How good is our straight line?

Bob O'Hara

5 February 2019

Last Week

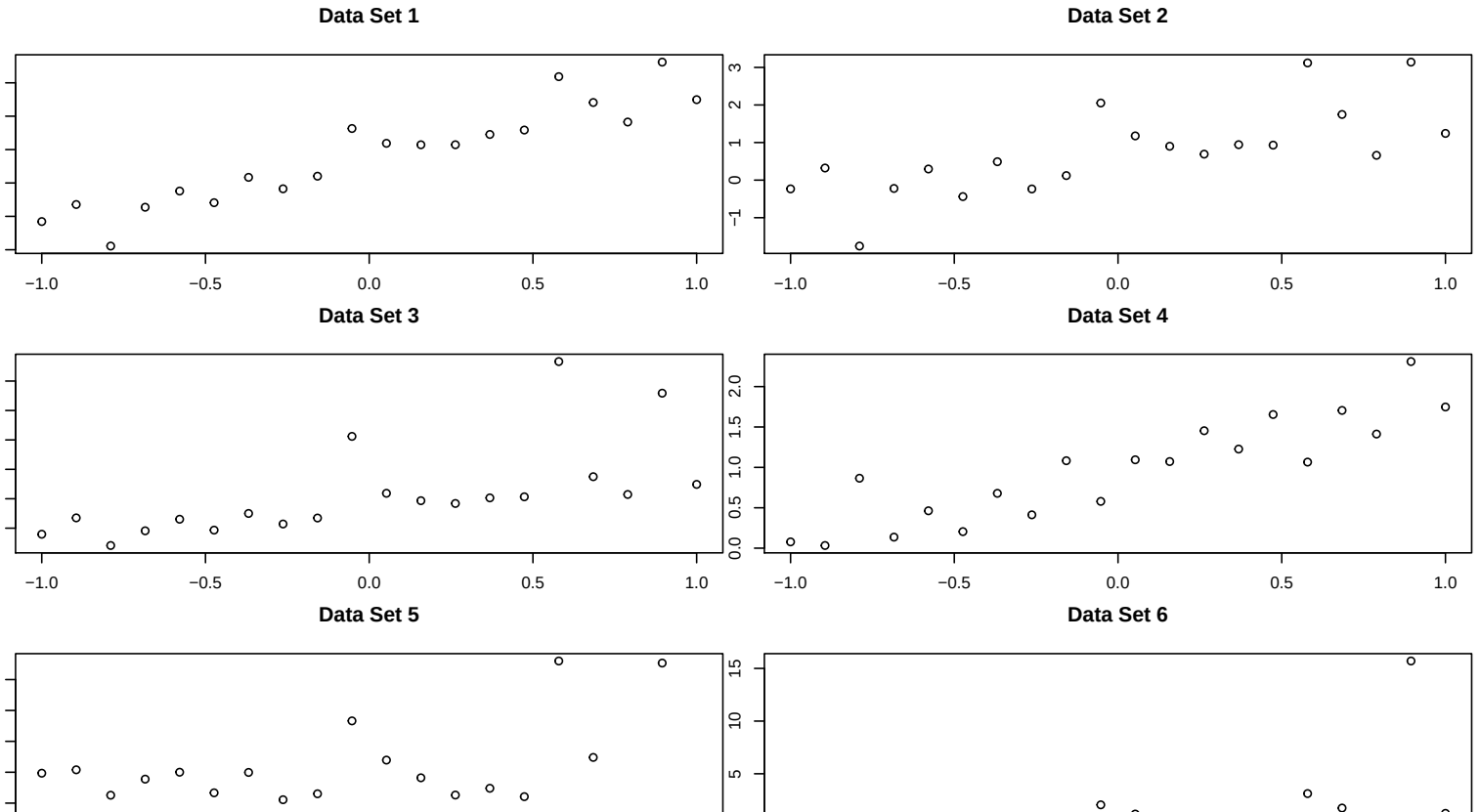
Last week we learned about regression: fitting straight lines

(show plot with residuals)

This Week: How good is my model?

Here are some simulated data sets. For all of them I used the the same errors, but manipulated the data in different ways. For each one, you should decide

- if you think a straight line would be a good fit to the data, and
- if it is not, can you do something simple to improve the fit? (for some you cannot, for some you can)



How good is my model? A Summary

- Model as fit + residuals
- R^2: How much variation does the model explain?
- Residual plots
  - curvature
  - outliers
  - heteroscedasticity
- Normal Probability Plots
- Influential Points

Another View of Regression

Model is systematic part + random part

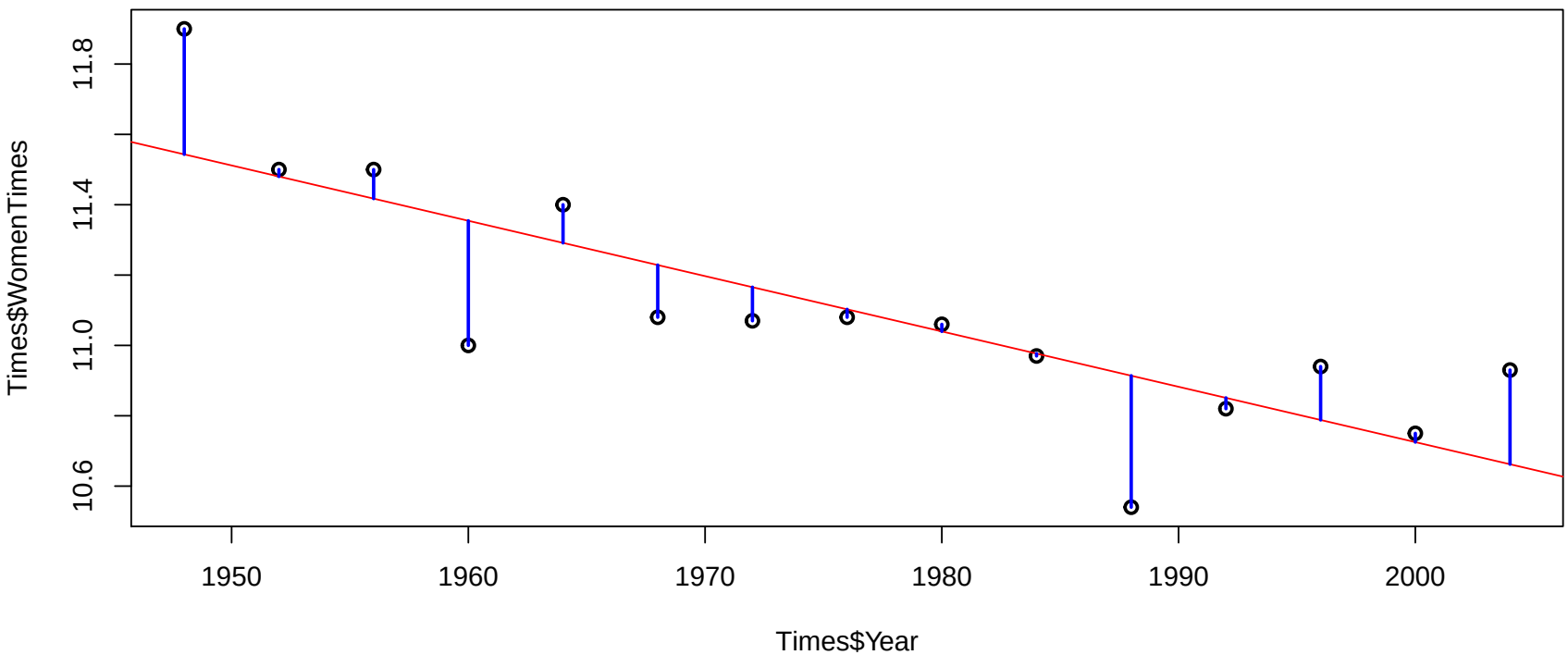
$$\begin{aligned} y_i &= \mu_i & +\varepsilon_i \\ &= \alpha + \beta x_i & +\varepsilon_i \end{aligned}$$

- Systematic part of model: a straight line
- Random part of model: residual error

All of the models we will see have this general form, but both parts can be more complicated

Women's times

```
Times <- read.csv("https://www.math.ntnu.no/emner/ST2304/2
WomenMod <- lm(WomenTimes~Year, data=Times)
plot(Times$Year, Times$WomenTimes, lwd=2)
abline(WomenMod, col=2)
segments(Times$Year, fitted(WomenMod), Times$Year, Times$Wo
```



How much variation does the model explain?

