Methods in Human Geography

Quantitative Methods: Statistical Analysis II

Dr Justin van Dijk
j.t.vandijk@ucl.ac.uk

UCL

This week

Part I

- Crosstabulation.
- Correlation.

Part II

- Regression.
- Assumptions.


V= 1
lim (1+1/n)^n = e
)= 1- 1
P

Crosstabulation

Crosstabulation

- We often encounter nominal or ordinal variables like gender, ethnic group, educational qualifications, income bands, or age groups.
- Information across two variables can be presented with a crosstab (contingency table) that organises data into a matrix, showing the frequency distribution across categories of two categorical variables.

Crosstabulation

	Large share population over 50		
Winner	**No**	**Yes**	***Total***
Conservative	25	91	116
Labour	243	131	374
Liberal Democrats	24	42	66
Other	8	11	19
Total	300	275	575

Crosstabulation

	Large share population over 50		
Winner	**No**	**Yes**	***Total***
Conservative	25 (21.6%)	91 (78.4%)	116 (20.2%)
Labour	243 (65%)	131 (35.0%)	374 (65.0%)
Liberal Democrats	24 (36.4%)	42 (63.6%)	66 (11.5%)
Other	8 (42.1%)	11 (57.9)	19 (3.3%)
Total	300	275	575

Chi-square test

- Chi-square test (χ^2) assesses if the observed frequencies differ significantly from expected frequencies:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- A significant chi-square result suggests an association, while a non-significant result implies independence between variables.

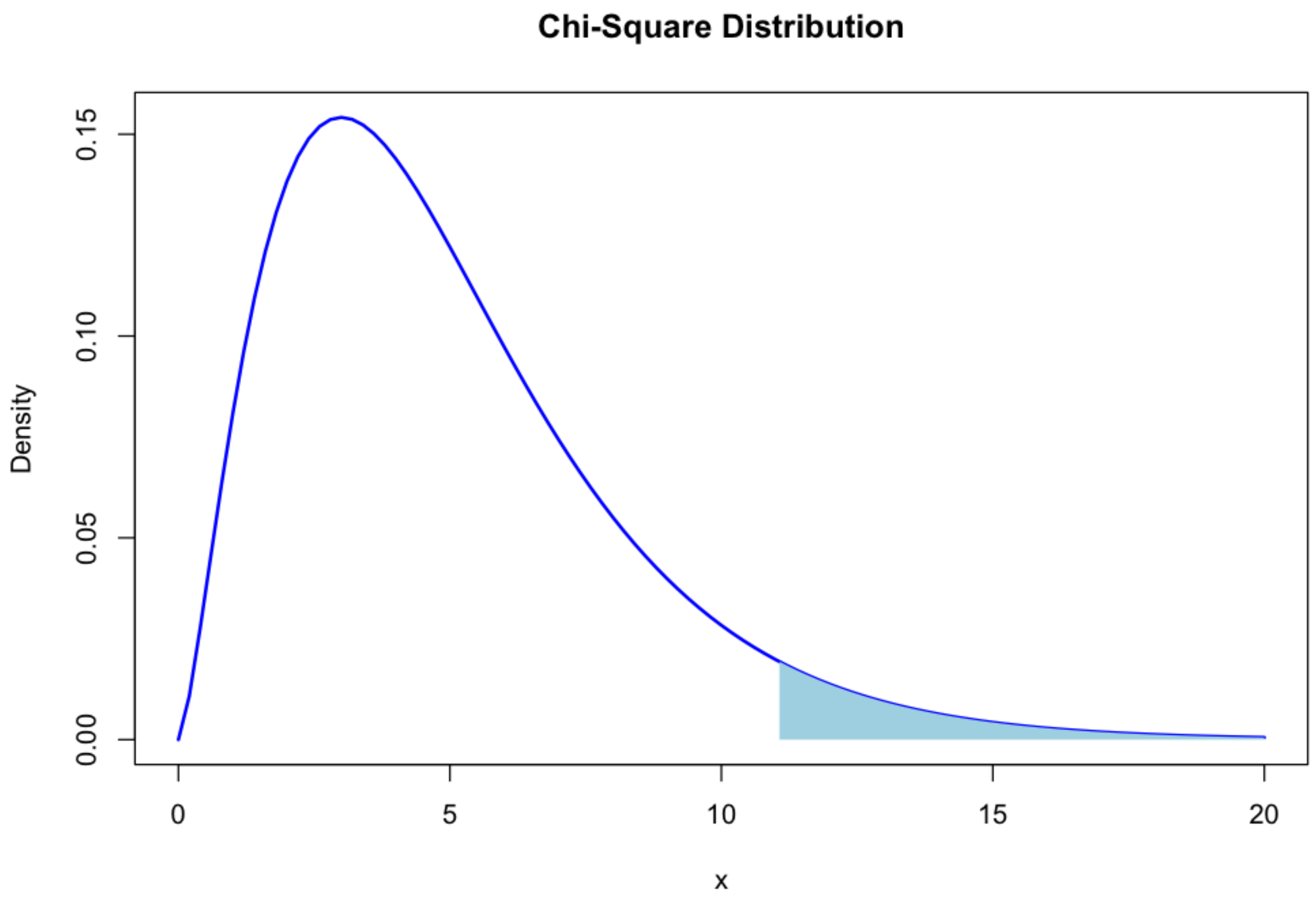
Chi-square test

	Large share over 50		
Winner	**No**	**Yes**	***Total***
Conservative	25 60.5	91 55.5	116
Labour	243 195.1	131 178.9	374
Liberal Democrats	24 34.4	42 31.6	66
Other	8 9.9	11 9.1	19
Total	300	275	575

Chi-square test

- Chi-square requires a sufficiently large sample size and expected frequency of at least five in each cell for valid results.
- It measures whether the associations (or not) in your data are different than random, but it cannot tell you strengths or directions of relationships.

Chi-square test

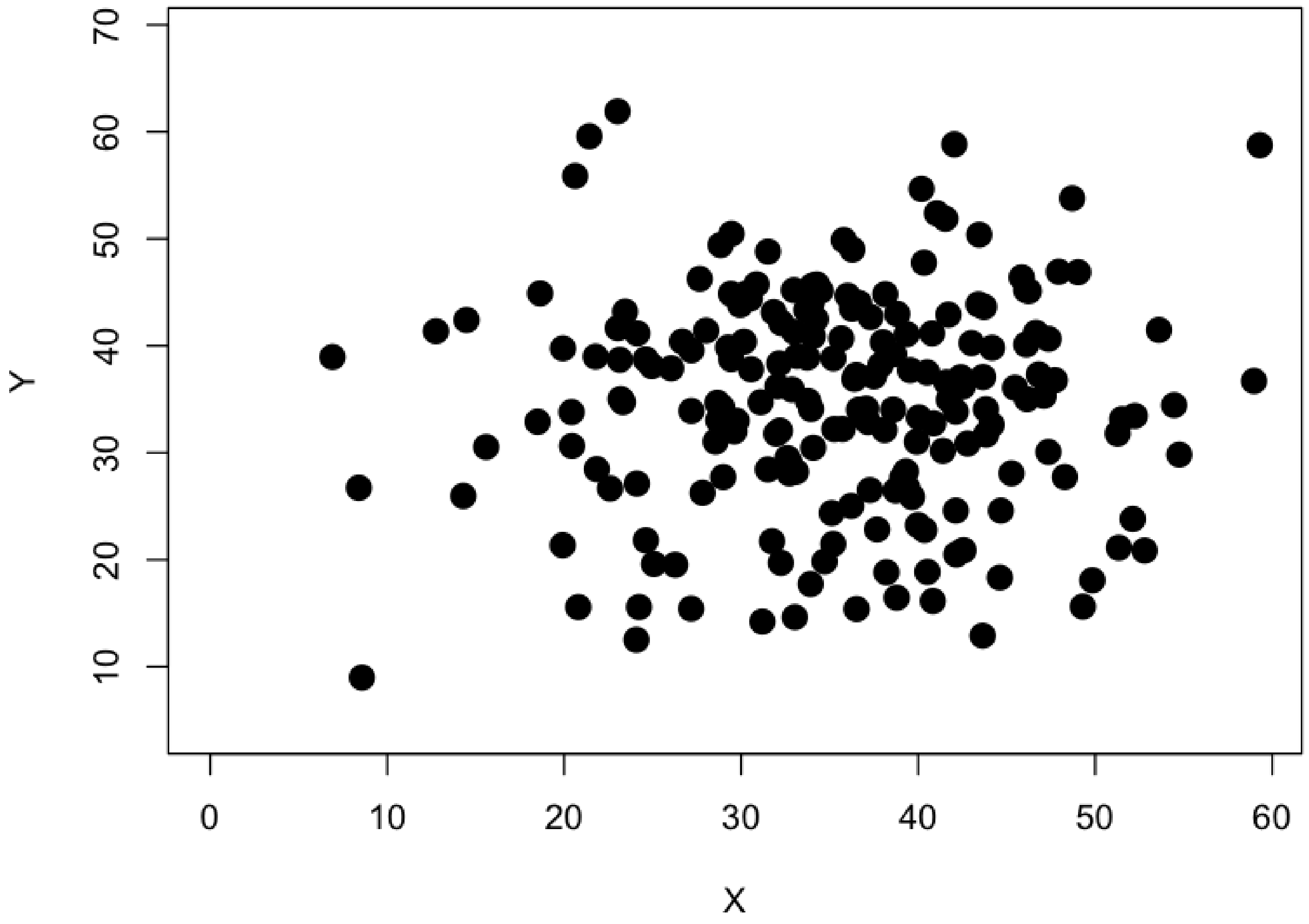


Correlation

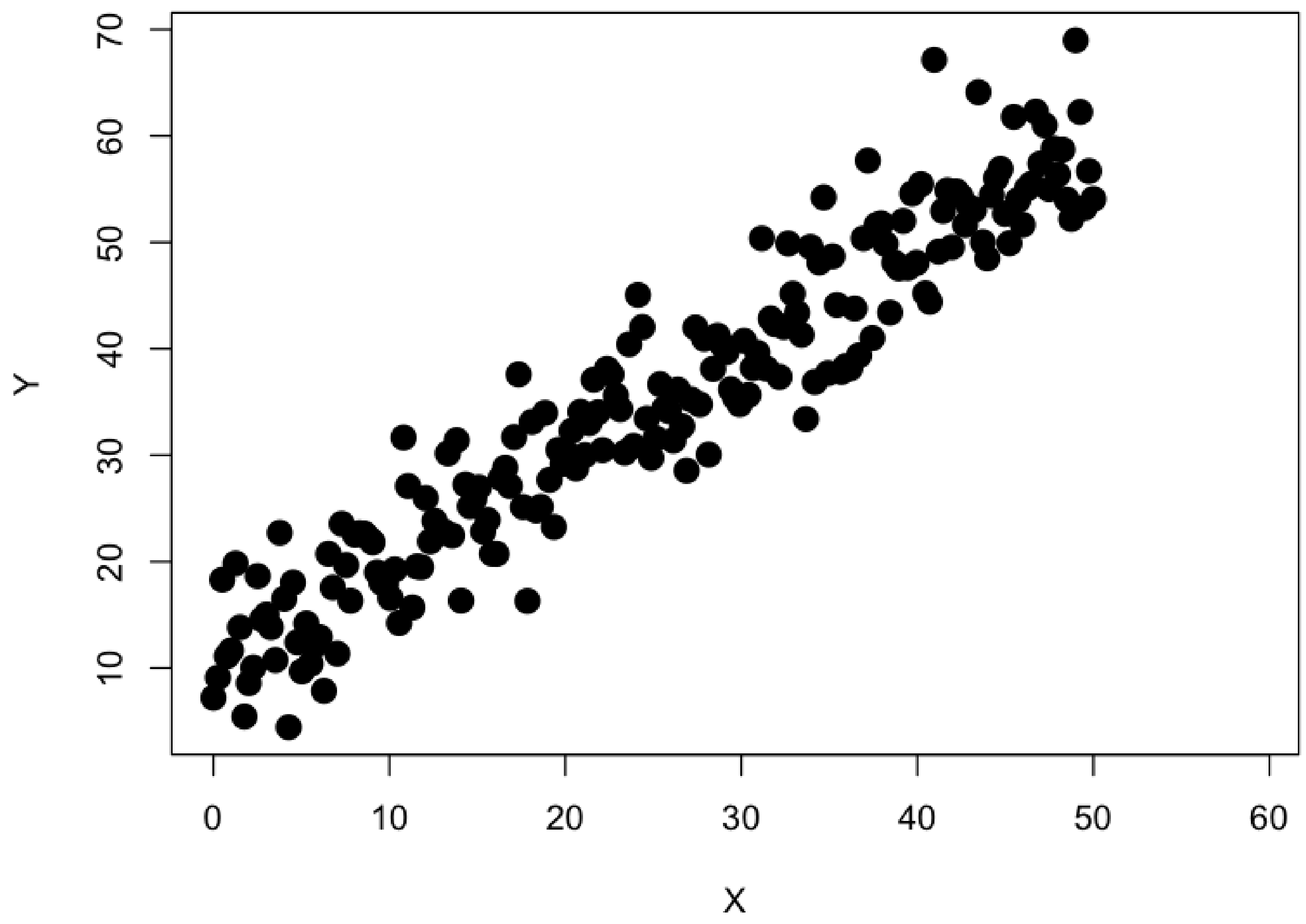
Scatterplot

- Crosstabulations are useful for examining bivariate relationships between nominal and ordinal data – but what about continuous data?
- Scatterplots offer a visual representation of the relationship between two continuous variables, though interpretation can be subjective.

