Week 6: Correlation

POP88162 Introduction to Quantitative Research Methods

Tom Paskhalis

Department of Political Science, Trinity College Dublin

So Far

- *Sampling distributions* of population distributions of any shape approximate *normal distributions*.
- *Confidence intervals* describe how certain we are about the point estimates of population parameters.
- For *hypothesis testing* we formulate *null* and *alternative hypotheses*.
- We collect the data and calculate the test statistic to try to reject the null hypothesis.
- *P-value* based on calculated test statistic describes the weight of evidence against the null.
- χ^2 and t-test are examples of statistical tests for cases where our independent variable is categorical.

Topics for Today

- Scatterplot
- Logarithmic transformation
- Covariance
- Correlation
- Slope of the Regression Line

Today's Plan


Two Variables
Scatterplot
Covariance
Correlation
Regression Coefficient

Review: Significance Test

- **Significance test** is used to summarize the evidence about a hypothesis.
- It does so by comparing the our estimates with those predicted by a null hypothesis.
- 5 components of a significance test:
  - **Assumptions**: scale of measurement, randomization, population distribution, sample size
  - **Hypotheses**: null H_0 and alternative H_a hypothesis
  - **Test statistic**: compares estimate to those under H_0
  - **P-value**: weight of evidence against H_0, smaller P indicate stronger evidence
  - **Conclusion**: decision to *reject* or *fail to reject* H_0.

Review: Error Types

<table>
<tr><td colspan="2" rowspan="2"></td><th colspan="2">Null hypothesis (H_0) is</th></tr>
<tr><th>True</th><th>False</th></tr>
<tr><th rowspan="2">Decision about Null hypothesis (H_0)</th><th>Don’t reject</th><td>Correct inference
(true negative)

(probability = $1-\alpha$)</td><td>Type II Error
(false negative)

(probability = β)</td></tr>
<tr><th>Reject</th><td>Type I Error
(false positive)

(probability = α)</td><td>Correct inference
(true positive)

(probability = $1-\beta$)</td></tr>
</table>

Review: Statistical Tests

		Dependent Variable	
		Nominal/ Ordinal	**Interval**
Independent Variable	**Nominal/ Ordinal**	χ^2 (chi-squared) test	Mean comparison test
	Interval	Logistic Regression	Linear Regression

Two Random Variables

Moving Beyond Categorical X

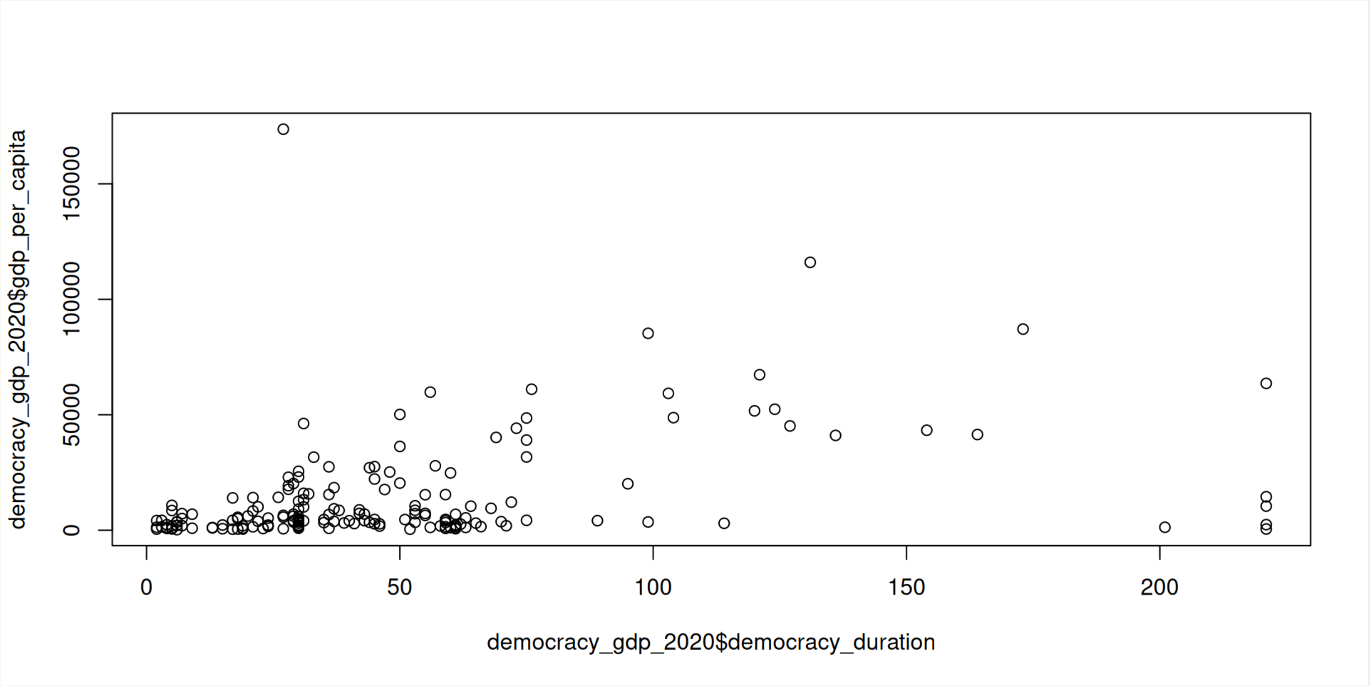
- So far, statistical tests for data, where our independent variable X is categorical (nominal/ordinal).
- But how can we measure and test the levels of association between quantitative variables?
- Three principal ways of measuring such association are:
  - **Covariance**
  - **Correlation coefficient**
  - **Slope of a regression line**
- At their core all these measures assume that 2 variables are related <u>linearly</u>.

Visualising Quantitative Data: Scatterplot

- Graphical plot which shows 2 quantitative (interval) variables alongside each other is called a **scatterplot**.
- It can be thought of as analogous to contingency table but for quantitative variables.
- It is an excellent tool for exploring bivariate (i.e. between 2 variables) relationships in quantitative data.
- We can think of variable on *x-axis* as our X and variable on *y-axis* as our Y.
- Many relationships between quantitative variables can be revealed just by plotting them against each other.

Example: Regime Longevity and GDP in 2020

```r
democracy_gdp_2020 <- read.csv("../data/democracy_gdp_2020.csv")
plot(democracy_gdp_2020$democracy_duration, democracy_gdp_2020$gdp_per_capita)
```



Source

World Bank, Boix, Miller and Rosato (2013), (2020)

Logarithmic Transformation

- **Logarithmic transformation** offers a convenient way of dealing with highly-skewed variables.
- In addition it offers a way to model non-linear relationships between two variables.
- They serve a lot of other important purposes in mathematics ans statistics, but we will focus on these two use cases for data analysis.

Extra

Benoit (2011)

Quick Primer on Logarithms

- Logarithm is the power to which the *base* should be raised to equal a given number.
- Or, to put it mathematically:

$$a^b = x$$
$$log_a x = b$$

- $log_a x$ is pronounced as "the logarithm of x to base a".
- We can think of a logarithm $log_a x$ as an inverse of exponentiation a^b.
- Some common bases include 2, 10 and e (2.71828...).

Logarithms in R

- We can use `^` operator in R to raise a given number to any power (exponentiate).
- For logarithms with base *e* (*natural logarithms*) we can use `exp()` function.
- To get use function `log()` to calculate a logarithm given a number and a base.
- *Natural logarithm* (with base *e*) is the default and most commonly used base.

```
exp(b)
log(x, base = exp(1))
2 ^ b
log2(x)
10 ^ b
log10(x)
```

Logarithms: Example

5^2

```r
5 ^ 2
```

```
[1] 25
```

$log_5 25$

```r
log(25, base = 5)
```

```
[1] 2
```

10^{-3}

```r
10 ^ -3
```

```
[1] 0.001
```

$log_{10} 0.001$

```r
log10(0.001)
```

```
[1] -3
```

Logarithms Examples Continued

Natural logarithm: $log_e 0.001$, where $e \approx 2.71828$

```r
log(0.001)
```

```
[1] -6.907755
```

e^5

```r
exp(5)
```

```
[1] 148.4132
```

Rate of change as the original variable is multiplied by 10:

```r
log(148)
```

```
[1] 4.997212
```

```r
log(1480)
```

