Model selection

Slides and code modified from Eric Allan

Why do model selection?

- You want to see which factors are most important out of a selection
- You don’t want an over parameterised model
- You have interactions and want to test if they matter
-

Different approaches

Maximising fit (e.g. R2)

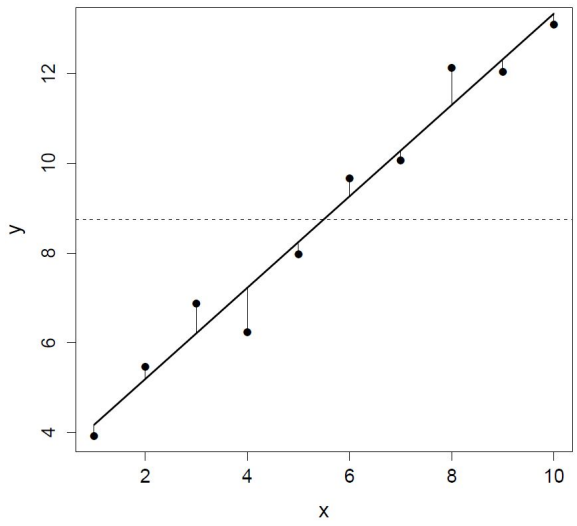
Model simplification (null hypothesis testing)

Model comparison (AIC, or other information criteria)

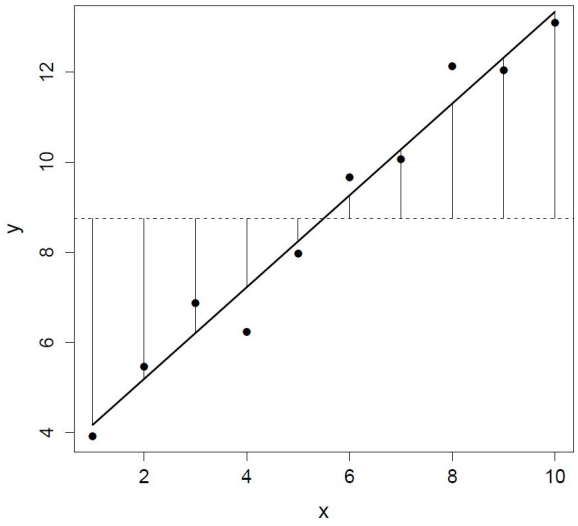
Measures of model fit: R^2

- The coefficient of determination R^2 is the "**proportion of variance explained by the model**", i.e. the proportion of variance in the response variable that is predictable from the explanatory variable(s).
- It is based on the Sum of Squares

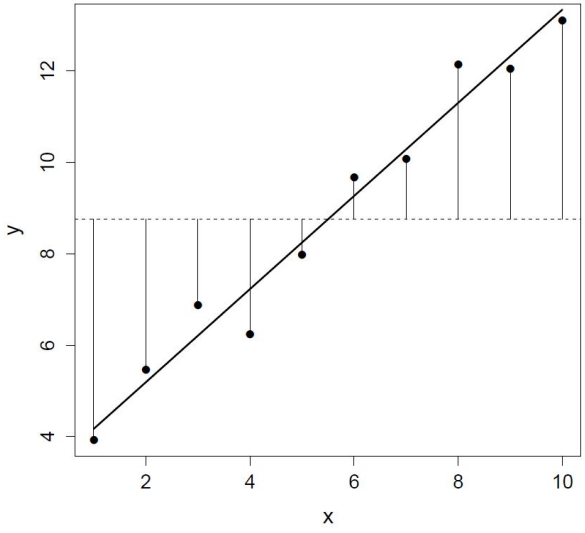
$SS_{residual} = \Sigma_i e_i^2 = \Sigma (y_i - \hat{y}_i)^2$



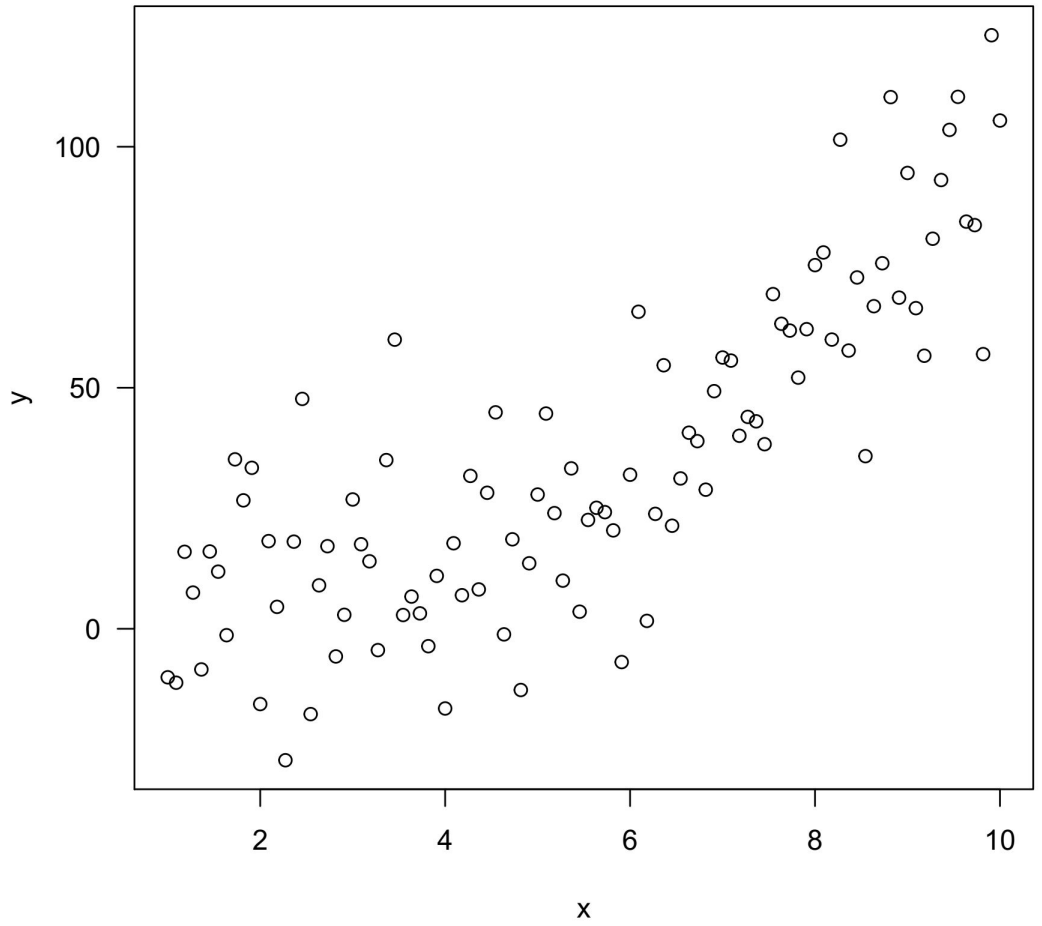
$SS_{regression} = \Sigma (\hat{y}_i - \bar{y})^2$



