Evaluating linear model assumptions

Max Joseph

March 14, 2017

Recap

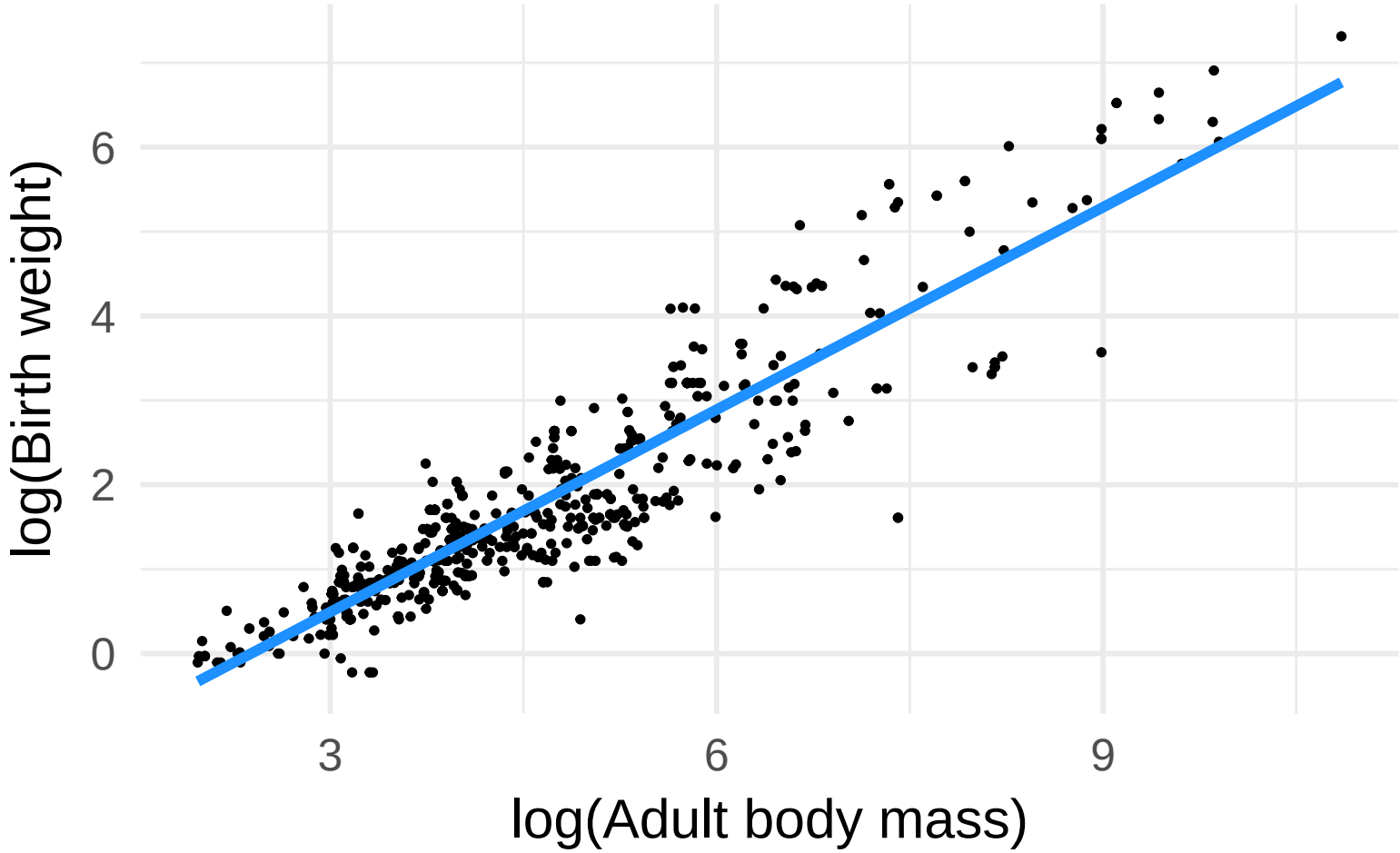
1. Linear relationship between x and y:

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

2. Normal error in y (variation around line):

$$y_i \sim \text{Normal}(\hat{y}_i, \sigma)$$

3. Homoskedasticity (errors have constant variance)
4. Observations are independent


log(Birth weight)
6
4
2
0
3
6
9
log(Adult body mass)

Today: evaluating assumptions

1. Linear relationship between x and y:

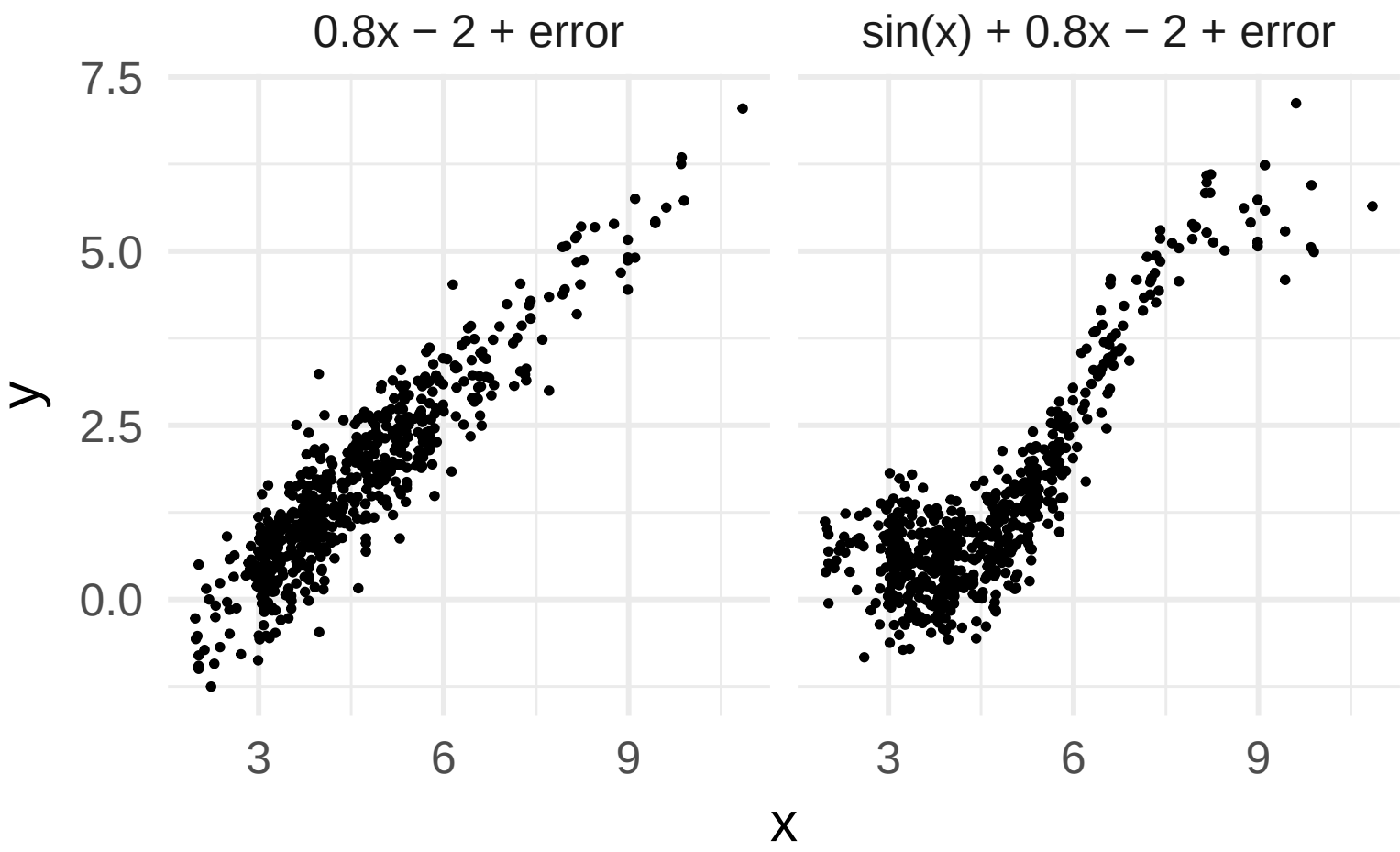
$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

2. Normal error in y (variation around line):

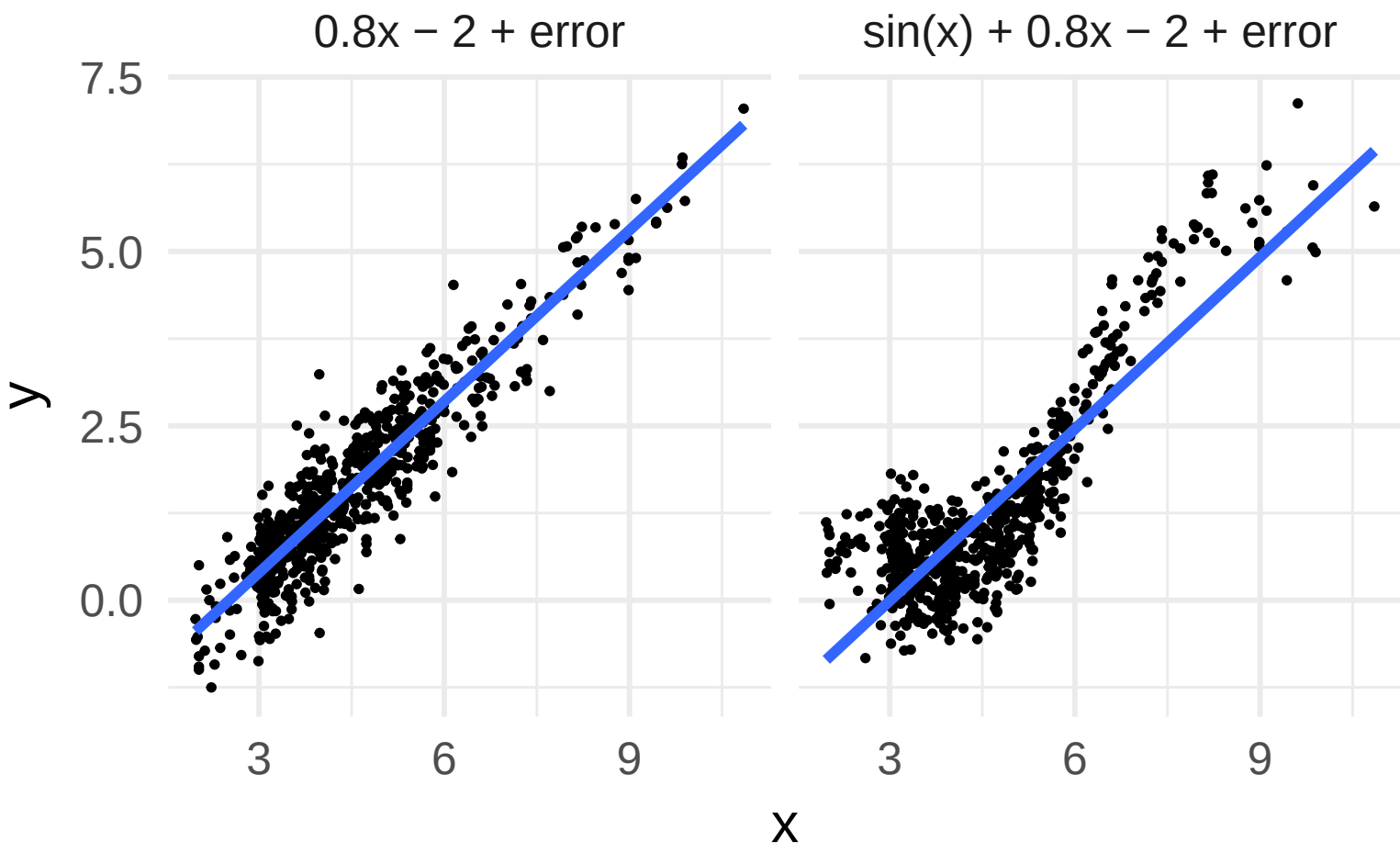
$$y_i \sim \text{Normal}(\hat{y}_i, \sigma)$$

