

# Class 4b: Simple Linear Regression diagnostics

Business Forecasting



# Roadmap

## This class

- Testing assumptions behind residuals
  - Linearity
  - Constant Variance
  - Uncorrelated Residuals
  - Normality

Let's revisit the assumptions behind linear model:

1. Model is linear in the parameter and with additive error term
2.  $E(u_i) = 0$
3.  $Var(u_i) = \sigma^2$
4.  $cov(u_i, u_j) = 0$

Additional assumption needed for hypothesis testing and confidence intervals:

1.  $u_i \sim N(0, \sigma)$

# Assumptions

Visualizing residuals, we can test:

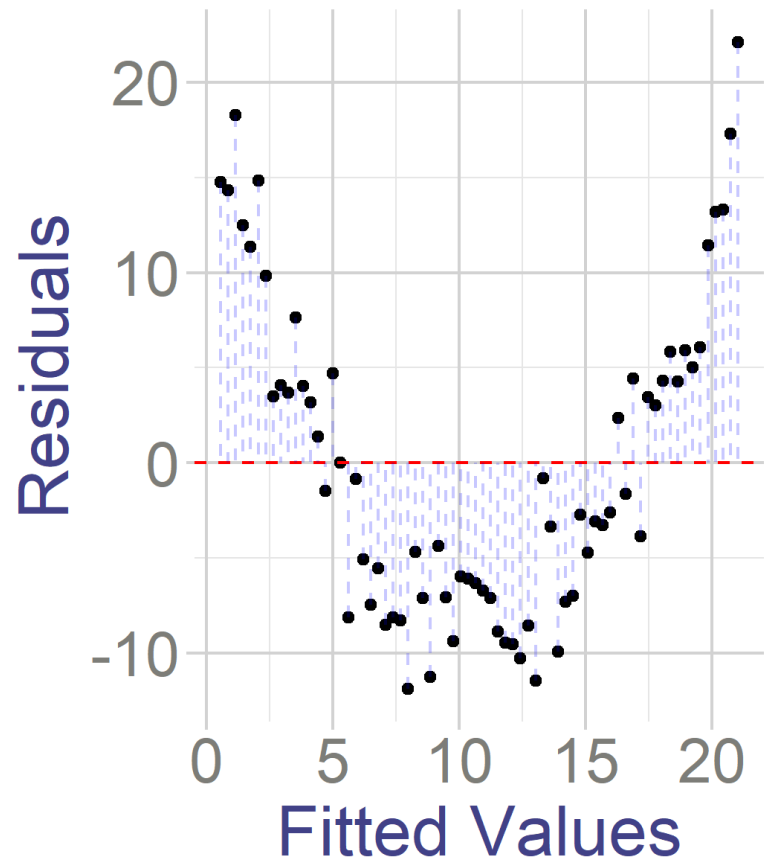
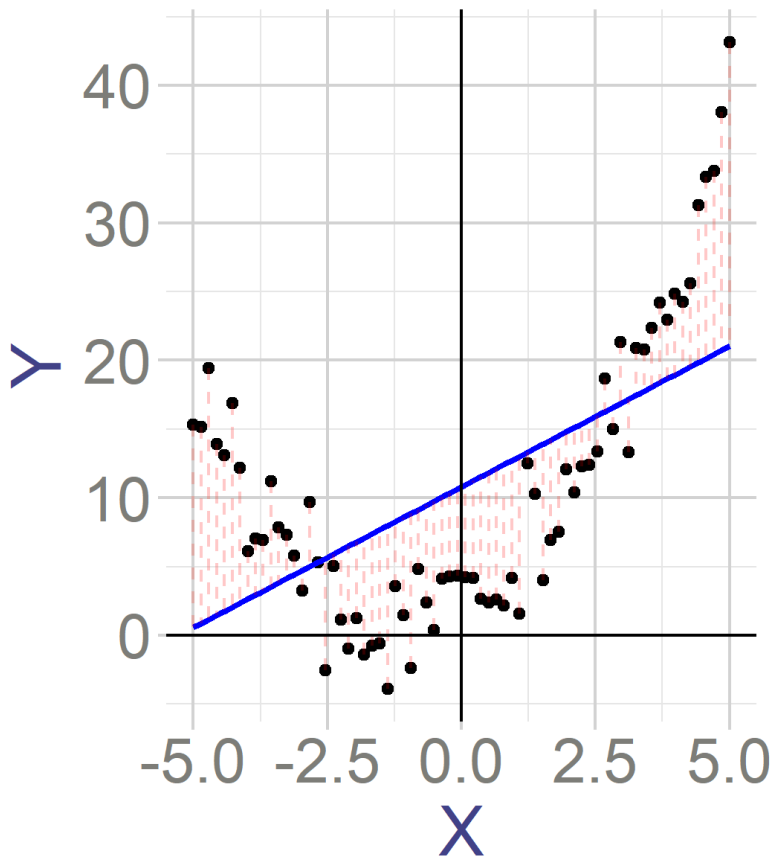
- Linearity
- Constant Variance (homoskedasticity)
- Uncorrelated errors

With vizuale and numerical tests we can analyze:

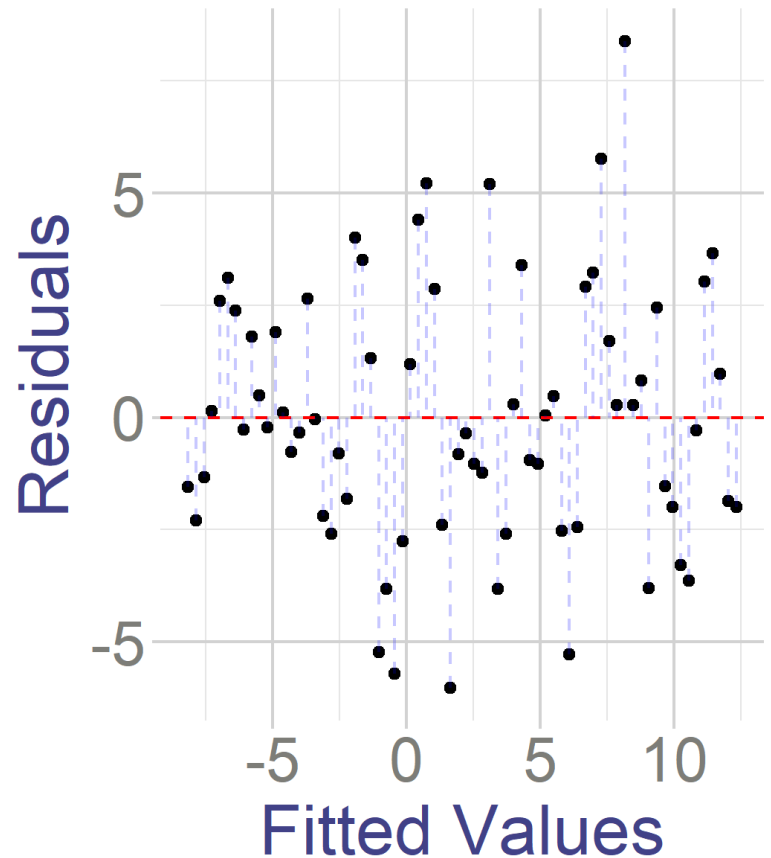
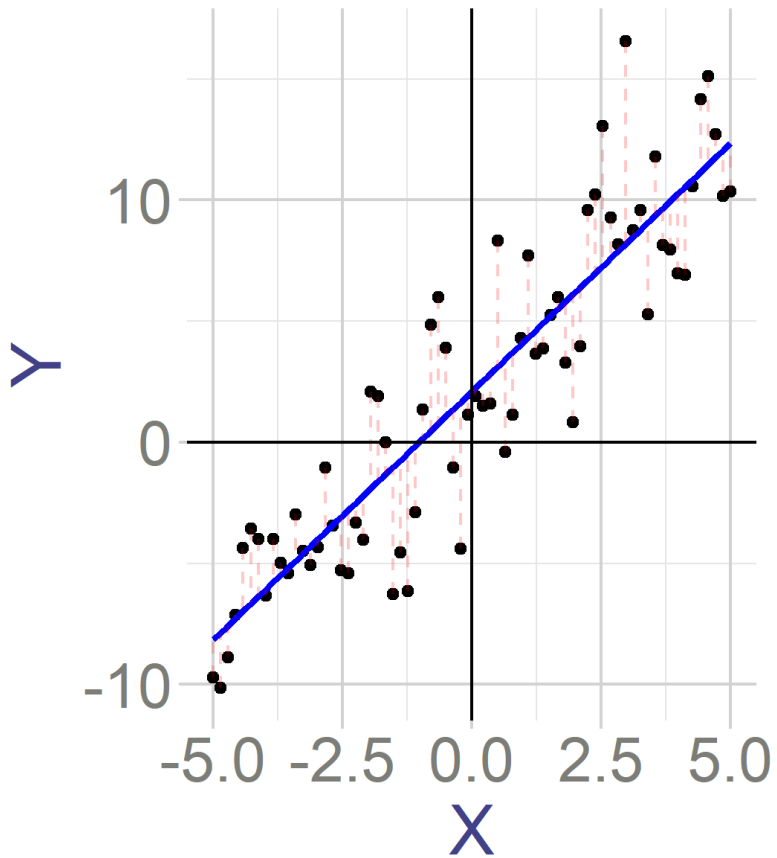
- Normality

