APTS Statistical Modelling

Lecture Discussion 1

Helen Ogden

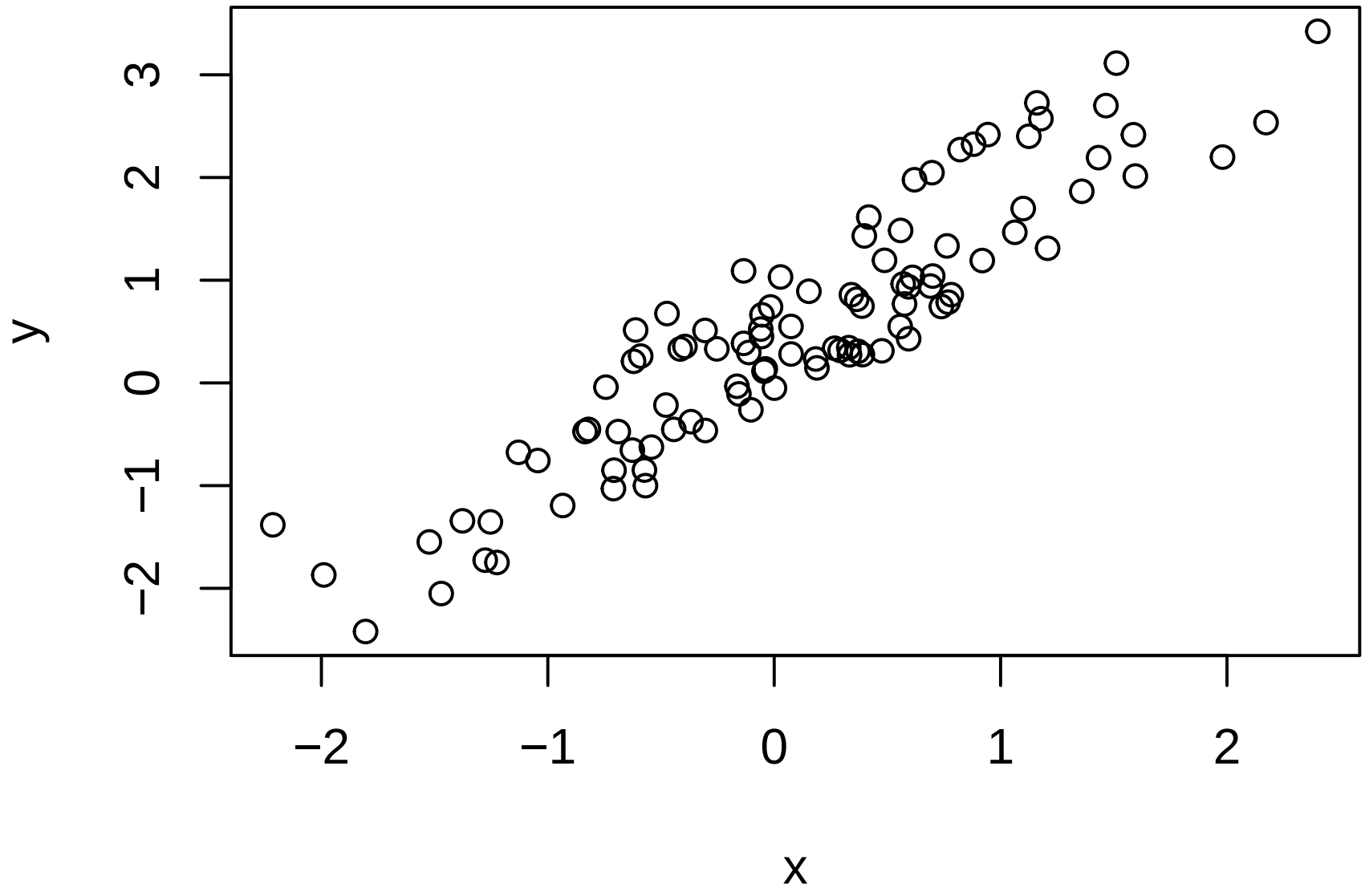
April 2021

Lecture discussion 1

All models are wrong, but some models are useful
– George Box (1919–2013)

- Please mute your microphone and turn off video, unless you are asking a question.
- Feel free to ask questions about any of the lecture content so far.
- Use the chat to ask questions directly, or to state that you have a question.
- For the two computer labs, if you would like to be assigned to a group, request this on the Statistical Modelling Teams channel by 13:00 today.

How should we model this data?



