Lecture 5: How Good is Our Regression

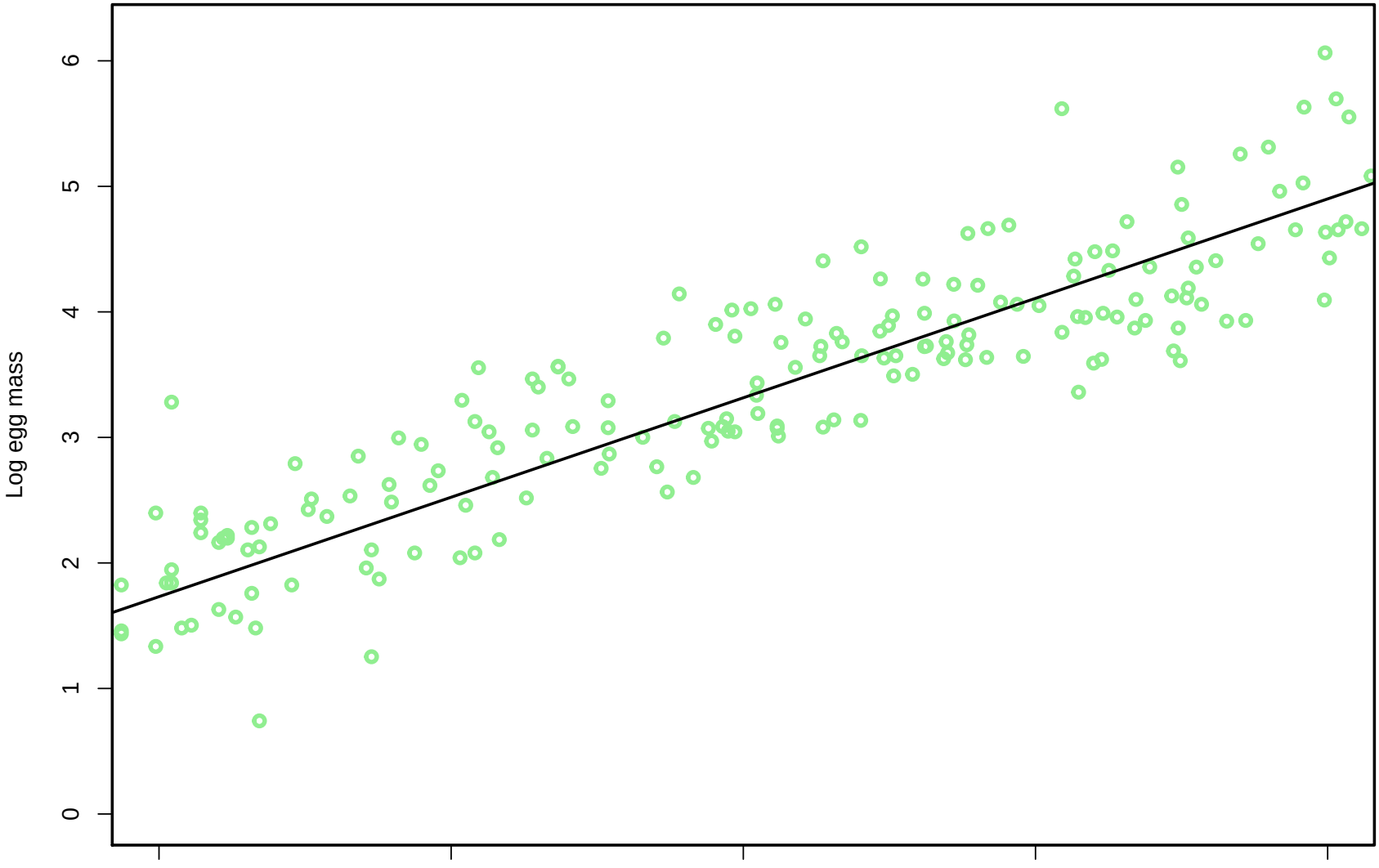
(Grafen & hails Chapter 9)

bob.ohara@ntnu.no

Regression: How good is the model?

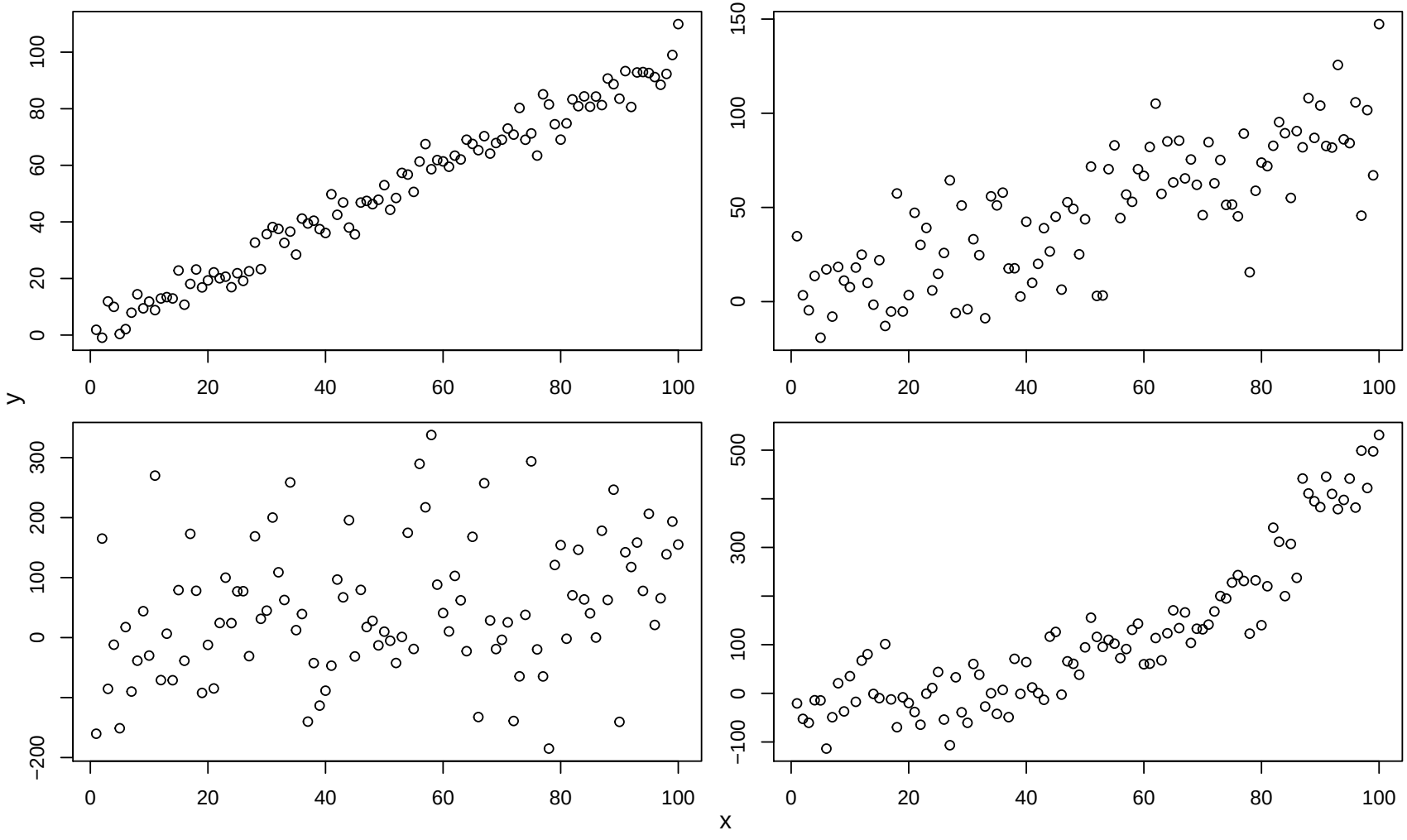
Last week we learned how to draw straight lines

$$y_i = \alpha + \beta x_i + \varepsilon_i$$



How well does it explain the data?

How could we summarise these?



How well does it explain the data? R^2

One way is to ask what proportion of the total variation is explained by the model

Informally

$$\frac{Var(\alpha + \beta X)}{Var(Y)} = 1 - \frac{Var(\varepsilon)}{Var(Y)}$$

More formally

$$Var(y) = \frac{1}{N-1}\sum(y_i - \bar{y})^2$$

$$Var(E(y)) = \frac{1}{N-1}\sum(\alpha + \beta x_i - \bar{y})^2 = \frac{1}{N-1}\beta^2\sum(x_i - \bar{x})^2$$