# Linearity diagnostic

1. Plot residuals against fitted values of  $y$   $\hat{y}$
2. Check if residuals have systematic non-linear pattern



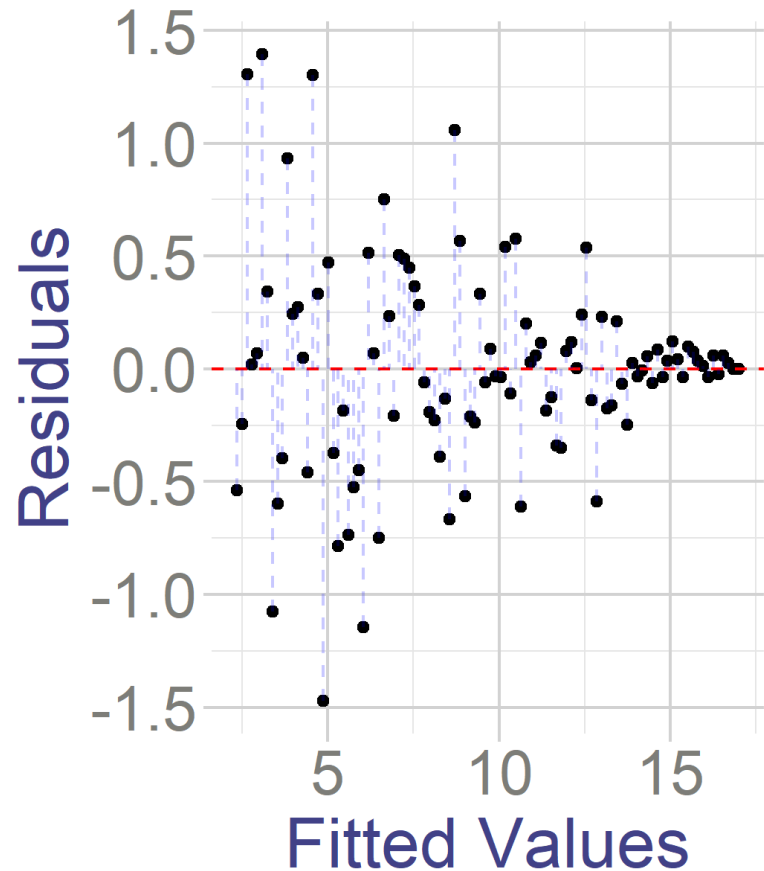
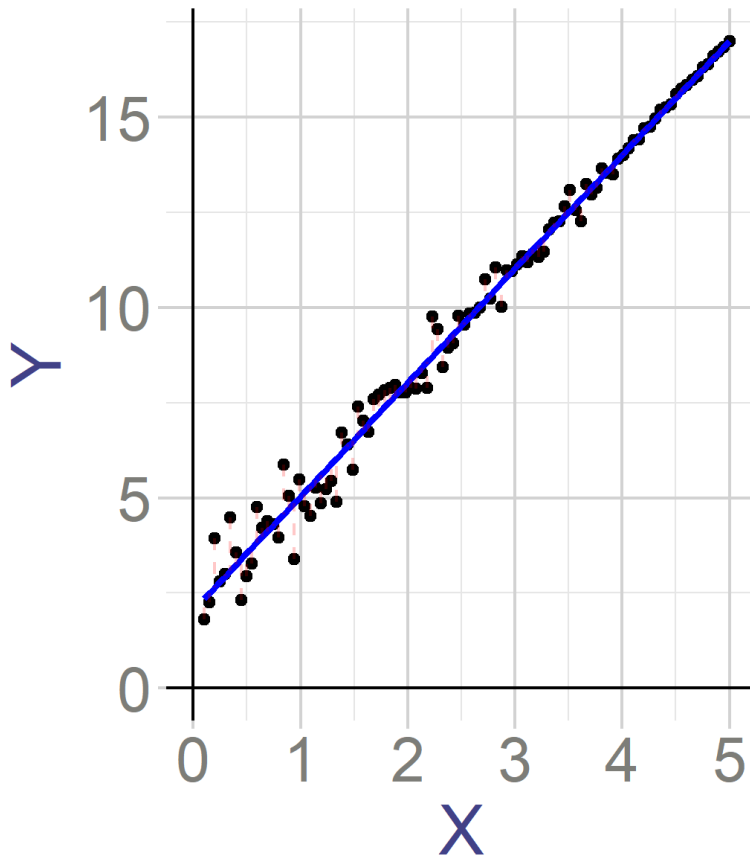
# Linearity diagnostic



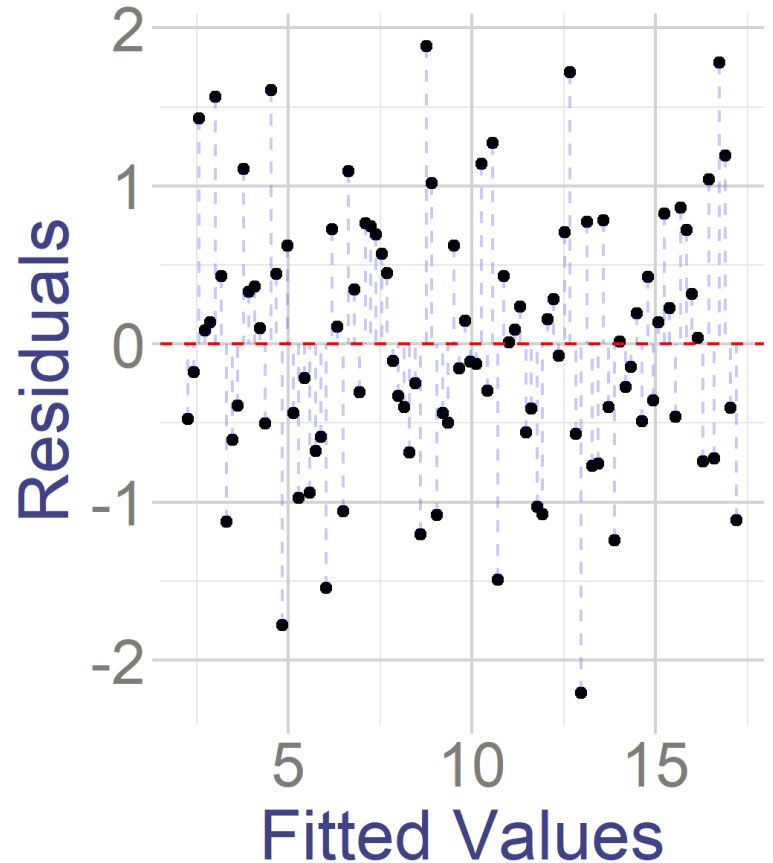
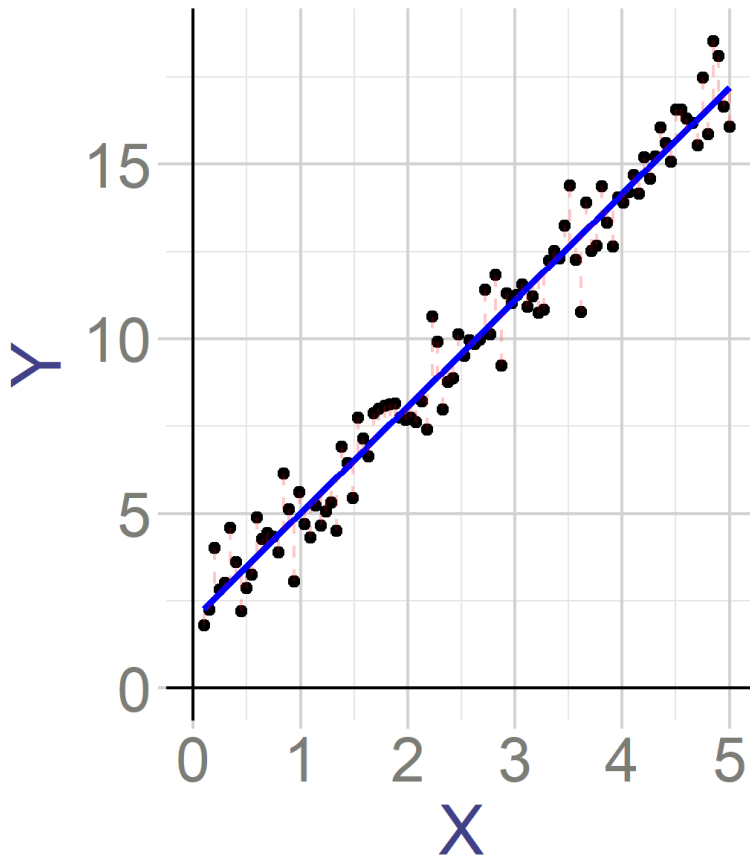


