

# STATS 2107

## Statistical Modelling and Inference II

### Solutions

### Workshop 7:

### Assumptions in simple linear regression.

Matt Ryan

Semester 2 2022

## Contents

<b>Simple linear regression</b>	<b>2</b>
The model . . . . .	2
What are the assumptions? . . . . .	2
How do we check the assumptions? . . . . .	2
5-point check . . . . .	2
<b>Your turn</b>	<b>3</b>
What to do . . . . .	3
<b>Linearity</b>	<b>3</b>
A linear model . . . . .	3
The residual vs fitted plot . . . . .	4
The true model . . . . .	4
A non-linear model . . . . .	4
The residual vs fitted plot . . . . .	5
The true model . . . . .	5
<b>Your turn</b>	<b>5</b>
What to do . . . . .	5
<b>Homoscedasticity</b>	<b>6</b>
A homoscedastic model . . . . .	6
The residual vs fitted plot . . . . .	7
The true model . . . . .	7
A heteroscedastic model . . . . .	7
The residual vs fitted plot . . . . .	8
The true model . . . . .	8
<b>Your turn</b>	<b>8</b>
What to do . . . . .	8
<b>Normality</b>	<b>9</b>
A normal model . . . . .	9
The residual vs fitted plot . . . . .	10

The residual QQ plot . . . . .	10
The true model . . . . .	10
A non-normal model . . . . .	11
The residual vs fitted plot . . . . .	11
The residual QQ plot . . . . .	12
The true model . . . . .	12
<b>Your turn</b>	<b>12</b>
What to do . . . . .	12
<b>One more assumption</b>	<b>13</b>
The plot . . . . .	13
The residual vs fitted plot . . . . .	14
The residual QQ plot . . . . .	14
The true model . . . . .	14
<b>Your turn</b>	<b>14</b>
What to do . . . . .	14

## Simple linear regression

### The model

Suppose you have data  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  where  $x_i, y_i \in \mathbb{R}$  for each  $i$ .

#### THE MODEL:

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

where  $\varepsilon_i \sim N(0, \sigma^2)$  independently for each  $i = 1, 2, \dots, n$ .

### What are the assumptions?

- **Linearity:**  $E[\varepsilon_i] = 0$
- **Homoscedasticity:**  $\text{Var}(\varepsilon_i) = \sigma^2$
- **Normality:**  $\varepsilon_i \sim N$
- **Independence:** Design assumption

### How do we check the assumptions?

We look at the residuals  $\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ :

- Use the `plot` command in R
  - **Linearity:** The Residuals vs Fitted plot
  - **Homoscedasticity:** The Residuals vs Fitted plot
  - **Normality:** The Normal QQ plot
  - **Independence:** Check the design/data collection

### 5-point check

When checking assumptions, answer:

- **What?**
- **Where?**
- **What do you expect?**

- What do you see?
- What do you conclude?

## Your turn

### What to do

1. Load the `cheddar.csv` dataset. Fit the simple linear regression of `taste` on either `acetic`, `h2s`, or `lactic`.

---

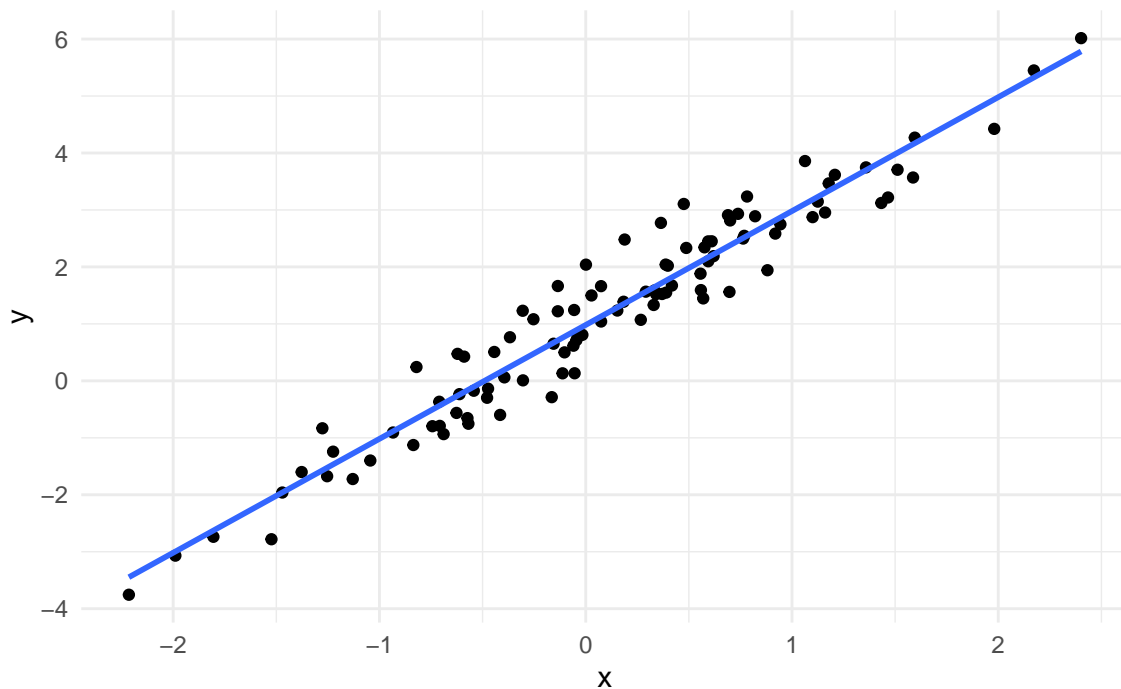
### Solutions:

For example, we could fit the model `taste ~ lactic`:

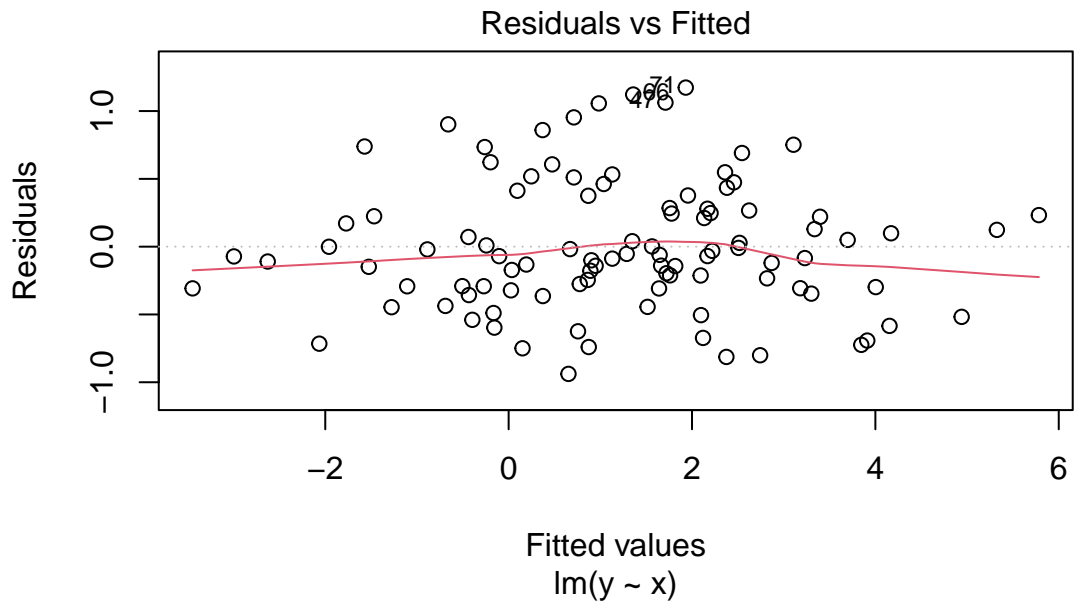
```
cheddar <- read_csv("../data/cheddar.csv", col_types = cols())  
cheddar_lm <- lm(taste ~ lactic, data = cheddar)
```

## Linearity

### A linear model



### The residual vs fitted plot

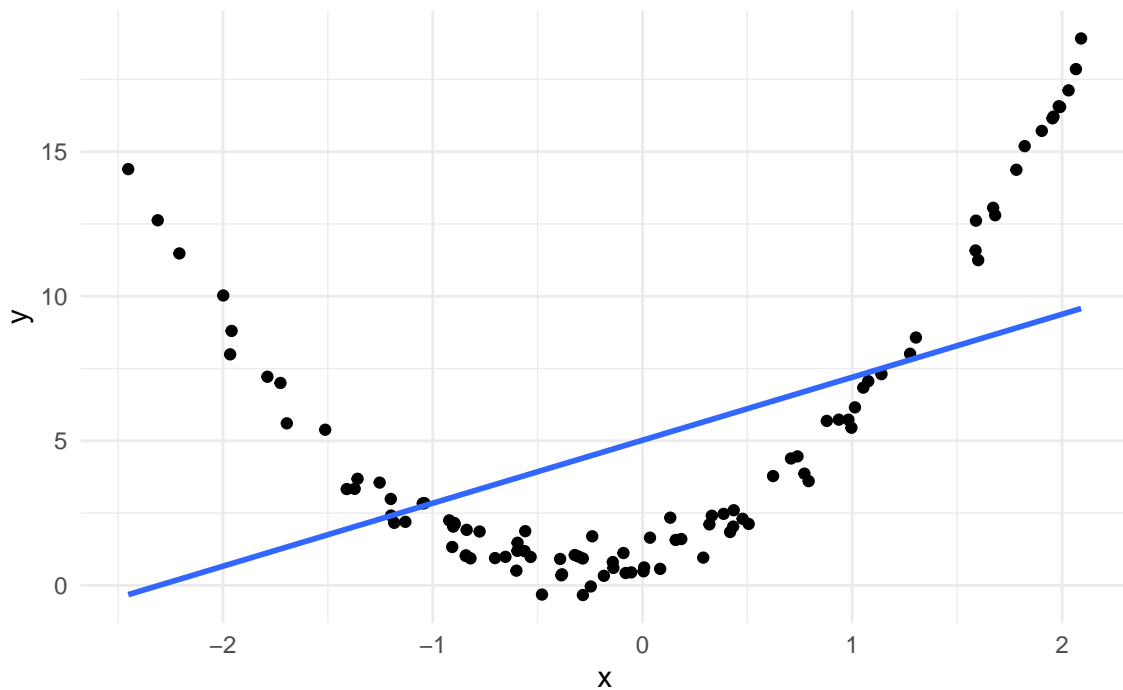


### The true model

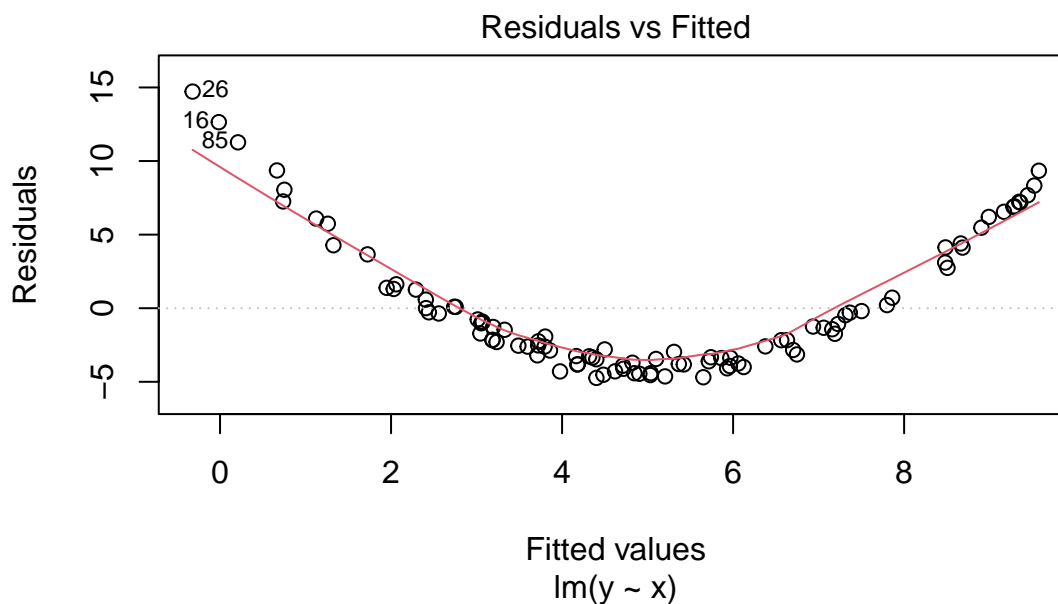
$$y_i = 1 + 2x_i + \varepsilon_i$$

where  $\varepsilon_i \sim N(0, 0.5^2)$ .

### A non-linear model



## The residual vs fitted plot



## The true model

$$y_i = 1 + 2x_i + 3x_i^2 + \varepsilon_i$$

where  $\varepsilon_i \sim N(0, 0.5^2)$ .

