Lecture 3: The Relationship Between 2 Variables

Intro Stats with Nathan Favero

American University (Washington, DC)

August 2, 2024

Associations among variables

- So far, we've mostly focused on learning tools to help us describe a single variable
- Now, we want to start thinking about describing two variables jointly
- We often use words like "association," "related," or "correlated" when describing two variables
- The statistical tools we'll use to describe the relationship between two variables will depend on whether our variables are quantitative or qualitative

Cheatsheet: Which measure of association?

- 1 quantitative and 1 qualitative variable: compare means/medians across groups
- 2 qualitative variables: use a contingency table (crosstabs) to compare rates across groups
- 2 quantitative variables: correlation/regression
- Ordinal variables can be treated as quantitative or qualitative
  - Calculating a mean or a correlation with an ordinal variable requires you to assign specific numerical values to the ordered response options and treat the variable as though it is discrete/continuous; this is often a fine approximation
  - Alternatively, you can always treat ordinal variables like qualitative variables, though this will often make the analysis slightly more complex

1 quantitative & 1 qualitative variable

Example: Charter vs traditional schools

- The pass rate is a quantitative (continuous) variable
- The school type (charter vs traditional) is a qualitative variable
- We can calculate a median (or mean) pass rate for each category (a median for charter schools, and a median for traditional schools)
- Then we can compare the medians (means) to see which of the two groups (charters vs traditional) has a higher typical pass rate and how big the difference is

1 quantitative & 1 qualitative variable

Example: Charter vs traditional schools

```
. sum passrate if charter==0

    Variable |        Obs        Mean    Std. Dev.
-------------+---------------------------------------
    passrate |      1,021    45.01175     11.4699

. sum passrate if charter==1

    Variable |        Obs        Mean    Std. Dev.
-------------+---------------------------------------
    passrate |        166    40.86747    16.44257
```

1 quantitative & 1 qualitative variable

Example: Charter vs traditional schools

- Based on this output, we would conclude that traditional schools have a higher average pass rate than charter schools
- More specifically, we would say that the pass rate is about four percentage points higher in traditional schools
- Thus, the two variables (pass rate and school type) are related
- Based on the standard deviations, we can also conclude that charter schools have a wider spread (less consistency) than traditional schools

1 quantitative & 1 qualitative variable

Example: Charter vs traditional schools

- Instead of using means, we could compare median pass rates
- Using medians might be better if we are worried that the difference in means might be driven by a few outliers

1 quantitative & 1 qualitative variable

Example: Charter vs traditional schools

```
. tabstat passrate, s(median) by(charter)

Summary for variables: passrate
     by categories of: charter

         charter |       p50
-----------------+----------
Traditional Scho |        44
 Charter Schools |      39.5
-----------------+----------
           Total |        43
---------------------------
```

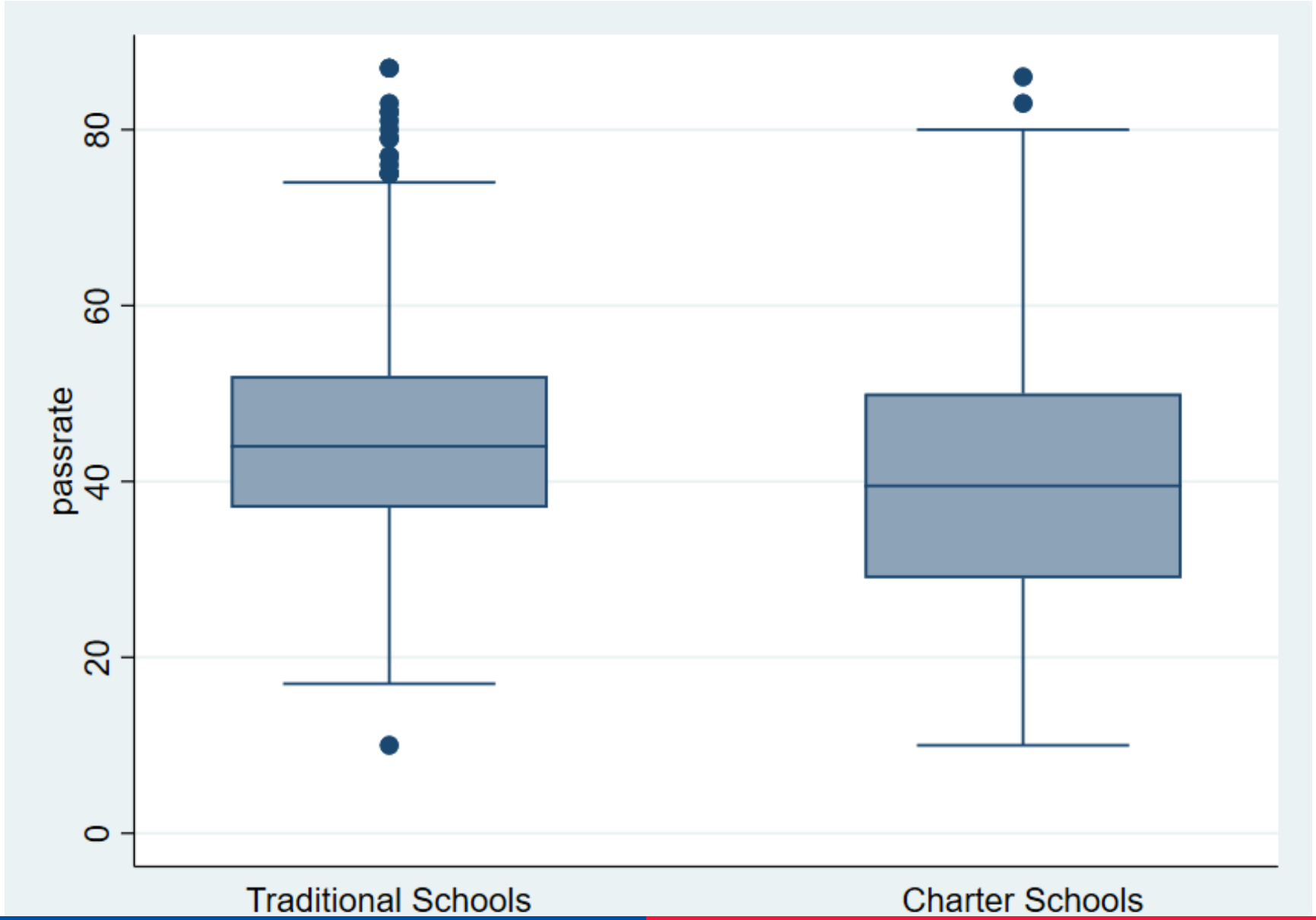
1 quantitative & 1 qualitative variable

Example: Charter vs traditional schools

- Comparing the medians provides very similar results to what we saw with the means
- According to the results on the prior slide, the median pass rate is 4.5 percentage points higher in traditional schools
- We can also study the association between pass rate and school type using a visual depiction of the data (and we'll reach a similar conclusion)

1 quantitative & 1 qualitative variable

Example: Charter vs traditional schools



2 qualitative variables: Crosstabs

- If we have two qualitative variables, we can create a contingency table showing the number of cases that fall into each cell
- Often a contingency table will show both raw numbers and proportions/percentages
  - For determining whether there is an association, focus on the proportions/percentages rather than the raw counts
  - Specifically, you want to note whether the percentage breakdown for one variable changes depending on the value of the other variable

2 qualitative variables: Crosstabs

Example: Are charter-school districts more likely to have only 1 school?

```
. tab oneschooldist charter, col

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| column percentage  |
+--------------------+

Single-Sch |
       ool |     District Type
  District | Tradition  Charter S |     Total
```

(continued on next slide)

2 qualitative variables: Crosstabs

Example: Are charter-school districts more likely to have only 1 school?

Single-School District	District Type: Tradition	Charter S	Total
0	819	109	928
	80.06	61.58	77.33
1	204	68	272
	19.94	38.42	22.67
Total	1,023	177	1,200
	100.00	100.00	100.00

2 qualitative variables: Crosstabs

Example: Who lives on campus?

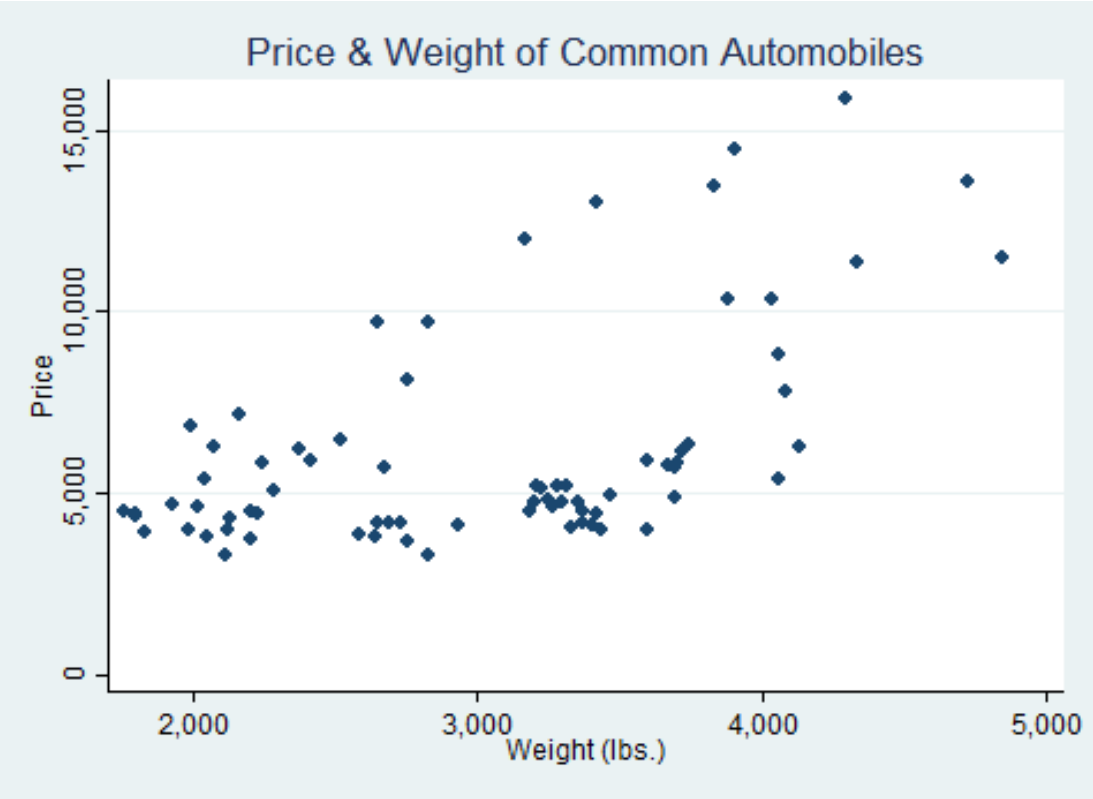
- Again, focus on the percentages listed, and the task is actually similar to comparing means
- Based on the table in the prior slide, 19.9% of traditional districts have only 1 school, while 38.4% of charter districts are single-school districts
- We conclude that there is an association between district type and the single-school district variable: charter districts are more likely to be single-school districts
- More specifically, charter districts are around 18 percentage points (38% - 20% = 18%) more likely to be single-school districts

2 quantitative variables: Correlation

- Correlation is a statistic we often use to describe the relationship between two quantitative variables
- Correlation will tell us whether higher values in one variable (X) predict higher values in another variable (Y)

2 quantitative variables: Correlation

- Graphs can help us explain correlation (for a more technical explanation of the math, see the appendix)
- We often graph two quantitative variables using a scatter plot (each axis corresponds to one of the variables)