Fitting a linear model

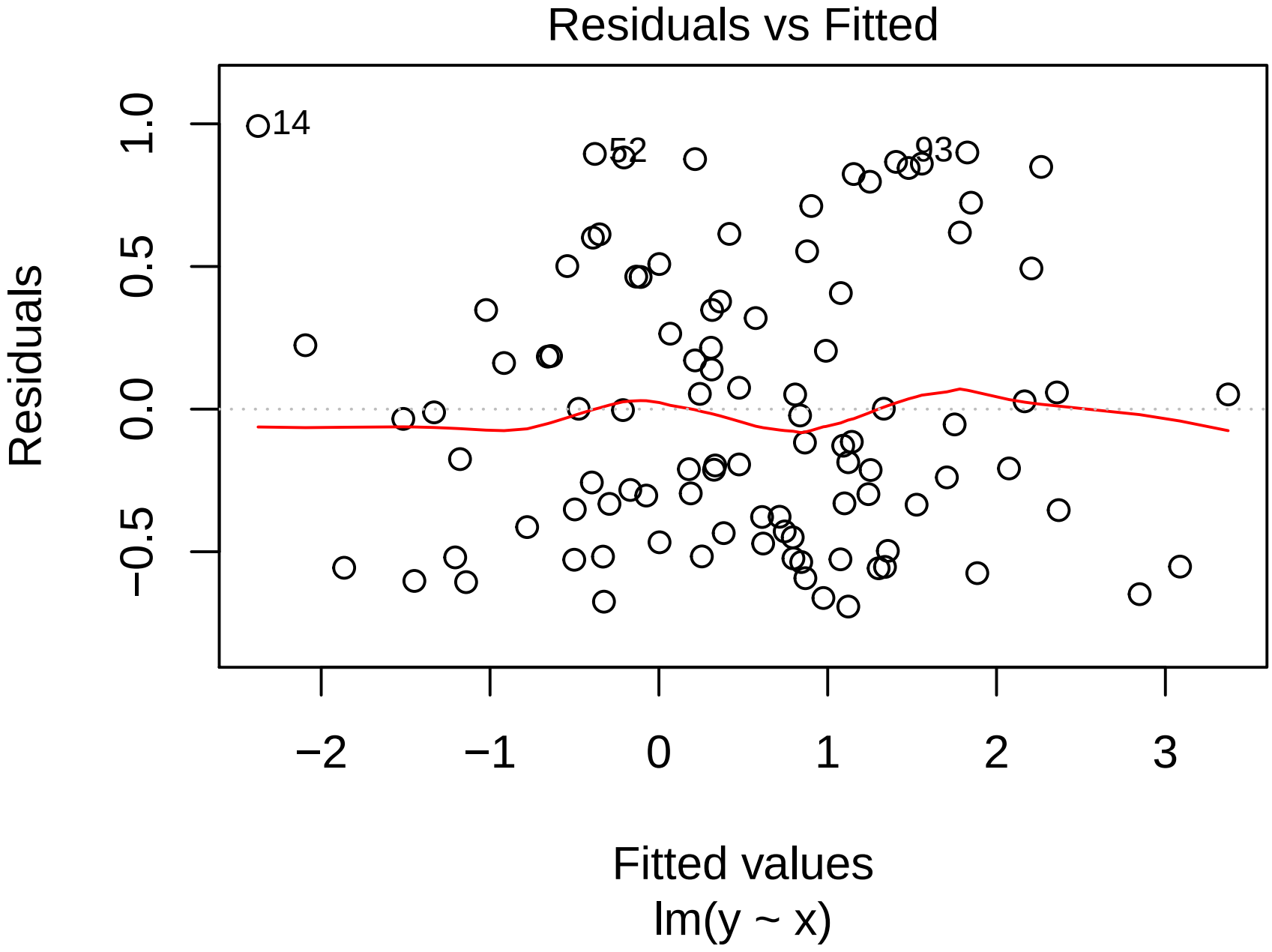
We can fit the normal linear model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

