STATS 2107
Statistical Modelling and Inference II

Workshop 7:
Assumptions in simple linear regression.

Matt Ryan

School of Mathematical Sciences, University of Adelaide

Semester 2 2022

Simple linear regression

The model

Suppose you have data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ 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 $i = 1, 2, \ldots, n$.

What are the assumptions?

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

How do we check the assumptions?

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

- 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?**

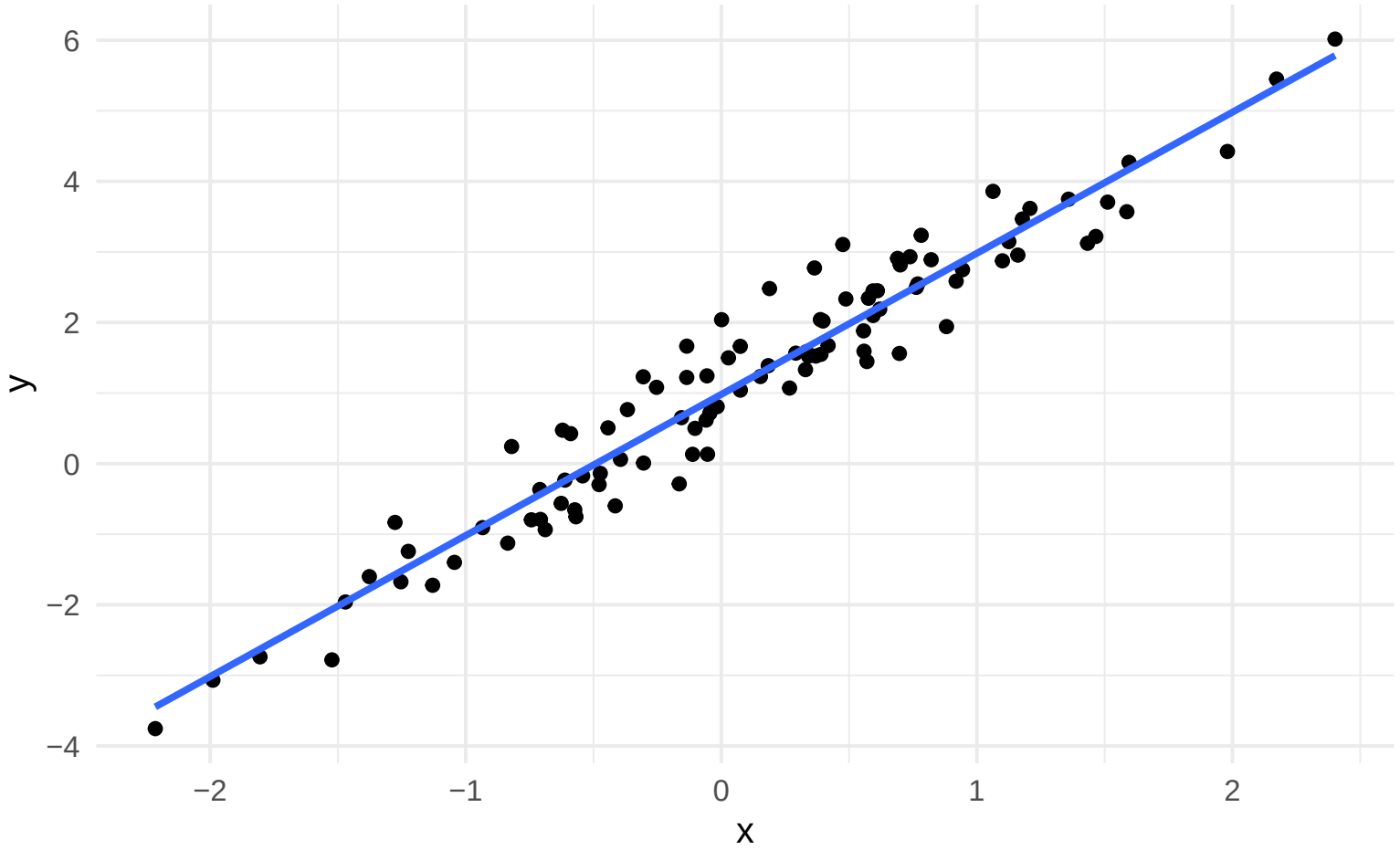
Your turn

What to do

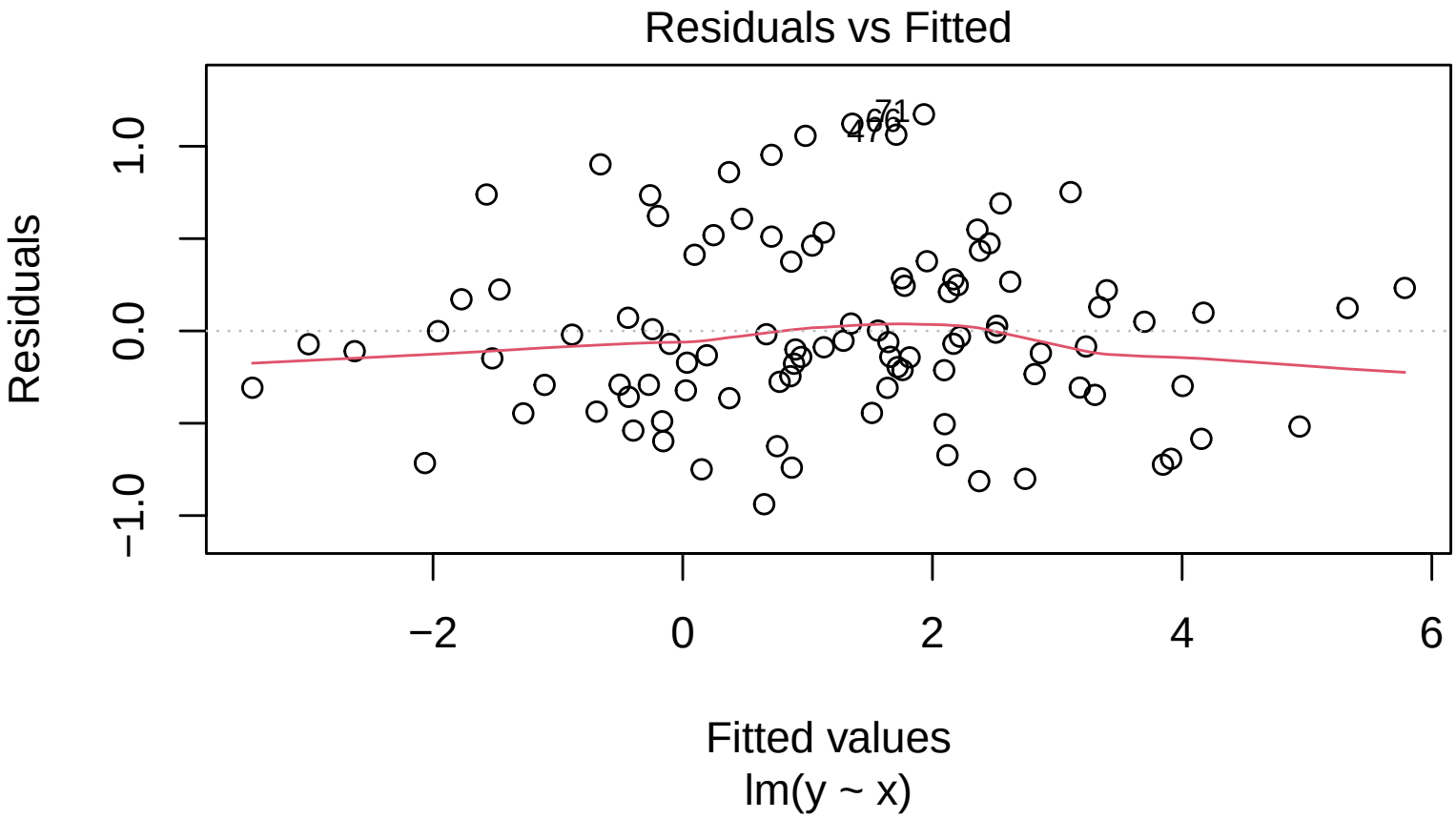
1. Load the `ceddar.csv` dataset. Fit the simple linear regression of `taste` on either `acetic`, `h2s`, or `lactic`.

