

An introduction to statistical inference

Francisco Rodriguez-Sanchez

http://bit.ly/frod_san

Why statistics?

To answer questions like...

- ▶ what's the probability that something occurs?

To answer questions like...

- ▶ what's the probability that something occurs?
- ▶ does X influence Y? How much?

To ensure correct inferences

11	451	97	80	46	83	74	29	15	8.03	1.3
439	164	94	45	73	98	99	29	73	34.0	1.3
235	166	172	54	91	85	40	79	73	34.0	1.3
1.433	896	2.132	2.390	3.860	2.175	1.980	3.000	3.000	3.000	1.3
1.870	2.845	1.001	1.920	1.760	2.981	3.000	3.000	3.000	3.000	1.3
2.427	1.335	1.230	1.250	1.715	2.524	2.956	2.956	2.956	2.956	1.3
2.424	2.697	1.001	1.250	1.705	2.520	2.756	2.756	2.756	2.756	1.3
1.692	84	2.05	2.390	3.860	2.175	1.980	3.000	3.000	3.000	1.3
1.199	2.032	1.198	2.415	3.860	2.175	1.980	3.000	3.000	3.000	1.3
35	290	92	430	268	159	324	324	324	324	1.3
35	243	249	277	175	324	324	324	324	324	1.3
74	249	301	47	3.809	2.450	2.450	2.450	2.450	2.450	1.3
94	301	47	6.308	2.450	2.450	2.450	2.450	2.450	2.450	1.3

Inference



Bolker et al 2009 TREE:

'311 out of 537 GLMM analyses (58%) used these tools
inappropriately'

To get answers to tough problems

For example. . .

How many seeds do trees produce?



Inferring tree fecundity



Course goals

- ▶ **Understand** statistical inference

Course goals

- ▶ **Understand** statistical inference
- ▶ Avoid **misconceptions**

Course goals

- ▶ **Understand** statistical inference
- ▶ Avoid **misconceptions**
- ▶ Promote **good practices**

Topics

- ▶ Descriptive statistics

Topics

- ▶ Descriptive statistics
- ▶ Graphics

Topics

- ▶ Descriptive statistics
- ▶ Graphics
- ▶ Sampling

Topics

- ▶ Descriptive statistics
- ▶ Graphics
- ▶ Sampling
- ▶ Experimental design

Topics

- ▶ Descriptive statistics
- ▶ Graphics
- ▶ Sampling
- ▶ Experimental design
- ▶ Hypothesis testing

Topics

- ▶ Descriptive statistics
- ▶ Graphics
- ▶ Sampling
- ▶ Experimental design
- ▶ Hypothesis testing
- ▶ Bayesian inference

Topics

- ▶ Descriptive statistics
- ▶ Graphics
- ▶ Sampling
- ▶ Experimental design
- ▶ Hypothesis testing
- ▶ Bayesian inference
- ▶ Linear models & GLMs

Topics

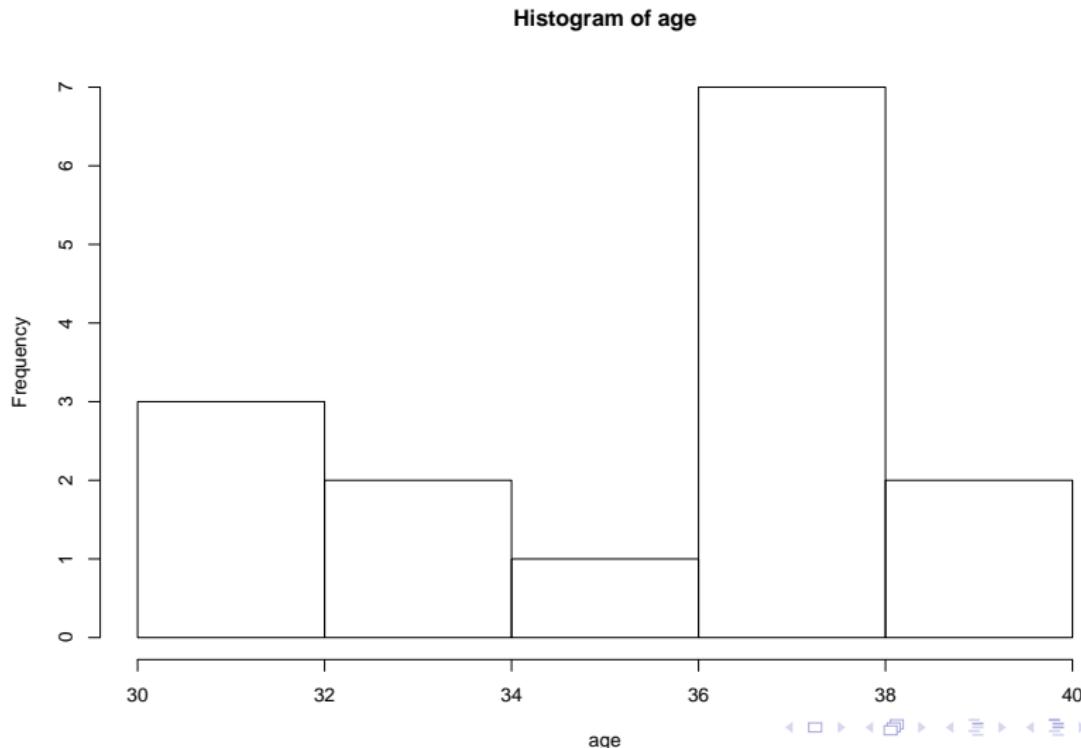
- ▶ Descriptive statistics
- ▶ Graphics
- ▶ Sampling
- ▶ Experimental design
- ▶ Hypothesis testing
- ▶ Bayesian inference
- ▶ Linear models & GLMs
- ▶ Model selection

Descriptive statistics

Guess my age

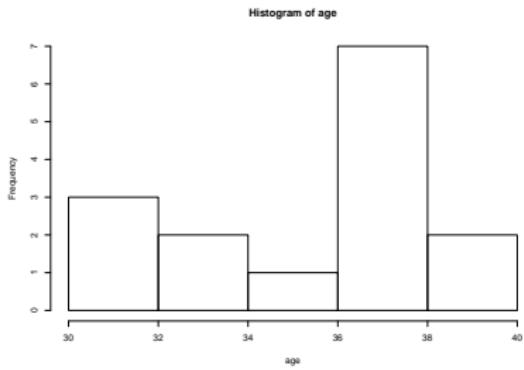
Graph your estimates

```
hist(age)
```



Summarise that distribution

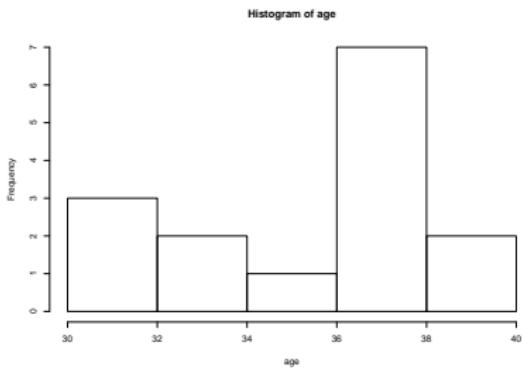
- ▶ Central tendency / location



Summarise that distribution

- ▶ Central tendency / location

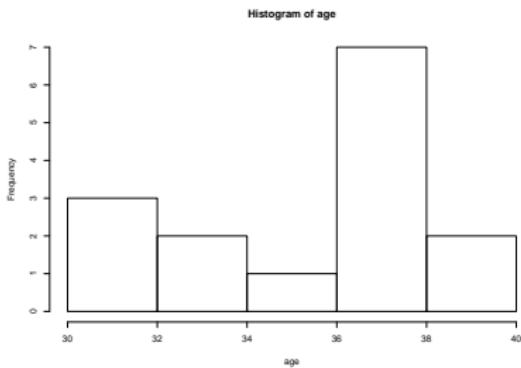
- ▶ mean



Summarise that distribution

► Central tendency / location

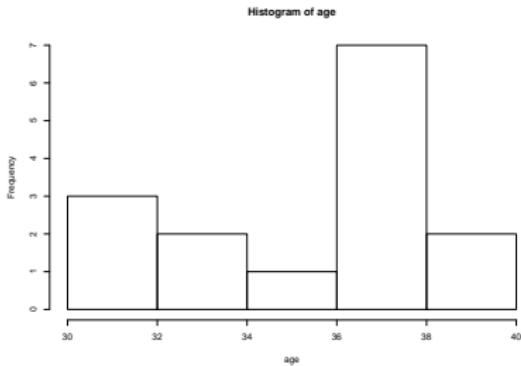
- ▶ mean
- ▶ median



Summarise that distribution

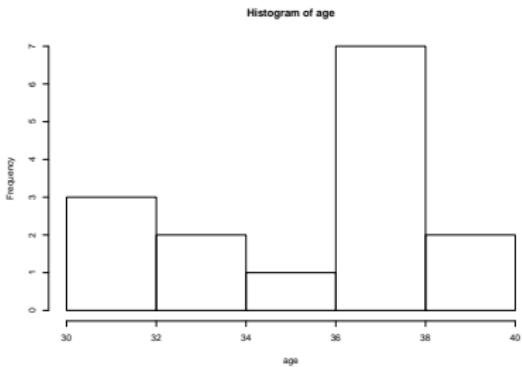
► Central tendency / location

- ▶ mean
- ▶ median
- ▶ mode



Summarise that distribution

- ▶ **Central tendency / location**
 - ▶ mean
 - ▶ median
 - ▶ mode
- ▶ **Variation / Spread**



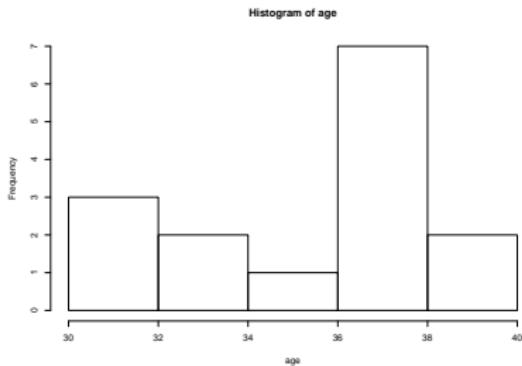
Summarise that distribution

- ▶ **Central tendency / location**

- ▶ mean
- ▶ median
- ▶ mode

- ▶ **Variation / Spread**

- ▶ min, max, range



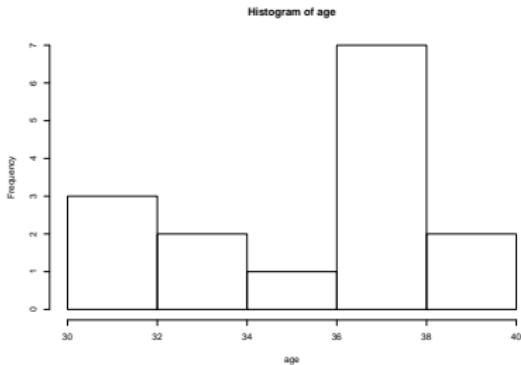
Summarise that distribution

- ▶ **Central tendency / location**

- ▶ mean
- ▶ median
- ▶ mode

- ▶ **Variation / Spread**

- ▶ min, max, range
- ▶ quantiles



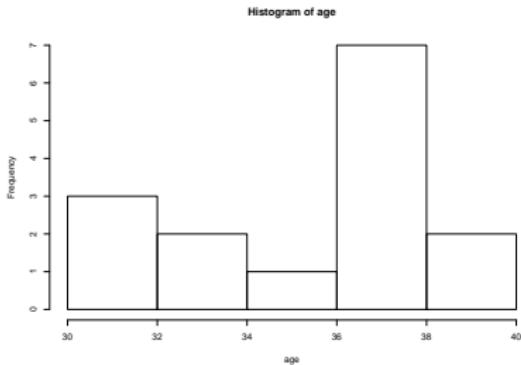
Summarise that distribution

- ▶ **Central tendency / location**

- ▶ mean
- ▶ median
- ▶ mode

- ▶ **Variation / Spread**

- ▶ min, max, range
- ▶ quantiles
- ▶ standard deviation



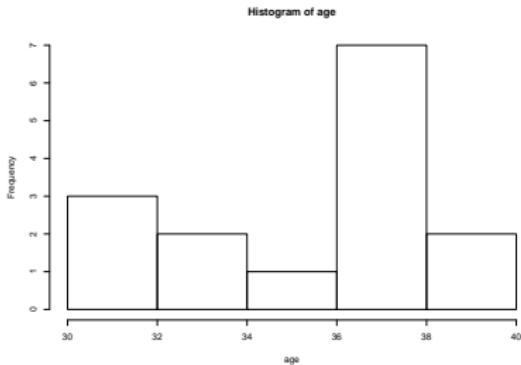
Summarise that distribution

- ▶ **Central tendency / location**

- ▶ mean
- ▶ median
- ▶ mode

- ▶ **Variation / Spread**

- ▶ min, max, range
- ▶ quantiles
- ▶ standard deviation
- ▶ standard error



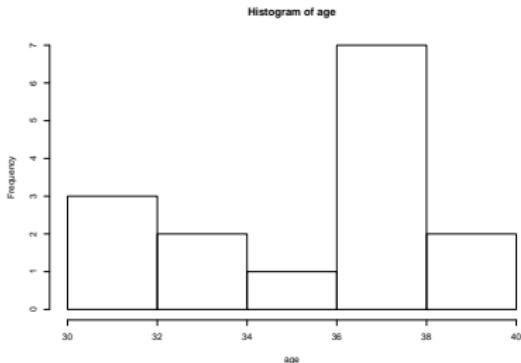
Summarise that distribution

- ▶ **Central tendency / location**

- ▶ mean
- ▶ median
- ▶ mode

- ▶ **Variation / Spread**

- ▶ min, max, range
- ▶ quantiles
- ▶ standard deviation
- ▶ standard error
- ▶ coefficient of variation



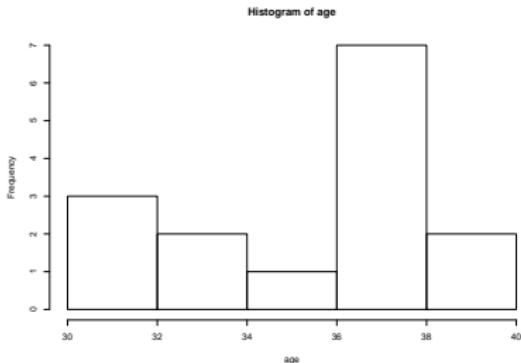
Summarise that distribution

- ▶ **Central tendency / location**

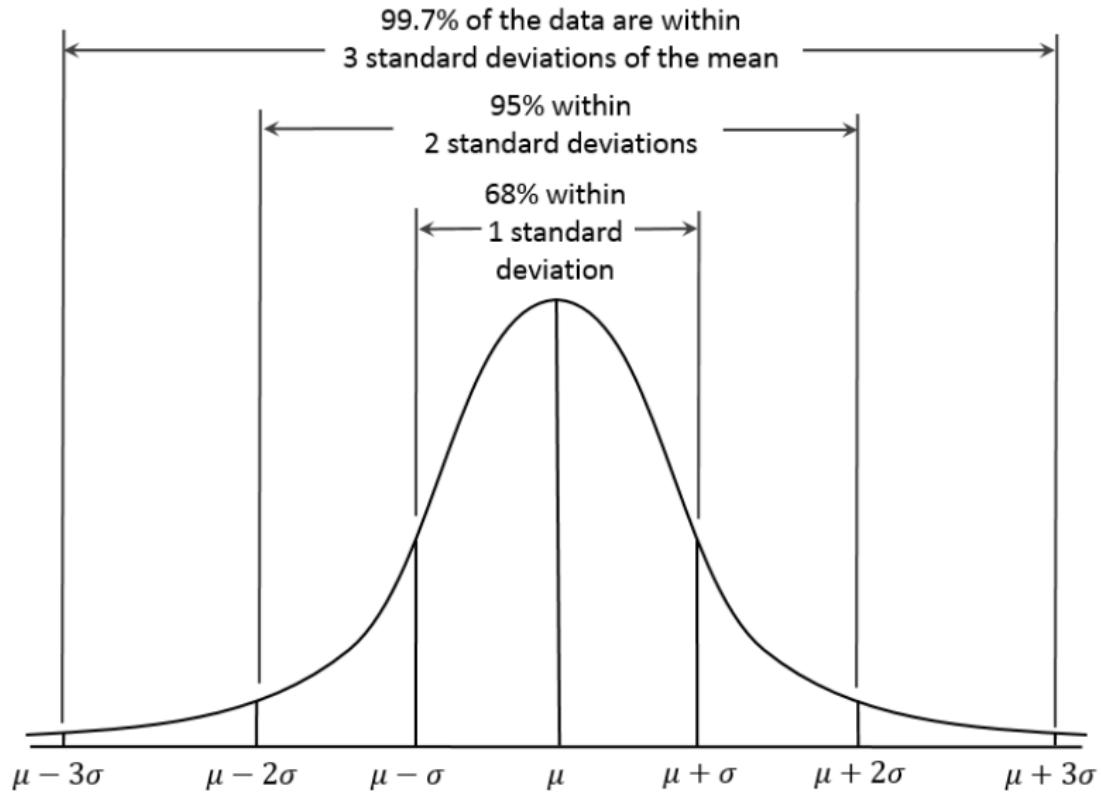
- ▶ mean
- ▶ median
- ▶ mode

- ▶ **Variation / Spread**

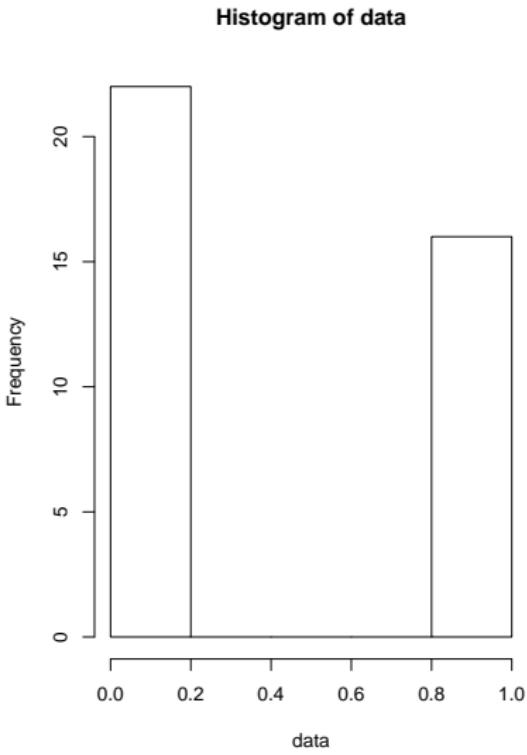
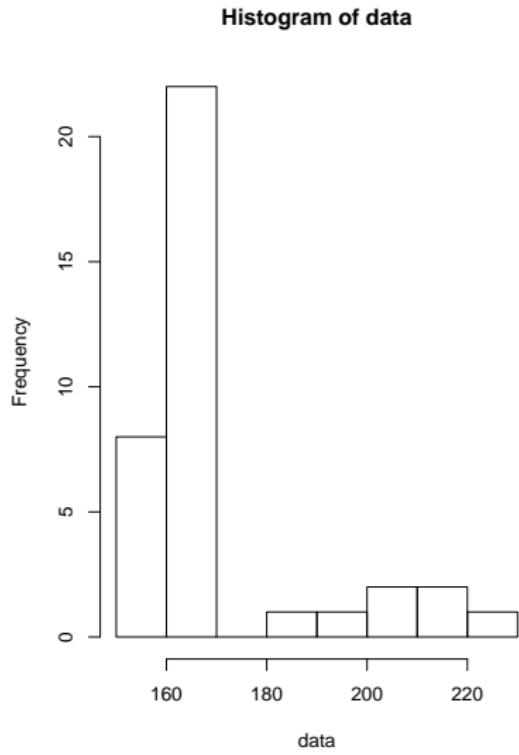
- ▶ min, max, range
- ▶ quantiles
- ▶ standard deviation
- ▶ standard error
- ▶ coefficient of variation
- ▶ confidence intervals



In a Normal distribution



What statistical descriptors are best? (and why)



Sampling

Inference: from samples to population

We rarely measure the whole **population**, but take **samples** instead.



What's the average height in this class?

1. Write down your height and place of origin (Sevilla or other) in a piece of paper and put it in the bag.

What's the average height in this class?

1. Write down your height and place of origin (Sevilla or other) in a piece of paper and put it in the bag.
2. Now everyone **sample** 5 individuals from the whole **population** of heights.

What's the average height in this class?

1. Write down your height and place of origin (Sevilla or other) in a piece of paper and put it in the bag.
2. Now everyone **sample** 5 individuals from the whole **population** of heights.
3. Calculate the mean and 95% CI for your sample
(<http://graphpad.com/quickcalcs/CImean1/>).

What's the average height in this class?

1. Write down your height and place of origin (Sevilla or other) in a piece of paper and put it in the bag.
2. Now everyone **sample** 5 individuals from the whole **population** of heights.
3. Calculate the mean and 95% CI for your sample (<http://graphpad.com/quickcalcs/CImean1/>).
4. Draw on blackboard.

What's the average height in this class?

1. Write down your height and place of origin (Sevilla or other) in a piece of paper and put it in the bag.
2. Now everyone **sample** 5 individuals from the whole **population** of heights.
3. Calculate the mean and 95% CI for your sample (<http://graphpad.com/quickcalcs/CImean1/>).
4. Draw on blackboard.
5. Do all CIs contain true mean height?

Understanding confidence intervals

- ▶ <http://rpsychologist.com/d3/CI/>

Understanding confidence intervals

- ▶ <http://rpsychologist.com/d3/CI/>
- ▶ A 95% CI is **NOT** 95% likely to contain the true parameter value!

Understanding confidence intervals

- ▶ <http://rpsychologist.com/d3/CI/>
- ▶ A 95% CI is **NOT** 95% likely to contain the true parameter value!
- ▶ Instead, 95% of the CIs obtained with this sampling will contain the true value.

Understanding confidence intervals

- ▶ <http://rpsychologist.com/d3/CI/>
- ▶ A 95% CI is **NOT** 95% likely to contain the true parameter value!
- ▶ Instead, 95% of the CIs obtained with this sampling will contain the true value.
- ▶ It's a frequentist, long-run property.

Understanding confidence intervals

- ▶ <http://rpsychologist.com/d3/CI/>
- ▶ A 95% CI is **NOT** 95% likely to contain the true parameter value!
- ▶ Instead, 95% of the CIs obtained with this sampling will contain the true value.
- ▶ It's a frequentist, long-run property.
- ▶ To read more: Morey et al (2015)

What happens if we increase sample size?

- ▶ CI width *decreases* . . .

What happens if we increase sample size?

- ▶ CI width *decreases* . . .
- ▶ but still 5% of CIs will NOT contain true mean!

Bayesian credible intervals

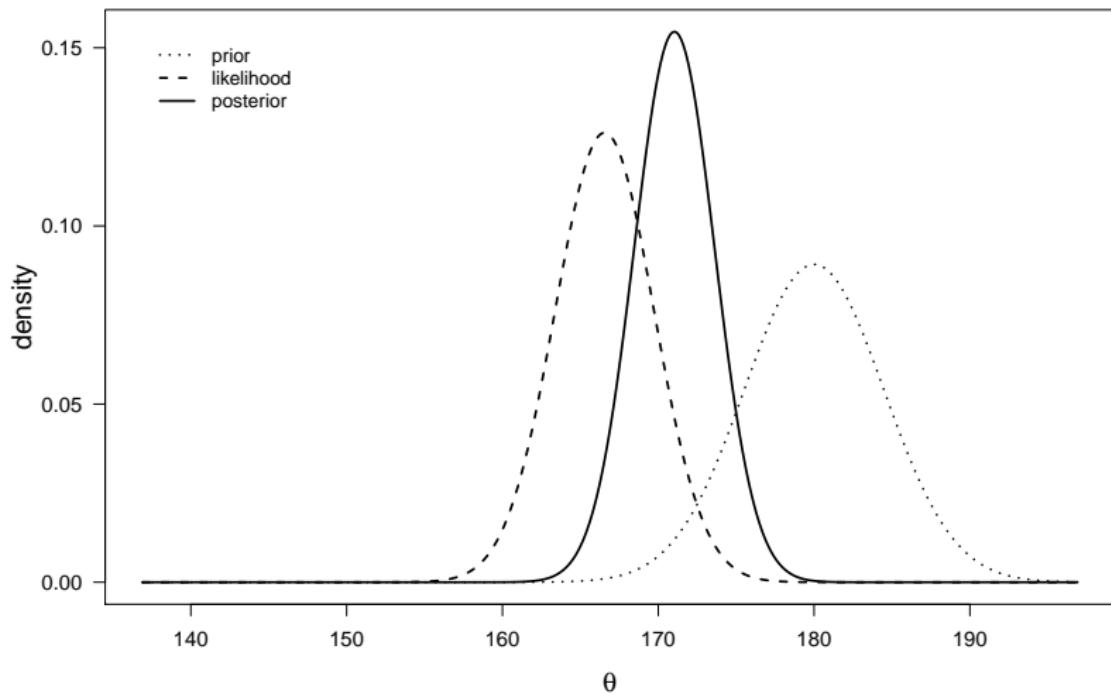
- ▶ Bayesian **credible** intervals do give the probability that true parameter value is contained within them.