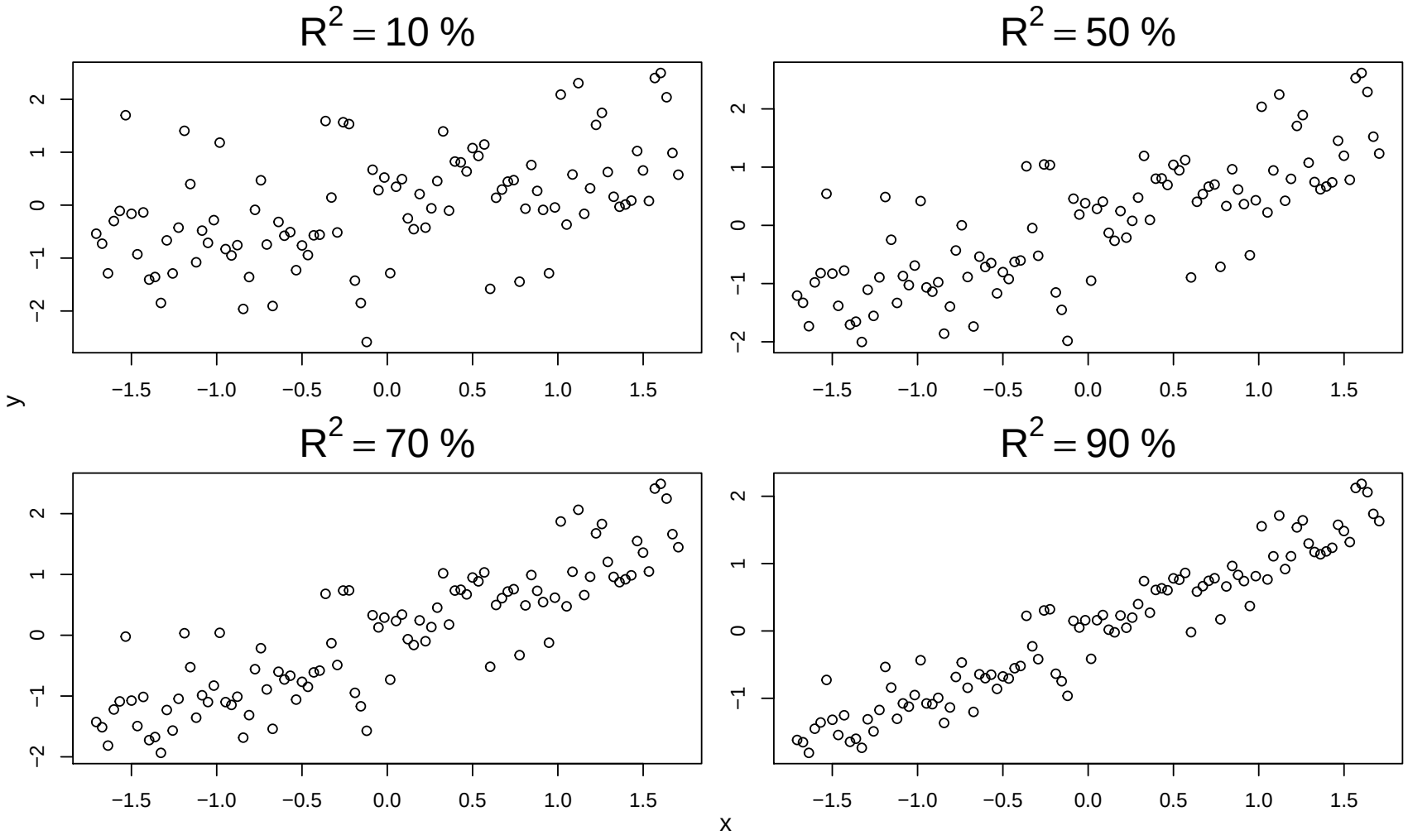
$$Var(\varepsilon) = \frac{1}{N-1}\sum(y_i - (\alpha + \beta x_i))^2$$

In practice, we calculate it as

$$R^2 = 1 - \frac{(\sum y_i - \mu_i)^2}{(\sum y_i - \bar{y})^2}$$

What is a good R^2?

It depends!



Model Checking

How could the model go wrong?

There are lots of ways the model go wrong

Discuss and list at least 3

- draw a plot of how it might look

The assumptions:

- linear in covariate
- error is normally distributed, with equal variance

Graphical Checks

The assumption is that the error term ε_i is unstructured error. So if we see structure, we might have a problem.

Easier to see what is going on with graphs: not just if there is a problem, but where it is

Residuals

An important statistic we use is the *residual*: the difference between the observed and expected values:

$$r_i = y_i - E(y_i) = y_i - (\hat{\alpha} + \hat{\beta} X_i)$$

If $\hat{\alpha} = \alpha$ and $\hat{\beta} = \beta$, then $r_i = \varepsilon_i$

We also sometimes standardise them:

$$t_i = \frac{y_i - E(y_i)}{\sqrt{var(r_i)}}$$

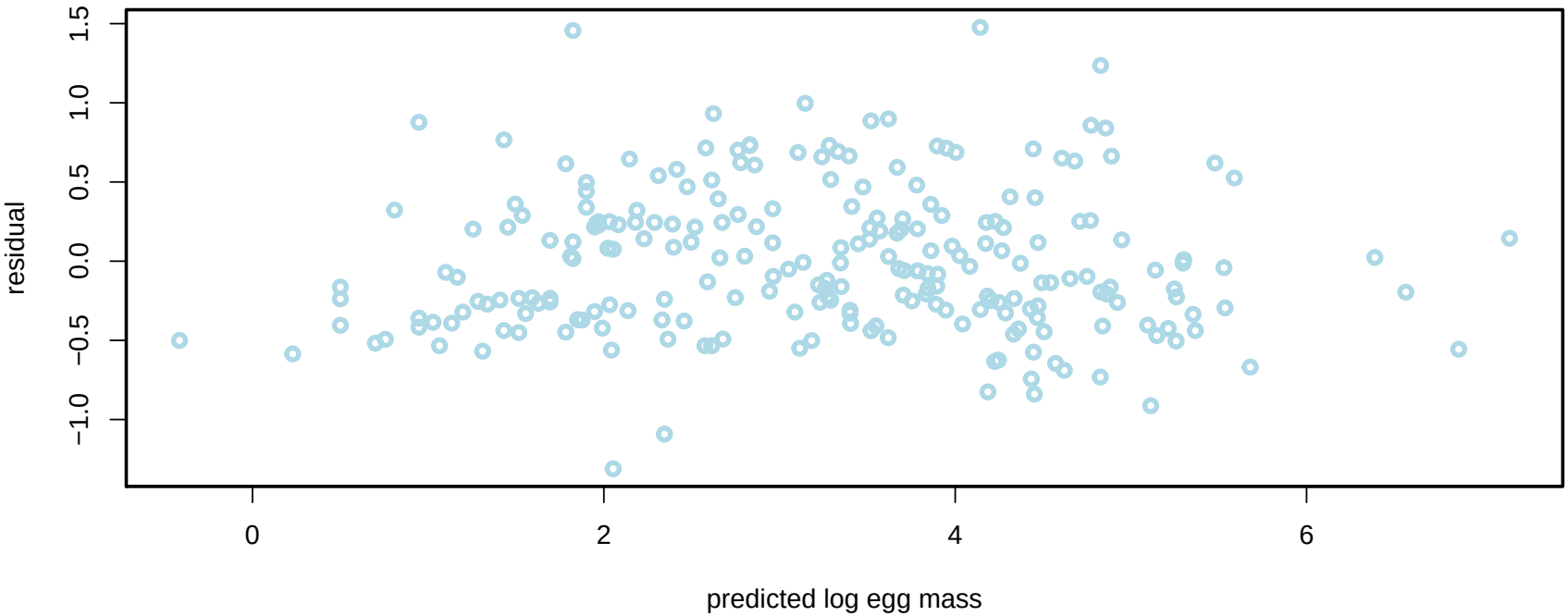
Why they are useful

The assumption is that the error term ε_i is unstructured error. So if we see structure, we might have a problem.

So we should check this

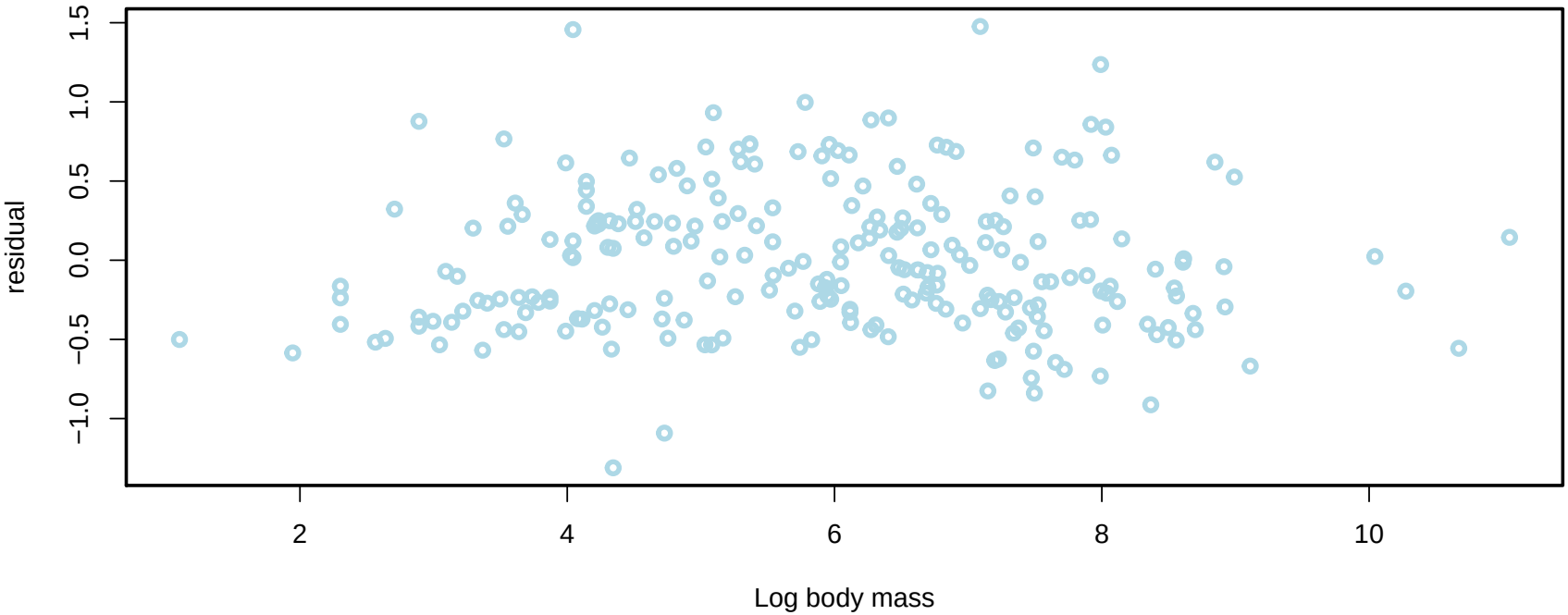
Plot against the predicted value

Plot r_i against the predicted value, $\hat{\mu}_i = \hat{\alpha} + \hat{\beta} x_i$