$SS_{total} = \Sigma (y_i - \bar{y})^2$



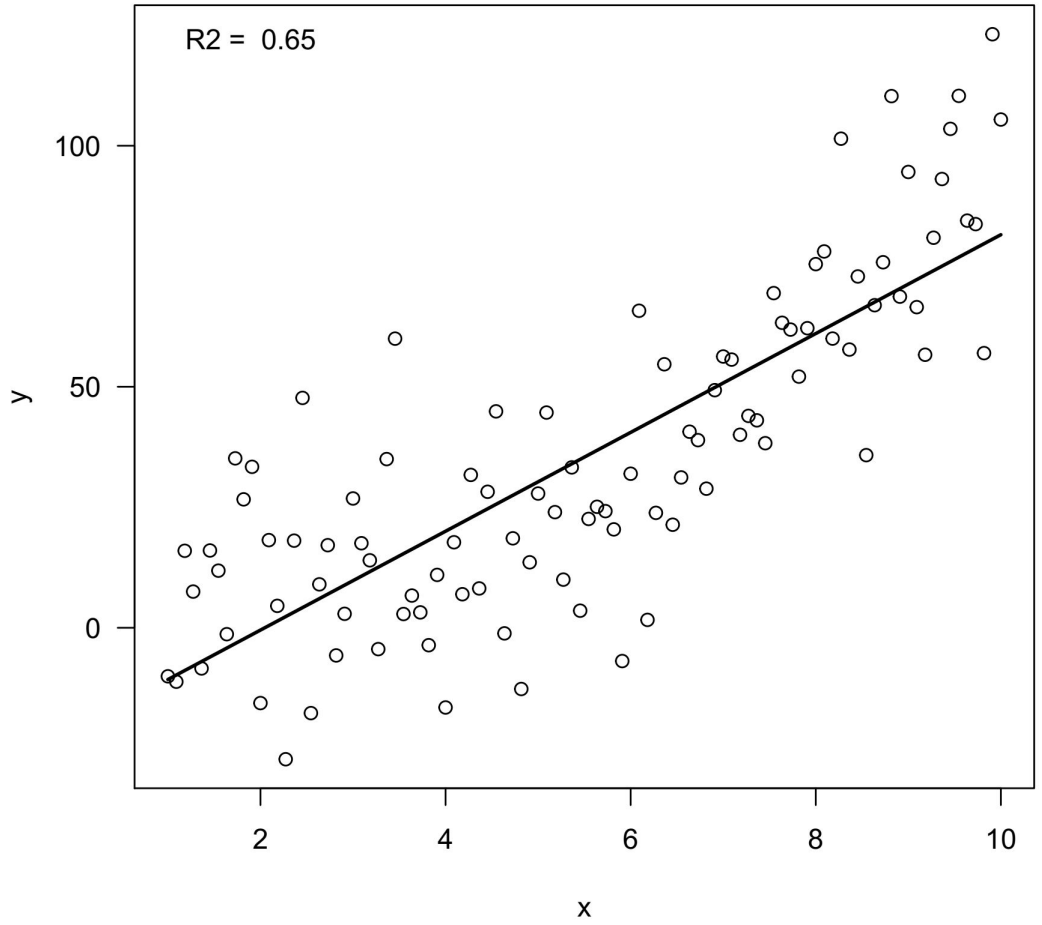
R^2 and model complexity



R^2 and model complexity

$y = a + bx$

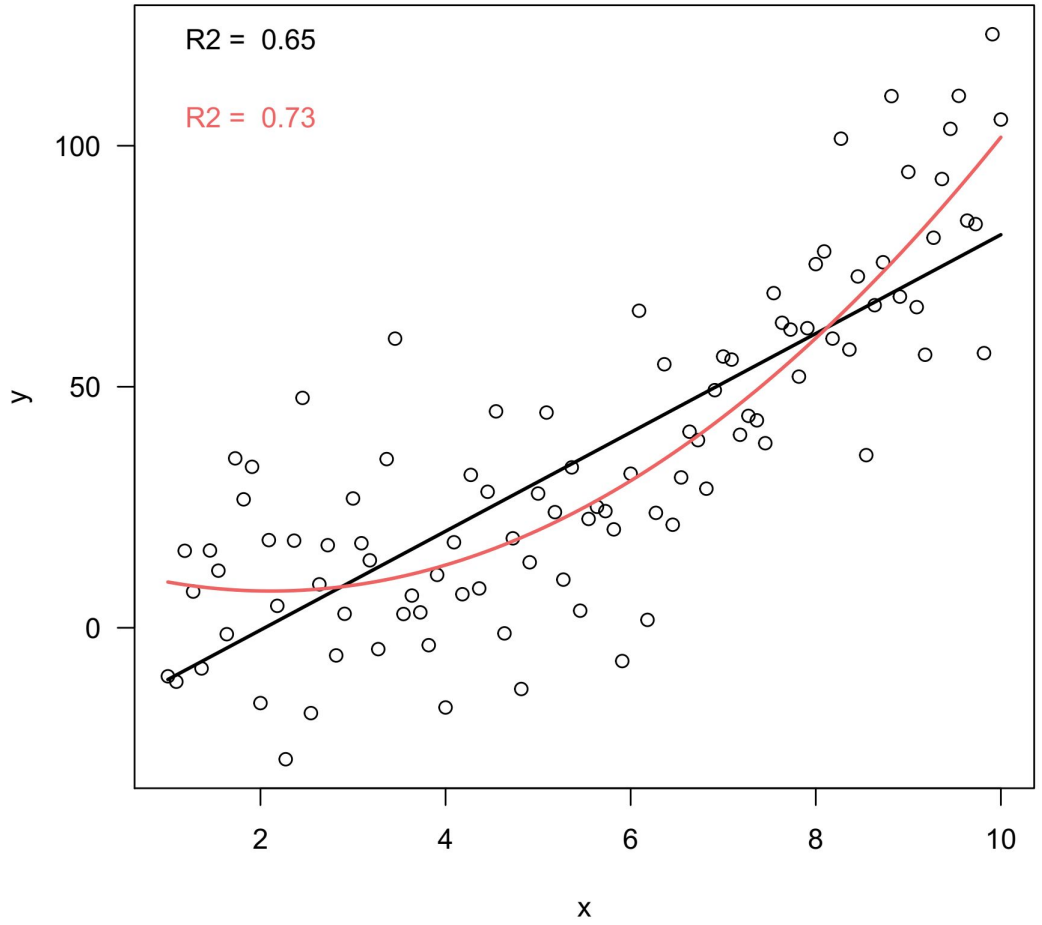
or `lm(y~x)`



R^2 and model complexity

$y = a + bx + cx^2$

