Module 6. Bivariate analysis: quantitative—quantitative

Data Science & AI

Sabine De Vreese Stijn Lievens Lieven Smits Bert Van Vreckem
2021–2022

HO
GENT

Contents

HO
GENT

Learning goals

- Determine the equation of the regression line and plot it;
- Calculate the covariance Cov, the correlation coefficient R and the coefficient of determination R^2
- Interpret these values using the correct terms;
- Visualization

HO
GENT

Bivariate analysis: overview

Independent	Dependent	Test/Metric
Qualitative	Qualitative	χ^2-test Cramér's V
Qualitative	Quantitative	two-sample t-test Cohen's d
Quantitative	Quantitative	— Regression, correlation

HO GENT

Data visualization

HO
GENT

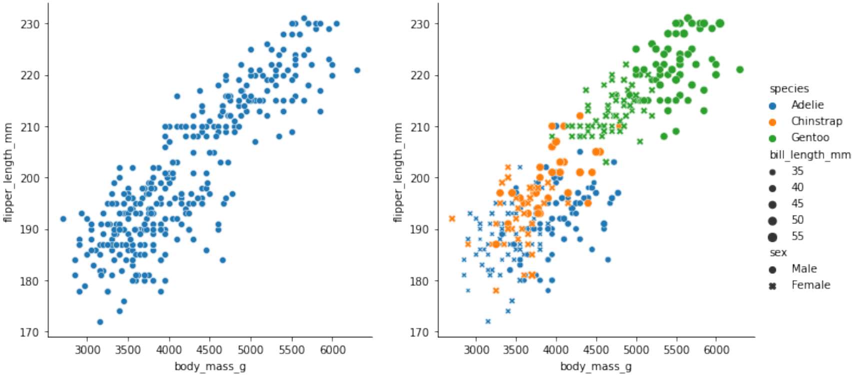
Data visualization

To visualize quantitative data, we use a *scatter plot*

- X-axis: independent variable
- Y-axis: dependent variable
- Each point corresponds to an observation

Data visualization

Scatterplot



Source: Horst A., et al. (2020) palmerpenguins: Palmer Archipelago (Antarctica) penguin data, `https://allisonhorst.github.io/palmerpenguins/`

Linear regression

HO
GENT

Linear Regression

With **regression** we will try to find a **consistent** and **systematic** relationship between two qualitative variables.

1. **Monotonic:** consistent direction of the relationship between the two variables: increasing or decreasing
2. **Non-monotonic:** value of dependent variable changes systematically with value of independent variable, but the direction is not consistent

HO
GENT

Linear Regression

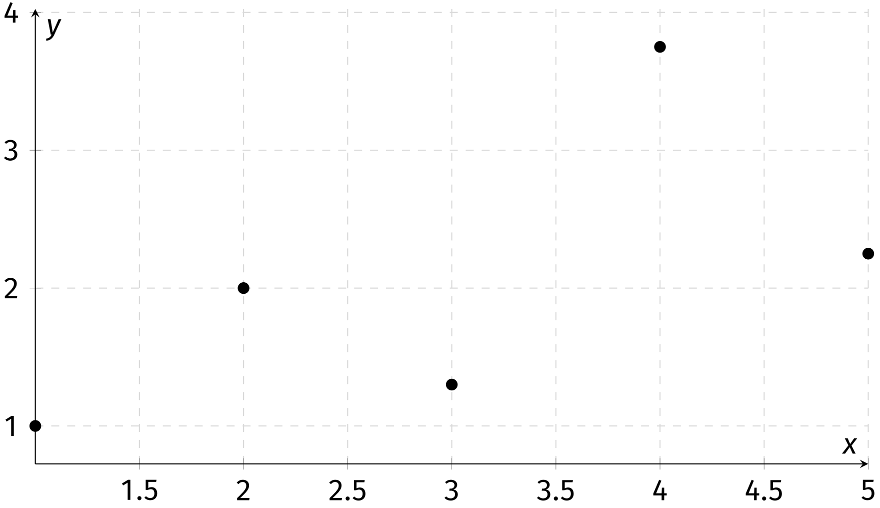
A *linear* relationship between an independent and dependent variable.

