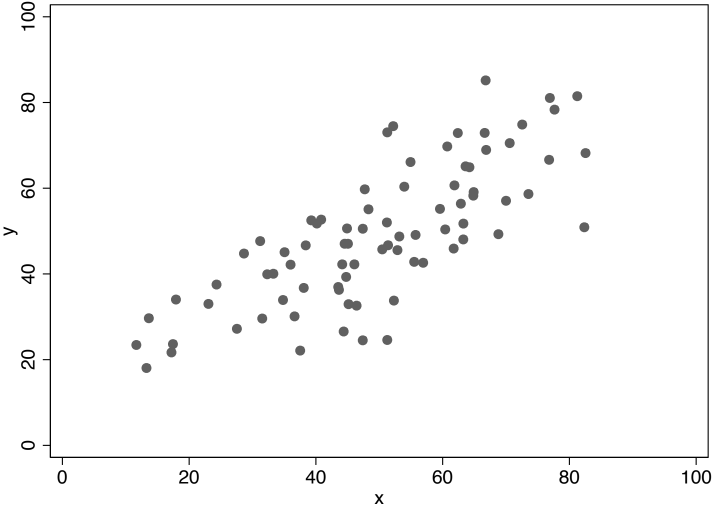
2. Calculating Pearson's r

3. Using Software

When to use Correlation Analysis

- Correlation analysis is appropriate when all of our variables have the interval-level of measurement.
- In this scenario, we cannot take means of one variable across the categories of another.
- Likewise, joint frequency distributions would be overwhelmed with too many values.
- Instead, we analyze the strength and direction (positive, negative) of the relationship between two or more variables.

Consider a Basic Scatterplot



We see a positive relationship between x and y. Correlation analysis provides a systematic way to measure this correlation.

Correlation Analysis

- The most common measure of correlation is the Pearson's r correlation coefficient.

- This coefficient, which ranges from -1 to $+1$, measures the direction and linearity of the relationship between two variables.

- The sign ($+$ or -) indicates whether the two variables are positively or negatively (i.e. inversely) correlated.

- At the extremes, -1 or $+1$, the variables are perfectly linear. At 0, there is no relationship.

Software Commands: Quick Version

- Stata

```
pwcorr var1 var2, sig
```

- R

```
cor.test(data$var1, data$var2)
```

Software reports the Pearson's *r* correlation coefficient and a *p*-value versus the null hypotheses that the true correlation is 0.

Example: Female Life Expectancy and Democracy

```
. pwcorr CIALIF_F democ11, sig

             | CIALIF_F  democ11
-------------+------------------
    CIALIF_F |   1.0000
             |
             |
     democ11 |   0.3751   1.0000
             |   0.0000
             |
```

The Pearson's *r* correlation coefficient is .3751, which is weak-to-moderately positive. When democracy is higher, female life expectancy tends to be higher also.

The *p*-value is .0000, so we reject the null hypothesis that $r = 0$.

Outline

1. Preliminaries

2. Calculating Pearson's r

3. Using Software

Pearson's r

A correlation coefficient with a value in the range of -1 to $+1$.

$$r \approx \frac{\text{covariation of } x \text{ and } y}{\text{total variation of } x \text{ and } y}$$

More specifically:

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

There are other formulas in the textbooks, but this is the easiest to interpret.

Consider the Formula

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

- The denominator includes the sums of squared deviations for x and y. A measure of total variation, it is always positive.
- The numerator can be positive or negative. It captures covariation of x and y.
- For each observation, the numerator calculates the product of the deviations of x and y from their means. These products are summed up for all n observations.

A Closer Look at the Numerator

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

- For an observation, if the deviations of x and y from their means are both positive, or both negative, their product will be positive.
- Conversely, if the deviation of one variable is positive, while the other is negative, their product will be negative.
- This has substantive meaning: if x is above its mean at the same time y is above its mean, that indicates positive correlation.
- Conversely, if x (or y) is above its mean while y (or x) is below its mean, that indicates negative correlation.