The total variation is

$$\begin{aligned} \text{Var}(y_i) &= \text{Var}(\alpha + \beta x_i) + \text{Var}(\varepsilon_i) \\ &= \beta^2 \text{Var}(x_i) + \sigma^2 \end{aligned}$$

- σ^2 is the residual variation

So we can ask how much of the total total variation is explained by the model

- if it only explains 4% then the model is not good

A poor model might be because it is wrong, or because the data come from a problem that is just too noisy

The Proportion of variance explained: R^2?

We can calculate the proportion of the total variation explained by the model

$$R^2 = \frac{\text{Variance Explained}}{\text{Total Variance}} = 1 - \frac{\text{Residual Variance}}{\text{Total Variance}}$$

After a bit of maths, we get

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

- $\sum(y_i - \mu_i)^2$ is the residual variance
  - squared difference from expected value
- $\sum(y_i - \bar{y})^2$ is the total variance
  - squared difference from grand mean

How do we calculate R^2 in R?

R calculates R^2 in a summary, so we can get it from this

```
R2 <- summary(WomenMod)$r.squared
R2
```

```
## [1] 0.6723703
```

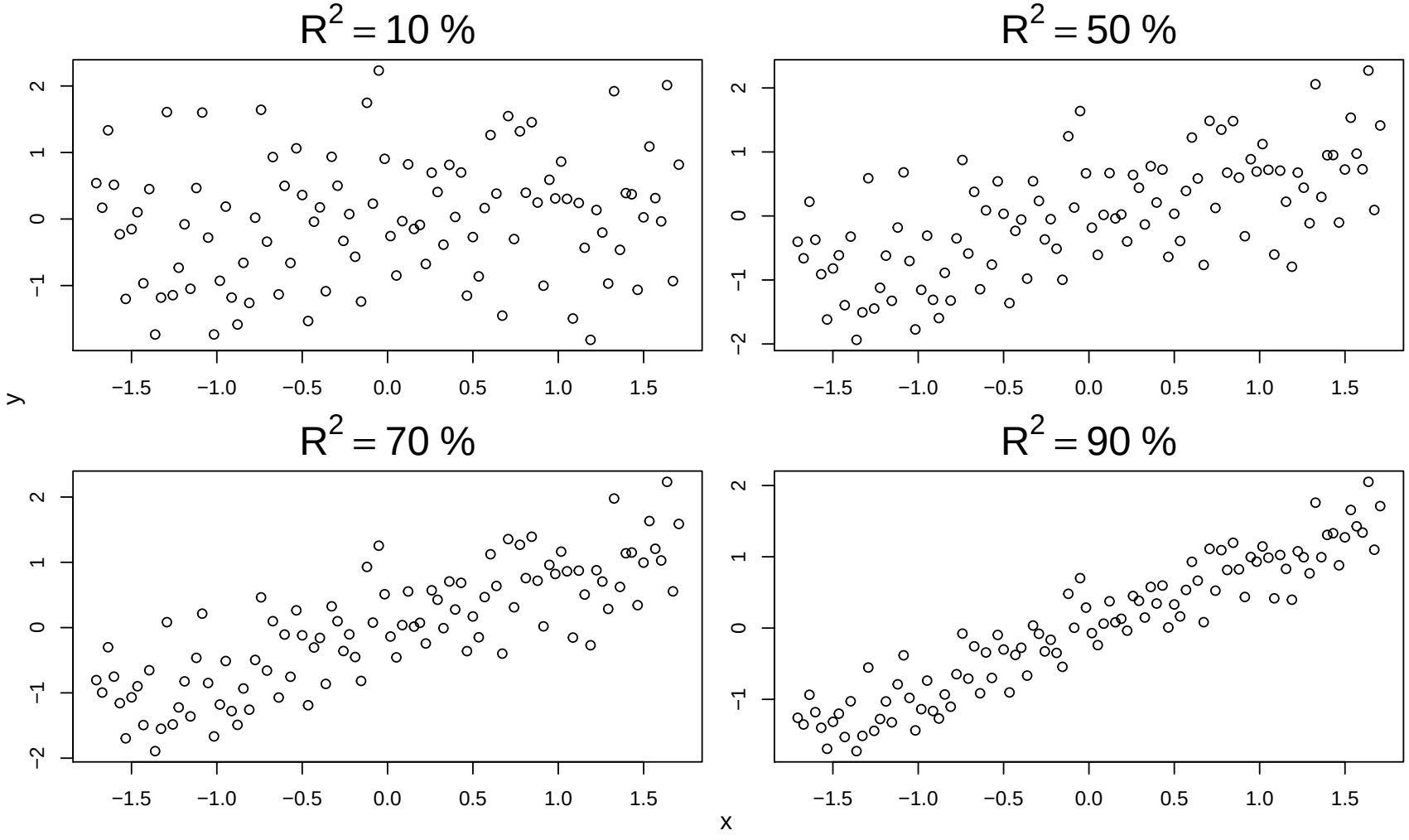
```
round(100*R2, 1)
```

```
## [1] 67.2
```

- we usually write R^2 as a percentage

What is a good R^2?

It depends!



Exercise

Exercise: calculate the R^2 for the 8 plots

You will need to read in the data, and fit the models.

`x` is the same for all `y`'s *except* `y7`

```
Data <- read.csv("https://www.math.ntnu.no/emner/ST2304/201
mod1 <- lm(y1 ~ x, data=Data)
mod7 <- lm(y7 ~ x7, data=Data)

summary(lm(y1 ~ x, data=Data))$r.squared
```

```
## [1] 0.8708701
```

Regression Assumptions

Model is systematic part + random part

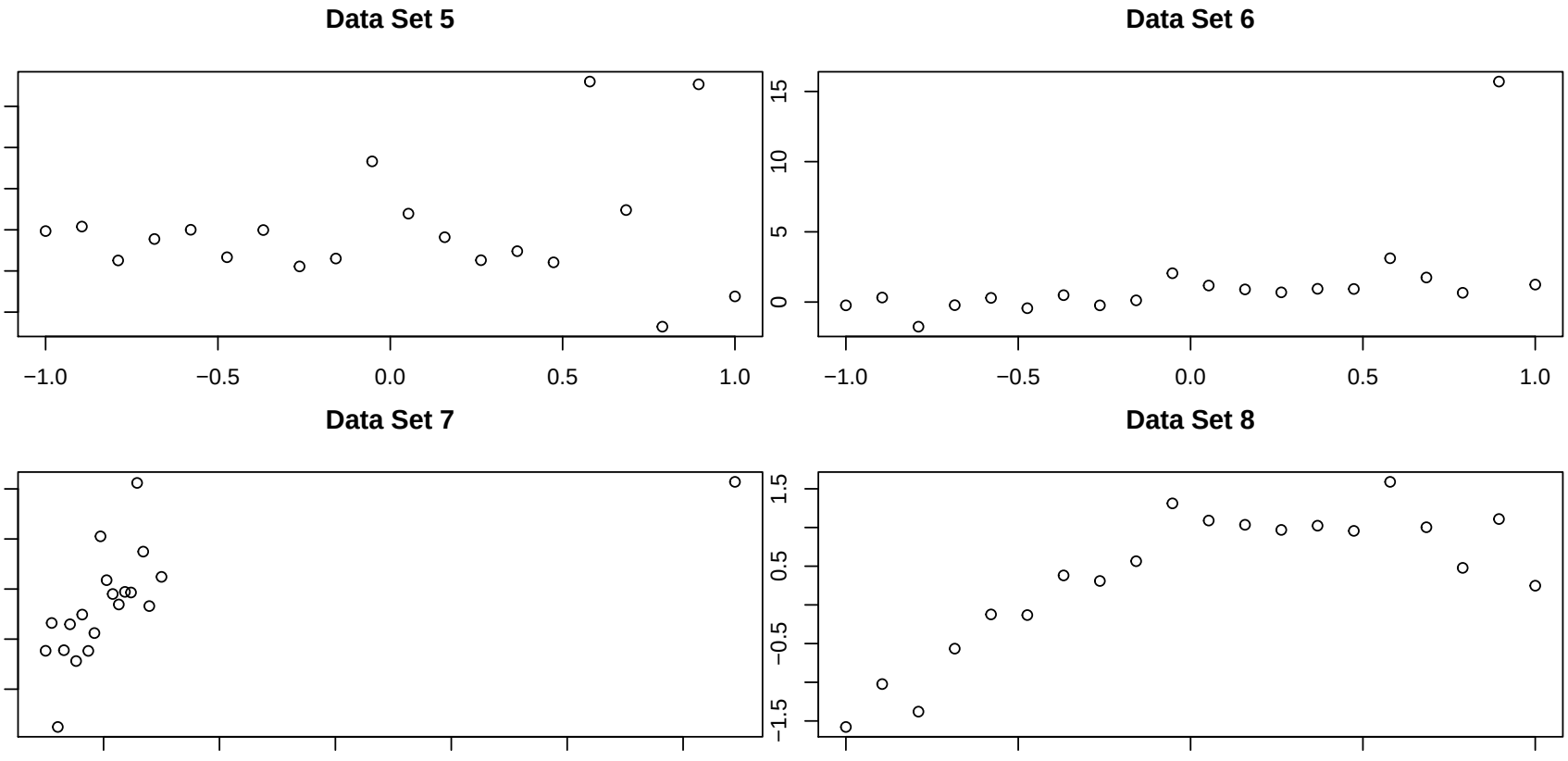
$$\begin{aligned} y_i &= \mu_i & +\varepsilon_i \\ &= \alpha + \beta x_i & +\varepsilon_i \end{aligned}$$