Characteristics:

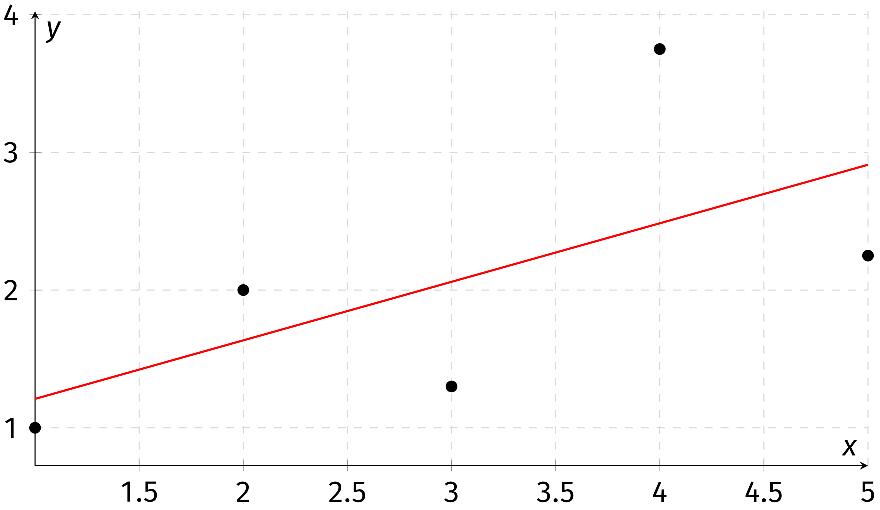
- Presence: is there a relationship?
- Direction: increasing or decreasing?
- Strength of the relationship: strong, moderate, weak, nonexistent, …

HO
GENT

Linear Regression



Linear Regression



Method of least squares

Example

Santa Claus wants to increase the weight of his reindeer. Is there a relationship between the protein content of the food and the weight gain of the reindeer?

Method of least squares

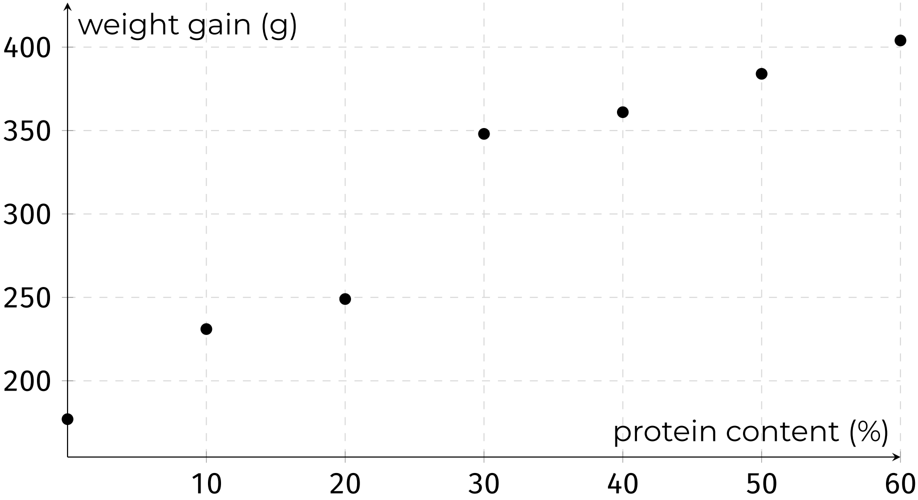
Example

Protein content%	Weight gain (grams)
0	177
10	231
20	249
30	348
40	361
50	384
60	404

HO
GENT

Method of least squares

Example



Method of least squares

Example

x	y	$x - \overline{x}$	$y - \overline{y}$	$(x - \overline{x})(y - \overline{y})$	$(x - \overline{x})^2$
0	177	-30	-130,71	3921,3	900
10	231	-20	-76,71	1534,2	400
20	249	-10	-58,71	587,1	100
30	348	0	40,29	0	0
40	361	10	53,29	532,9	100
50	384	20	76,29	1525,8	400
60	404	30	96,29	2888,7	900
				10990	2800

Tabel: Calculations required to apply the method of least squares.

HO
GENT

Method of least squares

Equation

The regression line has the following equation:

$$\hat{y} = \beta_1 x + \beta_0$$

with:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x - \overline{x})^2} = \frac{10990}{2800} = 3.925$$

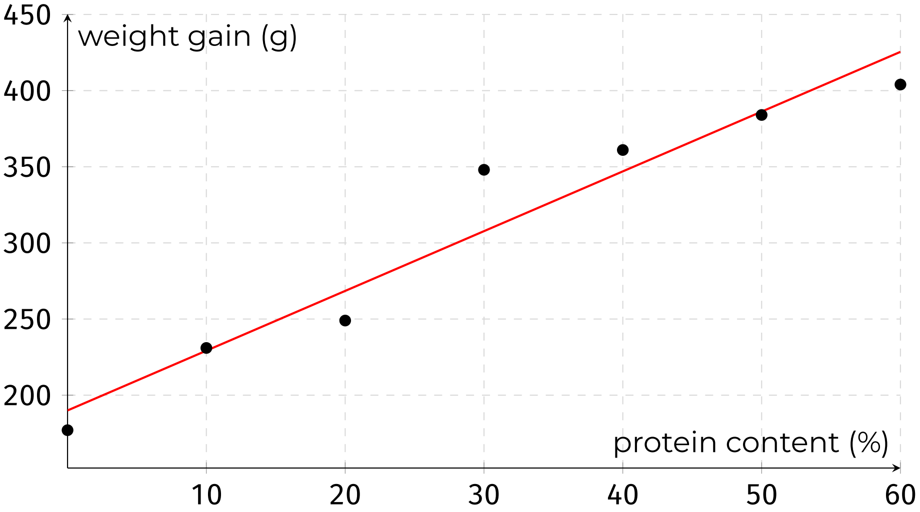
$$\beta_0 = \overline{y} - \beta_1 \overline{x} = 307.7143 - 3.925 \times 30 = 189.96$$