Linearity

A linear model



The residual vs fitted plot

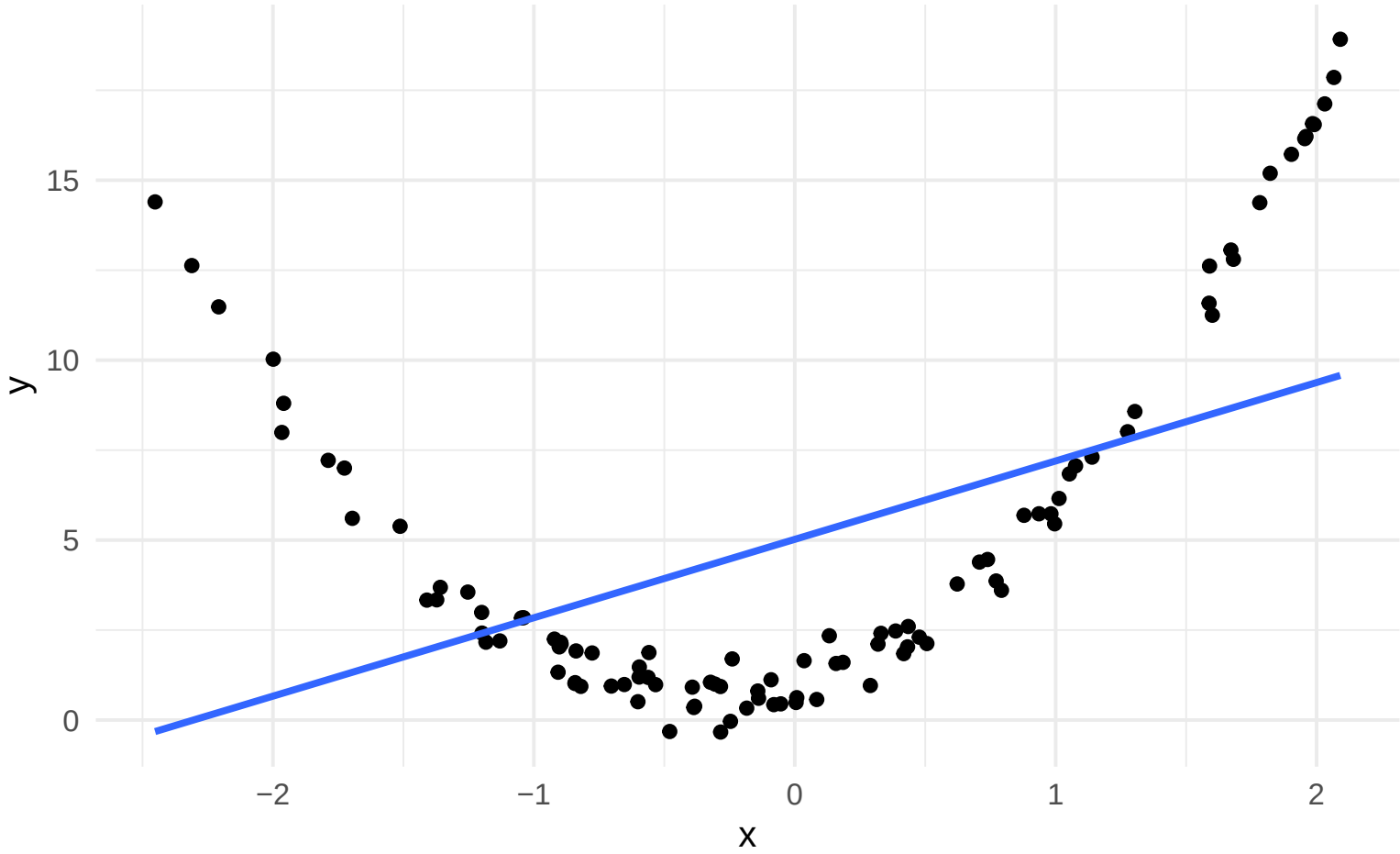


The true model

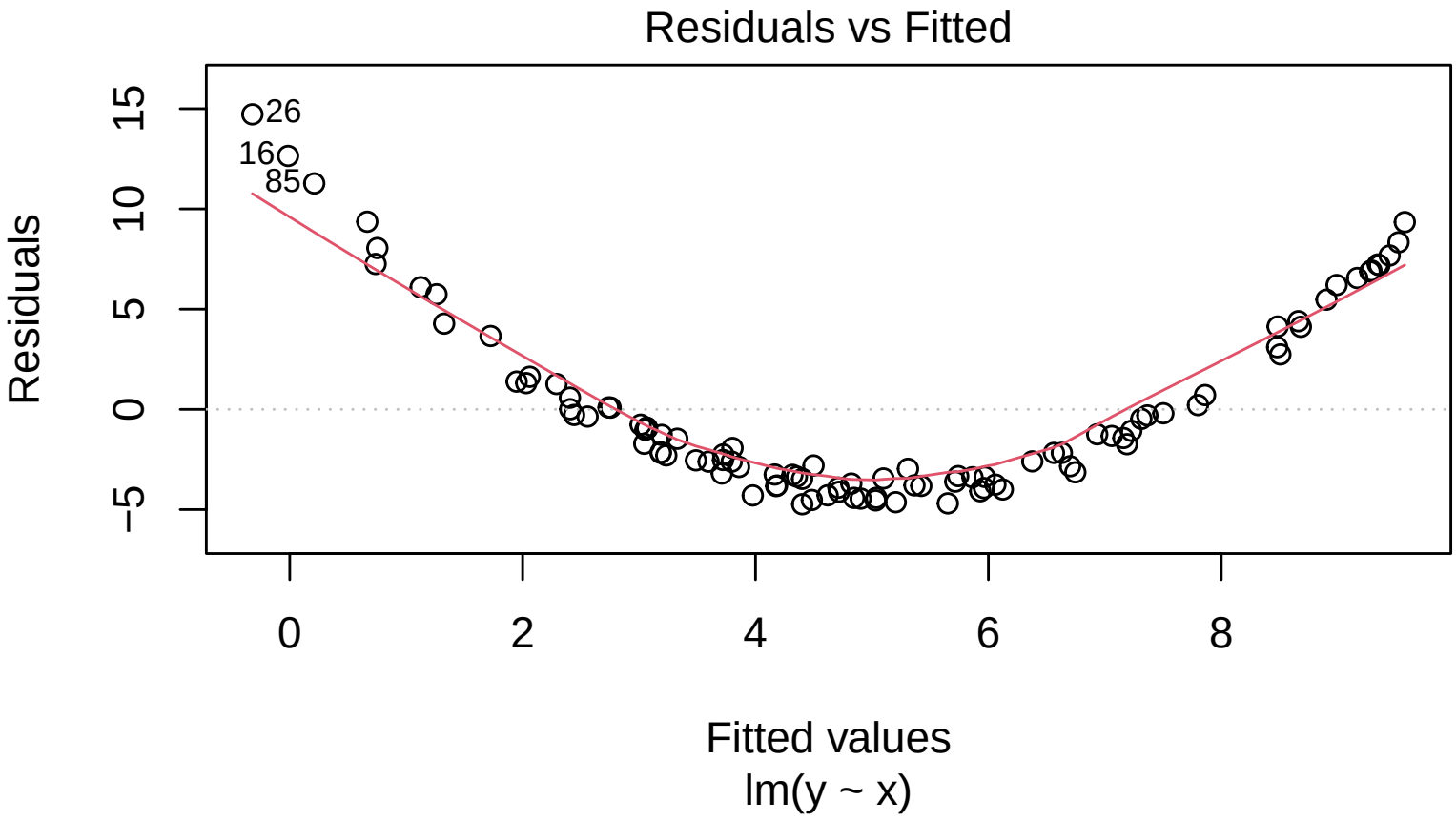
$$y_i = 1 + 2x_i + \varepsilon_i$$