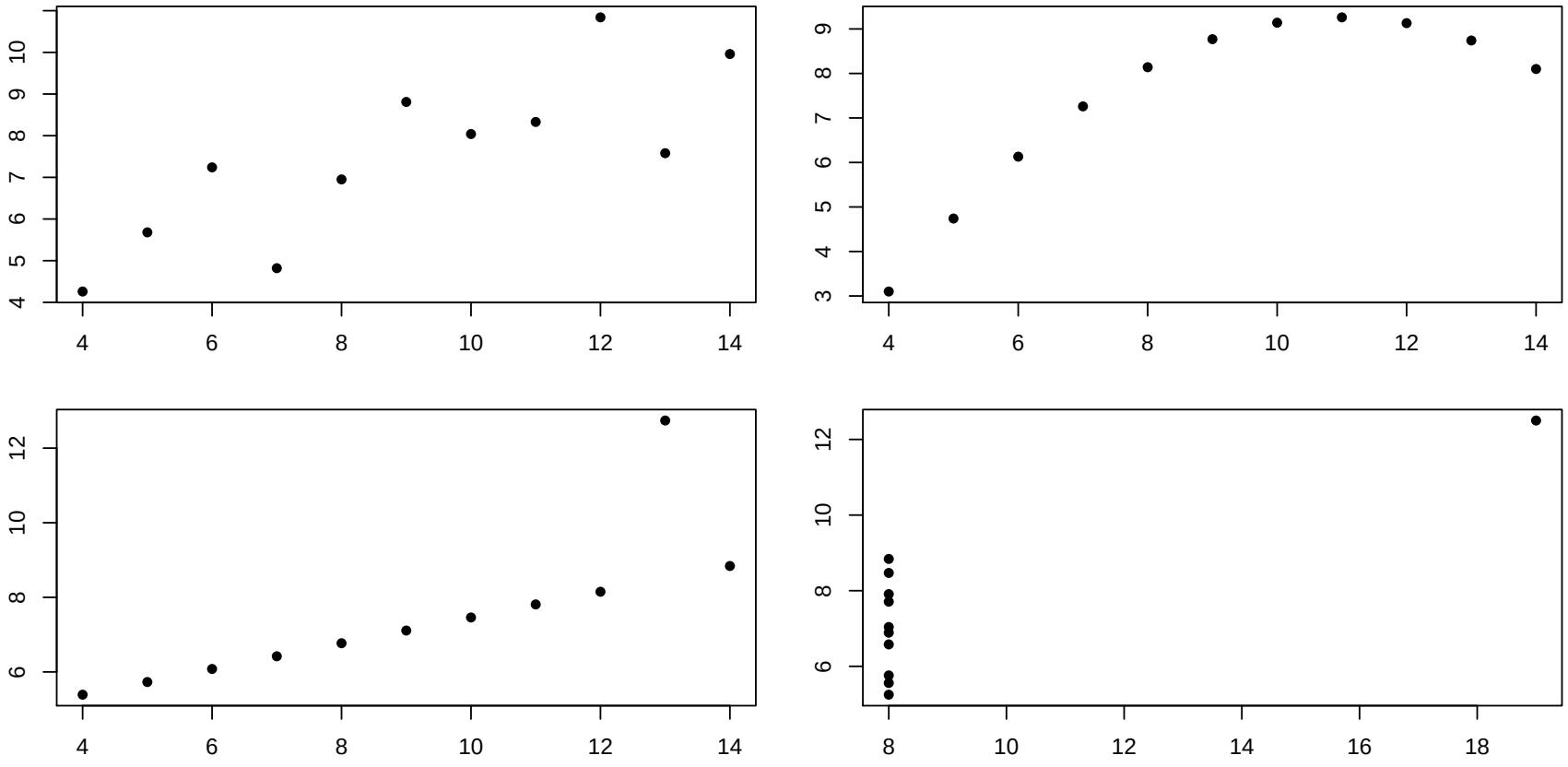
Plot against the covariate

Plot r_i against x_i

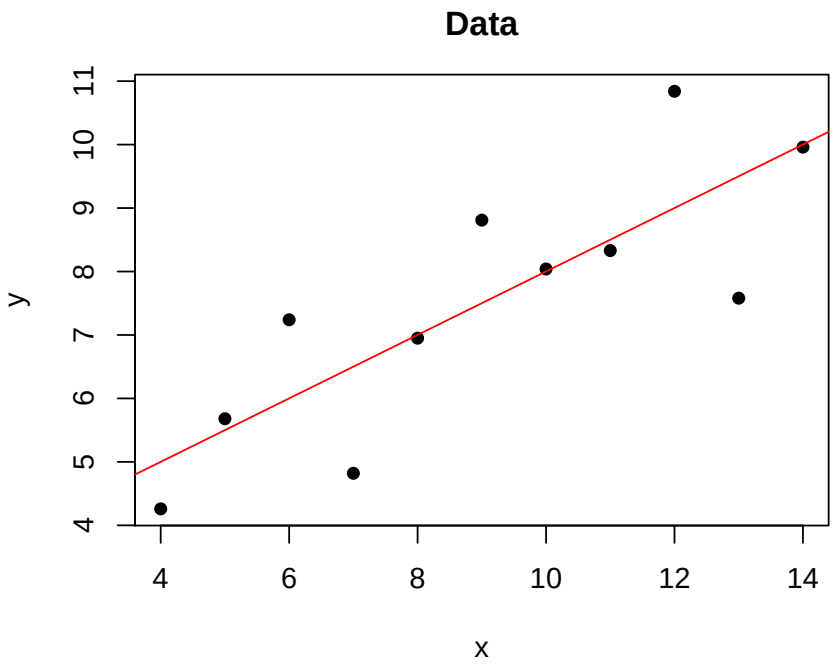


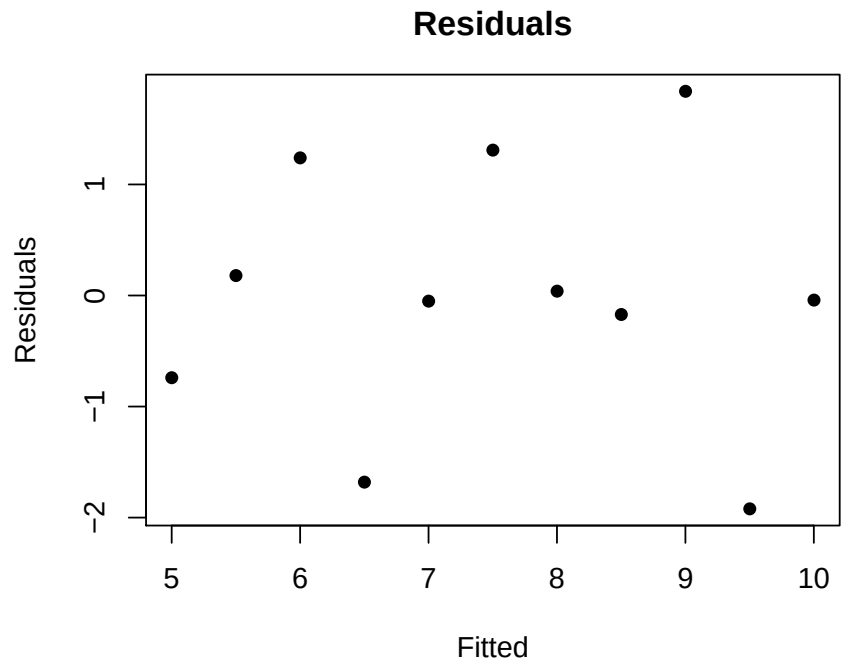
(same as last plot, but this will become useful with multiple regression)