# Constant variance (homoskedasticity) diagnostic

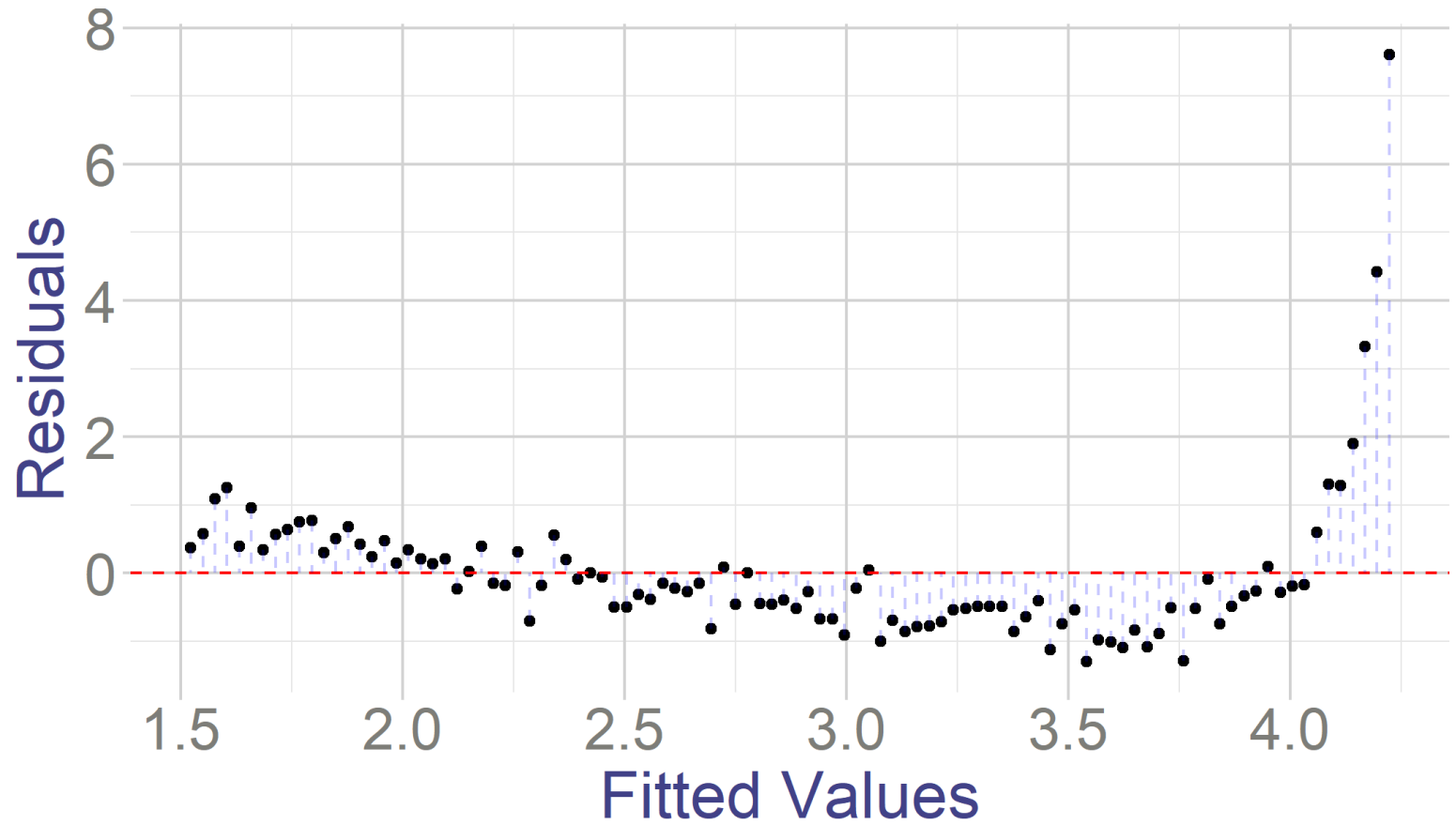
1. Plot residuals against fitted values of  $y$   $\hat{y}$
2. Check if variance changes as  $\hat{y}$  changes (heteroskedasticity)



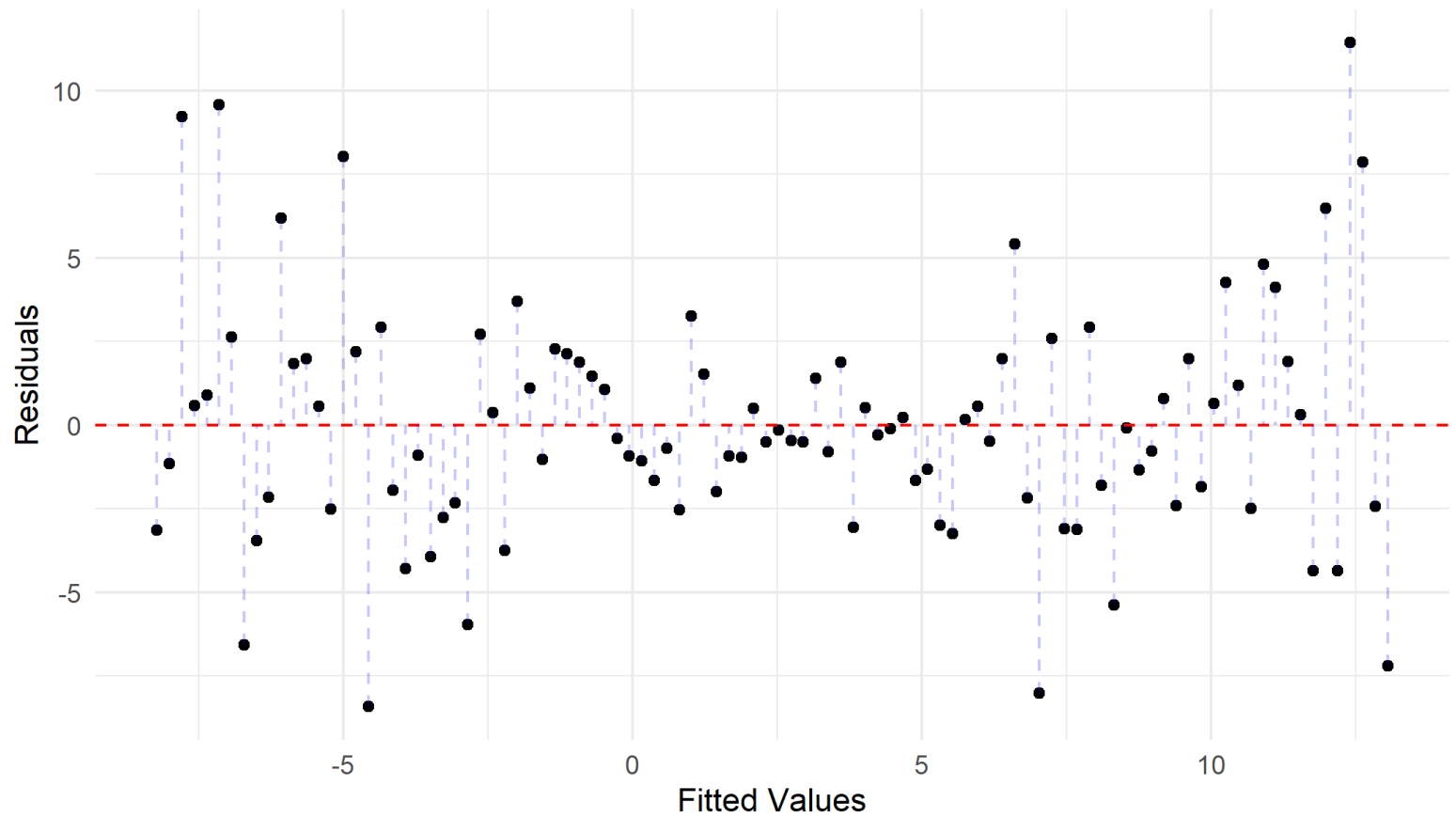
# Constant variance (homoskedasticity) diagnostic



What goes wrong here?