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

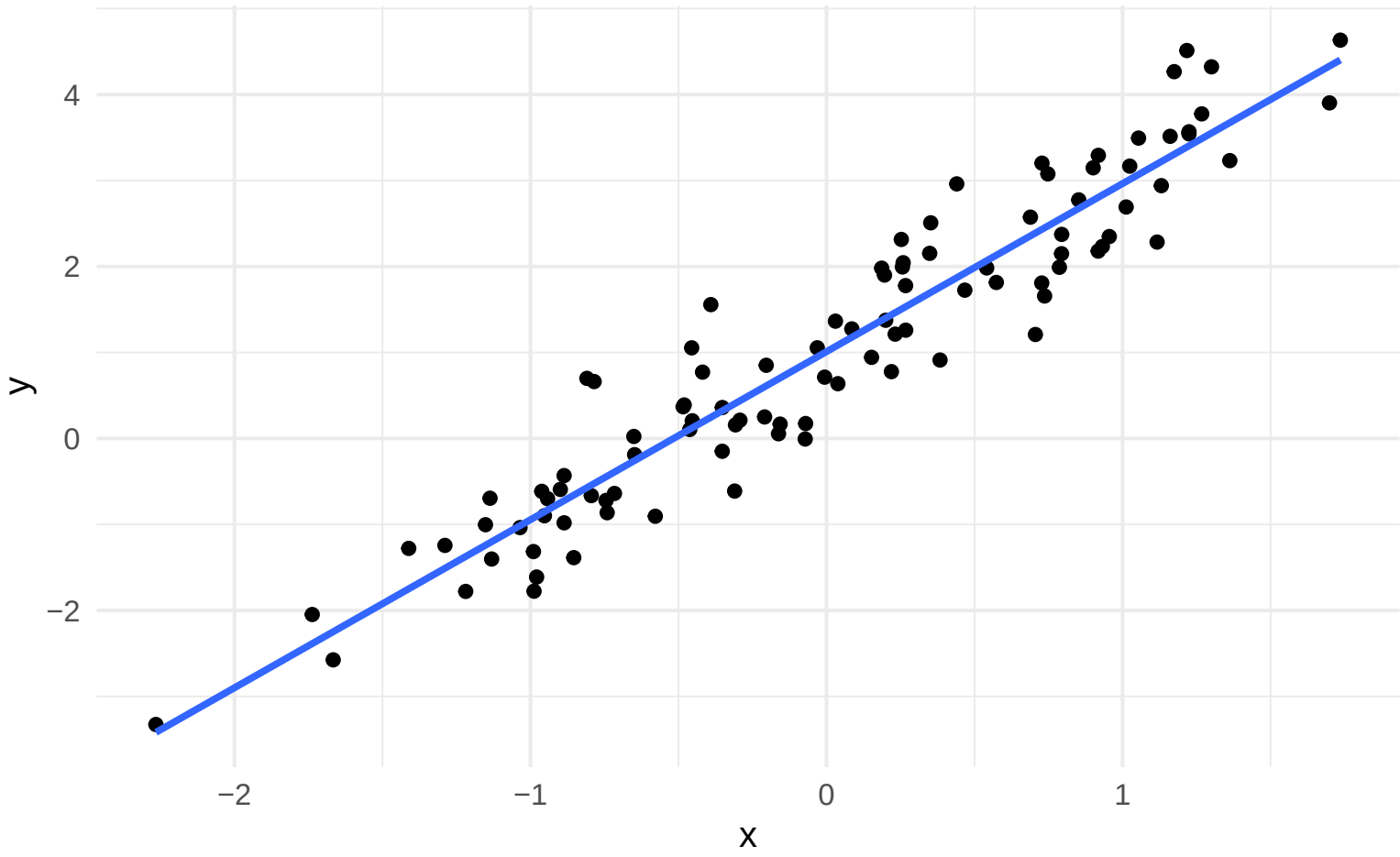
Your turn

What to do

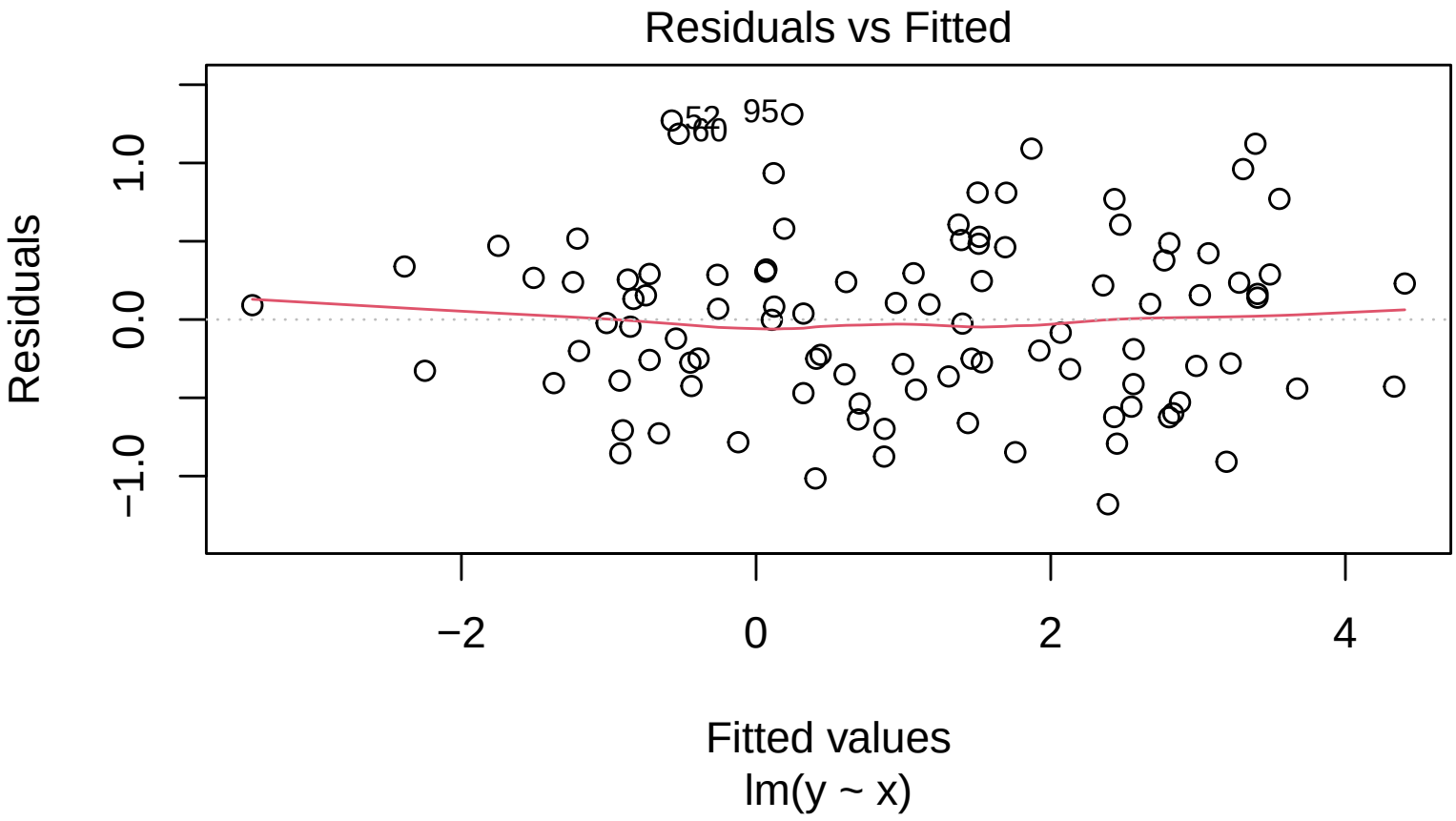
1. For you model fitted previously, test the linearity assumption.