Bayesian credible intervals

- ▶ Bayesian **credible** intervals do give the probability that true parameter value is contained within them.
- ▶ Frequentist CIs and Bayesian credible intervals can be similar, but not always.

Bayesian inference: prior, posterior, and likelihood

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$



Experimental design

How would you evaluate fertilizer effect?



Replication!



Replication

- ▶ Replication is key: we need several samples.

Replication

- ▶ Replication is key: we need several samples.
- ▶ How many? Determine n *a priori* according to wanted precision of estimates (*power analysis*).

Replication

- ▶ Replication is key: we need several samples.
- ▶ How many? Determine n *a priori* according to wanted precision of estimates (*power analysis*).
- ▶ Traditionally, ecology studies have had too low sample sizes.

Replication

- ▶ Replication is key: we need several samples.
- ▶ How many? Determine n *a priori* according to wanted precision of estimates (*power analysis*).
- ▶ Traditionally, ecology studies have had too low sample sizes.
- ▶ Hence missing many subtle effects, and prone to bias.

Replication

- ▶ Replication is key: we need several samples.
- ▶ How many? Determine n *a priori* according to wanted precision of estimates (*power analysis*).
- ▶ Traditionally, ecology studies have had too low sample sizes.
- ▶ Hence missing many subtle effects, and prone to bias.
- ▶ Complex models (w/ many predictors, interactions etc) require **high** sample sizes.

Sample size is very important

See *The evolution of correlations*

Stopping rules:

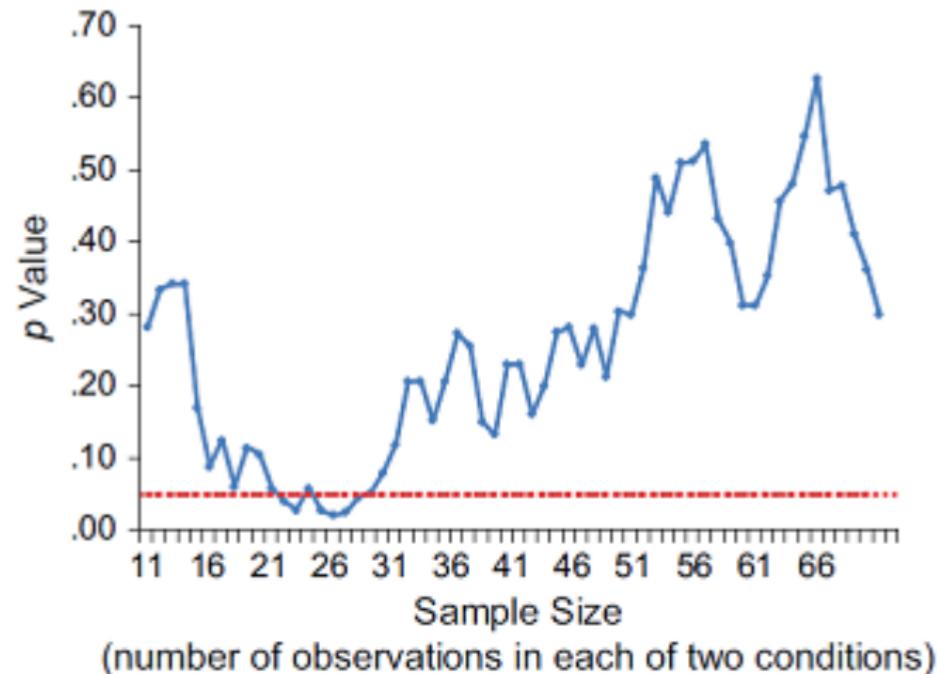


Fig. 2. Illustrative simulation of p values obtained by a researcher who continuously adds an observation to each of two conditions, conducting a t test after each addition. The dotted line highlights the conventional significance criterion of $p \leq .05$.

Randomization



Randomization

- ▶ Haphazard \neq Random

Randomization

- ▶ Haphazard \neq Random
- ▶ Stratify: randomize within groups (e.g. species, soil types)

Have controls

- ▶ Untreated individuals, plots... (assigned randomly, of course).

Have controls

- ▶ Untreated individuals, plots... (assigned randomly, of course).
- ▶ Must differ only in treatment (i.e. homogeneous environment).

Have controls

- ▶ Untreated individuals, plots... (assigned randomly, of course).
- ▶ Must differ only in treatment (i.e. homogeneous environment).
- ▶ Measure before & after treatment.

Have controls

- ▶ Untreated individuals, plots... (assigned randomly, of course).
- ▶ Must differ only in treatment (i.e. homogeneous environment).
- ▶ Measure before & after treatment.
- ▶ Consider blind designs to avoid observer bias.

Hypothesis testing

Does height differ between local and foreign students?

- ▶ Local people heights:

188 162 195 150 162

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
150.0	162.0	162.0	171.4	188.0	195.0

180 182 166 179 188 177 176 167 186 191

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
166.0	176.2	179.5	179.2	185.0	191.0

Does height differ between local and foreign students?

- ▶ Local people heights:

188 162 195 150 162

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
150.0	162.0	162.0	171.4	188.0	195.0

- ▶ Other heights:

180 182 166 179 188 177 176 167 186 191

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
166.0	176.2	179.5	179.2	185.0	191.0

Does height differ between local and foreign students?

- ▶ Local people heights:

188 162 195 150 162

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
150.0	162.0	162.0	171.4	188.0	195.0

- ▶ Other heights:

180 182 166 179 188 177 176 167 186 191

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
166.0	176.2	179.5	179.2	185.0	191.0

- ▶ We know what happens in **our samples**, but want to extrapolate to the whole **population**.

If we sample students' heights in this class. . .

- ▶ Can we extrapolate results to

If we sample students' heights in this class. . .

- ▶ Can we extrapolate results to
 - ▶ this class?

If we sample students' heights in this class. . .

- ▶ Can we extrapolate results to
 - ▶ this class?
 - ▶ this university?

If we sample students' heights in this class. . .

- ▶ Can we extrapolate results to
 - ▶ this class?
 - ▶ this university?
 - ▶ this city?

If we sample students' heights in this class. . .

- ▶ Can we extrapolate results to
 - ▶ this class?
 - ▶ this university?
 - ▶ this city?
 - ▶ the world?

If we sample students' heights in this class. . .

- ▶ Can we extrapolate results to
 - ▶ this class?
 - ▶ this university?
 - ▶ this city?
 - ▶ the world?
- ▶ What's the **suitable population** to make inferences given this sample?

NHST concepts

Null and alternative hypotheses

- ▶ Tell me...

Null and alternative hypotheses

- ▶ Tell me...
- ▶ **Null hypothesis:** heights don't differ.

Null and alternative hypotheses

- ▶ Tell me...
- ▶ **Null hypothesis:** heights don't differ.
- ▶ **Alternative hypothesis:** heights are different.

P value

► ...

P value

- ▶ ...
- ▶ Probability of observing data as or more extreme than these *if H₀ was true.*

P value

- ▶ ...
- ▶ Probability of observing data as or more extreme than these *if H₀ was true.*
- ▶ Hence **the lower P the more unlikely H₀** (i.e. more likely there's a true difference).

Are differences *significant*?

- ▶ If $p < 0.05$, we **reject** H_0 .

Are differences *significant*?

- ▶ If $p < 0.05$, we **reject** H_0 .
- ▶ If $p > 0.05$, we **fail to reject** H_0

Are differences *significant*?

- ▶ If $p < 0.05$, we **reject** H_0 .
- ▶ If $p > 0.05$, we **fail to reject** H_0
- ▶ (which is **not** the same as ' H_0 is true')

Let's do the test

```
t.test(h.sevi, h.out)
```

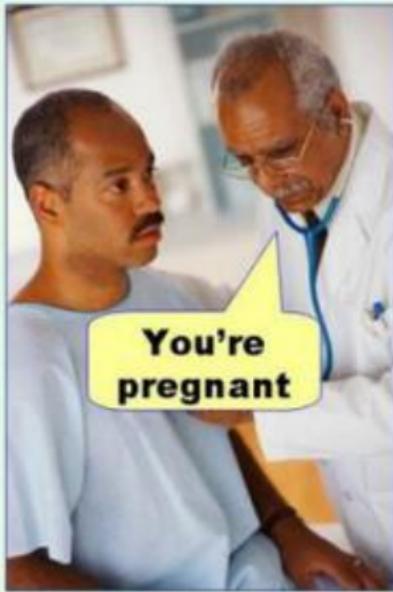
Welch Two Sample t-test

```
data: h.sevi and h.out
t = -0.87134, df = 4.7547, p-value = 0.4254
alternative hypothesis: true difference in means is not equal to
95 percent confidence interval:
-31.1728 15.5728
sample estimates:
mean of x mean of y
171.4      179.2
```

Are heights different then?

Rejecting hypotheses: two types of error

Type I error
(false positive)



Type II error
(false negative)



Rejecting hypotheses: two types of error

Statistics: Hypothesis Test	Null Hypothesis is True	Null Hypothesis is False
Reject Null Hypothesis	Type I Error	Correct
Fail to Reject Null Hypothesis	Correct	Type II Error

Power: Probability of detecting true difference (rejecting H₀ when it's false).

Understanding NHST

<http://rpsychologist.com/d3/NHST/>

Example: biased coin

```
[1] 0 0 1 0 0 1 1 1 1 0
```

```
1-sample proportions test without continuity correction
```

```
data: sum(coin) out of ntrials, null probability 0.5  
X-squared = 0, df = 1, p-value = 1
```

```
alternative hypothesis: true p is not equal to 0.5
```

```
95 percent confidence interval:
```

```
0.2365931 0.7634069
```

```
sample estimates:
```

```
p  
0.5
```

Correlation between variables

<http://rpsychologist.com/d3/correlation/>

Common pitfalls and good practice

Interesting reading

esa

ECOSPHERE

Applied statistics in ecology: common pitfalls and simple solutions

E. ASHLEY STEEL,^{1,†} MAUREEN C. KENNEDY,² PATRICK G. CUNNINGHAM,³ AND JOHN S. STANOVICK⁴

<http://dx.doi.org/10.1890/ES13-00160.1>
Also <http://www.statisticsonewrong.com/>

First things first

- ▶ Always

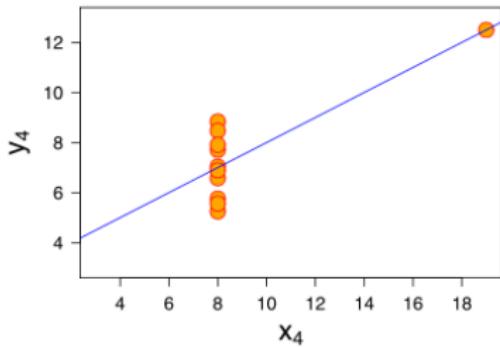
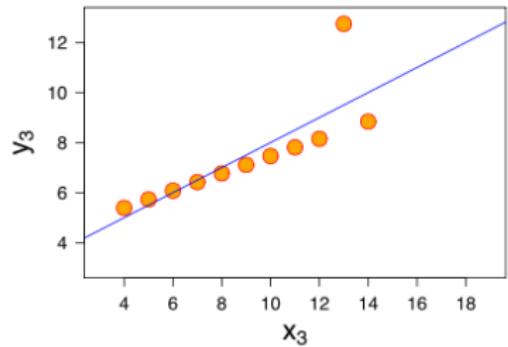
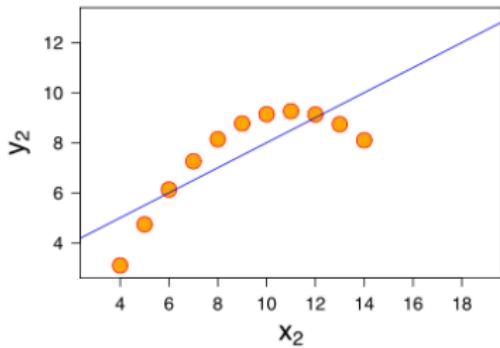
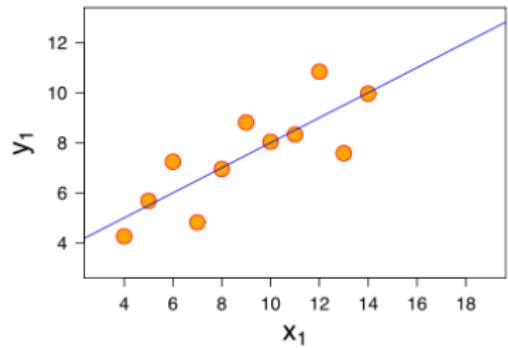
First things first

- ▶ Always
- ▶ Always

First things first

- ▶ Always
- ▶ Always
- ▶ Always

Plot data and models



Plot. Check models. Plot. Check assumptions. Plot.

Lavine 2014 *Ecology*

News: Hamburgers increase risk of heart attack

- ▶ In a sample of 10,000 people, it was found that people eating >2 hamburgers a week had 50% higher probability of heart attack.

News: Hamburgers increase risk of heart attack

- ▶ In a sample of 10,000 people, it was found that people eating >2 hamburgers a week had 50% higher probability of heart attack.
- ▶ **Do hamburgers increase heart attacks?**

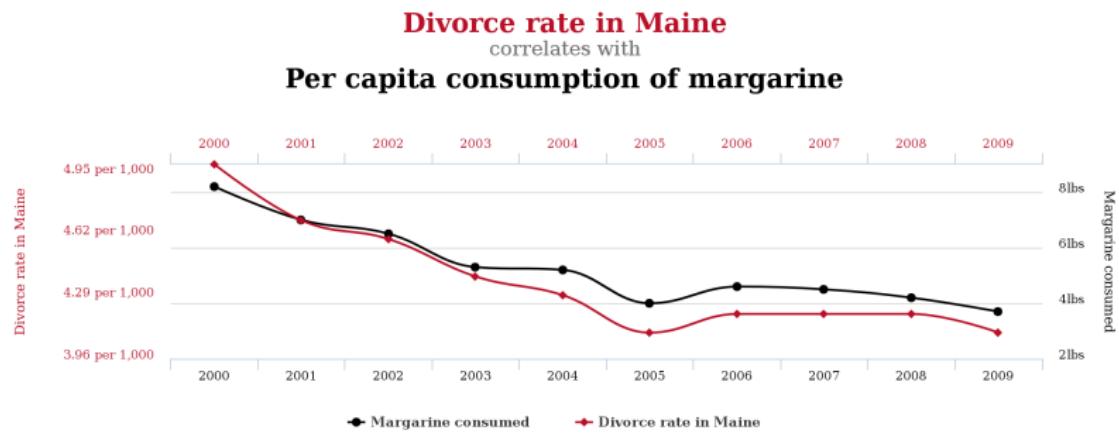
Bigger flowers increase reproductive success

- ▶ We found that plants with big flowers produced 30% more seeds... .

Bigger flowers increase reproductive success

- ▶ We found that plants with big flowers produced 30% more seeds...
- ▶ **Do big flowers increase reproductive success?**

Correlation vs Causation



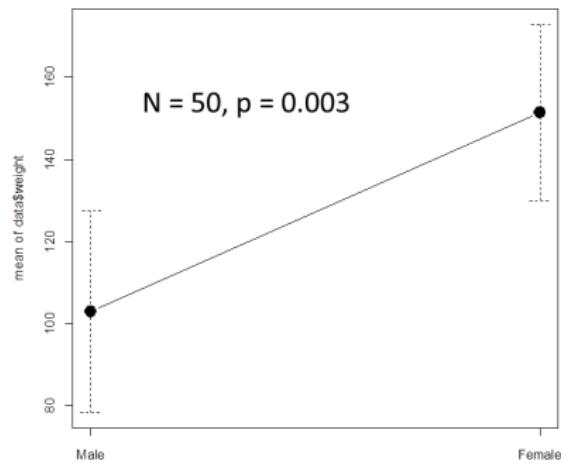
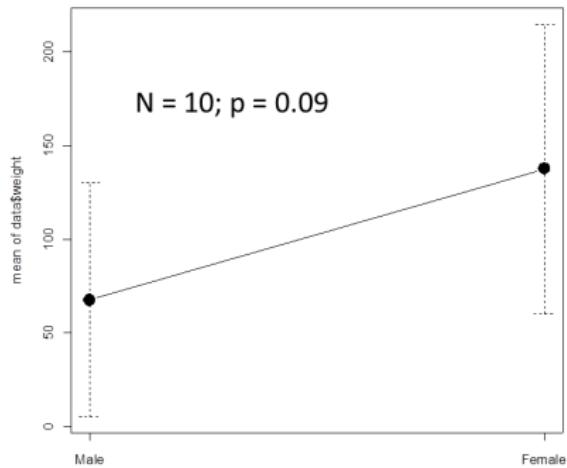
tylervigen.com

<http://tylervigen.com/spurious-correlations>

P-value depends on sample size

- ▶ Same real difference is detected as significant or not depending on sample size:

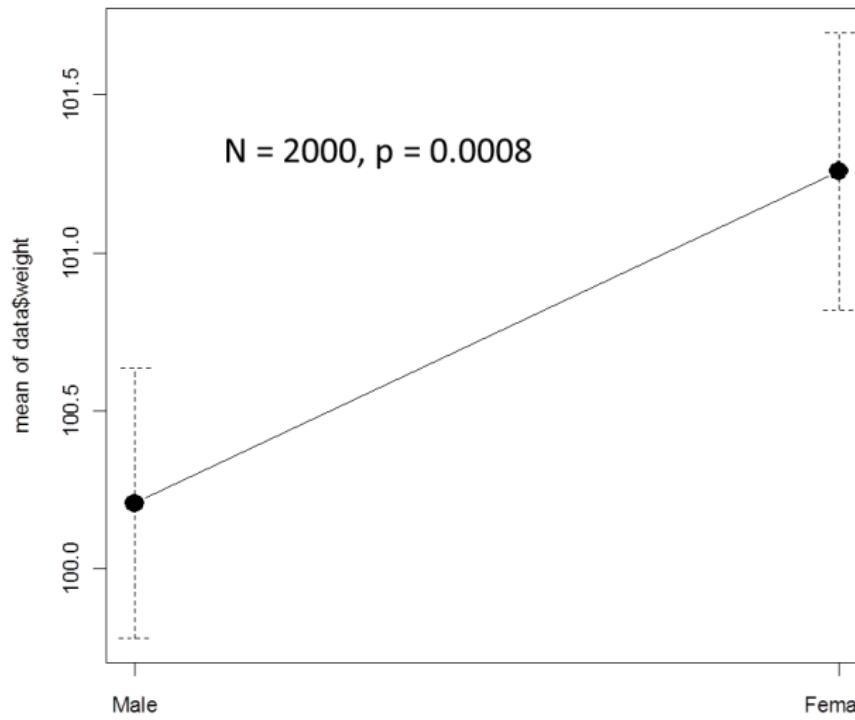
Real difference = 40 g



Statistically significant != biologically important

- With big sample size, we can find **highly significant but biologically unimportant** differences.

Real difference = 1 g



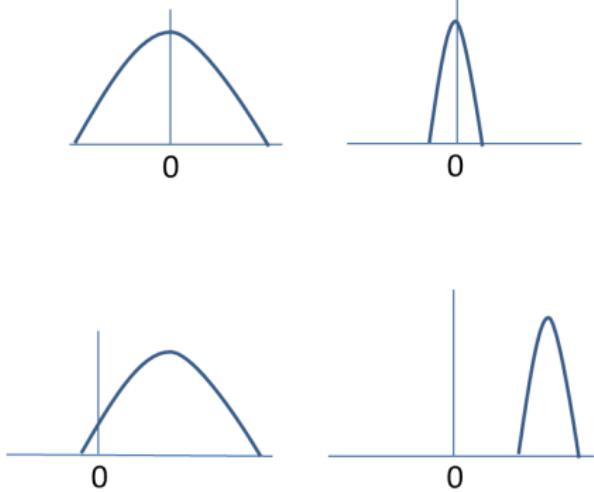
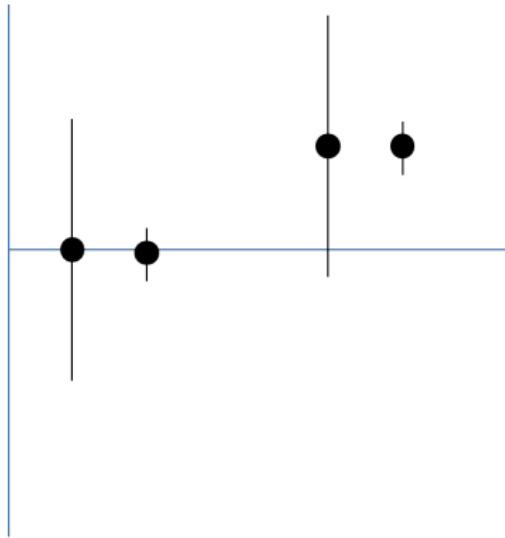
Statistically significant != biologically important

- ▶ Beyond significance, look at *effect sizes*.

Statistically significant != biologically important

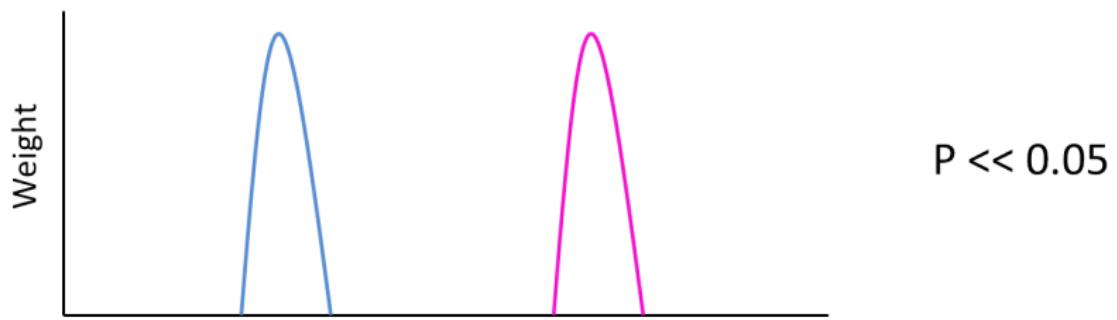
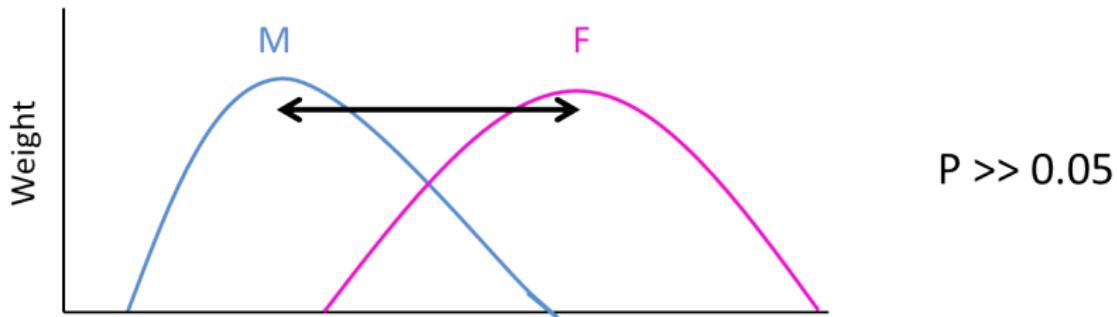
- ▶ Beyond significance, look at *effect sizes*.
- ▶ Suggested reading: *significantly misleading*

'Not significant' does NOT mean 'there is no effect'



- ▶ Absence of evidence \neq Evidence of absence

Failure to reject H₀ != H₀ is true



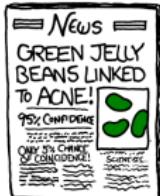
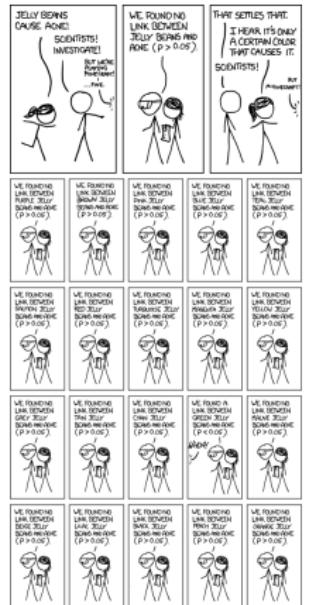
0.05 is an arbitrary threshold

The Difference Between “Significant” and “Not Significant” is not Itself Statistically Significant

Andrew GELMAN and Hal STERN

<http://dx.doi.org/10.1198/000313006X152649>

Multiple hypothesis testing



How to make your results significant

1. Test multiple variables, then report the ones that are significant.

How to make your results significant

1. Test multiple variables, then report the ones that are significant.
2. Artificially choose when to end your experiment.