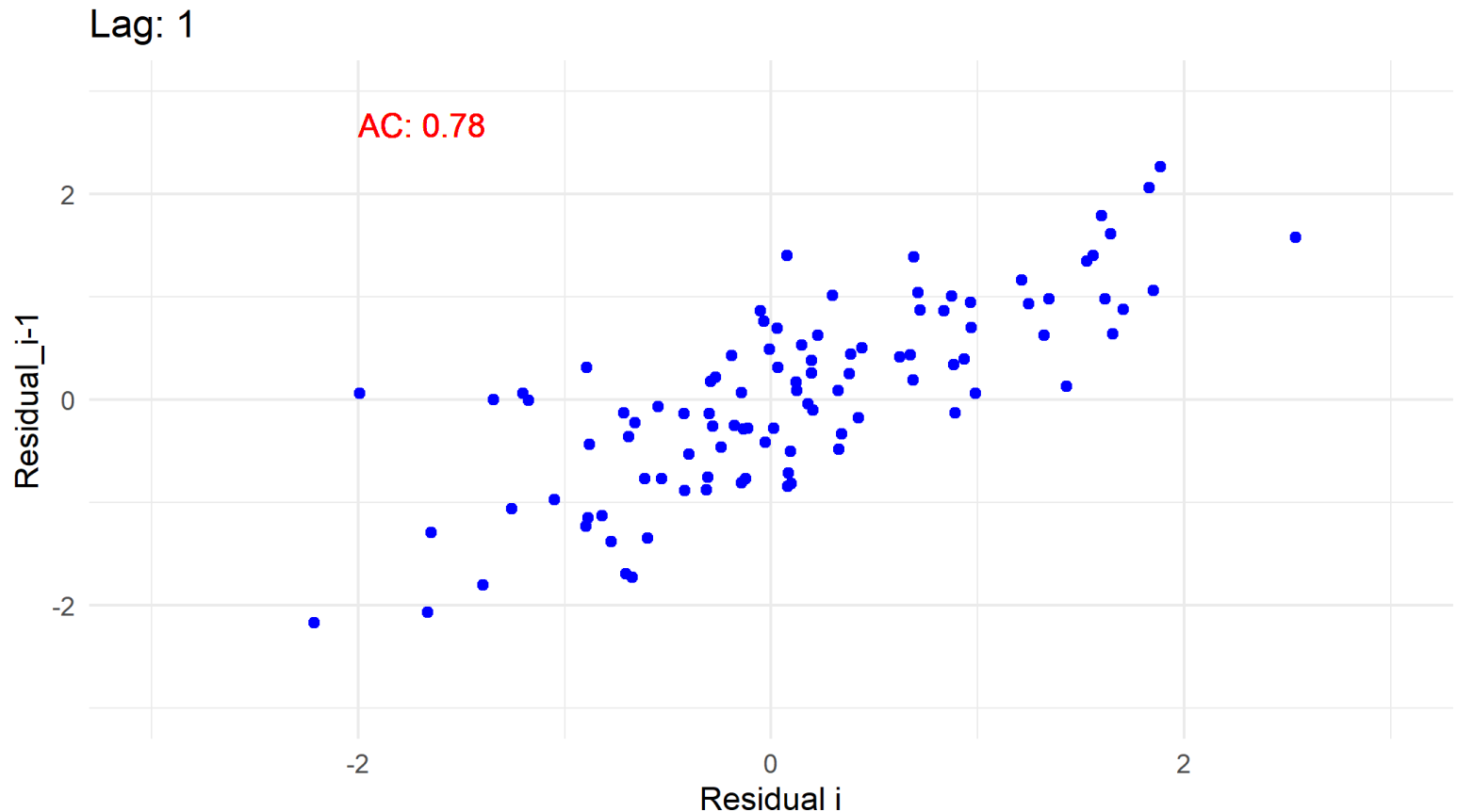


What goes wrong here?



# Correlation of error terms

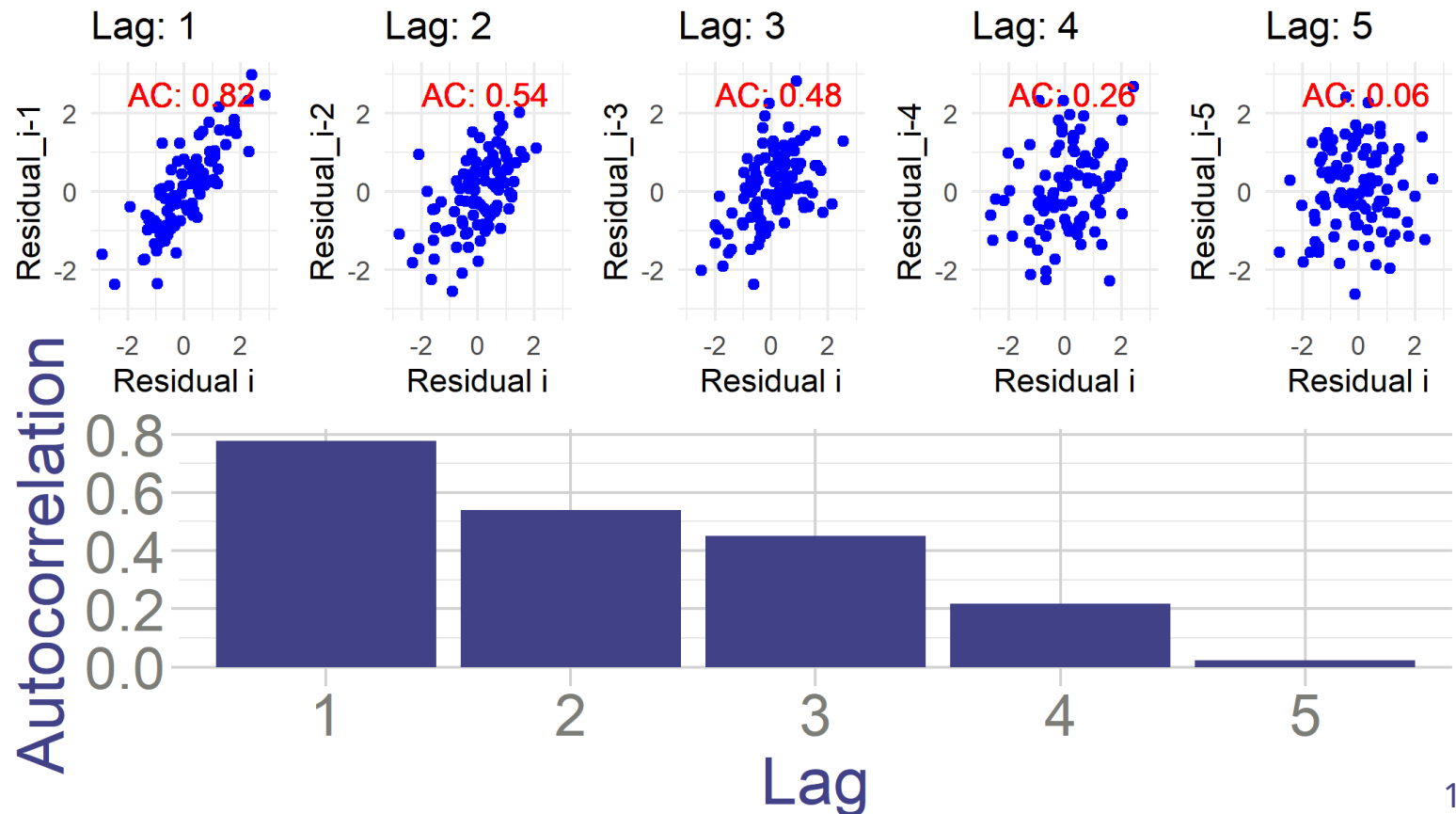
1. Plot fitted residuals vs their value in previous observation  $e_i$  vs  $e_{i-1}$
2. So  $e_2$  vs  $e_1$ ,  $e_5$  vs  $e_4$  etc



# Correlation of error terms

1. We can also calculate the correlation with other lags

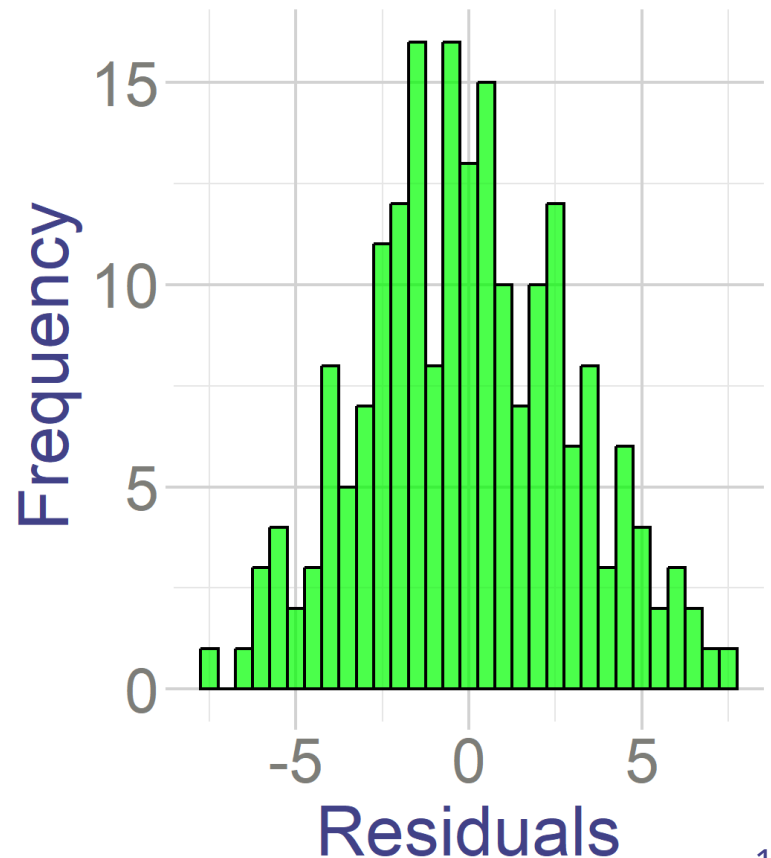
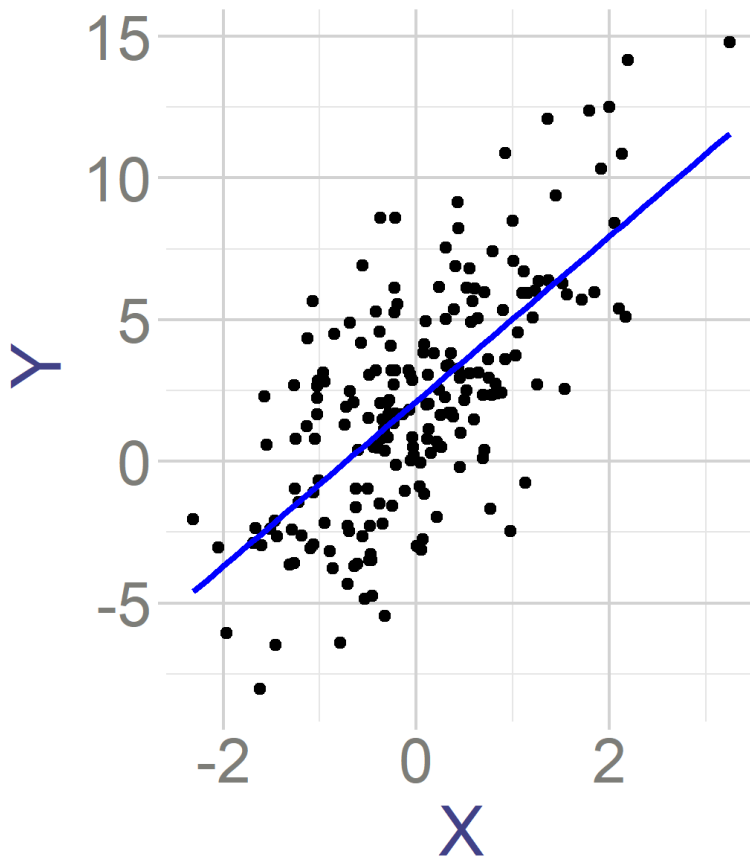
- Example:  $e_i$  vs  $e_{i-3}$
- Note down correlation at each lag:  $\rho(e_i, e_{i-1})$ ,  $\rho(e_i, e_{i-2})$ ,  $\rho(e_i, e_{i-3})$