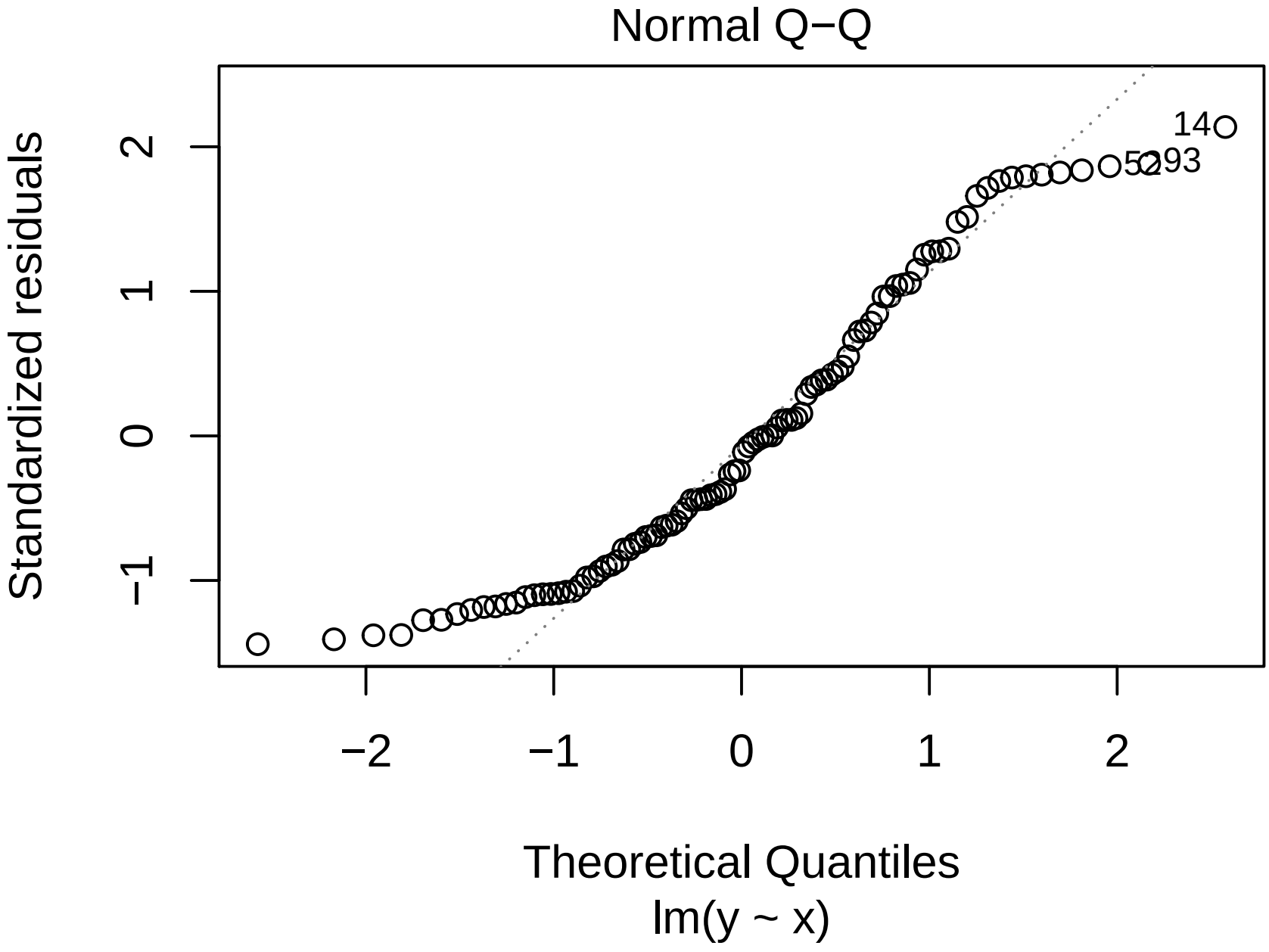
```
mod <- lm(y ~ x)
mod
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##      0.3826       1.2447
```

Checking model fit



Checking model fit



Checking model fit

Are you happy with the model fit?

- Yes, the model seems to fit well.
- No, there is some problem with the model fit.
- Not sure or need more information.

Robustness to model misspecification

In reality

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim U(-a, a),$$

(with $\beta_0 = 0.5$, $\beta_1 = 1.2$ and $a = 0.87$, chosen so that $\text{var}(\epsilon_i) = 0.25$.)

We fit the normal linear model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

Robustness to model misspecification

In reality

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim U(-a, a),$$

(with $\beta_0 = 0.5$, $\beta_1 = 1.2$ and $a = 0.87$, chosen so that $\text{var}(\epsilon_i) = 0.25$.)

We fit the normal linear model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

Will our inference be correct, even though the model is wrong?