Note: $\hat{y}$ indicates “an estimation for y”

HO
GENT

Method of least squares

Example



Covariance

HO
GENT

Covariance

Covariance

Covariance is a measure that indicates whether a relationship between two variables is increasing or decreasing.

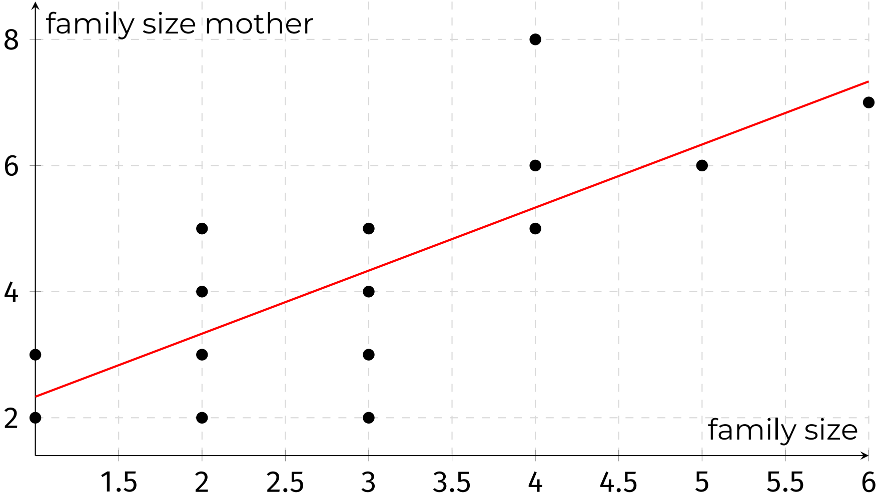
$$\text{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

Cov > 0: increasing
Cov ≈ 0: no relationship
Cov < 0: decreasing
Note Covariance of population (denominator n) vs. sample (denominator $n - 1$)

HO GENT

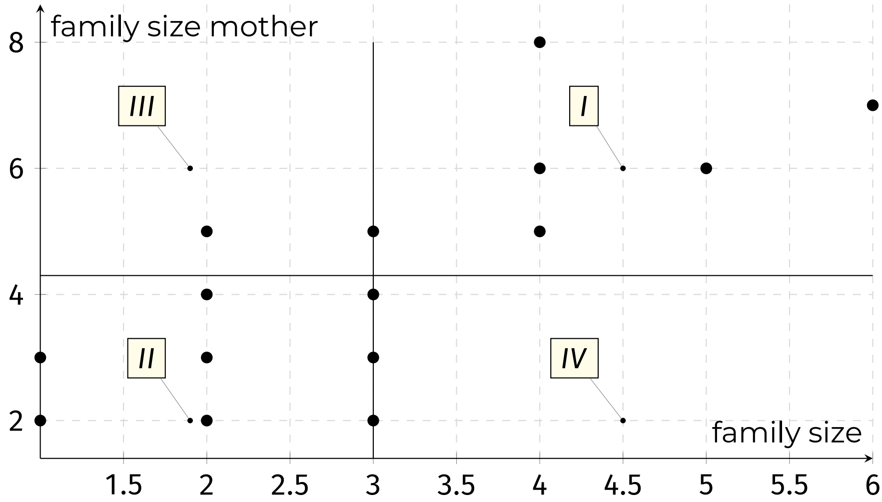
Covariance

Family size of 15 families compared to the family size of the mother.