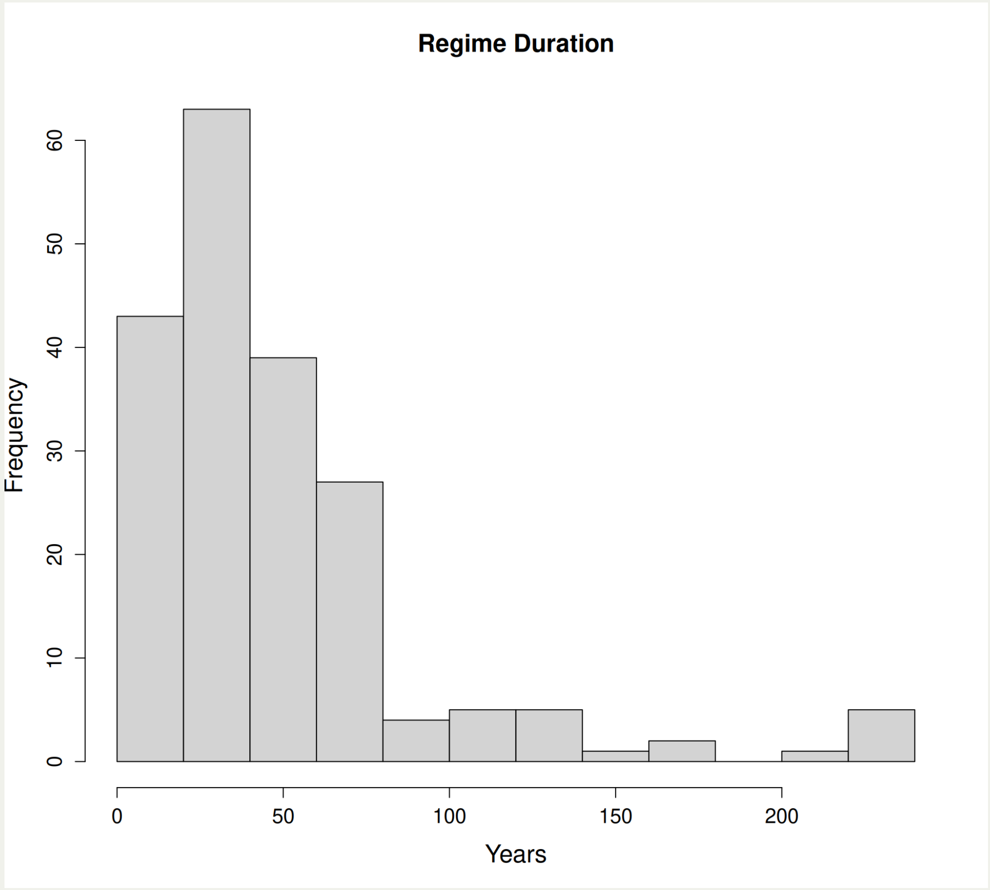
```
[1] 7.299797
```

```r
log(14800)
```