## Your turn

### What to do

1. For your model fitted previously, test the linearity assumption.

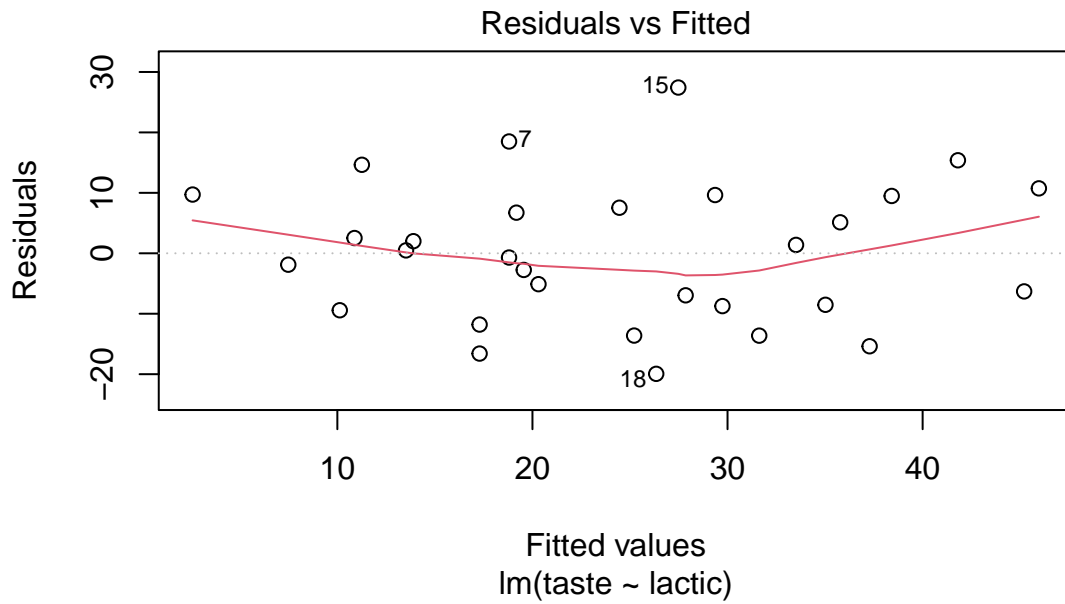
---

### Solutions:

See the plot below

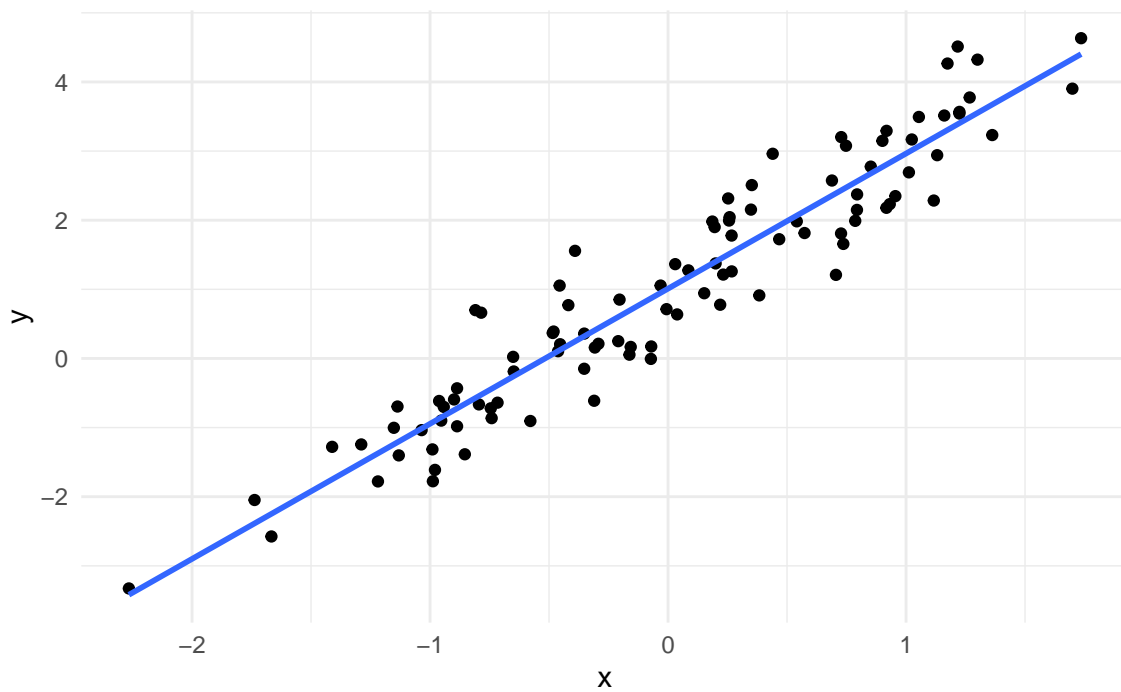
- **What?** Linearity.
  - **Where?** Residual vs fitted plot.
  - **What do you expect?** Random scatter above and below 0.
  - **What do you see?** Random scatter above and below 0.
  - **What do you conclude?** Linearity reasonable.
-

```
plot(cheddar_lm, which = 1)
```