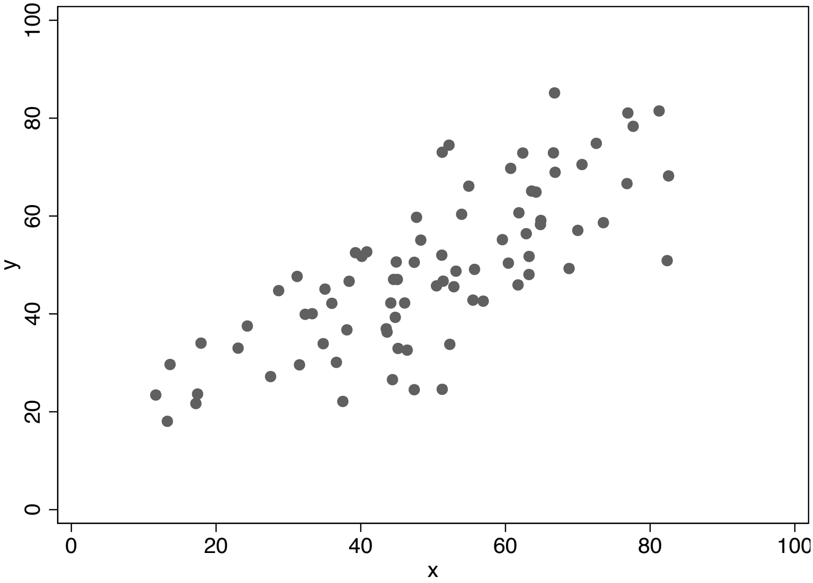
Example: Calculating the Numerator

Suppose $\bar{x} = 4$ and $\bar{y} = 6$.

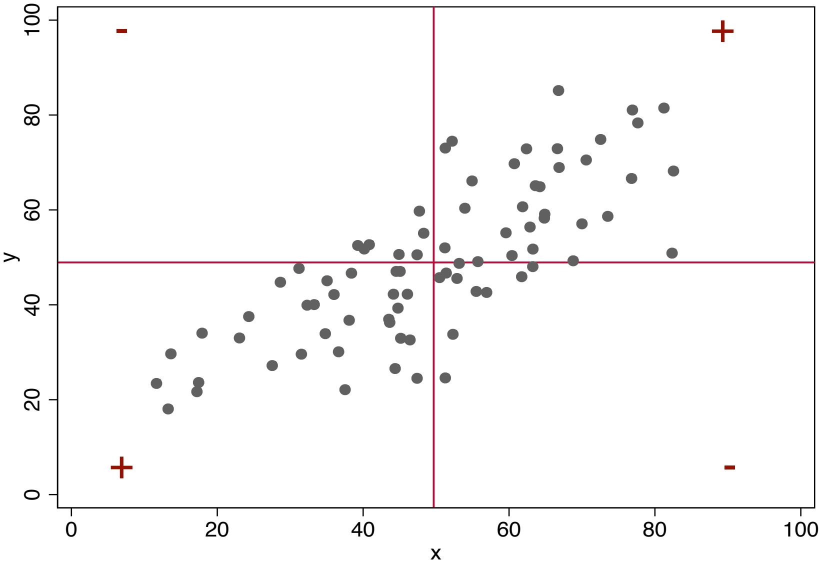
x	$x - \bar{x}$	y	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
6	2	8	2	4
3	-1	7	1	-1
4	0	4	-2	0
5	1	7	1	1
				4

The sum of the products is $+4$. Since the sum is positive, the tendency across the four cases is positive correlation.