- straight line
- errors are independent
- errors have the same variance
- errors are normally distributed
- errors have zero mean

How can these be wrong? (zero mean is forced by the maximum likelihood)

Regression Assumptions

For data sets 5 - 8, which assumption is wrong?



straight line, independence, same variance, normally distributed

How can we check these?

This will get more complicated later

We need some tools!

Residuals

The model is

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

We can mimic this with the fitted model

$$y_i = \hat{\alpha} + \hat{\beta} x_i + e_i$$

e_i are the ***residuals***

$\hat{\alpha}$ and $\hat{\beta}$ are the parameter estimates: $\hat{\alpha} + \hat{\beta} x_i$ is the prediction for y_i

Residuals

Residuals are estimates of the error

- they should have no structure
- they should be normally distributed

We often use standardised residuals

We also sometimes standardise them:

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

Residuals and Fitted Values

We can extract them in R like this:

```
Women.res <- residuals(WomenMod)
round(Women.res, 2)[1:5]
```

```
##     1     2     3     4     5
##  0.36  0.02  0.08 -0.35  0.11
```

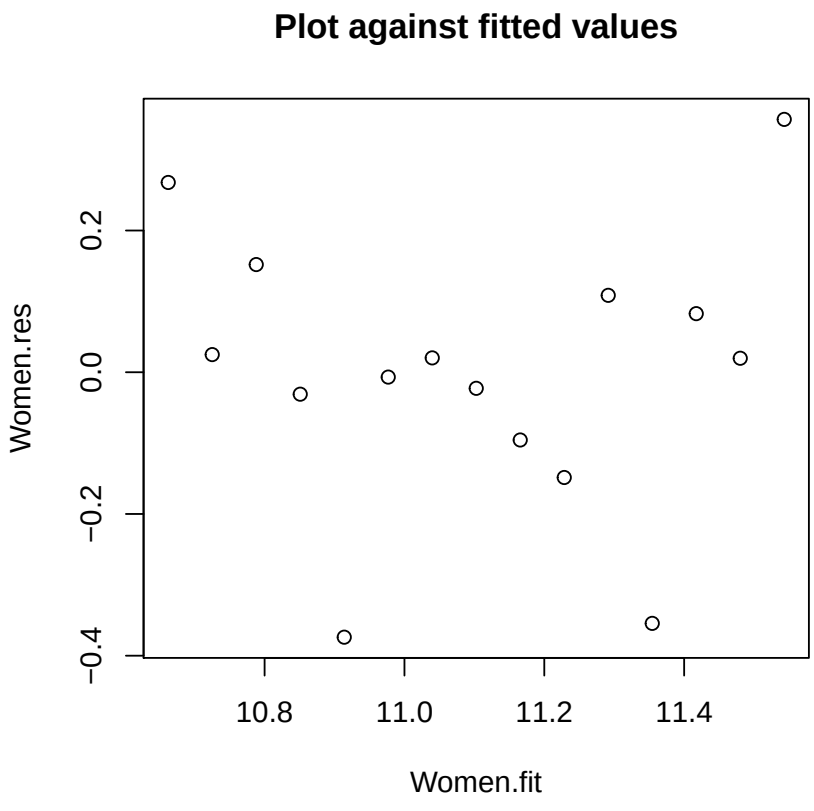
```
Women.fit <- fitted(WomenMod)
round(Women.fit, 2)[1:5]
```

```
##     1     2     3     4     5
## 11.54 11.48 11.42 11.35 11.29
```

We can stare at them, but it is more useful if we plot them

Residual plots

```
par(mfrow=c(1,2))
plot(Women.fit, Women.res, main = "Plot against fitted valu
plot(Times$Year, Women.res, main = "Plot against predictor"
```



(yes, these do look similar)

What Residual plots show

Residuals should not have any structure

With them we can see

- curvature
- outliers
- heteroscedasticity (variance changing)

Residual Exercise

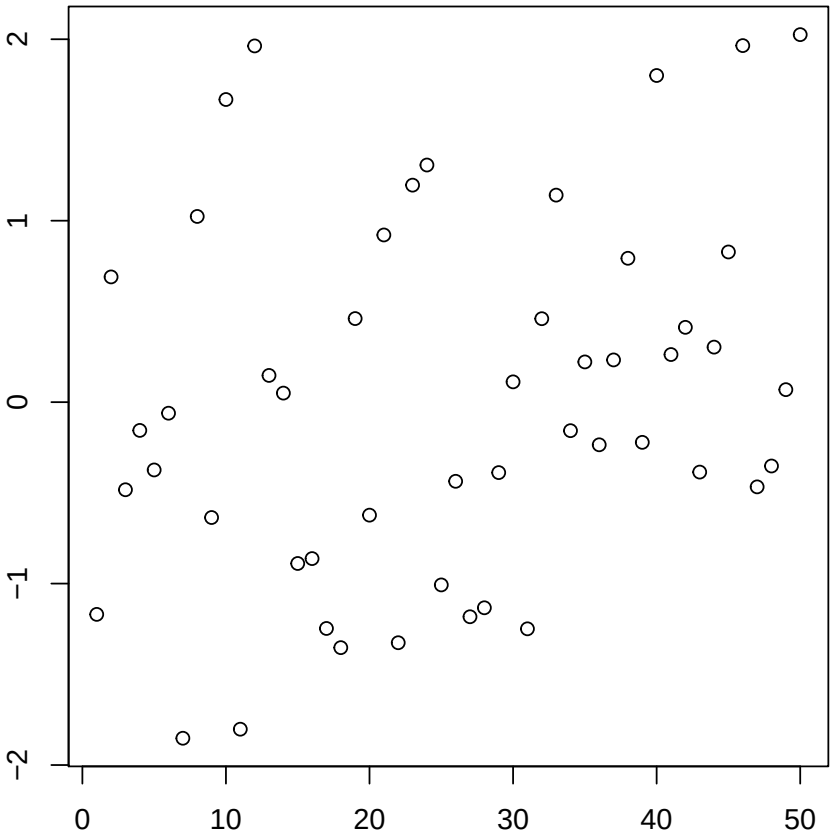
Plot the residuals against the fitted values for all 8 plots.