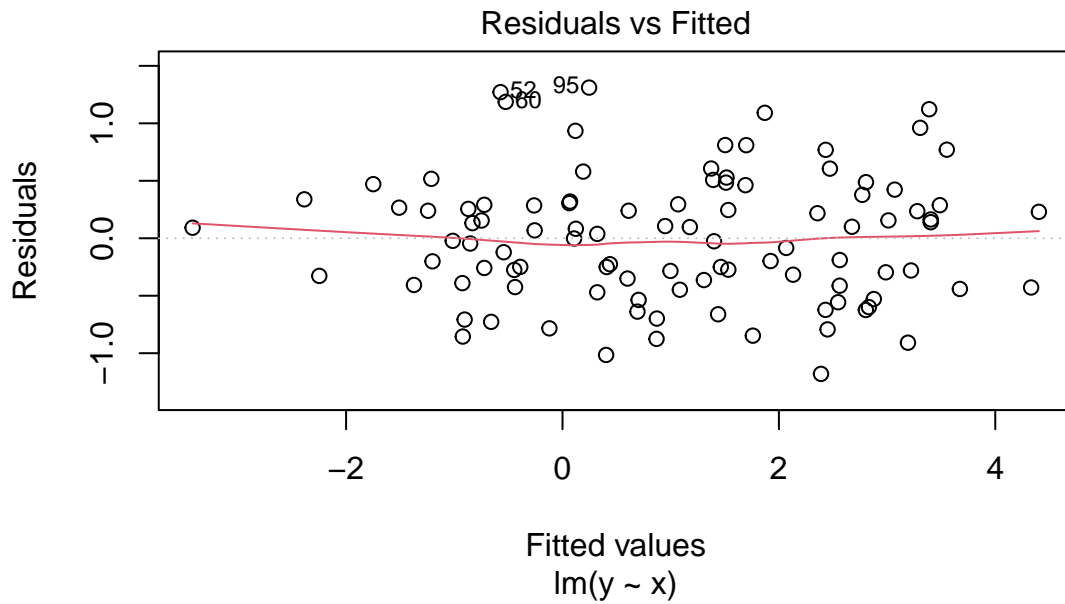


## Homoscedasticity

A homoscedastic model



## The residual vs fitted plot

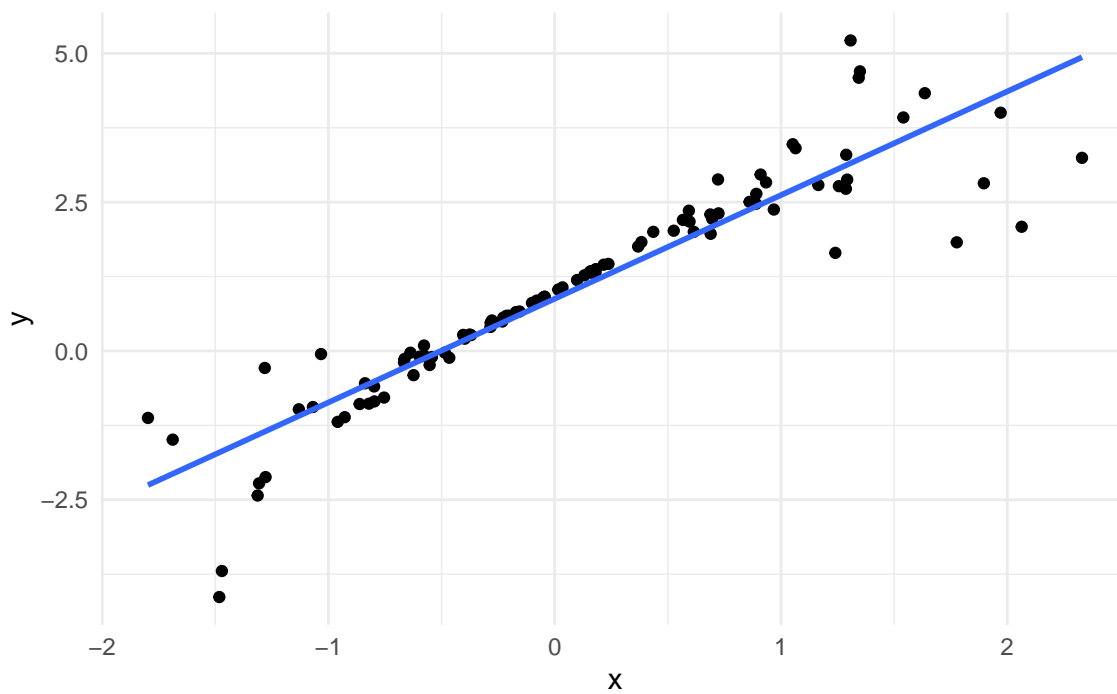


## The true model

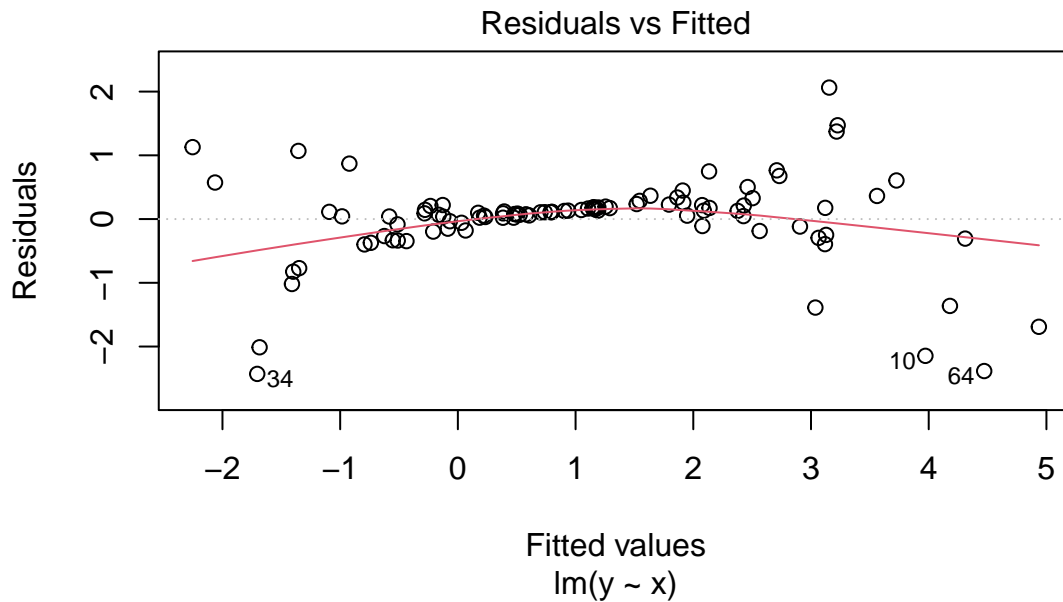
$$y_i = 1 + 2x_i + \varepsilon_i$$

where  $\varepsilon_i \sim N(0, 0.5^2)$ .

## A heteroscedastic model



## The residual vs fitted plot



## The true model

$$y_i = 1 + 2x_i + \varepsilon_i$$

where  $\varepsilon_i \sim N(0, 0.5^2 x_i^4)$ .