Anscombe's Quartet: when things gow rong



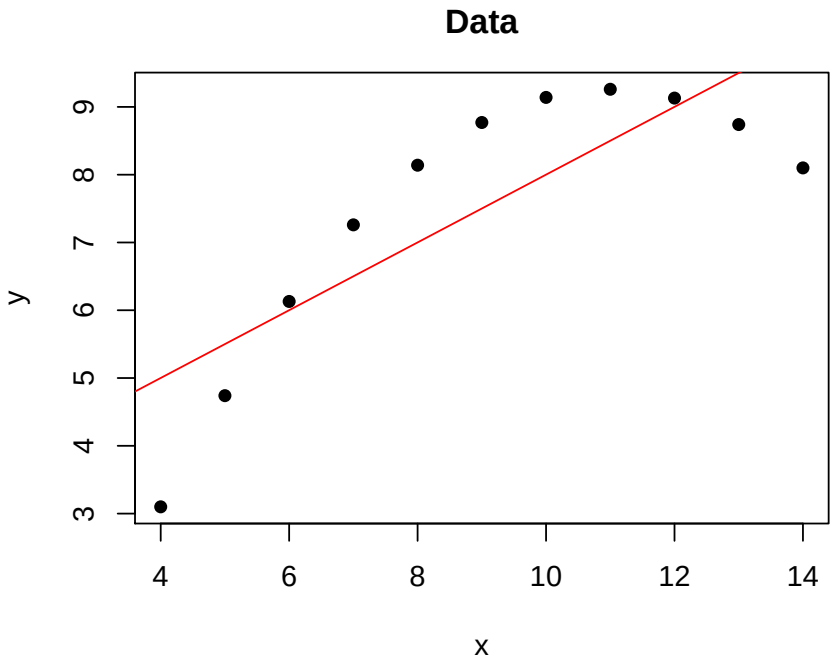
Anscombe's Residuals: 1

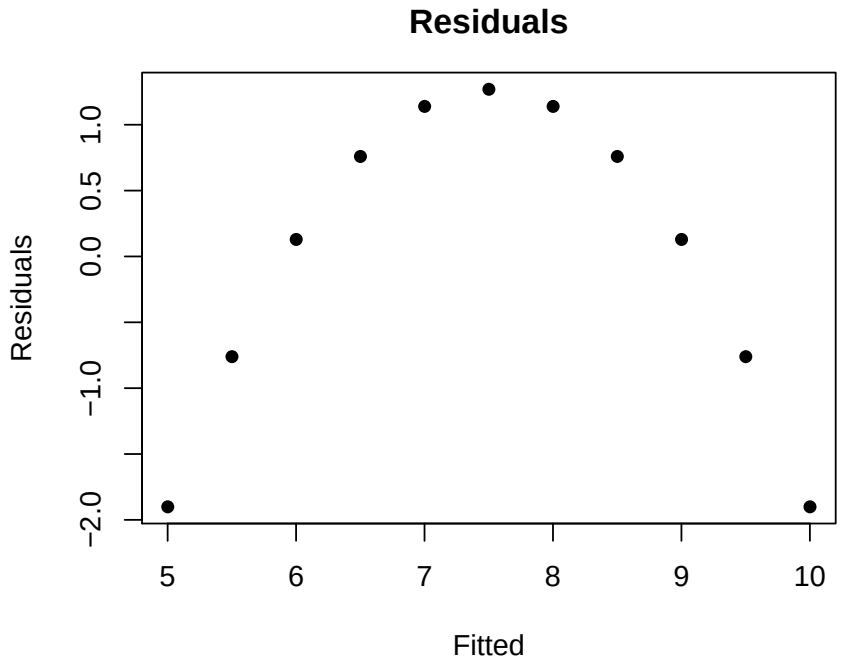




These look OK: no structure

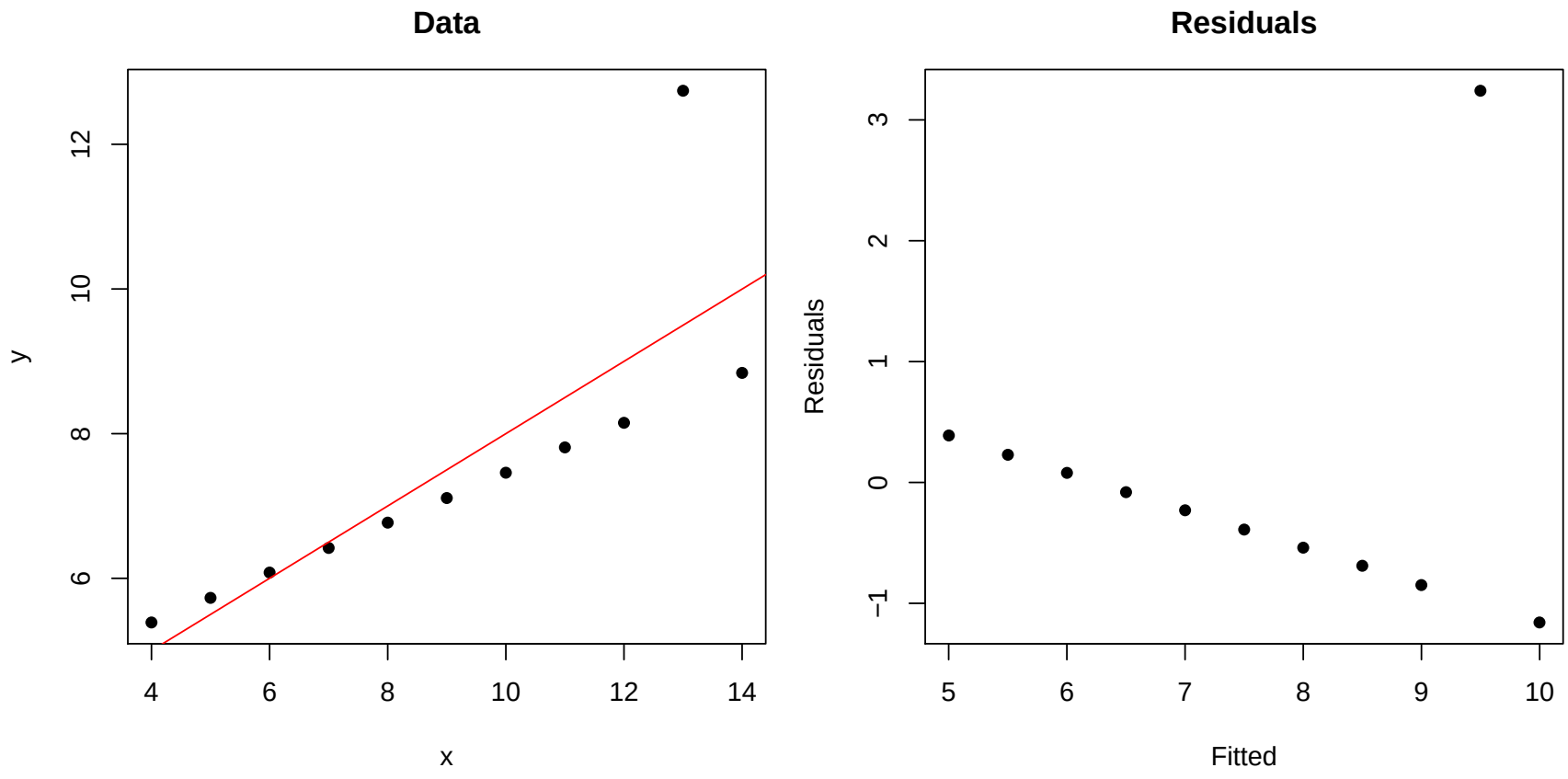
Anscombe's Residuals: 2





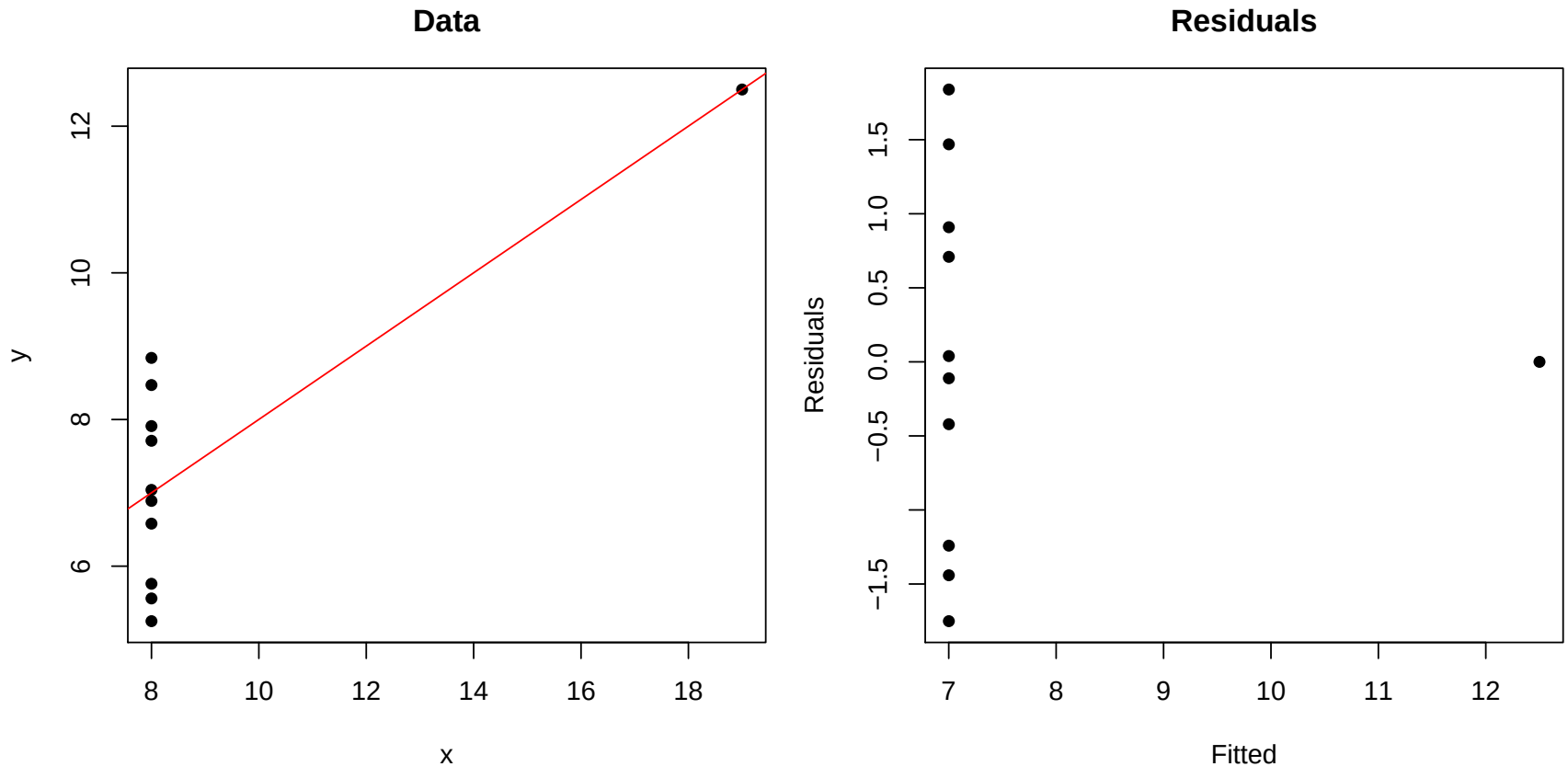
:-(

Anscombe's Residuals: 3



The outlier stands out, with a large residual