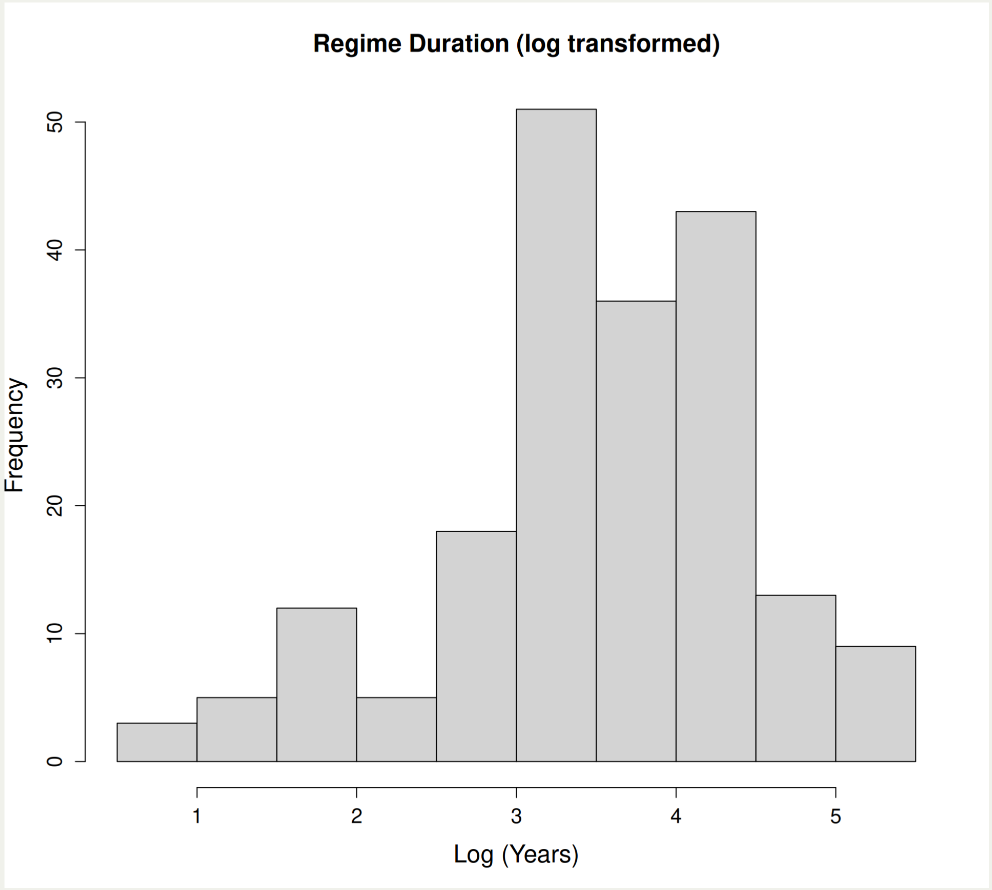
```
[1] 9.602382
```

Example: Log Transformation

Plot Code

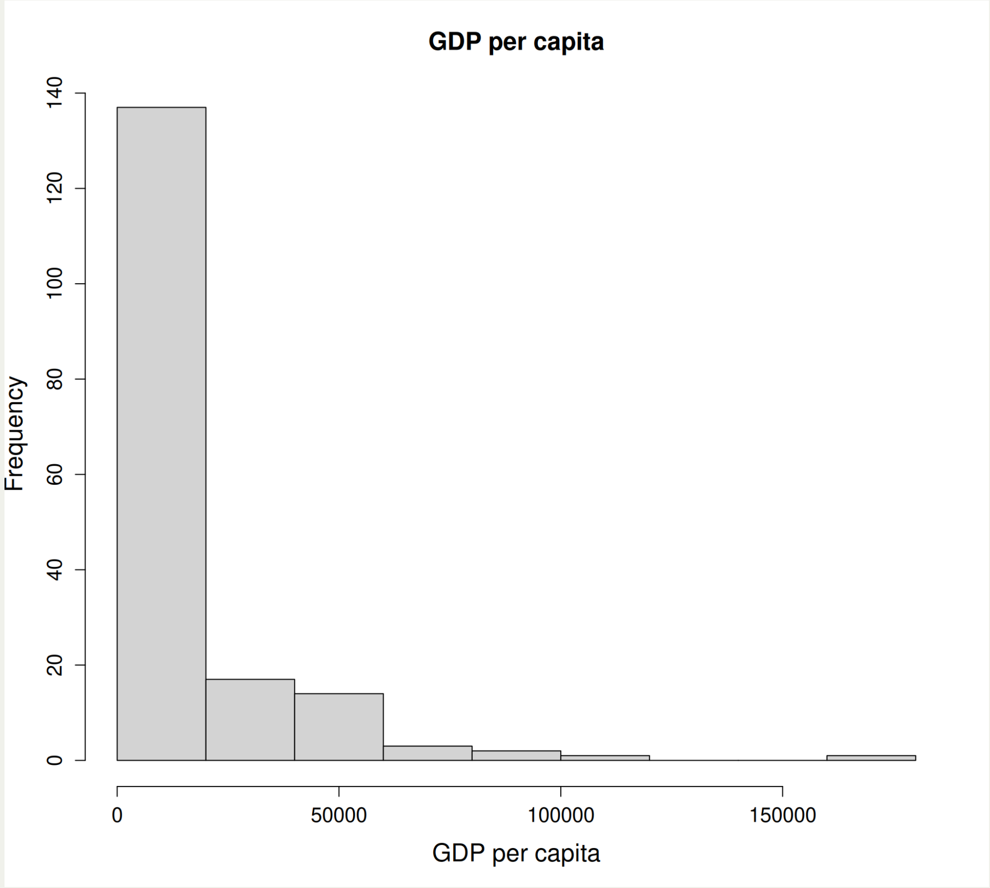


Plot Code

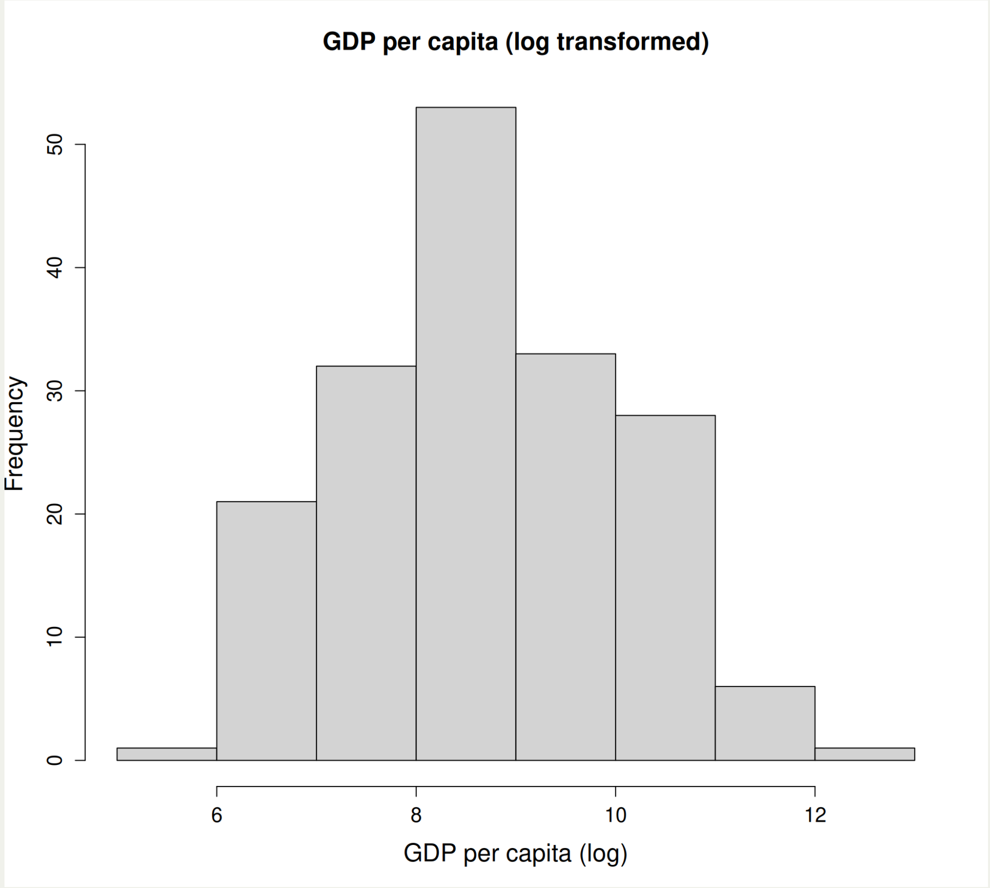


Example: Log Transformation

Plot | Code

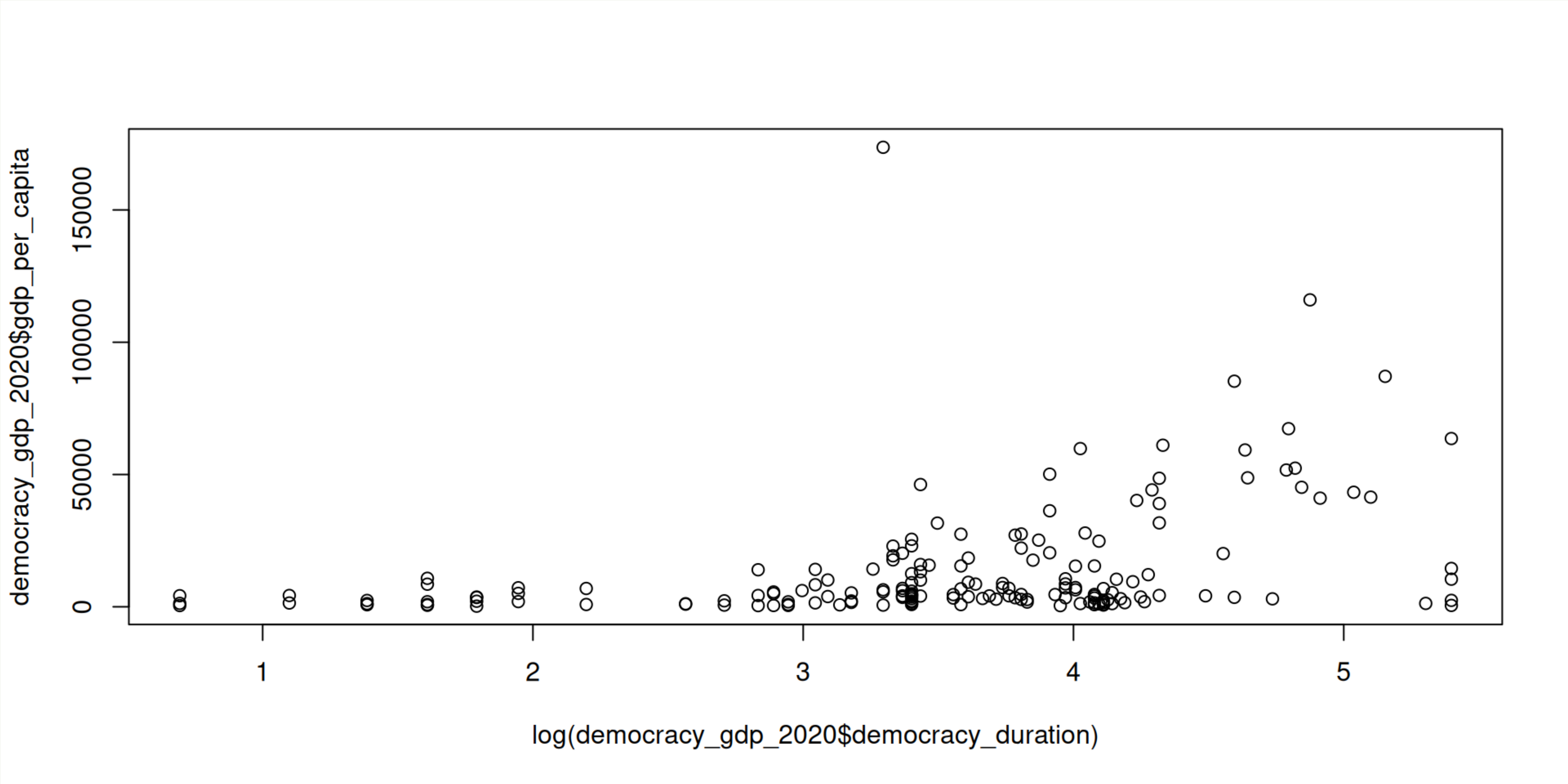


Plot | Code



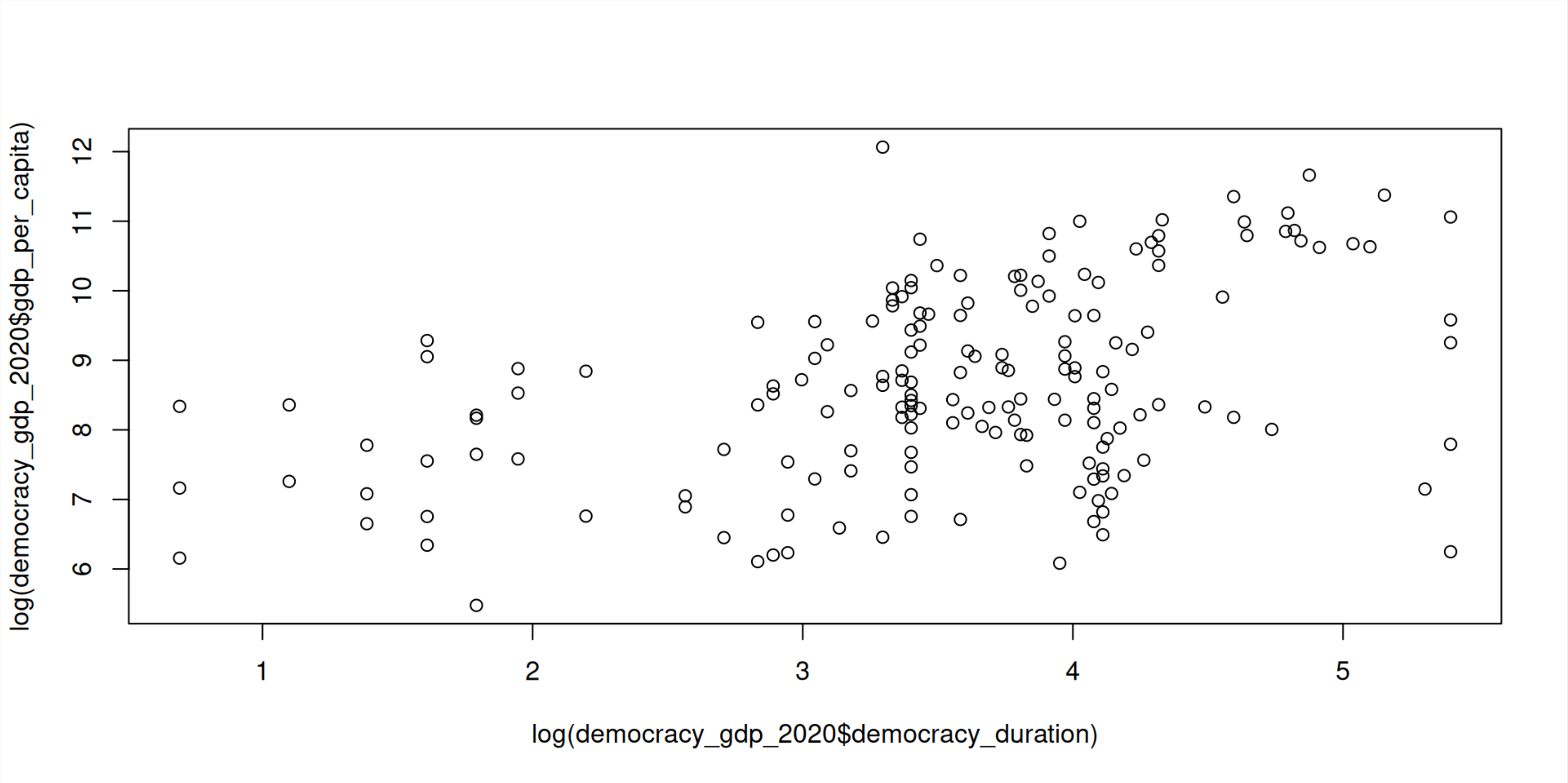
Scatterplot and Log Transformation

```r
plot(
  log(democracy_gdp_2020$democracy_duration),
  democracy_gdp_2020$gdp_per_capita
)
```