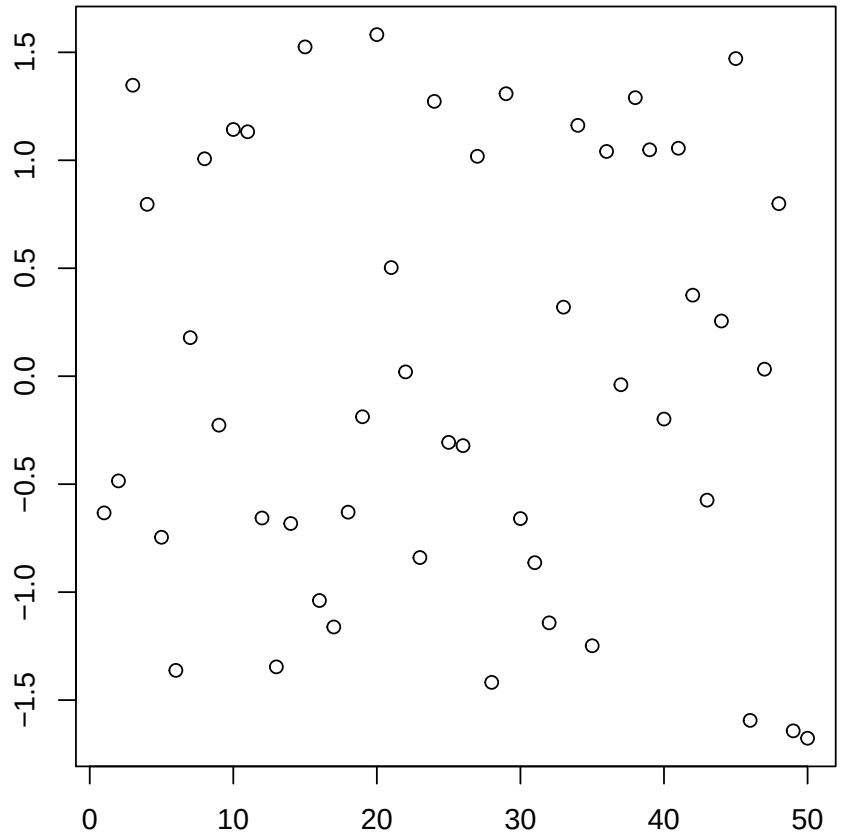
- For which data do they suggest a problem?
- What is the problem?
- Can you think of ways to improve these models?
  - no, you haven't been given the tools yet! So you can be creative

Normal Probability Plots

Residual plots can show some deviant patterns

But they are poor as a test of normality





Yesterday

Last week we looked at regrssion, yesterday we started to look at how good the model is

- does it explain a lot of variation?
- are the assumptions reasonable?

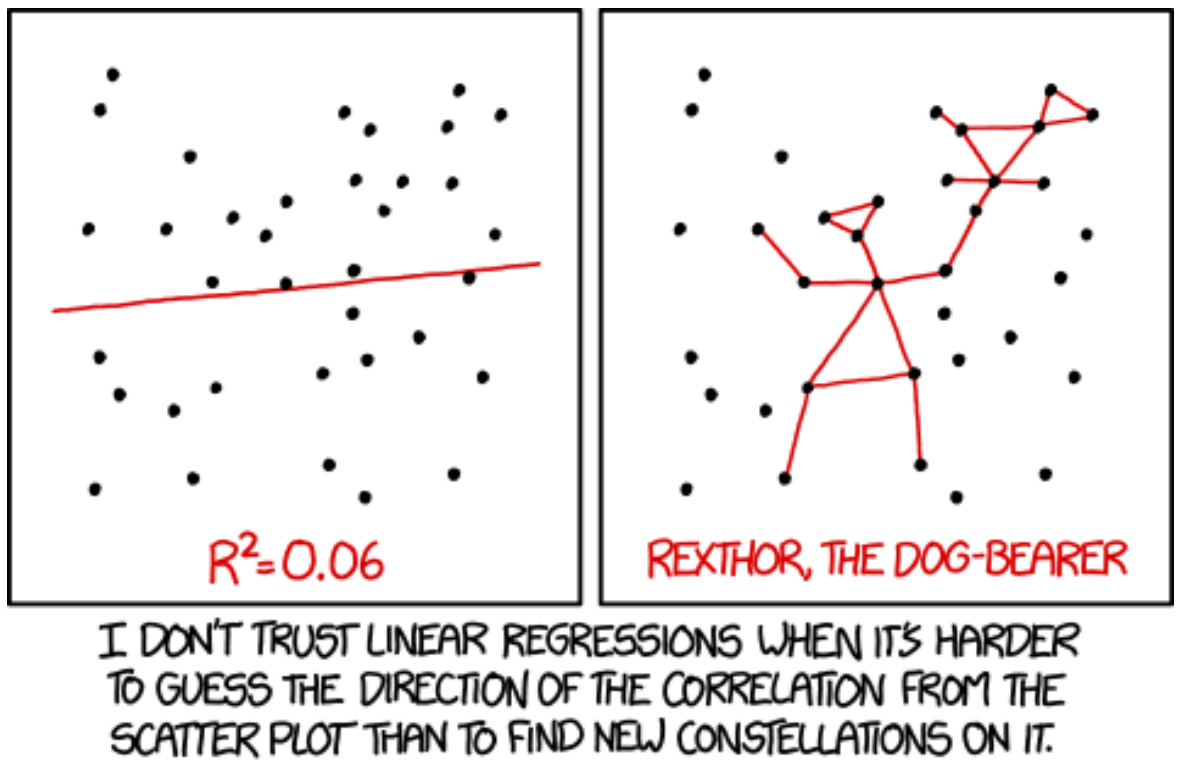


Figure 1: The 95% confidence interval suggests Rexthor's dog could also be a cat, or possibly a teapot.