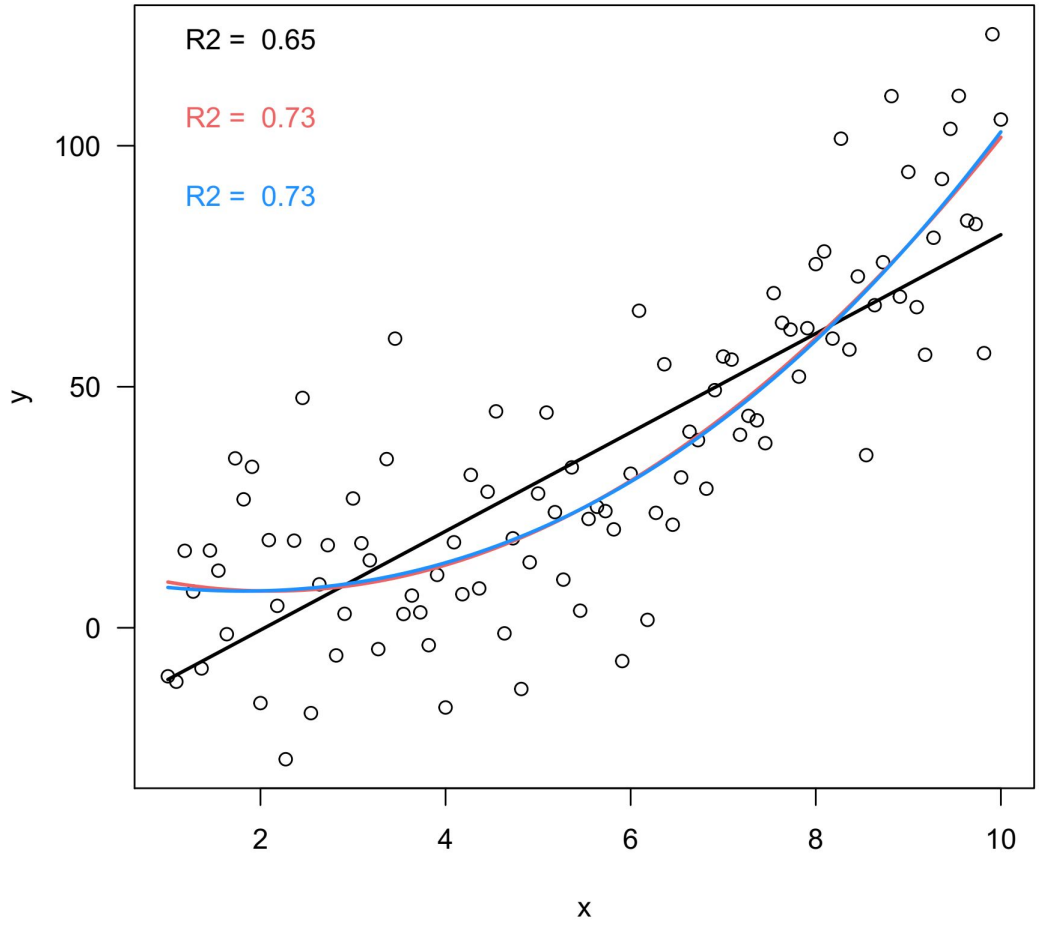
or `lm(y ~ x + I(x^2))` or `lm(y ~ poly(x, 2))`



R^2 and model complexity

$y = a + bx + cx^2 dx^3$

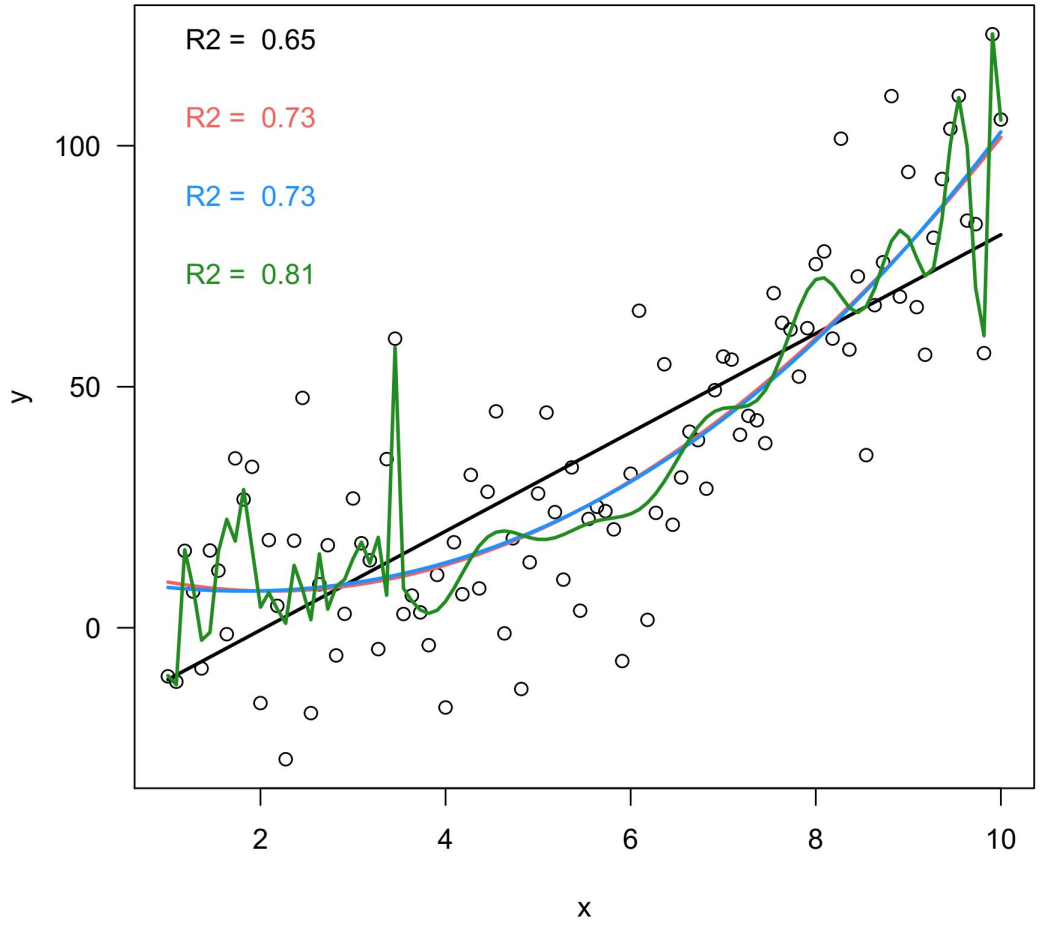
or `lm(y ~ x + I(x^2) + I(x^3))` or `lm(y ~ poly(x, 3))`