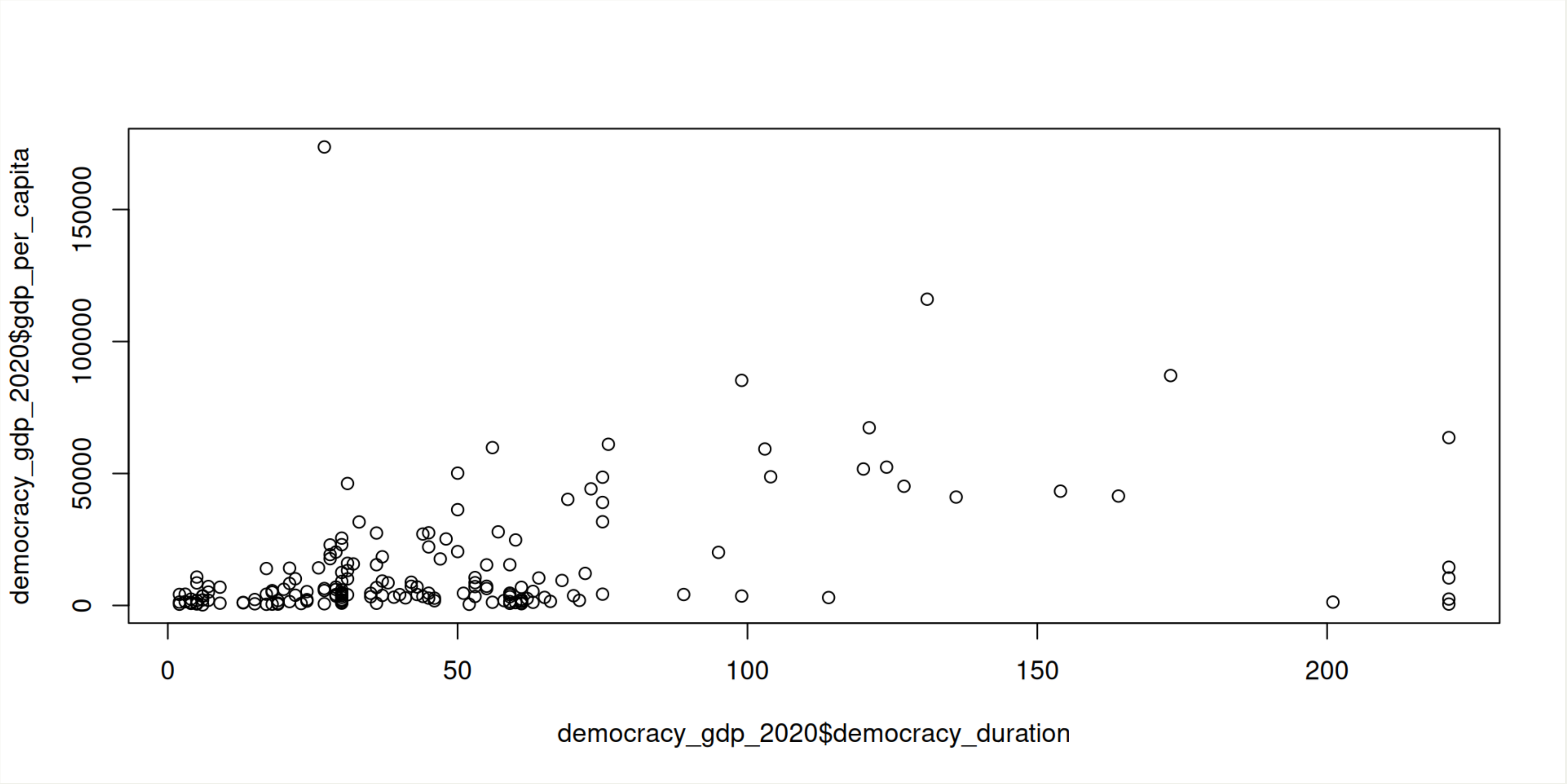
Scatterplot and Log Transformation

```r
plot(
  log(democracy_gdp_2020$democracy_duration),
  log(democracy_gdp_2020$gdp_per_capita)
)
```



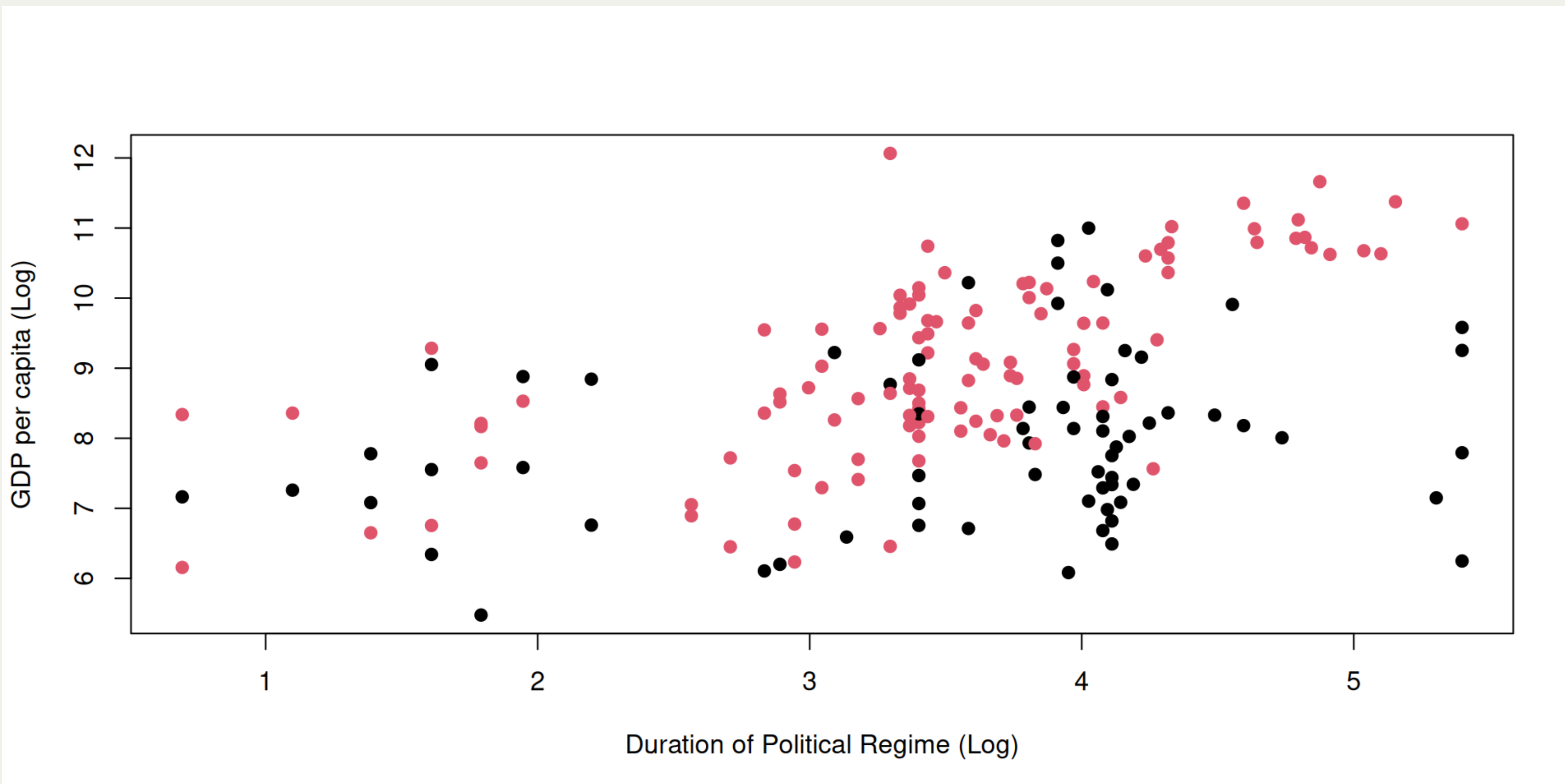
Compare To Before

```
plot(
  democracy_gdp_2020$democracy_duration,
  democracy_gdp_2020$gdp_per_capita
)
```



Prettifying Scatterplot

```
plot(
  log(democracy_gdp_2020$democracy_duration),
  log(democracy_gdp_2020$gdp_per_capita),
  xlab = "Duration of Political Regime (Log)", ylab = "GDP per capita (Log)
  pch = 19, col = democracy_gdp_2020$democracy + 1 # To avoid white colour
)
```



Correlation

Covariance

- Recall that the sum of all the deviations for one variable Y looks like this:

$$\sum_{i=1}^{n} Y_i - \bar{Y}$$

- Then, **covariance** is the average of the product of deviations of two quantitative variables:

$$cov(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

- Note that if each larger-than-average X_i co-occurs with larger-than-average Y_i, and vice versa the sum will be positive, indicating <u>positive association</u>.
- If, on the contrary, each if each larger-than-average X_i co-occurs with smaller-than-average Y_i, and vice versa the sum will be negative, indicating <u>negative association</u>.