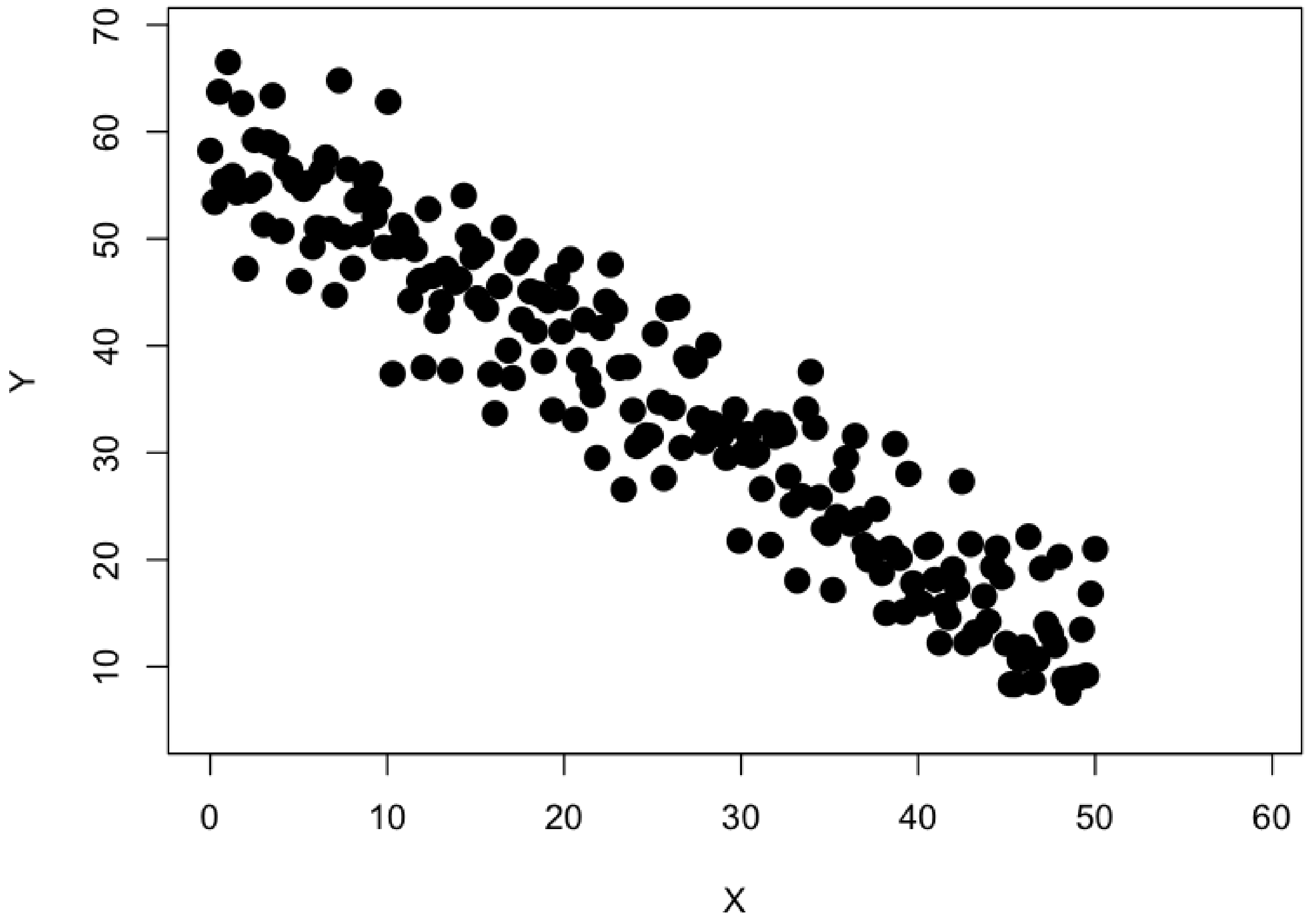
Correlation



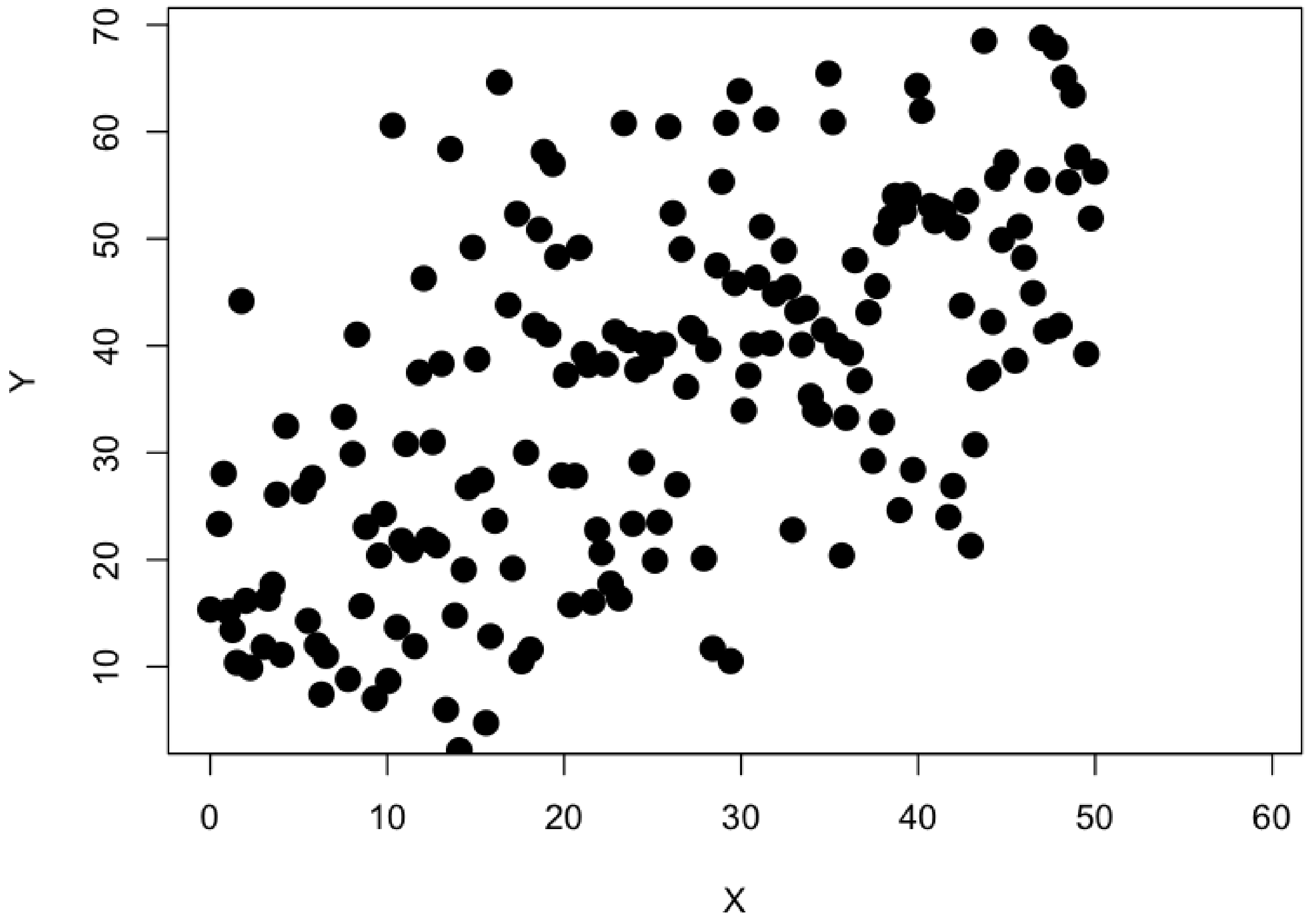
Correlation



Correlation



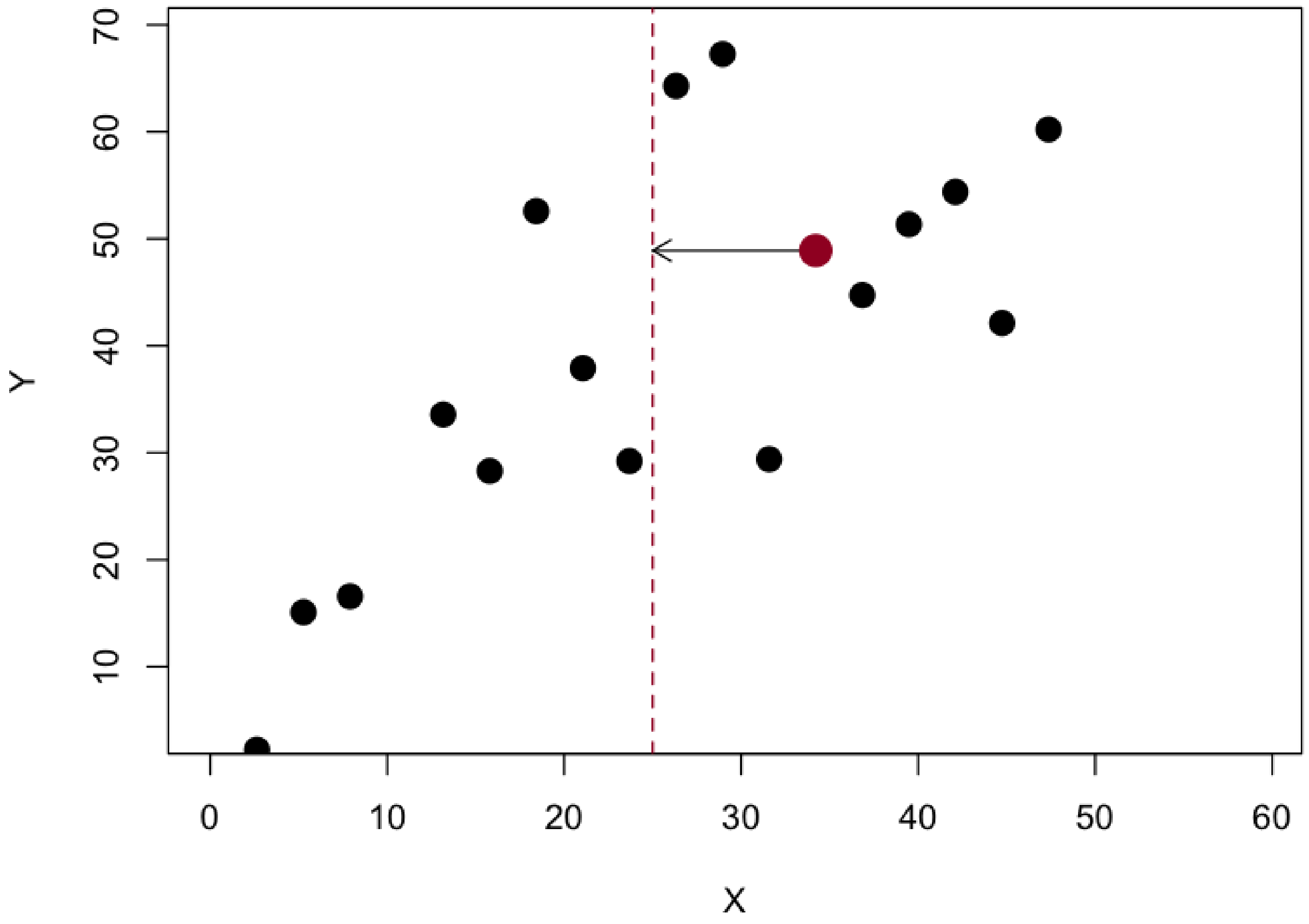
Correlation



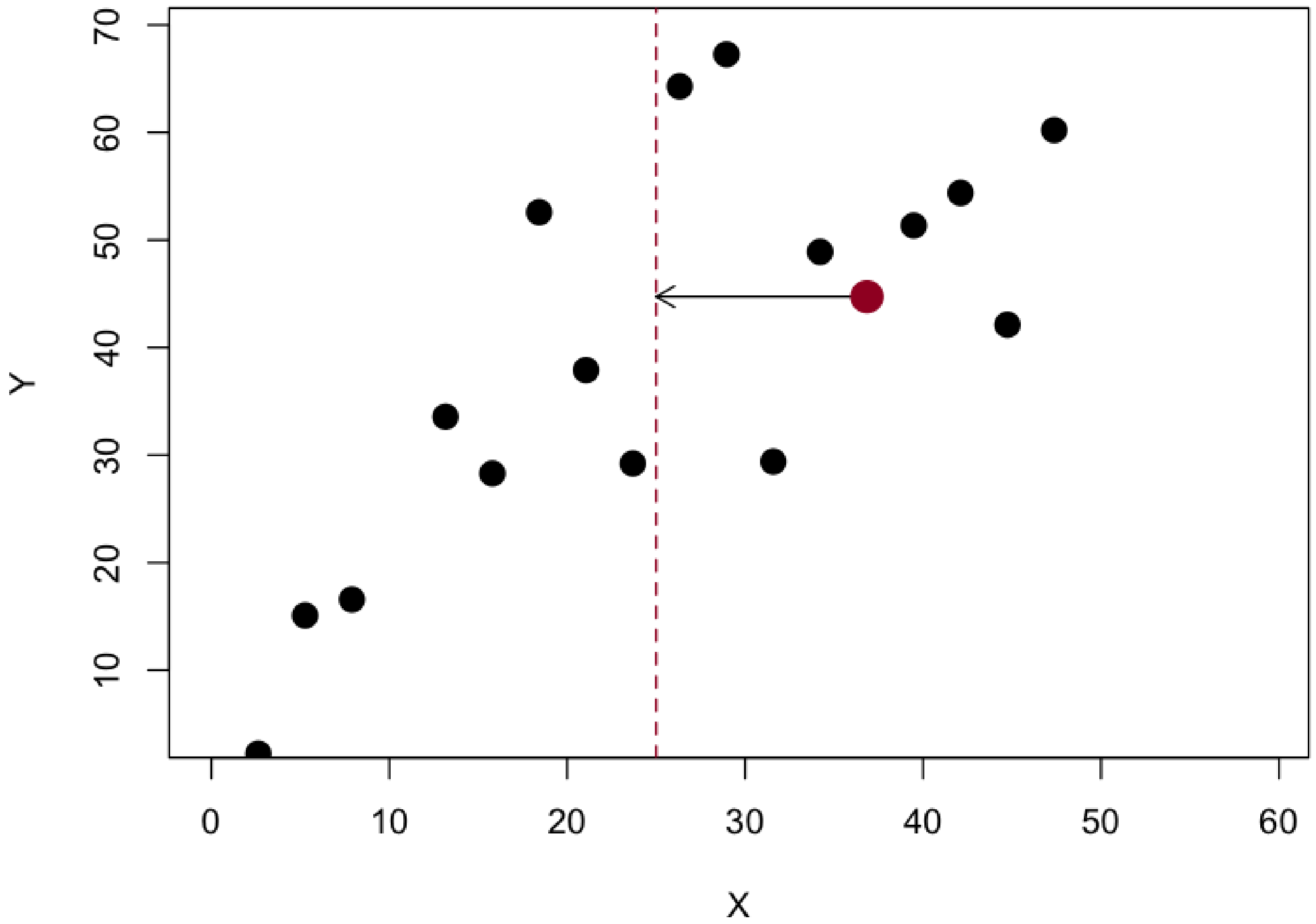
Correlation

- Correlation is the degree to which two variables ‘vary together’ or are ‘related’.
- Variables are correlated when there is some change in one variable at the same time as there is a change in another variable.
- Correlation quantifies both the strength and direction of the relationship between two variables.
- There are several measures of correlation, with Pearson’s correlation coefficient being the most commonly used.