R^2 and model complexity

$y = a + bx + cx^2 dx^3 + ex^4 + fx^5 \ldots zx^{25}$

or `lm(y ~ poly(x, 25))`



Model simplification

Is a simpler model **significantly** worse than a more complex one?

Always prefer a simpler model if it performs as well (principle of parsimony)

Start with a complex model and remove terms

- If the simpler model is significantly worse the term has to stay in
- Must compare NESTED models

```
y ~ x1 + x2 + x3
```

Vs.

```
y ~ x1 + x2
```

=

```
y ~ x1 + x2 + 0*x3
```

Model simplification

How to compare models

- Linear models: F-ratio test (compare the RSS of the two models)
- GLMs or mixed models: Likelihood ratio tests (compare ratio of log likelihoods to a χ^2 distribution)

Model simplification

Some rules:

- Always remove higher order terms or interactions first
- Respect “principle of marginality”
- Keep going until only significant terms remain
- Be careful that models are fitted to same dataset!

```
y ~ x1 + x2 + x3 + x2:x3
```

Vs.

```
y ~ x1 + x2 + x3
```

Not!

```
y ~ x2 + x3 + x2:x3
```

Model simplification example

What drives predatory insect diversity in grasslands?

```
lm(Predator_SpeciesRichness ~ Region + Fstd + Gstd + Mstd +
Plant_SpeciesRichness + Plant_biomass + Herbivore_SpeciesRichness +
Herbivore_biomass +
Mstd:Herbivore_SpeciesRichness + Plant_biomass:Herbivore_SpeciesRichness +
Mstd:Herbivore_biomass + Plant_biomass:Herbivore_biomass,
    data = dat)
```

```
Coefficients:
                                        Estimate    Std. Error t value Pr(>|t|)
(Intercept)                          6.5912891471  2.3094395533   2.854  0.00517 **
RegionSCH                           -0.4223224789  0.6886784962  -0.613  0.54100
RegionALB                           -0.6975464857  0.7703844012  -0.905  0.36722
Fstd                                 0.0758311097  0.2342755414   0.324  0.74680
Gstd                                -0.1572938757  0.2019442853  -0.779  0.43773
Mstd                                -0.8457580810  0.9743828383  -0.868  0.38730
Plant_SpeciesRichness               -0.0608167406  0.0437888681  -1.389  0.16771
Plant_biomass                       -0.0038467374  0.0035852834  -1.073  0.28567
Herbivore_SpeciesRichness            0.0244618616  0.0782082811   0.313  0.75505
Herbivore_biomass                    0.0000708773  0.0000937884   0.756  0.45145
Mstd:Herbivore_SpeciesRichness       0.0398443434  0.0329829256   1.208  0.22965
Plant_biomass:Herbivore_SpeciesRichness  0.0000569330  0.0001304427   0.436  0.66337
Mstd:Herbivore_biomass              -0.0000720048  0.0000421465  -1.708  0.09040 .
Plant_biomass:Herbivore_biomass      0.0000001585  0.0000001210   1.310  0.19289

Residual standard error: 2.523 on 109 degrees of freedom
Multiple R-squared:  0.3801,    Adjusted R-squared:  0.3061
F-statistic: 5.141 on 13 and 109 DF,  p-value: 0.0000004716
```

Model simplification example

What drives predatory insect diversity in grasslands?

```
lm(Predator_SpeciesRichness ~ Region + Fstd + Gstd + Mstd +
Plant_SpeciesRichness + Plant_biomass + Herbivore_SpeciesRichness +
Herbivore_biomass +
Mstd:Herbivore_SpeciesRichness + Plant_biomass:Herbivore_SpeciesRichness +
Mstd:Herbivore_biomass + Plant_biomass:Herbivore_biomass,
    data = dat)
```

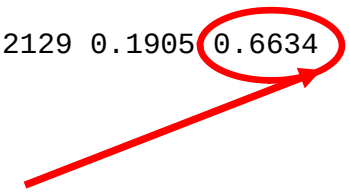
```
Coefficients:
                                          Estimate    Std. Error t value Pr(>|t|)
(Intercept)                           6.5912891471  2.3094395533   2.854  0.00517 **
RegionSCH                            -0.4223224789  0.6886784962  -0.613  0.54100
RegionALB                            -0.6975464857  0.7703844012  -0.905  0.36722
Fstd                                  0.0758311097  0.2342755414   0.324  0.74680
Gstd                                 -0.1572938757  0.2019442853  -0.779  0.43773
Mstd                                 -0.8457580810  0.9743828383  -0.868  0.38730
Plant_SpeciesRichness                -0.0608167406  0.0437888681  -1.389  0.16771
Plant_biomass                        -0.0038467374  0.0035852834  -1.073  0.28567
Herbivore_SpeciesRichness             0.0244618616  0.0782082811   0.313  0.75505
Herbivore_biomass                     0.0000708773  0.0000937884   0.756  0.45145
Mstd:Herbivore_SpeciesRichness        0.0398443434  0.0329829256   1.208  0.22965
Plant_biomass:Herbivore_SpeciesRichness  0.0000569330  0.0001304427   0.436  0.66337
Mstd:Herbivore_biomass               -0.0000720048  0.0000421465  -1.708  0.09040 .
Plant_biomass:Herbivore_biomass       0.0000001585  0.0000001210   1.310  0.19289

Residual standard error: 2.523 on 109 degrees of freedom
Multiple R-squared:  0.3801,     Adjusted R-squared:  0.3061
F-statistic: 5.141 on 13 and 109 DF,  p-value: 0.0000004716
```

