Fundamentals of Quantitative Biology

STATS 02: Correlation and Linear Regression

Winter semester 2023/2024

Multivariate distributions

Multivariate distributions

- So far we have dealt exclusively with univariate distributions, i.e. distributions are concerned with only one variable.
- But we are often interested in several characteristics on the same examination object (individual), i.e. multivariate distributions
- Multivariate distributions describe the behavior of multiple random variables simultaneously capturing relationships, dependencies, and interactions between these variables.

Example: Joint distribution

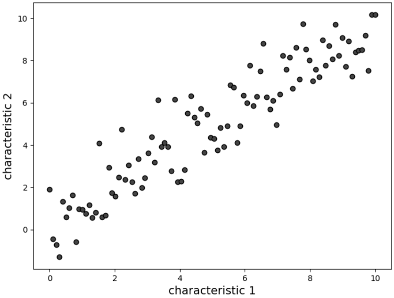
In scRNA-seq, to which gene and which cell does the transcript belong?

		cell	
		C_1	C_2
gene	A	$f(A \cap C_1)$	$f(A \cap C_2)$
	B	$f(B \cap C_1)$	$f(B \cap C_2)$
	$\vdots$	$\vdots$	$\vdots$
	X	$f(X \cap C_1)$	$f(X \cap C_2)$

This is called a **bivariate distribution.**

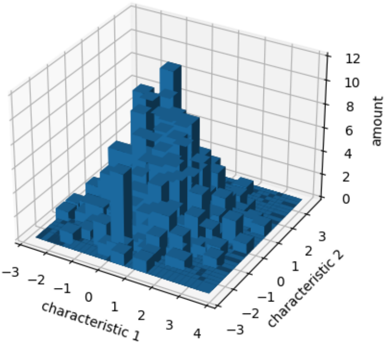
- methods for graphical representation of such distributions
- description of the relationship between two characteristics

Graphical representation of a bivariate distribution



scatter plot

Graphical representation of a bivariate distribution



bivariate histogram

Discrete case: joint distribution

Consider discrete random variables X and Y, where X takes on the values $x_1, \ldots, x_n$ and Y takes on the values $y_1, \ldots, y_m$. Then X and Y have the **joint distribution**

$$p_{ij} = \mathbb{P}(\{X = x_i\} \cap \{Y = y_j\}).$$

The **marginal probability distributions** $p_{X,i}$ for the random variable X and $p_{Y,j}$ for the random variable Y are

$$p_{X,i} = \sum_{j=1}^{m} p_{ij} \qquad \text{(row sum)}$$

$$p_{Y,j} = \sum_{i=1}^{n} p_{ij} \qquad \text{(column sum)}$$

Sample joint distribution

The joint distribution table lists transcripts per cell (characteristic A) and per gene (characteristic B).

		characteristic A				
		1	2	...	r	Σ
characteristic B	1	n_{11}	n_{12}	...	n_{1r}	$n_{1\cdot}$
	2	n_{21}	n_{22}		n_{2r}	$n_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	c	n_{c1}	n_{c2}	...	n_{cr}	$n_{c\cdot}$
	Σ	$n_{\cdot 1}$	$n_{\cdot 2}$		$n_{\cdot r}$	N

- sample joint distribution: n_{ij}
- marginal distributions:
  $n_{\cdot j} = \sum_{i=1}^{c} n_{ij}$
  (column sum)
  $n_{i\cdot} = \sum_{j=1}^{r} n_{ij}$
  (row sum)
- total sample size
  $N = \sum_{i,j} n_{ij}$

Independent random variables

Recall that two events A and B are independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

- Is there an equivalent concept for random variables?

Definition of independence of two discrete random variables

Consider two random variables X and Y, with probability mass functions $\mathbb{P}(X = x)$ and $\mathbb{P}(Y = y)$.

- If
$$\mathbb{P}(\{X = x\} \cap \{Y = y\}) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$$
for all x and y, then we say that X and Y are independent.
- If two random variables are not independent, then we say they are dependent.