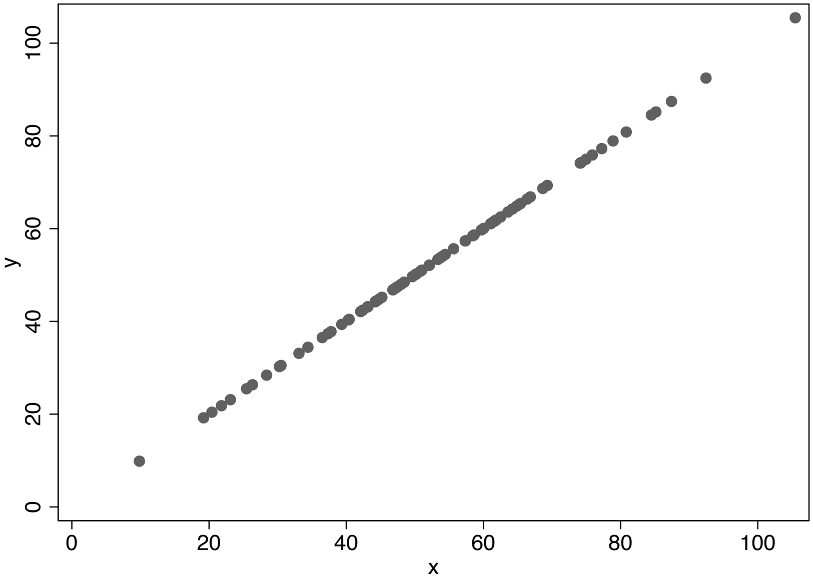
$r = .77$



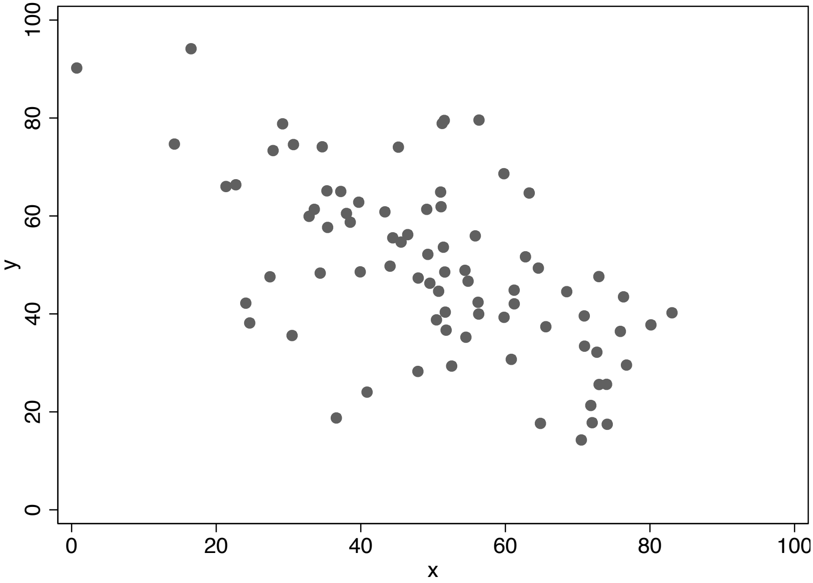
$r = .77$



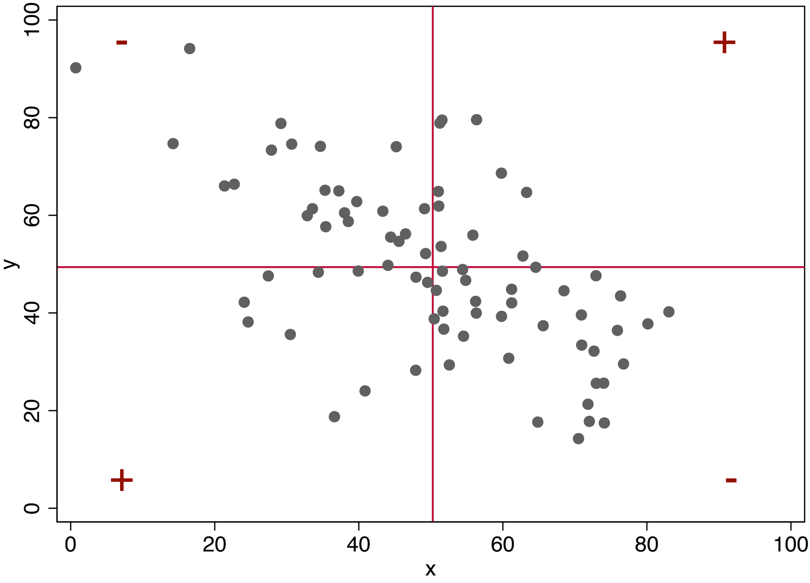
$r = 1.00$



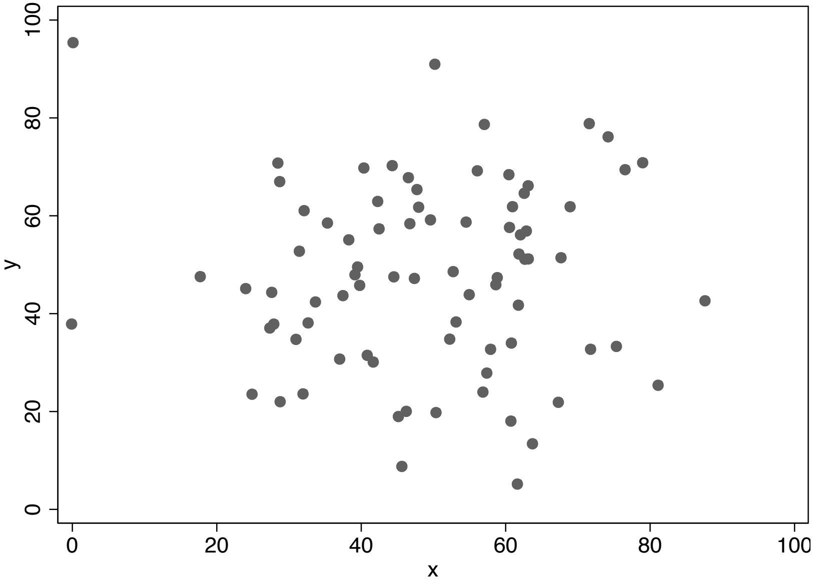
$r = -.6$



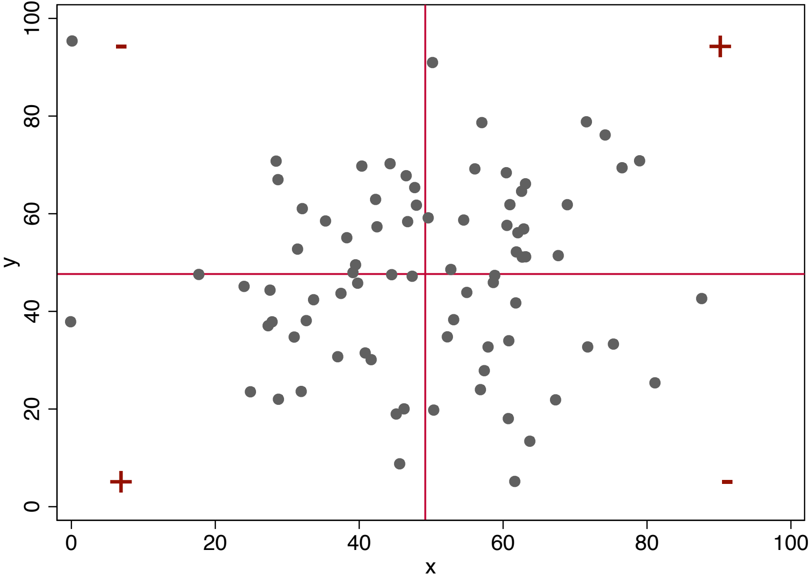
$r = \text{-}.6$



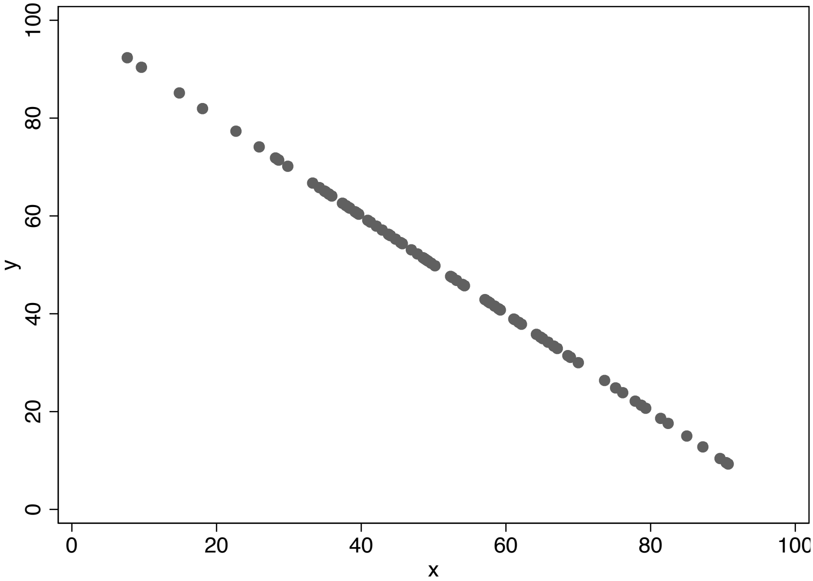
r = -.01



r = -.01



r = -1.00



Standard Error of Pearson's *r*

The sampling distribution of the correlation coefficient can be approximated by a *t* distribution with $n - 2$ degrees of freedom.

The formula for the standard error is:

$$\mathrm{se}_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

The *t*-statistic for *r* (vs. the null hypothesis that r=0) is:

$$t = \frac{r}{se_r}$$

Outline

1. Preliminaries

2. Calculating Pearson's r

3. Using Software

More Sofware Commands

- Commands for correlation coefficients between 2+ variables.

  Stata: `pwcorr var1 var2 var3 var4, sig`

  R: `cor(data[c("var1", "var2", "var3", "var4")], use = "pairwise.complete.obs")`