Remove the interaction with highest p-value and refit the model

```
m2 <- update(m, ~.-Plant_biomass:Herbivore_SpeciesRichness
anova(m, m2)
```

```
Model 1: Predator_SpeciesRichness ~ Region + Fstd + Gstd + Mstd + Plant_SpeciesRichness +
    Plant_biomass + Herbivore_SpeciesRichness + Herbivore_biomass +
    Mstd:Herbivore_SpeciesRichness + Plant_biomass:Herbivore_SpeciesRichness +
    Mstd:Herbivore_biomass + Plant_biomass:Herbivore_biomass

Model 2: Predator_SpeciesRichness ~ Region + Fstd + Gstd + Mstd + Plant_SpeciesRichness +
    Plant_biomass + Herbivore_SpeciesRichness + Herbivore_biomass +
    Mstd:Herbivore_SpeciesRichness +
    Mstd:Herbivore_biomass + Plant_biomass:Herbivore_biomass
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	109	693.98			
2	110	695.20	-1	-1.2129	0.1905	0.6634



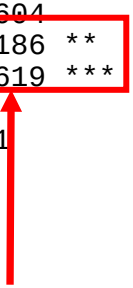
The models are not significantly different, so we prefer the simpler one

$$F = \frac{(RSSR - RSSF)/(dfR - dfF)}{RSSR/dfR}$$

$$\frac{(693.98-695.2)/(109-110)}{(693.98/109)}$$

If the interactions are out then simplify the main effects

```
Call:
lm(formula = Predator_SpeciesRichness ~ Region + Fstd + Gstd +
    Mstd + Plant_SpeciesRichness + Plant_biomass + Herbivore_SpeciesRichness +
    Herbivore_biomass, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1402 -1.4454 -0.3977  1.2254  8.3654

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                4.2055351  1.7400107   2.417 0.017237 *
RegionSchorfheide-Chorin  -0.7713757  0.6715471  -1.149 0.253103
RegionSchwäbische_Alb     -0.3545025  0.7367129  -0.481 0.631300
Fstd                      -0.1689098  0.1491063  -1.133 0.259669
Gstd                      -0.1179246  0.1958533  -0.602 0.548299
Mstd                      -0.2080174  0.3855000  -0.540 0.590522
Plant_SpeciesRichness     -0.0777282  0.0423478  -1.835 0.069042 .
Plant_biomass             -0.0001162  0.0010559  -0.110 0.912604
Herbivore_SpeciesRichness  0.0955918  0.0304921   3.135 0.002186 **
Herbivore_biomass          0.0001021  0.0000290   3.521 0.000619 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.559 on 114 degrees of freedom
Multiple R-squared:  0.3445,     Adjusted R-squared:  0.2927
F-statistic: 6.656 on 9 and 114 DF,  p-value: 1.303e-07
```



The significance of main effects can change when interactions are removed!

Keep going until only significant effects remain....

Minimal adequate model

```
Call:
lm(formula = Predator_SpeciesRichness ~ Region + Herbivore_SpeciesRichness +
    Herbivore_biomass, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5090 -1.6048 -0.3992  1.1872  7.8557

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.095e+00  8.394e-01   2.496  0.01393 *
RegionSchorfheide-Chorin  -6.096e-02  5.805e-01  -0.105  0.91654
RegionSchwäbische_Alb     -1.051e+00  5.844e-01  -1.799  0.07461 .
Herbivore_SpeciesRichness  7.754e-02  2.640e-02   2.937  0.00398 **
Herbivore_biomass          1.144e-04  2.683e-05   4.263 4.06e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.556 on 119 degrees of freedom
Multiple R-squared:  0.3171,  Adjusted R-squared:  0.2942
F-statistic: 13.82 on 4 and 119 DF,  p-value: 2.764e-09
```

I kept Region in because it is a type of “block” which you might just want to correct for

The principle of marginality

If you have an interaction you need to also keep the main effects

If interaction A:B is significant then you must keep the main effects of A + B as well

Also applies to higher order interactions: e.g. if you have A:B:C then you need the two way interaction A:B + B:C + A:C as well as main effects A + B + C

Other model simplification options

- Use `drop1` to drop each term from a full model (be careful with interactions)

- Do forward selection, i.e. keep adding terms to simple model

- `stepAIC` does a combination of forward and backward, compares with AIC (not very conservative, usually need to further simplify the models produced)

Model simplification issues

Multiple testing problem

- Carry out many statistical tests: some may be significant by chance

Problem with finding a best model

- With correlated predictors the order of deletion/addition may matter

It is often useful to try to simplify a model but be careful!

- Be aware of the problems and think about what you are doing
-

What is AIC?

In general:

$$\mathrm{AIC} = -2\ln(L) + 2k$$

Likelihood (probability of data given model)

Number of parameters

Absolute value are meaningless

Whichever model has the lower AIC is “better” (a better approximation)

What is AIC?

For linear models

$$\mathrm{AIC} = n\left[\ln\left(\frac{\mathrm{RSS}}{n}\right)\right] + 2k$$

Number of observations

Residual sum of squares

Number of parameters

Absolute value are meaningless

Whichever model has the lower AIC is “better” (a better approximation)

AICc

If sample sizes are low can correct further

Low sample size if n/k < 40

Or use AICc as default (it is the same as AIC at large sample sizes

$$\mathrm{AIC_c} = \mathrm{AIC} + \frac{2k(k+1)}{n-k-1}$$

Number of parameters

Number of observations

Comparing models

The model with the lowest (more negative) AIC is better

The degree of difference is also important

- Calculate the Δ AIC between two models (difference in AIC)
- It is sometimes considered that if Δ AIC < 2 models are equivalent
- If Δ AIC > 6 could consider model is rejected

Easy to calculate for any model