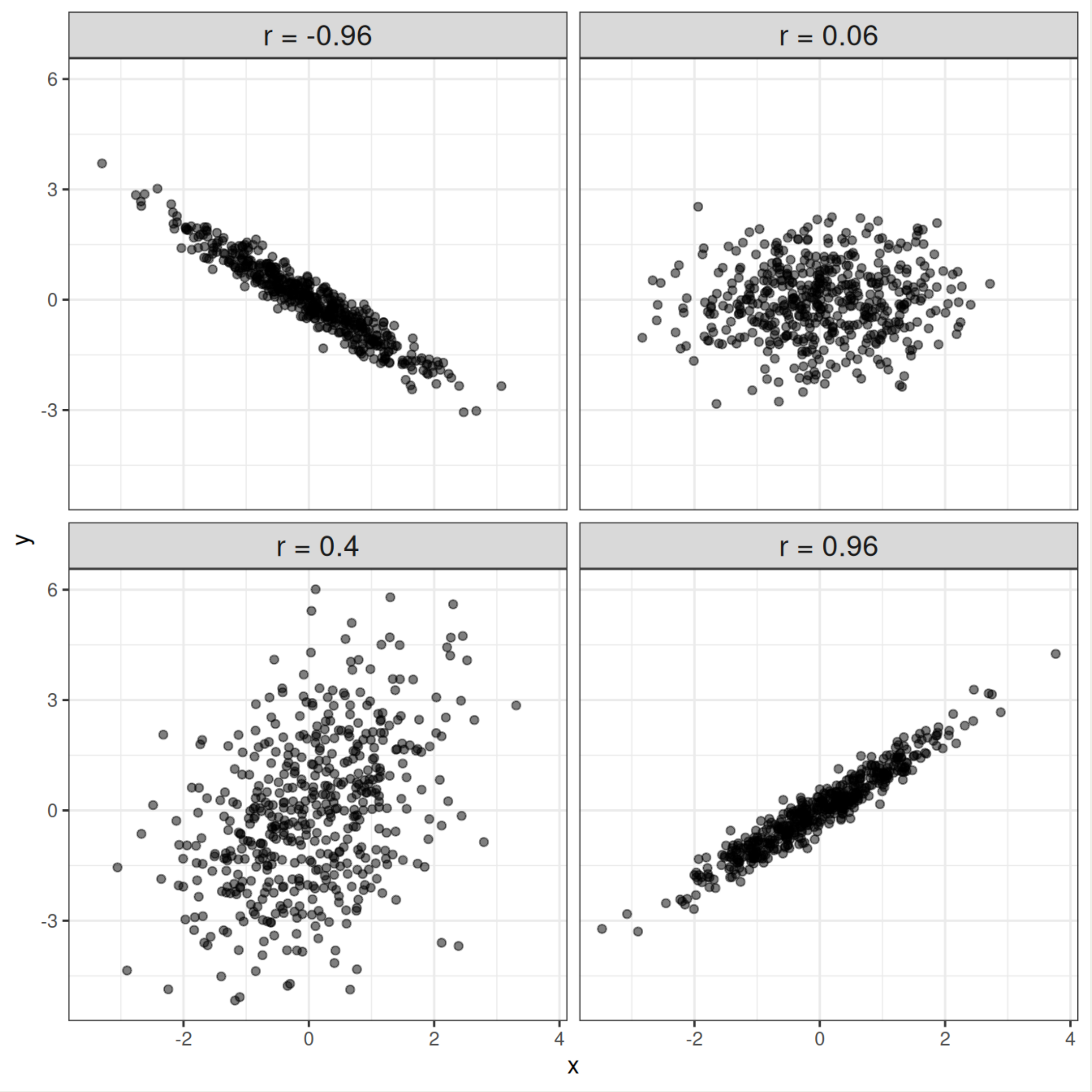
Correlation

- While the sign of the covariance has direct interpretation, its magnitude is driven by the units of measurement for X and Y.
- To standardise covariance we can divide it by the product of standard deviations of the two variables:

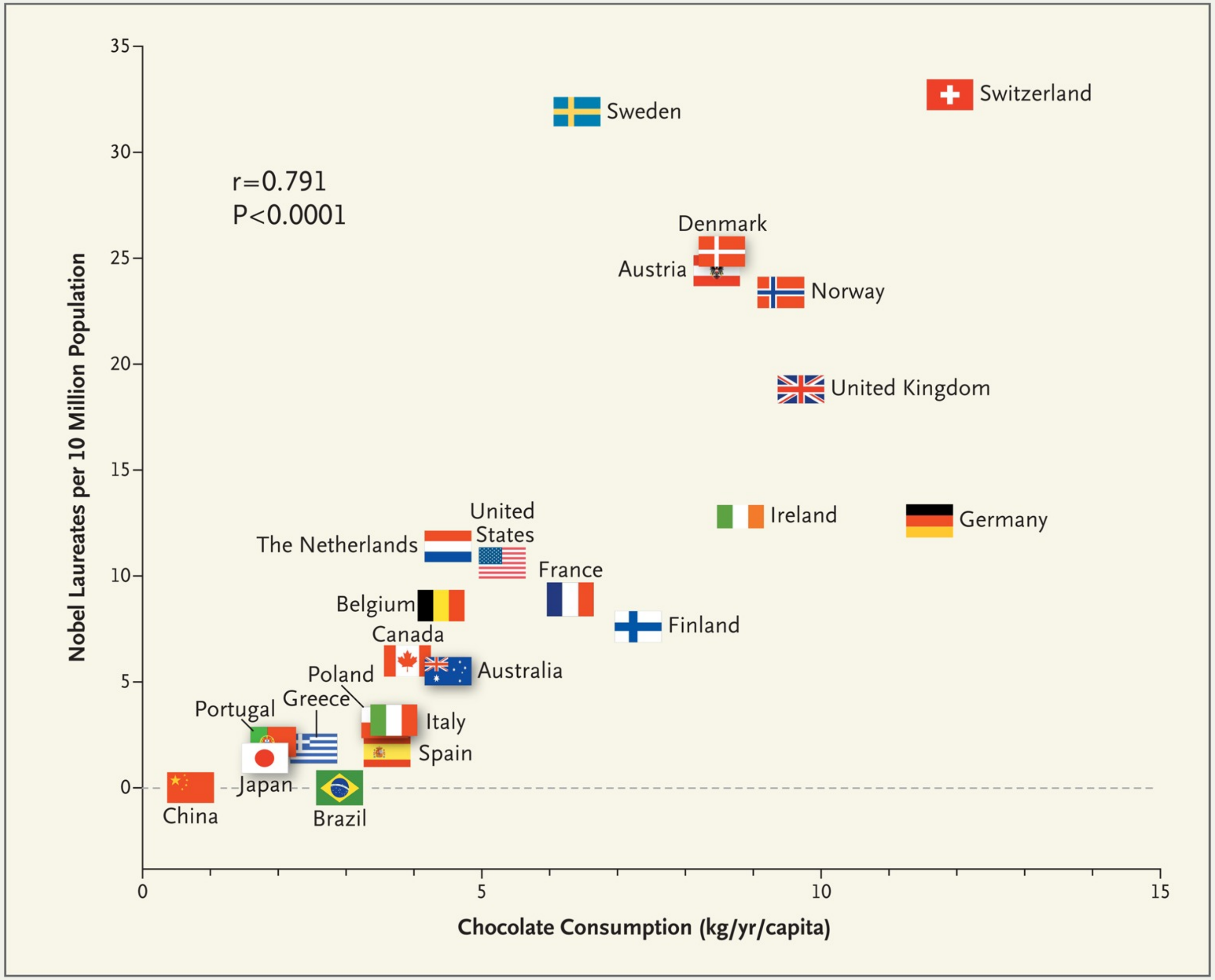
$$r = corr(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

- This is called a **correlation coefficient** (aka *Pearson correlation coefficient*, *Pearson's r*).
- It takes values between -1 and 1 (as opposed to any value potentially taken by covariance).
- A value of 0 indicates no correlation (no association).
- The larger is the absolute value of r, the stronger is the association between X and Y.