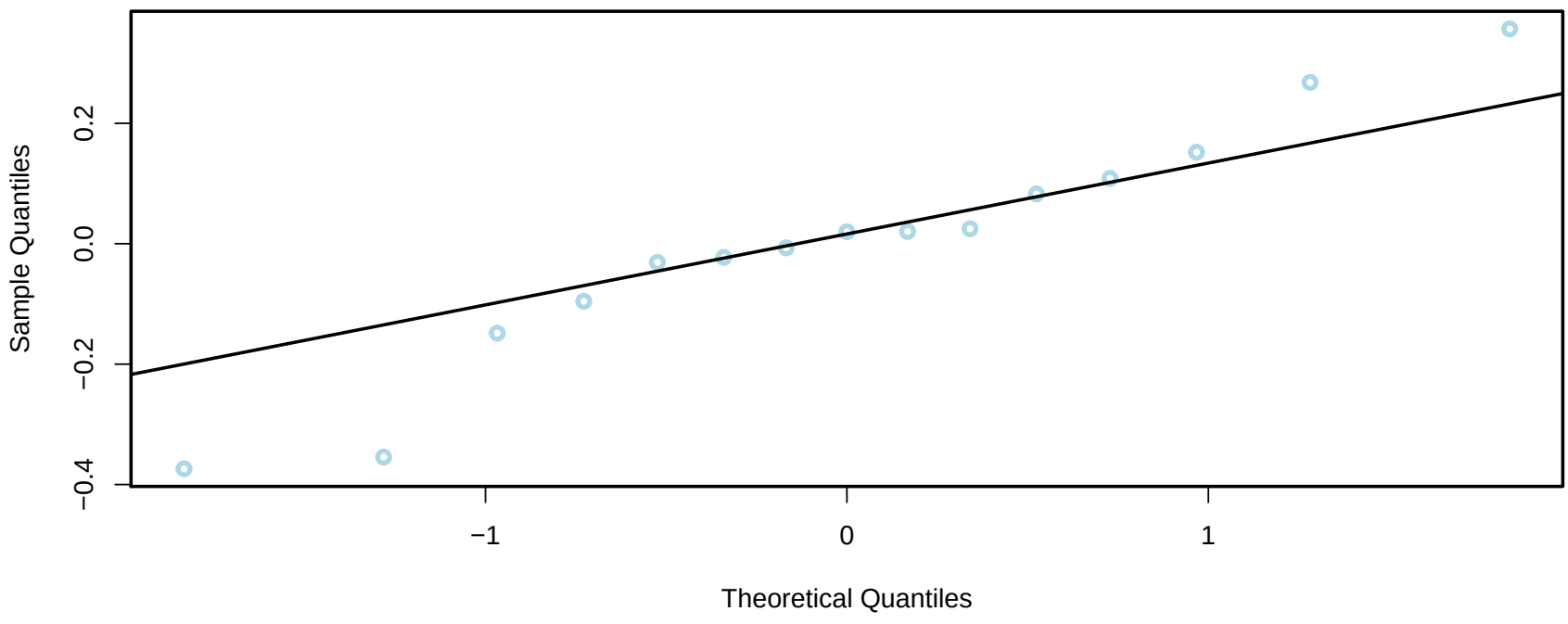
Today

- Normal Probability Plots
- Leverage
- What you can do to improve the model

Normal Probability Plots

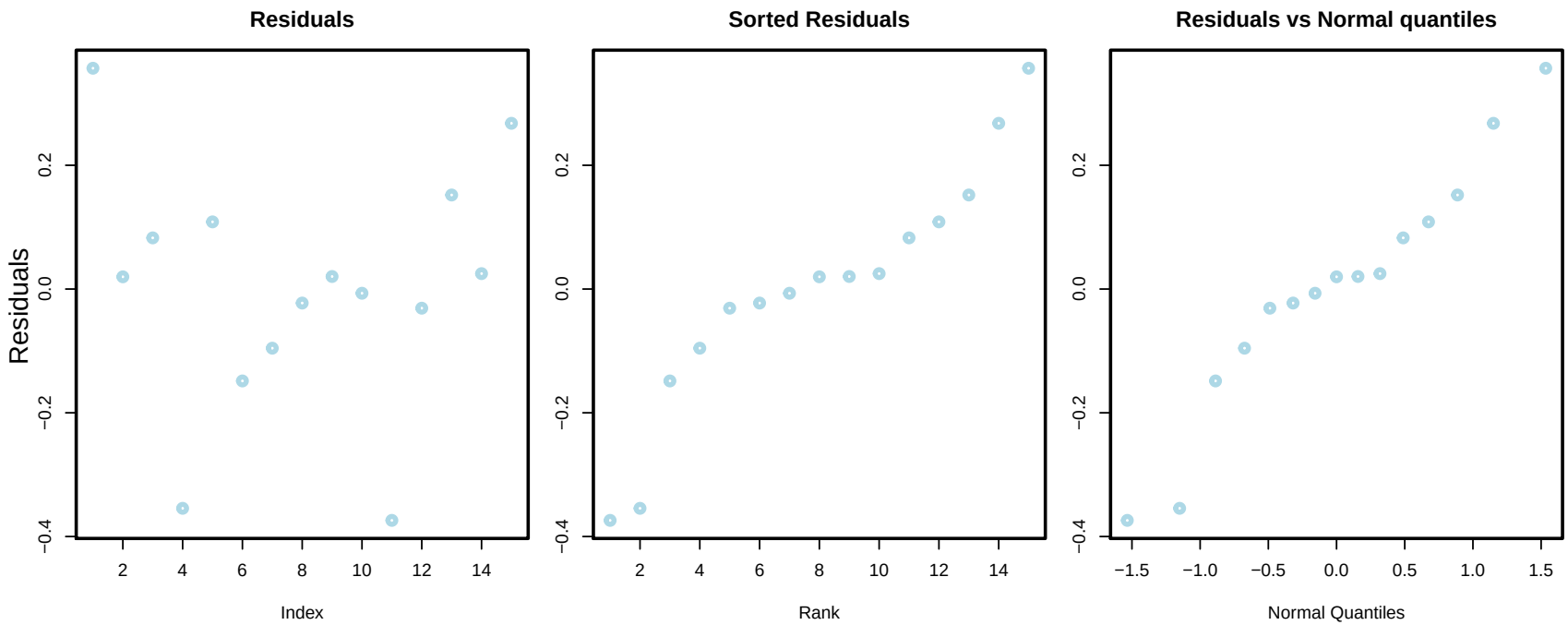
If we sort the data (smallest to largest), we can plot them against their expected values, i.e. plot r_i against the normal quantile

```
par(mar=c(4.1,4.1,1,1), lwd=2)
qqnorm(resid(WomenMod), main="", lwd=3, col="lightblue")
qqline(resid(WomenMod))
```

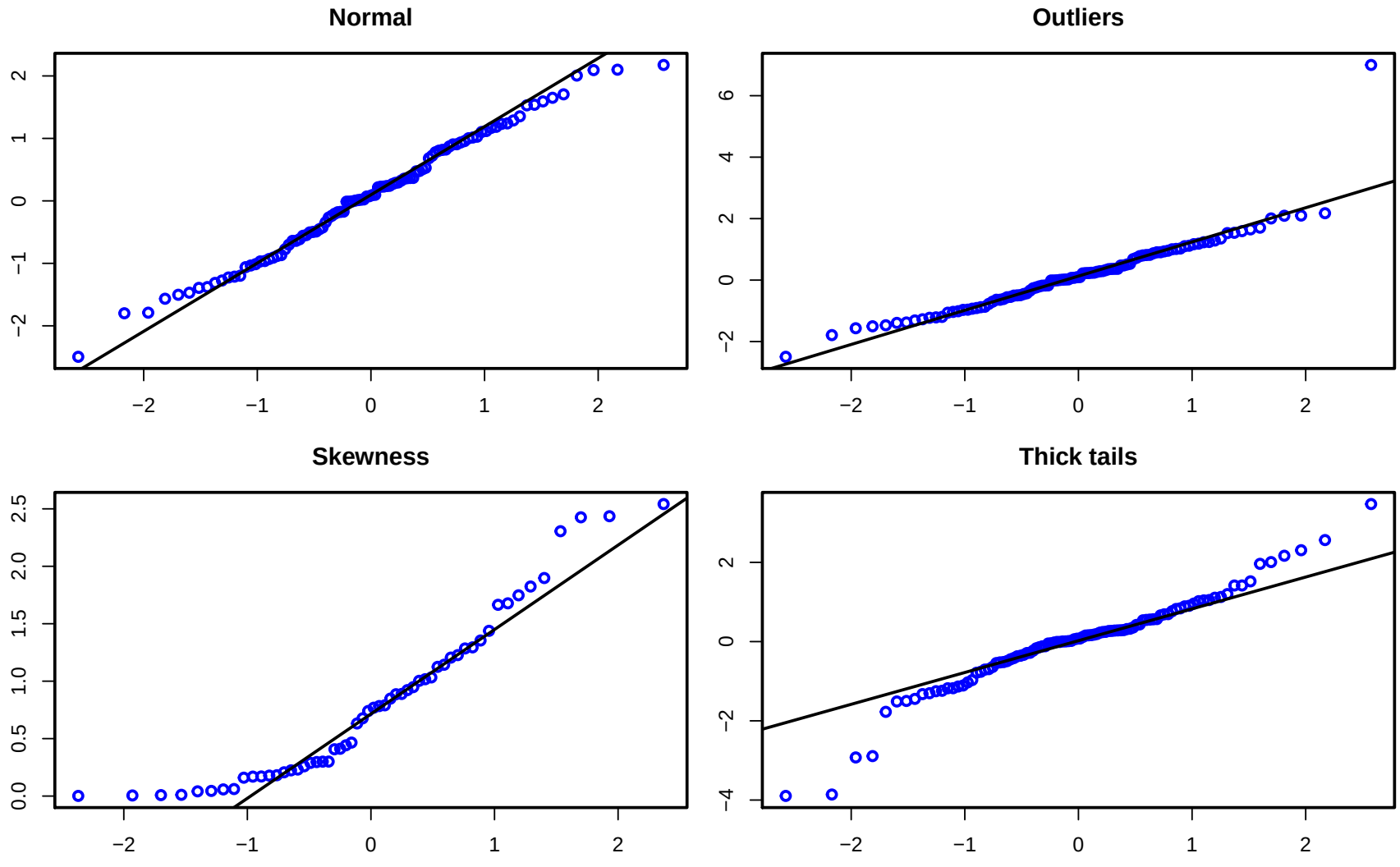


Constructing Probability Plots

```
NormQuants <- qnorm(1:length(Women.res)/
                        (1+length(Women.res)))
par(mfrow=c(1,3), mar=c(4.1,2.1,3,1), oma=c(0,2,0,0), lwd=2
plot(Women.res, lwd=3, col="lightblue", main="Residuals", y
plot(sort(Women.res), lwd=3, col="lightblue", main="Sorted
plot(NormQuants, sort(Women.res), lwd=3, col="lightblue", m
mtext("Residuals", 2, outer=TRUE)
```