Homoscedasticity

A homoscedastic model



The residual vs fitted plot

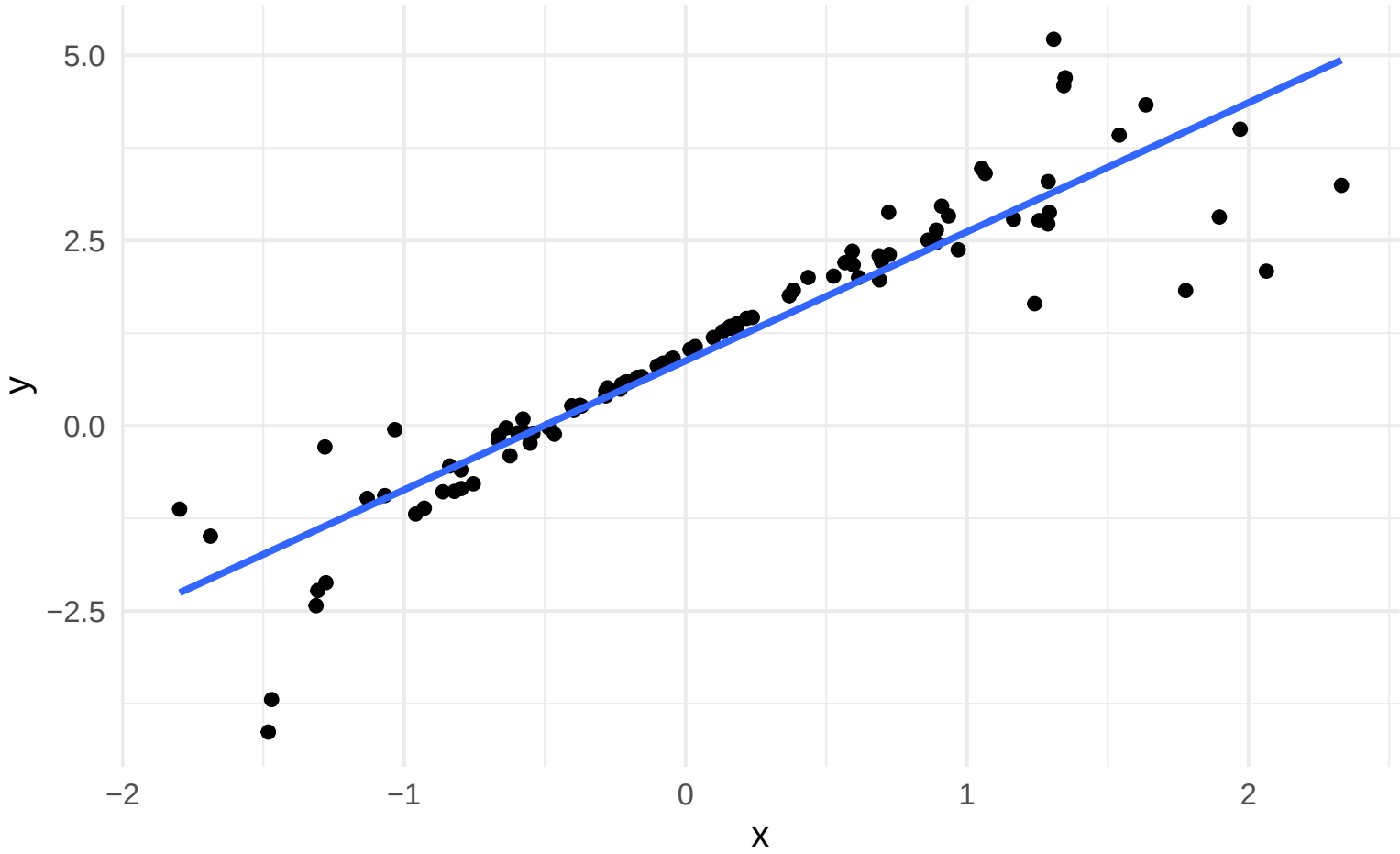


The true model

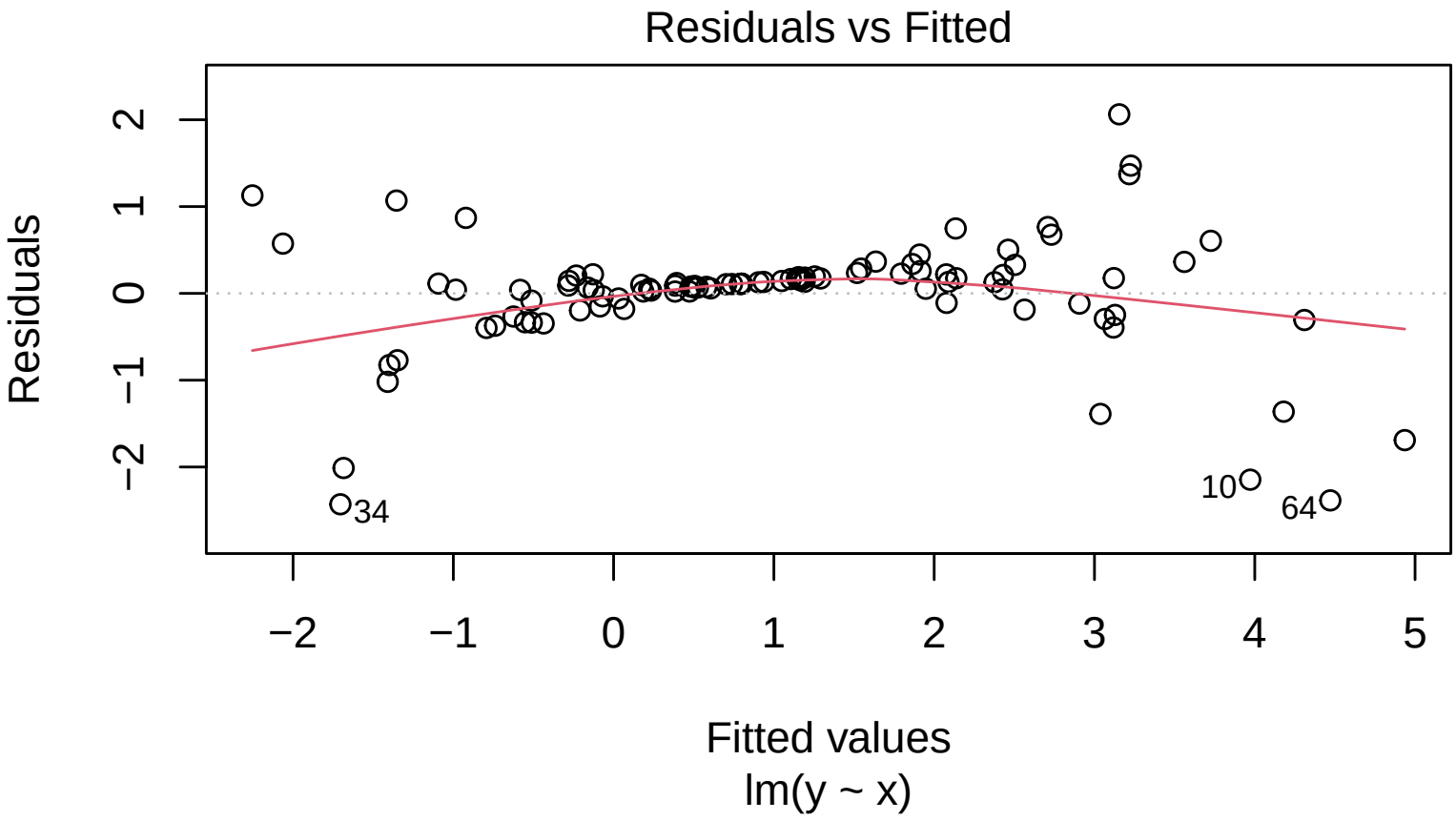
$$y_i = 1 + 2x_i + \varepsilon_i$$