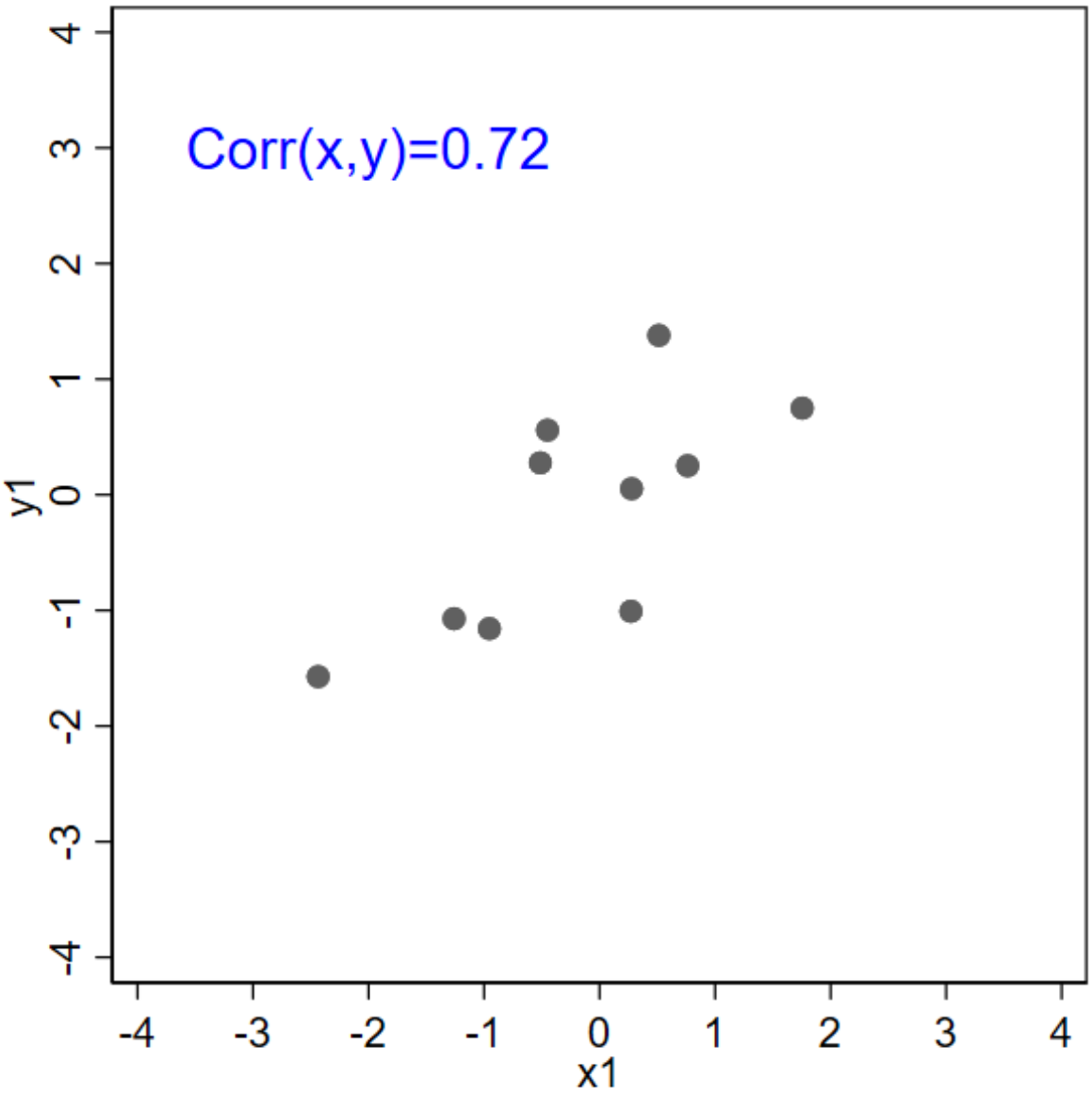


2 quantitative variables: Correlation

- Correlations always fall between -1 and 1
- One way to understand correlation is to think about drawing a straight line through the data and seeing how closely the data map to the line
- If the data falls perfectly along a straight line that points upward (going from left to right), the correlation will be 1
- If the data falls perfectly along a straight line that points downward (going from left to right), the correlation will be -1
- If the two variables are totally unrelated to one another, the correlation will be 0
- Many real-world correlations are non-zero but also don't fall perfectly along a line; we'll look at a couple examples now, but see the appendix for more details

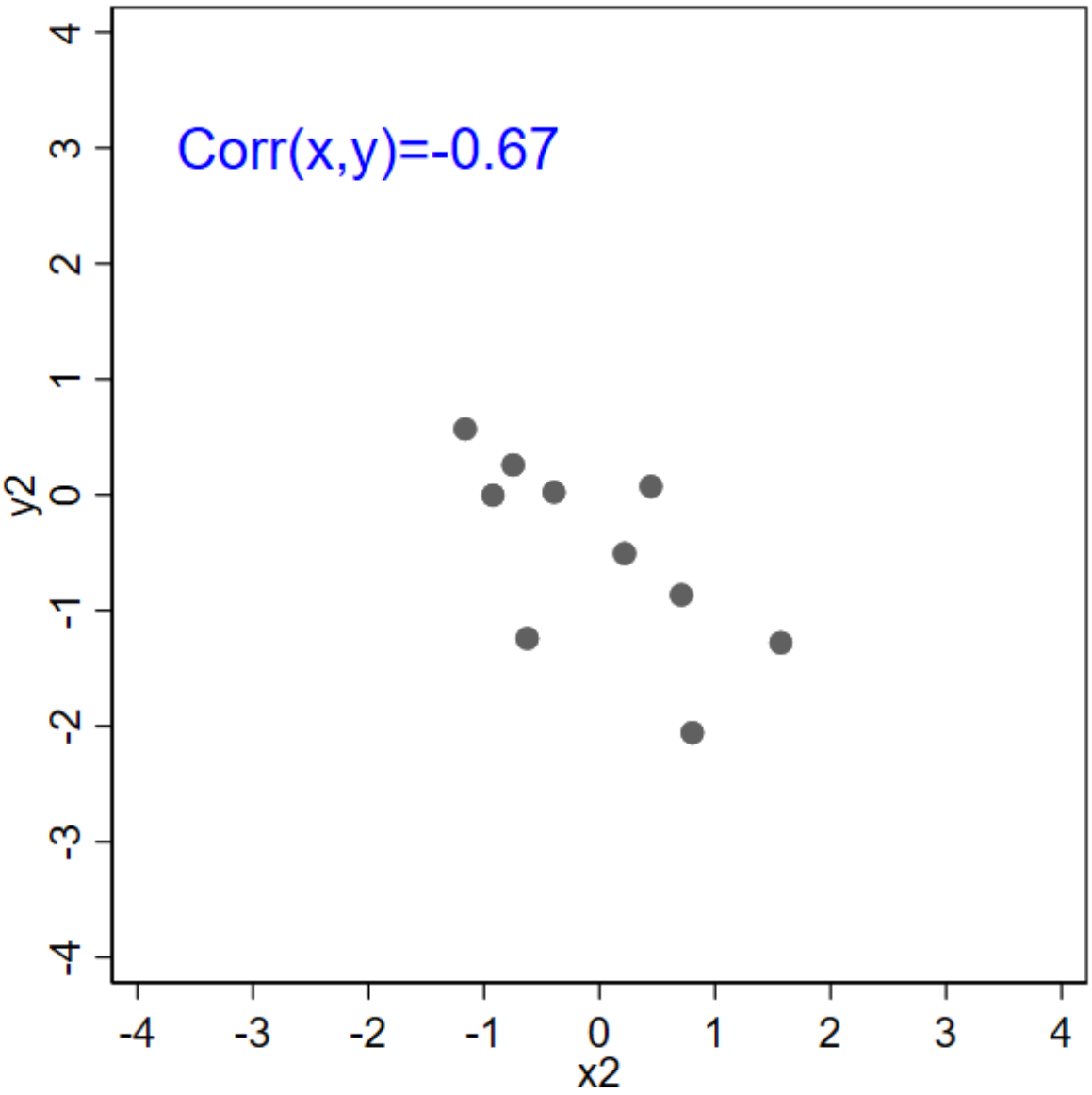
2 quantitative variables: Correlation

Positive values of correlation indicate a "positive relationship" (higher values of x are associated with higher values of y)



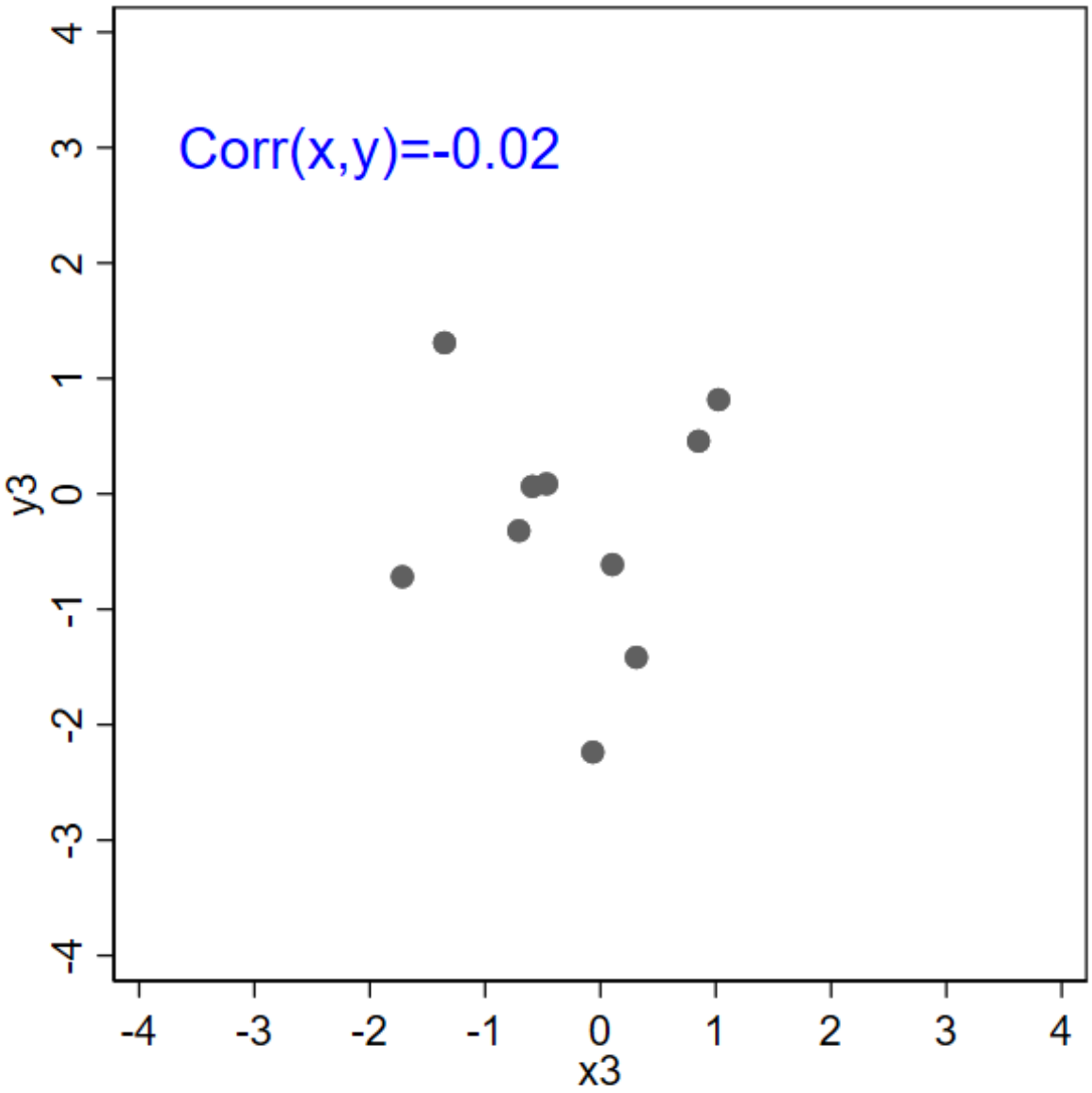
2 quantitative variables: Correlation

Negative values of correlation indicate a "negative relationship" (higher values of x are associated with lower values of y)



2 quantitative variables: Correlation

A correlation of zero (-0.02 is very close to 0) indicates no "linear" relationship between x and y
(More explanation of "linear" in the appendix)

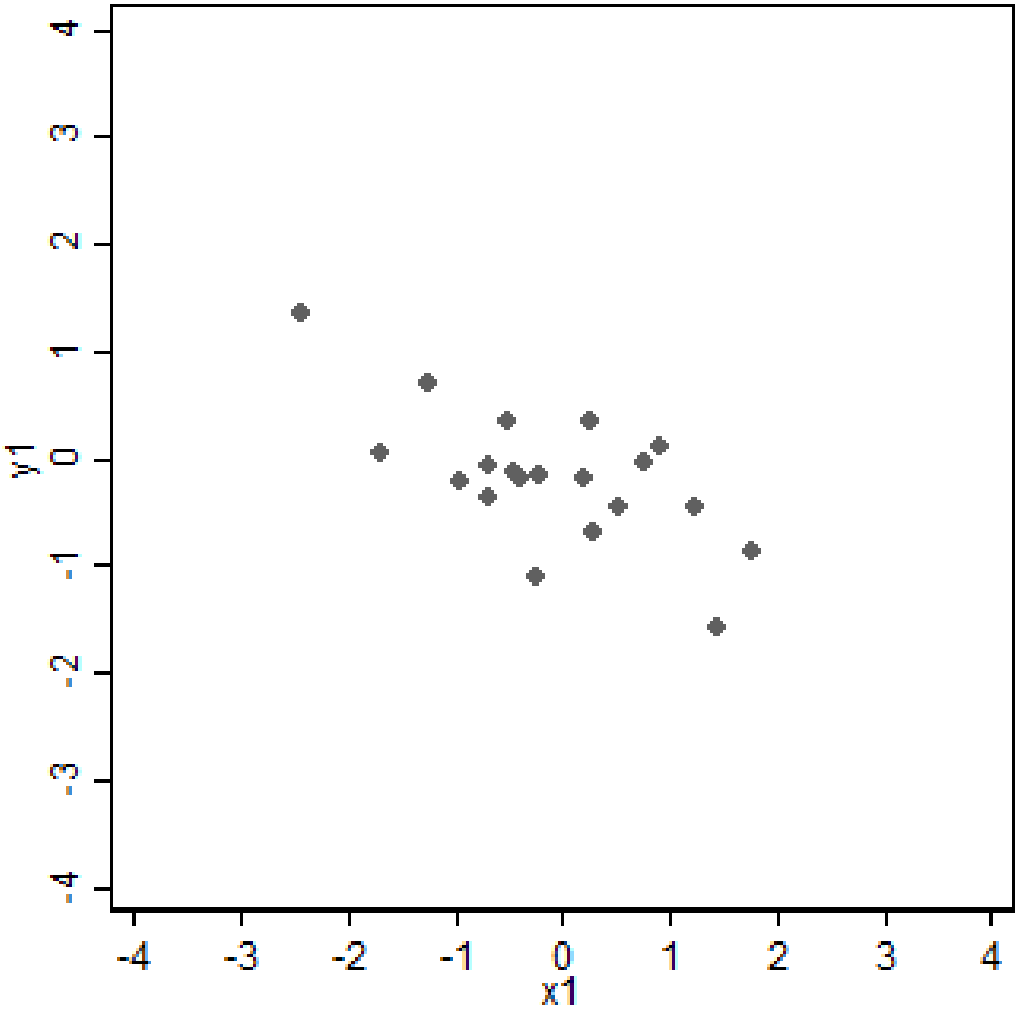


Linear regression

- For regression, it is important to distinguish between independent and dependent variables
- Independent variable: a variable that affects or explains another variable
- Dependent variable: a variable that is affected or explained by another variable
- Whenever we create graphs, the convention is to plot the independent variable (x) on the horizontal axis and the dependent variable (y) on the vertical axis