What you can see



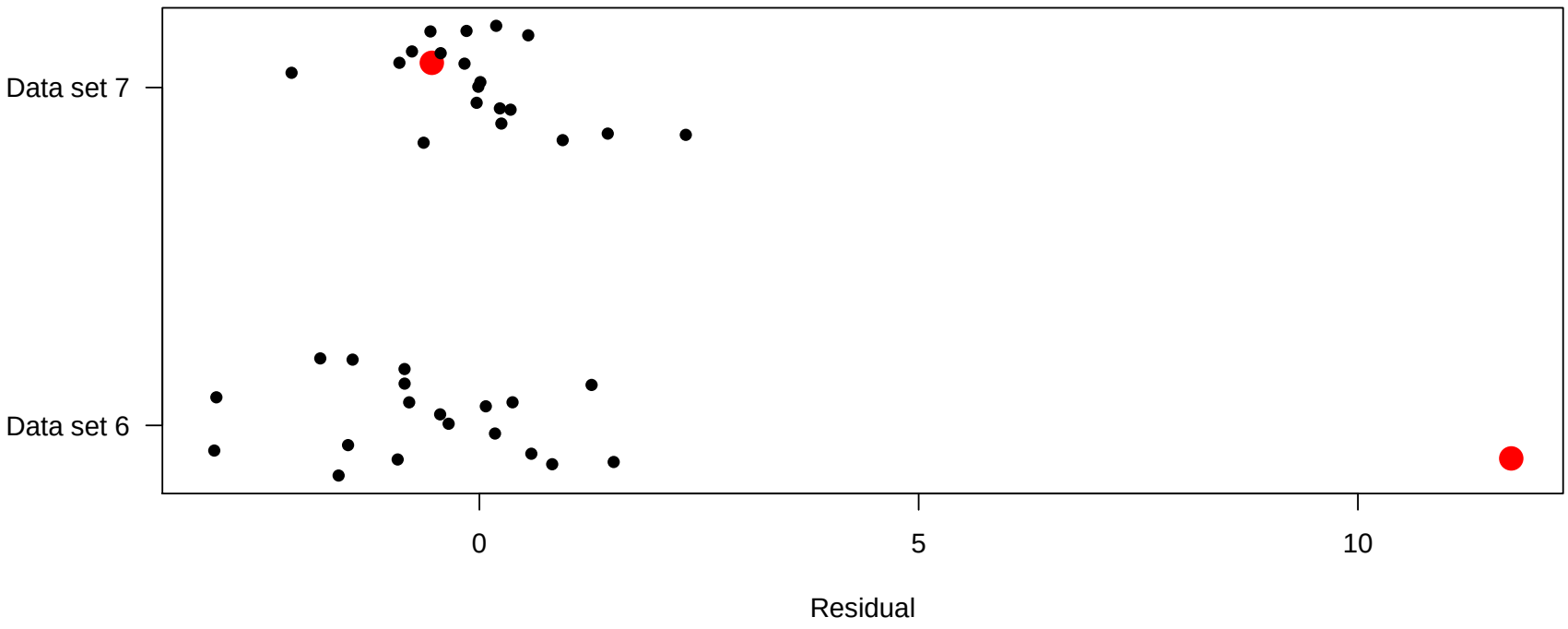
You Turn...

- Draw normal probability plots for the 8 data sets. Do any suggest problems?
- Try to draw normal probability plots that are normal, and then have outliers, skewness and thick tails
  - you will need to simulate data (e.g. with `rnorm()`), and then add points, or transform the data

Leverage

This is less well know, but can be a problem.

Let's look at the residuals for data sets 6 & 7:



In data set 7 there is an obvious weird point, but the residuals don't see it

Influence

Your task

Fit the model with and without the weird point

You can remove the point like this:

```
DataNotWeird <- SimData[SimData$x7<10,]
```

Look at the fitted models. How similar are they?

- check the parameter estimates
- plot the fitted lines on the data (with `abline()`)

Influence and Leverage: Cook's D

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

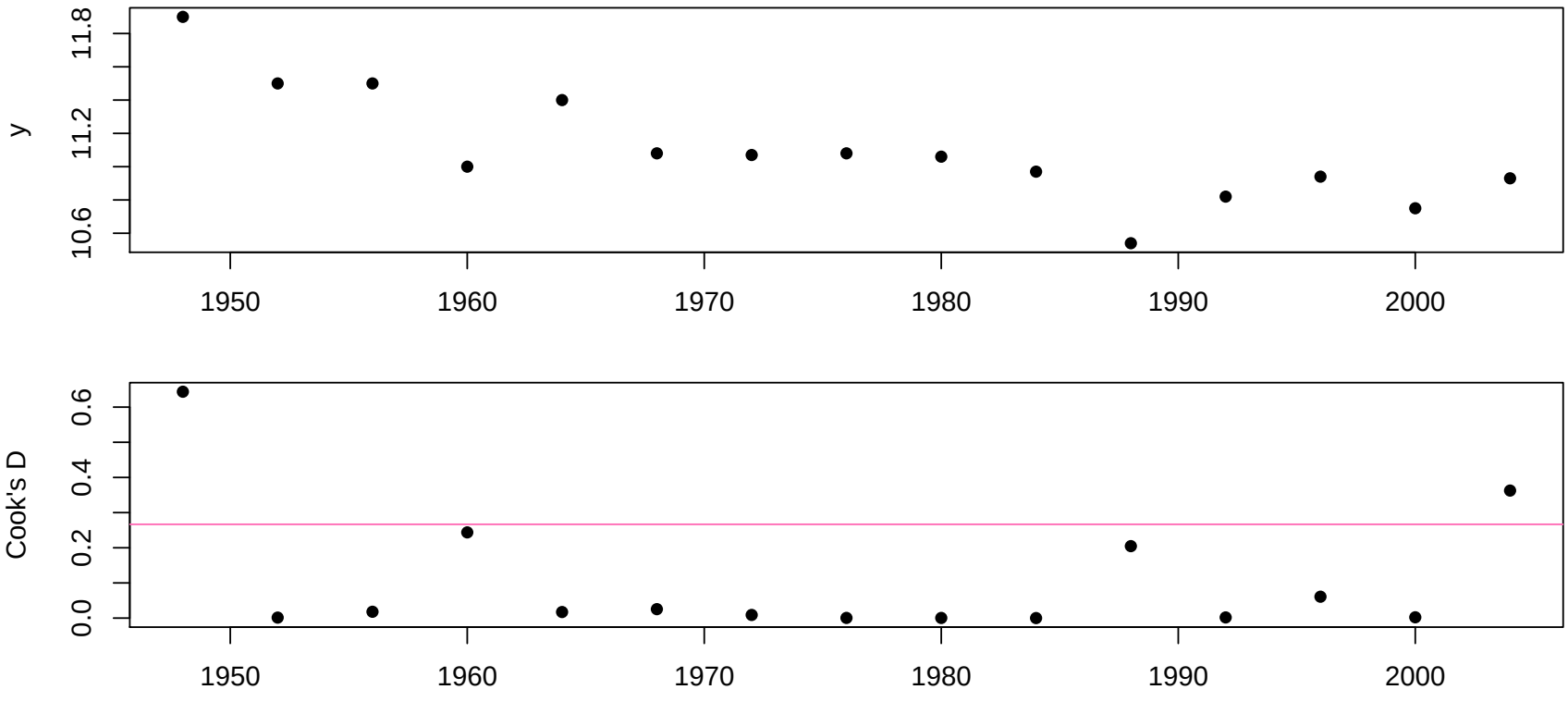
$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(-i)})^2}{s^2}$$

- $\hat{y}_j$ - prediction for full model
- $\hat{y}_{j(-i)}$ - prediction for model with data point i removed
- s^2 - residual variance
- for each data point take the difference in the predicted value for that point between the full model, and the model with that point removed
- sum the squares, and standardise by the residual variance

What is influential?

Large values of D_i mean a large influence

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



Your Turn

Calculate Cook's D for the different data sets, and plot them against `x`. Do you see any influential points?

```
cooks.distance(WomenMod)[1:5]
```

```
##           1           2           3           4
## 0.643742510 0.001416355 0.018007524 0.243659099 0.01724
```

How good is my model? A Summary

- Model as fit + residuals
- R^2: How much variation does the model explain?
- Residual plots
  - curvature
  - outliers
  - heteroscedasticity
- Normal Probability Plots
- Influential Points

How can we improve the model?

First, check the data and model for silly mistakes

- typos are common

Then, ask if if any misfit is a problem

- does it change the conclusions?
- will it change predictions?

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!