# Normality

We can start by looking at histograms

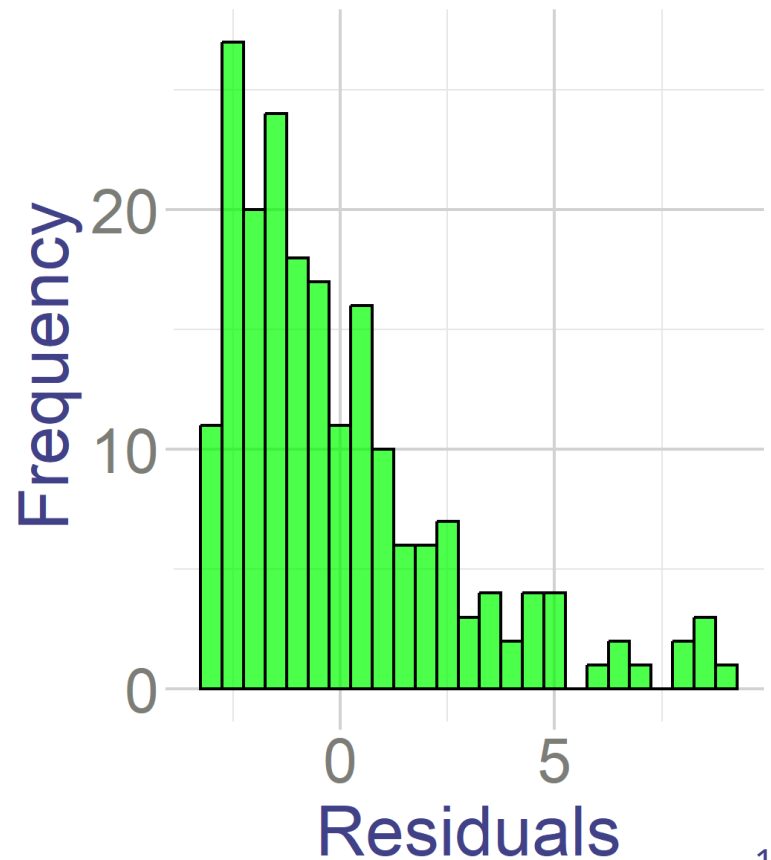
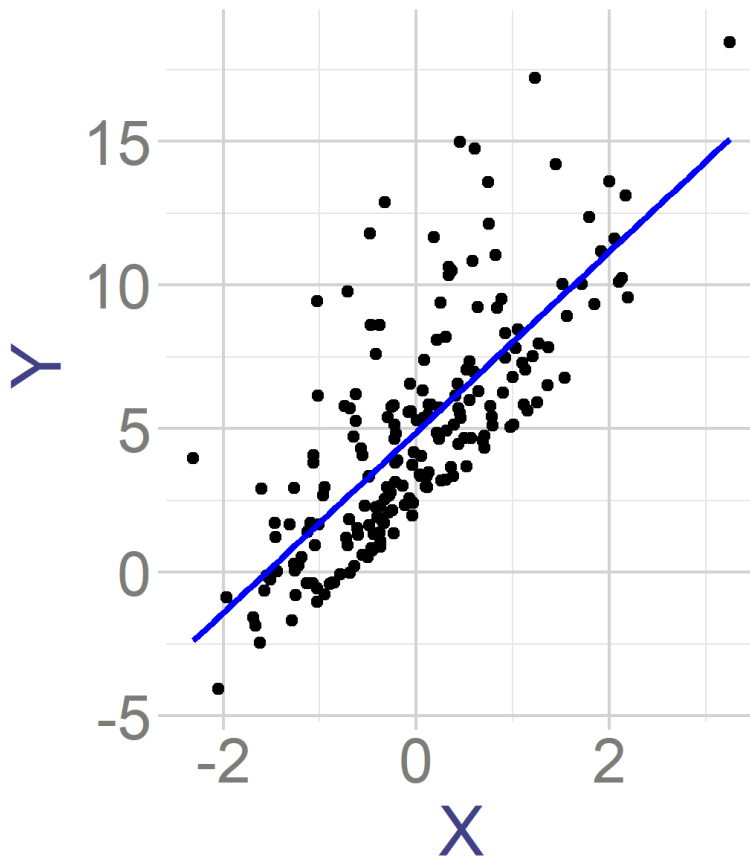
- Some are obvious



# Normality

We can start by looking at histograms

- Some are obvious

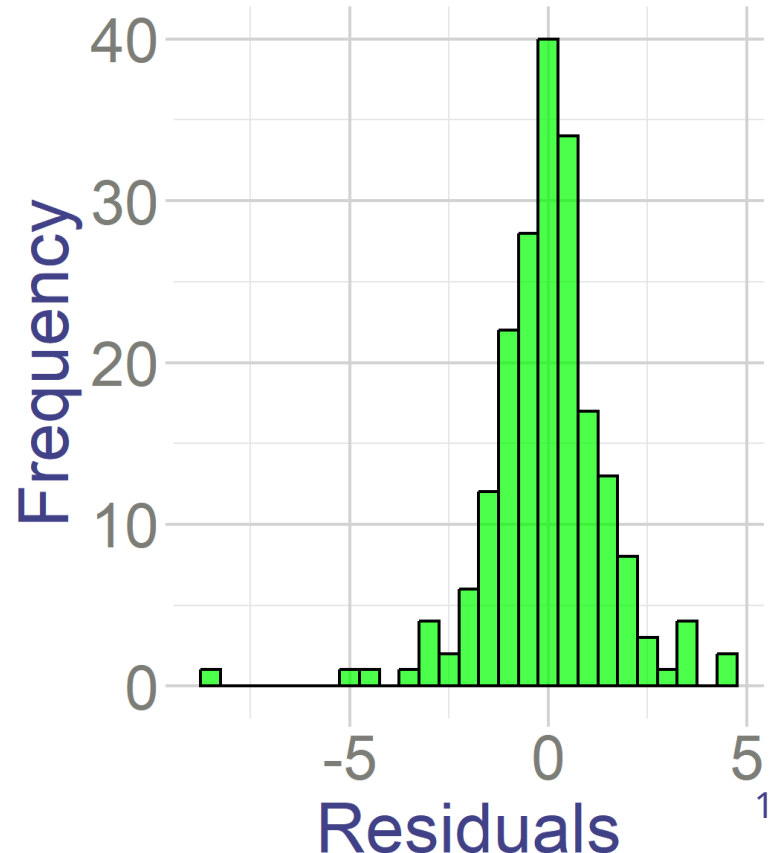
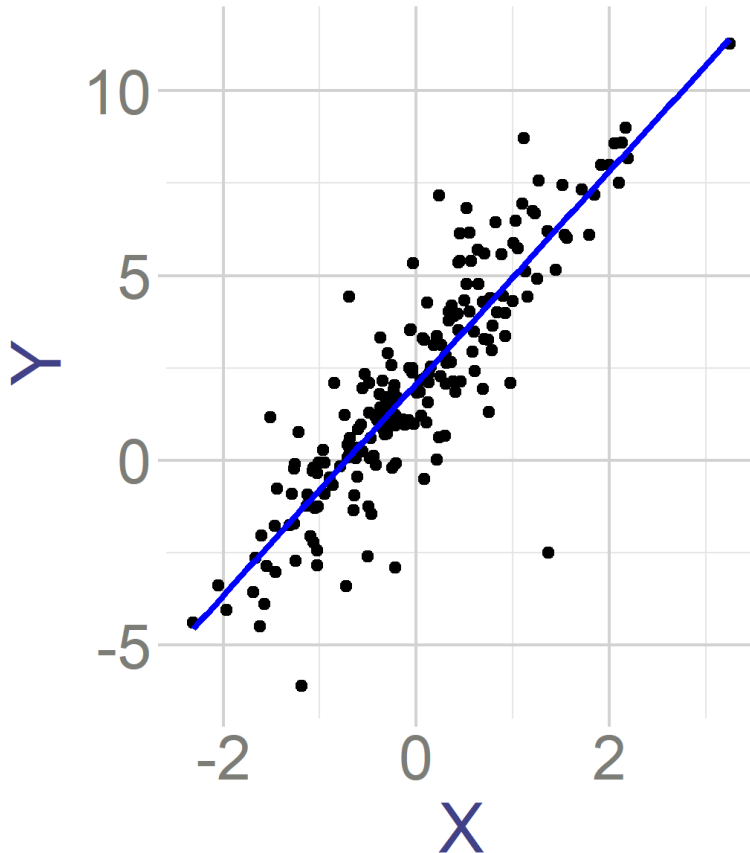