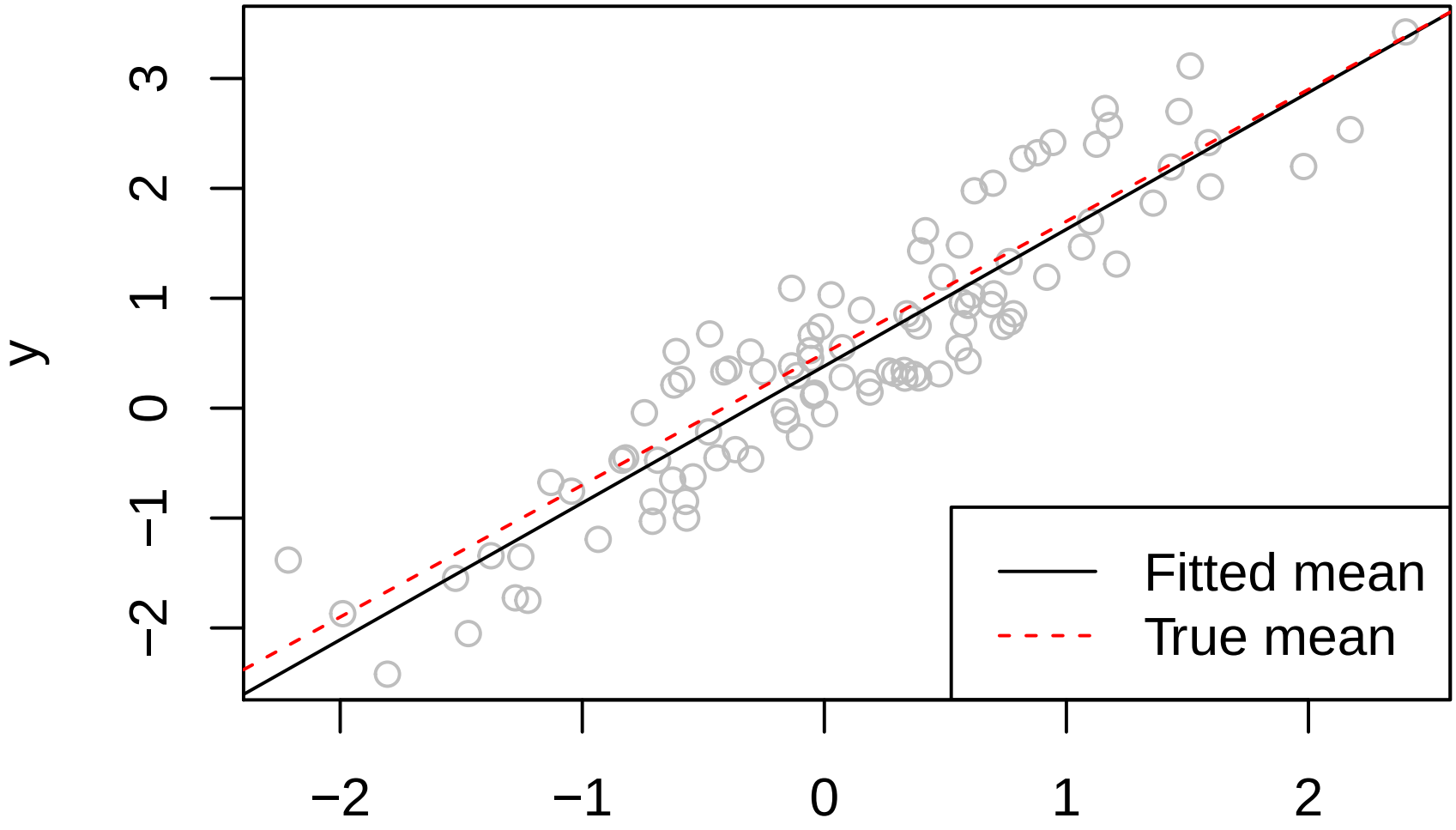
- Yes, the inference will be correct because the linear model is robust to misspecification of the error distribution.
- No, the model was incorrectly specified so the inference will be incorrect.
- Not sure or need more information.