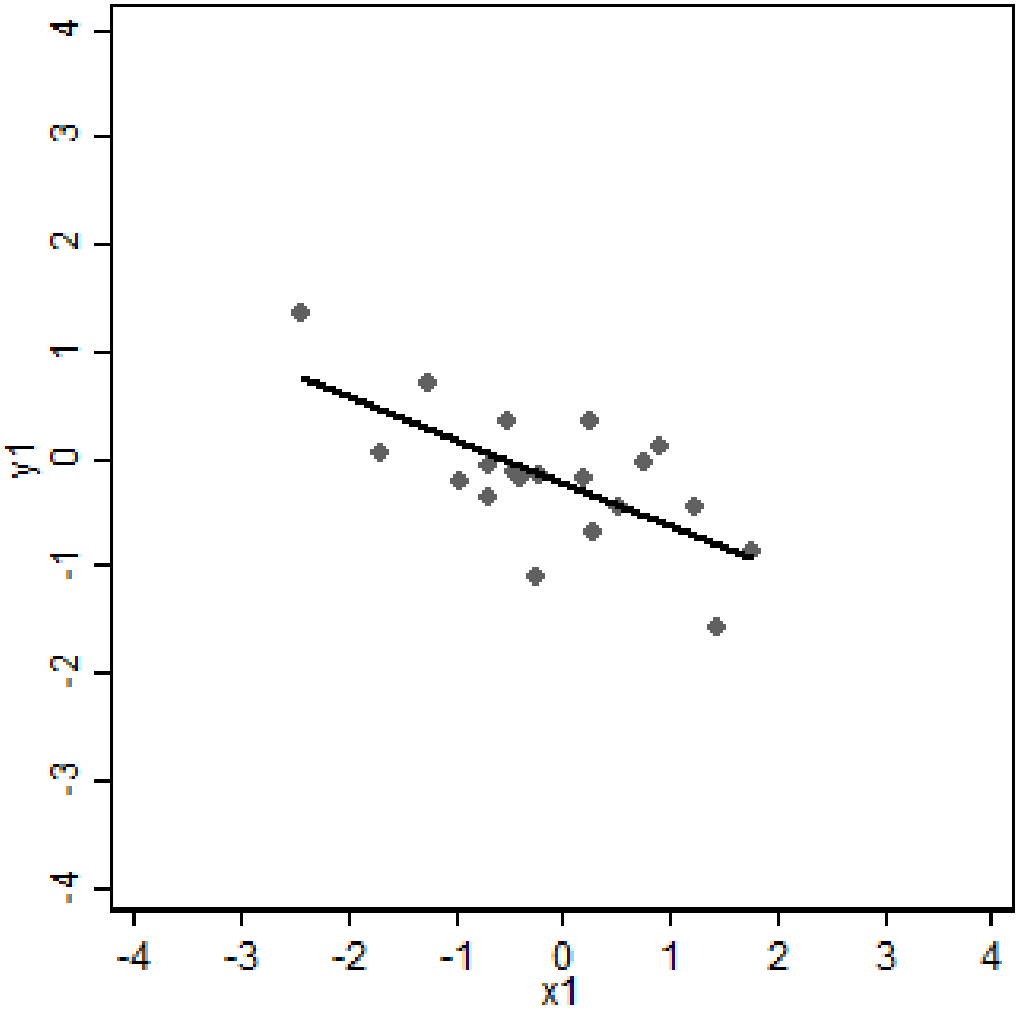
Linear regression

- Linear regression: drawing a line through some data



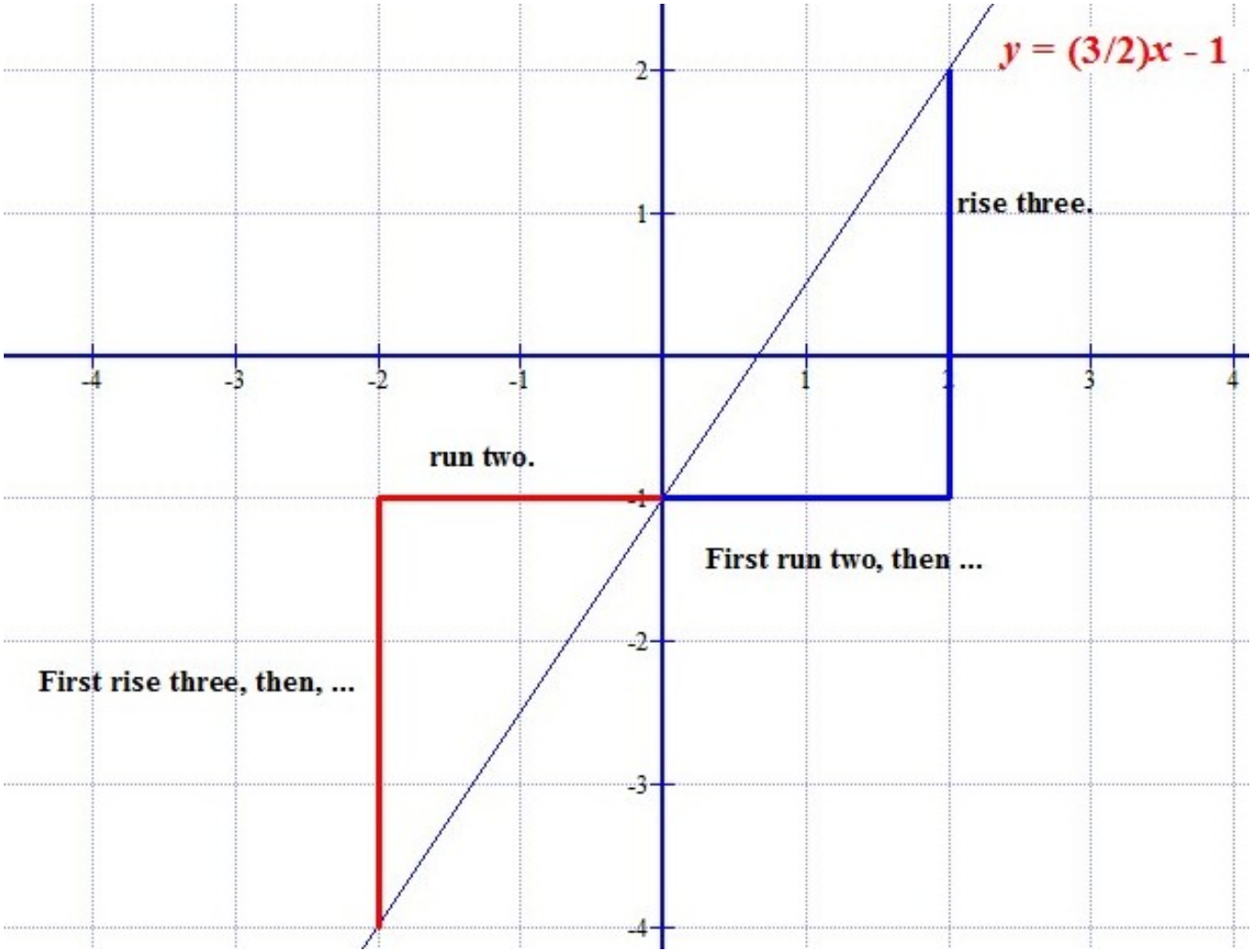
Linear regression

- Linear regression: drawing a line through some data



Linear functions

- To write the equation for a straight line, we need to know the y-intercept and the slope: $y = mx + b$



(Image: Public domain, Dino)

Linear functions

- In math class you might have seen lines represented as:

$$y = mx + b$$

- For regression we'll use the following notation:
  - a: the intercept (or constant, sometimes also written as α or β_0)
  - b: the slope (calculated as rise-over-run, sometimes also written as β or β_1)
  - It is assumed that Y is the dependent variable and X is the independent variable

$$\hat{Y} = a + bX$$

Linear regression

- Once we've estimated our line, what we usually care about most is the values of *a* and *b*
- We call *a* and *b* "regression coefficients"
- We typically care more about *b* (the slope) since it tells us something about how the two variables are related

Cheatsheet: Interpreting slope coefficients

- **Significance**: is relationship strong enough to be considered reliable?
  - If $p < .05$: relationship is significant; look at sign and size
  - If $p > .05$: results somewhat indeterminate; relationship could be caused by coincidence or by "chance."
- **Sign**: is relationship positive or negative?
  - Positive coefficient: when independent variable is bigger, dependent variable tends to be bigger
  - Negative coefficient: when independent variable is bigger, dependent variable tends to be smaller
- **Size**: how big is the effect?
  - If *b* is a positive coefficient, a one-unit increase in the independent variable (e.g., a 1-year increase in age) predicts a *b* -unit increase in the dependent variable
  - If *b* is negative, predicts a *b* -unit *decrease*

Linear regression

Example: Predicting university grades (`http://onlinestatbook.com/2/regression/intro.html`)

- DV: University grade point average (GPA; how good their grades are)
- IV: High school GPA
- How much better do I expect a student to do at university (in terms of GPA) if they had a high GPA in high school?

Linear regression

Example: Predicting university grades

- Type into Stata:

```
insheet using "https://onlinestatbook.com/2/case_studies/data/sat.txt", ///
   delimiter(" ")
twoway scatter univ_gpa high_gpa
corr univ_gpa high_gpa
```

- We can see there's a strong correlation (r=0.78), but this still doesn't tell us how much higher the university GPA is expected to be if high school GPA goes up by a point
- We still need to run a regression

Linear regression

Example: Predicting university grades

- Type into Stata:

```
reg univ_gpa high_gpa
twoway (scatter univ_gpa high_gpa) ///
    (lfit univ_gpa high_gpa)
```

- **Significance**: p=0.000, which is < .05, so we conclude **there's a reliable relationship between high school GPA and university GPA**
- **Sign**: The coefficient (.67) for *high_GPA* is positive, so **doing well in high school predicts doing well at university**
- **Size**: The coefficient for *high_GPA* is .67, so **a one-point increase in high school GPA predicts a 0.67-point increase in university GPA**

Linear regression

- Linear regression is a very powerful tool, so we have only scratched the surface with this brief explanation
- For example, the appendix shows how we can use linear regression to make predictions
- We can also have a regression with multiple independent variables; the interpretation of coefficients remains similar to what we've already seen, but there is more explanation in the appendix

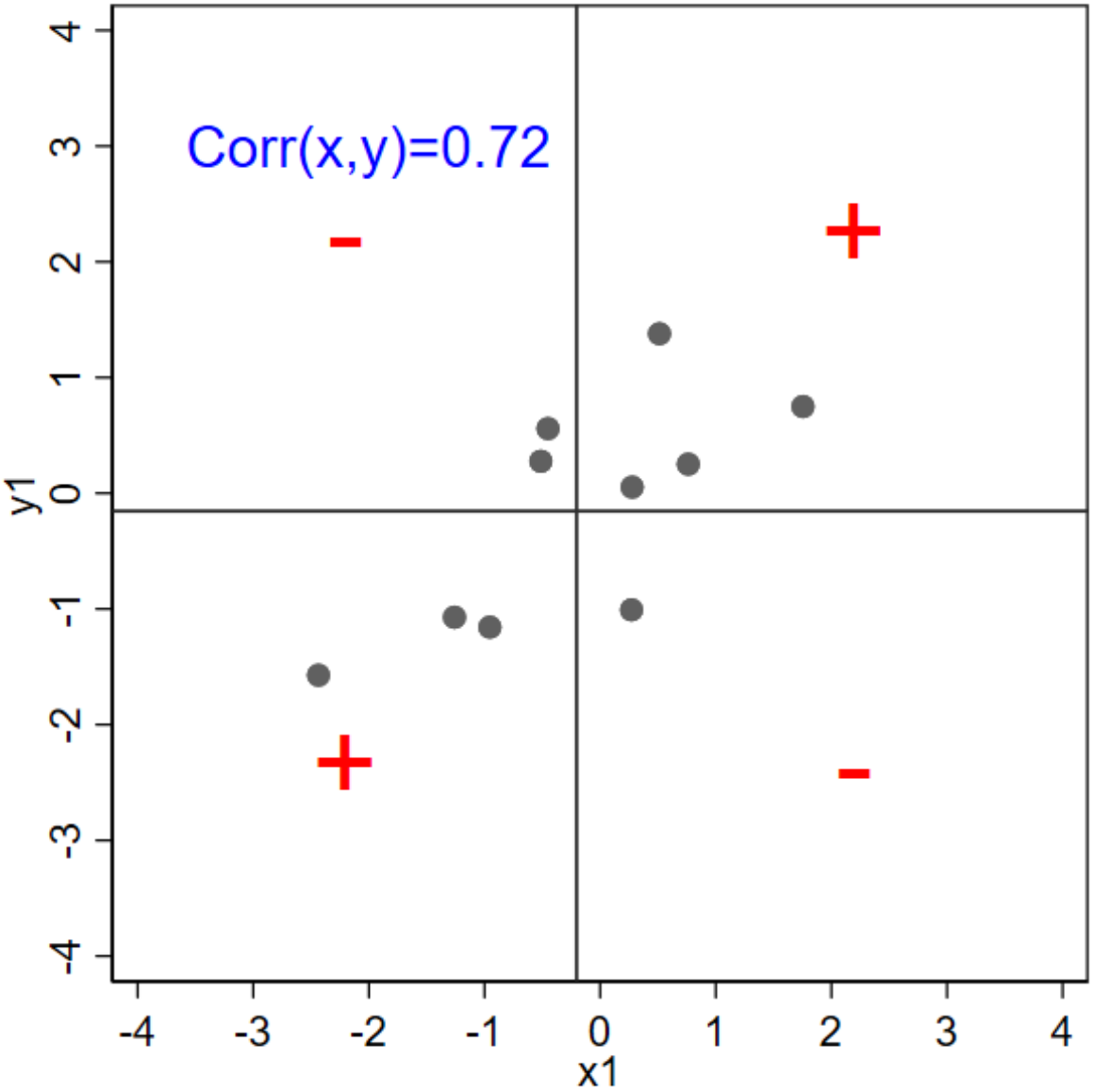
Correlation vs. regression

- Like the correlation coefficient, regression coefficients tell us something about the relationship between two variables
- But correlation and regression coefficients are interpreted differently
  - Correlation coefficient: tells us how closely the data fall along a straight line
  - Regression coefficients: tell us what the line looks like (*b* tells us how steep it is)
  - Both allow us to draw conclusions about whether two (continuous) variables are (linearly) associated with one another
  - The correlation coefficient and the regression slope coefficient will always have the same sign (positive or negative), but their magnitudes are usually different
  - Regression coefficients *aren't* bounded by -1 and 1