## Your turn

### What to do

1. For your model fitted previously, test the homoscedasticity assumption.

---

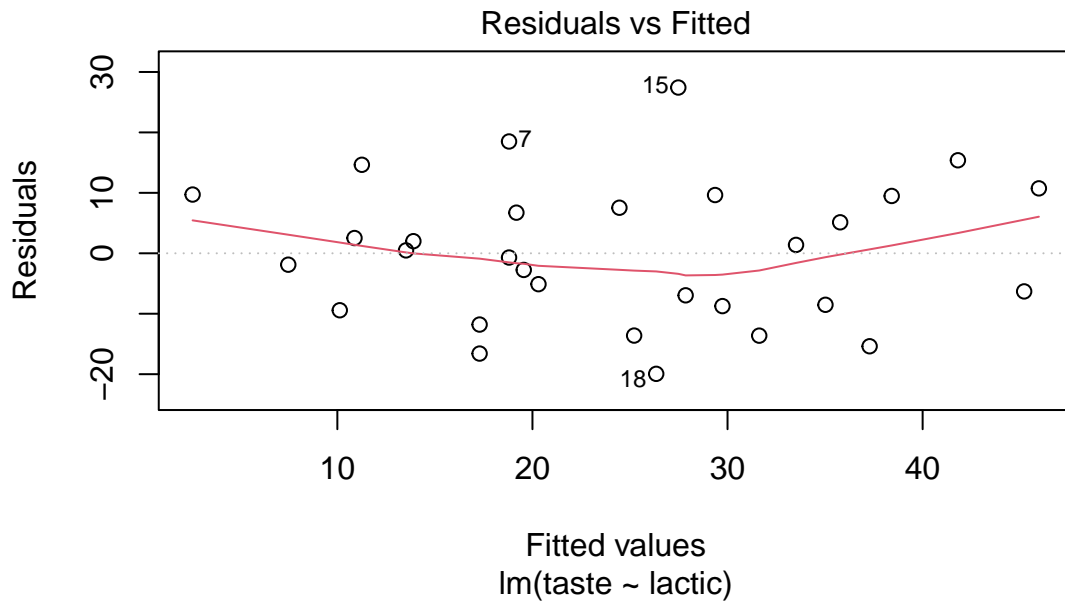
### Solutions:

See the plot below

- **What?** homoscedasticity
  - **Where?** Residual vs fitted plot.
  - **What do you expect?** Random scatter from left to right.
  - **What do you see?** Random scatter from left to right, given the data size
  - **What do you conclude?** homoscedasticity reasonable.
-