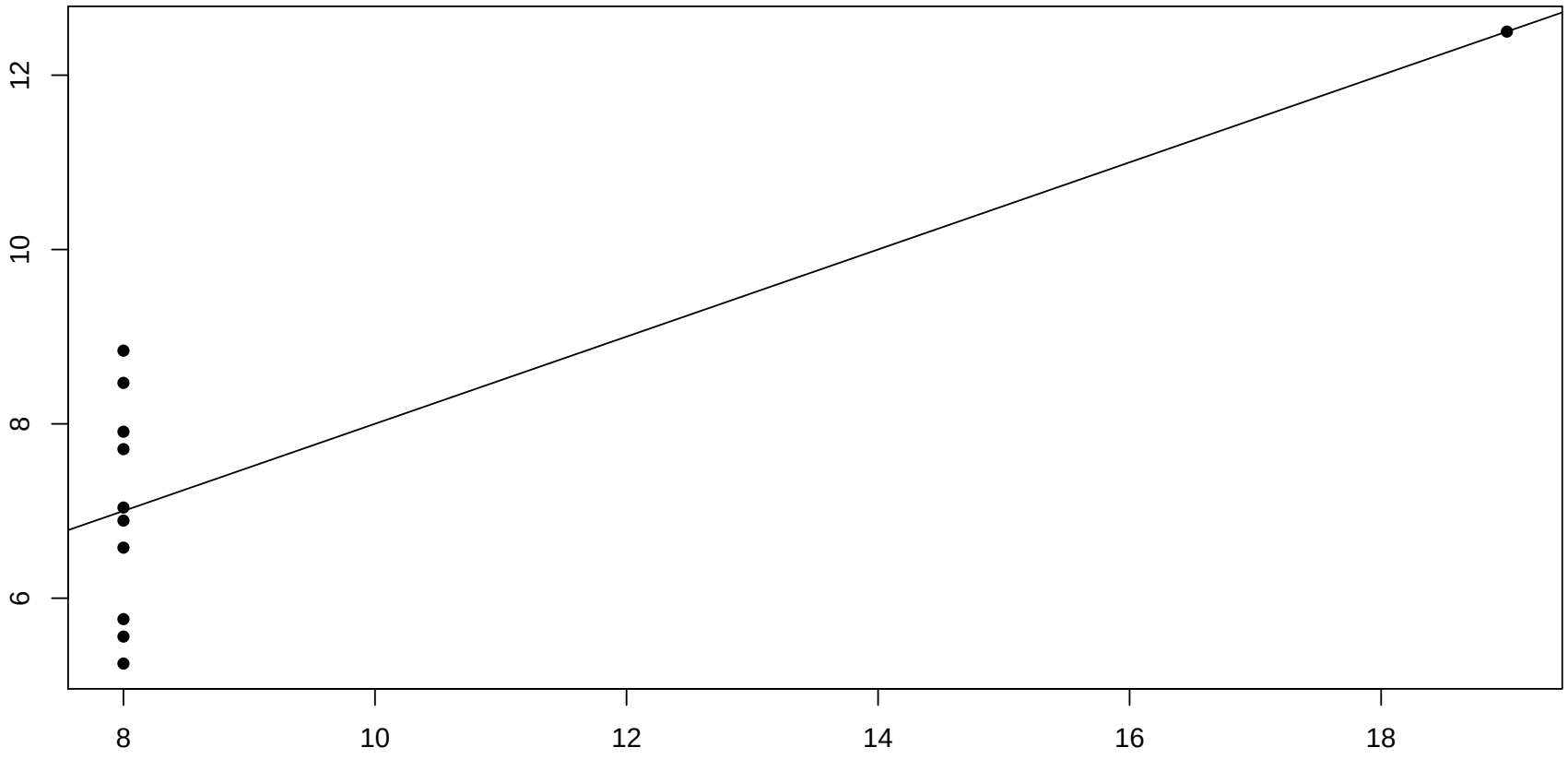
Anscombe's Residuals: 4



The outlier actually has a small residual (1.1102230246251 6e-16).

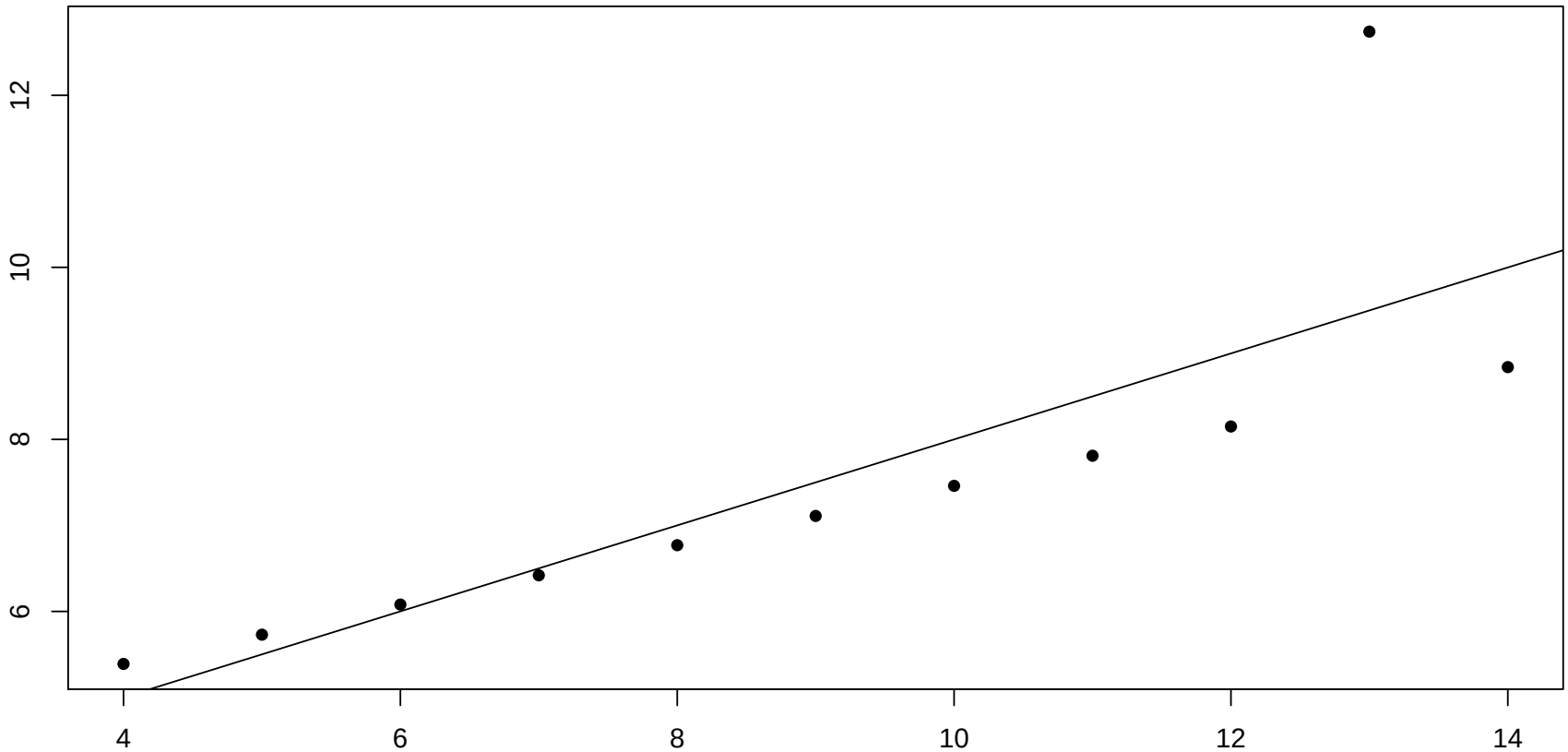
Influential Points

The fitted line goes through the point on the right



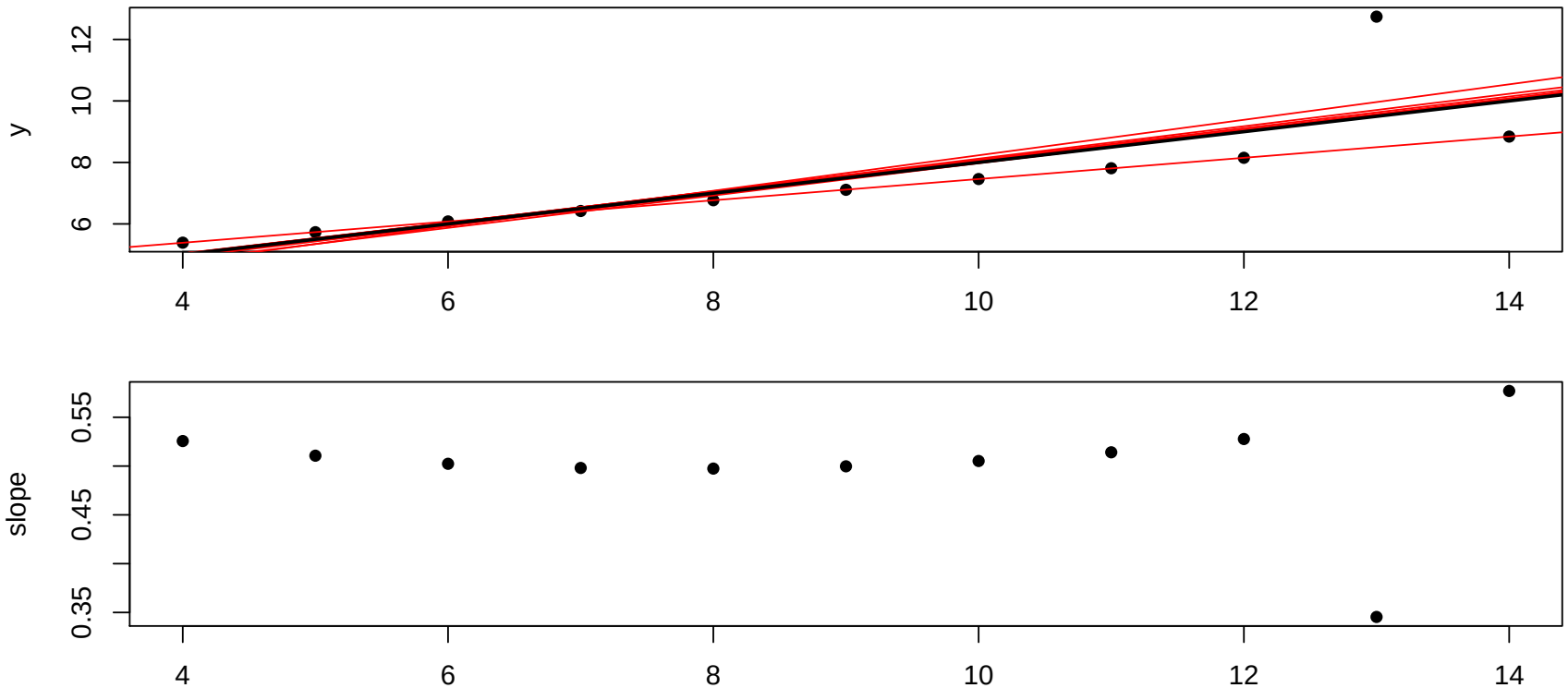
Influential Points

Here the outlier pulls the whole line up towards it



What is an influnetial point?