# Normality

We can start by looking at histograms

- Some are **NOT** obvious
- Then we need a different test



# Normality

## Q-Q Plot: Quantile-Quantile Plot

- Comparing quantiles of our data to what they should be if they come from normal distribution
- Procedure:
  - Standardize residuals
  - Sort them
  - Check which quantile they represent
  - This is equivalent to checking cumulative probability:  $CP_i = \frac{Index}{n+1}$
  - Where index is the number of observation if we sort if from smallest to largest
    - 1 is the smallest, 2 is the second smallest, ..., n is the largest
  - In other words, what share of data is smaller than this observation
  - Compare them to quantiles from standard normal
  - If our distribution is normal, the quantiles should be similar

# Q-Q plot

Show  entries

Residuals	Standardized_Residuals	Index	Cumulative_Probability	Quantile_Normal
-1.681	-0.713	23	0.228	-0.745
-0.691	-0.351	39	0.386	-0.29
4.676	1.609	95	0.941	1.563
0.212	-0.022	51	0.505	0.013
0.388	0.043	54	0.535	0.088
5.145	1.78	96	0.95	1.645
1.383	0.406	69	0.683	0.476
-3.795	-1.485	5.5	0.054	-1.607
-2.061	-0.852	19	0.188	-0.885
-1.337	-0.587	29	0.287	-0.562

Showing 1 to 10 of 100 entries

Previous

1

2

3

4

5

...

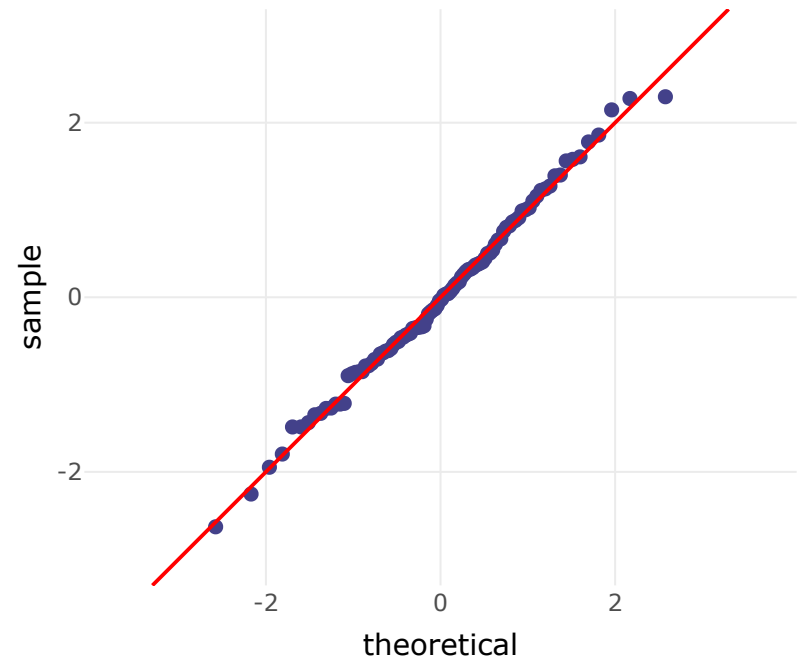
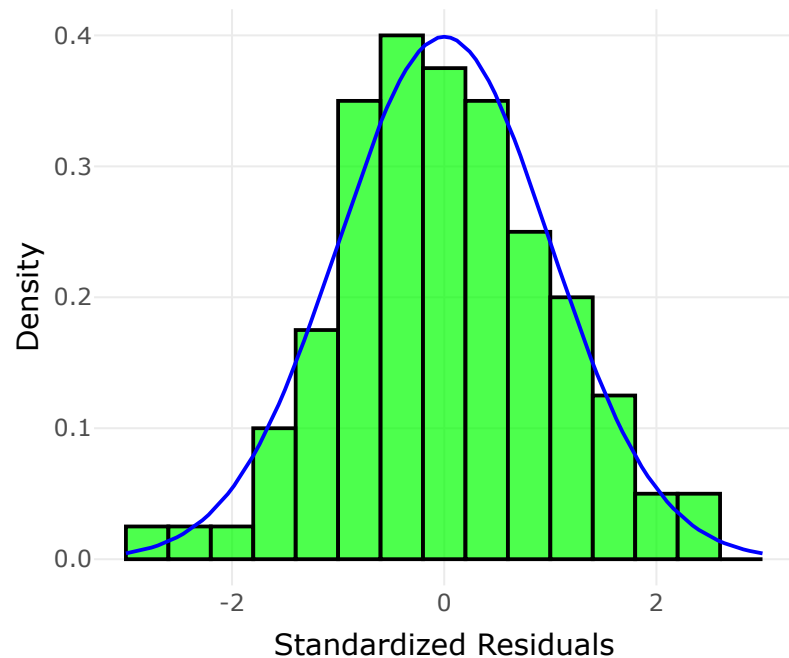
10

Next

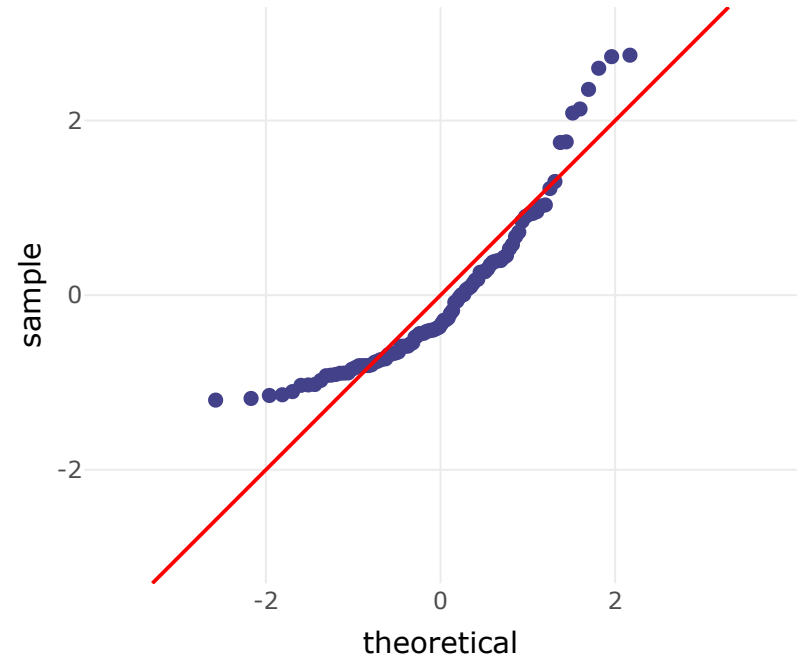
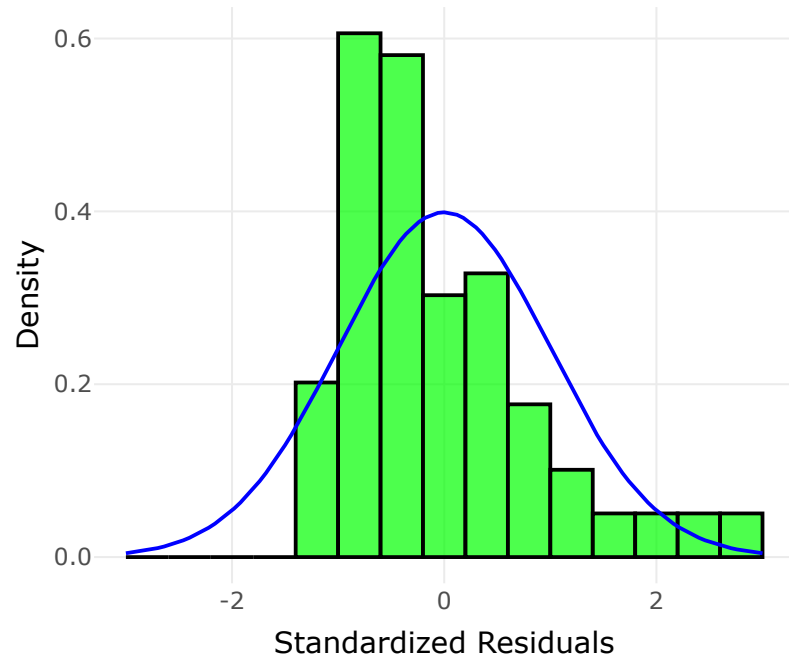
# Q-Q plot

- Next, we plot the sample quantiles vs the standard normal quantiles
- If they are similar, they should be on the straight line