How to make your results significant

1. Test multiple variables, then report the ones that are significant.
2. Artificially choose when to end your experiment.
3. Add covariates until effects are significant.

How to make your results significant

1. Test multiple variables, then report the ones that are significant.
2. Artificially choose when to end your experiment.
3. Add covariates until effects are significant.
4. Test different conditions (e.g. different levels of a factor) and report the ones you like.

How to make your results significant

1. Test multiple variables, then report the ones that are significant.
 2. Artificially choose when to end your experiment.
 3. Add covariates until effects are significant.
 4. Test different conditions (e.g. different levels of a factor) and report the ones you like.
- To read more: Simmons et al 2011

How to make your results significant

1. Test multiple variables, then report the ones that are significant.
 2. Artificially choose when to end your experiment.
 3. Add covariates until effects are significant.
 4. Test different conditions (e.g. different levels of a factor) and report the ones you like.
- ▶ To read more: Simmons et al 2011
 - ▶ NB: This is ironic!

The New Statistics

Aim for estimation of effects and their uncertainty.



General Article

The New Statistics: Why and How

Geoff Cumming

La Trobe University

Psychological Science
2014, Vol. 25(1) 7–29
© The Author(s) 2013
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797613504966
pss.sagepub.com



<http://dx.doi.org/10.1177/0956797613504966>

How many types of errors?

- ▶ **Type I:** incorrect rejection of null hypothesis.

How many types of errors?

- ▶ **Type I:** incorrect rejection of null hypothesis.
- ▶ **Type II:** failure to reject false null hypothesis.

How many types of errors?

- ▶ **Type I:** incorrect rejection of null hypothesis.
- ▶ **Type II:** failure to reject false null hypothesis.
- ▶ **Type S (Sign):** estimating effect in opposite direction.

How many types of errors?

- ▶ **Type I:** incorrect rejection of null hypothesis.
- ▶ **Type II:** failure to reject false null hypothesis.
- ▶ **Type S (Sign):** estimating effect in opposite direction.
- ▶ **Type M (Magnitude):** Misestimating magnitude of the effect (under or overestimating).

How many types of errors?

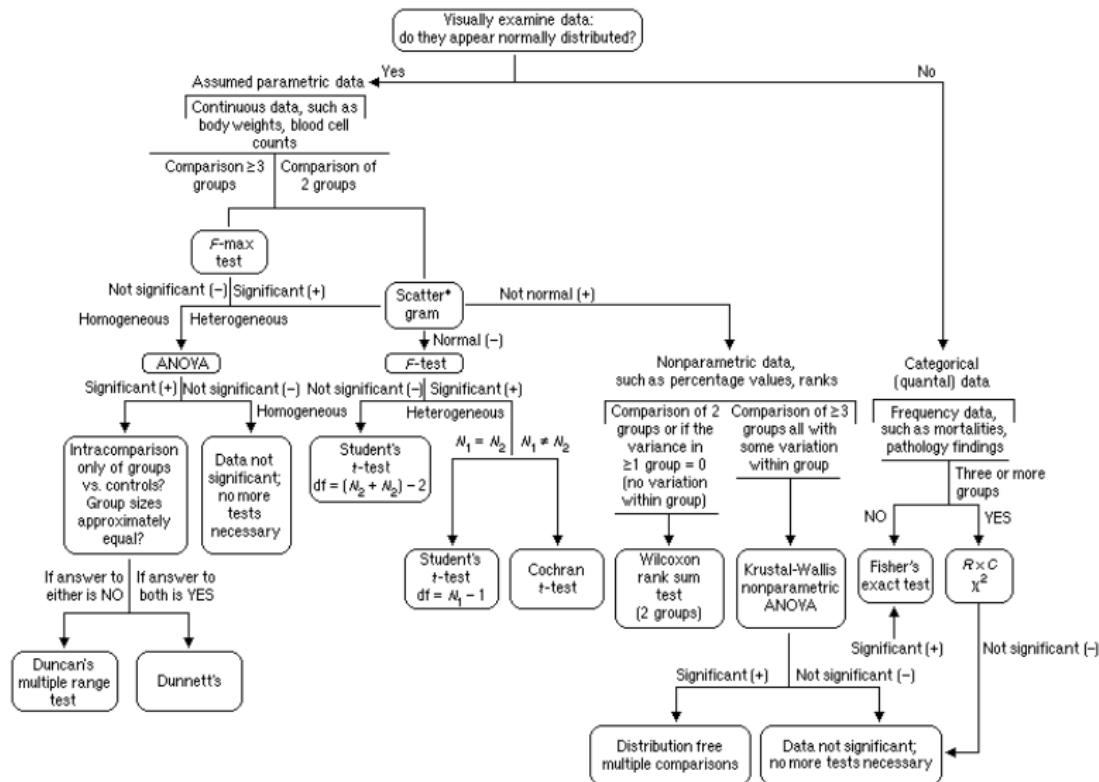
- ▶ **Type I:** incorrect rejection of null hypothesis.
- ▶ **Type II:** failure to reject false null hypothesis.
- ▶ **Type S (Sign):** estimating effect in opposite direction.
- ▶ **Type M (Magnitude):** Misestimating magnitude of the effect (under or overestimating).
- ▶ **Type III:** finding right answer to the wrong question!

Introduction to statistical modelling

The purpose of models is not to fit data but to sharpen thinking

Sam Karlin

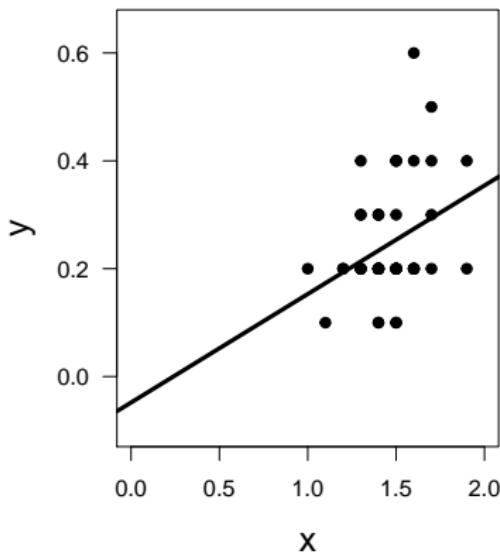
Modern statistics are easier than this



Our overarching regression framework

$$y_i = a + bx_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$



Data

y = response variable

x = predictor

Parameters

a = intercept

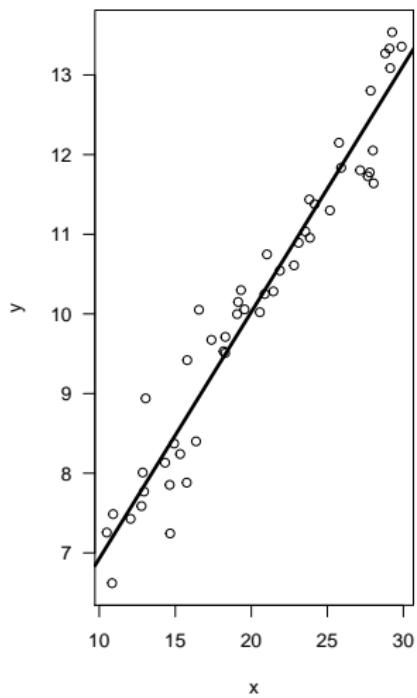
b = slope

σ = residual variation

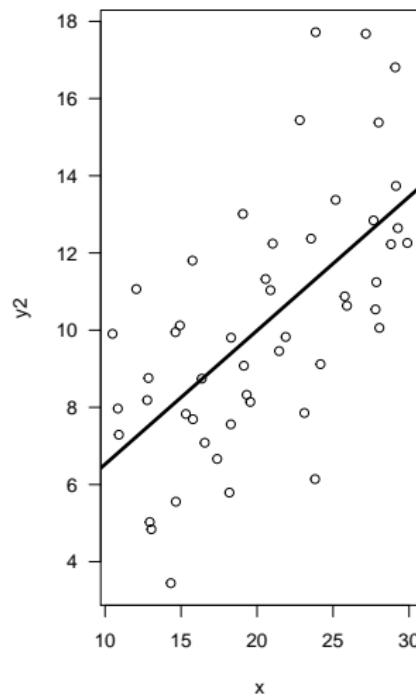
ε = residuals

Residual variation (error)

small



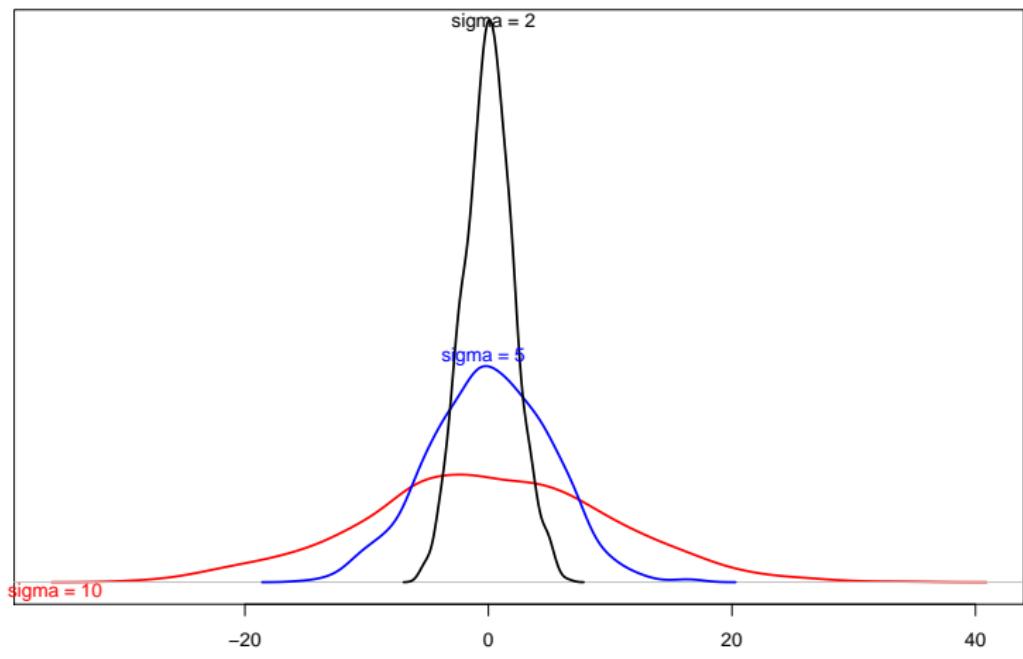
large



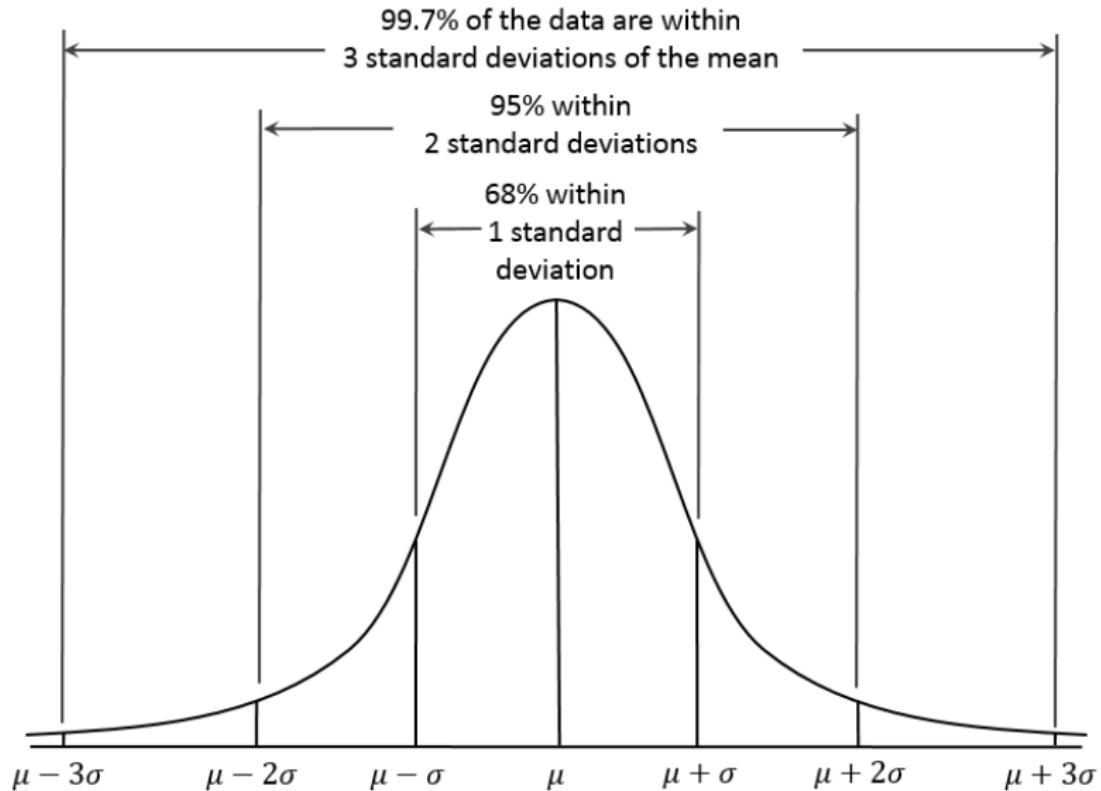
Residual variation

$$\varepsilon_i \sim N(0, \sigma^2)$$

Distribution of residuals



In a Normal distribution

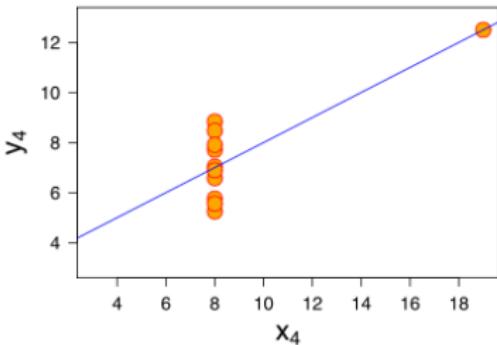
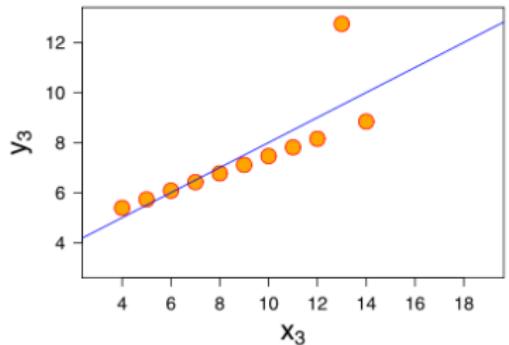
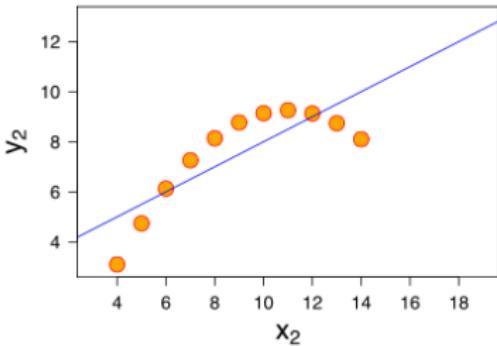
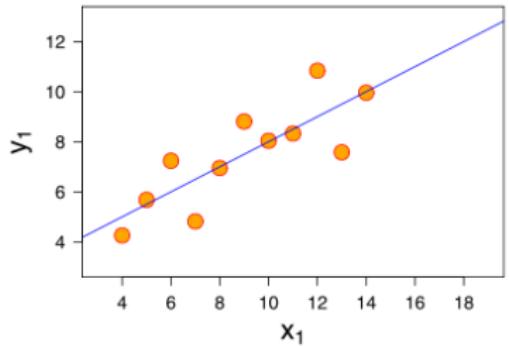


Let's do real data analysis

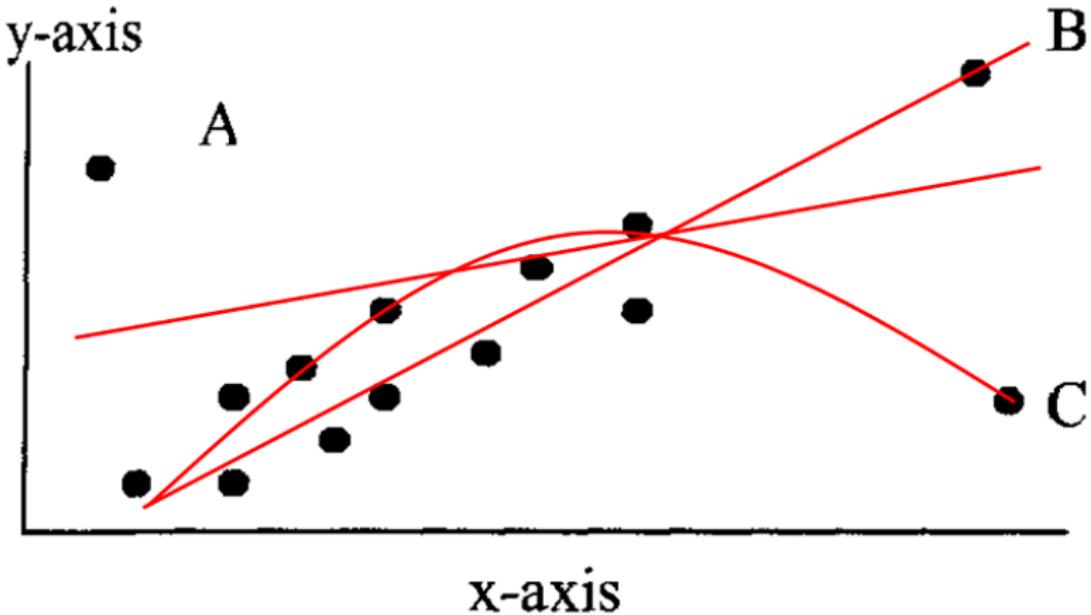
Q: What is the relationship between petal width and length in *Iris setosa*?

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

Always plot your data first!



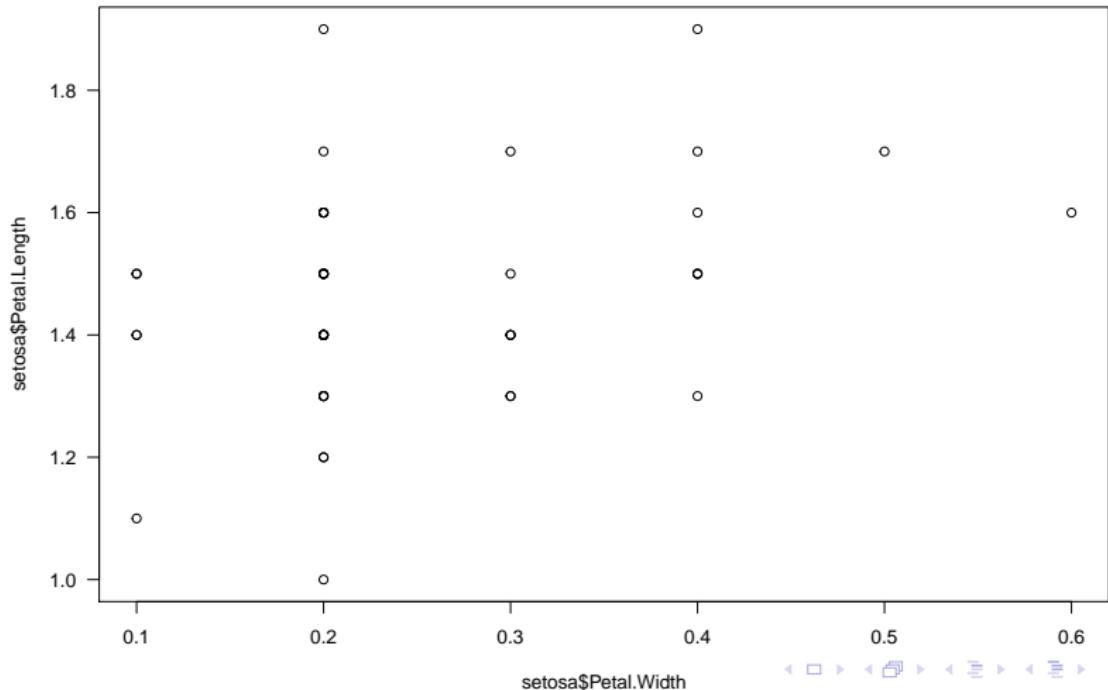
Outliers impact on regression



See <http://rpsychologist.com/d3/correlation/>

Scatterplot

```
plot(setosa$Petal.Width, setosa$Petal.Length, las = 1)
```



Now fit model

```
m1 <- lm(Petal.Length ~ Petal.Width, data = setosa)
```

What does this mean?