If we remove it, the model will change a lot



Influence and Leverage

We can generalise this idea by asking how much the fitted values for the other points change if we remove a data point

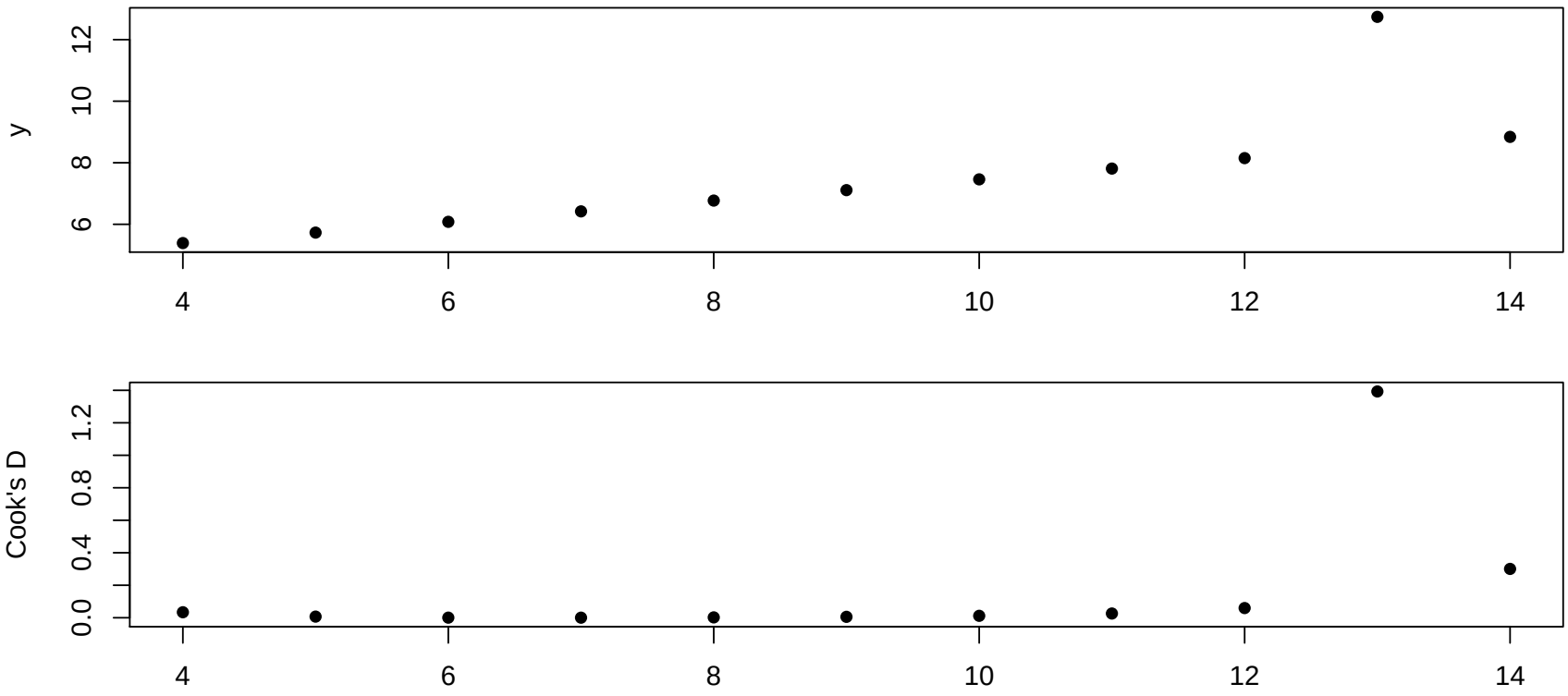
$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(-i)})^2}{s^2}$$

where $s^2 = 1/(n-1)\sum r_i^2$ is the *mean squared error*

What is influential?

Large values of D_i mean a large influence

- $D_i > 1$, or $4/n$?



Probability Plots

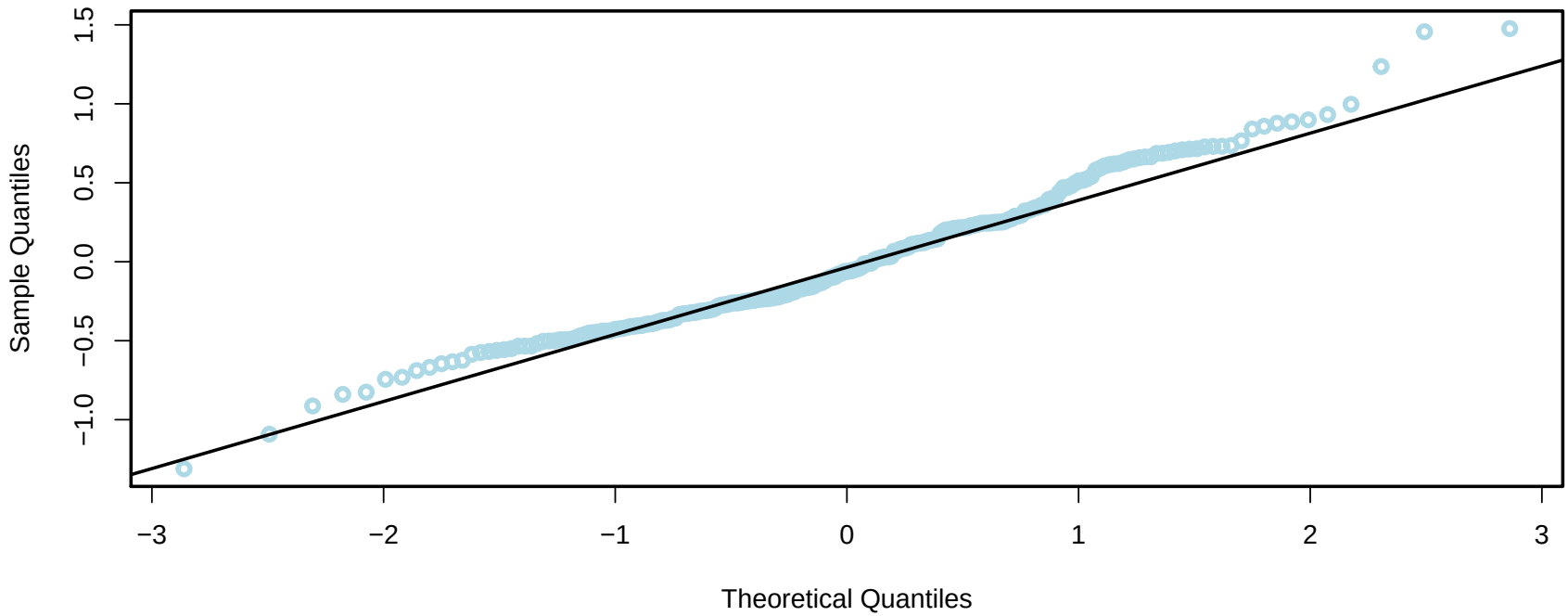
We are assuming that the residuals are normally distributed

We can check this with normal probability plots

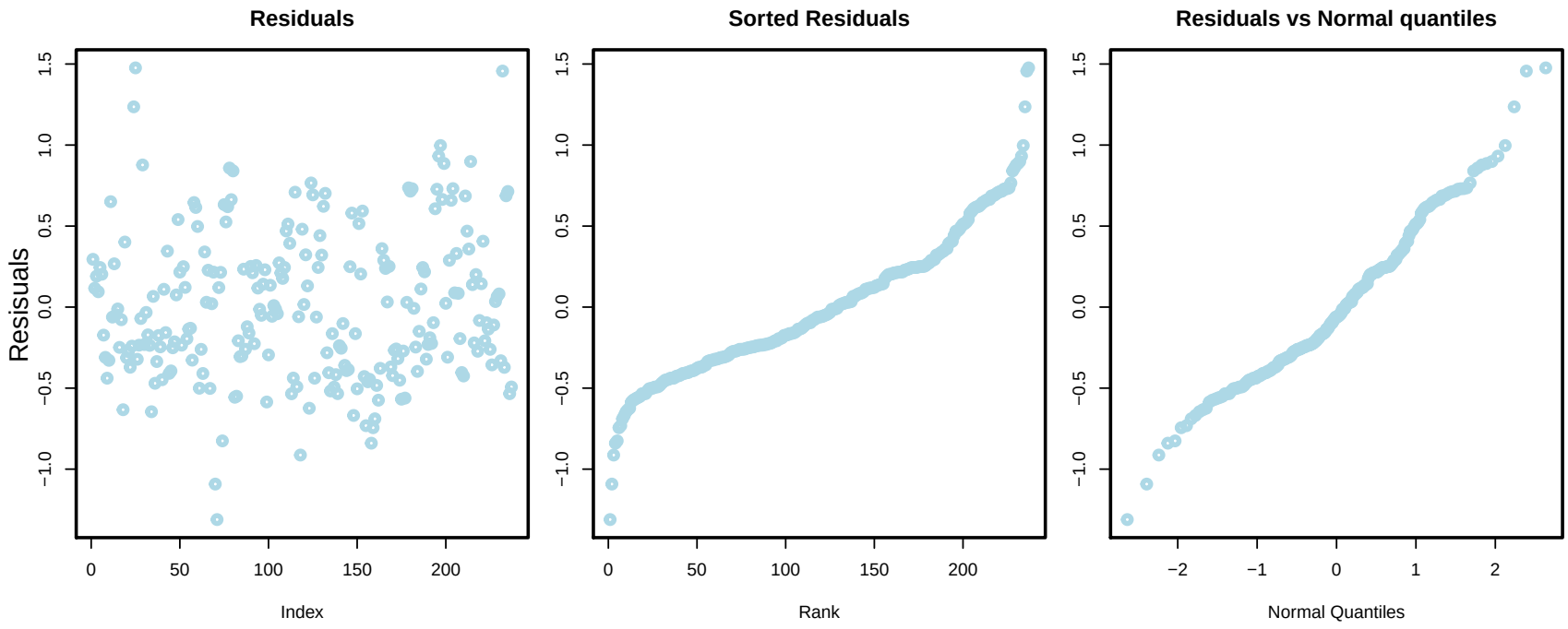
Plot the residual against its expected value (i.e. if the model is correct)