3. Homoskedasticity (errors have constant variance)
4. Observations are independent

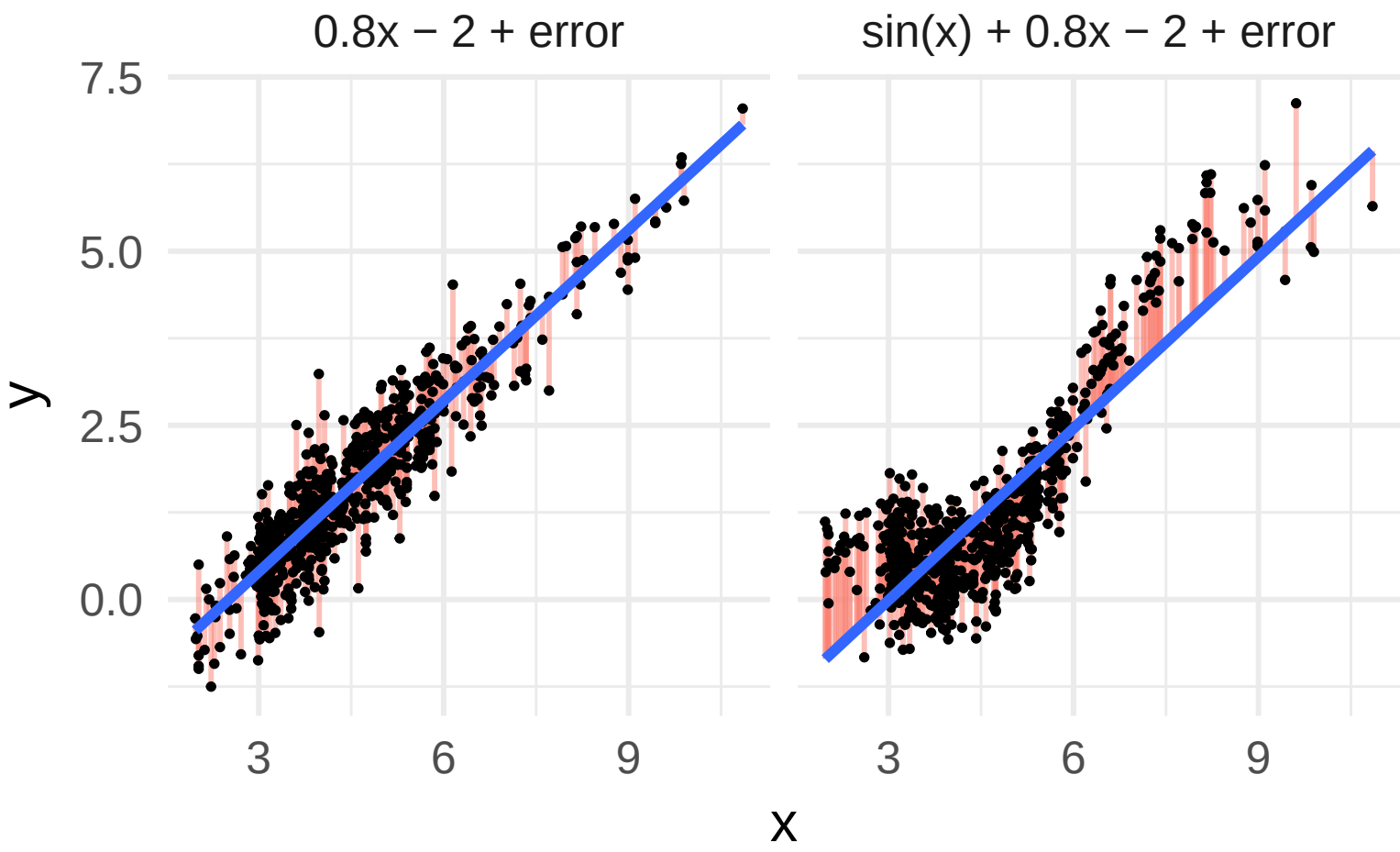
Evaluating linearity



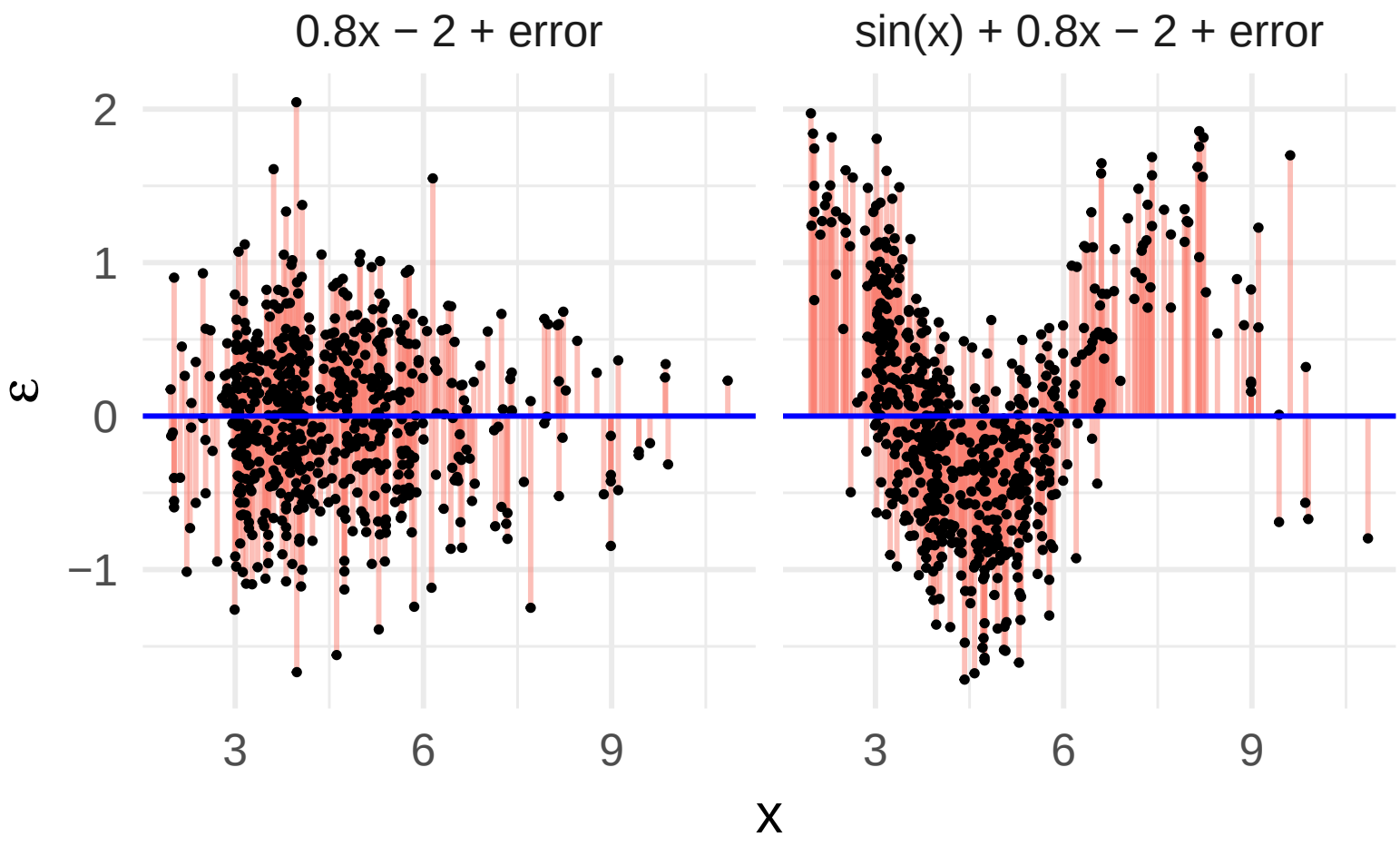
Evaluating linearity



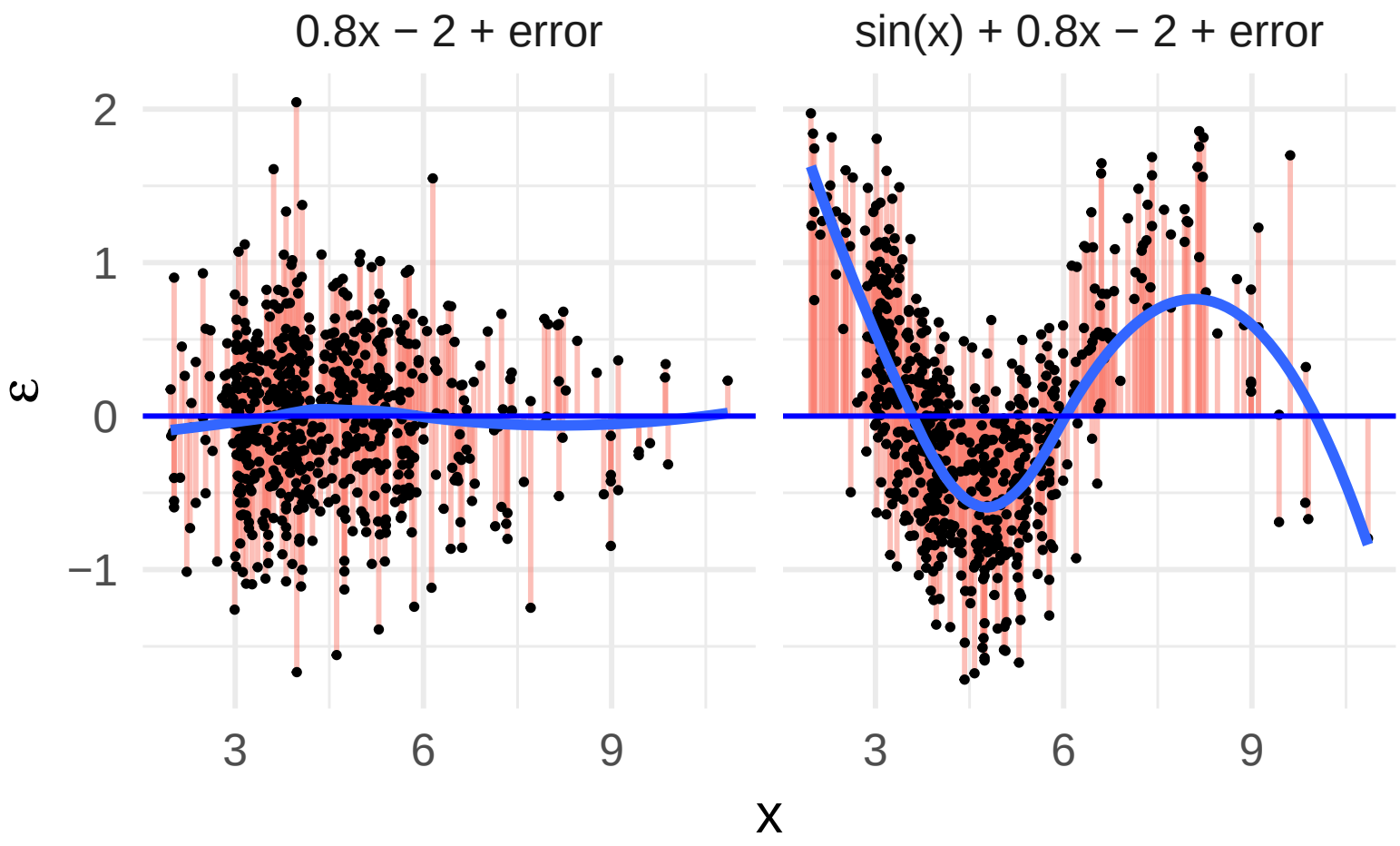
Adding residuals



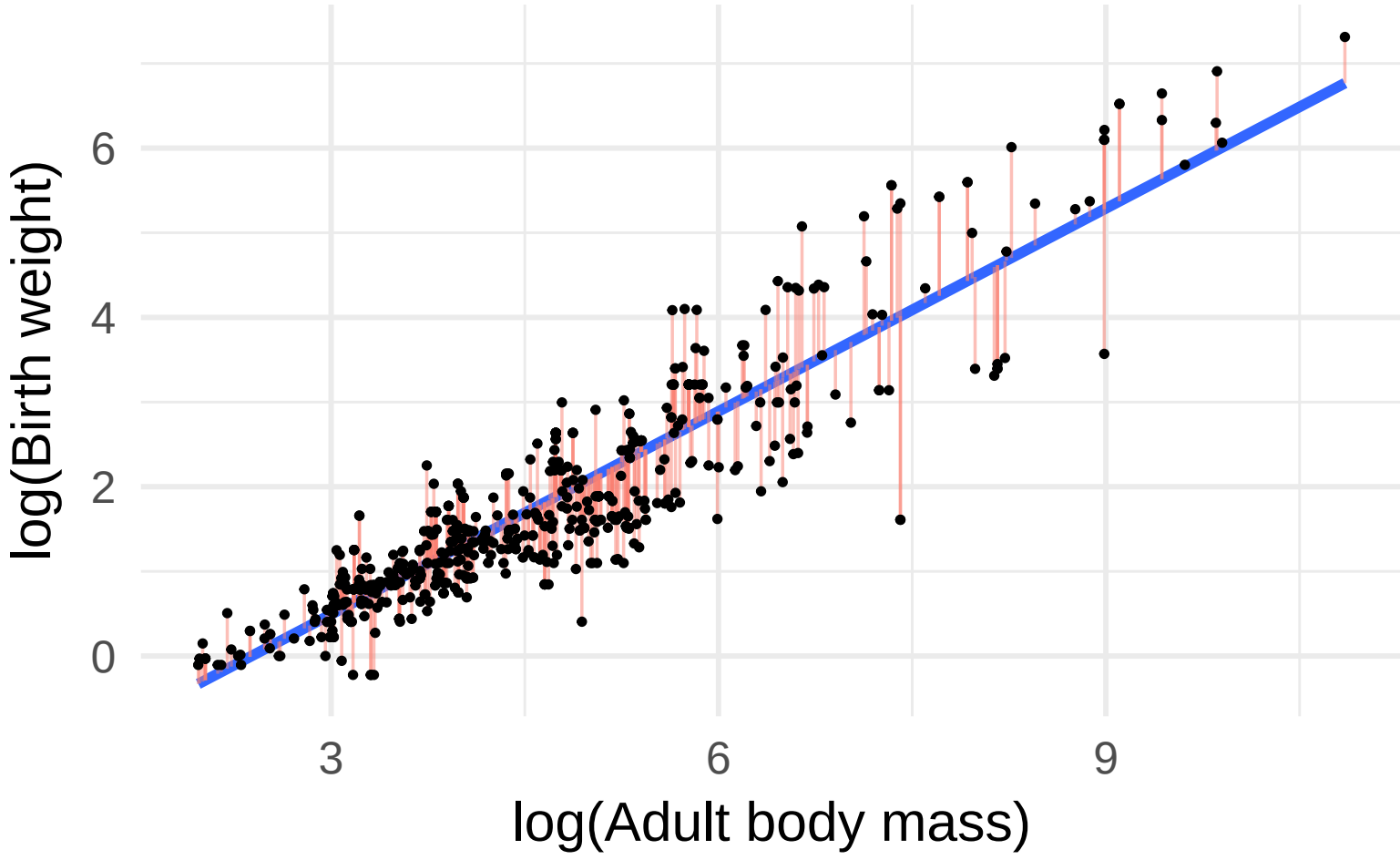