Direction and Magnitude



Spurious Correlation



Messerli (2012)

r^2 Statistic

- Once we estimate a correlation coefficient (aka r), we can also compute a square of it, known as r^2 statistic.
- r^2 statistic has a useful interpretation of a proportion of variation.
- As the correlation coefficient does not indicate which one of the two variables is dependent and which independent, we can say:
  - "The proportion of variation of Y explained by X is r^2", or
  - "The proportion of variation of X explained by Y is r^2"
- Since $-1 \leq r \leq 1$, r^2 falls between 0 and 1.

Covariance and Correlation in R

- In R we can calculate covariance and correlation using `cov()` and `cor()` functions, respectively.

```r
# Note that as with mean() function we need to take care of missing values to avoid NAs
cov(
  democracy_gdp_2020$democracy_duration,
  democracy_gdp_2020$gdp_per_capita,
  use = "complete.obs"
)
```

```
[1] 370314.5
```

```r
cor(
  democracy_gdp_2020$democracy_duration,
  democracy_gdp_2020$gdp_per_capita,
  use = "complete.obs"
)
```

```
[1] 0.3667133
```

Covariance and Correlation in R Continued

- Note that logarithmic transformation affects the estimated quantities.

```
cov(
  log(democracy_gdp_2020$democracy_duration),
  log(democracy_gdp_2020$gdp_per_capita),
  use = "complete.obs"
)
```

```
[1] 0.5610533
```

```
cor(
  log(democracy_gdp_2020$democracy_duration),
  log(democracy_gdp_2020$gdp_per_capita),
  use = "complete.obs"
)
```

```
[1] 0.416297
```

Statistical Inference for Correlation

- As with χ^2 test and t-test before, we can test the statistical significance of a correlation coefficient.
- Hypotheses:
  - Null Hypothesis $H_0 : \rho = 0$ or "in the population there is no association (relationship) between X and Y"
  - Alternative Hypothesis $H_a : \rho \neq 0$ or "in the population there is an association (relationship) between X and Y"
- The test statistic for testing $H_0 : \rho = 0$ is:

$$t = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}}$$

Example: Inference for Regime Longevity and GDP

- Let's work out the test statistic for the correlation between regime longevity and GDP.

$$t = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}} = \frac{0.3667}{\sqrt{(1 - 0.134)/(175 - 2)}} = \frac{0.3667}{0.07} \approx 5.23$$

- We can then find an associated two-tail *p*-value:

```
(1 - pnorm(5.23)) * 2
```

```
[1] 1.6951e-07
```

Example: Inference for Regime Longevity and GDP in R

- In R we can also carry out the entire test by using `cor.test()` function:

```
cor.test(
  democracy_gdp_2020$democracy_duration,
  democracy_gdp_2020$gdp_per_capita
)
```

```
	Pearson's product-moment correlation

data:  democracy_gdp_2020$democracy_duration and democracy_gdp_2020$gdp_per_capita
t = 5.1845, df = 173, p-value = 5.995e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2309328 0.4884833
sample estimates:
      cor 
0.3667133
```

Example: Significance Testing for Log Transformed Variables

- Let's see how logarithmic transformation affects our statistical inference.

```r
cor.test(
  log(democracy_gdp_2020$democracy_duration),
  log(democracy_gdp_2020$gdp_per_capita),
)
```

```
	Pearson's product-moment correlation

data:  log(democracy_gdp_2020$democracy_duration) and log(democracy_gdp_2020$gdp_per_capita)
t = 6.0222, df = 173, p-value = 1.005e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2855905 0.5317990
sample estimates:
     cor 
0.416297 
```

Regression Coefficient

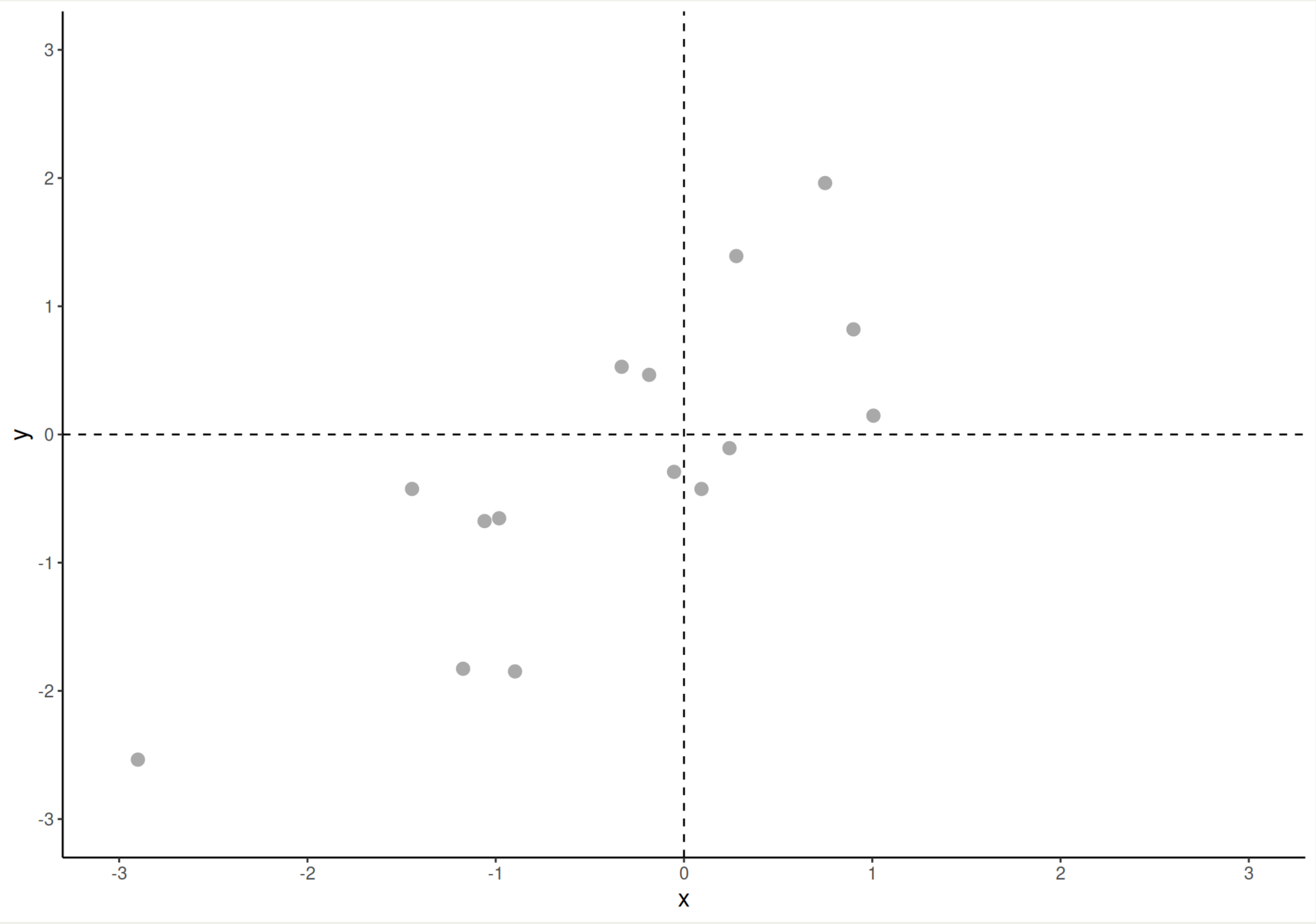
Slope of the Regression Line

- While interesting, measure of proportion of variation given by r^2 isn't very intuitive.
- It says nothing about the substantive importance or the size of this relationship.
- **Slope of the regression line** (aka **regression coefficient**) is the most common focus when analysing relationships between quantitative variables.
- *Regression line* minimises the sum of vertical distances between data points and itself.
- It can be expressed as covariance divided by the squared standard deviations of one of the two variables:

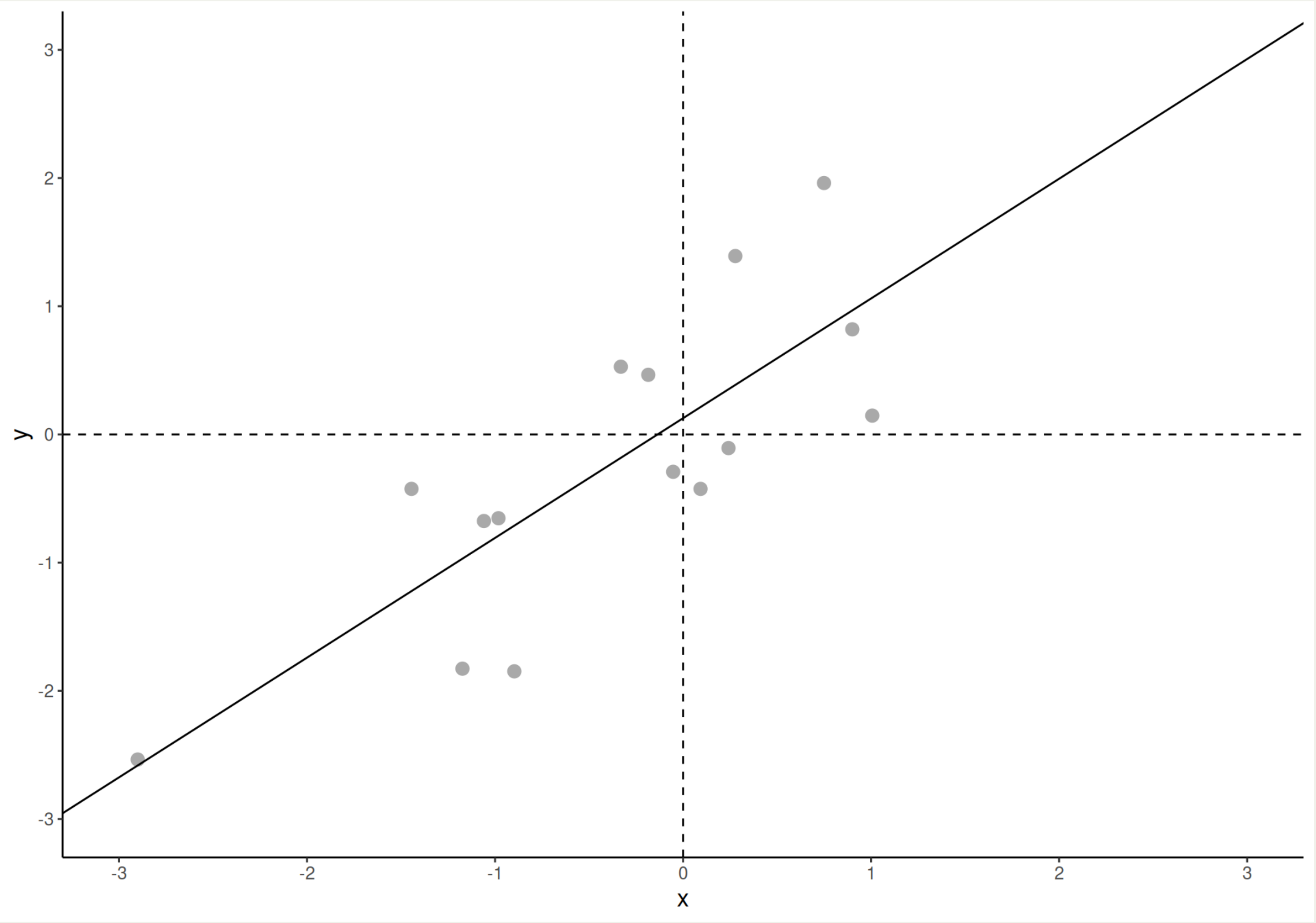
$$\beta_X = \frac{cov(X, Y)}{\sigma_X^2}$$

- Intuitively, this number tells us how much Y changes, on average, as X increases by one unit.

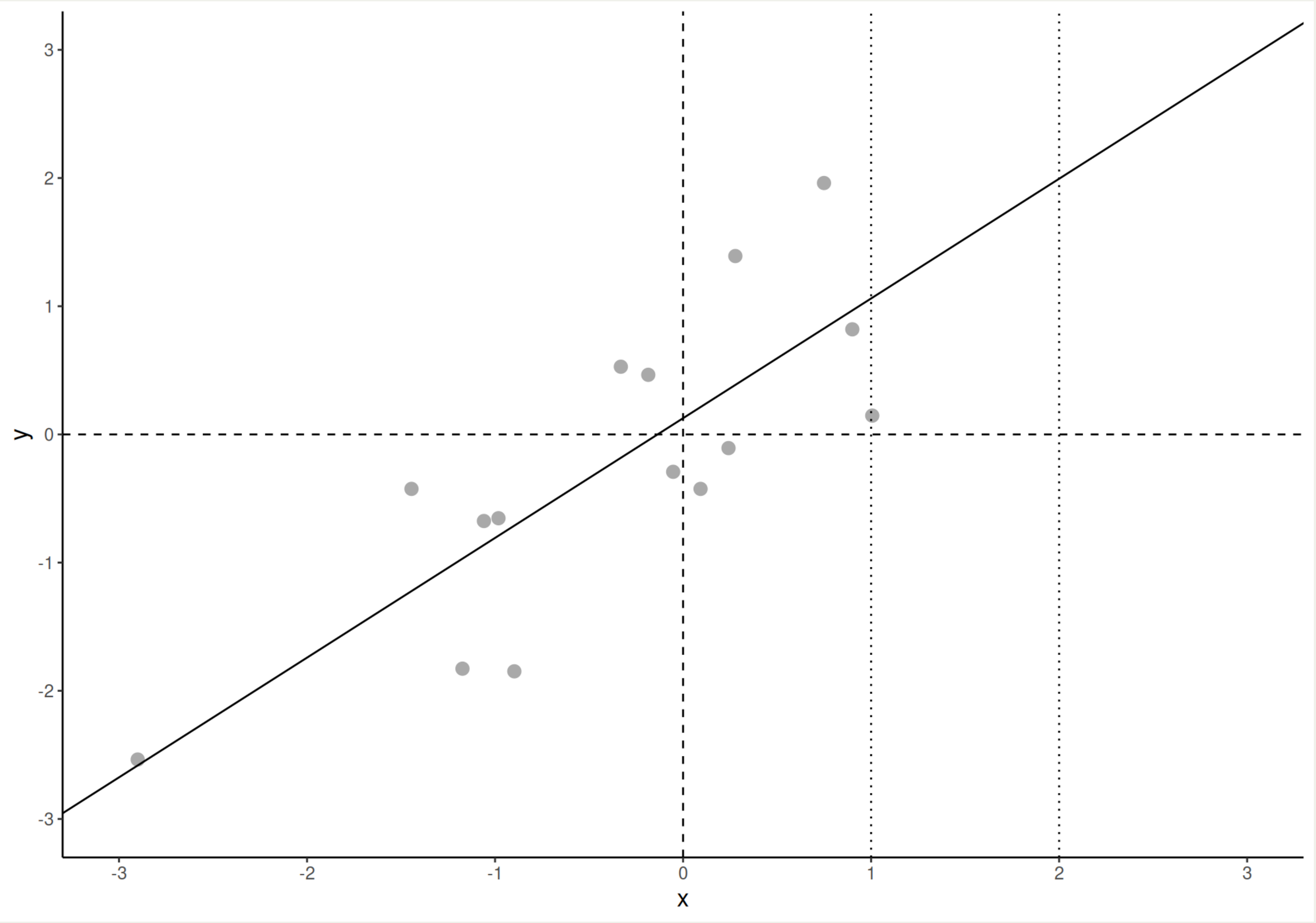
Geometric Interpretation of Regression Line



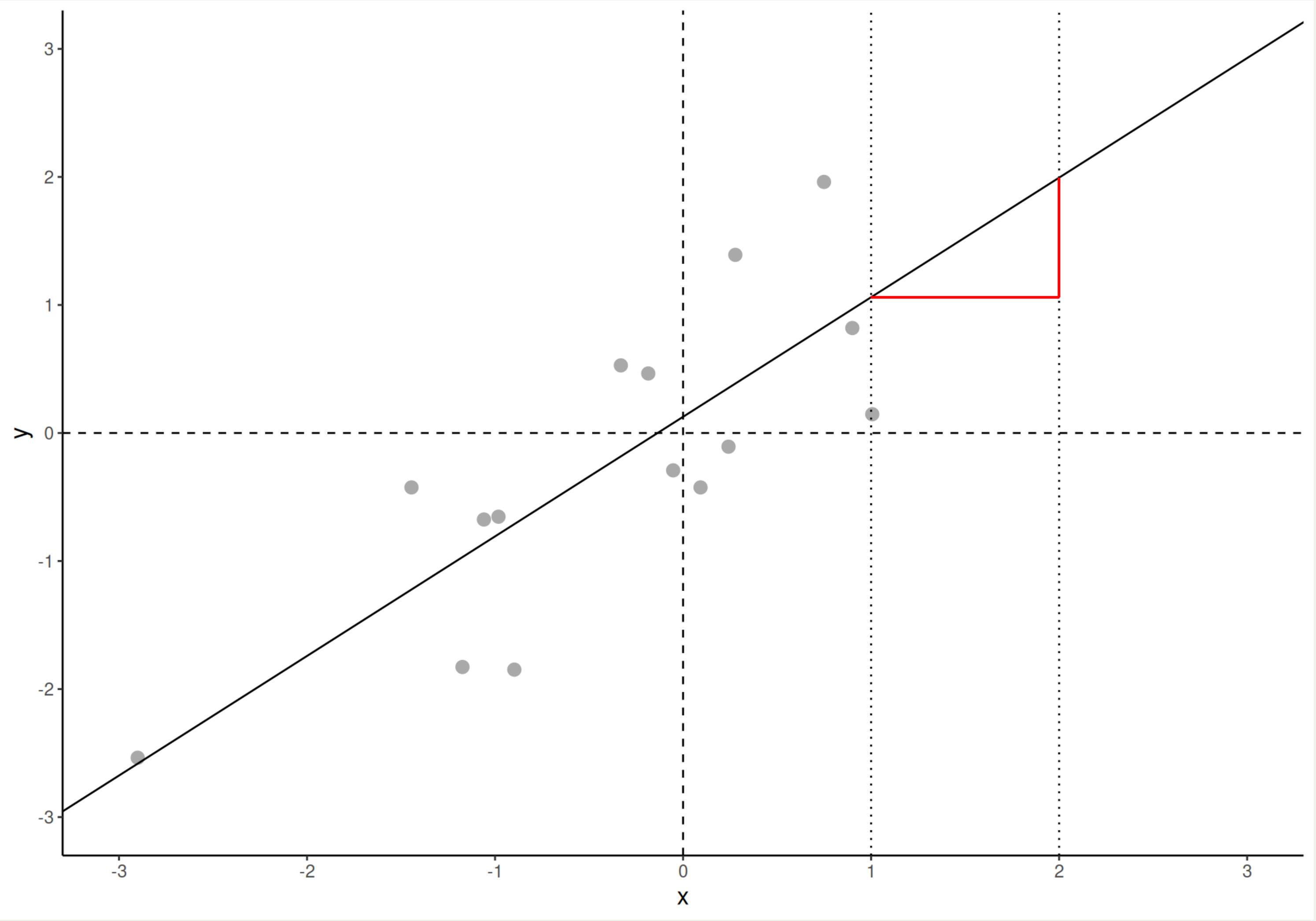
Geometric Interpretation of Regression Line