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

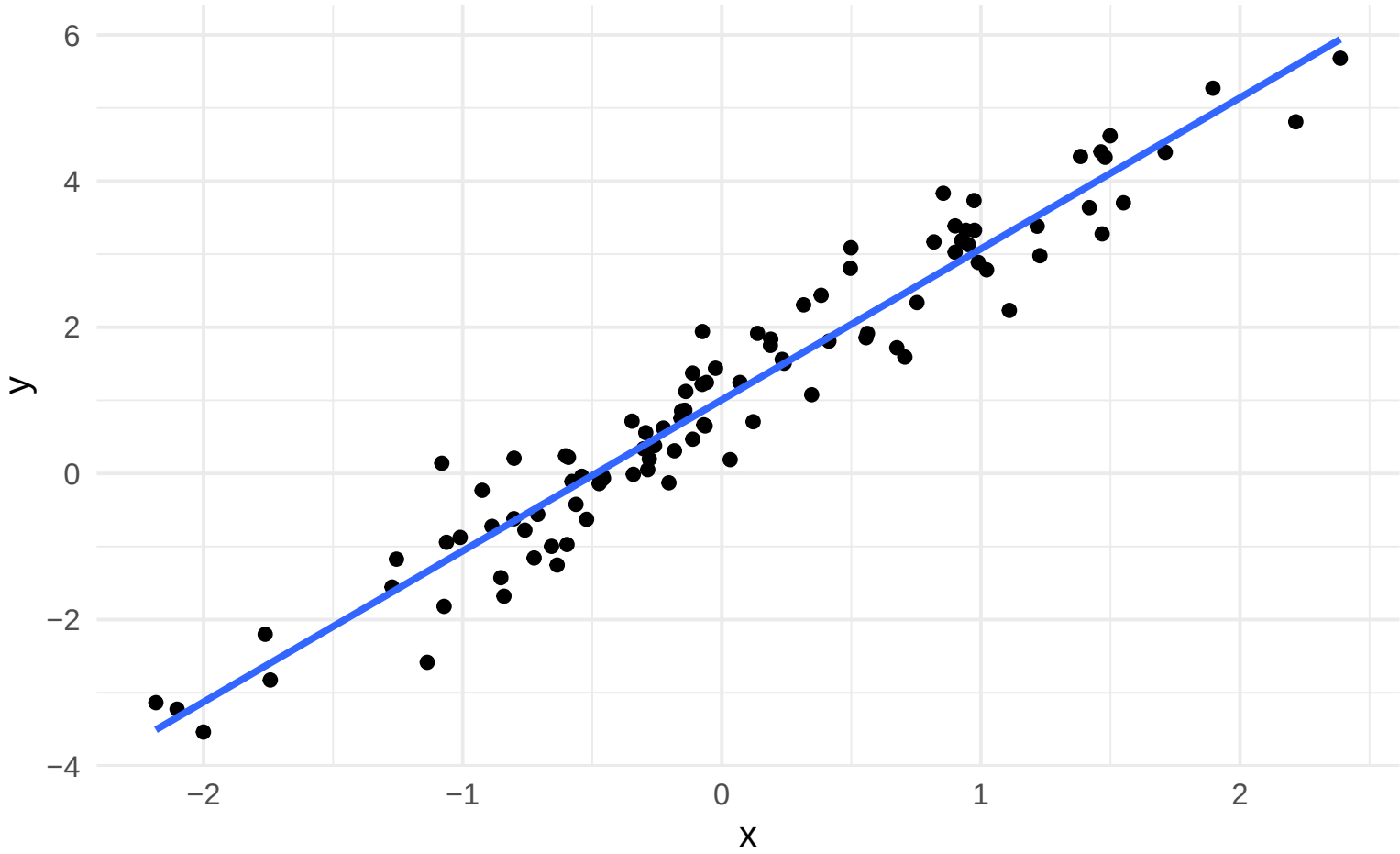
Your turn

What to do

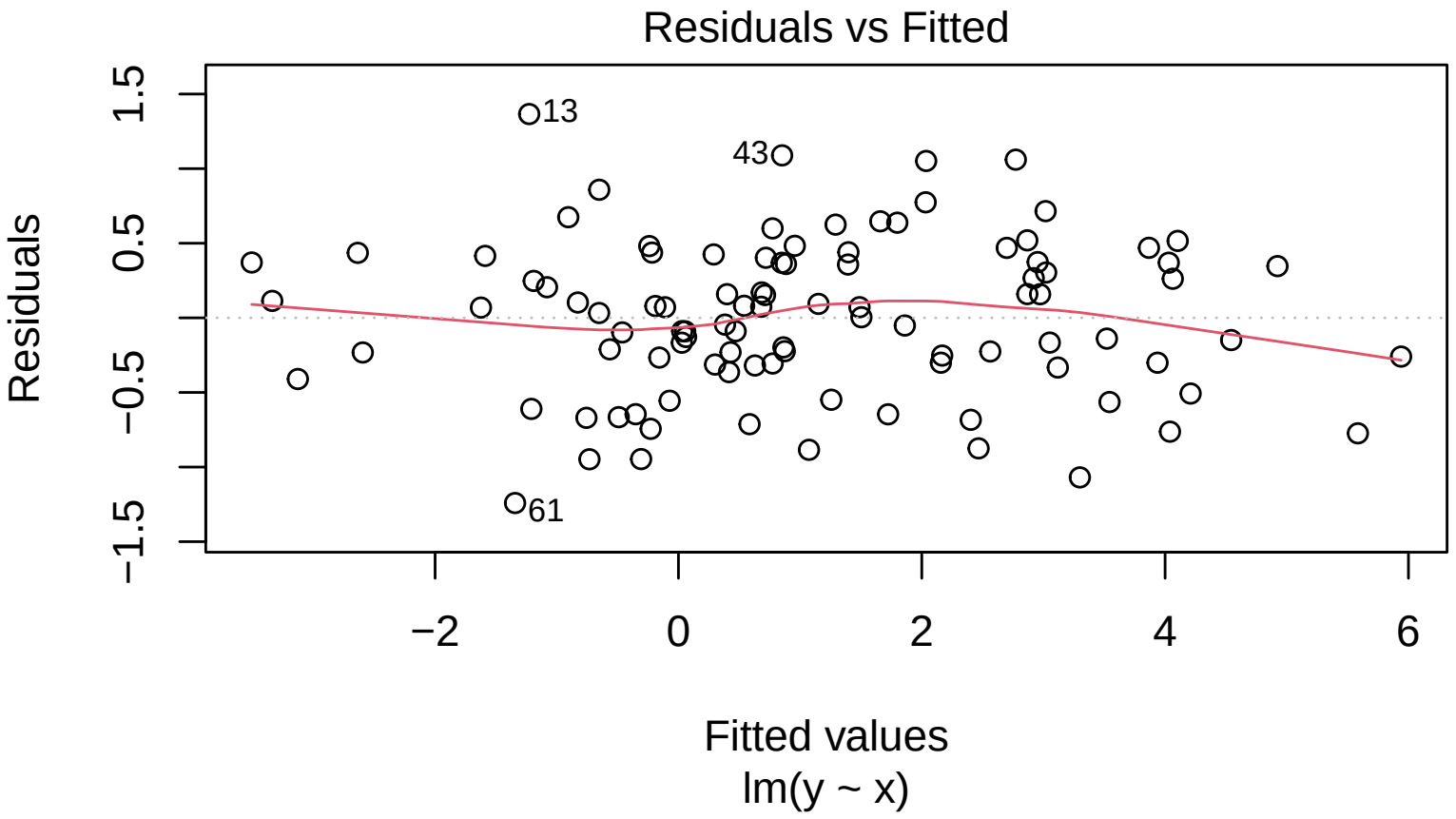
1. For you model fitted previously, test the homoscedasticity assumption.