Call:

```
lm(formula = Petal.Length ~ Petal.Width, data = setosa)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.43686	-0.09151	-0.03686	0.09018	0.46314

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.32756	0.05996	22.141	<2e-16 ***
Petal.Width	0.54649	0.22439	2.435	0.0186 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

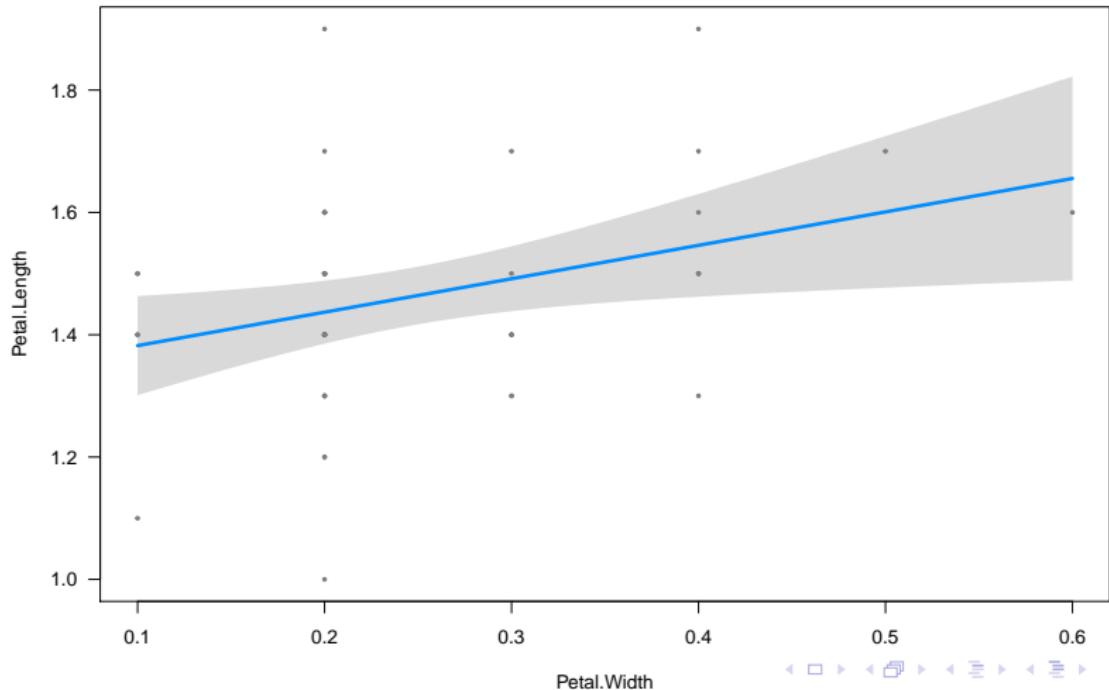
Residual standard error: 0.1655 on 48 degrees of freedom

Multiple R-squared: 0.11, Adjusted R-squared: 0.09144

F-statistic: 5.931 on 1 and 48 DF, p-value: 0.01864

Plot model (visreg)

```
visreg(m1)
```



Linear model assumptions

- ▶ Linearity (transformations, GAM...)

Linear model assumptions

- ▶ Linearity (transformations, GAM...)
- ▶ Residuals:

Linear model assumptions

- ▶ Linearity (transformations, GAM...)
- ▶ Residuals:
 - ▶ Independent

Linear model assumptions

- ▶ Linearity (transformations, GAM...)
- ▶ Residuals:
 - ▶ Independent
 - ▶ Equal variance

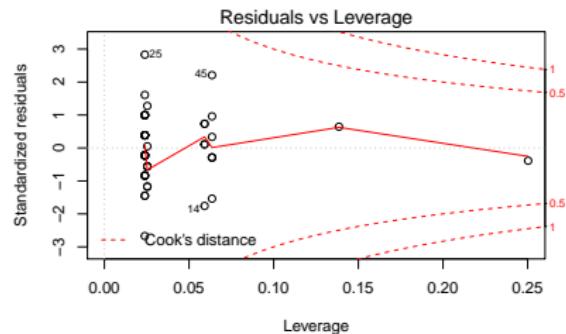
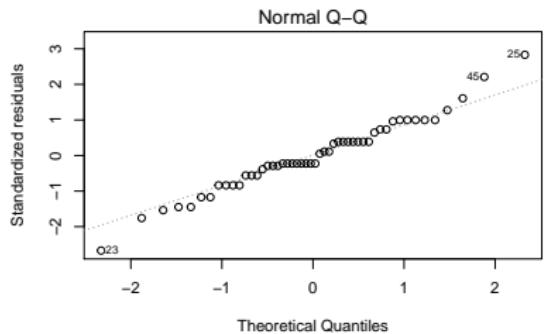
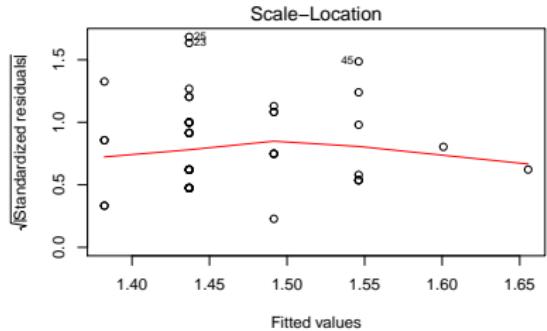
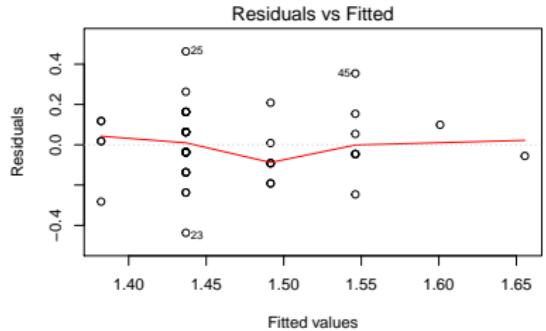
Linear model assumptions

- ▶ Linearity (transformations, GAM...)
- ▶ Residuals:
 - ▶ Independent
 - ▶ Equal variance
 - ▶ Normal

Linear model assumptions

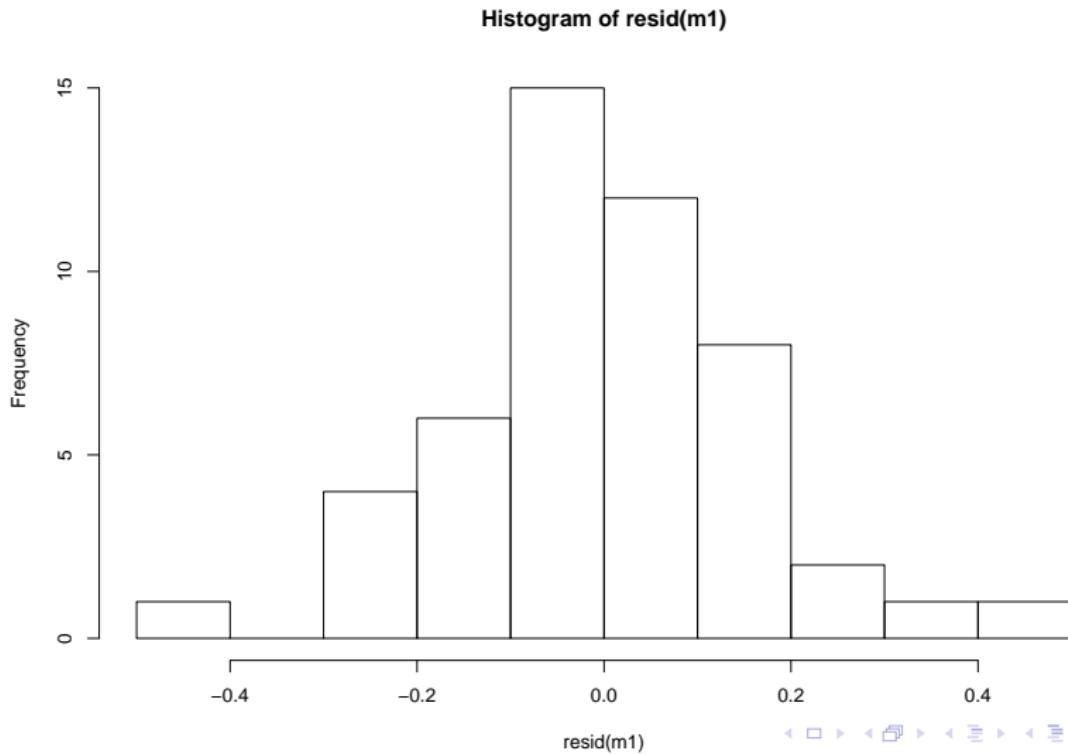
- ▶ Linearity (transformations, GAM...)
- ▶ Residuals:
 - ▶ Independent
 - ▶ Equal variance
 - ▶ Normal
- ▶ No measurement error in predictors

Model checking: residuals



Are residuals normal?

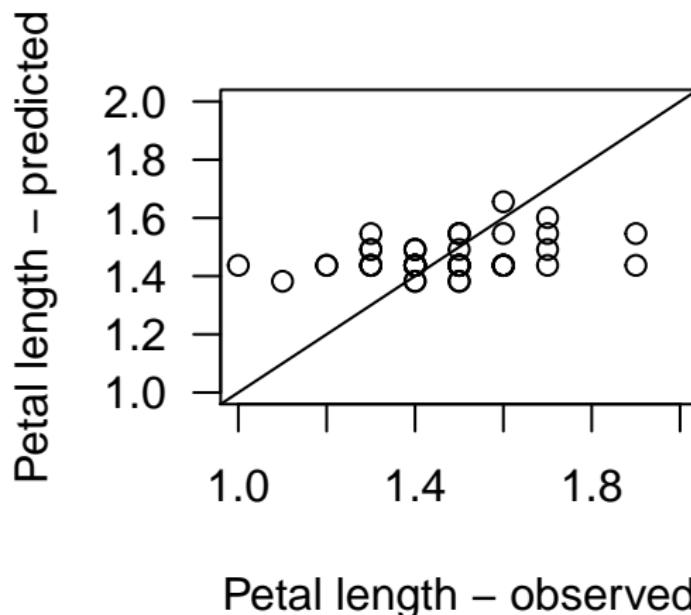
```
hist(resid(m1))
```



How good is the model in predicting petal length?

Observed vs Predicted values: use fitted.

```
plot(setosa$Petal.Length, fitted(m1), xlab = "Petal length - obs")
```



Using fitted model for prediction

Q: Expected petal length if width = 0.39?

Using fitted model for prediction

Q: Expected petal length if width = 0.39?

```
predict(m1, data.frame(Petal.Width = c(0.39)), se.fit = TRUE)
```

```
$fit
```

```
1
```

```
1.540695
```

```
$se.fit
```

```
[1] 0.03990149
```

```
$df
```

```
[1] 48
```

```
$residual.scale
```

```
[1] 0.1655341
```

Important functions

- ▶ plot

Important functions

- ▶ `plot`
- ▶ `summary`

Important functions

- ▶ `plot`
- ▶ `summary`
- ▶ `coef`

Important functions

- ▶ `plot`
- ▶ `summary`
- ▶ `coef`
- ▶ `confint`

Important functions

- ▶ `plot`
- ▶ `summary`
- ▶ `coef`
- ▶ `confint`
- ▶ `fitted`

Important functions

- ▶ `plot`
- ▶ `summary`
- ▶ `coef`
- ▶ `confint`
- ▶ `fitted`
- ▶ `resid`

Important functions

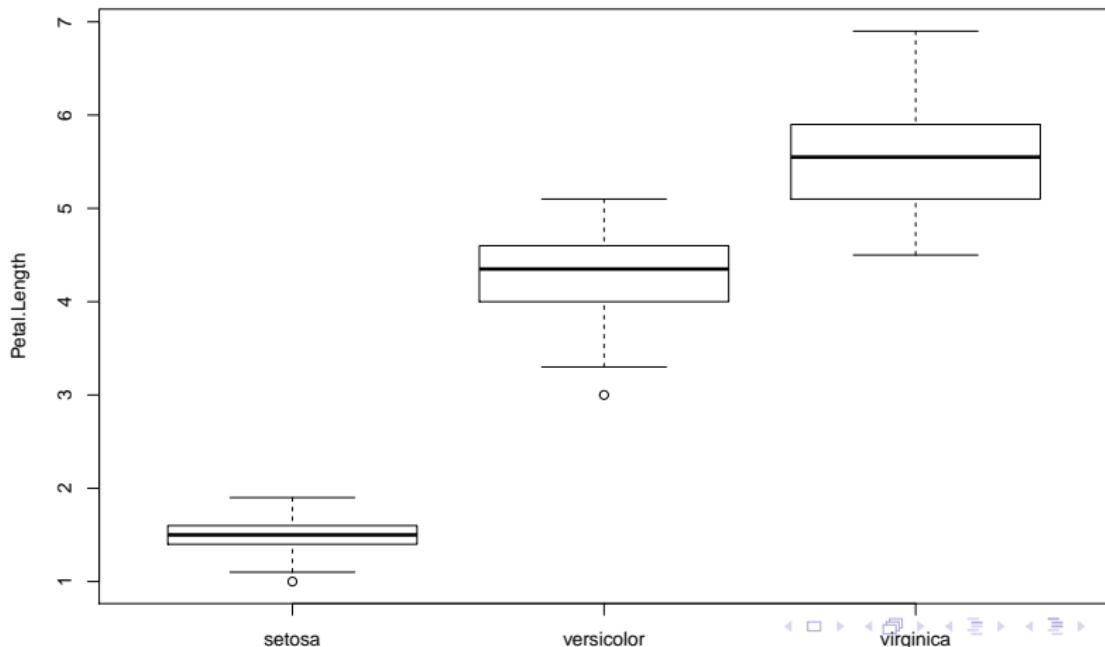
- ▶ plot
- ▶ summary
- ▶ coef
- ▶ confint
- ▶ fitted
- ▶ resid
- ▶ predict

Categorical predictors (factors)

Q: Does petal length vary among *Iris* species?

First, a plot:

```
plot(Petal.Length ~ Species, data = iris)
```



Linear model with categorical predictors

$$y_i = a + bx_i + \varepsilon_i$$

$$y_i = a + b_{versicolor} + c_{virginica} + \varepsilon_i$$

Model

```
m2 <- lm(Petal.Length ~ Species, data = iris)
```

Call:

```
lm(formula = Petal.Length ~ Species, data = iris)
```

Residuals:

Min 1Q Median 3Q Max
-1.260 -0.258 0.038 0.240 1.348

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.46200	0.06086	24.02	<2e-16	***
Speciesversicolor	2.79800	0.08607	32.51	<2e-16	***
Speciesvirginica	4.09000	0.08607	47.52	<2e-16	***

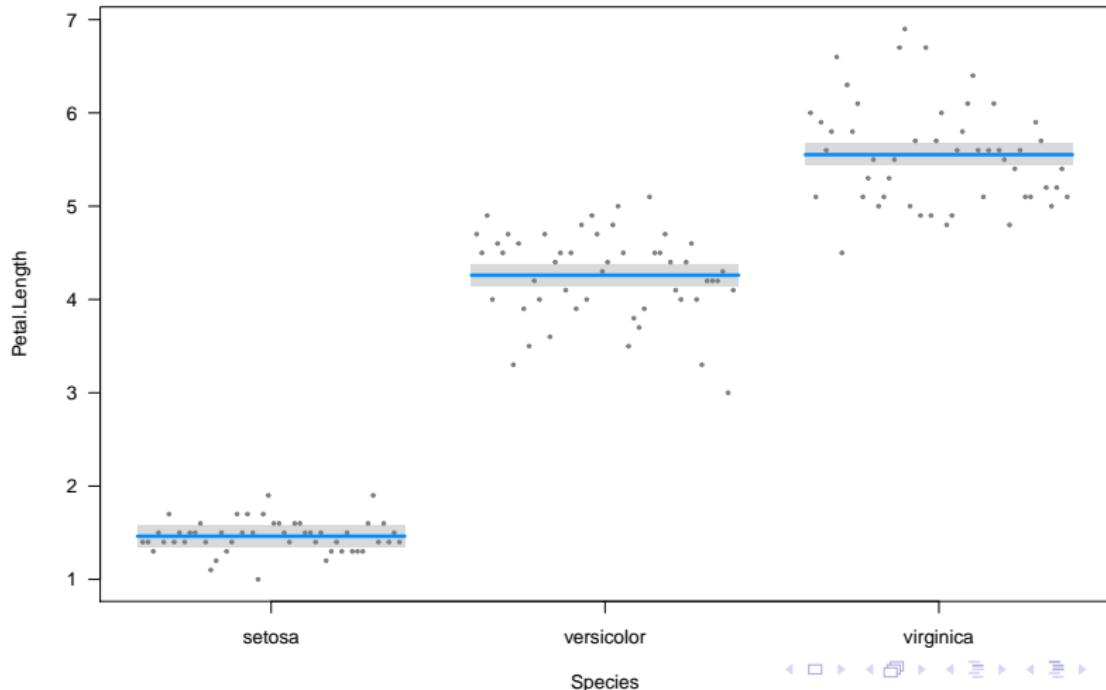
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4303 on 147 degrees of freedom

Multiple R-squared: 0.9414, Adjusted R-squared: 0.9406

Petal length differences across 3 *Iris* species

```
visreg(m2)
```



Are differences statistically significant?

Compare CIs

```
summary(allEffects(m2))
```

model: Petal.Length ~ Species

Species effect

Species

	setosa	versicolor	virginica
	1.462	4.260	5.552

Lower 95 Percent Confidence Limits

Species

	setosa	versicolor	virginica
	1.341729	4.139729	5.431729

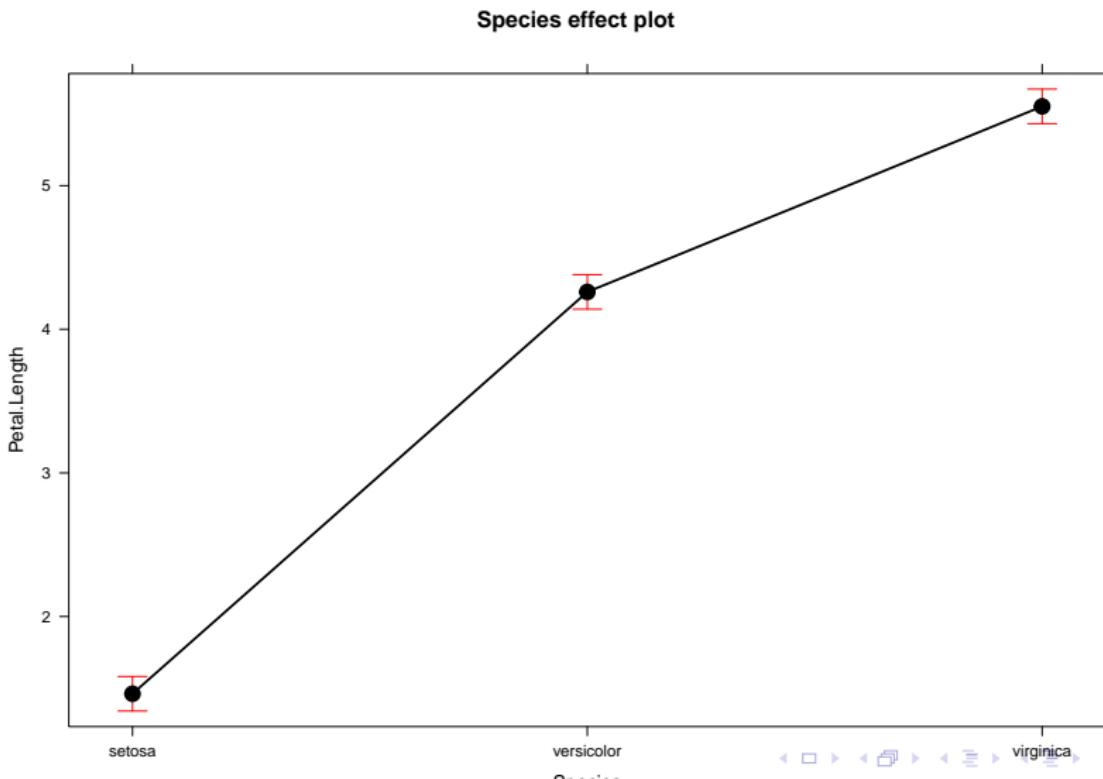
Upper 95 Percent Confidence Limits

Species

	setosa	versicolor	virginica
	1.582271	4.380271	5.672271

Plotting effects

```
plot(allEffects(m2))
```



Does height differ between local and foreign students?

Combining continuous and categorical predictors

Predicting *Iris* petal length according to species and petal width

$$y_i = a + bx_i + \varepsilon_i$$

$$y_i = a + b_{versicolor} + c_{virginica} + \varepsilon_i$$

$$y_i = a + b_{versicolor} + c_{virginica} + d \cdot PetalWidth_i + \varepsilon_i$$

Predicting *Iris* petal length according to species and petal width

Call:

```
lm(formula = Petal.Length ~ Species + Petal.Width, data = iris)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.02977	-0.22241	-0.01514	0.18180	1.17449

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.21140	0.06524	18.568	< 2e-16 ***
Speciesversicolor	1.69779	0.18095	9.383	< 2e-16 ***
Speciesvirginica	2.27669	0.28132	8.093	2.08e-13 ***
Petal.Width	1.01871	0.15224	6.691	4.41e-10 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3777 on 146 degrees of freedom

Multiple R-squared: 0.9551, Adjusted R-squared: 0.9542

Generalised Linear Models (GLMs)

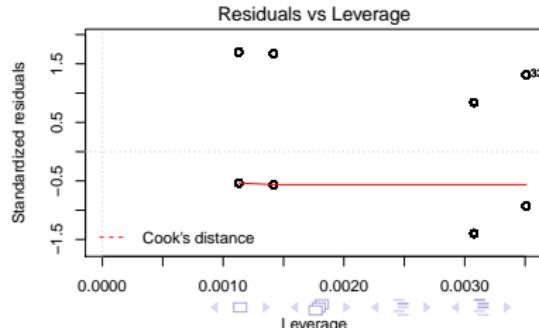
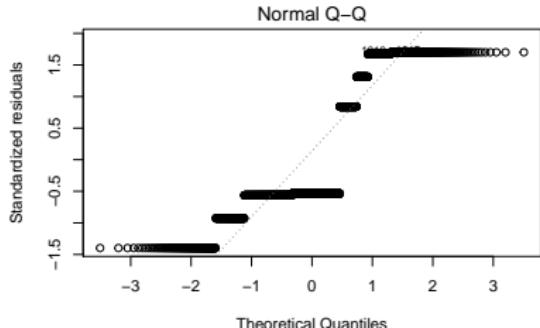
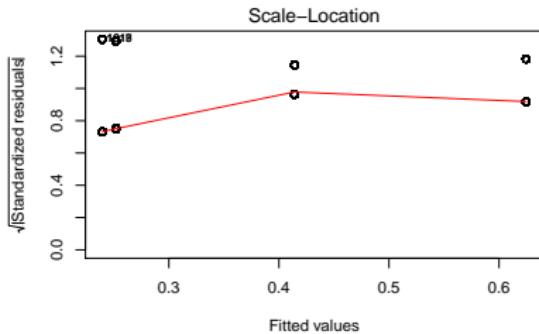
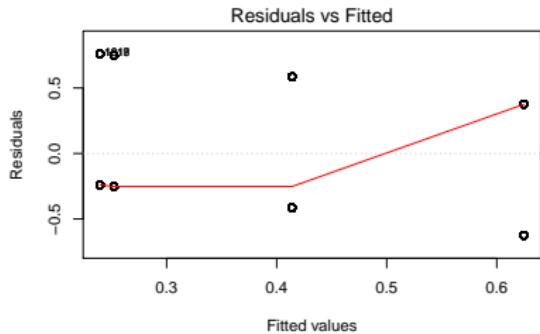
Q: Survival of passengers on the Titanic ~ Class

Read titanic_long.csv dataset.

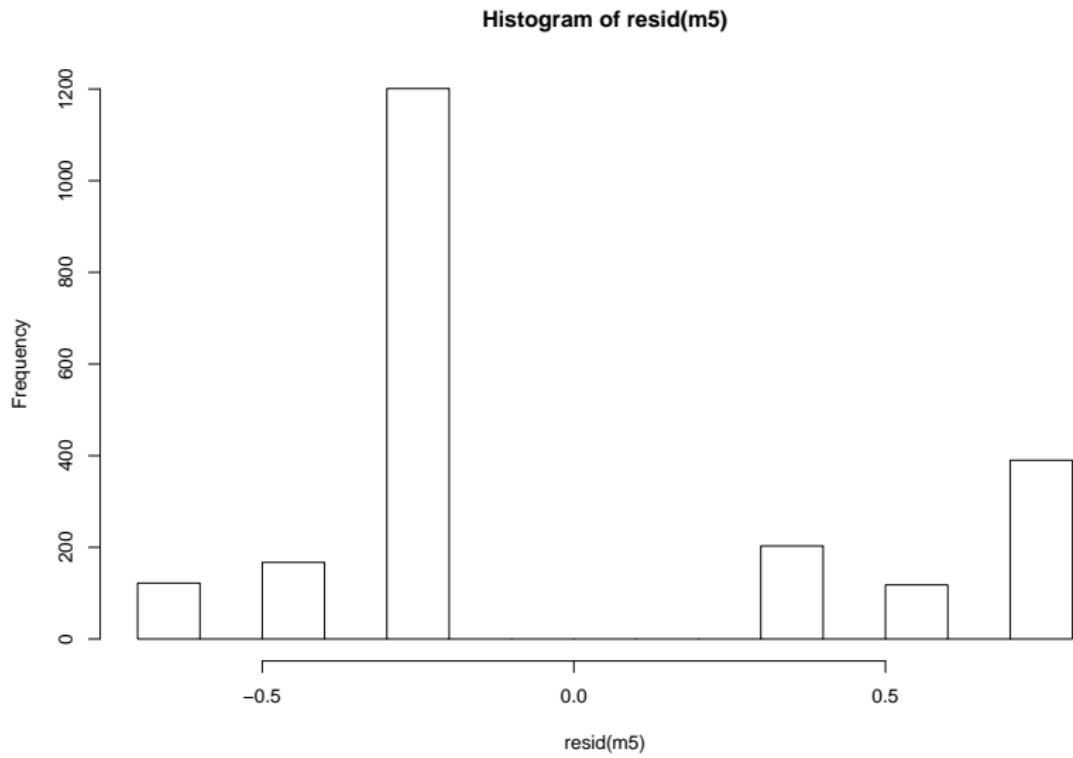
	class	age	sex	survived
1	first	adult	male	1
2	first	adult	male	1
3	first	adult	male	1
4	first	adult	male	1
5	first	adult	male	1
6	first	adult	male	1

Let's fit linear model:

```
m5 <- lm(survived ~ class, data = titanic)
```



Weird residuals!