Sums of independent random variables

Rules of sums of independent random variables:

- If X and Y are independent random variables, and $Z = X + Y$, then

$$\mathbb{E}[Z] = \mathbb{E}[X] + \mathbb{E}[Y] \quad \text{and} \quad \mathsf{Var}(Z) = \mathsf{Var}(X) + \mathsf{Var}(Y)\,.$$

Differences of independent random variables

This rule can be extended to the difference of independent random variables as following

- If X and Y are independent random variables, and $Z = X - Y$, then

$$\mathbb{E}[Z] = \mathbb{E}[X] - \mathbb{E}[Y] \quad \text{and} \quad \mathrm{Var}(Z) = \mathrm{Var}(X) + \mathrm{Var}(Y)\,.$$

Linear combination of two independent random variables

The general rule for two independent random variables is as follows

- If X and Y are independent random variables, a and b are constants, and $Z = aX + bY$, then

$$\mathbb{E}[Z] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y] \quad \text{and} \quad \mathrm{Var}(Z) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y)\,.$$

Correlation analysis

What about dependent random variables?

Similar rules apply for the cases where the random variables are dependent.

Consider the random variables X and Y and constants a and b.

- If $Z = aX + bY$, then

$$\begin{aligned} \mathbb{E}[Z] &= a\,\mathbb{E}[X] + b\,\mathbb{E}[Y] \quad \text{and} \\ \operatorname{Var}(Z) &= a^2 \operatorname{Var}(X) + b^2 \operatorname{Var}(Y) + 2ab \operatorname{cov}(X, Y). \end{aligned}$$

- Notice that if X and Y are independent, then $\operatorname{cov}(X, Y) = 0$ and we obtain the previous rules.

General rule

Finally consider n random variables, $X_1, X_2, \ldots, X_n$, and n constants $a_1, a_2, \ldots, a_n$ and Z is the linear combination

$$Z = \sum_{i=1}^{n} a_i X_i$$

then

$$\mathbb{E}[Z] = \sum_{i=1}^{n} a_i \mathbb{E}[X_i]$$

and

$$\mathrm{Var}(Z) = \sum_{i=1}^{n} a_i^2 \,\mathrm{Var}(X_i) + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ i \neq j}}^{n} a_i a_j \,\mathrm{cov}(X_i, X_j).$$

Covariance

Let X and Y be two random variables.

- The **covariance** of X and Y, denoted $\text{cov}(X, Y)$, is defined as

$$\begin{aligned}\text{cov}(X, Y) &= \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \\ &= \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y],\end{aligned}$$

  where $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$.

This definition applies both to discrete and continuous variables.

Example: Joint Probability Table

Consider the joint probability distribution of two random variables X and Y given by the following table:

$X \backslash Y$	$Y = 1$	$Y = 2$	$P(X)$
$X = 1$	0.1	0.2	0.3
$X = 2$	0.3	0.4	0.7
$P(Y)$	0.4	0.6	1.0

- Marginal probabilities:

$$\mathbb{E}[X] = 1 \cdot 0.3 + 2 \cdot 0.7 = 1.7,$$
$$\mathbb{E}[Y] = 1 \cdot 0.4 + 2 \cdot 0.6 = 1.6.$$

Calculating $\mathbb{E}[XY]$ and Covariance

To calculate the expected value of the product $\mathbb{E}[XY]$:

- Use the definition:

$$\mathbb{E}[XY] = \sum_x \sum_y x \cdot y \cdot P(X = x, Y = y).$$

- Substitute the values from the joint probability table:

$$\begin{aligned}\mathbb{E}[XY] &= (1 \cdot 1 \cdot 0.1) + (1 \cdot 2 \cdot 0.2) \\ &\quad + (2 \cdot 1 \cdot 0.3) + (2 \cdot 2 \cdot 0.4) \\ &= 2.7.\end{aligned}$$

Covariance Calculation:

- The covariance formula is:

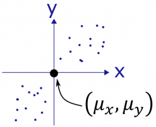
$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y].$$

- Substitute the values:

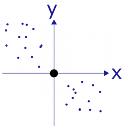
$$\mathrm{Cov}(X, Y) = -0.02.$$

Covariance

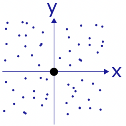
Intuition:



$\Rightarrow \text{cov}(X, Y) > 0$



$\Rightarrow \text{cov}(X, Y) < 0$



$\Rightarrow \text{cov}(X, Y) \approx 0$

Correlation

The **correlation coefficient** is denoted ρ and defined by

$$\rho = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y},$$

where

- σ_X is the standard deviation of X, and
- σ_Y is the standard deviation of Y.