Normality

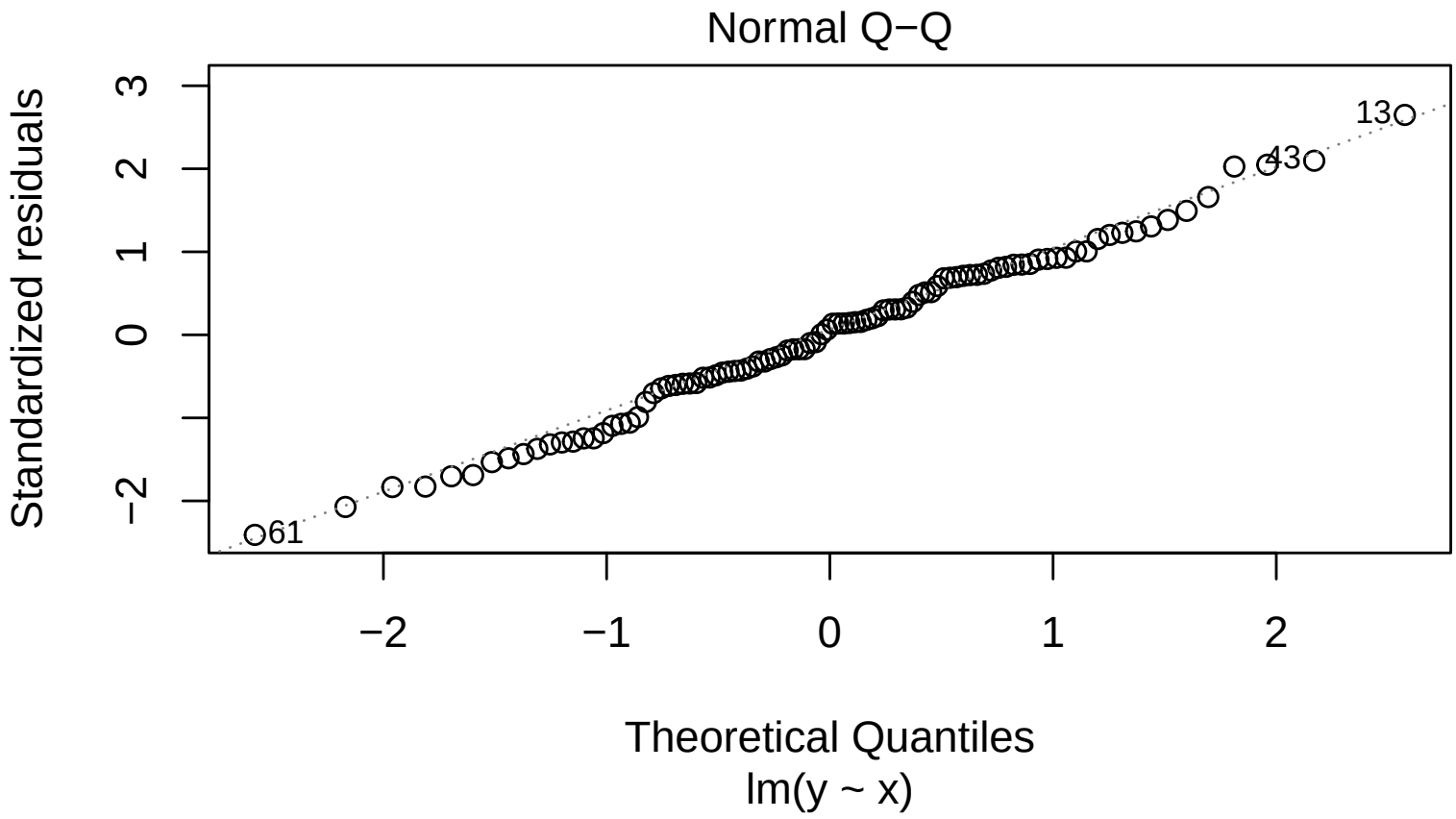
A normal model



The residual vs fitted plot



The residual QQ plot

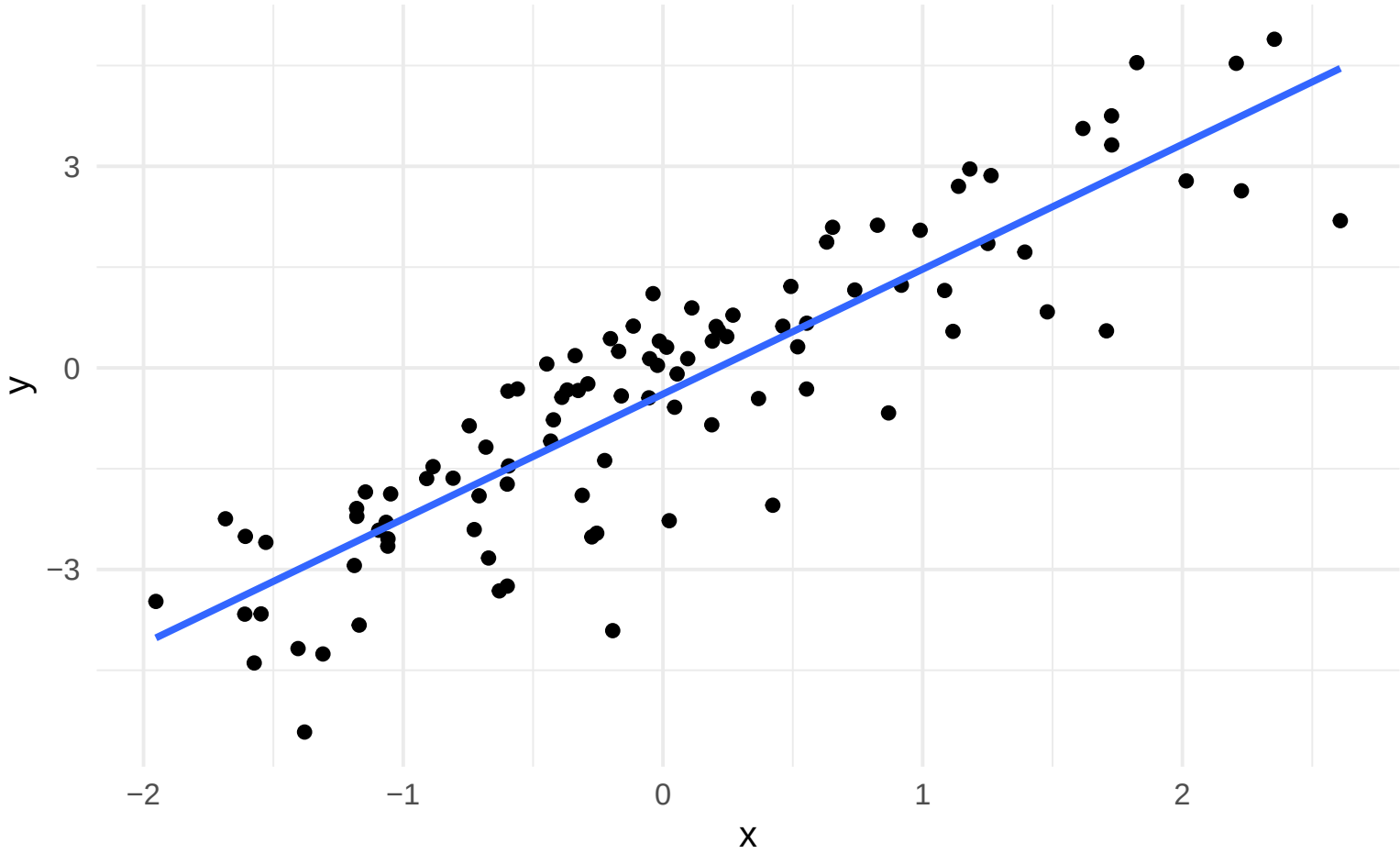


The true model

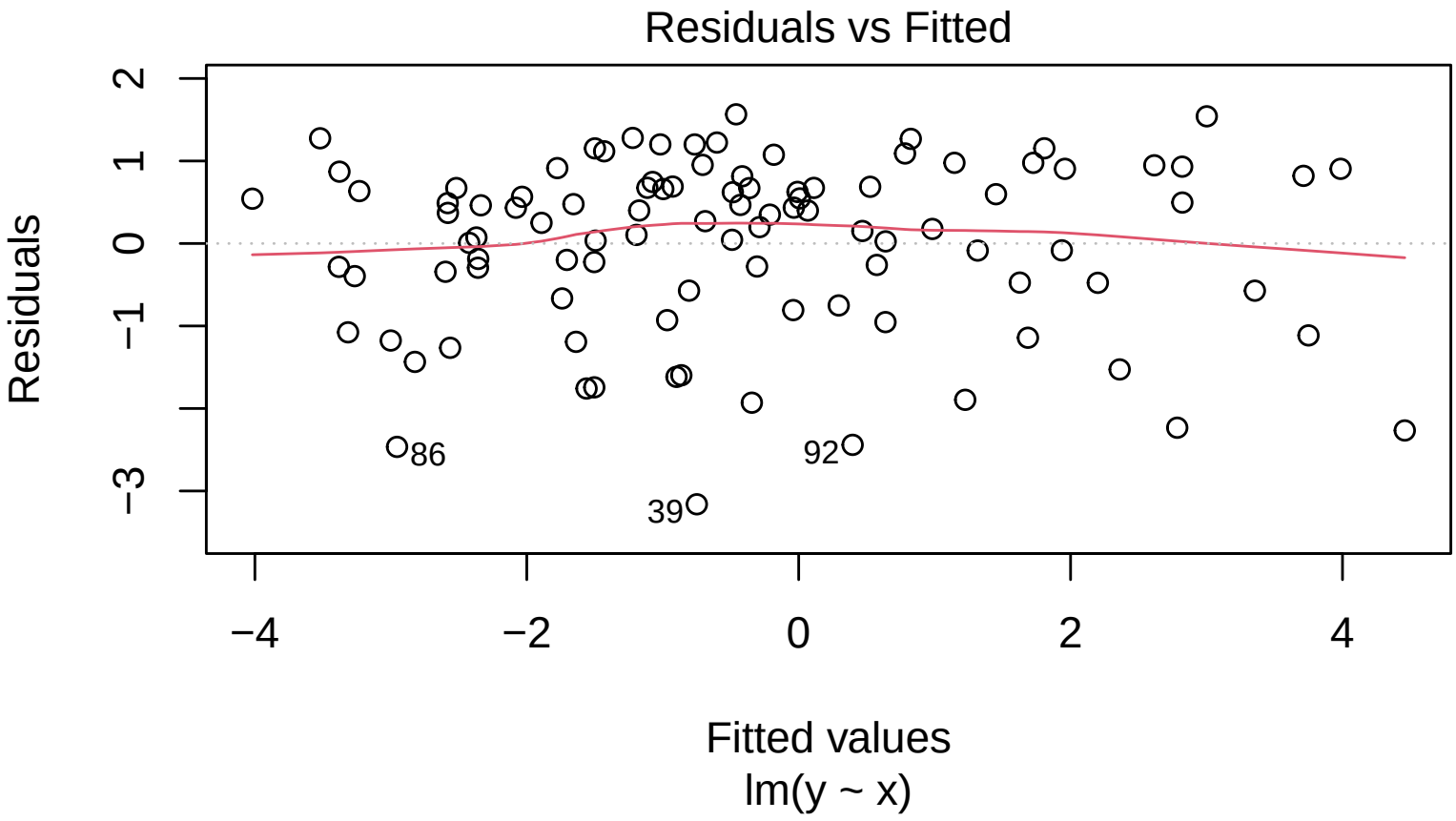
$$y_i = 1 + 2x_i + \varepsilon_i$$

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