```
AICc(m)
[1] 596.3672
AICc(m2)
[1] 594.2852
(delta.aic <- AICc(m) - AICc(m2))
[1] 2.082019
```

AIC issues

Despite penalising for parameters can still lead to overly complex models…

Check if any models being tested fit the data well (using R2)

If none do then AIC selection will give meaningless results…

Make sure you have a good reason for including all predictors, don’t blindly try everything you can think of!

You may also see...

BIC: Bayesian Information Criterion

Stronger penalty for number of parameters

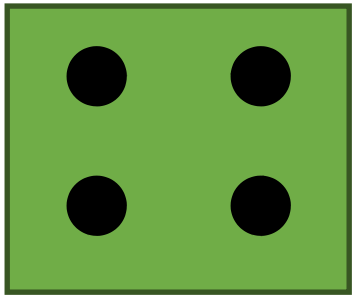
Also DIC: Deviance Information Criterion

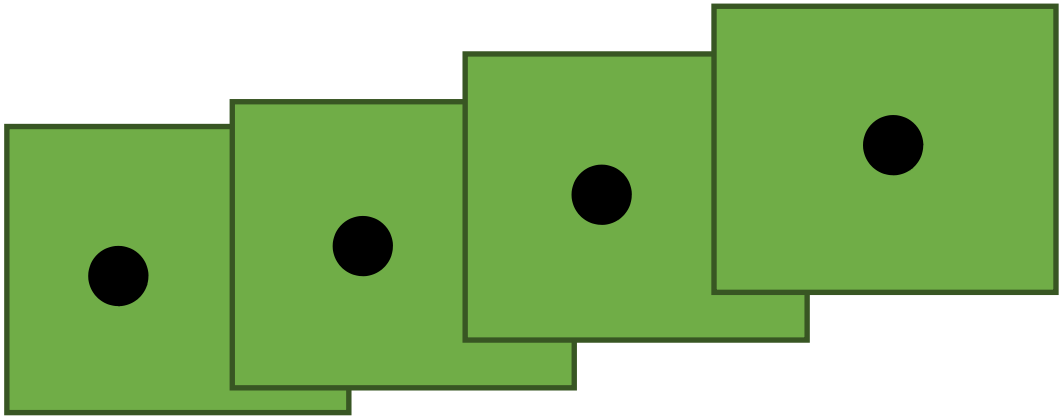
Mixed models

Non independence

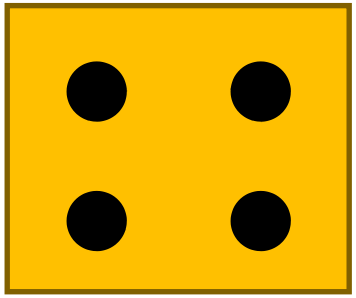
Pseudo replication remains a big issue in ecology

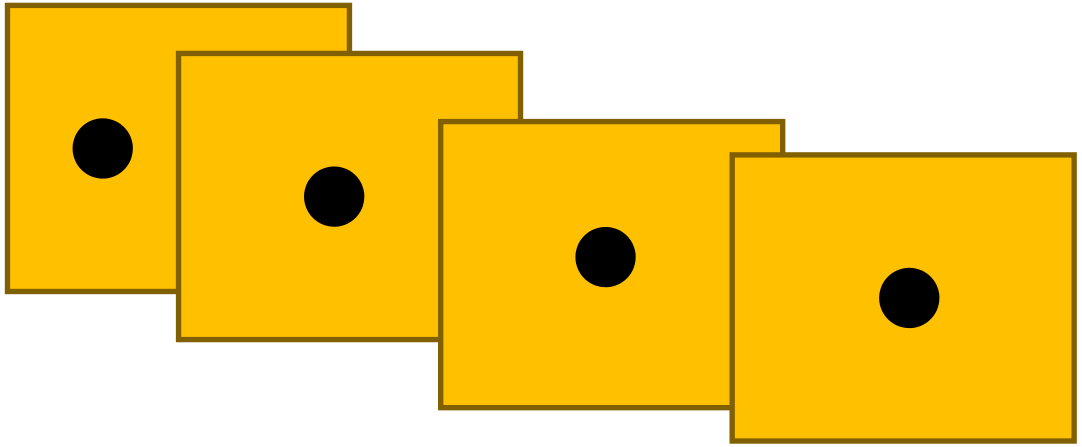
Low land use intensity





High land use intensity





Fixed and random effects

Fixed effects

- The effects you have dealt with so far
- Interested in effect size (slope or differences between levels)
- Levels often chosen deliberately or established by experiment
- Generally few levels

Random effects

- New type of effect: need mixed models for these
- Interested in accounting for variance due to the effect
- Levels often randomly selected
- Generally many levels (if <5, probably shouldn't be random)

Same variable could be fixed or random depending on inference you want to make, e.g. "species" or "region"

Why do we need the random effects?

Correct for additional sources of variance

- e.g. blocks, species
- Random effect takes fewer degrees of freedom than a fixed effect

Correct for pseudo-replication

- If designs are nested then ignoring this will inflate the sample size
- Mixed models have multiple different error levels not only one as before

$$Y_i = \alpha + \beta X_i + \varepsilon_i + \gamma_i$$

Types of random effects

Nested and crossed

- Nested as in split-plot
  - e.g. each subplot occurs in only one plot
- Crossed means factorial
  - e.g. each species occurs in every plot

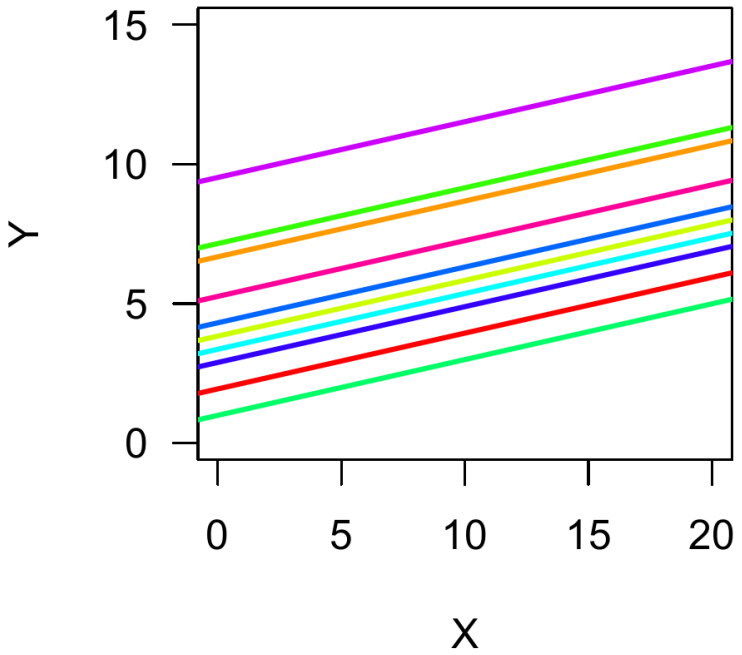
https://besjournals.onlinelibrary.wiley.com/doi/10.1111/j.2041-

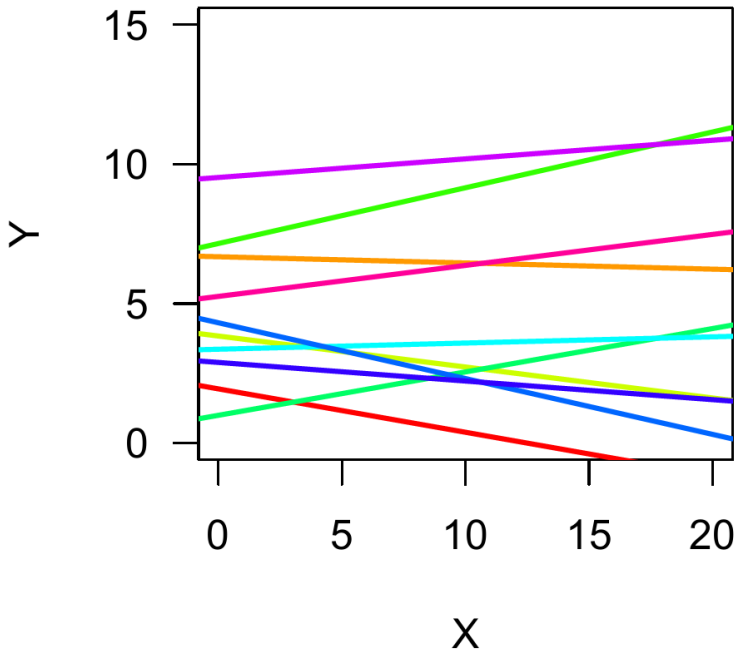
Types of random effect

Estimate variance between intercepts or slopes

→Cannot have a continuous random effect!

- Variance between intercepts or factor levels (account for variation due to plots, species, countries…)
- Variance between slopes (response to LUI, diversity, time etc. varies between species, plots, etc





Analysis of quantitative trait data

8 species collected on plots differing in LUI

- Each species on each plot is a population
- ~15 individuals per population
- Four grasses and four herbs (one of which was a legume)

Grown in a common garden to assess trait variation

- Biomass, height, flowering time

Has LUI caused changes in traits?

What happens if we don't fit random effects

wrong.mod <- lm(logbio ~ LUI * FG, data = intra2)
summary(wrong.mod)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.41651    0.16108   8.794  < 2e-16 ***
LUI         -0.53205    0.09405  -5.657 1.68e-08 ***
FGherb       1.72349    0.22070   7.809 7.73e-15 ***
LUI:FGherb  -0.40002    0.13222  -3.025   0.0025 **


Residual standard error: 2.536 on 3229 degrees of freedom
  (52 observations deleted due to missingness)
Multiple R-squared:  0.08771, Adjusted R-squared:  0.08687
F-statistic: 103.5 on 3 and 3229 DF,  p-value: < 2.2e-16
```