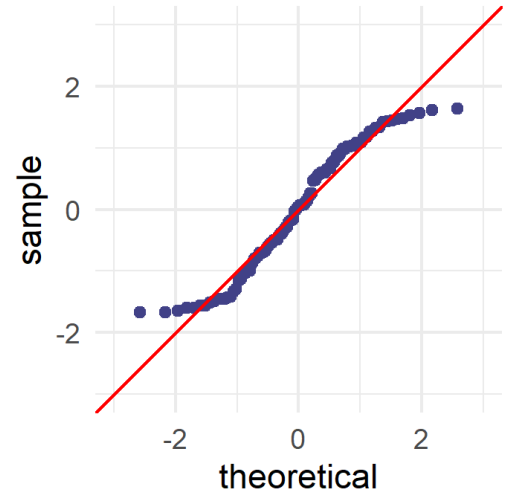
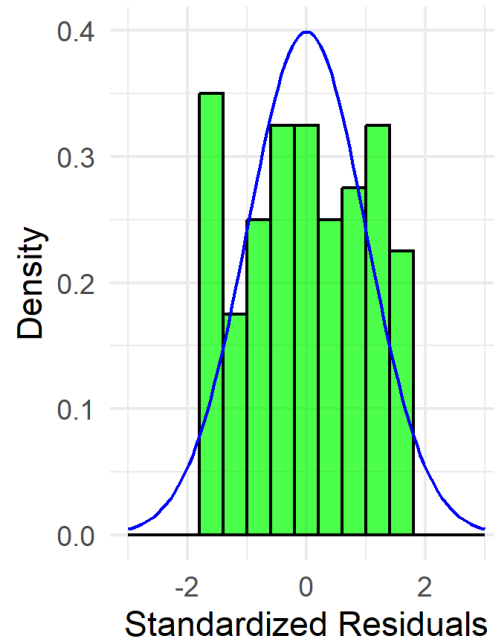
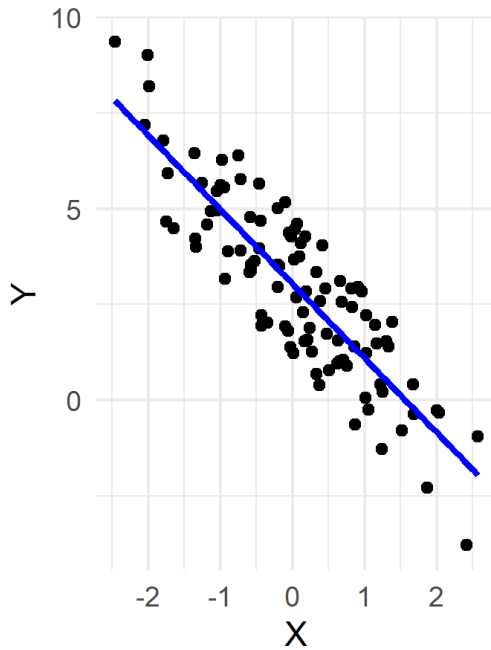
## Q-Q plot



# Q-Q plot



# Q-Q plot



# Jarque-Bera Test

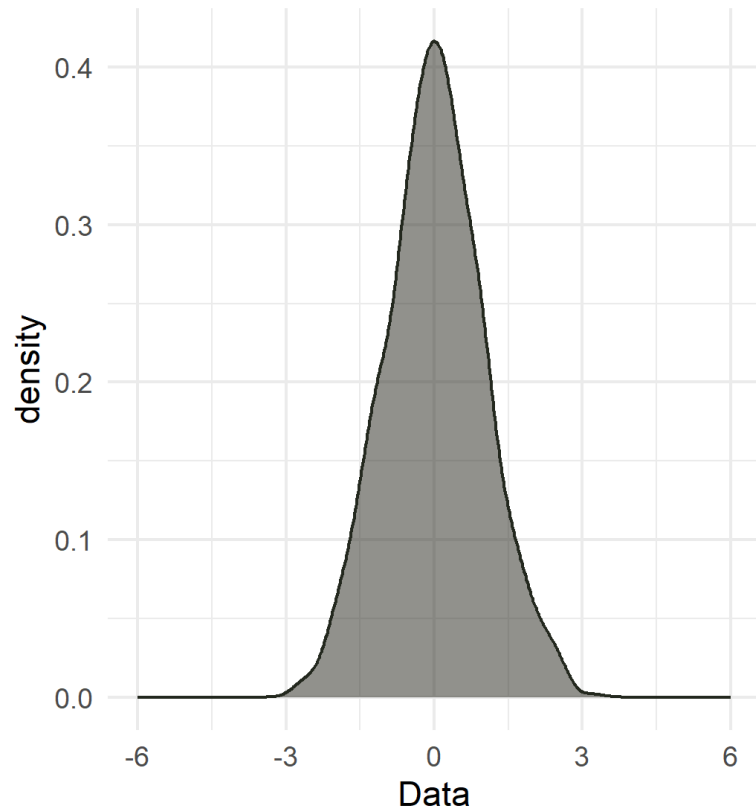
- We can also use the **Jarque-Bera Test** to see whether sample comes from a normal distribution

## Intuition

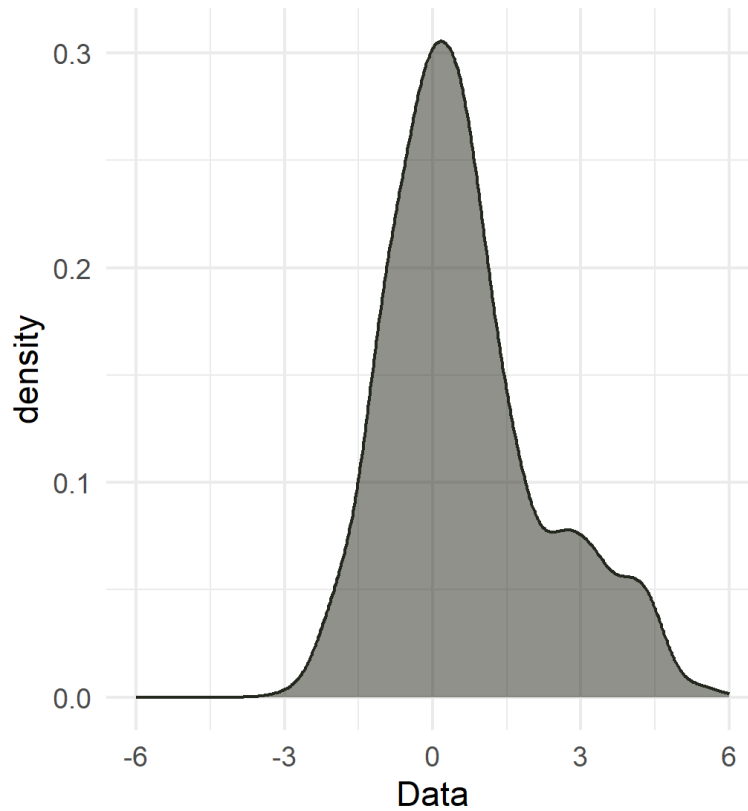
- Skeweness of normal distribution is 0 (it's symmetric)

$$S = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left( \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{\frac{3}{2}}}$$

Density Plot (Skewness = 0)



Density Plot (Positive Skewness)





# Jarque-Bera Test

- We can also use the **Jarque-Bera Test** to see whether sample comes from a normal distribution

## Intuition

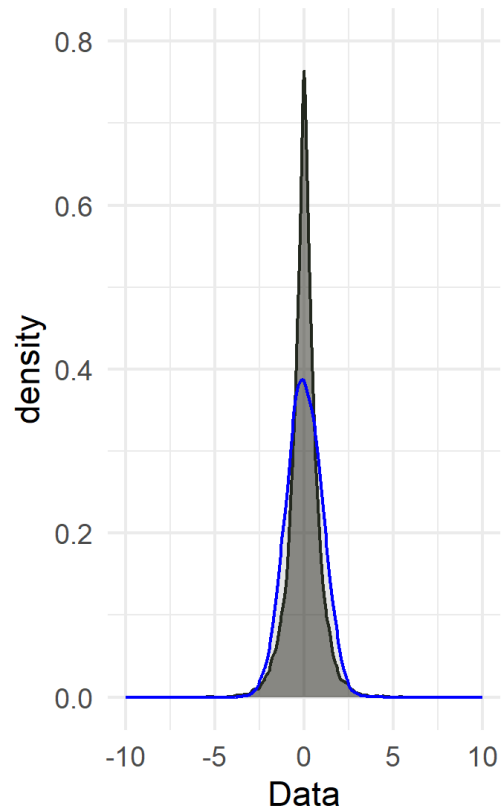
- Skeweness of normal distribution is 0 (it's symmetric)

$$S = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left( \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{\frac{3}{2}}}$$