Show residuals on y-axis



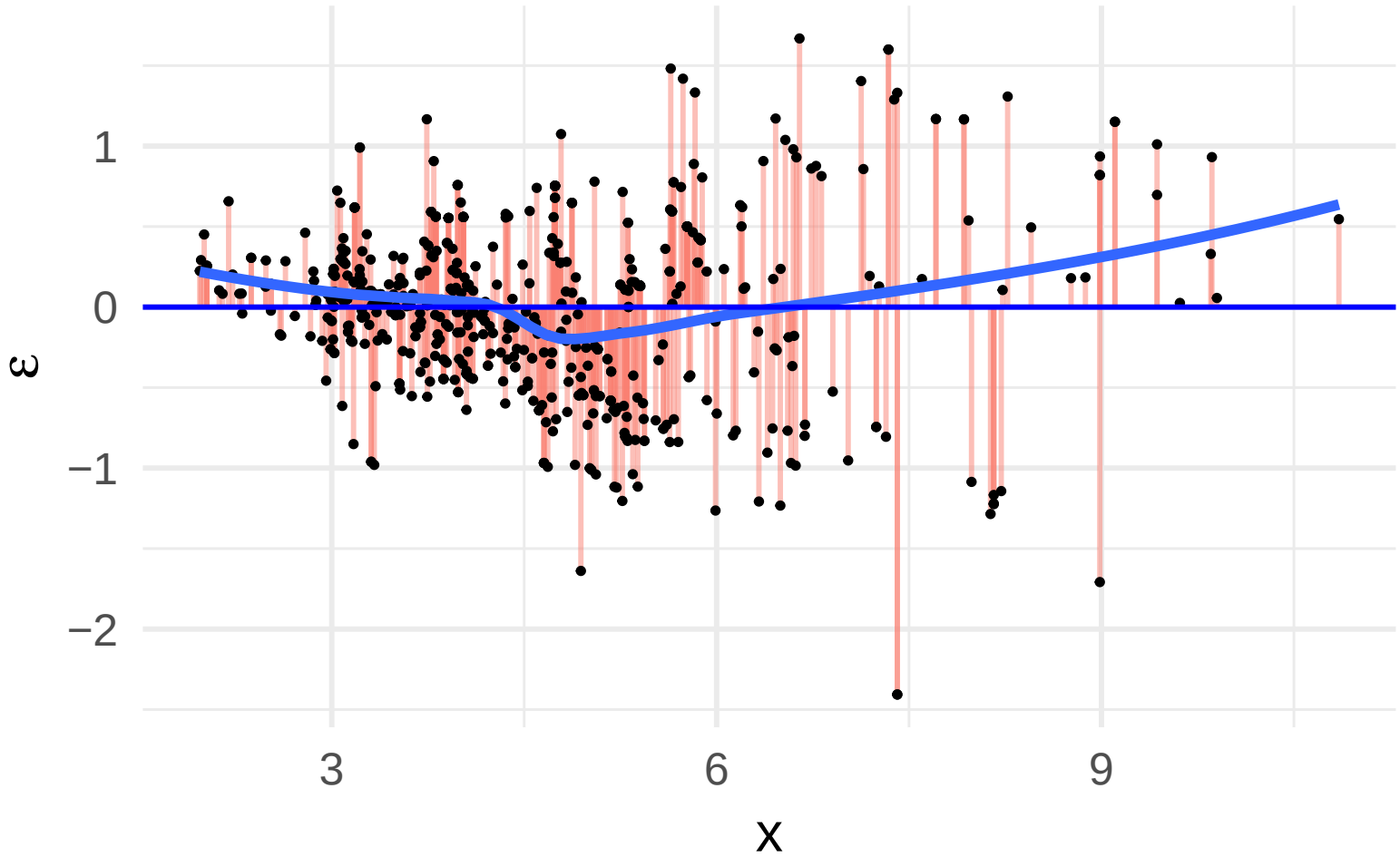
Look for patterns in the residuals



A real example



Rodent birth weight residuals

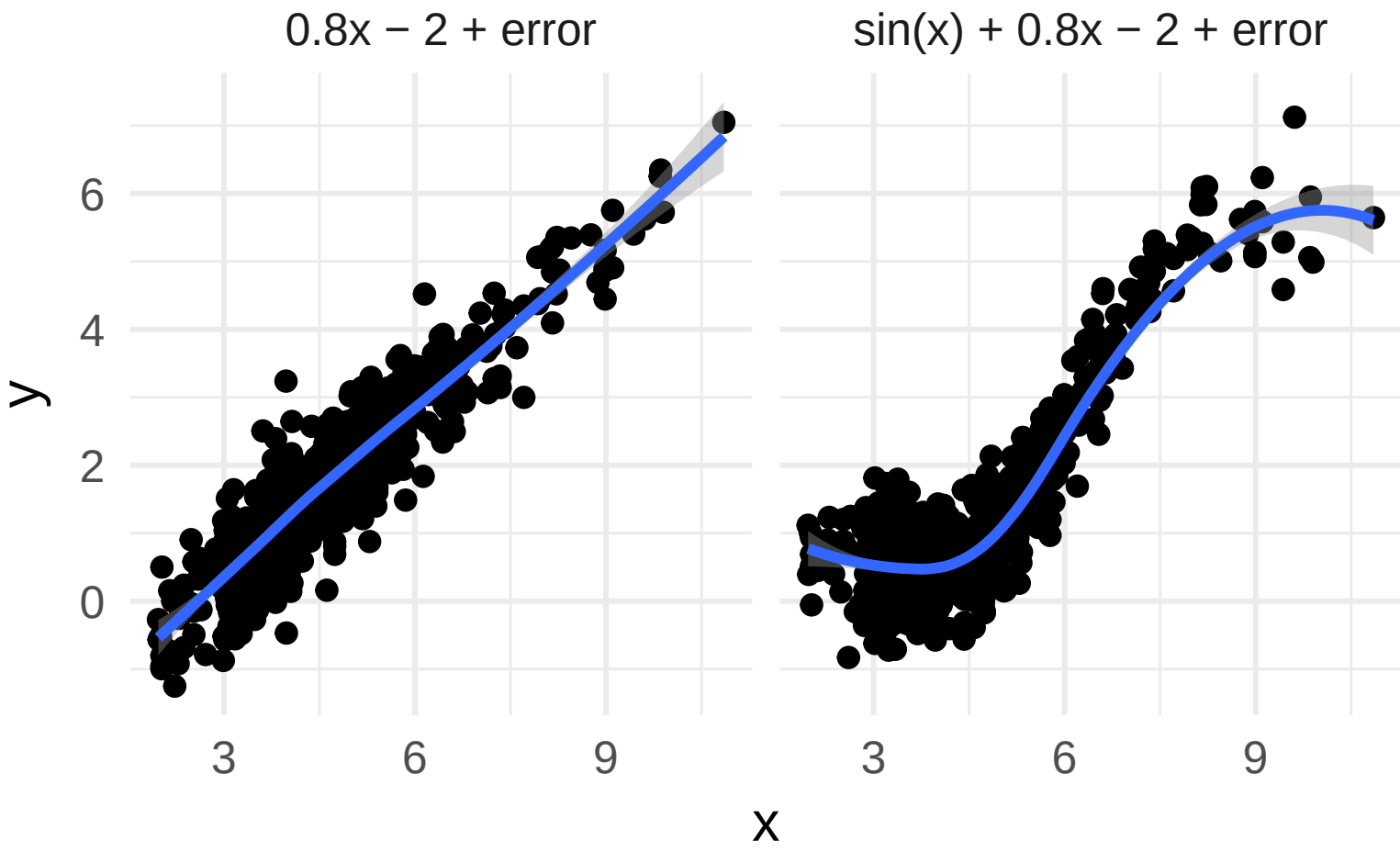


Recap: evaluating linearity

1. Plot x vs. y
2. Plot x vs. $y - \hat{y}$ (residuals)

Solution: what to do with nonlinearity

Model it!