Everything is highly significant!

http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#model-specification

How to specify random effects

formula	meaning
(1|group)	random group intercept
(x|group) = (1+x|group)	random slope of x within group with correlated intercept
(0+x|group) = (-1+x|group)	random slope of x within group: no variation in intercept
(1|group) + (0+x|group)	uncorrelated random intercept and random slope within group
(1|site/block) = (1|site)+(1|site:block)	intercept varying among sites and among blocks within sites (nested random effects)
site+(1|site:block)	*fixed* effect of sites plus random variation in intercept among blocks within sites
(x|site/block) = (x|site)+(x|site:block) = (1 + x|site)+(1+x|site:block)	slope and intercept varying among sites and among blocks within sites
(x1|site)+(x2|block)	two different effects, varying at different levels
x*site+(x|site:block)	fixed effect variation of slope and intercept varying among sites and random variation of slope and intercept among blocks within sites
(1|group1)+(1|group2)	intercept varying among crossed random effects (e.g. site, year)

A simple mixed model

Analyse just one species and correct for the fact that multiple measures were taken per plot

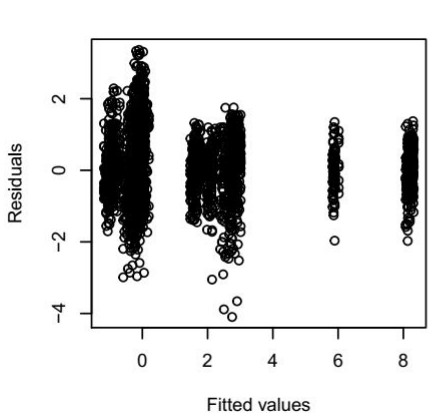
```
mixm1sp <- lmer(logbio ~ Region + LUI + (1|Plot),
subset = Species=="A_elatius", data = intra2)

summary(mixm1sp)

Linear mixed model fit by REML ['lmerMod']
Formula: logbio ~ Region + LUI + (1 | Plot)
   Data: intra2
 Subset: Species == "A_elatius"

REML criterion at convergence: 1250.1

Random effects:
 Groups   Name        Variance Std.Dev.
 Plot     (Intercept) 0.07273  0.2697
 Residual             0.97846  0.9892
Number of obs: 433, groups:  Plot, 48

Fixed effects:
            Estimate Std. Error t value
(Intercept)  3.03733    0.24014  12.648
RegionHAI    0.03316    0.17034   0.195
RegionSCH    0.28100    0.16414   1.712
LUI         -0.26956    0.12145  -2.219
```