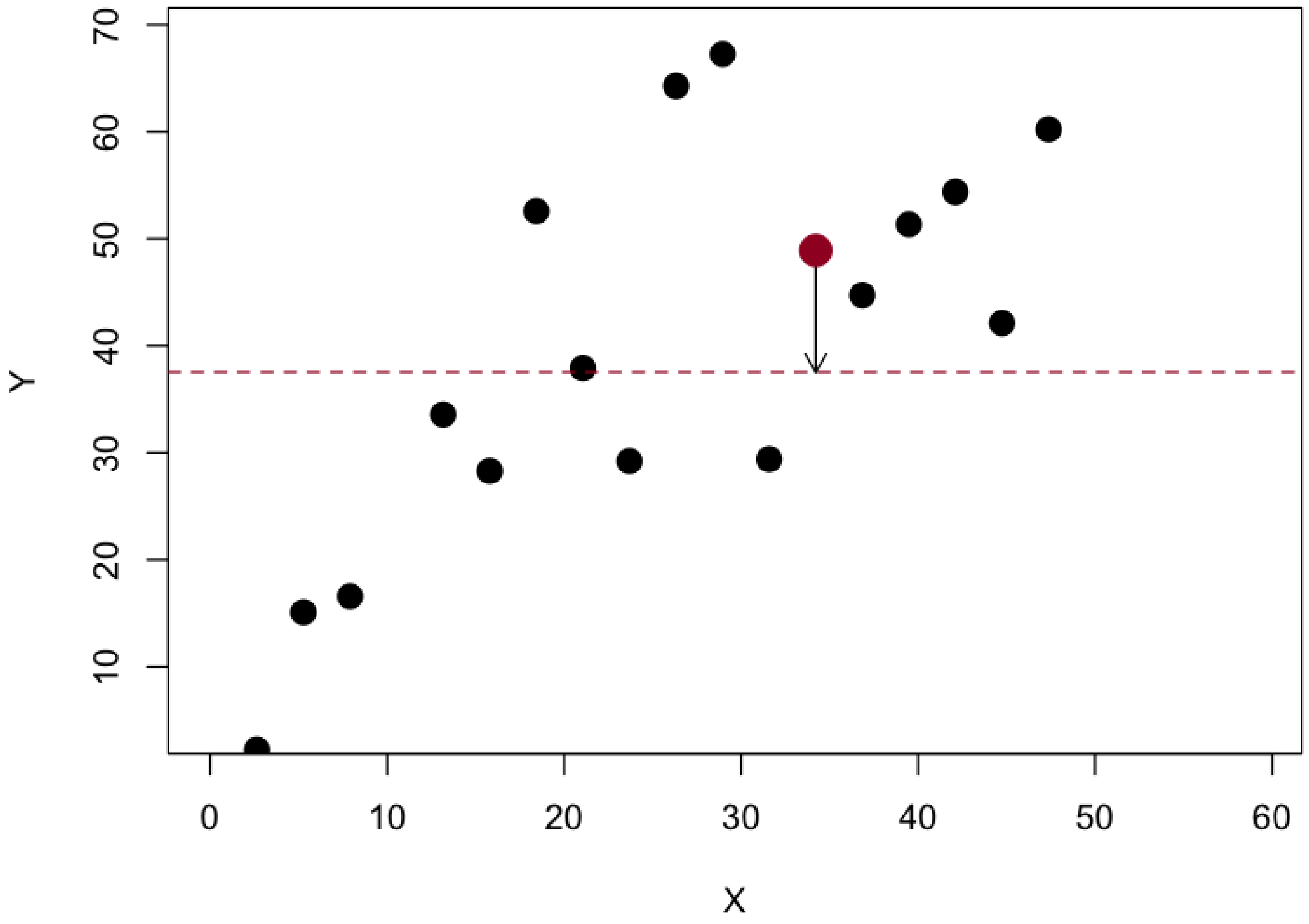
Variance



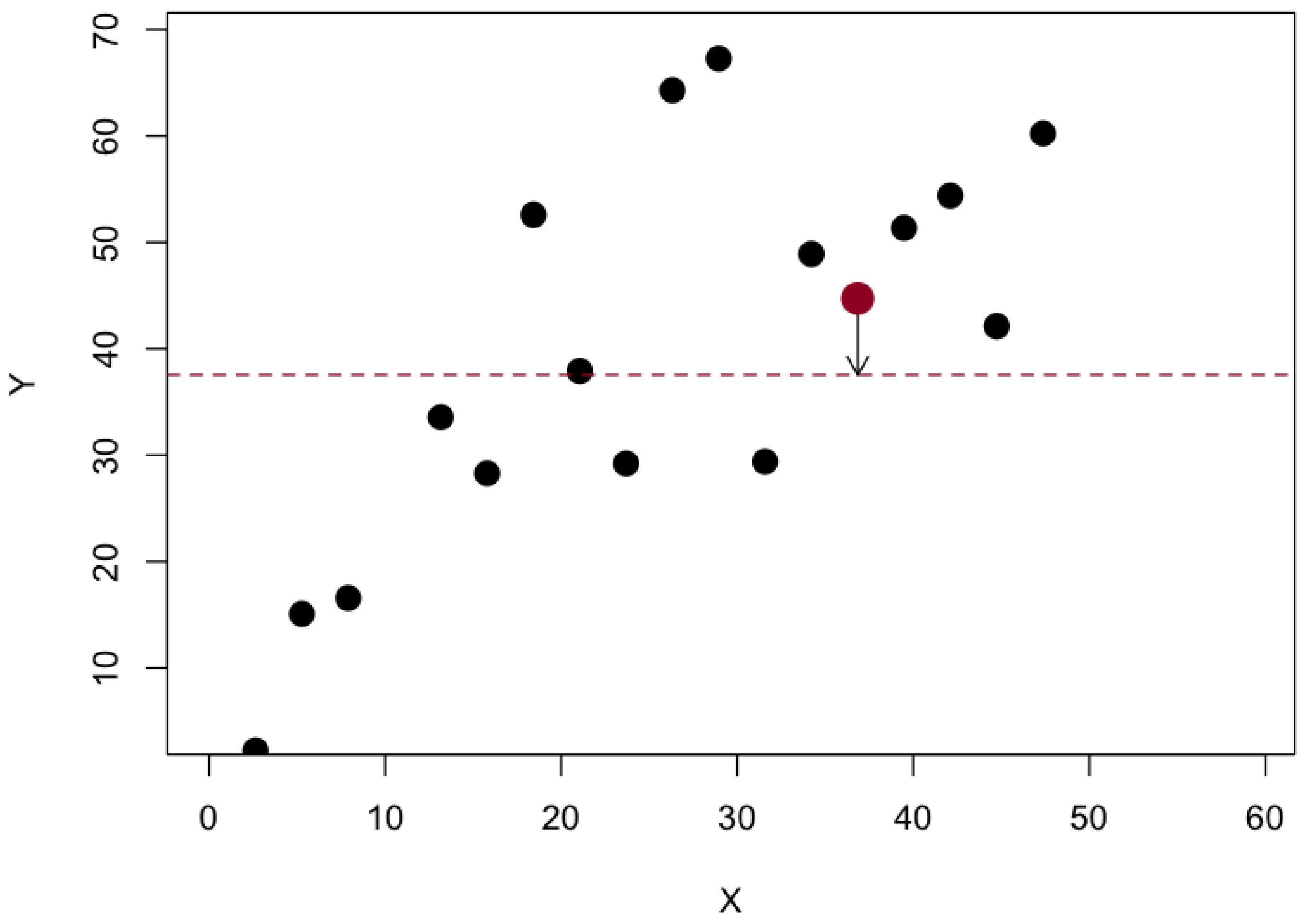
Variance



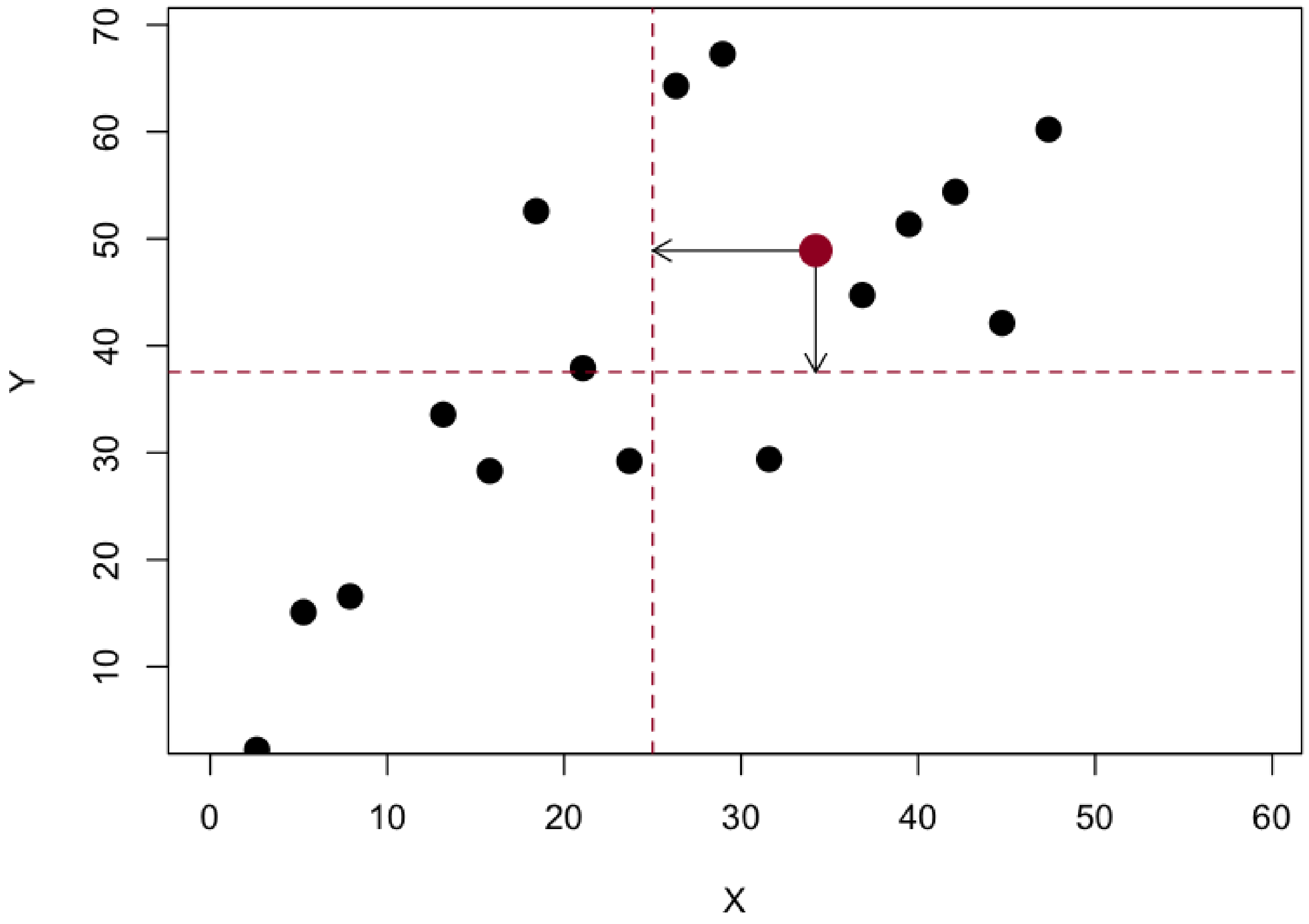
Variance



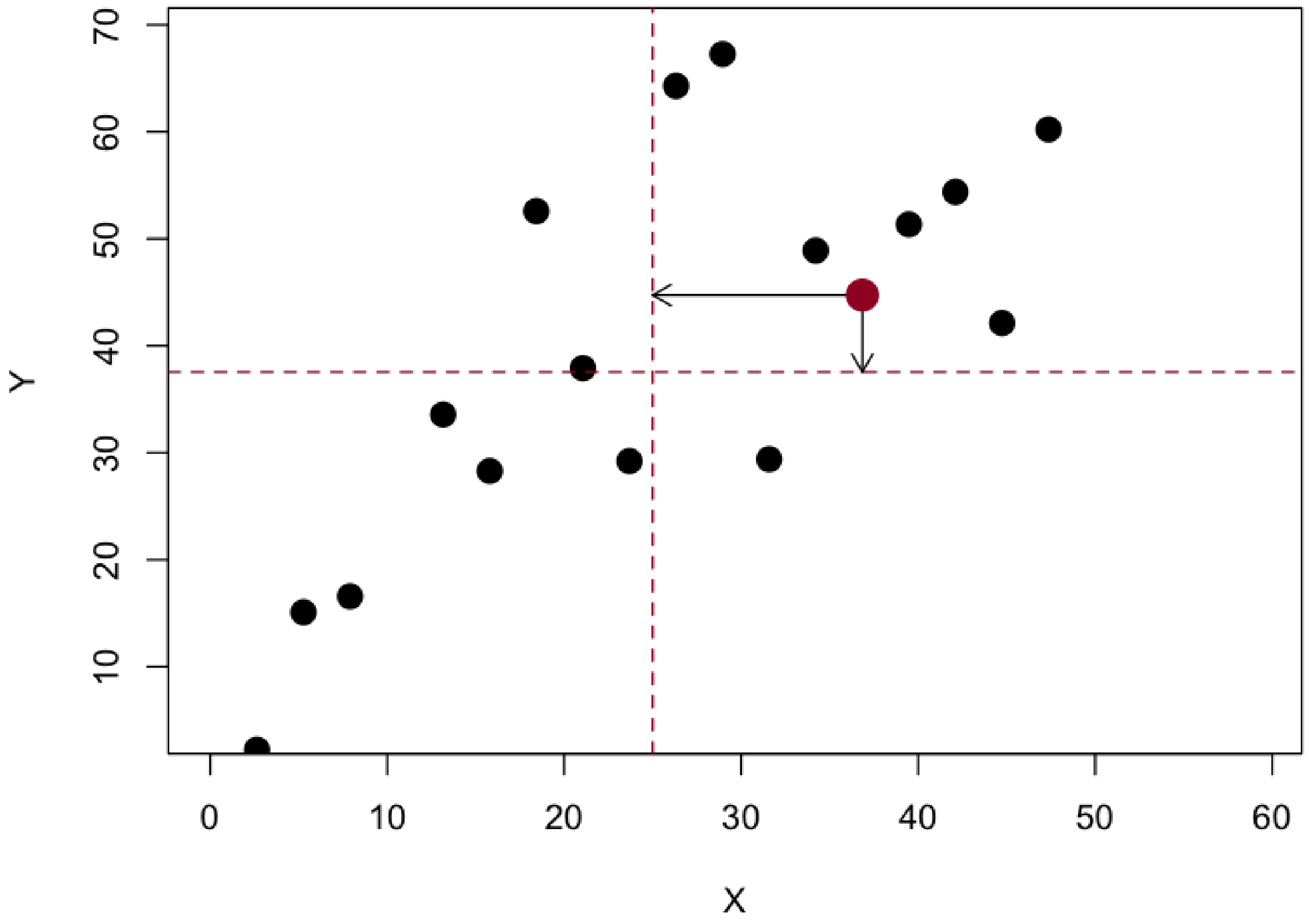
Variance



Covariance



Covariance



Variance

- Variance is a statistical measure that quantifies the dispersion of a variable's values around its mean, indicating how much the values differ from the average:

$$\sigma^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

Covariance

- Covariance assesses the degree to which two variables change together, showing whether increases in one variable correspond to increases or decreases in another:

$$cov_{x,y} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Correlation

- Pearson's correlation is a standardised measure that quantifies the strength and direction of the linear relationship between two variables, ranging from -1 to +1:

$$r = \frac{cov_{x,y}}{s_x s_y}$$

Correlation

Negative	Description	Positive
0.00	None	0.00
-0.19 – -0.01	Very weak	0.01 – 0.19
-0.39 – -0.20	Weak	0.20 – 0.39
-0.69 – -0.40	Modest	0.40 – 0.69
-0.89 – -0.70	Strong	0.70 – 0.89
-0.99 – -0.90	Very strong	0.90 – 0.99
-1.00	Perfect	1.00

More correlation

- Spearman's correlation is a non-parametric measure used to assess the strength and direction of the relationship between two ranked or ordinal variables.
- It is particularly useful when the data do not meet the assumptions of normality required for Pearson's correlation or when the relationship is not linear.
- Spearman's correlation calculates the degree to which the ranks of one variable correspond to the ranks of another.

Causation

- Correlation describes the association between variables.
- Correlation, however strong, does not imply causation, but it is one important aspect of inferring causality.
- A statistically significant relationship between two variables does not mean they are causally linked.

Causation

John Stuart Mill's conditions for establishing causality are:

- Temporal precedence: The cause must come before the effect.
- Covariance: The cause and effect must be related.
- Disqualification of alternative explanations: No other variable can explain the observed relationship.