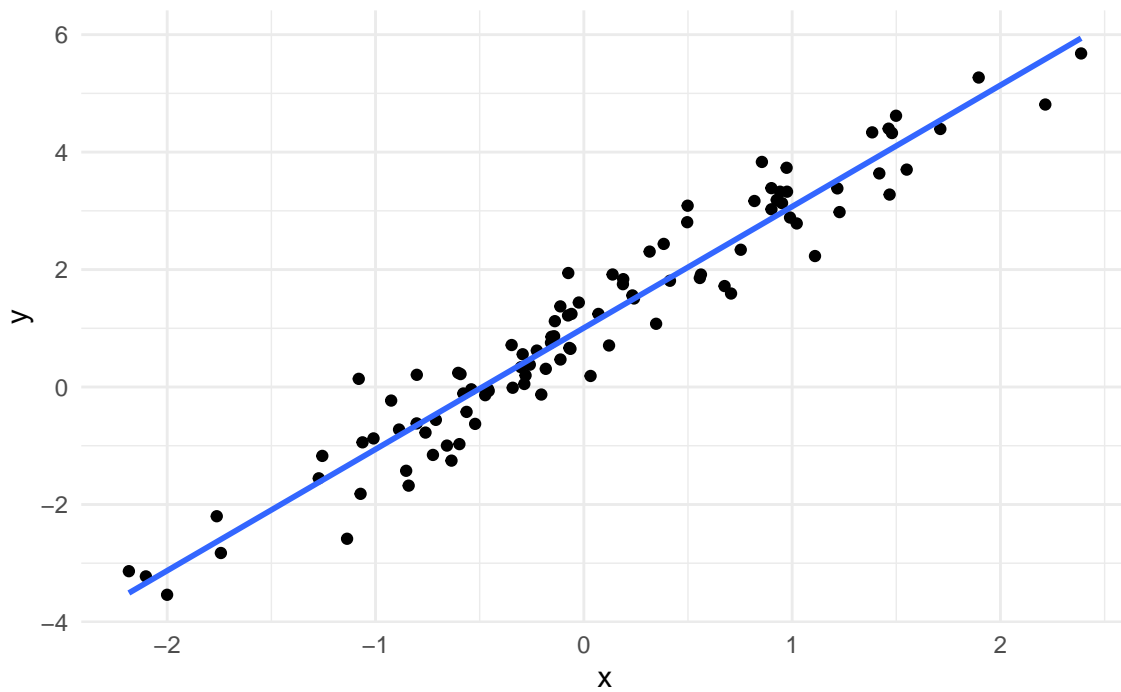


```
plot(cheddar_lm, which = 1)
```

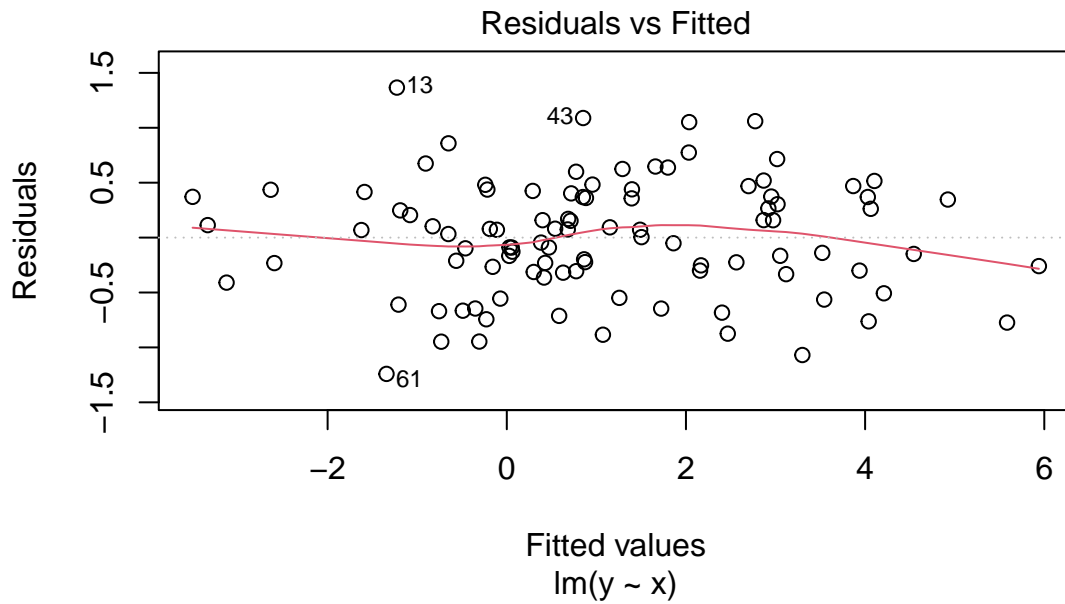


## Normality

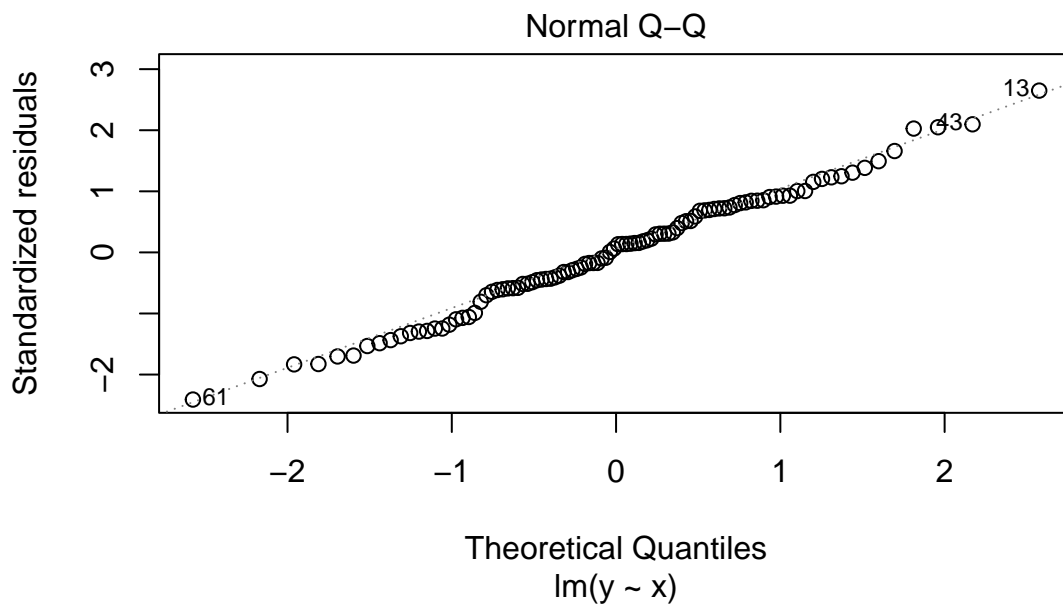
A normal model



The residual vs fitted plot



The residual QQ plot

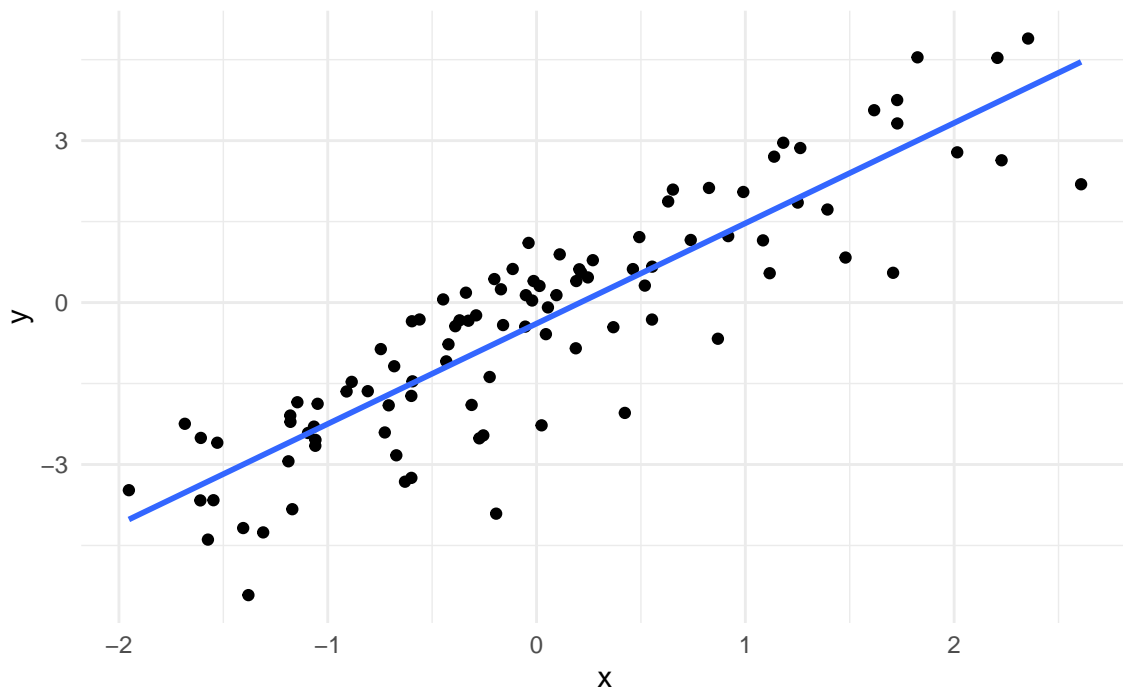


The true model

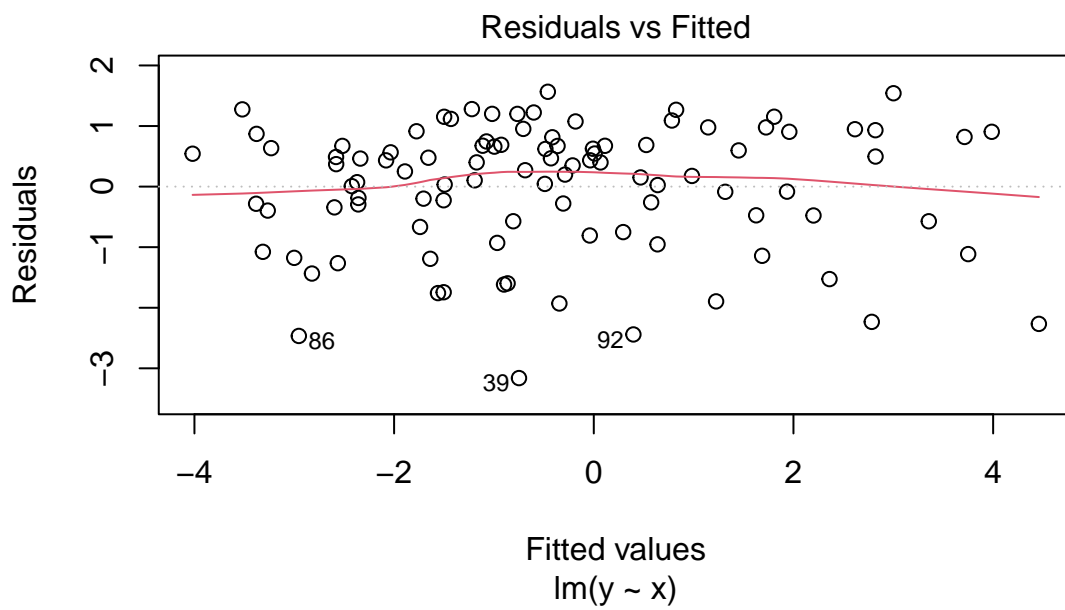
$$y_i = 1 + 2x_i + \varepsilon_i$$

where  $\varepsilon_i \sim N(0, 0.5^2)$ .