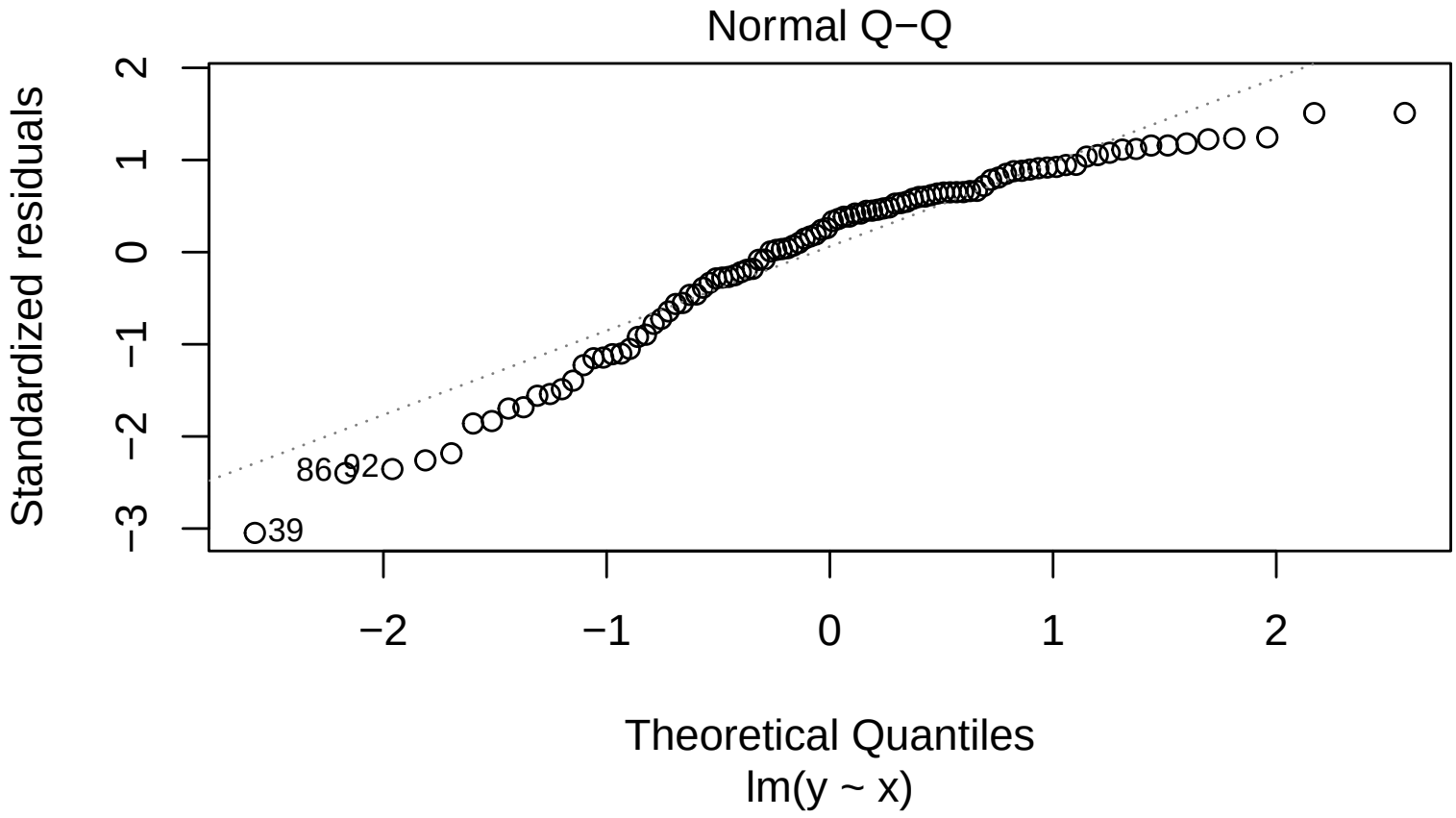
A non-normal model



The residual vs fitted plot



The residual QQ plot



The true model

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

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

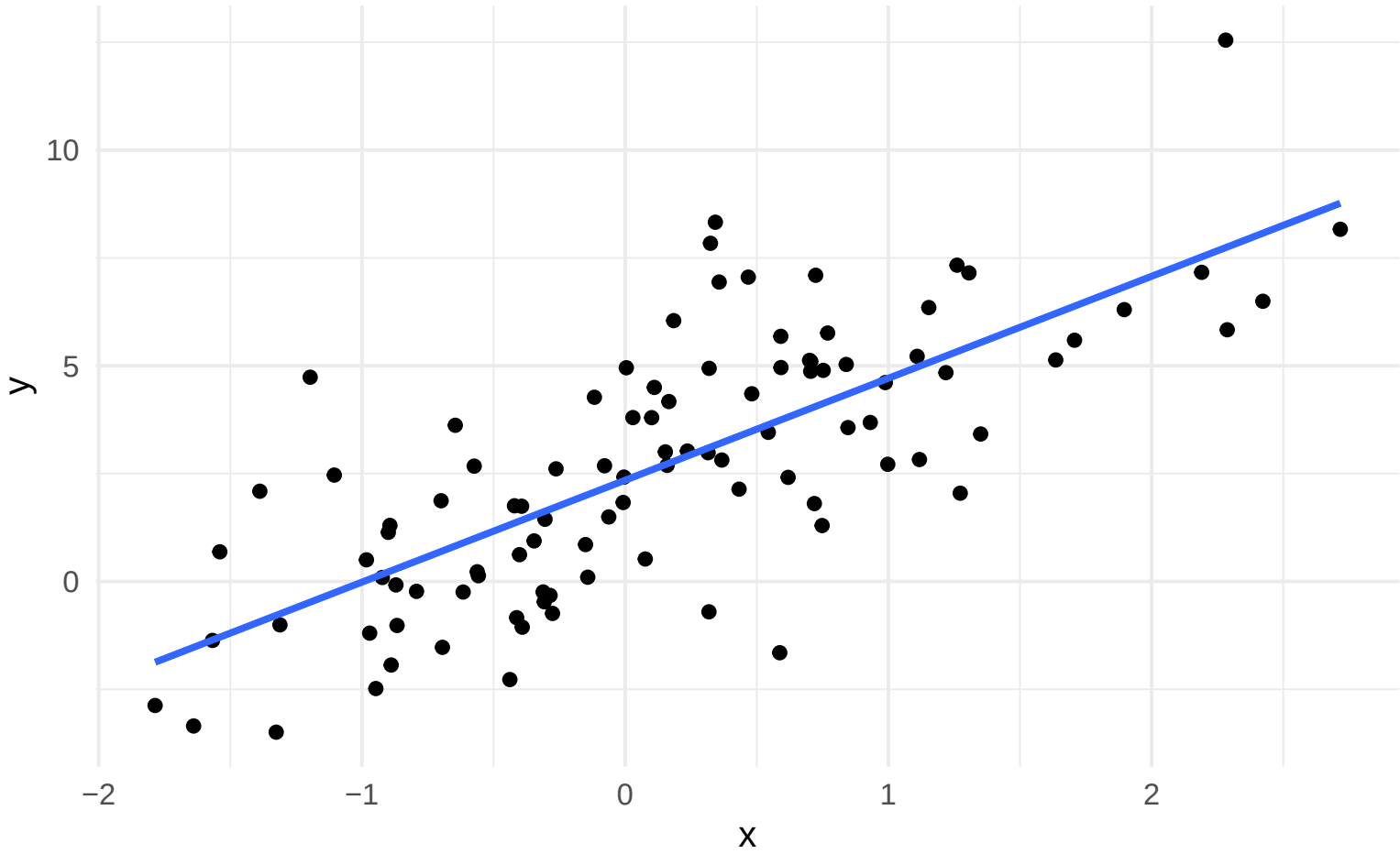
Your turn

What to do

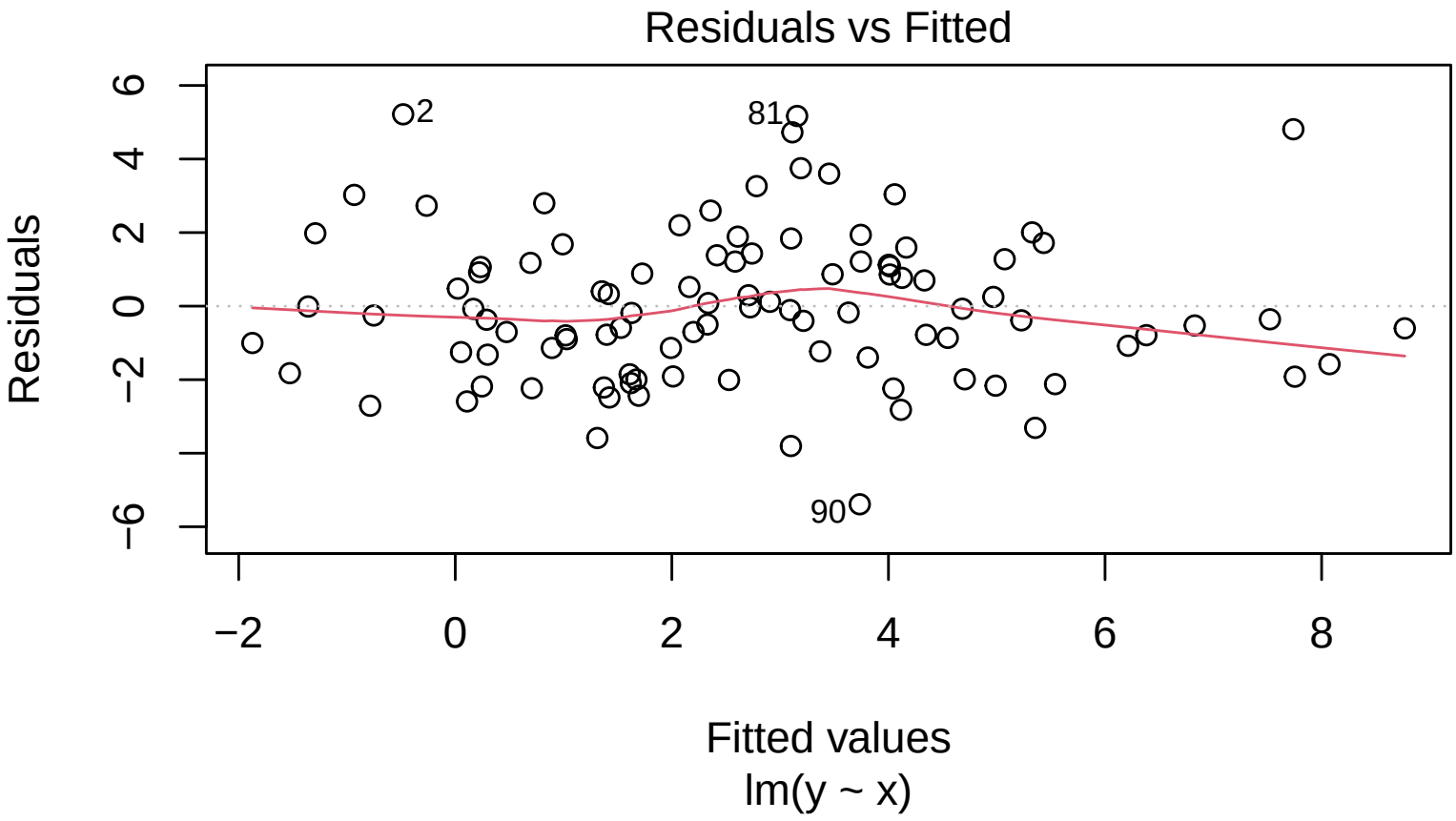