What if your residuals are clearly non-normal? | And variance not constant (heteroscedasticity)?

- ▶ Binary variables (0/1)

What if your residuals are clearly non-normal? | And variance not constant (heteroscedasticity)?

- ▶ Binary variables (0/1)
- ▶ Counts (0, 1, 2, 3, ...)

Generalised Linear Models

1. **Response variable** - distribution family

Generalised Linear Models

1. Response variable - distribution family

- ▶ Bernouilli - Binomial

Generalised Linear Models

1. Response variable - distribution family

- ▶ Bernouilli - Binomial
- ▶ Poisson

Generalised Linear Models

1. Response variable - distribution family

- ▶ Bernouilli - Binomial
- ▶ Poisson
- ▶ Gamma

Generalised Linear Models

1. Response variable - distribution family

- ▶ Bernouilli - Binomial
- ▶ Poisson
- ▶ Gamma
- ▶ etc

Generalised Linear Models

1. Response variable - distribution family

- ▶ Bernouilli - Binomial
- ▶ Poisson
- ▶ Gamma
- ▶ etc

2. Predictors (continuous or categorical)

Generalised Linear Models

1. Response variable - distribution family

- ▶ Bernouilli - Binomial
- ▶ Poisson
- ▶ Gamma
- ▶ etc

2. Predictors (continuous or categorical)

3. Link function

Generalised Linear Models

1. Response variable - distribution family

- ▶ Bernouilli - Binomial
- ▶ Poisson
- ▶ Gamma
- ▶ etc

2. Predictors (continuous or categorical)

3. Link function

- ▶ Gaussian: identity

Generalised Linear Models

1. Response variable - distribution family

- ▶ Bernouilli - Binomial
- ▶ Poisson
- ▶ Gamma
- ▶ etc

2. Predictors (continuous or categorical)

3. Link function

- ▶ Gaussian: identity
- ▶ Binomial: logit, probit

Generalised Linear Models

1. Response variable - distribution family

- ▶ Bernouilli - Binomial
- ▶ Poisson
- ▶ Gamma
- ▶ etc

2. Predictors (continuous or categorical)

3. Link function

- ▶ Gaussian: identity
- ▶ Binomial: logit, probit
- ▶ Poisson: log...

Generalised Linear Models

1. Response variable - distribution family

- ▶ Bernouilli - Binomial
- ▶ Poisson
- ▶ Gamma
- ▶ etc

2. Predictors (continuous or categorical)

3. Link function

- ▶ Gaussian: identity
- ▶ Binomial: logit, probit
- ▶ Poisson: log...
- ▶ See `family`.

Bernoulli - Binomial distribution (Logistic regression)

- ▶ Response variable: Yes/No (e.g. survival, sex, presence/absence)

$$\text{logit}(p) = \ln \left(\frac{p}{1-p} \right)$$

Then

$$Pr(\text{alive}) = a + bx$$

$$\text{logit}(Pr(\text{alive})) = a + bx$$

$$Pr(\text{alive}) = \text{invlogit}(a + bx) = \frac{e^{a+bx}}{1 + e^{a+bx}}$$

Bernoulli - Binomial distribution (Logistic regression)

- ▶ Response variable: Yes/No (e.g. survival, sex, presence/absence)
- ▶ Link function: logit (others possible, see `family`).

$$\text{logit}(p) = \ln \left(\frac{p}{1-p} \right)$$

Then

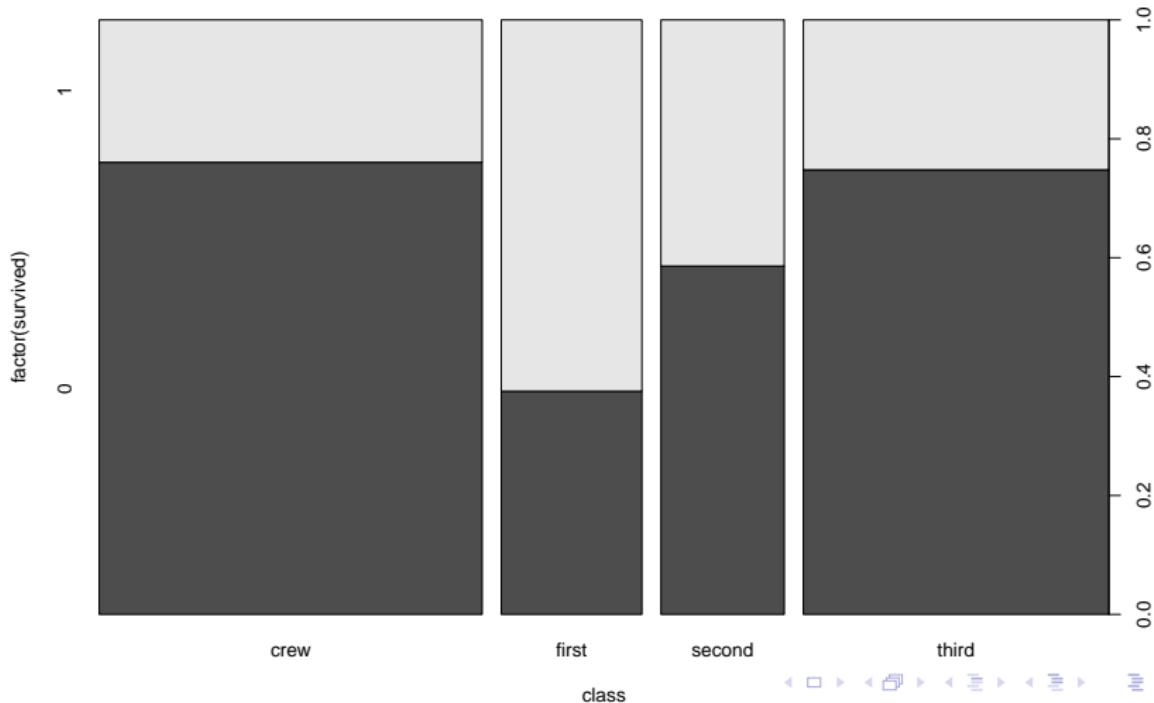
$$Pr(\text{alive}) = a + bx$$

$$\text{logit}(Pr(\text{alive})) = a + bx$$

$$Pr(\text{alive}) = \text{invlogit}(a + bx) = \frac{e^{a+bx}}{1 + e^{a+bx}}$$

Back to survival of Titanic passengers

```
plot(factor(survived) ~ class, data = titanic)
```



Fitting GLMs in R: `glm`

```
tit.glm <- glm(survived ~ class, data=titanic, family=binomial)
```

Call:

```
glm(formula = survived ~ class, family = binomial, data = titani
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.3999	-0.7623	-0.7401	0.9702	1.6906

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.15516	0.07876	-14.667	< 2e-16 ***
classfirst	1.66434	0.13902	11.972	< 2e-16 ***
classsecond	0.80785	0.14375	5.620	1.91e-08 ***
classthird	0.06785	0.11711	0.579	0.562

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)



Interpreting logistic regression output

Parameter estimates (logit-scale)

(Intercept)	classfirst	classesecond	classthird
-1.15515905	1.66434399	0.80784987	0.06784632

We need to back-transform: apply *inverse logit*
Crew probability of survival:

```
plogis(coef(tit.glm)[1])
```

(Intercept)
0.239548

Looking at the data, the proportion of crew who survived is

```
[1] 0.239548
```

Q: Probability of survival for 1st class passengers?

```
plogis(coef(tit.glm)[1] + coef(tit.glm)[2])
```

```
(Intercept)  
0.6246154
```

Needs to add intercept (baseline) to the parameter estimate. Again this value matches the data:

```
sum(titanic$survived[titanic$class == "first"]) /  
nrow(titanic[titanic$class == "first", ])
```

```
[1] 0.6246154
```

Model interpretation using effects package

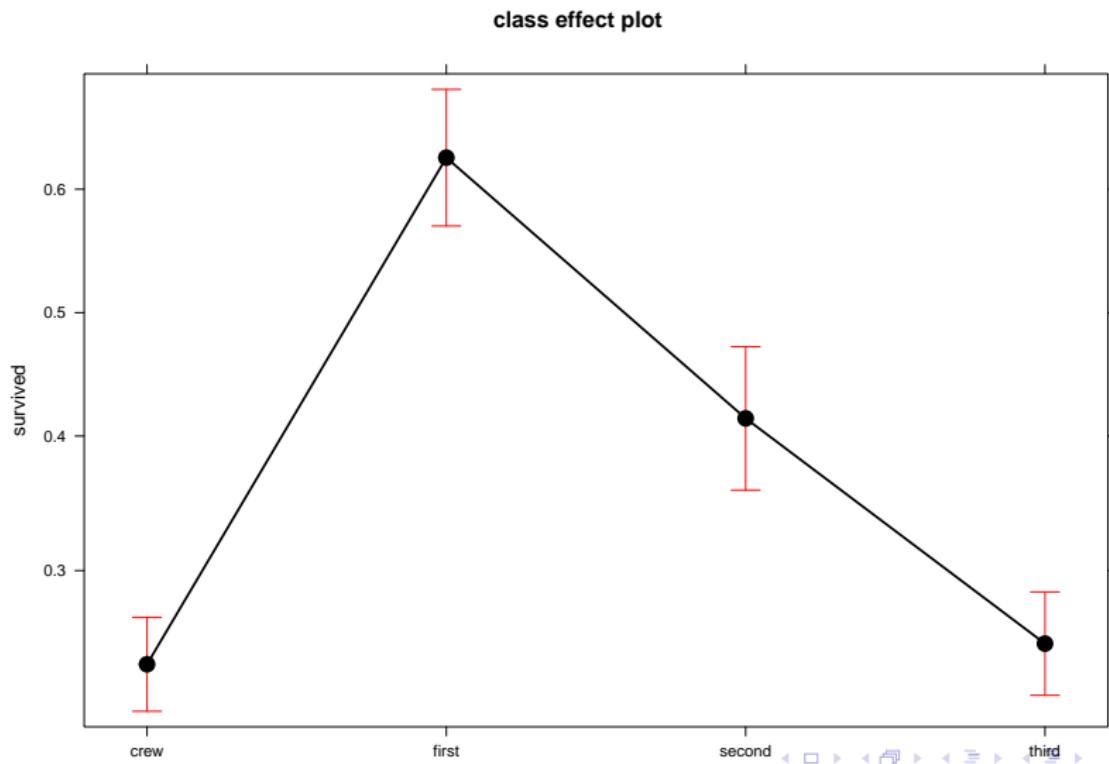
```
library(effects)
allEffects(tit.glm)

model: survived ~ class

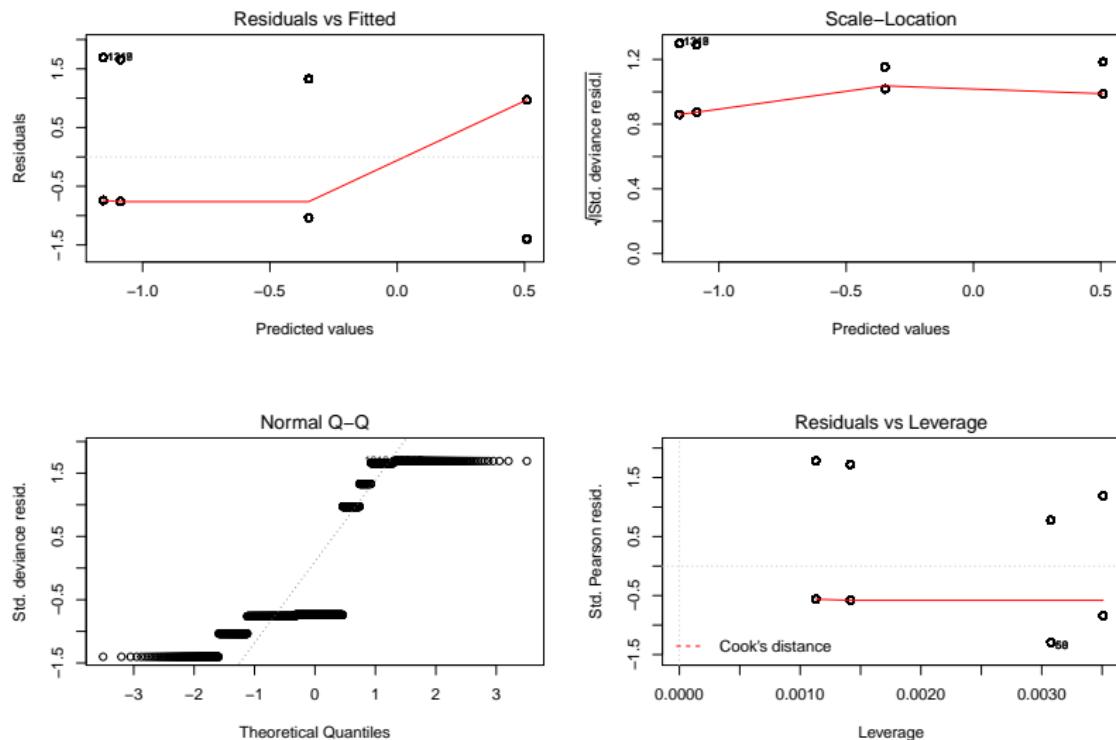
class effect
class
    crew      first     second     third
0.2395480 0.6246154 0.4140351 0.2521246
```

Effects plot

```
plot(allEffects(tit.glm))
```



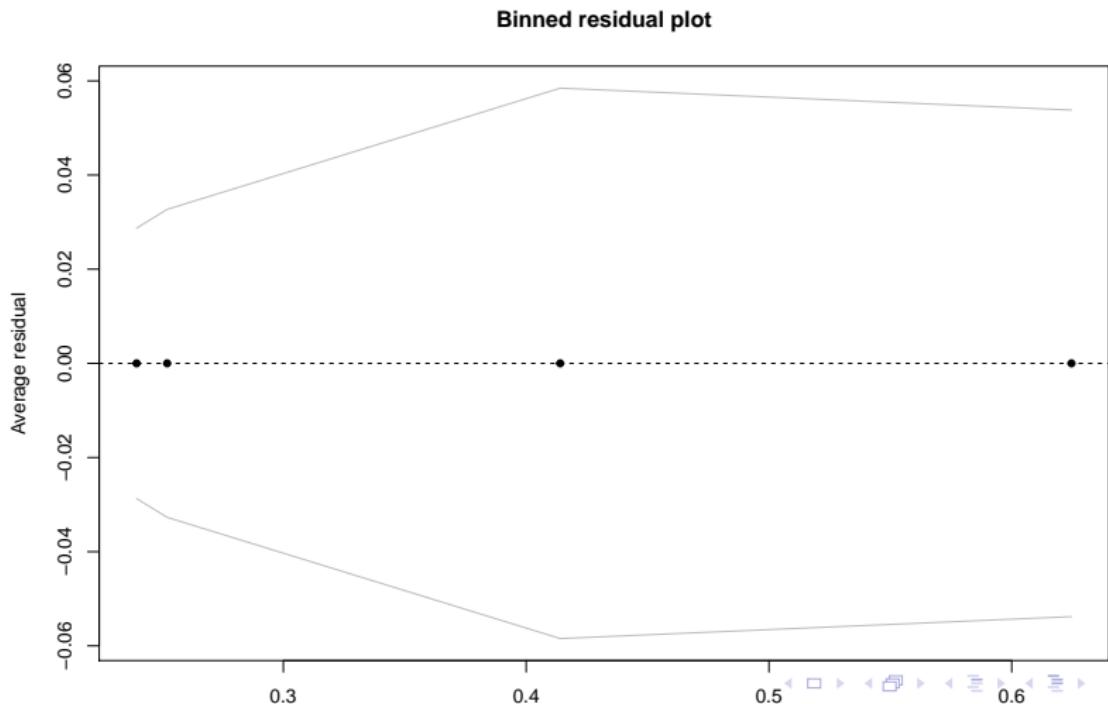
Logistic regression: model checking



Not very useful.

Binned residual plots for logistic regression

```
predvals <- predict(tit.glm, type="response")
arm::binnedplot(predvals, titanic$survived - predvals)
```



Recapitulating

1. Import data: `read.table` or `read.csv`

Recapitulating

1. Import data: `read.table` or `read.csv`
2. Check data: `summary`

Recapitulating

1. Import data: `read.table` or `read.csv`
2. Check data: `summary`
3. Plot data: `plot`

Recapitulating

1. Import data: `read.table` or `read.csv`
2. Check data: `summary`
3. Plot data: `plot`
4. Fit model: `glm`. Don't forget to specify `family`!

Recapitulating

1. Import data: `read.table` or `read.csv`
2. Check data: `summary`
3. Plot data: `plot`
4. Fit model: `glm`. Don't forget to specify `family!`
5. Examine models: `summary`

Recapitulating

1. Import data: `read.table` or `read.csv`
2. Check data: `summary`
3. Plot data: `plot`
4. Fit model: `glm`. Don't forget to specify `family!`
5. Examine models: `summary`
6. Use `plogis` to apply back-transformation (*invlogit*) to parameter estimates (`coef`). Alternatively, use `allEffects` from `effects` package.

Recapitulating

1. Import data: `read.table` or `read.csv`
2. Check data: `summary`
3. Plot data: `plot`
4. Fit model: `glm`. Don't forget to specify `family!`
5. Examine models: `summary`
6. Use `plogis` to apply back-transformation (*invlogit*) to parameter estimates (`coef`). Alternatively, use `allEffects` from `effects` package.
7. Plot model: `plot(allEffects(model))`. Or use `visreg`.

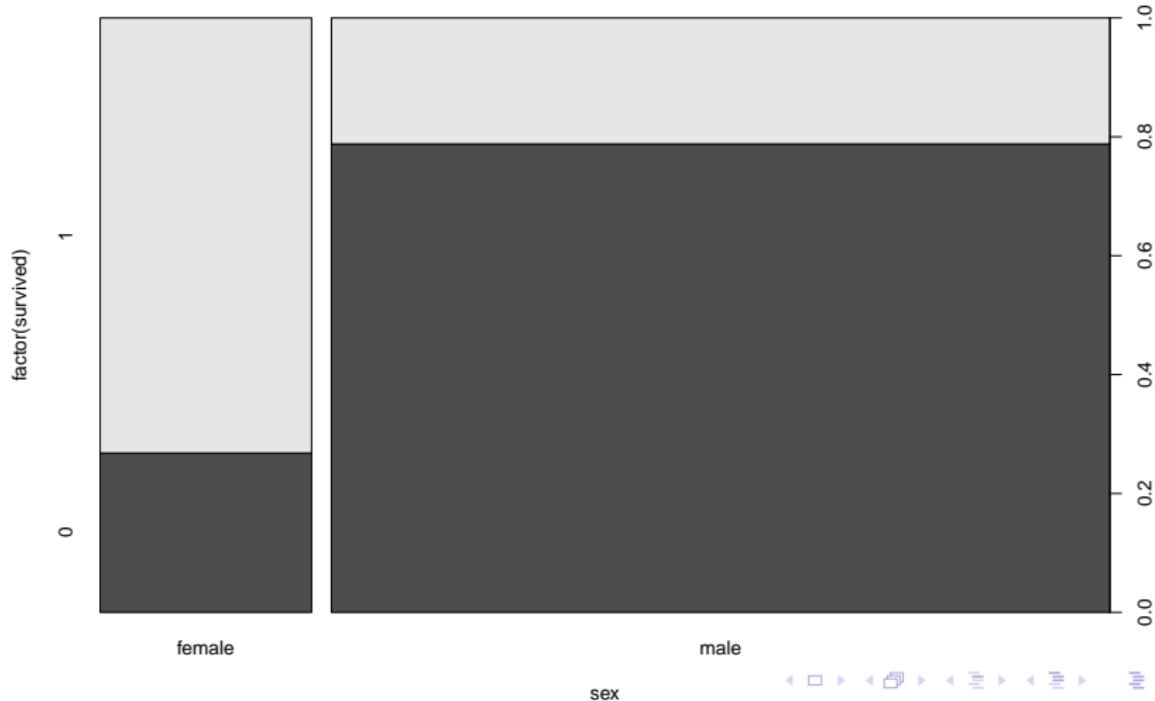
Recapitulating

1. Import data: `read.table` or `read.csv`
2. Check data: `summary`
3. Plot data: `plot`
4. Fit model: `glm`. Don't forget to specify `family!`
5. Examine models: `summary`
6. Use `plogis` to apply back-transformation (*invlogit*) to parameter estimates (`coef`). Alternatively, use `allEffects` from `effects` package.
7. Plot model: `plot(allEffects(model))`. Or use `visreg`.
8. Examine residuals: `binnedplot` from package `arm`. Use `predict` to obtain predicted values for each obs.

Q: Did men have higher survival than women?

Plot first

```
plot(factor(survived) ~ sex, data = titanic)
```



Fit model

```
tit.sex <- glm(survived ~ sex, data = titanic, family = binomial)
```

Call:

```
glm(formula = survived ~ sex, family = binomial, data = titanic)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.6226	-0.6903	-0.6903	0.7901	1.7613

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.0044	0.1041	9.645	<2e-16 ***
sexmale	-2.3172	0.1196	-19.376	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2769.5 on 2200 degrees of freedom ▶ 🔍 ↻

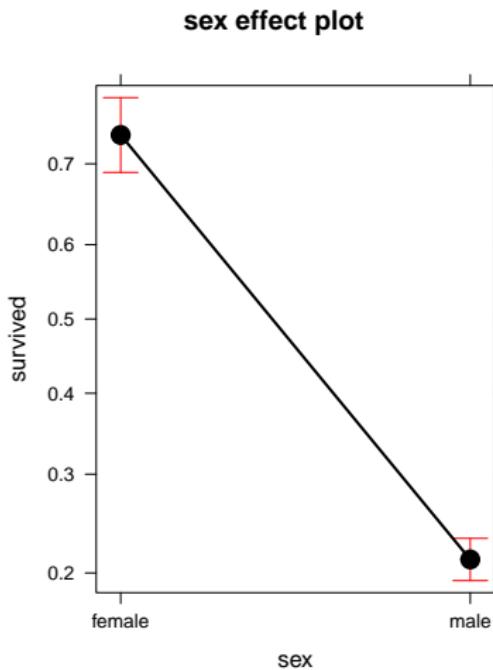
Effects

```
model: survived ~ sex
```

sex effect

sex

female	male
0.7319149	0.2120162



Q: Did women have higher survival because they travelled more in first class?

Let's look at the data

tapply

```
tapply(titanic$survived, list(titanic$class, titanic$sex), sum)
```

	female	male
crew	20	192
first	141	62
second	93	25
third	90	88

Mmmm...

Fit model with both factors (interactions)

```
tit.sex.class <- glm(survived ~ class * sex, data = titanic, family = binomial)

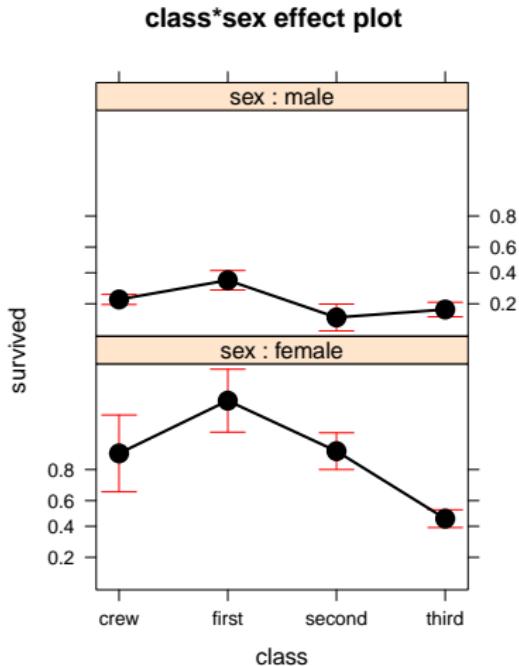
glm(formula = survived ~ class * sex, family = binomial, data = titanic)
    coef.est  coef.se
(Intercept)      1.90     0.62
classfirst       1.67     0.80
classsecond      0.07     0.69
classthird      -2.06     0.64
sexmale          -3.15     0.62
classfirst:sexmale -1.06     0.82
classsecond:sexmale -0.64     0.72
classthird:sexmale  1.74     0.65
---
n = 2201, k = 8
residual deviance = 2163.7, null deviance = 2769.5 (difference = 605.8)
```

Effects

```
model: survived ~ class * sex
```

class*sex effect

class	sex	
	female	male
crew	0.8695652	0.2227378
first	0.9724138	0.3444444
second	0.8773585	0.1396648
third	0.4591837	0.1725490



So, women had higher probability of survival than men, even within the same class.