Causation

$r = 0.7$

Number of ice scream sales x Number of people drowning

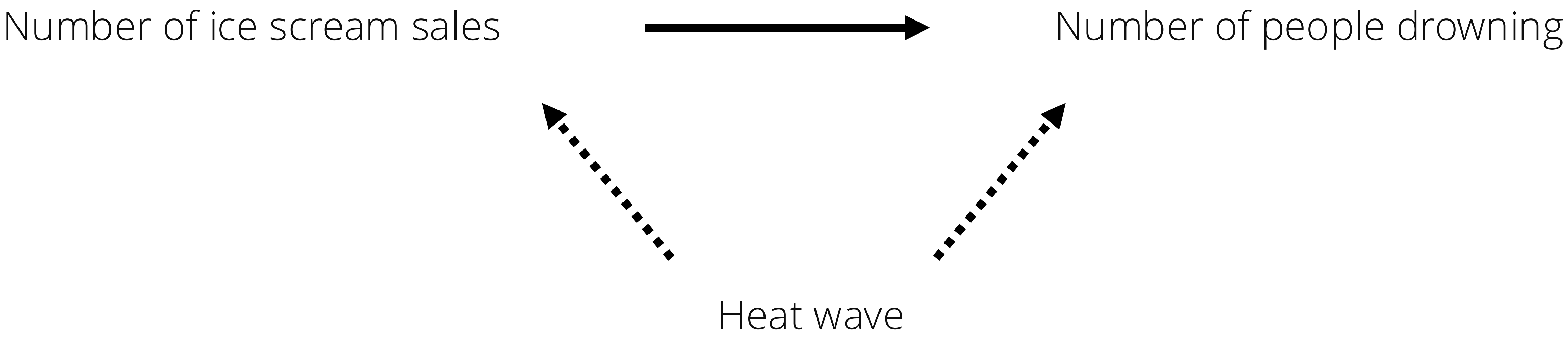
Causation

$r = 0.7$

Number of ice scream sales $\longrightarrow$ Number of people drowning

Causation

r = 0.7



Causation

$r = 0.8$

Number of fire fighters x Damage caused by the fire

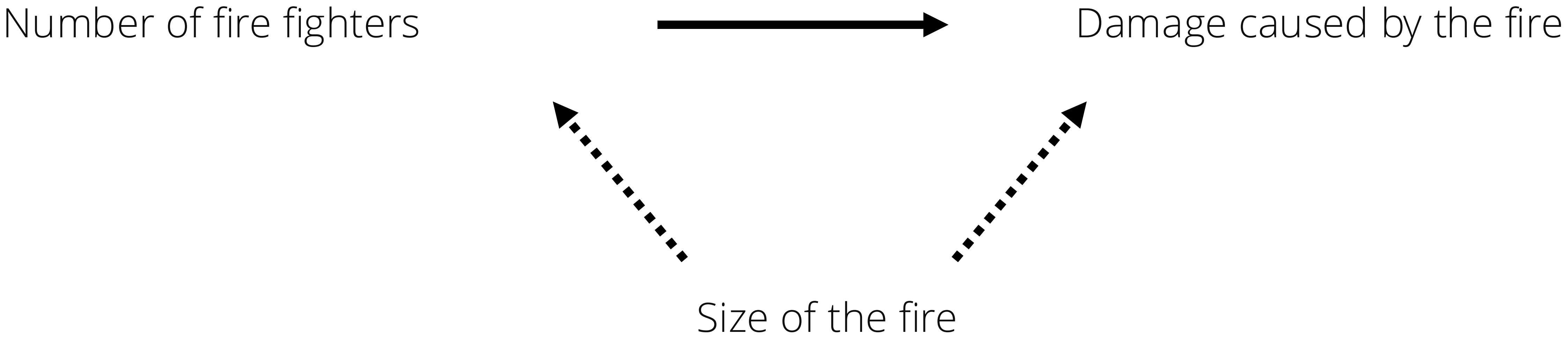
Causation

$r = 0.8$

Number of fire fighters $\longrightarrow$ Damage caused by the fire

Causation

$r = 0.8$



Mentimeter

- Go to www.menti.com.
- Use code: 3503 0583

Break

Regression

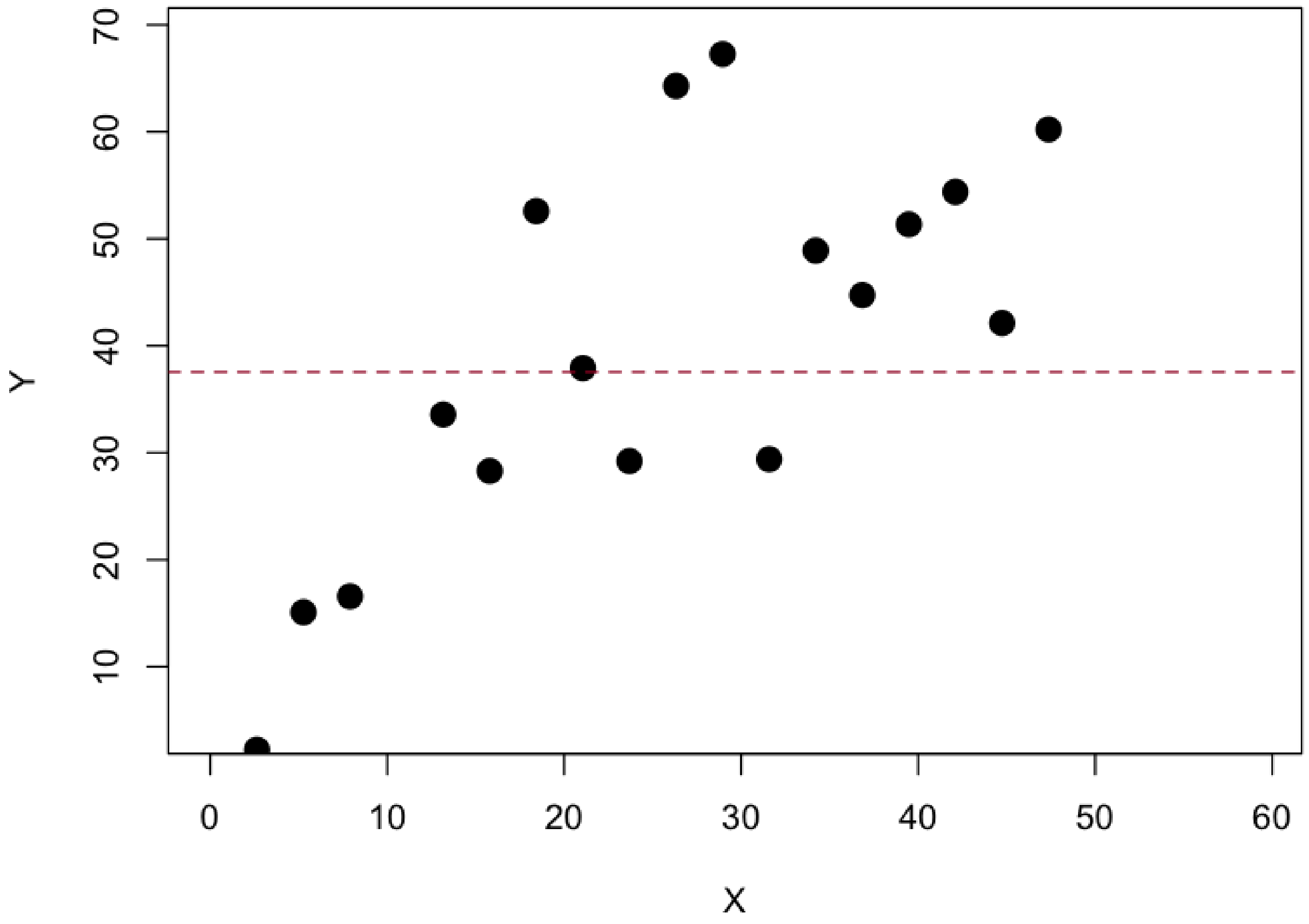
Bivariate regression

- We often want to know more than just whether two variables are related; we also want to predict how changes in one variable will affect another variable.
- To do this we can use a regression model to examine the relationship between a dependent variable (y) and one or more independent variables (x).

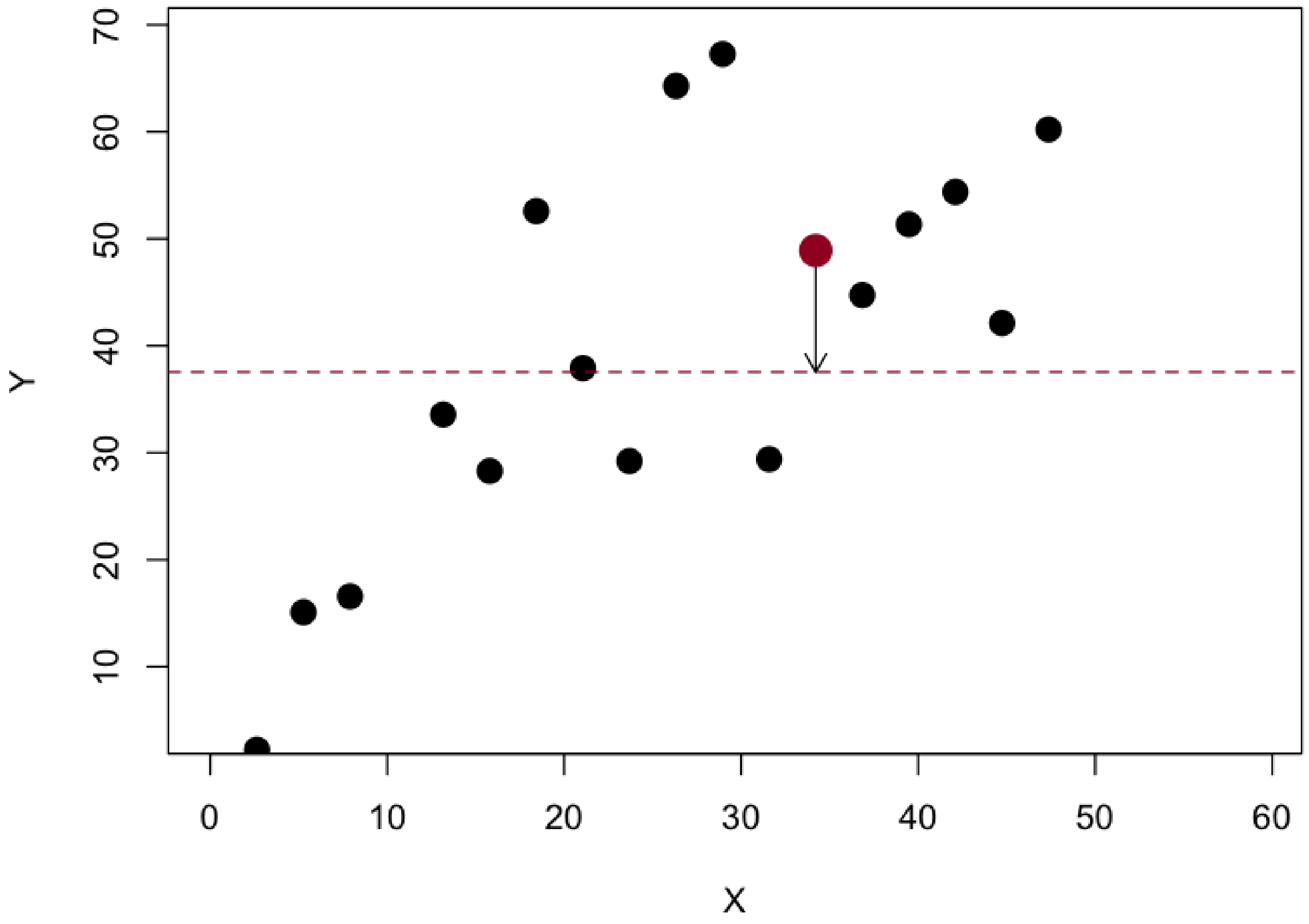
Bivariate regression

- Linear regression uses a line to summarise the relationship between x and y.
- The aim to find the line which best represents the relationships in the data.
- Typically, this line will not pass through every data point meaning we cannot predict y exactly.