Appendix 1: Quadrants for correlation

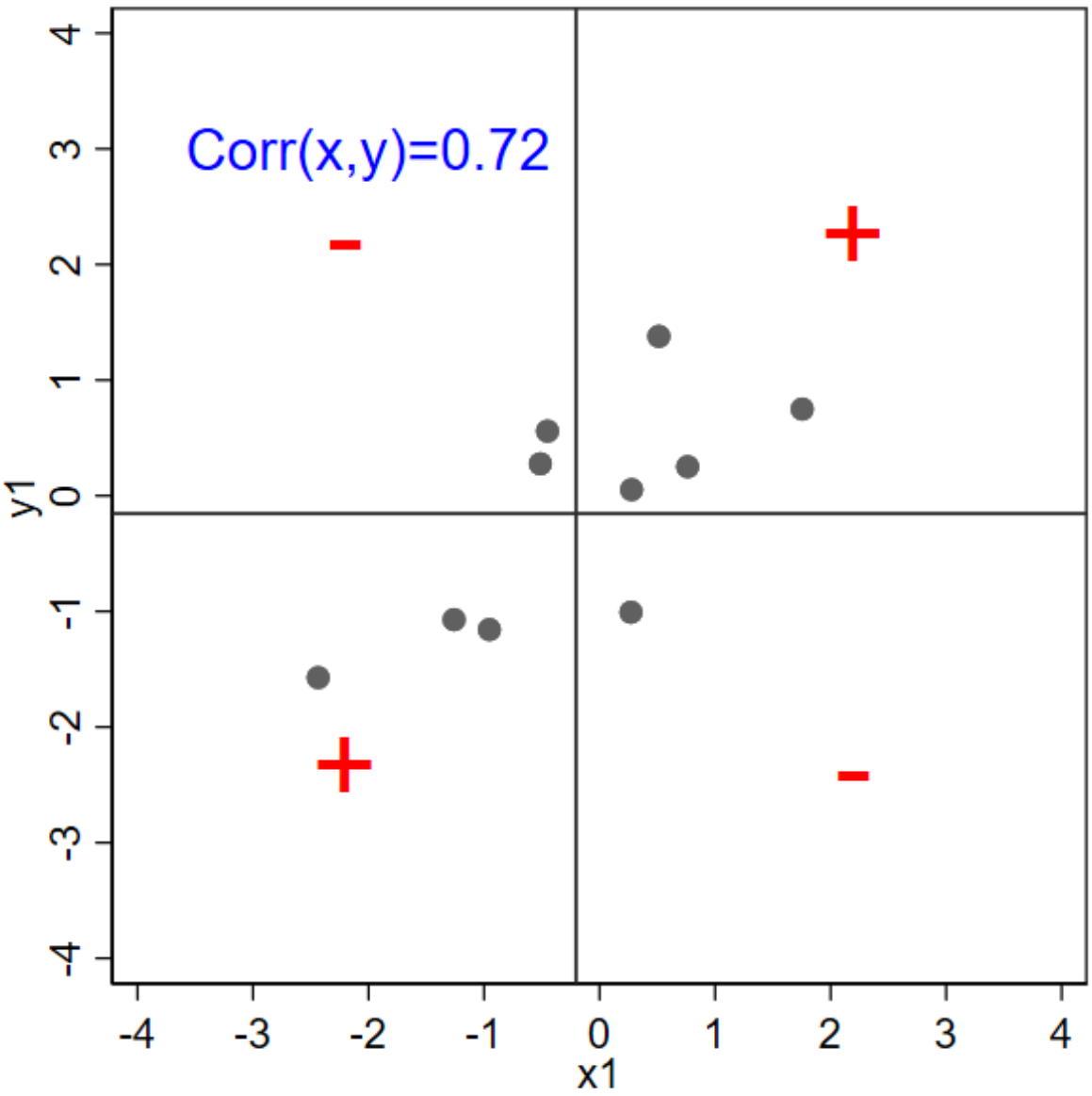
Return to main slides

Using lines for the means of x and y, we can divide the plot into 4 quadrants indicating whether each data point positively or negatively contributes to the correlation



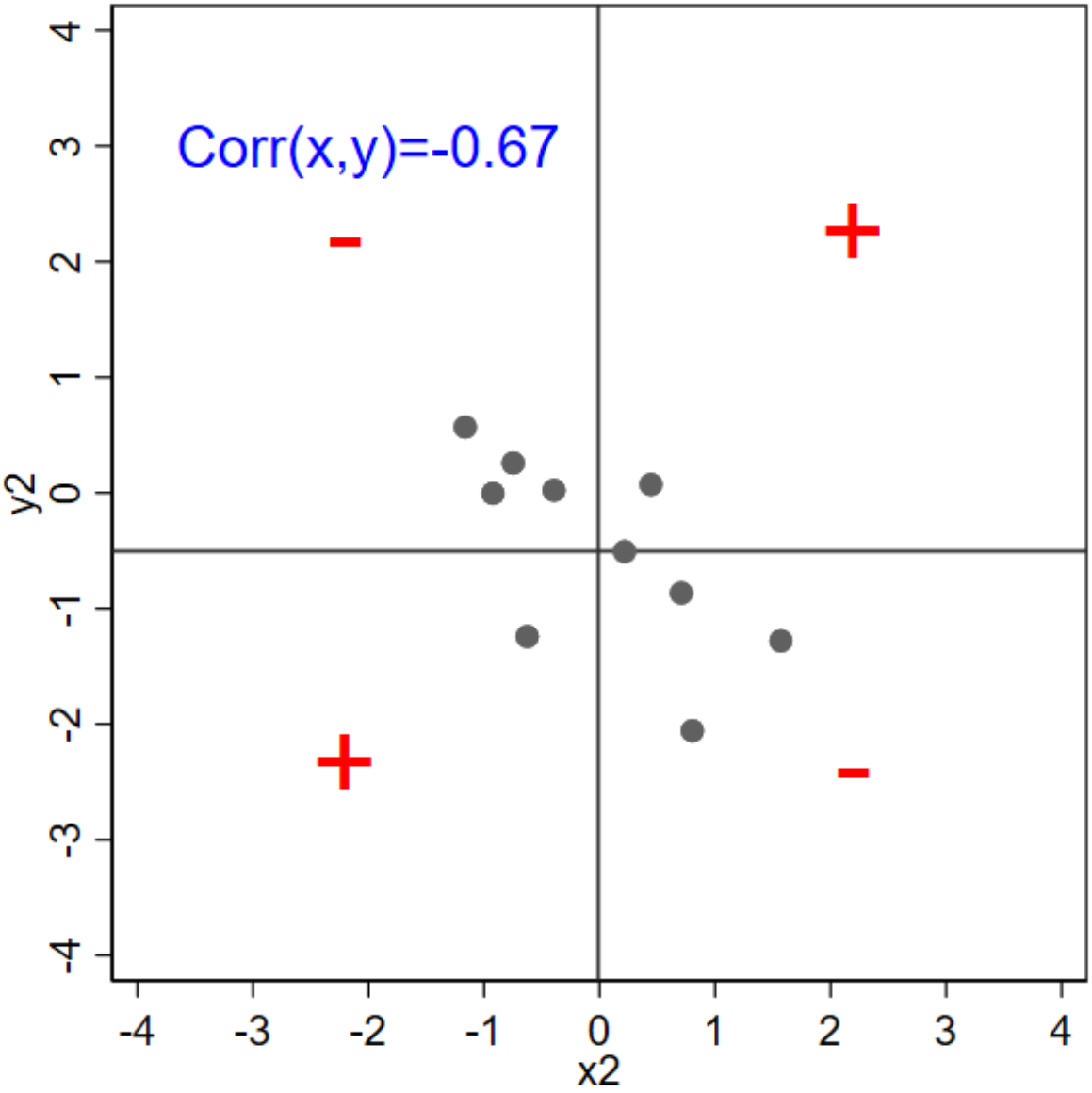
Appendix 1: Quadrants for correlation

If most data points are in the upper-right and lower-left quadrants, the correlation will be positive



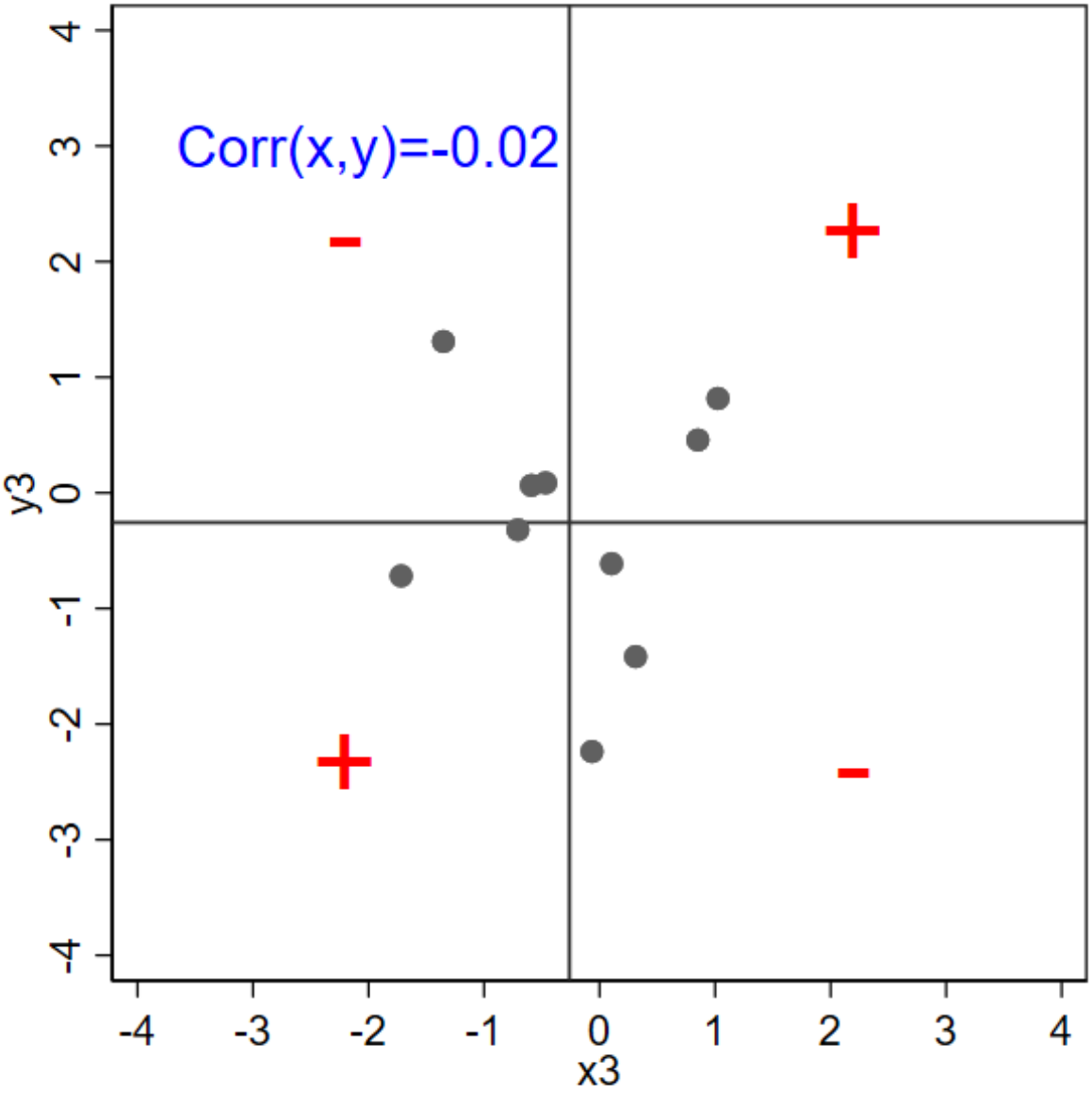
Appendix 1: Quadrants for correlation

If most data points are in the upper-left and lower-right quadrants, the correlation will be negative