Evaluating homoskedasticity

$$y_i \sim \text{Normal}(\hat{y}_i, \sigma)$$

$$\hat{y}_i = \alpha_i + \beta x_i$$

For all y_i, σ is assumed constant

$$\rightarrow$$

plot y_i vs. residuals

Activity

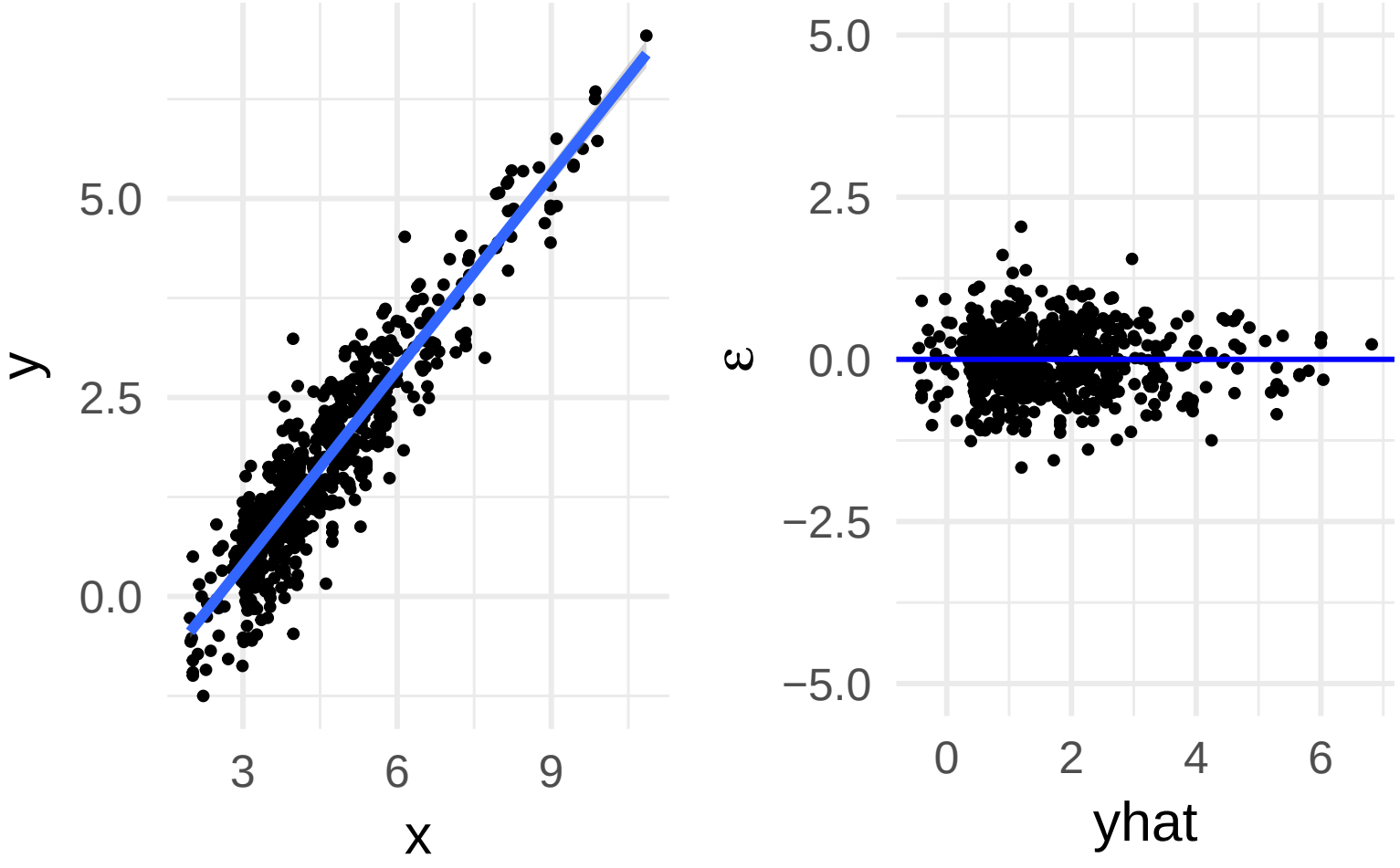
Drawing heteroskedasticity

- pair up
- draw 2 different cartoon examples of heteroskedasticity on the board in two coordinate systems (4 graphs total):