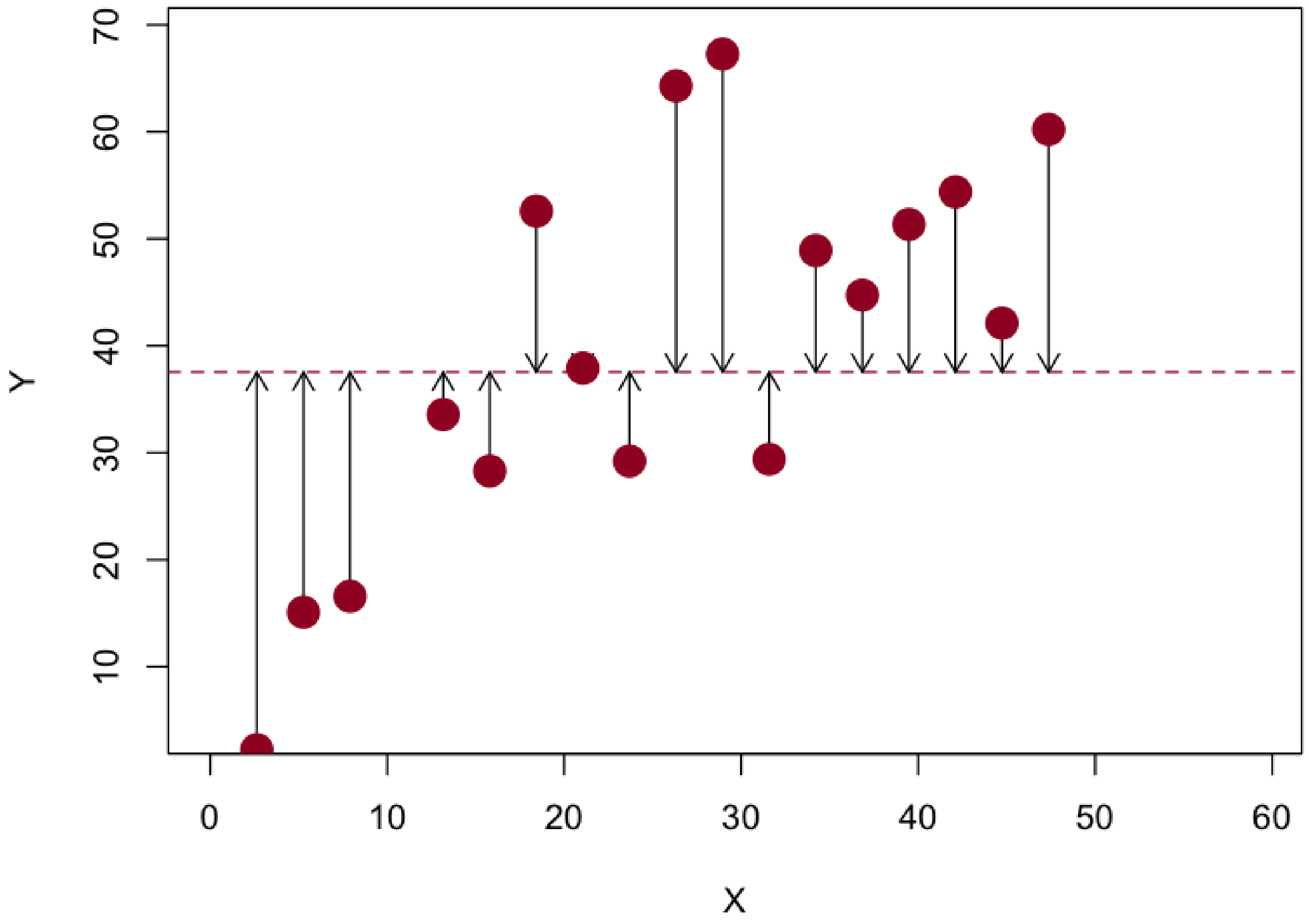
Null model



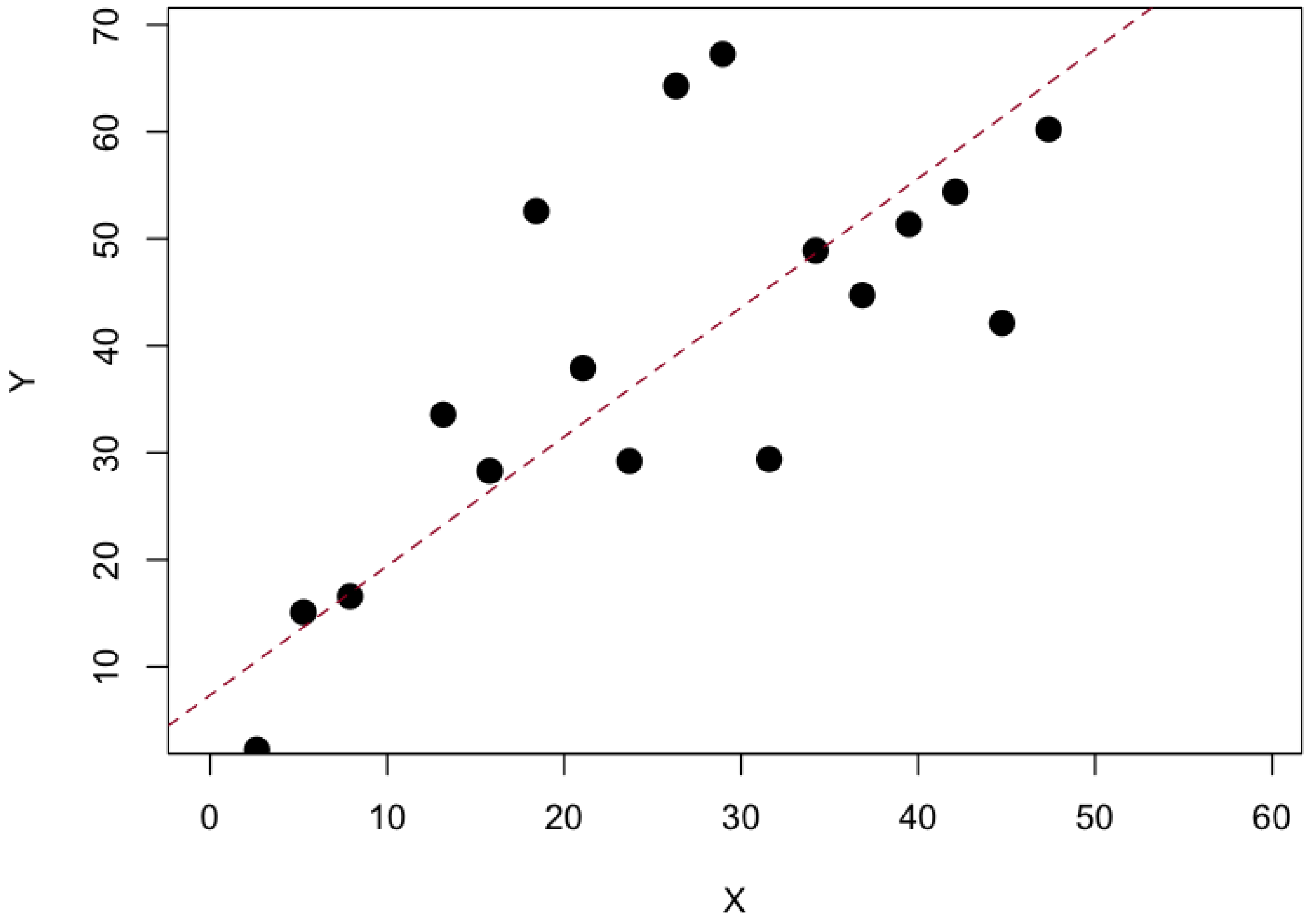
Sum of squared errors



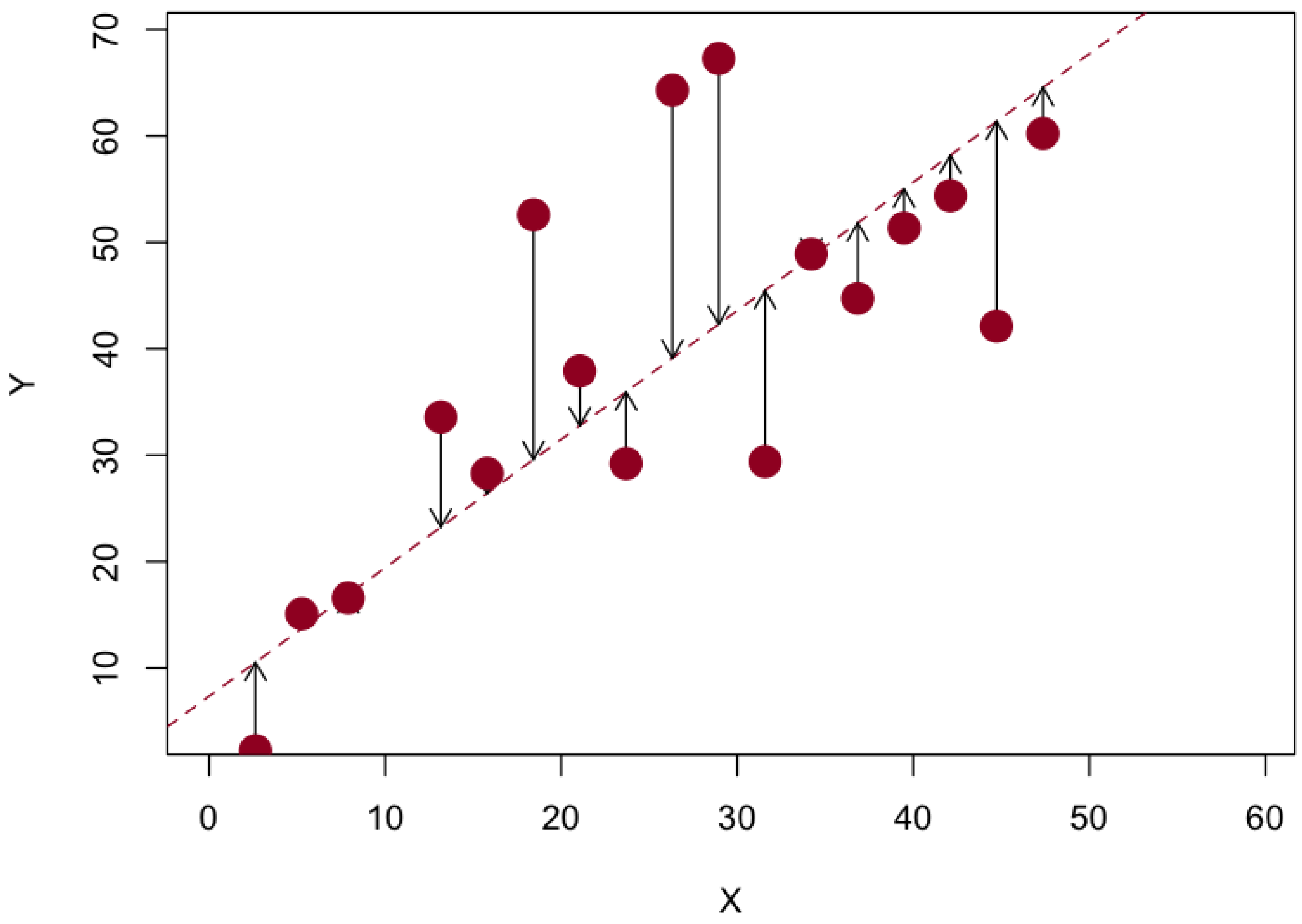
Sum of squared errors



Minimise sum of squared errors



Minimise sum of squared errors



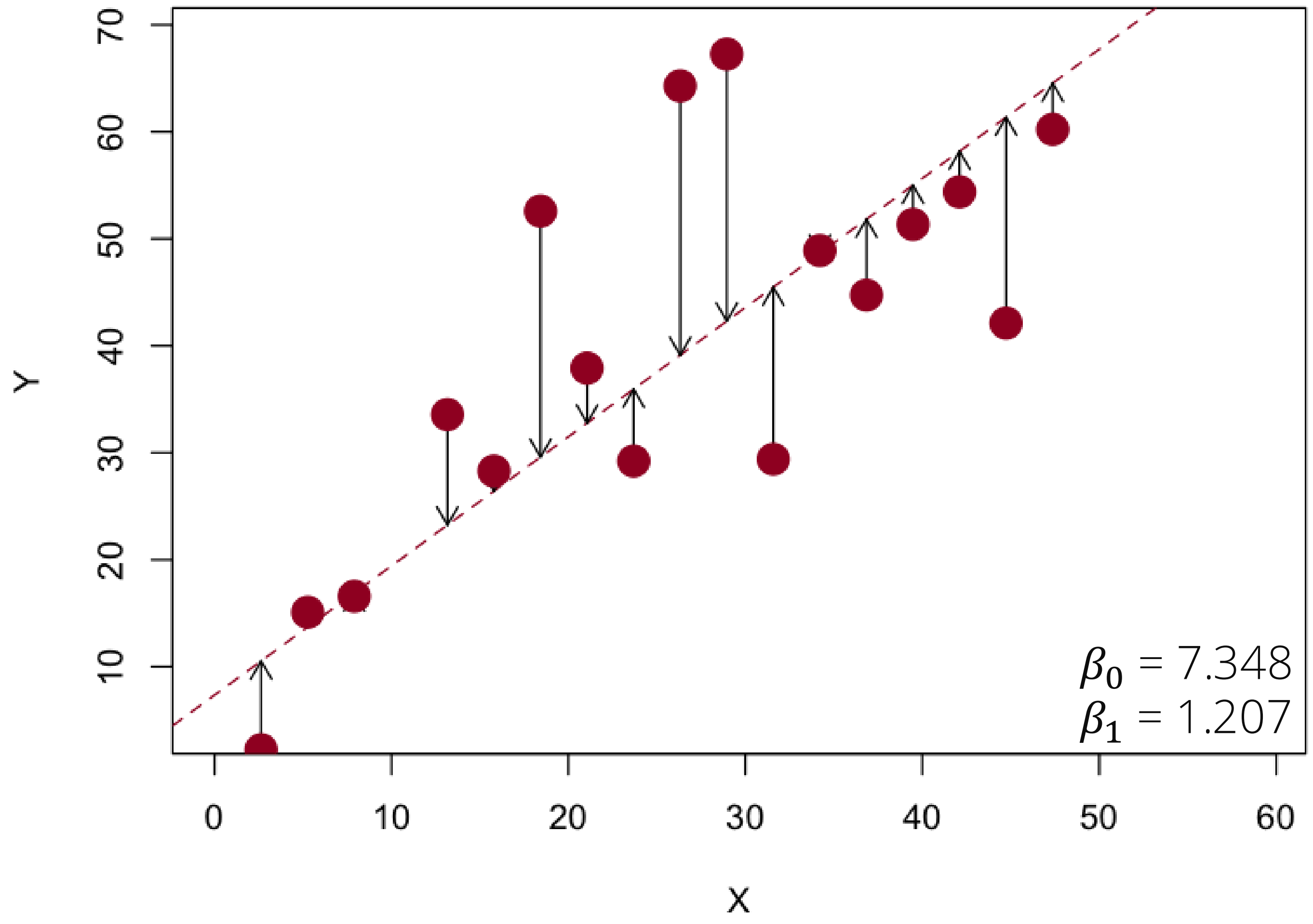
Bivariate regression

- Ordinary Least Squares (OLS) regression:

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

- The β terms are coefficients that define the regression line.
- The model estimates these parameters to find the line that gives the smallest sum of squared errors: Ordinary Least Squares (OLS) regression.

Minimise sum of squared errors



Regression with error term

- Ordinary Least Squares (OLS) regression:

$$y = \hat{y} + e = \ \beta_{0} + \beta_1 x + e$$

- The error term e captures the part of y that is not explained by the model, indicating how well the regression line fits the data.