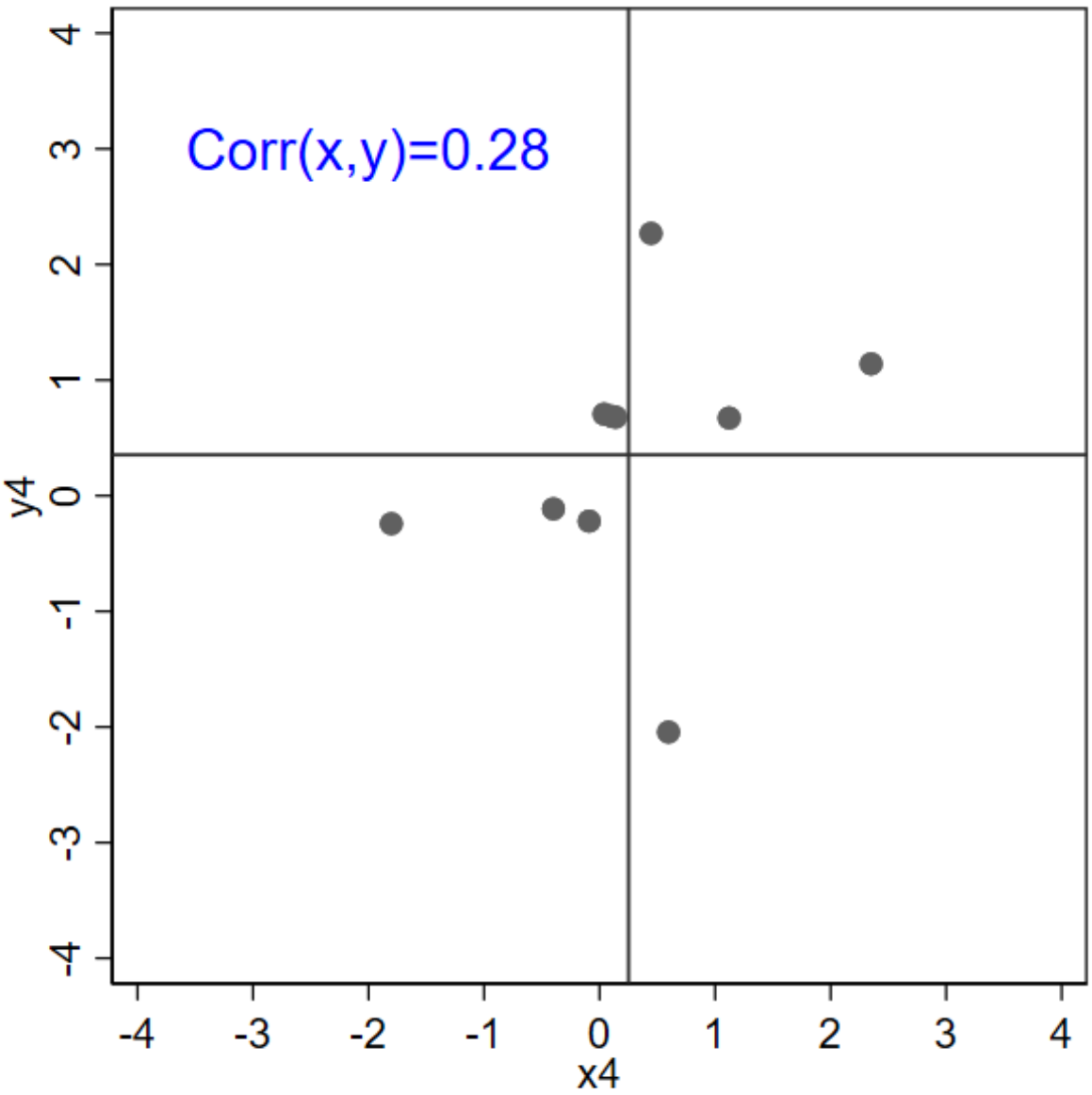
Appendix 1: Quadrants for correlation

If the data is spread fairly evenly among all quadrants, the correlation will be close to zero



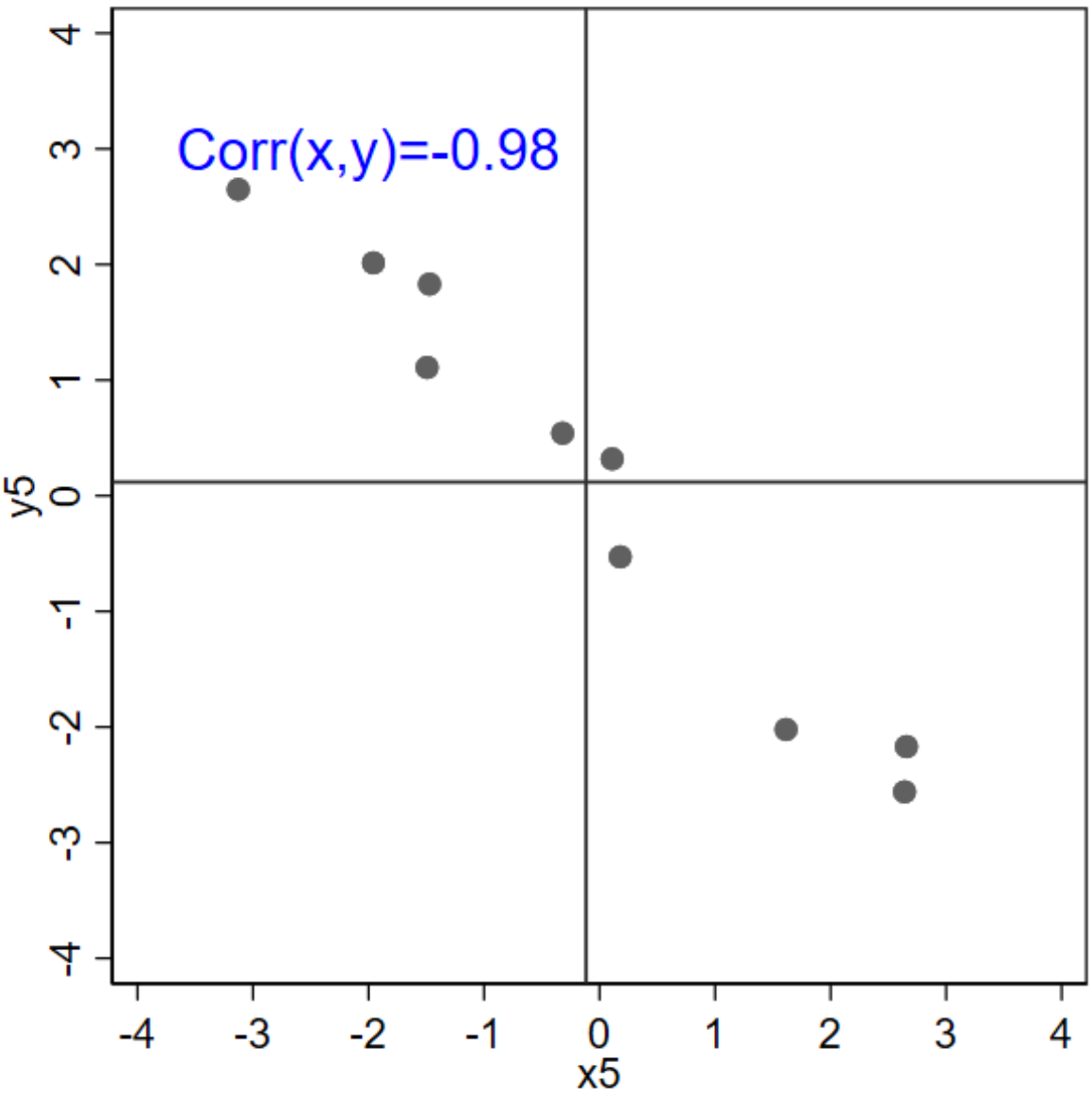
Appendix 1: Quadrants for correlation

Weaker correlations are harder to visually detect, but the quadrants can help



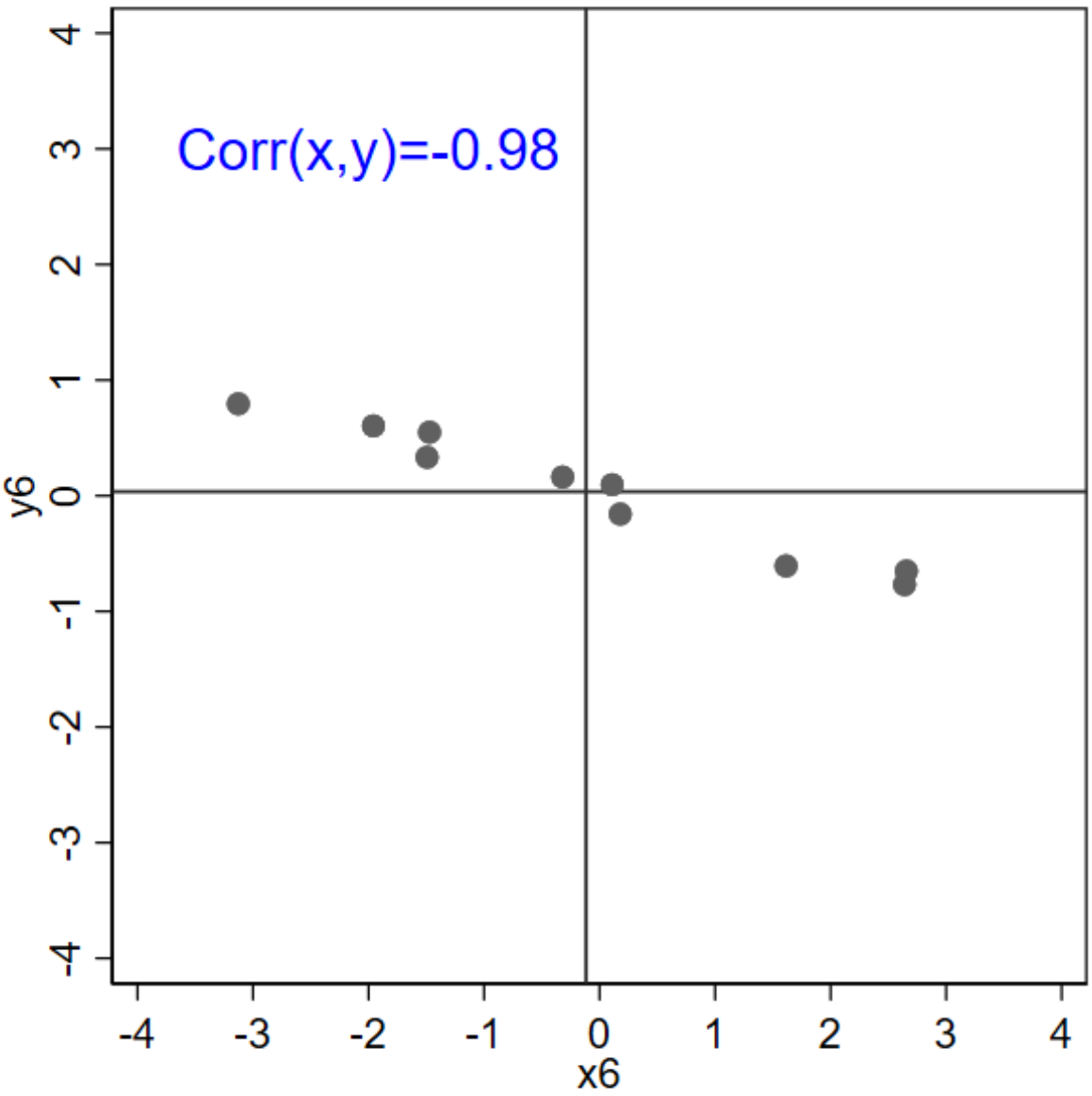
Appendix 1: Quadrants for correlation

The steepness of the line doesn't matter for correlation. Instead, focus on how closely the data fall along a line, or focus on the quadrants.



Appendix 1: Quadrants for correlation

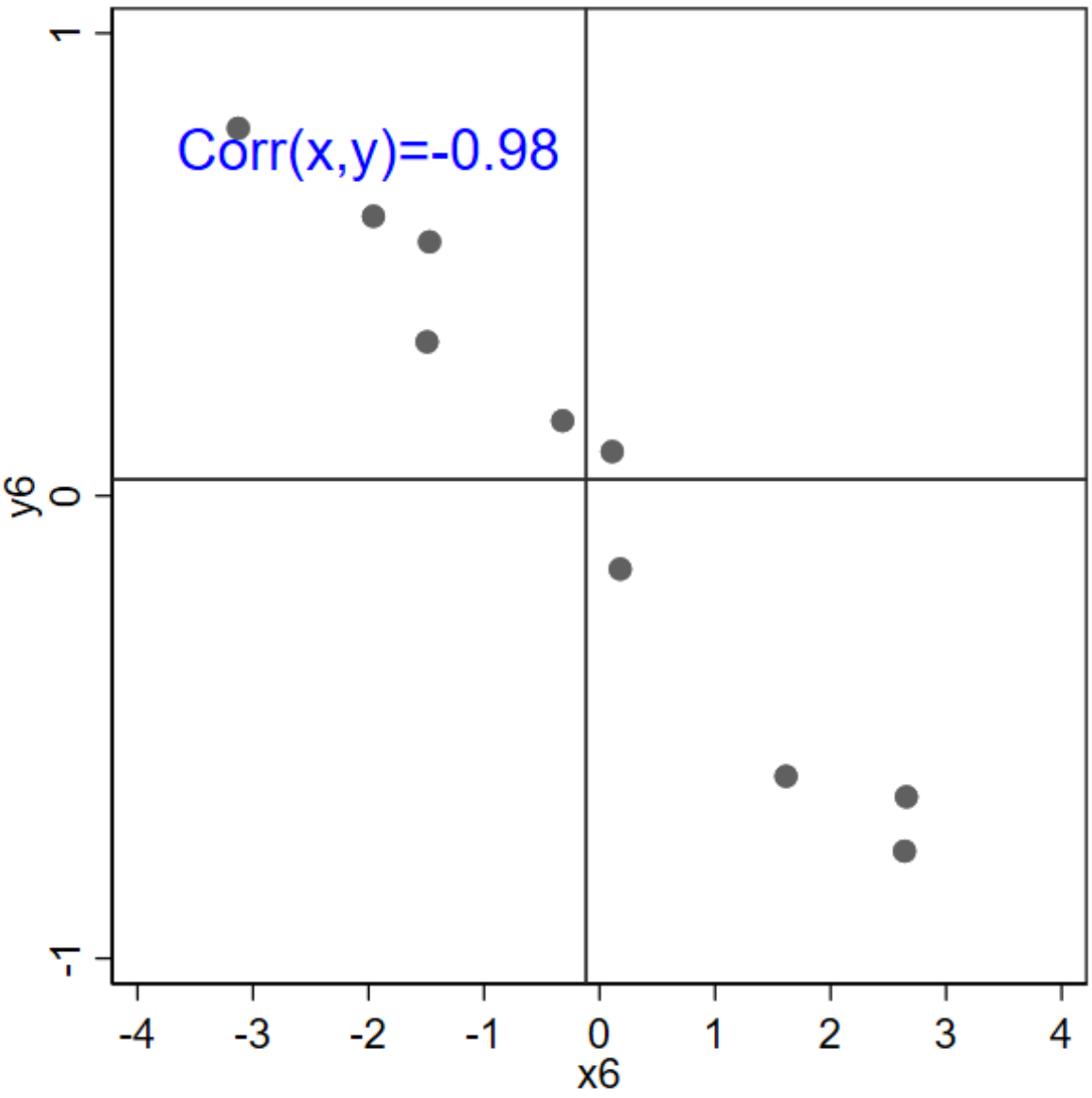
Notice the correlation hasn't changed though the line is now flatter