Pearson's correlation

Consider data

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

Then the sample correlation is calulated as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

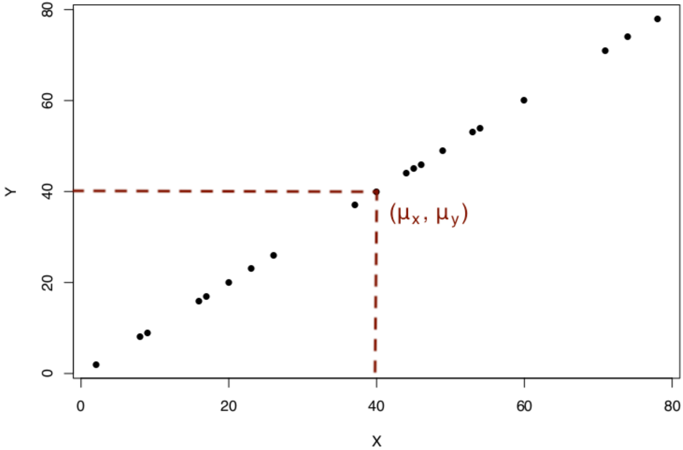
where

x_i	the measurement of feature X at the ith individual
y_i	the measurement of feature Y at the ith individual
$\bar{x}$ (or $\bar{y}$)	arithmetic mean of X (or Y)
n	number of data points (sample size)
i	index from 1 to n

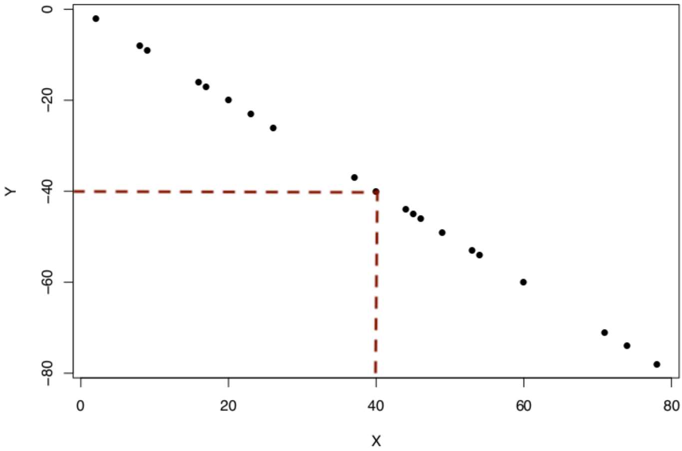
Properties of the correlation coefficient

1. $-1 \leq \rho \leq 1$
2. $\rho = 1$ if and only if the random variables X and Y satisfy a linear relationship $Y = aX + b$ for constants a and b with $a > 0$.
3. $\rho = -1$ if and only if the random variables X and Y satisfy a linear relationship $Y = aX + b$ for constants a and b with $a < 0$.
4. If X and Y are independent $\Rightarrow \rho = 0$.
5. $\rho = 0$ but X and Y are dependent.
6. If X and Y have a non-linear relationship then it can happen that $\rho = 0$.