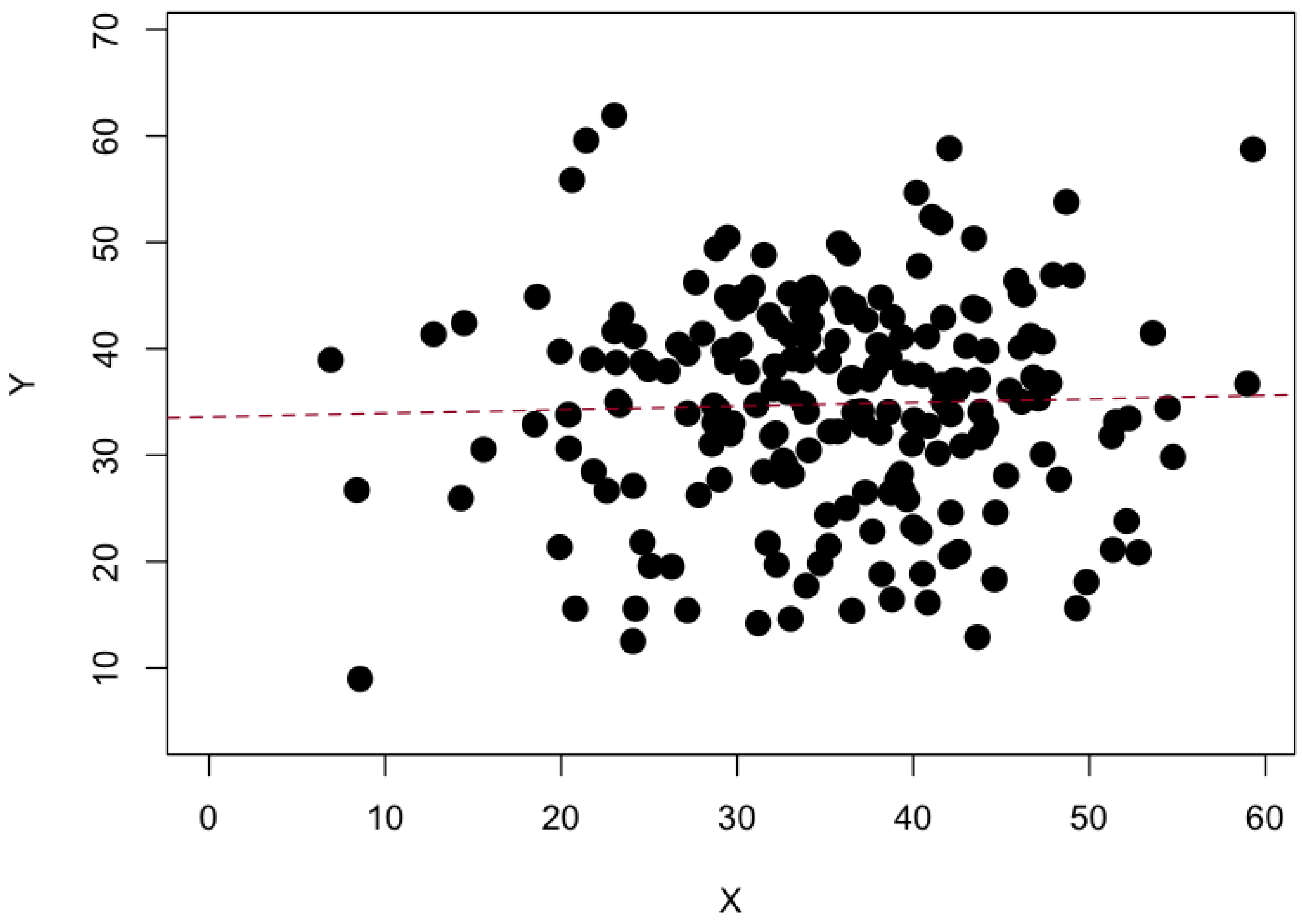
Model fit

- The R^2 measures the proportion of the variance in the dependent variable that is explained by the model.
- Values range from 0 to 1.
    - 0: The model explains none of the variance.
    - 1: The model explains all the variance.

Model significance

- The $\boldsymbol{F}$ statistic can be used to assess the significance of the model.
- Compares the sum of squared errors (SSE) of a baseline model (using only the mean) to the SSE of the proposed model to determine if the model provides a significantly better fit.
- A significant $\boldsymbol{F}$-test ($\boldsymbol{p}$-value < 0.05) suggests that the model explains more variance than the mean-only model and fits the data better.

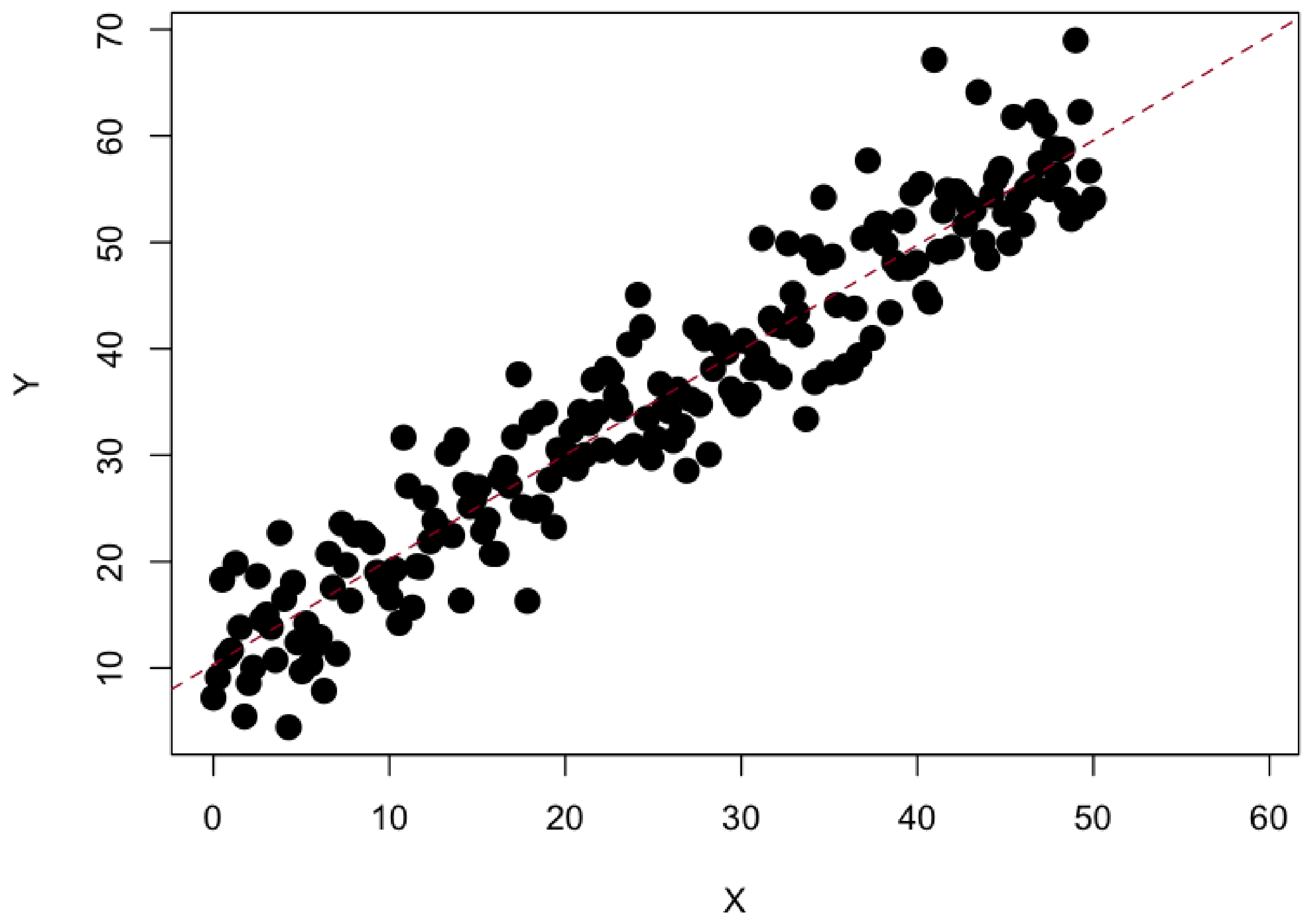
Bivariate regression



Model summary

	Dependent variable:
	y
x	0.034
	(0.075)
Constant	33.582***
	(2.761)
Observations	200
R^2	0.001
Adjusted R^2	−0.004
Residual Std. Error	10.267 (df = 198)
F Statistic	0.203 (df = 1; 198)
Note:	*p<0.1; **p<0.05; ***p<0.01

Bivariate regression



Model summary

	Dependent variable:
	y
x	0.985***
	(0.023)
Constant	10.340***
	(0.665)
Observations	200
R^2	0.902
Adjusted R^2	0.902
Residual Std. Error	4.722 (df = 198)
F Statistic	1,829.716*** (df = 1; 198)
Note:	*p<0.1; **p<0.05; ***p<0.01

Multiple regression

- Multiple regression allows for a more comprehensive explanation of variance in the dependent variable compared to bivariate models.
- Multiple predictors: It is often unrealistic to attribute variations in the dependent variable $\boldsymbol{y}$ to a single factor or independent variable $\boldsymbol{x}$.
- Separating effects: Multiple regression enables us to isolate and understand the individual effects of each predictor on the dependent variable.

Multiple regression

- As we add variables, the original regression equation changes:

$$\hat{y} = \beta_{0\,+}\ \beta_1\, x_1 + \beta_2\, x_2 + \cdots + \beta_n\, x_n$$

- Interpretation difference: In multiple regression, coefficients show the unique contribution of each predictor, isolating its effect from other predictors.

Model summary

	Dependent variable:
	y
x	0.984^{***}
	(0.023)
z	-0.016
	(0.035)
Constant	10.907^{***}
	(1.419)
Observations	200
R^2	0.902
Adjusted R^2	0.901
Residual Std. Error	4.732 (df = 197)
F Statistic	911.285^{***} (df = 2; 197)
Note:	*p<0.1; **p<0.05; ***p<0.01

LIVE

Assumptions

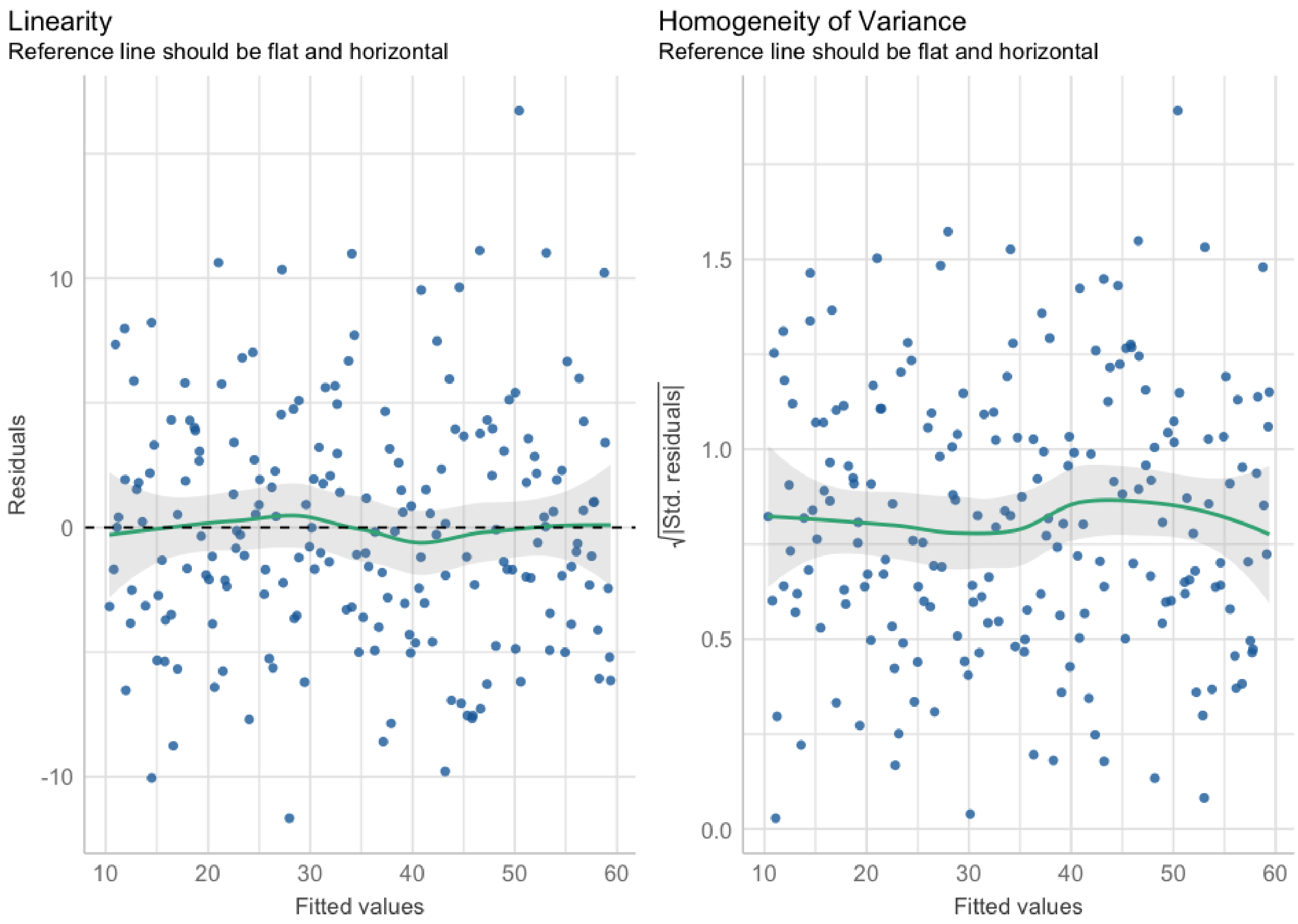
Assumptions

- There are several assumptions that must be satisfied in order to generalise our estimates beyond a sample.
- We can still use OLS when the assumptions are violated, but the estimates may be biased or inefficient.