Appendix 1: Quadrants for correlation

In fact, how steep the line appears depends on how the axes are drawn; this is the exact same data from the prior slide with a different y axis

Return to main slides



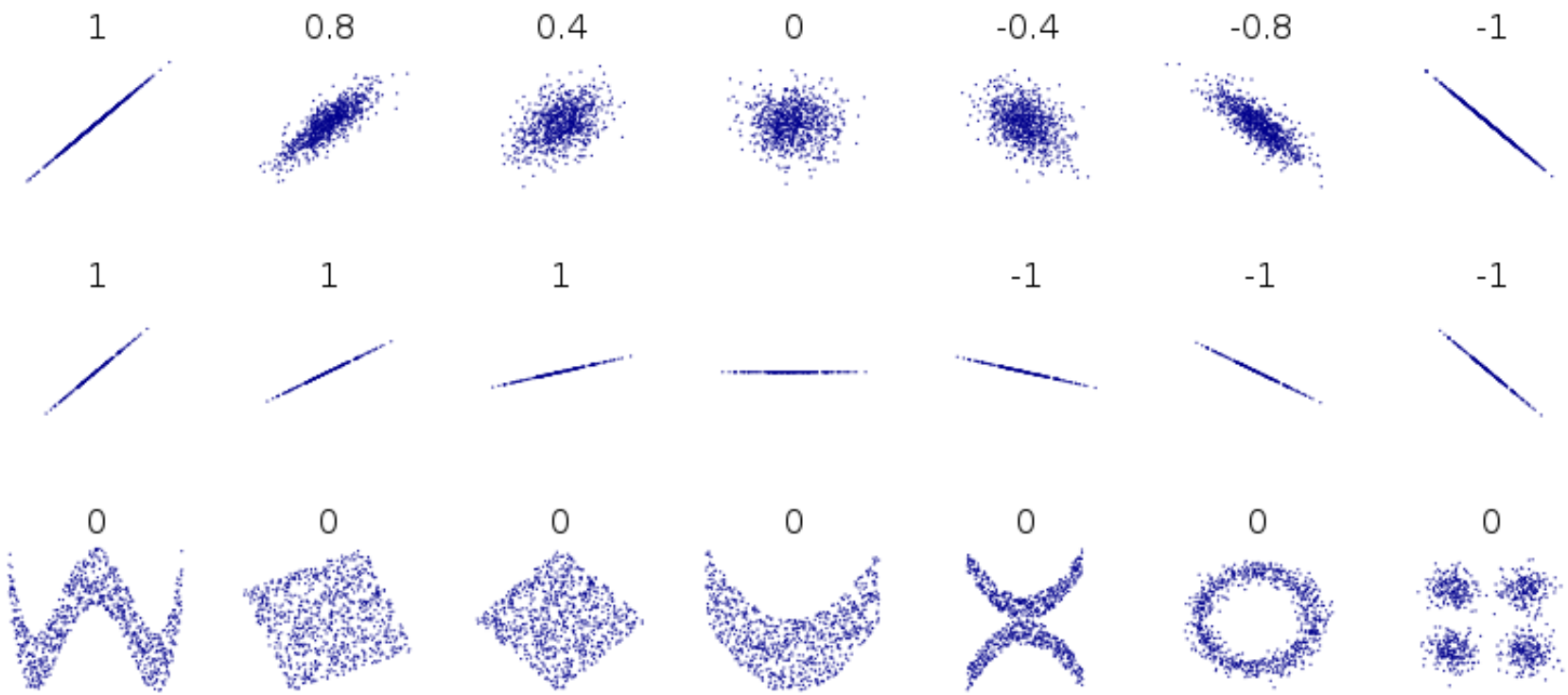
Appendix 2: Correlation, nonlinearity, etc.

Return to main slides

- There are multiple versions of the correlation formula which yield equivalent results (except that a population correlation differs from a sample correlation, but don't worry about that for now)
- The textbook version of the formula can be found by rearranging the version I just showed you (although the textbook notation is still a bit different; it will show X instead of x_i)
- Important note: If the data appears to follow a pattern but the pattern is not in the form of a straight line, there is a *nonlinear* relationship; nonlinear relationships may produce correlations of zero (or close to zero) even if the two variables are strongly related to one another

Appendix 2: Correlation, nonlinearity, etc.

Some more examples:



(Image: Public domain, Denis Boigelot)

Appendix 2: Correlation, nonlinearity, etc.

- There are some other types of correlation coefficients that are designed to handle limitations of the Pearson correlation coefficient
- Some are designed to handle data that isn't measured continuously (e.g., ordered responses to a question—"strongly disagree," "disagree," "agree," "strongly agree"—where we don't want to assume equal distance between each option)
- Some are able to detect certain types of nonlinear relationships
- If someone doesn't specify which correlation coefficient they're using, it's generally safe to assume they're referring to the Pearson correlation coefficient

Return to main slides

Appendix 3: The math behind correlation

Return to main slides

- Before examining correlation, we'll look at covariance since the formula is simpler:

$$Cov(x, y) = \sigma_{xy} = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N}$$

- μ_x represents the mean (average) of x while μ_y represents the mean of y
- Notice the similarity to the variance formula. In fact, the covariance of x with itself (covariance of x and x) is equal to the variance of x:

$$Cov(x, x) = \frac{\sum(x_i - \mu_x)(x_i - \mu_x)}{N} = \frac{\sum(x_i - \mu_x)^2}{N} = \sigma_x^2$$

Appendix 3: The math behind correlation

- Positive values of covariance indicate a “positive relationship” (higher values of x are associated with higher values of y)
- Negative values of covariance indicate a “negative relationship” (higher values of x are associated with lower values of y)
- A covariance of zero indicates no “linear” relationship between x and y (we’ll define “linear” later)

Appendix 3: The math behind correlation

$$Cov(x, y) = \sigma_{xy} = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N}$$

- Example:

i	x_i	y_i
1	5	5
2	2	4
3	2	0

Appendix 3: The math behind correlation

$$Cov(x, y) = \sigma_{xy} = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N}$$

- Example:

i	x_i	y_i	$\bar{x}$	$\bar{y}$
1	5	5	3	3
2	2	4	3	3
3	2	0	3	3

Appendix 3: The math behind correlation

$$Cov(x, y) = \sigma_{xy} = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N}$$

- Example:

i	x_i	y_i	$\bar{x}$	$\bar{y}$	$x_i - \bar{x}$	$y_i - \bar{y}$
1	5	5	3	3	2	2
2	2	4	3	3	-1	1
3	2	0	3	3	-1	-3

Appendix 3: The math behind correlation

$$Cov(x, y) = \sigma_{xy} = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N}$$

- Example:

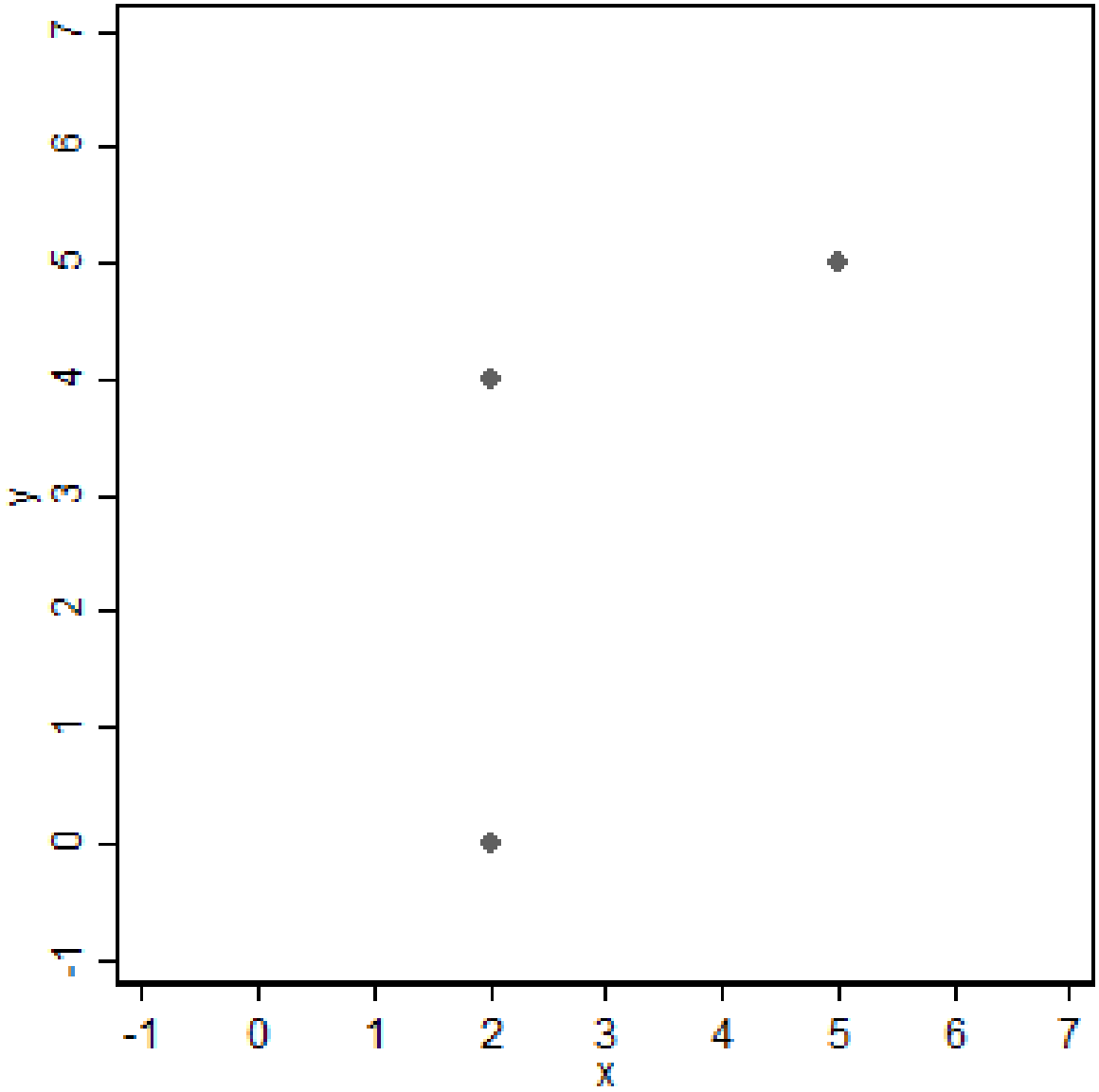
i	x_i	y_i	$\bar{x}$	$\bar{y}$	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	5	5	3	3	2	2	4
2	2	4	3	3	-1	1	-1
3	2	0	3	3	-1	-3	3

$$Cov(x, y) = \frac{4 - 1 + 3}{3} = \frac{6}{3} = 2$$

Appendix 3: The math behind correlation

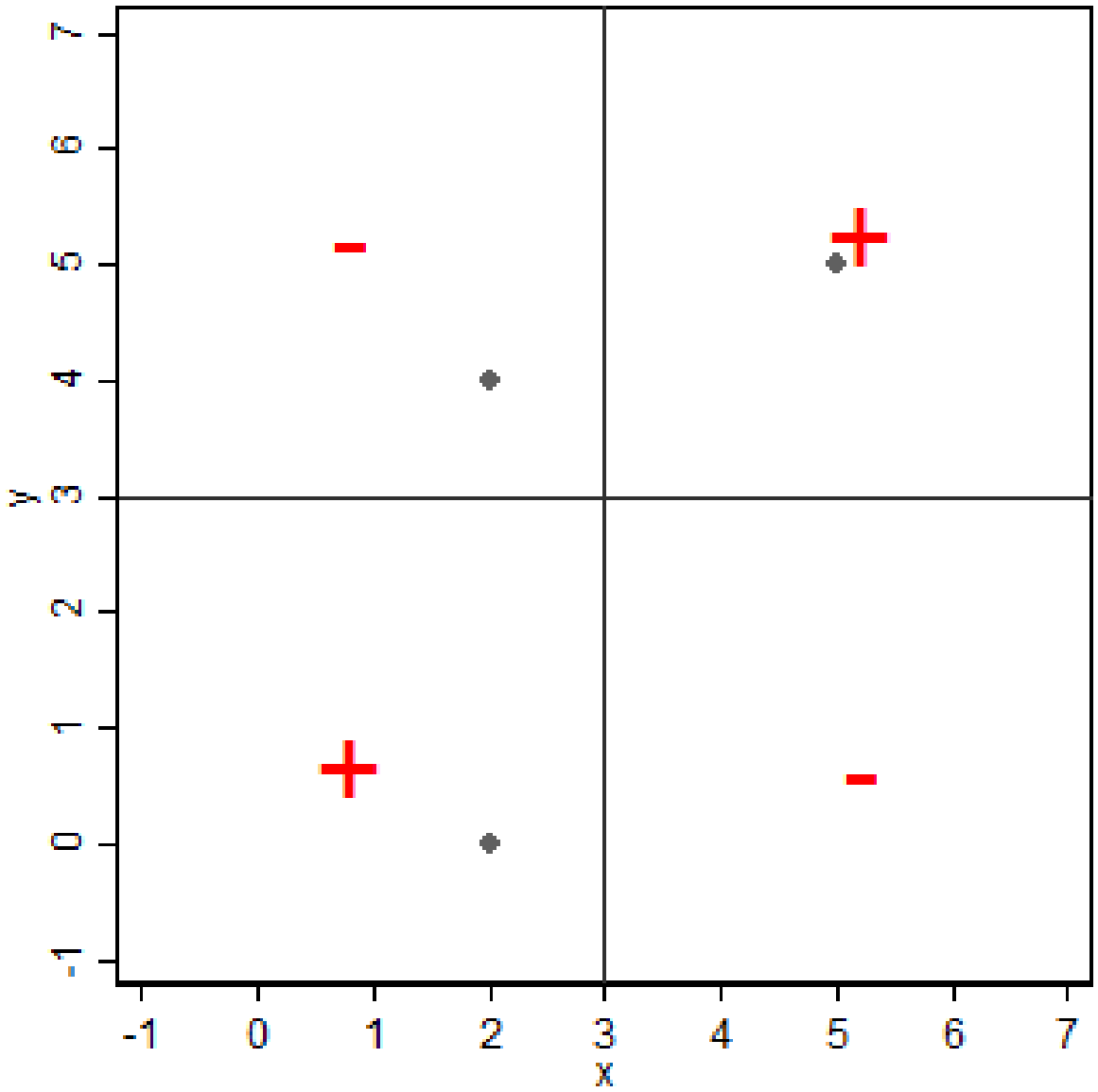
$$Cov(x, y) = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N}$$