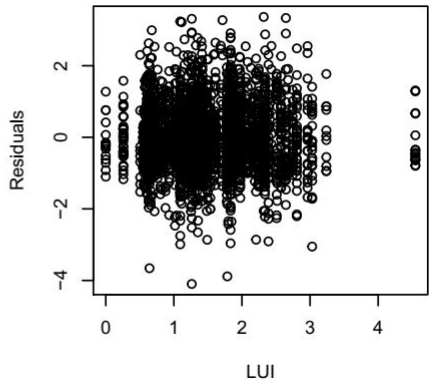
Variance between plots

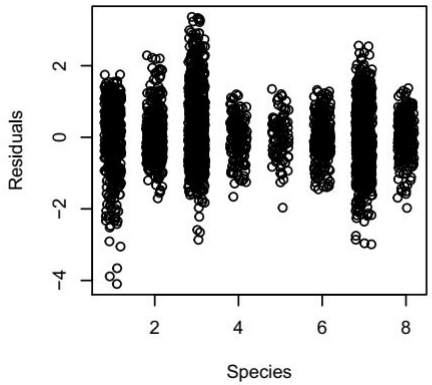
Deviance of the model, measure of fit
-2 * logLik

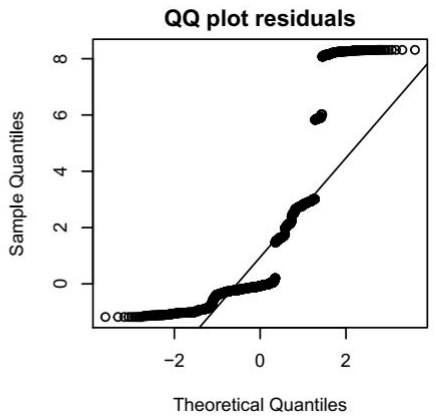
Residual variance (within plots)

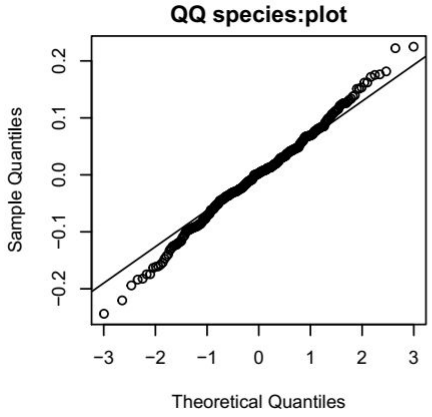
Model checking

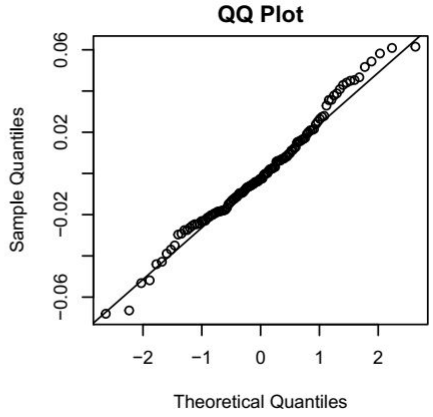


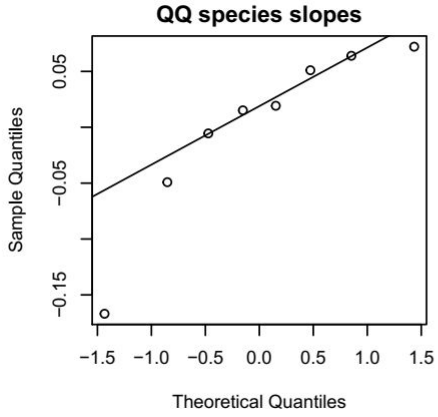


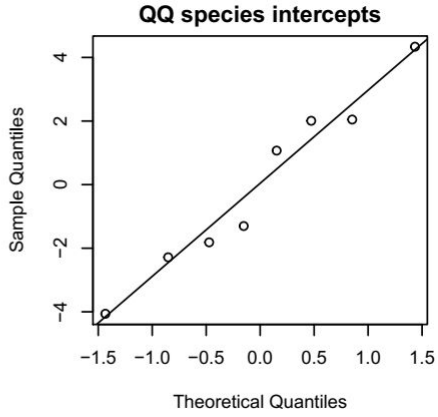












Same checks as before

Additionally need to check normality of all random effects (not just residuals)

Also good to plot residuals against x variables

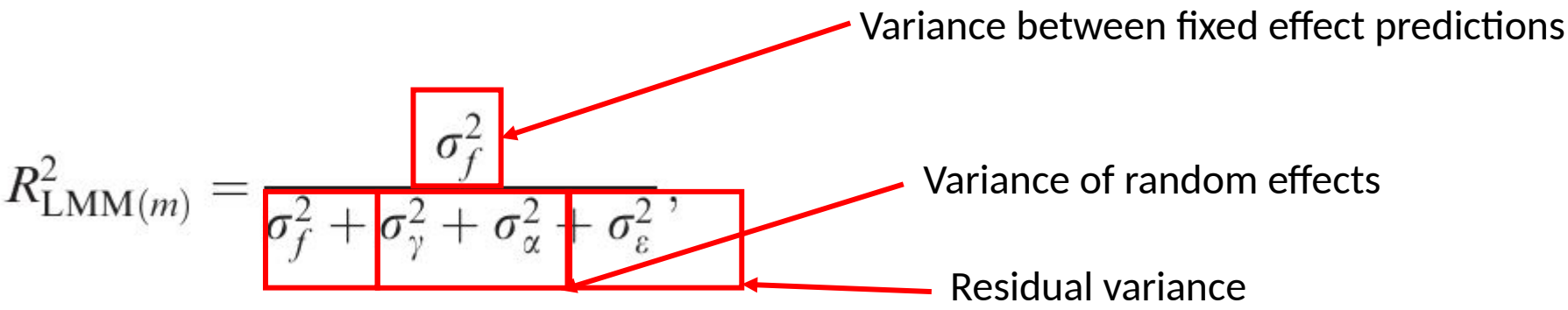
Measure of model fit

There is no way to precisely calculate a R^2 from a mixed model…

…however a "pseudo R^2" can be calculated

Either including the random and fixed effects "conditional"

Or only the fixed effects "marginal"



Nakagawa & Schielzeth 2013 MEE

Measures of model fit

MuMIn package has a function to calculate this:

```
r.squaredGLMM(mixmr1)
      R2m       R2c
0.1775251 0.9326873
```

Here the random effects explain a lot, the fixed less so

Both R2 should be reported in papers

GLMMs and other ways to fit mixed model

Can also fit a glmm easily, using `glmer` and `family = binomial` or `family = poisson`

Mixed models can also be fitted with `lme` in the `nlme` library

- Older package, no longer being developed
- Can fit unequal variances
- Can incorporate spatial/temporal autocorrelation
- Cannot deal with crossed random effects