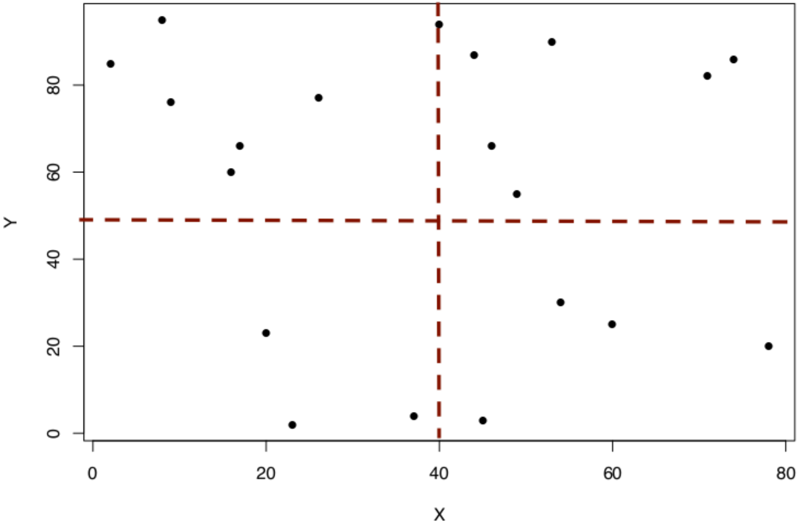
$\rho(X, Y) = 1$



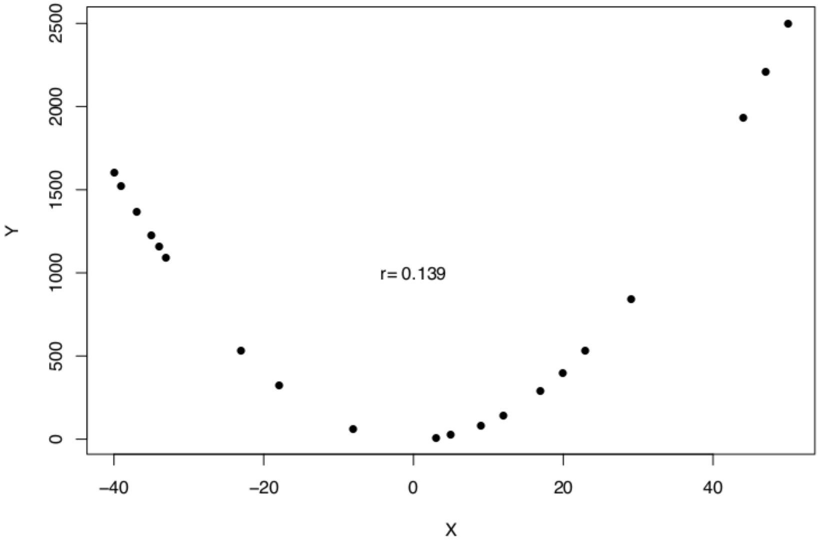
$\rho(X, Y) = -1$



$\rho(X, Y) = 0$



Counterexample with $\rho(X, Y) = 0$



Coefficient of determination

- The squared sample correlation coefficient r^2 is called the coefficient of determination.
- It can be interpreted as the proportion of the variation in the (response) variable Y that is explained by the linear relationship with the (explanatory) variable X,
- $r^2 \in [0, 1]$
- see regression

Spearman's Rank Correlation Coefficient

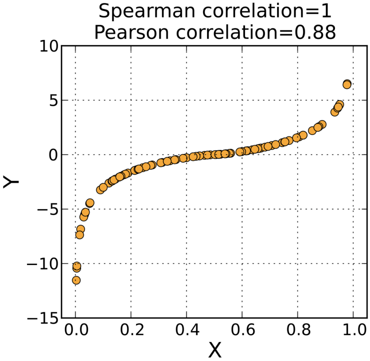
- The **Spearman correlation coefficient** ρ measures the strength and direction of a monotonic relationship between X and Y.
- It is computed using the formula:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)},$$

where $d_i = \text{rank}(X_i) - \text{rank}(Y_i)$ is the difference between the ranks of X_i and Y_i.

- ρ ranges from -1 (perfect negative monotonic relationship) to 1 (perfect positive monotonic relationship).
- Less sensitive to outliers compared to Pearson's correlation.
- Useful for non-linear but monotonic relationships.

Spearman's Rank Correlation Coefficient



Regression

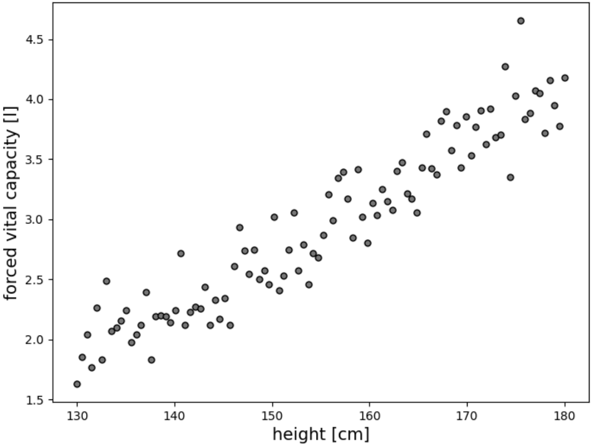
- The coefficients r and r^2 measures the strength of a relationship in bivariate distributions.
- We want to describe the relationship in a concise way.
- Let (x_i, y_i) be interval-scaled* pairs of measured values of the characteristics X and Y. We call X the independent (explanatory) and Y the dependent (response) variable.
- Regression determines a function to describe a relation between X and Y.

* Interval-scaled data means that differences between values can be meaningfully compared, whereas ordinal data may lead to inconsistencies.

Example - FVC data

To study lung function, the forced vital capacity (FVC) in litres and height in cm were measured on 127 twelve-year-old boys.