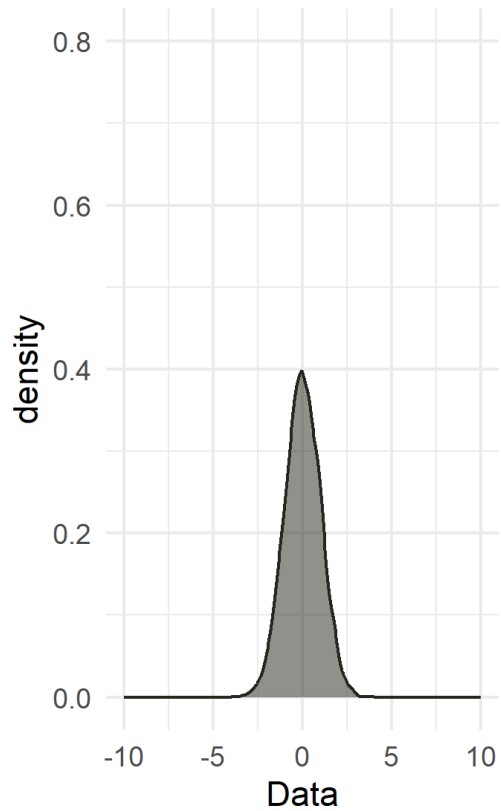
- Kurtosis of normal distribution is 3
- Excess kurtosis is 0

$$EK = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left( \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^2} - 3$$

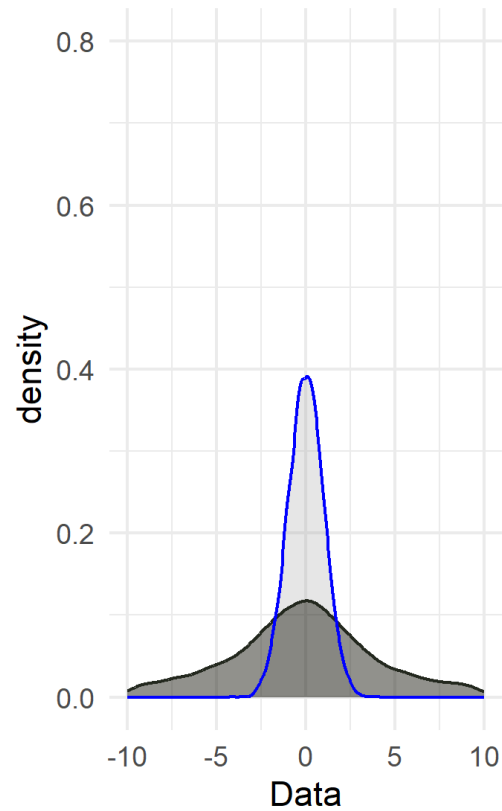
Kurtosis < 3



Kurtosis = 3



Kurtosis > 3



# Jarque-Bera Test

- We can also use the **Jarque-Bera Test** to see whether sample comes from a normal distribution

## Intuition

- Skeweness of normal distribution is 0 (it's symmetric)

$$S = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right)^{\frac{3}{2}}}$$

- Kurtosis of normal distribution is 3
- Excess kurtosis is 0

$$EK = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right)^2} - 3$$

- We will test, whether our sample has (more or less)
  - Skeweness of 0
  - Kurtosis of 3

# Jarque-Bera Test

The test statistic for our test is:

$$JB = \frac{n}{6} \left( \frac{S^2}{2} + \frac{EK^2}{4} \right)$$

- Its value will be high if:
  - Skeweness deviates significantly from 0
  - Kurtosis deviates significantly from 3
- If the data really comes from normal (that's our null hypothesis), then:

$$JB \sim \chi_2$$

- It follows the Chi-squared distribution with 2 degrees of freedom.

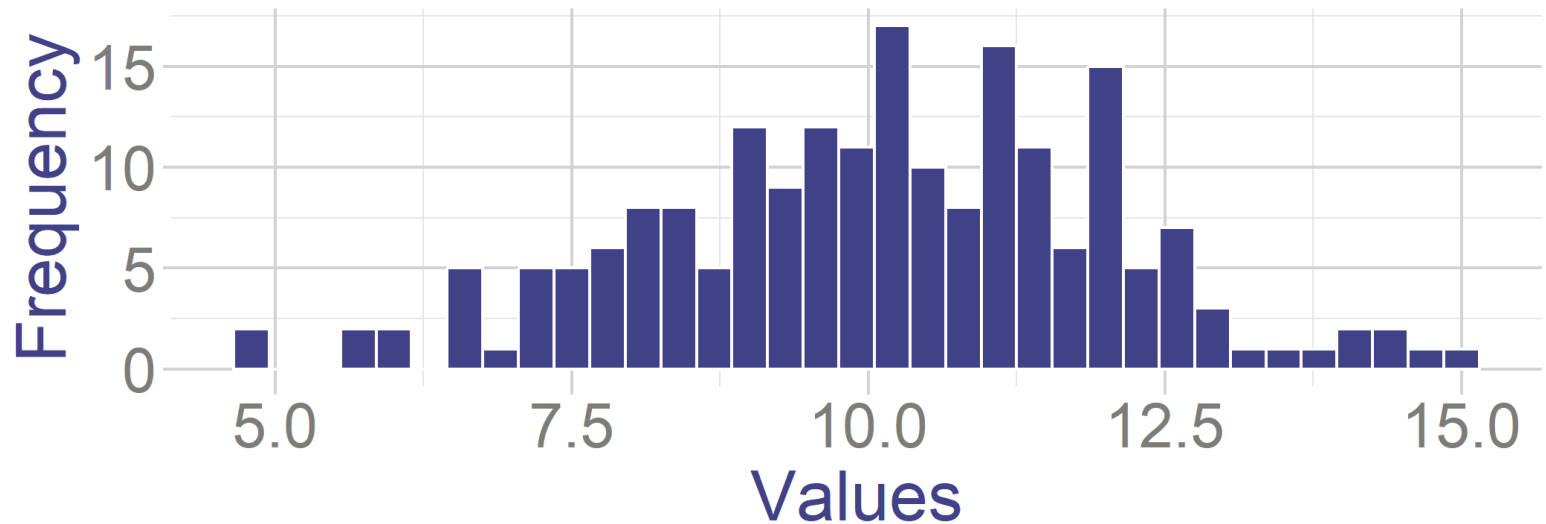
# Jarque-Bera Test

So in our usual testing setting:

- $H_0 : e_i$  comes from normal (JB is small)
- $H_A : e_i$  Does not come from normal (JB is large)
- It's a one sided test, so we reject at  $\alpha$  if

$$JB_{test} > \chi_{1-\alpha,2}$$

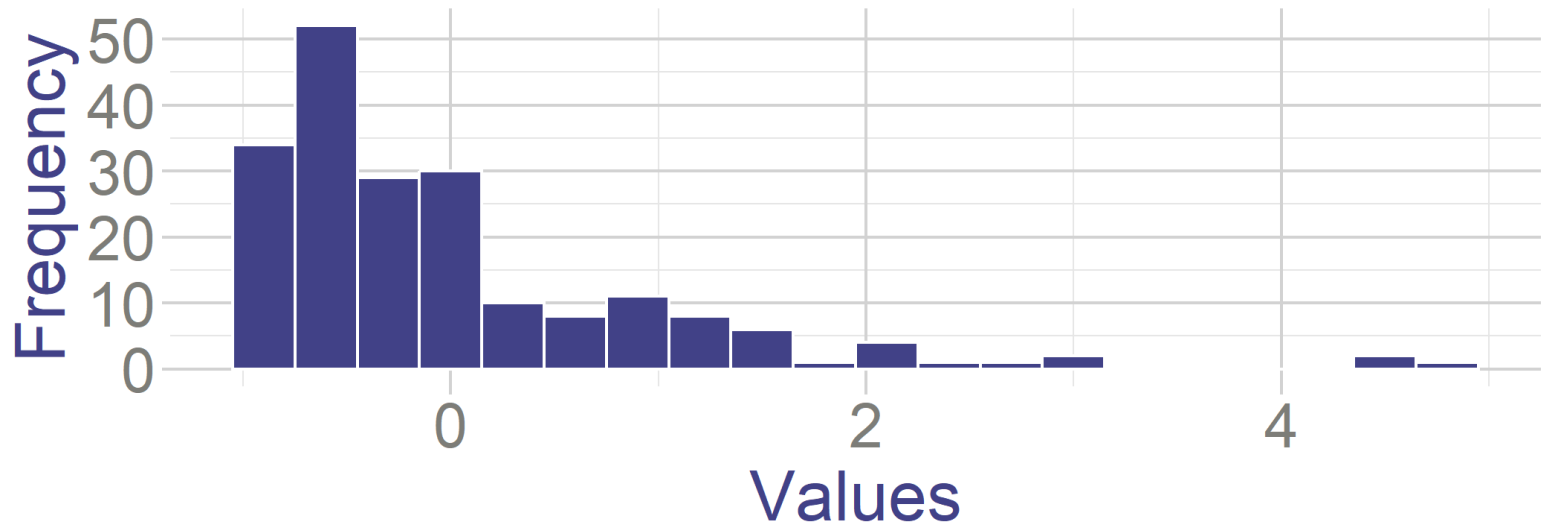
# Jarque-Bera Test



```
library(tseries)
jarque.bera.test(sample_data)
```

```
##
##      Jarque Bera Test
##
## data:  sample_data
## X-squared = 0.96619, df = 2, p-value = 0.6169
```

# Jarque-Bera Test



```
library(tseries)
jarque.bera.test(sample_data)
```

```
##
##      Jarque Bera Test
##
## data:  sample_data
## X-squared = 420.19, df = 2, p-value < 2.2e-16
```