Probability Plots

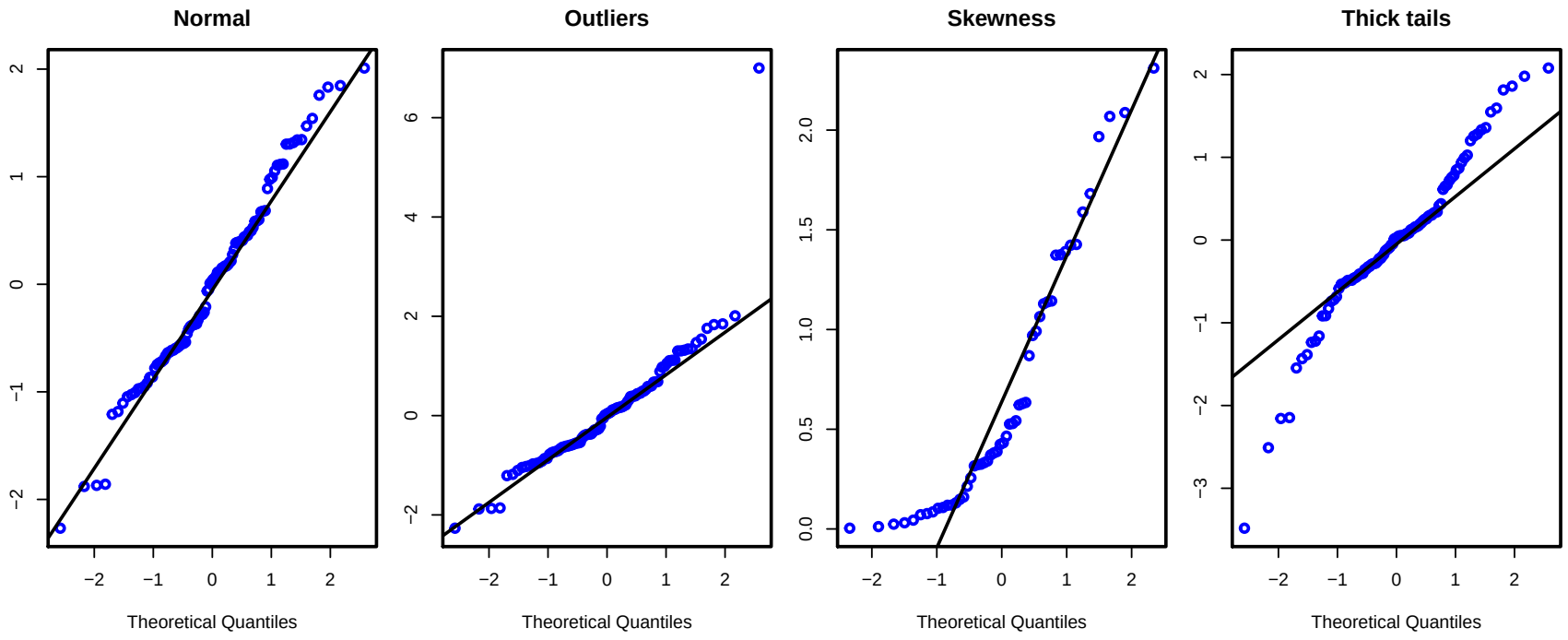
If we sort the data (smallest to largest), we can plot them against their expected values, i.e. plot r_i against the normal quantile



Constructing Probability Plots



What you can see



What to Do when your model is wrong

Break

Individual Data Points

Is your data point wrong?

- typos?
- real but unique

If it is wrong, correct, if it is right, might want to remove it & see if that makes a big difference

- if it does, be careful!