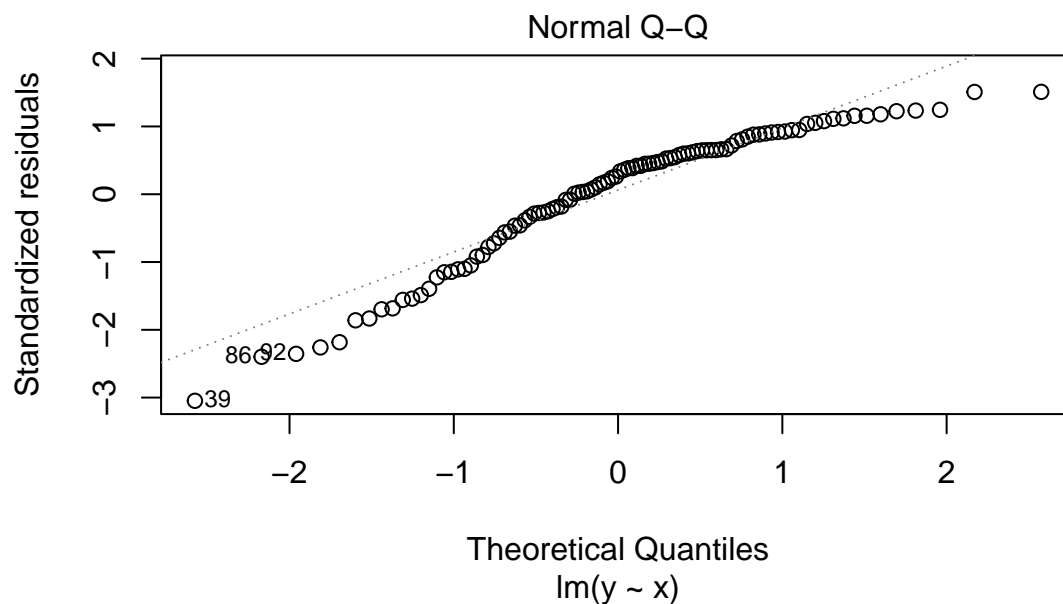
## A non-normal model



## The residual vs fitted plot



## The residual QQ plot



## The true model

$$y_i = 1 + 2x_i + \varepsilon_i$$

where  $\varepsilon_i \sim \log |N(0, 0.5^2)|$ .

