Module 6. Bivariate analysis: quantitative—quantitative

Data Science & AI

Sabine De Vreese **Lieven Smits** **Bert Van Vreckem**

2023–2024

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

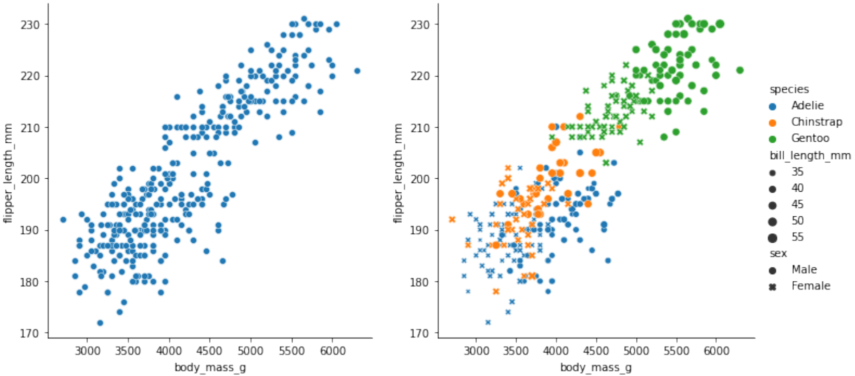
Data visualization

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

Linear Regression

With **regression** we will try to find a **consistent** and **systematic** relationship between two quantitative 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

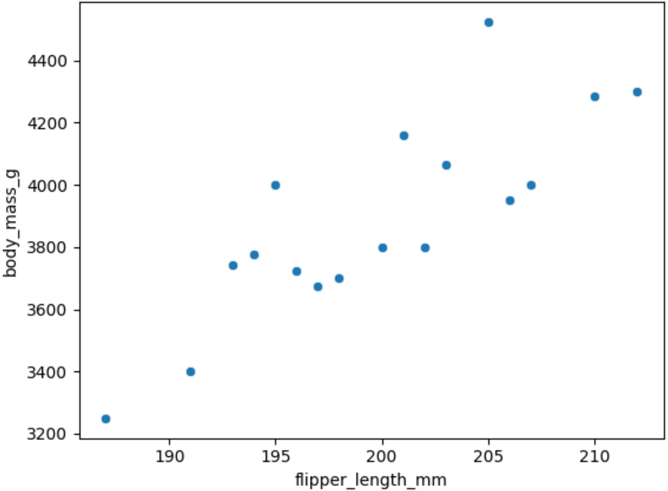
A *linear* relationship between an independent and dependent variable.

Characteristics:

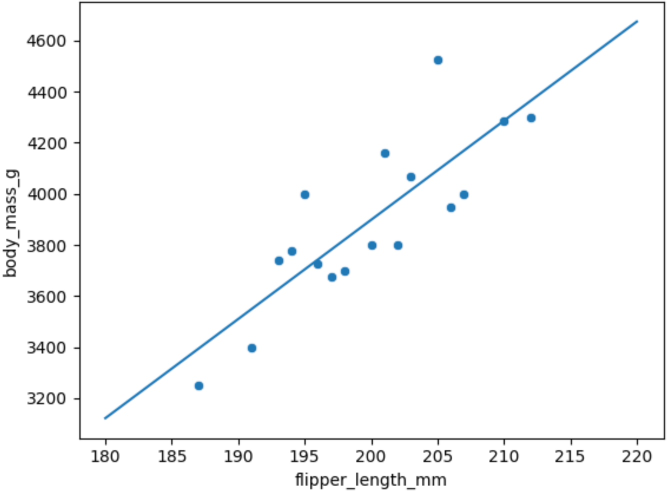
- Presence: is there a relationship?
- Direction: increasing or decreasing?
- Strength of the relationship: strong, moderate, weak, nonexistent, ...