Logistic regression for proportion data

Read Titanic data in different format

Read Titanic_prop.csv data.

X	Class	Sex	Age	No
Min. : 1.00	1st : 4	Female: 8	Adult: 8	Min. : 0.00
1st Qu.: 4.75	2nd : 4	Male : 8	Child: 8	1st Qu.: 0.00
Median : 8.50	3rd : 4			Median : 8.50
Mean : 8.50	Crew: 4			Mean : 93.12
3rd Qu.: 12.25				3rd Qu.: 96.25
Max. : 16.00				Max. : 670.00
Yes				
Min. : 0.00				
1st Qu.: 9.50				
Median : 14.00				
Mean : 44.44				
3rd Qu.: 75.25				
Max. : 192.00				

These are the same data, but summarized (see Freq variable).

Use cbind(n.success, n.failures) as response

```
prop.glm <- glm(cbind(Yes, No) ~ Class, data = tit.prop, family = binomial)
```

Call:

```
glm(formula = cbind(Yes, No) ~ Class, family = binomial, data = tit.prop)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-9.6404	-0.2915	1.5698	5.0366	10.1516

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.5092	0.1146	4.445	8.79e-06 ***
Class2nd	-0.8565	0.1661	-5.157	2.51e-07 ***
Class3rd	-1.5965	0.1436	-11.114	< 2e-16 ***
ClassCrew	-1.6643	0.1390	-11.972	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)



Effects

```
model: cbind(Yes, No) ~ Class  
  
Class effect  
Class  
    1st      2nd      3rd      Crew  
0.6246154 0.4140351 0.2521246 0.2395480
```

Compare with former model based on raw data:

```
model: survived ~ class  
  
class effect  
class  
    crew      first     second     third  
0.2395480 0.6246154 0.4140351 0.2521246
```

Same results!

Logistic regression with continuous predictors

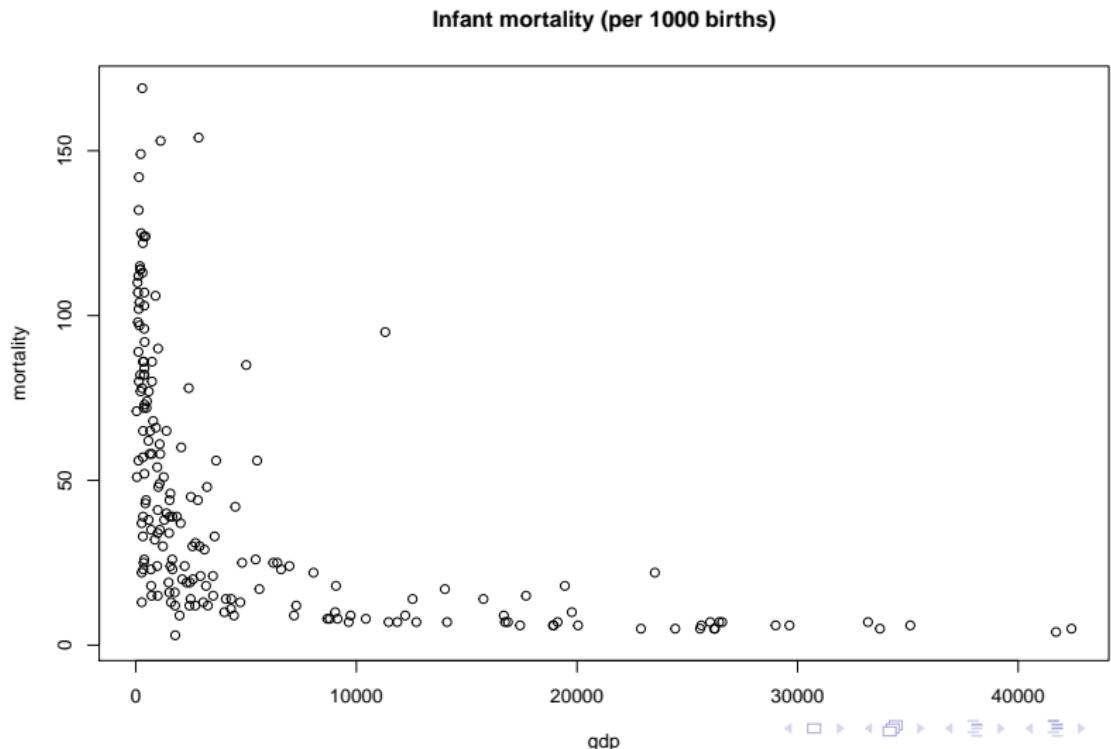
Example dataset: GDP and infant mortality

Read UN_GDP_infantmortality.csv.

	country	mortality	gdp
Afghanistan	: 1	Min. : 2.00	Min. : 36
Albania	: 1	1st Qu.: 12.00	1st Qu.: 442
Algeria	: 1	Median : 30.00	Median : 1779
American.Samoa	: 1	Mean : 43.48	Mean : 6262
Andorra	: 1	3rd Qu.: 66.00	3rd Qu.: 7272
Angola	: 1	Max. : 169.00	Max. : 42416
(Other)	: 201	NA's : 6	NA's : 10

EDA

```
plot(mortality ~ gdp, data = gdp, main = "Infant mortality (per
```



Fit model

```
gdp.glm <- glm(cbind(mortality, 1000 - mortality) ~ gdp,  
                 data = gdp, family = binomial)
```

Call:

```
glm(formula = cbind(mortality, 1000 - mortality) ~ gdp, family =  
     data = gdp)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-9.2230	-3.5163	-0.5697	2.4284	13.5849

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.657e+00	1.311e-02	-202.76	<2e-16 ***
gdp	-1.279e-04	3.458e-06	-36.98	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)



Effects

```
allEffects(gdp.glm)
```

```
model: cbind(mortality, 1000 - mortality) ~ gdp
```

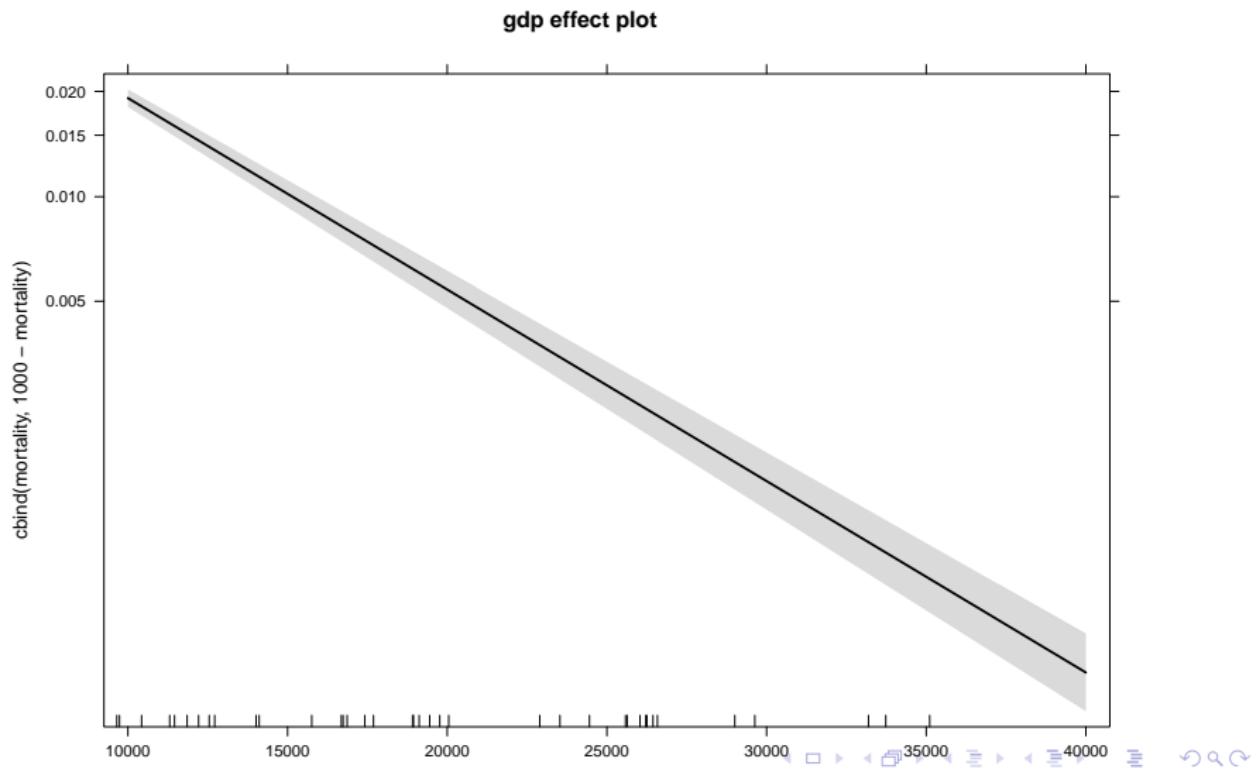
gdp effect

gdp

	10000	20000	30000	40000
0.0191438829	0.0054028095	0.0015096074	0.0004206154	

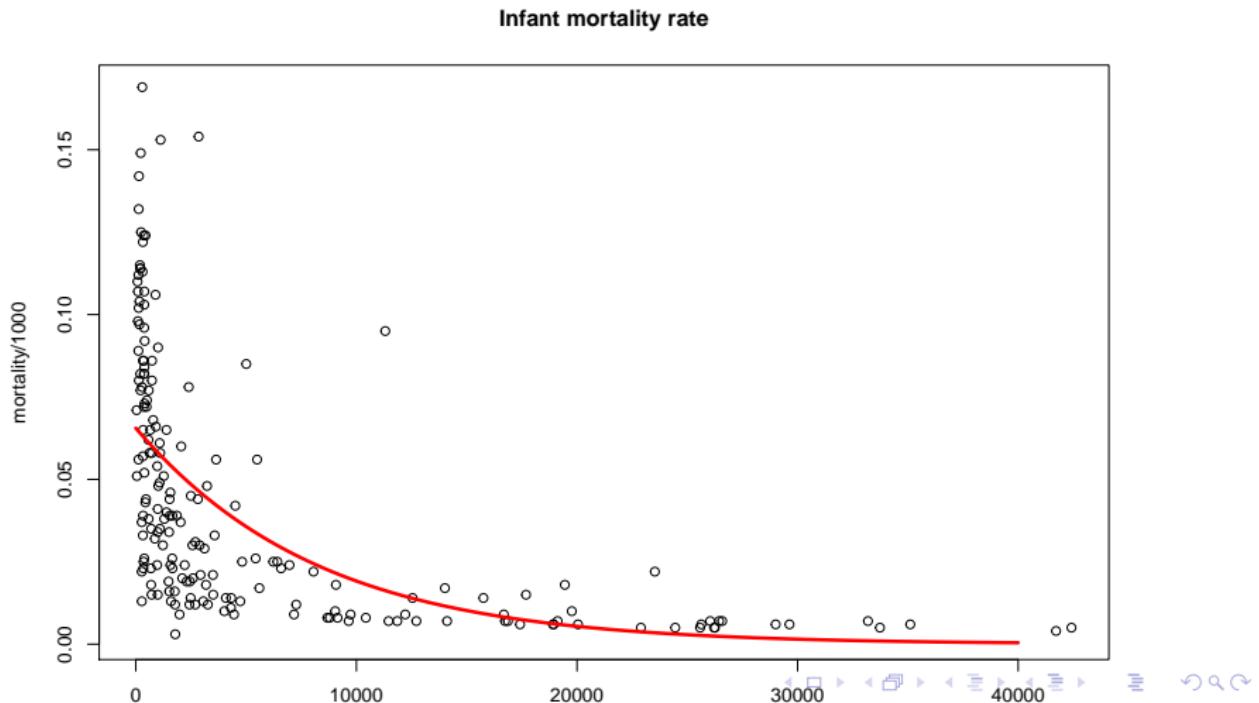
Effects plot

```
plot(allEffects(gdp.glm))
```



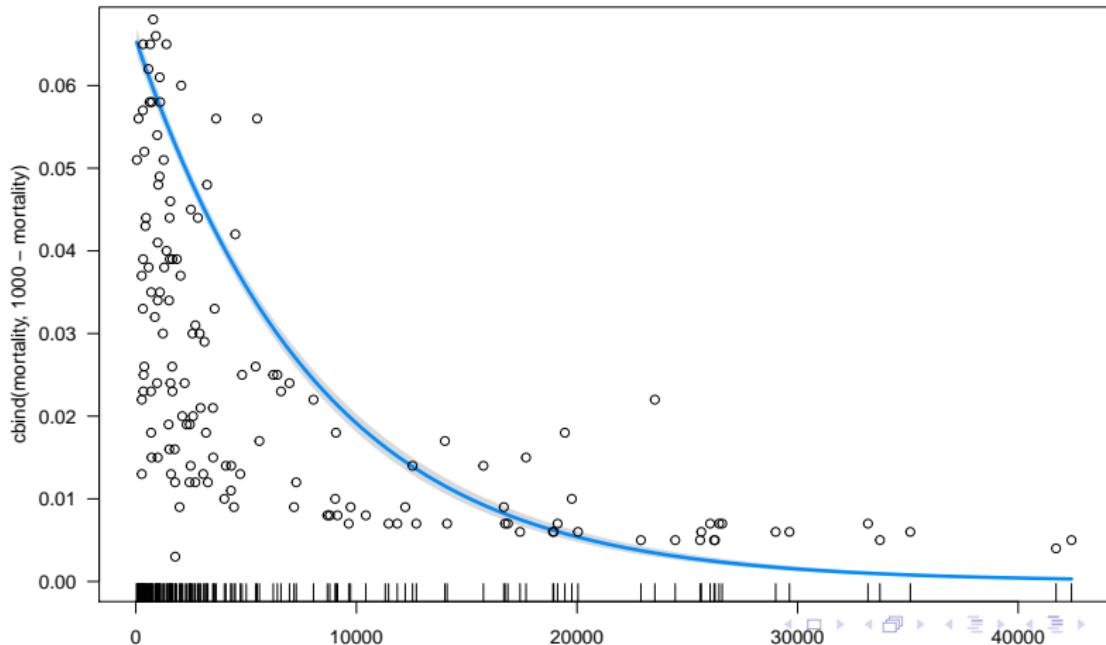
Plot model and data

```
plot(mortality/1000 ~ gdp, data = gdp, main = "Infant mortality  
curve(plogis(coef(gdp.glm)[1] + coef(gdp.glm)[2]*x), from = 0, t
```



Or using visreg:

```
visreg(gdp.glm, scale = "response")
points(mortality/1000 ~ gdp, data = gdp)
```



Overdispersion

Overdispersion in logistic regression with proportion data

```
gdp.overdisp <- glm(cbind(mortality, 1000 - mortality) ~ gdp,  
                     data = gdp, family = quasibinomial)
```

Call:

```
glm(formula = cbind(mortality, 1000 - mortality) ~ gdp, family =  
     data = gdp)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-9.2230	-3.5163	-0.5697	2.4284	13.5849

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-2.657e+00	5.977e-02	-44.465	< 2e-16 ***
gdp	-1.279e-04	1.577e-05	-8.111	5.96e-14 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 20.79)

Mean estimates do not change after accounting for overdispersion

```
model: cbind(mortality, 1000 - mortality) ~ gdp
```

```
gdp effect  
gdp
```

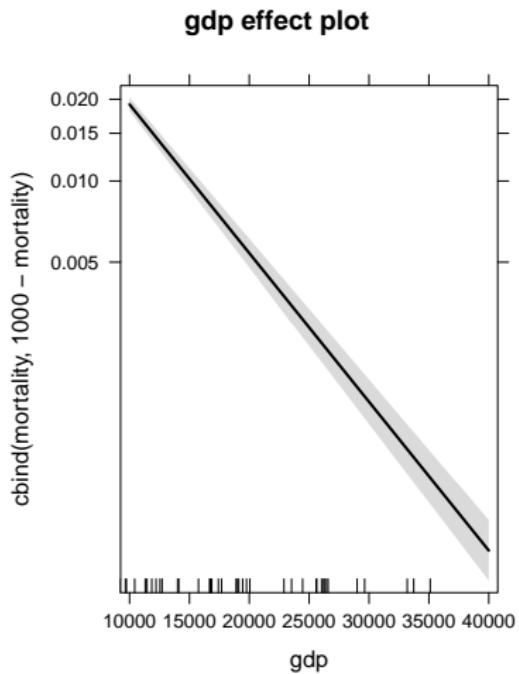
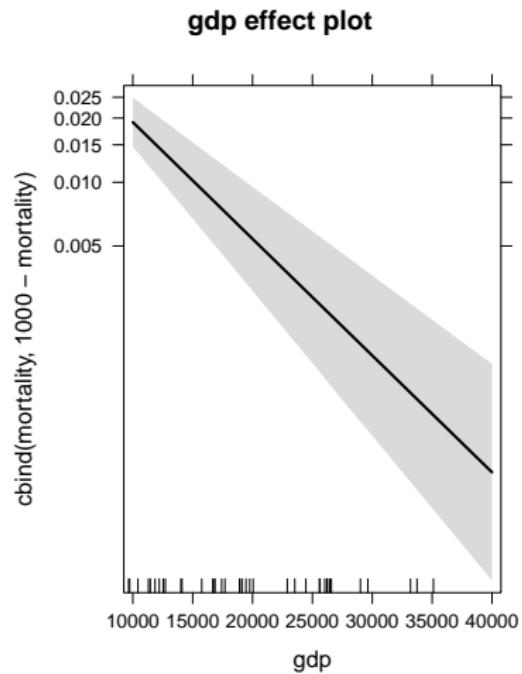
10000	20000	30000	40000
0.0191438829	0.0054028095	0.0015096074	0.0004206154

```
model: cbind(mortality, 1000 - mortality) ~ gdp
```

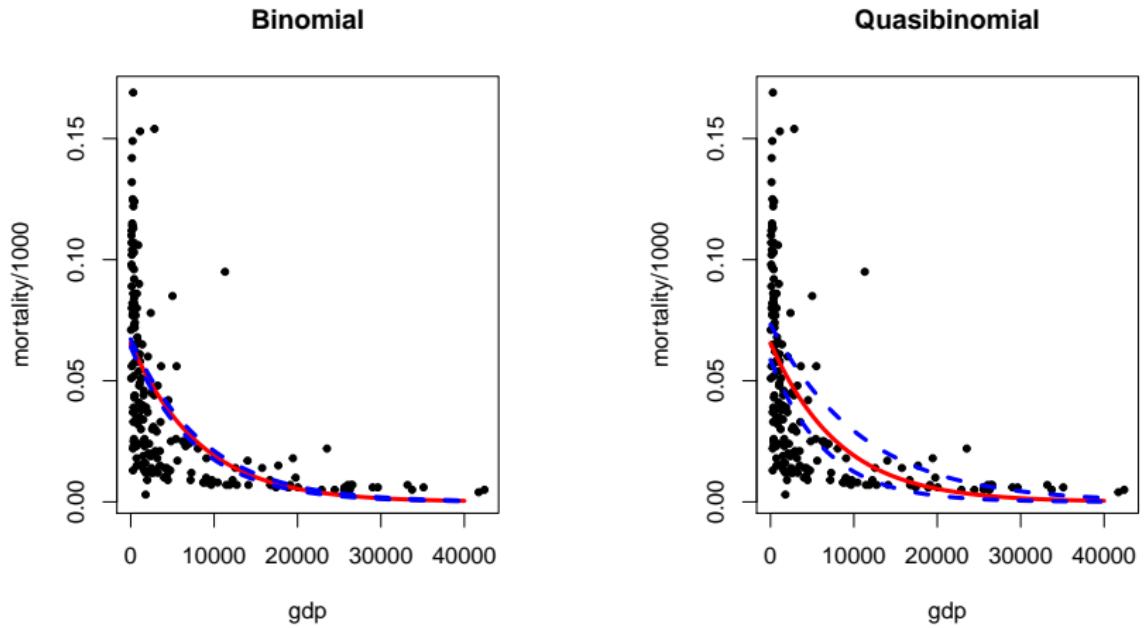
```
gdp effect  
gdp
```

10000	20000	30000	40000
0.0191438829	0.0054028095	0.0015096074	0.0004206154

But standard errors (uncertainty) do!



Plot model and data



Overdispersion

Whenever you fit logistic regression to **proportion** data, check family quasibinomial.

GLMs for count data: Poisson regression

Types of response variable

- ▶ Gaussian: lm

Types of response variable

- ▶ Gaussian: `lm`
- ▶ Bernouilli / Binomial: `glm` (family `binomial` / `quasibinomial`)

Types of response variable

- ▶ Gaussian: `lm`
- ▶ Bernouilli / Binomial: `glm` (family `binomial` / `quasibinomial`)
- ▶ Counts: `glm` (family `poisson` / `quasipoisson`)

Poisson regression

- ▶ Response variable: Counts (0, 1, 2, 3...) - discrete

Then

$$\log(N) = a + bx$$

$$N = e^{a+bx}$$

Poisson regression

- ▶ Response variable: Counts (0, 1, 2, 3...) - discrete
- ▶ Link function: log

Then

$$\log(N) = a + bx$$

$$N = e^{a+bx}$$

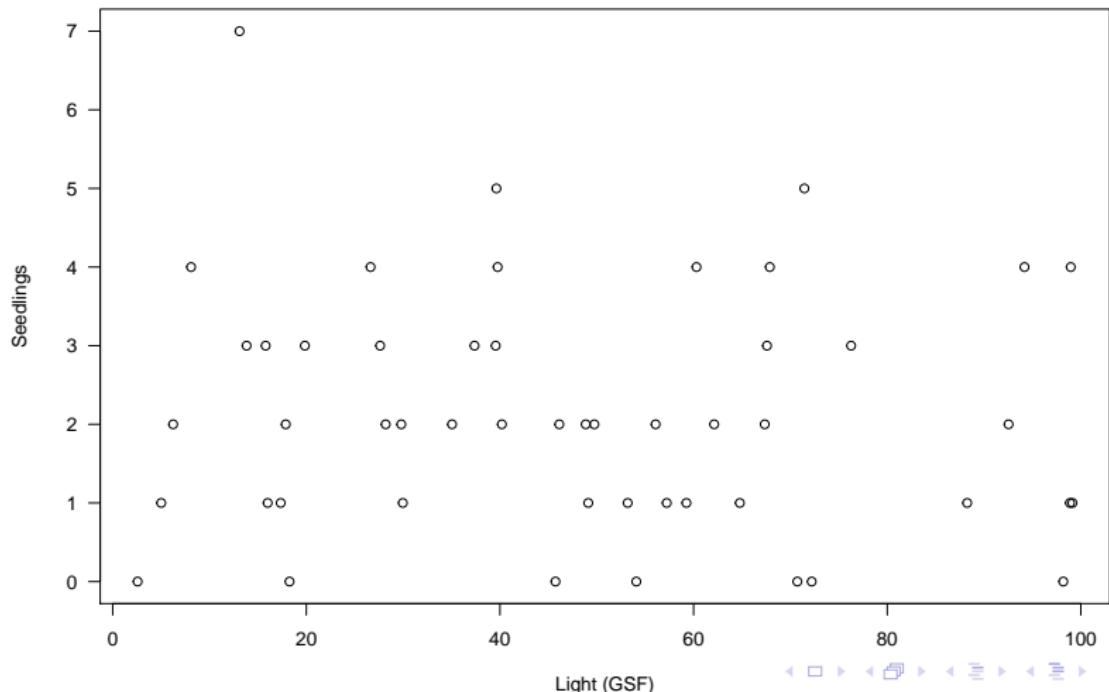
Example dataset: Seedling counts in 0.5 m² quadrats

```
seedl <- read.csv("data-raw/seedlings.csv")
```

X	count	row	col
Min. : 1.00	Min. : 0.00	Min. : 1	Min. : 1.0
1st Qu.:13.25	1st Qu.:1.00	1st Qu.:2	1st Qu.: 3.0
Median :25.50	Median :2.00	Median :3	Median : 5.5
Mean :25.50	Mean :2.14	Mean :3	Mean : 5.5
3rd Qu.:37.75	3rd Qu.:3.00	3rd Qu.:4	3rd Qu.: 8.0
Max. :50.00	Max. :7.00	Max. :5	Max. :10.0
light			
Min. : 2.571			
1st Qu.:26.879			
Median :47.493			
Mean :47.959			
3rd Qu.:67.522			
Max. :99.135			

Q: Relationship between Nseedlings and light?