1. x vs. y
2. $\hat{y}$ vs. ϵ

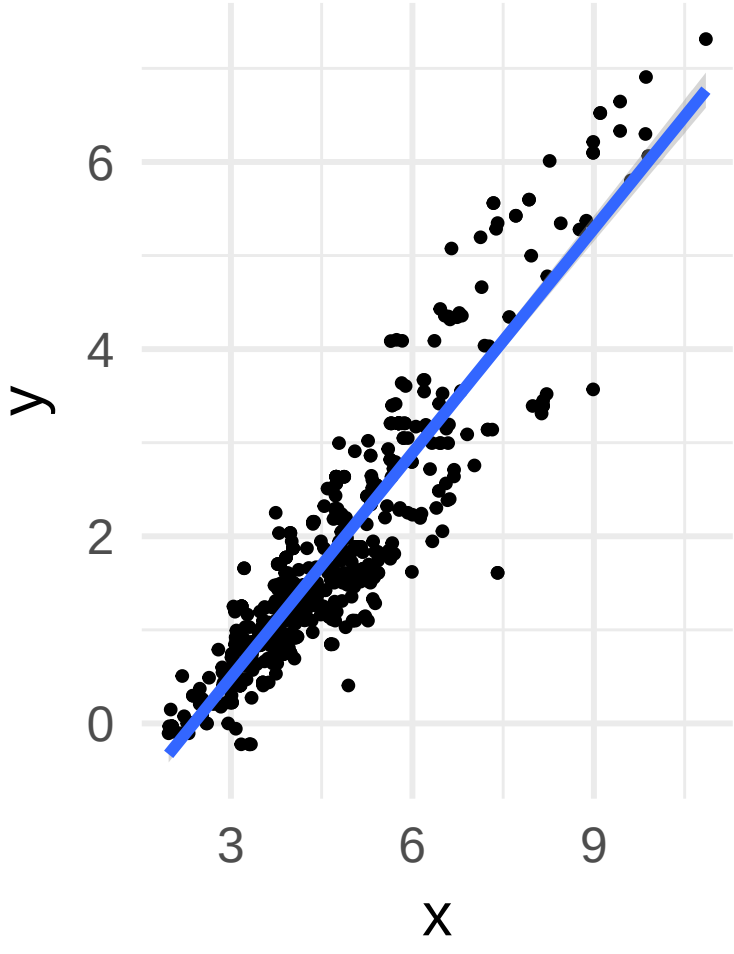
Evaluating homoskedasticity

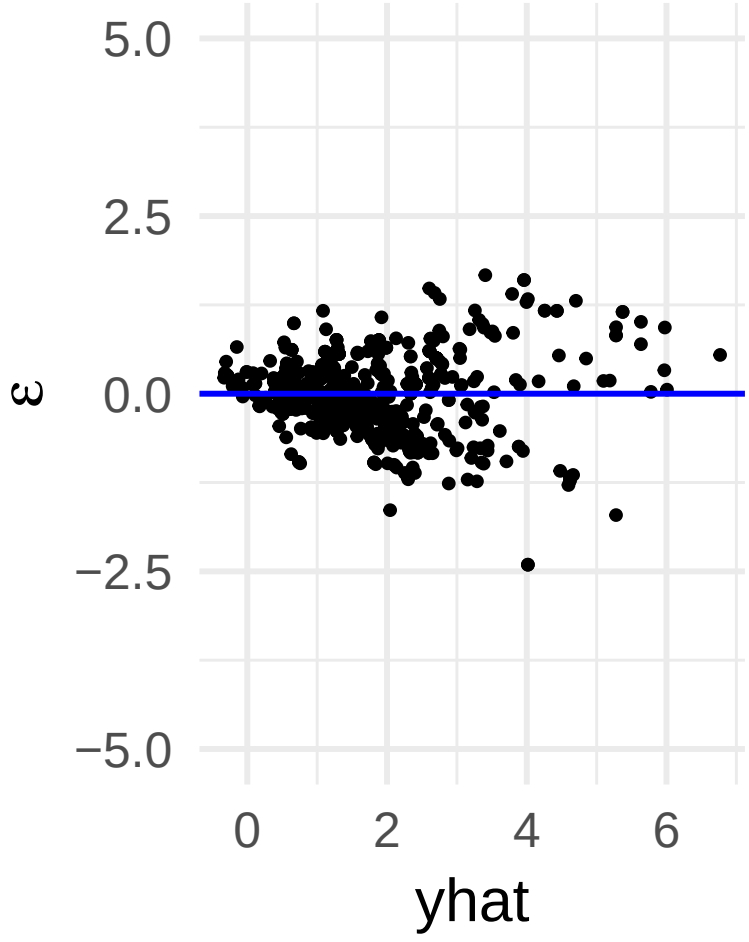
Simualted example



Evaluating homoskedasticity

Rodent example

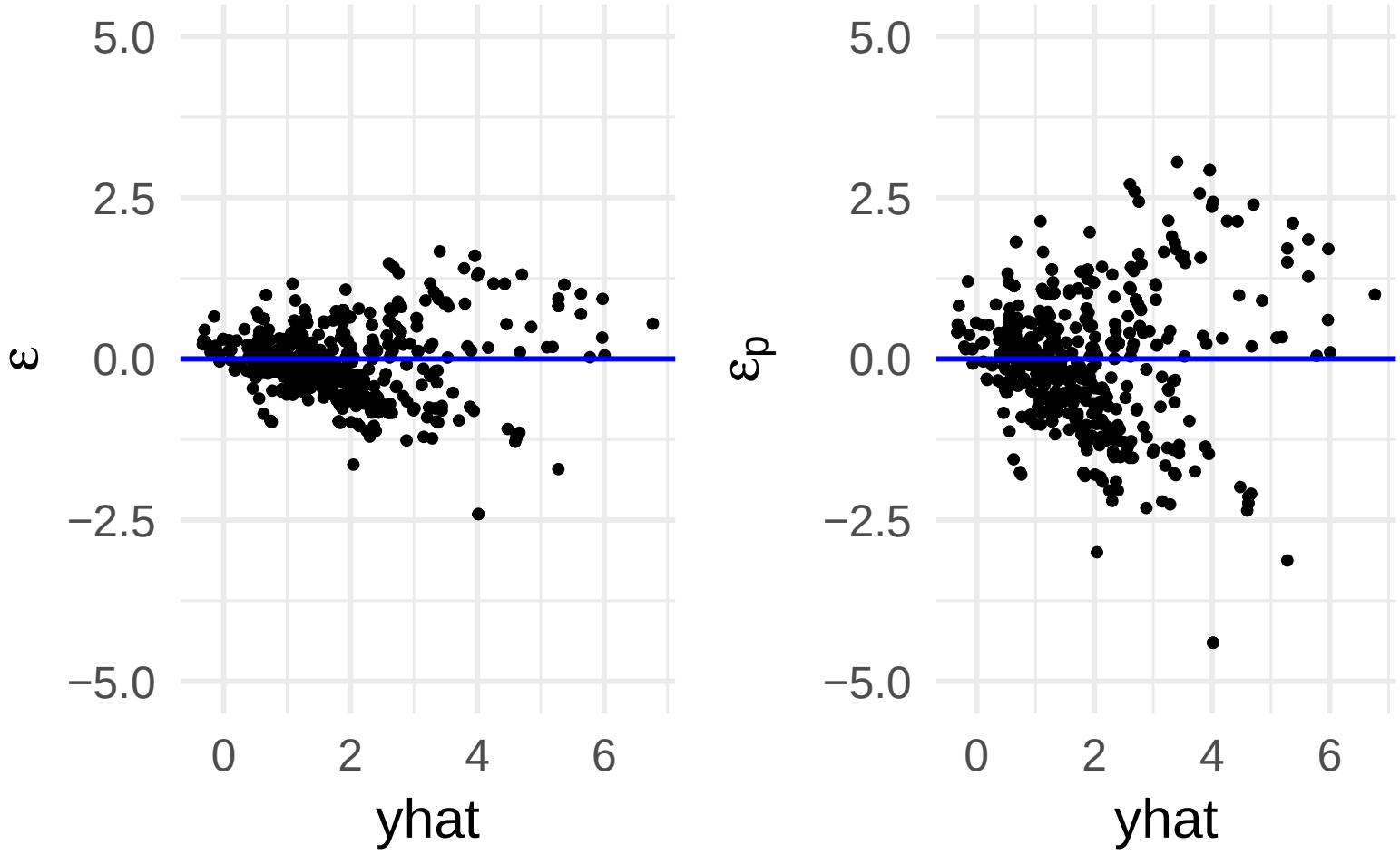




Pearson residuals

Rescaling $\rightarrow$ force standard deviation $= 1$

$$\epsilon_{\mathsf{p}} = \epsilon / \hat{\sigma}$$

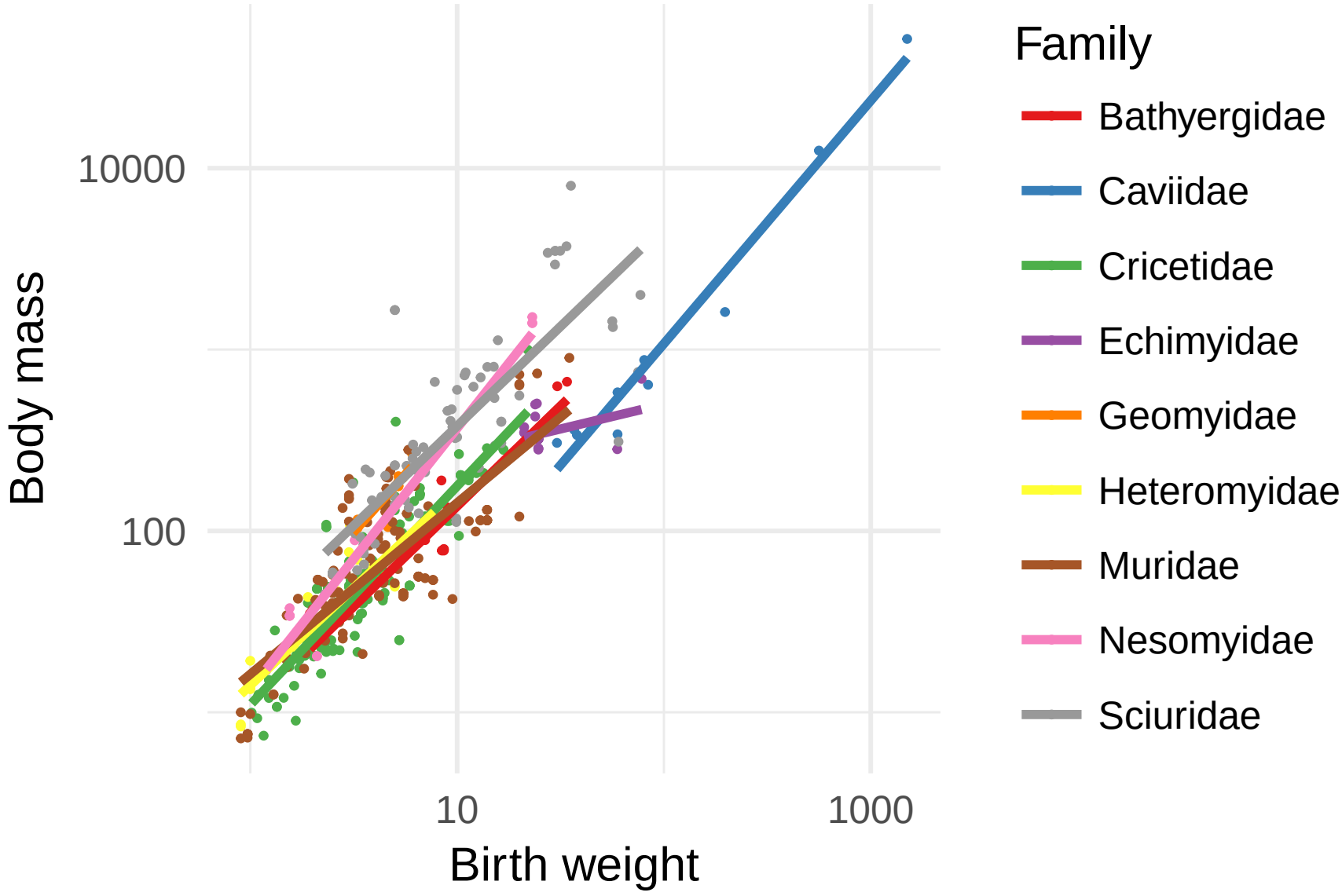


Dealing with heteroskedasticity

Most of the time:

- figure out what is causing the extra variation and model it

Example: Rodents in different families



Transformations of y

- $\log(y)$: variance increases with $\hat{y}$
- $y^{1/c}$: variance increases with $\hat{y}$
- y^c: variance decreases with $\hat{y}$

Problems with transformations

1. Inflexible
2. Consequences for model interpretation

Natural scale

$$\hat{y} = \alpha + \beta x$$

Log scale

$$\log(\hat{y}) = \alpha + \beta x$$

$$\rightarrow \hat{y} = e^{\alpha} e^{\beta x}$$

Recap: evaluating homoskedasticity

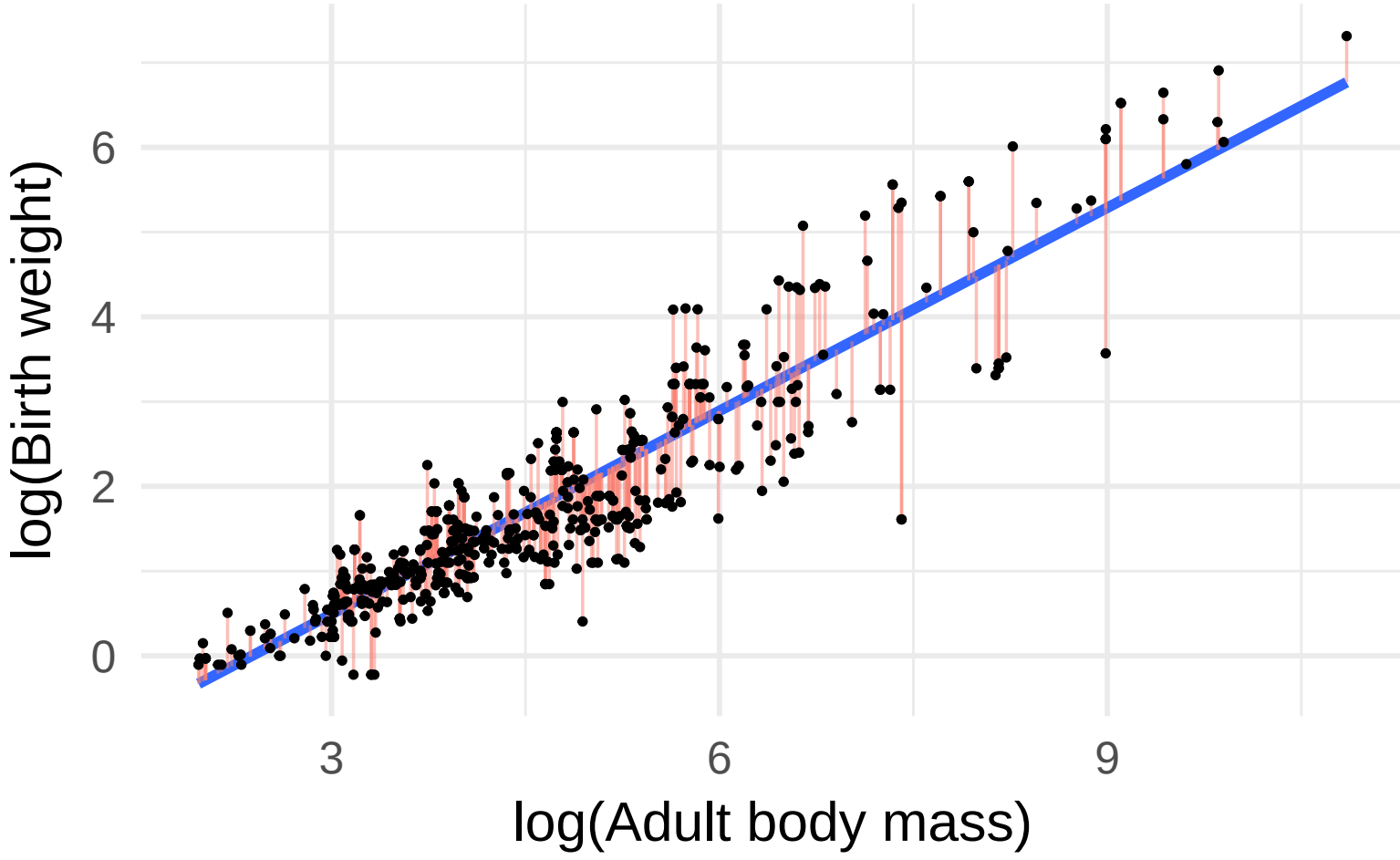
Diagnostics

- plot $\hat{y}$ vs. ϵ and look for "funnels"