FVC data - scatter plot



FVC data - scatter plot

Response variable (FVC) measures an outcome of a study.

Explanatory variable (height) explains or causes changes in the response variable. Sometimes referred to as the predictor variable.

Simple linear regression model

Consider n observations of the explanatory variable x_i and the response variable y_i,

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

The simple linear regression model is

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma)$ is the individual deviation (error-) value assumed to distributed normally.

Estimation of the regression parameters

Consider the function

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2.$$

- Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the values of β_0 and β_1 that minimise $Q(\beta_0, \beta_1)$.
- $\hat{\beta}_0$ and $\hat{\beta}_1$ are called the least squares estimates of β_0 and β_1
- The line
$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$
is called the line of best fit.

Example

The simple linear regression model is fitted to the fvc data and the resultant fit is

$$FVC = -4.2 + 0.05 \times \text{Height}.$$

- What does the value -4.2 mean?
- What does the value 0.05 mean?
- What FVC does this line predict for a twelve-year-old boy of height 150cm?

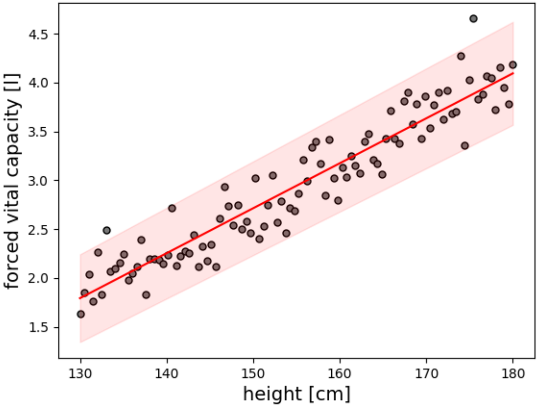
FVC data

```
import numpy as np

# Perform linear fit using numpy polyfit with covariance matrix
coefficients, covariance_matrix = np.polyfit(x, y, 1, cov=True)
slope, intercept = coefficients

# Extract standard errors from the covariance matrix
slope_error = np.sqrt(covariance_matrix[0, 0])
intercept_error = np.sqrt(covariance_matrix[1, 1])
print(f"Slope: {slope} ± {slope_error}")
print(f"Intercept: {intercept} ± {intercept_error}")
```

FVC data



Slope: 0.046 ± 0.002 Intercept: -4.19 ± 0.24

FVC data

For the FVC data $r = 0.95$.

- Thus $r^2 = 0.90$.
- i.e., 90% of the variation in FVC is explained by the linear relationship with Height.

Definition of residuals

Recall that the simple linear regression model is

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma)$ independently. We can obtain ε_i by

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_i).$$

As we do not know β_0 and β_1, we can estimate ε_i by

$$\hat{\varepsilon}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i).$$

The values $\hat{\varepsilon}_i$ which are calculated by observed - fitted are called the residuals.

Properties of residuals

The main use of the residuals is for model checking.

- If the regression model is true then

$$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$$

  must be independent $\mathcal{N}(0, \sigma)$ variables.

- In this case, the residuals

$$\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_n$$

  should look roughly like a sample of independent $\mathcal{N}(0, \sigma)$ variables.

Model checking

For the linear regression model, the assumptions can formulated as:

Modelling assumptions

1. Linearity: $\mathbb{E}[Y_i] = \beta_0 + \beta_1 x_i \Leftrightarrow \mathbb{E}[\varepsilon_i] = 0$.
2. Homoscedasticity (constant variance): $\mathrm{Var}(\varepsilon_i) = \sigma^2$ for all i.
3. Normality: $\varepsilon_i \sim \mathcal{N}(0, \sigma)$.