```
plot(seed1$light, seed1$count, las = 1, xlab = "Light (GSF)", yl
```



Let's fit model (Poisson regression)

```
seed1.glm <- glm(count ~ light, data = seed1, family = poisson)  
summary(seed1.glm)
```

Call:

```
glm(formula = count ~ light, family = poisson, data = seed1)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.1906	-0.8466	-0.1110	0.5220	2.4577

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.881805	0.188892	4.668	3.04e-06 ***
light	-0.002576	0.003528	-0.730	0.465

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Interpreting Poisson regression output

Parameter estimates (log scale):

```
coef(seed1.glm)
```

	light
(Intercept)	0.881805022
light	-0.002575656

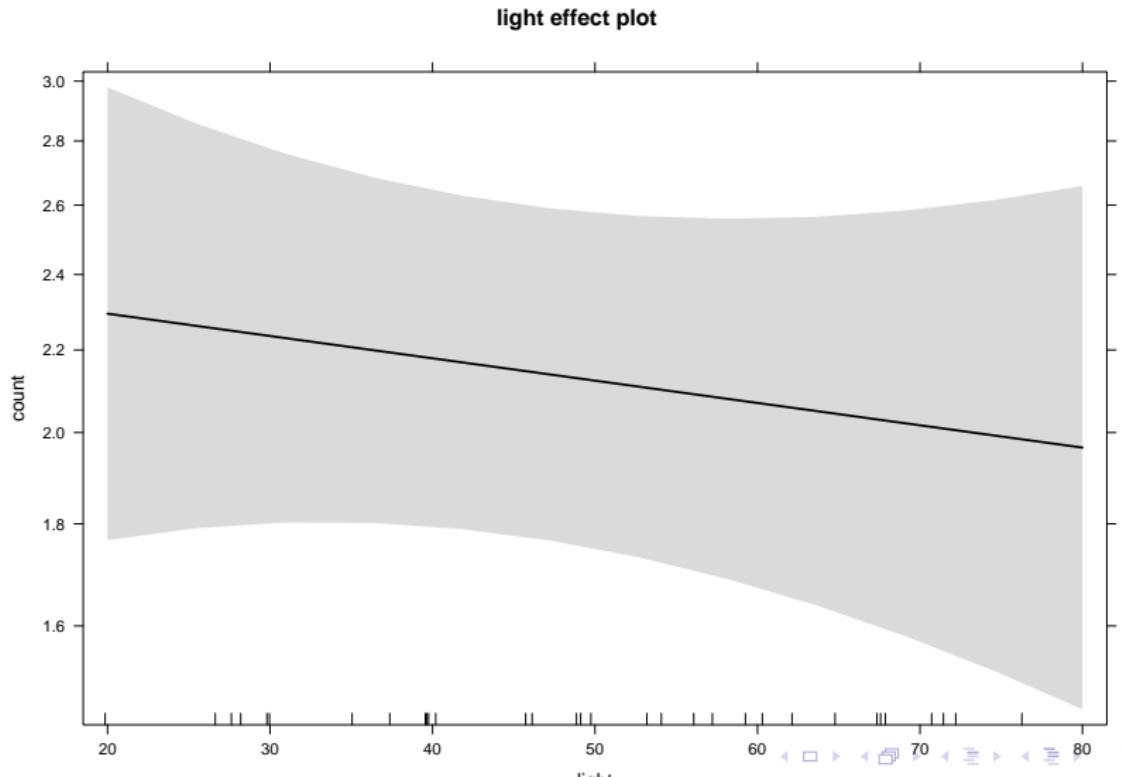
We need to back-transform: apply the inverse of the logarithm

```
exp(coef(seed1.glm))
```

	light
(Intercept)	2.4152554
light	0.9974277

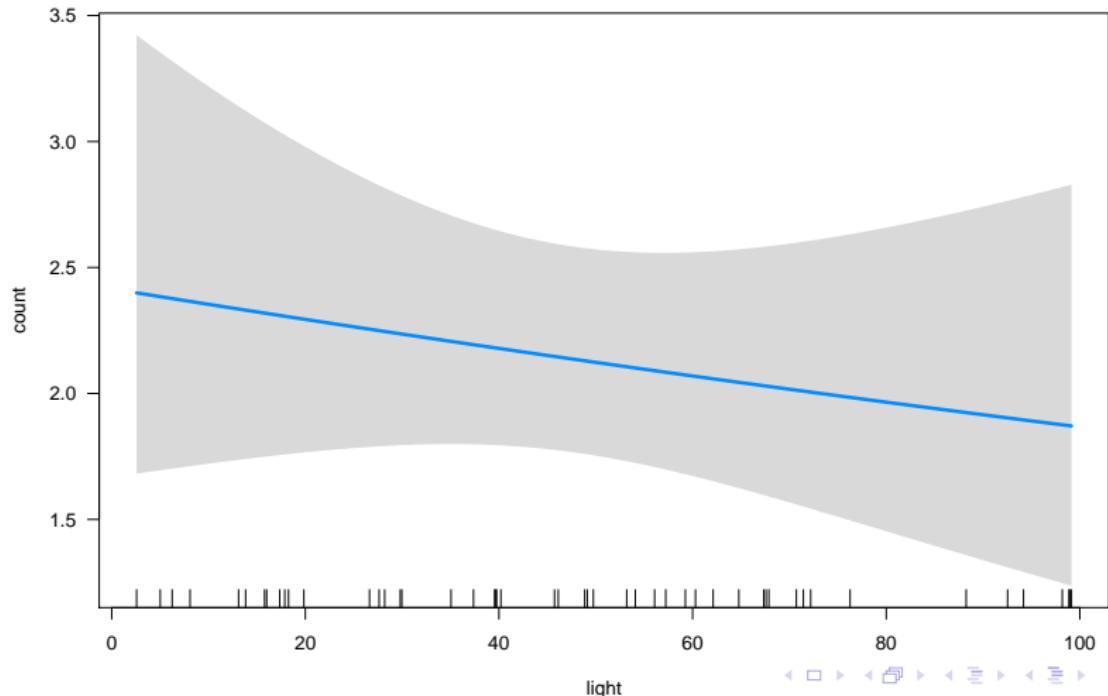
So what's the relationship between Nseedlings and light?

```
plot(allEffects(seed1.glm))
```

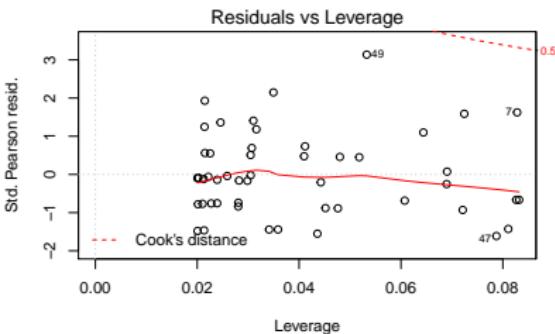
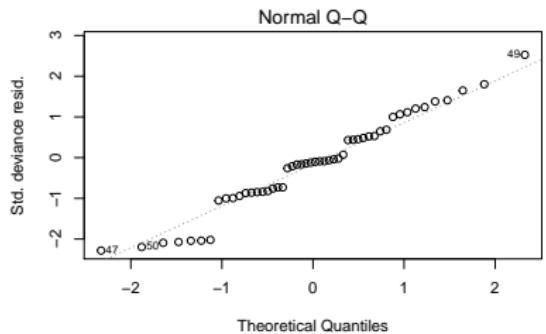
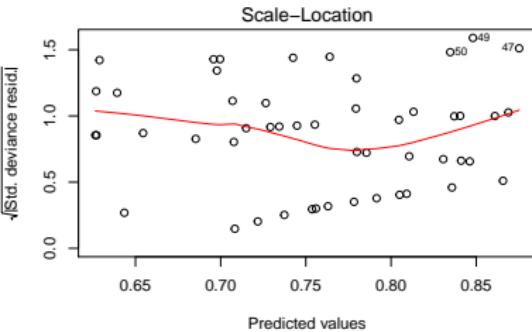
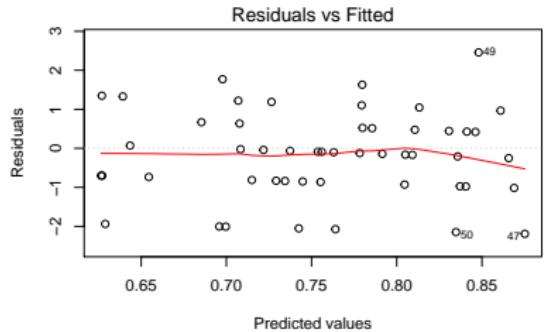


Using visreg

```
visreg(seed1.glm, scale = "response")
```

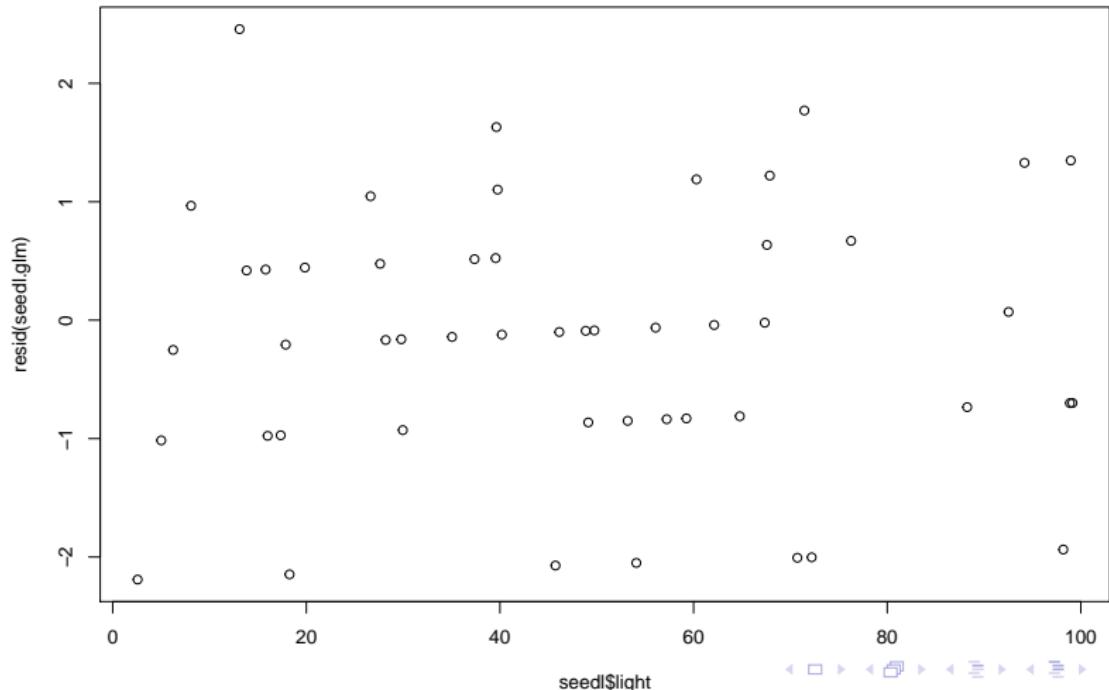


Poisson regression: model checking



Is there pattern of residuals along predictor?

```
plot(seed1$light, resid(seed1.glm))
```



Poisson regression: Overdispersion

Always check overdispersion with count data

Use family quasipoisson

Call:

```
glm(formula = count ~ light, family = quasipoisson, data = seedl)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.1906	-0.8466	-0.1110	0.5220	2.4577

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.881805	0.201230	4.382	6.37e-05 ***
light	-0.002576	0.003758	-0.685	0.496

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 1.1349)

Null deviance: 63.029 on 49 degrees of freedom

Residual deviance: 62.492 on 48 degrees of freedom

Mean estimates do not change after accounting for overdispersion

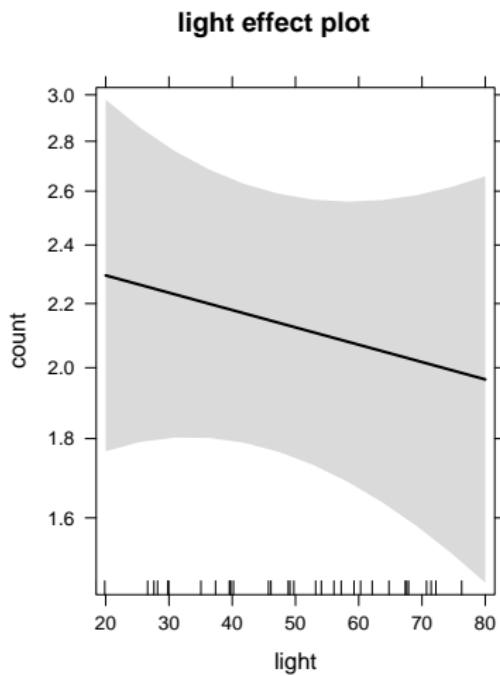
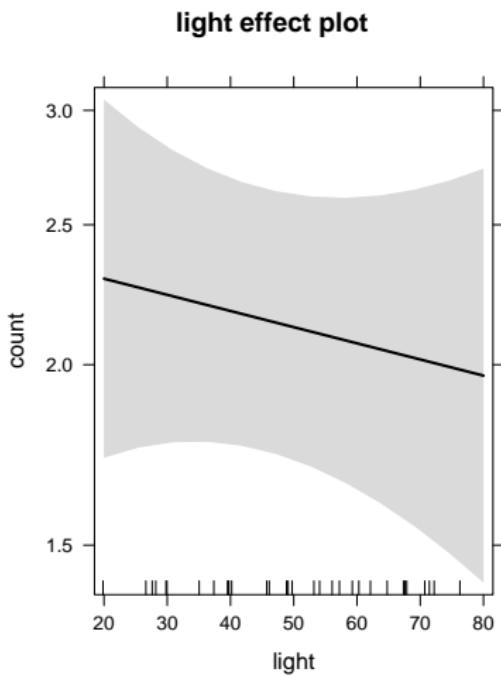
```
model: count ~ light
```

```
light effect  
light  
20      40      60      80  
2.293988 2.178810 2.069414 1.965512
```

```
model: count ~ light
```

```
light effect  
light  
20      40      60      80  
2.293988 2.178810 2.069414 1.965512
```

But standard errors may change



Mixed / Multilevel Models

Mixed models enable us to account for variability

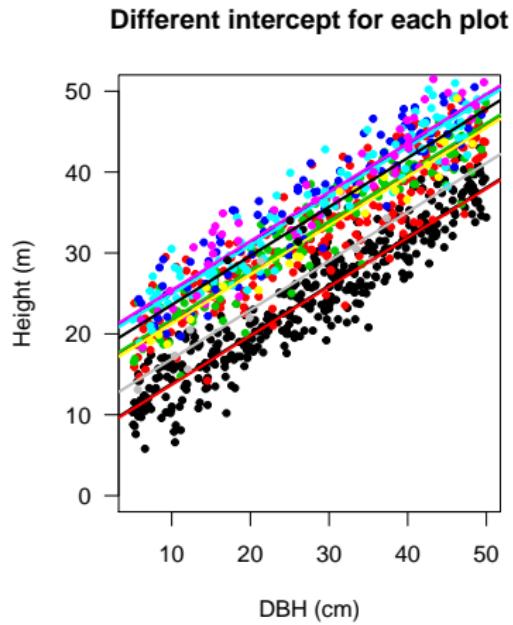
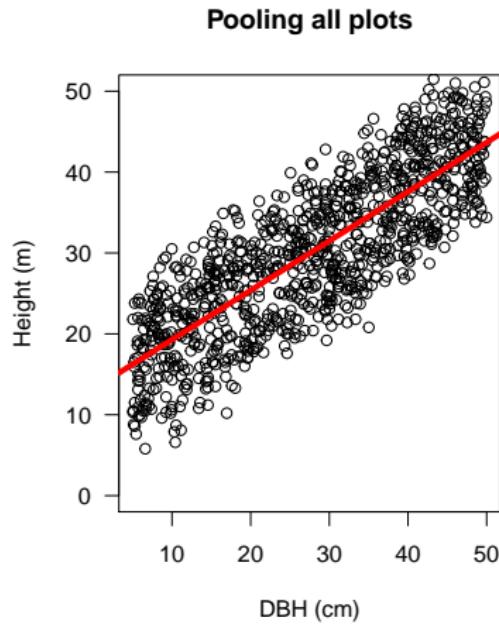
- ▶ Varying intercepts

Mixed models enable us to account for variability

- ▶ Varying intercepts
- ▶ Varying slopes

Single vs varying intercept

Dataset: 1000 trees from 10 plots (trees per plot: 4 - 392).



Fitting a varying intercepts model with lm

```
lm(formula = height ~ factor(plot) + dbh, data = trees)
            coef.est  coef.se
(Intercept)    7.79     0.24
factor(plot)2   7.86     0.24
factor(plot)3   7.95     0.32
factor(plot)4  11.48     0.33
factor(plot)5  11.05     0.32
factor(plot)6  11.55     0.43
factor(plot)7   7.41     0.63
factor(plot)8   3.05     0.97
factor(plot)9   9.73     1.45
factor(plot)10 -0.14     0.92
dbh             0.61     0.01
---
n = 1000, k = 11
residual sd = 2.89, R-Squared = 0.91
```

Mixed model with varying intercepts

$$y_i = a_j + bx_i + \varepsilon_i$$

$$a_j \sim N(0, \tau^2)$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

En nuestro ejemplo:

$$Height_i = plot_j + bDBH_i + \varepsilon_i$$

$$plot_j \sim N(0, \tau^2)$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

Fitting mixed/multilevel models

```
library(lme4)
mixed <- lmer(height ~ dbh + (1|plot), data = trees)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: height ~ dbh + (1 | plot)
Data: trees
```

REML criterion at convergence: 5007.6

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.84491	-0.65574	-0.02247	0.69295	3.09733

Random effects:

Groups	Name	Variance	Std.Dev.
plot	(Intercept)	19.834	4.454
	Residual	8.325	2.885

Number of obs: 1000, groups: plot, 10

Fixed effects:

Retrieve model coefficients

```
coef(mixed)
```

```
$plot  
(Intercept) dbh  
1 7.798373 0.6056549  
2 15.647613 0.6056549  
3 15.735397 0.6056549  
4 19.253661 0.6056549  
5 18.819467 0.6056549  
6 19.306574 0.6056549  
7 15.197908 0.6056549  
8 11.016485 0.6056549  
9 17.265447 0.6056549  
10 7.940715 0.6056549
```

```
attr(,"class")  
[1] "coef.mer"
```

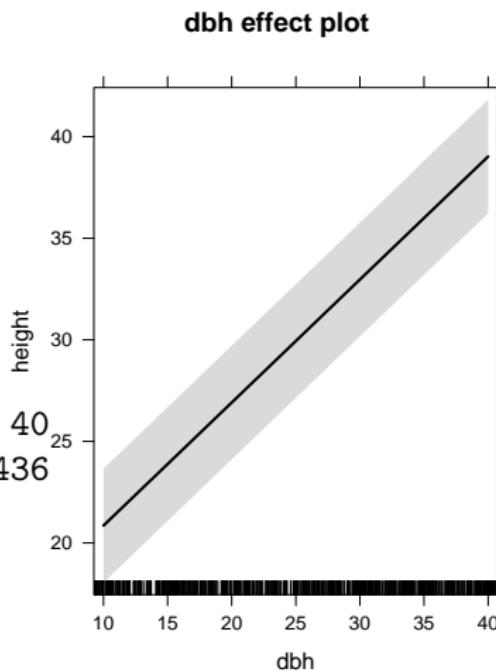
Visualising model: allEffects

model: height ~ dbh

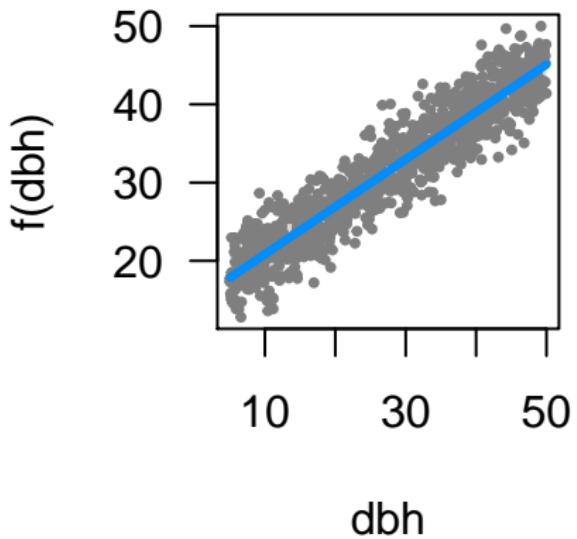
dbh effect

dbh

10	20	30	40
20.85471	26.91126	32.96781	39.02436

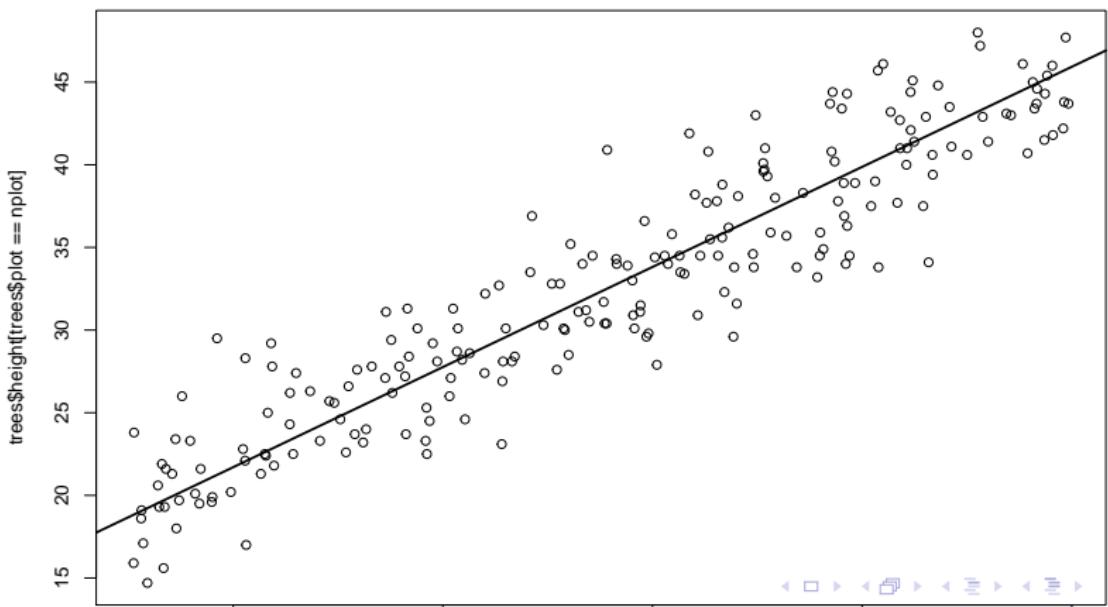


Visualising model: visreg



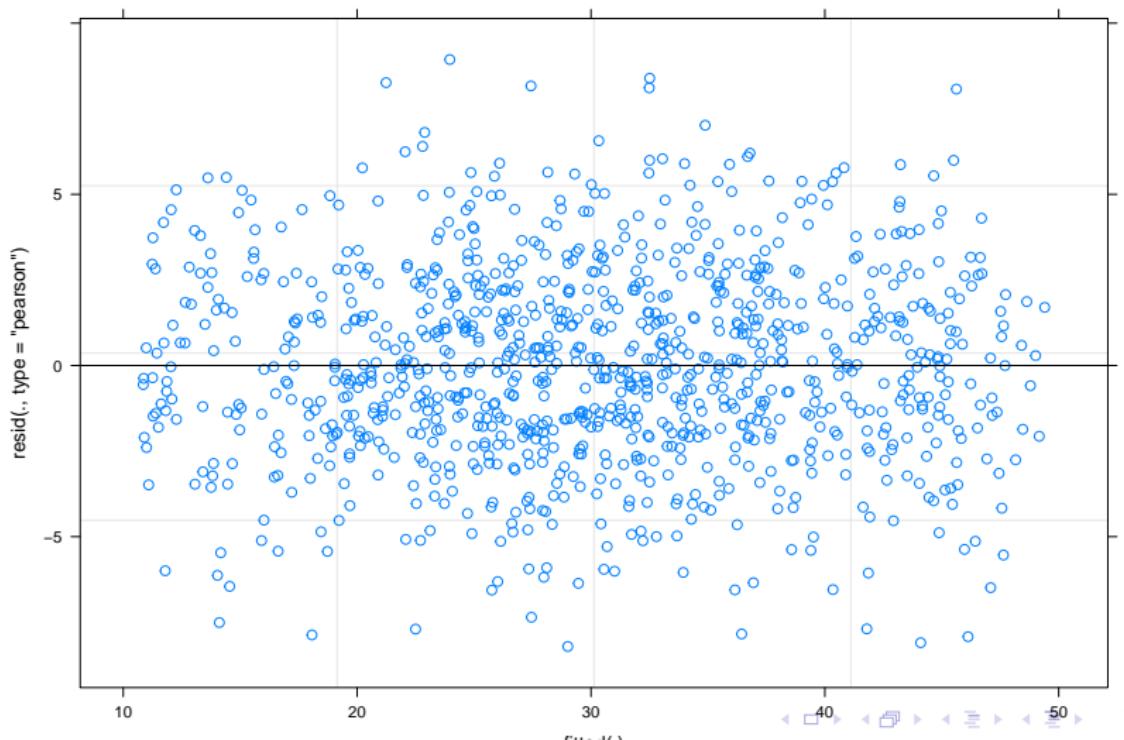
Plotting regression for individual forest plots