i	x_i	y_i
1	5	5
2	2	4
3	2	0



Appendix 3: The math behind correlation

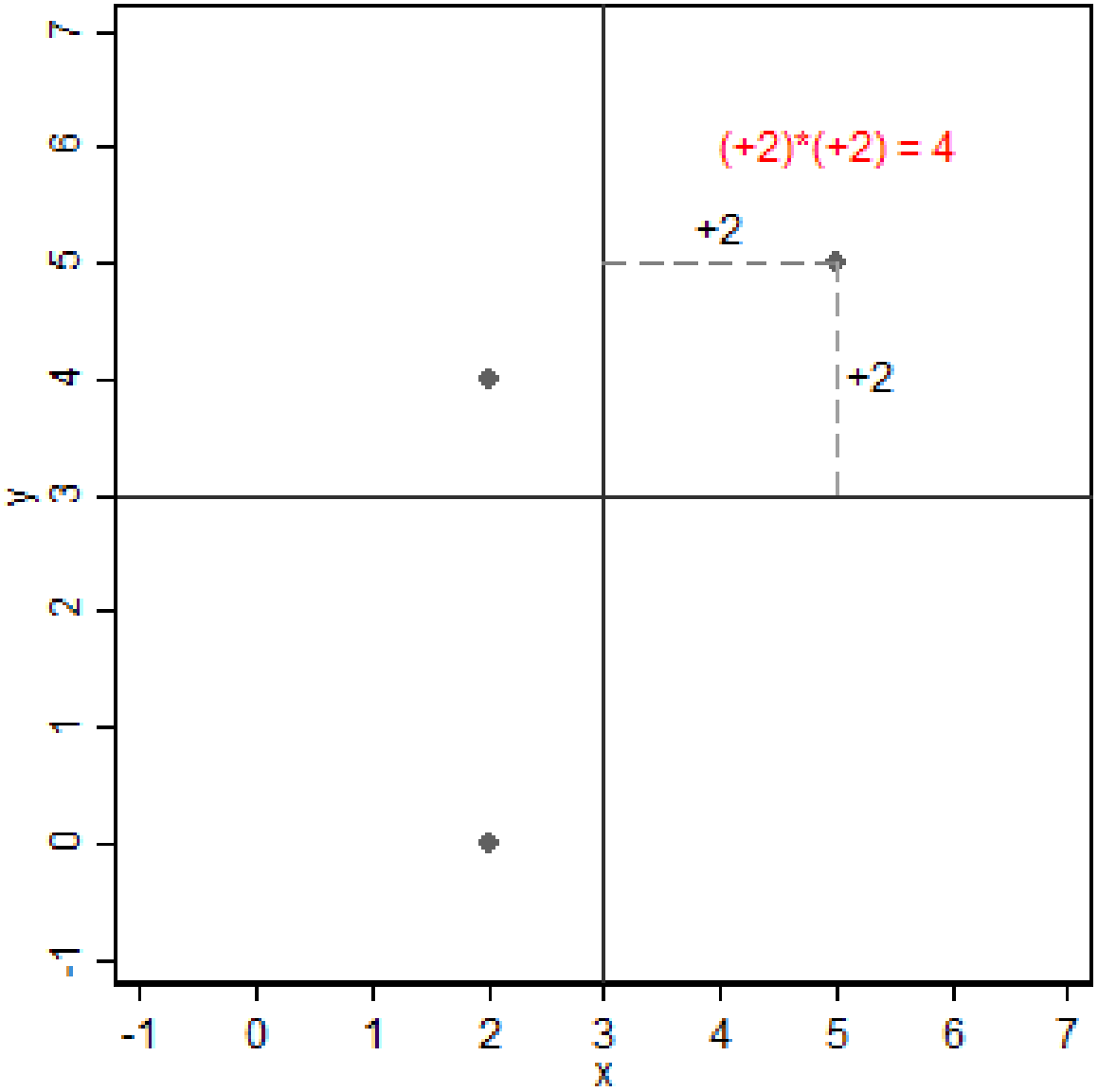
Using lines for $\bar{x}$ and $\bar{y}$ ($\bar{x} = 3; \bar{y} = 3$), we can divide the plot into 4 quadrants indicating whether data points positively or negatively contribute to covariance



Appendix 3: The math behind correlation

$$Cov(x, y) = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N}$$

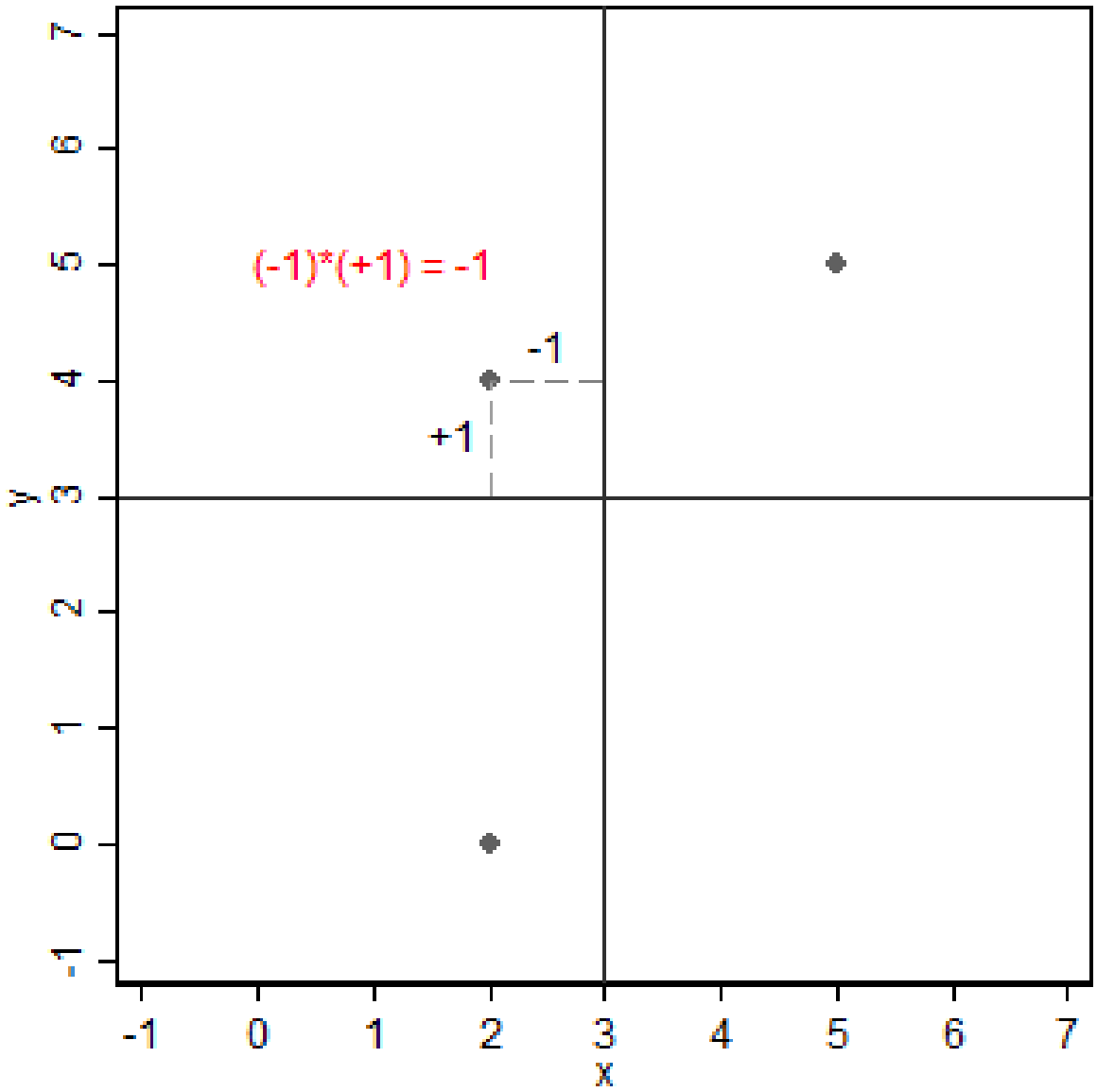
i	$(x_i - \bar{x})(y_i - \bar{y})$
1	4
2	-1
3	3



Appendix 3: The math behind correlation

$$Cov(x, y) = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N}$$

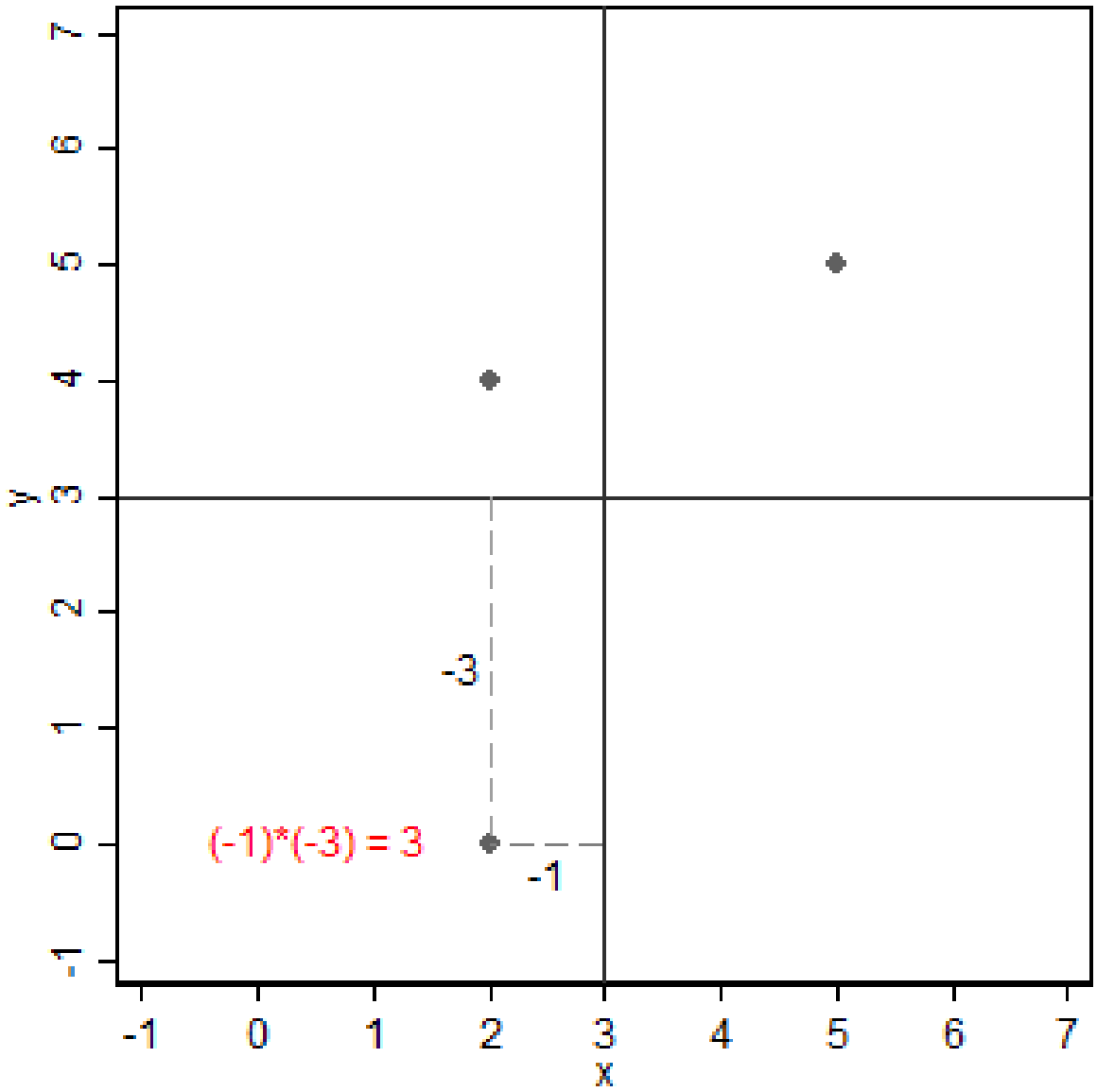
i	$(x_i - \bar{x})(y_i - \bar{y})$
1	4
2	-1
3	3