## Your turn

### What to do

1. For you model fitted previously, test the normality assumption.

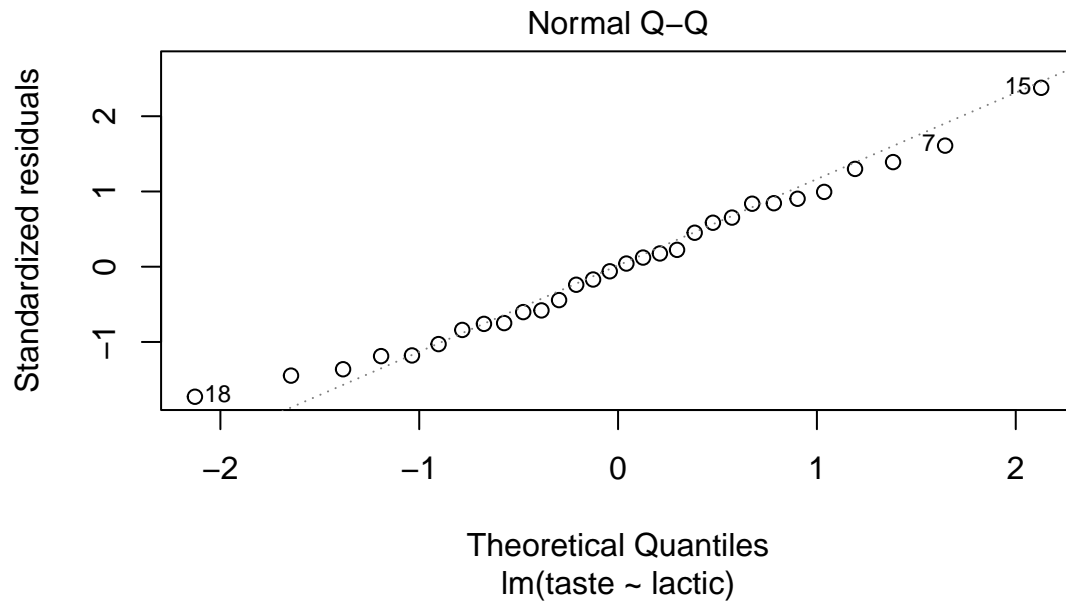
---

### Solutions:

See the plot below

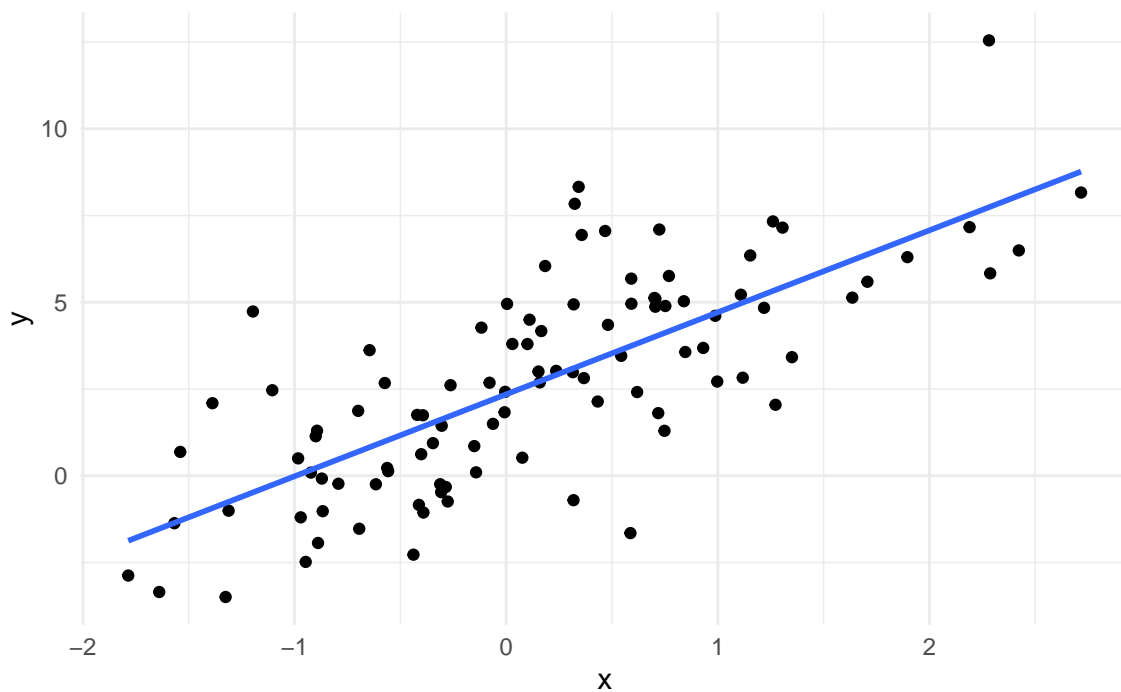
- **What?** normality
  - **Where?** normal QQ plot
  - **What do you expect?** A relative straight line
  - **What do you see?** A relative straight line
  - **What do you conclude?** normality reasonable.
-

```
plot(cheddar_lm, which = 2)
```

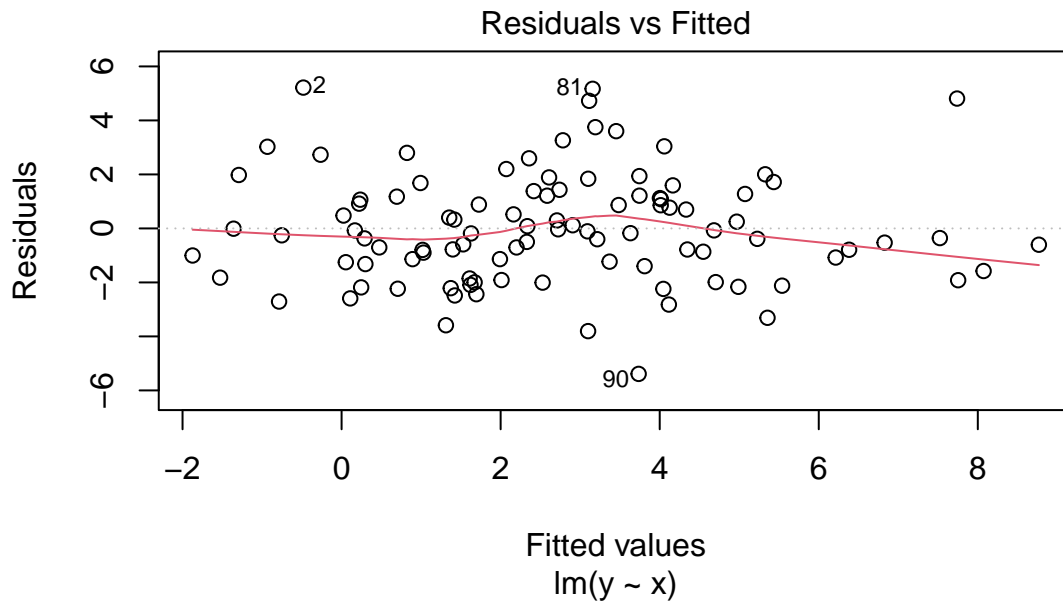


## One more assumption

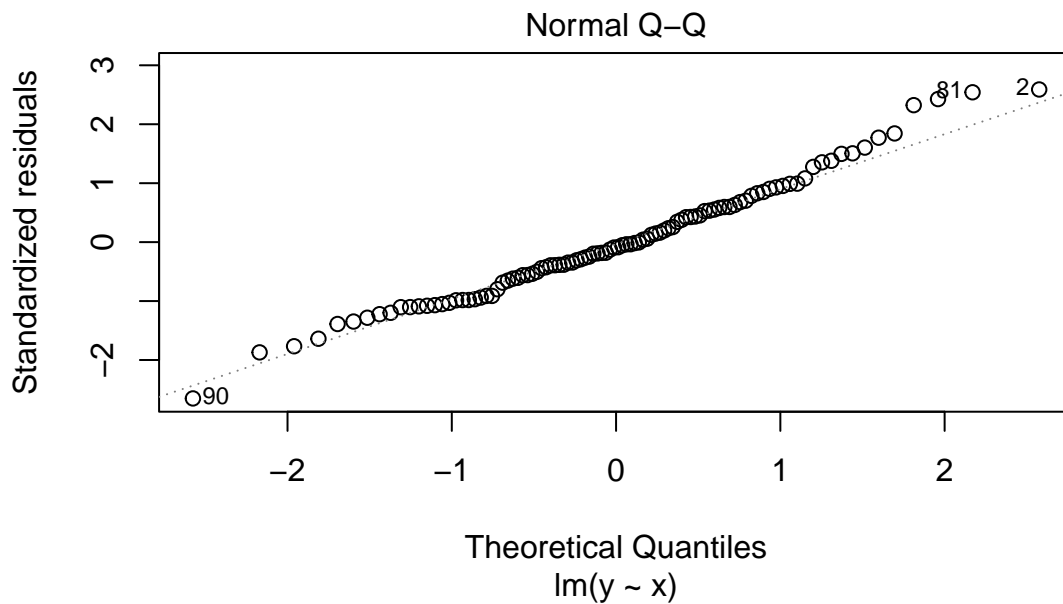
The plot



## The residual vs fitted plot



## The residual QQ plot



## The true model

$$y_i = 1 + 2x_i + y_{i-1} + \varepsilon_i$$

where  $\varepsilon_i \sim N(0, 0.5^2)$ .

## Your turn

### What to do

1. For your model fitted previously, test the independence assumption.

---

**Solutions:**

Must check data collection

- **What?** independence
  - **Where?** experiment design
  - **What do you expect?** random collections
  - **What do you see?** Overall taste scores were combined from several testers
  - **What do you conclude?** This is iffy. I would say not independent, we are averaging out over multiple tested, but we must assume that the same testers tasted each cheese.
-