Solutions

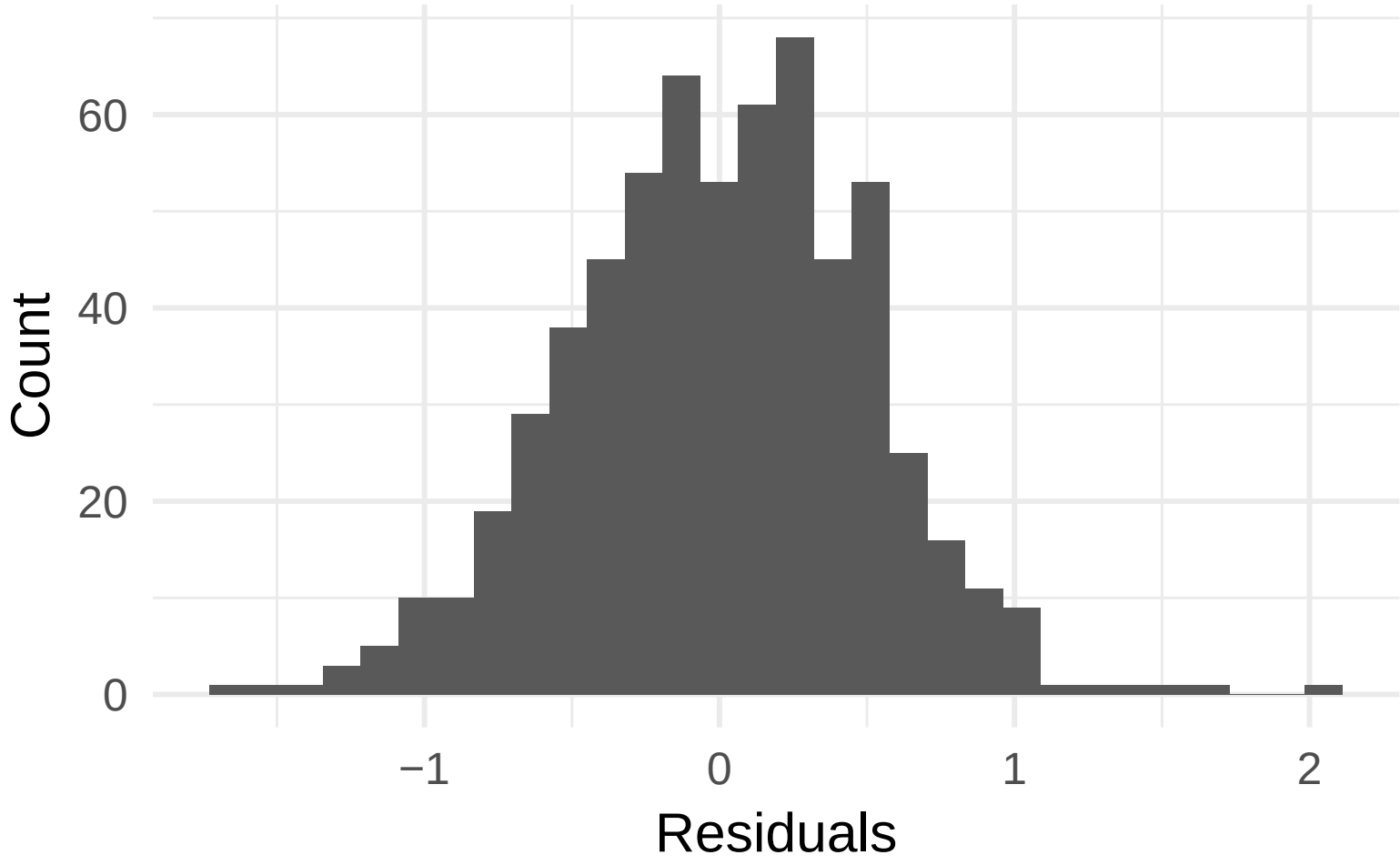
1. Model the source of extra variance
2. Consider transformations

Last: evaluating normality



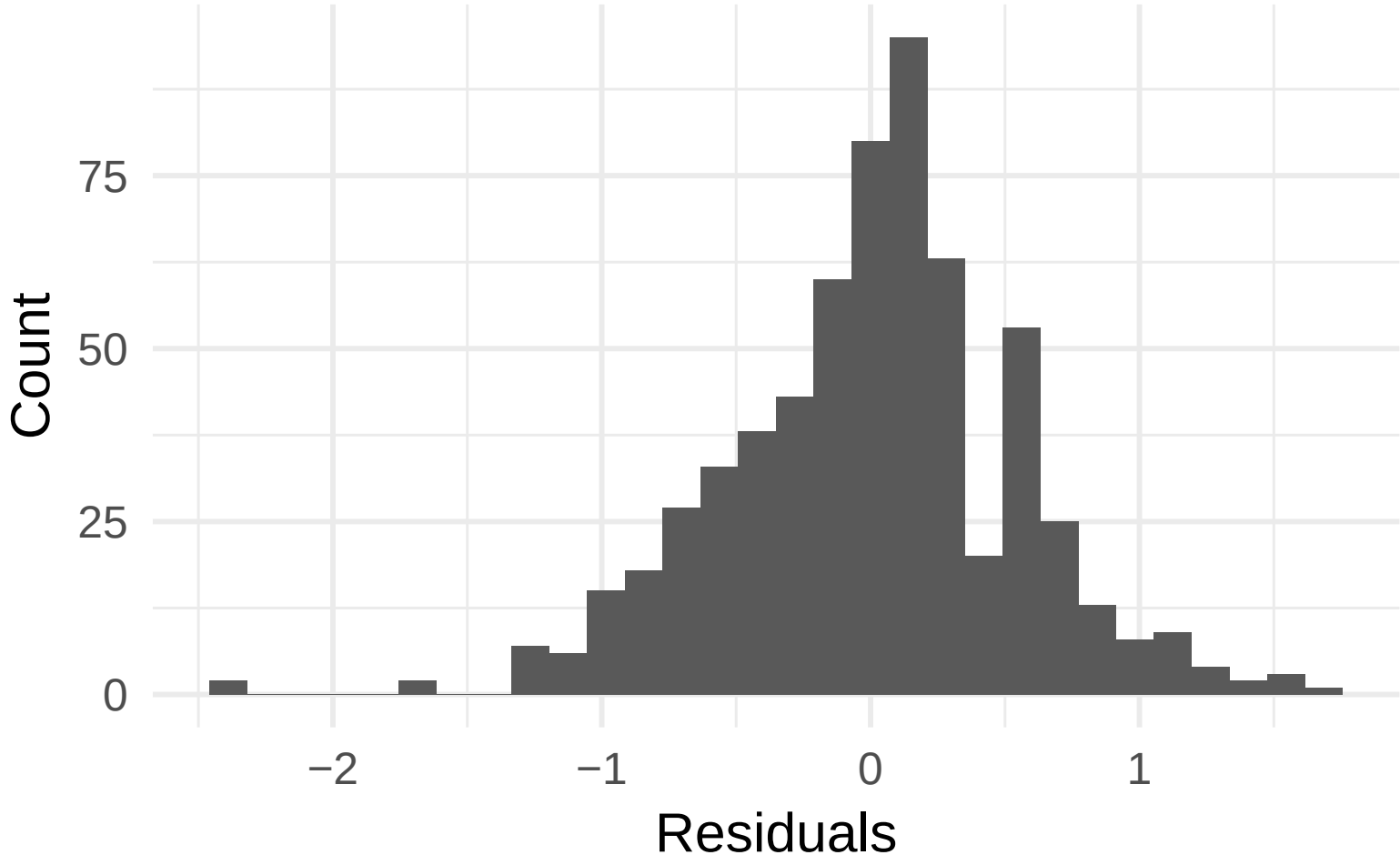
Histogram of residuals

Simulated data



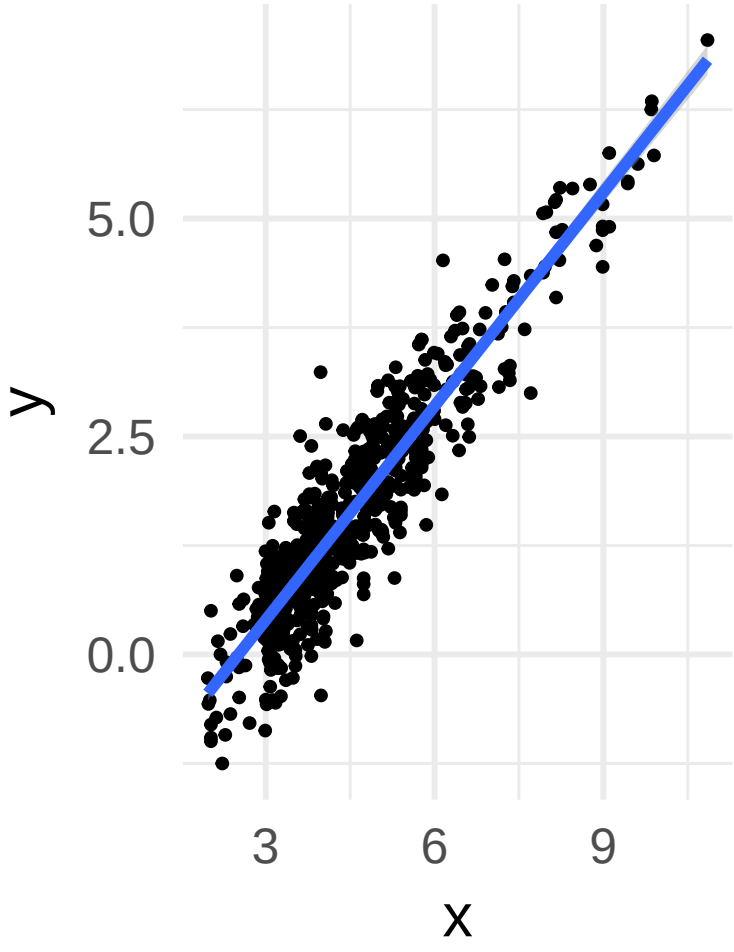
Histogram of residuals

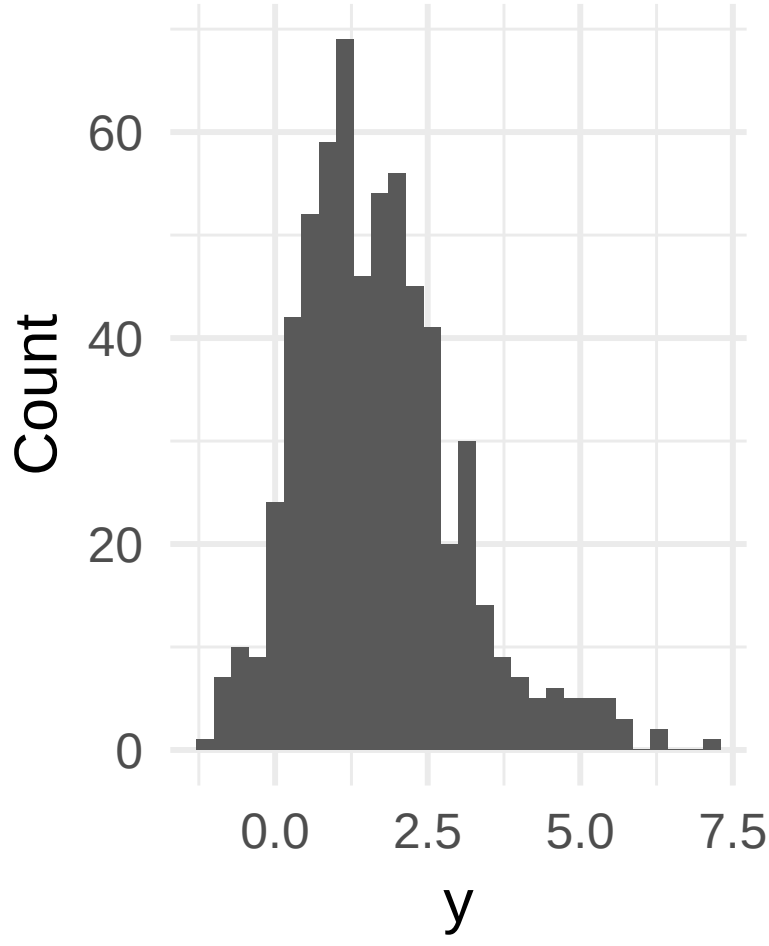
Rodent data



Common misconception

The normality assumption should be checked for y:





Crux

Normality assumption applies to:

Error

$y_i = \alpha + \beta x + \epsilon_i$

$\epsilon_i \sim \text{Normal}(0, \sigma)$

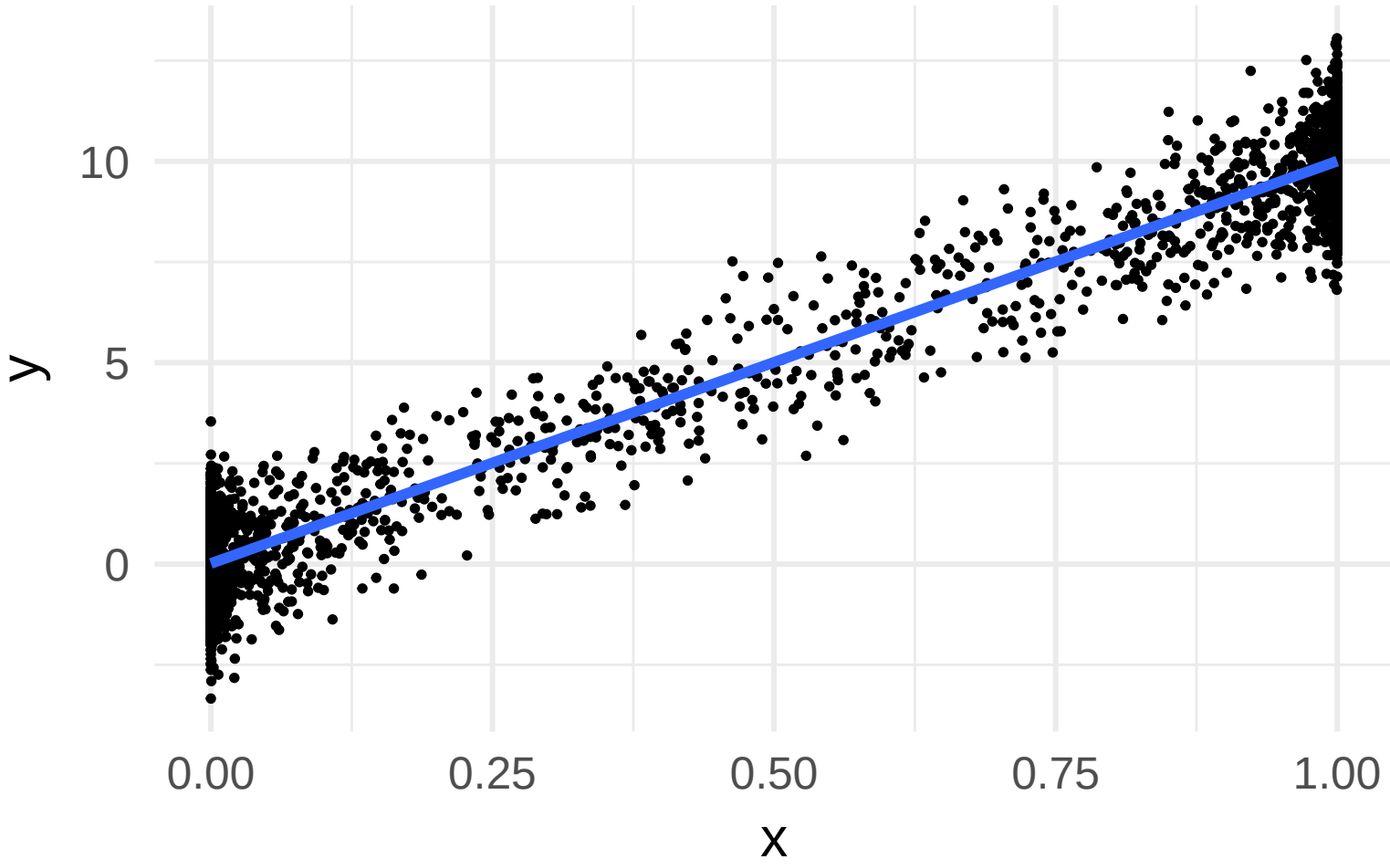
y after adjusting for the effect of x

$y_i \sim \text{Normal}(\alpha + \beta x, \sigma)$

but NOT y alone.