Appendix 3: The math behind correlation

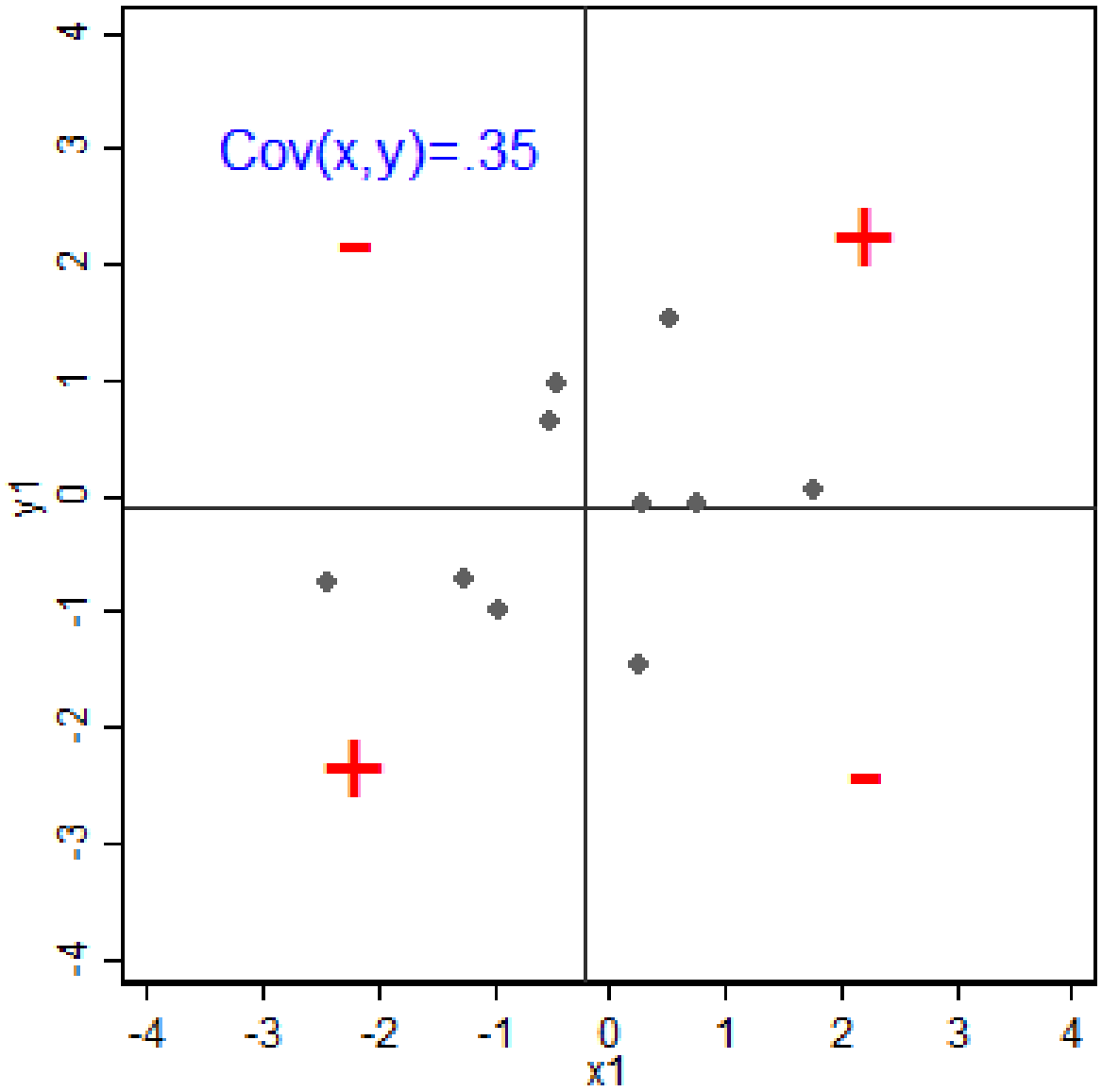
$$Cov(x, y) = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N}$$

i	$(x_i - \bar{x})(y_i - \bar{y})$
1	4
2	-1
3	3



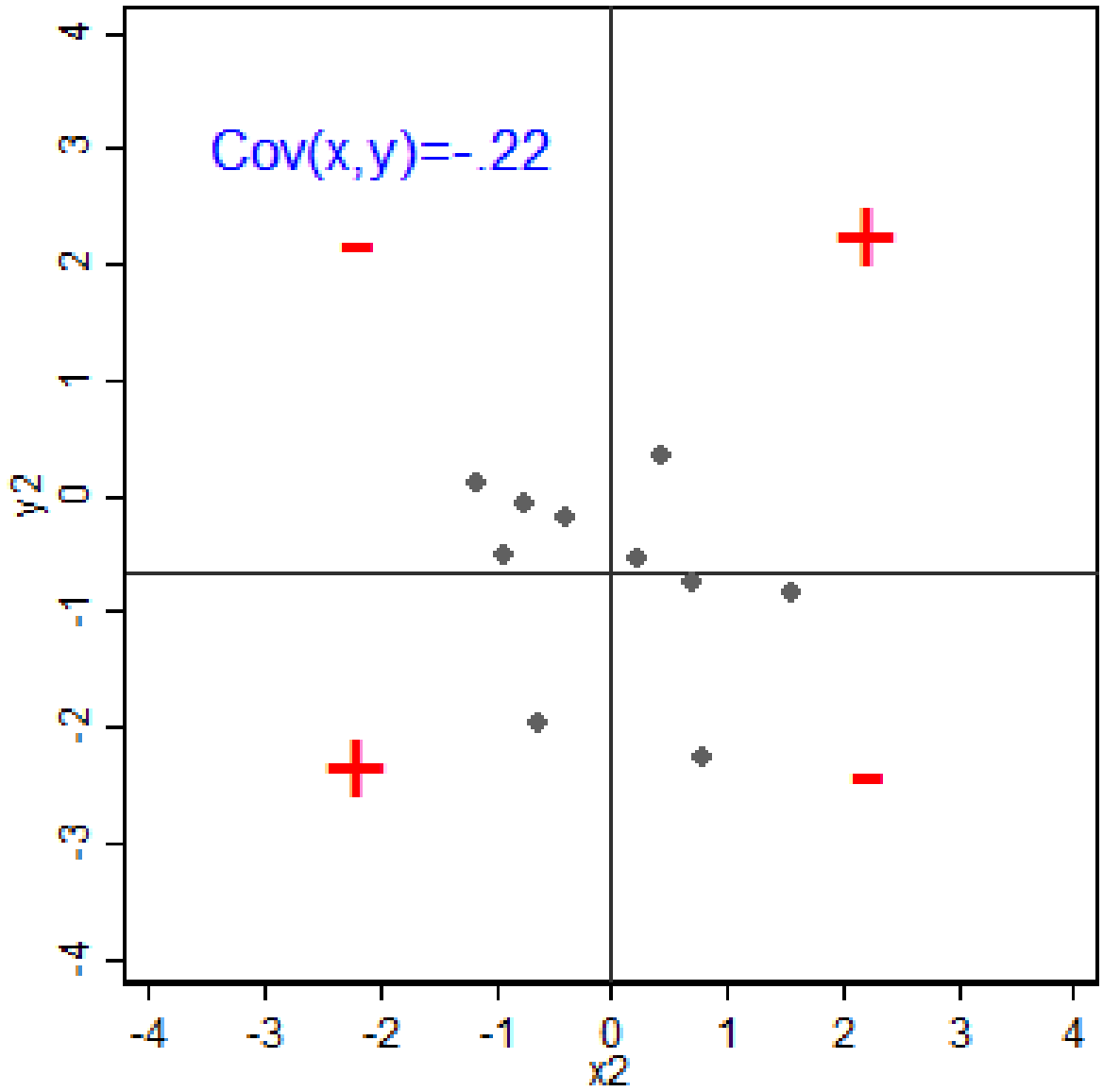
Appendix 3: The math behind correlation

If most data points are in the upper-right and lower-left quadrants, the covariance (and correlation) will be positive



Appendix 3: The math behind correlation

If most data points are in the upper-left and lower-right quadrants, the covariance (and correlation) will be negative



Appendix 3: The math behind correlation

- Correlation (more specifically, the Pearson correlation coefficient) is covariance scaled to the variance of the two variables
- Correlation is reported much more frequently than covariance since it can be used to compare variables measured in different units
- Correlations always fall between -1 and 1, with 0 indicating that the covariance is 0

$$r_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{Cov(x, y)}{\sqrt{Var(x)Var(y)}}$$

Return to main slides

Appendix 4: Using regression for prediction

Return to main slides

Example: Predicting university grades

- Continuing on with our regression example from the main slides...
- We can write the function for our estimated regression line by substituting our coefficient estimates into the regression model

$$\hat{Y} = a + bX$$

$$\widehat{college\_gpa} = 1.10 + 0.67 \times high\_gpa$$

Appendix 4: Using regression for prediction

Example: Predicting university grades

$$\widehat{college\_gpa} = 1.10 + 0.67 \times high\_gpa$$

- We can substitute in values to see the expected university GPA for different high school GPAs
- For someone with a high school GPA of 2.5 (between a B and C), we predict they'll have a GPA of 2.78 in college:

$$\widehat{college\_gpa} = 1.10 + 0.67 \times (2.5) = 2.78$$

- For someone with a high school GPA of 3.5 (between an A and B), we estimate they'll have a GPA of 3.45 in college:

$$\widehat{college\_gpa} = 1.10 + 0.67 \times (3.5) = 3.45$$

Return to main slides

Appendix 5: Multiple linear regression

Return to main slides

- If we have a second independent variable that we care about, we can rewrite the regression equation as follows:

$$\hat{Y} = a + b_1 X_1 + b_2 X_2$$

- Now, X_1 is our first independent variable and X_2 is our second independent variable
- Each variable has it's own coefficient: b_1 is the coefficient for X_1 and b_2 is the coefficient for X_2
- If we want to assume that X_1 and X_2 cause Y, we can say that b_1 is the effect of X_1 (on Y) and b_2 is the effect of X_2 (on Y)

Appendix 5: Multiple linear regression

Example: Predicting university grades

- Let's add another variable to our model: the student's SAT score
- Type into Stata:

```
gen sat = math_sat + verb_sat
twoway scatter univ_gpa sat
reg univ_gpa high_gpa sat
```

- We get:

$$\hat{Y} = a + b_1 X_1 + b_2 X_2$$

$$\hat{univ\_gpa} = 0.540 + 0.541 \times high\_gpa + 0.0008 \times sat$$

Appendix 5: Multiple linear regression

Example: Predicting university grades

- High school GPA
  - **Significance**: p=0.000, which is < .05, so we conclude **there's a reliable relationship between high school GPA and university GPA**
  - **Sign**: The coefficient (.54) for *high_GPA* is positive, so **doing well in high school predicts doing well at university**
  - **Size**: The coefficient for *high_GPA* is .54, so **a one-point increase in high school GPA predicts a 0.54-point increase in university GPA**
    - The effect size is smaller than what was estimated in the prior regression model because now we're effectively estimating the difference between two students with different high school GPAs but the same SAT score

Appendix 5: Multiple linear regression

Example: Predicting university grades

- SAT score
  - **Significance**: p=0.043, which is < .05, so we conclude **there's a reliable relationship between SAT score and university GPA**
  - **Sign**: The coefficient (.0008) for *sat* is positive, so **doing well on the SAT predicts doing well at university**
  - **Size**: The coefficient for *sat* is .0008, so **a one-point increase in SAT score predicts a 0.0008-point increase in university GPA**
    - This effect sounds tiny, but SAT scores range from 1034 to 1450 in this sample, so a one-point increase in SAT score is almost nothing
    - Since a one-point increase on the SAT predicts a 0.0008-point increase in university GPA, **a 200-point increase in SAT score predicts a 0.16-point increase in university GPA** ($200 \times .0008 = .16$)

Appendix 5: Multiple linear regression

- When we include multiple independent variables in the same regression:
  - We're asking: what is the effect of changing one factor at a time while holding all other factors constant?
  - The coefficient for X_1 tells me how much I expect Y to increase if X_1 goes up by one unit but all other independent variables (e.g., X_2) are left unchanged
  - Or, we can think about comparing two people (units) that are identical on all independent variables except there is a one-unit difference in X_1; the X_1 coefficient tells us how different we expect those two people to be

Appendix 5: Multiple linear regression

$$\hat{univ\_gpa} = 0.540 + 0.541 \times high\_gpa + 0.0008 \times sat$$

- Imagine two students, both with the same SAT score (1200) but different high school GPAs (2.5 and 3.5)
  - Student with 2.5 GPA in high school:

    $$\hat{univ\_gpa} = 0.540 + 0.541 \times (2.5) + 0.0008 \times (1200) = 2.853$$

  - Student with 3.5 GPA in high school:

    $$\hat{univ\_gpa} = 0.540 + 0.541 \times (3.5) + 0.0008 \times (1200) = 3.394$$

  - The student with the high school GPA that is one point higher is expected to have a university GPA that is 0.541 points higher ($3.394 - 2.853 = 0.541$)
  - Notice that this difference (0.541) is exactly equal to the *high_gpa* coefficient

Appendix 5: Multiple linear regression

$$\hat{univ\_gpa} = 0.540 + 0.541 \times high\_gpa + 0.0008 \times sat$$

- The same approach also applies to the interpretation of the *sat* coefficient
- We can imagine two students with the same high school GPA (3.0) and different SAT scores (1200 vs 1201)
  - Student with 1200 on SAT:
    $\hat{univ\_gpa} = 0.540 + 0.541 \times (3) + 0.0008 \times (1200) = 3.1230$
  - Student with 1201 on SAT:
    $\hat{univ\_gpa} = 0.540 + 0.541 \times (3) + 0.0008 \times (1201) = 3.1238$
  - The student with the SAT score that is one point higher is expected to have a university GPA that is 0.0008 points higher ($3.1238 - 3.1230 = 0.0008$)

Return to main slides