HO
GENT

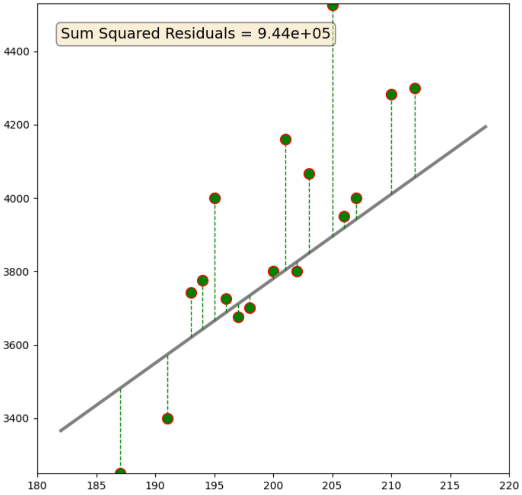
Linear Regression



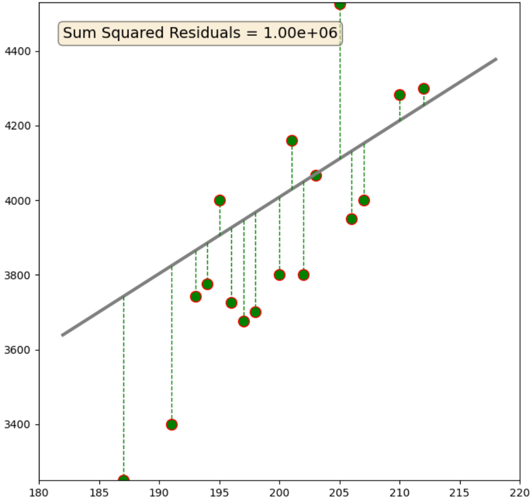
Linear Regression



Method of least squares



Method of least squares



HO
GENT

Method of least squares

	x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
0	187.0	3250.000000	-12.823529	-641.274510	8223.402537	164.442907
1	191.0	3400.000000	-8.823529	-491.274510	4334.775087	77.854671
2	193.0	3741.666667	-6.823529	-149.607843	1020.853518	46.560554
3	194.0	3775.000000	-5.823529	-116.274510	677.128028	33.913495
4	195.0	4000.000000	-4.823529	108.725490	-524.440600	23.266436
5	196.0	3725.000000	-3.823529	-166.274510	635.755479	14.619377
6	197.0	3675.000000	-2.823529	-216.274510	610.657439	7.972318
7	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

...

Method of least squares

Equation

The regression line has the following equation:

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

with:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x - \overline{x})^2} \approx \frac{29375.49}{756.47} \approx 38.83$$

$$\beta_0 = \overline{y} - \beta_1 \overline{x} \approx 3891.27 - 38.83 \times 199.82 \approx -3868.33$$

$$\hat{y} = 38.83x - 3868.33$$

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

HO
GENT

Covariance

Covariance

Covariance

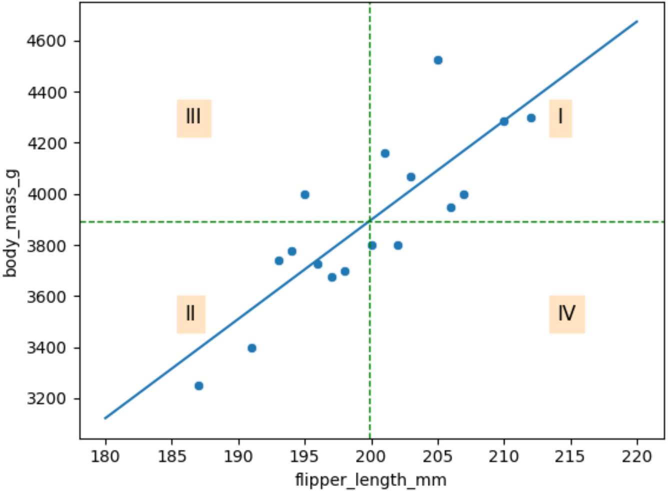
Covariance is a measure that indicates whether a relationship between two variables is increasing or decreasing.

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

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

HO GENT

Covariance



4600
4400
4200
4000
3800
3600
3400
3200
body_mass_g
180
185
190
195
200
205
210
215
220
flipper_length_mm
III
I
II
IV

Covariance

$$Cov(X, Y) \approx \frac{29375.49}{17 - 1} \approx 1835.97$$

- $Cov > 0 \Rightarrow$ increasing relationship
- What if body mass was expressed in kg instead of g?

Covariance has limited use as a measure of the relationship between two variables.

HO GENT

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

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

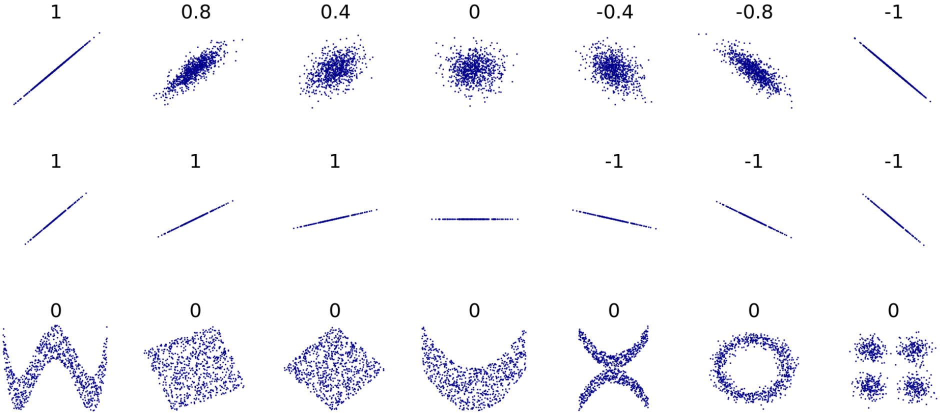
$$= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{2}$$

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

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

$\lvert R \rvert$	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

Example chinstrap penguins

$$Cov(X, Y) \approx \frac{29375.49}{17 - 1} \approx 1835.97$$

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

$$\approx \frac{1835.968}{6.876 \times 322.935} \approx 0.827$$

$$R^2 \approx 0.827^2 \approx 0.684$$

Conclusion: There is a strong linear and increasing relationship between flipper length and body mass of male chinstrap penguins. 68.4% of the variance in body mass can be explained by the variance in flipper length.

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.