Assumptions

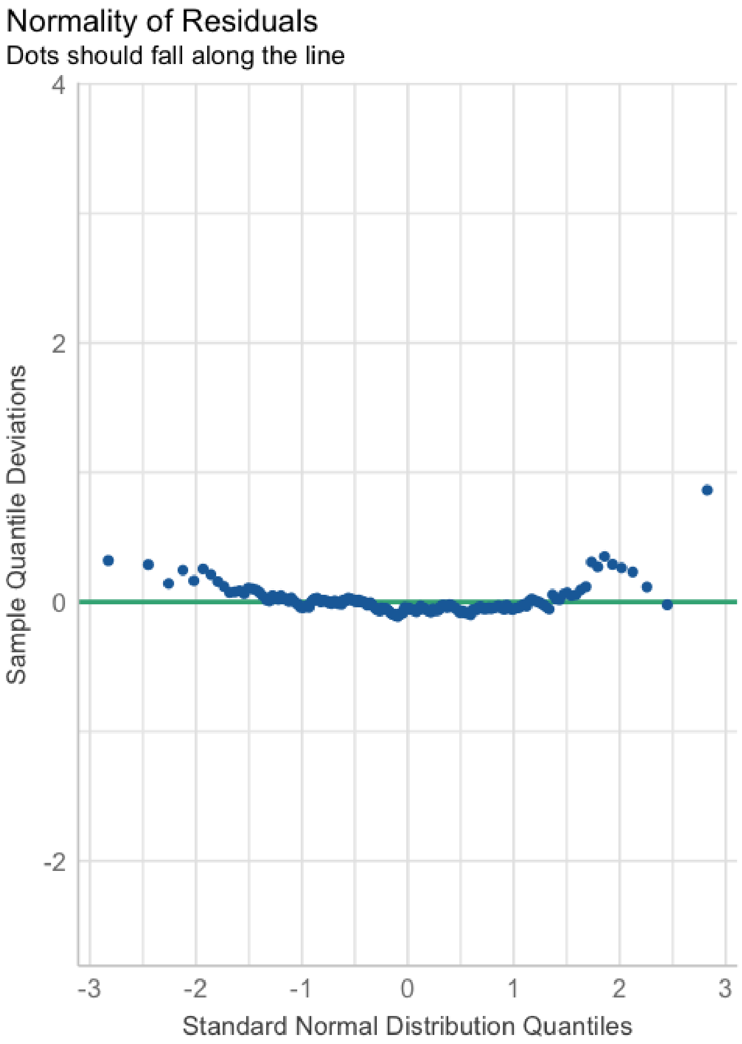
- Linearity: The relationship between the independent and dependent variables is linear.
- Independence: Observations are independent of each other; no correlation exists between the error terms.
- Homoscedasticity: The variance of error terms is constant across all levels of the independent variables (no heteroscedasticity).
- Normality of errors: The error terms are normally distributed.
- No multicollinearity: Independent variables are not strongly correlated with each other.

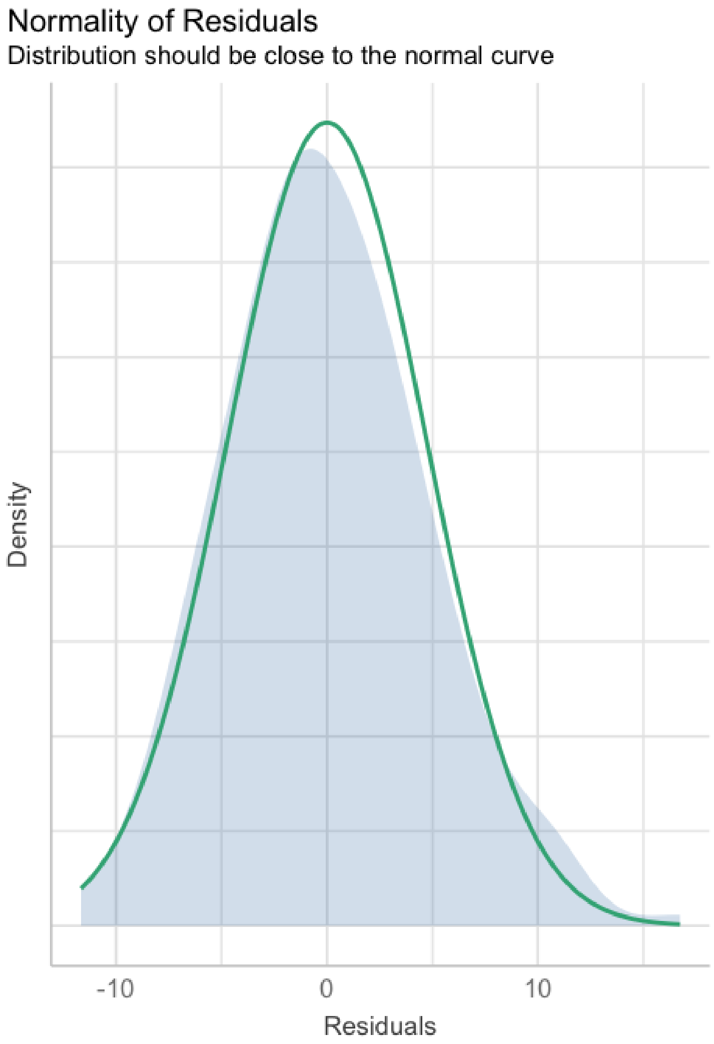
Assumptions



Output from the *easystats* R package.

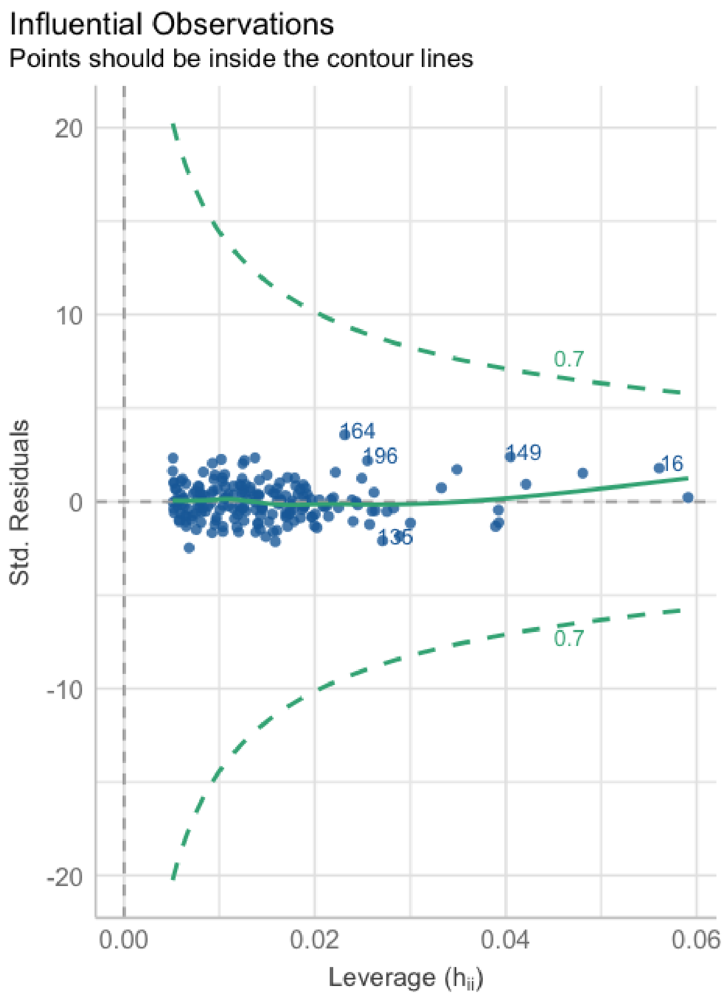
Assumptions

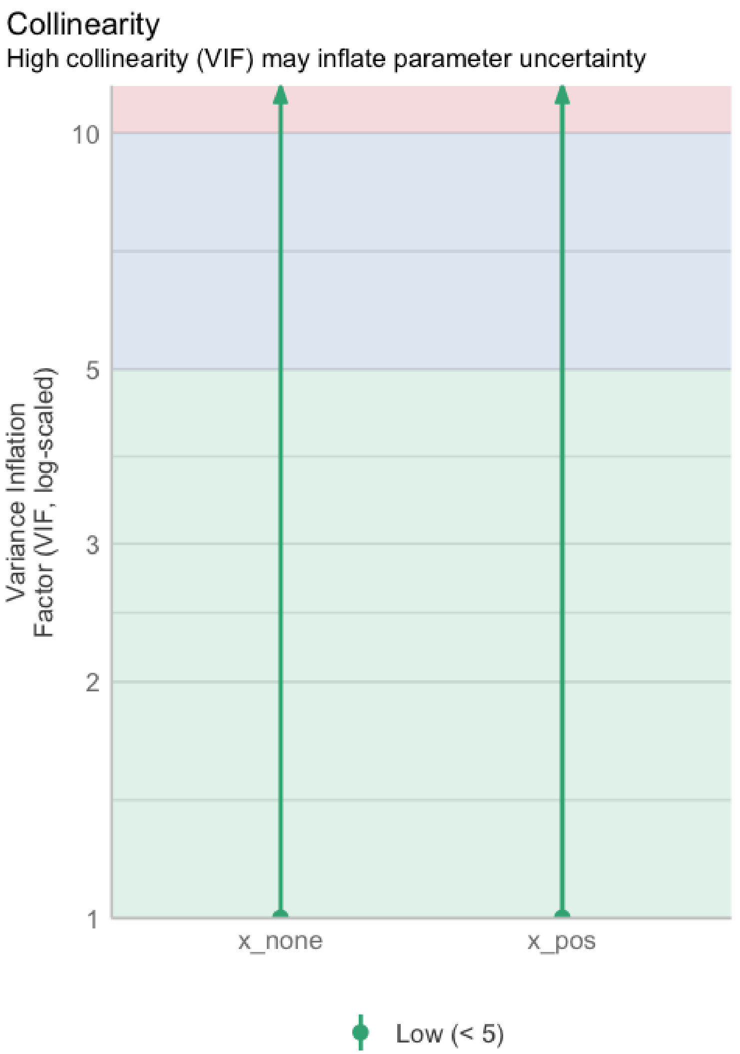




Output from the *easystats* R package.

Assumptions





Output from the *easystats* R package.

More regression?

- OLS regression on non-linear patterns (like curves): interaction terms, polynomials.
- Logistic regression to model the relationship between a binary dependent variable and one or more independent variables, estimating the probability of the outcome occurring.
- Poisson regression to model count data, where the dependent variable represents the number of events occurring within a fixed interval, assuming the events occur independently.

More regression?

- Time series regression for using historical data points collected over time to predict future values of a dependent variable, taking into account trends, seasonality, and autocorrelation.
- Multilevel regression to accounts for data with hierarchical or nested structures, allowing for the analysis of relationships at multiple levels.
- Geographically Weighted Regression to analyse spatial variations in relationships between variables by estimating coefficients for each observation based on its geographic location.

Summary

Summary

- Crosstabulation and chi-square statistic are useful tools for examining the association between categorical variables.
- For continuous variables, regression analysis is often used to quantify the relationship between variables.
- There are numerous variants of regression, each suited for different types of data and research questions.

Questions

Justin van Dijk

j.t.vandijk@ucl.ac.uk