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, $\log(x_i)$

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

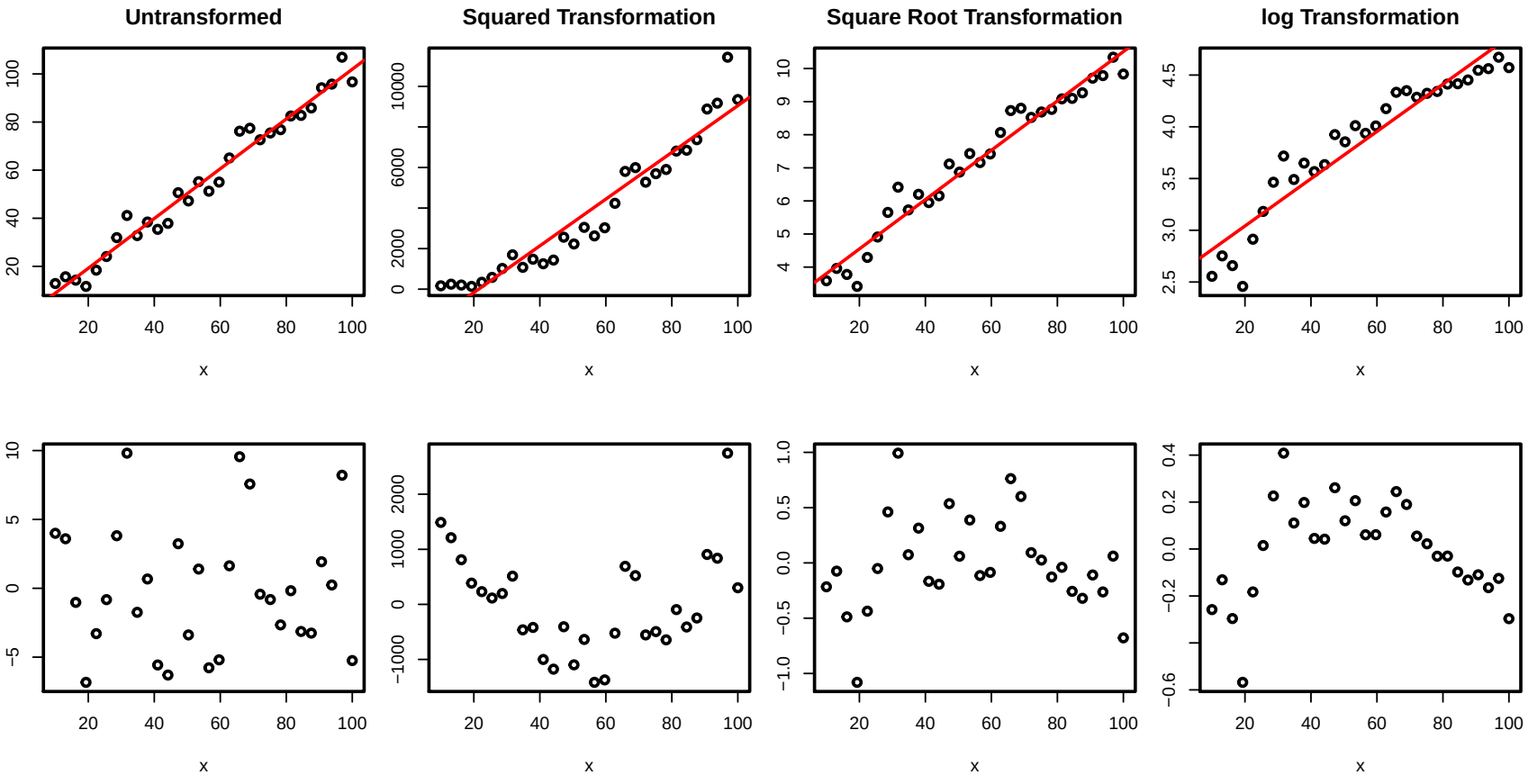
Box-Cox transformations

General Class of transformations

$$y_i \rightarrow y_i^p$$

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

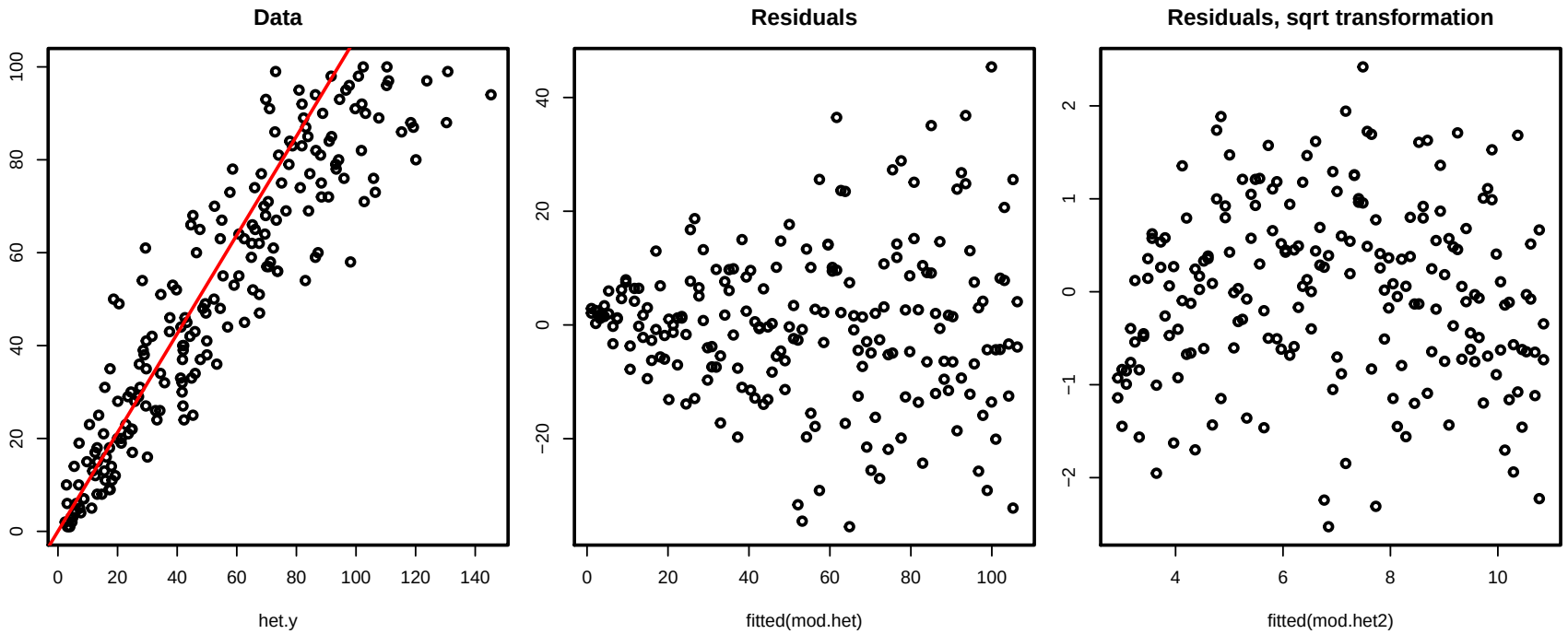
Using Box-Cox transformations



Heteroscedasticity

Variance changes with the mean

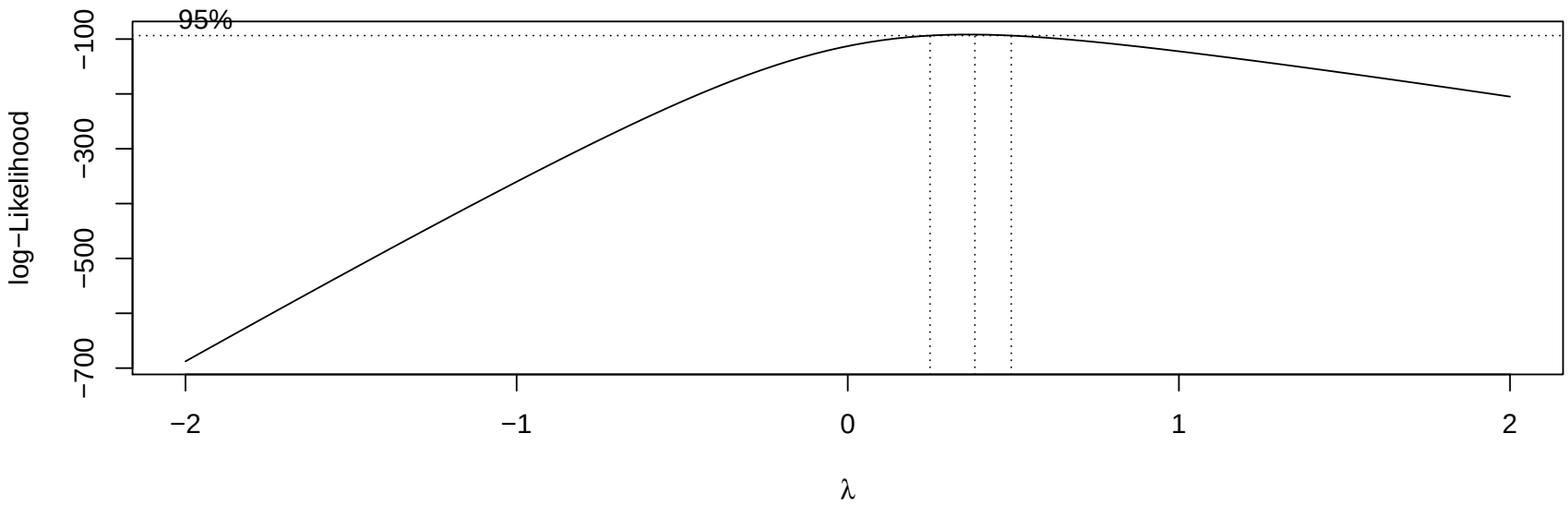
- Box-Cox can also solve this (or make it worse)



Box-Cox in R

R has a function to find the best Box-Cox transformation

```
library(MASS)
x <- 1:50
y <- rnorm(50,0.1*x, 1)^2
boxcox(lm(y ~ x)) # 0.5 is true transformation
```

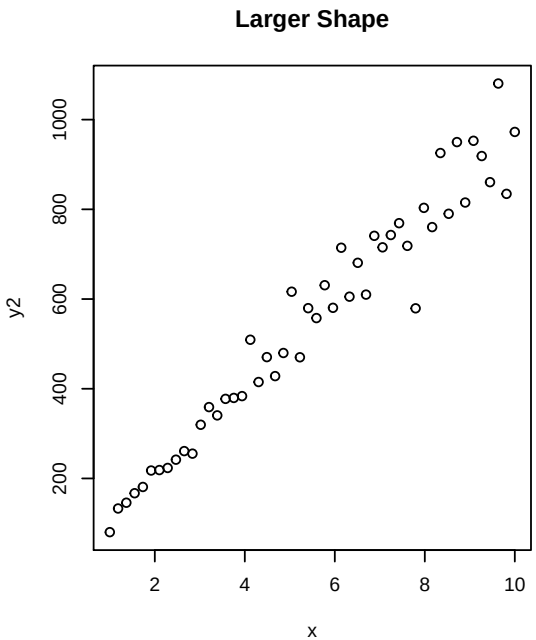


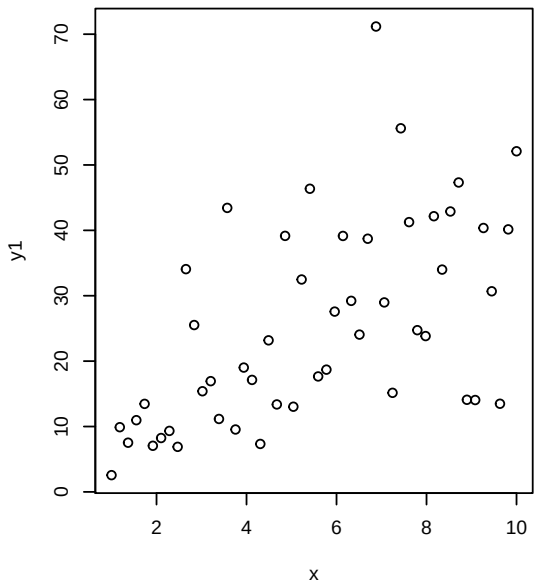
Your Turn

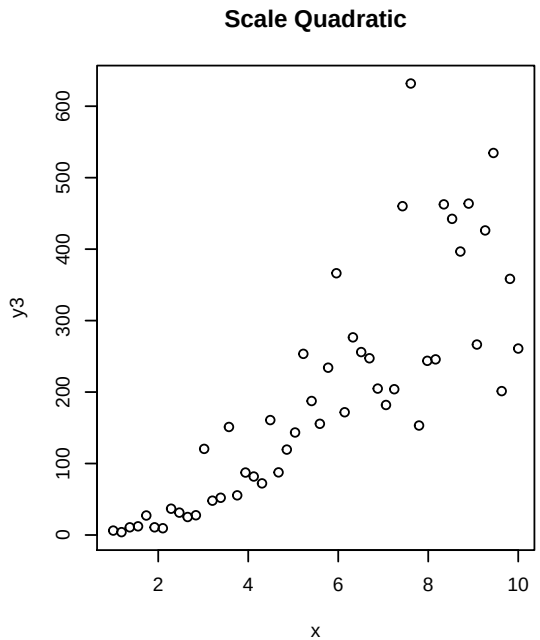
Unfortunately `boxcox()` needs positive responses, so we can't use the data we have already been using. Instead we can create data from a Gamma distribution with `rgamma()`. It has two parameters (after the number of points to simulate):

- Shape: controls skew: the higher, the more symmetrical
- Scale: this controls the mean (mean = shape*scale)