Assumptions

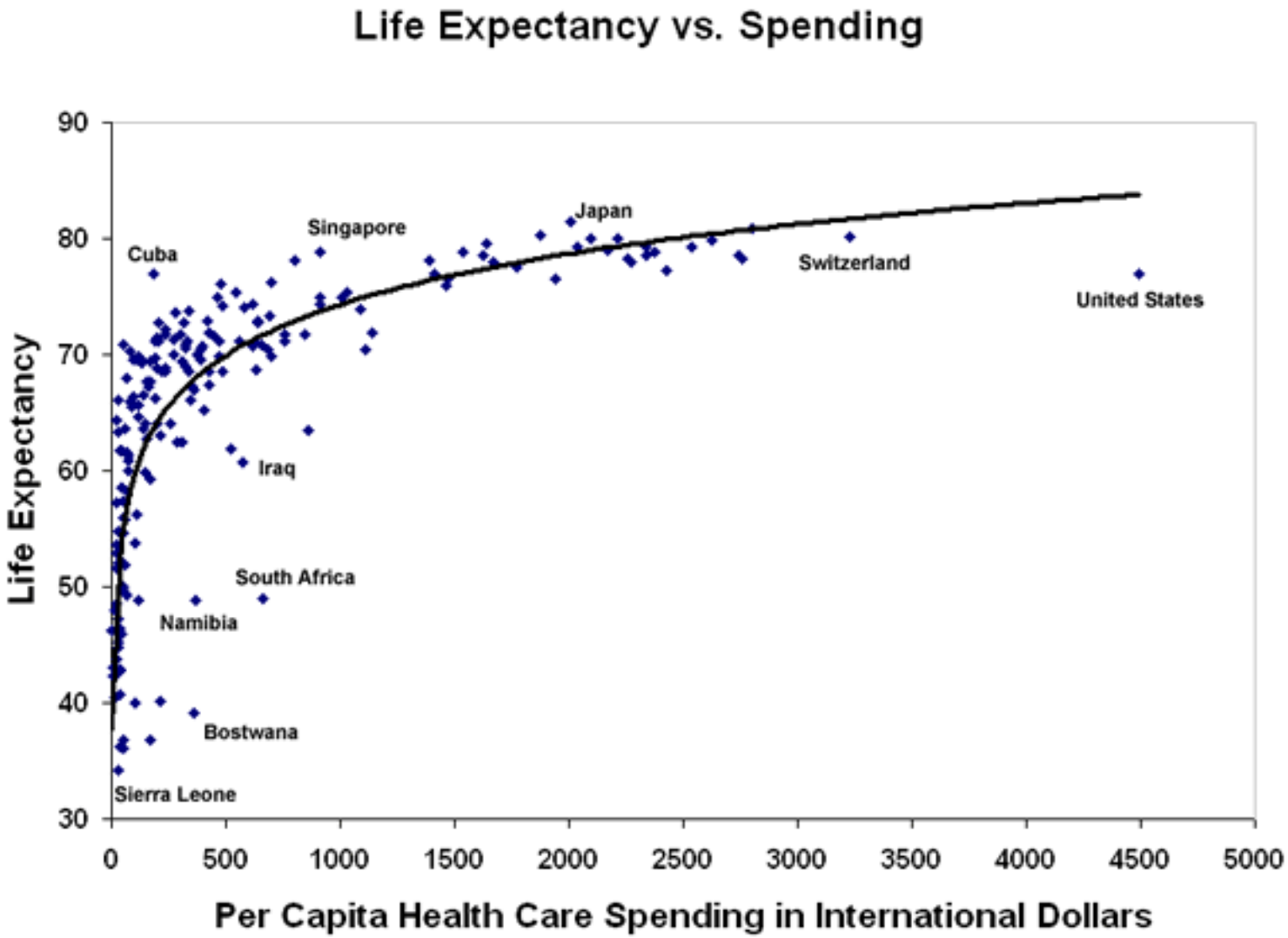


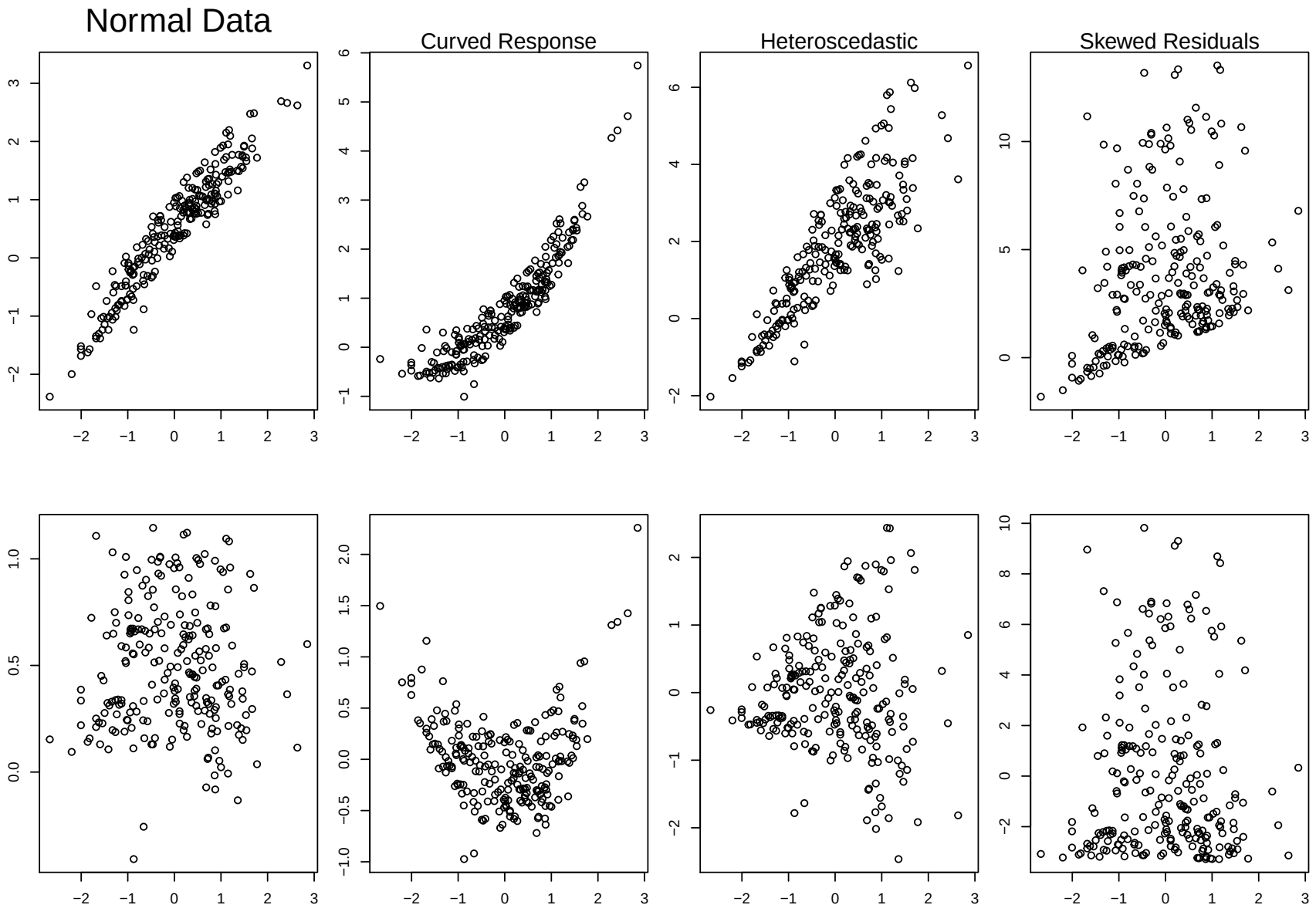
Figure 1:

Assumptions

The bigger picture

- curve
- skew
- heteroscedasticity

Problems


Normal Data
Curved Response
Heteroscedastic
Skewed Residuals

Possible Solutions

Transform the covariate

$$y_i = \alpha + \beta x_i^p + \varepsilon_i$$

e.g. $\sqrt{(}x_i)$, x_i^2, $\log(x_i)$,

Add more terms

- quadratic

$$y_i = \alpha + \beta x_i + \gamma x_i^2 + \varepsilon_i$$

More about this later

Transformations

Transform the response

e.g. $\sqrt{(}x_i)$, x_i^2, $\log(x_i)$

$$y_i^p = \alpha + \beta x_i + \varepsilon_i$$

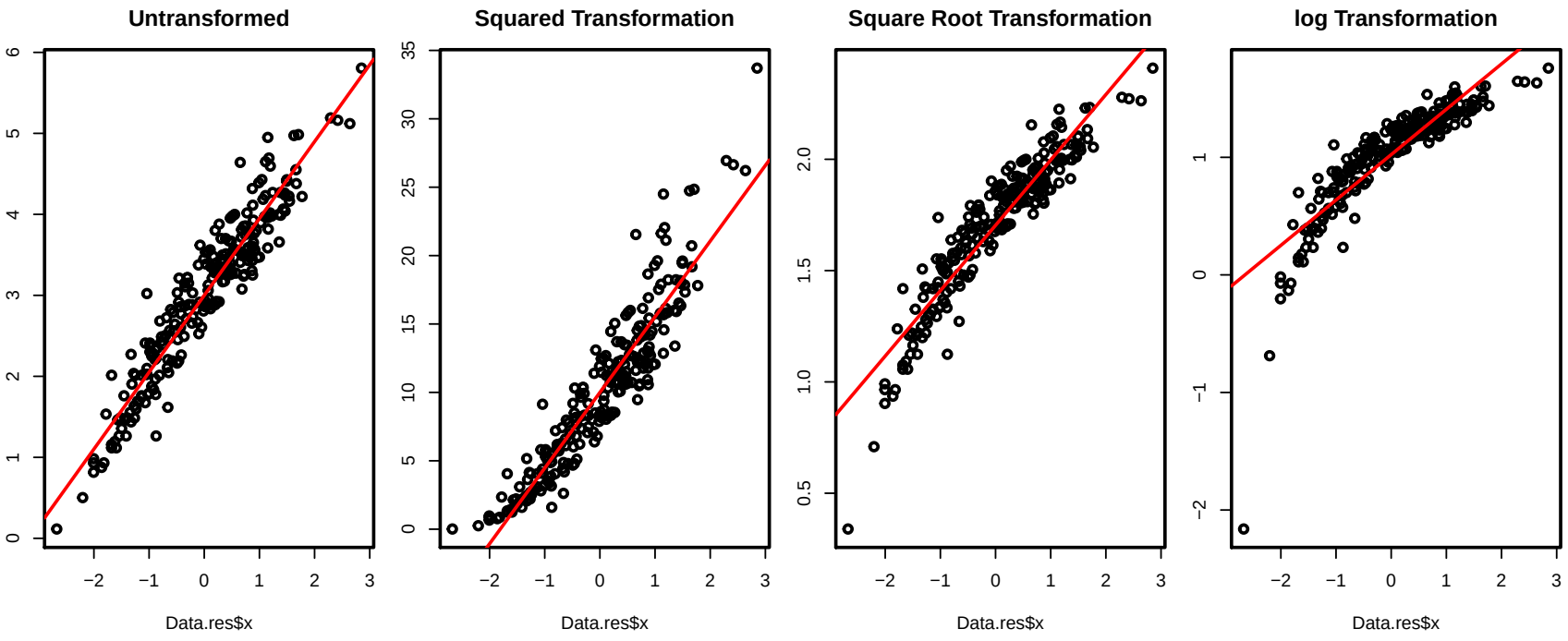
Box-Cox transformations

General Class of transformations

$$y_i \rightarrow y_i^p$$

if $p = 0$, use $log(y_i)$

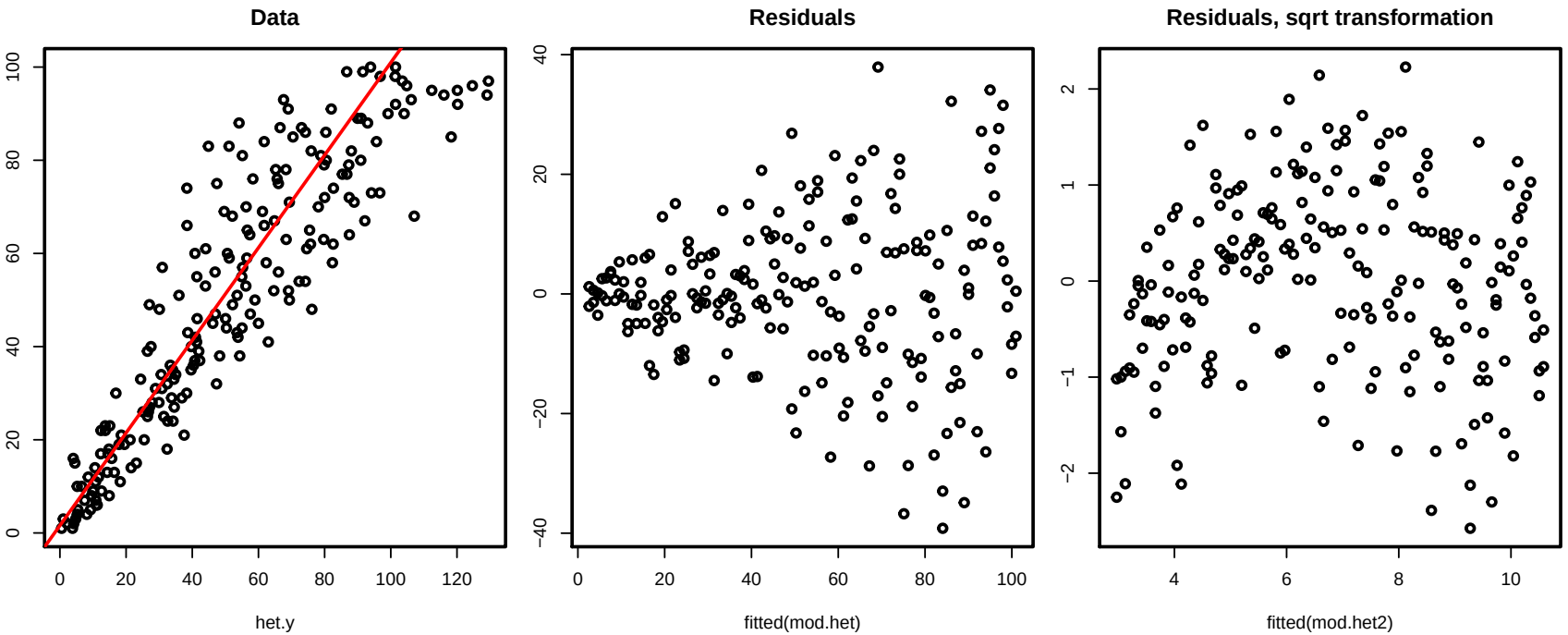
Using Box-Cox transformations



Heteroscedasticity

Variance changes with the mean

- Box-Cox can also solve this



Next Week

Multiple Regression

- more than 1 explanatory variable
- Grafen & Hails Chapter 4