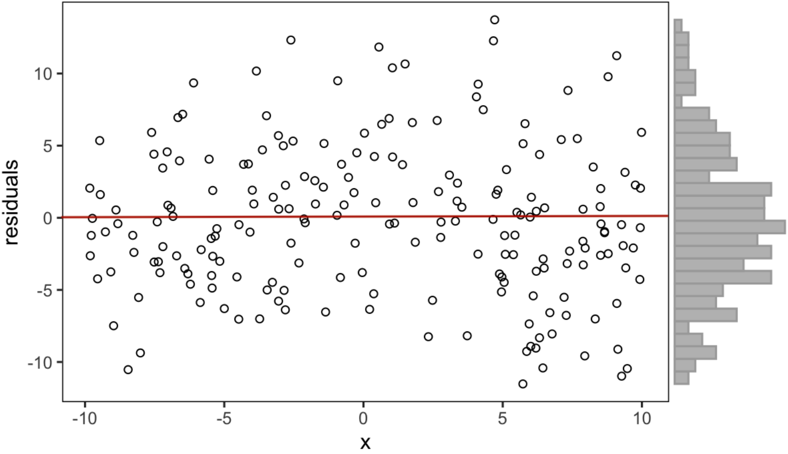
Design assumptions

4. Independence: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent.

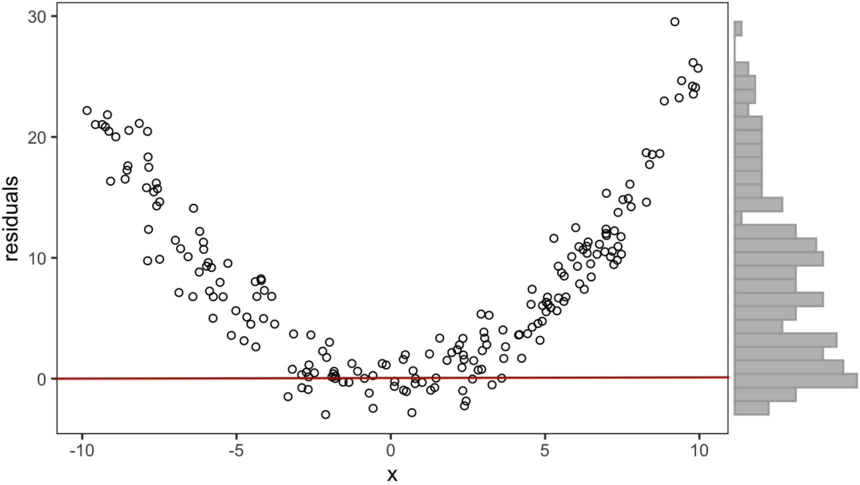
Linearity

- Check by looking at residual versus fitted plot
- If reasonable, expect to see points distributed symmetrically above and below the zero line.

Linearity valid



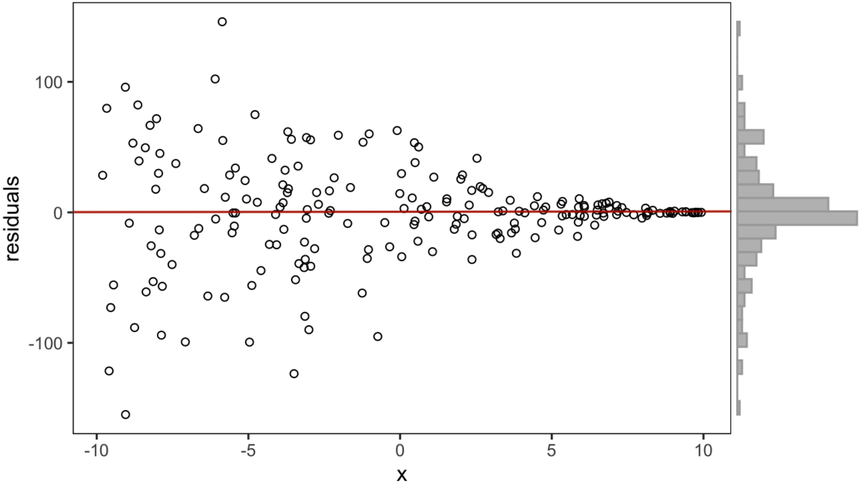
Linearity not valid



Homoscedasticity (constant spread)

- Check by looking at residual versus fitted plot
- If reasonable, expect to see roughly equal spread around zero line as we go from left to right.

Homoscedasticity not valid (Heteroscedasticity)



Q-Q Plot: Checking Normality

- A Q-Q (Quantile-Quantile) plot compares the observed residuals (errors) to the expected values from a normal distribution. It checks if residuals follow a normal distribution.
- **How is it made?**
  - **(1)** Sort the observed residuals from smallest to largest.
  - **(2)** Compute the theoretical quantiles.
  - **(3)** Plot the sorted residuals (y-axis) against the theoretical quantiles (x-axis).