```
nplot <- 2  
plot(trees$dbh[trees$plot==nplot], trees$height[trees$plot==nplot]  
abline(a=coef(mixed)$plot[nplot, 1], b=coef(mixed)$plot[nplot, 2])
```



Checking residuals

```
plot(mixed)
```



Model selection

Why model selection?

- ▶ *Nested models*: how much complexity is necessary to fit the data?

Why model selection?

- ▶ *Nested models*: how much complexity is necessary to fit the data?
- ▶ *Non-nested models*: compare fit of different models (e.g. alternative hypotheses)

Why model selection?

- ▶ *Nested models*: how much complexity is necessary to fit the data?
- ▶ *Non-nested models*: compare fit of different models (e.g. alternative hypotheses)
 - ▶ But building larger model might be better than choosing any of them!

Overfitting and balanced model complexity

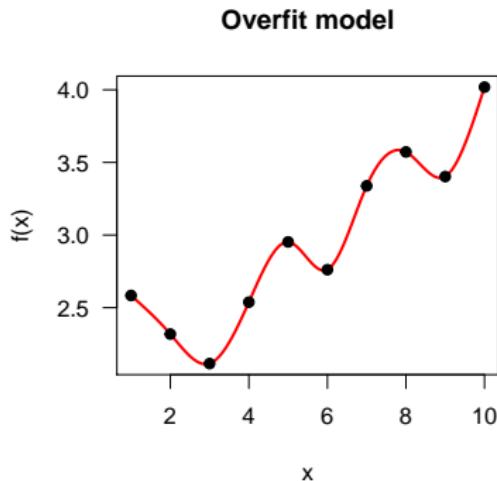


Figure 1: Overfitted model

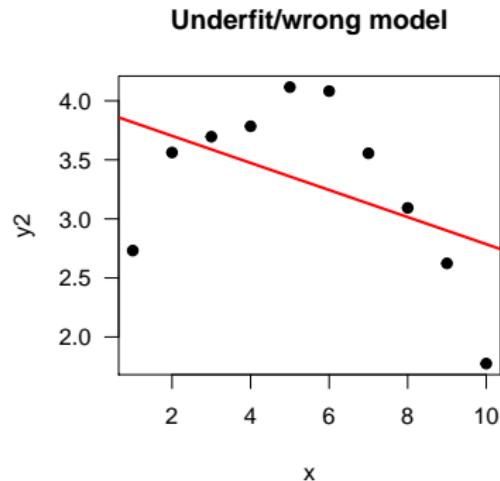
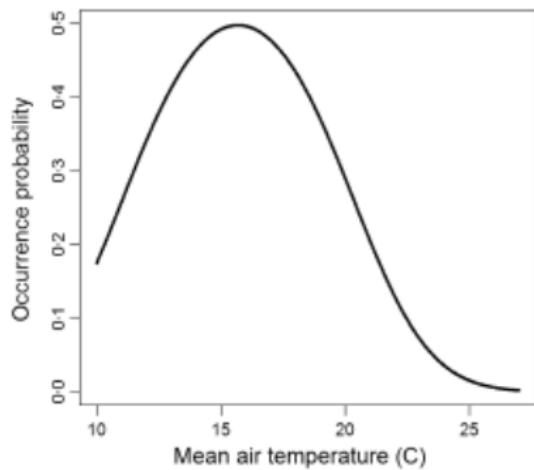


Figure 2: Wrong model

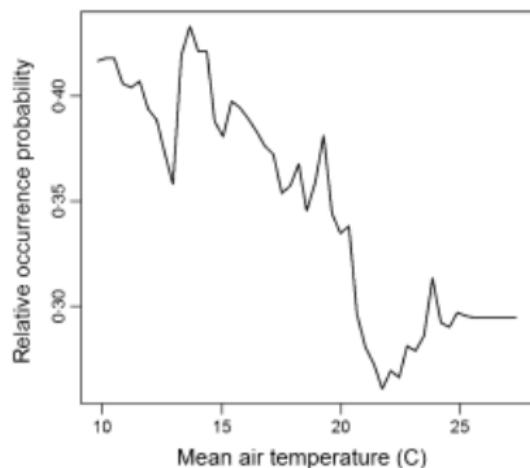
Overfitting: an example with niche modelling

Wenger & Olden (2012) Assessing transferability of ecological models: an underappreciated aspect of statistical validation. *Methods Ecol Evol.*

GLMM



Random forests (overfit)



So, two important aspects of model selection

- ▶ On one hand, we want to maximise fit.

So, two important aspects of model selection

- ▶ On one hand, we want to maximise fit.
- ▶ On the other hand, we want to avoid overfitting and overly complex models.

Evaluating models' predictive accuracy

- ▶ Cross-validation (k fold, leave one out...)

Evaluating models' predictive accuracy

- ▶ Cross-validation (k fold, leave one out...)
- ▶ Alternatives:

Evaluating models' predictive accuracy

- ▶ Cross-validation (k fold, leave one out...)
- ▶ Alternatives:
 - ▶ AIC

Evaluating models' predictive accuracy

- ▶ Cross-validation (k fold, leave one out...)
- ▶ Alternatives:
 - ▶ AIC
 - ▶ BIC

Evaluating models' predictive accuracy

- ▶ Cross-validation (k fold, leave one out...)
- ▶ Alternatives:
 - ▶ AIC
 - ▶ BIC
 - ▶ DIC

Evaluating models' predictive accuracy

- ▶ Cross-validation (k fold, leave one out...)
- ▶ Alternatives:
 - ▶ AIC
 - ▶ BIC
 - ▶ DIC
 - ▶ WAIC...

Evaluating models' predictive accuracy

- ▶ Cross-validation (k fold, leave one out...)
- ▶ Alternatives:
 - ▶ AIC
 - ▶ BIC
 - ▶ DIC
 - ▶ WAIC...
- ▶ All these attempt an impossible task:

Evaluating models' predictive accuracy

- ▶ Cross-validation (k fold, leave one out...)
- ▶ Alternatives:
 - ▶ AIC
 - ▶ BIC
 - ▶ DIC
 - ▶ WAIC...
- ▶ All these attempt an impossible task:
 - ▶ estimating out-of-sample prediction error without external data or further model fits!

Evaluating models' predictive accuracy

- ▶ Cross-validation (k fold, leave one out...)
- ▶ Alternatives:
 - ▶ AIC
 - ▶ BIC
 - ▶ DIC
 - ▶ WAIC...
- ▶ All these attempt an impossible task:
 - ▶ estimating out-of-sample prediction error without external data or further model fits!
- ▶ All these methods have flaws!

AIC

$$\text{AIC} = -2 \log p(y|\hat{\theta}_{\text{mle}}) + 2k$$

- ▶ First term: model fit (deviance, log likelihood)

AIC

$$\text{AIC} = -2 \log p(y|\hat{\theta}_{\text{mle}}) + 2k$$

- ▶ First term: model fit (deviance, log likelihood)
- ▶ k: number of estimated parameters (penalisation for model complexity)

AIC

$$\text{AIC} = -2 \log p(y|\hat{\theta}_{\text{mle}}) + 2k$$

- ▶ First term: model fit (deviance, log likelihood)
- ▶ k: number of estimated parameters (penalisation for model complexity)
- ▶ AIC biased towards complex models.

AIC

$$\text{AIC} = -2 \log p(y|\hat{\theta}_{\text{mle}}) + 2k$$

- ▶ First term: model fit (deviance, log likelihood)
- ▶ k: number of estimated parameters (penalisation for model complexity)
- ▶ AIC biased towards complex models.
- ▶ AICc recommended with ‘small’ sample sizes ($n/p < 40$). But see Richards 2005 Ecology.

AIC

$$\text{AIC} = -2 \log p(y|\hat{\theta}_{\text{mle}}) + 2k$$

- ▶ First term: model fit (deviance, log likelihood)
- ▶ k: number of estimated parameters (penalisation for model complexity)
- ▶ AIC biased towards complex models.
- ▶ AICc recommended with ‘small’ sample sizes ($n/p < 40$). But see Richards 2005 Ecology.
- ▶ Doesn’t work with hierarchical models or informative priors!

Problems of IC

- ▶ No information criteria is panacea: all have problems.

Problems of IC

- ▶ No information criteria is panacea: all have problems.
- ▶ They give average out-of-sample prediction error, but prediction errors can differ substantially within the same dataset (e.g. populations, species).

Problems of IC

- ▶ No information criteria is panacea: all have problems.
- ▶ They give average out-of-sample prediction error, but prediction errors can differ substantially within the same dataset (e.g. populations, species).
- ▶ Sometimes better models rank poorly (Gelman et al. 2013). So, combine with thorough model checks.

So which variables should enter my model?

Choosing predictors

- ▶ Choose variables based on **ecological understanding**, rather than throwing plenty of them in a fishing expedition.

Choosing predictors

- ▶ Choose variables based on **ecological understanding**, rather than throwing plenty of them in a fishing expedition.
- ▶ Propose single global model or small set (< 10 - 20) of **reasonable** candidate models.

Choosing predictors

- ▶ Choose variables based on **ecological understanding**, rather than throwing plenty of them in a fishing expedition.
- ▶ Propose single global model or small set (< 10 - 20) of **reasonable** candidate models.
- ▶ Number of variables balanced with sample size (at least 10 - 30 obs per param)

Choosing predictors

- ▶ Choose variables based on **ecological understanding**, rather than throwing plenty of them in a fishing expedition.
- ▶ Propose single global model or small set (< 10 - 20) of **reasonable** candidate models.
- ▶ Number of variables balanced with sample size (at least 10 - 30 obs per param)
- ▶ Assess collinearity between predictors (Dormann et al 2013)

Choosing predictors

- ▶ Choose variables based on **ecological understanding**, rather than throwing plenty of them in a fishing expedition.
- ▶ Propose single global model or small set (< 10 - 20) of **reasonable** candidate models.
- ▶ Number of variables balanced with sample size (at least 10 - 30 obs per param)
- ▶ Assess collinearity between predictors (Dormann et al 2013)
 - ▶ `pairs()` or similar

Choosing predictors

- ▶ Choose variables based on **ecological understanding**, rather than throwing plenty of them in a fishing expedition.
- ▶ Propose single global model or small set (< 10 - 20) of **reasonable** candidate models.
- ▶ Number of variables balanced with sample size (at least 10 - 30 obs per param)
- ▶ Assess collinearity between predictors (Dormann et al 2013)
 - ▶ `pairs()` or similar
 - ▶ If $|r| > 0.5 - 0.7$, consider leaving one variable out, but keep it in mind when interpreting model results.

Choosing predictors

- ▶ Choose variables based on **ecological understanding**, rather than throwing plenty of them in a fishing expedition.
- ▶ Propose single global model or small set (< 10 - 20) of **reasonable** candidate models.
- ▶ Number of variables balanced with sample size (at least 10 - 30 obs per param)
- ▶ Assess collinearity between predictors (Dormann et al 2013)
 - ▶ `pairs()` or similar
 - ▶ If $|r| > 0.5 - 0.7$, consider leaving one variable out, but keep it in mind when interpreting model results.
 - ▶ Or combine 2 or more in a synthetic variable (e.g. water deficit ~ Temp + Precip).

Choosing predictors

- ▶ Choose variables based on **ecological understanding**, rather than throwing plenty of them in a fishing expedition.
- ▶ Propose single global model or small set (< 10 - 20) of **reasonable** candidate models.
- ▶ Number of variables balanced with sample size (at least 10 - 30 obs per param)
- ▶ Assess collinearity between predictors (Dormann et al 2013)
 - ▶ `pairs()` or similar
 - ▶ If $|r| > 0.5 - 0.7$, consider leaving one variable out, but keep it in mind when interpreting model results.
 - ▶ Or combine 2 or more in a synthetic variable (e.g. water deficit ~ Temp + Precip).
 - ▶ Many methods available, e.g. sequential, ridge regression... (see Dormann et al)

Choosing predictors

- ▶ Choose variables based on **ecological understanding**, rather than throwing plenty of them in a fishing expedition.
- ▶ Propose single global model or small set (< 10 - 20) of **reasonable** candidate models.
- ▶ Number of variables balanced with sample size (at least 10 - 30 obs per param)
- ▶ Assess collinearity between predictors (Dormann et al 2013)
 - ▶ `pairs()` or similar
 - ▶ If $|r| > 0.5 - 0.7$, consider leaving one variable out, but keep it in mind when interpreting model results.
 - ▶ Or combine 2 or more in a synthetic variable (e.g. water deficit ~ Temp + Precip).
 - ▶ Many methods available, e.g. sequential, ridge regression... (see Dormann et al)
 - ▶ Measurement error can seriously complicate things (Biggs et al 2009; Freckleton 2011)

Choosing predictors

- ▶ Choose variables based on **ecological understanding**, rather than throwing plenty of them in a fishing expedition.
- ▶ Propose single global model or small set (< 10 - 20) of **reasonable** candidate models.
- ▶ Number of variables balanced with sample size (at least 10 - 30 obs per param)
- ▶ Assess collinearity between predictors (Dormann et al 2013)
 - ▶ `pairs()` or similar
 - ▶ If $|r| > 0.5 - 0.7$, consider leaving one variable out, but keep it in mind when interpreting model results.
 - ▶ Or combine 2 or more in a synthetic variable (e.g. water deficit ~ Temp + Precip).
 - ▶ Many methods available, e.g. sequential, ridge regression... (see Dormann et al)
 - ▶ Measurement error can seriously complicate things (Biggs et al 2009; Freckleton 2011)
- ▶ For predictors with large effects, consider interactions.

Choosing predictors

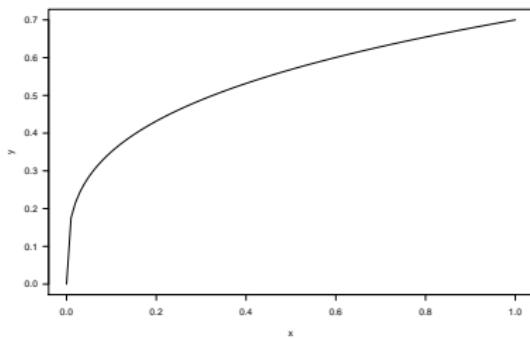
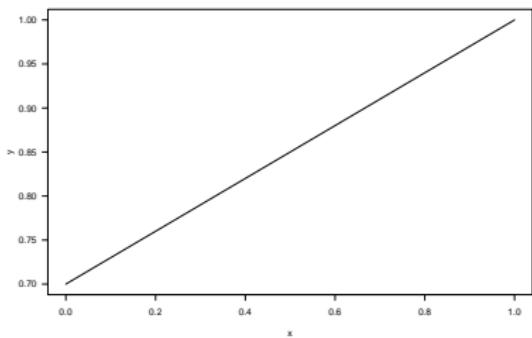
- ▶ Choose variables based on **ecological understanding**, rather than throwing plenty of them in a fishing expedition.
- ▶ Propose single global model or small set (< 10 - 20) of **reasonable** candidate models.
- ▶ Number of variables balanced with sample size (at least 10 - 30 obs per param)
- ▶ Assess collinearity between predictors (Dormann et al 2013)
 - ▶ `pairs()` or similar
 - ▶ If $|r| > 0.5 - 0.7$, consider leaving one variable out, but keep it in mind when interpreting model results.
 - ▶ Or combine 2 or more in a synthetic variable (e.g. water deficit ~ Temp + Precip).
 - ▶ Many methods available, e.g. sequential, ridge regression... (see Dormann et al)
 - ▶ Measurement error can seriously complicate things (Biggs et al 2009; Freckleton 2011)
- ▶ For predictors with large effects, consider interactions.
- ▶ See also Zuur et al 2010.

Think about the shape of relationships

$$y \sim x + z$$

Really? Not everything has to be linear! Actually, it often is not.

Think about shape of relationship. See chapter 3 in Bolker's book.



Removing predictors

Do not use stepwise regression

- ▶ Whittingham et al. (2006) Why do we still use stepwise modelling in ecology and behaviour? J. Animal Ecology.

Do not use stepwise regression

- ▶ Whittingham et al. (2006) Why do we still use stepwise modelling in ecology and behaviour? *J. Animal Ecology.*
- ▶ Mundry & Nunn (2009) Stepwise Model Fitting and Statistical Inference: Turning Noise into Signal Pollution. *Am Nat.*

Do not use stepwise regression

- ▶ Whittingham et al. (2006) Why do we still use stepwise modelling in ecology and behaviour? *J. Animal Ecology.*
- ▶ Mundry & Nunn (2009) Stepwise Model Fitting and Statistical Inference: Turning Noise into Signal Pollution. *Am Nat.*
- ▶ This includes stepAIC (e.g. Dahlgren 2010; Burnham et al 2011; Hegyi & Garamszegi 2011).

Gelman's criteria for removing predictors

(assuming only potentially relevant predictors have been selected a priori)

- ▶ NOT significant + expected sign = let it be.

Gelman's criteria for removing predictors

(assuming only potentially relevant predictors have been selected a priori)

- ▶ NOT significant + expected sign = let it be.
- ▶ NOT significant + NOT expected sign = remove it.

Gelman's criteria for removing predictors

(assuming only potentially relevant predictors have been selected a priori)

- ▶ NOT significant + expected sign = let it be.
- ▶ NOT significant + NOT expected sign = remove it.
- ▶ Significant + NOT expected sign = check... confounding variables?

Gelman's criteria for removing predictors

(assuming only potentially relevant predictors have been selected a priori)

- ▶ NOT significant + expected sign = let it be.
- ▶ NOT significant + NOT expected sign = remove it.
- ▶ Significant + NOT expected sign = check... confounding variables?
- ▶ Significant + expected sign = keep it!

The modelling process

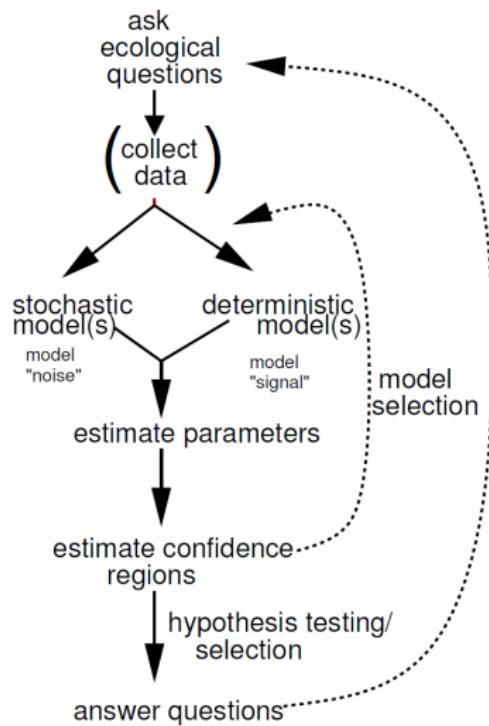


Figure 1.5 Flow of the modeling process.

Bolker 2008

Summary

1. Choose meaningful variables

Summary

1. Choose meaningful variables
 - ▶ Beware collinearity

Summary

1. Choose meaningful variables

- ▶ Beware collinearity
- ▶ Keep good n/p ratio

Summary

1. Choose meaningful variables
 - ▶ Beware collinearity
 - ▶ Keep good n/p ratio
2. Generate global model or (small) set of candidate models

Summary

1. Choose meaningful variables
 - ▶ Beware collinearity
 - ▶ Keep good n/p ratio
2. Generate global model or (small) set of candidate models
 - ▶ Avoid stepwise and all-subsets

Summary

1. Choose meaningful variables
 - ▶ Beware collinearity
 - ▶ Keep good n/p ratio
2. Generate global model or (small) set of candidate models
 - ▶ Avoid stepwise and all-subsets
 - ▶ Don't assume linear effects: think about appropriate functional relationships

Summary

1. Choose meaningful variables
 - ▶ Beware collinearity
 - ▶ Keep good n/p ratio
2. Generate global model or (small) set of candidate models
 - ▶ Avoid stepwise and all-subsets
 - ▶ Don't assume linear effects: think about appropriate functional relationships
 - ▶ Consider interactions for strong main effects

Summary

1. Choose meaningful variables
 - ▶ Beware collinearity
 - ▶ Keep good n/p ratio
2. Generate global model or (small) set of candidate models
 - ▶ Avoid stepwise and all-subsets
 - ▶ Don't assume linear effects: think about appropriate functional relationships
 - ▶ Consider interactions for strong main effects
3. If > 1 model have similar support, consider model averaging (or blending).

Summary

1. Choose meaningful variables
 - ▶ Beware collinearity
 - ▶ Keep good n/p ratio
2. Generate global model or (small) set of candidate models
 - ▶ Avoid stepwise and all-subsets
 - ▶ Don't assume linear effects: think about appropriate functional relationships
 - ▶ Consider interactions for strong main effects
3. If > 1 model have similar support, consider model averaging (or blending).
4. Always check thoroughly fitted models

Summary

1. Choose meaningful variables
 - ▶ Beware collinearity
 - ▶ Keep good n/p ratio
2. Generate global model or (small) set of candidate models
 - ▶ Avoid stepwise and all-subsets
 - ▶ Don't assume linear effects: think about appropriate functional relationships
 - ▶ Consider interactions for strong main effects
3. If > 1 model have similar support, consider model averaging (or blending).
4. Always check thoroughly fitted models
 - ▶ Residuals, goodness of fit...

Summary

1. Choose meaningful variables
 - ▶ Beware collinearity
 - ▶ Keep good n/p ratio
2. Generate global model or (small) set of candidate models
 - ▶ Avoid stepwise and all-subsets
 - ▶ Don't assume linear effects: think about appropriate functional relationships
 - ▶ Consider interactions for strong main effects
3. If > 1 model have similar support, consider model averaging (or blending).
4. Always check thoroughly fitted models
 - ▶ Residuals, goodness of fit...
 - ▶ Plot. Check models. Plot. Check assumptions. Plot. (Lavine 2014).

Summary

1. Choose meaningful variables
 - ▶ Beware collinearity
 - ▶ Keep good n/p ratio
2. Generate global model or (small) set of candidate models
 - ▶ Avoid stepwise and all-subsets
 - ▶ Don't assume linear effects: think about appropriate functional relationships
 - ▶ Consider interactions for strong main effects
3. If > 1 model have similar support, consider model averaging (or blending).
4. Always check thoroughly fitted models
 - ▶ Residuals, goodness of fit...
 - ▶ Plot. Check models. Plot. Check assumptions. Plot. (Lavine 2014).
5. Always report effect sizes

END

:)

Source code and materials:

<https://github.com/Pakillo/stats-intro>