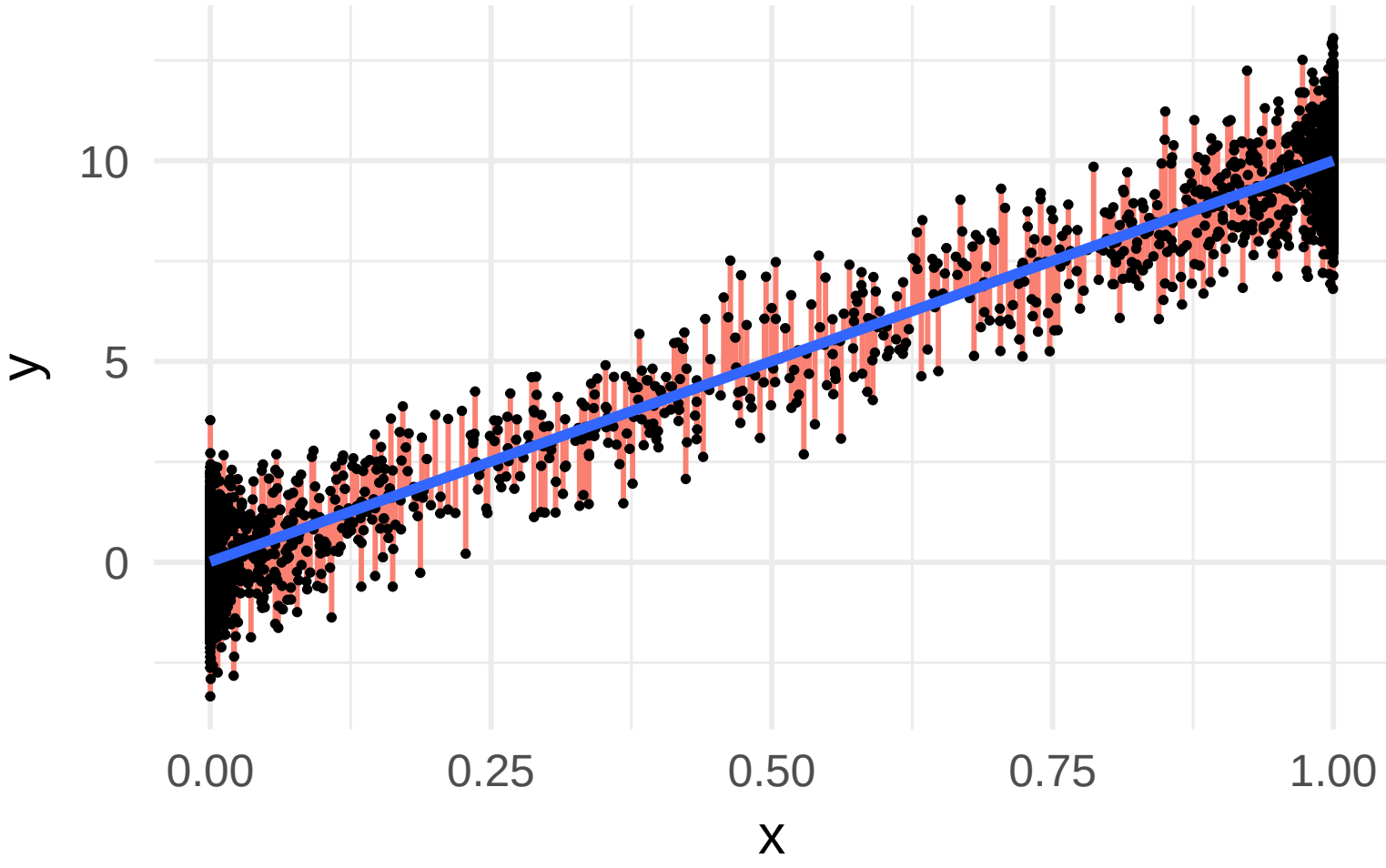
Simple example

$$y_i = 10x_i + \epsilon_i$$



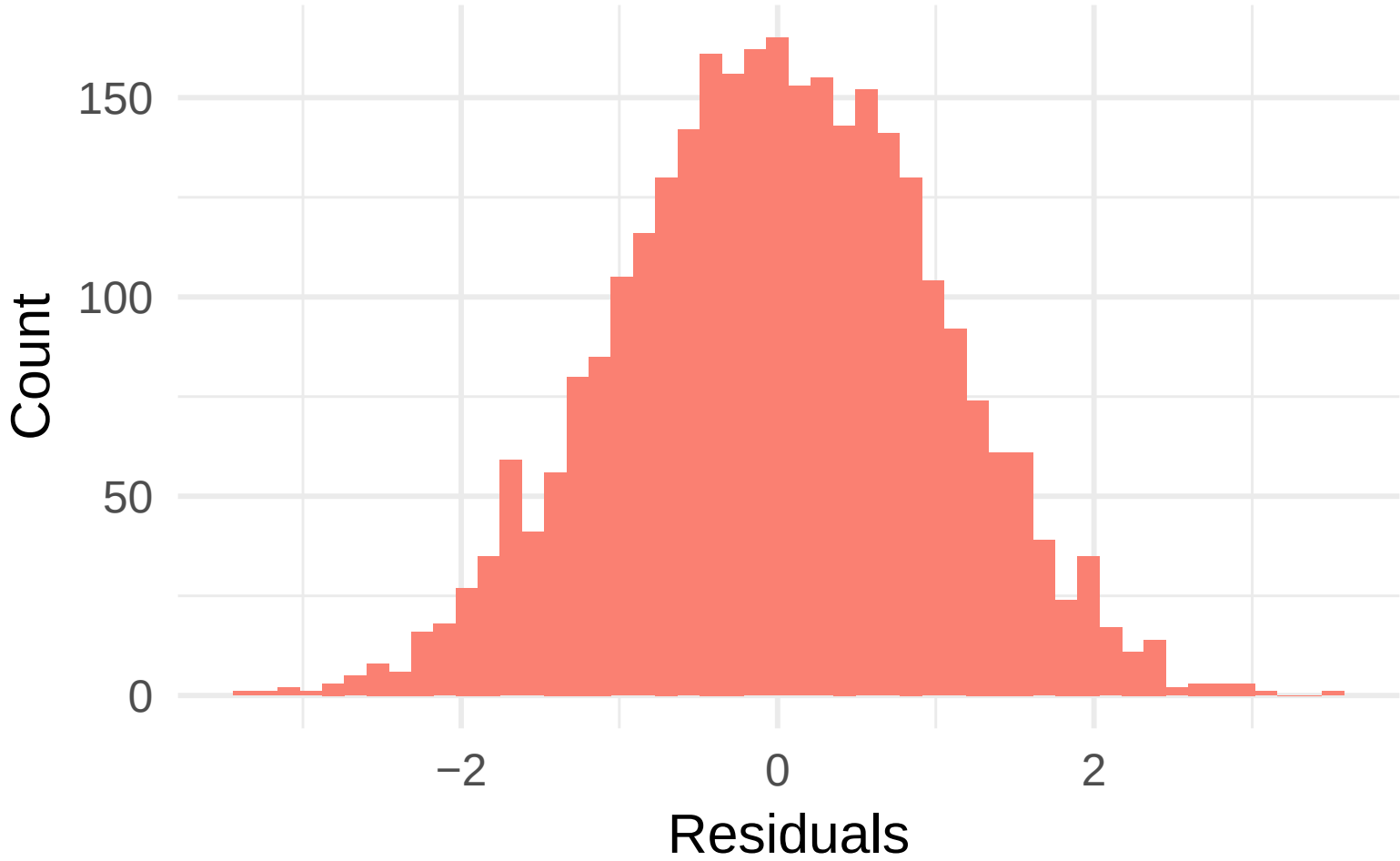
Plotting residuals

$$y_i = 10x_i + \epsilon_i$$



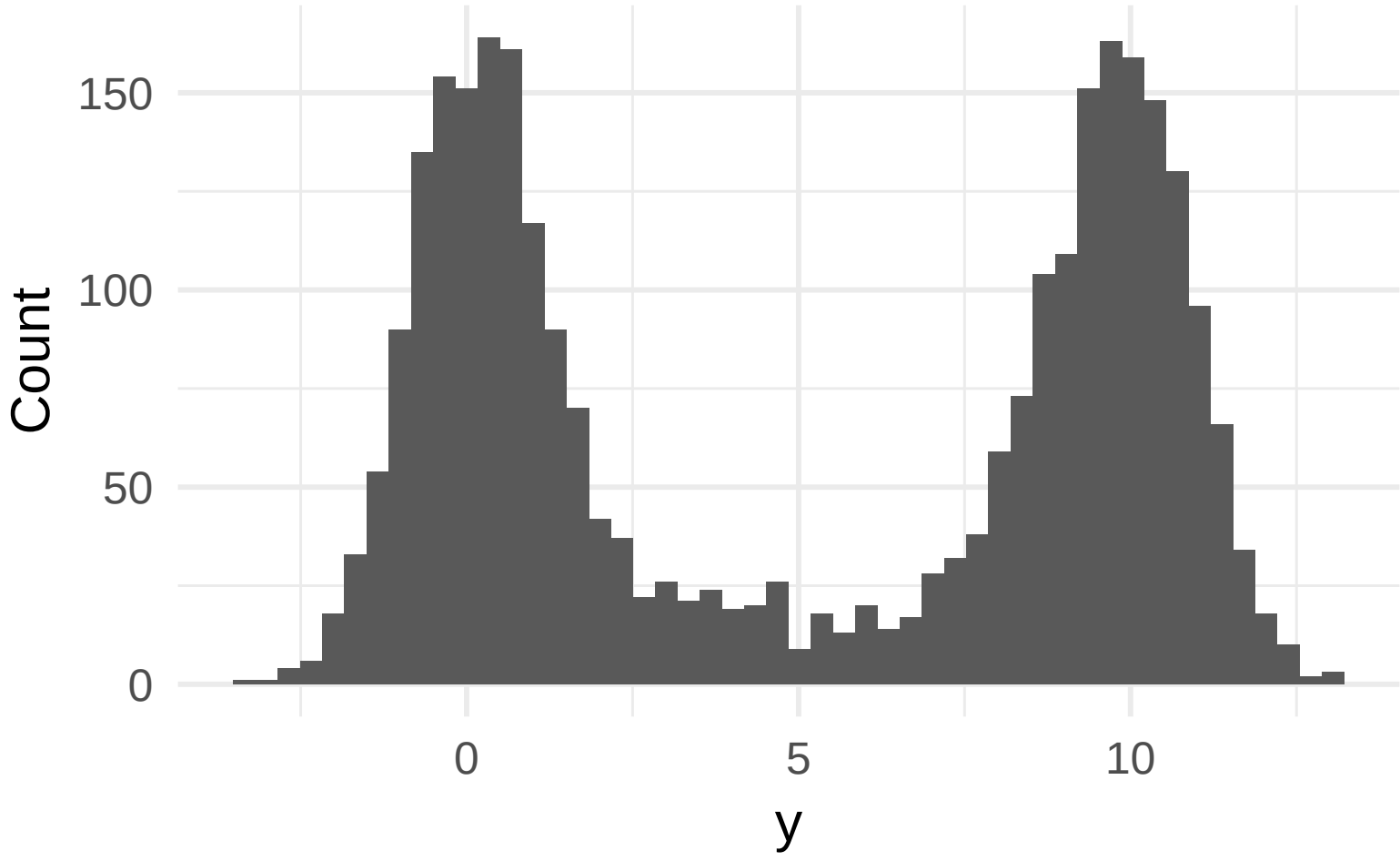
Distribution of residuals

$$y_i = 10x_i + \epsilon_i$$

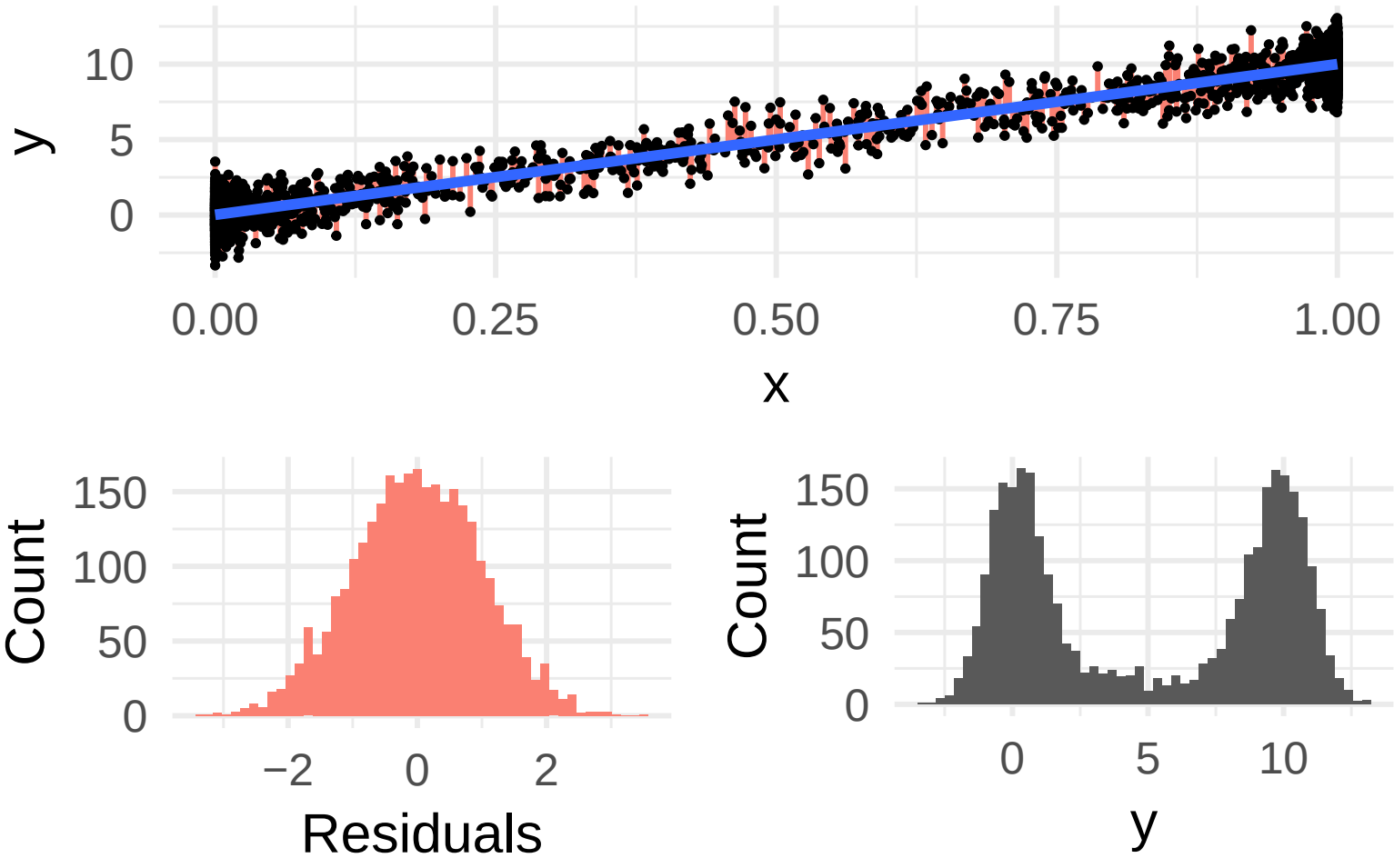


Distribution of y

$$y_i = 10x_i + \epsilon_i$$



All together now

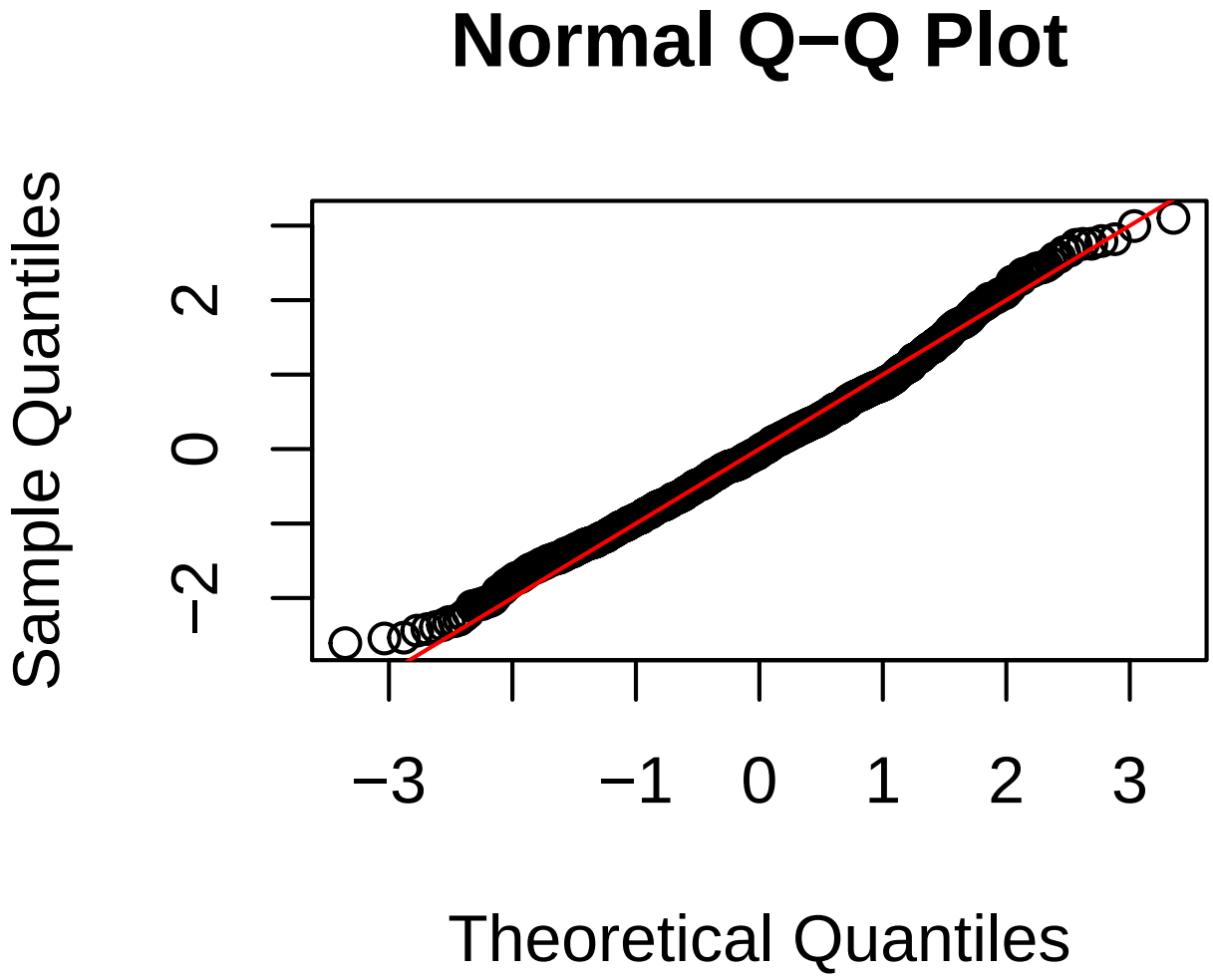


Quantile plots (Q-Q plots)

Another tool for evaluating residual normality

- compare quantiles of (Pearson) residuals to Normal quantiles

```
qqnorm(pearson_resid)
```



What to do when Normality fails

Generalized linear models (GLM)

Allow you to build models with other distributions

- counts
- proportions
- zero-inflation
- long tails
- and more!

Zooming out

Assumptions

1. Linear relationship between x and y:

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

2. Normal error in y (variation around line):

$$y_i \sim \text{Normal}(\hat{y}_i, \sigma)$$

3. Homoskedasticity (errors have constant variance)
4. Observations are independent

If assumptions are violated

1. Adjust your model*
2. Try to coerce your data to match assumptions

Evaluating assumptions for the rodent data

Demo: `3-model-checking/checking-assumptions.R`