1. For you model fitted previously, test the normality assumption.

One more assumption

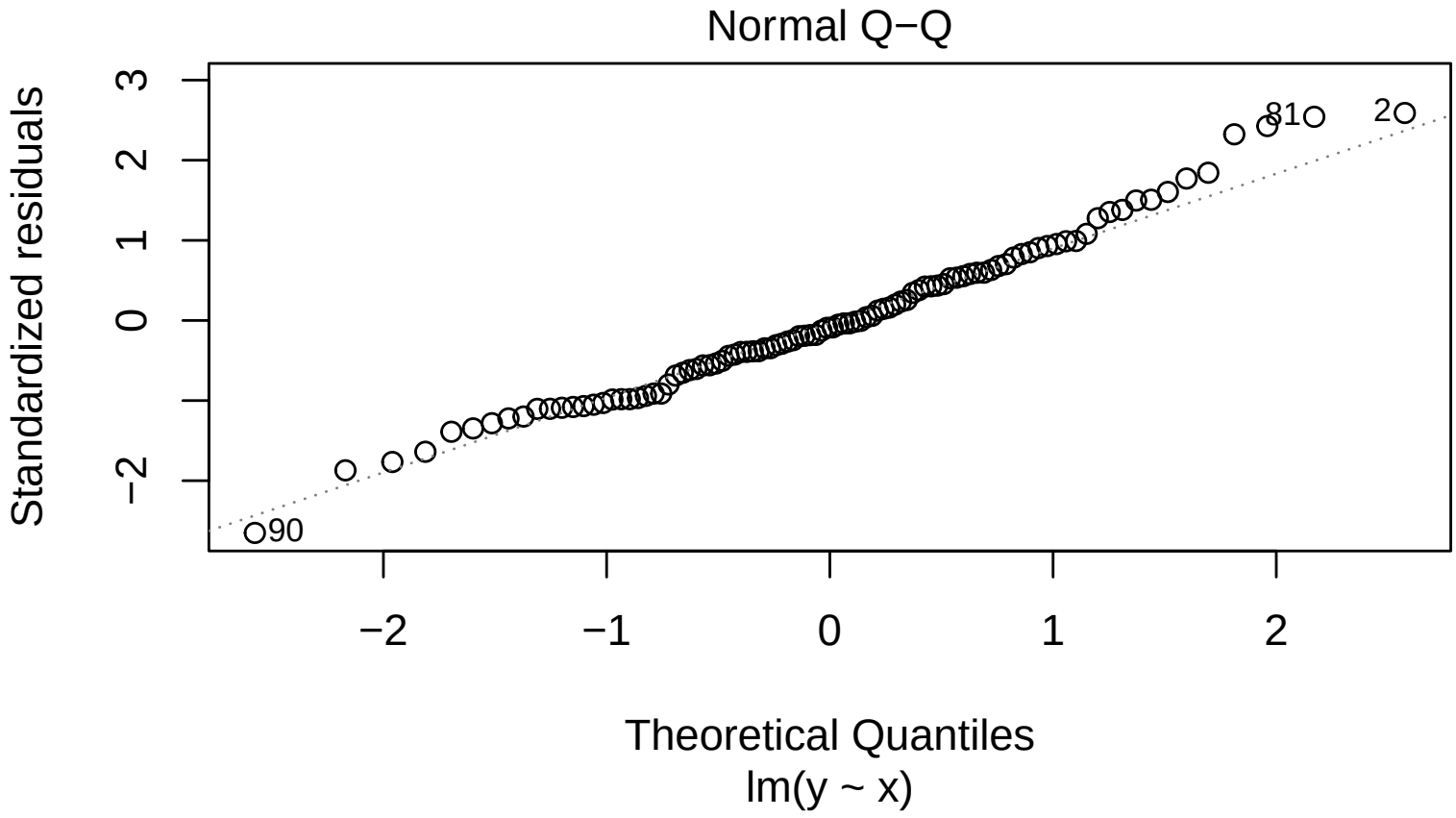
The plot



The residual vs fitted plot



The redisual 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 you model fitted previously, test the independence assumption.