We can create a (bad) model by keeping the shape constant but letting the scale vary with x:







Your Turn, in a moment

This is the code for the previous slide:

```
x <- seq(1,10, length=50)
y1 <- rgamma(length(x), shape=5, scale=x)
y2 <- rgamma(length(x), shape=100, scale=x)
y3 <- rgamma(length(x), shape=5, scale=x^2)
par(mfrow=c(1,3))
plot(x,y2, main="Larger Shape")
plot(x,y1)
plot(x,y3, main="Scale Quadratic")
```

Really Your Turn

- Look at the curves,
- Regress the y against X
- Check the residuals.
- See if a transformation helps, e.g.

```
gam.mod <- lm(y1 ~ x)
library(MASS)
boxcox(gam.mod)
```

- if a transformation is suggested, try it, and check the residuals again

A Word of Caution

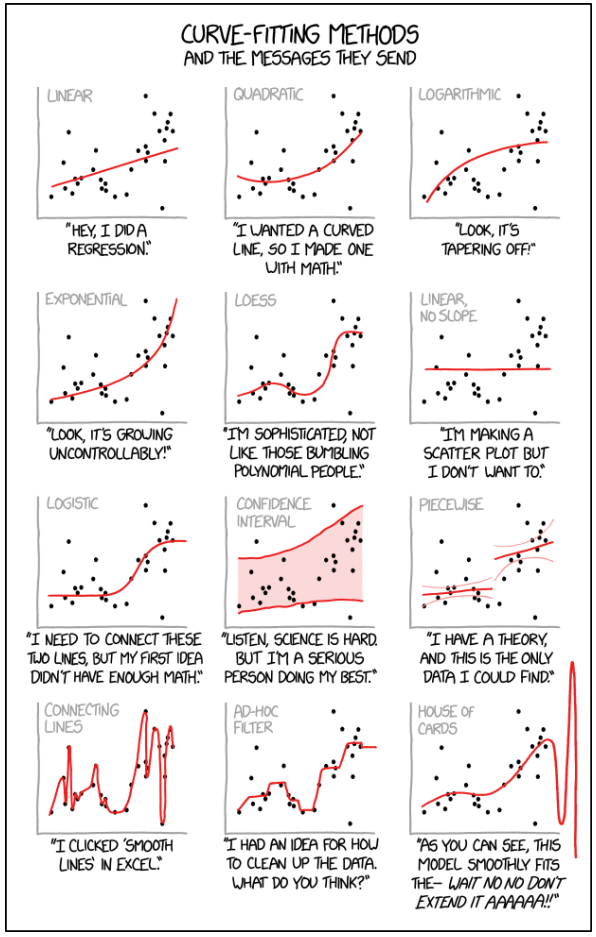


Figure 2: Don't overinterpret

Summary

We now know how to asses the model fit

- R^2 show how much variation the model explains
- Residual plots and Normal Probability Plots can show curvature, outliers, and varying variance
- Influential Points can be detected using Cook's D. These may not be large outliers!
- We should check outliers & other odd points - are they typos?
- We can try to transform the response to get a better model