Estimating the mean response

If we are just interested in the mean response, our fitted mean

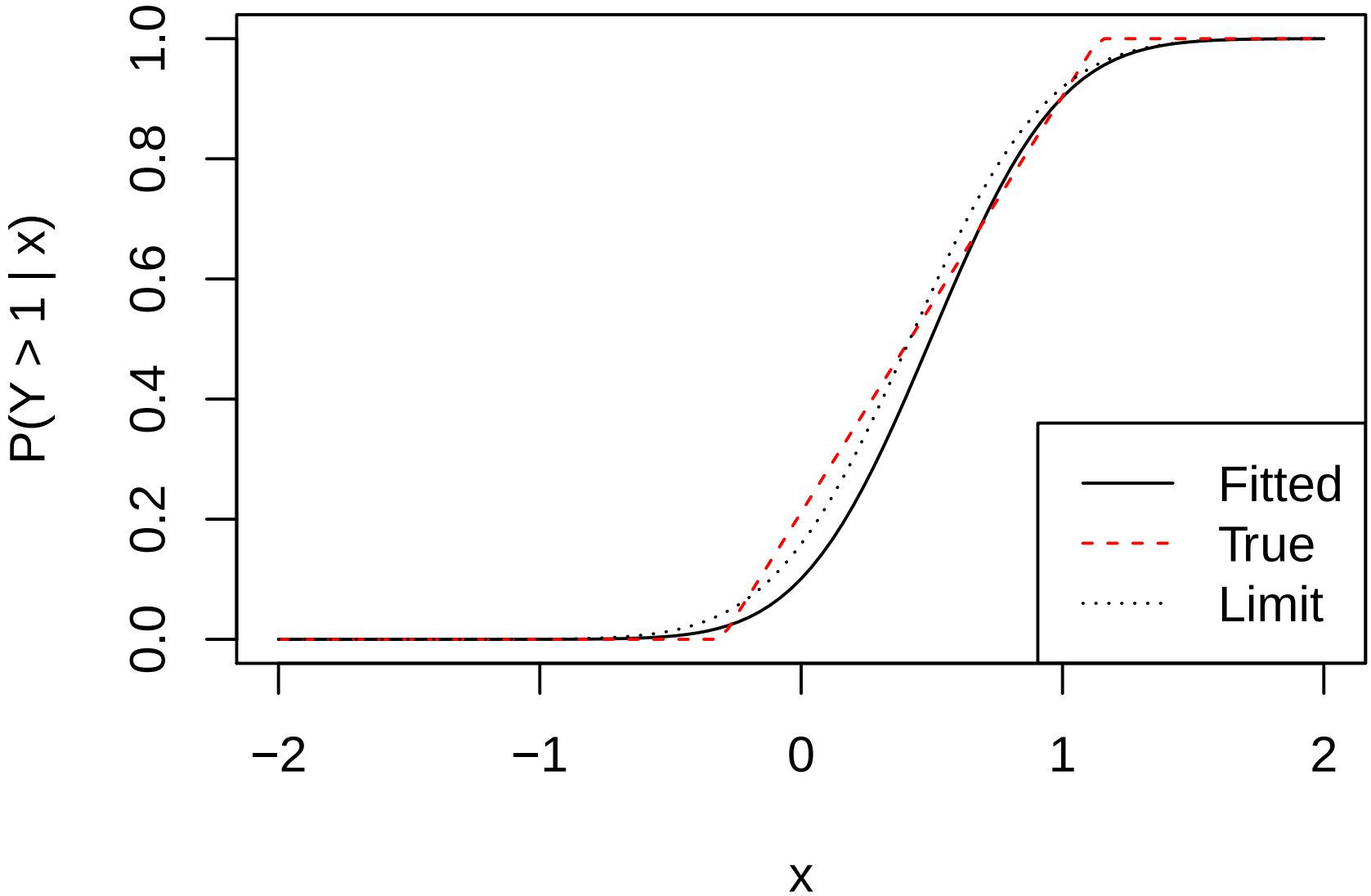
$$\hat{\mu}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

is a consistent estimator of the true mean $\mu(x)$.



Estimating tail probabilities

If we are interested in some aspect other than the mean response, the model misspecification matters. e.g. if we estimate $P(Y > 1|x)$, we will no longer get a consistent estimator from our model.



Comparing against a quadratic model

We could compare against a linear model with quadratic dependence on x:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

```
mod2 <- lm(y ~ poly(x, 2))
c(AIC(mod), AIC(mod2))
```

```
## [1] 142.4568 144.4419
```

Comparing against a quadratic model

We could compare against a linear model with quadratic dependence on x:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

```
mod2 <- lm(y ~ poly(x, 2))
c(AIC(mod), AIC(mod2))
```

```
## [1] 142.4568 144.4419
```

We prefer the simpler model, but that should **not** be used as evidence that the simpler model is correct.

Summary: impacts of model misspecification

- Robustness to model misspecification is dependent on the quantity of interest.
- Even if a model is very close to the truth (according to KL divergence), it can still give incorrect answers about some quantities of interest.
- Model selection criteria are not a substitute for model checking.