- Software reports out a correlation matrix, a square matrix with the same number of rows/columns as variables in the analysis.
- The correlation between any two variables is given by the row/column where they intersect.
- Since this matrix is symmetric, Stata reports only the lower left triangle.

Format of a Correlation Matrix

	var1	var2	var3	var4
var1	$r_{1,1}$	$r_{1,2}$	$r_{1,3}$	$r_{1,4}$
var2	$r_{1,2}$	$r_{2,2}$	$r_{2,3}$	$r_{2,4}$
var3	$r_{1,3}$	$r_{2,3}$	$r_{3,3}$	$r_{3,4}$
var4	$r_{1,4}$	$r_{2,4}$	$r_{3,4}$	$r_{4,4}$

Elements on the main diagonal, from upper left to lower right, always equal 1. The diagonal is each variable correlated with itself.

Symmetry makes the lower left and upper right triangles duplicative. Stata leaves out the upper right.

Stata Output

```
. pwcorr autoc ciaedex ciaunemp gini08 unions

             |    autoc  ciaedex ciaunemp   gini08   unions
-------------+---------------------------------------------
       autoc |   1.0000
     ciaedex |  -0.0802   1.0000
    ciaunemp |   0.0081   0.2015   1.0000
      gini08 |  -0.0154   0.0357   0.2129   1.0000
      unions |   0.1762   0.2567  -0.2112  -0.4306   1.0000
```

Variables from world.dta: `autoc` (autocracy score), `ciaedex` (% GDP spent on education), `ciaunemp` (unemployment rate), `gini08` (gini coefficient), and `unions` (union density).

The option `obs` would add *n* for each pair. The option `sig` would add the *p*-value.

Summary

- Correlation analysis is a simple, handy tool for measuring the relationship between two interval-level variables.
- Remember: the Pearson's correlation coefficient measures the direction and linearity of the relationship.
- Regression analysis also measures direction and linearity, but it adds magnitude and the ability to control for other factors.
- The correlation coefficient is unit-free. Regression analysis provides coefficients that measure the relationship between x and y in the units of y.