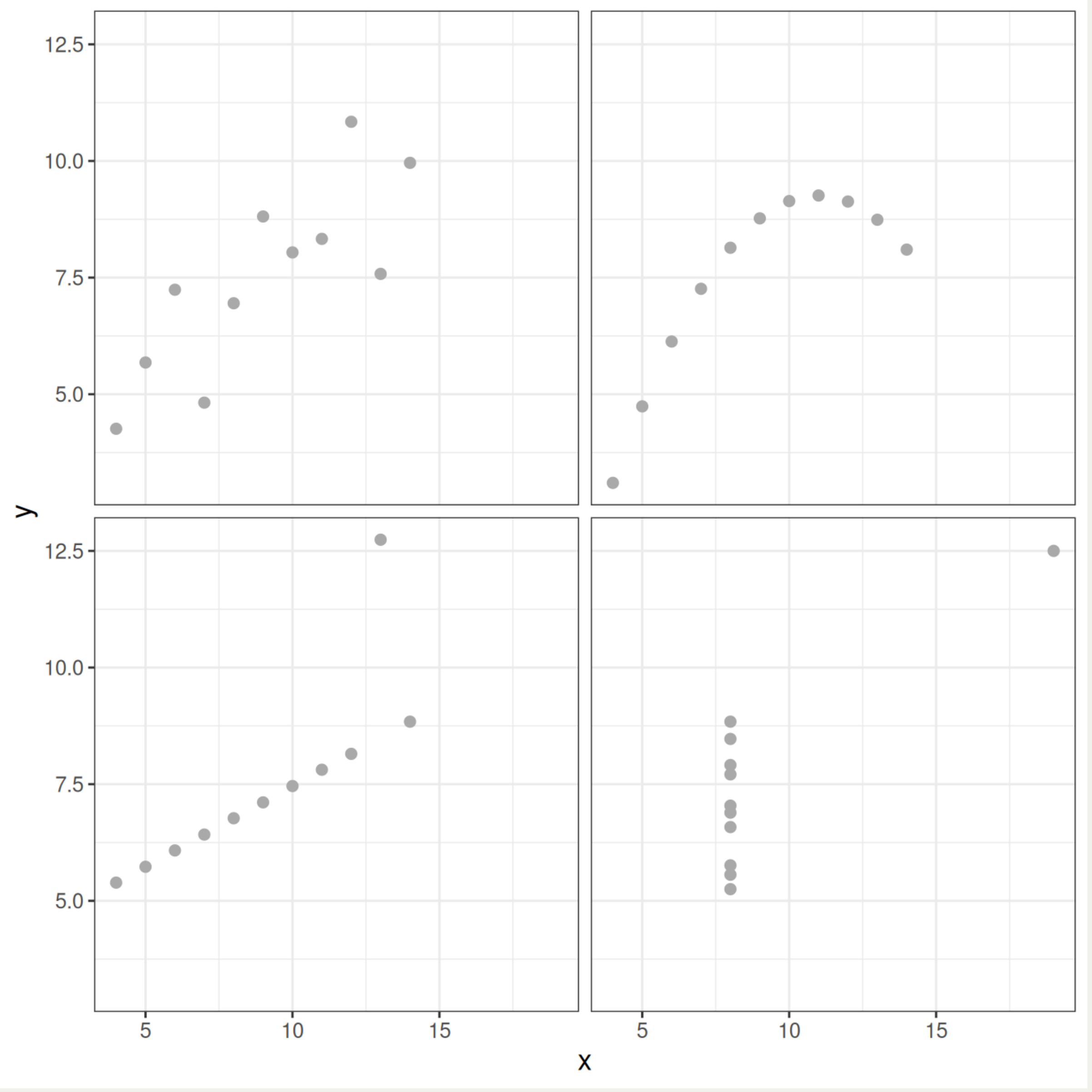
Geometric Interpretation of Regression Line



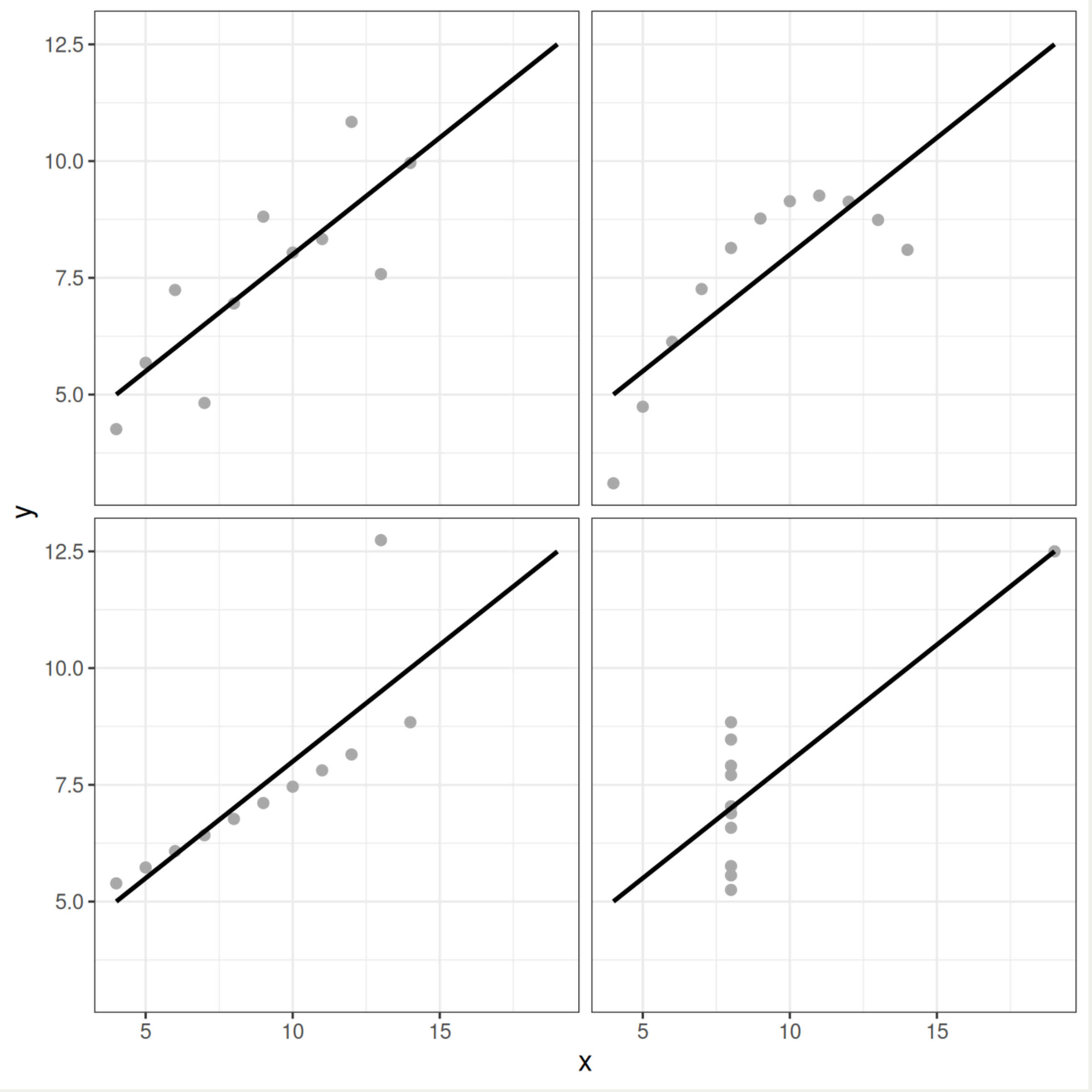
Geometric Interpretation of Regression Line



Anscombe's Quartet



Linearity Assumption



Next

- Workshop:
  - Visualisations
- Next week:
  - Reading week
- After reading week:
  - Linear regression