Theoretical quantiles

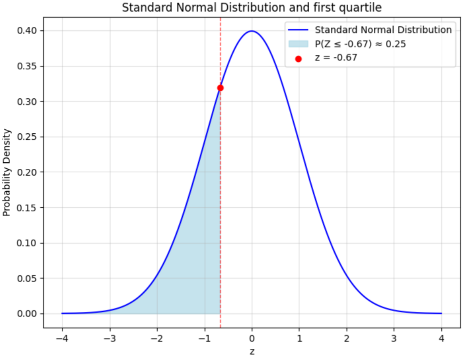
- The theoretical quantiles can be computed from the normal distribution directly. They correspond to the value of the cumulative probability given by

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \, dt$$

  where the mean $\mu = 0$ and the variance $\sigma = 1$.

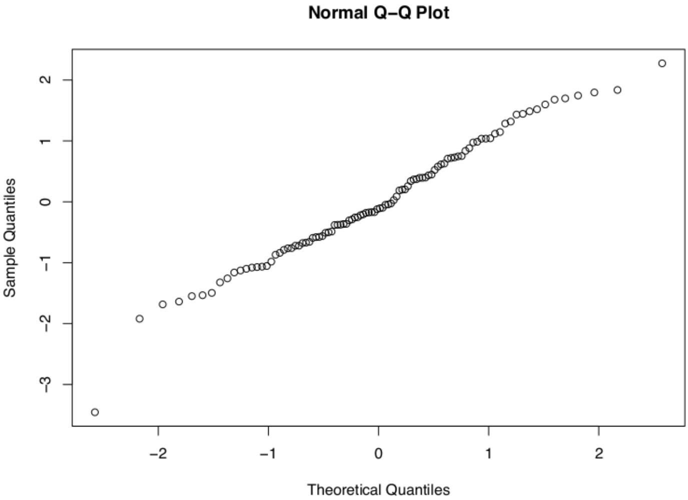
- To get the location z on the curve where the theoretical quantile q is reached we have to solve $\Phi(z) = q$ or $\Phi^{-1}(q) = z$.

Theoretical quantiles

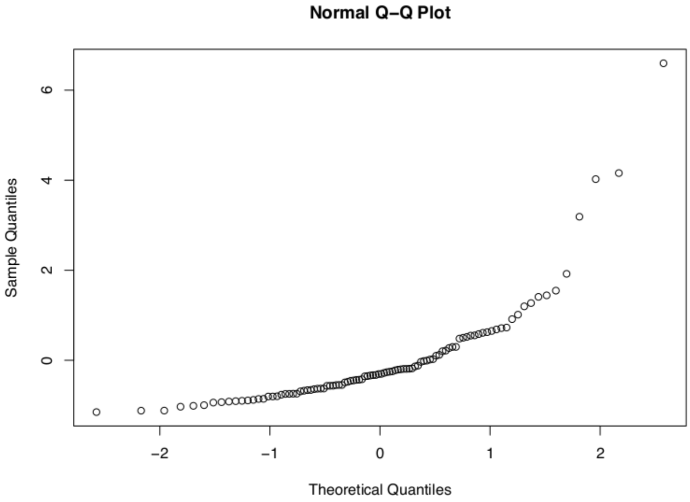


This Python function does the trick: `stats.norm.ppf([0.25])`

Normality valid



Normality not valid



Polynomial terms

We can also consider adding polynomial terms to the model.

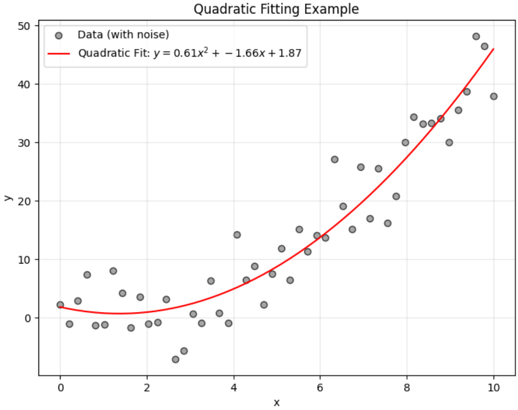
- e.g.,

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,1}^2 + \varepsilon_i,$$

  where $\varepsilon_i \sim \mathcal{N}(0, \sigma)$ independently.

Note: Regarding the relationship between X and Y, this is no longer a linear model. But the model is still linear in the parameters (i.e. the betas), so our estimation formulas still work!

Polynomial example



Here: $\beta_0 = 1.87$, $\beta_1 = -1.66$, and $\beta_2 = 0.61$,