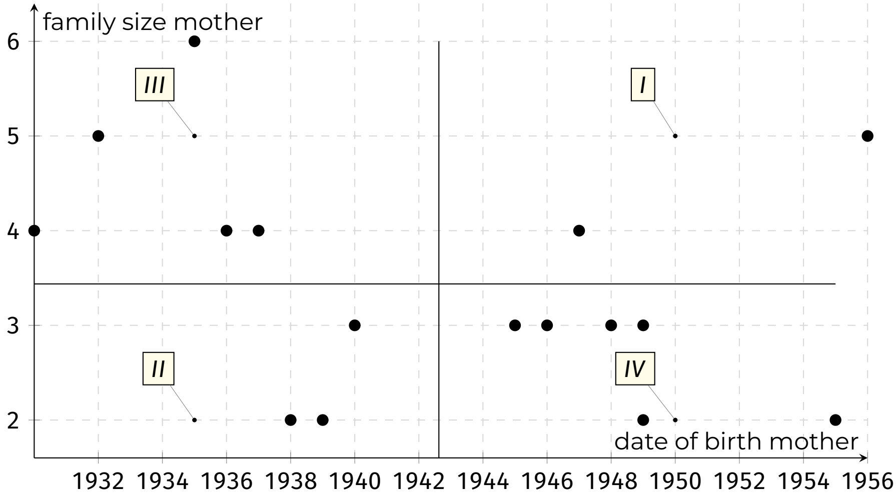


We have that $\overline{x} = 3$ and $\overline{y} = 4.3$.

Covariance


family size mother
8
6
4
2
III
I
II
IV
family size
1.5
2
2.5
3
3.5
4
4.5
5
5.5
6

Covariance for random variables



We have that $\overline{x}$ = 1942.625 and $\overline{y}$ = 3.4375.

Pearson's correlation coefficient

Pearson correlation coefficient

Pearson's Correlation Coefficient

Pearson's product-moment correlation coefficient R is a measure for the strength of a linear correlation between x and y

$$R = \frac{Cov(X, Y)}{\sigma_x \sigma_y} \tag{1}$$

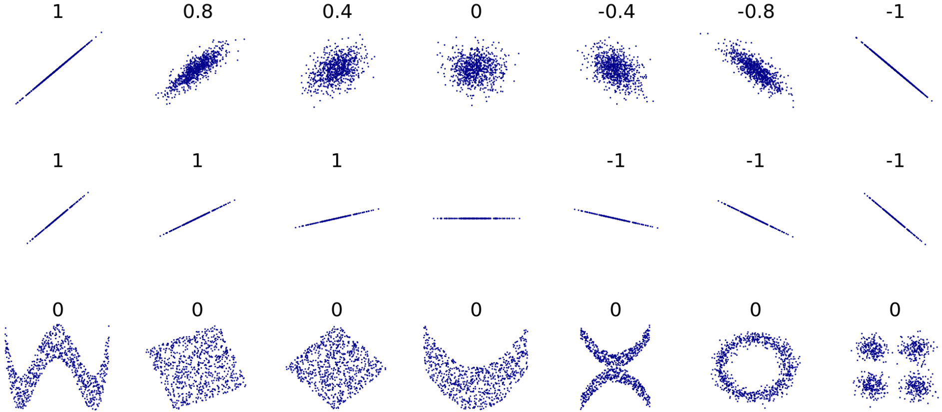
$$= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{2}$$

$$R \in [-1, +1]$$

HO
GENT

Correlation coefficient

Some datasets and their R-value



Source: Wikipedia `https://en.wikipedia.org/wiki/Pearson_correlation_coefficient`

Coefficient of determination

Coefficient of determination

Coefficient of determination

The coefficient of determination R^2 explains the percentage of the variance of the observed values relative to the regression line.

R^2: percentage variance observations explained by the regression line
$1 - R^2$: percentage variance observations *not* explained by regression

HO
GENT

Interpretation of R and R^2 values

$\|R\|$	R^2	Explained variance	Interpretation
< 0.3	< 0.1	< 10%	very weak
0.3 – 0.5	0.1 – 0.25	10 – 25%	weak
0.5 – 0.7	0.25 – 0.5	25 – 50%	moderate
0.7 – 0.85	0.5 – 0.75	50 – 75%	strong
0.85 – 0.95	0.75 – 0.9	75 – 90%	very strong
> 0.95	> 0.9	> 90%	exceptional(!)

HO
GENT

Strength of relationship reindeer

$(x - \overline{x})$	$(y - \overline{y})$	$(x - \overline{x})(y - \overline{y})$
−30	−130.714	3921.429
−20	−76.7143	1534.286
−10	−58.7143	587.1429
0	40.28571	0
10	53.28571	532.8571
20	76.28571	1525.714
30	96.28571	2888.571

$$\sum_i^n (x - \overline{x})(y - \overline{y}) = 10990$$

$$Cov = \frac{10990}{7} = 1570$$

$$\sigma_x = 20$$

$$\sigma_y = 81.03$$

$$R = \frac{1570}{20 \times 81.03} = 0.96$$

$$R^2 = 0.93$$

HO GENT

Considerations

- The correlation coefficient only looks at the relationship between two variables. Interactions with other variables are not considered.
- The correlation coefficient explicitly does not assume a causal relationship.
- Pearson’s correlation coefficient only expresses linear relationships.

HO
GENT